Information Theoretic Approaches to Social Computing

Aram Galstyan
University of Southern California
Information Sciences Institute

joint work with


Greg Ver Steeg


Shuyang Gao

USC VITERBI SCHOOL OF ENGINEERING

Multiscale Data-driven Models Workshop
IPAM, UCLA
May 9, 2016

ISI Information Sciences Institute

ISI
Information Sciences Institute

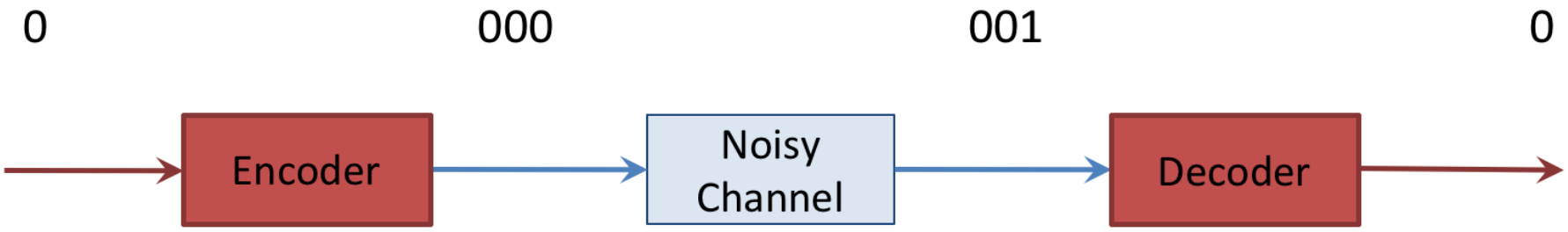
Reprinted with corrections from *The Bell System Technical Journal*,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

A Mathematical Theory of Communication

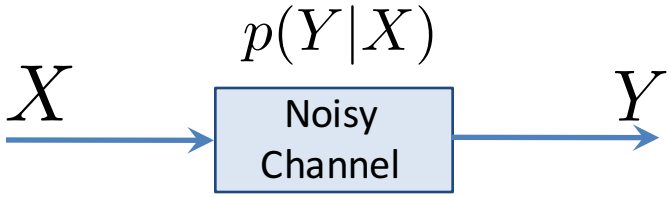
By C. E. SHANNON

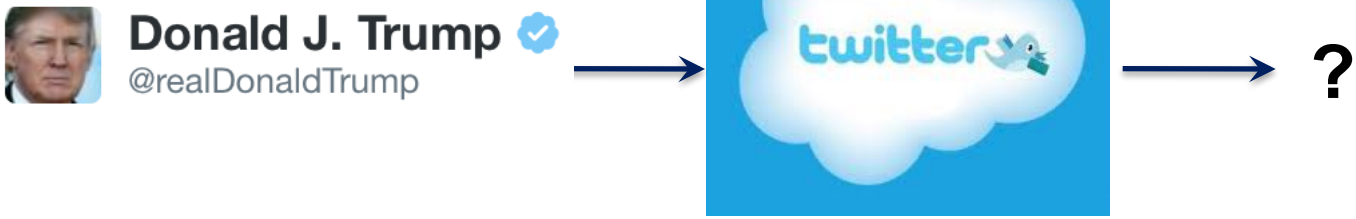
INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.



Noisy channel paradigm





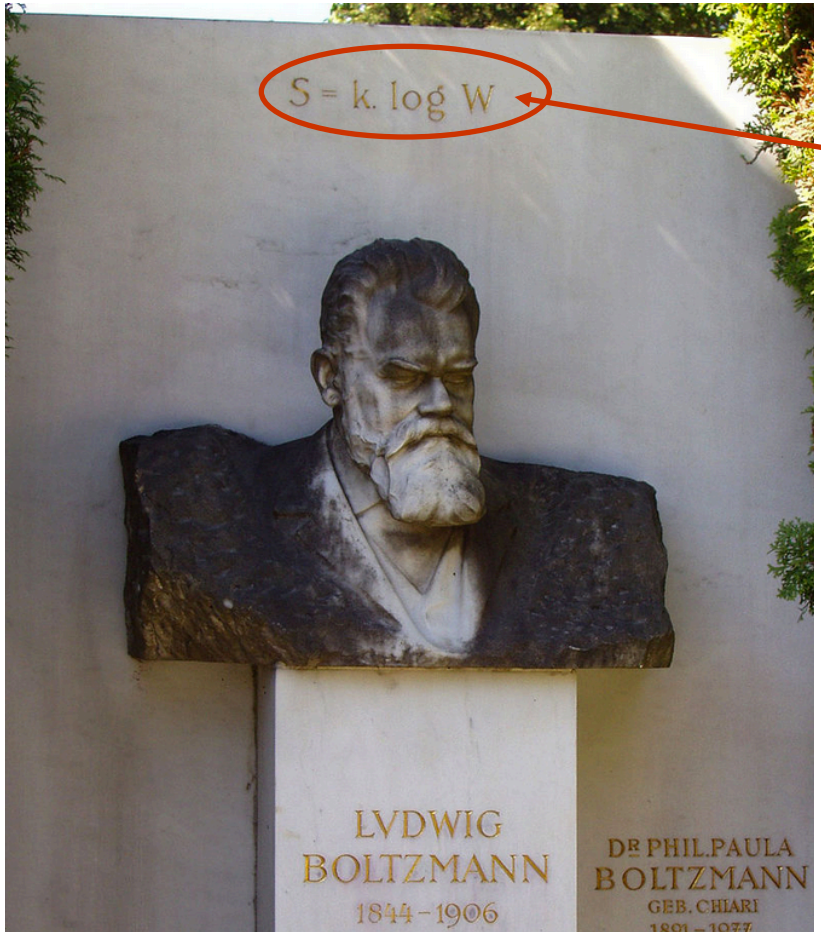
Outline

- Basic Information Theoretic Concepts
- Information Theoretic Measures of Social Influence
- Information Theoretic Representation Learning
  - For complex behavioral data
- Estimation of Entropic Measures
  - From limited data

Outline

- Basic Information Theoretic Concepts
- Information Theoretic Measures of Social Influence
- Information Theoretic Representation Learning
  - For complex behavioral data
- Estimation of Entropic Measures
  - From limited data

Good old entropy



Number of different micro-configurations corresponding to a given macro-state

Gibbs entropy $$S = -k_B \sum_i p_i \log p_i$$

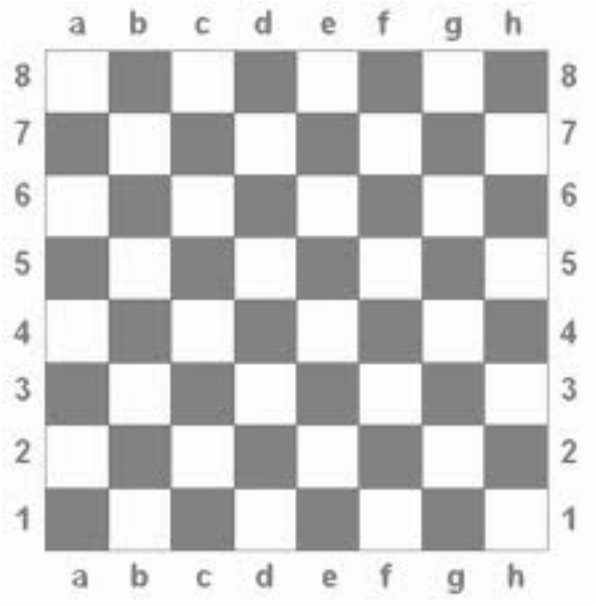
Entropy as a measure of information

- A random variable X, $p(X = x) = p(x) = 1/6$ $x = 1, \ldots, 6$

- Which functions quantify uncertainty?
  - **Continuous**
    - *a small change in p(x) should lead to a small change in our uncertainty*
  - **Increasing**
    - *If there are n equally likely outcomes, uncertainty goes up with n*
  - **Composition**
    - *The uncertainty for two independent coins should equal the sum of uncertainties for each coin)*

$$H(X) = \mathbb{E}(\log 1/p(x))$$
$$= -\sum_x p(x) \log p(x)$$

How many bits to encode a random variable?



Guess my square game:

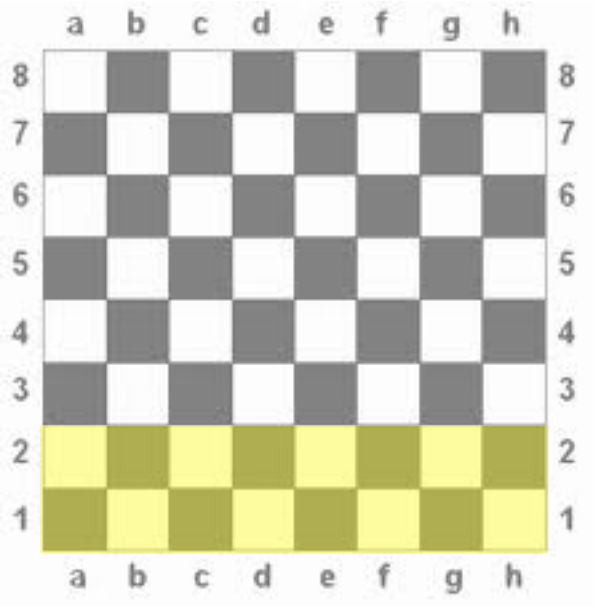
- I pick a square uniformly at random
- You can ask yes/no questions to determine the square

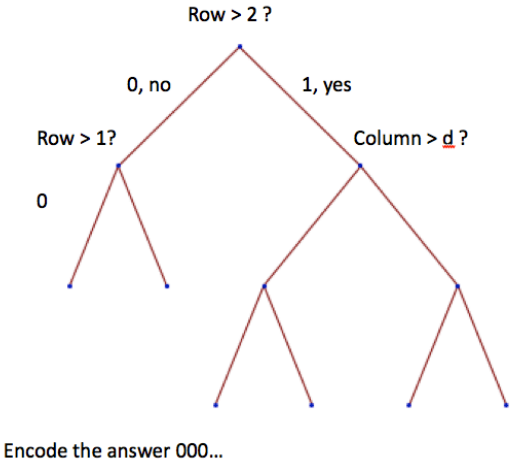
- How many questions are required?
- To distinguish between **N** squares, we need $\mathbf{\log_2 N}$ questions
  - Entropy of a uniform distribution over N outcomes

How many bits to encode a random variable?

- What if the distribution is not uniform?
  - E.g., I prefer the bottom two rows, and half the time pick one of those squares

- Find the correct square with fewer questions *on average*





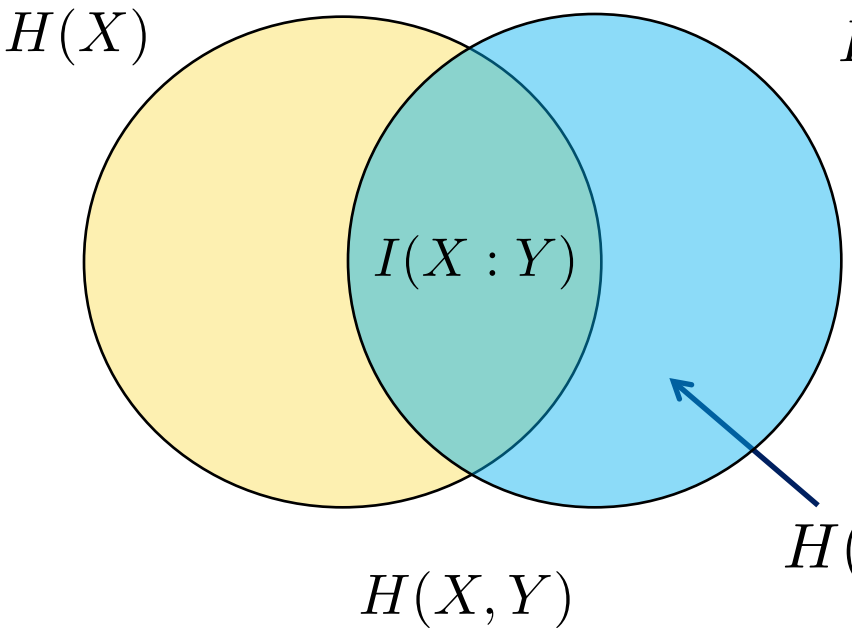
Encode the answer 000...

Mutual Information

$$I(X:Y) = \underbrace{H(X) + H(Y)}_{\text{Uncertainty if X and Y are independent}} - \underbrace{H(X,Y)}_{\text{Uncertainty considered as one system}}$$

- Things to note
  - Symmetric
  - Non-negative
  - Difference of entropic terms

Mutual Information

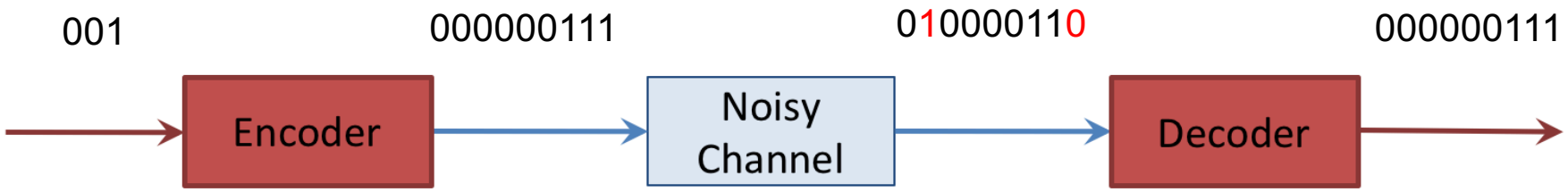


$$H(Y|X) = \sum_x p(x) H(Y|X=x)$$

$$\begin{aligned} I(X:Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

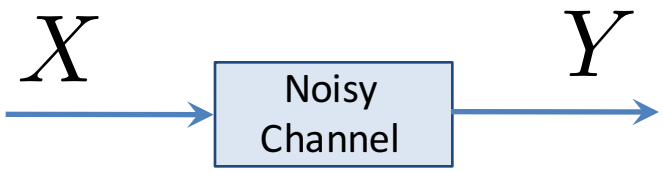
Channel Coding

How to communicate a message through a noisy channel?



1. Encode the message to introduce redundancy
   - Maps n-bits messaged to m-bit encoded message
   - Code rate R=n/m, e.g., R=1/3 for the above repetition code
2. Observe the transmitted message
3. Decode
   - E.g., via majority vote

Channel Coding & Mutual Information



channel capacity → $$C = \max_{p(X)} I(X : Y)$$

Channel Coding Theorem (Shannon, 1948)

For every code rate $R < C$, there are channel codes that allow almost error-free transmission of information.

- Does not say anything constructive about how to encode
- Decoding might be computationally expensive

Independence

$$I(X:Y) = \underbrace{H(X) + H(Y)}_{\text{Uncertainty if X and Y are independent}} - \underbrace{H(X,Y)}_{\text{Uncertainty considered as one system}}$$

$$H(X) = \mathbb{E}\left(\log 1/p(x)\right)$$

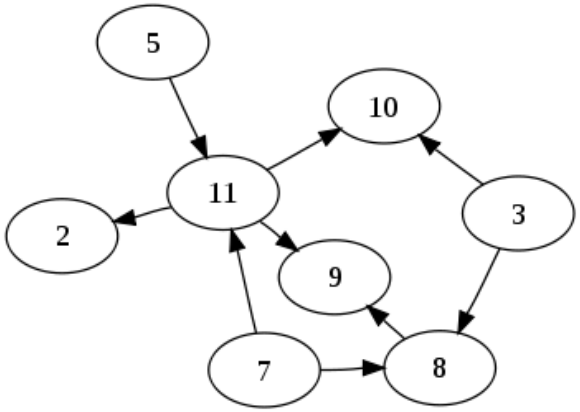
$$I(X:Y) = \mathbb{E}\left(\log 1/p(x) + \log 1/p(y) - \log 1/p(x,y)\right)$$

$$= \mathbb{E}\left(\log \frac{p(x,y)}{p(x)p(y)}\right)$$

$$I(X:Y) = 0 \longleftrightarrow p(x,y) = p(x)p(y)$$

Extends to Conditional Independence

- Bayesian networks, e.g., can be read as encoding a set of "conditional independence" relationships



$$p(X,Y|Z) = p(X|Z)p(Y|Z)\forall Z \longleftrightarrow X \perp Y|Z$$

$$X \perp Y|Z \longleftrightarrow I(X:Y|Z) = 0$$

$$I(X:Y|Z) = H(X|Z) - H(X|Z,Y)$$

- Basic Information Theoretic Concepts
- Information Theoretic Measures of Social Influence
- Information Theoretic Representation Learning
  - For complex behavioral data
- Estimation of Entropic Measures
  - From limited data

Social influence via predictability

- *Y* influences *X* if Y's past activity is a good predictor of X's future activity

- Quantified using information-theoretic concepts
  - E.g., *Transfer Entropy (Schreiber, 2000)*: How much our uncertainty about user X's future activity is reduced by knowing Y's past activity

$$TE_{Y \to X} = H(X^{\text{Future}}|X^{\text{Past}}) - H(X^{\text{Future}}|Y^{\text{Past}}, X^{\text{Past}})$$

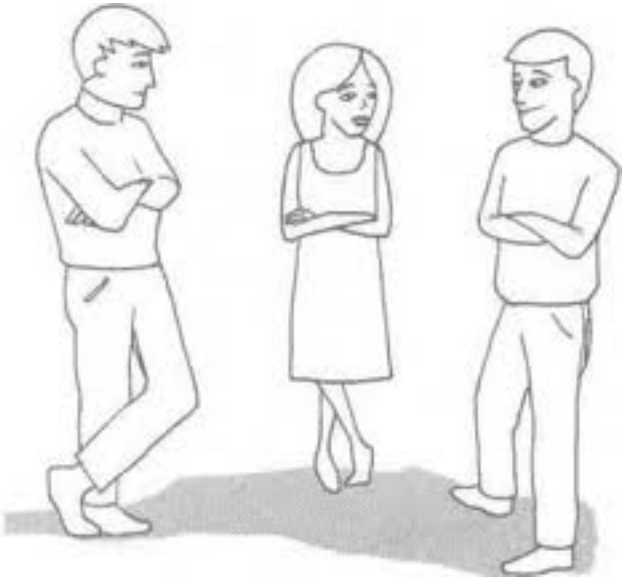
Uncertainty about X

Uncertainty about X, if you know Y's past activity

Model-free

***X, Y* can represent:**
Timing of activity (WWW'12)
Content (WSDM'13)
Style
Location
...

Behavioral mirroring



Coordination in communication

- Communication Accommodation Theory:
  - When conversing, people non-consciously adapt to one another's communicative behaviors [Giles, Coupland & Coupland 1991, Chartrand & Bargh, 1999]

Dimension	**Study**
Posture	Condon and Ogston, 1967
Head Nodding	Hale and Burgoon, 1984
Pause Length	Jaffe and Feldstein, 1970
Backchannels	White, 1984
Self-disclosure	Derlenga et al., 1973
Linguistic Style	Niederhoffer and Pennebaker, 2002
Linguistic Style **(Large Scale)**	Danescu-Niculescu-Mizil et al., 2011, 2012

Linguistic style coordination

- How things are said, rather what is said
- Example

  A: “What time are you available?”

  B: “Noon.”

Linguistic style coordination

- How things are said, rather what is said
- Example

  A: “What time are you available?”

  A: “At what time are you available?”

  B: “Noon.”

  B: “At noon.”

Linguistic style coordination

- How things are said, rather what is said
- Example

  A: "What time are you available?"

  A: "At what time are you available?"

  B: "Noon."

  B: "At noon."

