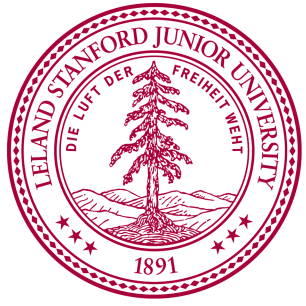
Sailing Through Data: Discoveries and Mirages

Emmanuel Candès, *Stanford University*



Baidu Research, Sunnyvale, May 2019

The Economist

OCTOBER 19TH–25TH 2013 Economist.com

Washington's lawyer surplus

How to do a nuclear deal with Iran

Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG.

99
Einsteinium

The replicability crisis

Begley and Ellis, *Nature* (2012)

Amgen could only replicate 6 of 53 studies they considered landmarks in basic cancer science

HealthCare could only replicate about 25% of 67 seminal studies

Systematic attempts to replicate widely cited priming experiments have failed

Media coverage...

SECTIONS | The New York Times | SUBSCRIBE NOW | LOG IN

A College Town Gets Ready for Its Moment Under No Sun

TRILOBITES
This Beautiful Parasitic Bird Could Soon Turn Up in Your Yard

TRILOBITES
Drumming Cockatoos and the Rhythms of Love

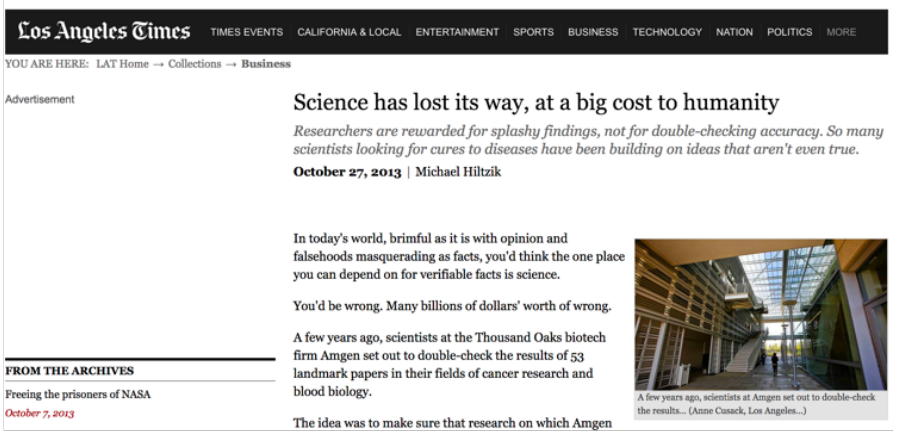
SCIENCE

New Truths That Only One Can See

George Johnson
RAW DATA JAN. 20, 2014

Carl Wiens

Since 1955, The Journal of Irreproducible Results has offered "spoofs, parodies, whimsies, burlesques, lampoons and satires" about life in the laboratory. Among its greatest hits: "Acoustic Oscillations in Jell-O, With and Without Fruit, Subjected to Varying Levels of Stress" and "Utilizing Infinite Loops to Compute an Approximate Value of Infinity." The good-natured jibes are a backhanded celebration of science. What really goes on in the lab is, by implication, of a loftier, more serious nature.



Los Angeles Times | TIMES EVENTS | CALIFORNIA & LOCAL | ENTERTAINMENT | SPORTS | BUSINESS | TECHNOLOGY | NATION | POLITICS | MORE

YOU ARE HERE: LAT Home → Collections → Business

Advertisement

Science has lost its way, at a big cost to humanity

Researchers are rewarded for splashy findings, not for double-checking accuracy. So many scientists looking for cures to diseases have been building on ideas that aren't even true.

October 27, 2013 | Michael Hiltzik

In today's world, brimful as it is with opinion and falsehoods masquerading as facts, you'd think the one place you can depend on for verifiable facts is science.

You'd be wrong. Many billions of dollars' worth of wrong.

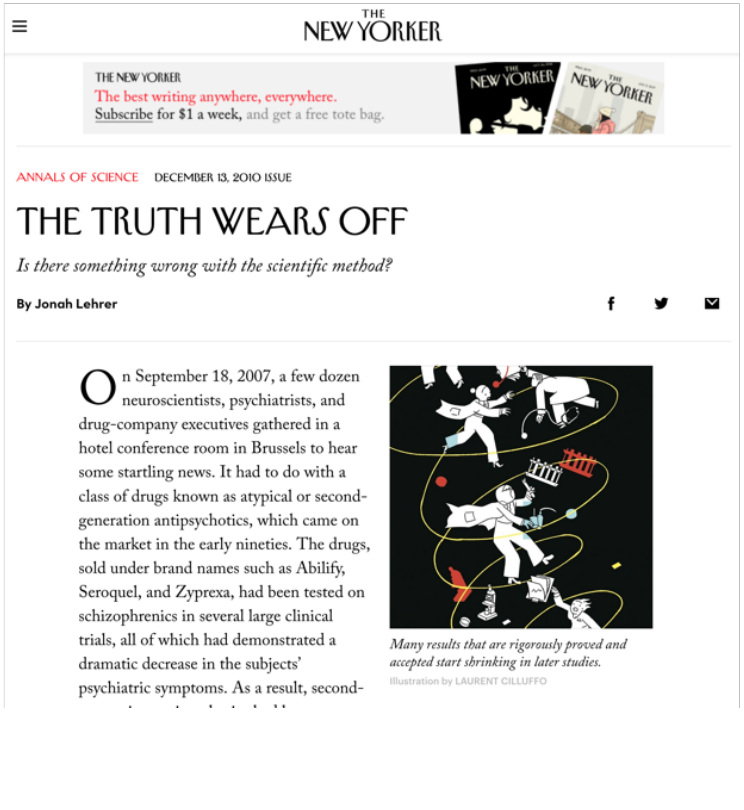
A few years ago, scientists at the Thousand Oaks biotech firm Amgen set out to double-check the results of 53 landmark papers in their fields of cancer research and blood biology.

The idea was to make sure that research on which Amgen

A few years ago, scientists at Amgen set out to double-check the results... (Anne Cusack, Los Angeles...)

FROM THE ARCHIVES

Freeing the prisoners of NASA
October 7, 2013



THE NEW YORKER

THE NEW YORKER
The best writing anywhere, everywhere.
Subscribe for $1 a week, and get a free tote bag.

ANNALS OF SCIENCE DECEMBER 13, 2010 ISSUE

THE TRUTH WEARS OFF

Is there something wrong with the scientific method?

By Jonah Lehrer

On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-

Many results that are rigorously proved and accepted start shrinking in later studies.

Illustration by LAURENT CILLUFFO

The New York Times

Essay: The Experiments Are Fascinating. But Nobody Can Repeat Them.

Science is mired in a "replication" crisis. Fixing it will not be easy.

New York Times, Science, November 19, 2018

The replicability problem

Early report (Kaplan, '08)

50% of Phase III FDA studies ended in failure

Personal and societal concern

Snippets from media

"Significance chasing"

"Publication bias"

"Selective reporting"

Personal and societal concern

Snippets from media

"Significance chasing"

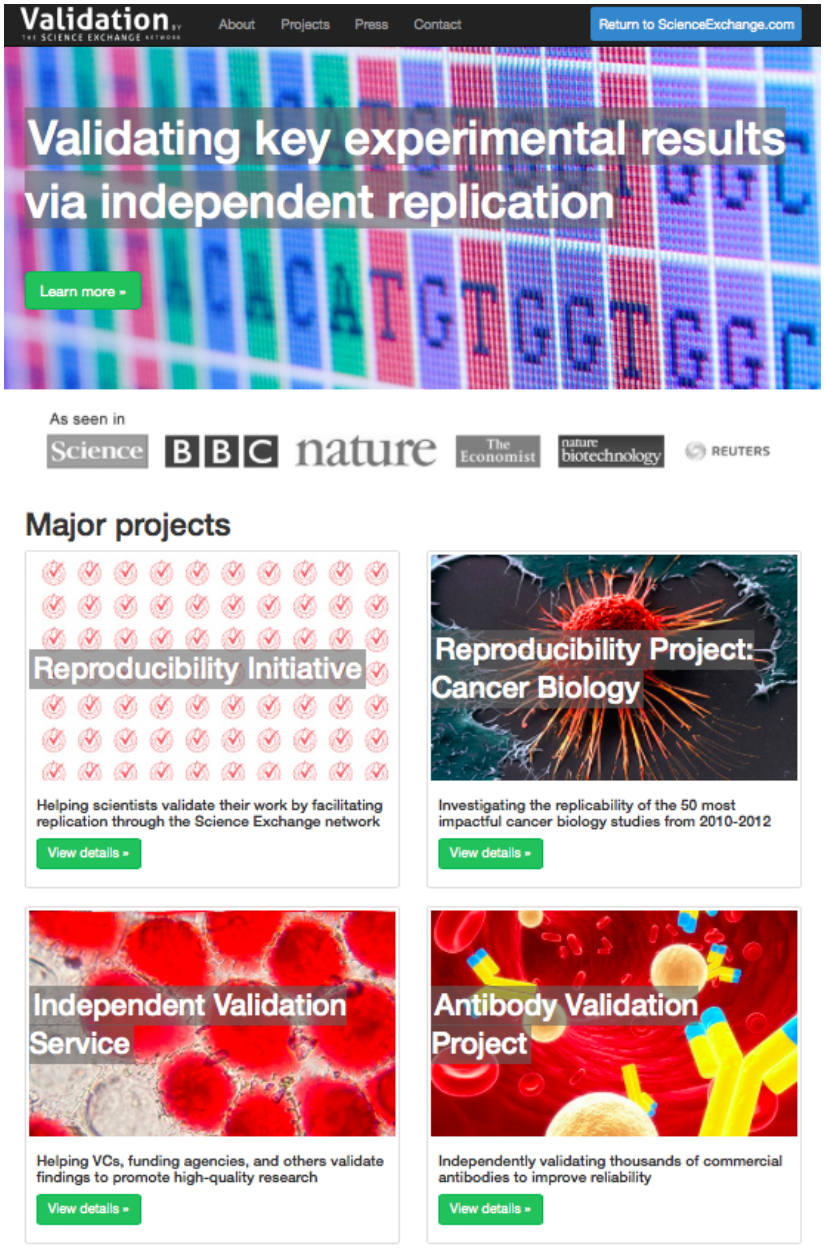
"Publication bias"

"Selective reporting"

Great danger in seeing erosion of public confidence in science

Scientific community is responding

Response: reproducibility intiatives



Reproducibility Initiative

`http://validation.scienceexchange.com/`

Response: editorial policies

ANNOUNCEMENT

Reducing our irreproducibility

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

From next month, *Nature* and the Nature research journals will introduce editorial measures to address the problem by improving the consistency and quality of reporting in life-sciences articles. To ease the interpretation and improve the reliability of published results we will more systematically ensure that key methodological details are reported, and we will give more space to methods sections. We will examine statistics more closely and encourage authors to be transparent, for example by including their raw data.

Central to this initiative is a checklist intended to prompt authors to disclose technical and statistical information in their submissions, and to encourage referees to consider aspects important for research reproducibility (go.nature.com/oloeip). It was developed after discussions with researchers on the problems that lead to irreproducibility, including workshops organized last year by US National Institutes of Health (NIH) institutes. It also draws on published concerns about reporting standards (or the lack of them) and the collective experience of editors at Nature journals.

The checklist is not exhaustive. It focuses on a few experimental and analytical design elements that are crucial for the interpretation of research results but are often reported incompletely. For example, authors will need to describe methodological parameters that can introduce bias or influence robustness, and provide precise characterization of key reagents that may be subject to biological variability, such as cell lines and antibodies. The checklist also consolidates existing policies about data deposition and presentation.

We will also demand more precise descriptions of statistics, and we will commission statisticians as consultants on certain papers, at the editor's discretion and at the referees' suggestion.

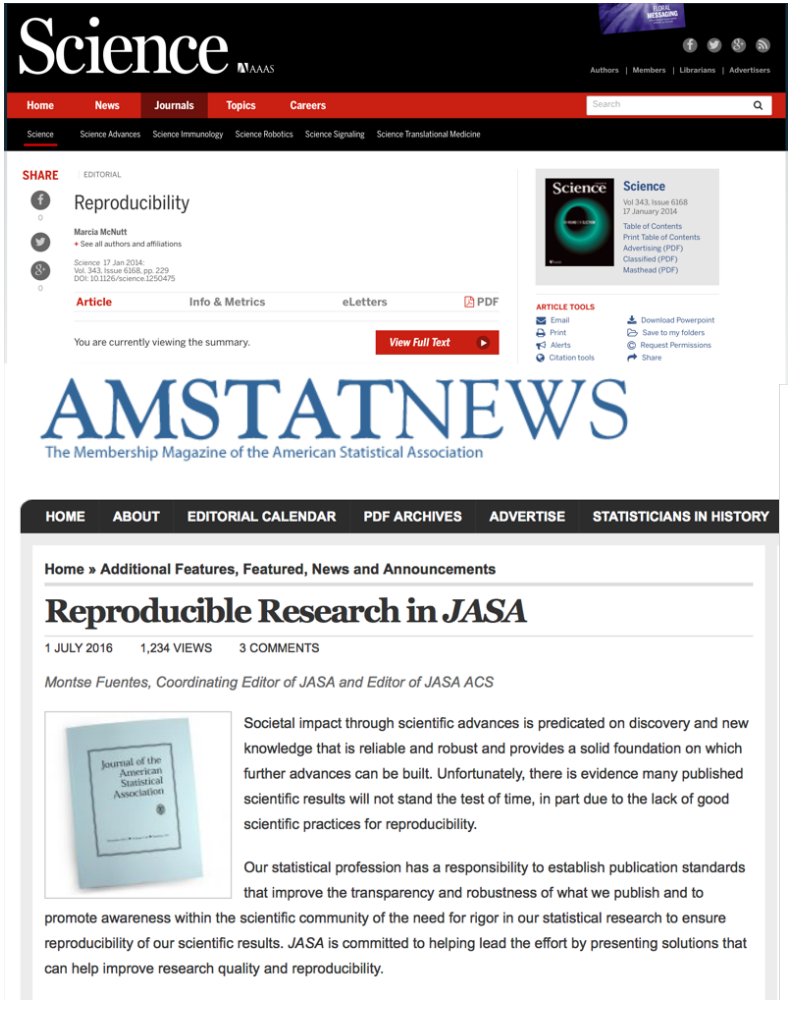
We recognize that there is no single way to conduct an experimental study. Exploratory investigations cannot be done with the same level of statistical rigour as hypothesis-testing studies. Few academic laboratories have the means to perform the level of validation required, for example, to translate a finding from the laboratory to the clinic. However, that should not stand in the way of a full report of how a study was designed, conducted and analysed that will allow reviewers and readers to adequately interpret and build on the results.

To allow authors to describe their experimental design and methods in as much detail as necessary, the participating journals, including *Nature*, will abolish space restrictions on the methods section.

To further increase transparency, we will encourage authors to provide tables of the data behind graphs and figures. This builds on our established data-deposition policy for specific experiments and large data sets. The source data will be made available directly from the figure legend, for easy access. We continue to encourage authors to share detailed methods and reagent descriptions by depositing protocols in Protocol Exchange (www.nature.com/protocolexchange), an open resource linked from the primary paper.

Renewed attention to reporting and transparency is a small step. Much bigger underlying issues contribute to the problem, and are beyond the reach of journals alone. Too few biologists receive adequate training in statistics and other quantitative aspects of their subject. Mentoring of young scientists on matters of rigour and transparency is inconsistent at best. In academia, the ever increasing pressures to publish and chase funds provide little incentive to pursue studies and publish results that contradict or confirm previous papers. Those who document the validity or irreproducibility of a published piece of work seldom get a welcome from journals and funders, even as money and effort are wasted on false assumptions.

Tackling these issues is a long-term endeavour that will require the commitment of funders, institutions, researchers and publishers. It is encouraging that NIH institutes have led community discussions on this topic and are considering their own recommendations. We urge others to take note of these and of our initiatives, and do whatever they can to improve research reproducibility. ■



Response: best practices

Research Reproducibility, Replicability, Reliability
A Speech by Ralph J. Cicerone, President
National Academy of Sciences
Presented at the Academy's 152nd Annual Meeting
April 27, 2015

President's Council of Advisors on Science and Technology (PCAST)
Public Meeting Agenda
January 31, 2014

National Academy of Sciences (NAS)
2101 Constitution Avenue, NW
Washington, DC

Lecture Room

9:00 am **Welcome from PCAST Co-Chairs**
John Holdren, Assistant to the President for Science and Technology; Director, Office of Science and Technology Policy (OSTP); Co-Chair, PCAST
Eric Lander, Co-Chair, PCAST

9:05 am **Improving Scientific Reproducibility in an Age of International Competition and Big Data I: Researchers**
Glenn Begley, Chief Scientific Officer and Senior Vice-President R&D, TetraLogic Pharmaceuticals
Donald Berry, Professor, Department of Biostatistics, University of Texas MD Anderson Cancer Center
Daniel MacArthur, Assistant Professor, Harvard Medical School and Massachusetts General Hospital and Associate Member, Broad Institute of Harvard and MIT

10:00 am **Improving Scientific Reproducibility in an Age of International Competition and Big Data II: Editors**
Marcia McNutt, Editor-In-Chief, *Science*
Philip Campbell, Editor-In-Chief, *Nature* and Nature Publishing Group
Véronique Kiermer, Executive Editor and Head of Researchers Services, Nature Publishing Group

Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results

A Workshop
of the Committee on Applied and Theoretical Statistics

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

The replicability issue

Many different components

1. Publishing culture
2. Granting agencies culture
3. Computational reproducibility
4. Statistics: how to choose a finding?
   Statistical methodology enhancing replicability

Can only do 3 and 4
1 and 2 above pay grade

Why is this happening? A new scientific paradigm

Collect data first $\Longrightarrow$ Ask questions later

Large data sets available prior to formulation of scientific hypotheses/theories

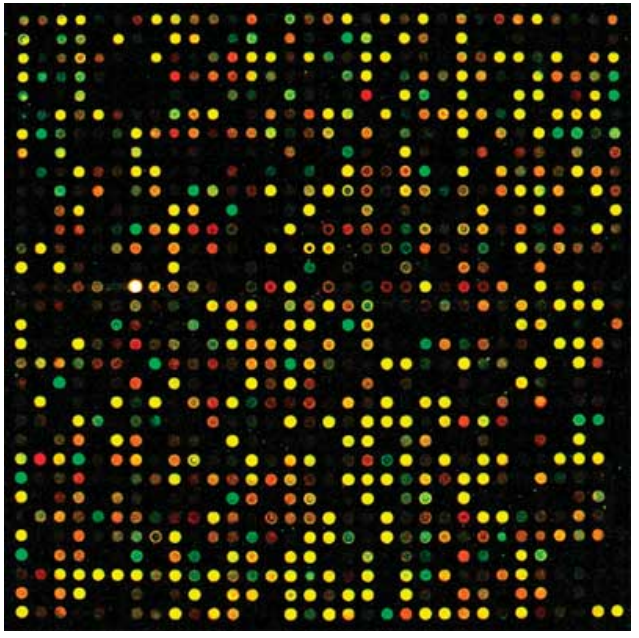
Very different from hypothesis-driven research

