www.sciencemag.org/cgi/content/full/334/6062/1518/DC1

Supporting Online Material for

Detecting Novel Associations in Large Data Sets

David N. Reshef,* Yakir A. Reshef,* Hilary K. Finucane, Sharon R. Grossman, Gilean McVean, Peter J. Turnbaugh, Eric S. Lander, Michael Mitzenmacher, Pardis C. Sabeti

*To whom correspondence should be addressed. E-mail: dnreshef@mit.edu (D.N.R.); yreshef@post.harvard.edu (Y.A.R.)

Published 16 December 2011, *Science* **334**, 1518 (2011)
DOI: 10.1126/science.1205438

This PDF file includes:

Contents

List of Figures

List of Tables

1 Overview

In this supplemental, we achieve the following:

- We define MIC, the characteristic matrix, and our additional statistics, and discuss statistical significance. (Section 2)
- We present our algorithm for approximating MIC. (Section 3)
- We discuss the methods used to generate our figures and analyze the datasets. (Section 4)
- We give instructions for how to use the MINE application. (Section 5)
- We formalize and prove the following statements: (Section 6)
  - The MIC of data sampled from a distribution (X, Y), where X and Y are continuous random variables, converges to 0 as sample size grows if and only if X and Y are statistically independent. (Theorem 1)
  - The MIC of a noiseless functional relationship converges to 1 as sample size grows, provided the function governing the relationship is nowhere-constant. (Theorem 3)
  - More generally, the MIC of data sampled a finite union of images of nowhere-flat, nowhere-vertical differentiable curves will approach 1 as sample size grows. (Theorem 4)
  - For any nowhere-constant function, a set of points drawn from the curve defined by the function and then vertically perturbed will receive an MIC that is lower bounded in terms of the amount of perturbation, given a large enough sample size. Moreover, this lower bound can be stated in terms of R^2. (Theorem 5)

2 Materials and methods: main definitions

2.1 Definitions of MIC and the characteristic matrix

Given a finite set D of ordered pairs, we can partition the x-values of D into x bins and the y-values of D into y bins, allowing empty bins. We call such a pair of partitions an *x-by-y grid* . Given a grid G, let $D|_G$ be the distribution induced by the points in D on the cells of G; that is, the distribution on the cells of G obtained by letting the probability mass in each cell be the fraction of points in D falling in that cell.

For a fixed D, different grids G result in different distributions $D|_G$. To exploit this fact in defining MIC, we first make the following definition.

Definition 2.1. For a finite set $D \subset \mathbb{R}^2$ and positive integers x, y, define

$$I^*(D, x, y) = \max I(D|_G)$$

where the maximum is over all grids G with x columns and y rows, and $I(D|_G)$ denotes the mutual information of $D|_G$.

We can now define the characteristic matrix and the MIC of D in terms of I^*.

Definition 2.2. The *characteristic matrix* $M(D)$ of a set D of two-variable data is an infinite matrix with entries

$$M(D)_{x,y} = \frac{I^*(D, x, y)}{\log \min\{x, y\}}.$$

Definition 2.3. The *Maximal Information Coefficient (MIC)* of a set D of two-variable data with sample size n and grid size less than $B(n)$ is given by

$$\text{MIC}(D) = \max_{xy<B(n)} \{M(D)_{x,y}\}.$$

where $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ for some $0 < \varepsilon < 1$.

Remark 2.4. Unless specified otherwise, in this paper we use $B(n) = n^{0.6}$, which we have found to work well in practice. We discuss the role of $B(n)$ further in Section 2.2.1.

Three elementary properties of MIC follow from properties of mutual information. First, since for an x-by-y grid G, $0 \leq I(D|_G) \leq \log \min\{x, y\}$, all entries of the characteristic matrix fall between 0 and 1. Second, the symmetry of mutual information ($I(X;Y) = I(Y;X)$) implies that the characteristic matrix remains the same when the x- and y-values of D are interchanged. In follows that MIC falls between 0 and 1 and is symmetric. Third, since the distribution $D|_G$ depends only on the rank-order of the data, the characteristic matrix is invariant under order-preserving transformations of the x- and y-values of D.

However, like the concept of statistical independence, MIC is not invariant under rotation of the coordinate axes. For example, the plot of a slightly noisy diagonal line exhibits statistical dependence, but if the diagonal line is rotated so that it is horizontal, the plot will exhibit statistical independence. Likewise, given sufficient sample size, the former plot will have a non-zero MIC while the latter plot will have an MIC very close to 0.

The space of grids that must be searched to compute each entry of the characteristic matrix grows exponentially with the number of data points, so for efficiency we use a heuristic dynamic programming algorithm to approximate MIC in practice; this algorithm is presented and discussed in Section 3.

2.2 Details of the definition of MIC

We now discuss in more detail two steps of the calculation of MIC: the parameter $B(n)$, which controls how many much of the characteristic matrix we search over, and the normalization in the definition of I^*.

2.2.1 The maximal grid size $B(n)$

The function $B(n)$ upper bounds the sizes of the grids we search over. Determining the appropriate choice of $B(n)$ is important: setting $B(n)$ too high can lead to non-zero scores even for random data because each data point gets its own cell, while setting $B(n)$ too low means we are searching only for simple patterns. We balance these competing considerations with a pair of proofs demonstrating that inflated scores are avoided precisely when $B(n)$ grows more slowly than n (see Theorems 1 and 2 in Section 6.2). In addition to this theoretical work, we also conducted empirical tests to establish that $B(n) = \Theta(n)$ is an inflection point above which statistically independent data receive scores bounded away from zero as sample size grows, and below which they receive scores approaching zero; Figure S1 shows the results of these tests.

Our default setting for $B(n)$ is $n^{0.6}$, and all analyses carried out in this paper use this setting unless noted otherwise.

2.2.2 Normalization

The optimal grids found for different sets of data need not have the same dimensions. For example, a line can be captured perfectly by a two-by-two grid, but no two-by-two grid can perfectly capture a parabola. This is problematic because grids with different dimensions have different maximal mutual information scores: the maximal possible mutual information of a distribution on an x-by-y grid is $\log \min\{x, y\}$ [38]. Thus, even once an optimal grid is found for the line and for the parabola, they will yield different mutual information scores even though both are noiseless functions. Normalizing by $\log \min\{x, y\}$ creates a score that can be compared across grids with different dimensions and therefore across different distributions. It also guarantees that almost all noiseless functions receive perfect scores (Theorem 3) and that the entries of the characteristic matrix range from zero to one.

2.3 Definitions of additional statistics

Recall that the characteristic matrix M of a set of data D is defined by

$$M(D)_{x,y} = \frac{I^*(D, x, y)}{\log \min\{x, y\}},$$

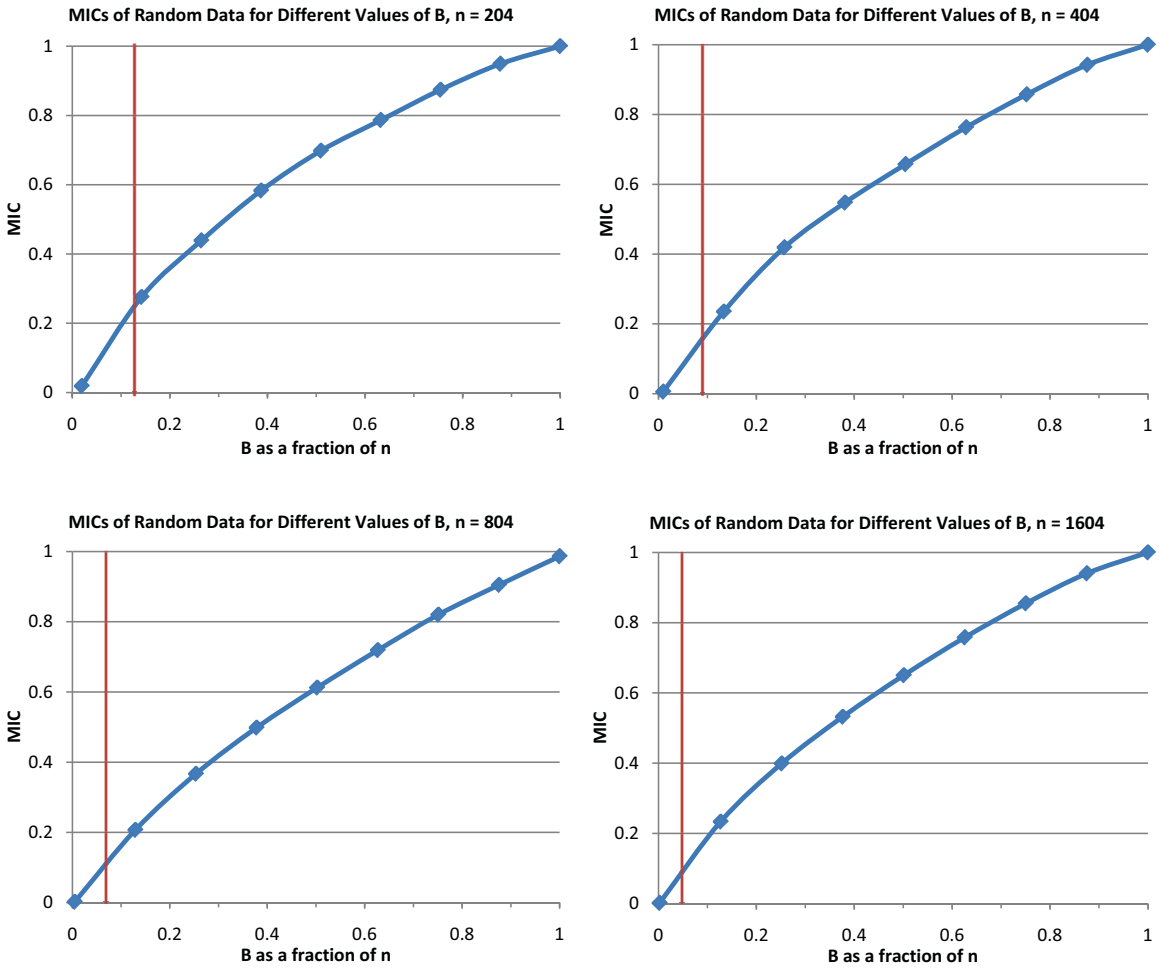


Figure S1: Empirical evidence demonstrating that $B(n) = O(n^{1-\varepsilon})$ results in MIC scores approaching zero for statistically independent data. Plots of MIC scores of single random clouds with different sample sizes n using different numbers $B(n)$ of bins. The scores depend only on the fraction $B(n)/n$. The red line represents the value of $B(n)$ corresponding in each case to $B(n) = n^{0.6}$. The MIC score moves toward 0 as n grows.

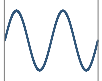
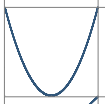
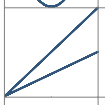
Data	MIC	MAS	MEV	MCN
	1.00	0.00	1.00	2.00
	1.00	0.74	1.00	3.00
	1.00	0.89	1.00	4.00
	1.00	0.69	1.00	2.56
	0.79	0.16	0.70	6.91
	0.71	0.03	0.32	6.87
	0.46	0.19	0.22	6.98

Table S1: MINE statistics calculated for some sample associations. Intuitively, MIC captures relationship strength; MAS captures departure from monotonicity; MEV captures closeness to being a function; and MCN captures complexity. Note that the circle and the pair of lines do not receive perfect MIC scores. This is because the scores of these associations approach 1 as sample size tends to infinity, while the table was generated with only $n = 1000$. The circle with a line through the middle gets an even lower score because the line is completely flat and so MIC considers it "noise" rather than "signal". (The flat line by itself would have an MIC of 0.) This is consistent with the requirement of nowhere flatness in Theorem 4. The characteristic matrices for these relationships are shown in Figure S7.

where $I^*(D, x, y)$ is the maximum mutual information achieved by any grid with x columns and y rows on the data D. MIC is the maximum value of this matrix, but the matrix contains more information than just its maximum value. We have developed a few other statistics which can be derived from the characteristic matrix; further exploration of such properties of the characteristic matrix appears to be a promising area for future research. Like MIC, these characteristics are unchanged when the x- and y-values of the data D are swapped and when order-preserving transformations are applied to the x- or y-values.

Table S1 shows a few different patterns and their corresponding scores using these statistics.

- **Non-monotonicity**

  The Maximum Asymmetry Score (MAS) is defined by

  $$\text{MAS}(D) = \max_{xy<B} |M(D)_{x,y} - M(D)_{y,x}|$$

  and measures deviation from monotonicity. MAS is never greater than MIC. For an illustration of the intuition behind MAS, see Figure S2.

- **Closeness to being a function**

  The Maximum Edge Value (MEV) is defined by

  $$\text{MEV}(D) = \max_{xy<B} \{M(D)_{x,y} : x = 2 \text{ or } y = 2\}$$

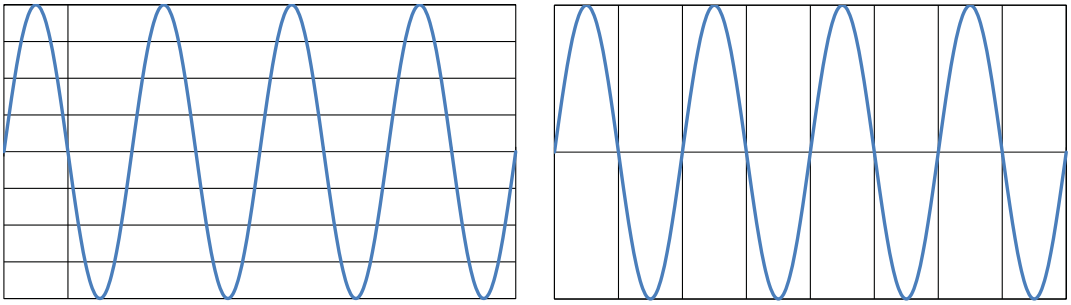


Figure S2: A demonstration of the intuition behind MAS. Left: A 4-period sinusoid shown with its optimal 8×2 grid (i.e. rows outnumber columns). This grid is only able to give a normalized score of 0.14. Right: The same 4-period sinusoid shown with its optimal 2×8 grid (i.e. columns outnumber rows). Because this grid has the property that each column contains exactly one non-empty cell, it gives a normalized score of 1. The reported MAS for this dataset is the difference between these scores: 0.86.

and measures the degree to which the dataset appears to be sampled from a continuous function. Like MIC and MAS, it ranges from 0 to 1 with a score of 1 suggesting a well-behaved function. Like in the case of MAS, we have MEV $\leq$ MIC always.

The intuition behind MEV is that if a set of data points "passes the vertical line test", in that for the imagined underlying distribution each vertical line can contain only one point, it should have a close-to-optimal grid with only 2 rows (see the proof of Proposition 6.14). Similarly we could consider a horizontal line test. On the other hand, when the underlying distribution for the data points does not pass either the vertical or hoirzontal line test, we have observed that the sampled datasets receive low scores when the grids are restricted to having only 2 rows or only 2 columns, as one would expect.

- **Complexity**

The Minimum Cell Number (MCN) is defined by

$$\text{MCN}(D, \epsilon) = \min_{xy<B} \{\log(xy) : M(D)_{x,y} \geq (1-\epsilon)\text{MIC}(D)\}$$

This statistic measures the complexity of the association, in terms of the number of cells required to reach the MIC score. For example, a simple function like $f(x) = x$ requires very few cells (four, in fact) to grid in an effective way while a complex function such as $f(x) = \sin(18\pi x)$ requires many (thirty-six) cells. The ϵ parameter provides robustness and should depend on the MIC of the relationship in question. In Table S1, ϵ is set to 0 because the functions considered are noiseless. For the more general case, something like $\epsilon = 1 - \text{MIC}(D)$ may be a more appropriate parametrization.

2.4 Determining statisticial significance

Our null hypothesis is that the variables X and Y are statistically independent. We compute the p-value of a given MIC score by selecting a probability α of false rejection, creating a set of $1/\alpha - 1$ surrogate datasets, and comparing the MIC of the real data with the MIC scores of the surrogate datasets [39].

Since MIC depends only on the rank-order of the data, we can create a surrogate instance of the null hypothesis for a given sample size by choosing a random permutation of Y with respect to X [40]. Since this does not depend on the specific relationship being tested, we have created tables of the p-values of various MIC scores at different sample sizes for public use. These are available at the MINE website: `exploredata.net`.

In datasets where the number of tested variable pairs is large, it may be necessary to address the multiple testing problem; in these cases, methods controlling the false discovery rate (FDR) are appropriate [41].

Controlling FDR is better suited to our goal of identifying and scoring many significant relationships while incurring a relatively low proportion of false positives than a more traditional Bonferroni correction, which controls familywise error rate [41, 42]. Adopting this approach also has the advantage that calculating FDRs at a reasonable resolution on very large datasets is computational feasible.

Unless noted otherwise, we controlled the FDR for all analyses using the Benjamini and Hochberg procedure [41]. For additional clarity, we present q-values for each relationship from the main text. In our context, the q-value of a relationship is the minimum FDR at which that relationship may be called significant [43]. These methods assume that all the hypotheses being tested are independent, an assumption which may not hold, for instance, when we are testing many relationships from the same dataset. However, it turns out that the Benjamini and Hochberg procedure works unchanged when the dataset satisfies a certain condition about positive dependencies between the variables, and that when this condition is not satisfied, a simple correction factor can be used [44].

3 Materials and methods: the approximation algorithm for generating the characteristic matrix

In this section, we describe our algorithm for heuristically generating the characteristic matrix of a set of two-variable data.

Remark 3.1. As noted in the main text, MIC is a rank-order statistic, meaning that if data are perturbed in a way that does not change the relative ranks of the x- and y-values, the MIC of the data will not change. Our approximation algorithm preserves this property since it effectively takes as input only the relative ranks of the x- and y-values of the data.

3.1 Overview

3.1.1 The idealized algorithm

We begin by outlining an idealized algorithm for generating the characteristic matrix. Algorithm 1 represents what we would use if efficiency were not a problem. Because we implement a heuristic approximation of the MaxMI subroutine (called ApproxMaxMI), our implementation of Algorithm 1 is actually a heuristic approximation of the characteristic matrix.

Algorithm 1 CharacteristicMatrix(D, B)

Require: D is a set of ordered pairs
Require: B is an integer greater than 3
1: **for** (x, y) such that $xy \leq B$ **do**
2: $\quad I_{x,y} \leftarrow \text{MaxMI}(D, x, y)$
3: $\quad M_{x,y} \leftarrow I_{x,y} / \min\{\log x, \log y\}$
4: **end for**
5: **return** $\{M_{x,y} : xy \leq B\}$

Line 3 is the normalization step discussed in Section 2.2.2. Using the scores returned by Algorithm 1, MIC, MAS, and the other statistics are straightforward to calculate from their mathematical definitions.

3.2 The MaxMI subroutine

The MaxMI function invoked in Algorithm 1 is meant to return the highest mutual information attainable using a grid with x columns and y rows on the data D. This is the portion of the procedure that we chose to implement as a heuristic approximation algorithm using dynamic programming. Our implementation, which we call ApproxMaxMI for clarity, can and should be replaced in the future if a method that efficiently finds solutions that are closer to optimal or even optimal is developed.

3.2.1 The core of ApproxMaxMI: finding an optimal x-axis partition

The portion of ApproxMaxMI that uses dynamic programming is a function called OptimizeXAxis. Given a fixed y-axis partition, OptimizeXAxis finds the x-axis partition that will lead to a grid with maximal mutual information. Before we describe this function, we first put forth some notational conventions. We will then prove the recursion behind the dynamic programming and present pseudocode for the function.

Preliminaries

We will assume that our set of ordered pairs D is sorted in increasing order by x-value. We denote various partitions of the x-axis by specifying the indices of the endpoints of their columns. Specifically, we will call an ordered list of integers $\langle c_0, \ldots, c_t \rangle$ with $c_0 < c_1 < \cdots < c_t$ an *x-axis partition* of size t of the the

(c_0+1)-st through c_t-th points of D. Given a partition $C = \langle c_0, \ldots, c_t \rangle$ and an integer a with $c_i < a < c_{i+1}$, we let $C \cup a$ denote $\langle c_0, \ldots, c_i, a, c_{i+1}, \ldots, c_t \rangle$. If $a = c_i$ for some i, we define $C \cup a := C$.

Since the dataset D is fixed for our purposes, we will abuse notation in the following way: given an x-axis partition P of m points of D, we let $H(P)$ be the entropy of the distribution induced by those m points on the columns of P. Similarly, $H(P, Q)$ will denote the entropy of the distribution induced by those m points in D on the cells of the grid formed by P and Q where Q is our (fixed) y-axis partition. We will use $I(P; Q)$ analogously. Since Q is fixed just like D, $H(Q)$ will always denote the entropy of the distribution induced by *all* the points of D on the rows of Q.

Description of the OptimizeXAxis function

We begin by proving the recursion that underpins our dynamic programming algorithm. For any discrete random vector (X, Y), a standard definition of $I(X; Y)$ is $I(X; Y) = H(X)+H(Y)-H(X, Y)$ where $H(\cdot)$ denotes Shannon entropy. This means that, using our notation, we have $I(P; Q) = H(P)+H(Q)-H(P, Q)$ for all partitions P and Q. However, since Q is fixed, the OptimizeXAxis function need only maximize $H(P) - H(P, Q)$ over all partitions P in order to maximize $I(P; Q)$. Proposition 3.2 below establishes how we will do so.

Proposition 3.2. *Fix a y-axis partition Q and a dataset D of size n. For every $m, \ell \in [n]$, define $F(m, \ell) = \max\{H(P) - H(P, Q)\}$ where the maximum is over all partitions of size up to ℓ of the first m points of D. We have the following recurrence for $\ell > 1$ and $1 < m \leq n$.*

$$F(m, \ell) = \max_{1 \leq i < m} \left\{ \frac{i}{m} F(i, \ell - 1) + \frac{m - i}{m} H(\langle i, m \rangle, Q) \right\}$$

Proof. Let $P = \langle 0 = c_0, \ldots, c_\ell = m \rangle$ be an x-axis partition maximizing $H(P) - H(P, Q)$ as in the definition of F. (We have assumed without loss of generality that P is of size exactly ℓ.) Let $\#_{*,j}$ denote the number of points in the j-th column of P and note that $\#_{*,j} = c_j - c_{j-1}$. Define $\#_{i,j}$ to be the number of points in the i-th row of Q and the j-th column of P. Using this notation, we have

$$\begin{aligned}
F(m, \ell) &= \sum_{j=1}^{\ell} \frac{\#_{*,j}}{m} \log \frac{m}{\#_{*,j}} - \sum_{j=1}^{\ell} \sum_{i=1}^{|Q|} \frac{\#_{i,j}}{m} \log \frac{m}{\#_{i,j}} \\
&= \sum_{j=1}^{\ell} \sum_{i=1}^{|Q|} \frac{\#_{i,j}}{m} \log \frac{\#_{i,j}}{\#_{*,j}} \\
&= \sum_{j=1}^{\ell-1} \sum_{i=1}^{|Q|} \frac{\#_{i,j}}{m} \log \frac{\#_{i,j}}{\#_{*,j}} + \sum_{i=1}^{|Q|} \frac{\#_{i,\ell}}{m} \log \frac{\#_{i,\ell}}{\#_{*,\ell}} \\
&= \frac{c_{\ell-1}}{m} \sum_{j=1}^{\ell-1} \sum_{i=1}^{|Q|} \frac{\#_{i,j}}{c_{\ell-1}} \log \frac{\#_{i,j}}{\#_{*,j}} + \frac{\#_{*,\ell}}{m} \sum_{i=1}^{|Q|} \frac{\#_{i,\ell}}{\#_{*,\ell}} \log \frac{\#_{i,\ell}}{\#_{*,\ell}} \\
&= \frac{c_{\ell-1}}{m} \left(H(P') - H(P', Q) \right) + \frac{\#_{*,\ell}}{m} H(\langle c_{\ell-1}, m \rangle, Q) \\
&= \frac{c_{\ell-1}}{m} F(c_{\ell-1}, \ell - 1) + \frac{m - c_{\ell-1}}{m} H(\langle c_{\ell-1}, m \rangle, Q)
\end{aligned}$$

where $P' = \langle c_0, \ldots, c_{\ell-1} \rangle$ and the last line is because if $H(P') - H(P', Q)$ did not equal $F(c_{\ell-1}, \ell - 1)$ then $F(m, \ell)$ could be increased by choosing different values for $c_1, \ldots, c_{\ell-1}$.

This establishes that $F(m, \ell)$ is in the set $\{\frac{i}{m} F(i, \ell - 1) + \frac{m-i}{m} H(\langle i, m \rangle, Q)\}$, and since $F(m, \ell)$ must be a maximal element of this set, we have the desired result. □

Theorem 3.2 gives rise to a natural dynamic programming algorithm that builds up a table of values of F and eventually returns $F(n, \ell)$ where ℓ is the desired partition size. The function that carries out this recursion, called OptimizeXAxis, is described in Algorithm 2.

