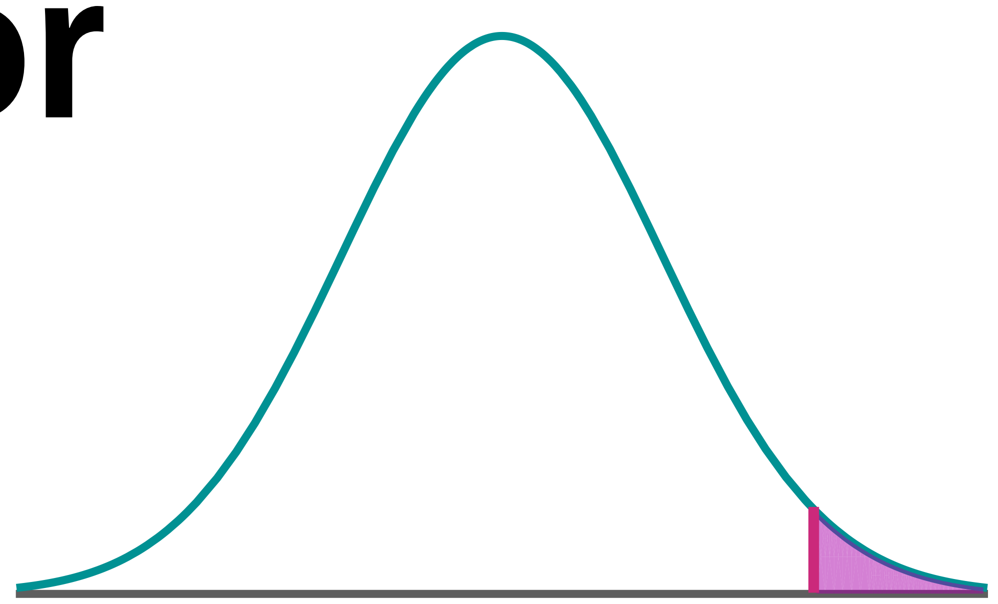
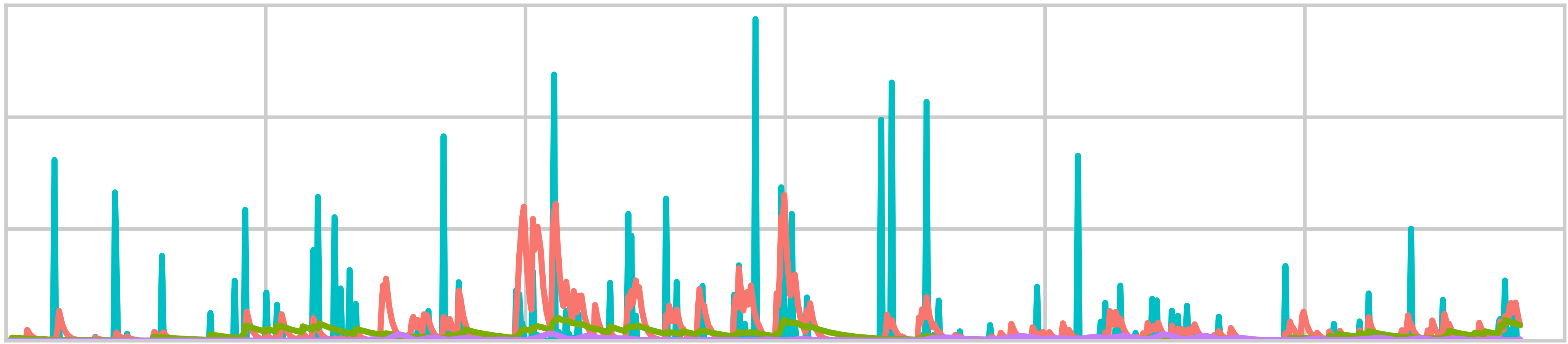
Statistical Learning and Model Criticism for Networks and Point Processes

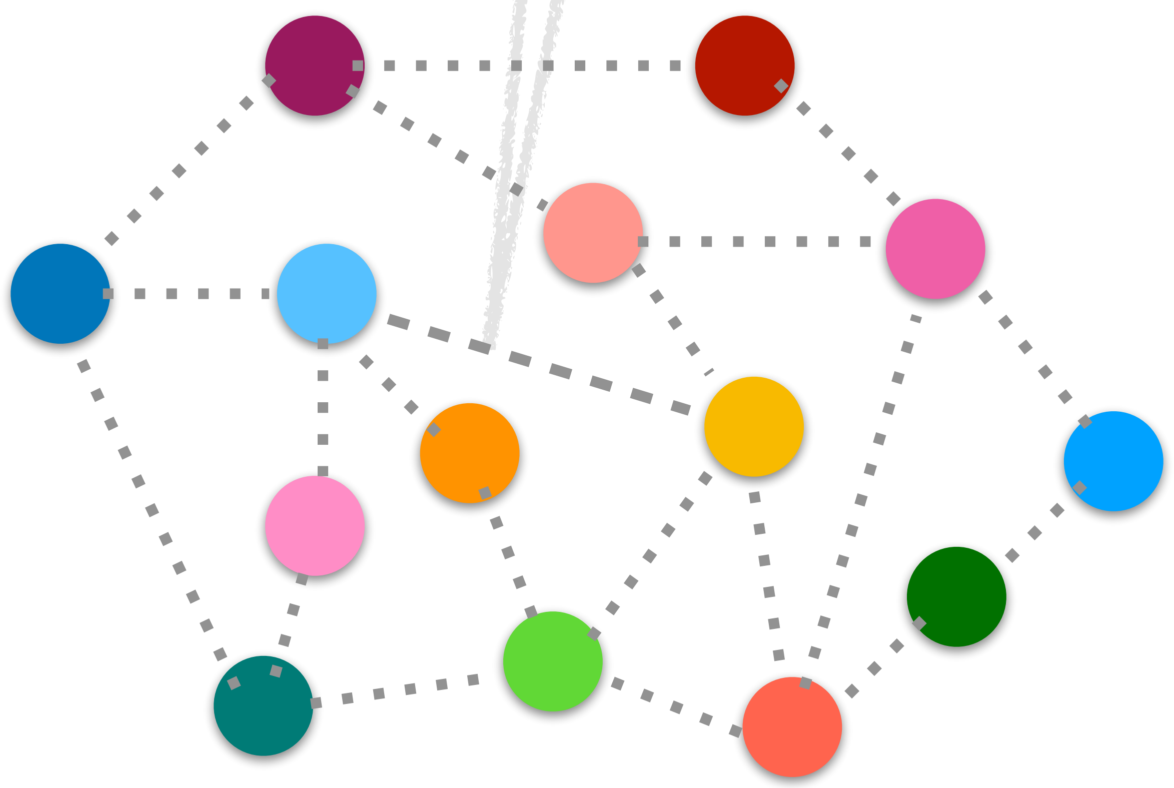


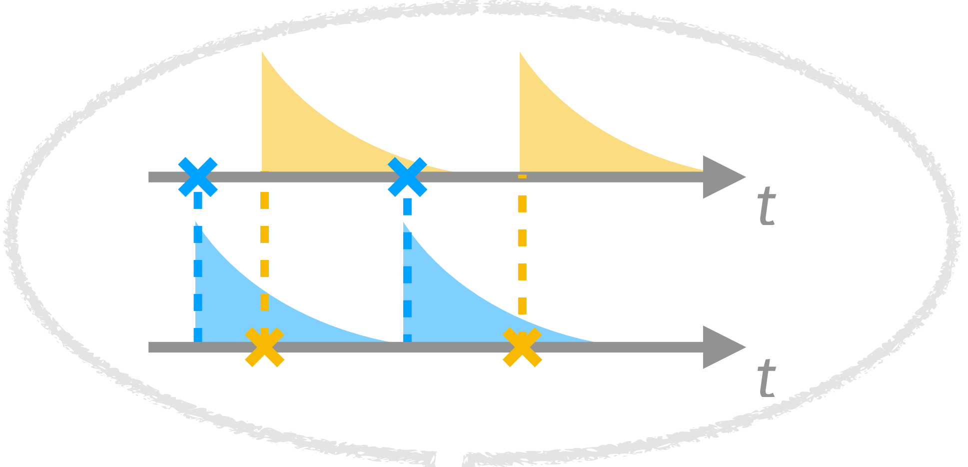
Jiasen Yang

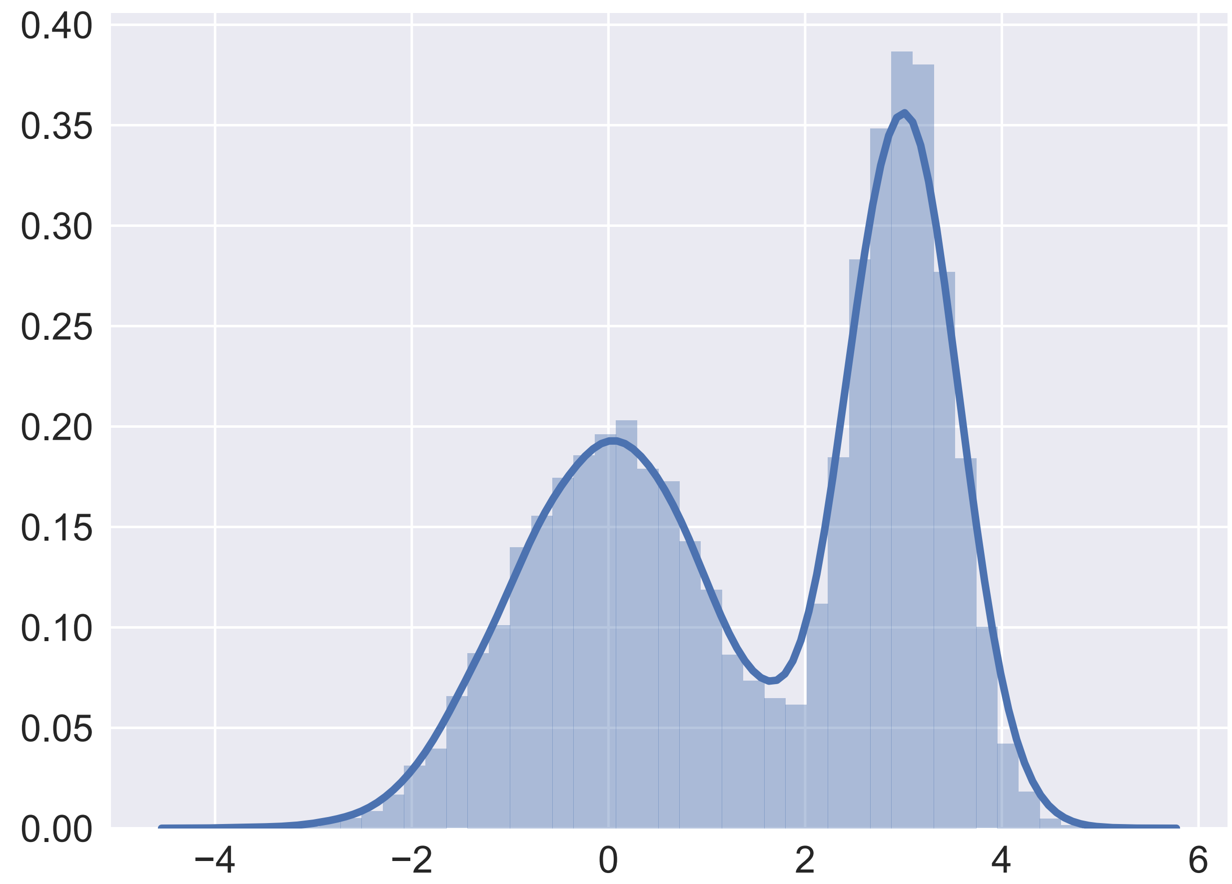
Purdue University

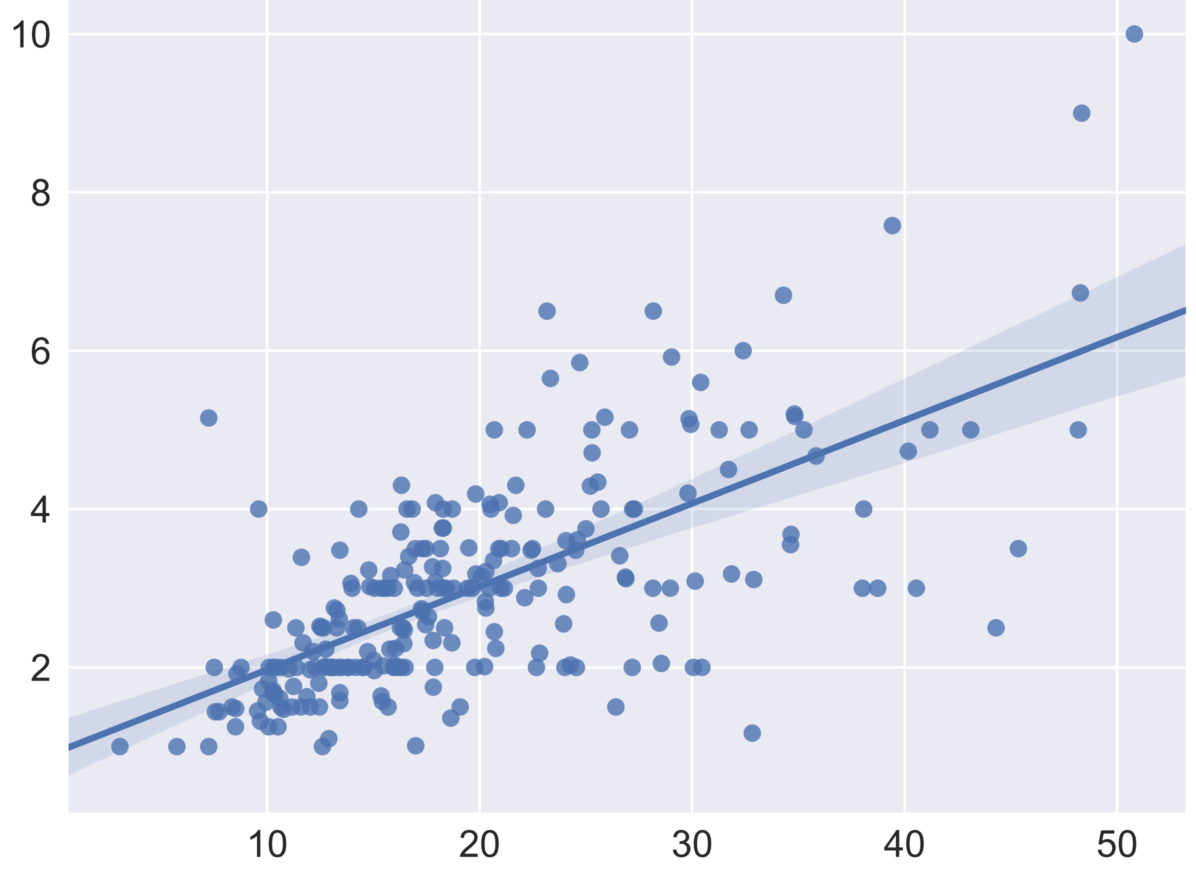
April 24, 2019

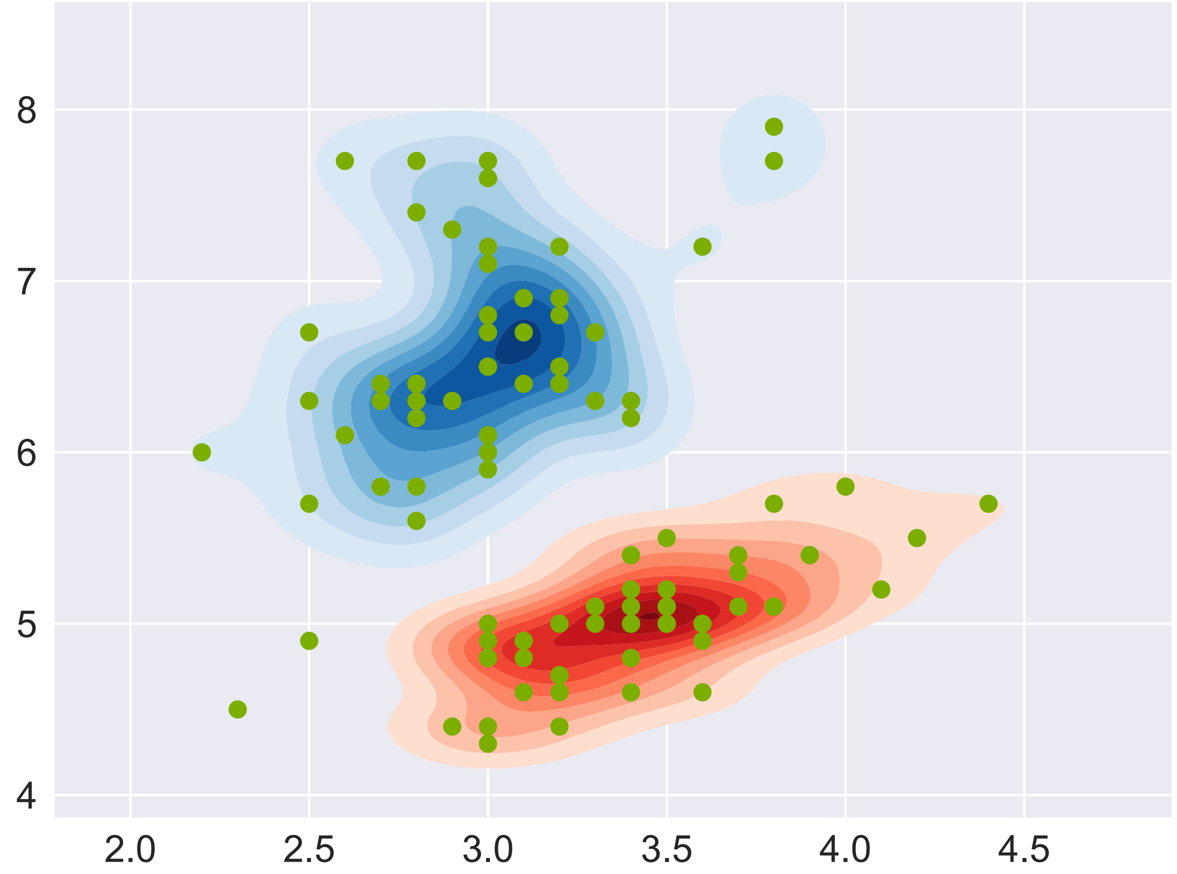










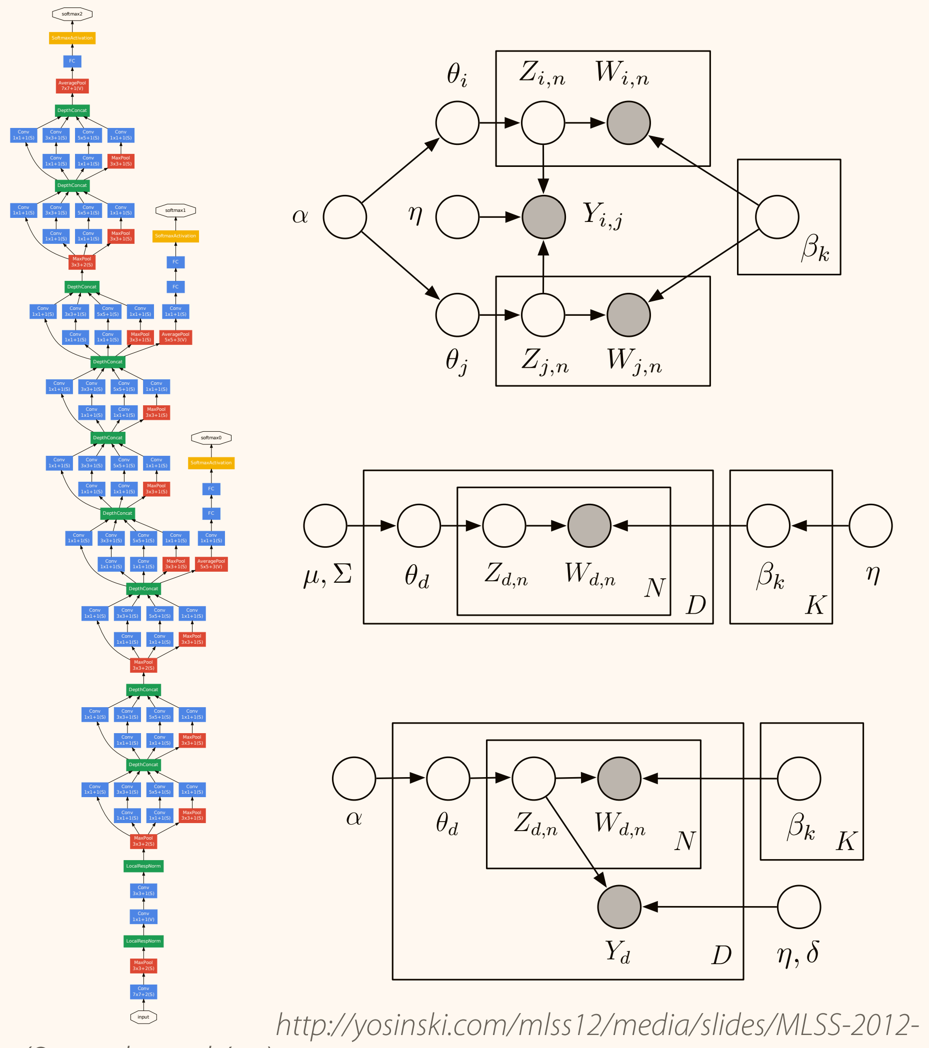


PURDUE UNIVERSITY

The Data Analysis Pipeline

https://cs.stanford.edu/people/karpathy/cnnembed/
https://commons.wikimedia.org/wiki/File:Protein_GC_PDB_1j78.png
https://www.b2bmarketing.net/en-gb/resources/blog/network-effect-what-b2b-comms-can-learn-facebook-revolution