Example from genomics

Historically, molecular biology was hypothesis-driven research

"Sometime in the 90's"

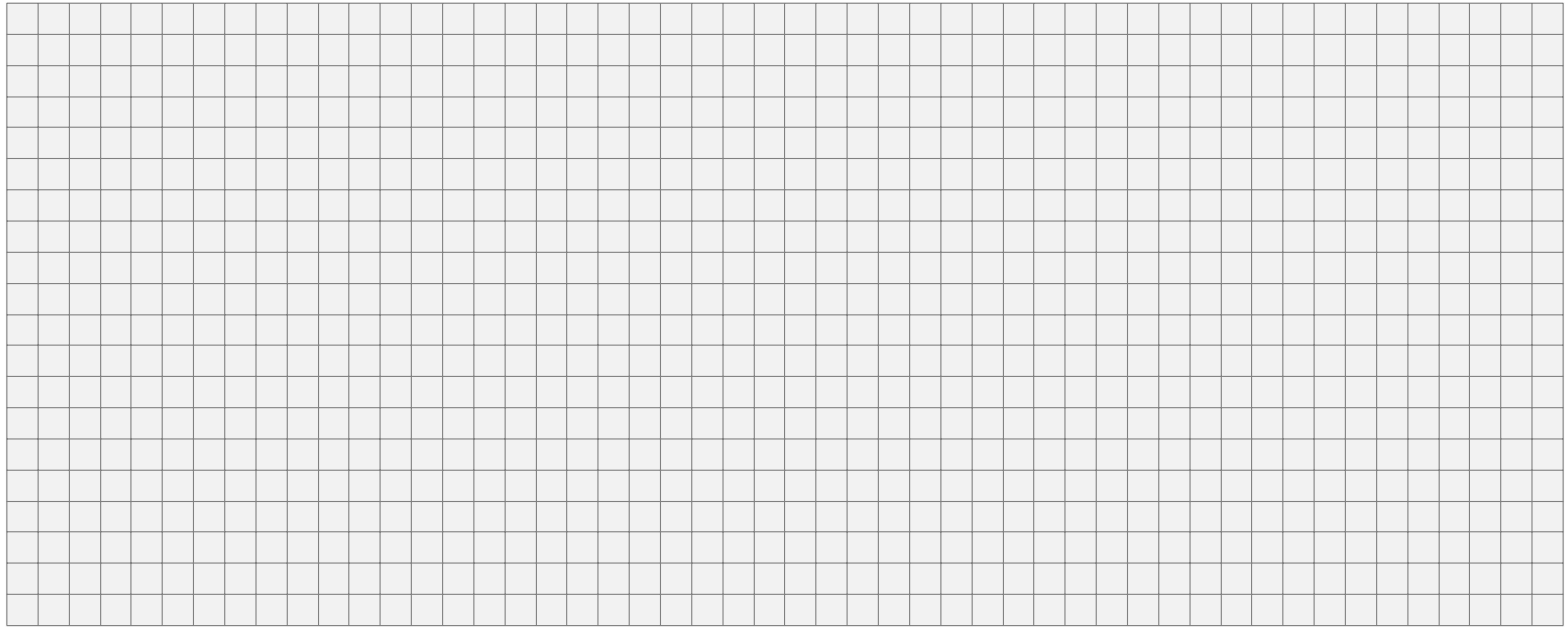
- Eruption of high-throughput technologies
- Enabled thousands of genes to be tested simultaneously for differential expression
  - Small # of samples
  - High # of variables
- Researchers begin to look everywhere



gene expression microarray

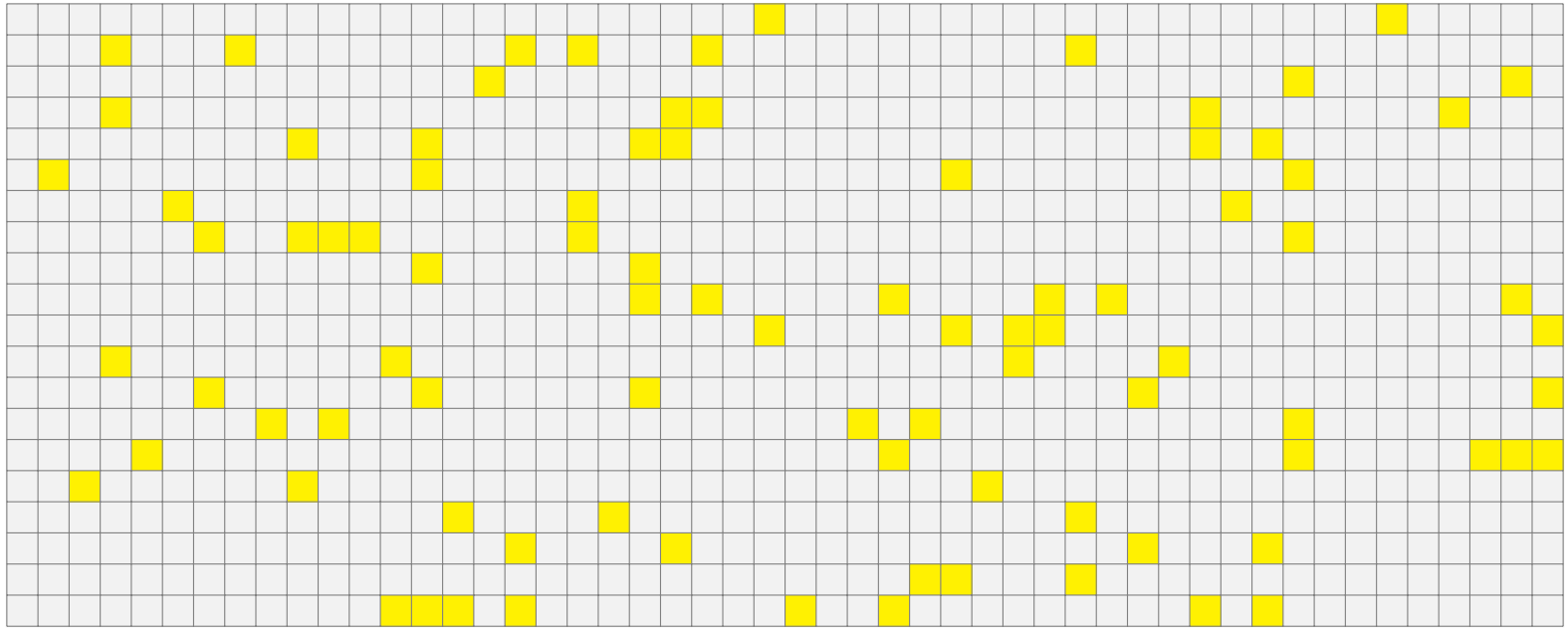
Complete revolution: from hypothesis- to data-driven research

False discovery rate (FDR), Benjamini-Hochberg '95



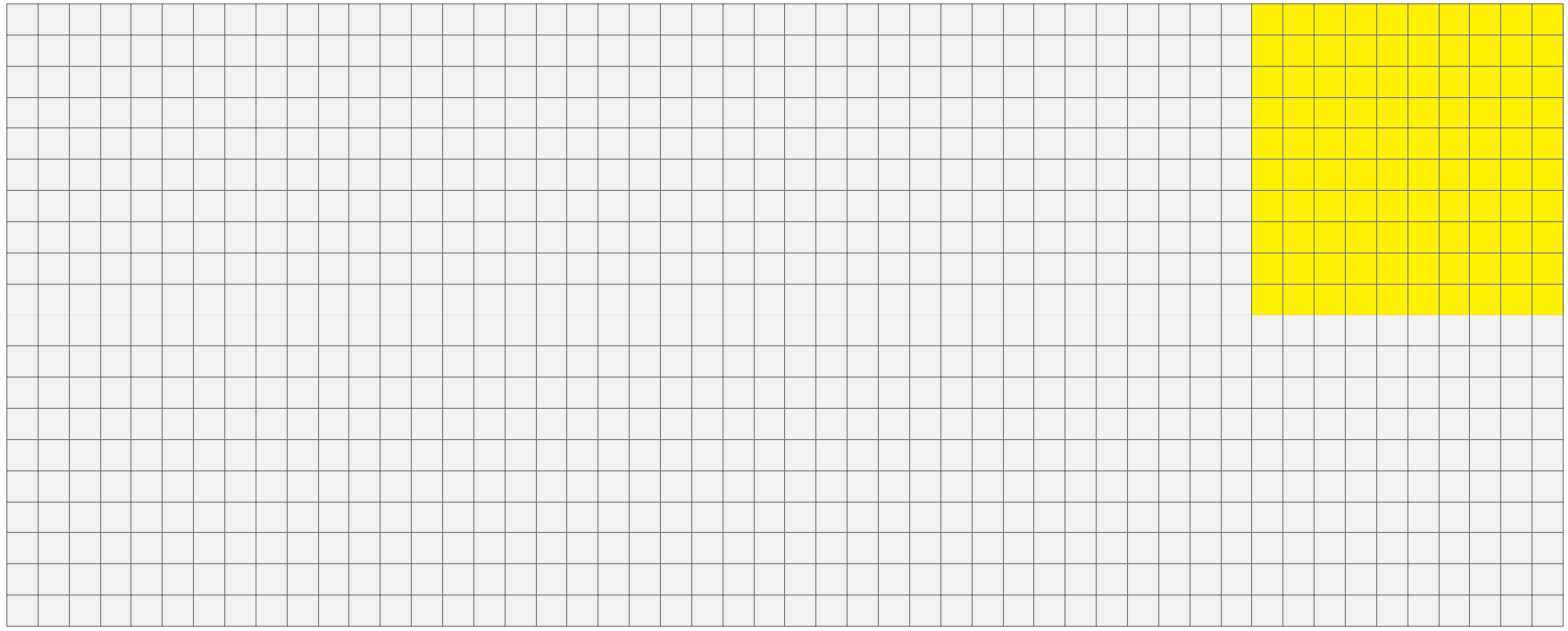
1000 hypotheses to test

False discovery rate (FDR), Benjamini-Hochberg '95



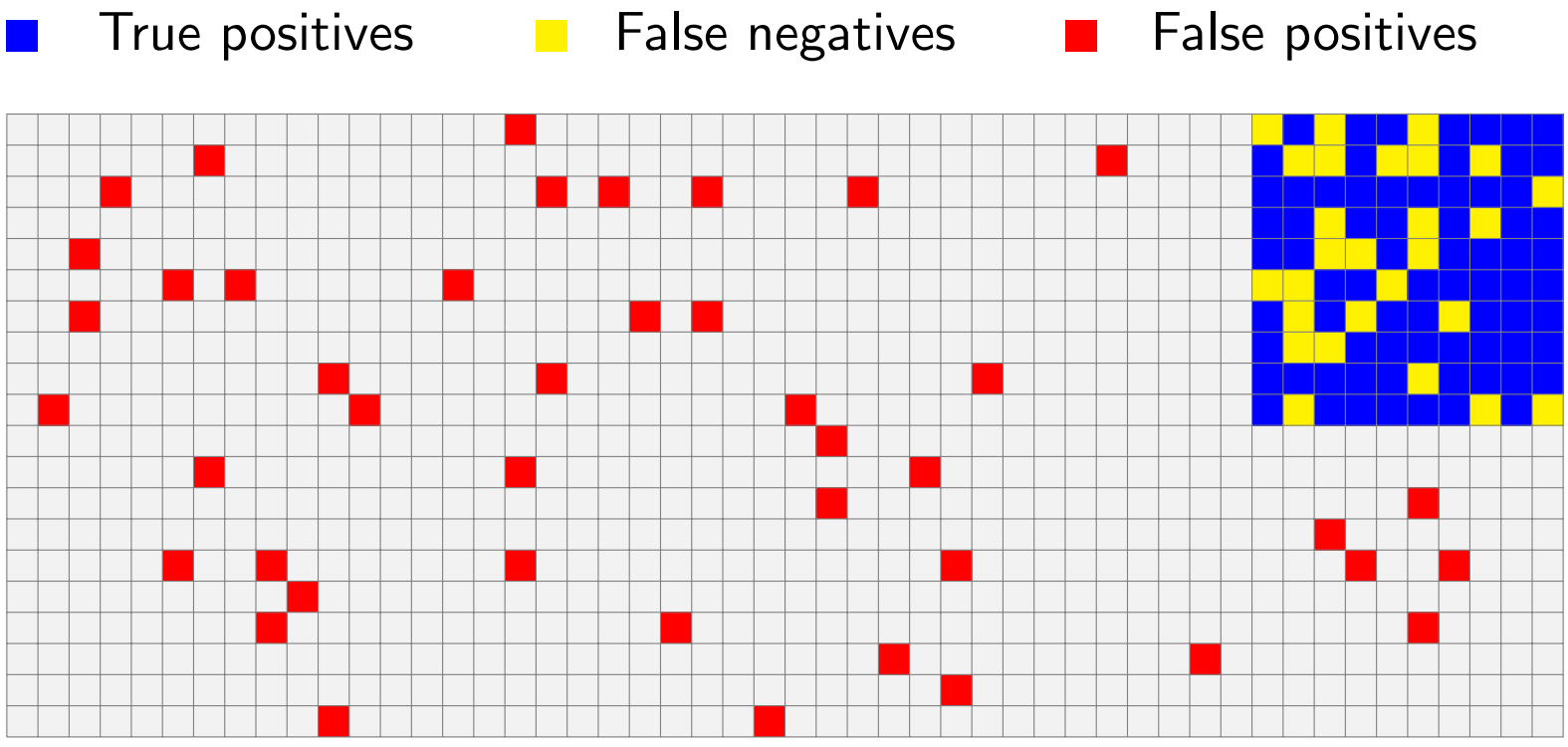
1000 hypotheses, 100 potential discoveries

False discovery rate (FDR), Benjamini-Hochberg '95



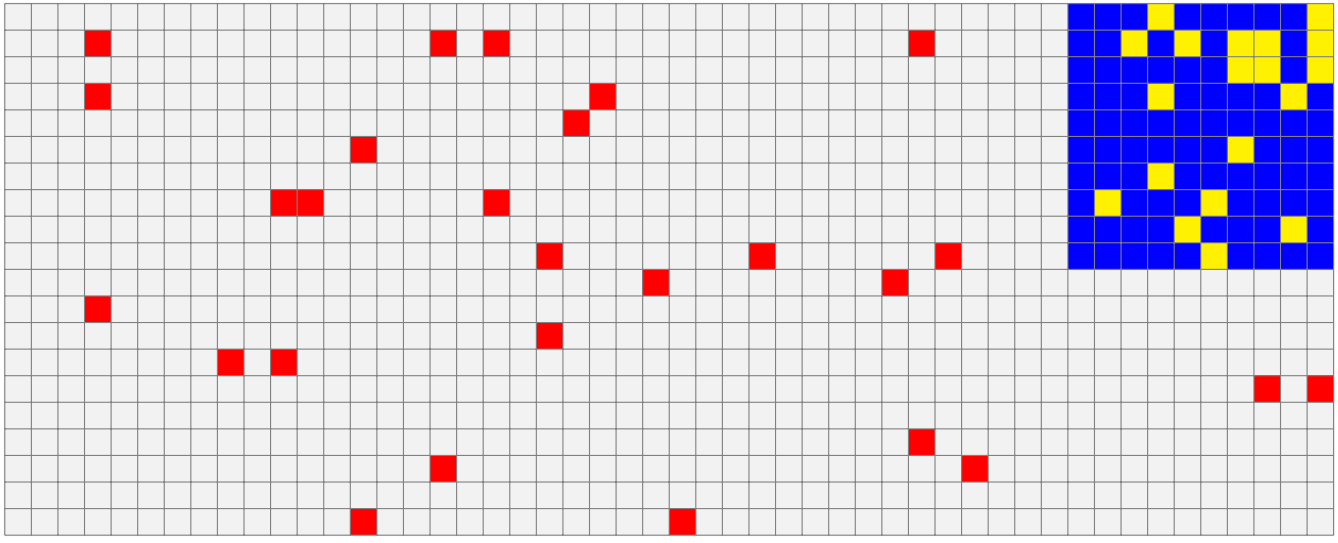
1000 hypotheses, 100 potential discoveries

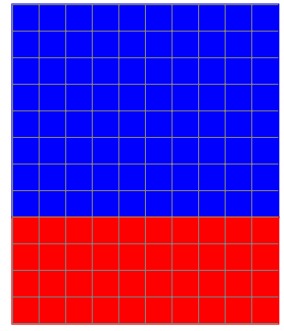
False discovery rate (FDR), Benjamini-Hochberg '95



False discovery rate (FDR), Benjamini-Hochberg '95

True positives False negatives False positives



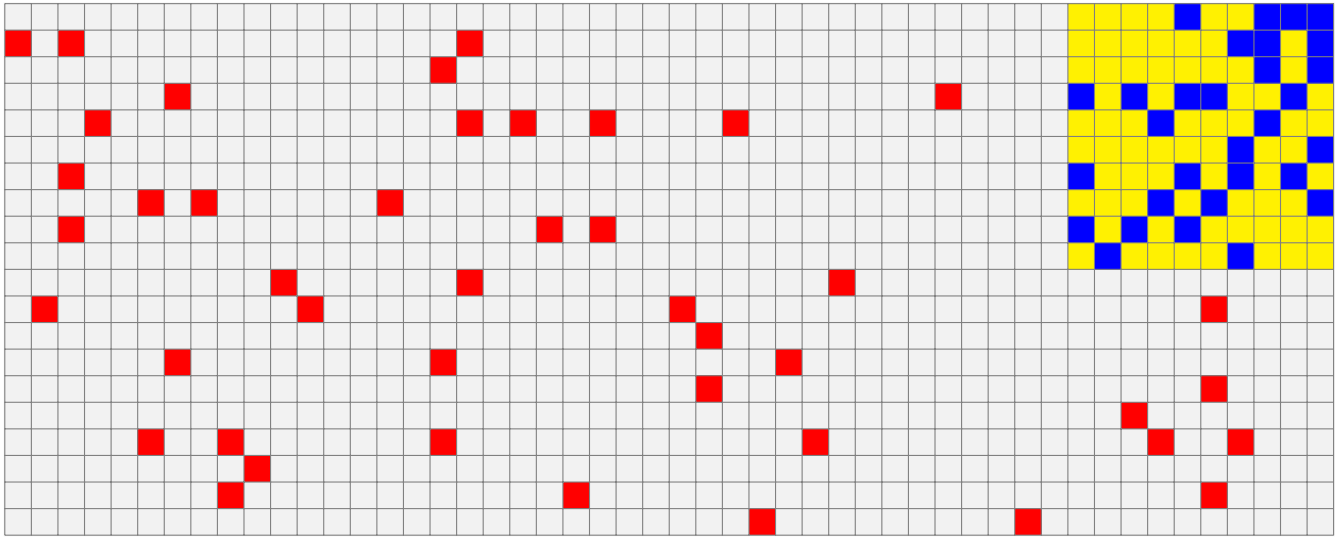


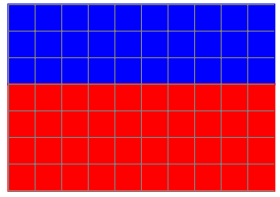
Selected

$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false positives}}{\#\text{selections}}\right]$$

False discovery rate (FDR), Benjamini-Hochberg '95

True positives False negatives False positives





Reported

$$\text{FDR} = \mathbb{E}\left[\frac{\#\text{false positives}}{\#\text{selections}}\right]$$

Knockoffs: Tools for Replicable Selections

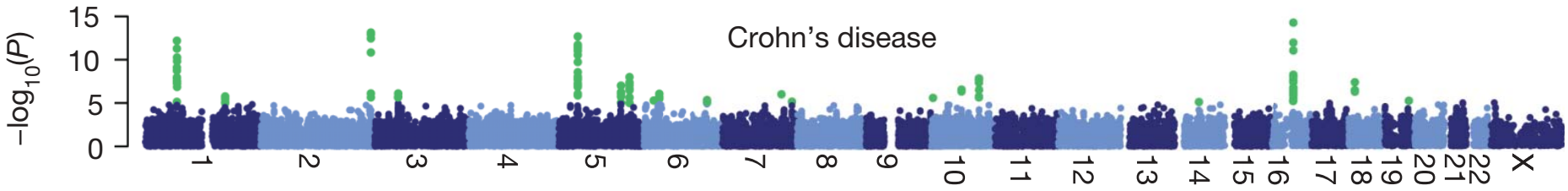
Joint with R. Barber
and Y. Fan, L. Janson and J. Lv

Some data-driven scientific problems

One response Y: phenotype; e.g. Crohn's disease status, cholesterol level

Hundreds of thousands of variables X: genotype information

Ex. 1: which genetic variations affect traits, e.g. the risk of a disease?