Our version of the OptimizeXAxis function restricts itself to drawing x-axis partition lines only between runs of consecutive points that fall in the same row of the y-axis partition Q (called *clumps*). This increases effiency without sacrificing optimality because no matter how a clump is split into columns, the contributions of those columns to $H(P) - H(P, Q)$ will be 0. In Algorithm 2, the GetClumpsPartition subroutine is responsible for finding the edges of clumps: it returns the minimal partition that separates every pair of points that lie in distinct clumps.

Algorithm 2 OptimizeXAxis(D, Q, x)

Require: D is a set of ordered pairs sorted in increasing order by x-value
Require: Q is a y-axis partition of D
Require: x is an integer greater than 1
Ensure: Returns a list of scores $(I_2, \ldots, I_x)$ such that each I_ℓ is the maximum value of $I(P; Q)$ over all partitions P of size ℓ.

$\langle c_0, \ldots, c_k \rangle \leftarrow$ GetClumpsPartition(D, Q)

{Find the optimal partitions of size 2}
for $t = 2$ to k **do**
  Find $s \in \{1, \ldots, t\}$ maximizing $H(\langle c_s, c_t \rangle) - H(\langle c_s, c_t \rangle, Q)$.
  $P_{t,2} \leftarrow \langle c_s, c_t \rangle$
  $I_{t,2} \leftarrow H(Q) + H(P_{t,2}) - H(P_{t,2}, Q)$
end for

{Inductively build the rest of the table of optimal partitions}
for $\ell = 3$ to x **do**
  for $t = 2$ to k **do**
    Find $s \in \{1, \ldots, t\}$ maximizing

$$F(s, t, \ell) := \frac{c_s}{c_t} \left(I_{s,\ell-1} - H(Q) \right) + \sum_{i=1}^{|Q|} \frac{\#_{i,\ell}}{c_t} \log \frac{\#_{i,\ell}}{\#_{*,\ell}}$$

    where $\#_{*,j}$ is the number of points in the j-th column of $P_{s,\ell-1} \cup c_t$ and $\#_{i,j}$ is the number of points in the j-th column of $P_{s,\ell-1} \cup c_t$ that fall in the i-th row of Q
    $P_{t,\ell} \leftarrow P_{s,\ell-1} \cup c_t$
    $I_{t,\ell} \leftarrow H(Q) + H(P_{t,\ell}) - H(P_{t,\ell}, Q)$
  end for
end for
return $(I_{k,2}, \ldots, I_{k,x})$

In Algorithm 2, each $P_{t,\ell}$ is an optimal partition of size ℓ of the first t clumps of D, and $I_{t,\ell}$ is defined such that it will contain the corresponding mutual information when t equals the total number of clumps k. The reason OptimizeXAxis returns an array of scores instead of just one is that, since we seek the maximal scores for every possible number of columns, the entire array is useful. Note that the partitions $P_{t,\ell}$ appear in our pseudocode for clarity alone; only the scores $I_{t,\ell}$ are actually used by the function.

3.2.2 Departures from optimality in the ApproxMaxMI algorithm

We now describe the two key features that both enable ApproxMaxMI to run in a reasonable amount of time and make it a heuristic approximation.

Equipartitioning one axis

If, given some number x of columns and some number y of rows, we could run the OptimizeXAxis function on every possible y-axis partition of size y, we would find an optimal grid. But the number of possible y-axis partitions makes this infeasible. Therefore, since the mutual information is bounded from above by the entropy of the less informative axis and the marginal entropies of the axes are maximized by equipartitions[1], a natural heuristic approach to this problem is to consider only grids for which at least one axis is equipartitioned. To this end, the ApproxMaxMI algorithm fixes an equipartition of size y on the y-axis and then runs OptimizeXAxis. Later, ApproxMaxMI is called again but with the axes switched. The maximum of the two scores obtained is used. Creating the equipartitions involves some tie-breaking when points have identical y-values. This is carried out by the EquipartitionYAxis function, described in Algorithm 3.

Algorithm 3 EquipartitionYAxis(D, y)

Require: D is a set of n ordered pairs
Require: y is an integer greater than 1
Ensure: Returns a map $Q\colon D \to \{1, \ldots, y\}$ such that $Q(p)$ is the row assignment of the point p and there is approximately the same number of points in each row

$D \leftarrow$ SortInIncreasingOrderByYValue(D)
$i \leftarrow 1$
currRow $\leftarrow 1$
desiredRowSize $\leftarrow n/y$
repeat
  $S \leftarrow \{(a_j, b_j) \in D : b_j = b_i\}$
  $\# = |\{(a_j, b_j) \in D : Q(a_j, b_j) = \text{currRow}\}|$
  if $\# = 0$ or $|\# + S - \text{desiredRowSize}| \leq |\# - \text{desiredRowSize}|$ **then**
    $Q((a_j, b_j)) \leftarrow$ currRow for every $(a_j, b_j) \in S$
    $i \leftarrow i + |S|$
    desiredRowSize $\leftarrow (n - i)/y$
  else
    currRow $\leftarrow$ currRow $+ 1$
  end if
until $i > n$
return Q

Restricting the number of clumps

If k is the number of clumps created by a given y-axis partition Q of size y imposed on the dataset D, the runtime of OptimizeXAxis(D, Q, x) is $O(k^2xy)$. This is often too much for large datasets, and so there is one additional parameter to which ApproxMaxMI responds: a maximum number of clumps $\hat{k}$ to allow in its analysis. When $k > \hat{k}$, the true clumps are merged into *superclumps* in a way that aims to have each superclump contain approximately the same number of points. The algorithm then forgets about the clumps and only considers drawing grid-lines between the superclumps. The parameter $\hat{k}$ allows for a standard efficiency vs. optimality tradeoff. It can be set high enough that $k < \hat{k}$ always holds, but the algorithm seems effective even when this condition is not met.

The subroutine that builds the superclumps, called GetSuperclumpsPartition, takes as input only a partition describing the boundaries of the clumps and the parameter $\hat{k}$. It uses the same logic as EquipartitionYAxis (Algorithm 3), only it considers points in the same clump to be a unit rather than points with the same y-value.

[1] An *equipartition* is a partition into either rows or columns such that each row/column contains the same number of points

3.2.3 The complete ApproxMaxMI algorithm

We now give the pseudocode of the ApproxMaxMI algorithm. This is followed by the pseudocode for the ApproxCharacteristicMatrix algorithm, which is slightly different from the outline given in Algorithm 1. This is because, for a fixed number y of rows, OptimizeXAxis returns optimal partitions of size x for every x. Additionally, the ApproxMaxMI algorithm calls the "ApproxOptimizeXAxis" function. This function is identical in every way to OptimizeXAxis (described in Algorithm 2) except that it takes as an argument the maximum number $\hat{k}$ of superclumps and calls GetSuperclumpsPartition instead of GetClumpsPartition.

Algorithm 4 ApproxMaxMI$(D, x, y, \hat{k})$

Require: D is a set of ordered pairs
Require: x, y, and $\hat{k}$ are integers greater than 1
Ensure: Returns a set of mutual information scores $(I_{2,y}, \ldots, I_{x,y})$ such that $I_{i,j}$ is heuristically close to the highest achievable mutual information score using i rows and j columns.

$Q \leftarrow$ EquipartitionYAxis(D, y)
$D \leftarrow$ SortInIncreasingOrderByYValue(D)
return ApproxOptimizeXAxis$(D, Q, x, \hat{k})$

Algorithm 5 ApproxCharacteristicMatrix(D, B, c)

Require: $D = \{(a_1, b_1), \ldots, (a_n, b_n)\}$ is a set of ordered pairs
Require: B is an integer greater than 3
Require: c is greater than 0

$D^{\perp} \leftarrow \{(b_1, a_1), \ldots, (b_n, a_n)\}$
for all $y \in \{2, \ldots, \lfloor B/2 \rfloor\}$ **do**
    $x \leftarrow \lfloor B/y \rfloor$
    $(I_{2,y}, \ldots, I_{x,y}) \leftarrow$ ApproxMaxMI(D, x, y, cx)
    $({I^{\perp}}_{2,y}, \ldots, {I^{\perp}}_{x,y}) \leftarrow$ ApproxMaxMI$(D^{\perp}, x, y, cx)$
end for
for (x, y) such that $xy \leq B$ **do**
    $I_{x,y} \leftarrow \max\{I_{x,y}, {I^{\perp}}_{y,x}\}$
    $M_{x,y} \leftarrow I_{x,y} / \min\{\log x, \log y\}$
end for
return $\{M_{x,y} : xy \leq B\}$

4 Materials and methods: creation of figures

In this section, we describe how we generated each of the figures in the main text, and how we analyzed the data sets.

4.1 Command-line arguments used for analyses

All analyses carried out throughout the paper used MINE's default arguments except for in the cases noted below. (For descriptions of these arguments, see Section 5.)

- Figures 2B and S4 (MIC on noisy functional relationships): c = 75
- Figure 4A-G (WHO dataset): exp = 0.65, cv = 0.25
- Figure 4I (WHO dataset): exp = 0.70
- Figure S10 (WHO spring graph): exp = 0.65, cv = 0.5
- Figure 5 (Gene expression): exp = 0.67, cv = 1.0
- Figure 6 (Microbiome dataset): exp = 0.551, c=10
- Major League Baseball statistics: exp = 0.7, cv = 0.25
- Comparison of MAS, Spearman test, Fisher test, and Ahdesmaki et al. statistic: c = 25

4.2 Analysis for Figure 2A (performance on noiseless functional relationships)

For each function f listed in Table S2, we generated a data series D^f containing 320 points evenly spaced along the curve described by f. We then ran each of the statistics in the table on each D^f, and reported the scores. (For the CorGC method of Delicado [19], we multiplied all x-values by 100 before running the method, as in Figure 2F, because we found this to lead to more accurate estimates of the principal curves of the data.)

For the random cloud, we calculated the scores of 1000 independent realizations of 320 points chosen uniformly at random from the unit box, and reported the average.

Relationship Name	**Description (The domain is $[0,1]$ for all functions.)**
Linear	$y = x$
Parabolic	$y = 4(x - \frac{1}{2})^2$
Cubic	$y = 128(x - \frac{1}{3})^3 - 48(x - \frac{1}{3})^2 - 12(x - \frac{1}{3}) + 2$
Exponential	$y = 10^{10x} - 1$
Linear/Periodic	$y = \sin(10\pi x) + x$
Sinusoidal (Fourier Frequency)	$y = \sin(16\pi x)$
Sinusoidal (non-Fourier Frequency)	$y = \sin(13\pi x)$
Sinusoidal (Varying Frequency)	$y = \sin(7\pi x(1 + x))$
Categorical	64 points chosen from the following set: $\{(1, 0.287), (2, 0.796), (3, 0.290), (4, 0.924), (5, 0.717)\}$
Random	random number generator

Table S2: Definitions of the functions analyzed in Figure 2A.

4.3 Analysis for Figures 2B-F, and Figures S3 and S4 (performance on noisy functional relationships)

Each relationship S listed in Table S4 consists of a function f_S described in Table S3 and a sample size n_S given in parentheses. For each relationship S, we generated a data series D_0^S of n_S points spaced evenly along the curve described by f_S. We then created 249 additional data series $\{D_i^S : 0 < i < 250\}$ by adding incrementally larger amounts of uniform vertical noise to D_0^S.

For each relationship S, we calculated, for all i,

1. The R^2 between D_i^S and D_0^S
2. The MIC of D_i^S
3. The Spearman correlation coefficient of D_i^S
4. The mutual information of D_i^S
5. The mean-squared error (MSE) between D_i^S and the estimated principal curve[2] of D_i^S
6. CorGC(D_i^S), where CorGC is the principal curve-based measure of non-linear dependence due to Delicado and Smrekar [19]. Prior to running the method, x-values were all multiplied by 100 because this was found to lead to more accurate estimates of the principal curves of the data.
7. The maximal correlation, as calculated by the method of alternating conditional expectations [15], of D_i^S
8. The distance correlation [21] of D_i^S

[Principal curves for (5) and (6) were estimated using the Hastie-Stuetzle algorithm (princurve R package) with default parameters [45]. Alternative algorithms from [46] and [18] yielded similar results.]

Figure 2B is a plot of (2) against (1) for all relationships S;
Figure 2C is a plot of (3) against (1) for all relationships S;
Figure 2D is a plot of (4) against (1) for all relationships S;
Figure 2E is a plot of (7) against (1) for all relationships S;
Figure 2F is a plot of (6) against (1) for all relationships S;
Figure S3A is a plot of (5) against (1) for all relationships S;
Figure S3B is a plot of (8) against (1) for all relationships S.

Figure S4 contains versions of Figures 2B and 2D generated with other sample sizes and methods of adding noise (choosing points to be evenly spaced along the x-axis, adding horizontal as well as vertical noise, etc.).

4.4 Analysis for Figure 2G (MIC on selected non-functional associations)

For each relationship type S in Figure 2G, a "noiseless" instance D_0^S was generated with a sample size of $10,000$. (For example, for the relationship that looks like an X, half of the points were chosen from the set $\{(x, x)\}$ and the other half were chosen from the set $\{x, 1 - x\}$ with x values evenly spaced in the interval $[0, 1]$ in both cases.) For each S, 199 additional data series $\{D_i^S : 0 < i < 200\}$ were created by adding incrementally larger amounts of independent uniform horizontal and vertical noise to D_0^S.

MIC was run on all the sets D_i^S. For each S, four elements were chosen from the set $\{D_i^S : 0 \leq i < 200\}$: one with an MIC of 0.8, one with an MIC of 0.65, one with an MIC of 0.5, and one with an MIC of 0.35. These datasets are the ones displayed in the table.

[2] It is not clear how to calculate R^2 relative to a principal curve since the curve is not always a function and may not be defined at all the x-values in the data; therefore MSE was used in lieu of an alternative obvious choice. Data were normalized such that projections onto either axis had unit variance to aid comparison of MSEs across different distributions.

Function Name	**Definition**	
Linear+Periodic, Low Freq	$y = \frac{1}{5}\sin(4(2x-1)) + \frac{11}{10}(2x-1)$	$x \in [0,1]$
Linear+Periodic, Medium Freq	$y = \sin(10\pi x) + x$	$x \in [0,1]$
Linear+Periodic, High Freq	$y = \frac{1}{10}\sin(10.6(2x-1)) + \frac{11}{10}(2x-1)$	$x \in [0,1]$
Linear+Periodic, High Freq 2	$y = \frac{1}{5}\sin(10.6(2x-1)) + \frac{11}{10}(2x-1)$	$x \in [0,1]$
Non-Fourier Freq [Low] Cosine	$y = \cos(7\pi x)$	$x \in [0,1]$
Cosine, High Freq	$y = \cos(14\pi x)$	$x \in [0,1]$
Cubic	$y = 4x^3 + x^2 - 4x$	$x \in [-1.3, 1.1]$
Cubic, Y-stretched	$y = 41(4x^3 + x^2 - 4x)$	$x \in [-1.3, 1.1]$
L-shaped	$y = \begin{cases} x/99 & \text{if } x \leq \frac{99}{100} \\ 1 & \text{if } x > \frac{99}{100} \end{cases}$	$x \in [0,1]$
Exponential [2^x]	$y = 2^x$	$x \in [0,10]$
Exponential [10^x]	$y = 10^x$	$x \in [0,10]$
Line	$y = x$	$x \in [0,1]$
Parabola	$y = 4x^2$	$x \in [-\frac{1}{2}, \frac{1}{2}]$
Random	random number generator	$x \in [0,1]$
Non-Fourier Freq [Low] Sine	$y = \sin(9\pi x)$	$x \in [0,1]$
Sine, Low Freq	$y = \sin(8\pi x)$	$x \in [0,1]$
Sine, High Freq	$y = \sin(16\pi x)$	$x \in [0,1]$
Sigmoid	$y = \begin{cases} 0 & \text{if } x \leq \frac{49}{100} \\ 50(x-\frac{1}{2})+\frac{1}{2} & \text{if } \frac{49}{100} \leq x \leq \frac{51}{100} \\ 1 & \text{if } x > \frac{51}{100} \end{cases}$	$x \in [0,1]$
Varying Freq [Medium] Cosine	$y = \sin(5\pi x(1+x))$	$x \in [0,1]$
Varying Freq [Medium] Sine	$y = \sin(6\pi x(1+x))$	$x \in [0,1]$
Spike	$y = \begin{cases} 20x & \text{if } x < \frac{1}{20} \\ -18x + \frac{19}{10} & \text{if } \frac{1}{20} \leq x < \frac{1}{10} \\ -\frac{x}{9} + \frac{1}{9} & \text{if } x \geq \frac{1}{10} \end{cases}$	$x \in [0,1]$
Lopsided L-shaped	$y = \begin{cases} 200x & \text{if } x < \frac{1}{200} \\ -198x + \frac{199}{100} & \text{if } \frac{1}{200} \leq x < \frac{1}{100} \\ -\frac{x}{99} + \frac{1}{99} & \text{if } x \geq \frac{1}{100} \end{cases}$	$x \in [0,1]$

Table S3: Definitions of the functions used in Figures 2B-F and Figures S3 and S4.

Table S4: The relationships analyzed in Figures 2B-F. Sample sizes are indicated next to each function type in parentheses. For definitions of the functions, see Table S3.

Remark 4.1. Note that the relationship called "non-coexistence" is not a plot of two statistically independent variables with skewed distributions. Rather, it is a superposition of two random clouds that are scaled in a way that imitates a situation in which, for instance, the presence of one gene de-activates another, or one bacterium suppresses another.

This figure is also included in the SOM as Figure S6A.

4.5 Analysis for Figure 3 (visualizations of characteristic matrices)

For each of the functions in Table S5, we generated a data series of 16,700 points and computed its characteristic matrix using MINE. Due to space limitations, we only visualize the matrices for the range $(1 < x \leq 30, 1 < y \leq 30)$; in fact, analyzing 16,700 data points using $B(n) = n^{0.7}$ gives up to $16,700^{0.7} = 903$ cells in a gridding, allowing the characteristic matrix to have entries, for example, at $(x = 450, y = 2)$.

Relationship Name	**Description** ($x \in [0, 1]$ **for all functions)**
Linear	$y = x$
Parabolic	$y = 4(x - \frac{1}{2})^2$
Sinusoidal (Varying Frequency)	$y = \sin(6\pi x(1 + x))$
Categorical	200 points chosen from the following set: $\{(1, 0.287), (2, 0.796), (3, 0.310), (4, 0.924), (5, 0.717)\}$
Circle	$\{(\cos t, \sin t) : t \in [0, 2\pi]\}$
Random	random number generator

Table S5: Definitions of the functions used for Figure 3.

4.6 Analysis for Figure 4 (WHO global indicators dataset)

We obtained two datasets, one from the World Health Organization Statistical Information Systems (WHOSIS) website [7], and one from the Gapminder website [25]. These two databases were joined on the 'Country' variable and countries with data in only one of the two datasets were thrown out. For each country, for each indicator, the data for the time period 1995-2005 was aggregated; only the most recent value from the time period was used. The resulting dataset had 357 global indicators for 202 countries around the world, from 1960 through 2005. This dataset was analyzed using MINE, the Pearson correlation coefficient, and the Kraskov *et al.* mutual information estimator [14].

The results of the MINE analysis were examined by sorting both by MIC and by the quantity MIC $- \rho^2$. Table S6 lists the ranks, uncorrected p-values, and q-values of the relationships in parts C-H of Figure 4. A false discovery rate of 5% was used under a null hypothesis of statistical independence to determine which relationships were significant according to each method. The lines of best fit for the relationships in parts C-H of Figure 4 were generated manually using regression on each trend.

The non-linear associations identified using MIC$-\rho^2$ could not be detected as easily—or in some cases at all—using other methods. The Pearson correlation coefficient did not work because it does not differentiate between highly non-linear associations and unrelated variable pairs. Mutual information estimators did not work either: because the scores they assign do not roughly equal R^2 (as the MIC scores do), there is no natural analogue for MIC $- \rho^2$ using these statistics. Both Kraskov $- \rho^2$ as well as Kraskov $- \bar{K}\rho^2$ where $\bar{K}$ is the highest mutual information score in the dataset were tried, and neither detected the relationships found using MIC $- \rho^2$. Presumably, the preference of these methods for linear relationships, as discussed in the main text (see Figure 2), also contributes to the difficultly of using them to identify non-linear relationships.

Table S9 shows the top 1% of relationships by MIC from an abbreviated version of the global indicators dataset. This modified dataset only includes 114 of the less redundant variables from the larger dataset. The variables in the modified dataset were chosen by separating the variables from the full dataset into groups of closely related (redundant) variables, and choosing one representative from each each group. The representative was the variable that was involved in the most relationships out of the top $1/3$ of all relationships.

Relationship	rank_{MIC}	$\text{rank}_{MIC-\rho^2}$	rank_{MI}	rank_{ρ}	p_{MIC}	q_{MIC}
C	49136	-	42989	33998	1	-
D	2379	-	3418	1106	1.54×10^{-7}	1.02×10^{-6}
E	1146	7	5303	33022	1.54×10^{-7}	1.02×10^{-6}
F	9655	942	5363	43427	9.23×10^{-7}	1.02×10^{-6}
G	5986	393	9251	31119	5.38×10^{-4}	1.02×10^{-6}
H	120	-	183	874	1.54×10^{-7}	1.02×10^{-6}

Table S6: Ranks, uncorrected p-values, and q-values of relationships in Figure 4C-H. (The q-value of a relationship is the minimum false discovery rate at which that relationship may be called significant.)

Relationship	p_{MIC}	q_{MIC}
C	$< 2.3 \times 10^{-8}$	1.4×10^{-6}
D	$< 2.3 \times 10^{-8}$	1.4×10^{-6}
E	$< 1.1 \times 10^{-3}$	4.3×10^{-2}
F	$< 2.3 \times 10^{-8}$	1.4×10^{-6}
G	$< 2.3 \times 10^{-8}$	1.4×10^{-6}

Table S7: P-values (uncorrected) and q-values of relationships from Figure 5C-G. (The q-value of a relationship is the minimum false discovery rate at which that relationship may be called significant.)

Tables S10 and S11 list specific countries that follow the minority trend in Figures 4F and 4H.

4.7 Analysis for Figure 5 (gene expression dataset)

We analyzed the cdc15 expression data from Spellman et al. (1998) [26]. The sampling time in this dataset was not always equidistant. As in [24], we considered the missing sampling times as missing values, and then used linear interpolation to fill in any missing value for which both adjacent time points did have (non-interpolated) data. We also removed the first and last three time-points, interpolated or not, for each gene, as they consistently appeared unrelated to the rest of each respective time series. The resulting data file had 23 time-points.

We ran MINE on this file, telling it to compare each time series against time (rather than compare the time series against each other) and then merged the scores obtained by Spellman et al. with the results. (Note that because we used the MINE parameter cv=1.0, only genes with no missing timepoints after truncation and interpolation were analyzed. There were 4381 genes meeting this criterion.) A false discovery rate of 5% was used under a null hypothesis of statistical independence to determine which relationships had significant MIC scores. Table S7 contains uncorrected p-values as well as q-values of the relationships highlighted in Figure 5.

Note: our comparison of MINE against the other methods discussed in the main text on this dataset was limited in the following ways: [26] did not determine a threshold of statistical significance but rather heursitically set a threshold based on the scores of genes known to be periodic; [22] did not correct for multiple testing; and [24] made publicly available only the top third of the genes found to be statistically significant.

4.8 Analysis for Figure 6 (microbiome dataset)

We obtained a dataset of bacterial abundance levels for 6,696 species-level operational taxonomic units (OTU's) in humanized mice ($n = 675$ readings). For each reading, the dataset also contained the donor microbiota (fresh human fecal sample, frozen human fecal sample, or 'second generation' transfer from a humanized mouse donor), the sex of the host mouse, the diet fed to the host mouse, the collection method of the sample (luminal contents or mucosal scraping), as well as the location of sampling along the gastroin-

testinal tract (stomach, small intestine, cecum, colon, or feces) [29]. We ran MINE on this dataset, asking it to compare all the variables against each other. Due to the large number of comparisons, false discovery rates were calculated directly from an empirical null distribution rather than from uncorrected p-values [43]. The relationships deemed significant were those with FDRs below 5%.

The threshold of 0.2 for the non-linearity score MIC$-\rho^2$ was chosen heuristically based on identification of a long tail in the distribution of non-linearity scores, as well as visual inspection of several plots with a range of non-linearity scores around the beginning of this tail. As in the case of the WHO data, the non-linear relationships found using MIC $-\ \rho^2$ could not be detected as easily using existing methods because these methods give certain association types higher scores than others (see Section 4.6).

We used the following heuristic A to determine which relationships were affected by each auxiliary variable: given a set of abundance data for two OTUs D, we partition D into disjoint sets $D_1, \ldots, D_k$, one for each value of the auxiliary variable. For example, D_1 might be the data points for which the diet is western and D_2 the data points for which the diet is LFPP. For each D_i, we then calculated the average Euclidian distance Δ_i from the point (x_i, y_i) where x_i was the median x-value in D_i and y_i was the median y-value. We set

$$A = \sum_i \frac{|D_i|}{|D|} \Delta_i.$$

The heuristic A was calculated for a given relationship, and then calculated on 1,000 instances of the null hypothes consisting of an identical dataset in which each data point was randomly assigned to a set D_i in a manner that preserved the original sizes of the D_i. A relationship in which the value of A was greater than 95% of the scores obtained by the instances of the null hypotheses were said to be affected by the auxiliary variable in question.