- Quantified using function words (LIWC)
  - Reflect psychological processes [Chung & Pennebaker, 2007]
  - In this study: *articles*, *auxiliary verbs*, *conjunctions*, *adverbs*, *impersonal pronouns*, *personal pronouns*, *prepositions*, *quantifiers*

Linguistic style coordination

Alice:	dfasdf to the dafgaf	(1,1)
Bob:	by dfa at dafsdf the dagfg	(1,1)
Alice:	dfasgfge of dfsd gaf dgevm	(1,0)
Bob:	drgt for dag fgfd	(1,0)
Alice:	dasf to dagftef an erfsadfa	(1,1)
Bob:	dfasd dag ad dagf dafs	(0,0)

……

……

red: prepositions blue: articles

Linguistic style coordination

Alice:	dfasdf to the dafgaf	(1,1)
Bob:	by dfa at dafsdf the dagfg	(1,1)
Alice:	dfasgfge of dfsd gaf dgevm	(1,0)
Bob:	drgt for dag fgfd	(1,0)
Alice:	dasf to dagftef an erfsadfa	(1,1)
Bob:	dfasd dag ad dagf dafs	(0,0)

- Coordination: Is Bob more likely to use a particular feature in his response, if Alice used that feature in her post?

$$\text{Coord}(Bob \rightarrow Alice) = p(m_b = 1 | m_a = 1) - p(m_b = 1)$$

Prior results

- Observation of statistically significant coordination
  - Laboratory experiments [Pennebaker, 1999]
  - Large-scale experiments [Danescu-Niculescu-Mizil, 2012]
    - *Data from Supreme court transcripts & Wikipedia discussions*

- Stylistic coordination can be used to predict different behavioral outcomes
  - Relationship stability [Ireland, 2010]
  - Power relationship/social status [Danescu-Niculescu-Mizil, 2012]
  - Presidential debates & polling numbers [Romero 2015]

Alternative measure of stylistic coordination

- Given two users Alice and Bob and their corresponding feature sequence, we define stylistic coordination using (time-shifted) mutual information

m_A	m_B
0	0
1	0
0	1
0	0
1	0
...	...

$$\mathrm{Coord}(Bob \rightarrow Alice) = I(m_b^t : m_a^{t-1})$$

- For independent sequences the measure is identically zero

- Allows to consider possible confounders
  - E.g., length of utterances, conversation topic, etc

$$\mathrm{Coord}(Bob \rightarrow Alice) = I(m_b^t : m_a^{t-1} | Z)$$

Experiments

U.S. Supreme Court Oral arguments:

-50,000 verbal exchanges

-between **Justices** and **Lawyers**

Wikipedia Community of editors:

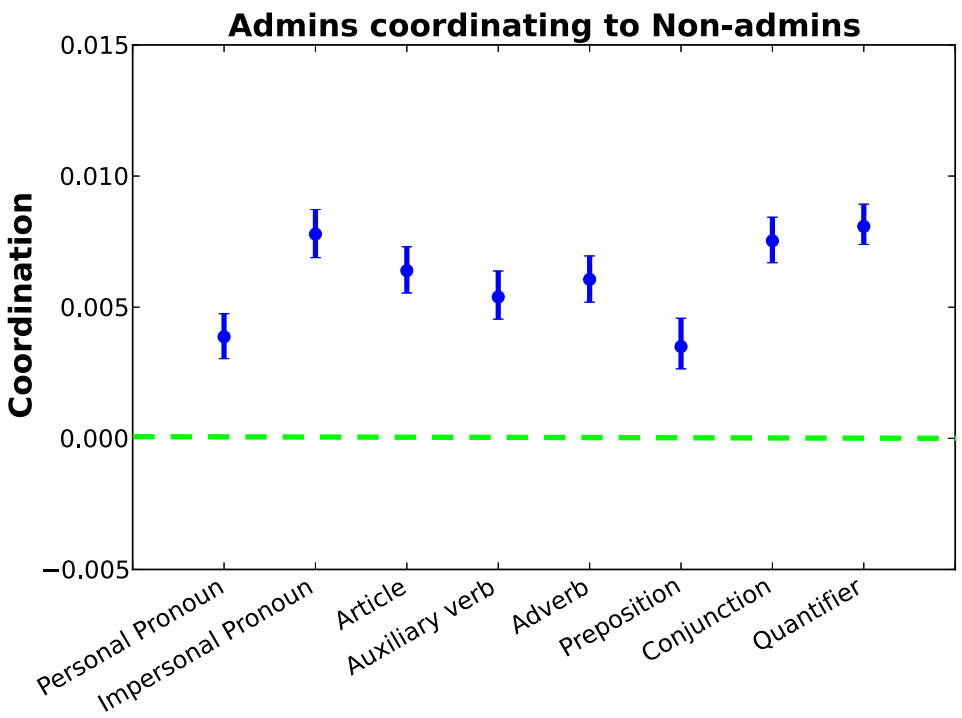
-240,000 conversational exchanges of discussions

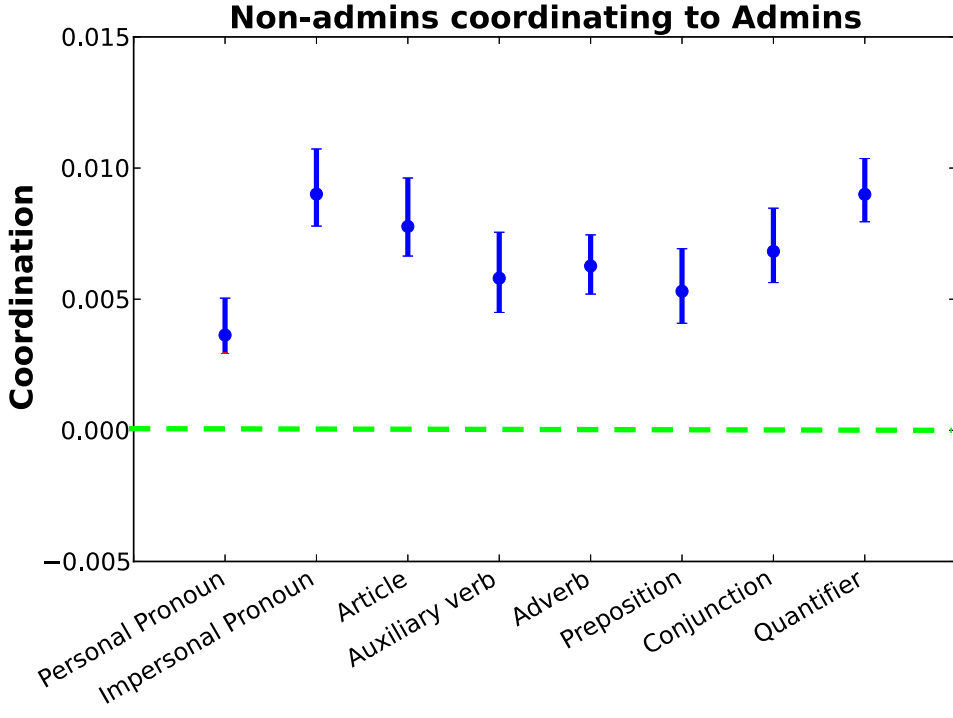
-users are either **admins** or **non-admins**

WIKIPEDIA
The Free Encyclopedia

Results

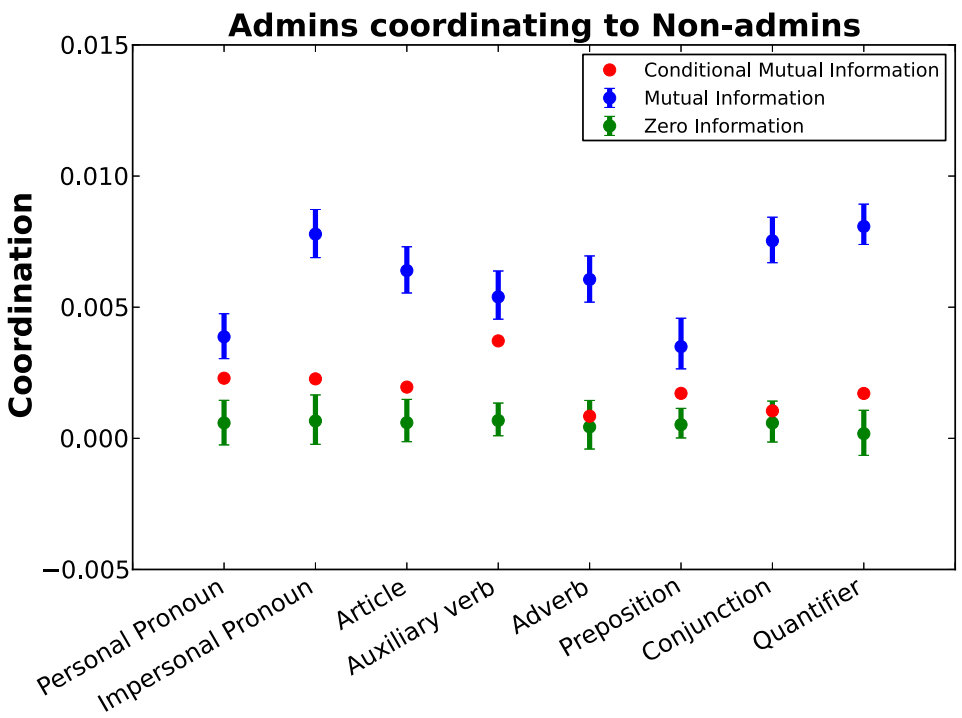
Wikipedia:

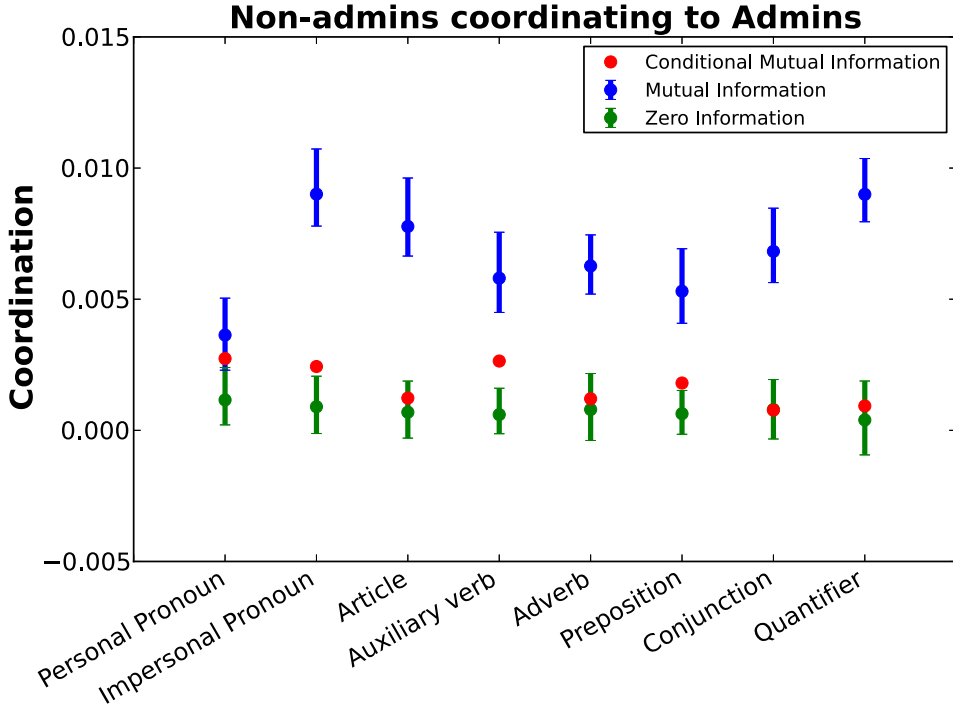




Results

Wikipedia: green error bars are obtained via shuffling the sequences

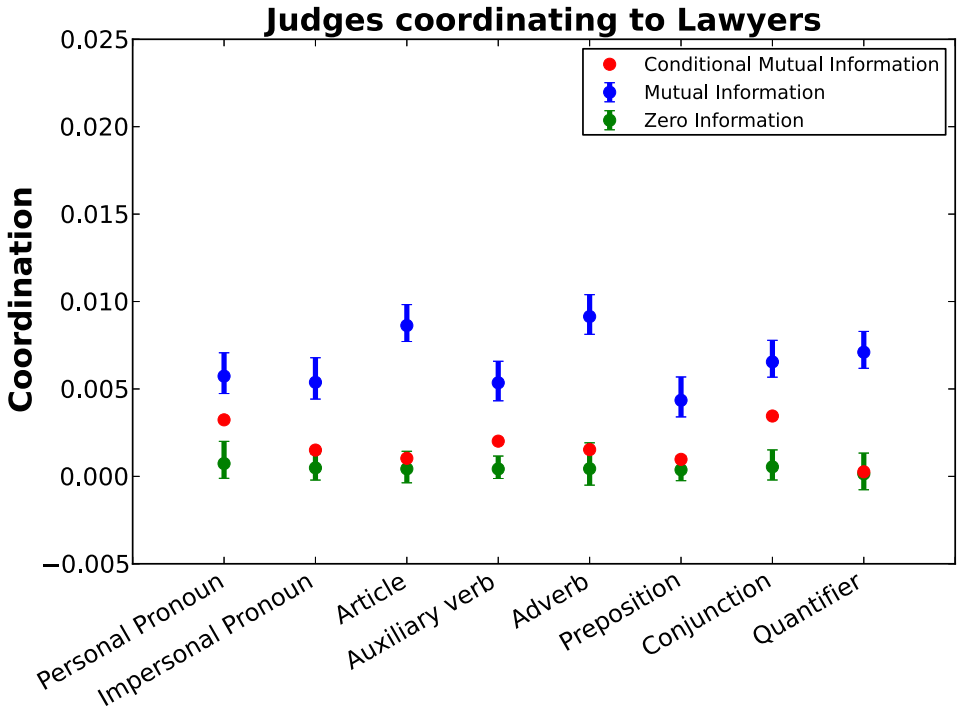


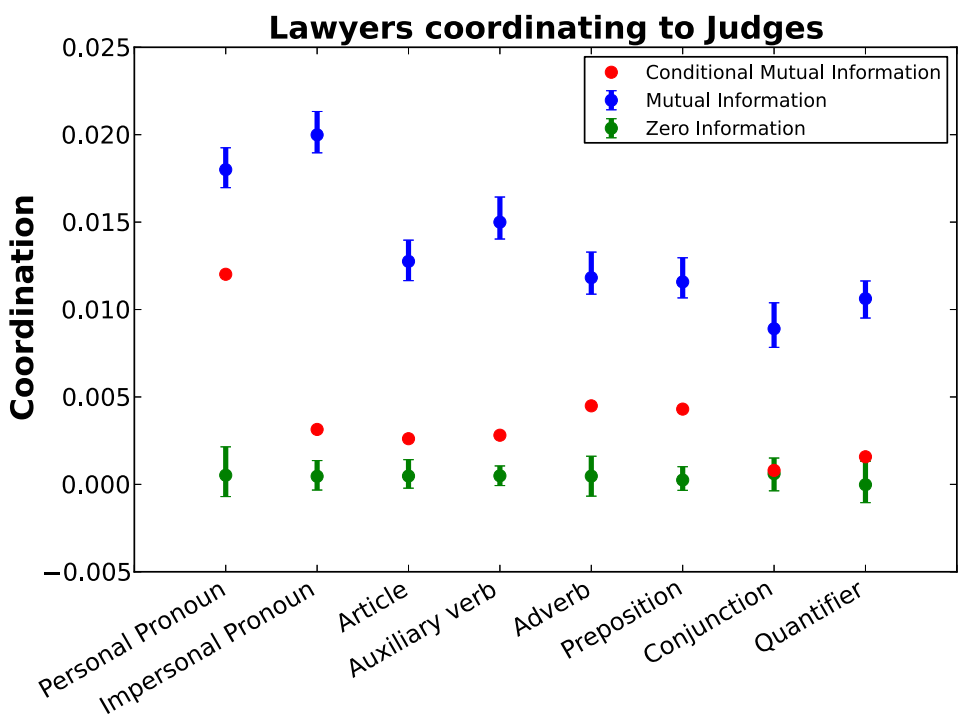


most "stylistic" coordination is "explained away" by length

Results

Supreme Court:

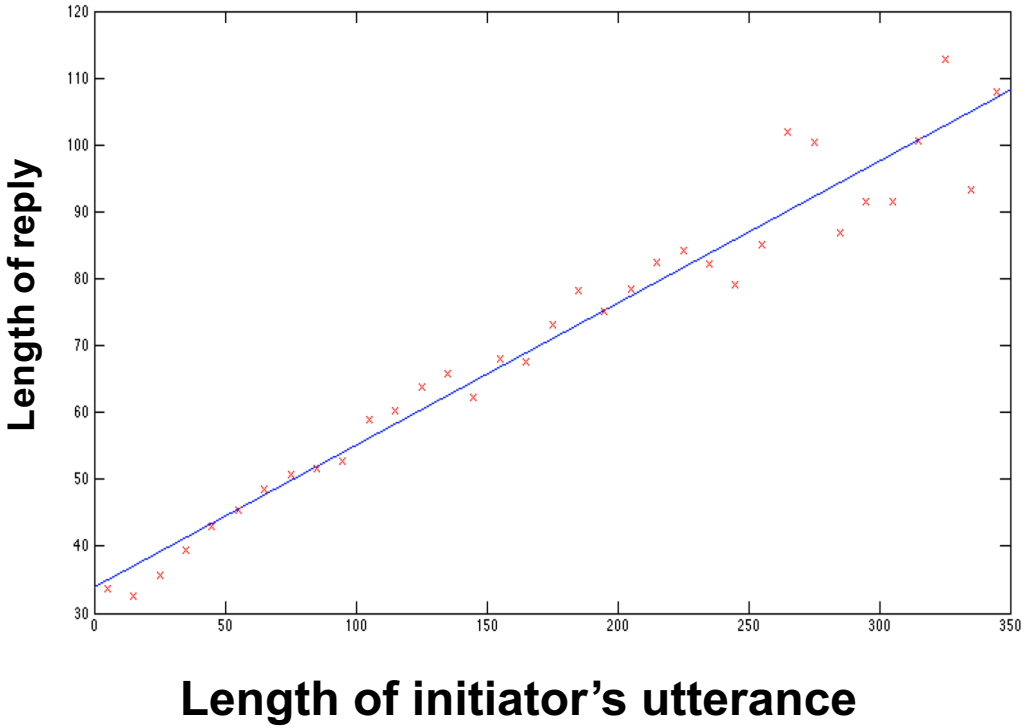




most "stylistic" coordination is "explained away" by length

Length as a confounding factor

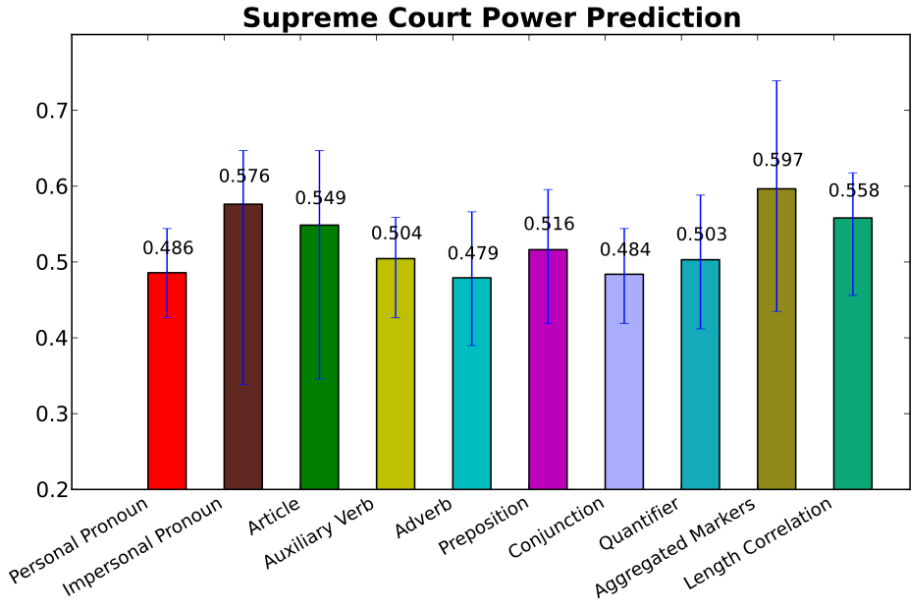
Wikipedia:

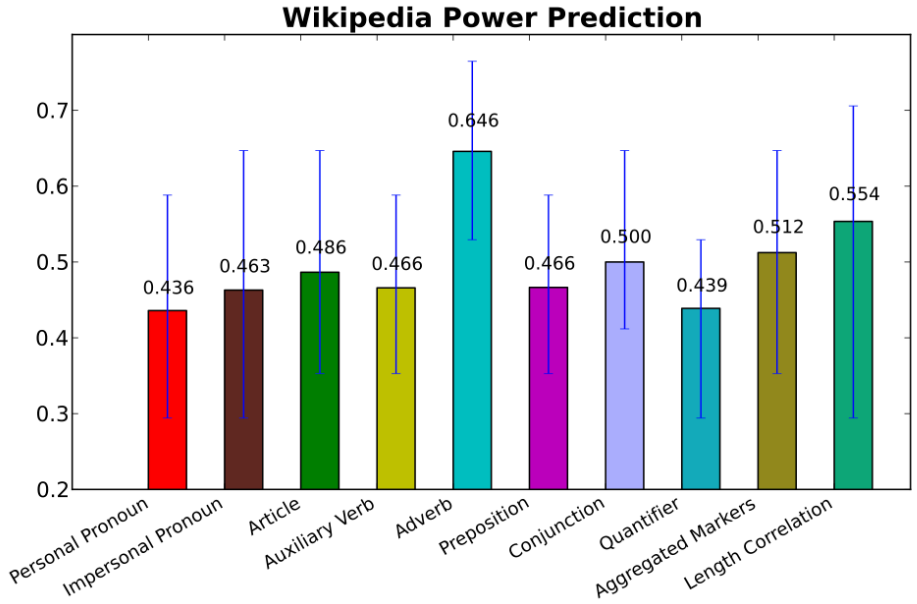


Longer utterances solicit longer response, producing spurious correlations in other features, e.g., # of occurrences of letter "r"

Stylistic coordination and social status

- Can we use asymmetry in stylistic coordination to predict power relationship?
  - Justices vs. lawyers, admin vs. non-admins

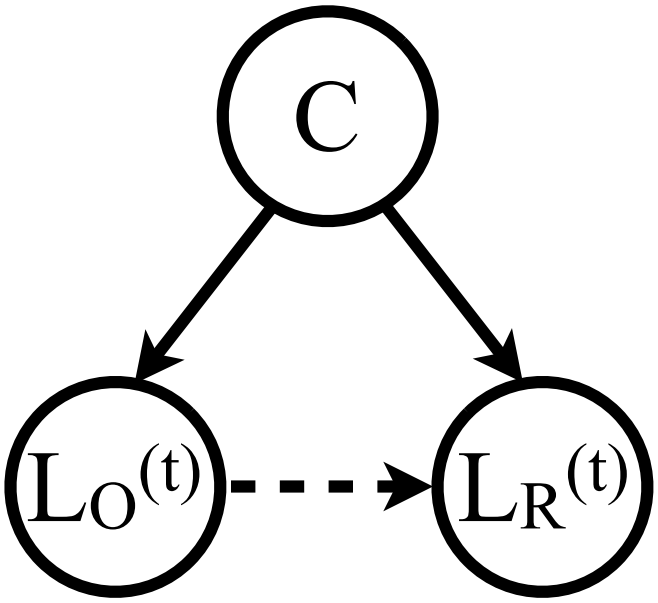




- Not really: observed asymmetry in stylistic coordination diminishes after conditioning on length

Understanding Length Coordination

- Bayesian Network for length coordination:



- Contextual factor: **C**
- Contextual influence: **C→L_O C→L_R**
- Turn-by-turn length coordination: **L_O→L_R**

Turn-by-turn Length Coordination Test

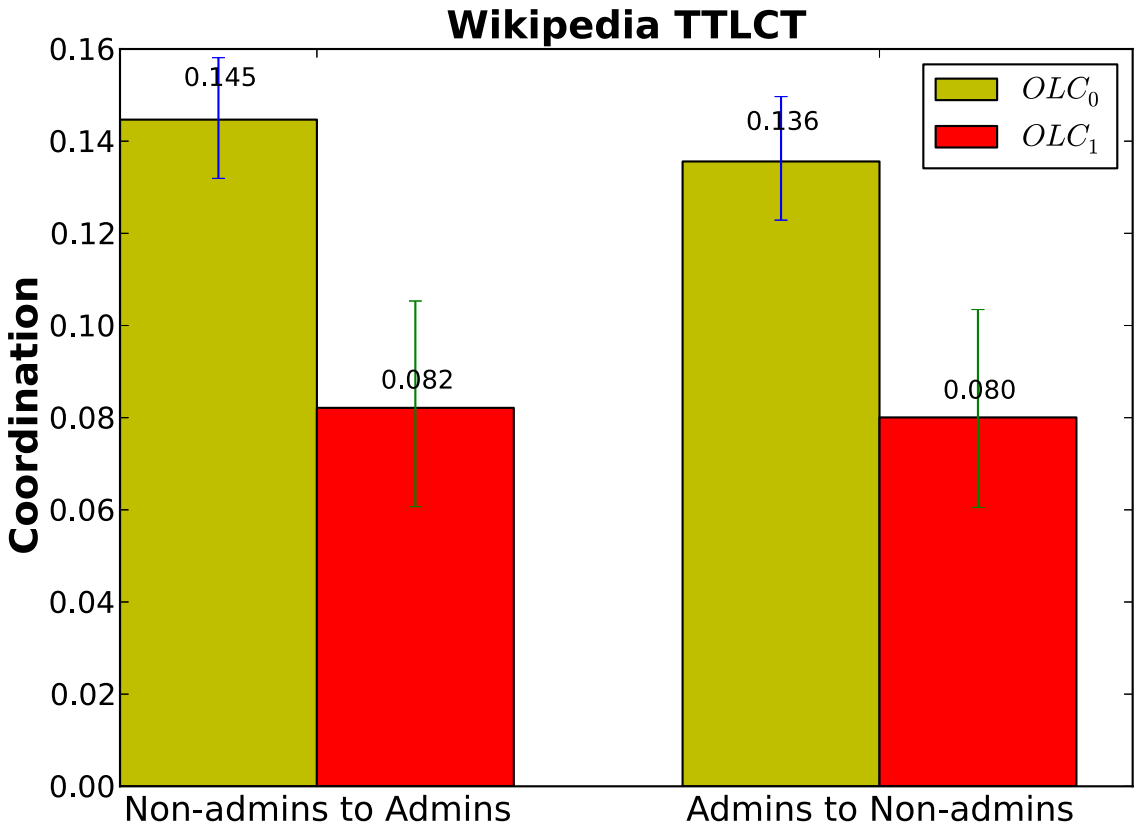
- A Conditional Monte Carlo Test

- *Overall Length Coordination*: OLC =I(L_O:L_R)
  - OLC_0: Original OLC
  - OLC_1: After shuffling utterances within each conversation

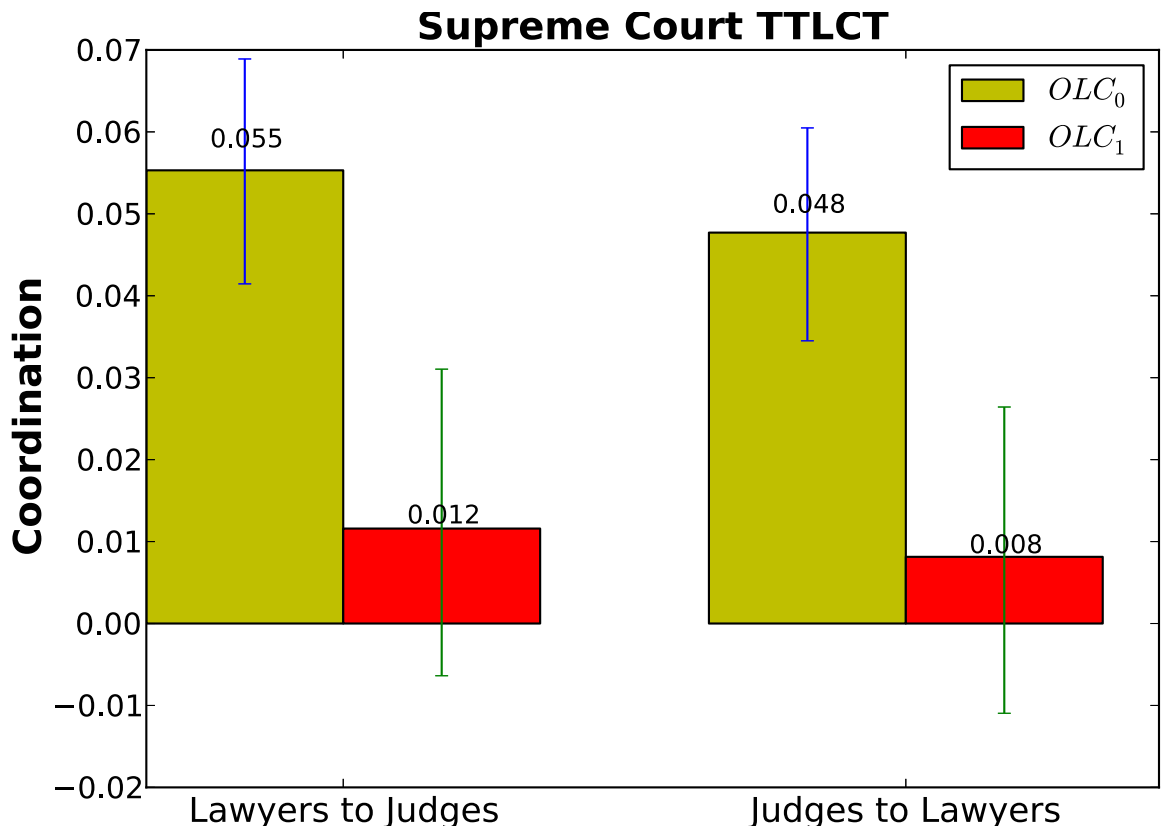
- Test: OLC_0=OLC_1?
  - If yes, then there is no turn-by-turn coordination

L_O	L_R	L_R
6	10	7
4	7	10
5	8	16
10	16	8

Turn-by-turn Length Coordination Test

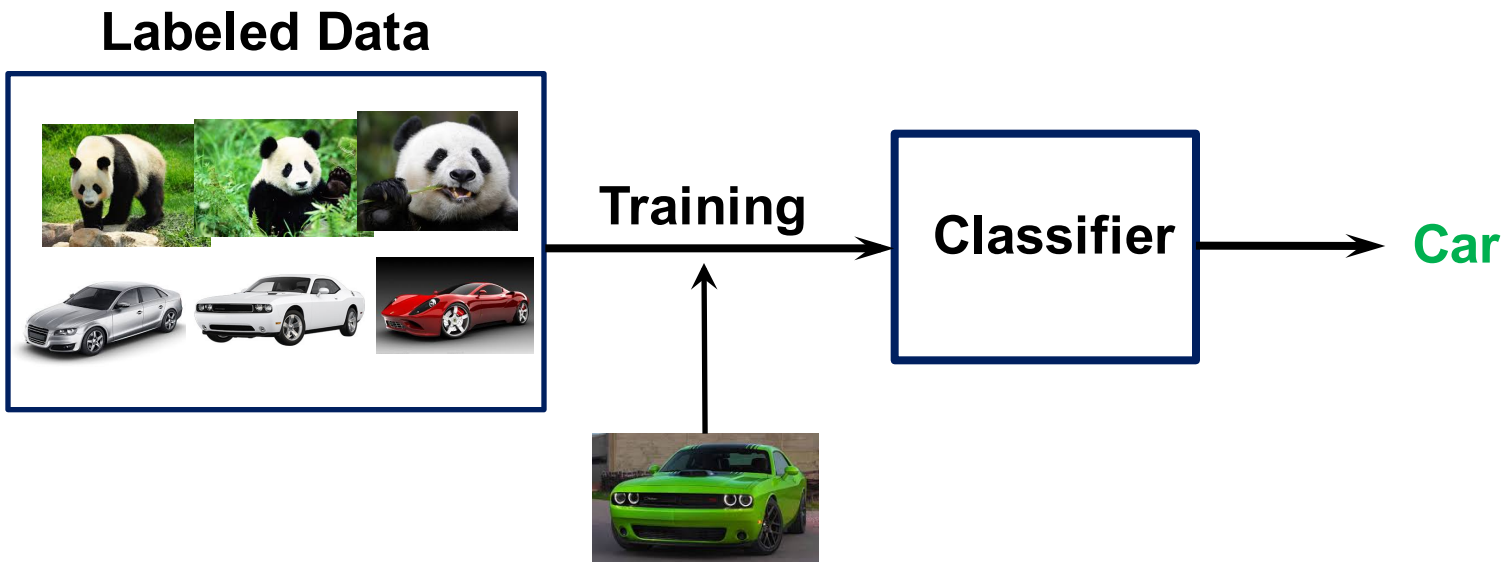


Turn-by-turn Length Coordination Test

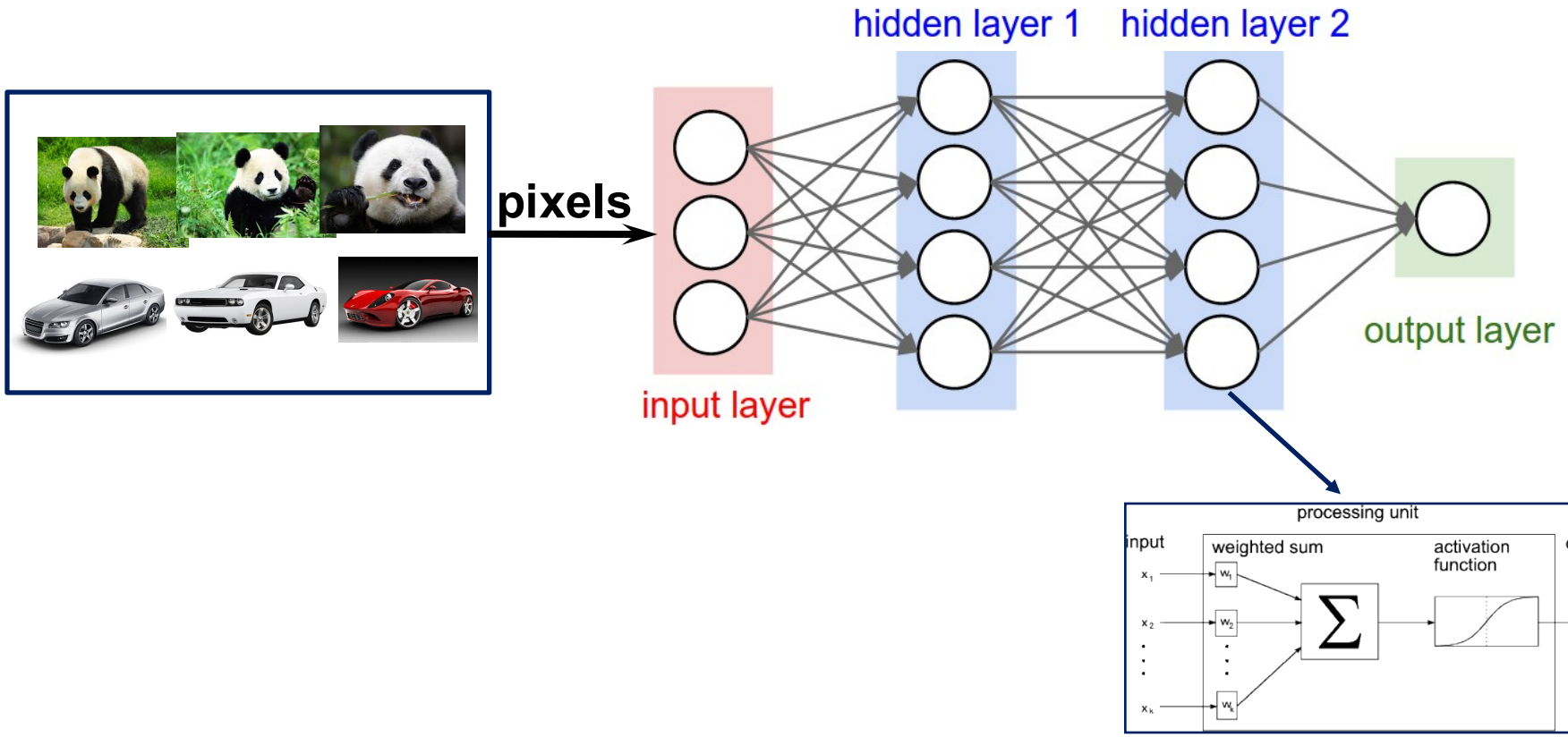


- Basic Information Theoretic Concepts
- Information Theoretic Measures of Social Influence
- Information Theoretic Representation Learning
  - For complex behavioral data
- Estimation of Entropic Measures
  - From limited data

Supervised Classification



Classification via Neural Networks



- Training corresponds to readjusting the weights of connection
- Inspired by neural computations in brain
- Extensive research in 80s-90s

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

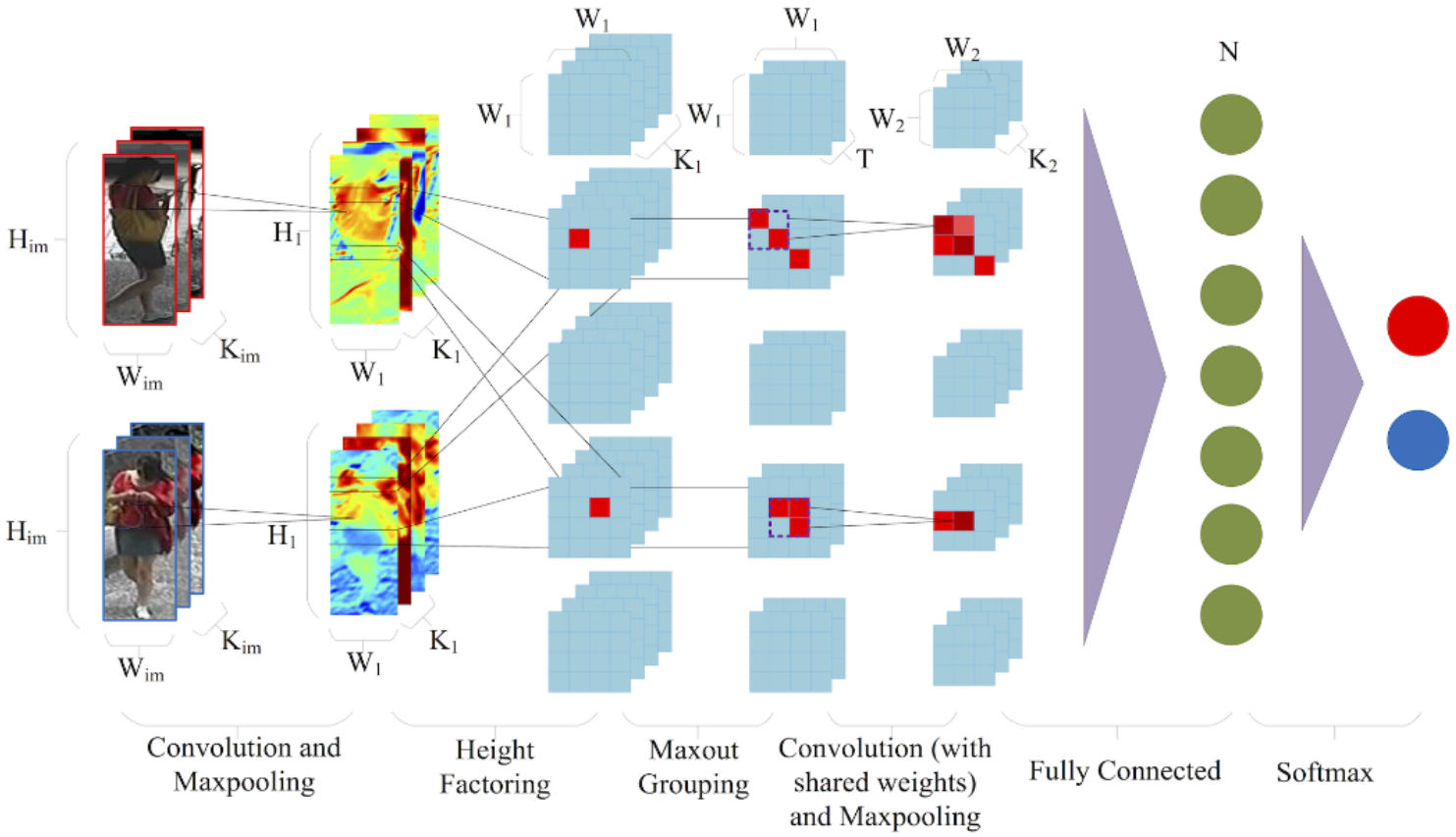
High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

2006 VOL 313 **SCIENCE** www.sciencemag.org

*..It has been obvious since the 1980s that backpropagation through deep autoencoders would be very effective for nonlinear dimensionality reduction, provided that computers were **fast** enough, data sets were **big** enough, and the initial weights were close enough to a good solution. **All three conditions are now satisfied**.*

Deep Learning



Unparalleled success in image processing, NLP, etc

Why is it Impressive?