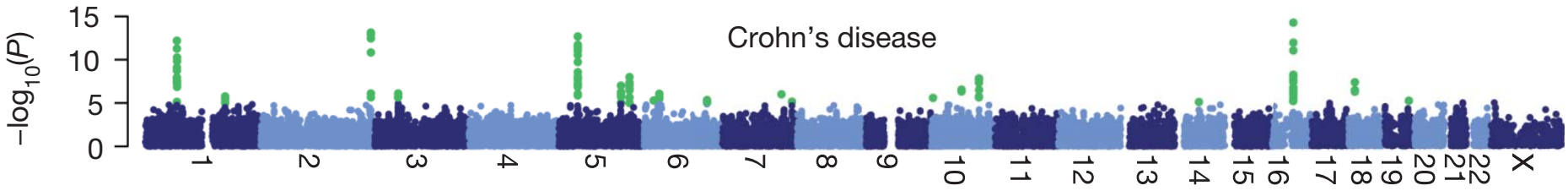


Some data-driven scientific problems

One response Y: phenotype; e.g. Crohn's disease status, cholesterol level

Hundreds of thousands of variables X: genotype information

Ex. 1: which genetic variations affect traits, e.g. the risk of a disease?



Ex. 2: which gene expression profiles help determine severity of a tumor?

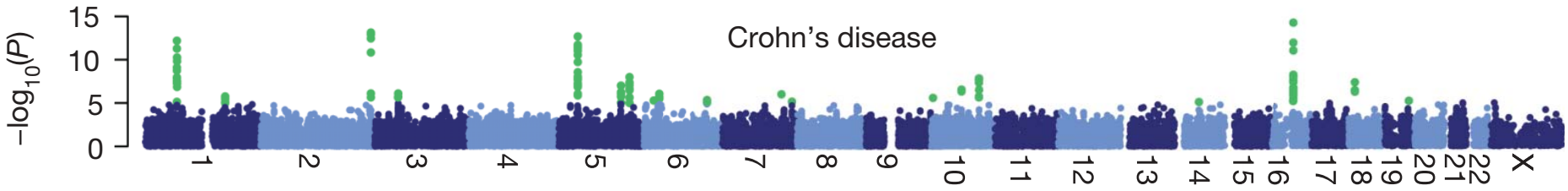
Ex. 3: which factors/variables help determine whether a loan will be repaid?

Some data-driven scientific problems

One response Y: phenotype; e.g. Crohn's disease status, cholesterol level

Hundreds of thousands of variables X: genotype information

Ex. 1: which genetic variations affect traits, e.g. the risk of a disease?



Ex. 2: which gene expression profiles help determine severity of a tumor?

Ex. 3: which factors/variables help determine whether a loan will be repaid?

How can we select variables without too many false positives?
$\rightsquigarrow$ do not run into problem of irreproducibility

Formalizing the selection problem

Each with their X and Y

variables response

Thousands/millions of variables X

Which ones are important?

Distribution of $Y \mid X$ depends on X through which variables?

Formalizing the selection problem

Each with their X and Y

variables response

Thousands/millions of variables X

Which ones are important?

Distribution of $Y \mid X$ depends on X through which variables?

Variable is a discovery if

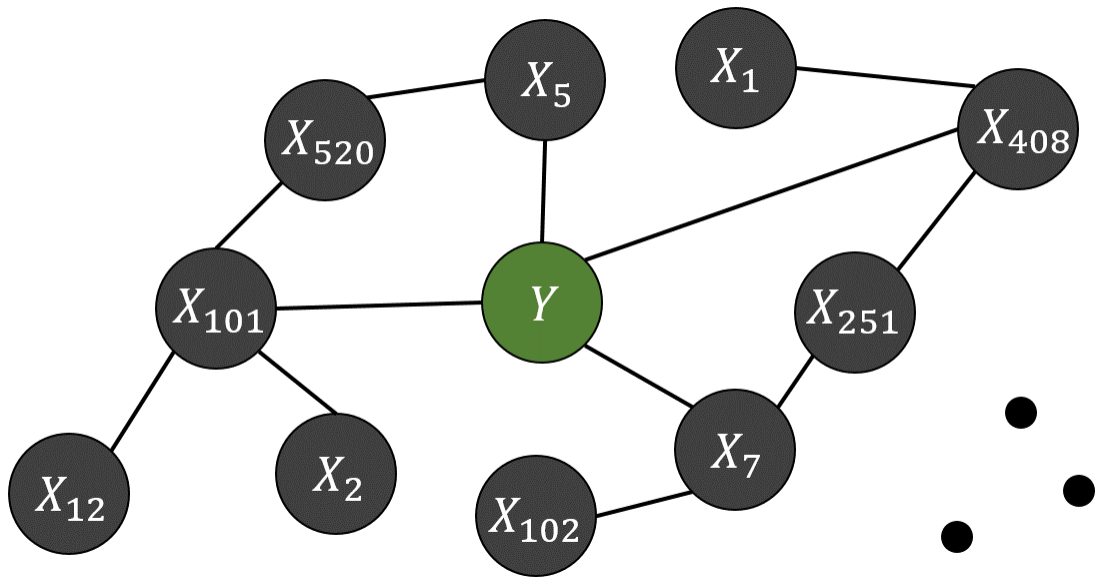
$$p(\text{response} \mid \text{variable}, \text{others}) \neq p(\text{response} \mid \text{others})$$

(Formally) j null iff $Y \perp\!\!\!\perp X_j \mid X_{-j}$

Conditional testing

$$j \text{ null iff } Y \perp\!\!\!\perp X_j \,|\, X_{-j}$$

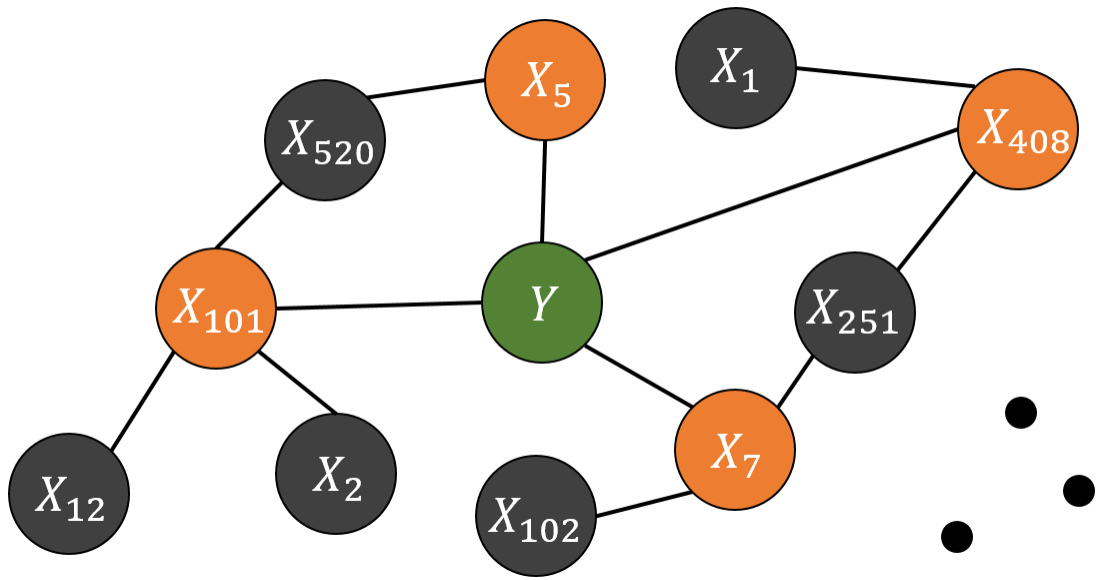
Local Markov property $\Longrightarrow$ non nulls form Markov blanket of Y



Conditional testing

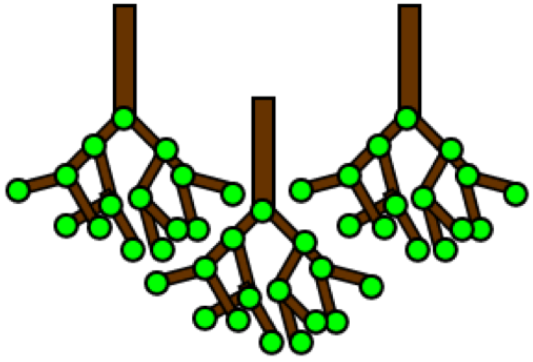
$$j \text{ null iff } Y \perp\!\!\!\perp X_j \,|\, X_{-j}$$

Local Markov property $\Longrightarrow$ non nulls form Markov blanket of Y

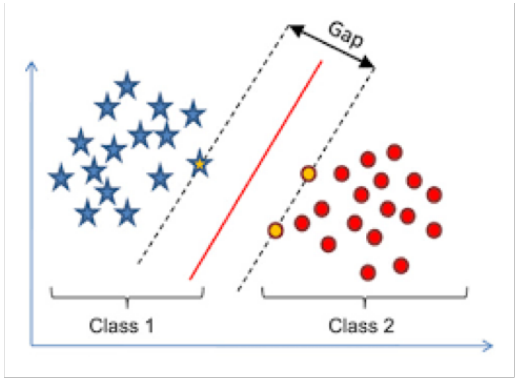


Selection in the computer age

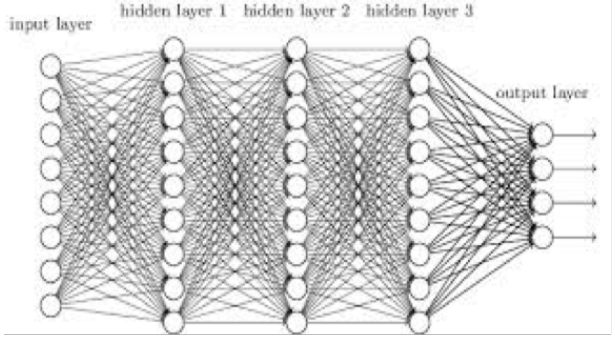
Many sophisticated tools to measure strength of dependence



Random forests

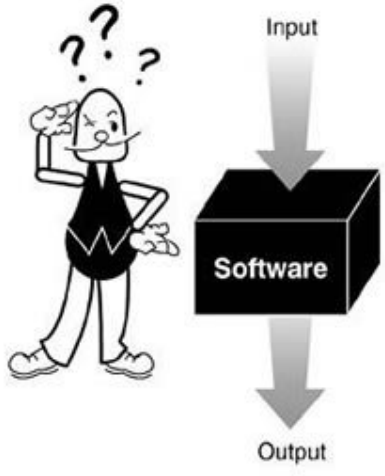


SVM



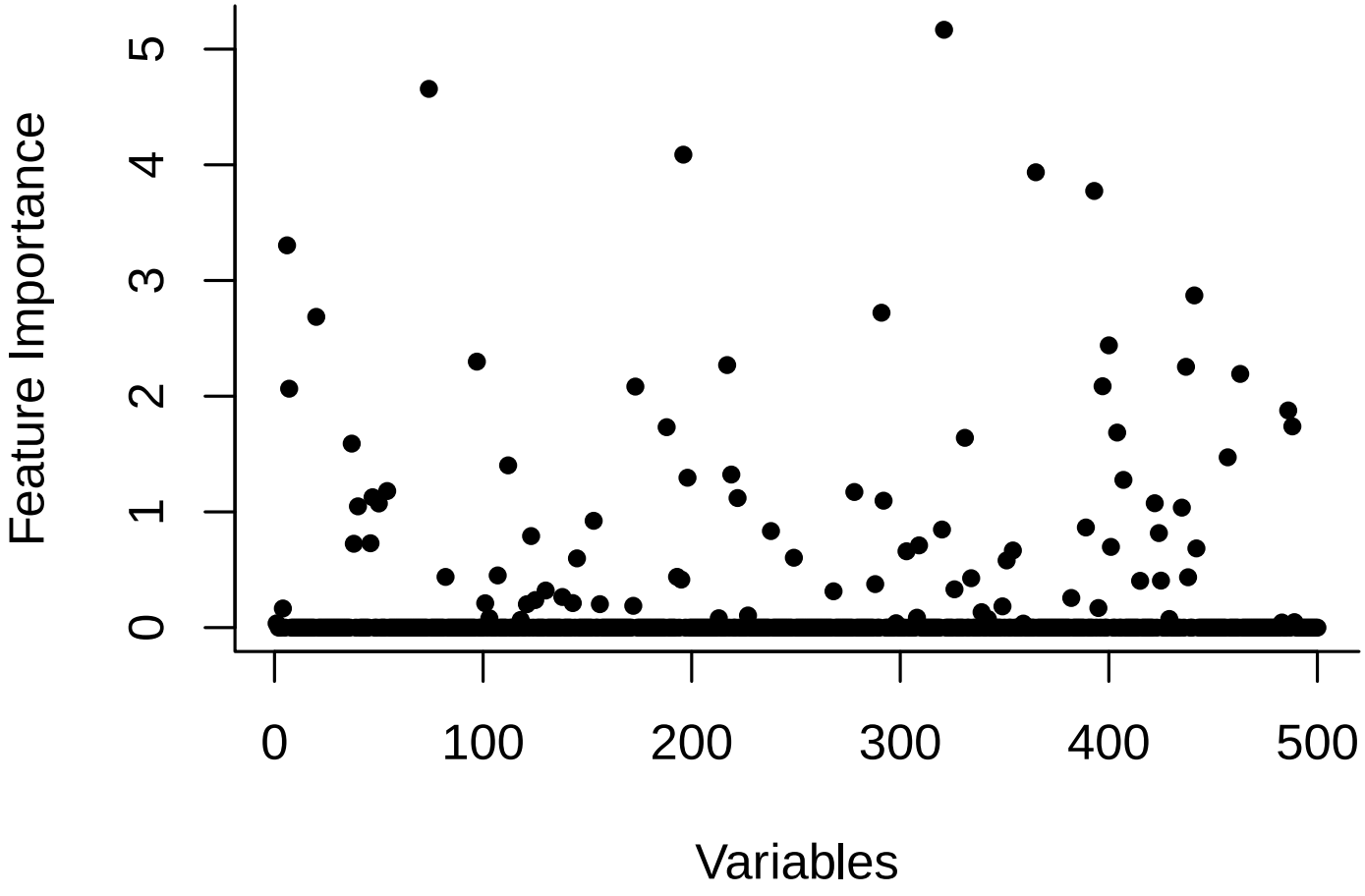
Deep nets

Bayes posteriors (MCMC)