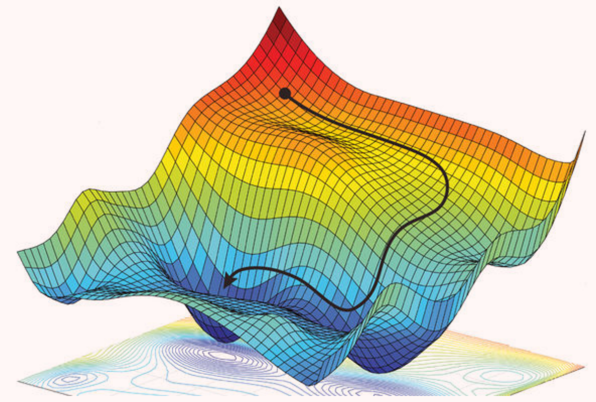
Data



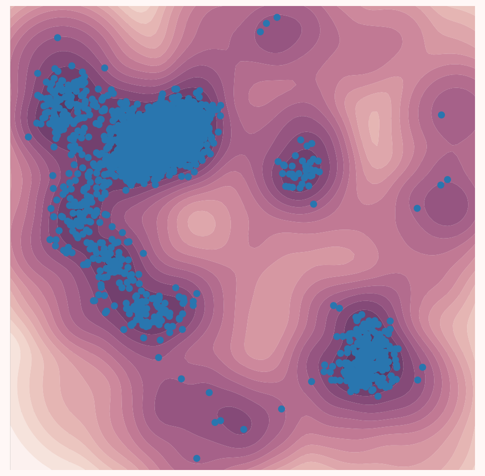
(Szegedy et al. '15)

http://yosinski.com/mlss12/media/slides/MLSS-2012-Blei-Probabilistic-Topic-Models.pdf

Design Model



http://www.sciencemag.org/news/2018/05/ai-researchers-allege-machine-learning-alchemy



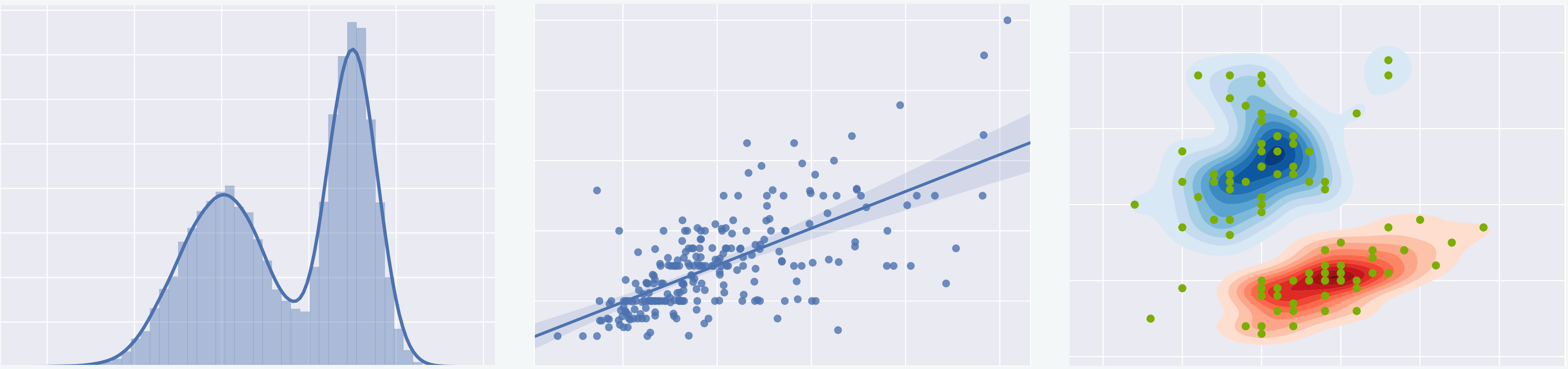
Infer Parameters

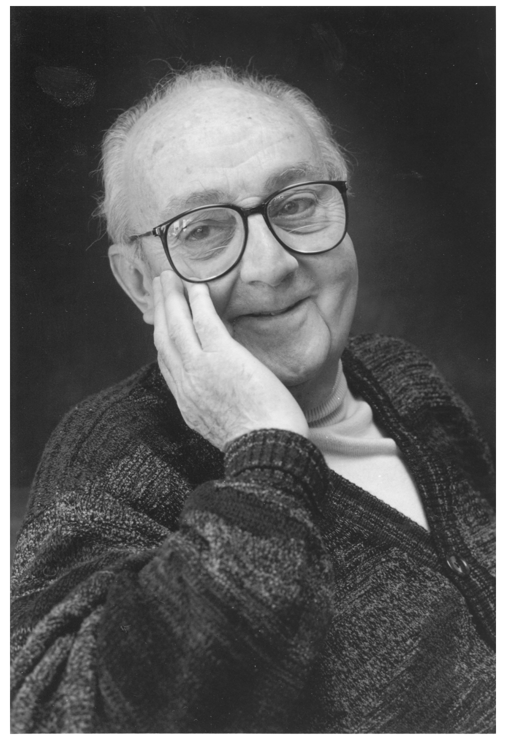
(Nickel & Kiela '17)

Apply Model

Criticize Model

- Predictive performance
- Statistical hypothesis tests
- Posterior predictive checks





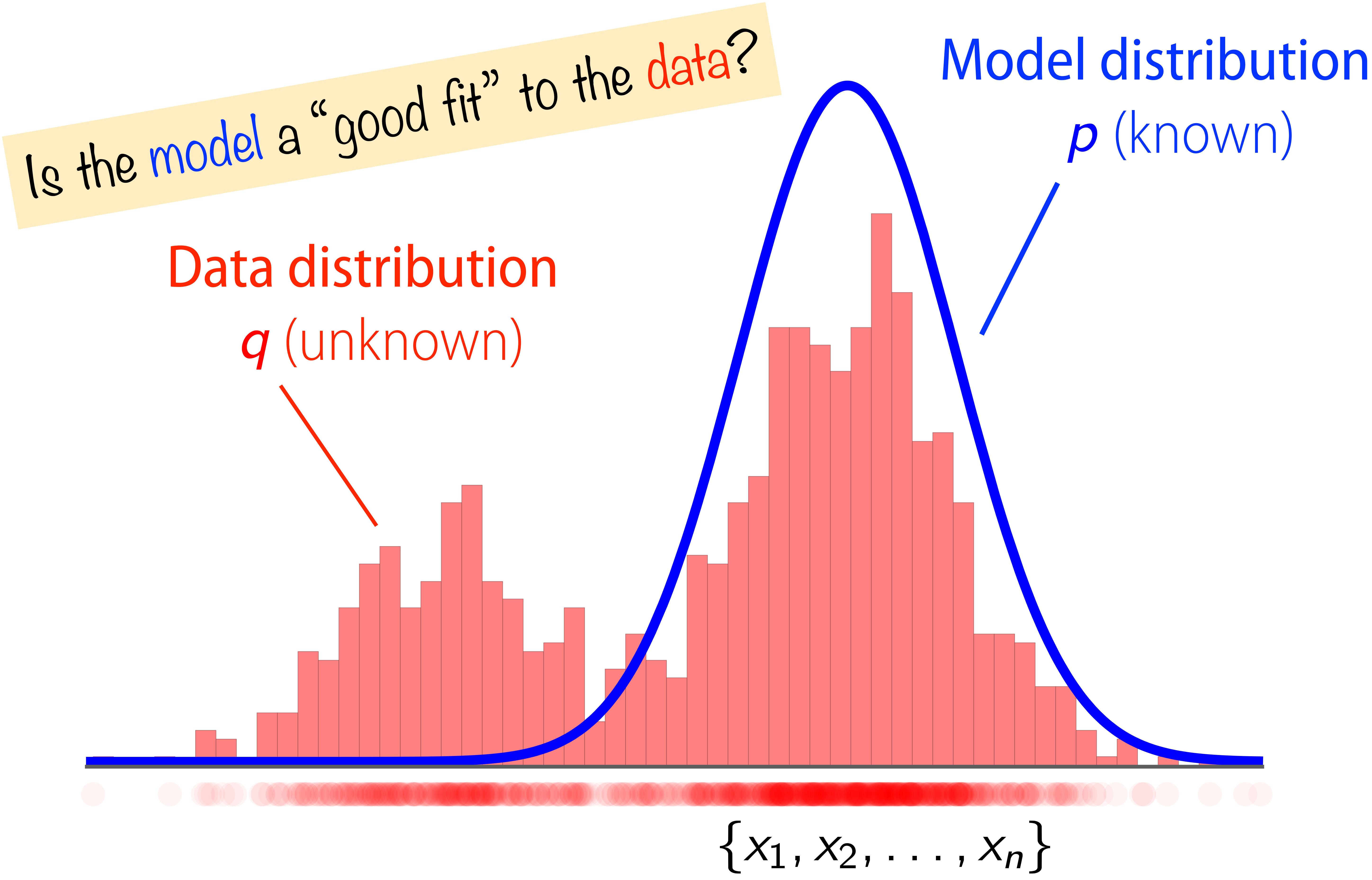
George E. P. Box (1976):
"All models are wrong, but some are useful."

"Box's Loop" *(Blei' 14)*

Goodness-of-Fit Testing

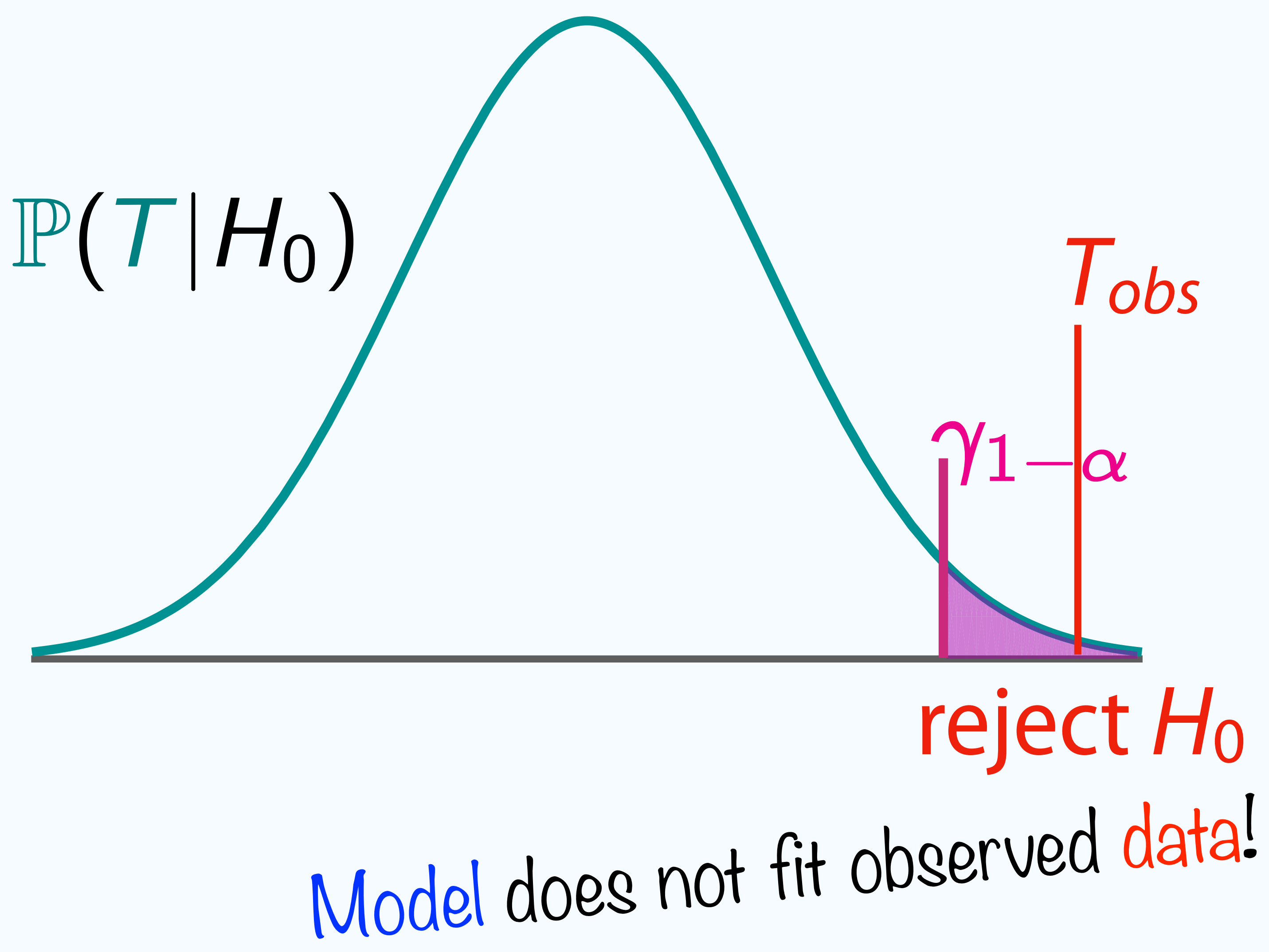
Given a probability distribution p on $\mathcal{X}^d$ and *data samples* $\{\mathbf{x}_i\}_{i=1}^n \sim q$, test

$$H_0 : p = q \qquad \text{vs.} \qquad H_1 : p \neq q$$



Goodness-of-Fit Test

- Construct test statistic T
- Compute critical value $\gamma_{1-\alpha}$



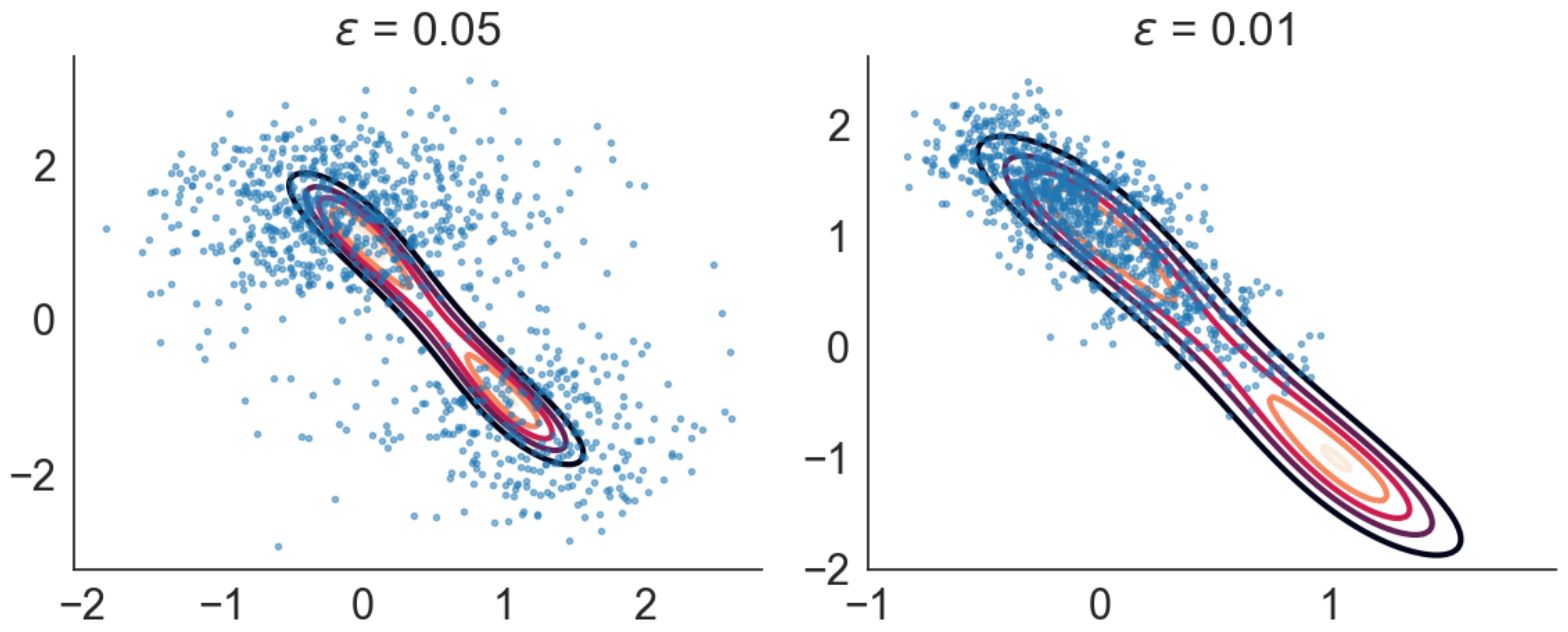
Goodness-of-Fit Testing (Cont'd)

Given a probability distribution p on $\mathcal{X}^d$ and *data samples* $\{\mathbf{x}_i\}_{i=1}^n \sim q$, test

$$H_0 : p = q \qquad \text{vs.} \qquad H_1 : p \neq q$$

Applications

- **Model criticism & evaluation**: checking model assumptions, etc.
- **Measuring sample quality**: Markov chain diagnostics, etc.
- **Selecting hyper-parameters** (for model or inference algorithm).



Effect of step-size in SGLD (Huggins & Mackey '18)

Classical approaches:

- Chi-squared test *(Pearson, 1900)*
- Kolmogorov–Smirnov test *(Kolmogorov, 193*
- Cramér–von Mises test *(Cramér, 1928,*
- Anderson–Darling test *(Anderson & Darling, 1954)*

Require p to be **tractable!**

K. Pearson

A. Kolmogorov

R. A. Fisher

Goodness-of-Fit Testing (Cont'd)

Given a probability distribution p on $\mathcal{X}^d$ and *data samples* $\{\mathbf{x}_i\}_{i=1}^n \sim q$, test

$$H_0 : p = q \qquad \text{vs.} \qquad H_1 : p \neq q$$

Modern applications:

Model dist. *un-normalized*

$$p(\mathbf{x}) = \frac{1}{Z}\tilde{p}(\mathbf{x}) \propto \tilde{p}(\mathbf{x})$$

Normalization constant

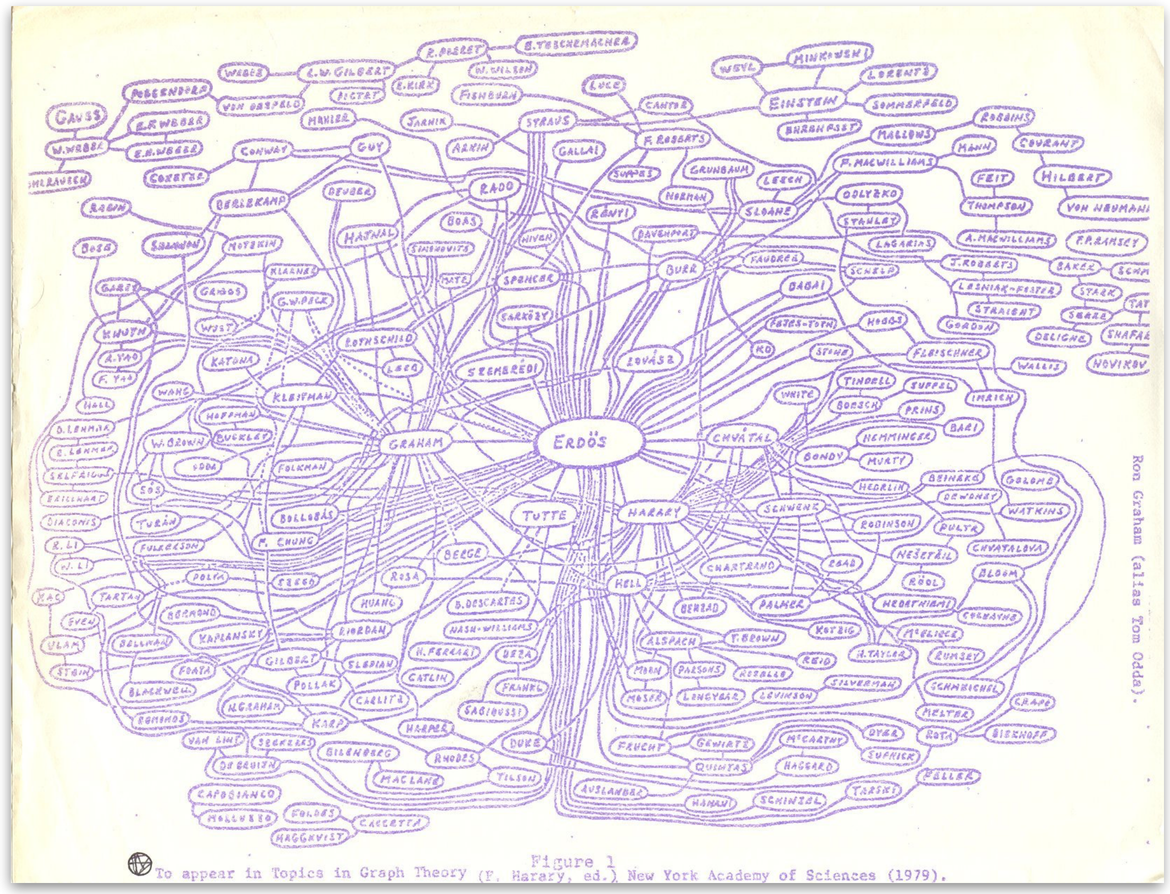
$$Z = \sum \tilde{p}(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$Z = \int_{\mathbf{x}\in\mathcal{X}^d} \tilde{p}(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

Intractable!

