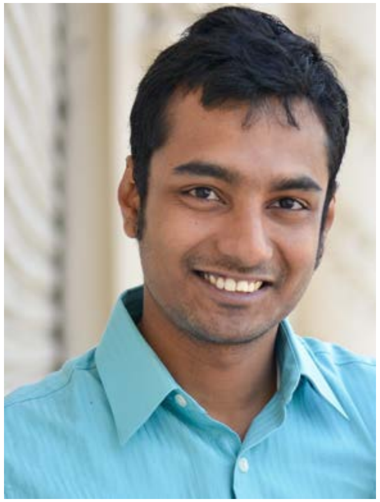
Testing Distribution Properties

Constantinos (Costis) Daskalakis

CSAIL and EECS, MIT



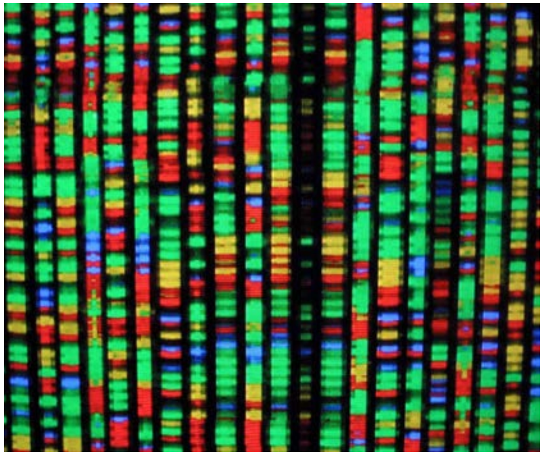
Jayadev Acharya
Cornell

Nishanth Dikkala
CSAIL, MIT

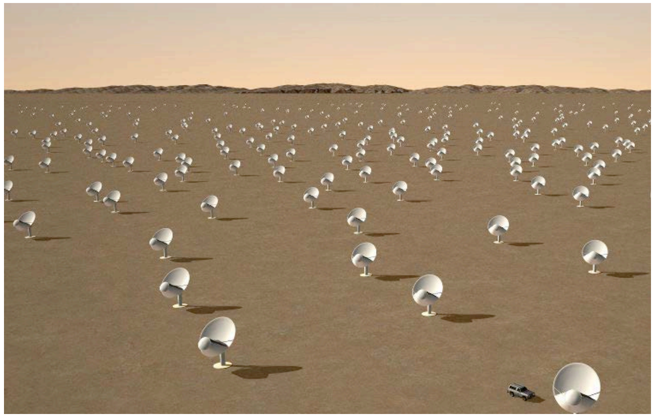
Gautam Kamath
CSAIL, MIT

BIG Data

Facebook: **20 petabytes** images daily



Human genome: **40 exabytes** storage by 2025

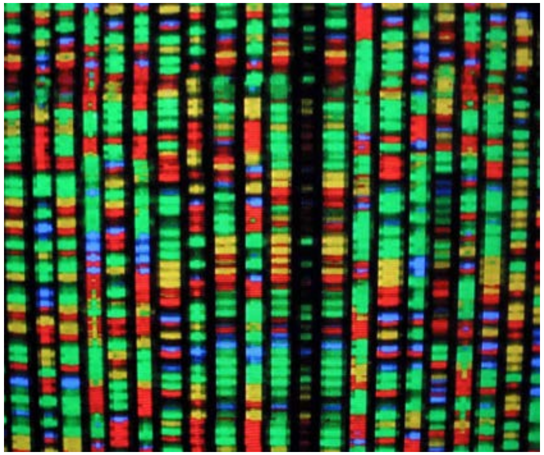


SKA Telescope: **1 exabyte** daily

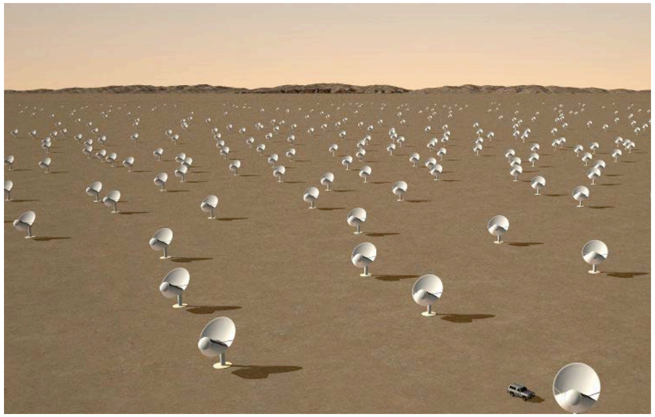
~~BIG~~ Data

small

Facebook: **20 petabytes** images daily



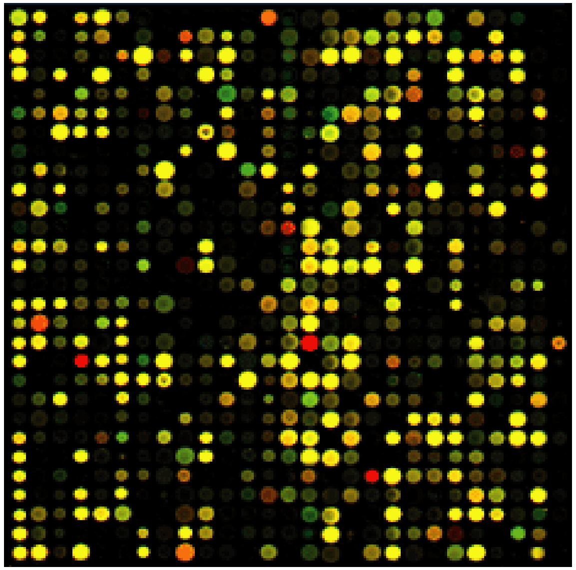
Human genome: **40 exabytes** storage by 2025



SKA Telescope: **1 exabyte** daily

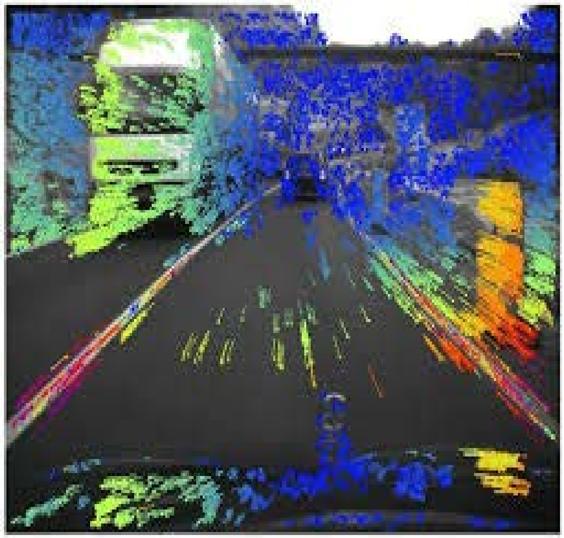
High-dimensional **Expensive**

DNA microarray



Experimental drugs

Computer vision



Financial records

What properties do your BIG distributions have?

e.g.1: play the lottery?

e.g.1: play the lottery?

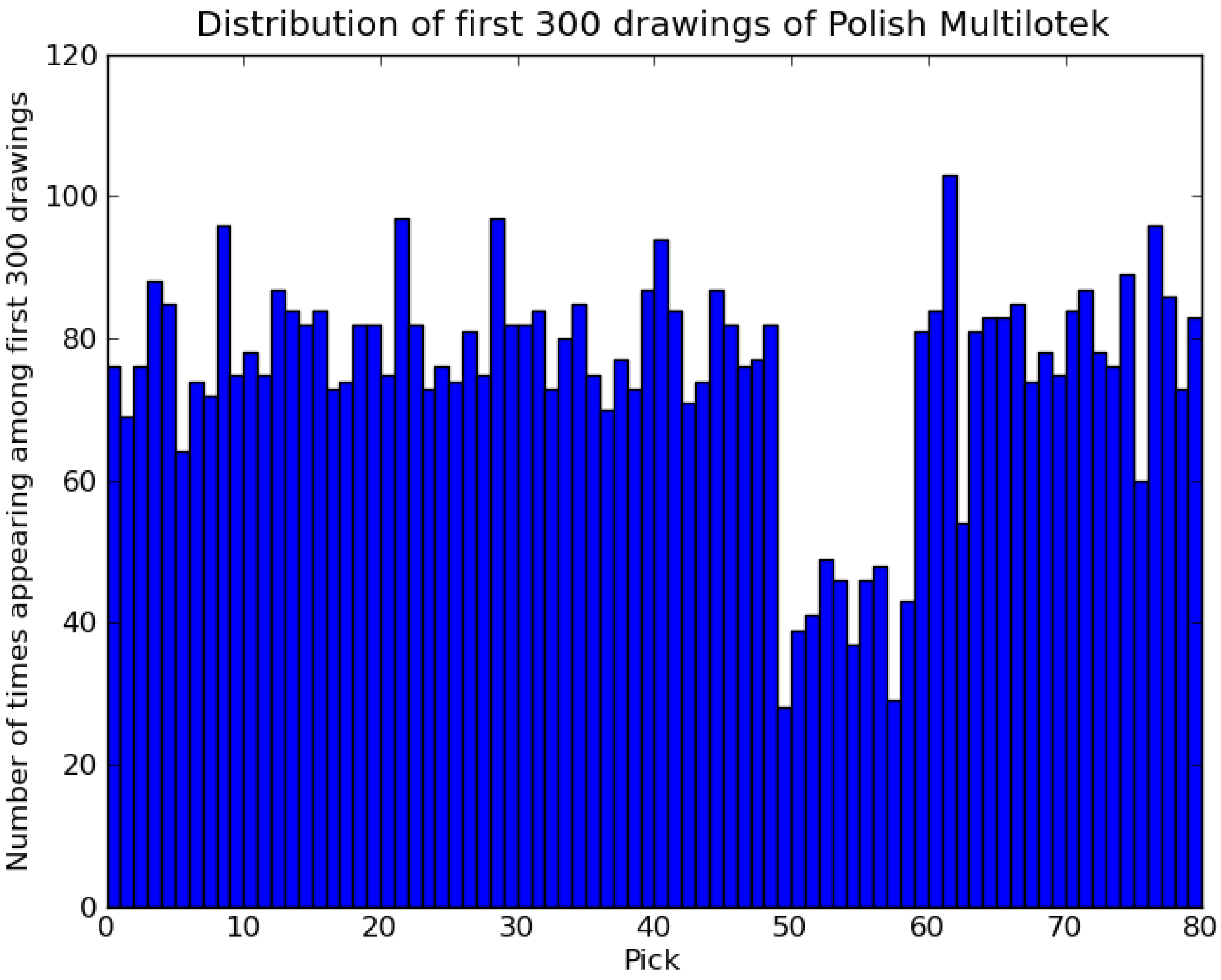
e.g. 1.1: Polish MultiLotek

e.g. 1.1: Polish MultiLotek

- Chooses “uniformly” at random distinct 20 numbers in {1,…,80}.

e.g. 1.1: Polish MultiLotek

- Chooses "uniformly" at random distinct 20 numbers in {1,…,80}.
- Initial machine biased



Thanks to Krzysztof Onak (pointer) and Eric Price (graph)

e.g. 1.2: New Jersey Pick 3,4 Lottery

- New Jersey Pick k ($=3,4$) Lottery.
  - Pick k random digits in order.
  - 10^k possible values.

e.g. 1.2: New Jersey Pick 3,4 Lottery

- New Jersey Pick *k (=3,4)* Lottery.
  - Pick *k* random digits in order.
  - 10^k possible values.
- Data:
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00

e.g. 1.2: New Jersey Pick 3,4 Lottery

- New Jersey Pick *k (=3,4)* Lottery.
  - Pick *k* random digits in order.
  - 10^k possible values.
- Data:
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00
    - χ^2-test (on Excel) answers "42% confidence"

e.g. 1.2: New Jersey Pick 3,4 Lottery

- New Jersey Pick *k (=3,4)* Lottery.
  - Pick *k* random digits in order.
  - 10^k possible values.
- Data:
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00
    - χ^2-test (on Excel) answers "42% confidence"
  - Pick 4 - 6544 results from 9/1/77 to 10/15/00.

e.g. 1.2: New Jersey Pick 3,4 Lottery

- New Jersey Pick *k (=3,4)* Lottery.
  - Pick *k* random digits in order.
  - 10^k possible values.
- Data:
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00
    - χ^2-test (on Excel) answers "42% confidence”
  - Pick 4 - 6544 results from 9/1/77 to 10/15/00.
    - fewer results than possible values
    - not a good idea to run χ^2 test

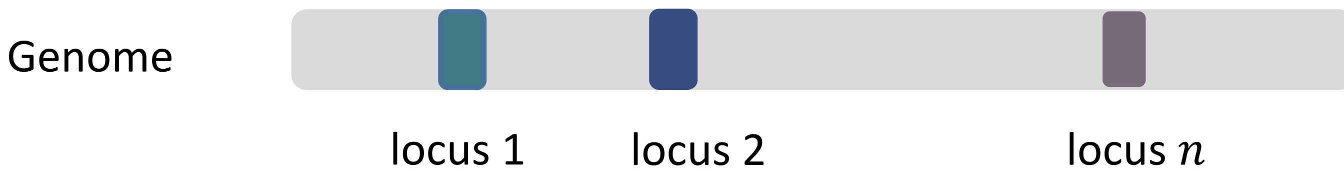
e.g. 1.2: New Jersey Pick 3,4 Lottery

- New Jersey Pick *k (=3,4)* Lottery.
  - Pick *k* random digits in order.
  - 10^k possible values.
- Data:
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00
    - χ^2-test (on Excel) answers "42% confidence"
  - Pick 4 - 6544 results from 9/1/77 to 10/15/00.
    - fewer results than possible values
    - not a good idea to run χ^2 test
    - convergence to χ^2 distribution won't kick in for small sample size

e.g. 1.2: New Jersey Pick 3,4 Lottery

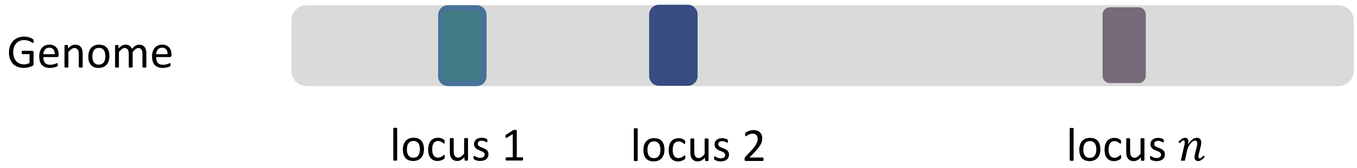
- New Jersey Pick *k (=3,4)* Lottery.
  - Pick *k* random digits in order.
  - 10^k possible values.
- Data:
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00
    - χ^2-test (on Excel) answers "42% confidence"
  - Pick 4 - 6544 results from 9/1/77 to 10/15/00.
    - fewer results than possible values
    - not a good idea to run χ^2 test
    - convergence to χ^2 distribution won't kick in for small sample size
    - **(textbook) rule of thumb:** expected number of at least 5 observations of each element in the domain under the null hypothesis to run χ^2

e.g.2: Linkage Disequilibrium



Single nucleotide polymorphisms, are they independent?

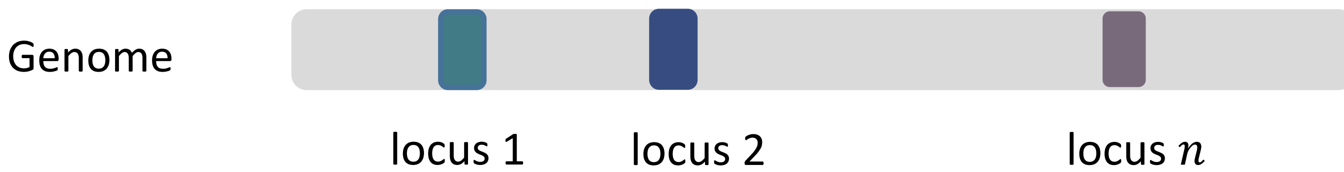
e.g.2: Linkage Disequilibrium



Single nucleotide polymorphisms, are they independent?

Suppose n loci, 2 possible states each, then:

e.g.2: Linkage Disequilibrium

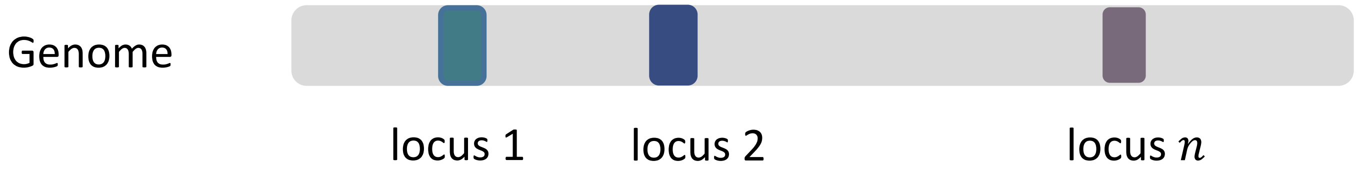


Single nucleotide polymorphisms, are they independent?