Of the top 500 non-linear relationships, 312 were found to be affected by one or more auxiliary variables: 135 by host diet, 103 by host sex, 202 by the identity of the human donor, 161 by collection method, and 143 by location in the gastrointestinal tract.

Of the relationships with statistically significant MIC scores, the top 300 relationships by non-linearity score were placed in a spring graph similar to the graph constructed for the WHO data (see Section 4.10).

We developed a stricter heursitic to identify relationships, such as the relationship in Figure 6A, that appear almost entirely explained by diet. We called a relationship between two OTUs (call them X and Y) explained by diet if it met the following criteria:

- The heuristic A identified the relationship as being affected by diet.
- Both X and Y individually were statistically significantly related to diet according to MIC.
- The mean abundance level of X under the western diet was higher than that of Y under the western diet and the reverse was true was under the LFPP diet, or vice versa.

The motivation behind this heuristic was that we were trying to identify relationships in which one strain is suppressed under the western diet and the other strain is suppressed under the LFPP diet.

Table S13 contains a list of the 77 relationships out of the top 500 non-linear relationships that this stricter heuristic found to be almost entirely explained by diet.

Table S14 contains a list of the 188 relationships out of the top 500 non-linear relationships that the heuristic A indicated were not affected by any auxiliary variable.

4.9 Analysis of Major League Baseball statistics from 2008 season

We downloaded the salaries of Major League Baseball players for the 2008 season from The Baseball Archive [28], and the collection of 131 other offensive statistics for all players for the 2008 season from Baseball Prospectus [27]. We joined the two databases on 'Player Name', and took the following pre-processing steps:

- We removed all players with fewer than 40 plate appearances.

Relationship	p_ρ	q_ρ
salary vs. walks	$< 7.6 \times 10^{-7}$	7.8×10^{-7}
salary vs. intentional walks	$< 7.6 \times 10^{-7}$	7.8×10^{-7}
salary vs. RBI	$< 7.6 \times 10^{-7}$	7.8×10^{-7}

Relationship	p_{MIC}	q_{MIC}
salary vs. hits	5.6×10^{-4}	3.6×10^{-2}
salary vs. total bases	1.1×10^{-3}	4.2×10^{-2}
slaray vs. RPMLV	3.8×10^{-4}	3.6×10^{-2}

Table S8: P-values (uncorrected) and q-values of relationships from Major League Baseball dataset. (The q-value of a relationship is the minimum false discovery rate at which that relationship may be called significant.)

- We removed all pitchers because we were examining relationships between salary and offensive statistics, and pitchers are primarily paid for their defensive, rather than offensive, production. (Note: this applied almost exclusively to National League pitchers, as most American League pitchers were already removed for having fewer than 40 plate appearances.)
- Because we were looking for relationships between salary and offensive statistics, we removed all players who were forced to earn approximately the Major League minimum (salary $< 400,000$ USD). These players were not eligible for free agency (due to having just joined the Major Leagues) and were thus paid a fixed salary that was not based on performance.

We calculated the MIC as well as the Pearson correlation coefficient of all relationships between salary and another variable. A false discovery rate of 5% was used under a null hypothesis of statistical independence to determine which associations were significant according to each method. Table S12 lists the 50 variables most closely related to salary according to MIC. Figure S12 shows a few of the strongest non-linear distributions identified by MIC, as well as a few of the distributions of the strongest correlates with salary according to ρ. Table S8 contains uncorrected p-values, as well as q-values of the relationships mentioned in the main text.

4.10 Analysis for Figures 6E and S10 (interactive spring graphs)

To enable an intuitive and efficient interpretation of how the different variables in a dataset are related to each other, we developed a graph visualization tool based on common techniques [47], which can be applied to the results of our analysis to create an interactive visualization of datasets that we call a *spring graph*.

In a spring graph, the variables in the data set are represented by nodes, and the relationships between them are represented by edges. The graph is a dynamic physical equilibrium based on a Hookean spring model, which is governed by forces that vary in proportion to either the MIC or the other characteristics of relationships between variables. Thus, every aspect of the spring system between nodes is dependent on one of the properties of the characteristic matrix generated by MINE, allowing the user to examine several of these properties simultaneously. To determine the layout of the nodes, numerical integration of these forces is employed until the dynamics converge to one of many potential stable equilibria [48]. User-controlled perturbations can be applied in an attempt to identify lower energy equilibria. It is important to note that while this may be a useful exploratory tool, this visualization could exhibit degenerate energy landscapes and that the particular arrangement of any non-trivial number of vertices in two-dimensional space will contain some artifactual structure relative to the global minimum energy landscape.

Users can interact with this graph by adding or removing variables or relationships between variables and seeing how the physical model reacts to these changes. For example, a user can create a spring graph that contains only a few variables (nodes) and relationships (edges), and can progressively add more relationships into the graph, to see how the physical equilibrium shifts. The Spring Graph is intended to make groups of

interrelated variables immediately obvious and therefore easy for the investigator to pick out of a huge initial list of variables and relationships.

Figure S10 is a screenshot of an interactive Spring Graph generated from the output of the MINE analysis of the global indicators dataset. The graph depicted is the largest connected component (100 edges) that emerges when the top 450 relationships in this dataset (ranked by MIC) are added sequentially. For ease of display, redundant variables have been removed (for example, only one of the under-five mortality rates from the several reporting institutions was kept). In this example, variables were clustered only by MIC—spring equilibrium length, edge thickness, and coloring are all proportional to MIC score. (The shorter the spring equilibrium length and thicker the edge, the higher the MIC score; the more red a region, the more tightly correlated the variables in it, relative to the rest of the graph, while the more blue a region is, the less tightly correlated the variables in it are relative to the rest of the graph.)

Figure 6E is a screenshot of an interactive Spring Graph generated from the MINE analysis of the microbiome dataset, and includes the 300 most non-linear relationships out of all relationships with significant MIC scores. Again, the equilibrium length of each spring is inversely proportional to MIC score; the size of each node is proportional to the number of these relationships in which a given OTU is found; and the color of the node is proportional to fraction of relationships involving that node that are explained by diet. Black edges represent relationships explained by diet.

5 How to use MINE

MINE is written in Java can be downloaded as a JAR from exploredata.net. The only mandatory parameters are the name of the file containing the data and a specification of which variable pairs to analyze. It is invoked as follows:

```
java -jar MINE.java   infile   masterVariable
```

The mandatory parameters may be set as follows

- infile : A path to a comma-separated values (csv) file containing the data. The variable names can either be in the first line of the file (making each row a record), or the first column in the file (making each column an entry).
- masterVariable : Set this to '-allPairs' to compare all pairs of variables against each other or to '-adjacentPairs' to compare consecutive pairs of variables only. Set to a number i to compare all variables only against the i-th variable.

In addition, the following optional parameters/flags are provided

- cv : A floating point number indicating which percentage of the records need to have data in them for both variables before those two variables are compared. Default value is 0.
- exp : The exponent in the equation $B(n) = n^{\alpha}$. Default value is 0.6.
- c : Determines by what factor clumps may outnumber columns when OptimizeXAxis is called. When trying to partition the x-axis into x columns, the algorithm will start with at most cx clumps. Default value is 15.
- gc : The number of variable pairs to analyze before forcing a Java garbage collection. This should not be necessary unless sample size is very small and there are very many variable pairs. Default value is Integer.MAX_VALUE.
- -permute : Instructs MINE to permute the dataset before running it. If masterVariable is set to '-adjacentPairs', then every other variable will be permuted. If it is set to a variable id, all variables except for the master variable will be permuted. If it is set to '-allPairs', every variable will be permuted independently of the others. This flag is unset by default.
- jobID : A string to identify this job. The program will produce two files; one is called [infile] ,[jobID],Results.csv, and the other is called [infile],[jobID],Status.txt (always contains the name of the variable being analyzed). The default jobID is B=n^[exp],k=[c]x[-permute].

5.1 Example

```
java -jar MINE.jar "path/to/data.txt"  0  cv=0.1  exp=0.6  c=10  fewBoxes
```

This will run MINE on the file path/to/data.txt. The only variable pairs that will be analyzed are the first variable against the rest of the variables. Also, a variable pair will be ignored if less than 10% of the records have values for both the variables in question. The program will use $B(n) = n^{0.6}$ and will have the maximal number of clumps allowed be $k = 10x$ when attempting to draw a grix with x columns. Two output files will be created:

- 'path/to/data.txt,fewBoxes,Results.csv', which contains the results of the analysis, and
- 'path/to/data.txt,fewBoxes,Status.txt', which contains the name of the variable being analyzed while MINE runs.

6 Proofs about MIC

In this section, we prove the following statements from the main text:

- The MIC of data sampled from a distribution (X, Y), where X and Y are continuous random variables, converges in probability to 0 as sample size grows if and only if X and Y are statistically independent. (Theorem 1)
- The MIC of data sampled from a distribution $(X, f(X))$, where X is a continuous random variable and f is a nowhere-constant function, converges to 1 almost surely as sample size grows. (Theorem 3)
- More generally, the MIC of data sampled from a finite union of images of nowhere-flat, nowhere-vertical differentiable curves converges to 1 almost surely as sample size grows. (Theorem 4)
- For any nowhere-constant function, a set of points drawn from the curve defined by the function and then vertically perturbed will receive an MIC that is lower bounded almost surely in terms of the amount of perturbation, given a large enough sample size. Moreover, this lower bound can be stated in terms of R^2. (Theorem 5)

We also show that the restriction in the definition of MIC (Definition 2.3) that the maximal number of grid cells $B(n)$ grow no faster than $n^{1-\varepsilon}$ for some $0 < \varepsilon < 1$ is necessary in the sense that without this condition data drawn from (X, Y) with X and Y statistically independent will achieve non-trivial scores with high probability. (Theorem 2)

6.1 Preliminaries

There are a few lemmas and definitions that we will use in many of the proofs in this section. The first is the following multiplicative Chernoff bound, which we state without proof. (It appears in many standard texts; see, e.g. [Corollary 4.6] of [49].)

Lemma 6.1 (Multiplicative Chernoff Bound)**.** *Let $X_1, \ldots, X_n$ be independent random variables that each equal 1 with probability p and 0 otherwise. Then for every $0 < \varepsilon \leq 1$ we have*

$$\Pr\left[\left|\sum X_i - pn\right| \geq \varepsilon pn\right] \leq 2^{-\Omega(pn\varepsilon^2)}$$

The next lemma states that if a grid is imposed on a set of points sampled from some prior distribution and each cell contains approximately the expected number of points then the measured mutual information will be close to the mutual information of the distribution.

Lemma 6.2. *Let G be a grid with B cells imposed on a set D of n points in $[0, 1] \times [0, 1]$. Given some probability distribution (X, Y) on the cells of G, define $\varepsilon_{i,j}$ to be the difference between the fraction of points in the i, j-th cell of G and the probability of choosing that cell from the distribution (X, Y). Define $\varepsilon_{i,*}$ for the i-th row and $\varepsilon_{*,j}$ for the j-th column analogously. We have*

$$|I(D|_G) - I(X; Y)| \leq \frac{B}{2n} + \sum_i O({\varepsilon_{i,*}}^3 + {\varepsilon_{*,i}}^3) + \sum_{i,j} O({\varepsilon_{i,j}}^3)$$

Proof. See Equations 10, 11, 12, 13, 26, and 27 in [50]. □

The following lemma gives the mutual information of a distribution on a grid in terms of the average of the entropies of the columns making up the grid. Before stating it we will develop some notation that will be useful later in this section: for a distribution D on the cells of a grid G, let $H(D)$ denote the Shannon entropy of D; let $H^Y(D)$ denote the Shannon entropy of the marginal distribution on the rows of G and define $H^X(D)$ analogously; let $H^Y_j(D)$ denote the Shannon entropy of the distribution on the rows of G induced only by the probability mass in the j-th column, and similarly let $H^X_i(D)$ denote the corresponding based on the distribution of the columns of G; let $p(\cdot, \cdot)$ be the probability mass function of D; and let p_X and p_Y be the marginal probability mass functions of p.

Lemma 6.3. *Let D be a distribution on the cells of a grid G with y rows and x columns. We have that*

$$I(D) = H^Y(D) - \sum_{j=1}^{x} p_X(j) H_j^Y(D)$$

Proof. The lemma follows from the fact that for two jointly distributed random variables X and Y, $I(X;Y) = H(Y) - H(Y|X)$. If we let Y be the marginal distribution on the rows of G and X the marginal distribution on the columns of G, then the $I(X;Y)$ term becomes $I(D)$ and $H(Y) = H^Y(D)$. What does the $H(Y|X)$ term equal? We know that $H(Y|X = j)$ equals $H_j^Y(D)$. Thus,

$$H(Y|X) = E_X[H(Y|X)] = \sum_{j=1}^{x} p_X(j) H_j^Y(D)$$

which gives the result. □

We now define a few simple terms. Recall that an *equipartition* is a partition into either rows or columns such that each row/column contains the same number of points. Since the mutual information of a distribution is upper-bounded by the minimum of the Shannon entropies of the marginal distributions, we will often be concerned with whether we are able to make our grids such that the axis with fewer rows/columns can be partitioned in a way that is close to an equipartition. For this reason, we will need the following two definitions.

Definition 6.4. A finite set $S \subset \mathbb{R}$ is (m, α)*-partitionable* if it can be partitioned into at most m bins such that the discrete random variable induced on the bins has Shannon entropy at least $\alpha \log m$.

Definition 6.5. A finite set $S \subset \mathbb{R}$ is m*-equipartitionable* if it is $(m, 1)$-partitionable.

For example, every even-sized set of distinct numbers is 2-equipartitionable. If m does not divide the size of a set of distinct numbers, then the set will not be m-equipartitionable. However, our final lemma shows that this effect is negligible.

Lemma 6.6. *For every $m \in \mathbb{N}$ and every $0 < \varepsilon \leq 1$, and for sufficiently large n, every set of n distinct numbers is $(m, 1-\varepsilon)$-partitionable.*

6.2 MIC approaches 0 if and only if X and Y are statistically independent

We now show that because $\omega(1) < B(n) \leq O(n^{1-\varepsilon})$ in the definition of MIC, the MIC of data sampled from a distribution (X, Y) converges in probability to 0 if and only if X and Y are statistically independent. This is done in Section 6.2.1 and Section 6.2.2, whose main results give us the following theorem. (In one direction, the theorem is only proven for our heuristic approximation of MIC.)

Theorem 1. *Let D be a set of n independent samples from a distribution (X, Y) over $[0,1] \times [0,1]$ where X and Y are continuous random variables. The following statements hold:*

1. *If X and Y are statistically independent, ApproxMIC(D) converges to 0 in probability as $n \to \infty$.*
2. *If X and Y are not statistically dependent, there exists a constant $\zeta > 0$ such that MIC$(D) \geq \zeta$ for sufficiently large n almost surely.*

Proof. The result follows from Proposition 6.8 and Proposition 6.12. □

After proving Proposition 6.8 and Proposition 6.12, we will show that our constraint on $B(n)$ in the definition of MIC is tight in the sense that if we instead used $B(n) = \Omega(n^{1+\varepsilon})$ then statistically independent points would almost surely have MICs of 1. This is done in Section 6.2.3.

6.2.1 MIC is bounded away from 0 almost surely for statistically dependent data

To prove that MIC is bounded away from 0 for statistically dependent data, will need the fact that if two variables each taking values in $[0, 1]$ are not statistically independent then there exists a two-by-two grid such that the joint distribution induced on the cells of the grid by the probability mass in each cell has non-zero mutual information. This is stated in the following lemma.

Lemma 6.7. *Let X and Y be random variables each taking values in $[0, 1]$. For any $\alpha \in [0, 1]$, define $\chi_\alpha : [0, 1] \to \{0, 1\}$ to be 0 on the interval $[0, \alpha)$ and 1 on the interval $[\alpha, 1]$. Then if $I(\chi_a(X); \chi_b(Y)) = 0$ for all $a, b \in [0, 1]$, then X and Y are statistically independent.*

Proof. Because in general $I(X; Y) = 0$ only if X and Y are statistically independent, our assumption implies that $\chi_a(X)$ and $\chi_b(Y)$ are independent for all $a, b \in [0, 1]$. The claim then follows from Theorem 2.1.3 of [51] with $\mathcal{A}_1 = \{\{(x, y) \in [0, 1] \times [0, 1] : x < a\} : a \in [0, 1]\}$ and $\mathcal{A}_2 = \{\{(x, y) \in [0, 1] \times [0, 1] : y < a\} : a \in [0, 1]\}$. □

The above lemma implies that when X and Y are *not* statistically independent, we have a two-by-two gridding of the unit box such that the distribution induced on the grid cells by the distribution (X, Y) has non-zero mutual information. The following proposition uses this to lower-bound the MIC of data drawn from (X, Y) in this case.

Proposition 6.8. *Let D be a set of n independent samples from a distribution (X, Y) over $[0, 1] \times [0, 1]$. If X and Y are not statistically independent, then there exists a constant $\zeta > 0$ such that MIC$(D) \geq \zeta$ for sufficiently large n almost surely.*

Proof. To establish our claim, we need to exhibit a grid with few enough cells whose normalized score is bounded away from 0 with high probability as n grows. Because MIC is the maximum over the normalized scores of all grids with at most $B(n)$ cells, this will lower-bound MIC. We will exhibit a grid with 4 cells that serves our purpose.

The contrapositive of Lemma 6.7 gives us our grid because it says that there exist a and b such that $I(\chi_a(X); \chi_b(Y)) \gneq 0$. Consider the grid G that partitions the x-axis at a into 2 columns and the y-axis at b into 2 rows (independently of any data points). As n grows, the strong law of large numbers gives that the distribution $D|_G$ induced by D on the cells of G will approach $(\chi_a(X); \chi_b(Y))$ almost surely. Lemma 6.2 then gives that $D|_G$ will have mutual information greater than some constant ζ as n grows almost surely. Because this grid has 2 rows and 2 columns, its normalized score will equal its mutual information, giving the result. □

6.2.2 MIC converges to 0 in probability for statistically independent data

Before we prove this result, we establish some basic facts about randomly drawn points in the unit box. The following lemma will allow us to think of a set of n points drawn from a distribution (X, Y) on $[0, 1] \times [0, 1]$ with X and Y statistically independent as a random permutation on n elements.

Lemma 6.9. *There exists a map φ from sets of n points in $[0, 1] \times [0, 1]$ to the symmetric group S_n with the property that for every distribution (X, Y) on $[0, 1] \times [0, 1]$ with X and Y statistically independent continuous random variables, and for every $\sigma \in S_n$, we have $\Pr[\varphi(D) = \sigma] = 1/n!$ where the probability is over the choice of the set D of n independent samples from (X, Y).*

Proof. We construct φ. Because X and Y are continuous, the points in D will almost surely have distinct x- and y-values. Thus, we need only define φ on such sets, because the value of φ on sets with non-distinct x- or y-values will not change $\Pr[\varphi(D) = \sigma]$.

Any set D of points that all have distinct x- and y-values can be written as $\{(a_1, b_{\sigma(1)}), \ldots, (a_n, b_{\sigma(n)})\}$ where the lists $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ are sorted and σ is a random permutation in S_n. For any such set D, we therefore define $\varphi(D) = \sigma$.

By statistical independence, the y-values of each of the n points in D are drawn from the same distribution. This implies that any ordering among them is equally likely. Since there are $n!$ possible orderings, the probability of ending up with any particular one is therefore $1/n!$. □

The next lemma will be used later to show that a random set of points that is equipartitioned on the y-axis will not contain too many consecutive points in the same row. In the lemma, we call a q-ary string of length n *balanced* if each value in $[q]$ appears at least $\lfloor n/q \rfloor$ times in the string. Also, for any q-ary string, we call a set of consecutive digits that are equal a *clump*.

Lemma 6.10. *The probability that a randomly chosen balanced q-ary string of length n contains a clump of length at least a is at most n/q^{a-1}.*

Proof. For simplicity, we assume $q \mid n$. We can think of choosing successive digits of a balanced q-ary string of length n as successively choosing balls without replacement from a set of q bins that each start out containing n/q points. The probability that the first a digits of such a string equal zero is at most $1/q^a$ because the first digit is zero with probability at most $1/q$ and successive digits only become less likely to equal zero as long as only zeros are being chosen.

We now take successive union bounds, first over all q possible values of the digit that makes up this clump and then over the at most n possible starting positions for the clump. The result is that the probability of having any clump in the string is at most n/q^{a-1}, as desired. □

The last lemma we prove before establishing our result shows that if the gridding subroutine of the MIC algorithm equipartitioned along each axis then most of the cells would contain close to the expected number of points.

Lemma 6.11. *Let S be a randomly chosen permutation on $[n]$, and define $D = \{(i, S(i)\}$. Let G be a $y \times k$ grid that contains equal numbers of points in every row and column. For every $1 \le i \le y$ and $1 \le j \le k$, define $\varepsilon_{i,j}$ relative to the uniform distribution on the cells of G as in Lemma 6.2. If $ky = O(n^{1-\varepsilon})$, then with probability at least $1 - kyn^{-\Omega(\log n)}$ we have*

$$|\varepsilon_{i,j}| \le \frac{\log n}{\sqrt{kyn}}$$

for all i and j and sufficiently large n.

Proof. We can write the $\varepsilon_{i,j}$ in terms of the number of points that fall in each cell as

$$\varepsilon_{i,j} = \frac{n_{i,j}}{n} - \frac{1}{ky}$$

where $n_{i,j}$ is the number of points in the i,j-th cell of our grid.

We first bound each $\varepsilon_{i,j}$ individually and then use a union bound. Without loss of generality we will consider $\varepsilon_{1,1}$, which only depends on the distribution of the n/k points in the first column of our grid. Choosing which row each of these points falls in involves assigning a value in $[y]$ to each of the x-values $\{1, \ldots, n/k\}$. Because we are choosing a random permutation and the partition of the y-axis needs to be an equipartition, these assignments are not independent of each other: every row, or value in $[y]$, will appear exactly n/ky times. However, in such a setting we can still apply Chernoff bounds; this is a consequence of the theory of negative dependence of random variables. (See, e.g., [Chapter 3.1] of [52].)

The variable $n_{1,1}$ is a sum of $n' = n/k$ trials each having success probability $p = 1/y$. Although the trials are not independent we can apply the Chernoff bound stated in Lemma 6.1 with $\varepsilon = \log n/\sqrt{pn'} = \sqrt{ky}\log n/\sqrt{n}$, which is less than 1 for sufficiently large n because of the assumption that $ky = O(n^{1-\varepsilon})$. Doing so gives

$$\Pr\left[\left|n_{1,1} - \frac{n}{ky}\right| > \frac{\sqrt{n}\log n}{\sqrt{ky}}\right] \le 2^{-\Omega(\log^2 n)}$$

$$\Rightarrow \quad \Pr\left[|\varepsilon_{1,1}| > \frac{\log n}{\sqrt{kyn}}\right] \le 2^{-\Omega(\log^2 n)}$$

where the second inequality is obtained from the first by dividing the condition in the $\Pr[\cdots]$ by n.

To get the bound above to apply to all the $\varepsilon_{i,j}$ simultaneously, we use a union bound: since there are ky cells in total, the probability that the condition is not satisfied even for one of them is at most $ky2^{-\Omega(\log^2 n)}$, as desired. □

We now state the main result of this subsection. The result is only proven for our heuristic approximation of MIC.

Proposition 6.12. *Let D be a set of n independent samples from a distribution (X, Y) over $[0, 1] \times [0, 1]$ with X and Y statistically independent continuous random variables. Then ApproxMIC(D) converges in probability to 0 as $n \to \infty$.*

Proof. Recall that, for every number of rows and columns that it considers, the ApproxMIC algorithm begins by drawing an equipartition of, without loss of generality, the y-axis into y rows for some $y > 0$. It then attempts to equipartition the x-axis into $k = \lfloor cB(n)/y \rfloor$ columns for some constant c, subject to the restriction that no clumps are split in this process. This restriction is enforced greedily: the algorithm runs through a sorted list of x-values and repeatedly adds successive clumps to its current column when doing so brings the current column size closer to the column size of an equipartition. Last, the algorithm chooses a set of at most $\lfloor B(n)/y \rfloor - 1$ of the $k - 1$ lines partitioning the x-axis that maximizes the resultant mutual information.