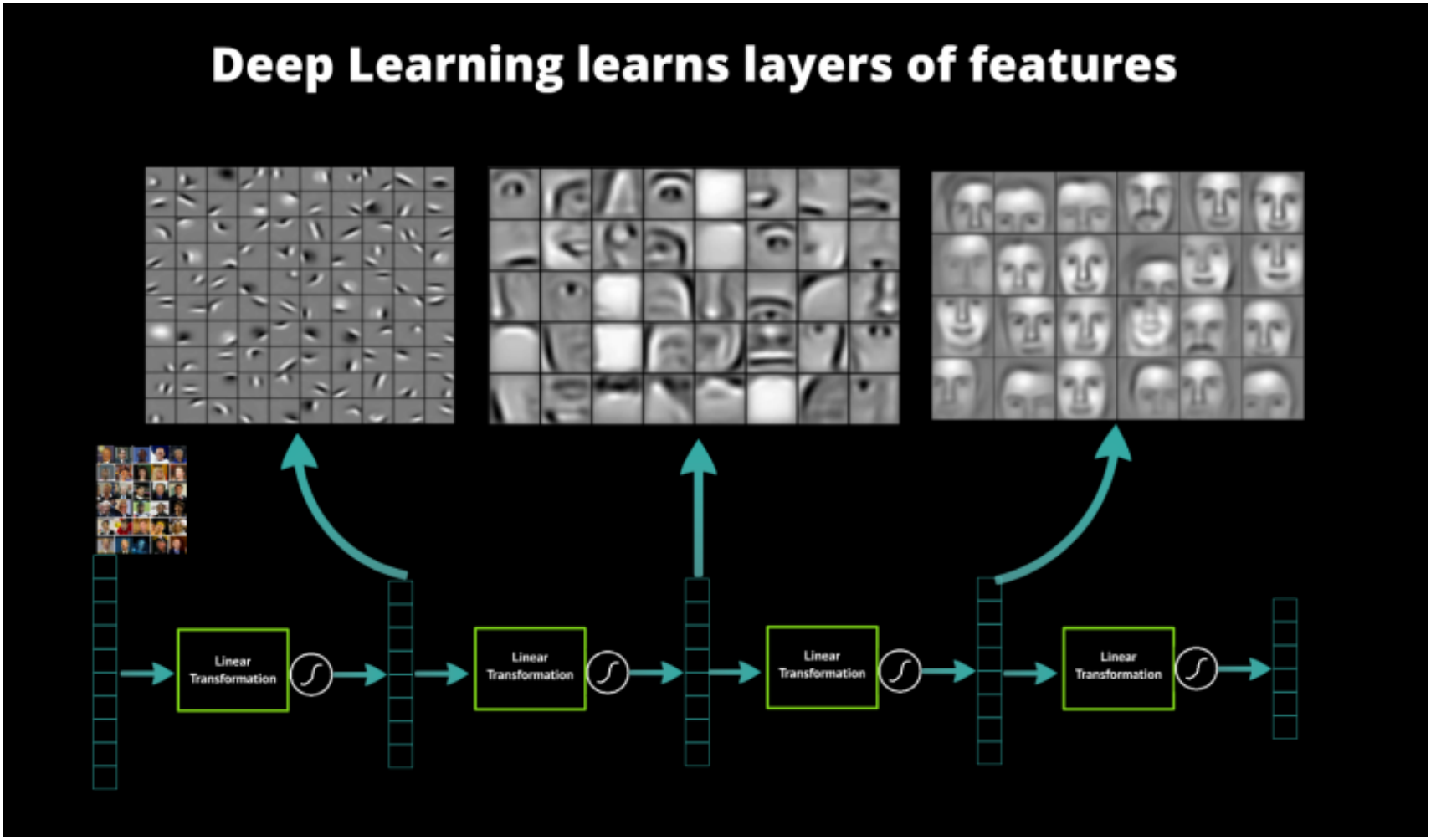
This is what we see

194	210	201	212	199	213	215	195	178	158	182	209
180	189	190	221	209	205	191	167	147	115	129	163
114	126	140	188	176	165	152	140	170	106	78	88
87	103	115	154	143	142	149	153	173	101	57	57
102	112	106	131	122	138	152	147	128	84	58	66
94	95	79	104	105	124	129	113	107	87	69	67
68	71	69	98	89	92	98	95	89	88	76	67
41	56	68	99	63	45	60	82	58	76	75	65
20	43	69	75	56	41	51	73	55	70	63	44
50	50	57	69	75	75	73	74	53	68	59	37
72	59	53	66	84	92	84	74	57	72	63	42
67	61	58	65	75	78	76	73	59	75	69	50

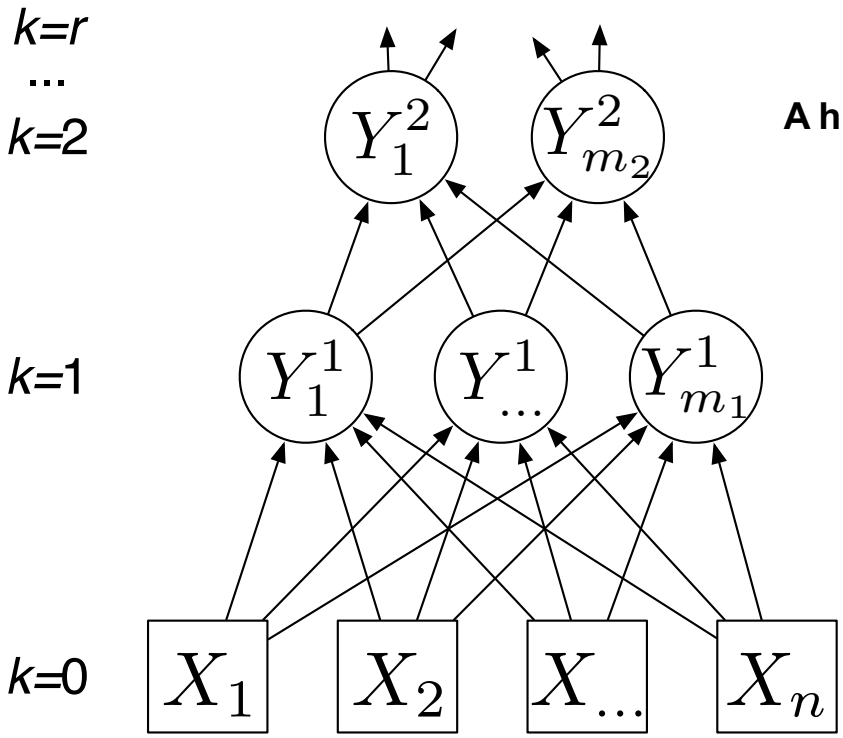
This is what the machine sees

[from Andrew Ng's slides]

Deep Representation Learning



Unsupervised Representation Learning



A hierarchical representation of X

Arbitrary probabilistic functions: a representation of X

How good is this representation?

High-dim. observations

Should be maximally informative about input data

Mutual Information

- The *reduction* of our uncertainty about X_1 if we know X_2

$$\begin{aligned} I(X_1; X_2) &= H(X_1) - H(X_1|X_2) \\ &= D_{KL}\left(p(x_1, x_2)||p(x_1)p(x_2)\right) \end{aligned}$$

- KL-divergence:

$$D_{KL}(p(x)||q(x)) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Multivariate Information

- Total correlation or multivariate information in X

$$TC(X) = \sum_i H(X_i) - H(X)$$

$$= D_{KL}(p(X) \prod p_i(X_i))$$

- If Y explains all the dependencies in X_i -s, then

$$TC(X|Y) = 0$$

$$TC(X|Y) = D_{KL}(p(X|Y)||p_i(X_i|Y))$$

- Reduction in dependence is quantified by

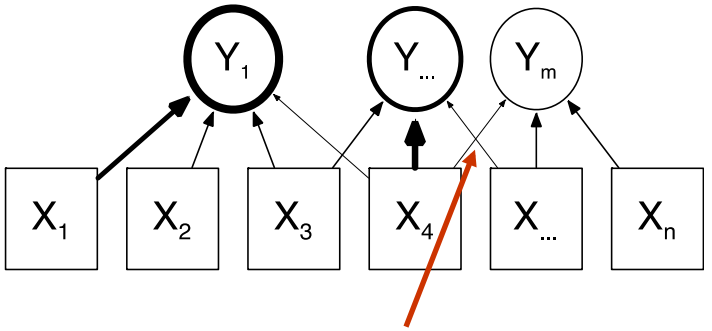
$$TC(X;Y) = TC(X) - TC(X|Y)$$

$$= \sum_i I(X_i : Y) - I(X : Y)$$

Correlation Explanation (CorEx)

$$\max_{p(y_j|x)} TC(X|Y)$$

Optimize over all probabilistic functions



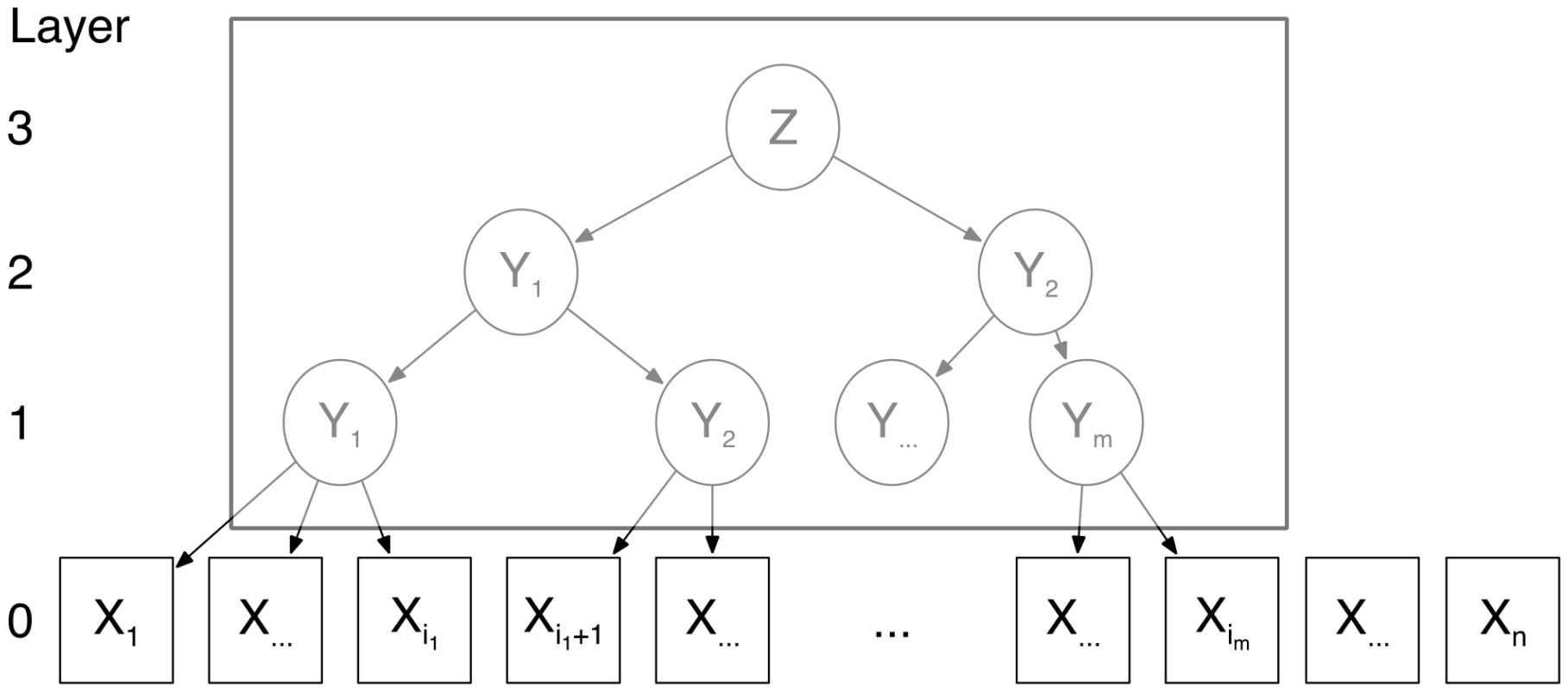
$$p(y_j|x) = \frac{p(y_j)}{Z_j(x)} \prod_{i=1}^{n} \left(\frac{p(y_j|x_i)}{p(y_j)} \right)^{\alpha_{i,j}}$$

- Efficient iterative solution
  - Linear scaling in # of variables, fewer samples required
- Theoretical guarantees
  - Discovered structure is maximally informative about the data
- Rich set of results
  - Structure, latent factors, anomalies

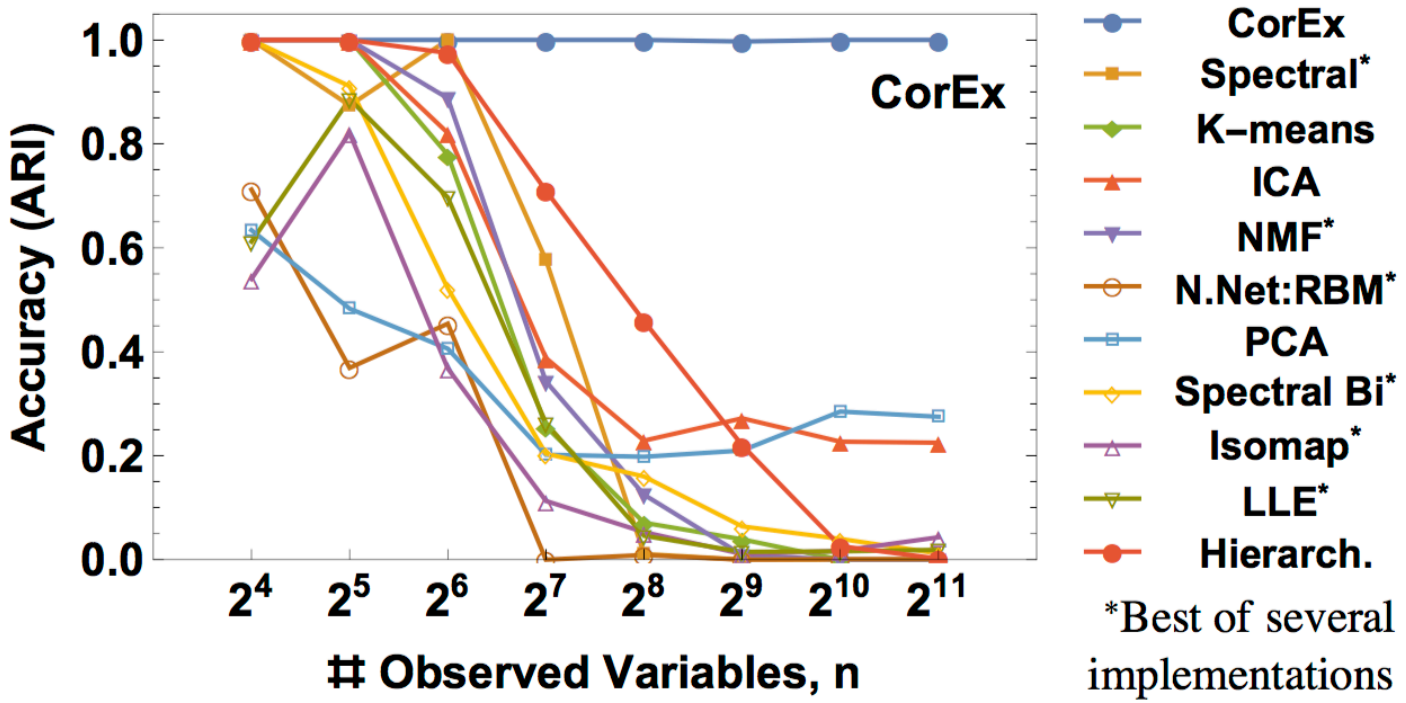
[Ver Steeg & Galstyan, NIPS'14, AISTATS'15, ICML'16]

Reconstructing Latent Tree Structure

Reconstructing Latent Tree Structure



Reconstructing Latent Tree Structure



There are also specialized techniques dedicated to latent tree learning: the complexity of these are $O(n^3)$ – $O(n^5)$, none could run on these examples with thousands of variables