Suppose n loci, 2 possible states each, then:

- state of one's genome $\in \{0,1\}^n$

e.g.2: Linkage Disequilibrium

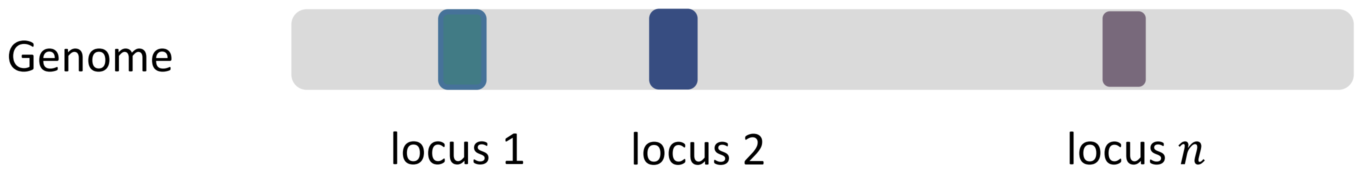


Single nucleotide polymorphisms, are they independent?

Suppose n loci, 2 possible states each, then:

- state of one's genome $\in \{0,1\}^n$
- **humans:** some distribution p over $\{0,1\}^n$

e.g.2: Linkage Disequilibrium



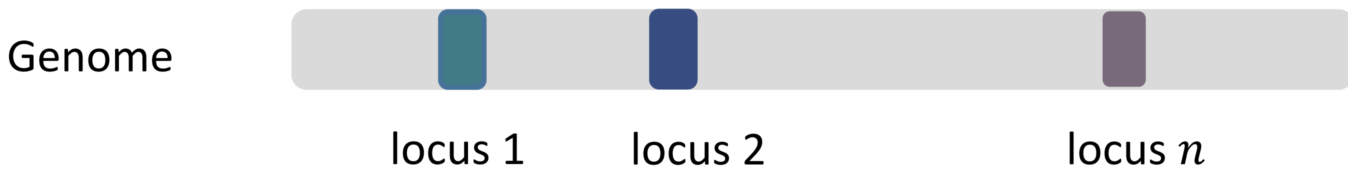
Single nucleotide polymorphisms, are they independent?

Suppose n loci, 2 possible states each, then:

- state of one's genome $\in \{0,1\}^n$
- **humans:** some distribution p over $\{0,1\}^n$

Question: Is p a product dist'n OR ***far*** from all product dist'ns?

e.g.2: Linkage Disequilibrium



Single nucleotide polymorphisms, are they independent?

Suppose n loci, 2 possible states each, then:

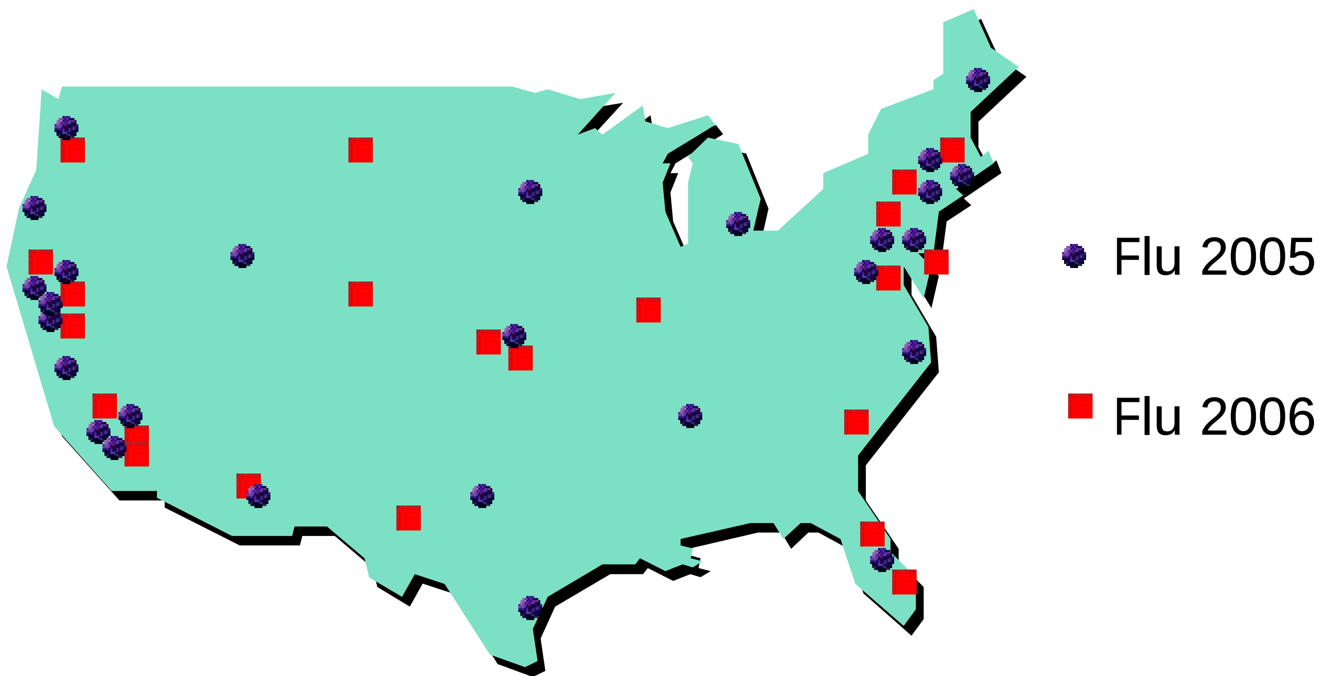
- state of one's genome $\in \{0,1\}^n$
- **humans:** some distribution p over $\{0,1\}^n$

Question: Is p a product dist'n OR ***far*** from all product dist'ns?

should we expect the genomes from the 1000 human genomes project to be sufficient? up to how many loci?

e.g. 3: Outbreak of diseases

- Similar patterns in different years?
- More prevalent near large airports?



Old questions, new challenges

Classical Setting

Modern Setting

Old questions, new challenges

Classical Setting

Domain: 

1000 tosses

Modern Setting

Domain: 

One human genome

Old questions, new challenges

Classical Setting

Domain: 

1000 tosses

Small domain D

$n\ large, |D|\ small$

(comparatively)

Modern Setting

Domain: 

One human genome

Large domain D

$n\ small, |D|\ large$

Old questions, new challenges

Classical Setting

Domain: 

1000 tosses

Small domain D

$n \; large, |D| \; small$

(comparatively)

Asymptotic analysis
Computation **not crucial**

Modern Setting

Domain: 

One human genome

Large domain D

$n \; small, |D| \; large$

Old questions, new challenges

Classical Setting

Domain: 

1000 tosses

Small domain D

$n\ large, |D|\ small$

(comparatively)

Asymptotic analysis
Computation **not crucial**

Modern Setting

Domain: 

One human genome

Large domain D

$n\ small, |D|\ large$

New challenges:

samples, computation, communication, storage

A Key Question

- How many samples do you need in terms of domain size?

A Key Question

- How many samples do you need in terms of domain size?
  - Do you need to estimate the probabilities of each domain item?

  -- OR --

  - Can sample complexity be *sublinear* in size of the domain?

A Key Question

- How many samples do you need in terms of domain size?
  - Do you need to estimate the probabilities of each domain item?

  -- OR --

  - Can sample complexity be *sublinear* in size of the domain?

Rules out standard statistical techniques

The Menu

- **Motivation**
- Problem Formulation
- Uniformity Testing, Goodness of Fit
- Testing Properties of Distributions
- Testing in High Dimensions
- Conclusion

The Menu

- **Motivation**
- **Problem Formulation**
- Uniformity Testing, Goodness of Fit
- Testing Properties of Distributions
- Testing in High Dimensions
- Conclusion

Problem formulation

Model

$\mathcal{P}$: family of distributions over D

Problem formulation

Model

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave

Problem formulation

Model

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave

Problem

Given: samples from **unknown** p

with probability 0.9, distinguish

$$p \in \mathcal{P} \quad \textbf{vs} \quad d(p, \mathcal{P}) > \varepsilon$$

Problem formulation

Model

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave

Problem

Given: samples from **unknown** p

with probability 0.9, distinguish

$$p \in \mathcal{P} \quad \textbf{vs} \quad d(p, \mathcal{P}) > \varepsilon$$

Objective

Minimize samples

Computational efficiency

Problem formulation

discrete

Model

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave

Problem

Given: samples from **unknown** p

with probability 0.9, distinguish

$$p \in \mathcal{P} \quad \textbf{vs} \quad d(p, \mathcal{P}) > \varepsilon$$

Objective

Minimize samples

Computational efficiency

Problem formulation

Model

discrete

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave

Problem

Given: samples from **unknown** p

with probability 0.9, distinguish

$$p \in \mathcal{P} \quad \textbf{vs} \quad d(p, \mathcal{P}) > \varepsilon$$

$$\min_{q \in \mathcal{P}} \frac{\ell_1(p,q)}{2}$$

Objective

Minimize samples

Computational efficiency

Problem formulation

Model

discrete

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave

Problem

Given: samples from **unknown** p

with probability 0.9, distinguish

$$p \in \mathcal{P} \quad \textbf{vs} \quad d(p, \mathcal{P}) > \varepsilon$$

$$\min_{q \in \mathcal{P}} \underbrace{\frac{\ell_1(p,q)}{2}}$$

Objective

Minimize samples

Computational efficiency

$$\max_{events\ \mathcal{E}} |p(\mathcal{E}) - q(\mathcal{E})| \equiv d_{TV}(p,q)$$

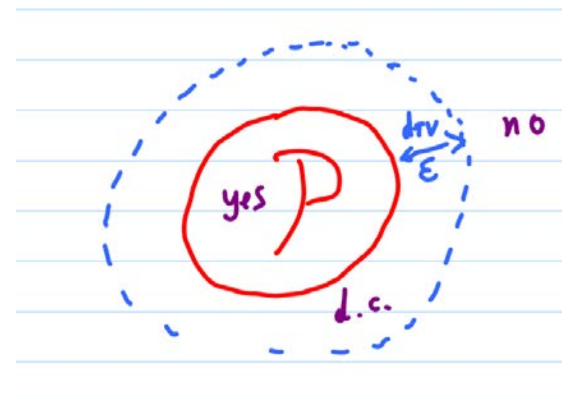
Problem formulation

Model

discrete

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave



Problem

Given: samples from **unknown** p

with probability 0.9, distinguish

$$p \in \mathcal{P} \quad \textbf{vs} \quad d(p, \mathcal{P}) > \varepsilon$$

$$\min_{q \in \mathcal{P}} \underbrace{\frac{\ell_1(p,q)}{2}}$$

Objective

Minimize samples

Computational efficiency

$$\max_{events\ \mathcal{E}} |p(\mathcal{E}) - q(\mathcal{E})| \equiv d_{TV}(p,q)$$

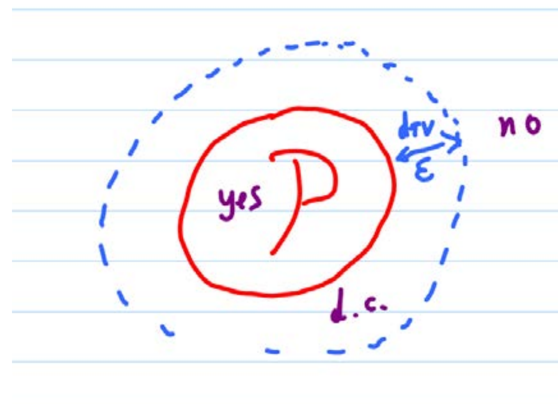
Problem formulation

Model

discrete

$\mathcal{P}$: family of distributions over D

may be non-parametric, e.g. unimodal, product, log-concave



Problem

Given: samples from **unknown** p

with probability 0.9, distinguish

$$p \in \mathcal{P} \quad \textbf{vs} \quad d(p, \mathcal{P}) > \varepsilon$$

$$\min_{q \in \mathcal{P}} \underbrace{\frac{\ell_1(p,q)}{2}}$$

Objective

Minimize samples

Computational efficiency

Sublinear in $|D|$?

$$\max_{events\ \mathcal{E}} |p(\mathcal{E}) - q(\mathcal{E})| \equiv d_{TV}(p,q)$$

Well-studied Problem

(Composite) hypothesis testing

- Neyman-Pearson test
- Kolmogorov-Smirnov test
- Pearson’s chi-squared test
- Generalized likelihood ratio test
- …

Well-studied Problem

(Composite) hypothesis testing

- Neyman-Pearson test
- Kolmogorov-Smirnov test
- Pearson’s chi-squared test
- Generalized likelihood ratio test
- …

Quantities of Interest

$$P_F = \Pr(\boldsymbol{accept}\ when\ hypothesis\ \boldsymbol{false})$$
$$P_M = \Pr(\boldsymbol{reject}\ when\ hypothesis\ \boldsymbol{true})$$

Well-studied Problem

(Composite) hypothesis testing

- Neyman-Pearson test
- Kolmogorov-Smirnov test
- Pearson's chi-squared test
- Generalized likelihood ratio test
- …

Quantities of Interest

$$P_F = \Pr(\boldsymbol{accept}\ when\ hypothesis\ \boldsymbol{false})$$
$$P_M = \Pr(\boldsymbol{reject}\ when\ hypothesis\ \boldsymbol{true})$$

Focus

Consistency

Error exponents: $\exp(-s \cdot R)$ as $s \to \infty$

Well-studied Problem

(Composite) hypothesis testing

- Neyman-Pearson test
- Kolmogorov-Smirnov test
- Pearson's chi-squared test
- Generalized likelihood ratio test
- …

Quantities of Interest

$$P_F = \Pr(\boldsymbol{accept}\ when\ hypothesis\ \boldsymbol{false})$$
$$P_M = \Pr(\boldsymbol{reject}\ when\ hypothesis\ \boldsymbol{true})$$

Focus

Consistency

Error exponents: $\exp(-s \cdot R)$ as $s \to \infty$

Asymptotic regime: Results kick in when $s \gg |D|$

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

Article | Authors | Metrics | Comments | Related Content

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

oblem

$$\textit{thesis } \boldsymbol{false})$$

$$P_M = \Pr(\boldsymbol{reject} \textit{ when hypothesis } \boldsymbol{true})$$

Focus

Consistency

Error exponents: $\exp(-s \cdot R)$ as $s \to \infty$

Asymptotic regime: Results kick in when $s \gg |D|$

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

roblem

Article

Abstract

Modeling the Framew
for False Positive
Findings

Bias

Testing by Several
Independent Teams

Corollaries

Most Research Findin
Are False for Most
Research Designs and
for Most Fields

Claimed Research
Findings May Often Be
Simply Accurate
Measures of the
Prevailing Bias

Study delivers bleak verdict on validity of psychology experiment results

Of 100 studies published in top-ranking journals in 2008, 75% of social psychology experiments and half of cognitive studies failed the replication test

Psychology experiments are failing the replication test - for good reason

There are many reasons why an experiment might fail to replicate, but more than this, the study has highlighted some issues with academic publishing and modern science. Photograph: Pere Sanz / Alamy/Alamy

A major investigation into scores of claims made in psychology research journals has delivered a bleak verdict on the state of the science.

$esis$ $\boldsymbol{false}$)

$esis$ $\boldsymbol{true}$)

Focu

C

E

Asymptotic regime: Results kick in when $s \gg |D|$

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

·oblem

Article

Abstract

Modeling the Framew for False Positive Findings

Bias

Testing by Several Independent Teams

Corollaries

Most Research Findin Are False for Most Research Designs anc for Most Fields

Claimed Research Findings May Often B Simply Accurate Measures of the Prevailing Bias

Study delivers bleak verdict on validity of psychology experiment results

Of 100 studies published in top-ranking journals in 2008, 75% of social psychology experiments and half of cognitive studies failed the replication test

Psychology experiments are failing the replication test - for good reason

There are many reasons why an experiment might fail to replicate, but more than this, the study has highlighted some issues with academic publishing and modern science. Photograph: Pere Sanz / Alamy/Alamy

A major investigation into scores of claims made in psychology research journals has delivered a bleak verdict on the state of the science.

esis $\boldsymbol{false}$)

esis $\boldsymbol{true}$)

Focu **- Sublinear in $|D|$?**

C

E

Asymptotic regime: Results kick in when $s \gg |D|$

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • http://dx.doi.org/10.1371/journal.pmed.0020124

Article

·oblem

Study delivers bleak verdict on validity of psychology experiment results

Of 100 studies published in top-ranking journals in 2008, 75% of social psychology experiments and half of cognitive studies failed the replication test

Psychology experiments are failing the replication test – for good reason

There are many reasons why an experiment might fail to replicate, but more than this, the study has highlighted some issues with academic publishing and modern science. Photograph: Pere Sanz / Alamy/Alamy

A major investigation into scores of claims made in psychology research journals has delivered a bleak verdict on the state of the science.

esis ***false***)

esis ***true***)

Focu

C

E

- Sublinear in $|D|$?