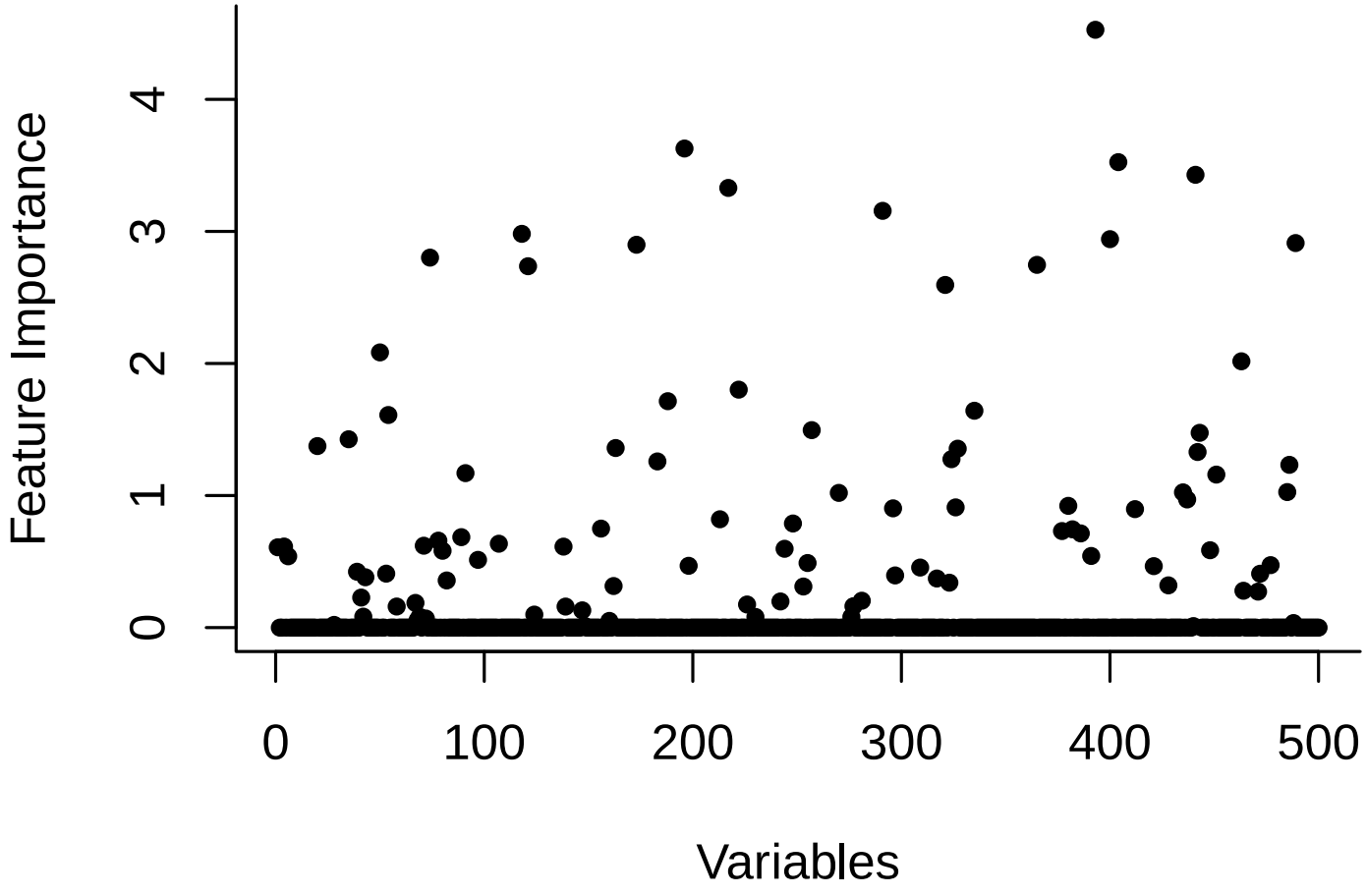
Black-box algorithm

‘Black box’ produces measures of strength of dependence



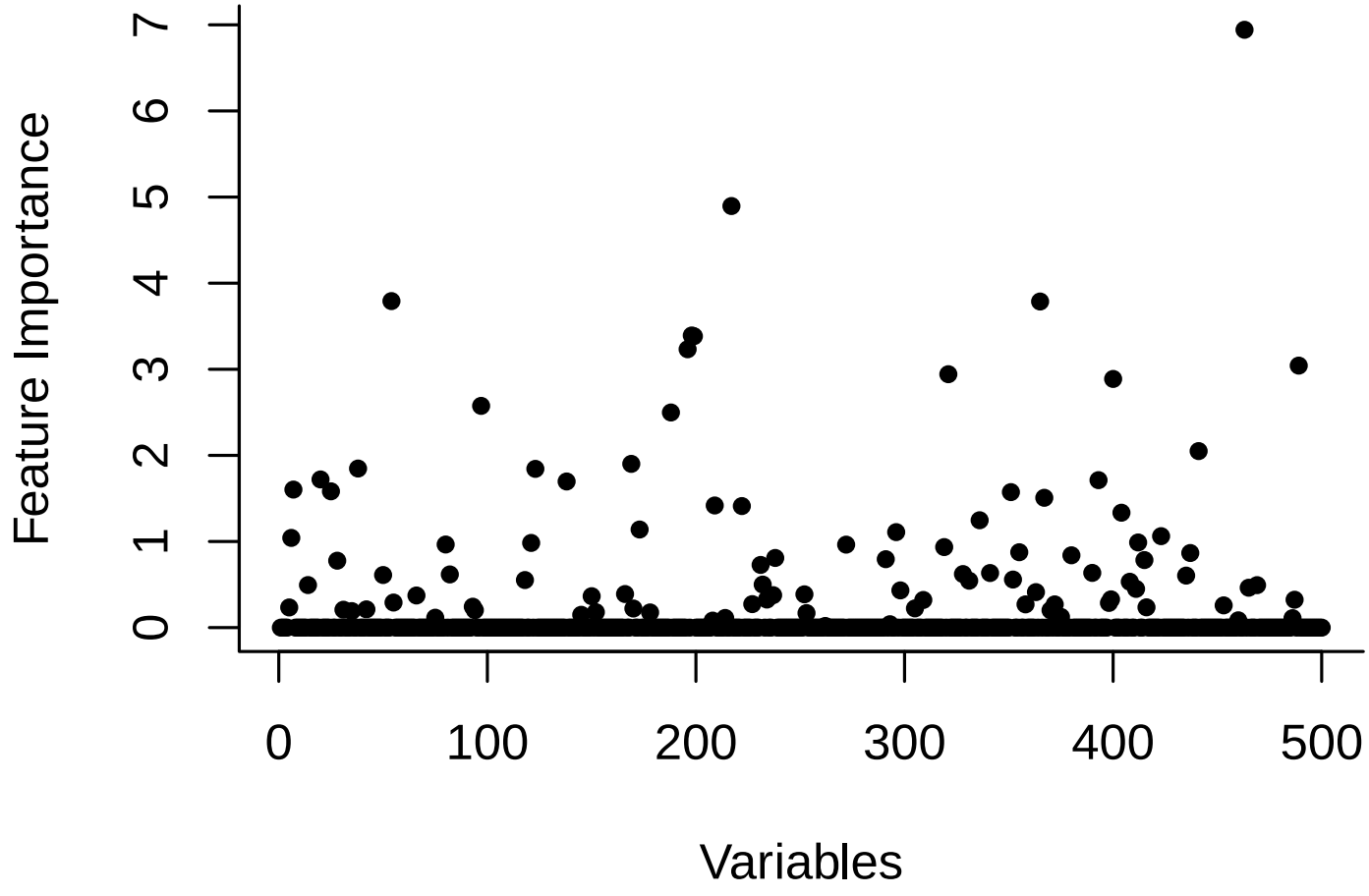
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



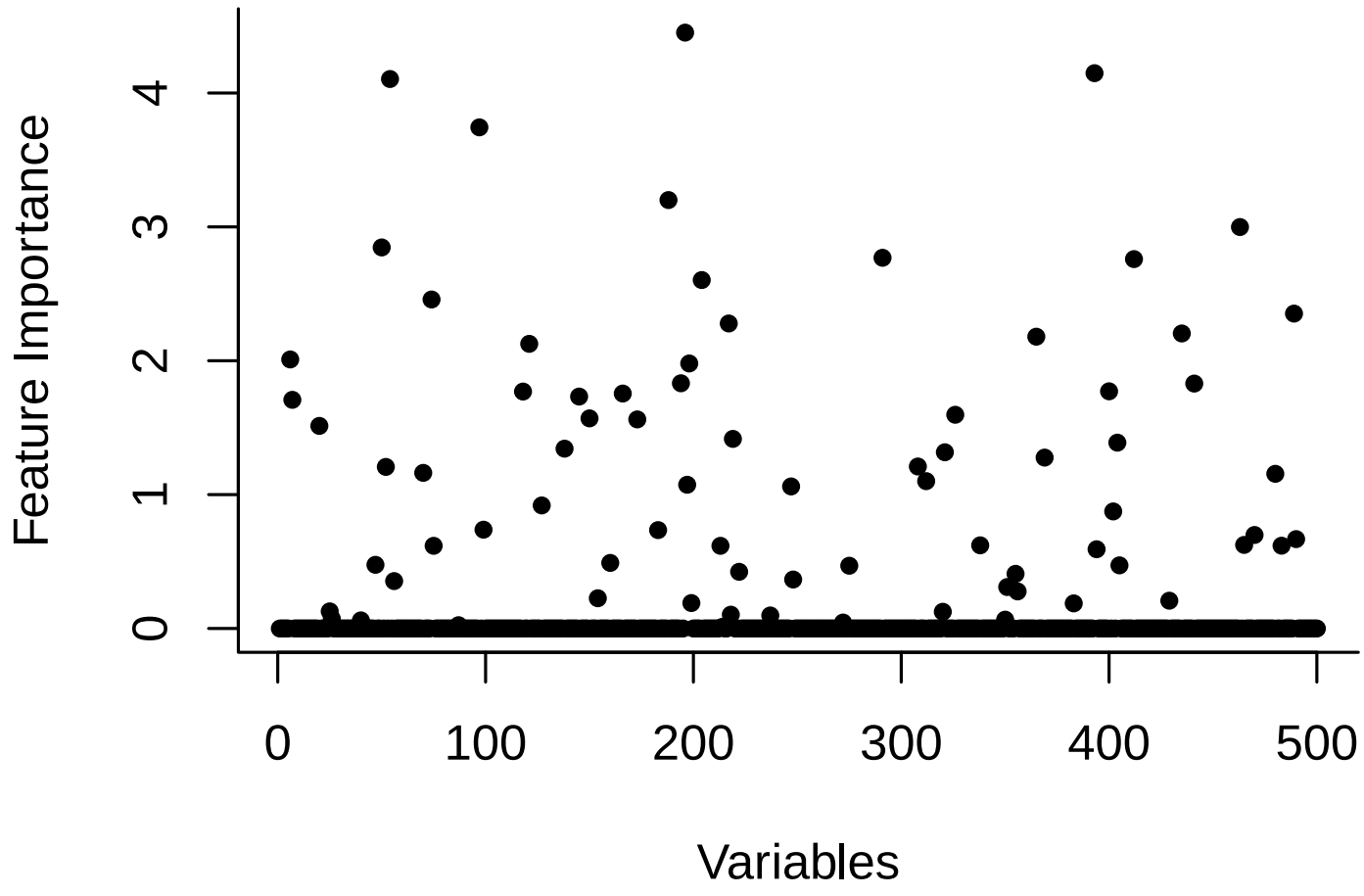
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



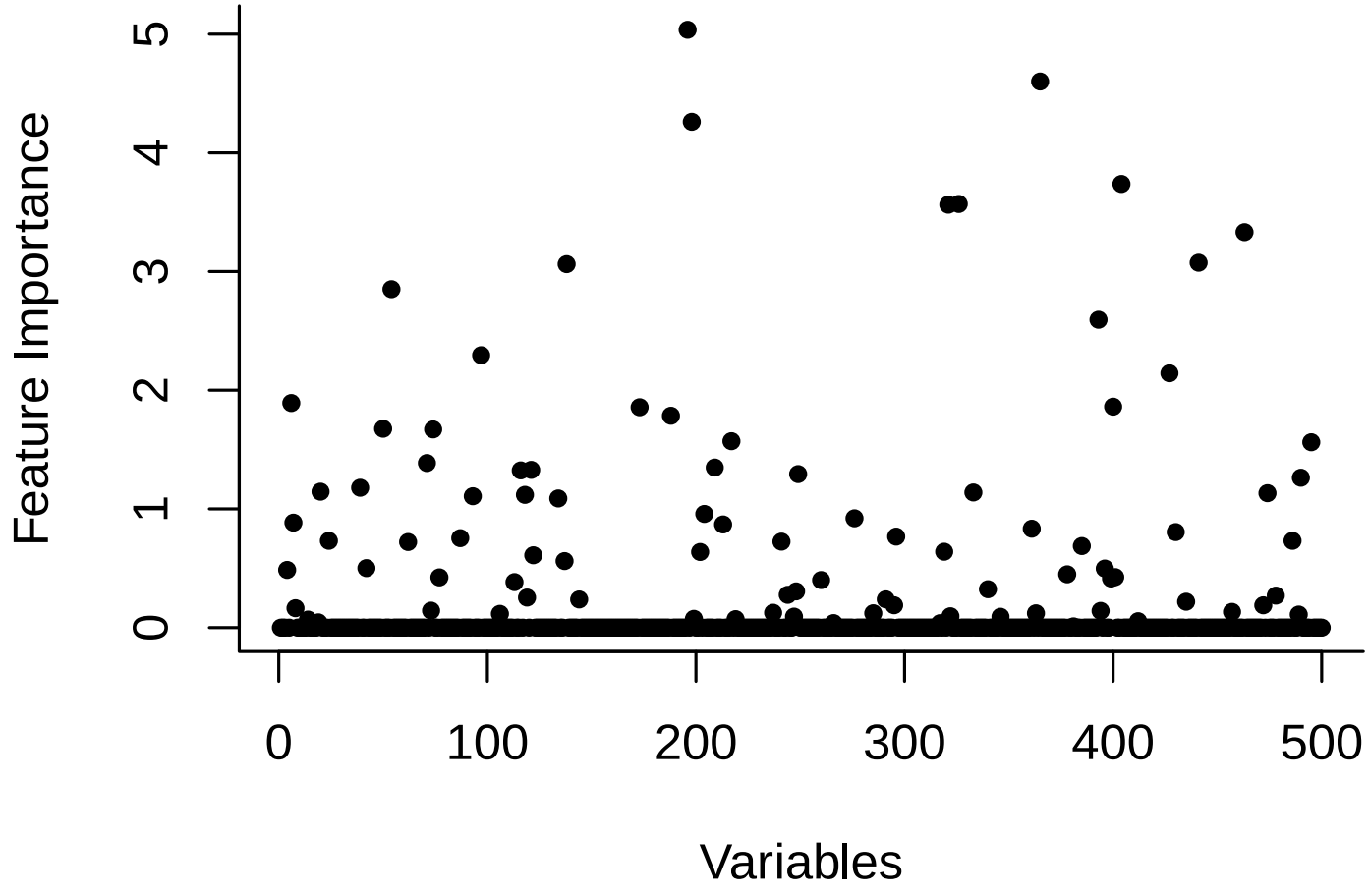
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



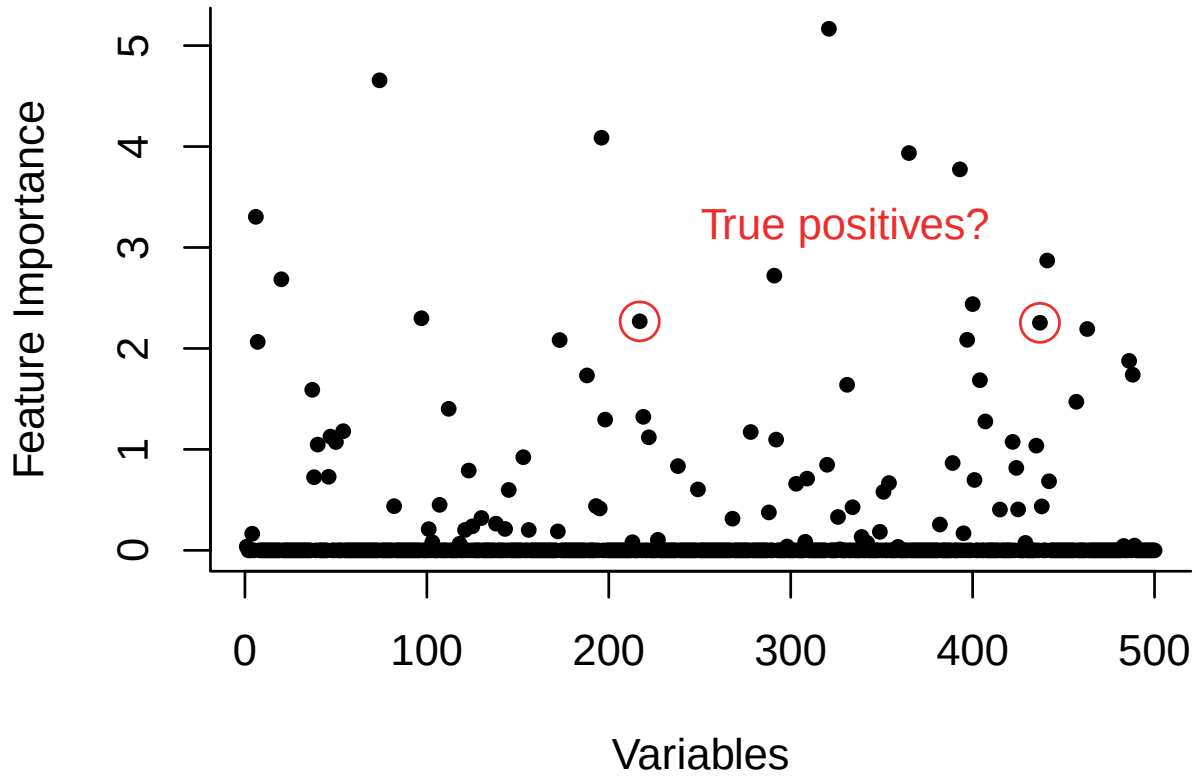
How do we make reliable decisions in the face of unknown statistical variability?

‘Black box’ produces measures of strength of dependence



How do we make reliable decisions in the face of unknown statistical variability?

Only one dataset: what should we report?



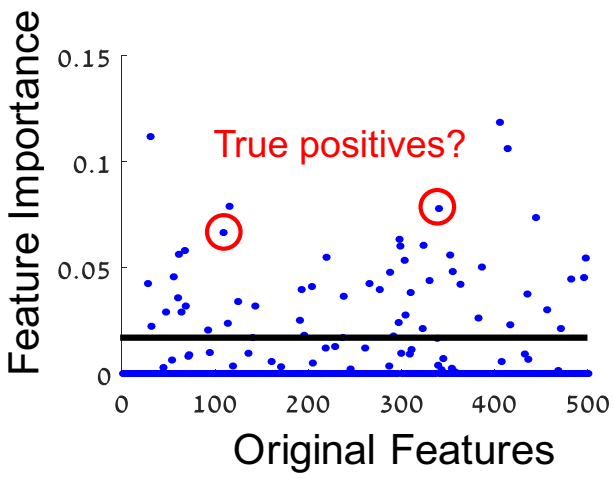
Modern science faces the problem of selection of promising findings from the noisy estimates of many.