The Big-5 personality test

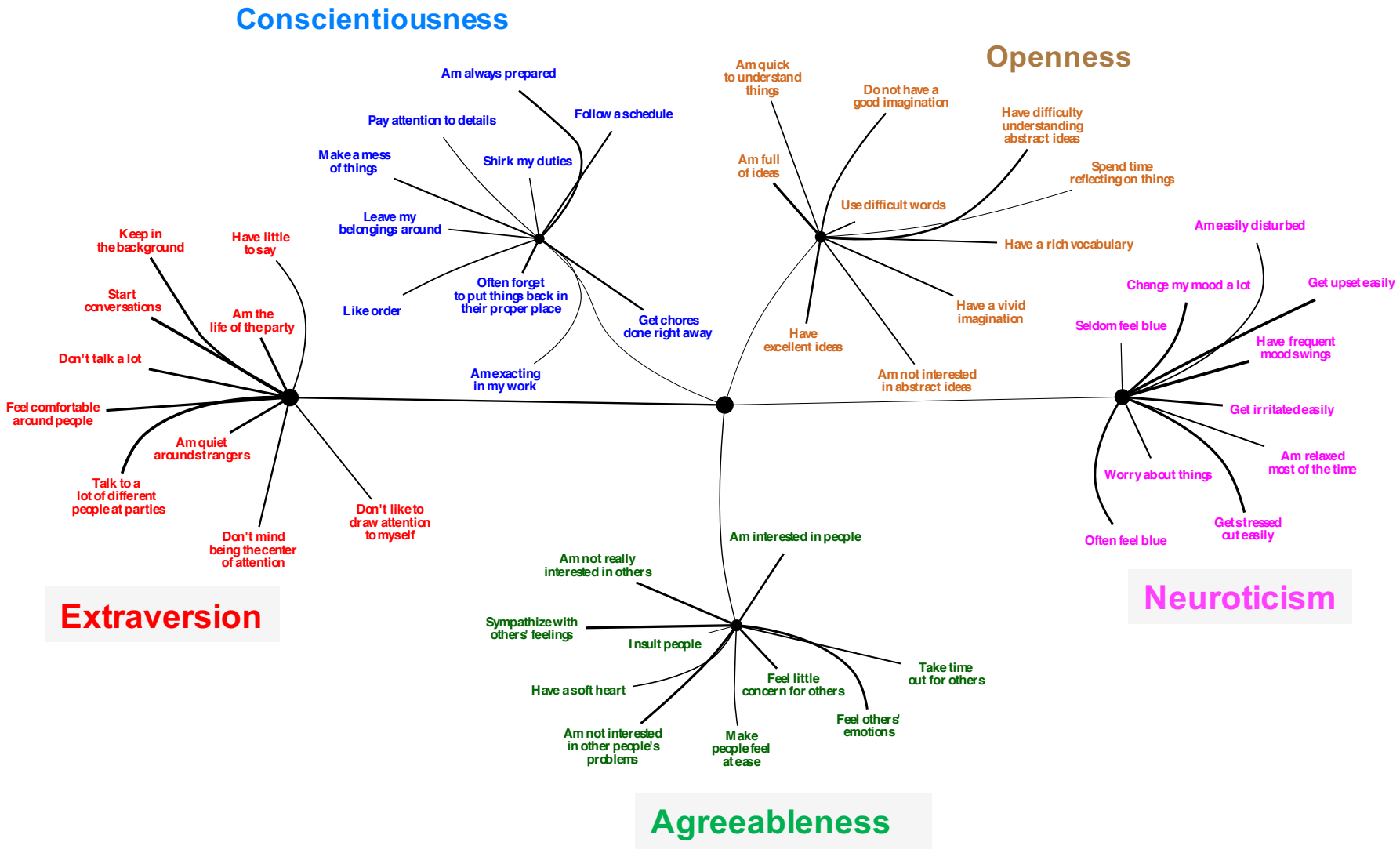
Q31: I am the life of the party

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

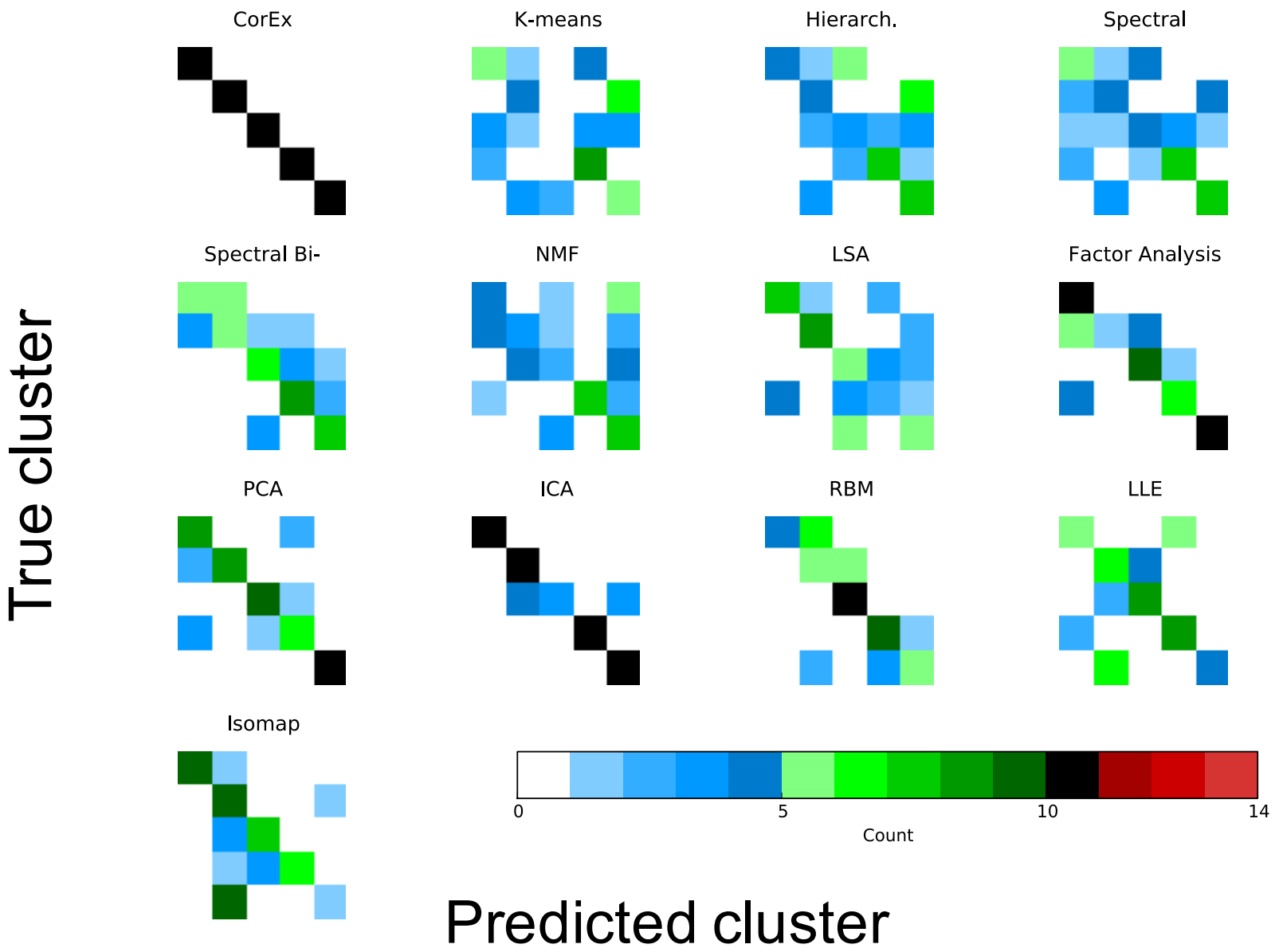
According to psychologists, this question measures ***Extroversion***, one of the "Big 5" personality traits.

Given answers to many questions, can we reverse engineer personality types?

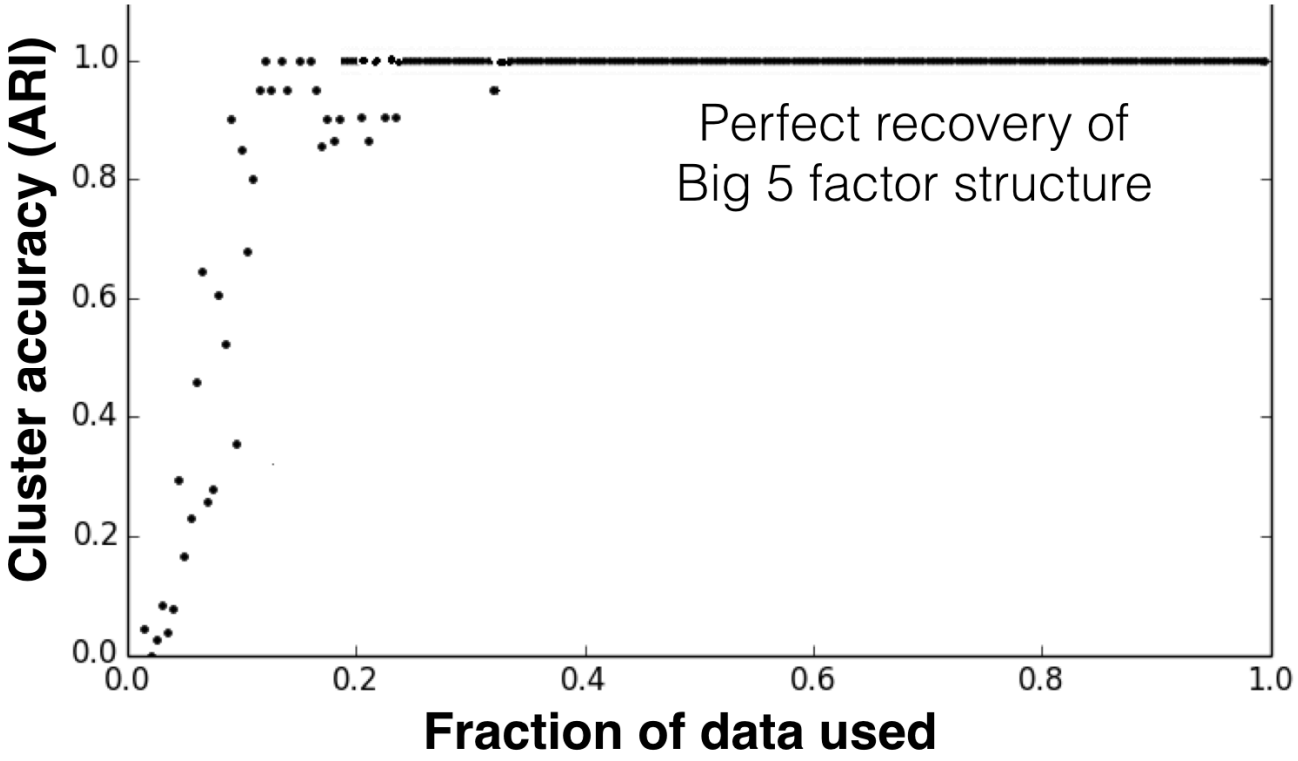
	Q1	Q2	Q3	…	Q50
Person 1	5	2	4		1
…					
Person N	2	2	5		5


Conscientiousness
Am always prepared
Follow a schedule
Pay attention to details
Make a mess of things
Shirk my duties
Leave my belongings around
Like order
Often forget to put things back in their proper place
Get chores done right away
Am exacting in my work
Openness
Am quick to understand things
Do not have a good imagination
Have difficulty understanding abstract ideas
Am full of ideas
Spend time reflecting on things
Use difficult words
Have a rich vocabulary
Have a vivid imagination
Have excellent ideas
Am not interested in abstract ideas
Keep in the background
Have little to say
Start conversations
Am the life of the party
Don't talk a lot
Feel comfortable around people
Am quiet around strangers
Talk to a lot of different people at parties
Don't mind being the center of attention
Don't like to draw attention to myself
Extraversion
Am easily disturbed
Change my mood a lot
Get upset easily
Seldom feel blue
Have frequent mood swings
Get irritated easily
Am relaxed most of the time
Worry about things
Get stressed out easily
Often feel blue
Neuroticism
Am interested in people
Am not really interested in others
Sympathize with others' feelings
Insult people
Have a soft heart
Feel little concern for others
Take time out for others
Am not interested in other people's problems
Make people feel at ease
Feel others' emotions
Agreeableness

Comparison with Other Methods

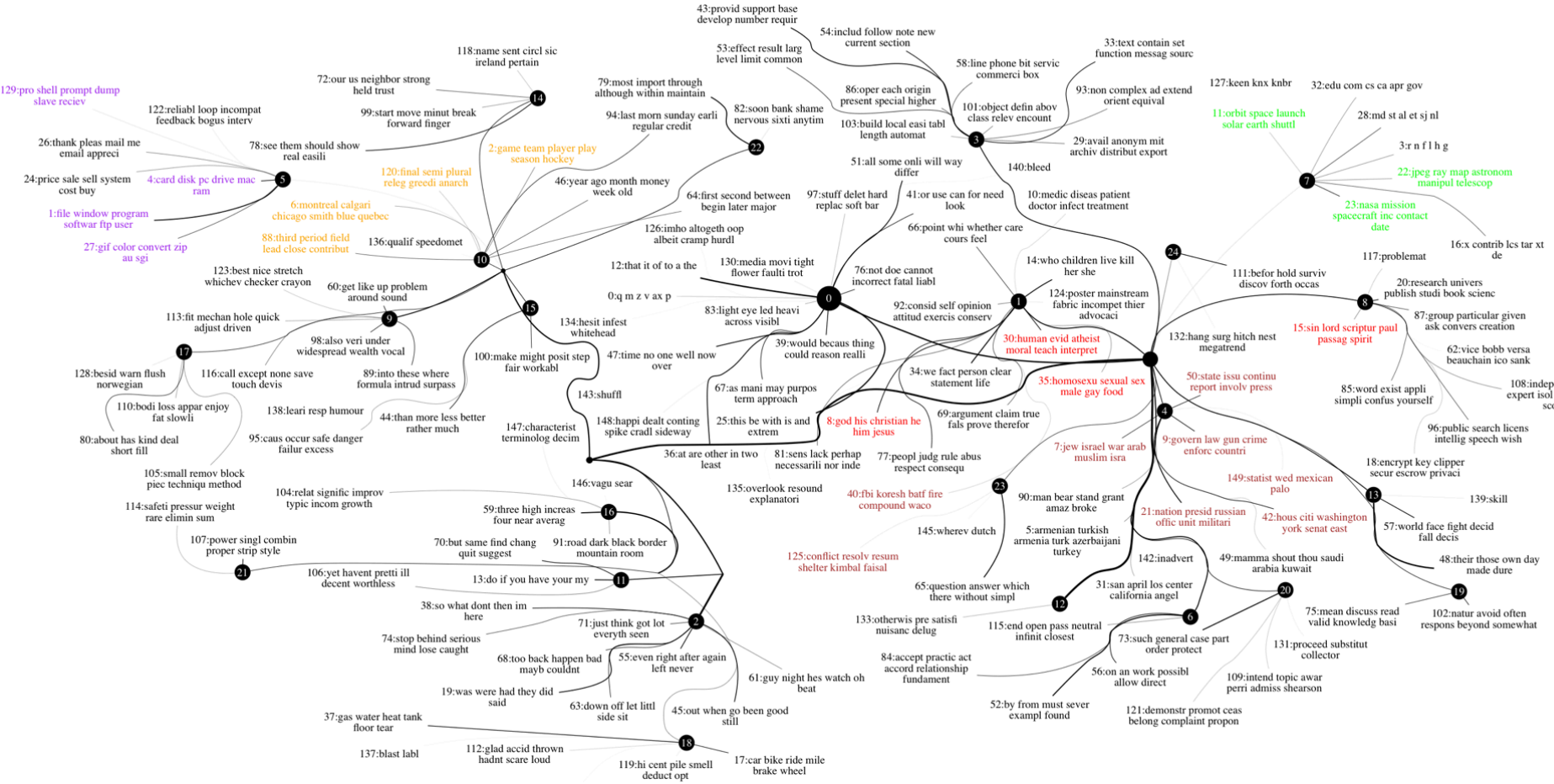


How many questions do we need?

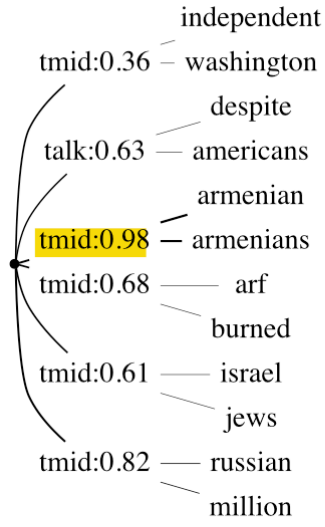


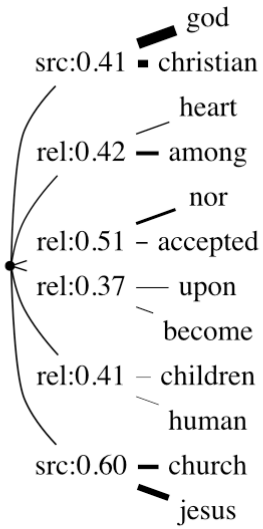
Unsupervised topic discovery

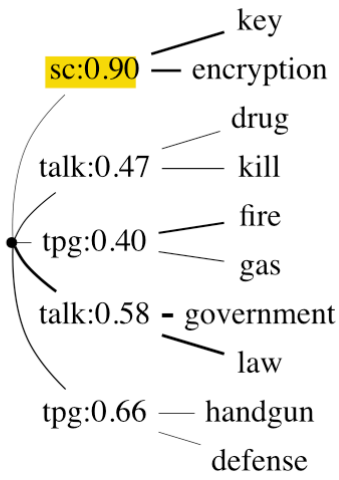
- Data from 20 newsgroups
- Each document is a sample, each variable is a word

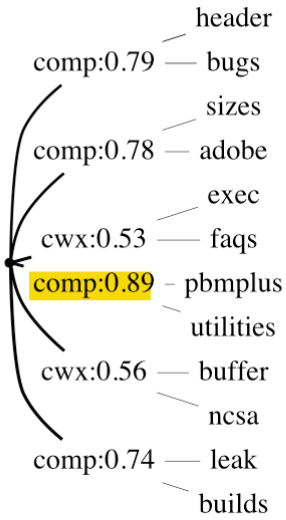


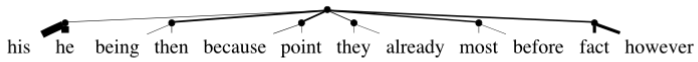
Unsupervised topic discovery

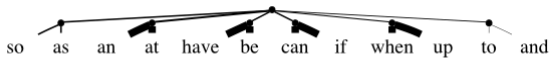












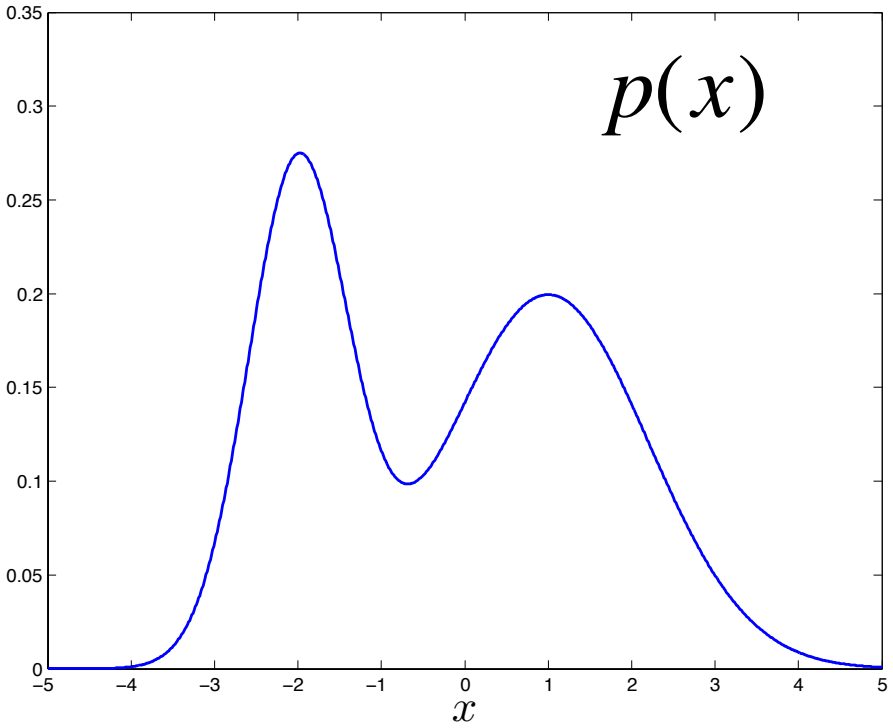
alt.atheism	aa	rel
comp.graphics	cg	comp
comp.os.ms-windows.misc	cms	comp
comp.sys.ibm.pc.hardware	cpc	comp
comp.sys.mac.hardware	cmac	comp
comp.windows.x	cwx	comp
misc.forsale	mf	misc
rec.autos	ra	vehic
rec.motorcycles	rm	vehic
rec.sport.baseball	rsb	sport
rec.sport.hockey	rsh	sport
sci.crypt	sc	sci
sci.electronics	se	sci
sci.med	sm	sci
sci.space	ss	sci
soc.religion.christian	src	rel
talk.politics.guns	tpg	talk
talk.politics.mideast	tmid	talk
talk.politics.misc	tmisc	talk
talk.religion.misc	trm	rel