- Strong control for false negatives?

Asymptotic regime: Results kick in when $s \gg |D|$

The Menu

- **Motivation**
- **Problem Formulation**
- Uniformity Testing, Goodness of Fit
- Testing Properties of Distributions
- Testing in High Dimensions
- Conclusion

The Menu

- **Motivation**
- **Problem Formulation**
- **Uniformity Testing**, Goodness of Fit
- Testing Properties of Distributions
- Testing in High Dimensions
- Conclusion

Testing the Fairness of a Coin

- $\boldsymbol{b}$: unknown probability of a 
- **Question:** Is $\boldsymbol{b} = 0.5$ OR $|\boldsymbol{b} - 0.5| \geq \epsilon$?
- **Goal:** Toss coin several times, deduce correct answer w/ prob. ≥ 0.99

Testing the Fairness of a Coin

- $\mathcal{P} = \left\{\text{Bernoulli}\left(\frac{1}{2}\right)\right\}$
- $p = \text{Bernoulli}(\boldsymbol{b})$
- $p \in \mathcal{P}$ **vs** $d_{\text{TV}}(p, \mathcal{P}) > \varepsilon$

- $\boldsymbol{b}$: unknown probability of a 
- **Question:** Is $\boldsymbol{b} = 0.5$ OR $|\boldsymbol{b} - 0.5| \geq \epsilon$?
- **Goal:** Toss coin several times, deduce correct answer w/ prob. ≥ 0.99

Testing the Fairness of a Coin

- $\mathcal{P} = \left\{\text{Bernoulli}\left(\frac{1}{2}\right)\right\}$
- $p = \text{Bernoulli}(\boldsymbol{b})$
- $p \in \mathcal{P}$ **vs** $d_{\text{TV}}(p, \mathcal{P}) > \varepsilon$

- $\boldsymbol{b}$: unknown probability of a 
- **Question:** Is $\boldsymbol{b} = 0.5$ OR $|\boldsymbol{b} - 0.5| \geq \epsilon$?
- **Goal:** Toss coin several times, deduce correct answer w/ prob. ≥ 0.99
- Number of samples required?
  - Tight answer $\Theta\left(\frac{1}{\epsilon^2}\right)$

Testing the Fairness of a Coin

- $\mathcal{P} = \left\{\text{Bernoulli}\left(\frac{1}{2}\right)\right\}$
- $p = \text{Bernoulli}(\boldsymbol{b})$
- $p \in \mathcal{P}$ **vs** $d_{\text{TV}}(p, \mathcal{P}) > \varepsilon$

- $\boldsymbol{b}$: unknown probability of a 
- **Question:** Is $\boldsymbol{b} = 0.5$ OR $|\boldsymbol{b} - 0.5| \geq \epsilon$?
- **Goal:** Toss coin several times, deduce correct answer w/ prob. ≥ 0.99
- Number of samples required?
  - Tight answer $\Theta\left(\frac{1}{\epsilon^2}\right)$
  - **Upper bound:** compare average to 0.5, reject if farther than $\frac{\epsilon}{2}$

Testing the Fairness of a Coin

- $\mathcal{P} = \left\{\text{Bernoulli}\left(\frac{1}{2}\right)\right\}$
- $p = \text{Bernoulli}(\boldsymbol{b})$
- $p \in \mathcal{P}$ **vs** $d_{\text{TV}}(p, \mathcal{P}) > \varepsilon$

- $\boldsymbol{b}$: unknown probability of a 
- **Question:** Is $\boldsymbol{b} = 0.5$ OR $|\boldsymbol{b} - 0.5| \geq \epsilon$?
- **Goal:** Toss coin several times, deduce correct answer w/ prob. ≥ 0.99
- Number of samples required?
  - Tight answer $\Theta\left(\frac{1}{\epsilon^2}\right)$
  - **Upper bound:** compare average to 0.5, reject if farther than $\frac{\epsilon}{2}$
  - **Lower bound:** a sleek one uses the subadditivity of Hellinger^2 distance

Testing Uniformity

- p: unknown distribution over D
  - sample access to p
- **Question:** is $p = U_D$ or $d_{\mathrm{TV}}(p, U_D) > \epsilon$?

Testing Uniformity

- p: unknown distribution over D
  - sample access to p
- **Question:** is $p = U_D$ or $d_{\mathrm{TV}}(p, U_D) > \epsilon$?

- $\mathcal{P} = \{\mathrm{Uniform}_D\}$
- p : unknown
- $p \in \mathcal{P}$ **vs** $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$

Testing Uniformity

- $\mathcal{P} = \{\text{Uniform}_D\}$
- p : unknown
- $p \in \mathcal{P}$ **vs** $d_{\text{TV}}(p, \mathcal{P}) > \varepsilon$

- p: unknown distribution over D
  - sample access to p
- **Question:** is $p = U_D$ or $d_{\text{TV}}(p, U_D) > \epsilon$?
- **[Paninski'03]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time

Testing Uniformity

- p: unknown distribution over D
  - sample access to p
- **Question:** is $p = U_D$ or $d_{\mathrm{TV}}(p, U_D) > \epsilon$?
- **[Paninski'03]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time

- $\mathcal{P} = \{\mathrm{Uniform}_D\}$
- p : unknown
- $p \in \mathcal{P}$ **vs** $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$

"Intuition:"

- **Lower Bound:** Suppose q is uniform distribution over $\{1, \ldots, m\}$ and p is uniform on random $m/2$ size subset of $\{1, \ldots, m\}$

Testing Uniformity

- p: unknown distribution over D
  - sample access to p
- **Question:** is $p = U_D$ or $d_{\mathrm{TV}}(p, U_D) > \epsilon$?
- **[Paninski'03]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time

- $\mathcal{P} = \{\mathrm{Uniform}_D\}$
- p : unknown
- $p \in \mathcal{P}$ **vs** $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$

"Intuition:"

- **Lower Bound:** Suppose q is uniform distribution over $\{1, \ldots, m\}$ and p is uniform on random $m/2$ size subset of $\{1, \ldots, m\}$
  - Unless $\Omega(\sqrt{m})$ samples are observed, there are no collisions, hence cannot distinguish between q or p chosen as above

Testing Uniformity

- p: unknown distribution over D
  - sample access to p
- **Question:** is $p = U_D$ or $d_{\mathrm{TV}}(p, U_D) > \epsilon$?
- **[Paninski'03]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time

"Intuition:"

- **Lower Bound:** Suppose q is uniform distribution over $\{1, \ldots, m\}$ and p is uniform on random $m/2$ size subset of $\{1, \ldots, m\}$
  - Unless $\Omega(\sqrt{m})$ samples are observed, there are no collisions, hence cannot distinguish between q or p chosen as above
- **Upper Bound:** Collision statistics suffice to distinguish

The Menu

- **Motivation**
- **Problem Formulation**
- **Uniformity Testing, Goodness of Fit**
- Testing Properties of Distributions
- Testing in High Dimensions
- Conclusion

Identity Testing ("goodness of fit")

- p, q: distributions over D
  - q: given; sample access to p
- **Question:** is $p = q$ or $d_{\mathrm{TV}}(p, q) > \epsilon$?

Identity Testing ("goodness of fit")

- p, q: distributions over D
  - q: given; sample access to p
- **Question:** is $p = q$ or $d_{\mathrm{TV}}(p, q) > \epsilon$?

Identity Testing (“goodness of fit”)

- p, q: distributions over D
  - q: given; sample access to p
- **Question:** is $p = q$ or $d_{\mathrm{TV}}(p, q) > \epsilon$?

> - $\mathcal{P} = \{q\}$
> - p : unknown
> - $p \in \mathcal{P}$ **vs** $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$

- **[Batu-Fisher-Fortnow-Kumar-Rubinfeld-White’01]…**
- **[Paninski’08, Valiant-Valiant’14]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time

Identity Testing ("goodness of fit")

- p, q: distributions over D
  - q: given; sample access to p
- **Question:** is $p = q$ or $d_{\mathrm{TV}}(p, q) > \epsilon$?

- $\mathcal{P} = \{q\}$
- p : unknown
- $p \in \mathcal{P}$ **vs** $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$

- **[Batu-Fisher-Fortnow-Kumar-Rubinfeld-White'01]…**
- **[Paninski'08, Valiant-Valiant'14]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time
- **[w/ Acharya-Kamath NIPS'15]:** a *tolerant* goodness of fit test with same sample size can distinguish: $\chi^2(p, q) \le \epsilon^2/2$ vs $\ell_1^2(p, q) > \epsilon^2$

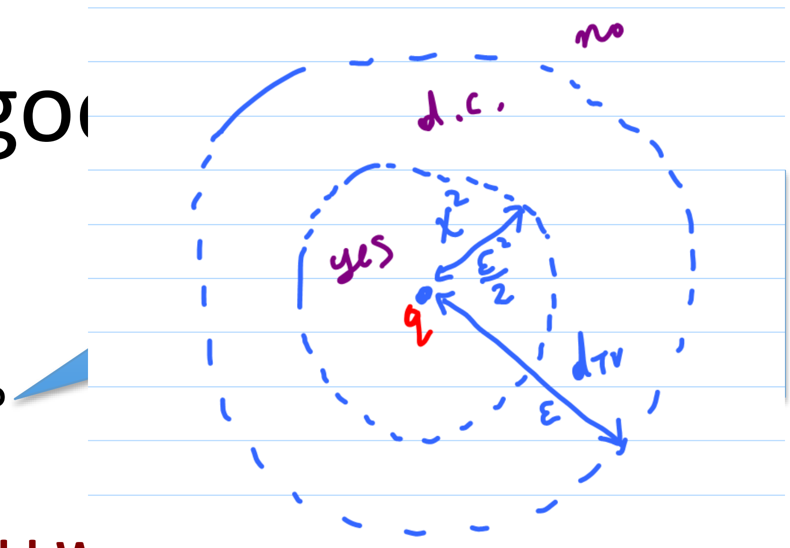
Identity Testing ("goodness of fit")

- $\mathcal{P} = \{q\}$
- p : unknown
- $p \in \mathcal{P}$ **vs** $d_{\mathrm{TV}}(p, \mathcal{P}) > \varepsilon$

- p, q: distributions over D
  - q: given; sample access to p
- **Question:** is $p = q$ or $d_{\mathrm{TV}}(p, q) > \epsilon$?

- **[Batu-Fisher-Fortnow-Kumar-Rubinfeld-White'01]…**
- **[Paninski'08, Valiant-Valiant'14]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time
- **[w/ Acharya-Kamath NIPS'15]:** a *tolerant* goodness of fit test with same sample size can distinguish: $\chi^2(p, q) \leq \epsilon^2/2$ vs $\ell_1^2(p, q) > \epsilon^2$
  - $\chi^2(p, q) \coloneqq \sum_{i \in D} \frac{(p_i - q_i)^2}{q_i}$
  - Cauchy-Schwarz: $\chi^2(p, q) \geq \ell_1(p, q)^2$

Identity Testing (“go



- p, q: distributions over D
  - q: given; sample access to p
- **Question:** is $p = q$ or $d_{\mathrm{TV}}(p, q) > \epsilon$?

- **[Batu-Fisher-Fortnow-Kumar-Rubinfeld-W**
- **[Paninski’08, Valiant-Valiant’14]:** $\Theta\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time
- **[w/ Acharya-Kamath NIPS’15]:** a *tolerant* goodness of fit test with same sample size can distinguish: $\chi^2(p, q) \leq \epsilon^2/2$ vs $\ell_1^2(p, q) > \epsilon^2$
  - $\chi^2(p, q) := \sum_{i \in D} \frac{(p_i - q_i)^2}{q_i}$
  - Cauchy-Schwarz: $\chi^2(p, q) \geq \ell_1(p, q)^2$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:
  Case 1: $\chi^2(p, q) \leq \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p, q) \geq \epsilon^2$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:
  Case 1: $\chi^2(p,q) \le \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \ge \epsilon^2$
- **Approach:** Draw m samples from p
  - N_i: # of appearances of symbol $i \in D$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:
  Case 1: $\chi^2(p,q) \leq \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \geq \epsilon^2$
- **Approach:** Draw m samples from p
  – N_i: # of appearances of symbol $i \in D$

- **Statistic:** $Z = \sum_i \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:
  Case 1: $\chi^2(p,q) \leq \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \geq \epsilon^2$
- **Approach:** Draw m samples from p
  - N_i: # of appearances of symbol $i \in D$
- **Statistic:** $Z = \sum_i \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$

Want:

- Z small $\rightarrow$ Case 1
- Z large $\rightarrow$ Case 2

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:
  Case 1: $\chi^2(p,q) \le \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \ge \epsilon^2$
- **Approach:** Draw $\mathbf{Poisson}(m)$ **many** samples from p
  - N_i: # of appearances of symbol $i \in D$
- **Statistic:** $Z = \sum_i \frac{(N_i - m\cdot q_i)^2 - N_i}{m\cdot q_i}$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:
  Case 1: $\chi^2(p,q) \le \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \ge \epsilon^2$
- **Approach:** Draw **$\mathbf{Poisson}(m)$ many** samples from p
  - N_i: # of appearances of symbol $i \in D$
  - $N_i \sim \mathrm{Poisson}(m \cdot p_i)$
  - $(N_i)_{i\in D}$ independent random variables
- **Statistic:** $Z = \sum_i \frac{(N_i - m\cdot q_i)^2 - N_i}{m\cdot q_i}$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:

  Case 1: $\chi^2(p,q) \le \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \ge \epsilon^2$
- **Approach:** Draw **$\mathbf{Poisson}(m)$ many** samples from p
  - N_i: # of appearances of symbol $i \in D$
  - $N_i \sim \mathrm{Poisson}(m \cdot p_i)$
  - $(N_i)_{i \in D}$ independent random variables
- **Statistic:** $Z = \sum_i \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$
  - $E[Z] = m \cdot \chi^2(p,q)$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:

  Case 1: $\chi^2(p,q) \leq \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \geq \epsilon^2$
- **Approach:** Draw **Poisson(m) many** samples from p
  - N_i: # of appearances of symbol $i \in D$
  - $N_i \sim \text{Poisson}(m \cdot p_i)$
  - $(N_i)_{i \in D}$ independent random variables
- **Statistic:** $Z = \sum_i \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$
  - $E[Z] = m \cdot \chi^2(p,q)$
  - *Case 1:* $E[Z] \leq m \cdot \epsilon^2/2$; *Case 2:* $E[Z] \geq m \cdot \epsilon^2$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:

  Case 1: $\chi^2(p, q) \le \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p, q) \ge \epsilon^2$
- **Approach:** Draw **Poisson(m) many** samples from p
  - N_i: # of appearances of symbol $i \in D$
  - $N_i \sim \text{Poisson}(m \cdot p_i)$
  - $(N_i)_{i \in D}$ independent random variables
- **Statistic:** $Z = \sum_i \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$
  - $E[Z] = m \cdot \chi^2(p, q)$
  - *Case 1:* $E[Z] \le m \cdot \epsilon^2/2$; *Case 2:* $E[Z] \ge m \cdot \epsilon^2$
  - chug chug chug…bound variance of Z

    $\to O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples suffice to distinguish

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:

  Case 1: $\chi^2(p,q) \le \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \ge \epsilon^2$
- **Approach:** Draw **Poisson(m) many** samples from p
  - N_i: # of appearances of symbol $i \in D$
  - $N_i \sim \text{Poisson}(m \cdot p_i)$
  - $(N_i)_{i\in D}$ independent random variables
- **Statistic:** $Z = \sum_i \frac{(N_i - m\cdot q_i)^2 - N_i}{m\cdot q_i}$
  - $E[Z] = m \cdot \chi^2(p,q)$
  - *Case 1:* $E[Z] \le m \cdot \epsilon^2/2$; *Case 2:* $E[Z] \ge m \cdot \epsilon^2$
  - chug chug chug…bound variance of Z

    $\rightarrow O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples suffice to distinguish

Side-Note:

- Pearson's χ^2 test uses statistic $\sum_i \frac{(N_i - m\cdot q_i)^2}{m\cdot q_i}$

An Improved χ^2- Test

- **Goal:** given q and sample access to p distinguish:

  Case 1: $\chi^2(p,q) \le \epsilon^2/2$ vs *Case 2:* $\ell_1^2(p,q) \ge \epsilon^2$

- **Approach:** Draw **Poisson(m) many** samples from p
  - N_i: # of appearances of symbol $i \in D$
  - $N_i \sim \text{Poisson}(m \cdot p_i)$
  - $(N_i)_{i \in D}$ independent random variables

- **Statistic:** $Z = \sum_i \frac{(N_i - m \cdot q_i)^2 - N_i}{m \cdot q_i}$
  - $E[Z] = m \cdot \chi^2(p,q)$
  - *Case 1:* $E[Z] \le m \cdot \epsilon^2/2$; *Case 2:* $E[Z] \ge m \cdot \epsilon^2$
  - chug chug chug…bound variance of Z

  $\rightarrow O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples suffice to distinguish

Side-Note:

- Pearson's χ^2 test uses statistic $\sum_i \frac{(N_i - m \cdot q_i)^2}{m \cdot q_i}$
- Subtracting N_i in the numerator gives an unbiased estimator and importantly may hugely decrease variance

The Menu

- **Motivation**
- **Problem Formulation**
- **Uniformity Testing, Goodness of Fit**
- Testing Properties of Distributions
- Testing in High Dimensions
- Conclusion

The Menu

- **Motivation**
- **Problem Formulation**
- **Uniformity Testing, Goodness of Fit**
- **Testing Properties of Distributions**
- Testing in High Dimensions
- Conclusion

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly
- **natural extension:** test structural properties

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly
- **natural extension:** test structural properties
  - monotonicity: "PDF is monotone," e.g. cancer vs radiation

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly
- **natural extension:** test structural properties
  - monotonicity: "PDF is monotone," e.g. cancer vs radiation
  - unimodality: "PDF is single-peaked," e.g. single source of disease

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly
- **natural extension:** test structural properties
  - monotonicity: "PDF is monotone," e.g. cancer vs radiation
  - unimodality: "PDF is single-peaked," e.g. single source of disease
  - log-concavity: "$\log \mathrm{PDF}$ is concave"

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly
- **natural extension:** test structural properties
  - monotonicity: “PDF is monotone,” e.g. cancer vs radiation
  - unimodality: “PDF is single-peaked,” e.g. single source of disease
  - log-concavity: “$\log \mathrm{PDF}$ is concave”
  - monotone-hazard rate: “$\log\left(1-\mathrm{CDF}\right)$ is concave”

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly
- **natural extension:** test structural properties
  - monotonicity: "PDF is monotone," e.g. cancer vs radiation
  - unimodality: "PDF is single-peaked," e.g. single source of disease
  - log-concavity: "$\log \mathrm{PDF}$ is concave"
  - monotone-hazard rate: "$\log\left(1-\mathrm{CDF}\right)$ is concave"
  - product distribution, e.g. testing linkage disequilibrium

Testing Properties of Distributions

- so far $\mathcal{P}$ ={single distribution}
  - **restrictive**, as rarely know hypothesis distribution exactly
- **natural extension:** test structural properties
  - monotonicity: "PDF is monotone," e.g. cancer vs radiation
  - unimodality: "PDF is single-peaked," e.g. single source of disease
  - log-concavity: "$\log \mathrm{PDF}$ is concave"
  - monotone-hazard rate: "$\log(1-\mathrm{CDF})$ is concave"
  - product distribution, e.g. testing linkage disequilibrium
- Example question:
  - $\mathcal{P} = \{unimodal\ distributions\ over\ [m]\}$
  - Sample access to p
  - Is p unimodal OR is p ϵ-far from or unimodal distributions?

Testing Properties of Distributions

[w/ Acharya and Kamath NIPS'15]:

1. Testing identity, monotonicity, log-concavity, monotone hazard-rate, unimodality for distributions over (ordered set) D **is doable** w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.

Testing Properties of Distributions

[w/ Acharya and Kamath NIPS'15]:

1. Testing identity, monotonicity, log-concavity, monotone hazard-rate, unimodality for distributions over (ordered set) D **is doable** w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
2. Testing monotonicity/independence of a distribution over $D = [m]^d$ **is doable** w/ $O\left(\frac{m^{d/2}}{\epsilon^2}\right) \equiv O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
   - previous best for monotonicity testing: $\tilde{O}\left(\frac{m^{d-0.5}}{\epsilon^4}\right)$ **[Bhattacharya-Fisher-Rubinfeld-Valiant'11]**
   - previous best for independence: d=2, worse bounds **[Batu et al.'01]**

Testing Properties of Distributions

[w/ Acharya and Kamath NIPS'15]:

1. Testing identity, monotonicity, log-concavity, monotone hazard-rate, unimodality for distributions over (ordered set) D **is doable** w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
2. Testing monotonicity/independence of a distribution over $D = [m]^d$ **is doable** w/ $O\left(\frac{m^{d/2}}{\epsilon^2}\right) \equiv O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
   - previous best for monotonicity testing: $\tilde{O}\left(\frac{m^{d-0.5}}{\epsilon^4}\right)$ **[Bhattacharya-Fisher-Rubinfeld-Valiant'11]**
   - previous best for independence: d=2, worse bounds **[Batu et al.'01]**
3. All bounds above are **optimal**
   - i.e. matching lower bounds for both 1 and 2 via Le Cam Inequality.

Testing Properties of Distributions

[w/ Acharya and Kamath NIPS'15]:

1. Testing identity, monotonicity, log-concavity, monotone hazard-rate, unimodality for distributions over (ordered set) D **is doable** w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
2. Testing monotonicity/independence of a distribution over $D = [m]^d$ **is doable** w/ $O\left(\frac{m^{d/2}}{\epsilon^2}\right) \equiv O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
   - previous best for monotonicity testing: $\tilde{O}\left(\frac{m^{d-0.5}}{\epsilon^4}\right)$ **[Bhattacharya-Fisher-Rubinfeld-Valiant'11]**
   - previous best for independence: d=2, worse bounds **[Batu et al.'01]**
3. All bounds above are **optimal**
   - i.e. matching lower bounds for both 1 and 2 via Le Cam Inequality.
4. Unified approach, computationally efficient tests, based on new χ^2-tolerant tester

Testing Properties of Distributions

[w/ Acharya and Kamath NIPS'15]:

1. Testing identity, monotonicity, log-concavity, monotone hazard-rate, unimodality for distributions over (ordered set) D **is doable** w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
2. Testing monotonicity/independence of a distribution over $D = [m]^d$ **is doable** w/ $O\left(\frac{m^{d/2}}{\epsilon^2}\right) \equiv O\left(\frac{\sqrt{|D|}}{\epsilon^2}\right)$ samples and time.
   - previous best for monotonicity testing: $\tilde{O}\left(\frac{m^{d-0.5}}{\epsilon^4}\right)$ **[Bhattacharya-Fisher-Rubinfeld-Valiant'11]**
   - previous best for independence: d=2, worst bounds **[Batu et al.'01]**
3. All bounds above are **optimal**
   - i.e. matching lower bounds for both 1 and 2 via Le Cam Inequality.
4. Unified approach, computationally efficient tests, based on new χ^2-tolerant tester

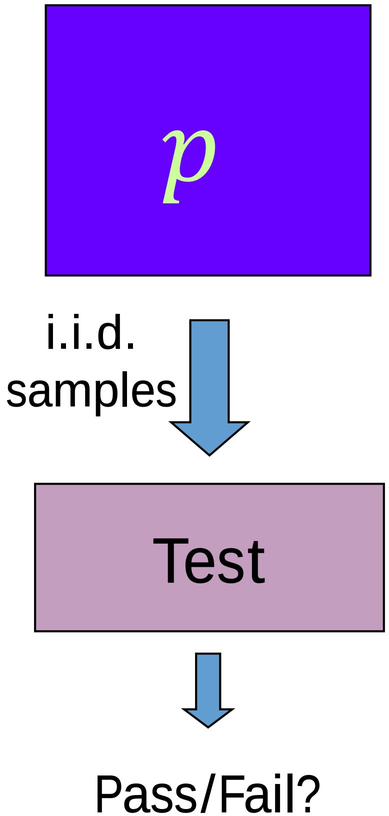
N.B. Contemporaneous work of **[Canonne et al'2015]** provide a different unified approach for testing structure but their results are suboptimal.

Summary so far

- Hypothesis Testing in the small sample regime.

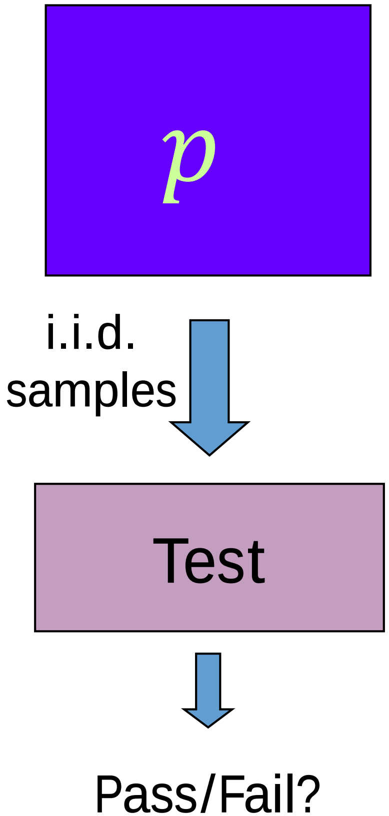
Summary so far

- Hypothesis Testing in the small sample regime.



Summary so far

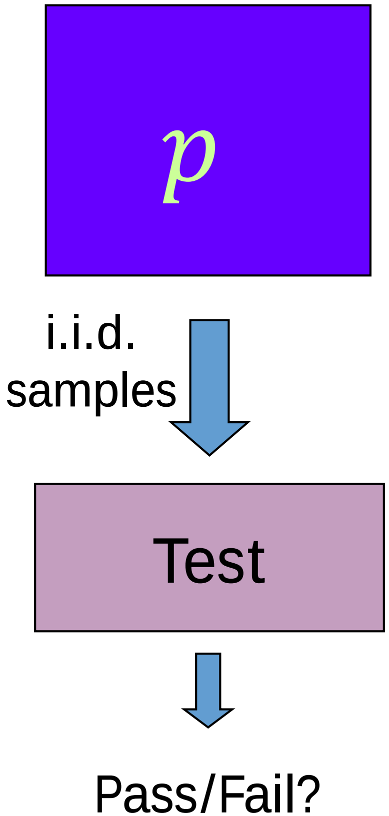
- Hypothesis Testing in the small sample regime.



- p *unknown* distribution over some *discrete* set D
- $\mathcal{P}$: set of distributions over D
- **Given:** ϵ, δ, sample access to p
- **Goal:** w/ prob $\geq 1 - \delta$ tell $p \in \mathcal{P}$ vs $\ell_1(p, \mathcal{P}) > \epsilon$

Summary so far

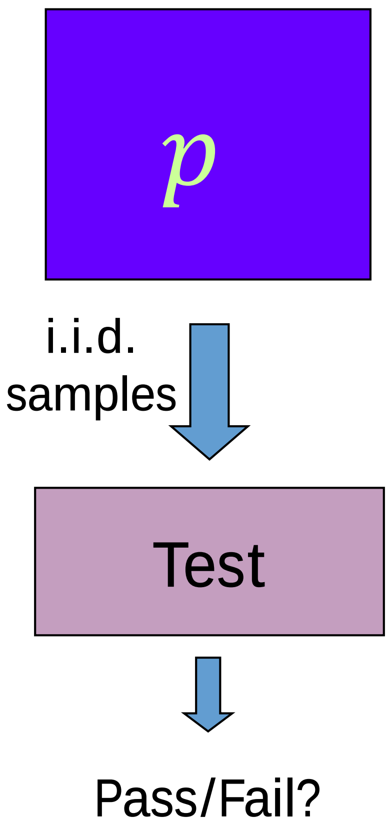
- Hypothesis Testing in the small sample regime.



- p *unknown* distribution over some *discrete* set D
- $\mathcal{P}$: set of distributions over D
- **Given:** ϵ, $\boldsymbol{\delta}$, sample access to p
- **Goal:** w/ prob $\geq 1 - \boldsymbol{\delta}$ tell $p \in \mathcal{P}$ vs $\ell_1(p, \mathcal{P}) > \epsilon$
- Properties of interest: Is p uniform? unimodal? log-concave? MHR? product measure?

Summary so far

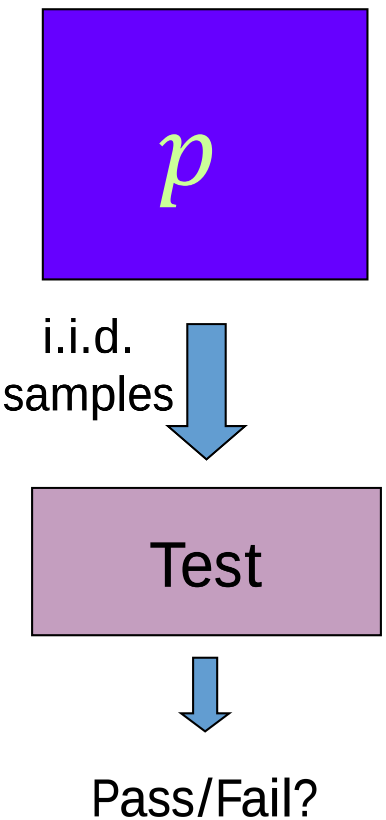
- Hypothesis Testing in the small sample regime.



- p *unknown* distribution over some *discrete* set D
- $\mathcal{P}$: set of distributions over D
- **Given:** ϵ, $\boldsymbol{\delta}$, sample access to p
- **Goal:** w/ prob $\geq 1-\boldsymbol{\delta}$ tell $p \in \mathcal{P}$ vs $\ell_1(p, \mathcal{P}) > \epsilon$
- Properties of interest: Is p uniform? unimodal? log-concave? MHR? product measure?
- All above properties can be tested w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2} \cdot \log\frac{1}{\delta}\right)$ samples and time

Summary so far

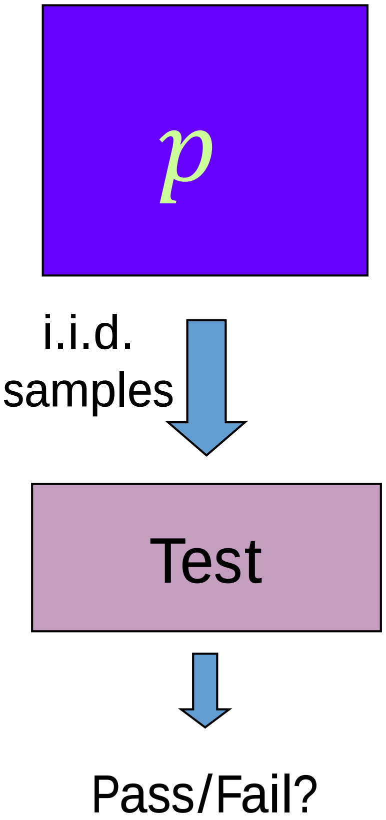
- Hypothesis Testing in the small sample regime.



- p *unknown* distribution over some *discrete* set D
- $\mathcal{P}$: set of distributions over D
- **Given:** ϵ, $\boldsymbol{\delta}$, sample access to p
- **Goal:** w/ prob $\geq 1 - \boldsymbol{\delta}$ tell $p \in \mathcal{P}$ vs $\ell_1(p, \mathcal{P}) > \epsilon$
- Properties of interest: Is p uniform? unimodal? log-concave? MHR? product measure?
- All above properties can be tested w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2} \cdot \log\frac{1}{\delta}\right)$ samples and time
- Unified approach based on ***modified*** Pearson's goodness of fit test: statistic $Z = \sum_{i \in D} \frac{(N_i - E_i)^2 - N_i}{E_i}$

Summary so far

- Hypothesis Testing in the small sample regime.