Now suppose that X and Y are statistically independent. We upper-bound the MIC of points drawn from (X, Y) by showing that no step the ApproxMIC algorithm takes allows it to produce a distribution on the cells of the grid that is far from uniform. Lemma 6.9, together with the fact that the grids drawn depend only on the order of the points and not on their actual positions, means that instead of proving the claim for our set of n points chosen from (X, Y), we can prove it for the plot of a randomly chosen permutation on n elements. In this case, Lemma 6.11 gives us that if the algorithm ignored clumps in the way it drew the first partition into k columns, we would have $|\varepsilon_{i,j}| \leq \log n/\sqrt{kyn}$ for every cell with high probability. Lemma 6.10 implies that with probability at least $1 - 2/n$ there are no clumps of length greater than $2 \log n$. When both of these conditions are met, we have that even if every adjustment of the equipartition to accommodate clumps brings every cell farther away from uniform, the adjusted grid satisfies

$$|\varepsilon_{i,j}| \leq \frac{\log(n)}{\sqrt{nky}} + \frac{O(\log(n))}{n}$$

Now that we know that the first partition gives a distribution close to uniform, we analyze the sub-partition that the algorithm actually uses. To do this, we note that the error term $\hat{\varepsilon}_{i,j}$ for any cell in the sub-partition that results from the merging of two cells is at most the sum of the error terms of those two cells. Since any cell of the sub-partition will be a combination of at most k old cells, the magnitudes of all the $\hat{\varepsilon}_{i,j}$ in our sub-partition are at most

$$k\left(\frac{\log(n)}{\sqrt{nky}} + \frac{O(\log(n))}{n}\right)$$

with high probability.

We can apply the same reasoning to the variables $\hat{\varepsilon}_{i,*}$ and $\hat{\varepsilon}_{*,j}$. Each is the sum of some subset of the variables $\{\varepsilon_{i,j}\}$. Since the size of $\{\varepsilon_{i,j}\}$ is ky, the maximal absolute value of each of the $\hat{\varepsilon}_{i,j}$ is therefore at most ky times the bound on the $\varepsilon_{i,j}$.

Because these bounds imply that $\hat{\varepsilon}_{i,j}$, $\hat{\varepsilon}_{i,*}$, and $\hat{\varepsilon}_{*,j}$ all vanish for large n and the bounds hold with probability at least $1 - ky2^{-\Omega(\log^2 n)} - 2/n$, we can apply Lemma 6.2 to obtain the result. □

6.2.3 The requirement $B(n) \leq O(n^{1-\varepsilon})$ in Definition 2.3 is tight

We now turn to showing that our requirement in the definition of MIC that $B(n) \leq O(n^{1-\varepsilon})$ is tight. We mean this in the sense that if $B(n)$ were allowed to grow as $n^{1+\varepsilon}$, the MIC of n randomly chosen points would grow non-trivially. Our lower bound rests on a simple argument that considers a partition into a few rows and many (up to n) columns. The following lemma is proven for the optimal MIC algorithm but also

holds for our approximation algorithm since the partition into rows is an equipartition. It will also be central to our results about the high MIC scores of functions in the next section.

Lemma 6.13. *Let D be a set of n ordered pairs with unique x-values whose y-values are (y, h)-partitionable. Let z be the number of pairs of points with consecutive x-values whose y-values are in different bins of the partition. Then MIC$(D) \geq h$ if $B(n) \geq y(z+1)$.*

Proof. We construct a grid G that achieves a score of h. We begin with the equipartition into rows. Next, we order the points of D by their x-value and add a vertical line between every two consecutive points whose y-values are in different bins. This grid has at most y rows and $z+1$ columns, making for at most $y(z+1)$ total cells, as required.

By Lemma 6.3, we have that $I(D|_G) = H^Y(D|_G) \geq h \log y$ because $H_j^Y(D|_G) = 0$ for all j. And since the normalization step is a division by $\log y$, the MIC is at least h. □

Applying Lemma 6.13 to statistically independent distributions shows us that if we did allow $B(n)$ to grow like $n^{1+\varepsilon}$ then the MIC of statistically independent data would be 1 for sufficiently large n. We show this below.

Theorem 2. *Let (X, Y) be a joint distribution over $[0,1] \times [0,1]$ where X and Y are continuous random variables. Let D be a set of n points drawn from (X, Y) with n even. If MIC were defined in a way that allowed $B(n) = \Omega(n^{1+\varepsilon})$, then we would have MIC$(D) = 1$ almost surely for sufficiently large n.*

Proof. Since X and Y are continuous, the x- and y-values of D are all unique almost surely. It follows that the y-values are 2-equipartitionable almost surely. The result then follows from application of Lemma 6.13 with $y = 2$ together with the fact that for any collection of points with unique x- and y-values, $z \leq n - 1$ and so $2(z+1) \leq 2n = o(n^{1+\varepsilon})$. □

6.3 Most noiseless functions have MICs approaching 1

We now show that data drawn from a distribution $(X, f(X))$ where f is a nowhere-constant function and X is a continuous random variable will receive MIC scores approaching 1 as sample size grows. By nowhere-constant we mean that the function in question is not constant on any open interval. We note that this restriction is reasonable in the sense that the set of nowhere-constant functions is dense in the set of all functions; that is, any function can be approximated arbitrarily well by nowhere-constant functions.

Our result is an easy consequence of Lemma 6.13. The crux of the argument is that since the y-values of the data are related by a function to their corresponding x-values, the number of columns required to ensure that each column contains only one non-empty cell is small. In particular, this number depends only on the function in question so that as sample size increases MIC will detect the function almost surely.

The following proposition is at the heart of our statement about noiseless functions. It is followed by our main theorem and an additional proposition that make its consequences concrete.

Proposition 6.14. *Fix a function f on the unit interval and let D be a set of n distinct ordered pairs contained in the set $\{(x, f(x)) : x \in [0,1]\}$ whose y-values are $(B(n)^\alpha, h)$-partitionable for some $\alpha < 1/2$. Then MIC$(D) \geq h$ for sufficiently large n.*

Proof. There are at most $kB(n)^\alpha$ pairs of points with consecutive x-values with y-values in different bins of the equipartition, where k is a constant that depends on f. Therefore, when n is sufficiently large, the result follows from application of Lemma 6.13 with $y = B(n)^\alpha$ since $y(z+1) = o(B(n))$. □

The following theorem is our main result about the MIC scores of noiseless functions. It is stated in terms of distributions and takes into account the fact that when sample size is odd, the lack of perfect equipartitionability of the y-values means that the MIC cannot quite equal 1.

Theorem 3. *Let D be a set of n independent samples from some distribution $(X, f(X))$ where f is a nowhere-constant function on $[0,1]$ and X is a continuous random variable. Then MIC$(D) \to 1$ as $n \to \infty$ almost surely.*

Proof. Since X is continuous, the points in D will have unique x-values almost surely and since f is never constant the points will therefore also have unique y-values almost surely. Lemma 6.6 implies that for all $0 < \varepsilon \leq 1$, the y-values in D will be $(2, 1-\varepsilon)$-partitionable almost surely for sufficiently large n. Application of Proposition 6.14 then gives that $\text{MIC}(D) \to 1$ for sufficiently large n almost surely. □

The following proposition, whose proof is similar to that of Theorem 3, shows that functions that are constant can also be detected by MIC. It is trickier to characterize exactly which ones, but the clear case is that of step functions.

Proposition 6.15. *Let f be a step function defined on the unit interval with k steps of equal sizes. Let D be a set of n independent samples from the distribution $(X, f(X))$ where X is the uniform distribution on $[0, 1]$. Then MIC$(D) \to 1$ as $n \to \infty$ almost surely.*

6.4 Most finite unions of differentiable curves have MICs approaching 1

In this section, we use a generalization of Proposition 6.14 (Proposition 6.17 below) to prove that for finite unions of differentiable curves which are nowhere flat and nowhere vertical, the MIC of points drawn from the union of the images of the curves approaches 1 almost surely as sample size grows.

By a *curve* we mean a continuous map $c : [0, 1] \to [0, 1] \times [0, 1]$. A *differentiable curve* is a map $c(t) = (x(t), y(t))$ such that dx/dt and dy/dt exist everywhere (including a right-derivative at 0 and a left-derivative at 1). We say that a differentiable curve $c(t) = (x(t), y(t))$ is *nowhere flat (resp. vertical)* if dx/dt (resp. dy/dt) equals 0 at finitely many points.

Remark 6.16. The nowhere-flat and nowhere-vertical conditions are analogous to the nowhere-constant condition imposed on functions in Proposition 6.14. They ensure that no non-trivial distribution on the union of the images of the curves is statistically independent. The following argument shows this: take any distribution (X, Y) on the image of a curve c whose support contains an open subset of the image of c. This means that the support of (X, Y) contains some $U \cap c([0, 1])$ where U is open in $[0, 1] \times [0, 1]$. The continuity of C then implies that $c^{-1}(U)$ contains an open set in $[0, 1]$ and therefore contains some interval (a, b) which is in the preimage of the support of D.

Now, since $x(t)$ is continuous and dx/dt vanishes in finitely many places, $x((a, b))$ contains some interval on the x-axis. In other words, the projection of the support of (X, Y) onto the x-axis contains an interval. Applying the same argument with $y(t)$ instead gives that the projection of the support of (X, Y) onto the y-axis contains an interval as well. Now, assuming that X and Y are statistically independent, we get that the support of (X, Y) contains the Cartesian product of these two intervals. But this is a contradiction because the support of (X, Y), being contained in a finite union of images of differentiable curves, has measure 0 in $[0, 1] \times [0, 1]$.

We now prove Proposition 6.17, which we will use to prove Theorem 4.

Proposition 6.17. *Let $f_1 \ldots f_\ell$ be functions on the unit interval, and let D be a set of n ordered pairs with distinct x-values such that for all $(x, y) \in D$, $y = f_i(x)$ for some i, and whose y-values are $(B(n)^\alpha, h)$-partitionable for some $\alpha < 1/2$. Then for all $\epsilon > 0$, MIC$(D) \geq h - \epsilon$ for sufficiently large n.*

Proof. We lower bound MIC(D) by constructing a grid G as follows. We start with the partition of the y-values into at most $B(n)^\alpha$ rows that guarantees $H^Y(D) \geq h \log(B(n)^\alpha)$. Because the x-values are distinct, we see that as in the proof of Proposition 6.14 there is a constant k depending only on our functions such that it is possible to partition the x-axis into $kB(n)^\alpha$ columns, each with at most ℓ non-empty bins. Lemma 6.3 then gives that

$$\begin{aligned} I(D) &= H^Y(D) - \sum_{j=1}^{x} p_X(j) H_j^Y(D) \\ &\geq h \log(B(n)^\alpha) - \log \ell \end{aligned}$$

where the second line follows from the first because $H^Y(D) \geq h \log(B(n)^\alpha)$ and $H_j^Y(D) \leq \log \ell$ for all j. Since the normalization step is a division by at most $\log(B(n)^\alpha)$, we therefore have

$$\text{MIC}(D) \geq h - \frac{\log \ell}{\log(B(n)^\alpha)}$$

from which the result follows. □

We now present the main result about the MIC scores of finite unions of curves.

Theorem 4. *Let $c_1 \dots c_\ell$ be nowhere flat, nowhere vertical, differentiable curves, and let D be a set of n ordered pairs lying in $\cup_i c_i([0,1])$ with distinct x- and y-values. Then for all $\epsilon > 0$, MIC$(D) \geq 1 - \epsilon$ for sufficiently large n.*

Proof. To establish the result, we just need to exhibit a finite set of nowhere-constant functions $\mathcal{F} = \{f_1, \dots, f_k\}$ on $[0,1]$ such that each point in D falls in some set $\{(x, f_i(x)) : x \in [0,1]\}$. We can then apply Proposition 6.17 since the y-values of D are $B(n)^\alpha$-equipartitionable for every $\alpha < 1/2$ and the x-values of D are distinct.

To build the set $\mathcal{F}$, we show that for each curve c_i, we have

$$c_i([0,1]) \subseteq \cup_{f \in \mathcal{F}_i} \{(x, f(x)) : x \in [0,1]\}$$

where $|\mathcal{F}_i| < \infty$. Since there are finitely many c_i, this will give that $\mathcal{F} = \cup \mathcal{F}_i$ is finite, thereby implying the result.

To build $\mathcal{F}_i$, we use the following procedure: write $c_i(t) = (x(t), y(t))$ and let $T = \{t \in [0,1] : y'(t) = 0\}$. Because c_i is nowhere vertical, there are at most finitely many of these and so, letting $\{I_a\}$ denote the closed intervals delimited by the points in $\{0,1\} \cup T$, we have that $|\{I_a\}| < \infty$. On each interval I_a, $y(t)$ can be re-written as a function f_a of $x(t)$ on I_a. Moreover, by the nowhere flatness of c_i we have that f_a is nowhere constant. Therefore, there exists a nowhere-constant function $g_a : [0,1] \to \mathbb{R}$ that equals f_a on I_a.

Now, since at every $t \in [0,1]$ we have $c_i(t) = (x(t), f_a(x(t)))$ for some f_a, letting $\mathcal{F}_i = \{f_a\}$ establishes our claim. □

6.5 A lower bound on the MIC of noisy functional distributions in terms of R^2

We now move on to analyzing the MIC of a distribution obtained by drawing points at random from $[0,1]$, passing them through a function, and then perturbing each of them by some amount of noise. Specifically, we will prove a lower bound on the MIC of such a distribution in terms of the R^2 of that distribution with respect to the noiseless function used to create it.

By R^2, we mean the squared Pearson correlation coefficient between the perturbed y-values and the true y-values of our data. The Pearson correlation coefficient is defined as follows.

Definition 6.18. Given two distributions X and Y, the *Pearson correlation coefficient $\rho_{X,Y}$ between X and Y* is defined by

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y respectively and μ_X and μ_Y are the means of X and Y respectively.

There are two sources of error that we need to address in this analysis. The first is the error introduced by the perturbation of the points; the second is the sampling error that we have had all along. Specifically, it may be that our sample looks far from a typical random sample of the function. We begin by addressing this second problem, showing that for any specified grid, the deviation of the number of points in each cell from the expectation is small with high probability. Together with Lemma 6.2, the following Lemma implies that sampling error affects the MIC negligibly for our purposes.

Lemma 6.19. *Let G be a $x \times y$ grid with x and y constant, and let D be a set of n samples drawn from some probability distribution (X, Y) over $[0,1] \times [0,1]$. For every $1 \le i \le y$ and $1 \le j \le x$, define $\varepsilon_{i,j}$ as in Lemma 6.2 relative to the distribution (X, Y) induces on the cells of G. For all i and j, and for sufficiently large n, we have that*

$$|\varepsilon_{i,j}| < \frac{\log n}{n}$$

with probability at least $1 - xy2^{-\Omega(\log^2 n)} = 1 - xyn^{-\Omega(\log n)}$.

Proof. The number of points $n_{i,j}$ in the i,j-th cell of G is the sum of n independent Bernoulli trials each having probability $p_{i,j}$ where $p_{i,j}$ is the probability mass of (X, Y) that lies in the i,j-th cell. We therefore use the Chernoff bound of Lemma 6.1 with $\varepsilon = \log n / \sqrt{p_{i,j} n}$ to bound $|\varepsilon_{i,j}|$ by the desired quantity and then perform a union bound over the xy cells of G. □

The remainder of this section is devoted to proving the following main result. In doing so, we will analyze an idealized continuous distribution instead of worrying about the number of points in each cell as a random variable. That is, we will assume that every cell of our grids contains the expected number of points. Thus, when we write MIC(X, Y) for some distribution (X, Y), we will mean the limit of the MIC of sets of samples from D as sample size grows. (The same note holds for R^2.) Lemmas 6.19 and 6.2 then imply that our results are off by an additive factor of at most $o(1)$.

Theorem 5. *Fix a function $f\colon [0,1] \to [0,1]$. Let F_h be the distribution $(X, f(X) + E_h)$ where X is the uniform distribution on $[0,1]$ and E_h is the uniform distribution on $[-h, h]$. If D is a set of n points drawn from F_h, then with probability at least $1 - n^{-\Omega(\log n)}$, we have*

$$MIC(D_n) \ge 1 - \frac{c}{s_{min}} \sqrt{\frac{3s_{max} - 2}{s_{max}}} \sqrt{\frac{1}{R^2} - 1} - o(1).$$

where s_{max} is the maximum slope, in absolute value, of f on the unit interval, c is the number of intervals on which $|f(x) - y_0| \le h$, s_{min} is the minimum slope of f on those intervals, and R is the Pearson correlation coefficient between $f(X)$ and $f(X) + E_h$.

To prove this theorem, we will upper-bound R^2, and lower bound MIC.

6.5.1 Upper-bounding R^2 of F_h

We can actually calculate R^2 exactly using the fact that R is the Pearson correlation coefficient between $f(X)$ and $f(X) + E_h$. The following two lemmas accomplish this.

Lemma 6.20. $cov(f(X), f(X) + E_h) = \sigma^2_{f(X)}$

Proof. It is easily verified that the mean of $f(X)$ equals the mean of $f(X) + E_h$. Letting μ denote this mean, we have

$$\begin{aligned}
\operatorname{cov}(f(X), f(X) + E_h) &= \int_0^1 \frac{1}{2h} \int_{-h}^{h} (f(x) - \mu)(f(x) + y - \mu) dy dx \\
&= \int_0^1 \frac{1}{2h} \int_{-h}^{h} \left((f(x) - \mu)^2 + y(f(x) - \mu) \right) dy dx \\
&= \int_0^1 (f(x) - \mu)^2 dx \\
&= \sigma^2_{f(X)}
\end{aligned}$$

□

Lemma 6.21. *The standard deviation of $f(X) + E_h$ satisfies*

$$\sigma^2_{f(X)+E_h} = \sigma^2_{f(X)} + \frac{h^2}{3}$$

Proof. Using μ as in the previous lemma, we have

$$\begin{aligned}
\sigma^2_{f(X)+E_h} &= \int_0^1 \frac{1}{2h} \int_{-h}^{h} (f(x) + y - \mu)^2 dy dx \\
&= \int_0^1 \frac{1}{2h} \int_{-h}^{h} \left((f(x) - \mu)^2 + 2y(f(x) - \mu) + y^2 \right) dy dx \\
&= \int_0^1 \left((f(x) - \mu)^2 + \frac{h^2}{3} \right) dx \\
&= \sigma^2_{f(X)} + \frac{h^2}{3}
\end{aligned}$$

□

The previous two lemmas together give the following result

Proposition 6.22. *For every function f on the unit interval and any $h > 0$, we have*

$$R(h)^2 = \frac{\sigma^2}{\sigma^2 + \frac{h^2}{3}} = \frac{1}{1 + \frac{h^2}{3\sigma^2}}$$

where $R(h)$ denotes the Pearson correlation coefficient of $f(X)$ and $f(X) + E_h$, and σ denotes the standard deviation of $f(X)$.

Corollary 6.23. *For any function $f : [0, 1] \to [0, 1]$, we have*

$$R(h)^2 \leq \frac{1}{1 + \frac{4 s_{max} h^2}{3 s_{max} - 2}}$$

Where s_{max} is the maximum slope, in absolute value, of f on the unit interval.

Proof. The function with maximum slope s_{max} that maximizes σ^2 is the concatenation of a horizontal line at $y = 0$, a line with slope s_{max}, and a horizontal line at $y = 1$. This function has variance $\frac{1}{4} - \frac{1}{6 s_{max}}$. □

6.5.2 Lower-bounding MIC

We now proceed to lower-bounding the MIC of F_h. We will do so using Lemma 6.3, which relates the mutual information of the distribution induced on the cells of any grid by F_h to the weighted average of the entropies of the columns of that distribution. We then construct a grid with the property that most of its columns have low entropy.

Lemma 6.24. *Fix a nowhere-constant function f and a noise level h, and let y_0 be the y-value such that $1/2$ the probability mass of F_h is above y_0 and half is below it. Let $\ell(f, h)$ be the fraction of the unit interval on which $|f(x) - y_0| \leq h$. Then we have*

$$MIC(F_h) \geq 1 - \ell(f, h)$$

.

Proof. Draw a horizontal gridline at $y = y_0$. Every time f enters or exits the strip $y_0 \pm h$, draw a vertical line. If the j-th column of our grid has $|f(x) - y_0| > h$, then $H^Y_j(F_h|_G) = 0$. On the other hand, if it has $|f(x) - y_0| \leq h$, then we still have $H^Y_j(F_h|_G) \leq 1$ because binary entropy never exceeds one. The result follows from Lemma 6.3. □

Corollary 6.25. *Let c be the number of intervals on which $|f(x) - y_0| \leq h$, and let s_{min} be the minimum slope of f on those intervals. Then*

$$MIC(F_h) \geq 1 - \frac{2c}{s_{min}} h$$

Proof. $\ell(f, h) \leq \frac{2c}{s_{min}} h$. □

Theorem 5 then follows from combining Corollary 6.23 and Corollary 6.25 with Lemmas 6.19 and 6.2.

Main Text References

[1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2009.

[2] S. Staff, "Challenges and opportunities," *Science*, vol. 331, p. 692, 2011.

[3] *[Endnote]*.

[4] A. C. et al., "Influence of life stress on depression: Moderation by a polymorphism in the 5-htt gene," *Science*, vol. 301, no. 5631, p. 386, 2003.

[5] R. Clayton and T. Mayeda, "Oxygen isotope studies of achondrites," *Geochimica et Cosmochimica Acta*, vol. 60, no. 11, pp. 1999–2017, 1996.

[6] T. Algeo and T. Lyons, "Mo-total organic carbon covariation in modern anoxic marine environments: Implications for analysis of paleoredox and paleohydrographic conditions," *Paleoceanography*, vol. 21, no. 26, p. PA1016, 2006.

[7] World-Health-Organization, "World health organization statistical information systems (whosis)," 2009. http://www.who.int/whosis/en/.

[8] A. Rényi, "On measures of dependence," *Acta Mathematica Hungarica*, vol. 10, no. 3, pp. 441–451, 1959.

[9] C. Stone, "Consistent nonparametric regression," *The Annals of Statistics*, p. 595, 1977.

[10] W. Cleveland and S. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1998.

[11] *[Endnote]*.

[12] Y. Moon, B. Rajagopalan, and U. Lall, "Estimation of mutual information using kernel density estimators," *Physical Review E*, vol. 52, no. 3, pp. 2318–2321, 1995.

[13] G. Darbellay and I. Vajda, "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1315–1321, 1999.

[14] A. Kraskov, H. Stogbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, 2004.

[15] L. Breiman and J. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, vol. 80, no. 391, pp. 580–598, 1985.

[16] T. Hastie and W. Steutzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[17] R. Tibshirani, "Principal curves revisited," *Statistics and Computing*, vol. 2, no. 4, pp. 183–190, 1992.

[18] B. Kégl, A. Krzyzak, T. Linder, and K. Zeger, "A polygonal line algorithm for constructing principal curves," *Advances in Neural Information Processing Systems*, vol. 10, pp. 501–507, 1999.

[19] P. Delicado and M. Smrekar, "Measuring non-linear dependence for two random variables distributed along a curve," *Statistics and Computing*, vol. 19, no. 3, pp. 255–269, 2009.

[20] *[Endnote]*.

[21] G. Szekely and M. Rizzo, "Brownian distance covariance," *Annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[22] B. Tu, A. Kudlicki, M. Rowicka, and S. McKnight, "Logic of the yeast metabolic cycle: Temporal compartmentalization of cellular processes," *Science*, vol. 310, no. 5751, p. 1152, 2005.

[23] R. Fisher, "Tests of significance in harmonic analysis," in *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 125, pp. 54–59, 1929.

[24] M. Ahdesmaki, H. Lahdesmaki, R. Pearson, H. Huttunen, and O. Yli-Harja, "Robust detection of periodic time series measured from biological systems," *BMC Bioinformatics*, vol. 6, no. 1, p. 117, 2005.

[25] H. Rosling, "Indicators in gapminder world," 2008. http://www.gapminder.org/data/.

[26] P. S. et al., "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the cell*, vol. 9, no. 12, p. 3273, 1998.

[27] "Baseball prospectus statistics reports," 2009. http://www.baseballprospectus.com/sortable/.

[28] S. Lahman, "The baseball archive," 2009. http://baseball1.com/statistics/.

[29] P. Turnbaugh, V. Ridaura, J. Faith, F. Rey, R. Knight, and J. Gordon, "The effect of diet on the human gut microbiome: A metagenomic analysis in humanized gnotobiotic mice," *Science Translational Medicine*, vol. 1, no. 6, p. 6ra14, 2009.

[30] L. C. et al., "Human resources for health: Overcoming the crisis," *The Lancet*, vol. 364, no. 9449, pp. 1984–1990, 2004.

[31] S. Desai and S. Alva, "Maternal education and child health: Is there a strong causal relationship?," *Demography*, vol. 35, no. 1, pp. 71–81, 1998.

[32] S. Gupta and M. Verhoeven, "The efficiency of government expenditure: Experiences from africa," *Journal of Policy Modeling*, vol. 23, no. 4, pp. 443–467, 2001.

[33] T. G. et al., "Obesity in the pacific: Too big to ignore," *Noumea, New Caledonia: World Health Organization Regional Office for the Western Pacific, Secretariat of the Pacific Community*, 2002.

[34] *[Endnote]*.

[35] P. T. et al., "The human microbiome project," *Nature*, vol. 449, no. 7164, pp. 804–810, 2007.

[36] R. L. et al., "Evolution of mammals and their gut micobes," *Science*, vol. 320, no. 5883, p. 1647, 2008.

[37] *The World Factbook 2009*. Washington, DC: Central Intelligence Agency, 2009.

Supplemental References

[38] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 2006.

[39] R. Steuer, J. Kurths, C. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, pp. 231–240, 2002.