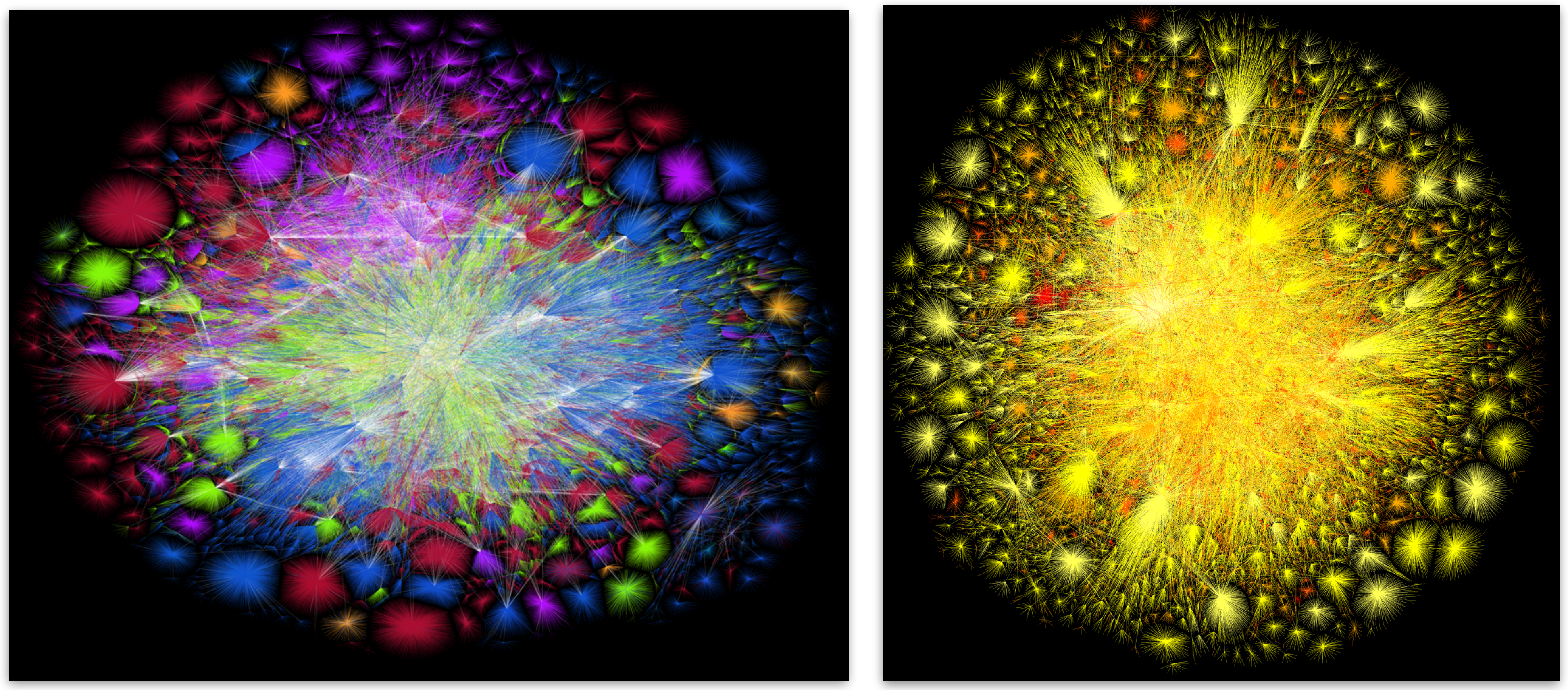
	Continuous distributions	Discrete distributions	Point processes
Normalized	Kolmogorov–Smirnov test Cramér–von Mises test Anderson–Darling test	Chi-squared test	(mainly Poisson-type)
Unnormalized	Kernelized Stein discrepancy *(Chwialkowski, Strathmann, Gretton. ICML'16)* *(Liu, Lee, Jordan. ICML'16)*	*(**Y**, Liu, Rao, Neville. ICML'18)*	*(**Y**, Rao, Neville. AISTATS'19)*

Networks and Point Processes



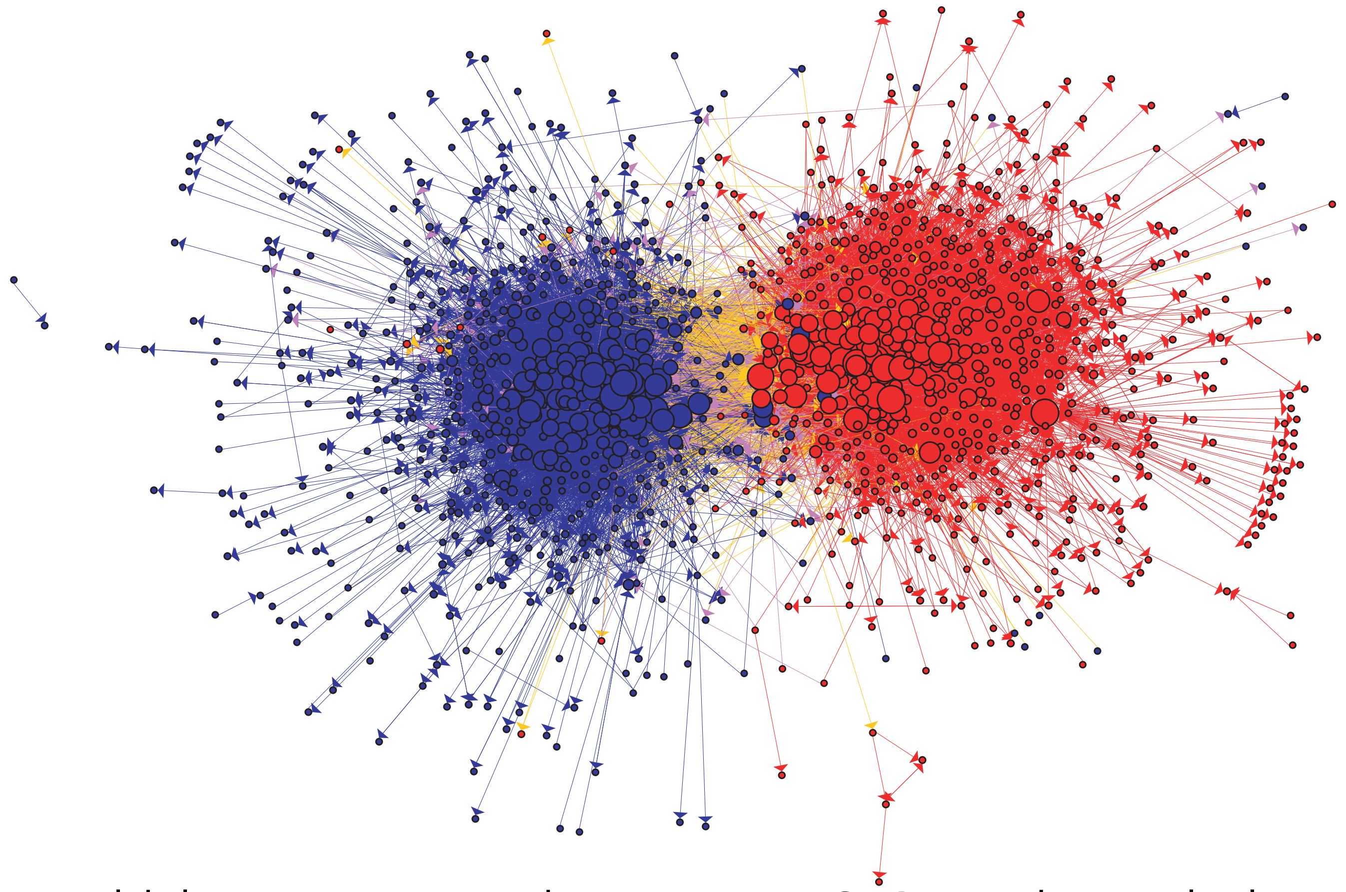
Collaboration graph centered on Erdős

https://oakland.edu/enp/trivia/



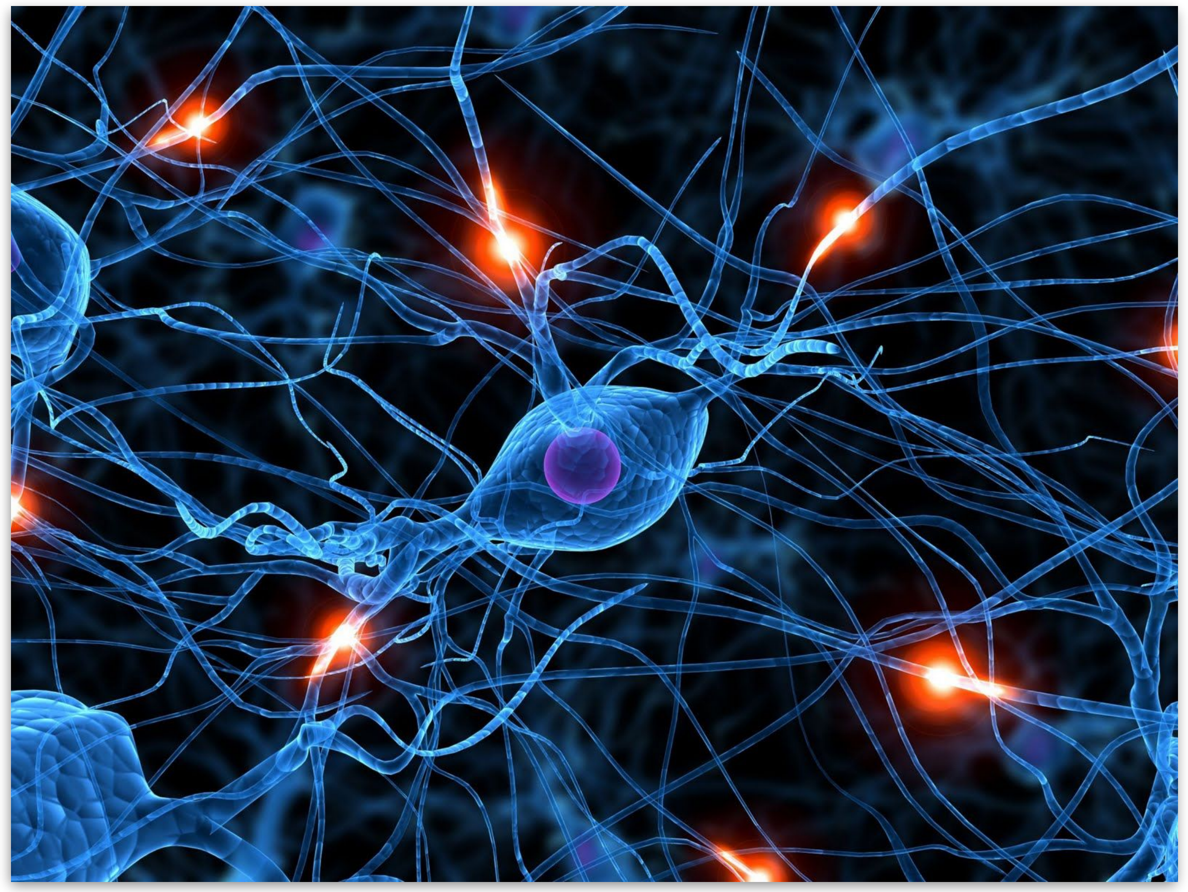
The Internet in 2005 and 2010

http://www.opte.org/the-internet/



Political blogs prior to the 2004 U.S. Presidential Election

(Adamic & Glance '05)

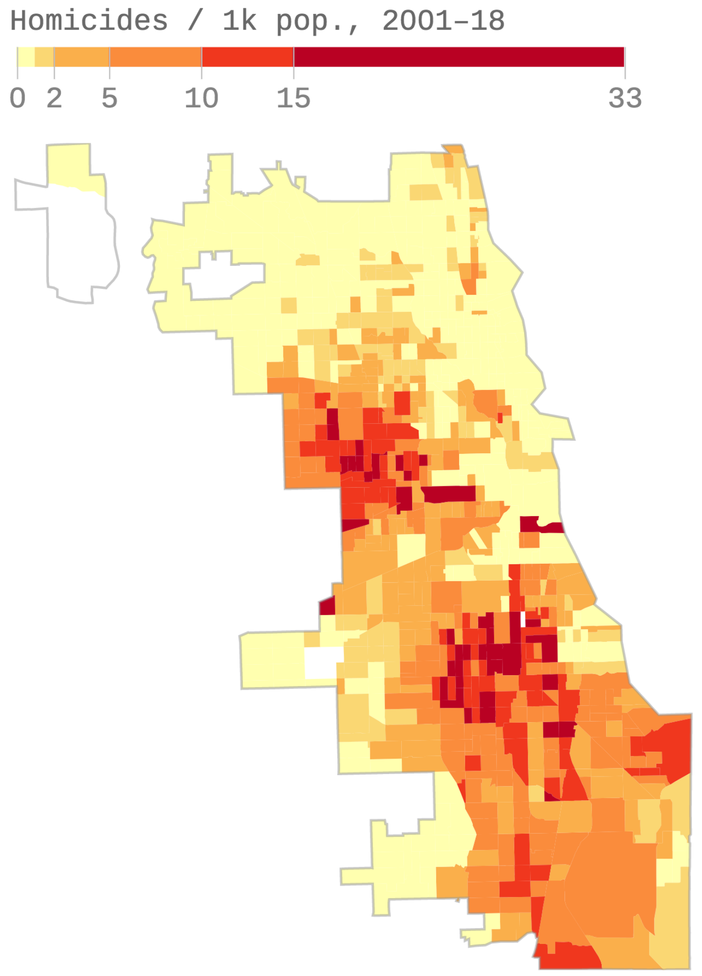


Neuron-firing patterns in the brain

https://www.pinterest.com/pin/394557617332618358/

Locations of trees in a forest

http://archive.stats.govt.nz/browse_for_stats/environment/environmental-reporting-series/environmental-indicators/Home/Land/distribution-indigenous-trees.aspx

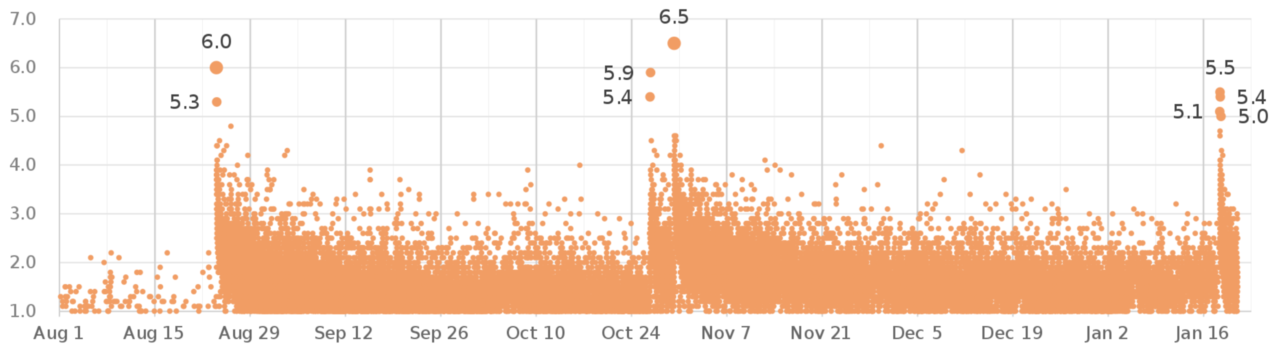


Homicides in Chicago

https://www.axios.com/chicago-gun-violence-murder-rate-statistics-4addeeec-d8d8-4ce7-a26b-81d428c14836.html

Distribution of earthquake aftershocks

http://www.earthquakepredict.com/2016/09/italy-earthquake-aerial-photos-show.html



https://en.wikipedia.org/wiki/January_2017_Central_Italy_earthquakes

Exponential Random Graph Model

(Wasserman and Pattison '96)

Distribution over graphs (adjacency matrices):

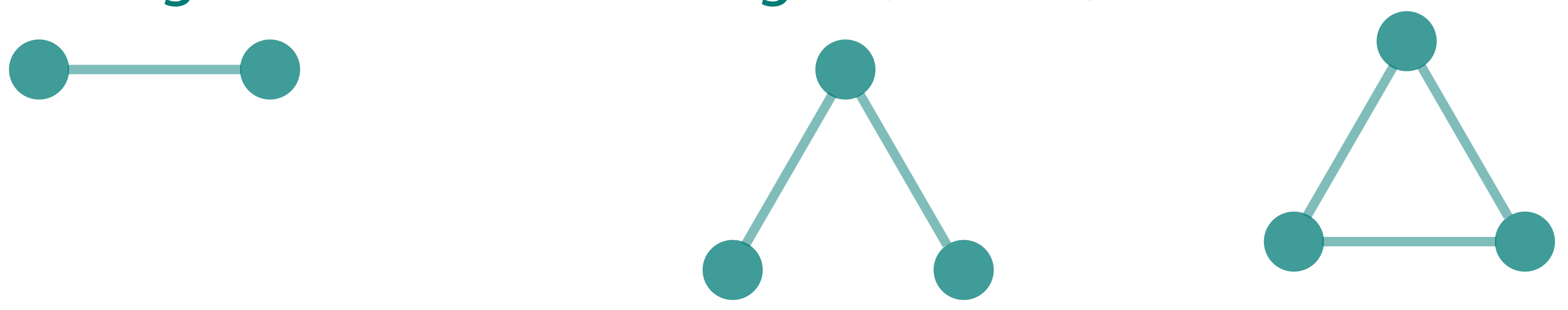
$$p(\mathbf{G}) = \frac{1}{Z} \exp\left\{\theta_1 E(\mathbf{G}) + \theta_2 S_2(\mathbf{G}) + \tau T(\mathbf{G})\right\}, \quad \mathbf{G} \in \{0,1\}^{n \times n}$$

Computing Z requires summing over 2^{n^2} configurations!