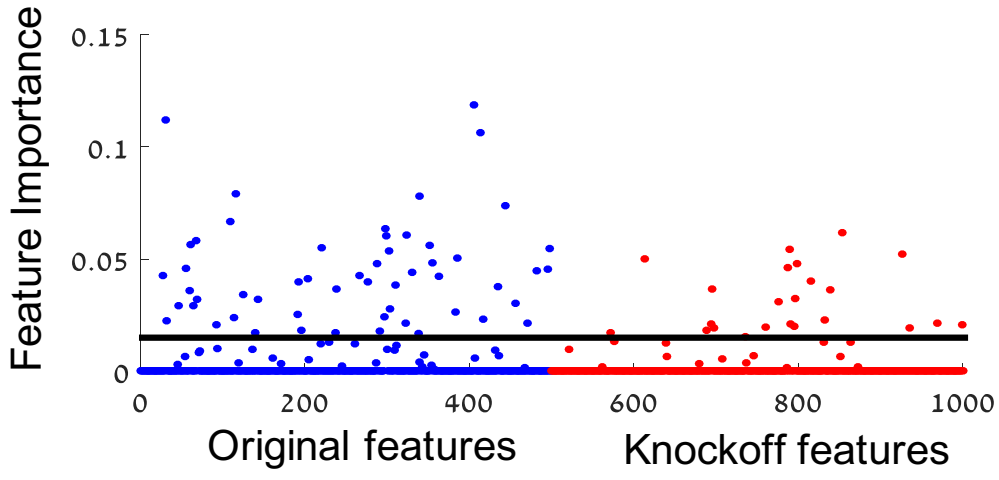
Y. Benjamini and Y. Hechtlinger

Knockoffs (Barber and Candès, 2015)

For each variable (e.g. SNP) X_j, make a knockoff version (e.g. fake SNP) $\tilde{X}_j$

Run scoring procedure on features and knockoffs 'serving as controls'

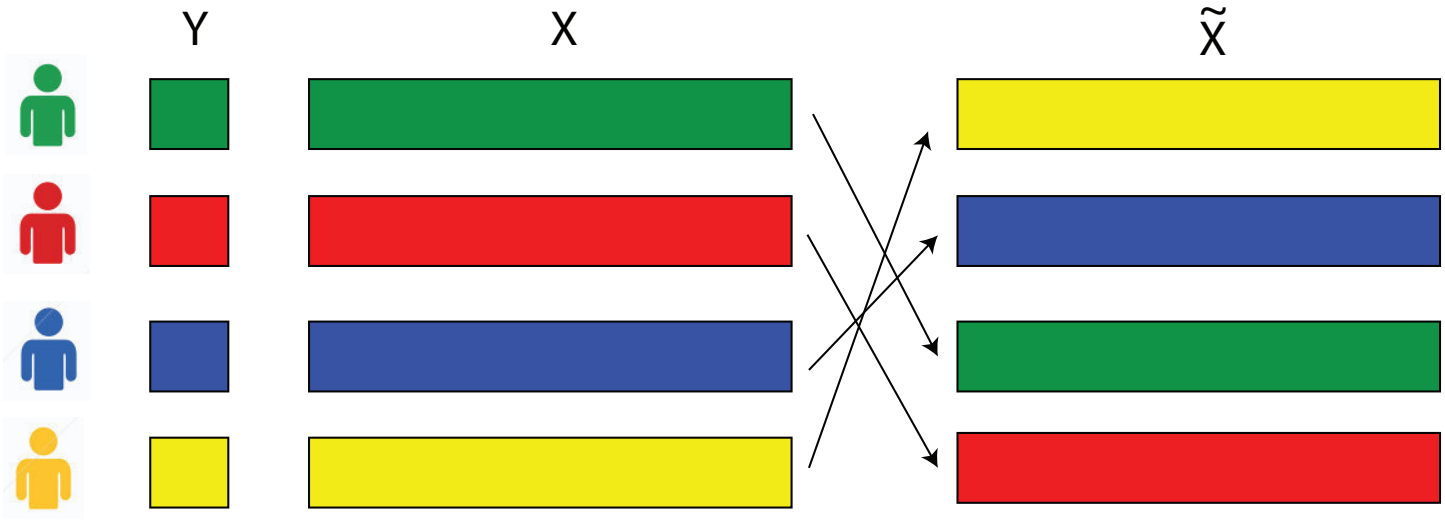




Black box selects 49 original features & 24 knockoff features

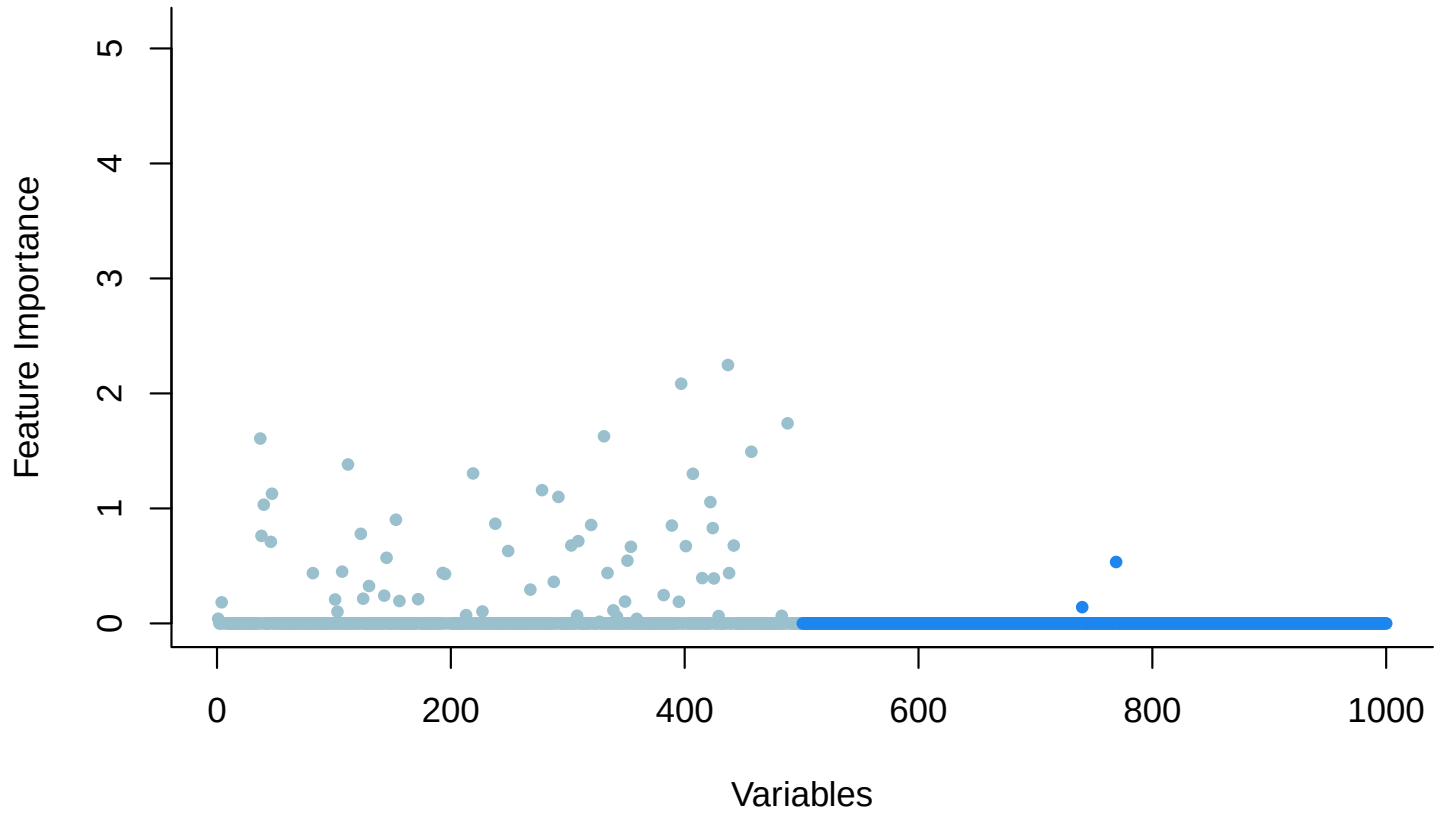
$\Longrightarrow$ probably $\approx$ 24 false positives among 49 original features

How? By permutation?

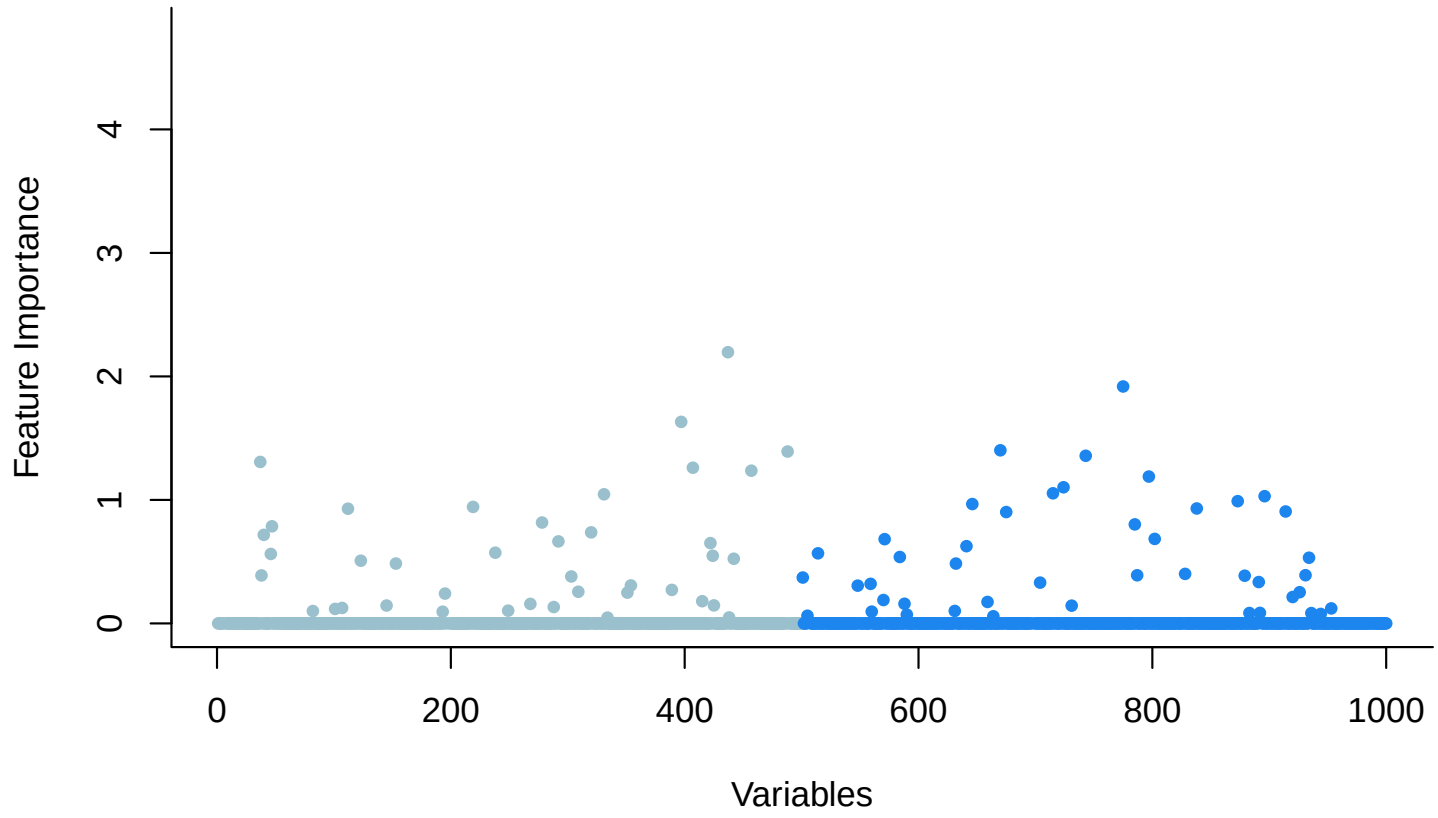


Can I use someone else's genetic info as control?

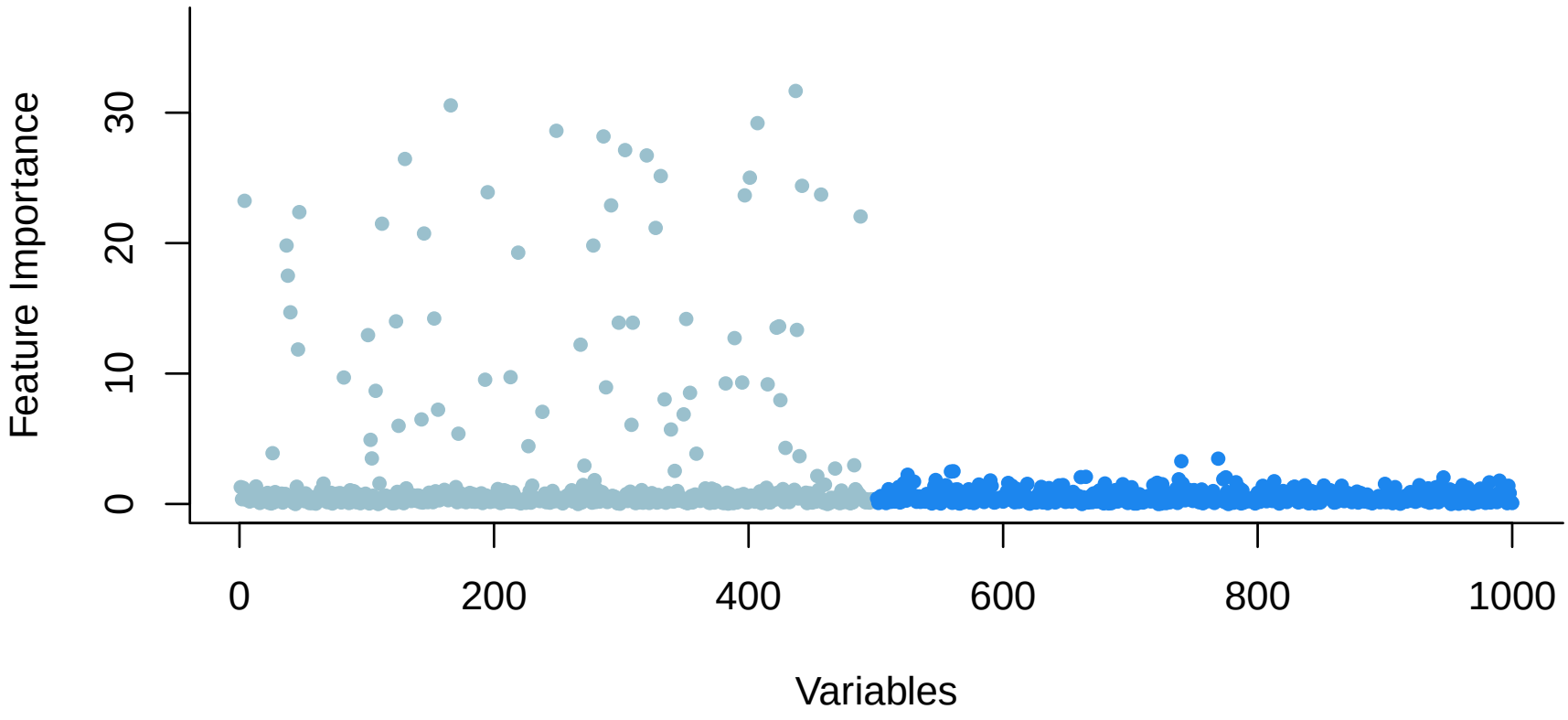
Permuted dummies do not work!



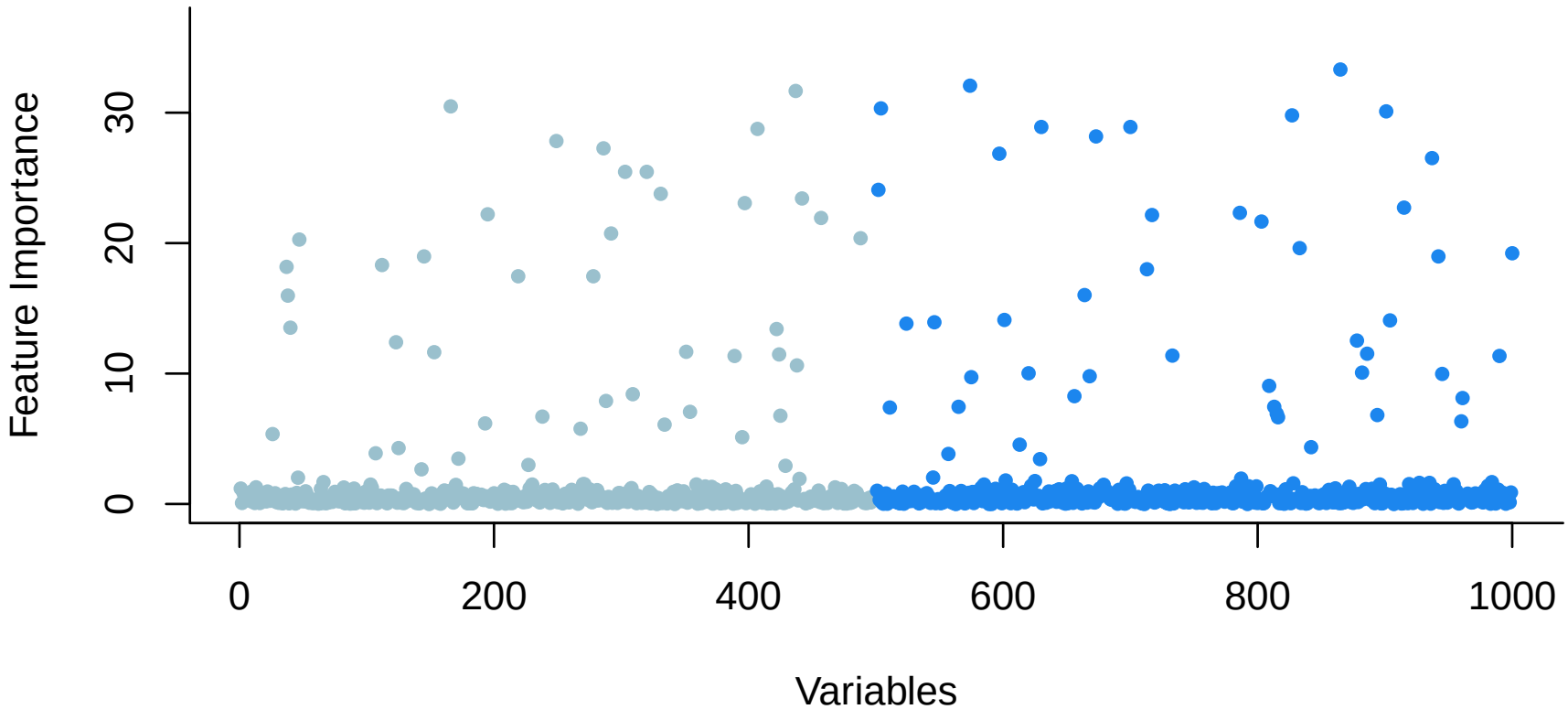
Knockoff dummies work!



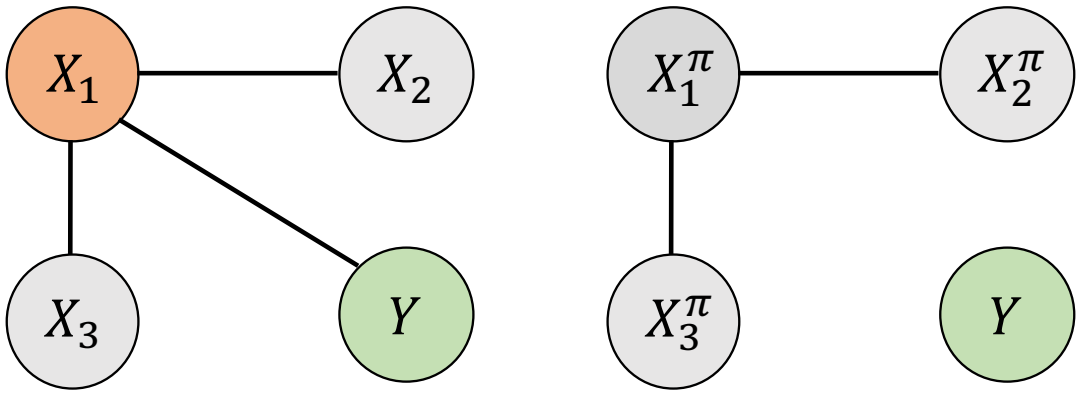
Permuted dummies: other feature importance Z_j



Knockoff dummies: other feature importance Z_j



What's wrong?