[40] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer, "Testing for nonlinearity in time series: The method of surrogate data," *Physica D*, vol. 58, p. 77, 1992.

[41] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate - a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B - Methodological*, vol. 57, pp. 289–300, 1995.

[42] B. Devlin, K. Roeder, and L. Wasserman, "Statistical genetics: False discovery or missed discovery?," *Heredity*, vol. 91, pp. 537–538, 2003.

[43] W. Noble, "How does multiple testing correction work?," *Nature Biotechnology*, vol. 27, no. 12, pp. 1135–1137, 2009.

[44] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *The Annals of Statistics*, vol. 29, no. 4, pp. pp. 1165–1188, 2001.

[45] M. Weingessel, *Package 'Princurve'*. 2009. http://cran.r-project.org/web/packages/princurve/.

[46] P. Delicado and M. Huerta, "Principal curves of oriented points: theoretical and computational improvements," *Computational Statistics*, vol. 18, no. 2, 2003.

[47] T. Fruchterman and E. Reingold, "Graph drawing by force-directed placement," *Software—Practice and Experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[48] J. T. Bernstein, "Traer physics 3.0," 2009. http://code.google.com/p/traer-physics/.

[49] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. New York: Cambridge Univ. Press, 2005.

[50] M. Roulston, "Estimating the errors on measured entropy and mutual information," *Physica D: Nonlinear Phenomena*, vol. 125, no. 3-4, pp. 285–294, 1999.

[51] R. Durret, *Probability: Theory and Examples*. New York: Cambridge Univ. Press, 2010.

[52] D. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. New York: Cambridge Univ. Press, 2009.

[53] D. Xiang and G. Wahba, "A generalized approximate cross validation for smoothing splines with non-gaussian data," *Statistica Sinica*, vol. 6, pp. 675–692, 1996.

[54] W. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

Figure S3: Additional measures of dependence applied to 27 different functional relationship types and graphed against R^2 for each relationship type as in Figure 2. The statistics on the y-axes are (a) The mean squared error (MSE) relative to the estimated principal curve of the data and (b) Distance correlation [21]. For more information on how these plots were created, see Section 4.3.

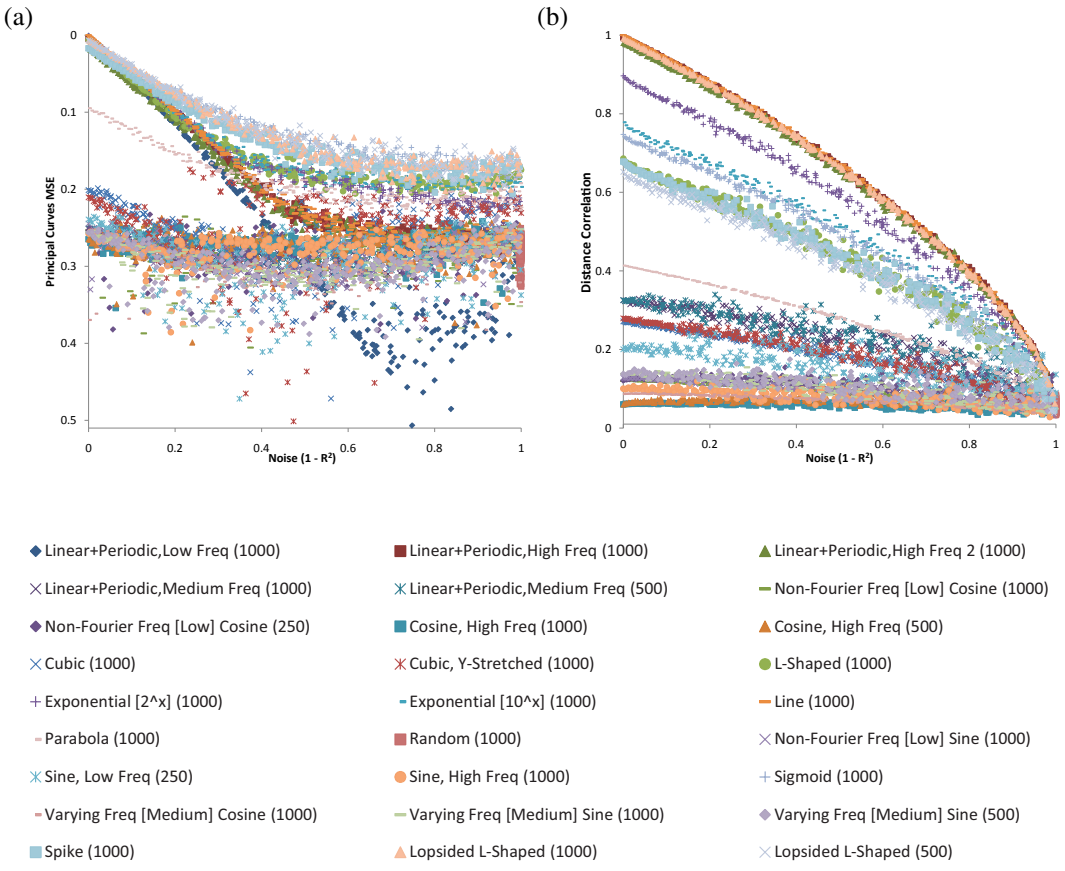
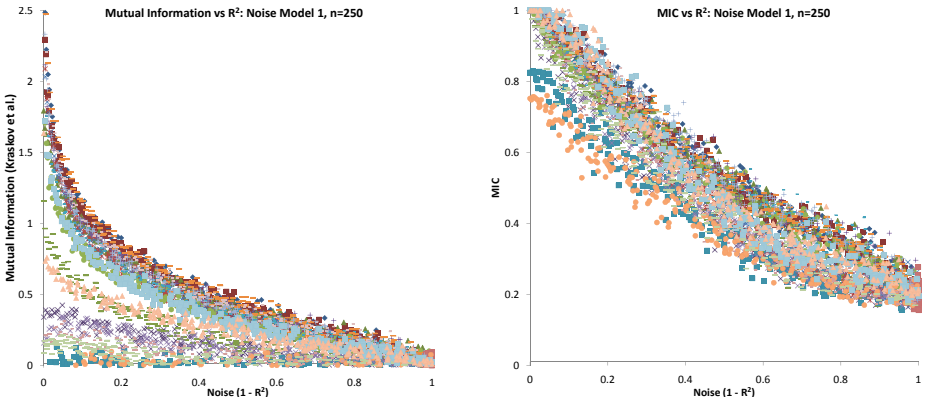
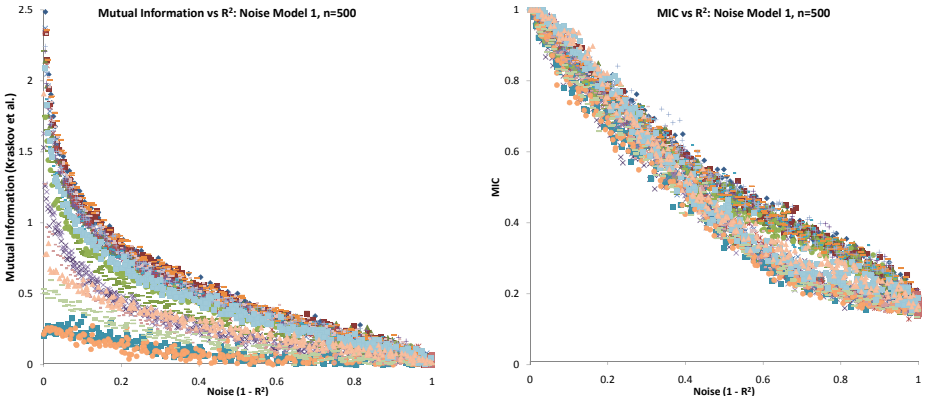


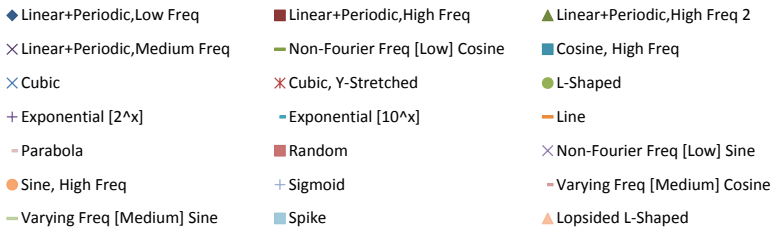
Figure S4: MIC and mutual information vs. R^2 as in Figure 2, for various noise models and sample sizes as follows. For parts (a) and (b), points were chosen evenly along the curve $\{(t, f(t)\}$ described by each function f, and noise was added only in the y direction (as in Figure 2). In part (a), the sample size is always 250, and in part (b), the sample size is always 500. Parts (c), (d), and (e) show results for different noise models: for part (c), points are chosen evenly along the curve described by the function, and noise is added in both the x and y directions; for part (d), points are chosen evenly along the x-axis, and noise is added only in the y direction; and for part (e), points are chosen evenly along the x axis, and noise is added in both the x and y directions. The legend for parts (c), (d), and (e) specifies the sample size used for each function in parentheses next to the name of that function. For more information on how these plots were created, see Section 4.3, and for descriptions of the specific functions used see Table S3. For figures (c), (d), and (e), functions with very steep portions are omitted and the "Exponential [2^x]" function has $x \in [0, 2]$ rather than $x \in [0, 10]$. This is because adding x noise to a steep function distorts its R^2, and because sampling uniformly along the x axis is also inappropriate for steep functions.

(a)

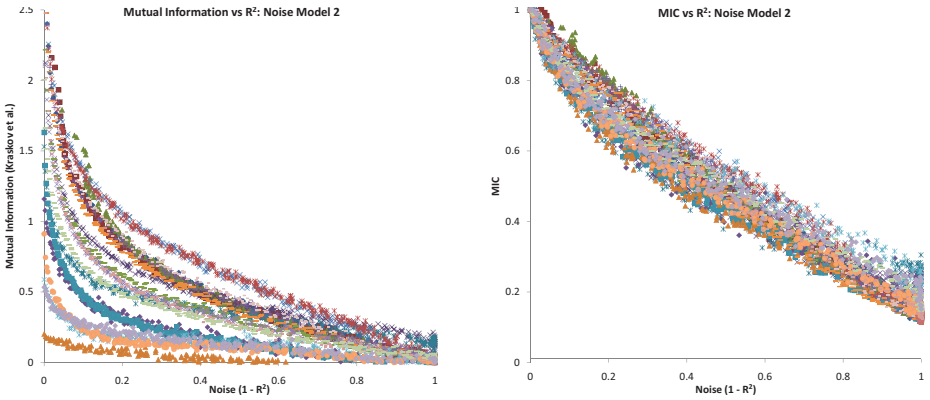


(b)

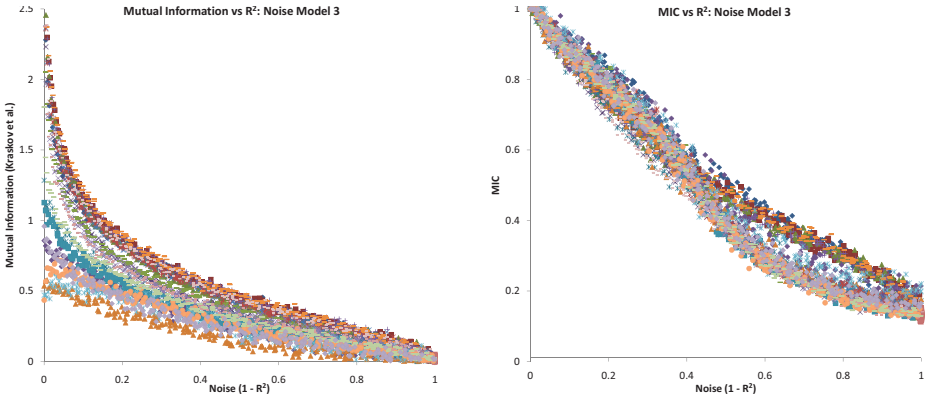




(c)


Mutual Information vs R²: Noise Model 2
Mutual Information (Kraskov et al.)
Noise (1 - R²)
MIC vs R²: Noise Model 2
MIC
Noise (1 - R²)

(d)


Mutual Information vs R²: Noise Model 3
Mutual Information (Kraskov et al.)
Noise (1 - R²)
MIC vs R²: Noise Model 3
MIC
Noise (1 - R²)

(e)

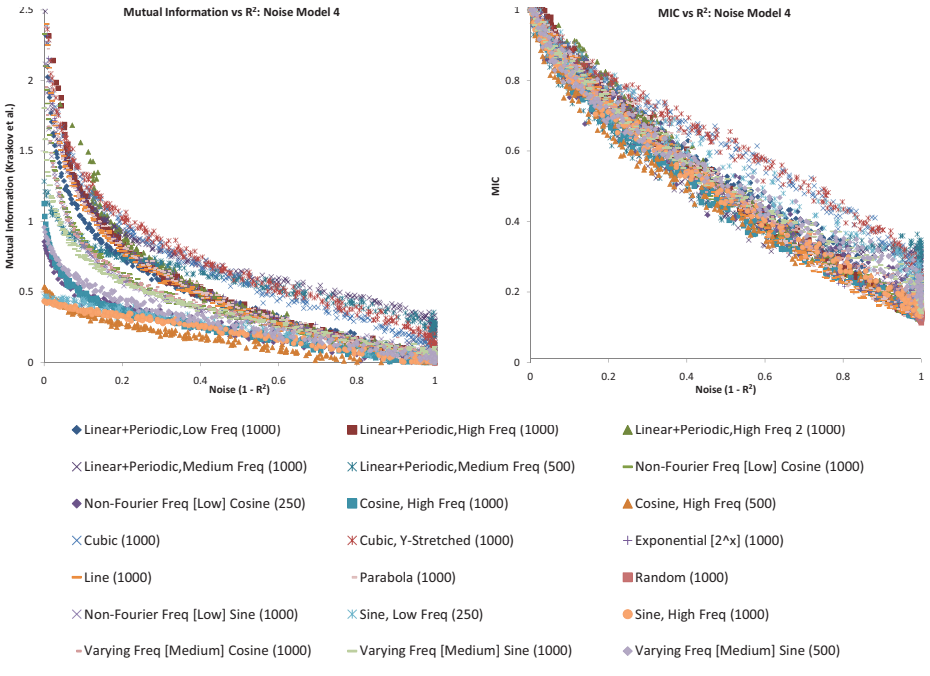
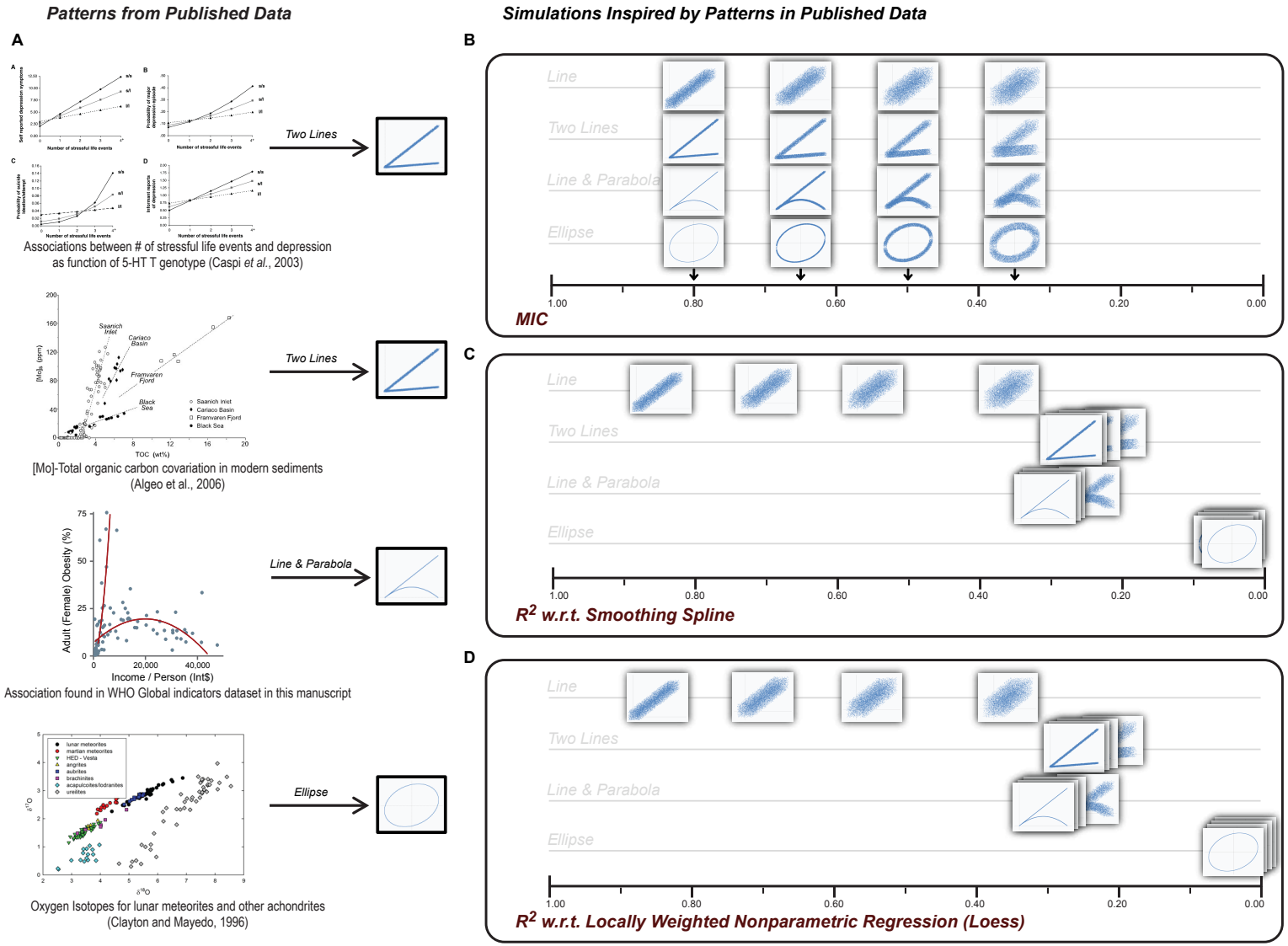
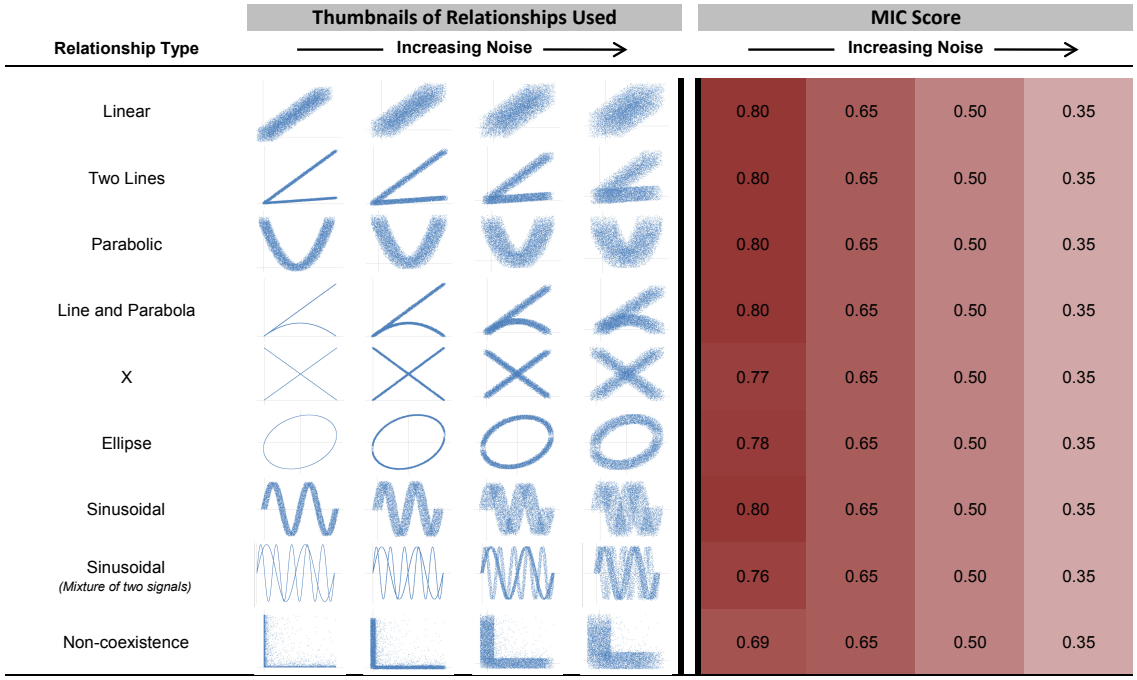

Mutual Information vs R²: Noise Model 4
Mutual Information (Kraskov et al.)
Noise (1 - R²)
MIC vs R²: Noise Model 4
MIC
Noise (1 - R²)
Linear+Periodic,Low Freq (1000)
Linear+Periodic,High Freq (1000)
Linear+Periodic,High Freq 2 (1000)
Linear+Periodic,Medium Freq (1000)
Linear+Periodic,Medium Freq (500)
Non-Fourier Freq [Low] Cosine (1000)
Non-Fourier Freq [Low] Cosine (250)
Cosine, High Freq (1000)
Cosine, High Freq (500)
Cubic (1000)
Cubic, Y-Stretched (1000)
Exponential [2^x] (1000)
Line (1000)
Parabola (1000)
Random (1000)
Non-Fourier Freq [Low] Sine (1000)
Sine, Low Freq (250)
Sine, High Freq (1000)
Varying Freq [Medium] Cosine (1000)
Varying Freq [Medium] Sine (1000)
Varying Freq [Medium] Sine (500)

Figure S5: Performance of MIC and competing methods on non-functional associations in which neither variable has strong predictive power for the other. (A) Selected non-functional relationships from the scientific literature in which neither variable has strong predictive power for the other. (B,C,D) Scores given by MIC, spatially adaptive smoothing splines (R^2) [53],[1], and loess nonparametric regression (R^2) [10],[54], respectively, to a simple linear relationship, as well as three simulated associations inspired by the relationships in (A) ($n = 10,000$). Each row contains four instances of one of these patterns with progressively more uniform horizontal and vertical noise added. The horizontal location of each thumbnail represents the score given to that set of data. As the patterns get noisier, their MIC scores degrade in an intuitive way. However, this is not the case for the other two methods, which assign the noisiest linear relationship a higher score than any instance of the other three relationships. (This is the expected behavior of these methods, since they try to find a function that describes the data).





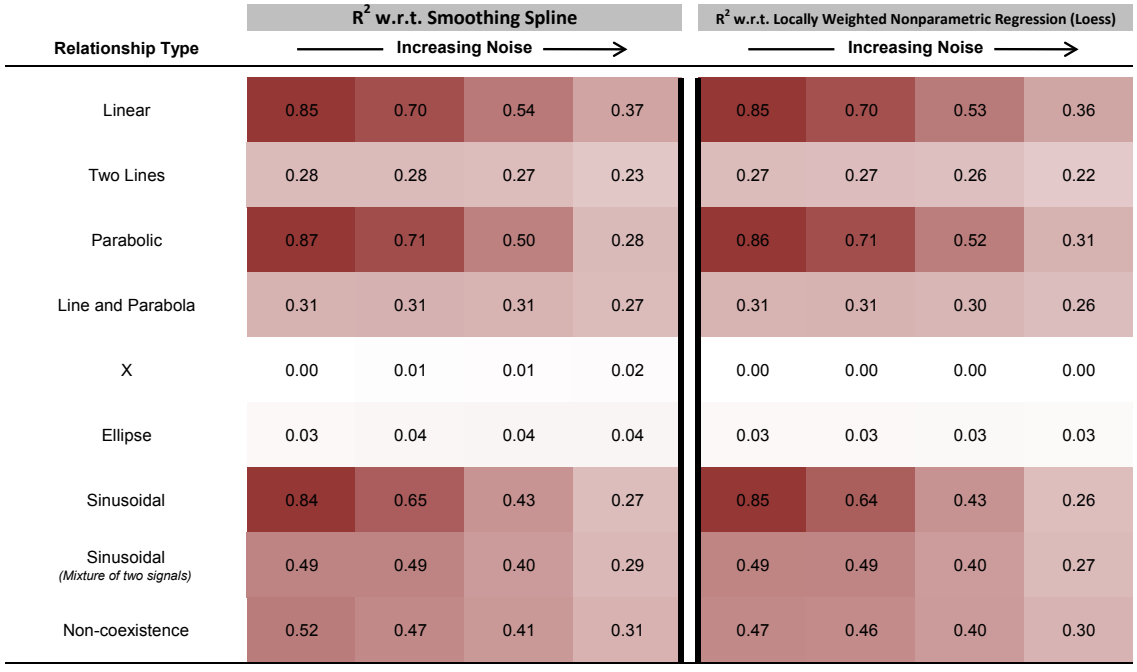


Figure S6: Performance of MIC and competing methods on non-functional associations in which neither variable has strong predictive power for the other. As in Figure S5, several relationship types were generated ($n = 10,000$) and increasing amounts of uniform horizontal and vertical noise were added to each one. Each relationship was scored with MIC, smoothing splines (R^2) [53],[1], and loess nonparametric regression (R^2) [10],[54]. (Top Left) Plots of the relationships used. (Rest) Scores assigned to each of these relationships, with cells colored according to the magnitude of the score. The MIC scores of the different relationships correspond intuitively to noise; the other two methods assign much higher scores to the functional relationships (as they are intended to do).