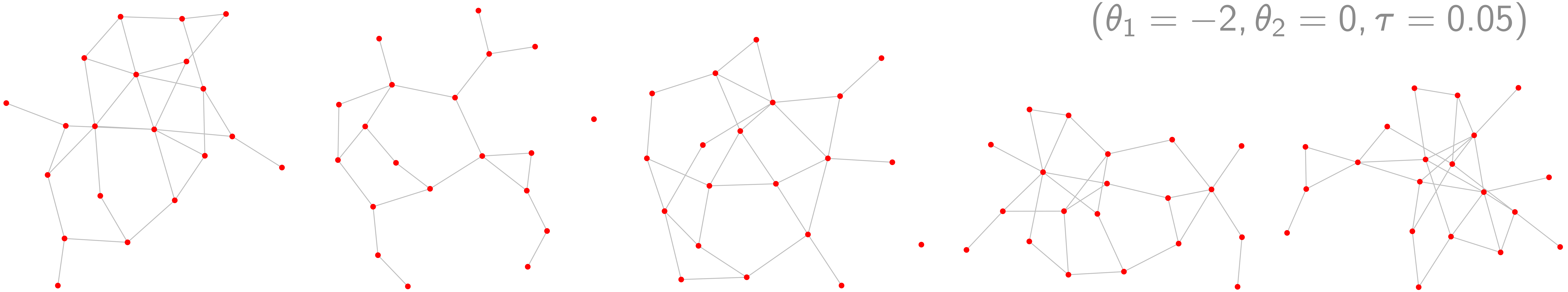
#Edges

#Wedges (2-stars)

#Triangles



$(\theta_1 = -2, \theta_2 = 0, \tau = 0.05)$

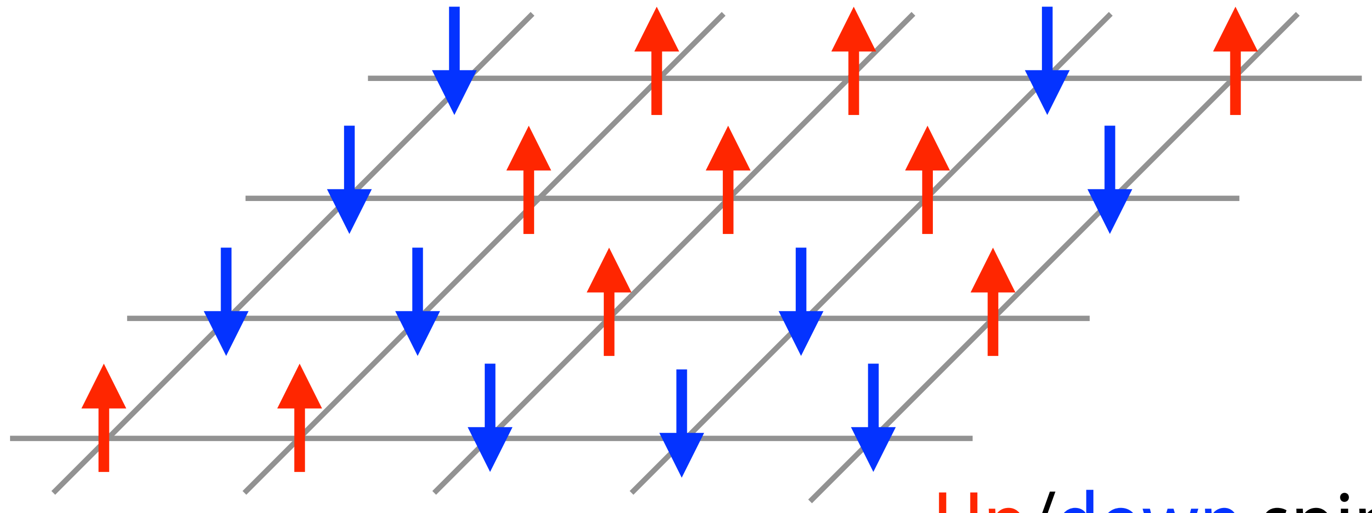


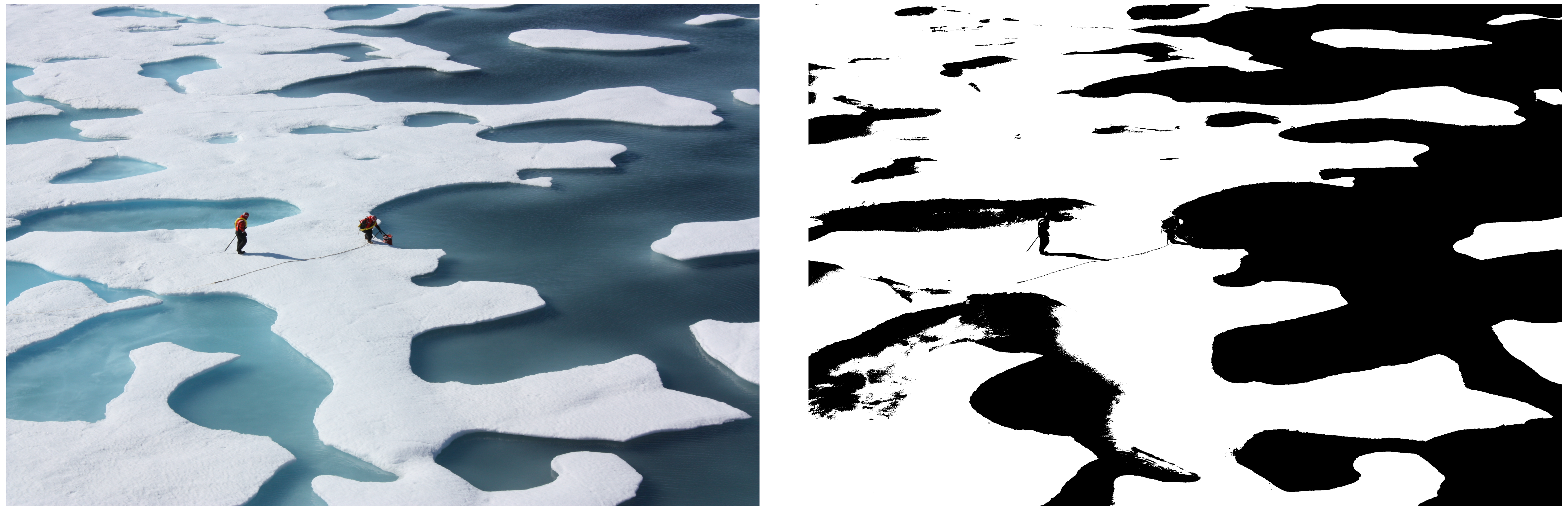
Ising Model

Given a 2-D lattice graph $G = (V, E)$,

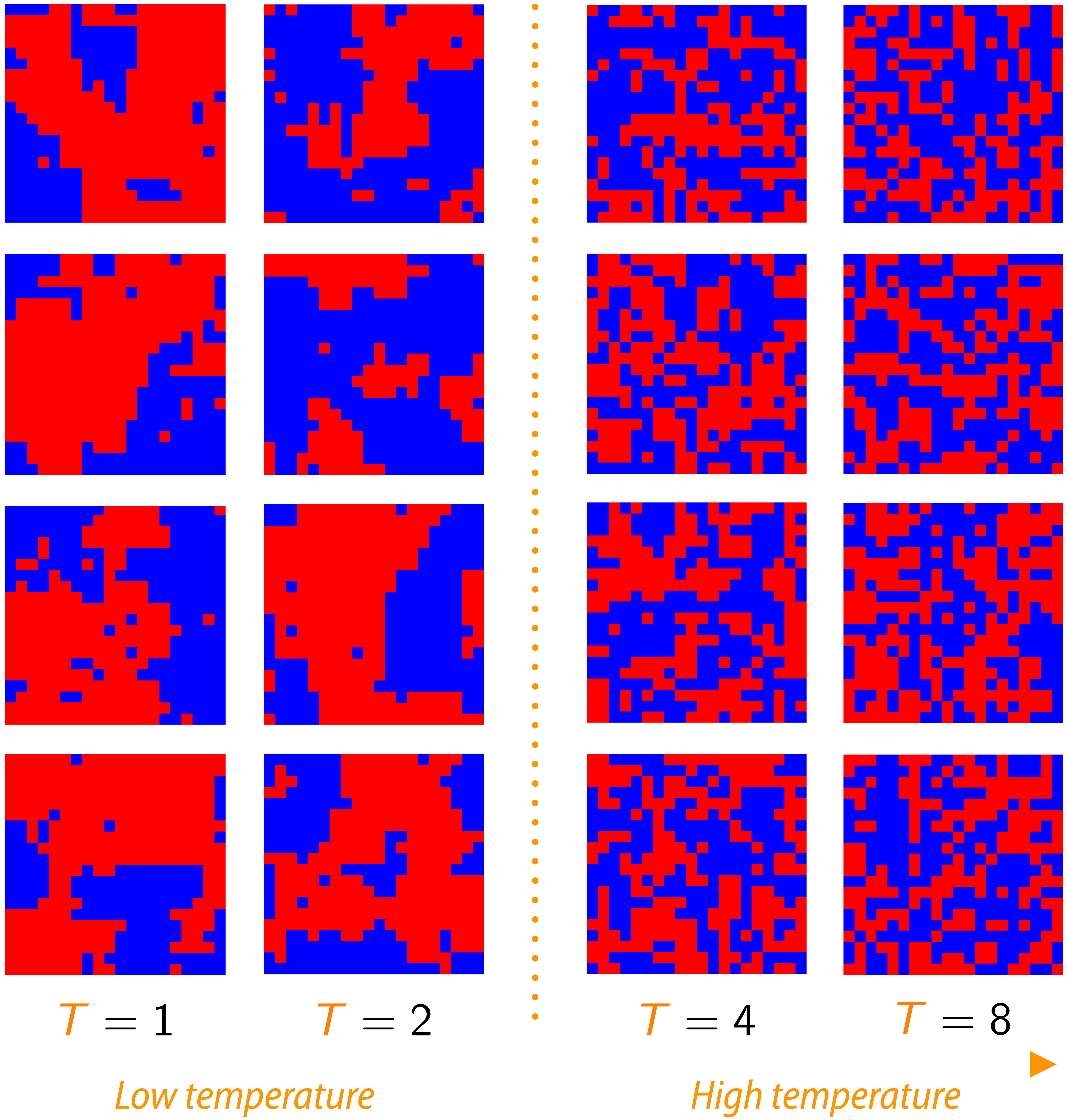
$$p(\mathbf{x}) = \frac{1}{Z} \exp\left\{ \sum_{(i,j)\in E} \frac{x_i x_j}{T} \right\}, \quad \mathbf{x} \in \{\pm 1\}^d$$

Computing Z requires summing over 2^d configurations!



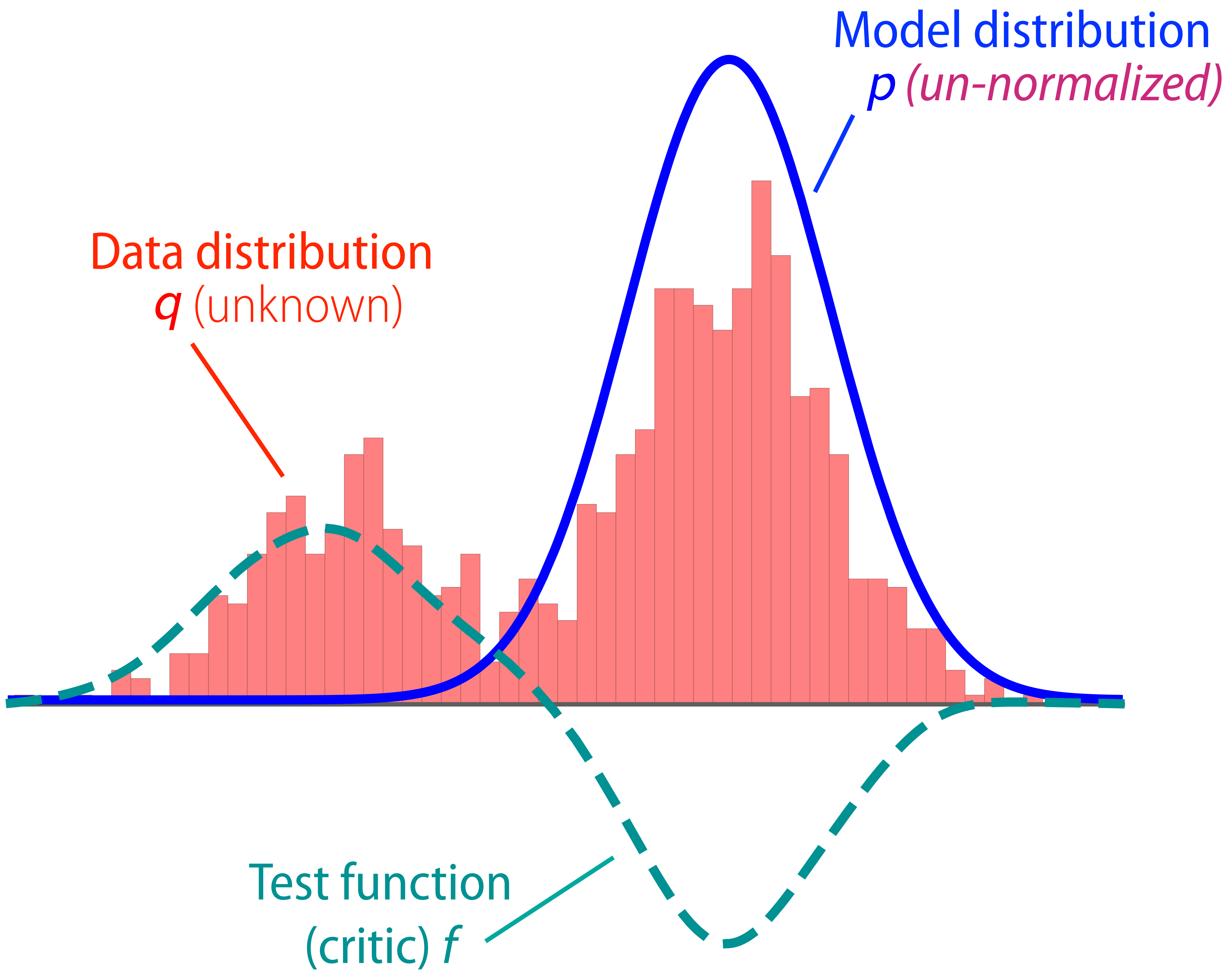


https://en.wikipedia.org/wiki/Melt_pond



Based on slides by Constantinos Daskalakis:
http://www.cs.columbia.edu/~ccanonne/workshop-focs2017/files/slides-workshop-daskalakis.pptx

Comparing Probability Distributions



Integral Probability Metrics (IPMs)

$$\sup_{f \in \mathcal{F}} \; \mathbb{E}_{\mathbf{x} \sim q}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p}[f(\mathbf{x})]$$

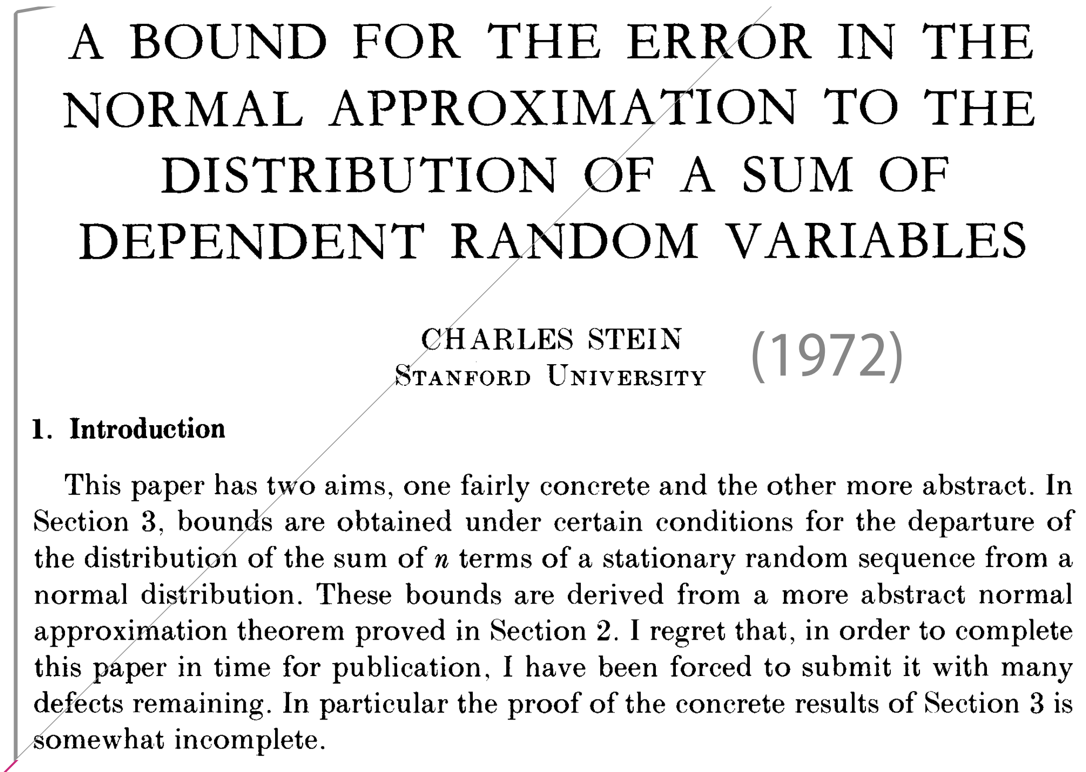
"test functions"

can estimate using samples 😀

cannot compute if p un-normalized! ☹️

$\mathcal{F}$	Metric
$\{f : \|f\|_\infty \leq 1\}$	Total variation distance
$\{\mathbf{1}_{(-\infty, t]} : t \in \mathbb{R}\}$	Kolmogorov distance
$\{f : \|f\|_L \leq 1\}$	Kantorovich metric (L_1-Wasserstein distance)[1]
$\{f : \|f\|_\infty + \|f\|_L \leq 1\}$	Dudley metric
$\{f : \|f\|_{\mathcal{H}} \leq 1\}$	Maximum mean discrepancy (Gretton et al. '12)

Comparing Unnormalized Distributions

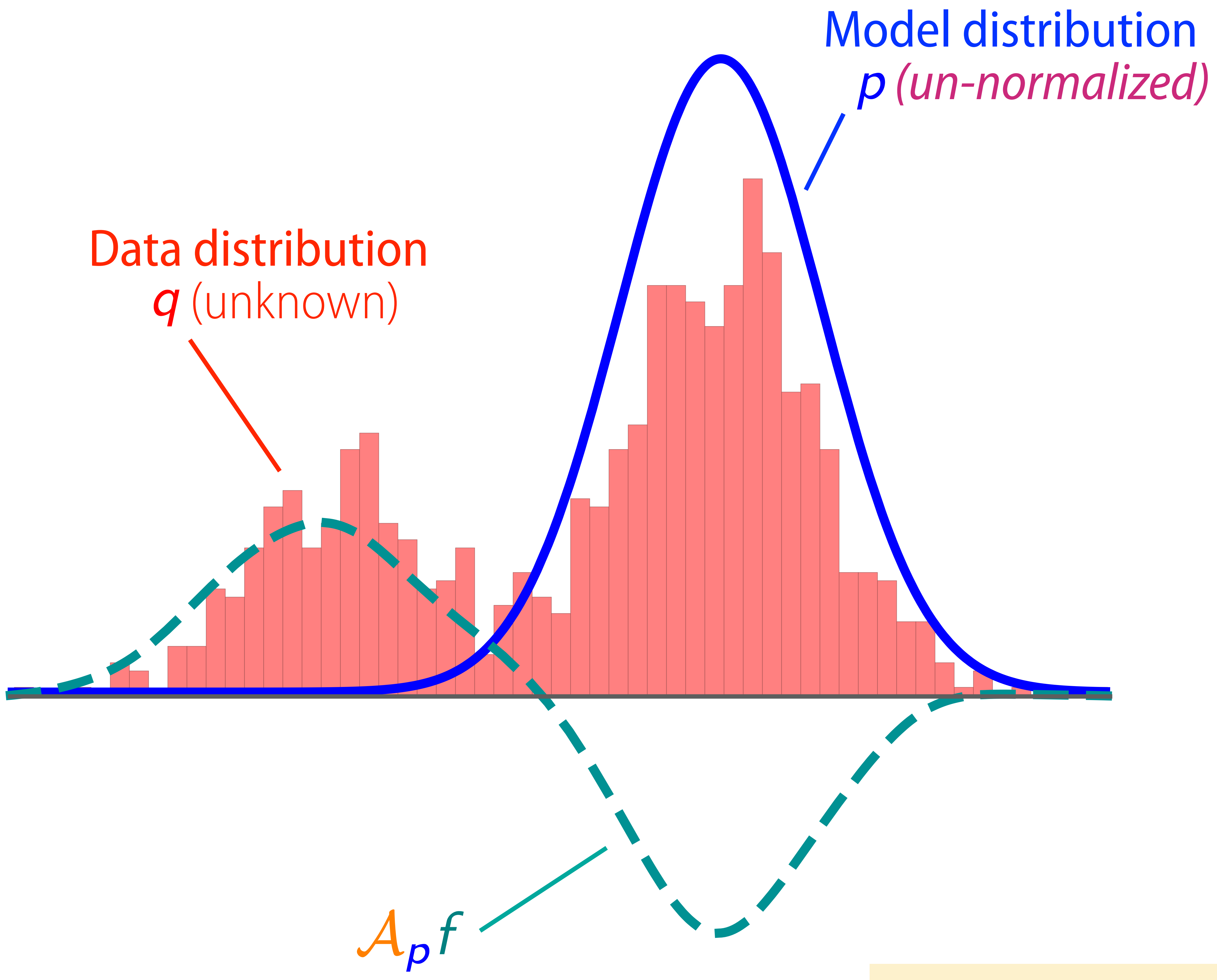

A BOUND FOR THE ERROR IN THE NORMAL APPROXIMATION TO THE DISTRIBUTION OF A SUM OF DEPENDENT RANDOM VARIABLES

CHARLES STEIN
STANFORD UNIVERSITY

1. Introduction

This paper has two aims, one fairly concrete and the other more abstract. In Section 3, bounds are obtained under certain conditions for the departure of the distribution of the sum of n terms of a stationary random sequence from a normal distribution. These bounds are derived from a more abstract normal approximation theorem proved in Section 2. I regret that, in order to complete this paper in time for publication, I have been forced to submit it with many defects remaining. In particular the proof of the concrete results of Section 3 is somewhat incomplete.