Original features

Permuted features

Model-X knockoffs

C., Fan, Janson and Lv ('16)

i.i.d. samples from P_{XY}

- P_X known
- $P_{Y|X}$ completely unknown

Model-X knockoffs

C., Fan, Janson and Lv ('16)

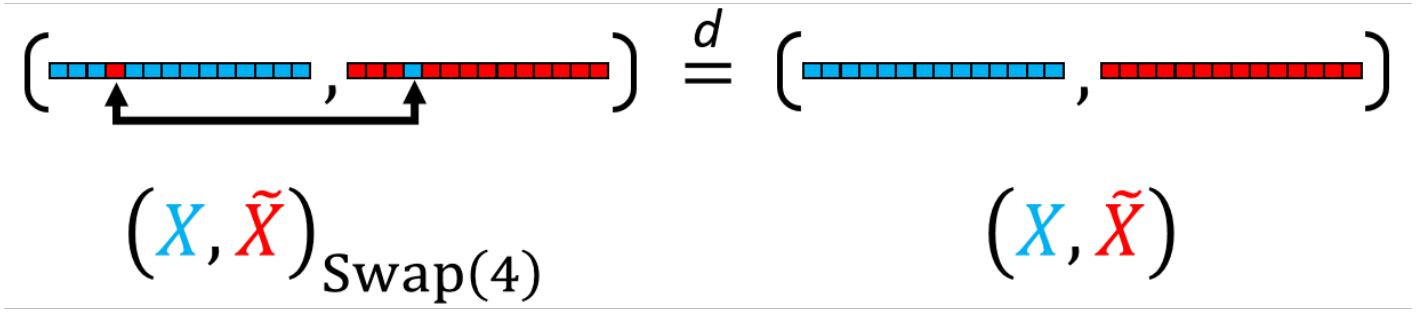
i.i.d. samples from P_{XY}

- P_X known
- $P_{Y|X}$ completely unknown

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

(1) Pairwise exchangeability: for any null j

$$\tilde{X}_j, X_j, X_{-j}, \tilde{X}_{-j} \stackrel{d}{=} X_j, \tilde{X}_j, X_{-j}, \tilde{X}_{-j}$$



Model-X knockoffs

C., Fan, Janson and Lv ('16)

i.i.d. samples from P_{XY}

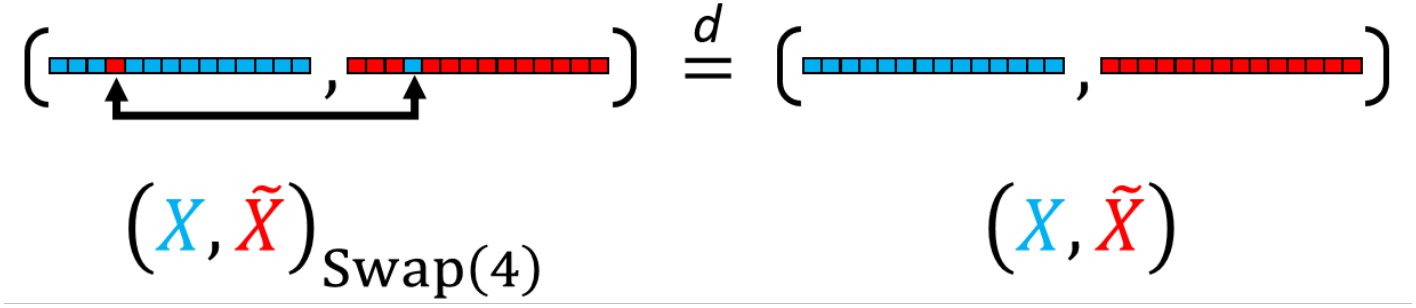
- P_X known
- $P_{Y|X}$ completely unknown

- Originals $X = (X_1, \ldots, X_p)$
- Knockoffs $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_p)$

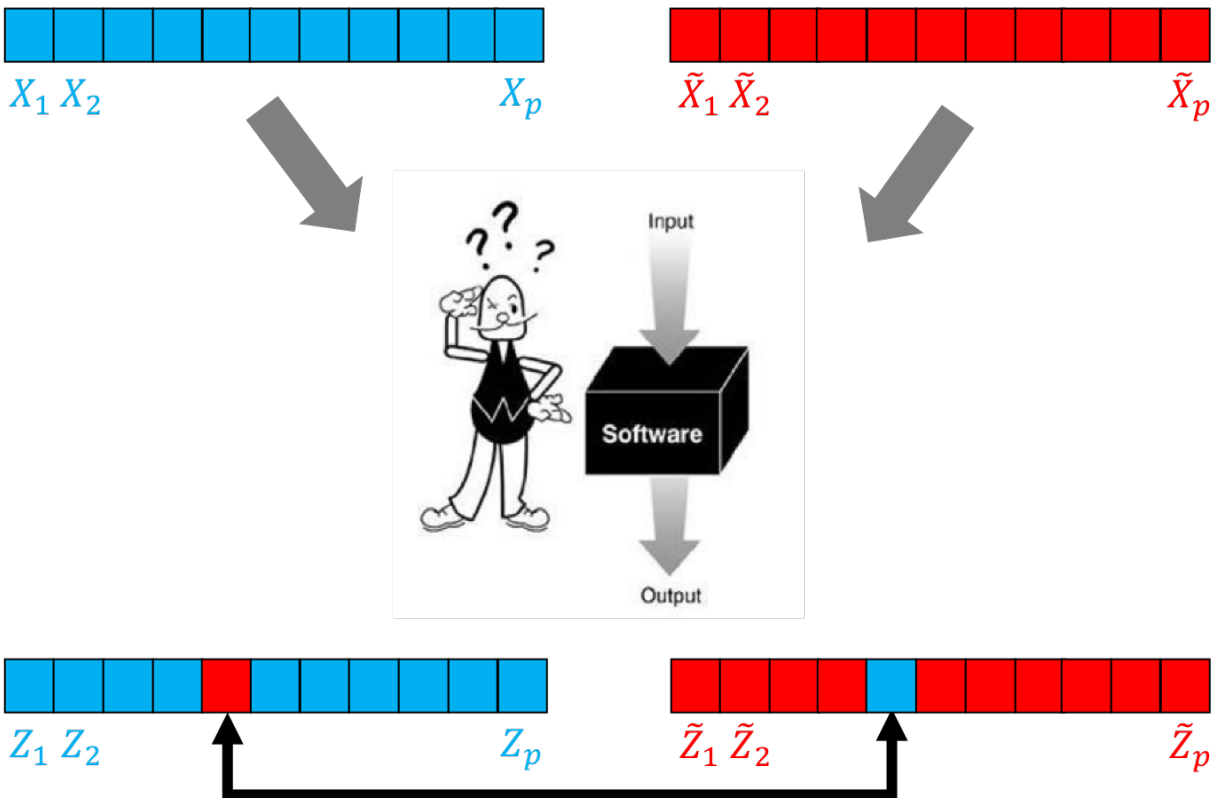
(1) Pairwise exchangeability: for any null j

$$\tilde{X}_j, X_j, X_{-j}, \tilde{X}_{-j} \stackrel{d}{=} X_j, \tilde{X}_j, X_{-j}, \tilde{X}_{-j}$$

(2) Ignore Y when constructing knockoffs: $\tilde{X} \perp\!\!\!\perp Y \,|\, X$



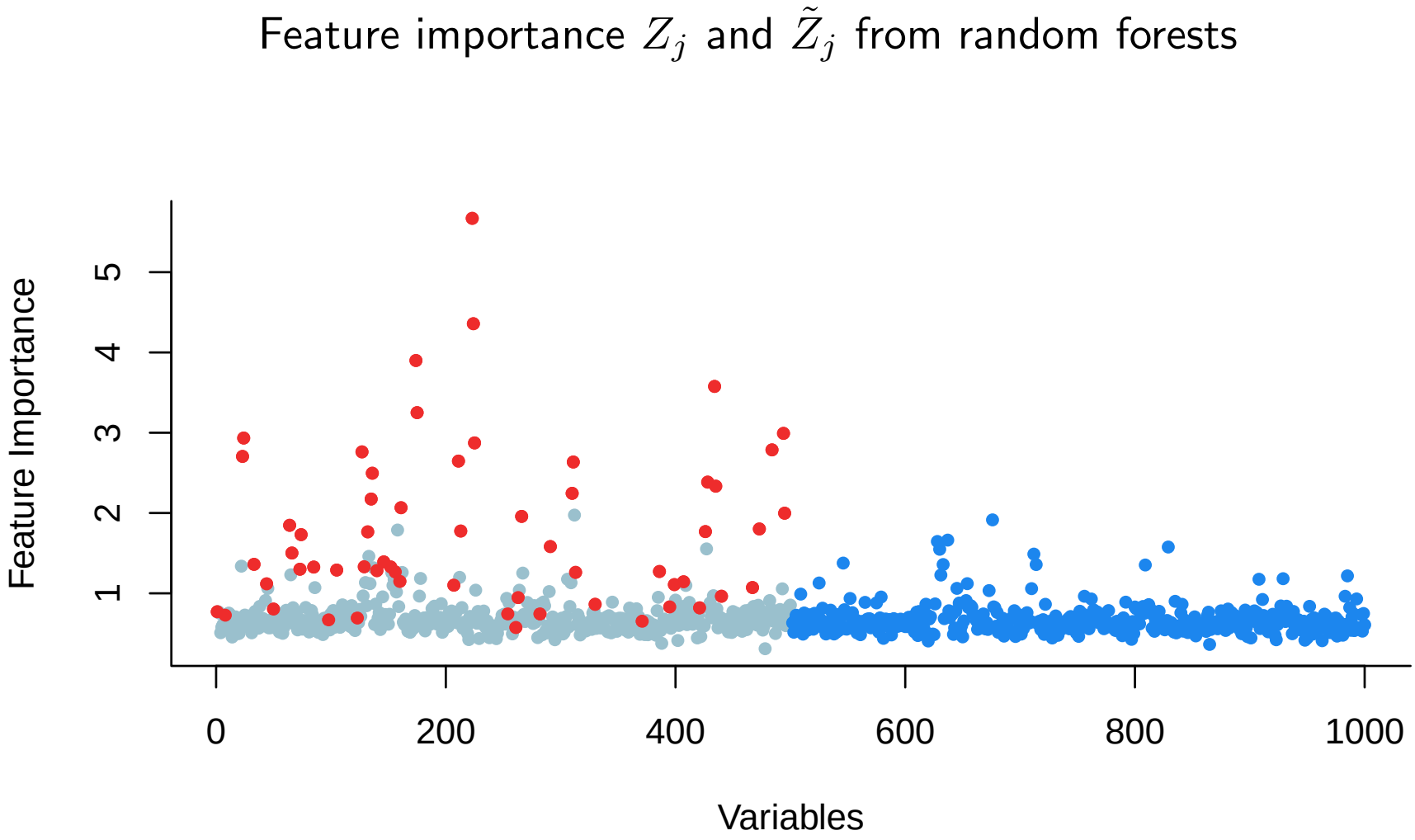
Knockoffs as negative controls



All null scores are exchangeable

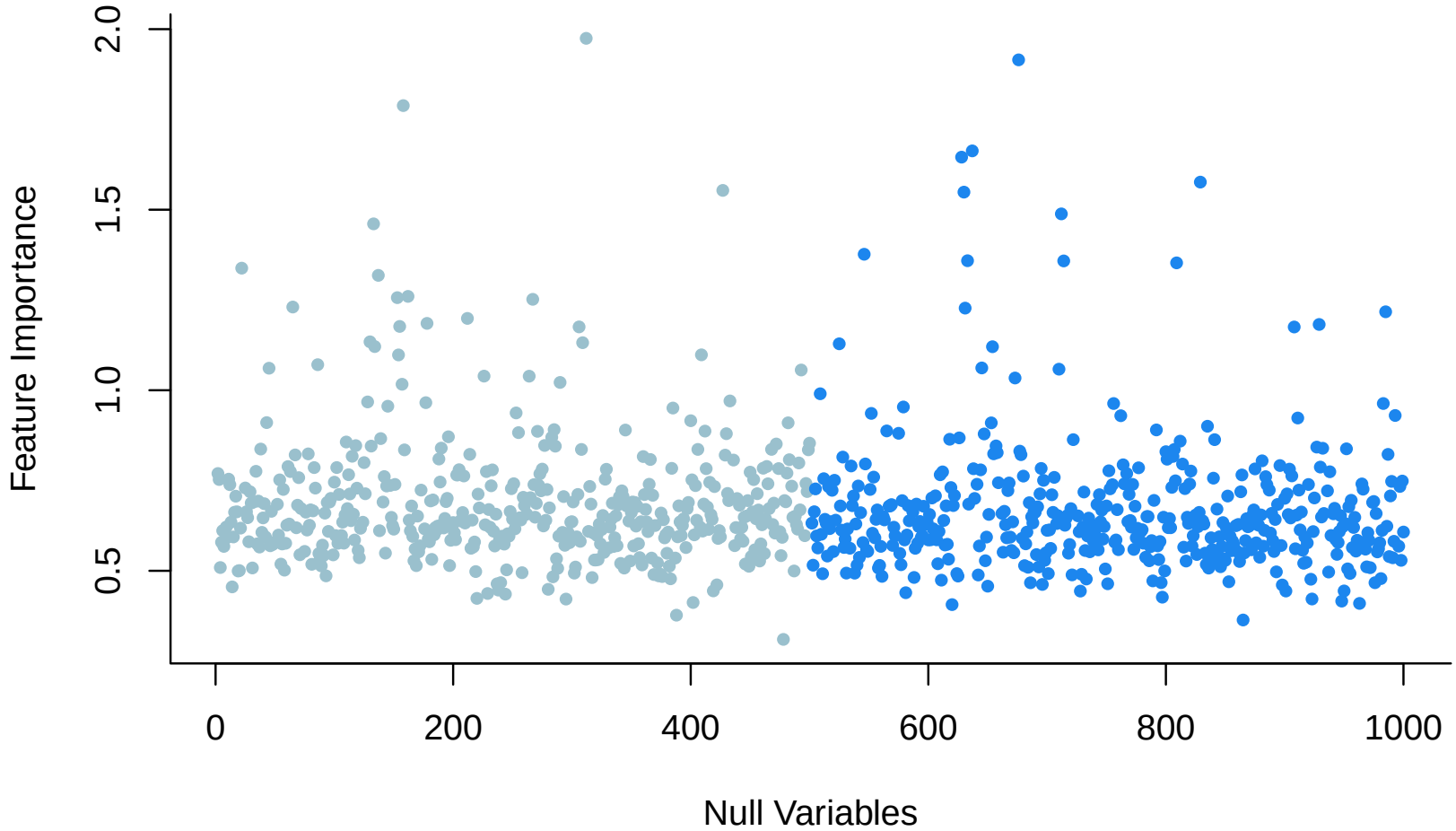
$$(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$$

Knockoffs with binary response



Knockoffs with binary response

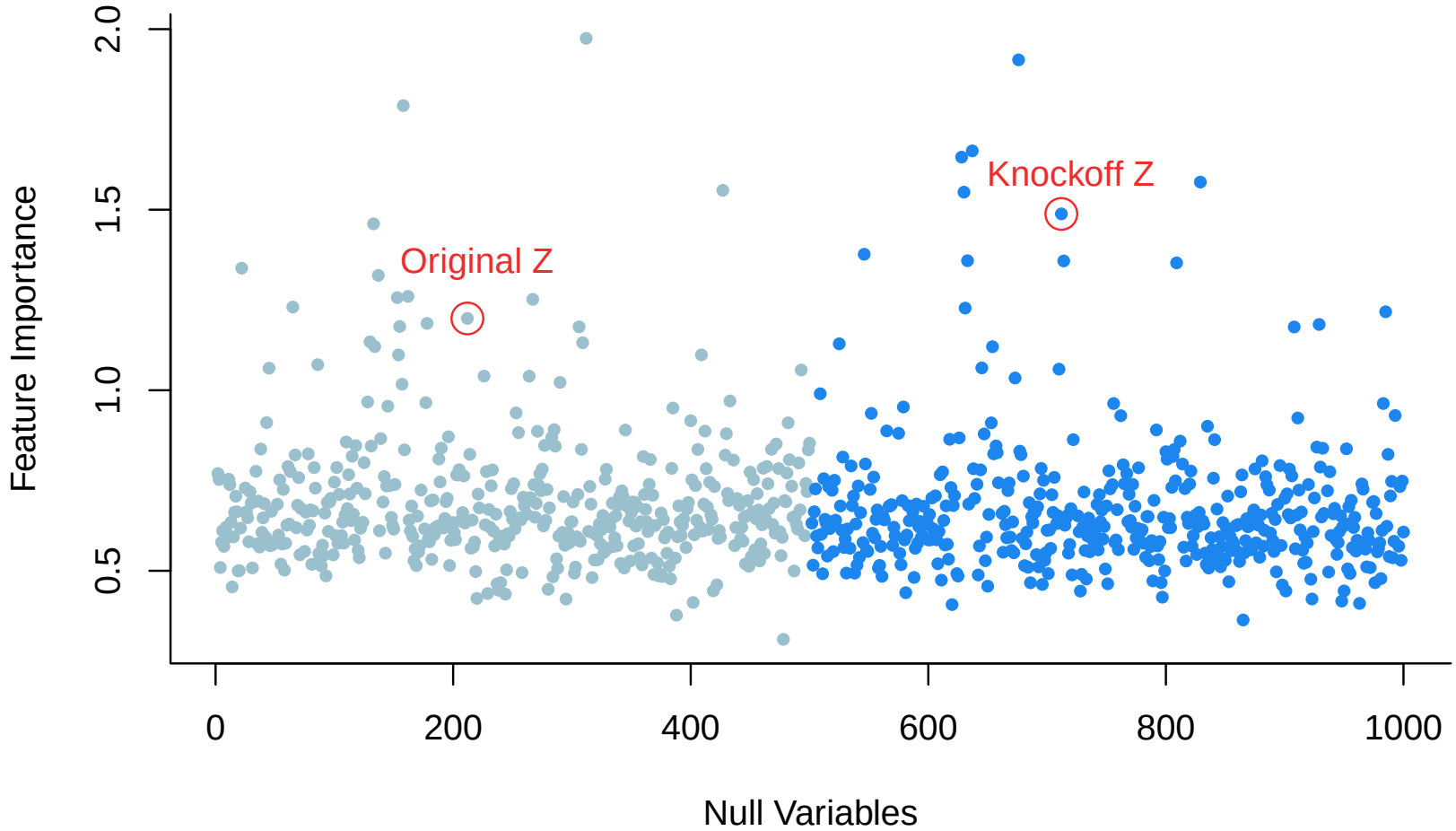
Feature importance Z_j and $\tilde{Z}_j$ from random forests