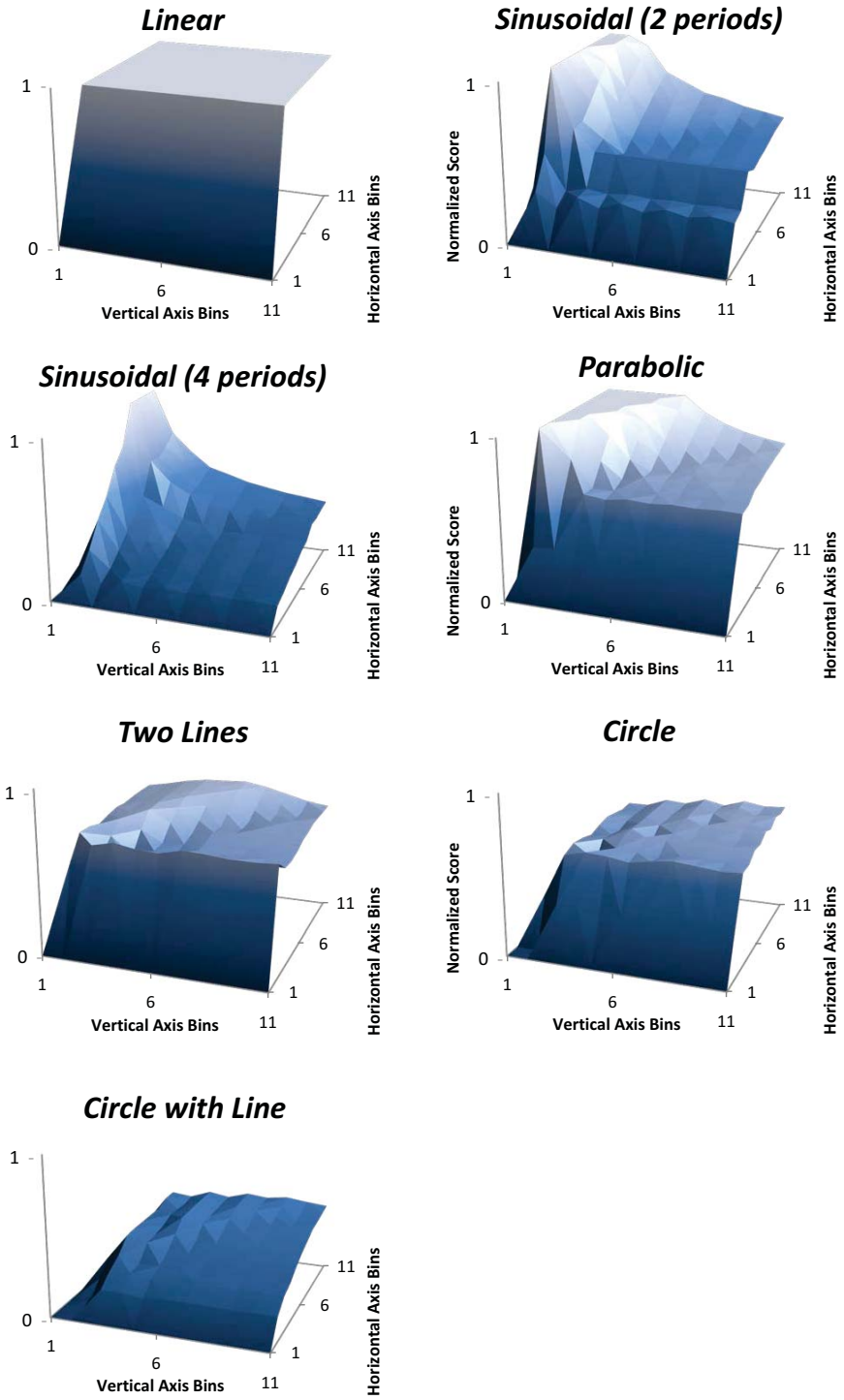


Figure S7: The surfaces derived from the characteristic matrices of the data sets presented in Table S1. Each surface's x- and y- axes represent the number of columns and rows respectively used to partition the two variables being analyzed, and the z-axis represents the normalized score of the data under those grid dimensions. The five different statistics presented in Table S1 are calculated from the characteristic matrices of these data. All relationships analyzed had a sample size of 1000.

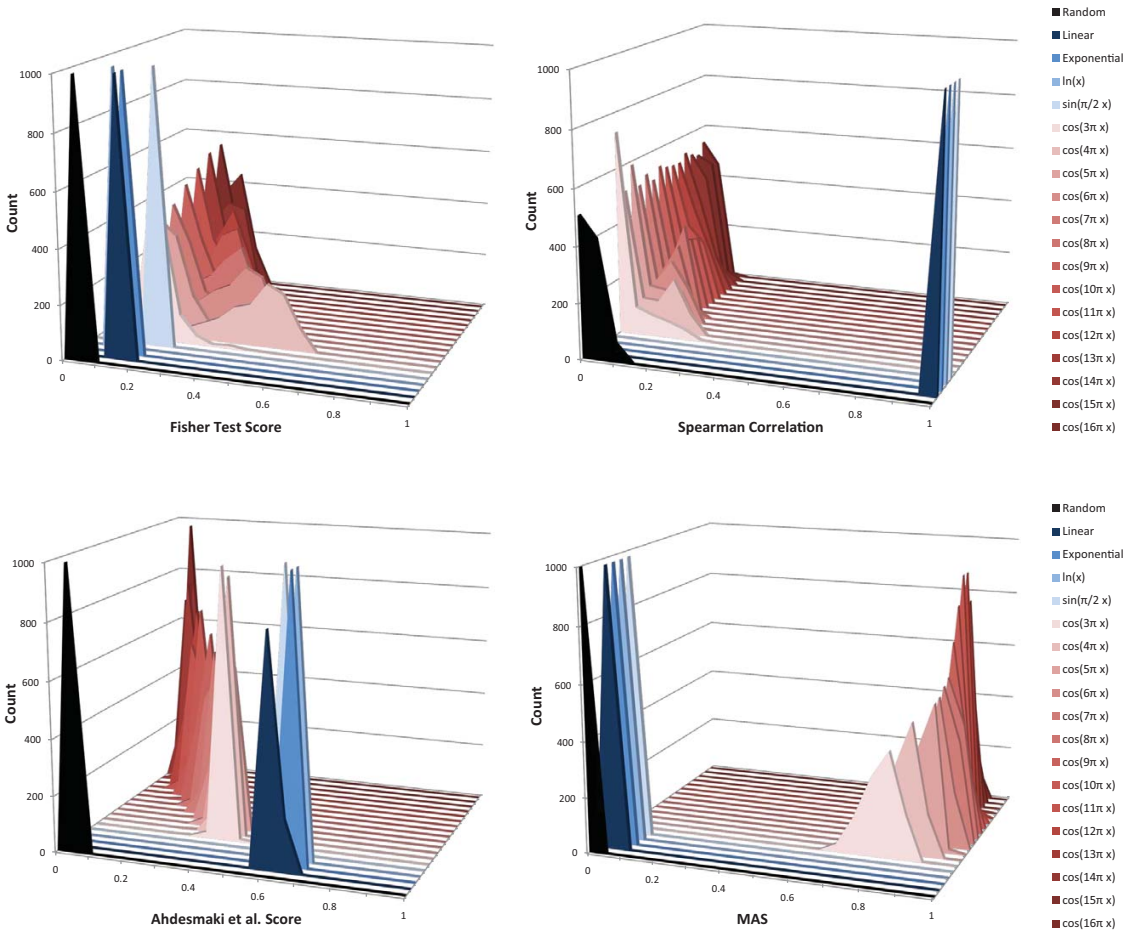
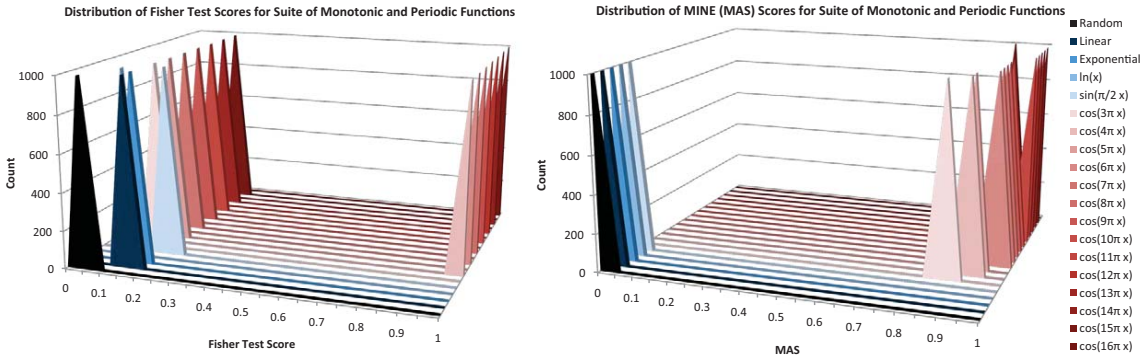
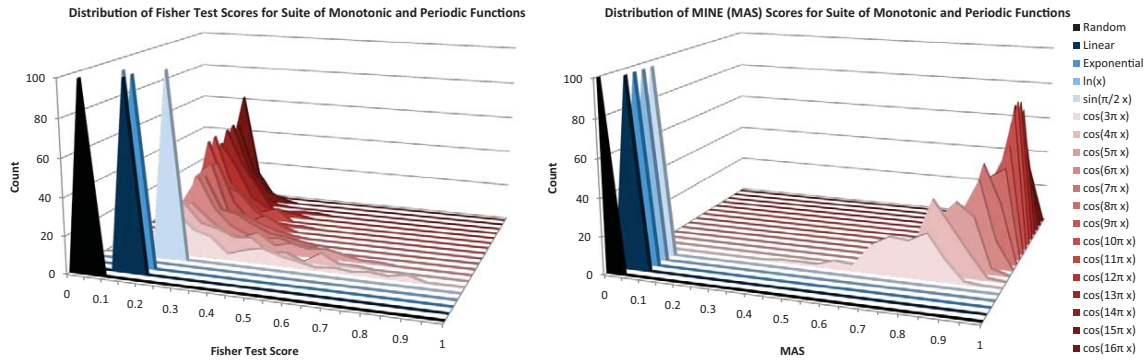


Figure S8: A comparison of MAS, Fisher test (Fourier analysis), Ahdesmaki et al. test, and Spearman correlation coefficient on a suite of monotonic and periodic functions. For each periodic function f in the legend, 1000 functions $f_1, \ldots, f_{1000}$ were generated with average wavelength equal to the wavelength of f (each period of each f_i has length $L + dL$ where L is the wavelength of f and d was chosen from a normal distribution with standard deviation 0.1.) 1000 points were chosen from each f_i, and the x- and y-values of each point were perturbed by a number chosen uniformly from $[-h, h]$, where h is a noise level set to 0.05. We did the same for each monotonic function shown, but with all of the f_i equal to f. We then calculated Fisher, Spearman, Ahdesmaki, and a combined MIC/MAS score (the score of data with MIC < 0.4 was 0; otherwise, the score was the MAS of the data) for each of the f_i and for 1000 random clouds, giving us a distribution of 1000 scores for each f and for each test. Each distribution is graphed as a histogram on the horizontal line corresponding to f, and the histograms are color-coded by function type (red = periodic functions, blue = non-periodic functions, and black = randomly generated data). This procedure is repeated, focusing only on Fisher and MAS scores and varying d (the period-length perturbation factor) and the noise level h. The results are shown in Figure S9.

(a)



(b)



(c)

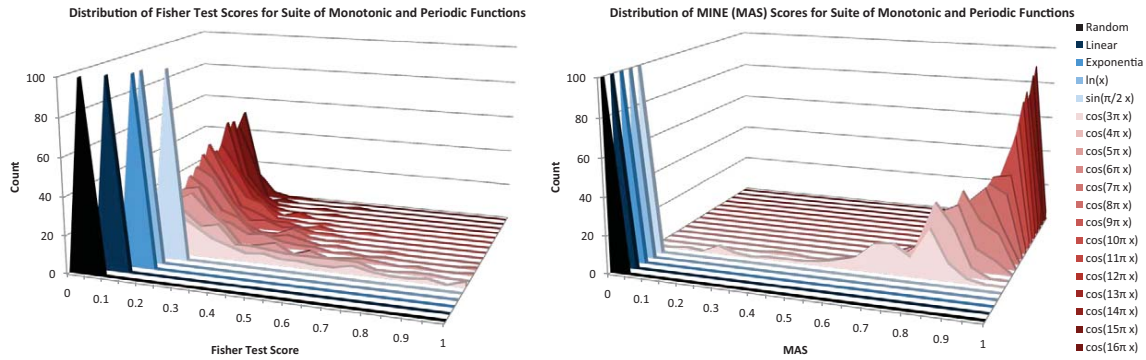


Figure S9: A comparison of MAS and Fourier analysis as in Figure S8 on a suite of noisy monotonic and periodic functions with varying period-length perturbation factors and noise levels: (a) period-length perturbation factor = 0, added noise level = 0; (b) period-length perturbation factor = 0.3, added noise level = 0.005; (c) period-length perturbation factor = 0.5, added noise level = 0.0.

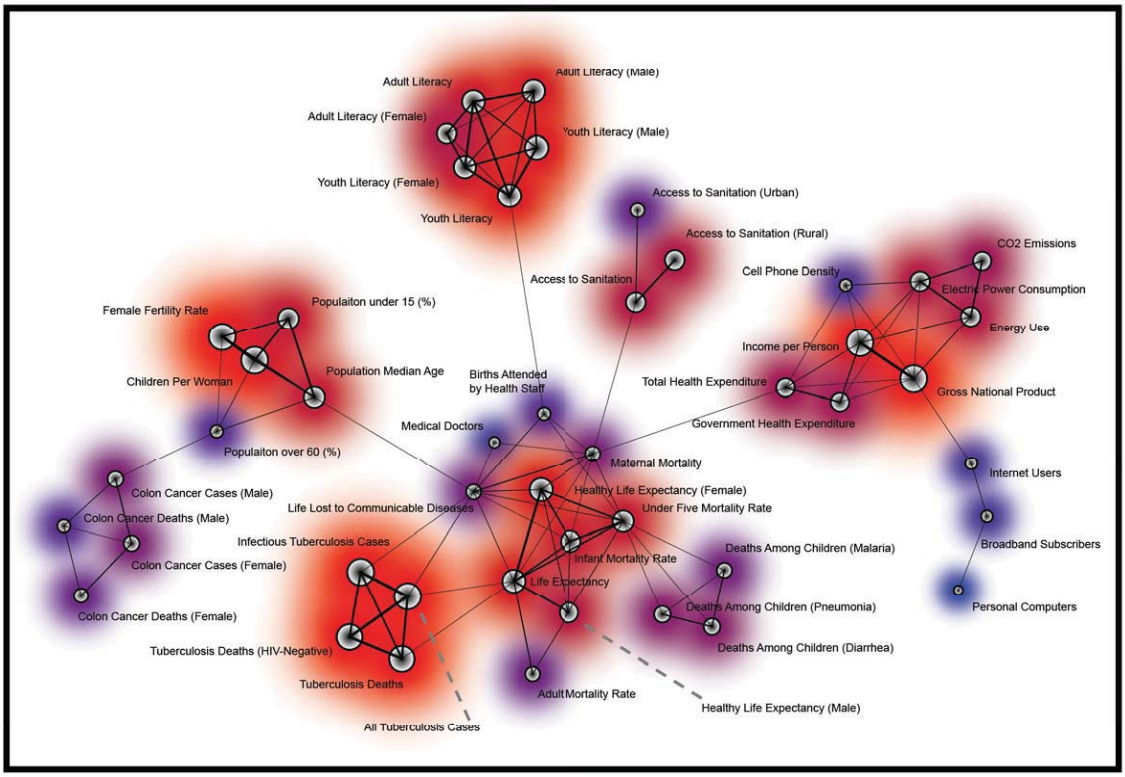
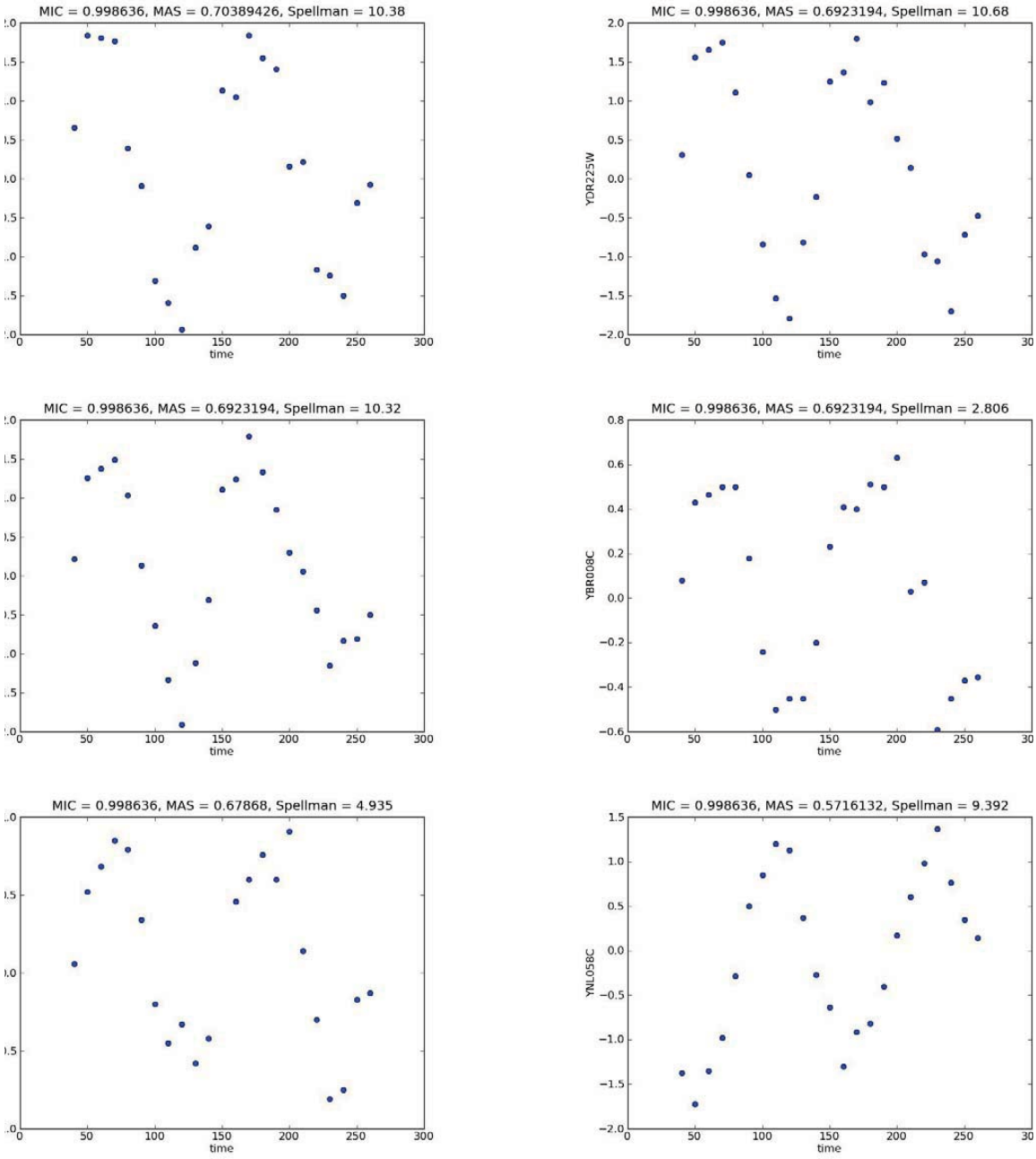


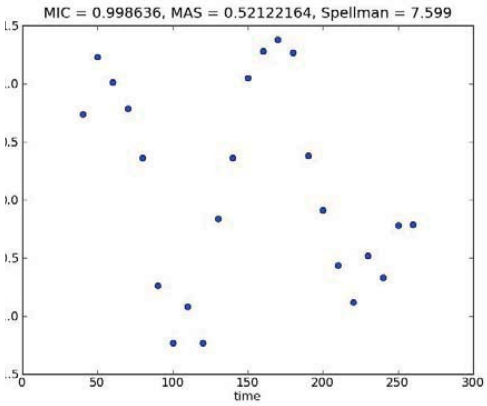
Figure S10: A screenshot of an interactive spring graph generated from the output of the MINE analysis of the global indicators dataset, reflecting the high-level structure of this dataset according to MINE. This figure is discussed in Section 4.10.

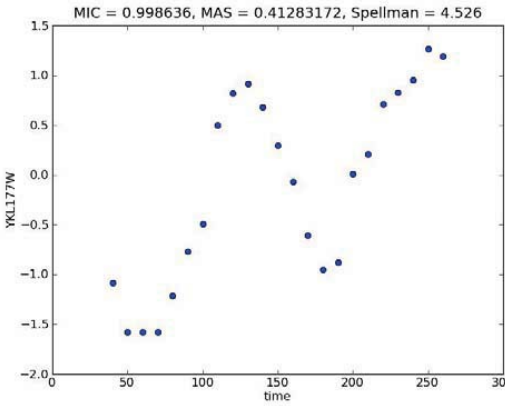
Figure S11: A sample of genes from Spellman et al. (1998) whose MICs were significant using a false discovery rate of 0.05, sorted by MAS. Each plot contains the given time series' MIC, MAS, and the score assigned to it by Spellman et al. Note that among these high-MIC genes, periodicity generally decreases as MAS decreases. (a) Six of the top scoring (MAS) genes. (b) Eight typical genes sampled from throughout the range of MAS scores, sorted by MAS (thus approximately in order of decreasing periodicity). (c) Six of the lowest scoring (MAS) genes.