(1972)



Stein Discrepancy

(Gorham & Mackey '15, Chwialkowski et al. '16, Liu et al. '16)

$$\sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{x}\sim q}[\mathcal{A}_p f(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim p}[\mathcal{A}_p f(\mathbf{x})]$$

Find *Stein operator* $\mathcal{A}_p$ s.t.

$$\mathbb{E}_{\mathbf{x}\sim q}[\mathcal{A}_p f(\mathbf{x})] = 0, \ \forall f \in \mathcal{F}$$

(Stein identity)

if and only if $p = q$.

- For a smooth density p on $\mathbb{R}^d$, set

$$\mathcal{A}_p f(\mathbf{x}) := \nabla_{\mathbf{x}} \log p(\mathbf{x}) \cdot f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})$$

(can still be evaluated when p is un-normalized!)

Applies only to **continuous distributions** with **smooth densities**!

What About Discrete Distributions?

Gradients $\nabla_{\mathbf{x}}$ are no longer available!

Consider a finite set $\mathcal{X}$: $\nabla_{\mathbf{x}} = (\ldots, \frac{\partial}{\partial x_i}, \ldots)^\top$ is not defined on $\mathcal{X}^d$!

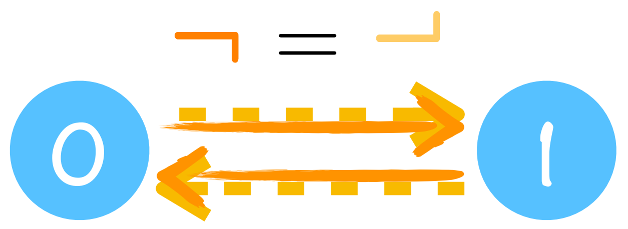
💡 Difference operator For any $\mathbf{x} \in \mathbb{R}^d$ and function $f : \mathcal{X}^d \to \mathbb{R}$,

$$\Delta f(\mathbf{x}) := (\ldots, f(\mathbf{x}) - f(\neg_i \mathbf{x}), \ldots)^\top \qquad \Delta^* f(\mathbf{x}) := (\ldots, f(\mathbf{x}) - f(\lrcorner_i \mathbf{x}), \ldots)^\top$$

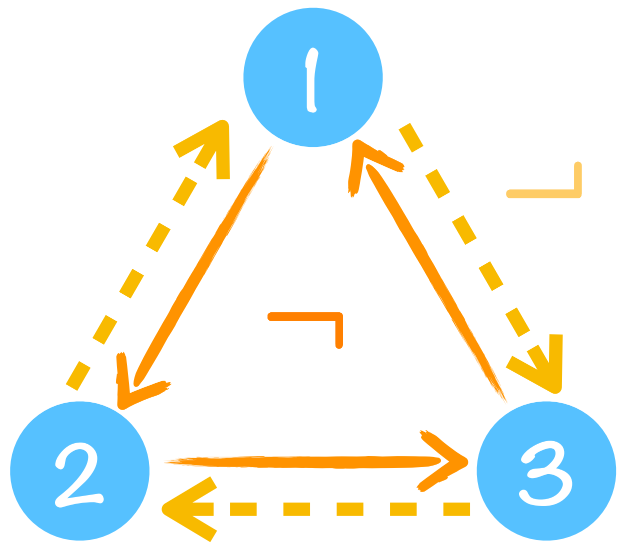
💡 Difference Stein operator For any function f and pmf p,

$$\mathcal{A}_p f(\mathbf{x}) := \frac{\Delta p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) - \Delta^* f(\mathbf{x})$$

normalization constant in p cancels out!



$\mathcal{X} = \{0, 1\}$



$\mathcal{X} = \{1, 2, 3\}$

Recall: Continuous case:

$$\mathcal{A}_p f(\mathbf{x}) = \frac{\nabla p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) + \nabla f(\mathbf{x})$$

Theorem (Difference Stein identity) For any function f and pmf p, $\mathbb{E}_{\mathbf{x}\sim p}[\mathcal{A}_p f(\mathbf{x})] = 0$.

Theorem For *positive* pmfs p and q, $\mathbb{E}_{\mathbf{x}\sim q}[\mathcal{A}_p f(\mathbf{x})] = 0, \forall f$ *iff.* $p = q$.

Characterization of Stein Operators

Theorem For any positive pmf p on $\mathcal{X}^d$, a linear operator $\mathcal{T}_p$ satisfies

$$\mathbb{E}_{\mathbf{x}\sim p}[\mathcal{T}_p f(\mathbf{x})] = 0 \qquad \text{(Stein identity)}$$

for all functions $f \in \mathcal{F}$ if and only if there exist linear operators

$$\mathcal{L} f(\mathbf{x}) = \sum_{\mathbf{x}' \in \mathcal{X}^d} g(\mathbf{x}, \mathbf{x}')\, f(\mathbf{x}'), \quad \mathcal{L}^* f(\mathbf{x}) = \sum_{\mathbf{x}' \in \mathcal{X}^d} g(\mathbf{x}', \mathbf{x})\, f(\mathbf{x}'), \quad \forall f \in \mathcal{F}$$

for some bivariate function g on $\mathcal{X}^d \times \mathcal{X}^d$, s.t.

$$\mathcal{T}_p f(\mathbf{x}) = \frac{\mathcal{L} p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) - \mathcal{L}^* f(\mathbf{x})$$

holds for all $\mathbf{x} \in \mathcal{X}^d$ and functions $f \in \mathcal{F}$.

"adjoint operators"

- Continuous case:

$$\mathcal{L} = \nabla, \ \mathcal{L}^* = -\nabla\cdot$$

- Discrete case:

$$\mathcal{L} = \Delta, \ \mathcal{L}^* = \Delta^*$$

- General recipe:
  - Graph-based construction (e.g., via Laplacian)

Discrete Stein Discrepancy

Kernelized Discrete Stein Discrepancy (KDSD)

For some space $\mathcal{F}$ of functions $\mathbf{f} : \mathcal{X}^d \to \mathbb{R}^d$,

$$\mathbb{D}(q \,\|\, p) := \sup_{\mathbf{f} \in \mathcal{H}^d,\ \|\mathbf{f}\|_{\mathcal{H}^d} \le 1} \mathbb{E}_{\mathbf{x} \sim q}\left[\mathrm{tr}\left(\mathcal{A}_p \mathbf{f}(\mathbf{x})\right)\right]$$

$\mathcal{H}$: reproducing kernel Hilbert space (RKHS) with kernel $k(\cdot, \cdot)$

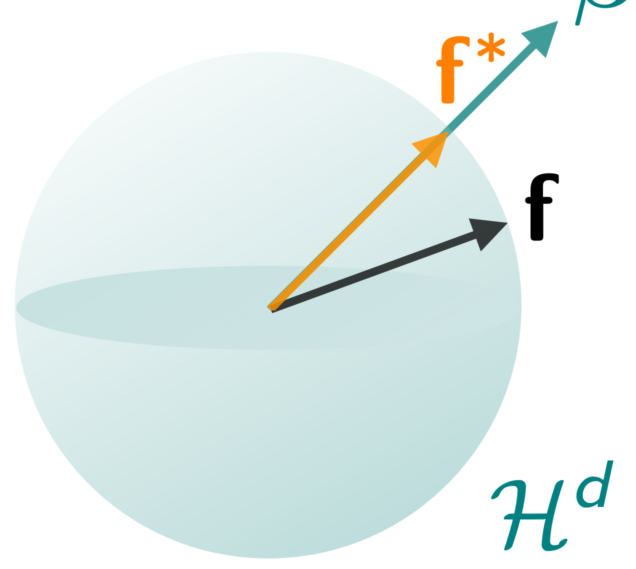
- Exponentiated Hamming kernel

$$k(\mathbf{x}, \mathbf{x}') = e^{-H(\mathbf{x}, \mathbf{x}')} \quad \left(H(\mathbf{x}, \mathbf{x}') := \frac{1}{d} \sum_{i=1}^{d} \mathbb{I}\{x_i \neq x_i'\}\right)$$

- Kernels for structured data

  Graph kernels, string kernels, etc.

$$\boldsymbol{\beta} := \mathbb{E}_{\mathbf{x} \sim q}\left[\mathcal{A}_p k(\cdot, \mathbf{x})\right]$$



$$\mathbb{D}^2(q \,\|\, p) = \|\boldsymbol{\beta}\|^2_{\mathcal{H}^d}$$

Theorem Optimizing over RKHS yields closed-form solution:

$$\mathbb{D}^2(q \,\|\, p) = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim q}\left[\kappa_p(\mathbf{x}, \mathbf{x}')\right]$$

where $\kappa_p(\mathbf{x}, \mathbf{x}') := \mathbf{s}_p(\mathbf{x})^\top k(\mathbf{x}, \mathbf{x}')\, \mathbf{s}_p(\mathbf{x}') - \mathbf{s}_p(\mathbf{x})^\top \Delta^*_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') - \Delta^*_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \mathrm{tr}(\Delta^*_{\mathbf{x}, \mathbf{x}'} k(\mathbf{x}, \mathbf{x}'))$

$\left(\mathbf{s}_p(\mathbf{x}) := \Delta p(\mathbf{x}) / p(\mathbf{x})\right)$

- Estimate from samples $\{\mathbf{x}_i\}_{i=1}^n \sim q$:

$$\widehat{\mathbb{D}^2}(q \,\|\, p) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \kappa_p(\mathbf{x}_i, \mathbf{x}_j)$$

Use as test statistic!

KDSD Goodness-of-Fit Test

Given a probability distribution p on $\mathcal{X}^d$ and *data samples* $\{\mathbf{x}_i\}_{i=1}^n \sim q$, test

$$H_0 : p = q \qquad \text{vs.} \qquad H_1 : p \neq q$$

💡 **Goodness-of-Fit Test**

- Compute KDSD test statistic

$$\widehat{\mathbb{D}^2}(q \,\|\, p) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \kappa_p(\mathbf{x}_i, \mathbf{x}_j)$$

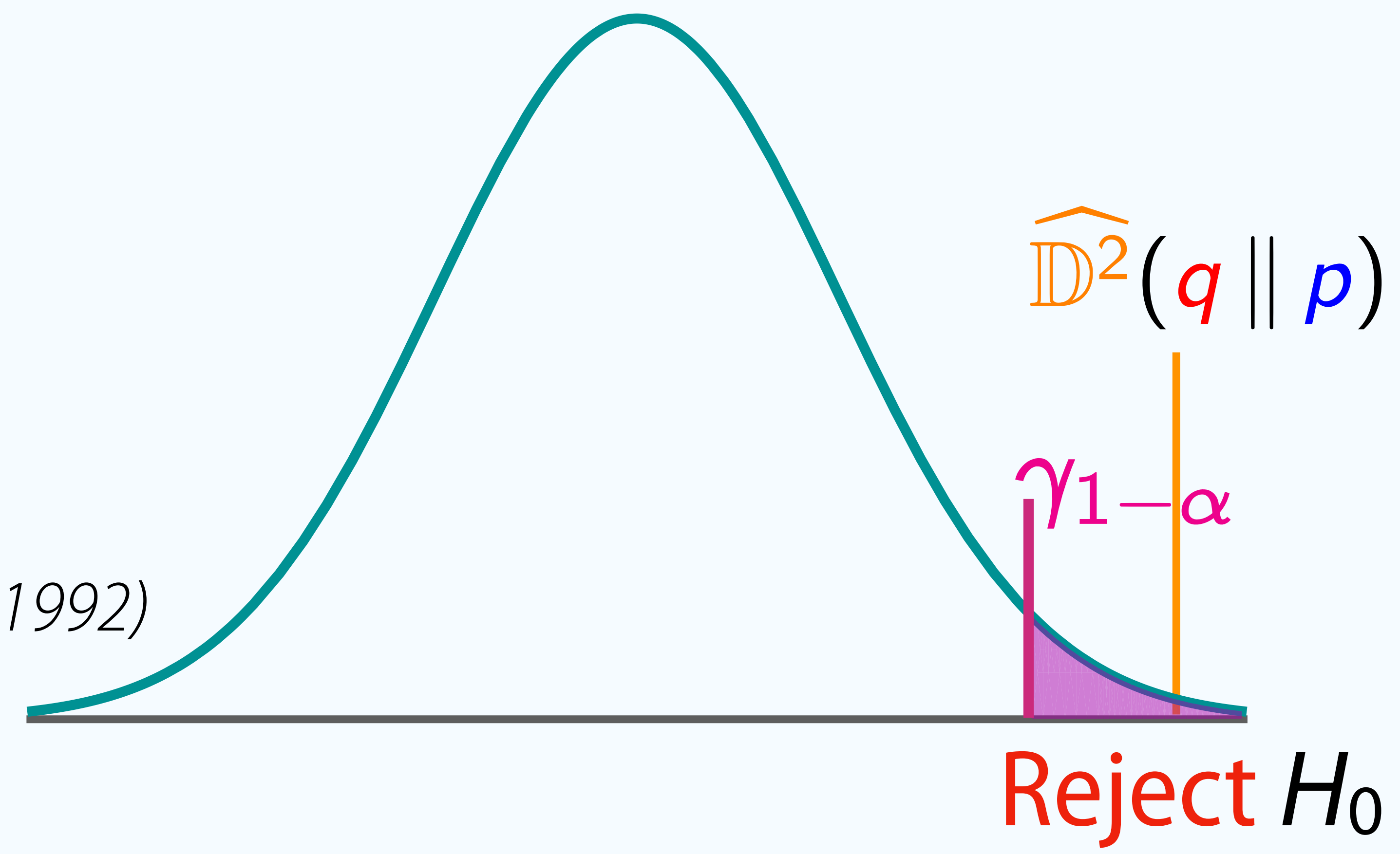
- Compute critical value $\gamma_{1-\alpha}$ via generalized bootstrap

(Arcones & Gine, 1992)

$w_1, \ldots, w_n \sim \text{Mult}(1/n, \ldots, 1/n)$
$\widetilde{w}_i = (w_i - 1)/n$

$$\widetilde{\mathbb{D}^2}(q \,\|\, p) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \widetilde{w}_i \, \widetilde{w}_j \, \kappa_p(\mathbf{x}_i, \mathbf{x}_j)$$

- Decision rule: Reject H_0 if $\widehat{\mathbb{D}^2}(q \,\|\, p) > \gamma_{1-\alpha}$



Model does not fit observed data!

Example: KDSD GoF Test for Ising Model

$$p(\mathbf{x}) \propto \exp\left\{ \sum_{(i,j)\in E} \frac{x_i x_j}{T_0} \right\}$$

$$q(\mathbf{x}) \propto \exp\left\{ \sum_{(i,j)\in E} \frac{x_i x_j}{T} \right\}$$

Given samples $\{\mathbf{x}_i\}_{i=1}^n \sim q$ on $\{\pm 1\}^d$, test

$$H_0 : T = T_0 \quad \text{vs.} \quad H_1 : T \neq T_0$$

- Compute KDSD test statistic

$$\widehat{\mathbb{D}^2}(q \,\|\, p) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \kappa_p(\mathbf{x}_i, \mathbf{x}_j)$$

$$k(\mathbf{x}, \mathbf{x}') = e^{-H(\mathbf{x}, \mathbf{x}')}$$

$$\mathbf{s}_p(\mathbf{x}) = \Delta p(\mathbf{x}) / p(\mathbf{x}) = \left(1 - \exp\left\{ -2x_i \sum_{j \in \mathcal{N}_i} \frac{x_j}{T_0} \right\} \right)_{i=1}^d$$

where $\kappa_p(\mathbf{x}, \mathbf{x}') := \mathbf{s}_p(\mathbf{x})^\top k(\mathbf{x}, \mathbf{x}')\, \mathbf{s}_p(\mathbf{x}') - \mathbf{s}_p(\mathbf{x})^\top \Delta^*_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') - \Delta^*_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \mathrm{tr}(\Delta^*_{\mathbf{x}, \mathbf{x}'} k(\mathbf{x}, \mathbf{x}'))$

- Compute critical value $\gamma_{1-\alpha}$ via generalized bootstrap

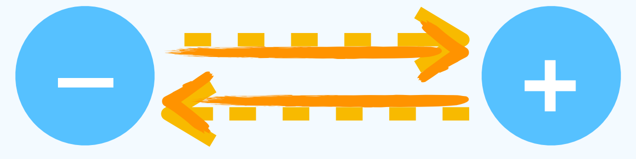
(Arcones & Gine, 1992)

$$\widehat{\mathbb{D}^2}(q \,\|\, p) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \widetilde{w}_i \, \widetilde{w}_j \, \kappa_p(\mathbf{x}_i, \mathbf{x}_j)$$

$$\Delta f(\mathbf{x}) := (\ldots, f(\mathbf{x}) - f(\neg_i \mathbf{x}), \ldots)^\top$$

$$\Delta^* f(\mathbf{x}) := (\ldots, f(\mathbf{x}) - f(\lnot_i \mathbf{x}), \ldots)^\top$$

$\neg = \lrcorner$



- Decision rule: Reject H_0 if $\widehat{\mathbb{D}^2}(q \,\|\, p) > \gamma_{1-\alpha}$

Empirical Evaluation

MMD two-sample test:

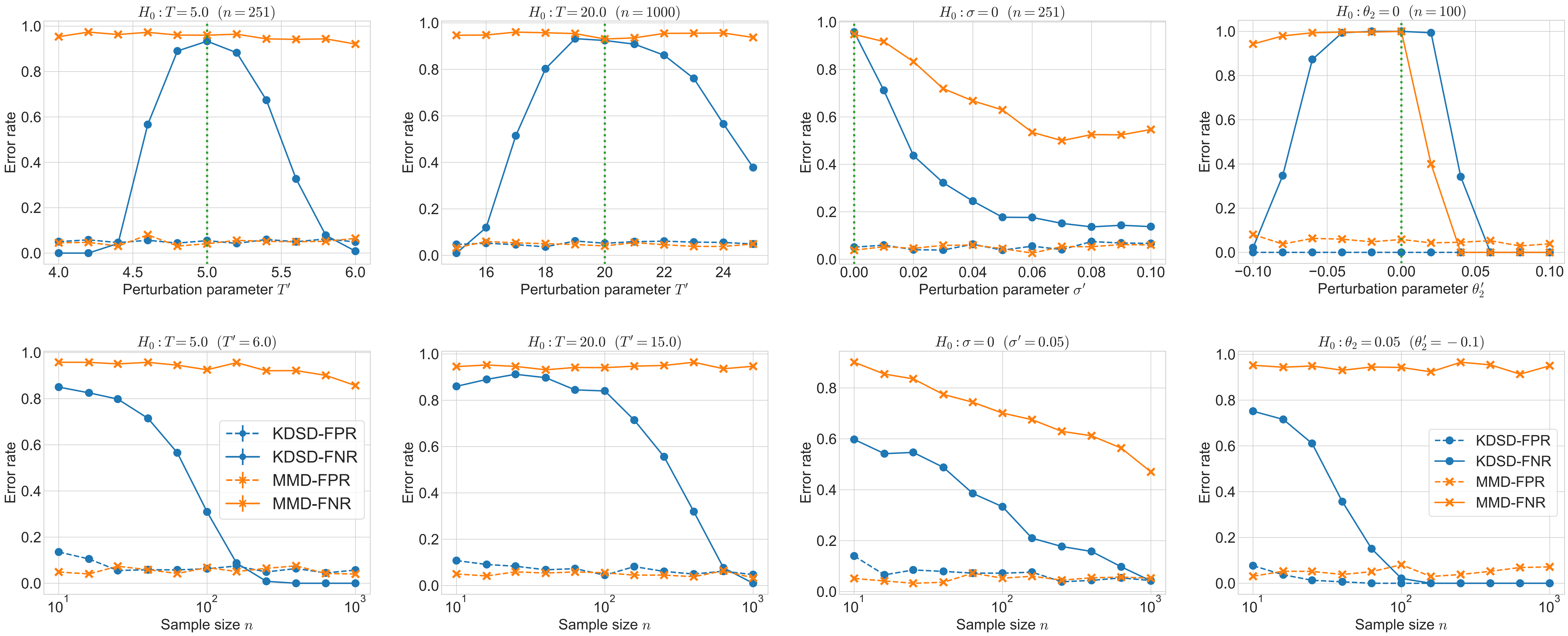
Requires samples from both p and q!