- Basic Information Theoretic Concepts
- Information Theoretic Measures of Social Influence
- Information Theoretic Representation Learning
  - For complex behavioral data
- Estimation of Entropic Measures
  - From limited data

Estimating Entropic Measures

$$H(X) = -\int dx p(x) \log p(x)$$

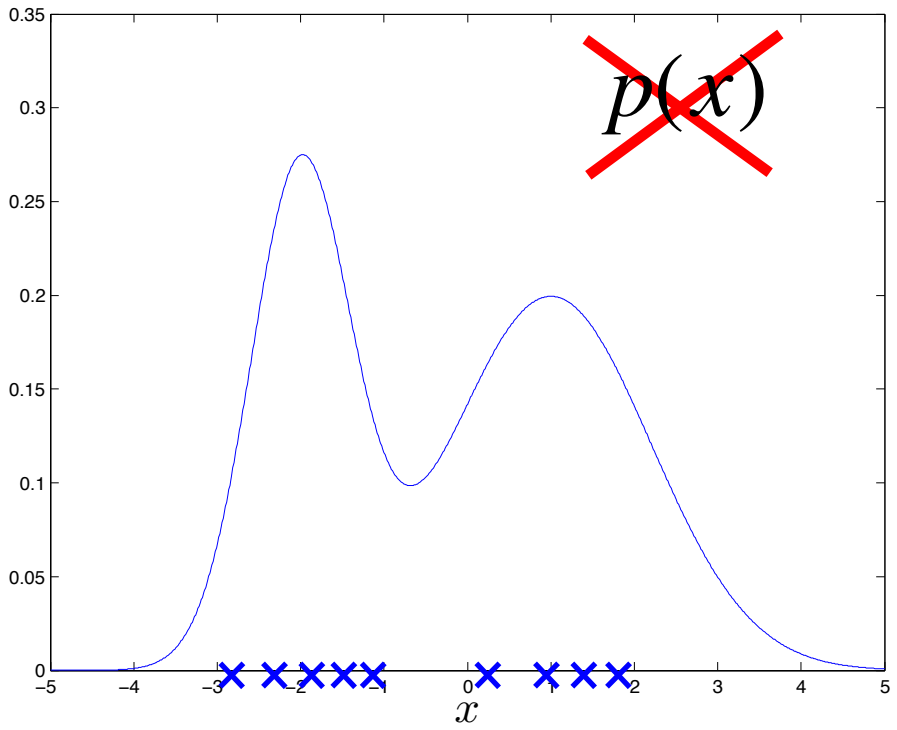
- Straightforward (kind of) if we know p(x)



Estimating Entropic Measures

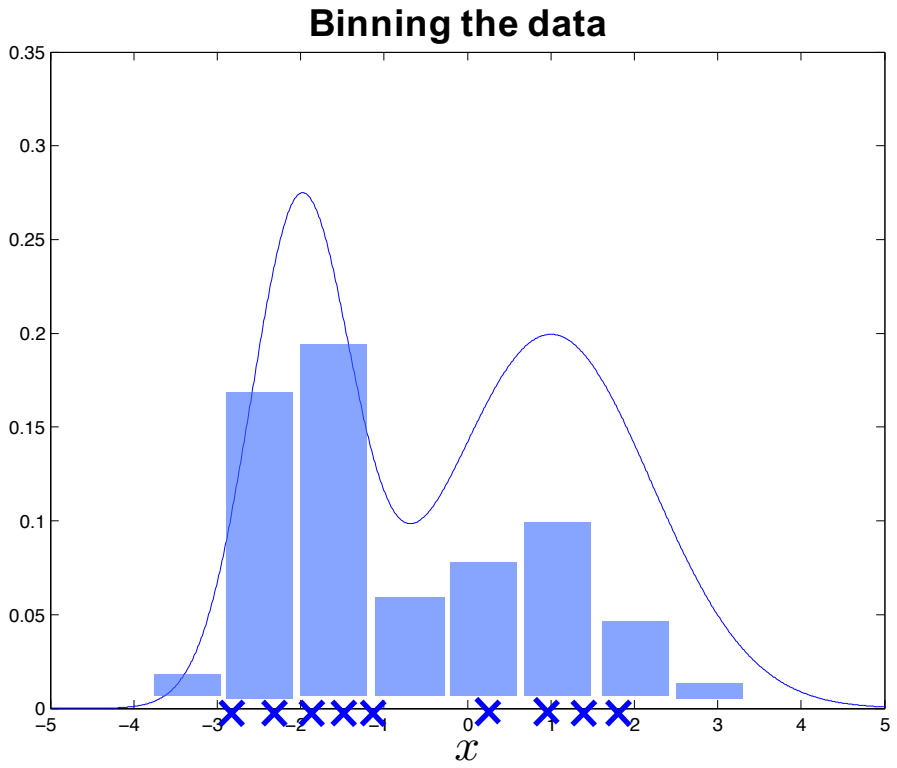
$$H(X) = -\int dx p(x) \log p(x)$$

- Usually we don't know p(x) (have samples $x_i \sim p(x)$)



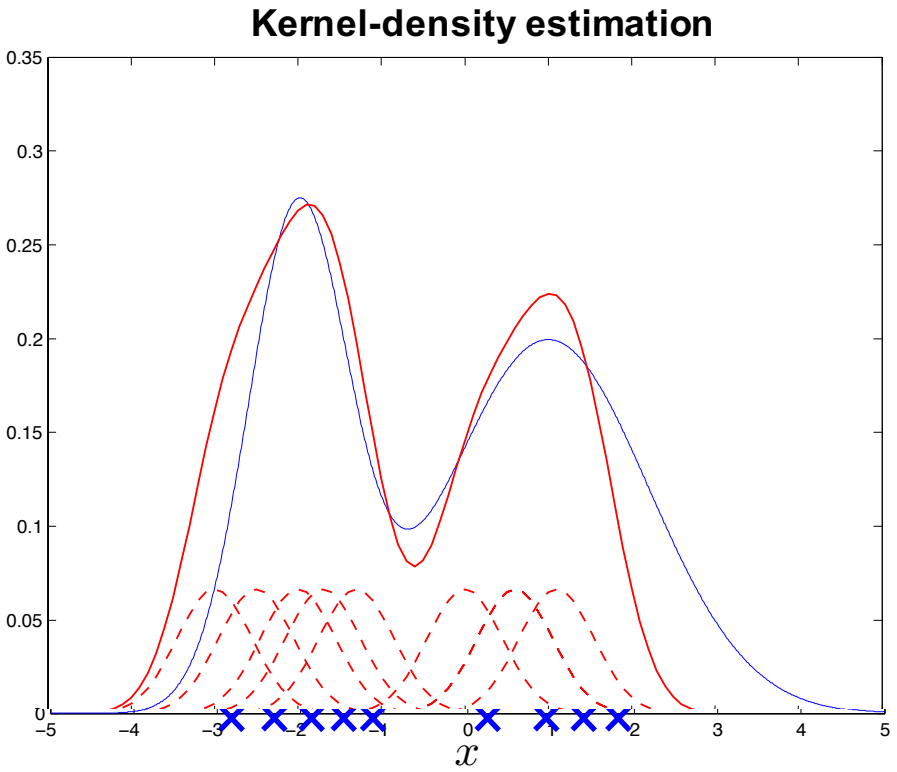
Plug-in Estimators

- Estimate p(x) and calculate the integral



Plug-in Estimators

- Estimate p(x) and calculate the integral



Does not work in high-dimensional, under-sampled settings

Binless Entropy Estimation

- One way to write entropy:

$$H(x) = \mathbb{E}_x[-\log p(x)]$$

- Given some samples $x_i \sim p(x)$,

$$\approx -\frac{1}{N}\sum_i \log p(x_i)$$

- We still don't know p(x)

- However, we need to estimate p(x) only at points x_i

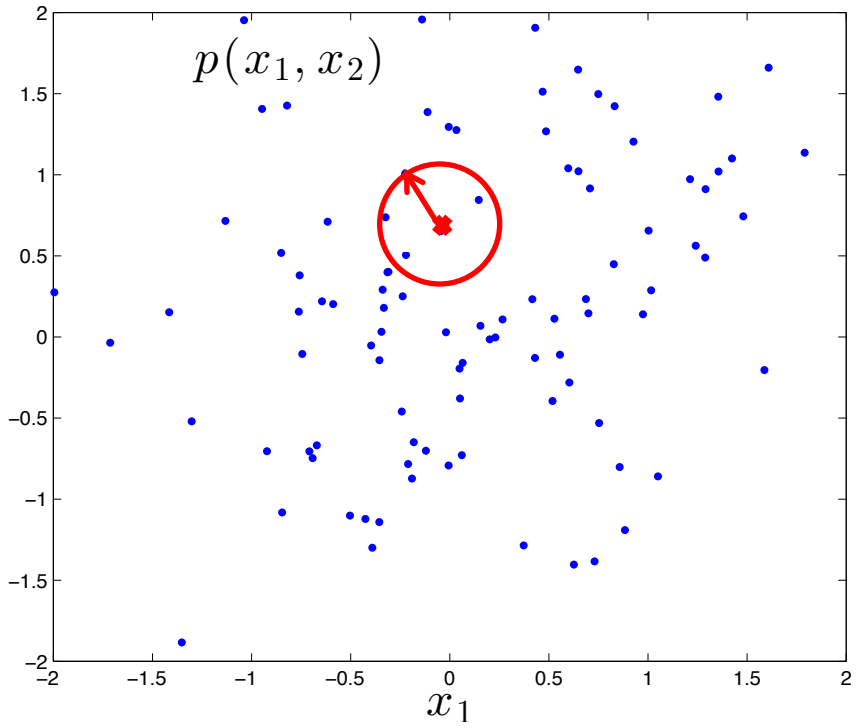
kNN Density Estimation for p(x)

- How to estimate the density p(x) at point $x^{(i)}$
  - Construct the k-nearest neighbor ball centered at $x^{(i)}$
  - **<u>Central Assumption</u>:**
    p(x) is uniform within the ball
  - Estimate

$$\hat{p}\left(\mathbf{x}^{(i)}\right) = \frac{probability\ mass\ of\ ball\ i}{Volume\ of\ ball\ i} = \frac{\%\ points\ in\ ball\ i}{Volume\ of\ ball\ i}$$

  - E.g. for d=2,k=4

$$\hat{p}_{k=4}\left(\mathbf{x}^{(i)}\right) = \frac{4/\left(N-1\right)}{\pi r_i^2}$$



$$\hat{H}\left(\mathbf{x}\right) = -\frac{1}{N}\sum_{i=1}^{N}\log\hat{p}\left(\mathbf{x}^{(i)}\right) = \frac{2}{N}\sum_{i=1}^{N}\log r_i + \log\left(N-1\right) - \log k$$

From Entropy to Mutual Information

- Mutual information is written as:

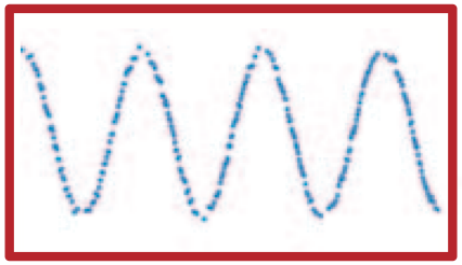
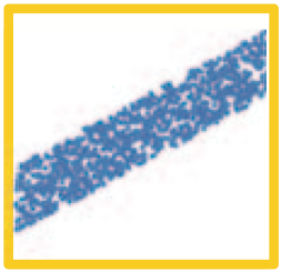
$$I(\mathbf{x}) = \sum_{i=1}^{d} H(x_i) - H(\mathbf{x})$$

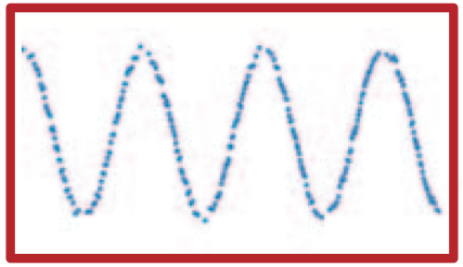
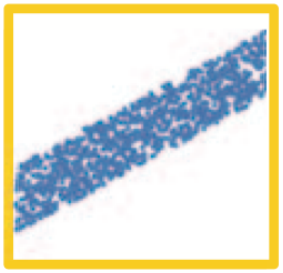
- A simple MI estimator:

$$\hat{I}(\mathbf{x}) = \sum_{i=1}^{d} \widehat{H}(x_i) - \widehat{H}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{\hat{p}\left(\mathbf{x}^{(i)}\right)}{\hat{p}\left(x_1^{(i)}\right) \hat{p}\left(x_2^{(i)}\right) \ldots \hat{p}\left(x_d^{(i)}\right)}$$

Limitations of MI estimators

Reshef et al., "Detecting novel associations in large data sets." Science, 2011

$MI($  $) \neq MI($  $) = 1.0$

$\widehat{MI}($  $) \approx \widehat{MI}($  $) \approx 1.0$

Mutual Information

Equitability, mutual information, and the maximal information coefficient

Justin B. Kinney[1] and Gurinder S. Atwal

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724

Edited* by David L. Donoho, Stanford University, Stanford, CA, and approved January 21, 2014 (received for review May 24, 2013)

How should one quantify the strength of association between two random variables without bias for relationships of a specific form? Despite its conceptual simplicity, this notion of statistical "equitability" has yet to receive a definitive mathematical formalization. Here we argue that equitability is properly formalized by a self-consistency condition closely related to Data Processing Inequality.

dependencies without bias for relationships of one type or another. And although it was proposed in the context of modeling communications systems, mutual information has been repeatedly shown to arise naturally in a variety of statistical problems (6–8).

The use of mutual information for quantifying associations in continuous data is unfortunately complicated by the fact that it requires an estimate (explicit or implicit) of the probability dis-

MI is just fine: one only needs more data points for accurate estimation

Cleaning up the record on the maximal information coefficient and equitability

Although we appreciate Kinney and Atwal's interest in equitability and maximal information coefficient (MIC), we believe they misrepresent our work. We highlight a few of our main objections below.

Regarding our original paper (1), Kinney instead that we look for approximations and solutions in restricted cases, an impossibility result about perfect equitability provides focus for further research, but does not mean that useful solutions are unattainable. Similarly, as others have noted far will allow researchers in the area to most productively and collectively move forward.

David N. Reshef[a,b,1,2], *Yakir A. Reshef*[b,1,2], *Michael Mitzenmacher*[c,3], *and Pardis C. Sabeti*[d,e,3]

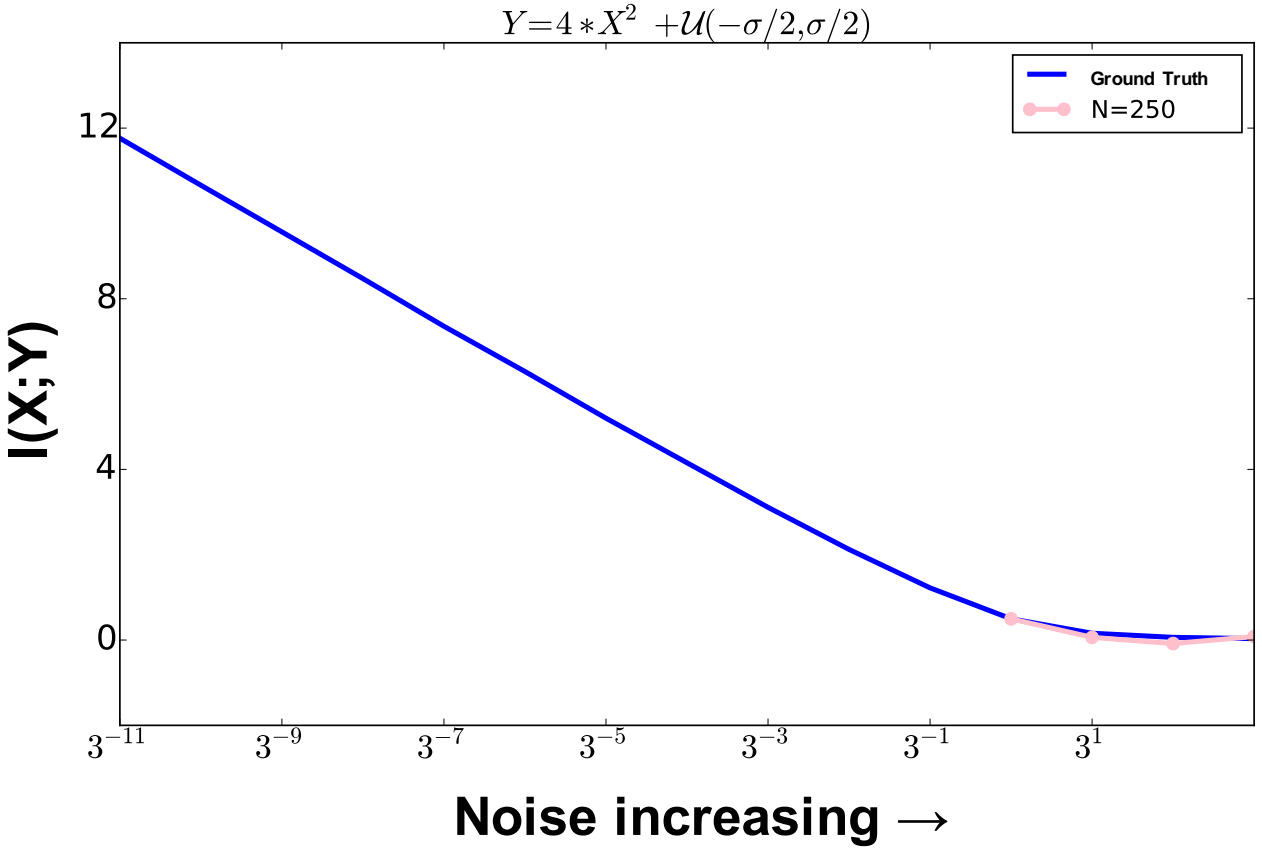
Reply to Reshef et al.: Falsifiability or bust

The term "equitability" was introduced by Reshef et al. in ref. 1 to describe measures of statistical dependence that "give similar scores to equally noisy relationships of different types." Their paper also introduced a new statistic, the "maximal information coefficient" (MIC), that was said to satisfy this equitability criterion. There has since been

the claimed equitability of MIC was only intended to describe a qualitative tendency that they observed when analyzing some data that they themselves simulated. We find this objection of theirs troubling, as it implies that the central claim of ref. 1—that MIC is equitable—was never meant to be falsifiable.

mately satisfy R^2-equitability better than do certain estimates of mutual information. The relevance of these select simulations is unclear. As proven in our paper, neither MIC nor mutual information satisfies R^2-equitability in any mathematical sense. The question of whether estimates of these quantities are approximately R^2-equitable is therefore nei-