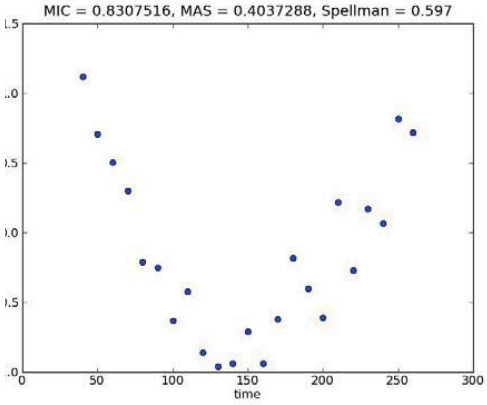
(a)

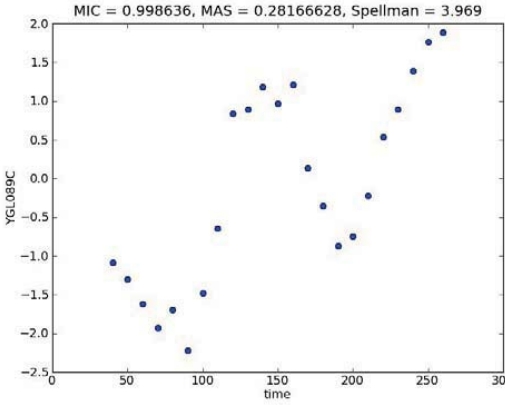


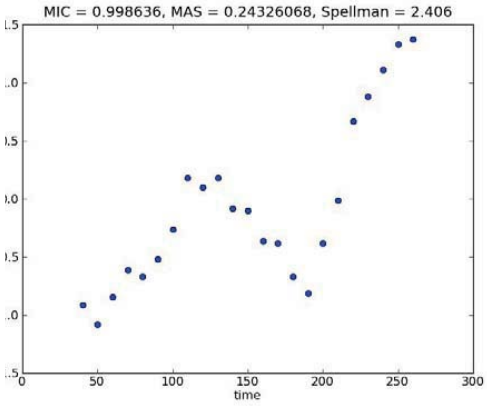
(b)

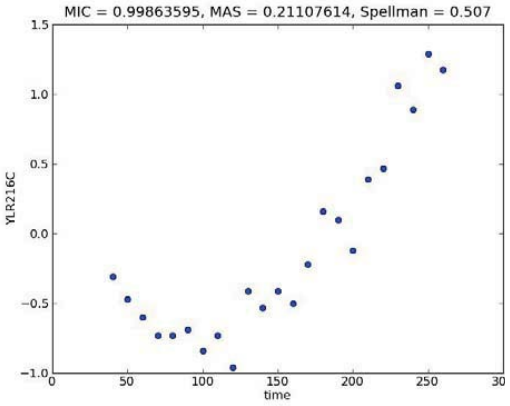

MIC = 0.998636, MAS = 0.52122164, Spellman = 7.599
time

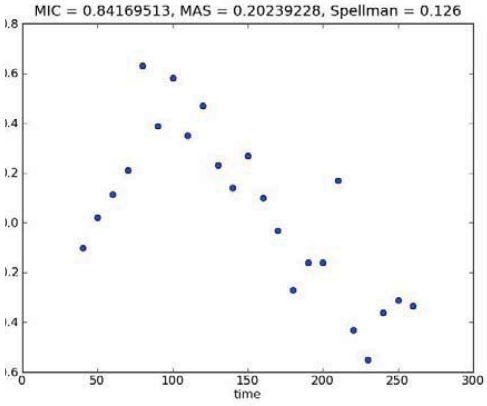

MIC = 0.998636, MAS = 0.41283172, Spellman = 4.526
YKL177W
time

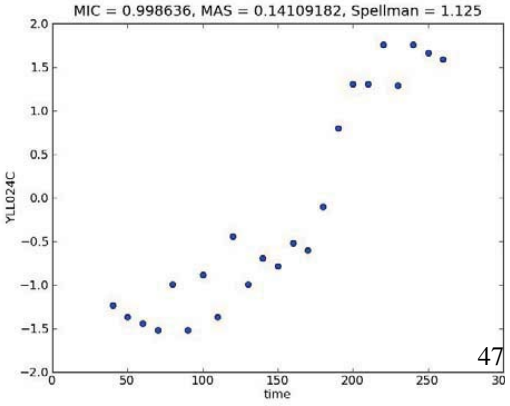

MIC = 0.8307516, MAS = 0.4037288, Spellman = 0.597
time


MIC = 0.998636, MAS = 0.28166628, Spellman = 3.969
YGL089C
time

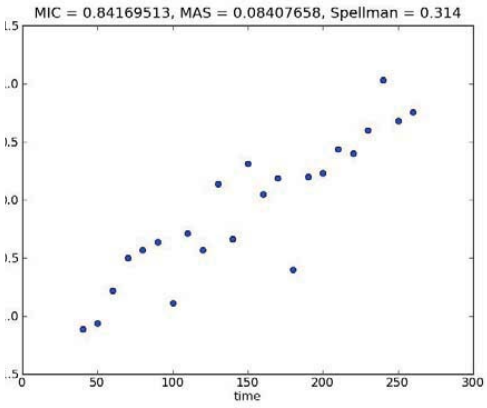

MIC = 0.998636, MAS = 0.24326068, Spellman = 2.406
time

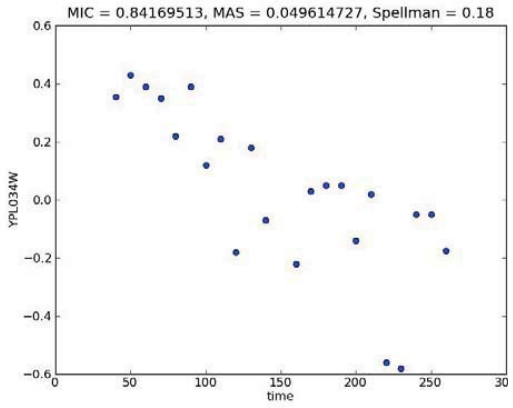

MIC = 0.99863595, MAS = 0.21107614, Spellman = 0.507
YLR216C
time

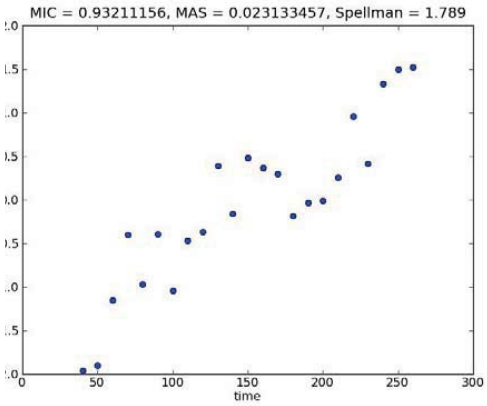

MIC = 0.84169513, MAS = 0.20239228, Spellman = 0.126
time

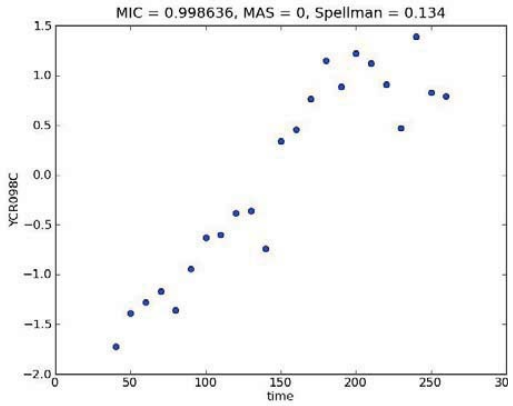

MIC = 0.998636, MAS = 0.14109182, Spellman = 1.125
YLL024C
time

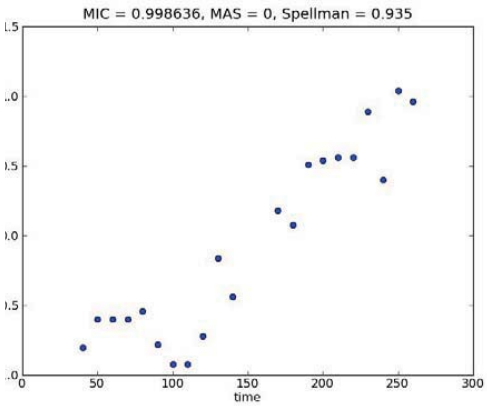
(c)

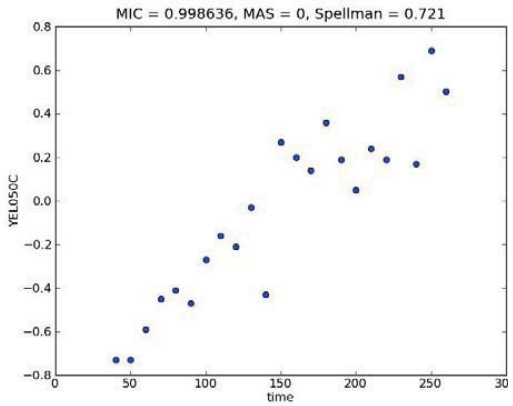

MIC = 0.84169513, MAS = 0.08407658, Spellman = 0.314
time


MIC = 0.84169513, MAS = 0.049614727, Spellman = 0.18
YPL034W
time


MIC = 0.93211156, MAS = 0.023133457, Spellman = 1.789
time


MIC = 0.998636, MAS = 0, Spellman = 0.134
YCR098C
time


MIC = 0.998636, MAS = 0, Spellman = 0.935
time


MIC = 0.998636, MAS = 0, Spellman = 0.721
YEL050C
time

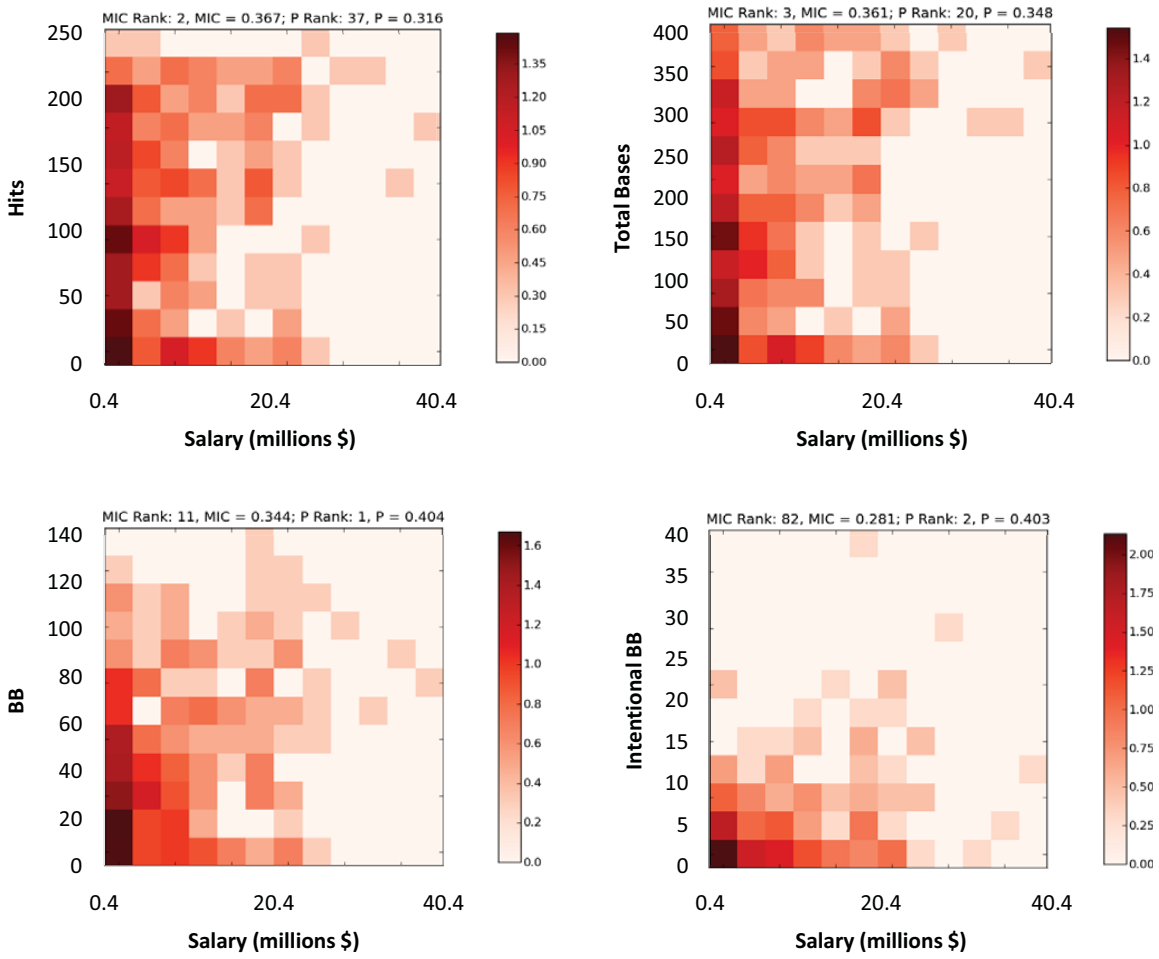


Figure S12: Histograms of joint distributions from several of the strongest associations with player salary according to MIC and ρ from the 2008 Major League Baseball season. Histograms are colored on a $\log_{10}$ scale. The plots of hits vs. salary and total bases vs. salary appear to be governed by joint distributions that are classified as weaker by ρ because they are non-linear.

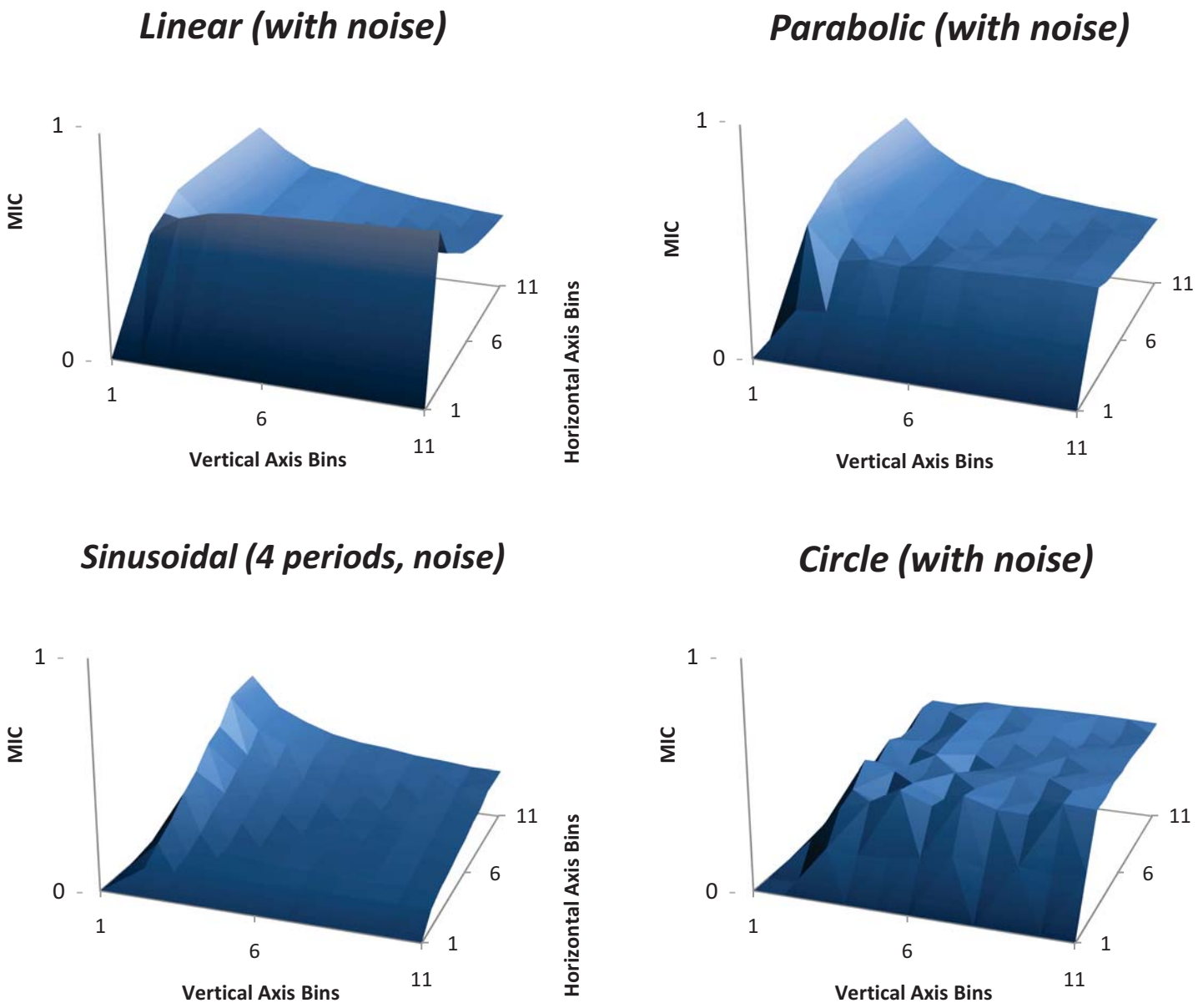


Figure S13: Illustrations of the characteristic matrix for noisy versions of some of the functions in Figure 3 from the main text, with $R^2 = .75$

Table S9: Top scoring 1% of relationships (by MIC) from a modified version of the global indicators dataset. This dataset included only 114 of the less redundant variables from original global indicators dataset. Each row includes the MIC score resulting from MINE analysis for a pair of variables, their Pearson product-moment correlation coefficient ρ, and the non-linearity score MIC$-\rho^2$. The ranks for each of the relationships are out of the 6,441 relationships in the reduced dataset.

Var1	Var2	MIC	MIC Rank	Pearson (ρ)	Pearson Rank	MIC - ρ^2	MIC - ρ^2 Rank
Children per woman	Total fertility rate (per woman)	1.000	1	0.999	1	0.001	5806
Energy use	Primary energy consumption per person	1.000	2	0.942	11	0.113	5501
Oil consumption per person	Income per person	1.000	3	0.733	207	0.463	173
Prevalence of tuberculosis (per 100 000 population)	Deaths from TB per 100 000 estimated	0.983	4	0.926	16	0.126	5429
CO2 emissions	Energy use	0.956	5	0.911	21	0.127	5425
Per capita total expenditure on health (PPP int. $)	Health expenditure per person	0.950	6	0.971	7	0.007	5793
Electric power consumption	Energy use	0.950	7	0.847	48	0.233	3192
Health expenditure per person	Income per person	0.943	8	0.827	63	0.260	2355
Number of laboratory health workers	External debt total DOD current US dollars	0.932	9	0.821	69	0.258	2412
Primary energy consumption per person	Personal computers per 100 people	0.931	10	0.342	1656	0.814	1
CO2 emissions	Electric power consumption	0.929	11	0.659	387	0.495	107
Women who have had mammography (%)	Women who have had PAP smear (%)	0.927	12	0.835	56	0.229	3339
Children per woman	Population median age (years)	0.924	13	-0.839	54	0.220	3645
Under five mortality from IHME	Life expectancy at birth	0.919	14	-0.877	33	0.149	5254
Women who have had mammography (%)	Medical Doctors	0.908	15	0.780	134	0.298	1463
Oil consumption per person	Life expectancy at birth	0.906	16	0.490	1000	0.667	5
Children per woman	Population living below the poverty line (% living on < US$1 per day)	0.903	17	0.712	252	0.396	413
Prevalence of HIV among adults aged >=15 years (per 100 000 population)	Deaths due to HIV/AIDS (per 100 000 population per year)	0.894	18	0.979	4	-0.065	5868
Under five mortality from IHME	Under-5 mortality rate (Probability of dying aged < 5 years per 1 000 live births) lowest wealth quintile	0.892	19	0.931	14	0.026	5765
Electric power consumption	Primary energy consumption per person	0.891	20	0.731	211	0.356	697
Inequality index	Income share held by lowest 20pct	0.884	21	-0.953	10	-0.024	5836
External debt total DOD current US dollars	Total income	0.882	22	0.852	43	0.156	5194
Energy use	Income per person	0.878	23	0.841	51	0.170	5003
Cervical cancer deaths per 100 000 women	Oil consumption per person	0.877	24	-0.396	1411	0.720	2
Under-5 mortality rate (Probability of dying aged < 5 years per 1 000 live births) lowest wealth quintile	Children per woman	0.868	25	0.848	46	0.149	5260
Per capita total expenditure on health (PPP int. $)	Income per person	0.867	26	0.825	64	0.186	4690
Electric power consumption	Income per person	0.860	27	0.849	44	0.139	5331
Maternal mortality ratio (per 100 000 live births)	Population living below the poverty line (% living on < US$1 per day)	0.848	28	0.744	188	0.294	1558
Number of laboratory health workers	Total income	0.847	29	0.980	3	-0.113	5884

Total fertility rate (per woman)	Population median age (years)	0.845	30	-0.841	52	0.138	5335
Women who have had mammography (%)	Income per person	0.843	31	0.807	85	0.193	4527
Per capita total expenditure on health (PPP int. $)	Cell phones per 100 people	0.843	32	0.729	216	0.311	1257
Registration coverage of births (%)	Maternal mortality ratio (per 100 000 live births)	0.840	33	-0.797	98	0.206	4143
Electric power consumption	Oil consumption per person	0.839	34	0.619	507	0.456	187
Registration coverage of births (%)	Prevalence of tuberculosis (per 100 000 population)	0.839	35	-0.632	467	0.439	238
Population living below the poverty line (% living on < US$1 per day)	Population with sustainable access to improved sanitation (%) total	0.838	36	-0.780	137	0.230	3309
Oil consumption per person	Health expenditure per person	0.836	37	0.373	1515	0.696	3
Births attended by skilled health staff	Maternal mortality ratio (per 100 000 live births)	0.834	38	-0.805	89	0.186	4686
Under-5 mortality rate (Probability of dying aged < 5 years per 1 000 live births) lowest wealth quintile	Total fertility rate (per woman)	0.834	39	0.848	47	0.116	5485
Under five mortality from IHME	Maternal mortality ratio (per 100 000 live births)	0.833	40	0.933	13	-0.038	5853
Under-5 mortality rate (Probability of dying aged < 5 years per 1 000 live births) lowest wealth quintile	Contraceptive prevalence (%)	0.831	41	-0.790	112	0.207	4114
Years of life lost to non-communicable diseases (%)	Maternal mortality ratio (per 100 000 live births)	0.829	42	-0.798	96	0.192	4546
Electric power consumption	Health expenditure per person	0.826	43	0.782	129	0.214	3865
Under five mortality from IHME	Registration coverage of births (%)	0.825	44	-0.805	88	0.177	4885
Women who have had mammography (%)	Health expenditure per person	0.823	45	0.748	183	0.263	2223
Women who have had mammography (%)	Internet users	0.823	46	0.795	101	0.191	4576
Years of life lost to non-communicable diseases (%)	Life expectancy at birth	0.821	47	0.890	27	0.029	5761
Registration coverage of births (%)	Births attended by skilled health staff	0.816	48	0.829	59	0.129	5404
Women who have had mammography (%)	Cell phones per 100 people	0.816	49	0.805	87	0.167	5042
Income per person	Cell phones per 100 people	0.814	50	0.782	131	0.202	4253
Life expectancy at birth	Deaths from TB per 100 000 estimated	0.812	51	-0.800	92	0.172	4983
Agriculture contribution to economy	Income per person	0.811	52	-0.620	505	0.427	284
Electric power consumption	Cell phones per 100 people	0.808	53	0.680	329	0.345	801
Number of dentistry personnel	Number of pharmaceutical personnel	0.806	54	0.554	752	0.499	105
Women who have had mammography (%)	Breast cancer new cases per 100 000 women	0.805	55	0.863	40	0.060	5697
Registration coverage of births (%)	Years of life lost to non-communicable diseases (%)	0.805	56	0.716	244	0.292	1592
Women who have had mammography (%)	Per capita total expenditure on health (PPP int. $)	0.803	57	0.823	67	0.125	5430
CO2 emissions	Income per person	0.803	58	0.719	233	0.287	1692
Women who have had PAP smear (%)	Medical Doctors	0.801	59	0.818	73	0.132	5379
Energy use	Maternal mortality ratio (per 100 000 live births)	0.801	60	-0.432	1240	0.615	13
Women who have had PAP smear (%)	Cell phones per 100 people	0.801	61	0.780	135	0.192	4556
External debt total DOD current US dollars	Patents in force	0.800	62	0.864	39	0.054	5710
Under-5 mortality rate (Probability of dying aged < 5 years per 1 000 live births) lowest wealth quintile	Population with sustainable access to improved drinking water sources (%) total	0.799	63	-0.711	256	0.294	1557

Country	Prevalence of Adult Females Who are Obese (%)	Income per Person (GDP/Capita, Inflation-Adjusted International $)
Tonga	74.9	5,135
Samoa	66.3	4,872
Cook Islands	65.7	9,000
Nauru	60.5	2,500
Egypt	46.6	5,049
Iraq	38.2	3,200
Fiji	26.4	4,209
Vanuatu	25.2	3,477

Table S10: Countries constituting the minority trend in the relationship between income per person (GDP/capita, inflation-adjusted $) and the prevalence of adult ($\geq$ 15 years old) female obesity (%) in the global indicators dataset (Figure 4F). These are the eight countries with the highest adult female obesity and an income per person of less than $10,000, and they appear to follow the steep linear trend rather than the parabolic trend. Of these eight countries, six are Pacific Island countries, where culturally, large physical size is considered a sign of beauty and status [33].

Country	Gross National Income per Capita (PPP International $)	Health Expenditure per Person (PPP International $)
Brunei Darussalam	49,900	519
Kuwait	48,310	687
Bahrain	34,310	710
United Arab Emirates	31,190	833
Saudi Arabia	22,300	448
Oman	19,740	312

Table S11: Countries leading the minority trend in the relationship between gross national income per capita (international dollars, using purchasing power parity) and health expenditure per person (international dollars, using purchasing power parity) in the global indicators dataset (Figure 4H). These are the six countries with the highest gross national income per capita and a health expenditure per person of less than $850, and they appear to follow the flat linear trend rather than the steep exponential trend. All six of these countries have economies that rely significantly on oil[37].