$$(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$$

Knockoffs with binary response

Feature importance Z_j and $\tilde{Z}_j$ from random forests



$$(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$$

Knockoffs-adjusted scores

Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

e.g. $\qquad W_j = Z_j - \tilde{Z}_j$

Knockoffs-adjusted scores

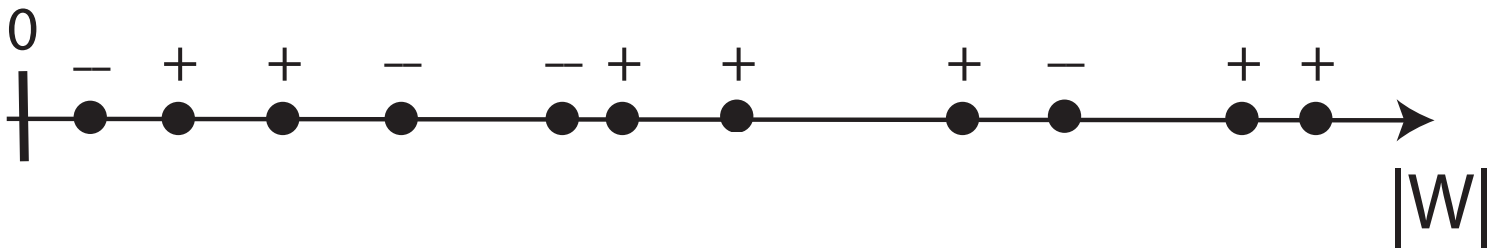
Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

$$\text{e.g.} \qquad W_j = Z_j - \tilde{Z}_j$$

Conditional on $|W|$, signs of null W_j's are i.i.d. coin flips



Knockoffs-adjusted scores

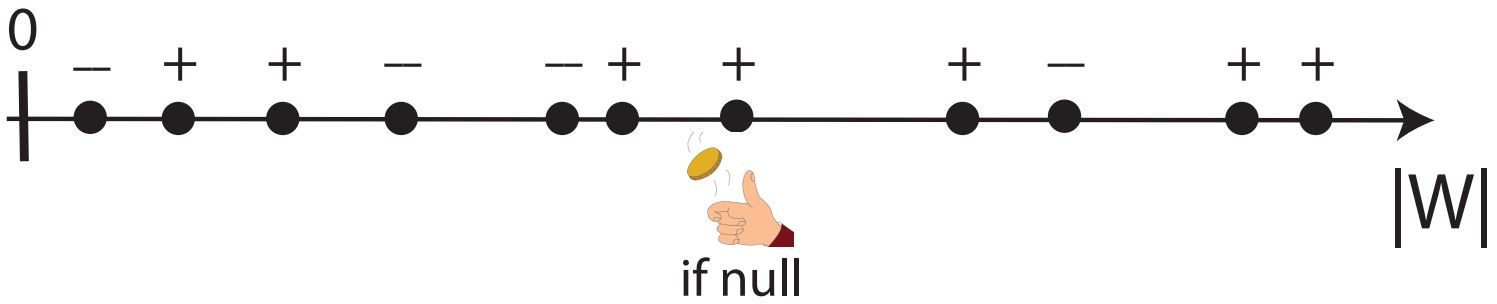
Adjusted scores W_j with flip-sign property

Combine Z_j and $\tilde{Z}_j$ into single (knockoff) score W_j

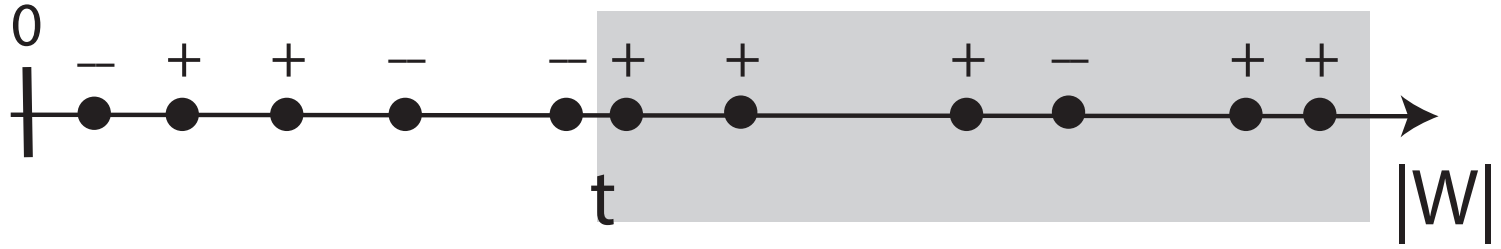
$$W_j = w_j(Z_j, \tilde{Z}_j) \qquad w_j(\tilde{Z}_j, Z_j) = -w_j(Z_j, \tilde{Z}_j)$$

$$\text{e.g.} \qquad W_j = Z_j - \tilde{Z}_j$$

Conditional on $|W|$, signs of null W_j's are i.i.d. coin flips

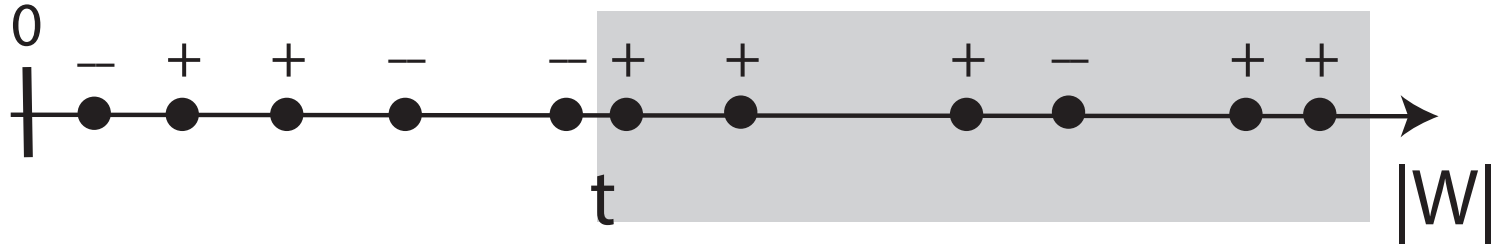


Knockoff estimate of FDR



Interested in selecting $\{j : W_j \geq t\}$

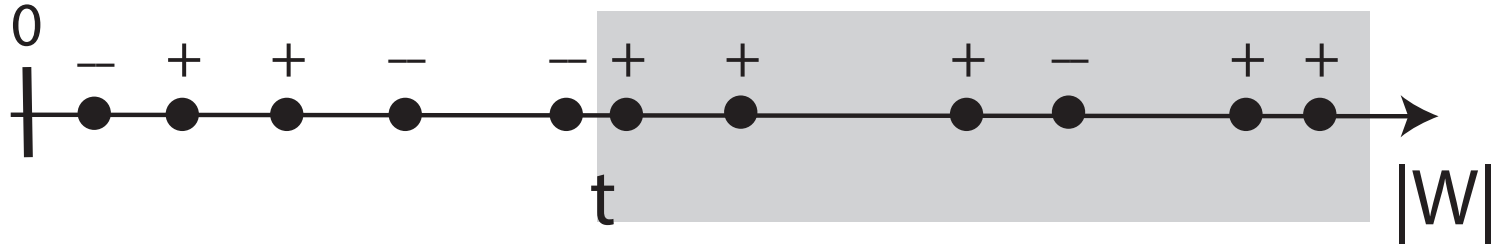
Knockoff estimate of FDR



Interested in selecting $\{j : W_j \geq t\}$

$$\text{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1}$$

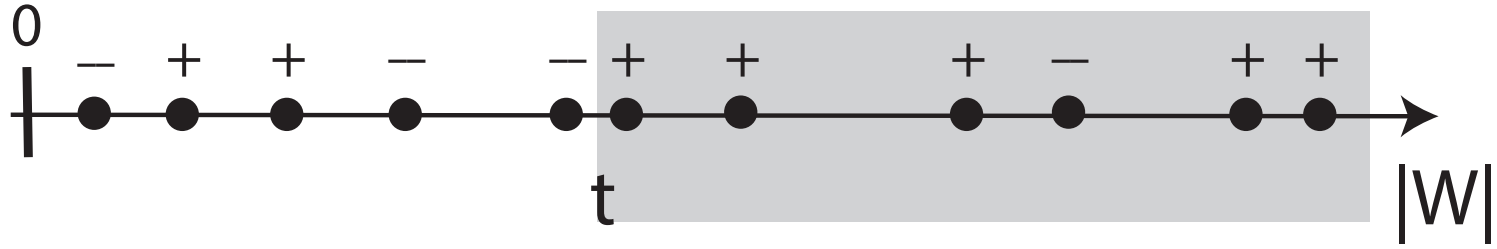
Knockoff estimate of FDR



Interested in selecting $\{j : W_j \geq t\}$

$$\text{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j \text{ null } : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

Knockoff estimate of FDR

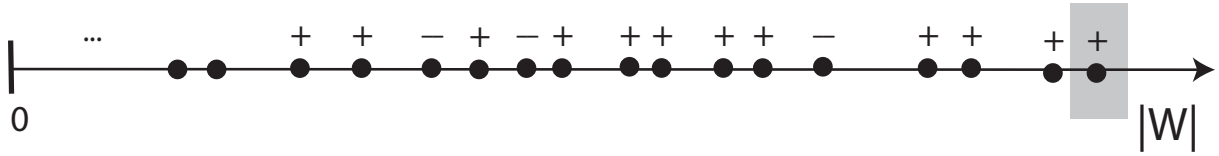


Interested in selecting $\{j : W_j \geq t\}$

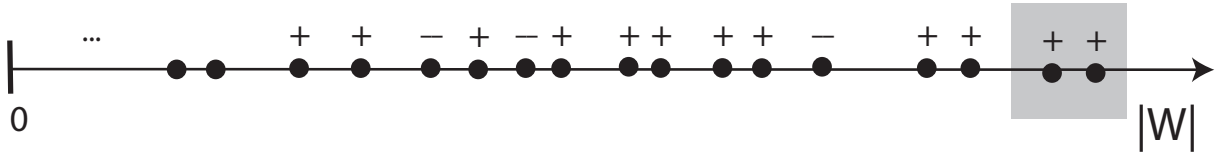
$$\mathsf{FDP}(t) = \frac{\#\{j \text{ null } : W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j \text{ null } : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

$$\leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} := \widehat{\mathsf{FDP}}(t)$$

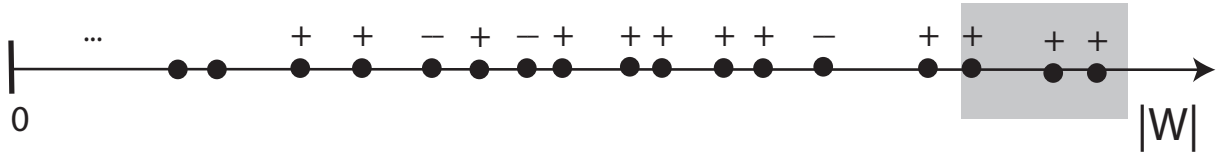
Selection (via sequential testing)



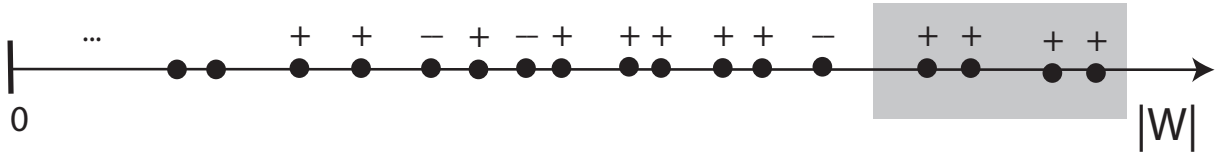
Selection (via sequential testing)



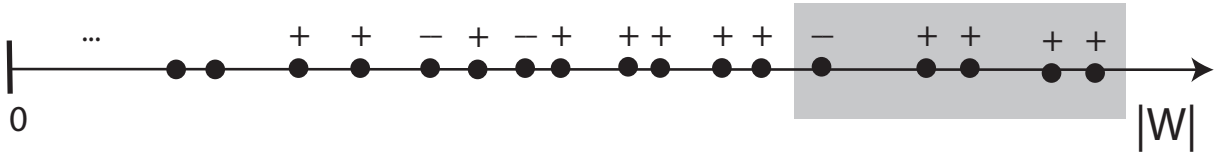
Selection (via sequential testing)



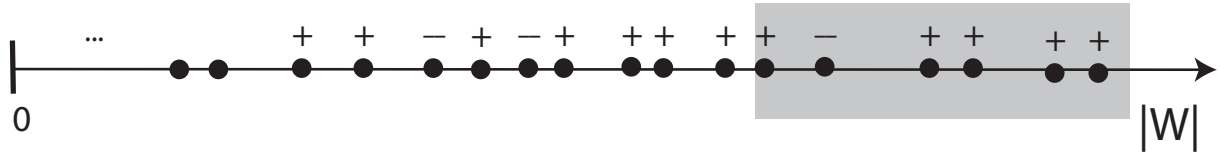
Selection (via sequential testing)



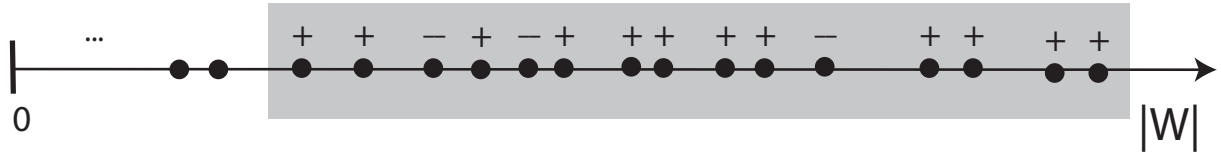
Selection (via sequential testing)



Selection (via sequential testing)

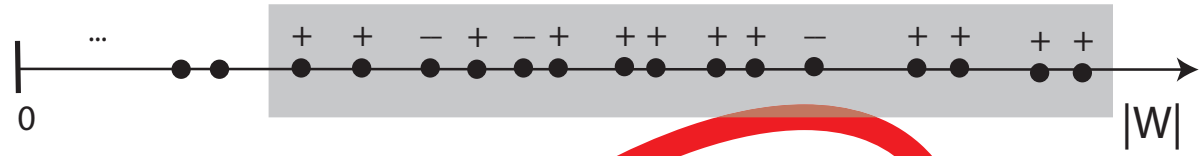


Selection (via sequential testing)

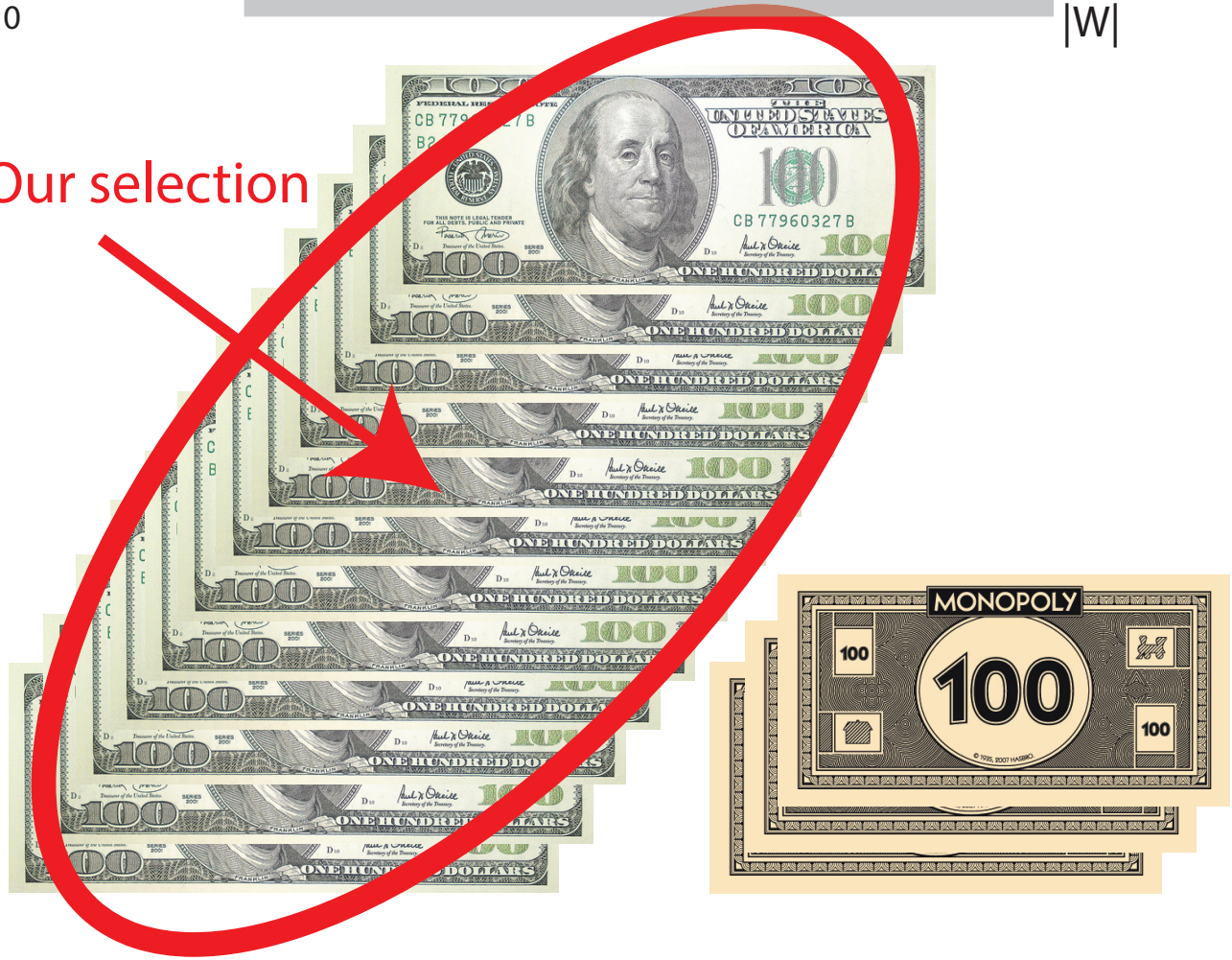


Step-up rule: stop last time ratio between '-' and '+' below target FDR level

Selection (via sequential testing)