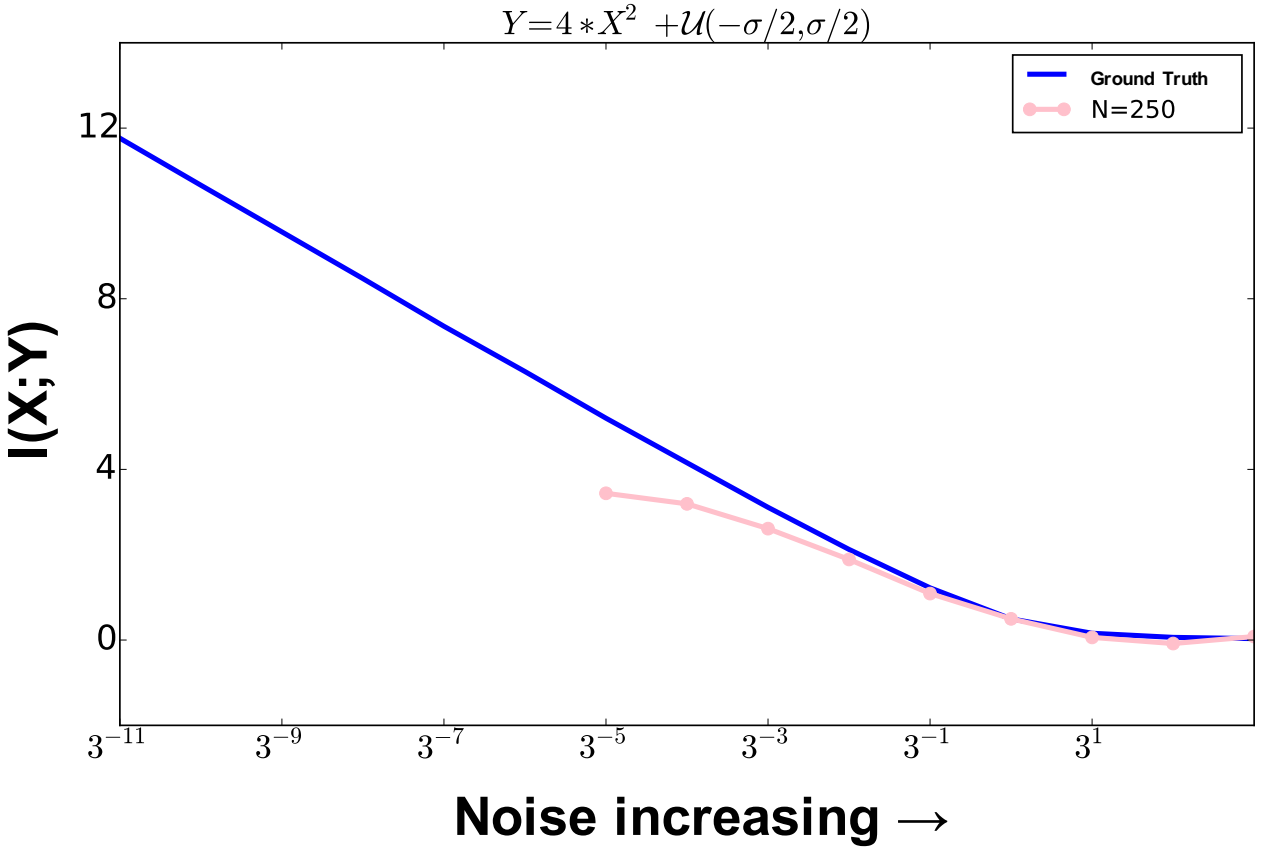
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

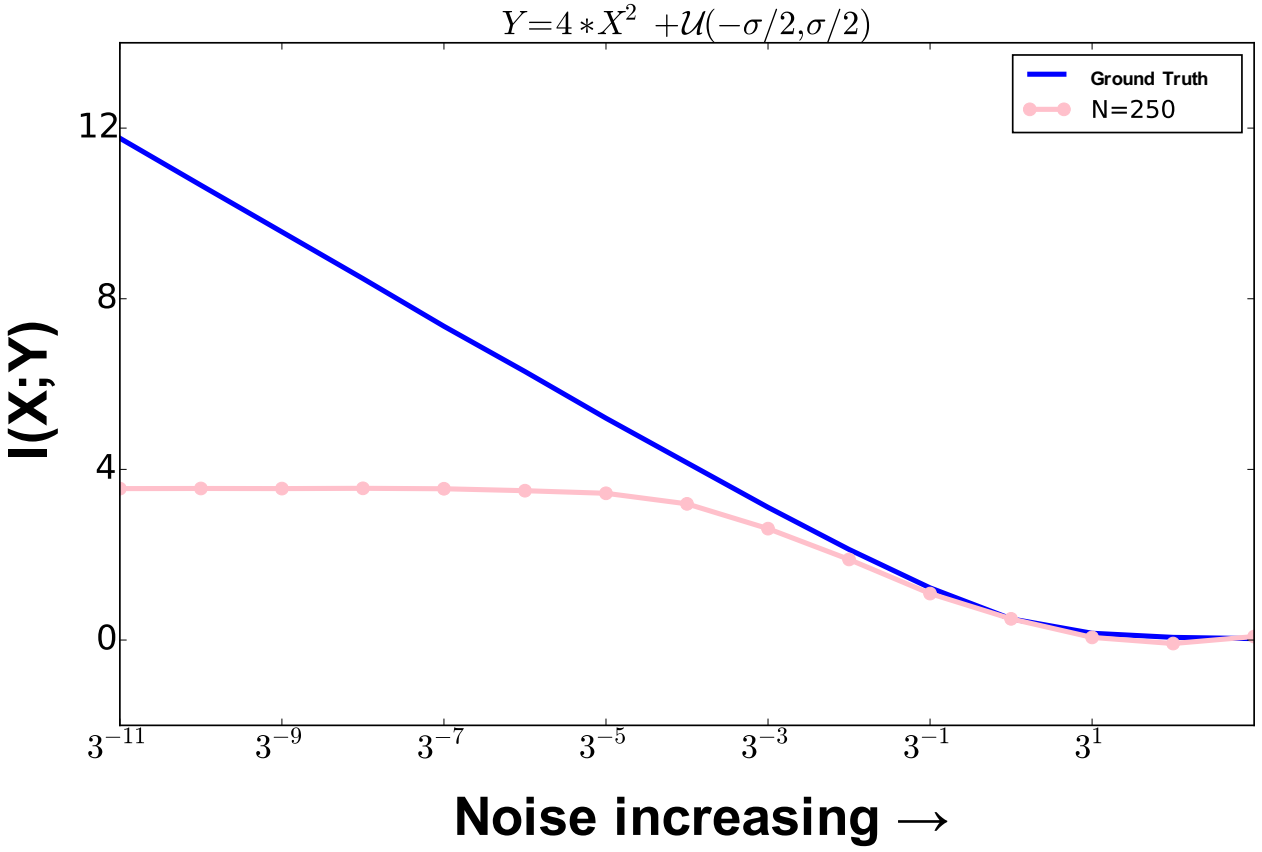
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

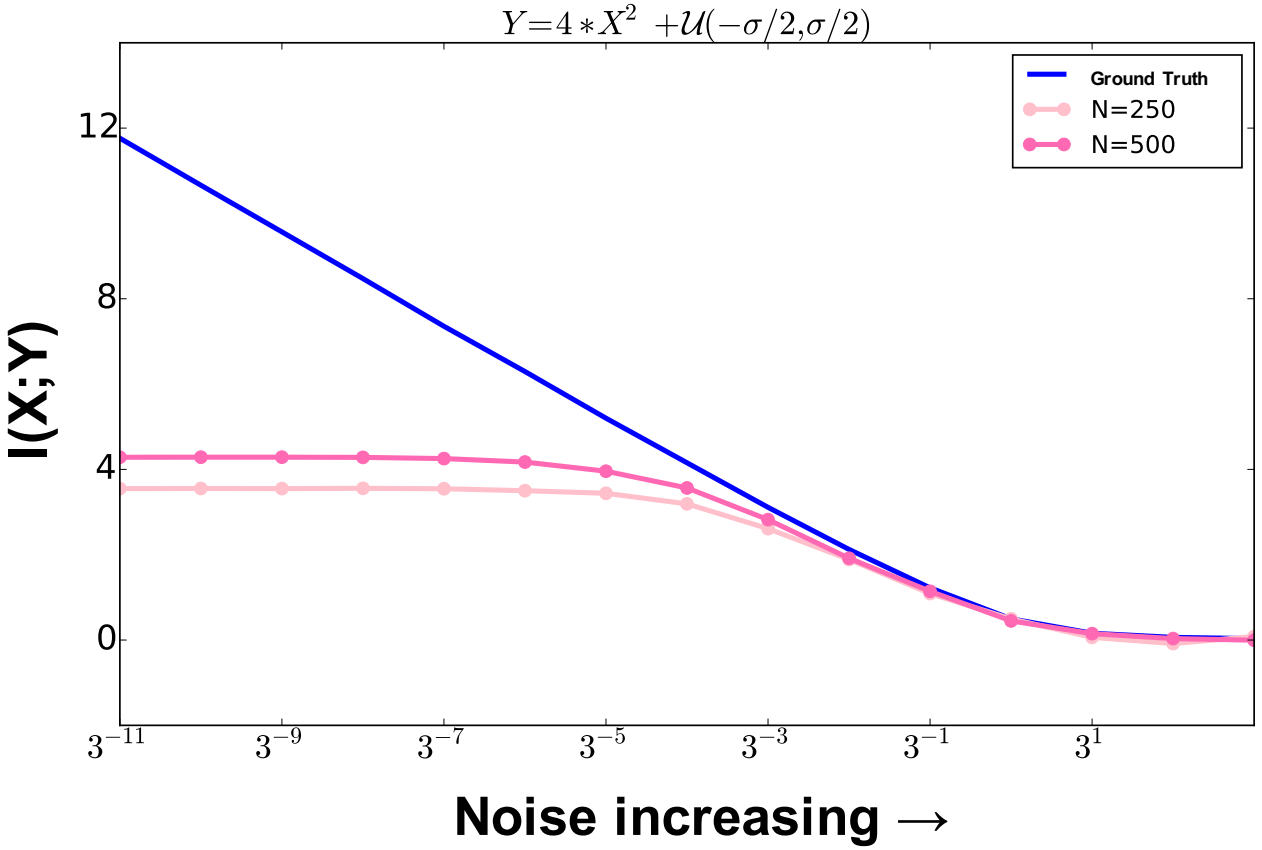
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

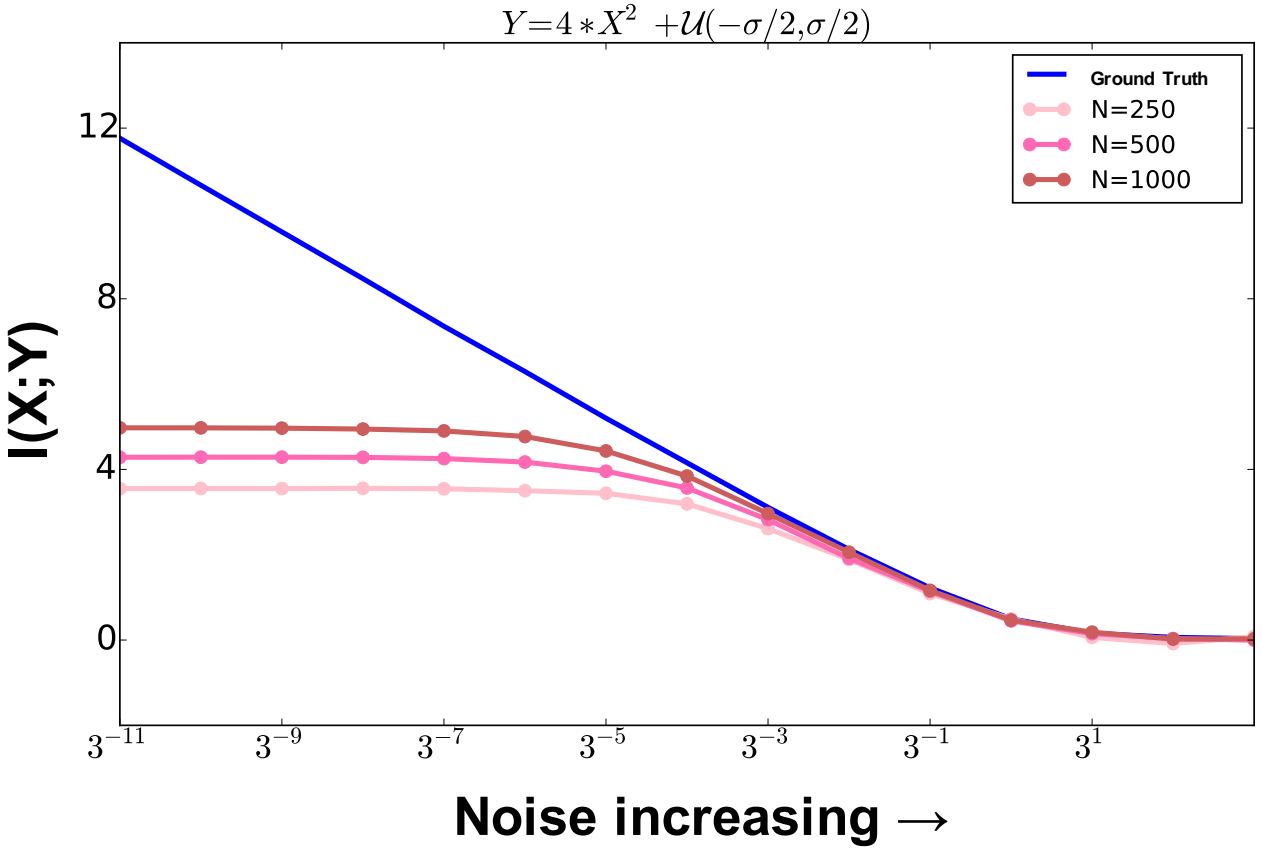
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

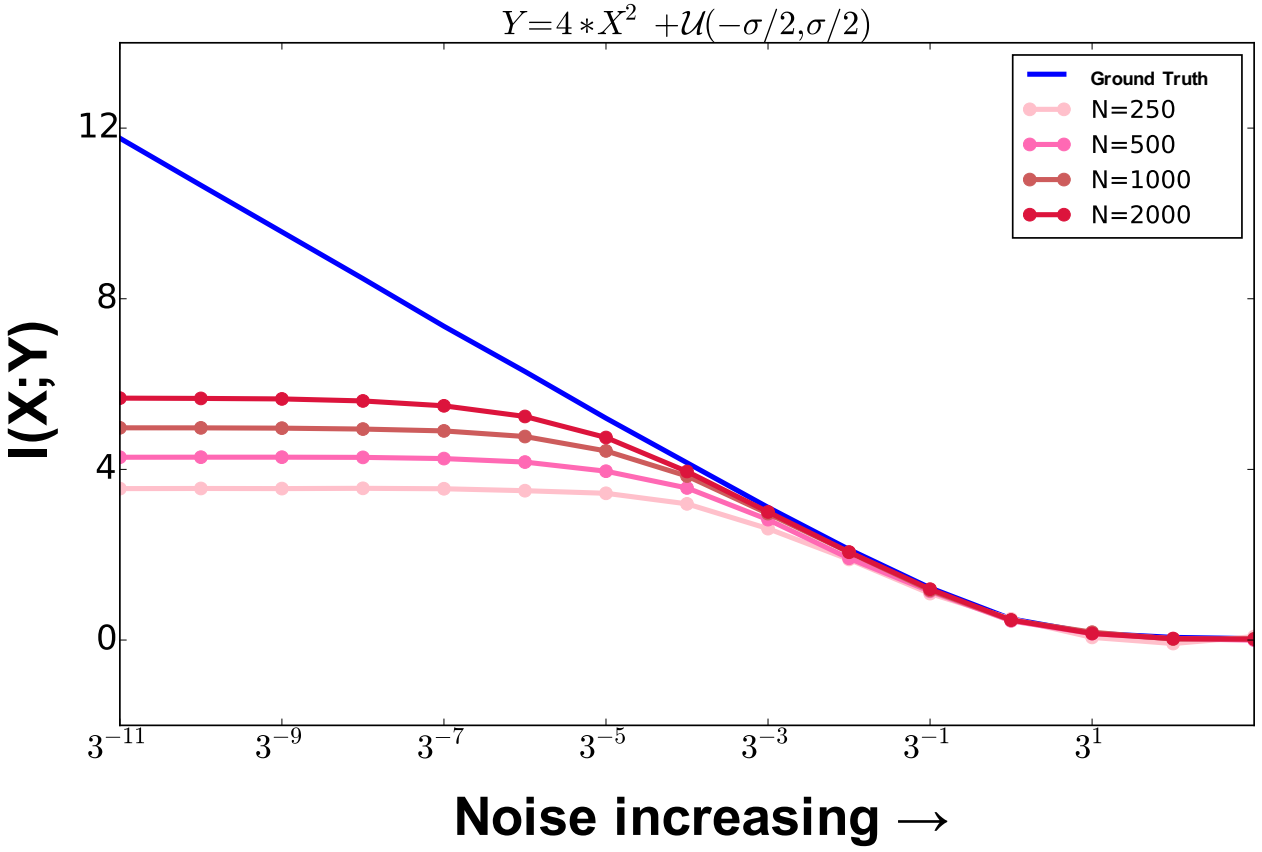
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

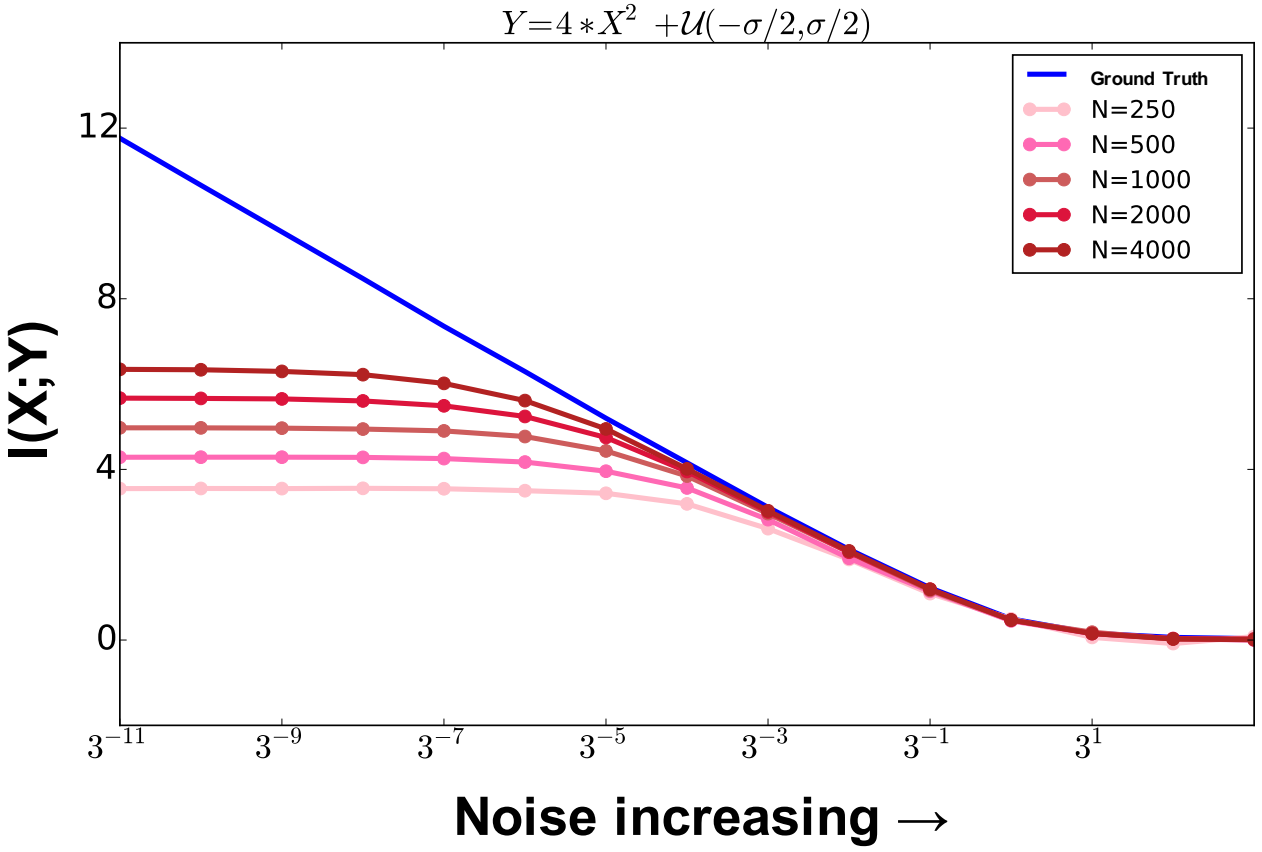
Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