- p *unknown* distribution over some *discrete* set D
- $\mathcal{P}$: set of distributions over D
- **Given:** ϵ, $\boldsymbol{\delta}$, sample access to p
- **Goal:** w/ prob $\geq 1 - \boldsymbol{\delta}$ tell $p \in \mathcal{P}$ vs $\ell_1(p, \mathcal{P}) > \epsilon$
- Properties of interest: Is p uniform? unimodal? log-concave? MHR? product measure?
- All above properties can be tested w/ $O\left(\frac{\sqrt{|D|}}{\epsilon^2} \cdot \log\frac{1}{\delta}\right)$ samples and time
- Unified approach based on ***modified*** Pearson's goodness of fit test: statistic $Z = \sum_{i \in D} \frac{(N_i - E_i)^2 - N_i}{E_i}$
  - tight control for false positives: want to be able to both assert and reject the null hypothesis
  - accommodate sublinear sample size

The Menu

- **Motivation**
- **Problem Formulation**
- **Uniformity Testing, Goodness of Fit**
- **Testing Properties of Distributions**
- Testing in High Dimensions
- Conclusion

The Menu

Motivation

Problem Formulation

Uniformity Testing, Goodness of Fit

Testing Properties of Distributions

Testing in High Dimensions

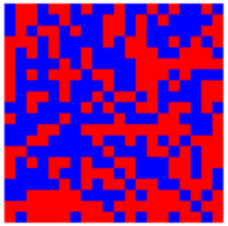
Conclusion

High-Dimensional Distn’s

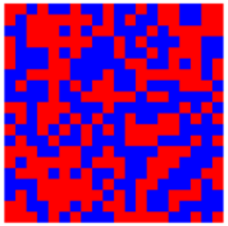
- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …

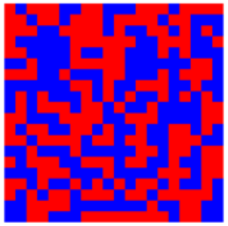
High-Dimensional Distn's

- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …



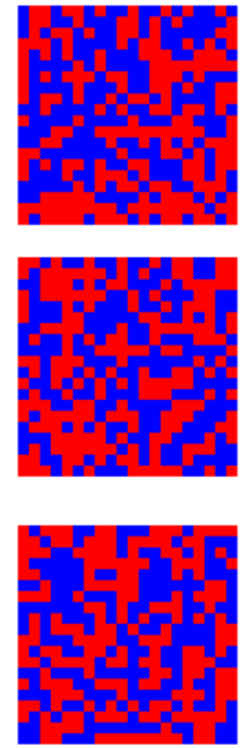
400 bit images





High-Dimensional Distn’s

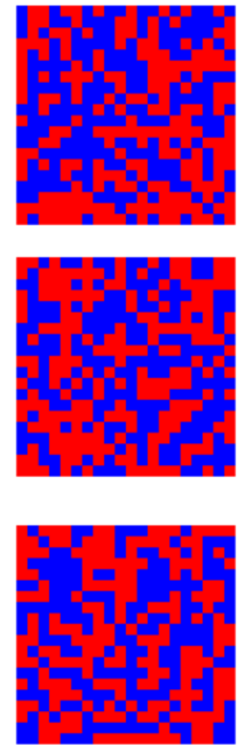
- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …
- Are bits/pixels independent?
  - Our algorithms require $\Theta\left(\frac{2^{n/2}}{\epsilon^2}\right)$ samples



400 bit images

High-Dimensional Distn’s

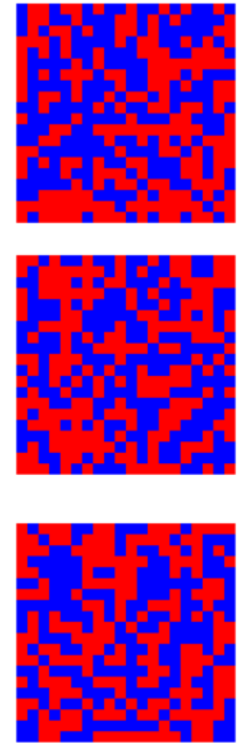
- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …
- Are bits/pixels independent?
  - Our algorithms require $\Theta\left(\frac{2^{n/2}}{\epsilon^2}\right)$ samples
- Is source generating graphs over n nodes Erdos-Renyi $G\left(n, \frac{1}{2}\right)$?
  - Our algorithms require $\Theta\left(\frac{2^{\binom{n}{2}/2}}{\epsilon^2}\right)$ samples



400 bit images

High-Dimensional Distn's

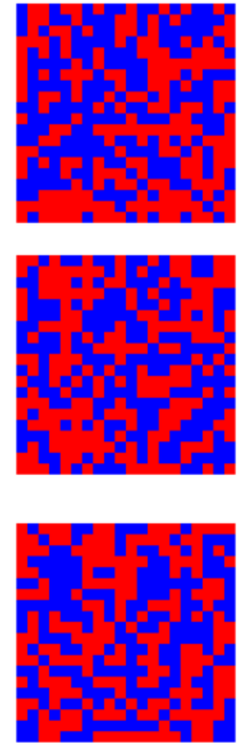
- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …
- Are bits/pixels independent?
  - Our algorithms require $\Theta\left(\frac{2^{n/2}}{\epsilon^2}\right)$ samples
- Is source generating graphs over n nodes Erdos-Renyi $G\left(n, \frac{1}{2}\right)$?
  - Our algorithms require $\Theta\left(\frac{2^{\binom{n}{2}/2}}{\epsilon^2}\right)$ samples
- Exponential dependence on n unsettling, but necessary
  - Lower bound exploits high possible correlation among bits



400 bit images

High-Dimensional Distn's

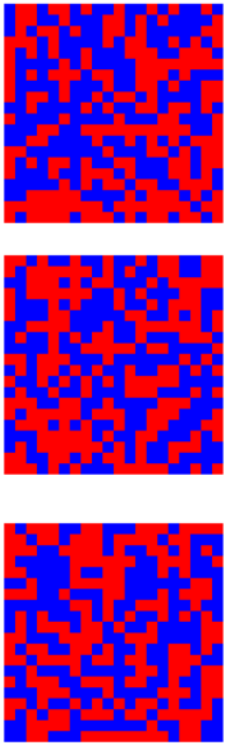
- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …
- Are bits/pixels independent?
  - Our algorithms require $\Theta\left(\frac{2^{n/2}}{\epsilon^2}\right)$ samples
- Is source generating graphs over n nodes Erdos-Renyi $G\left(n, \frac{1}{2}\right)$?
  - Our algorithms require $\Theta\left(\frac{2^{\binom{n}{2}/2}}{\epsilon^2}\right)$ samples
- Exponential dependence on n unsettling, but necessary
  - Lower bound exploits high possible correlation among bits
- Nature is not adversarial
  - Often high dimensional systems have structure, e.g. Markov random fields fields (MRFs), graphical models (Bayes nets), etc



400 bit images

High-Dimensional Distn's

- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010101 (sample 1)
  - 0101001110 (sample 2)
  - 0011110100 (sample 3)
  - …
- Are bits/pixels independent?
  - Our algorithms require $\Theta\left(\frac{2^{n/2}}{\epsilon^2}\right)$ samples
- Is source generating graphs over n nodes Erdos-Renyi $G\left(n, \frac{1}{2}\right)$?
  - Our algorithms require $\Theta\left(\frac{2^{\binom{n}{2}/2}}{\epsilon^2}\right)$ samples
- Exponential dependence on n unsettling, but necessary
  - Lower bound exploits high possible correlation among bits
- Nature is not adversarial
  - Often high dimensional systems have structure, e.g. Markov random fields fields (MRFs), graphical models (Bayes nets), etc



400 bit images

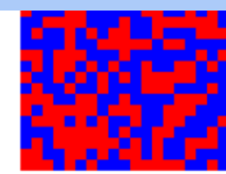
Testing high-dimensional distributions with combinatorial structure?

High-Dimensional Distn's

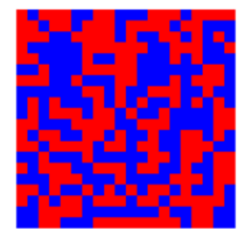
- Consider source generating n-bit strings $\in \{0,1\}^n$
  - 0011010
  - 0101001
  - 0011110
  - …

[w/ Dikkala, Kamath'16]: If unknown p is known to be an Ising model, then $\mathrm{poly}\left(n, \frac{1}{\epsilon}\right)$ samples suffice to test independence, goodness-of-fit. (extends to MRFs)

bit
ges



- Are bits/pixels independent?
  - Our algorithms require $\Theta\left(\frac{2^{n/2}}{\epsilon^2}\right)$ samples
- Is source generating graphs over n nodes Erdos-Renyi $G\left(n, \frac{1}{2}\right)$?
  - Our algorithms require $\Theta\left(\frac{2^{\binom{n}{2}/2}}{\epsilon^2}\right)$ samples



- Exponential dependence on n unsettling, but necessary
  - Lower bound exploits high possible correlation among bits
- Nature is not adversarial
  - Often high dimensional systems have structure, e.g. Markov random fields fields (MRFs), graphical models (Bayes nets), etc

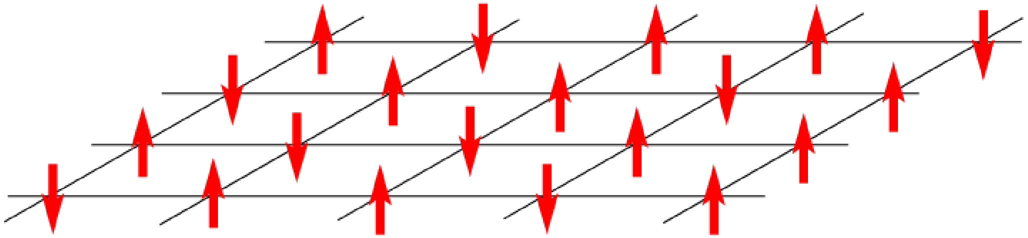
Testing high-dimensional distributions with combinatorial structure?

Ising Model

- Statistical physics, computer vision, neuroscience, social science

Ising Model

- Statistical physics, computer vision, neuroscience, social science
- Ising model:
  - Probability distribution defined in terms of a graph $G = (V, E)$, edge potentials θ_e, node potentials θ_v

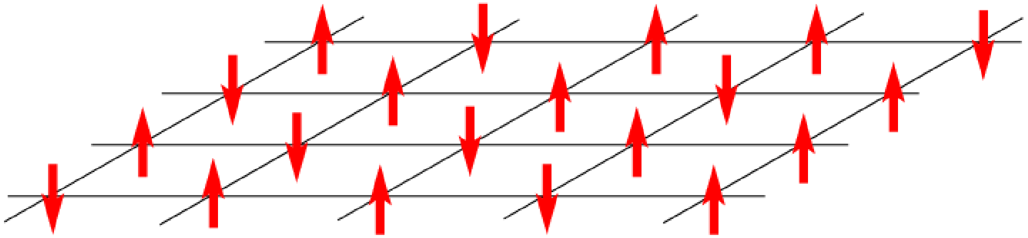


2–D Ising Model

Ising Model

- Statistical physics, computer vision, neuroscience, social science
- Ising model:
  - Probability distribution defined in terms of a graph $G = (V, E)$, edge potentials θ_e, node potentials θ_v
  - State space $\{\pm 1\}^V$

$$p_\theta(x) \propto \exp\left(\sum_{e=(u,v)\in E} \theta_e x_u x_v + \sum_{v \in V} \theta_v x_v\right)$$

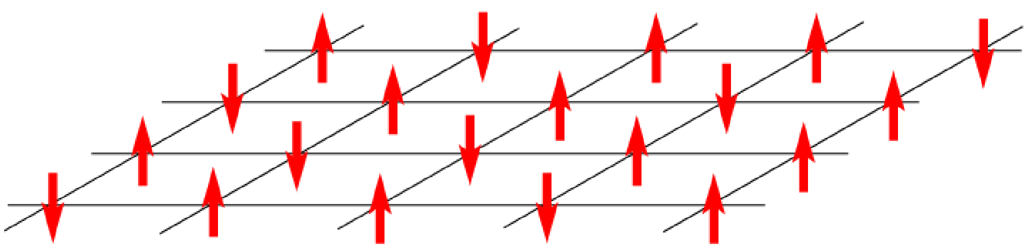


2–D Ising Model

Ising Model

- Statistical physics, computer vision, neuroscience, social science
- Ising model:
  - Probability distribution defined in terms of a graph $G = (V, E)$, edge potentials θ_e, node potentials θ_v
  - State space $\{\pm 1\}^V$

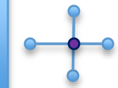
$$p_\theta(x) \propto \exp\left(\sum_{e=(u,v)\in E} \theta_e x_u x_v + \sum_{v\in V} \theta_v x_v \right)$$