Var1	Var2	MIC	MIC Rank	Pearson (ρ)	Pearson Rank
Player Salary	RPMLV	0.369	1	0.357	14
Player Salary	H	0.367	2	0.316	37
Player Salary	TB	0.361	3	0.348	20
Player Salary	PA	0.360	4	0.324	31
Player Salary	BALLS	0.356	5	0.369	8
Player Salary	LD	0.354	6	0.308	40
Player Salary	PA%	0.350	7	0.323	32
Player Salary	TOB	0.349	8	0.368	9
Player Salary	FB	0.346	9	0.285	52
Player Salary	STRIKES	0.345	10	0.306	41
Player Salary	BB	0.344	11	0.404	1
Player Salary	PITCHES	0.344	12	0.335	26
Player Salary	UBB	0.343	13	0.371	6
Player Salary	Batted Balls	0.342	14	0.296	48
Player Salary	RPMLVr	0.341	15	0.323	33
Player Salary	SIT_DP	0.340	16	0.349	19
Player Salary	PA_PH	0.339	17	-0.267	59
Player Salary	G_PH	0.339	18	-0.266	60
Player Salary	AB	0.339	19	0.304	43
Player Salary	EqR	0.338	20	0.368	10
Player Salary	PMLV	0.336	21	0.296	47
Player Salary	OBI	0.335	22	0.363	11
Player Salary	RAR	0.334	23	0.377	4
Player Salary	OUT	0.331	24	0.293	49
Player Salary	R3	0.327	25	0.323	34
Player Salary	RBIR	0.327	26	0.282	55
Player Salary	MLVr	0.327	27	0.345	22
Player Salary	PA_ROB	0.326	28	0.356	15
Player Salary	ROB	0.326	29	0.351	18
Player Salary	MLV	0.326	30	0.355	16
Player Salary	VORPr	0.325	31	0.324	30
Player Salary	TBP	0.324	32	0.289	50
Player Salary	LEADOFF_PA	0.324	33	0.240	65
Player Salary	R2	0.323	34	0.354	17
Player Salary	RARP	0.322	35	0.334	27
Player Salary	VORP	0.322	36	0.359	13
Player Salary	SH	0.321	37	-0.227	68
Player Salary	DP	0.321	38	0.361	12
Player Salary	OUTS_EQ	0.319	39	0.288	51
Player Salary	EqA	0.318	40	0.310	39
Player Salary	R1	0.317	41	0.348	21
Player Salary	GB	0.317	42	0.260	62
Player Salary	R3_BI	0.317	43	0.315	38
Player Salary	SLG	0.316	44	0.329	29
Player Salary	RBI	0.314	45	0.382	3
Player Salary	1B	0.313	46	0.272	58
Player Salary	BBr	0.312	47	0.282	54
Player Salary	HR	0.311	48	0.370	7
Player Salary	GIDP	0.311	49	0.342	23
Player Salary	2B	0.310	50	0.279	56

Table S12: The 50 variables most closely related to player salary among 2008 Major League Baseball individual performance statistics, according to MIC. For baseball statistic glossary see: http://www.baseballprospectus.com/glossary/

OTU X	OTU Y	Family of OTU X	Family of OTU Y	MIC	Nonlinearity
OTU4435	OTU4496	Erysipelotrichaceae	Lachnospiraceae	0.506	0.506
OTU1462	OTU4496	Lachnospiraceae	Lachnospiraceae	0.464	0.464
OTU4496	OTU6224	Lachnospiraceae	Lachnospiraceae	0.438	0.433
OTU155	OTU4496	Ruminococcaceae	Lachnospiraceae	0.425	0.425
OTU4496	OTU5417	Lachnospiraceae	--	0.414	0.413
OTU675	OTU5937	Bacteroidaceae	Veillonellaceae	0.414	0.406
OTU2728	OTU4496	Lachnospiraceae	Lachnospiraceae	0.408	0.403
OTU5417	OTU5937	--	Veillonellaceae	0.408	0.380
OTU4273	OTU4496	Eubacteriaceae	Lachnospiraceae	0.374	0.371
OTU675	OTU6256	Bacteroidaceae	Lachnospiraceae	0.362	0.362
OTU1629	OTU6256	Lachnospiraceae	Lachnospiraceae	0.374	0.357
OTU2970	OTU4496	Ruminococcaceae	Lachnospiraceae	0.358	0.356
OTU4273	OTU5937	Eubacteriaceae	Veillonellaceae	0.366	0.354
OTU710	OTU4496	Erysipelotrichaceae	Lachnospiraceae	0.354	0.349
OTU4257	OTU6256	Rikenellaceae	Lachnospiraceae	0.346	0.344
OTU4257	OTU4496	Rikenellaceae	Lachnospiraceae	0.339	0.334
OTU1193	OTU4496	Ruminococcaceae	Lachnospiraceae	0.336	0.334
OTU2642	OTU2728	Lachnospiraceae	Lachnospiraceae	0.385	0.329
OTU1373	OTU4496	--	Lachnospiraceae	0.322	0.320
OTU4273	OTU6256	Eubacteriaceae	Lachnospiraceae	0.329	0.320
OTU1462	OTU3991	Lachnospiraceae	Erysipelotrichaceae	0.326	0.319
OTU2941	OTU4496	Lachnospiraceae	Lachnospiraceae	0.332	0.317
OTU2728	OTU5490	Lachnospiraceae	Lachnospiraceae	0.353	0.316
OTU4496	OTU5420	Lachnospiraceae	Bacteroidaceae	0.312	0.312
OTU1629	OTU4496	Lachnospiraceae	Lachnospiraceae	0.305	0.305
OTU5937	OTU6224	Veillonellaceae	Lachnospiraceae	0.301	0.299
OTU2350	OTU4496	Rikenellaceae	Lachnospiraceae	0.302	0.295
OTU1347	OTU6256	Bacteroidaceae	Lachnospiraceae	0.329	0.294
OTU675	OTU4496	Bacteroidaceae	Lachnospiraceae	0.296	0.293
OTU4496	OTU5117	Lachnospiraceae	Lachnospiraceae	0.291	0.291
OTU1285	OTU6256	Desulfovibrionaceae	Lachnospiraceae	0.287	0.286
OTU1347	OTU4496	Bacteroidaceae	Lachnospiraceae	0.291	0.285
OTU4496	OTU5370	Lachnospiraceae	Lachnospiraceae	0.291	0.283
OTU3994	OTU5937	Bacteroidaceae	Veillonellaceae	0.285	0.283
OTU3994	OTU4496	Bacteroidaceae	Lachnospiraceae	0.294	0.282
OTU4257	OTU5937	Rikenellaceae	Veillonellaceae	0.379	0.281
OTU710	OTU5937	Erysipelotrichaceae	Veillonellaceae	0.296	0.277
OTU1373	OTU5937	--	Veillonellaceae	0.281	0.277
OTU1548	OTU6256	Ruminococcaceae	Lachnospiraceae	0.277	0.276
OTU453	OTU4496	Enterococcaceae	Lachnospiraceae	0.282	0.276
OTU1347	OTU5937	Bacteroidaceae	Veillonellaceae	0.300	0.275
OTU2728	OTU5937	Lachnospiraceae	Veillonellaceae	0.276	0.270
OTU2970	OTU6256	Ruminococcaceae	Lachnospiraceae	0.289	0.268
OTU1629	OTU3991	Lachnospiraceae	Erysipelotrichaceae	0.292	0.267
OTU4865	OTU5937	--	Veillonellaceae	0.321	0.264
OTU3991	OTU4273	Erysipelotrichaceae	Eubacteriaceae	0.266	0.262
OTU2350	OTU6256	Rikenellaceae	Lachnospiraceae	0.272	0.261
OTU2941	OTU6256	Lachnospiraceae	Lachnospiraceae	0.280	0.259
OTU5420	OTU5937	Bacteroidaceae	Veillonellaceae	0.432	0.259
OTU4496	OTU4501	Lachnospiraceae	Ruminococcaceae	0.259	0.258
OTU1285	OTU4496	Desulfovibrionaceae	Lachnospiraceae	0.256	0.256
OTU5407	OTU5420	Porphyromonadaceae	Bacteroidaceae	0.272	0.255
OTU3991	OTU4435	Erysipelotrichaceae	Erysipelotrichaceae	0.251	0.251
OTU5826	OTU6256	Ruminococcaceae	Lachnospiraceae	0.251	0.251
OTU1285	OTU5937	Desulfovibrionaceae	Veillonellaceae	0.296	0.250
OTU4496	OTU5826	Lachnospiraceae	Ruminococcaceae	0.255	0.250
OTU4865	OTU6256	--	Lachnospiraceae	0.249	0.249
OTU6256	OTU6484	Lachnospiraceae	Bacteroidaceae	0.247	0.247
OTU4496	OTU6484	Lachnospiraceae	Bacteroidaceae	0.247	0.246
OTU1629	OTU5937	Lachnospiraceae	Veillonellaceae	0.266	0.245
OTU4865	OTU5407	--	Porphyromonadaceae	0.245	0.244
OTU4496	OTU4865	Lachnospiraceae	--	0.252	0.244
OTU2399	OTU2728	Lachnospiraceae	Lachnospiraceae	0.244	0.244
OTU2399	OTU5420	Lachnospiraceae	Bacteroidaceae	0.250	0.243
OTU2516	OTU4496	Lachnospiraceae	Lachnospiraceae	0.253	0.242
OTU5117	OTU6256	Lachnospiraceae	Lachnospiraceae	0.243	0.241
OTU5826	OTU5937	Ruminococcaceae	Veillonellaceae	0.264	0.240
OTU4435	OTU5937	Erysipelotrichaceae	Veillonellaceae	0.244	0.239
OTU2728	OTU5407	Lachnospiraceae	Porphyromonadaceae	0.239	0.238
OTU2036	OTU6256	--	Lachnospiraceae	0.236	0.235
OTU774	OTU1347	Ruminococcaceae	Bacteroidaceae	0.236	0.235
OTU2970	OTU3991	Ruminococcaceae	Erysipelotrichaceae	0.236	0.235
OTU3991	OTU5370	Erysipelotrichaceae	Lachnospiraceae	0.235	0.235
OTU155	OTU6256	Ruminococcaceae	Lachnospiraceae	0.251	0.232
OTU4501	OTU6256	Ruminococcaceae	Lachnospiraceae	0.232	0.232
OTU453	OTU3991	Enterococcaceae	Erysipelotrichaceae	0.241	0.232
OTU1548	OTU4496	Ruminococcaceae	Lachnospiraceae	0.234	0.231

Table S13: Non-coexistence relationships in the microbiome dataset explained by diet; under one diet OTU X dominates, while under another diet, OTU Y dominates. The relationships in this table are sorted by non-linearity (MIC $-\ \rho^2$) and not by MIC.

Table S14: Relationships unaffected by any of the auxiliary variables recorded in the dataset of [29] (see Section 4.8).

OTU X	OTU Y	Family of OTU X	Family of OTU Y	MIC	Nonlinearity
OTU1177	OTU4154	Lachnospiraceae	Prevotellaceae	0.455	0.439
OTU2728	OTU3349	Lachnospiraceae	Porphyromonadaceae	0.461	0.432
OTU453	OTU1373	Enterococcaceae	--	0.461	0.429
OTU1100	OTU3350	Lachnospiraceae	Lachnospiraceae	0.479	0.427
OTU5417	OTU5948	--	Bacteroidaceae	0.441	0.422
OTU2728	OTU5499	Lachnospiraceae	Lachnospiraceae	0.445	0.415
OTU453	OTU5691	Enterococcaceae	Enterococcaceae	0.546	0.376
OTU453	OTU2970	Enterococcaceae	Ruminococcaceae	0.498	0.374
OTU3732	OTU5499	Lachnospiraceae	Lachnospiraceae	0.391	0.368
OTU2728	OTU5319	Lachnospiraceae	Lachnospiraceae	0.428	0.368
OTU5417	OTU5429	--	Bacteroidaceae	0.368	0.367
OTU3855	OTU4154	--	Prevotellaceae	0.406	0.363
OTU1901	OTU3102	Erysipelotrichaceae	Lachnospiraceae	0.358	0.357
OTU2399	OTU3406	Lachnospiraceae	Bacteroidaceae	0.386	0.352
OTU1193	OTU1373	Ruminococcaceae	--	0.512	0.350
OTU5417	OTU6256	--	Lachnospiraceae	0.346	0.346
OTU1373	OTU2970	--	Ruminococcaceae	0.545	0.341
OTU6224	OTU6256	Lachnospiraceae	Lachnospiraceae	0.348	0.340
OTU708	OTU2171	Lachnospiraceae	Lachnospiraceae	0.340	0.339
OTU1901	OTU6372	Erysipelotrichaceae	Porphyromonadaceae	0.342	0.335
OTU254	OTU1779	Erysipelotrichaceae	Lachnospiraceae	0.332	0.332
OTU1373	OTU5691	--	Enterococcaceae	0.358	0.329
OTU1812	OTU2728	Lachnospiraceae	Lachnospiraceae	0.328	0.326
OTU708	OTU5499	Lachnospiraceae	Lachnospiraceae	0.363	0.325
OTU3406	OTU6224	Bacteroidaceae	Lachnospiraceae	0.371	0.324
OTU675	OTU5948	Bacteroidaceae	Bacteroidaceae	0.321	0.320
OTU1779	OTU3711	Lachnospiraceae	Lachnospiraceae	0.322	0.319
OTU4257	OTU5948	Rikenellaceae	Bacteroidaceae	0.327	0.317
OTU3406	OTU4273	Bacteroidaceae	Eubacteriaceae	0.338	0.315
OTU2211	OTU4154	Prevotellaceae	Prevotellaceae	0.382	0.314
OTU453	OTU1193	Enterococcaceae	Ruminococcaceae	0.536	0.312
OTU2839	OTU4486	Verrucomicrobiaceae	Coriobacteriales	0.320	0.309
OTU2079	OTU2728	Porphyromonadaceae	Lachnospiraceae	0.318	0.306
OTU710	OTU6256	Erysipelotrichaceae	Lachnospiraceae	0.324	0.304
OTU2371	OTU6372	Lachnospiraceae	Porphyromonadaceae	0.328	0.303
OTU3708	OTU4435	Lachnospiraceae	Erysipelotrichaceae	0.302	0.302
OTU2516	OTU6256	Lachnospiraceae	Lachnospiraceae	0.326	0.302
OTU3994	OTU6256	Bacteroidaceae	Lachnospiraceae	0.304	0.300
OTU3349	OTU5499	Porphyromonadaceae	Lachnospiraceae	0.588	0.299
OTU708	OTU4852	Lachnospiraceae	Porphyromonadaceae	0.359	0.298
OTU1177	OTU6256	Lachnospiraceae	Lachnospiraceae	0.317	0.298
OTU4273	OTU5948	Eubacteriaceae	Bacteroidaceae	0.342	0.295
OTU3263	OTU6256	Lachnospiraceae	Lachnospiraceae	0.301	0.295
OTU1373	OTU6256	--	Lachnospiraceae	0.311	0.295
OTU618	OTU5499	Erysipelotrichaceae	Lachnospiraceae	0.305	0.295
OTU5948	OTU6484	Bacteroidaceae	Bacteroidaceae	0.297	0.294
OTU5420	OTU5948	Bacteroidaceae	Bacteroidaceae	0.341	0.294
OTU3520	OTU3991	Erysipelotrichaceae	Erysipelotrichaceae	0.304	0.294
OTU3406	OTU6256	Bacteroidaceae	Lachnospiraceae	0.329	0.292
OTU2211	OTU3592	Prevotellaceae	--	0.360	0.290
OTU1150	OTU5948	--	Bacteroidaceae	0.289	0.288
OTU1484	OTU4380	--	Ruminococcaceae	0.401	0.288
OTU556	OTU4512	Lachnospiraceae	Coriobacteriales	0.286	0.285
OTU4154	OTU6256	Prevotellaceae	Lachnospiraceae	0.326	0.284
OTU1462	OTU6256	Lachnospiraceae	Lachnospiraceae	0.286	0.284
OTU5319	OTU5417	Lachnospiraceae	--	0.281	0.280
OTU4865	OTU5948	--	Bacteroidaceae	0.279	0.279

OTU2320	OTU5499	Lachnospiraceae	Lachnospiraceae	0.294	0.279
OTU3102	OTU6372	Lachnospiraceae	Porphyromonadaceae	0.430	0.279
OTU4083	OTU5948	Streptococcaceae	Bacteroidaceae	0.301	0.278
OTU1177	OTU1541	Lachnospiraceae	Lachnospiraceae	0.278	0.278
OTU5429	OTU6224	Bacteroidaceae	Lachnospiraceae	0.300	0.278
OTU453	OTU4435	Enterococcaceae	Erysipelotrichaceae	0.367	0.278
OTU155	OTU1373	Ruminococcaceae	--	0.393	0.276
OTU2371	OTU3102	Lachnospiraceae	Lachnospiraceae	0.345	0.275
OTU1347	OTU4154	Bacteroidaceae	Prevotellaceae	0.284	0.274
OTU1044	OTU3406	Lachnospiraceae	Bacteroidaceae	0.385	0.273
OTU5319	OTU6256	Lachnospiraceae	Lachnospiraceae	0.282	0.272
OTU5948	OTU6224	Bacteroidaceae	Lachnospiraceae	0.334	0.272
OTU1779	OTU4822	Lachnospiraceae	Lachnospiraceae	0.272	0.271
OTU3592	OTU3855	--	--	0.311	0.271
OTU2839	OTU4273	Verrucomicrobiaceae	Eubacteriaceae	0.274	0.271
OTU770	OTU6256	Lachnospiraceae	Lachnospiraceae	0.292	0.271
OTU254	OTU4473	Erysipelotrichaceae	Bacteroidaceae	0.289	0.270
OTU155	OTU1333	Ruminococcaceae	Clostridiaceae	0.293	0.270
OTU5948	OTU6256	Bacteroidaceae	Lachnospiraceae	0.297	0.269
OTU708	OTU1629	Lachnospiraceae	Lachnospiraceae	0.271	0.268
OTU770	OTU1098	Lachnospiraceae	Lachnospiraceae	0.299	0.267
OTU1177	OTU1779	Lachnospiraceae	Lachnospiraceae	0.266	0.266
OTU3406	OTU5948	Bacteroidaceae	Bacteroidaceae	0.681	0.266
OTU2970	OTU5691	Ruminococcaceae	Enterococcaceae	0.384	0.265
OTU1629	OTU3406	Lachnospiraceae	Bacteroidaceae	0.294	0.265
OTU675	OTU1062	Bacteroidaceae	Lactobacillaceae	0.271	0.265
OTU710	OTU774	Erysipelotrichaceae	Ruminococcaceae	0.320	0.264
OTU4273	OTU5429	Eubacteriaceae	Bacteroidaceae	0.286	0.264
OTU708	OTU4154	Lachnospiraceae	Prevotellaceae	0.296	0.264
OTU708	OTU6372	Lachnospiraceae	Porphyromonadaceae	0.265	0.263
OTU1347	OTU5948	Bacteroidaceae	Bacteroidaceae	0.265	0.263
OTU3263	OTU5370	Lachnospiraceae	Lachnospiraceae	0.376	0.262
OTU3994	OTU5429	Bacteroidaceae	Bacteroidaceae	0.266	0.262
OTU453	OTU5948	Enterococcaceae	Bacteroidaceae	0.333	0.261
OTU4273	OTU4435	Eubacteriaceae	Erysipelotrichaceae	0.365	0.261
OTU453	OTU2516	Enterococcaceae	Lachnospiraceae	0.374	0.261
OTU710	OTU3263	Erysipelotrichaceae	Lachnospiraceae	0.267	0.260
OTU1100	OTU1901	Lachnospiraceae	Erysipelotrichaceae	0.262	0.260
OTU774	OTU2399	Ruminococcaceae	Lachnospiraceae	0.345	0.259
OTU345	OTU1062	--	Lactobacillaceae	0.276	0.259
OTU453	OTU2350	Enterococcaceae	Rikenellaceae	0.259	0.259
OTU708	OTU4473	Lachnospiraceae	Bacteroidaceae	0.263	0.258
OTU1462	OTU3708	Lachnospiraceae	Lachnospiraceae	0.297	0.257
OTU1901	OTU5063	Erysipelotrichaceae	Lachnospiraceae	0.257	0.256
OTU2350	OTU6224	Rikenellaceae	Lachnospiraceae	0.285	0.256
OTU708	OTU1177	Lachnospiraceae	Lachnospiraceae	0.257	0.256
OTU710	OTU5691	Erysipelotrichaceae	Enterococcaceae	0.298	0.256
OTU4154	OTU4810	Prevotellaceae	Prevotellaceae	0.378	0.255
OTU2350	OTU2970	Rikenellaceae	Ruminococcaceae	0.288	0.255
OTU4877	OTU5319	--	Lachnospiraceae	0.302	0.255
OTU1333	OTU5117	Clostridiaceae	Lachnospiraceae	0.271	0.255
OTU708	OTU2941	Lachnospiraceae	Lachnospiraceae	0.254	0.254
OTU3855	OTU4380	--	Ruminococcaceae	0.257	0.254
OTU2399	OTU5948	Lachnospiraceae	Bacteroidaceae	0.302	0.253
OTU155	OTU2350	Ruminococcaceae	Rikenellaceae	0.272	0.253
OTU1373	OTU3406	--	Bacteroidaceae	0.322	0.253
OTU4435	OTU5948	Erysipelotrichaceae	Bacteroidaceae	0.323	0.252
OTU2941	OTU3991	Lachnospiraceae	Erysipelotrichaceae	0.280	0.252
OTU710	OTU770	Erysipelotrichaceae	Lachnospiraceae	0.271	0.252

OTU1062	OTU1098	Lactobacillaceae	Lachnospiraceae	0.253	0.248
OTU254	OTU4969	Erysipelotrichaceae	Erysipelotrichaceae	0.251	0.248
OTU1779	OTU4473	Lachnospiraceae	Bacteroidaceae	0.282	0.247
OTU2400	OTU5319	Lachnospiraceae	Lachnospiraceae	0.278	0.247
OTU170	OTU708	--	Lachnospiraceae	0.249	0.247
OTU1779	OTU4154	Lachnospiraceae	Prevotellaceae	0.247	0.246
OTU4154	OTU4837	Prevotellaceae	Ruminococcaceae	0.261	0.246
OTU2728	OTU6256	Lachnospiraceae	Lachnospiraceae	0.252	0.246
OTU3406	OTU4257	Bacteroidaceae	Rikenellaceae	0.259	0.245
OTU1193	OTU3406	Ruminococcaceae	Bacteroidaceae	0.335	0.244
OTU1462	OTU5319	Lachnospiraceae	Lachnospiraceae	0.254	0.244
OTU1373	OTU5370	--	Lachnospiraceae	0.425	0.244
OTU774	OTU3994	Ruminococcaceae	Bacteroidaceae	0.255	0.244
OTU1462	OTU5862	Lachnospiraceae	Lachnospiraceae	0.267	0.244
OTU5319	OTU5429	Lachnospiraceae	Bacteroidaceae	0.316	0.243
OTU1373	OTU4435	--	Erysipelotrichaceae	0.348	0.243
OTU4435	OTU6256	Erysipelotrichaceae	Lachnospiraceae	0.261	0.243
OTU3406	OTU4083	Bacteroidaceae	Streptococcaceae	0.256	0.242
OTU1177	OTU5499	Lachnospiraceae	Lachnospiraceae	0.246	0.242
OTU4435	OTU5429	Erysipelotrichaceae	Bacteroidaceae	0.268	0.242
OTU708	OTU3994	Lachnospiraceae	Bacteroidaceae	0.243	0.242
OTU708	OTU1779	Lachnospiraceae	Lachnospiraceae	0.241	0.241
OTU708	OTU5948	Lachnospiraceae	Bacteroidaceae	0.243	0.241
OTU1373	OTU5429	--	Bacteroidaceae	0.274	0.241
OTU2516	OTU3406	Lachnospiraceae	Bacteroidaceae	0.303	0.241
OTU149	OTU5499	Lachnospiraceae	Lachnospiraceae	0.248	0.241
OTU5830	OTU6486	Lachnospiraceae	Helicobacteraceae	0.246	0.241
OTU4154	OTU5499	Prevotellaceae	Lachnospiraceae	0.246	0.240
OTU4154	OTU4512	Prevotellaceae	Coriobacteriales	0.242	0.240
OTU1629	OTU5948	Lachnospiraceae	Bacteroidaceae	0.270	0.239
OTU4154	OTU5407	Prevotellaceae	Porphyromonadaceae	0.243	0.239
OTU2839	OTU5429	Verrucomicrobiaceae	Bacteroidaceae	0.241	0.239
OTU675	OTU2728	Bacteroidaceae	Lachnospiraceae	0.239	0.239
OTU1062	OTU3895	Lactobacillaceae	--	0.262	0.239
OTU5429	OTU6256	Bacteroidaceae	Lachnospiraceae	0.305	0.238
OTU1098	OTU2245	Lachnospiraceae	Ruminococcaceae	0.261	0.238
OTU1098	OTU3110	Lachnospiraceae	--	0.257	0.238
OTU5499	OTU6325	Lachnospiraceae	Lachnospiraceae	0.319	0.237
OTU708	OTU3102	Lachnospiraceae	Lachnospiraceae	0.242	0.237
OTU1373	OTU5948	--	Bacteroidaceae	0.316	0.237
OTU1541	OTU1779	Lachnospiraceae	Lachnospiraceae	0.237	0.237
OTU3711	OTU4473	Lachnospiraceae	Bacteroidaceae	0.245	0.236
OTU453	OTU5370	Enterococcaceae	Lachnospiraceae	0.436	0.236
OTU3406	OTU6484	Bacteroidaceae	Bacteroidaceae	0.244	0.236
OTU1901	OTU4435	Erysipelotrichaceae	Erysipelotrichaceae	0.241	0.236
OTU3406	OTU4154	Bacteroidaceae	Prevotellaceae	0.236	0.236
OTU254	OTU336	Erysipelotrichaceae	Ruminococcaceae	0.252	0.235
OTU774	OTU3406	Ruminococcaceae	Bacteroidaceae	0.414	0.235
OTU1177	OTU3711	Lachnospiraceae	Lachnospiraceae	0.235	0.235
OTU710	OTU3110	Erysipelotrichaceae	--	0.247	0.235
OTU336	OTU1779	Ruminococcaceae	Lachnospiraceae	0.234	0.234
OTU3349	OTU5348	Porphyromonadaceae	Ruminococcaceae	0.244	0.234
OTU1062	OTU5763	Lactobacillaceae	--	0.238	0.234
OTU3406	OTU5319	Bacteroidaceae	Lachnospiraceae	0.327	0.233
OTU4154	OTU4380	Prevotellaceae	Ruminococcaceae	0.434	0.233
OTU2728	OTU5673	Lachnospiraceae	Lachnospiraceae	0.238	0.232
OTU1062	OTU3046	Lactobacillaceae	Lachnospiraceae	0.244	0.232
OTU1779	OTU5499	Lachnospiraceae	Lachnospiraceae	0.239	0.232
OTU336	OTU3711	Ruminococcaceae	Lachnospiraceae	0.261	0.232