$\{\mathbf{x}_i\}_{i=1}^m \sim p$

$\{\mathbf{y}_i\}_{i=1}^n \sim q$

$$\mathrm{MMD}_u^2 = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n(n-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(\mathbf{x}_i, \mathbf{y}_j)$$

$H_0 : T = 5$ vs. $H_1 : T \neq 5$ | $H_0 : T = 20$ vs. $H_1 : T \neq 20$ | $H_0 : \sigma = 0$ vs. $H_1 : \sigma \neq 0$ | $H_0 : \theta_2 = 0$ vs. $H_1 : \theta_2 \neq 0$



Ising model | Ising model | Bernoulli RBM | ERGM (Use W-L graph kernel)

So Far…

GoF testing for distributions over fixed-length vectors (∇, Δ defined only for vectors).

	Continuous distributions	Discrete distributions	Point processes
Normalized	Kolmogorov–Smirnov test Cramér–von Mises test Anderson–Darling test	Chi-squared test	(mainly Poisson-type)
Unnormalized	Kernelized Stein discrepancy *(Chwialkowski, Strathmann, Gretton. ICML'16)* *(Liu, Lee, Jordan. ICML'16)*	✔	**?**

But point processes are distributions over **sets** containing an **arbitrary** number of points!

Need a new set of tools!

Towards a Stein Operator for Point Processes

Gibbs processes

$\psi_k > 0\,(k \geq 2) \Rightarrow$ Repulsion

k-th order interaction potential

Density

$$f(\phi) = \frac{1}{Z} \exp\left\{ -\sum_{k=1}^{|\phi|} \sum_{\omega \subseteq \phi,\, |\omega|=k} \psi_k(\omega) \right\}$$

Point pattern (**set** of points)

Intractable!

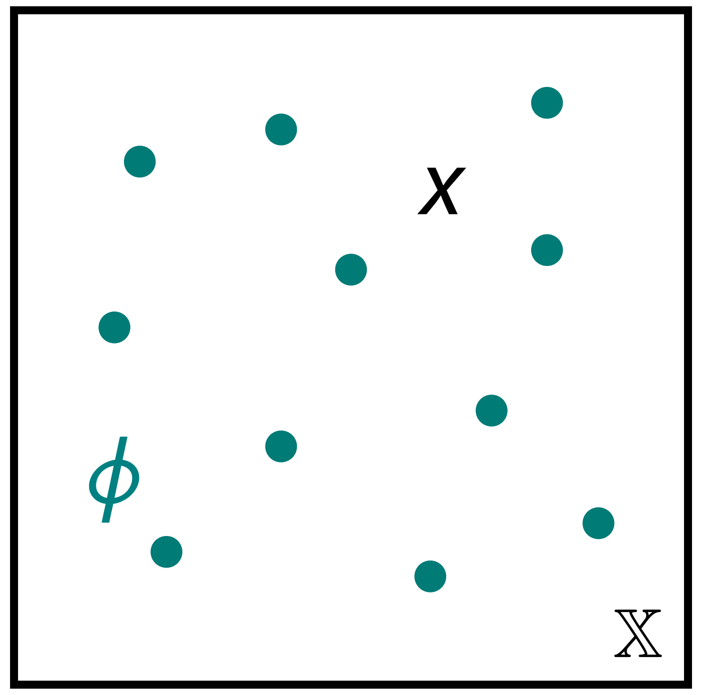
Intensity function $\lambda(x)$ is also intractable! ☹

Poisson process: $\psi_k \equiv 0, \ \forall k \geq 2$

Strauss process: $\psi_1(\{x\}) \equiv -\beta$

$\psi_2(\{x, y\}) = -(\log \gamma) \cdot \mathbb{I}\{\|x - y\|_2 \leq r\}$

Papangelou conditional intensity



Z's cancel out!

$$\rho(x|\phi) = \begin{cases} \dfrac{f(\phi \cup \{x\})}{f(\phi)}, & x \notin \phi \\ \dfrac{f(\phi)}{f(\phi \backslash \{x\})}, & x \in \phi \end{cases}$$

Gibbs process:

$$\rho(x|\phi) = \exp\left\{ -\sum_{k=1}^{|\phi|} \sum_{\omega \subseteq \phi,\, |\omega|=k-1} \psi_k(\{x\} \cup \omega) \right\}$$

Poisson process: $\rho(x|\phi) \equiv \lambda(x)$

Strauss process: $\rho(x|\phi) = \beta \gamma^{t_r(x,\phi)}$

$t_r(x, \phi) := \sum_{y \in \phi} \mathbb{I}\{\|x - y\|_2 \leq r\}$

A General Stein Operator for Point Processes

Stein–Papangelou operator For any function h and Papangelou intensity ρ, define

$$(\mathcal{A}_\rho h)(\phi) = \int_{\mathbb{X}} \underbrace{[h(\phi \cup \{x\}) - h(\phi)]}_{\text{“forward” difference}} \rho(x|\phi)\, \mathrm{d}x + \sum_{x \in \phi} \underbrace{[h(\phi \backslash \{x\}) - h(\phi)]}_{\text{“backward” difference}}$$

Recall: Difference Stein operator $\mathcal{A}_p f(\mathbf{x}) := \frac{\Delta p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) - \Delta^* f(\mathbf{x})$

Theorem (Stein identity) $\Phi \sim \rho \;\Rightarrow\; \mathbb{E}[\mathcal{A}_\rho h(\Phi)] = 0$ for all bounded functions h.

Proof Uses the *Georgii–Nguyen–Zessin (GNZ) formula* from point process theory.

For Poisson processes:
- $\rho(x|\phi) \equiv \lambda(x)$; recovers previously known result *(Barbour & Brown, 1992)*
- $\mathbb{E}[\mathcal{A}_\rho h(\Phi)] = 0,\ \forall h \;\Rightarrow\; \Phi \sim \rho$

(May be insufficient for non-Poisson processes.)

Kernelized Stein Discrepancy for Point Processes

Kernelized Stein Discrepancy

$$\mathbb{D}(\eta \,\|\, \rho) := \sup_{h \in \mathcal{H},\, \|h\|_{\mathcal{H}} \le 1} \mathbb{E}_{\Phi \sim \eta}\left[\mathcal{A}_\rho h(\Phi)\right]$$

$\mathcal{H}$: reproducing kernel Hilbert space (RKHS) with kernel $k(\cdot,\cdot)$

- An MMD-based kernel for point processes

$$k_{\mathcal{M}}(\phi, \psi) := \exp\{-\widehat{d^2}(\phi, \psi)\}$$

$$\widehat{d^2}(\phi, \psi) := \frac{1}{|\phi|^2} \sum_{x \in \phi} \sum_{x' \in \phi} k_{\mathbb{X}}(x, x') + \frac{1}{|\psi|^2} \sum_{y \in \psi} \sum_{y' \in \psi} k_{\mathbb{X}}(y, y') - \frac{2}{|\phi| \cdot |\psi|} \sum_{x \in \phi} \sum_{y \in \psi} k_{\mathbb{X}}(x, y) \quad \text{(MMD } V\text{-statistic estimate)}$$

Theorem $\mathbb{D}(\eta \,\|\, \rho) = \mathbb{E}_{\Phi, \Psi \sim \eta}\left[\kappa_\rho(\Phi, \Psi)\right]$

where

$$\begin{aligned}
\kappa_\rho(\phi, \psi) = &\int_{\mathbb{X}} \int_{\mathbb{X}} \Big[k(\phi \cup \{u\}, \psi \cup \{v\}) - k(\phi, \psi \cup \{v\}) - k(\phi \cup \{u\}, \psi) + k(\phi, \psi)\Big] \rho(u|\phi)\, \rho(v|\psi)\, \mathrm{d}u\, \mathrm{d}v \\
&+ \int_{\mathbb{X}} \Big[\sum_{x \in \phi} \big[k(\phi \backslash \{x\}, \psi \cup \{v\}) - k(\phi \backslash \{x\}, \psi)\big] - |\phi| \cdot \big[k(\phi, \psi \cup \{v\}) - k(\phi, \psi)\big]\Big] \rho(v|\psi)\, \mathrm{d}v \\
&+ \int_{\mathbb{X}} \Big[\sum_{y \in \psi} \big[k(\phi \cup \{u\}, \psi \backslash \{y\}) - k(\phi, \psi \backslash \{y\})\big] - |\psi| \cdot \big[k(\phi \cup \{u\}, \psi) - k(\phi, \psi)\big]\Big] \rho(u|\phi)\, \mathrm{d}u \\
&+ \Big[\sum_{x \in \phi} \sum_{y \in \psi} k(\phi \backslash \{x\}, \psi \backslash \{y\}) - |\phi| \cdot \sum_{y \in \psi} k(\phi, \psi \backslash \{y\}) - |\psi| \cdot \sum_{x \in \phi} k(\phi \backslash \{x\}, \psi) + |\phi| \cdot |\psi| \cdot k(\phi, \psi)\Big]
\end{aligned}$$

Require numerical integration

Goodness-of-Fit Test for Point Processes

Given a Papangelou conditional intensity ρ and *point patterns* $\{\mathcal{X}_i\}_{i=1}^n \sim \eta$, test

(point-sets)

$$H_0 : \eta = \rho \qquad \text{vs.} \qquad H_1 : \eta \neq \rho$$

💡 **Goodness-of-Fit Test**

- Compute KDSD test statistic

$$\widehat{\mathbb{D}^2}(\eta \,\|\, \rho) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \kappa_\rho(\mathcal{X}_i, \mathcal{X}_j)$$

- Compute critical value $\gamma_{1-\alpha}$ via generalized bootstrap

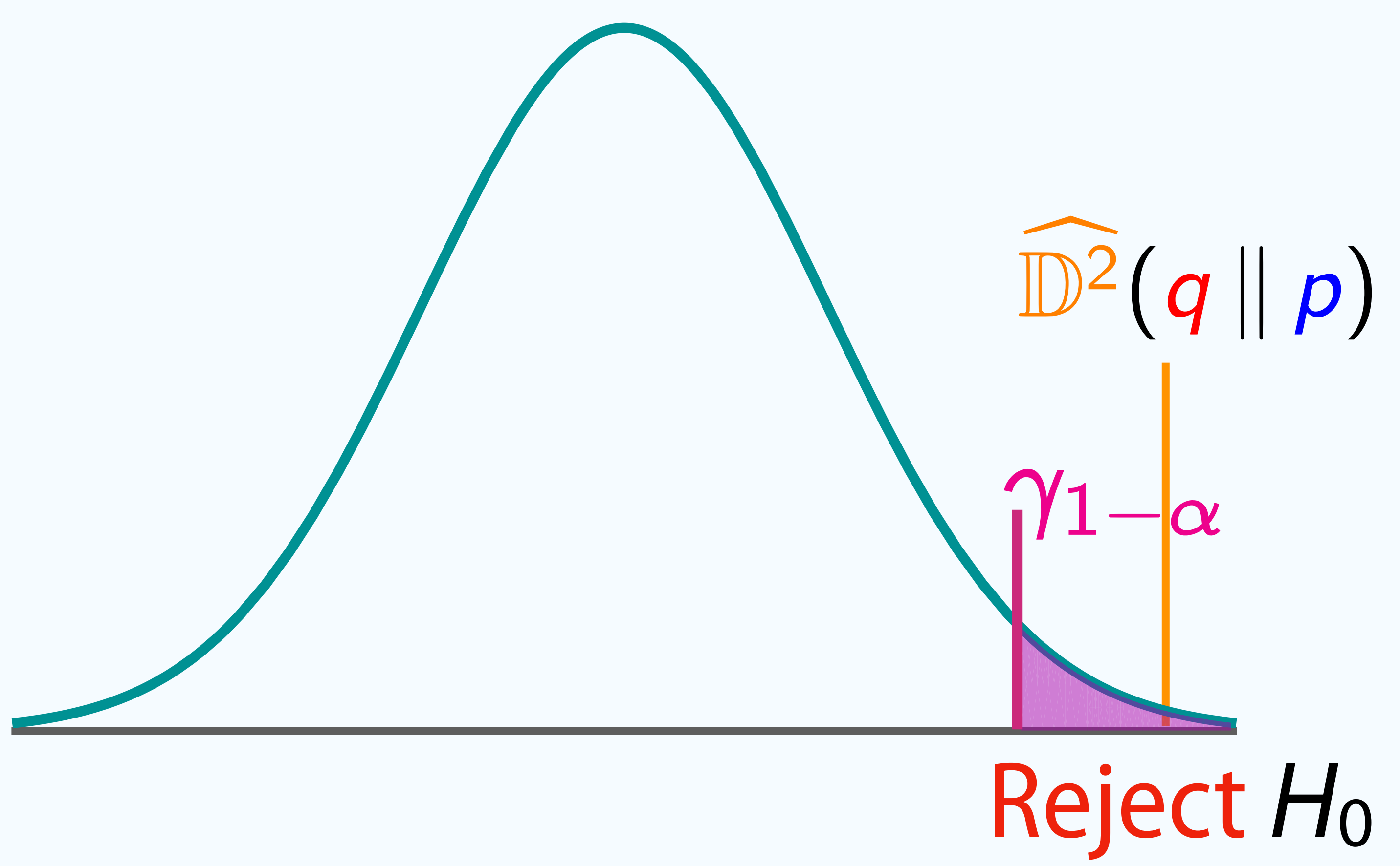
(Arcones & Gine, 1992)

$w_1, \ldots, w_n \sim \mathrm{Mult}(1/n, \ldots, 1/n)$

$\widetilde{w}_i = (w_i - 1)/n$

$$\widetilde{\mathbb{D}^2}(\eta \,\|\, \rho) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \widetilde{w}_i \, \widetilde{w}_j \, \kappa_\rho(\mathcal{X}_i, \mathcal{X}_j)$$

- Decision rule: Reject H_0 if $\widehat{\mathbb{D}^2}(\eta \,\|\, \rho) > \gamma_{1-\alpha}$



Model does not fit observed data!

Empirical Evaluation

MMD two-sample test:

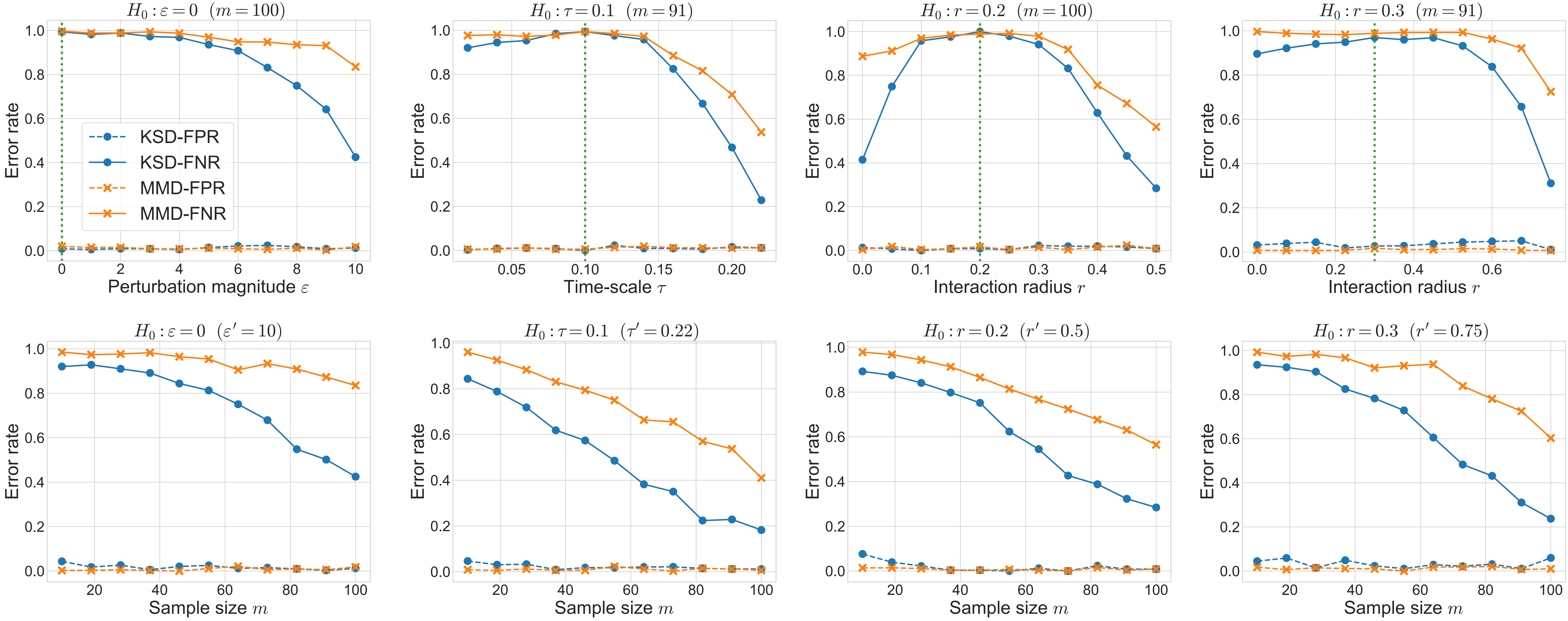
$\{\mathcal{X}_i\}_{i=1}^m \sim \rho$

$\{\mathcal{Y}_i\}_{i=1}^n \sim \eta$

$$\mathrm{MMD}_u^2 = \frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathcal{X}_i, \mathcal{X}_j) + \frac{1}{n(n-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} k(\mathcal{Y}_i, \mathcal{Y}_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(\mathcal{X}_i, \mathcal{Y}_j)$$

Requires samples from both p and q!

$H_0 : \varepsilon = 0$ vs. $H_1 : \varepsilon \neq 0$ | $H_0 : \tau = 0.1$ vs. $H_1 : \varepsilon \neq 0.1$ | $H_0 : r = 0.2$ vs. $H_1 : r \neq 0.2$ | $H_0 : r = 0.3$ vs. $H_1 : r \neq 0.3$



Poisson process ($d = 2$) | Hawkes process ($d = 1$) | Strauss process ($d = 1$) | Strauss process ($d = 2$)

Conclusion and Other Topics

Summary

	Continuous distributions	Discrete distributions	Point processes
Normalized	Kolmogorov–Smirnov test Cramér–von Mises test Anderson–Darling test	Chi-squared test	(mainly Poisson-type)
Unnormalized	*(Chwialkowski, Strathmann, Gretton. ICML'16)* *(Liu, Lee, Jordan. ICML'16)* $\mathcal{A}_p f(\mathbf{x}) = \frac{\nabla p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) + \nabla f(\mathbf{x})$	*(**Y**, Liu, Rao, Neville. ICML'18)* $\mathcal{A}_p f(\mathbf{x}) := \frac{\Delta p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) - \Delta^* f(\mathbf{x})$	*(**Y**, Rao, Neville. AISTATS'19)* $(\mathcal{A}_\rho h)(\phi) = \int_{\mathbb{X}} [h(\phi \cup \{x\}) - h(\phi)]\, \rho(x\vert\phi)\, \mathrm{d}x + \sum_{x\in\phi} [h(\phi \backslash \{x\}) - h(\phi)]$

Goodness-of-Fit Testing via Kernelized Stein Discrepancy

- Construct a Stein operator (prove Stein identity) (using the unnormalized density).
- Define a positive-definite kernel on the underlying space.
- Establish a kernelized Stein discrepancy measure.
- Computation of the test statistic; bootstrapping procedure.