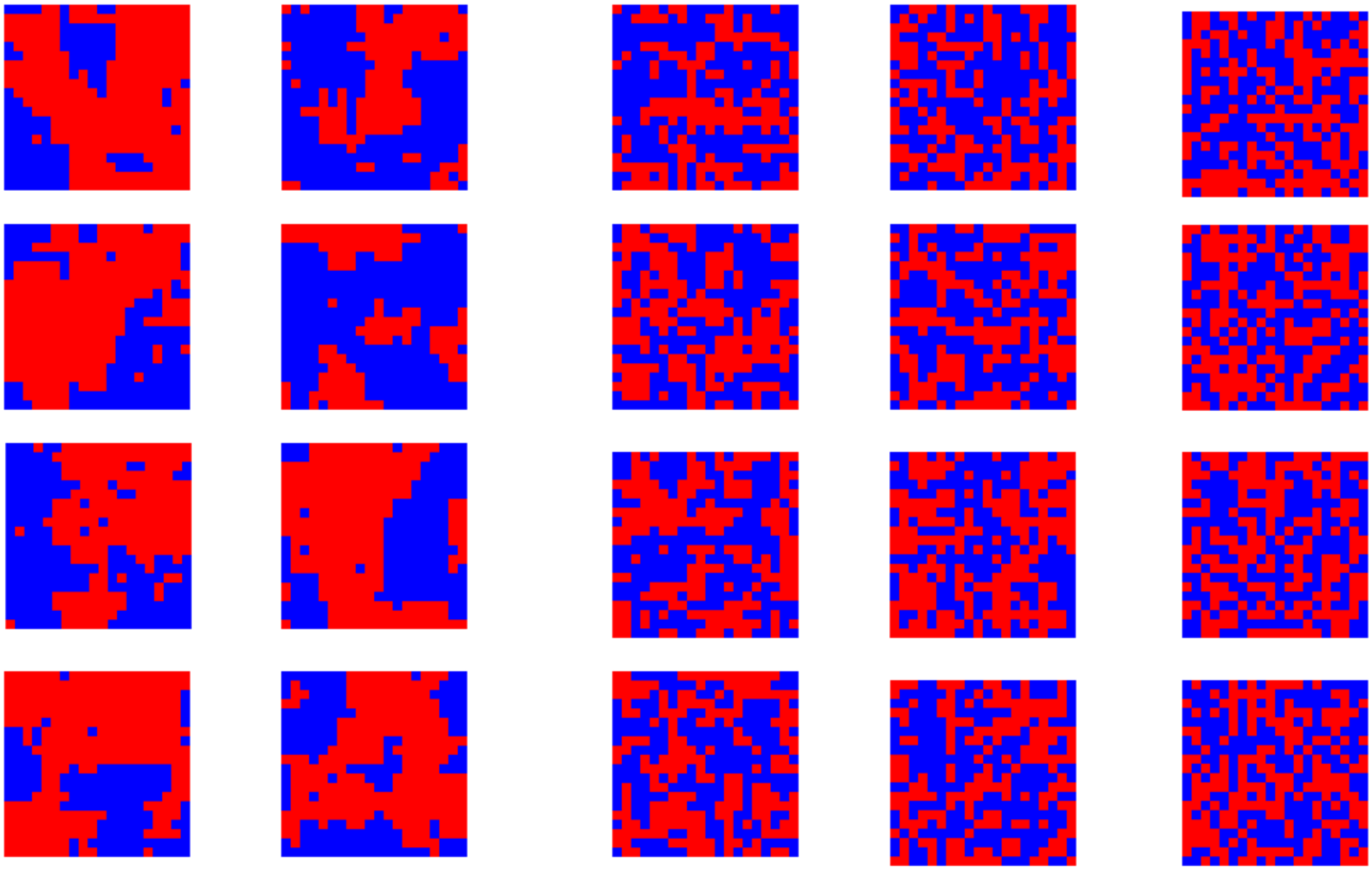
2–D Ising Model

  - High $|\theta_e|$'s $\Longrightarrow$ strongly (anti-)correlated spins

$\theta_v = 0$



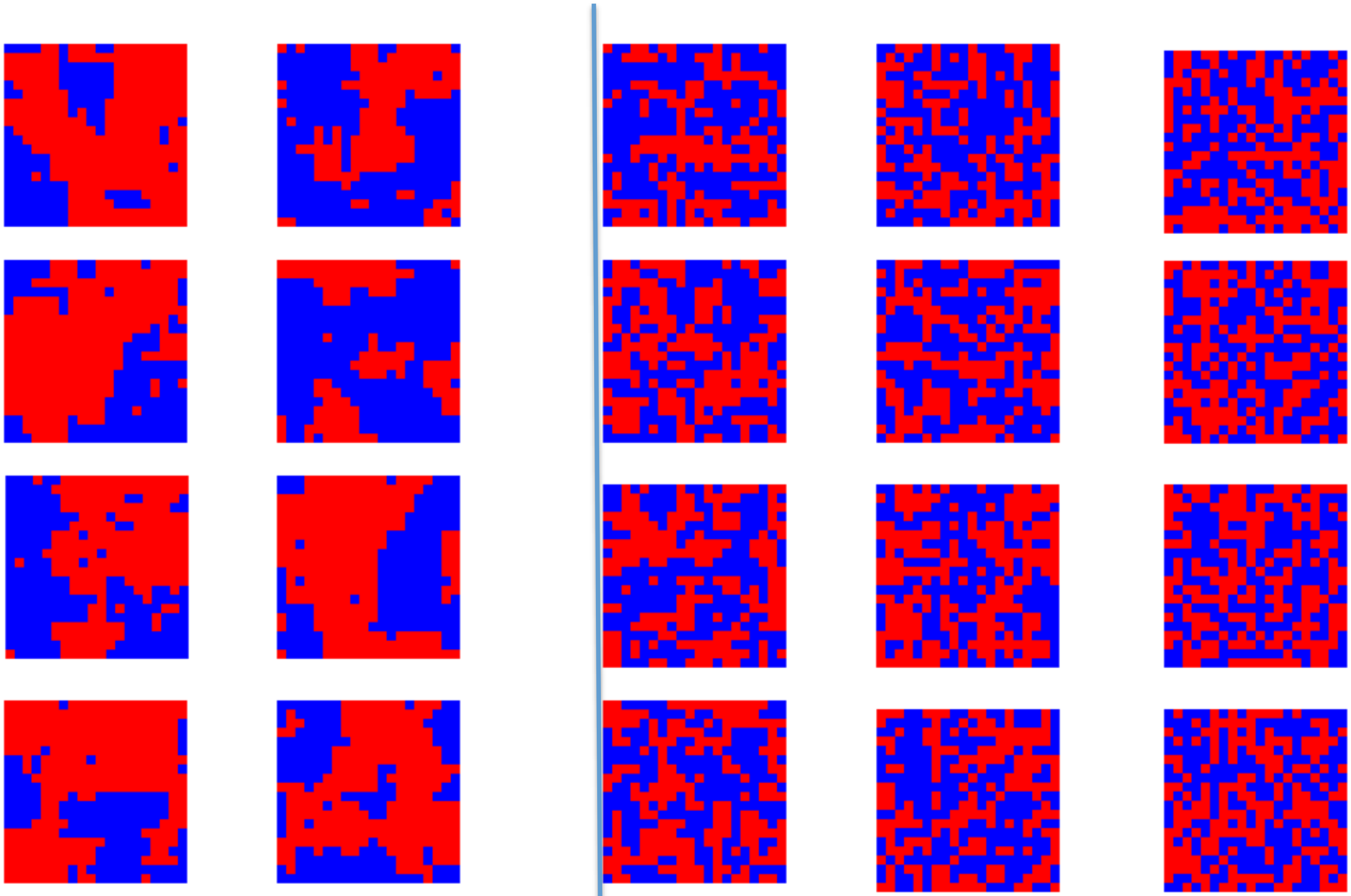
Ising Model: Strong vs weak ties



$\theta_e = 1$ $\theta_e = 0.5$ $\theta_e = 0.25$ $\theta_e = 0.125$ $\theta_e = 0$

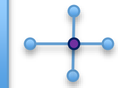
$\theta_v = 0$

Ising Model: Strong vs weak ties



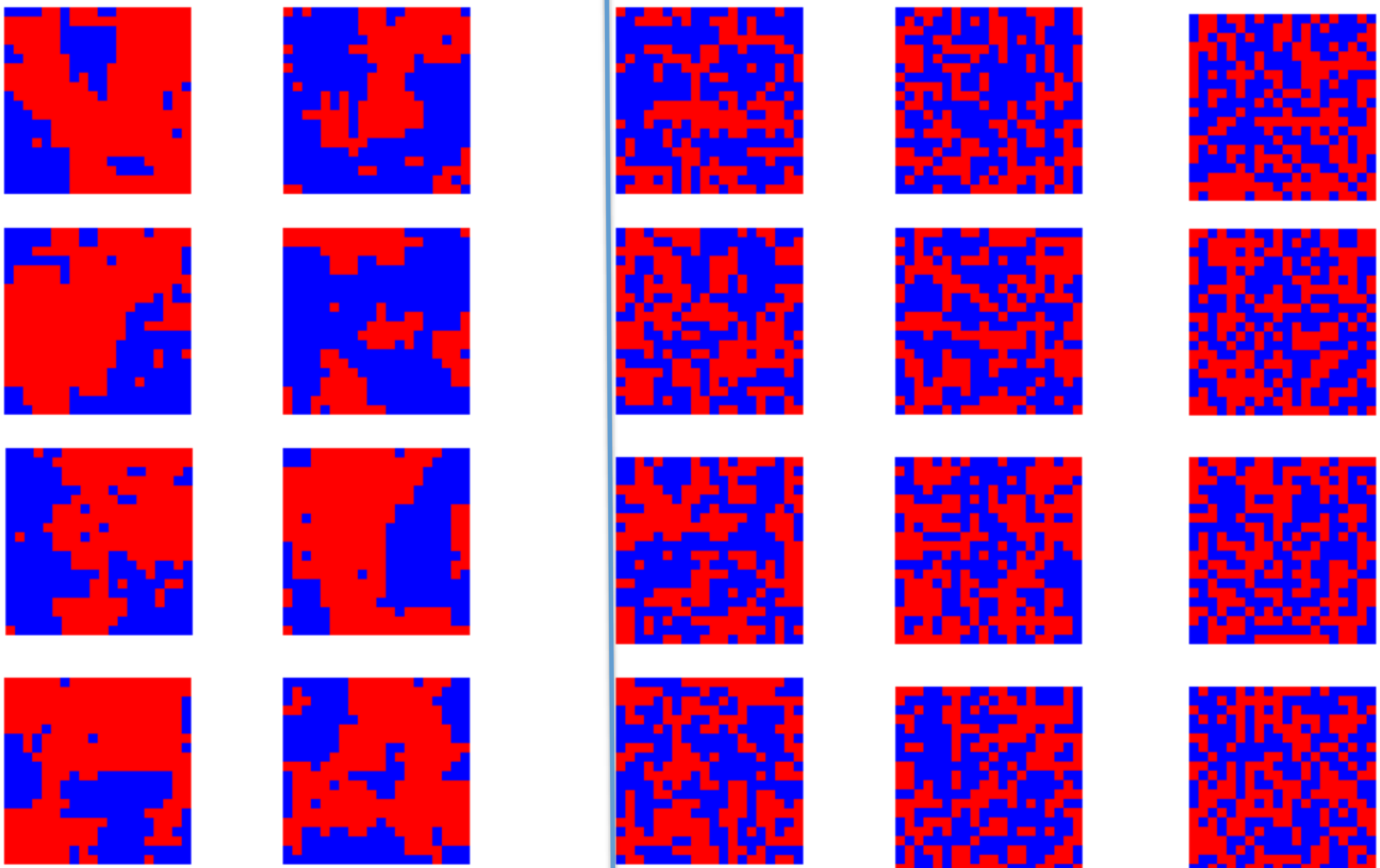
$\theta_e = 1$ $\theta_e = 0.5$ $\theta_e = 0.25$ $\theta_e = 0.125$ $\theta_e = 0$

$\theta_v = 0$



Ising Model: Strong vs weak ties

"low temperature regime" | "high temperature regime"

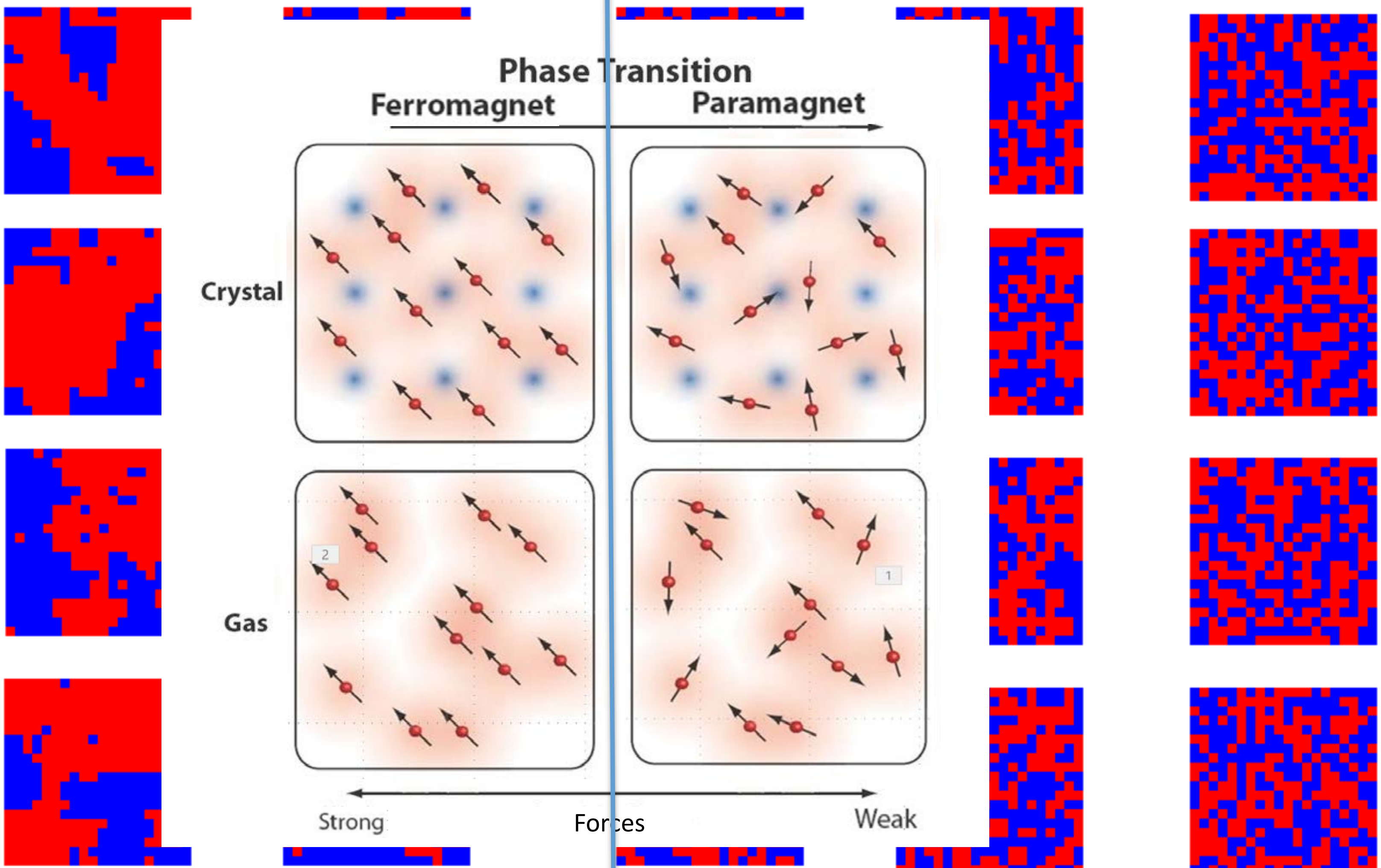


$\theta_e = 1$ | $\theta_e = 0.5$ | $\theta_e = 0.25$ | $\theta_e = 0.125$ | $\theta_e = 0$

$\theta_v = 0$

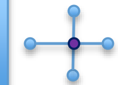
Ising Model: Strong vs weak ties

“low temperature regime” “high temperature regime”



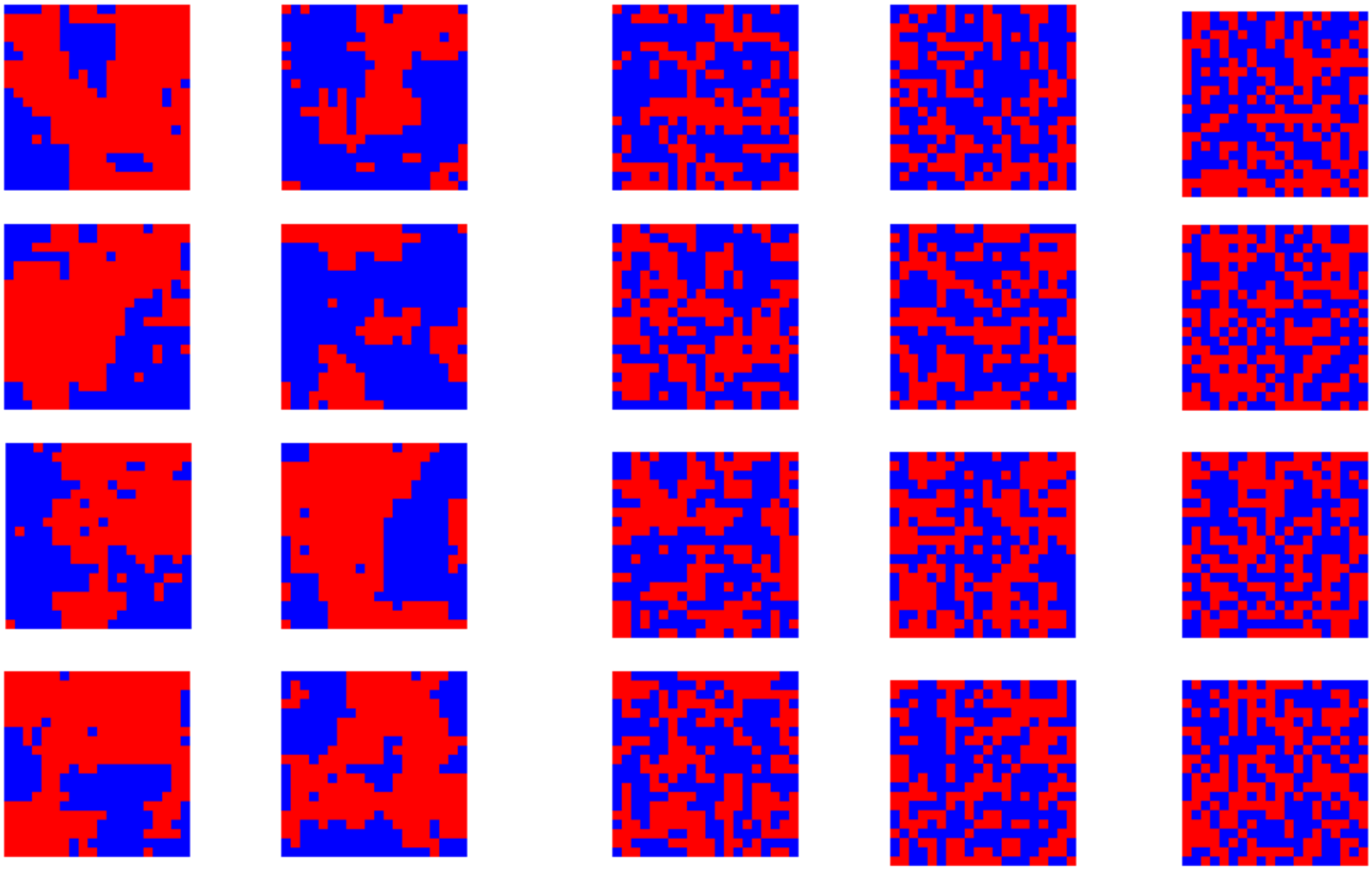
$\theta_e = 1$ $\theta_e = 0.5$ $\theta_e = 0.25$ $\theta_e = 0.125$ $\theta_e = 0$

$\theta_v = 0$



Ising Model: Strong vs weak ties

"low temperature regime" "high temperature regime"



$\theta_e = 1$ $\theta_e = 0.5$ $\theta_e = 0.25$ $\theta_e = 0.125$ $\theta_e = 0$