Mutual Information as a Function of Noise



Kraskov, Stögbauer, & Grassberger, Physical Review E, 2004

$$\hat{I}_{KSG,k}(\mathbf{x}) = (d-1)\psi(N) + \psi(k) - (d-1)/k - \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\psi\left(n_{x_j}(i)\right)$$

kNN Estimator Limitations

Theorem [Gao, Ver Steeg, & Galstyan, AISTATS'15]

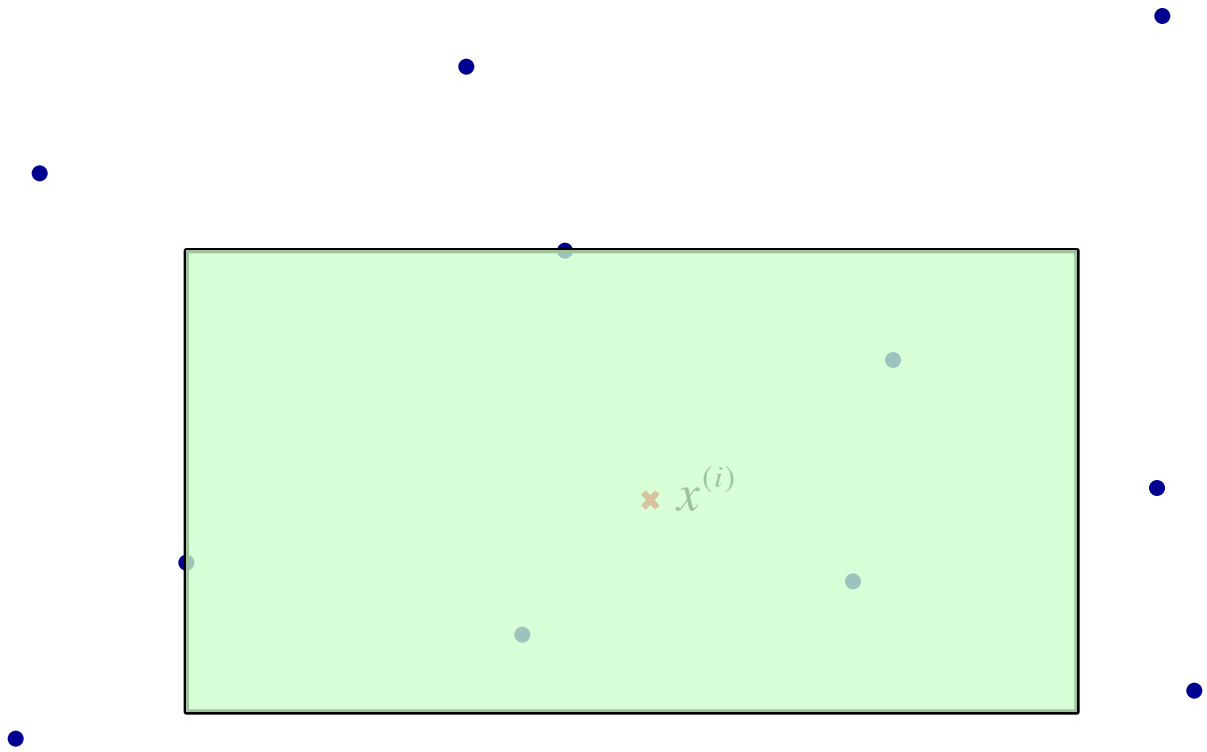
For a certain class of k-NN estimators, estimating mutual information within ε of its true value, $\left|\hat{I}(\mathbf{x}) - I(\mathbf{x})\right| \leq \varepsilon$, requires that the number of samples, N, is at least:

$$N \geq C \exp\left(\frac{I(\mathbf{x}) - \varepsilon}{d - 1}\right) + 1$$

Strong relationships require exponentially many samples to measure

kNN Estimator Limitations

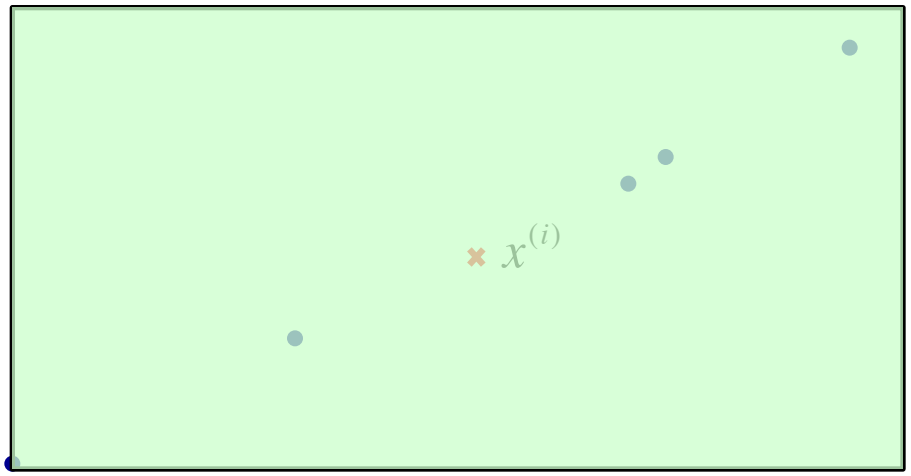
k=5



Works well for weakly correlated distributions

kNN Estimator Limitations

k=5



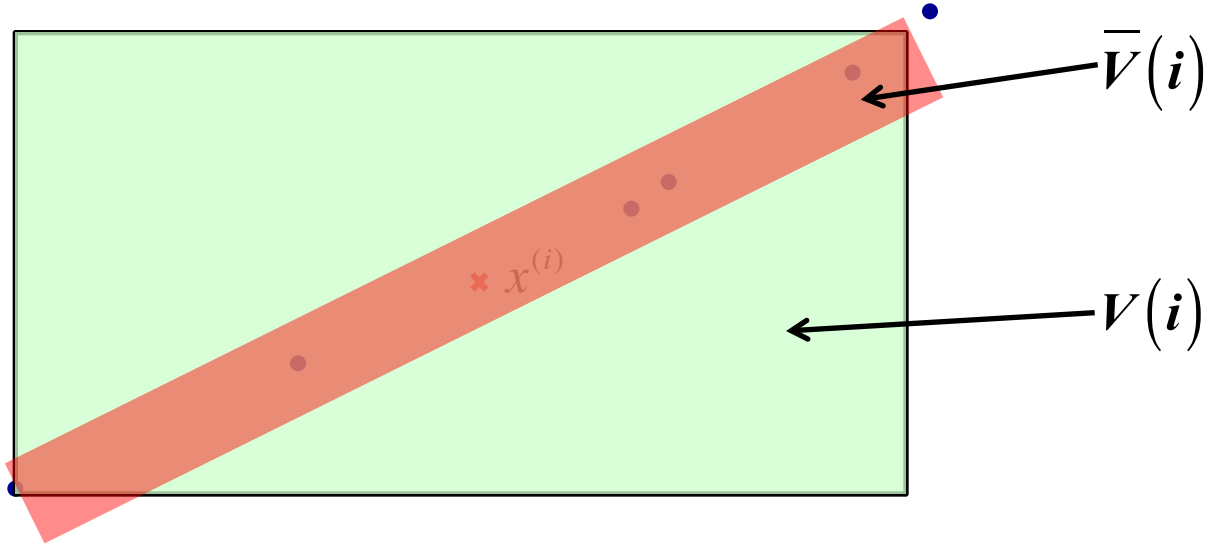
Works bad for strongly correlated distributions

Put a lot more probability mass out of the support

Relax Local Uniformity Condition

k=5

Non-axis aligned bounding rectangle



$$\hat{I}_{LNC}(\mathbf{x}) = \hat{I}(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N} \log \frac{\bar{V}(i)}{V(i)}$$

Local Non-Uniform Correction Algorithm

Algorithm 1 Mutual Information Estimation with Local Nonuniform Correction

$correction = 0$

for each point $\mathbf{x}^{(i)}$ **do**

Find k nearest neighbors of $\mathbf{x}^{(i)}$

Calculate volume of kNN rectangle $V(i)$

Apply PCA on k neighbors, obtain volume $\overline{V}(i)$

if $\overline{V}(i)/V(i) < \alpha_{k,d}$ **then**

$correction = correction + \log \frac{\overline{V}(i)}{V(i)}$

end if

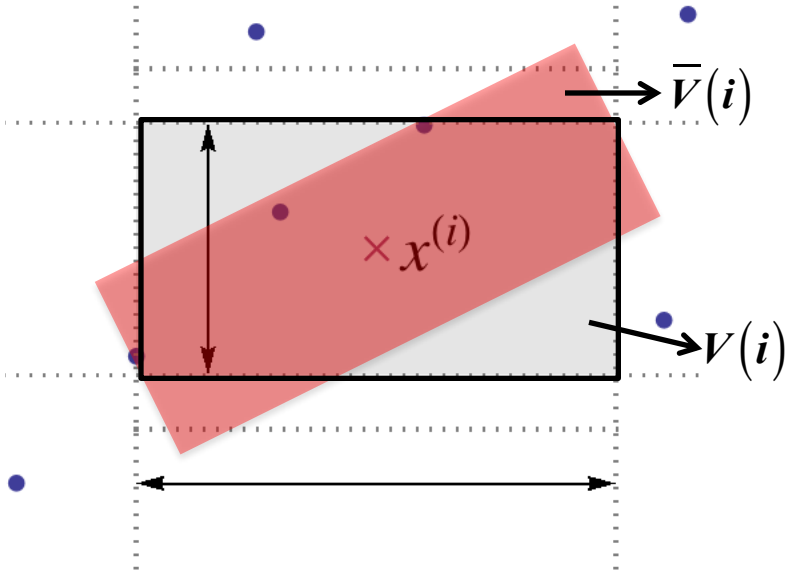
end for

$\hat{I}_{LNC}(\mathbf{x}) = \hat{I}(\mathbf{x}) - \frac{1}{N} * correction$

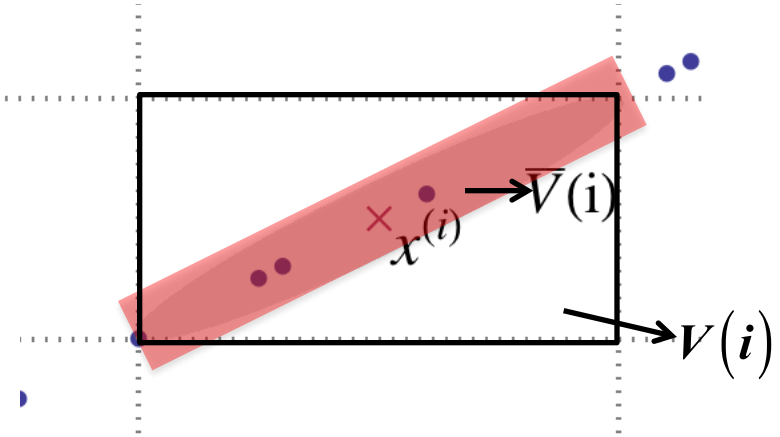
Non-Uniformity Checking

Test for Local Non-Uniformity

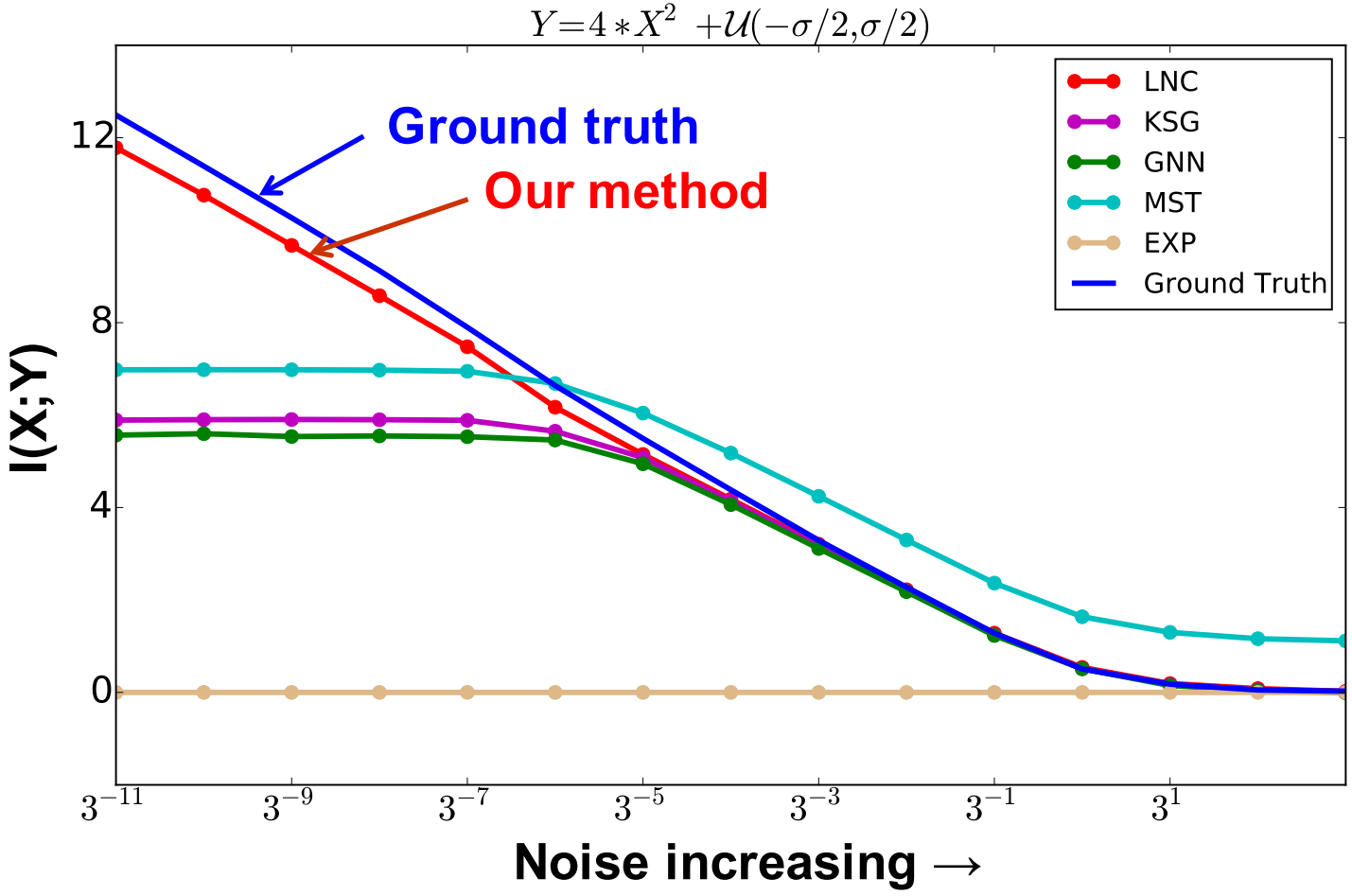
$\overline{V}(i)/V(i) \geq \alpha_{k,d}$



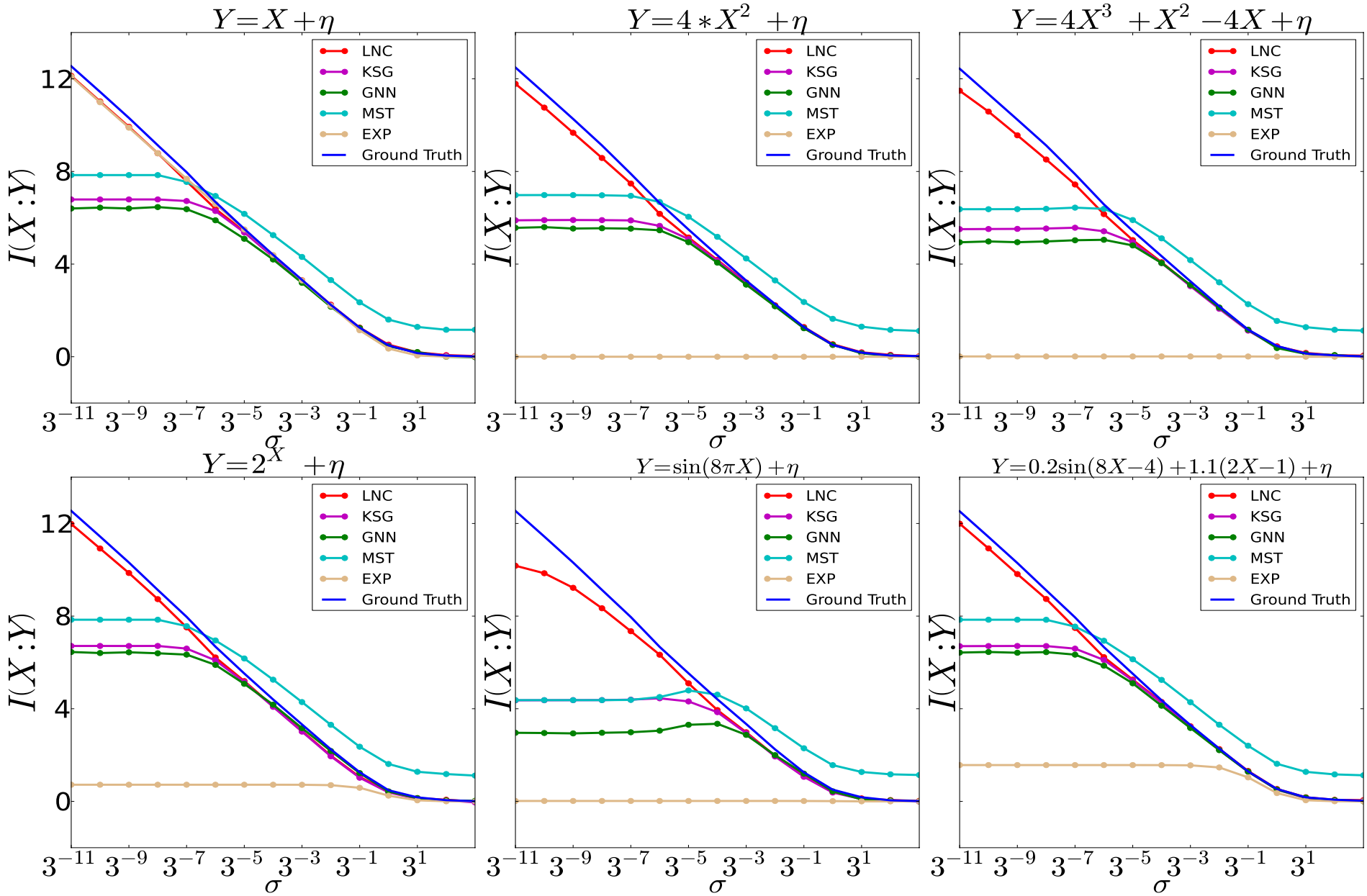
$\overline{V}(i)/V(i) < \alpha_{k,d}$



Functional Relationships



Functional Relationships



Summary

- Information theory is a general but challenging way to measure the strength of relationships
  - Suitable for hard to model domains, like social dynamics
- For medium or low-dimensional problems, careful estimation solves most of our problems
  - Bias correction for discrete data
  - Direct (binless) estimators for continuous signals
- For very high-dimensional systems, we can use information decomposition (CorEx)
  - Learning succinct representations of complex data in an unsupervised way
  - Practical: works on high-d data with few samples and no assumptions about data-generating process

CorEx Info

Contact:
gregv@isi.edu, galstyan@isi.edu

Papers, open source code, interactive visualizations:
http://bit.ly/corex_info