Open Questions and Future Directions

Immediate Questions

- KSD tests for very high-dimensional distributions?
- Stein operator that fully characterizes a general point processes? $\mathbb{E}\left[\mathcal{A}_\rho h(\Phi)\right] = 0,\ \forall h \ \overset{?}{\nRightarrow}\ \Phi \sim \rho$
- More efficient computation of Stein–Papangelou test statistic.

$$\kappa_\rho(\phi, \psi) = \int_{\mathbb{X}}\int_{\mathbb{X}} \Big[k(\phi \cup \{u\}, \psi \cup \{v\}) - k(\phi, \psi \cup \{v\}) - k(\phi \cup \{u\}, \psi) + k(\phi, \psi$$
$$+ \int_{\mathbb{X}} \Big[\sum_{x\in\phi} \left[k(\phi\backslash\{x\}, \psi \cup \{v\}) - k(\phi\backslash\{x\}, \psi)\right] - |\phi| \cdot \left[k(\phi, \psi \cup \{v\}) -\right.$$
$$+ \int_{\mathbb{X}} \Big[\sum_{y\in\psi} \left[k(\phi \cup \{u\}, \psi\backslash\{y\}) - k(\phi, \psi\backslash\{y\})\right] - |\psi| \cdot \left[k(\phi \cup \{u\}, \psi) -\right.$$
$$+ \Big[\sum_{x\in\phi}\sum_{y\in\psi} k(\phi\backslash\{x\}, \psi\backslash\{y\}) - |\phi| \cdot \sum_{y\in\psi} k(\phi, \psi\backslash\{y\}) - |\psi| \cdot \sum_{x\in\phi} k(\phi\backslash\{$$

Future Directions

- Composite hypothesis testing / latent variable models: $H_0 : q \in \mathcal{P}_\theta$ vs. $H_1 : q \notin \mathcal{P}_\theta$
- Stein discrepancy *beyond* KSD *(cf. Gorham & Mackey '15; Jitkrittum et al. '17; Huggins & Mackey '18)*
- Stein's method for approximate inference *(cf. Liu & Wang '16; Liu & Lee ' 17; Han & Liu' 18; Chen et al. '18)*
- Interpretable features for model criticism *(cf. Jitkrittum et al. '18)*
- Sketching for kernel hypothesis testing *(cf. Zhao & Meng '14; Huggins & Mackey '18))*

Thesis Organization

Chapter 2

Models for Networks and Point Processes

2.1 Statistical Network Models

2.2 Point Processes

- 2.2.1 Temporal Point Processes
- 2.2.2 General Point Processes

Chapter 3

Nonparametric Hypothesis Testing

3.1 Reproducing Kernel Hilbert Spaces

3.2 Maximum Mean Discrepancy and Two-Sample Tests

3.3 Stein Discrepancy and Goodness-of-Fit Tests

Chapter 4 *(UAI'17)*

Decoupling Homophily and Reciprocity with Latent Space Network Models

Chapter 5 *(ICML'18)*

Goodness-of-Fit Testing for Discrete Distributions via Stein Discrepancy

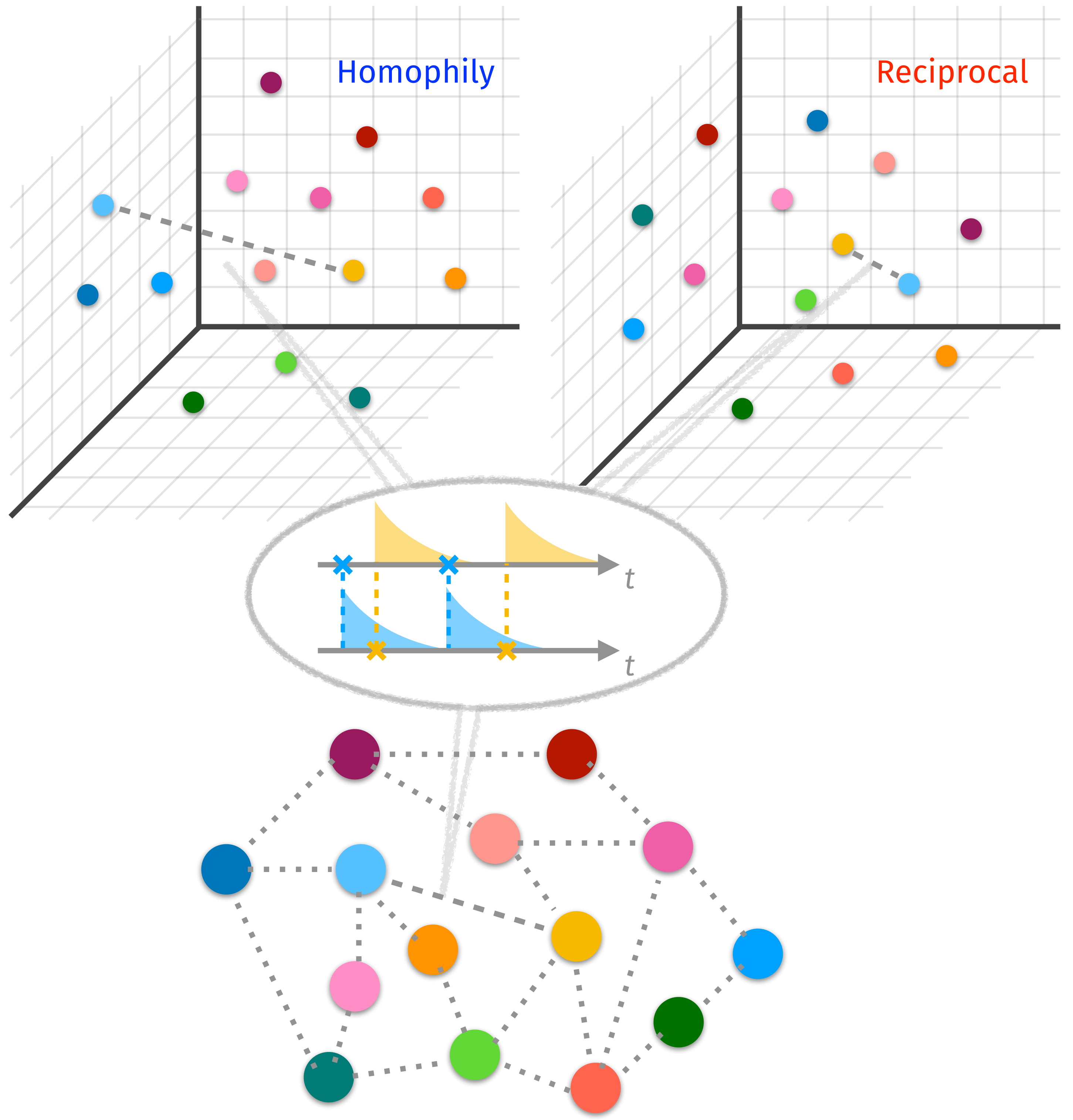
Chapter 6 *(AISTATS'19)*

A Stein–Papangelou Goodness-of-Fit Test for Point Processes

Decoupling Homophily and Reciprocity with Latent Space Network Models

(Y, Rao, Neville. UAI'17)

From	To	Date	Time
26e0c6c4d3b36b2594739fa30eb564b4	881938e1e49fe9d7ee0da580ae4e5946	7/7/11	12:00:02 AM
6d3d380010755cfbeded2e4ac008e8a5	6d3d380010755cfbeded2e4ac008e8a5	7/7/11	12:00:03 AM
eabf2d3d8b046cca7360e2caf1c03362	6cdca3d8290d6c9478a8c9cf6f702a0a	7/7/11	12:00:04 AM
bfd64e0e849d4ca70b719b25a34fa89a	bfd64e0e849d4ca70b719b25a34fa89a	7/7/11	12:00:09 AM
c0784dd06a9a556870d6cdf320b2042f	c0784dd06a9a556870d6cdf320b2042f	7/7/11	12:00:15 AM
589cbf69870575256743976503236bc0	43d451c9016fe24183ea83177e051cb5	7/7/11	12:00:17 AM
b311ecf15238219200d45940208b27c9	7362199de1973c3ce6fd0a8182879642	7/7/11	12:00:17 AM
8bff3584b3038174ed5064465d97578b	79de1530de5760996642b502dc41ab10	7/7/11	12:00:24 AM
6d3d380010755cfbeded2e4ac008e8a5	6d3d380010755cfbeded2e4ac008e8a5	7/7/11	12:00:46 AM
a49df0f8fd9f04ddffc35ddd75910b57	c7e640211096aa609b5c5568ded4a8c4	7/7/11	12:00:46 AM
697ac859008ce24a12343422b4188bea	697ac859008ce24a12343422b4188bea	7/7/11	12:00:49 AM
be986fcb0a18a639f2a730a2485f580d	be986fcb0a18a639f2a730a2485f580d	7/7/11	12:00:53 AM
10c1e0c9ed1a17411e784ad554db3b55	994cba4c9a4a9fbb2aedd1ec03270cb1	7/7/11	12:01:09 AM
547d16750b6ace87de48a5af95ce9d11	69c88cdaaadba8fd783e4612cc29f824	7/7/11	12:01:22 AM
6d3d380010755cfbeded2e4ac008e8a5	6d3d380010755cfbeded2e4ac008e8a5	7/7/11	12:01:27 AM
b13ec563a5e89c7a97b04662a0c43d31	b13ec563a5e89c7a97b04662a0c43d31	7/7/11	12:01:28 AM
19e74a1ab165c956d27a863854085485	2afda4ee8fc22fa741037304f43119af	7/7/11	12:02:25 AM
2acc1711a000ef1c512e8c0a00b7c122	924f4817dec2429f36b555bb0a27f059	7/7/11	12:02:26 AM
02e04607877ab6c02a0c7e93402952ca	70bbb51866b05f9b6d5a5f8ba042e2d3	7/7/11	12:02:34 AM
6808e627469fa5bd91de4cacf09d5619	6808e627469fa5bd91de4cacf09d5619	7/7/11	12:03:00 AM
ae5224db998ef857d02ccdea732f3191	66ee18050233a847ca6c81417bf0e642	7/7/11	12:03:05 AM
19e74a1ab165c956d27a863854085485	5a479e27f86d379aa5e5cae81ccaec4c	7/7/11	12:03:07 AM
e9cc27bb3cdd97b7e6c269b118169e83	e9cc27bb3cdd97b7e6c269b118169e83	7/7/11	12:03:10 AM
6a1827a939cf86b0271db68aa4f1cb43	6a1827a939cf86b0271db68aa4f1cb43	7/7/11	12:03:25 AM
0411c4c432071a78e5e081c859763408	082e77e6ed67753109ded73e2651e635	7/7/11	12:03:36 AM
feb6e82c08cd7fec916bab95e2fc4111	d1fc082e471464120e1e85ddb1da8cd1	7/7/11	12:03:46 AM
6cdca3d8290d6c9478a8c9cf6f702a0a	6cdca3d8290d6c9478a8c9cf6f702a0a	7/7/11	12:03:48 AM
332b355737cb2b7790400f2dcd4e8496	1825199e4a20c757b37df5ee0c4877e0	7/7/11	12:04:18 AM
16d35ae8ba06c11ee315a301599dcc10	96601467b1ce7c1498a36a28353ef056	7/7/11	12:04:55 AM



Hawkes Dual Latent Space (DLS) Model

$$z_v \sim \mathcal{N}(0, \sigma^2 \mathrm{I}_{d\times d}) \qquad \forall v \in V$$

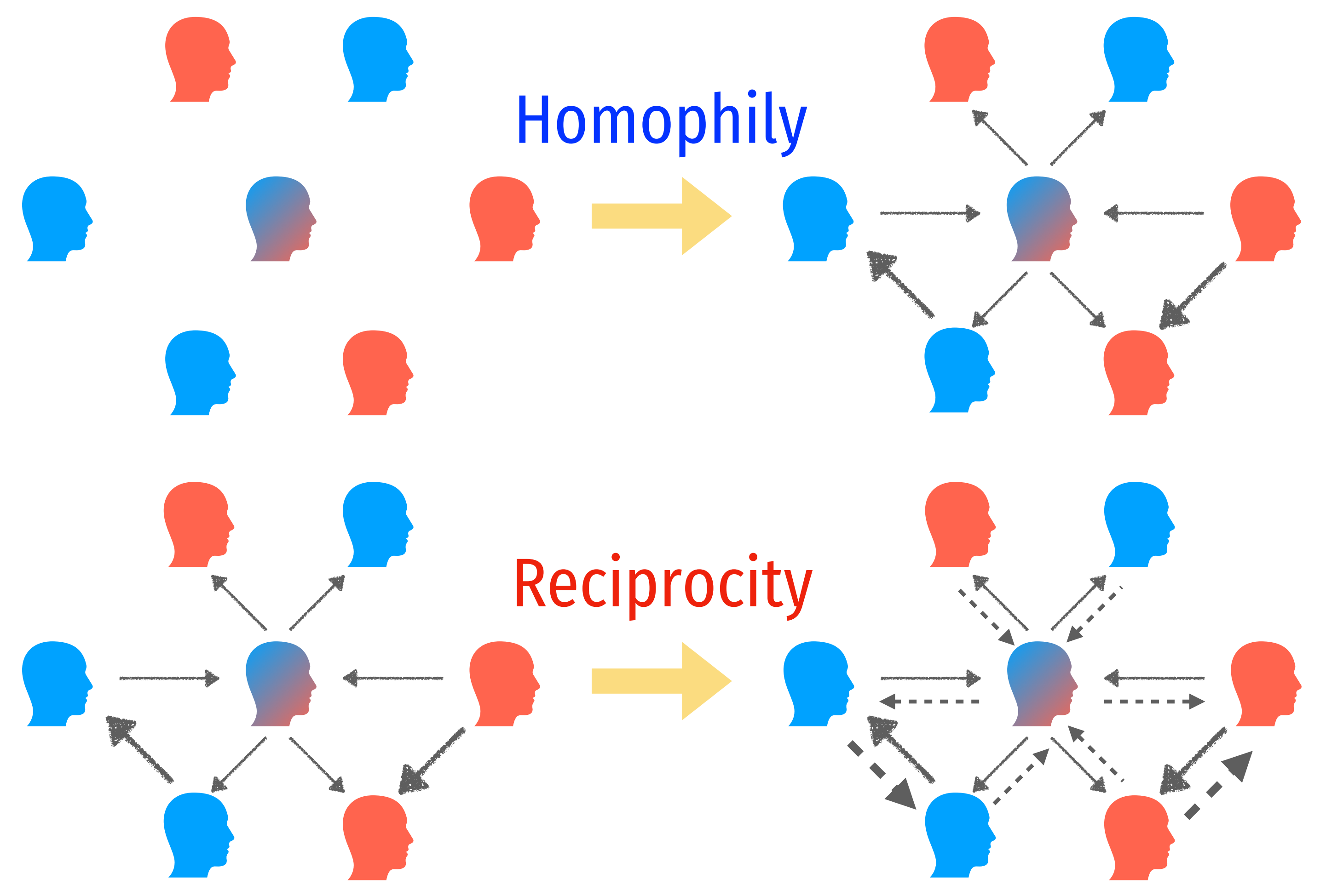
$$\mu_v \sim \mathcal{N}(0, \sigma_\mu^2 \mathrm{I}_{d\times d}) \qquad \forall v \in V$$

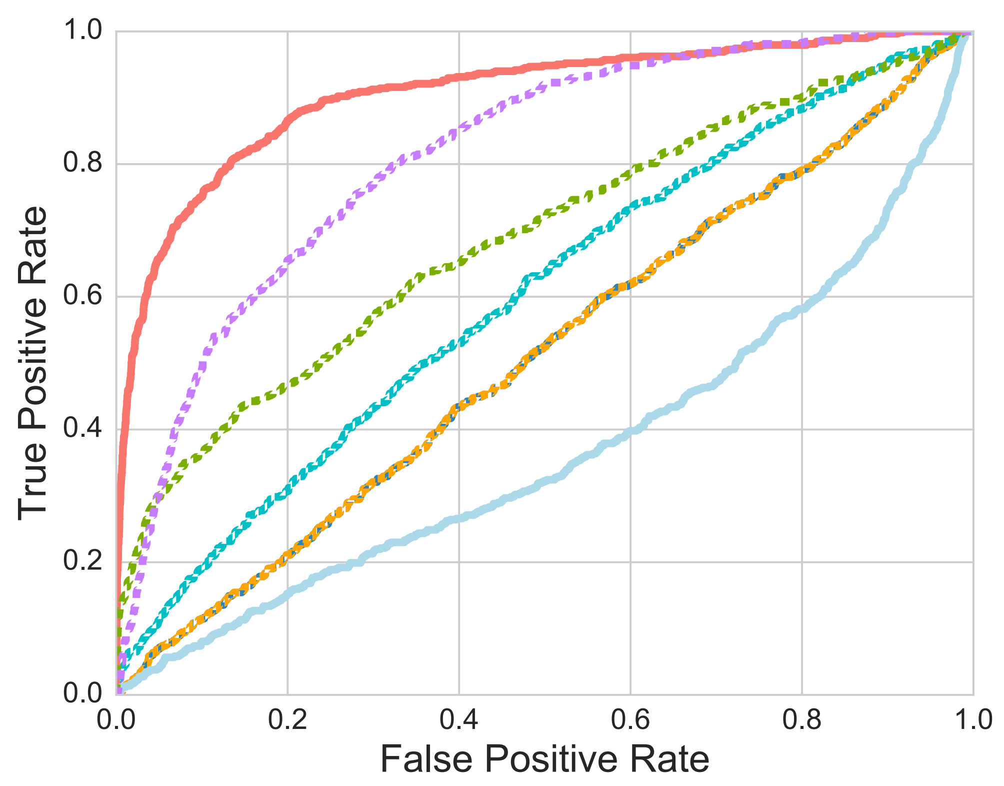
$$\varepsilon_v^{(b)} \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathrm{I}_{d\times d}) \qquad \forall v \in V,\ b = 1, \dots, B$$

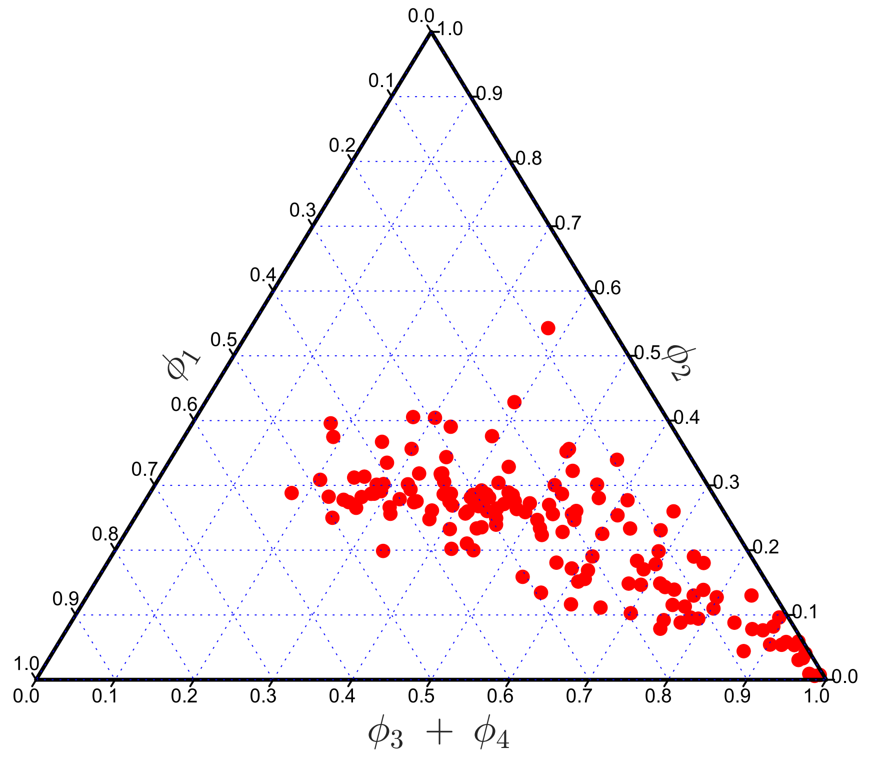
$$x_v^{(b)} \sim \mu_v + \varepsilon_v^{(b)} \qquad \forall v \in V,\ b = 1, \dots, B$$

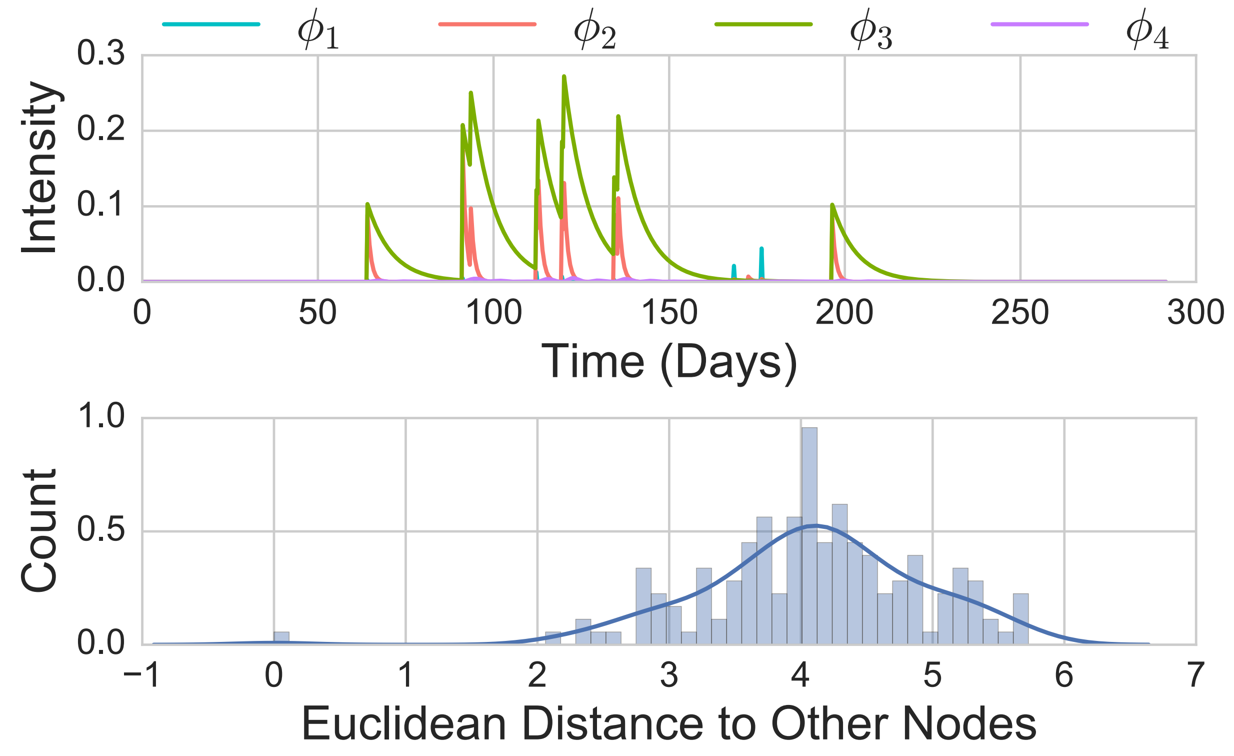
$$\lambda_{uv}(t) = \underbrace{\gamma\, e^{-\|z_u - z_v\|_2^2}}_{\text{Homophily base-rate}} + \underbrace{\sum_{k:\, t_k^{vu} < t} \sum_{b=1}^{B} \beta\, e^{-\|x_u^{(b)} - x_v^{(b)}\|_2^2} \phi_b(t - t_k^{vu})}_{\text{Reciprocal component}}$$

$$N_{uv}(\cdot) \sim \mathrm{HawkesProcess}(\lambda_{uv}(\cdot)) \qquad \forall u \neq v$$