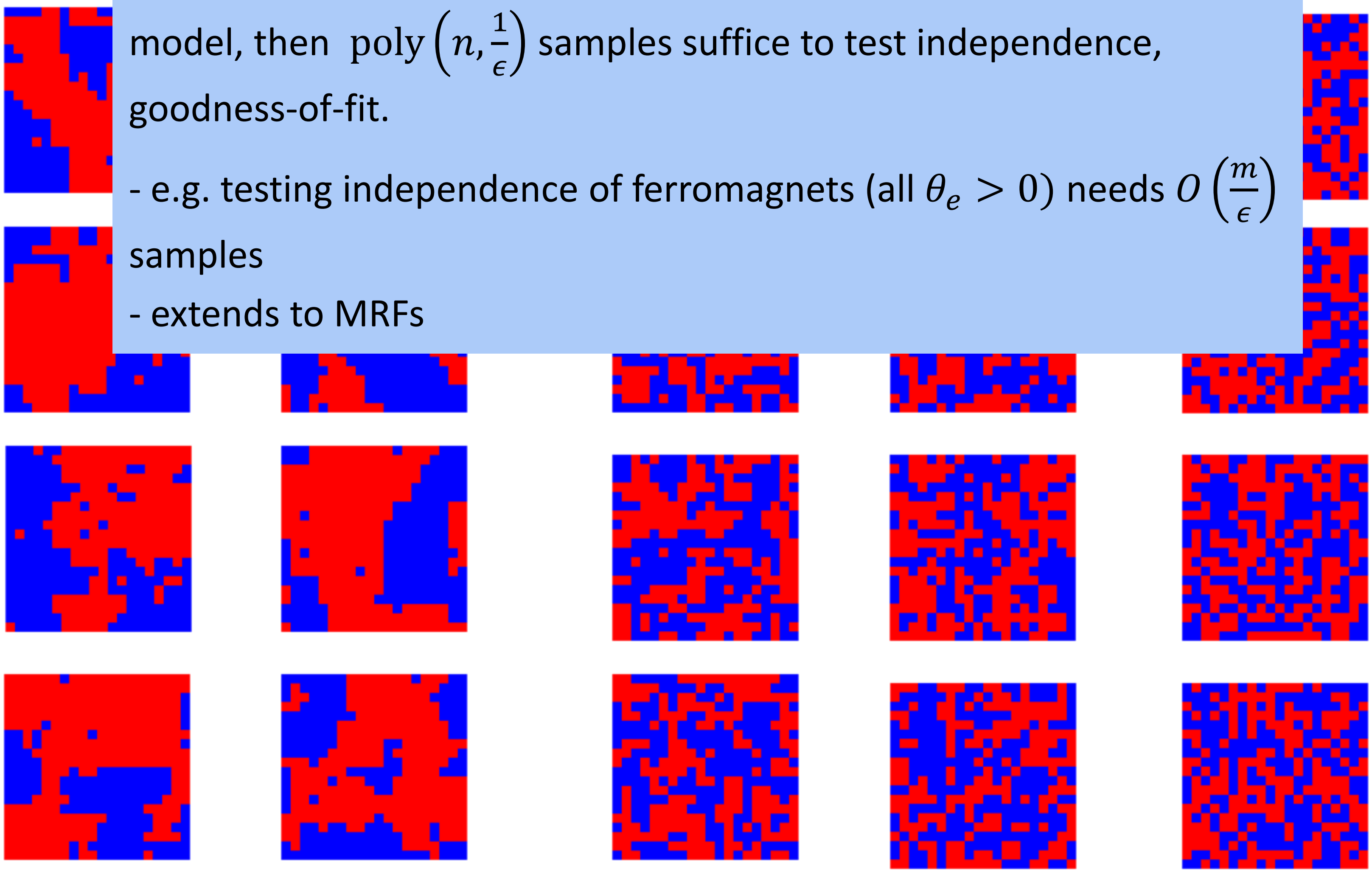
$\theta_v = 0$

Ising Model: Strong vs weak ties

"low

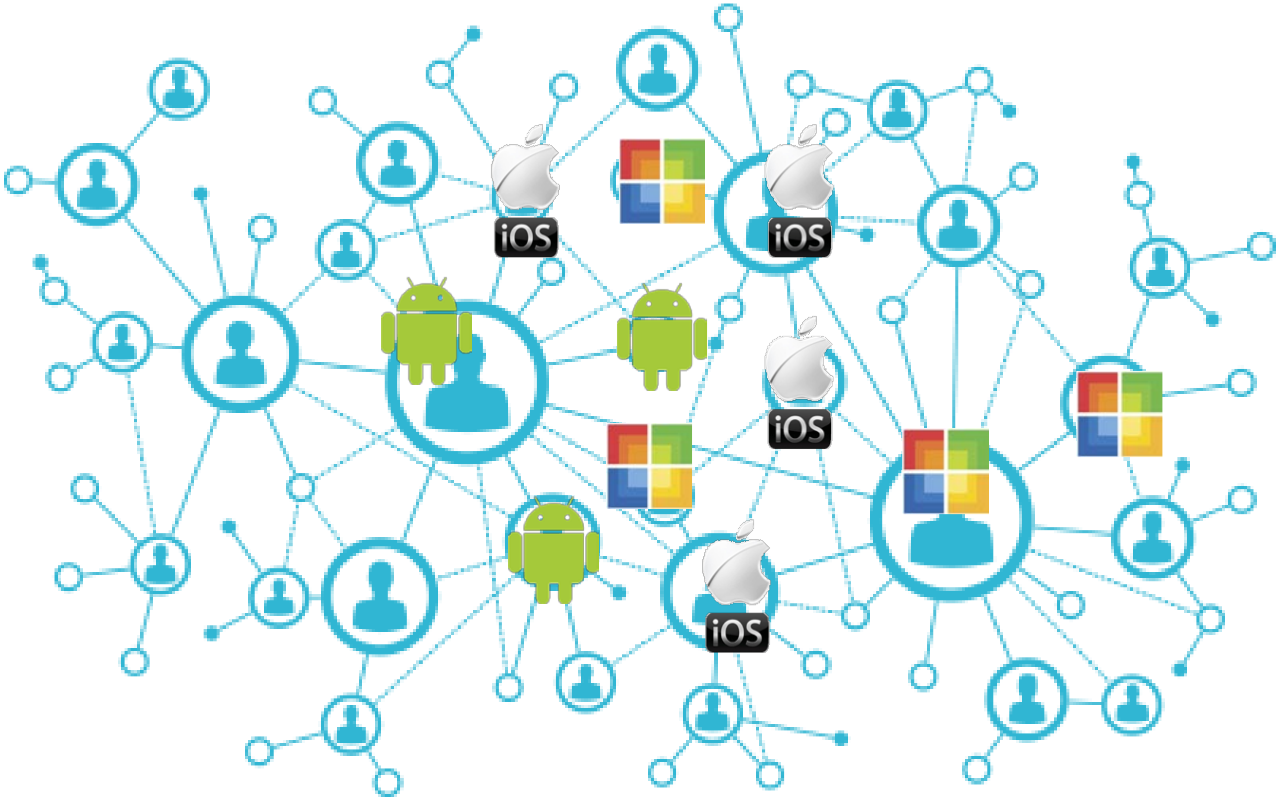
[w/ Dikkala, Kamath'16]: If unknown p is known to be an Ising model, then $\text{poly}\left(n, \frac{1}{\epsilon}\right)$ samples suffice to test independence, goodness-of-fit.

- e.g. testing independence of ferromagnets (all $\theta_e > 0$) needs $O\left(\frac{m}{\epsilon}\right)$ samples
- extends to MRFs

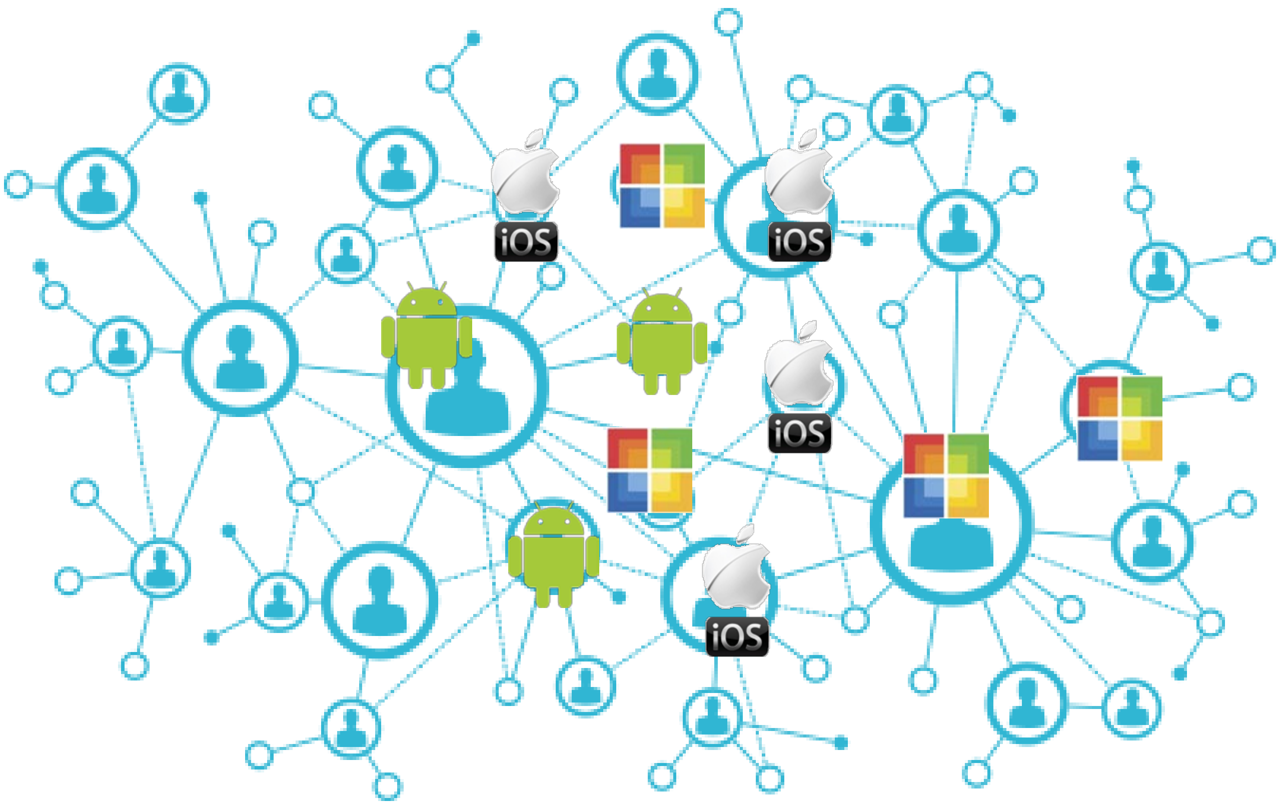


$\theta_e = 1$ $\quad$ $\theta_e = 0.5$ $\quad$ $\theta_e = 0.25$ $\quad$ $\theta_e = 0.125$ $\quad$ $\theta_e = 0$

e.g.4: Behavior in a Social Network



e.g.4: Behavior in a Social Network



Question: Are adopted technologies a product distribution or are they far from being from a product distribution?

The Menu

- **Motivation**
- **Problem Formulation**
- **Uniformity Testing, Goodness of Fit**
- **Testing Properties of Distributions**
- **Testing in High Dimensions**
- Conclusion

The Menu

- Motivation
- Problem Formulation
- Uniformity Testing, Goodness of Fit
- Testing Properties of Distributions
- Testing in High Dimensions
- Conclusion

Conclusions

- [w/ Acharya, Kamath'15]: Improved χ^2-test, requiring $O\left(\frac{\sqrt{D}}{\epsilon^2}\right)$ samples
  - implies testers of various distributional properties (independence, unimodality, logconcavity, etc) from same number of samples

Conclusions

- [w/ Acharya, Kamath'15]: Improved χ^2-test, requiring $O\left(\frac{\sqrt{D}}{\epsilon^2}\right)$ samples
  - implies testers of various distributional properties (independence, unimodality, logconcavity, etc) from same number of samples
- Testing properties of high-dimensional distributions requires exponentially many samples

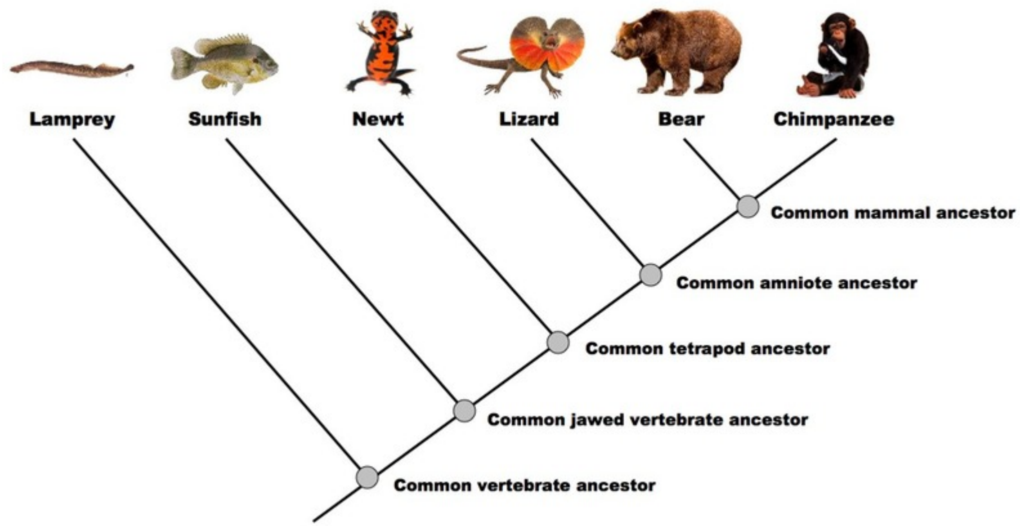
Conclusions

- [w/ Acharya, Kamath'15]: Improved χ^2-test, requiring $O\left(\frac{\sqrt{D}}{\epsilon^2}\right)$ samples
  - implies testers of various distributional properties (independence, unimodality, logconcavity, etc) from same number of samples
- Testing properties of high-dimensional distributions requires exponentially many samples
- Making assumptions about the distribution being sampled gives leverage

Conclusions

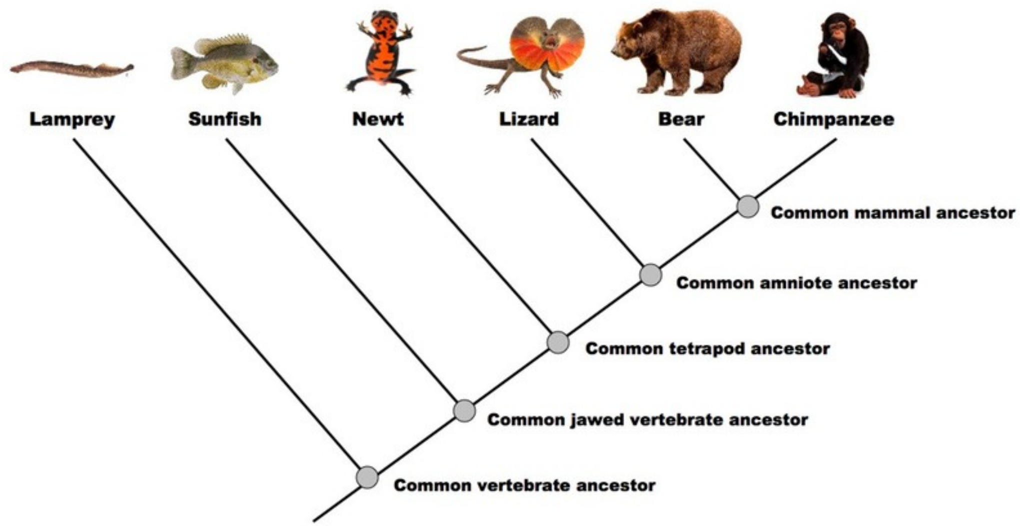
- [w/ Acharya, Kamath'15]: Improved χ^2-test, requiring $O\left(\frac{\sqrt{D}}{\epsilon^2}\right)$ samples
  - implies testers of various distributional properties (independence, unimodality, logconcavity, etc) from same number of samples
- Testing properties of high-dimensional distributions requires exponentially many samples
- Making assumptions about the distribution being sampled gives leverage
- [w/ Dikkala, Kamath'16]: Testing independence and goodness-of-fit in Ising models can be done with polynomially many samples

Testing Combinatorial Structure



Testing Combinatorial Structure

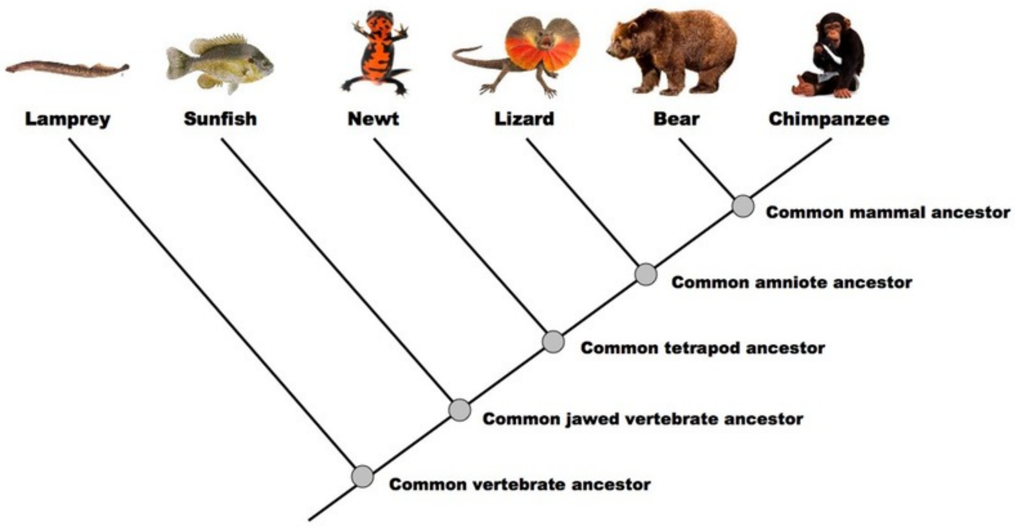
Is the phylogenic tree assumption true?