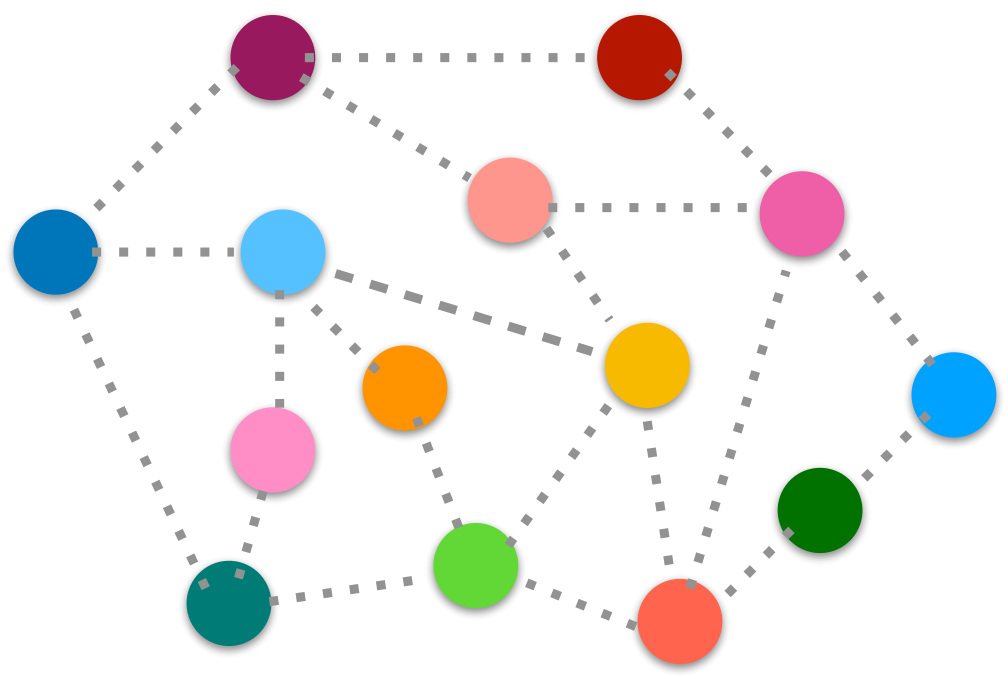




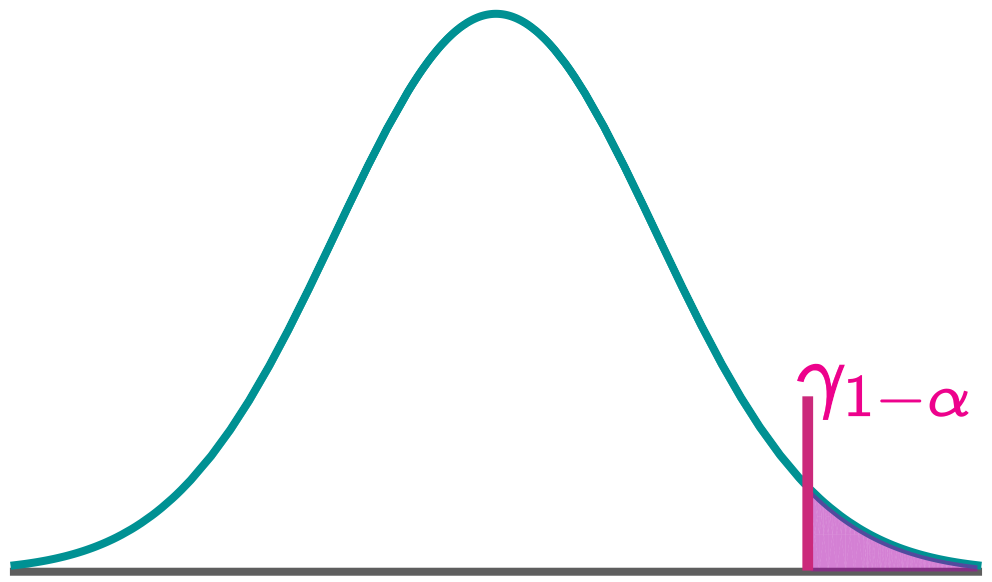


Publications

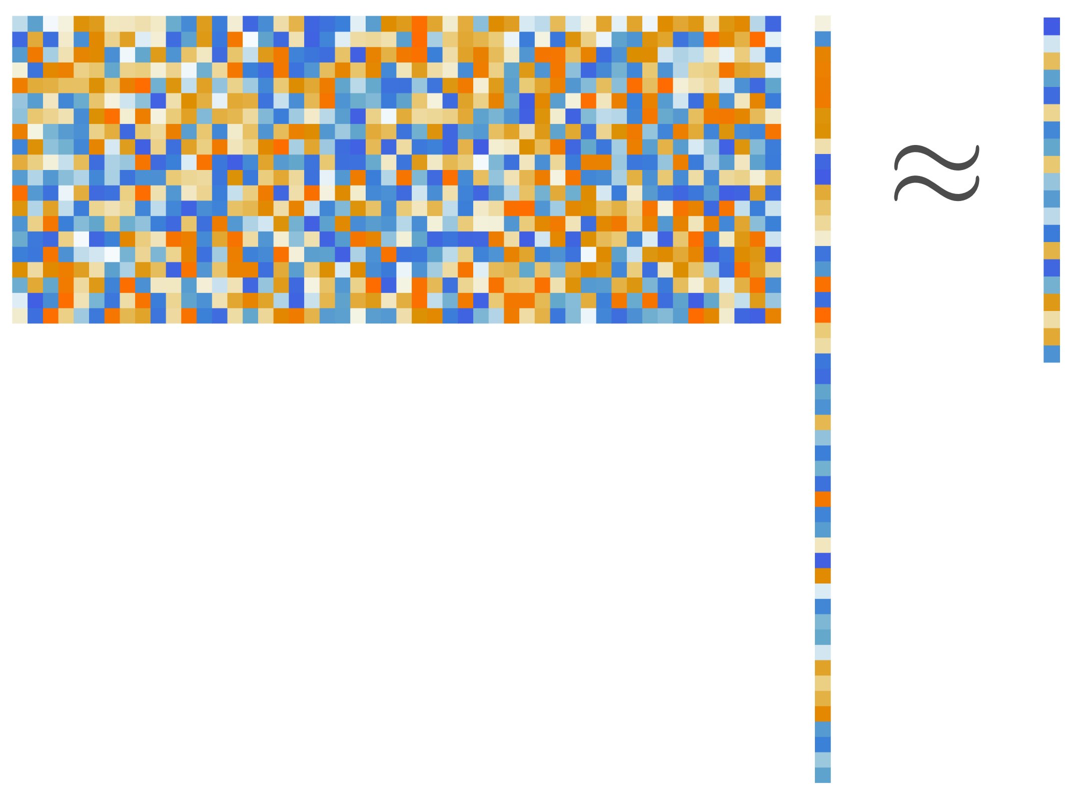
- Learning with Networks and Point Processes
  - *Y*, Rao, and Neville. Decoupling homophily and reciprocity with latent space network models. *UAI*, 2017.
  - *Y*, Ribeiro, and Neville. Stochastic gradient descent for relational logistic regression via partial network crawls. *StarAI, 2017*.
  - *Y*, Ribeiro, and Neville. Should we be confident in peer effects estimated from partial crawls of social networks? *ICWSM, 2017*.



- Statistical Model Criticism for Intractable Distributions
  - *Y*, Rao, and Neville. A Stein–Papangelou goodness-of-fit test for point processes. *AISTATS*, 2019.
  - *Y*, Liu, Rao, and Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. *ICML*, 2018.



- Randomized Sketching Methods for Scalable Computations
  - Chowdhury, *Y*, and Drineas. Randomized iterative algorithms for Fisher discriminant analysis. *Under review*, 2019.
  - Chowdhury, *Y*, and Drineas. Structural conditions for projection-cost preservation via randomized matrix multiplication. *LAA*, 2019.
  - Chowdhury, *Y*, and Drineas. An iterative, sketching-based framework for ridge regression. *ICML*, 2018.



Acknowledgements

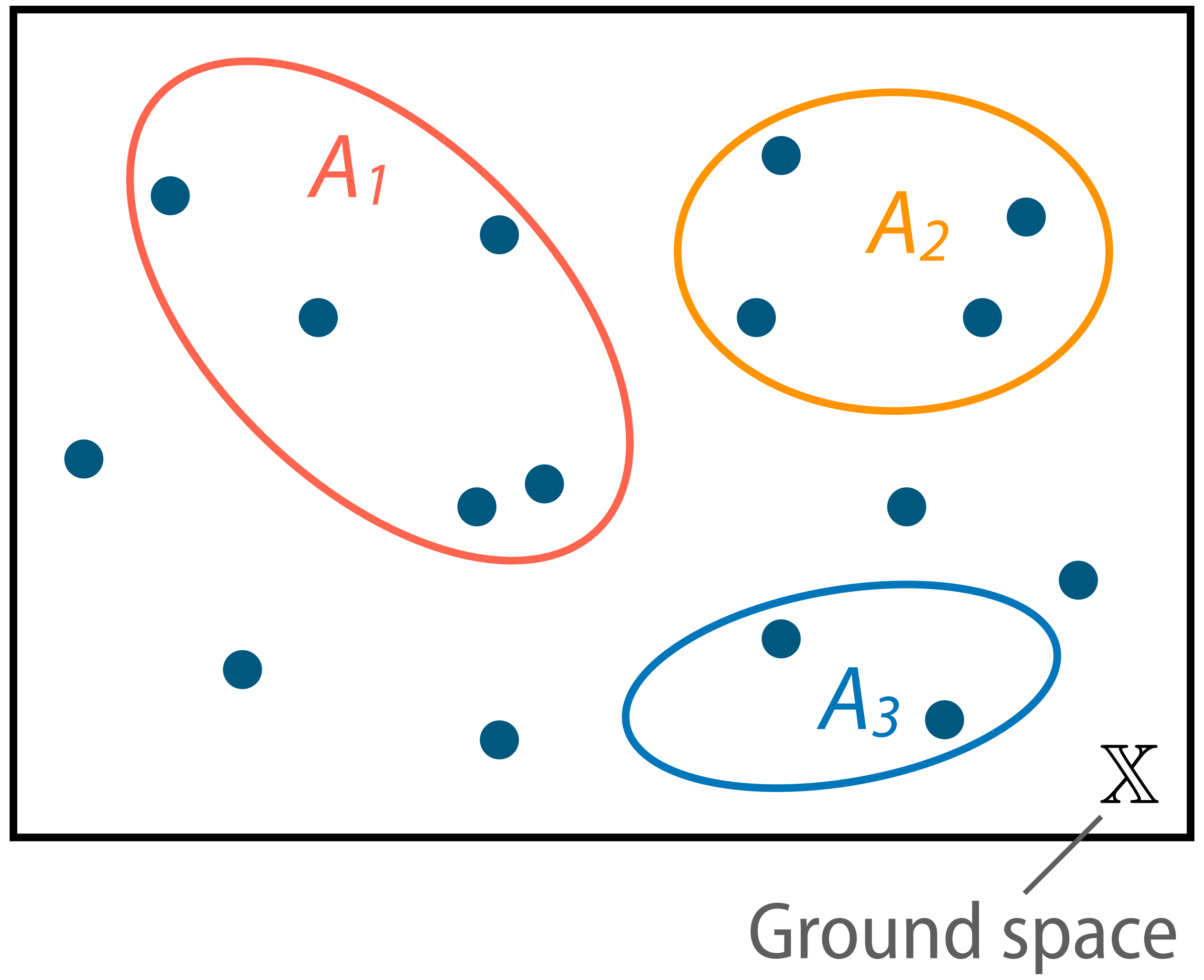




Thank You!

jiaseny@purdue.edu
www.stat.purdue.edu/~yang768

Point Processes



Point process

Φ: random counting measure

Mean measure $\mu(A) := \mathbb{E}[\Phi(A)] = \int_A \lambda(x)\,\mathrm{d}x$

Intensity function

Poisson process

- $A_1, \ldots, A_k$ *disjoint* $\Rightarrow$ $\Phi(A_1), \Phi(A_2), \ldots, \Phi(A_k)$ *independent*
- $\Phi(A) \sim \mathrm{Poi}(\mu(A))$

Complete randomness

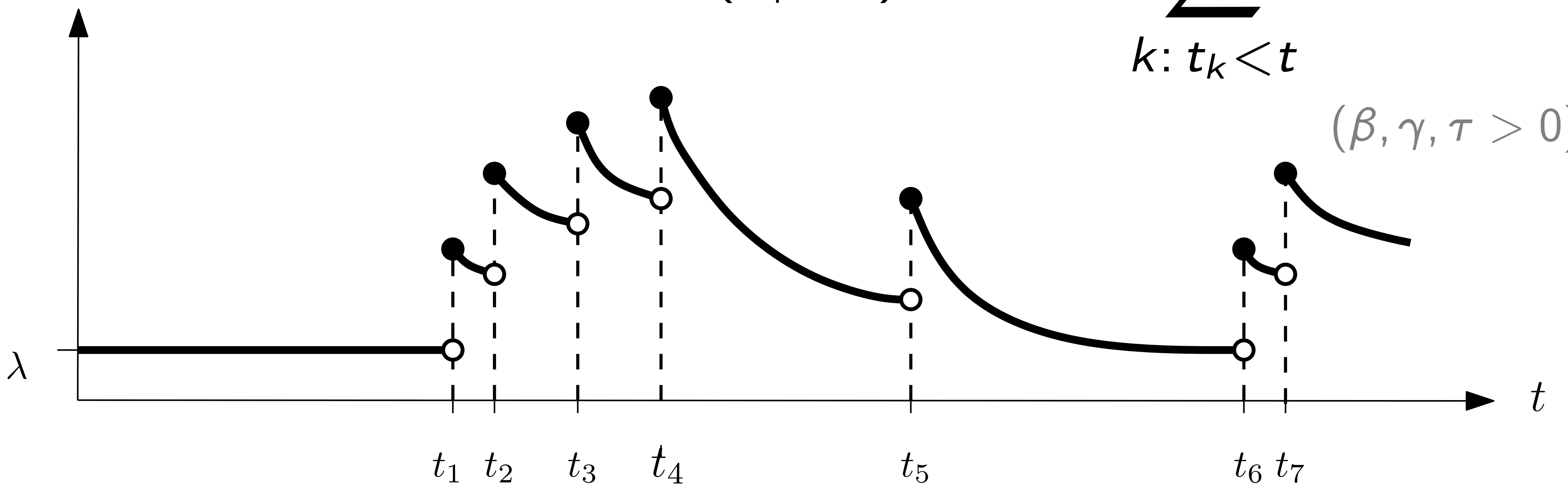
Hawkes process

Excitation

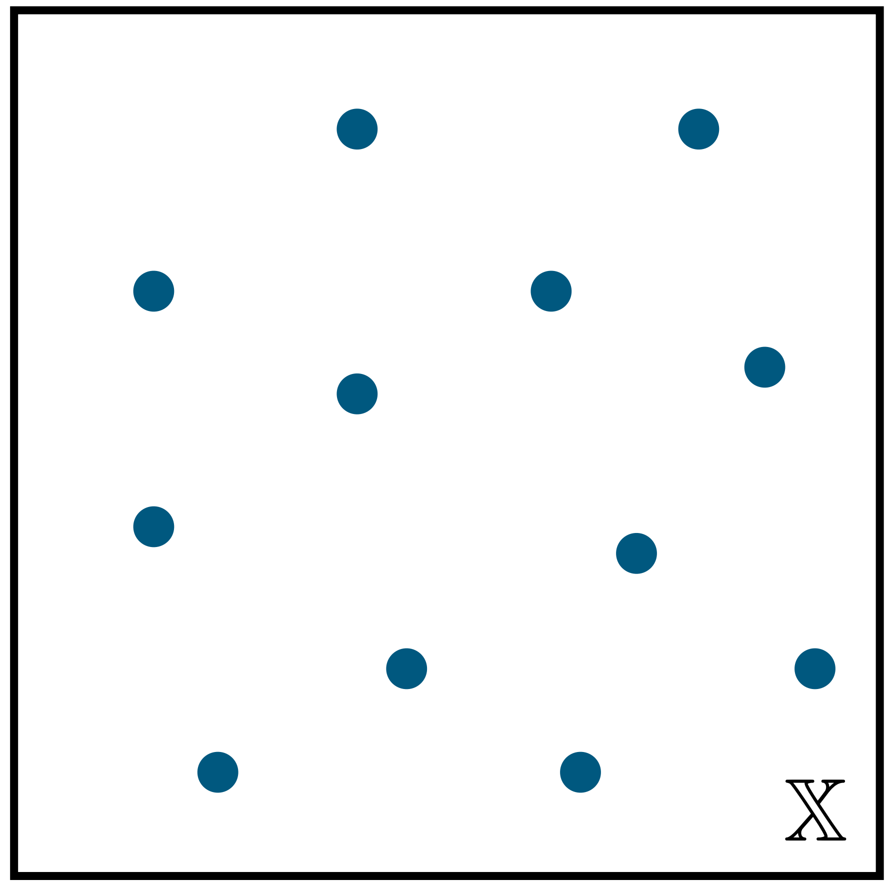
History up to time t

$$\lambda(t|\mathcal{H}_t) = \gamma + \sum_{k:\, t_k < t} \beta e^{-\frac{t-t_k}{\tau}}$$

$(\beta, \gamma, \tau > 0)$



Strauss process



Repulsion

$$s_r(\phi) = \sum_{\mathbf{x},\mathbf{y} \in \phi} \mathbb{I}\{\|\mathbf{x} - \mathbf{y}\|_2 < r\}$$

Density

$$f(\phi) = \frac{1}{Z} \beta^{|\phi|} \gamma^{s_r(\phi)}$$

$(0 < \gamma \leq 1;\ \beta, r > 0)$

Observed point pattern

Asymptotic Null Distribution of KSD Test Statistic

Theorem 5.4.1 (Adapted from Theorem 4.1 of Liu et al. (2016)). *Let* $k(\boldsymbol{x}, \boldsymbol{x}')$ *be a strictly positive definite kernel on* $\mathcal{X}^d$*, and assume that* $\mathbb{E}_{\boldsymbol{x},\boldsymbol{x}'\sim q}\left[\kappa_p(\boldsymbol{x},\boldsymbol{x}')^2\right] < \infty$. *We have the following two cases:*

(i) If $q \neq p$*, then* $\widehat{\mathbb{S}}(q \,\|\, p)$ *is asymptotically normal:*

$$\sqrt{n}\left(\widehat{\mathbb{S}}(q \,\|\, p) - \mathbb{S}(q \,\|\, p)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2),$$

where $\sigma^2 = \mathrm{Var}_{\boldsymbol{x}\sim q}(\mathbb{E}_{\boldsymbol{x}'\sim q}\left[\kappa_p(\boldsymbol{x},\boldsymbol{x}')\right]) > 0.$

(ii) If $q = p$*, then* $\sigma^2 = 0$*, and the* U*-statistic is degenerate:*

$$n\widehat{\mathbb{S}}(q \,\|\, p) \xrightarrow{\mathcal{D}} \sum_j c_j (Z_j^2 - 1),$$

where $\{Z_j\} \overset{iid}{\sim} \mathcal{N}(0,1)$ *and* $\{c_j\}$ *are the eigenvalues of the kernel* $\kappa_p(\cdot,\cdot)$ *under* q.