Testing Combinatorial Structure

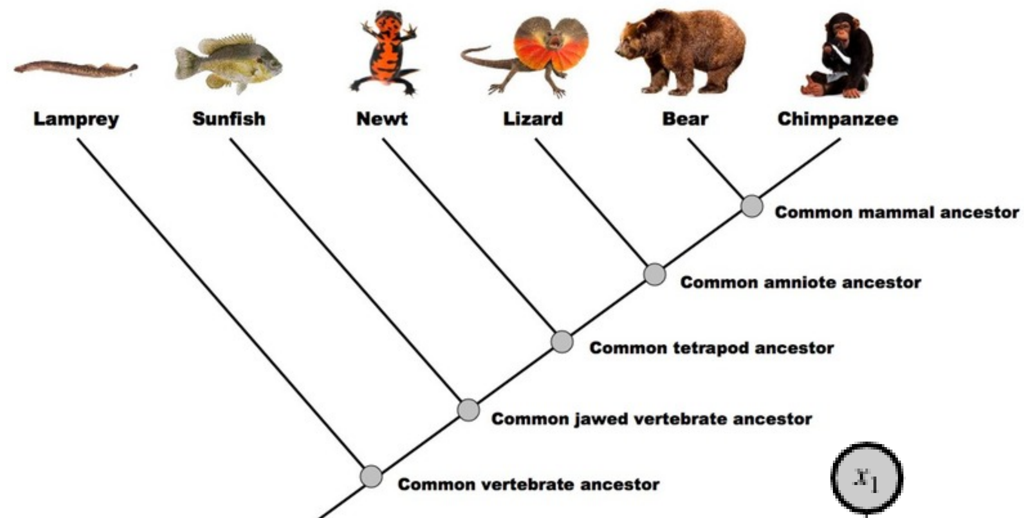
Is the phylogenic tree assumption true?

Sapiens-Neanderthal early interbreeding [Slatkin et al’13]



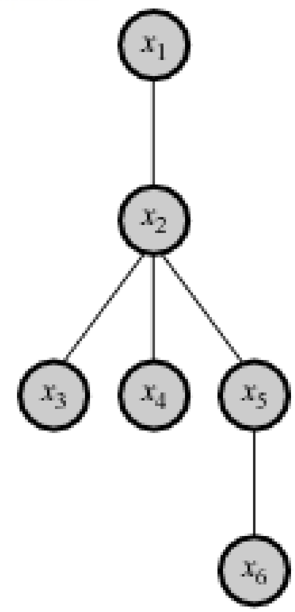
Testing Combinatorial Structure

Is the phylogenic tree assumption true?

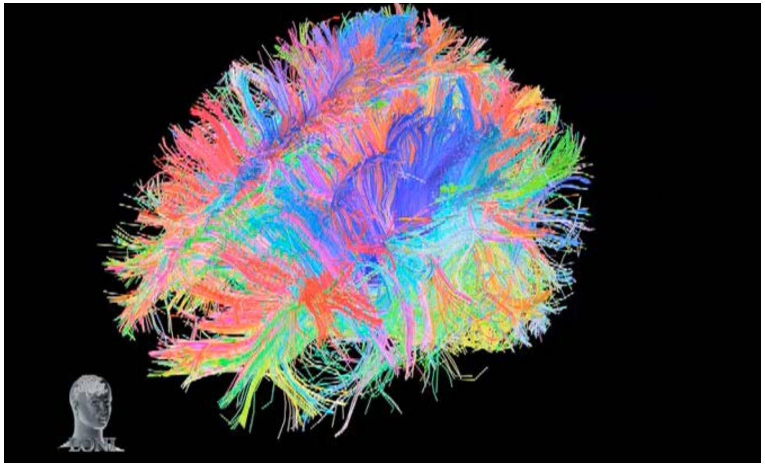


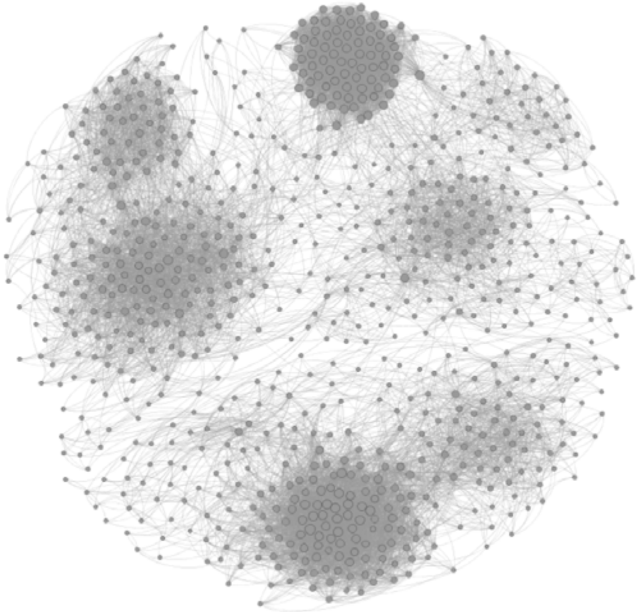
Sapiens-Neanderthal early interbreeding [Slatkin et al'13]

Is a graphical model a tree?
[ongoing work with Acharya,Bresler]



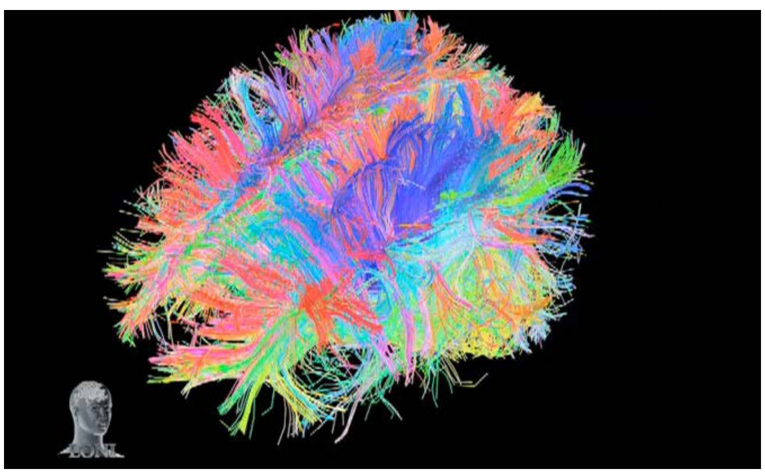
Testing from a Single Sample

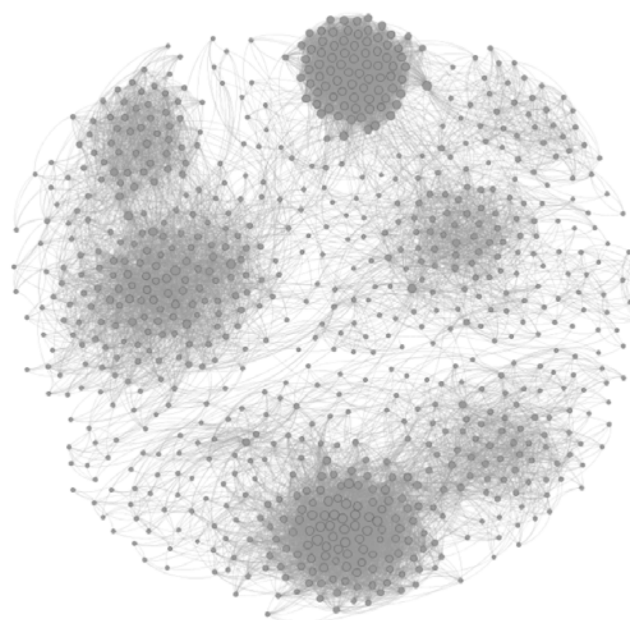




Testing from a Single Sample

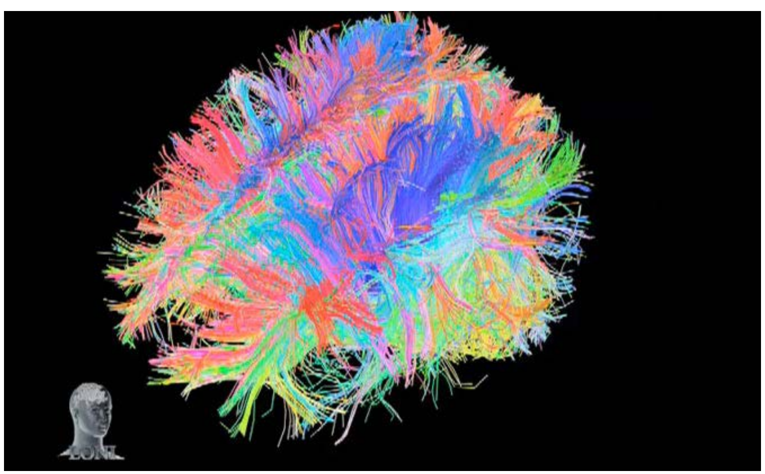
- Given **one** social network, **one** brain, etc., how can we test the validity of a certain generative model?

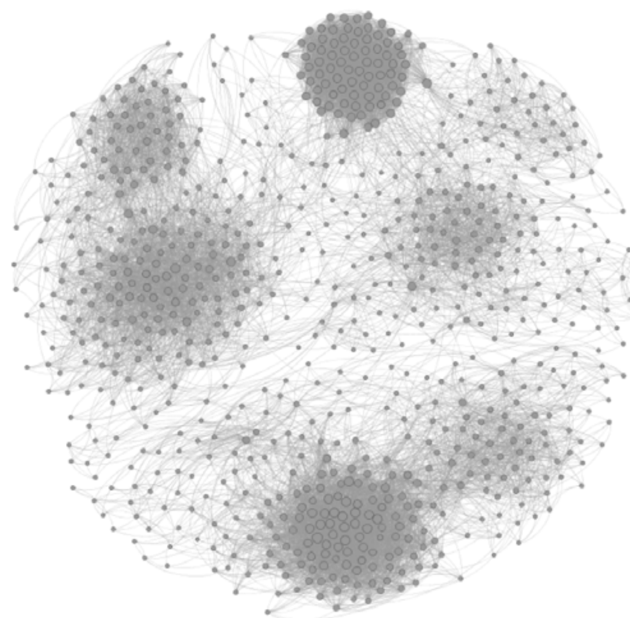




Testing from a Single Sample

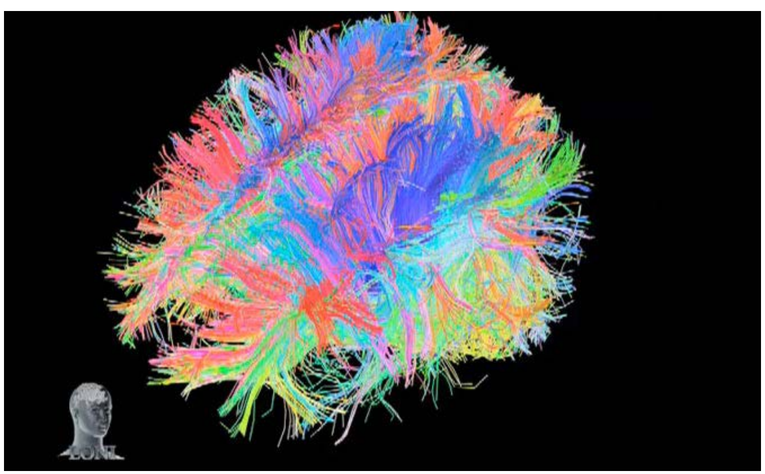
- Given **one** social network, **one** brain, etc., how can we test the validity of a certain generative model?
- Get many samples from one sample?

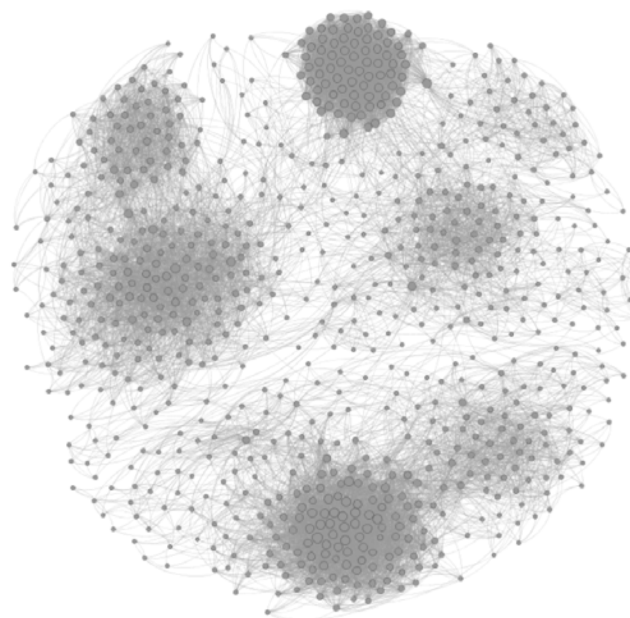




Testing from a Single Sample

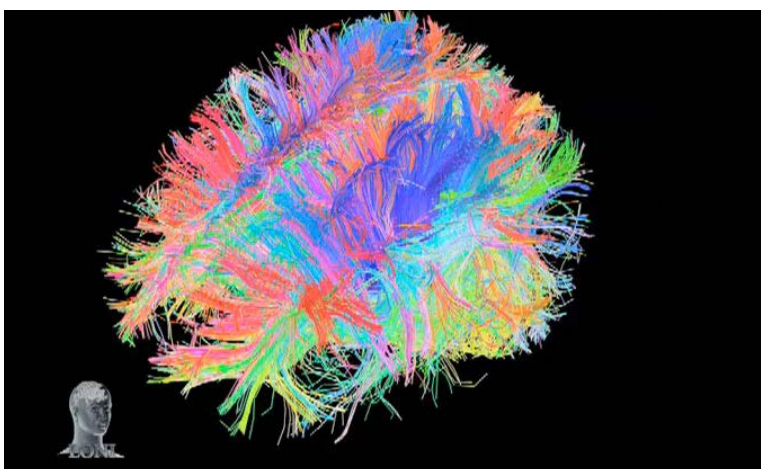
- Given **one** social network, **one** brain, etc., how can we test the validity of a certain generative model?
- Get many samples from one sample?
- Ongoing with Rubinfeld

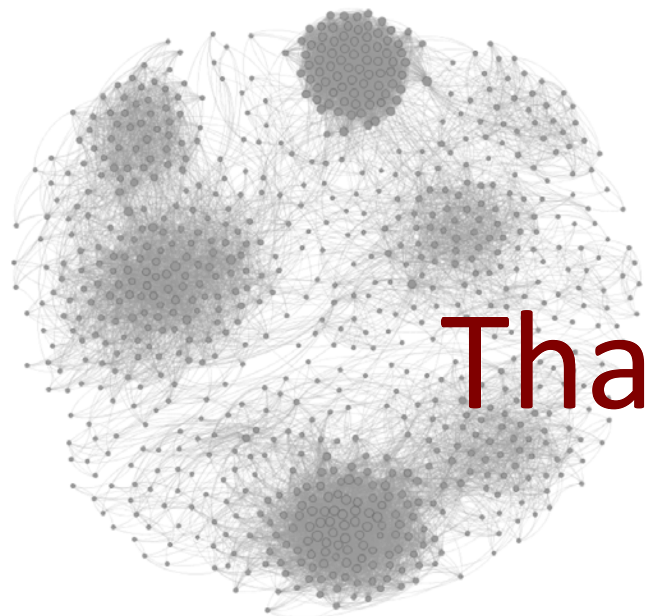




Testing from a Single Sample

- Given **one** social network, **one** brain, etc., how can we test the validity of a certain generative model?
- Get many samples from one sample?
- Ongoing with Rubinfeld





Thanks!