Our selection



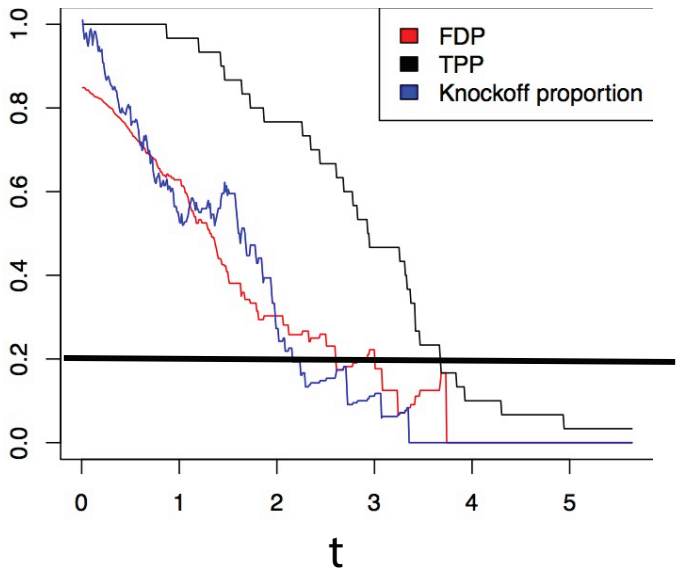
Select '+'s

Step-up rule: stop last time ratio between '-' and '+' below target FDR level

FDR control

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$

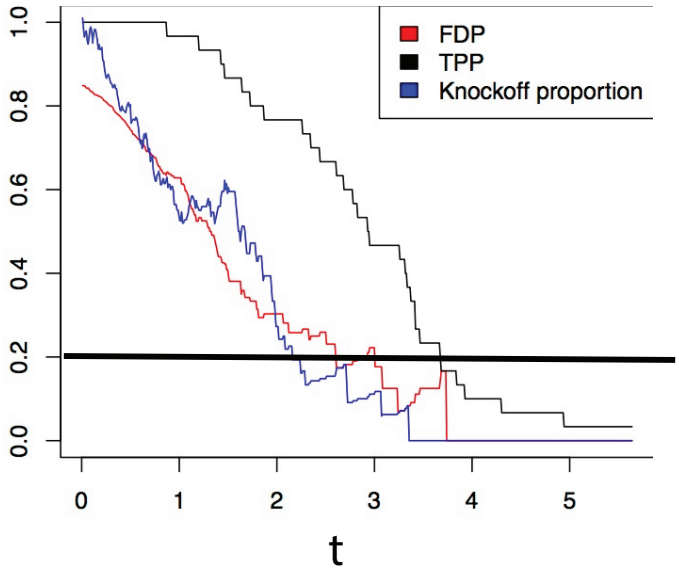
$$\tau = \min\left\{t : \widehat{\mathsf{FDP}}(t) \leq q\right\}$$



FDR control

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$

$$\tau = \min\left\{t : \widehat{\mathsf{FDP}}(t) \leq q\right\}$$



Theorem (Barber and C, '15)

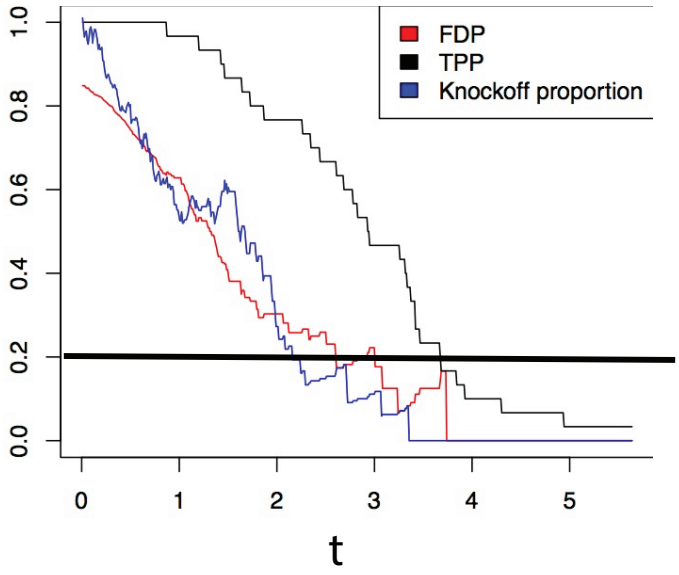
For knockoff+

$$\mathsf{FDR} = \mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections}}\right] \leq q$$

FDR control

$$\hat{\mathcal{S}} = \{W_j \geq \tau\}$$

$$\tau = \min\left\{t : \widehat{\mathsf{FDP}}(t) \leq q\right\}$$



Theorem (Barber and C, '15)

For knockoff+

$$\mathsf{FDR} = \mathbb{E}\left[\frac{\#\text{ false positives}}{\#\text{ selections}}\right] \leq q$$

Robust extension (Barber, C. and Samworth, '18):
P_X not known exactly and knockoffs are exchangeable only w.r.t. $Q_X \approx P_X$

Knockoffs framework

Always FDR control

- ✓ Under finite sample
- ✓ Any dimension (including $p > n$)
- ✓ Any model for $Y \mid X$
- ✓ Any black-box

The cost?

How to construct knockoffs?

Need access to P_X (not $P_{Y|X}$)

Let's Make Knockoffs! (Only a Taste)

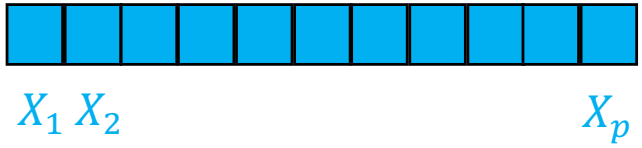
P_X *known:*	*with M. Sesia and C. Sabatti*
	with S. Bates, L. Janson and W. Wang
P_X *unknown:*	*with Y. Romano and M. Sesia*

How to make knockoffs?

Input

Features X

Dist. P_X

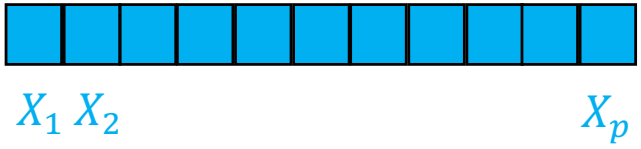


How to make knockoffs?

Input

Features X

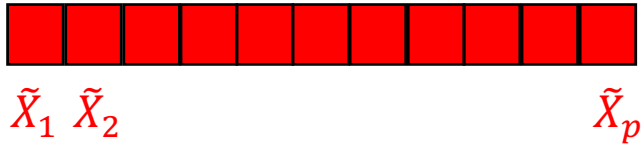
Dist. P_X



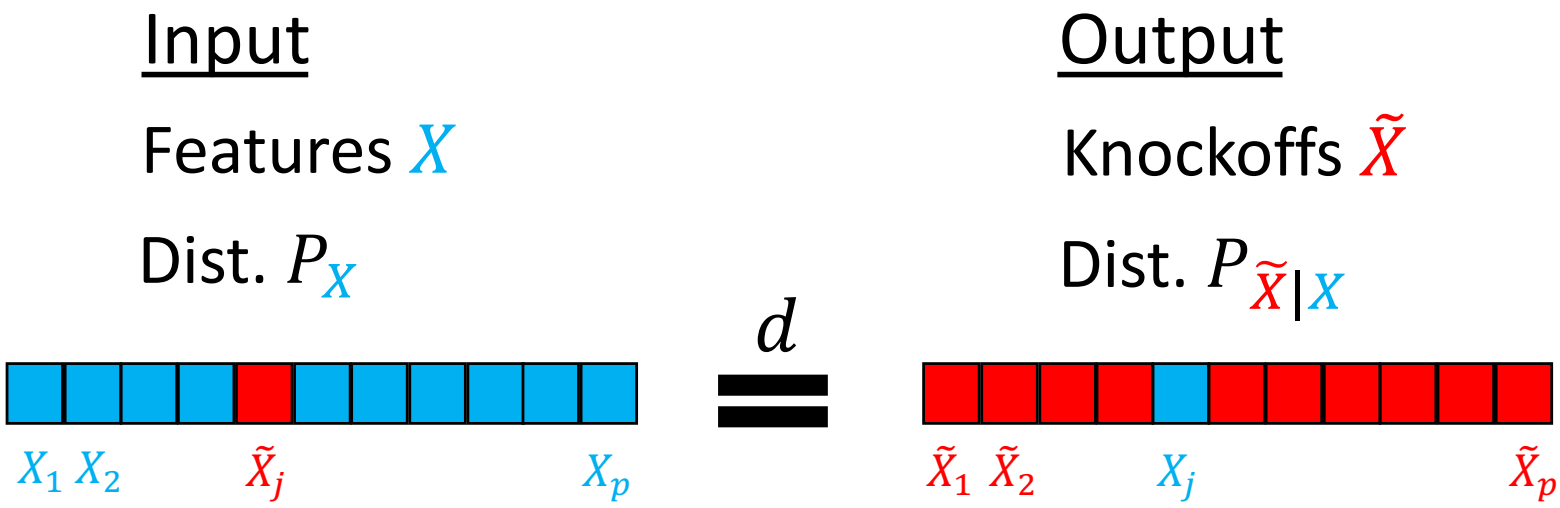
Output

Knockoffs $\tilde{X}$

Dist. $P_{\tilde{X}|X}$



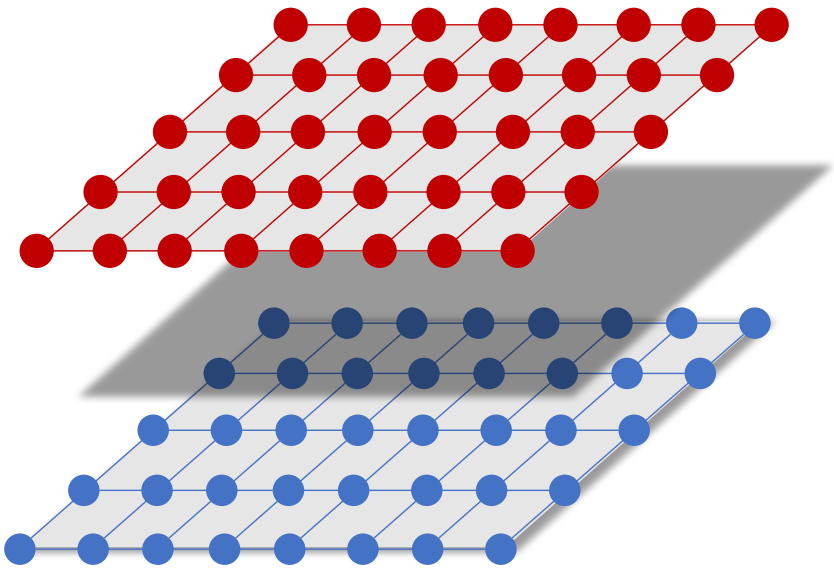
How to make knockoffs?



Exchangeability

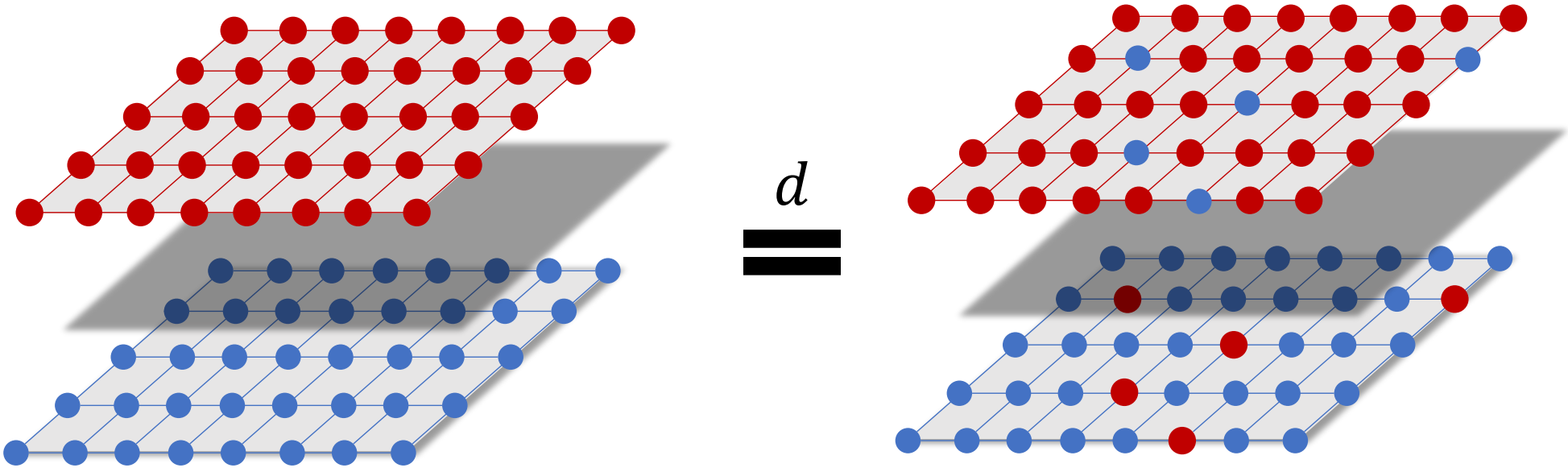
Challenges

$$P_X = \frac{1}{Z} e^{-\sum_{i,j} \beta_{ij} X_i X_j}$$



Challenges

$$P_X = \frac{1}{Z} e^{-\sum_{i,j} \beta_{ij} X_i X_j}$$



$$\overset{d}{=}$$

How to make $P_{\widetilde{X}|X}$?

The knockoff factory

- Gaussian variables

  C., Fan, Janson & Lv '16
- Hidden Markov models

  C., Sabatti & Sesia '17
- Some Bayesian networks

  Gimenez, Ghorbani & Zou '18
- Pretty much anything (e.g. any graphical model)

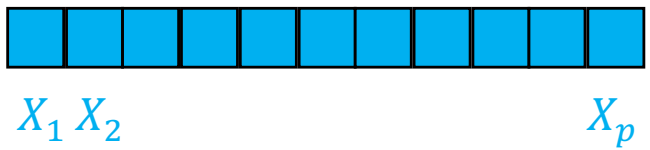
  Bates, C., Janson & Wang '19

Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

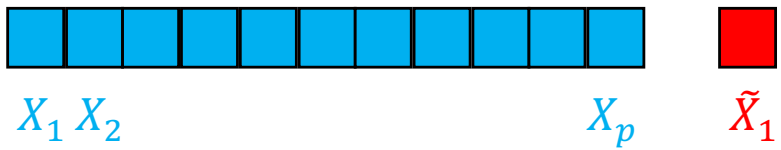


Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

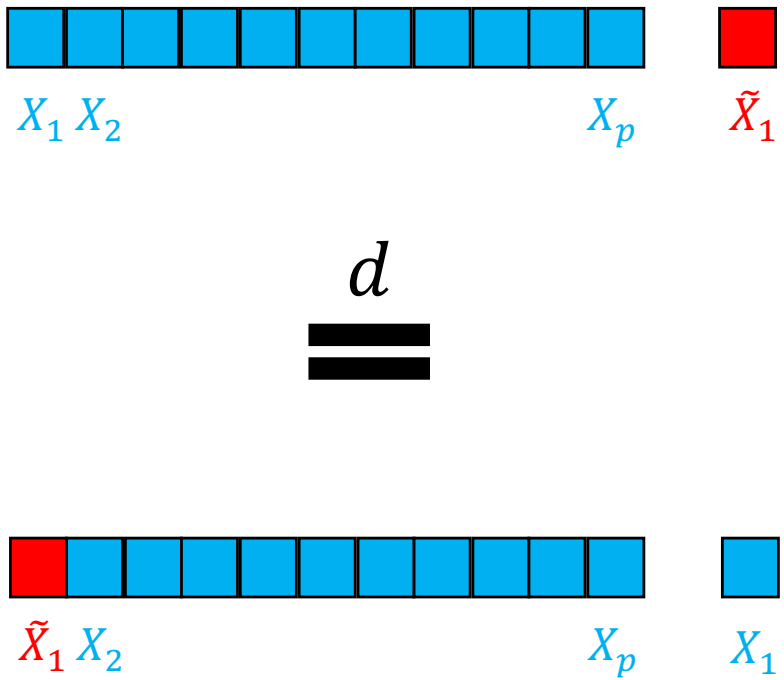


Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

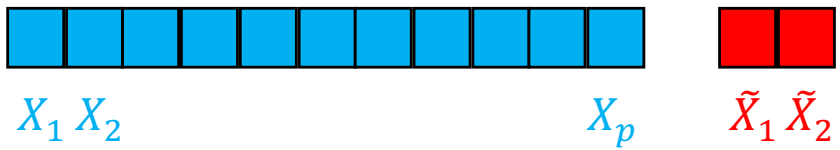


Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

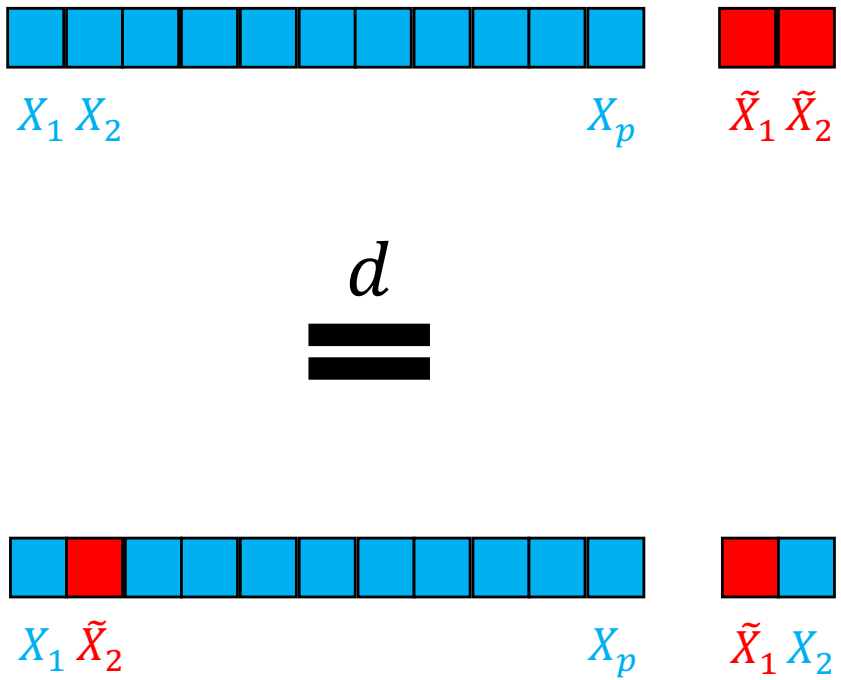


Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

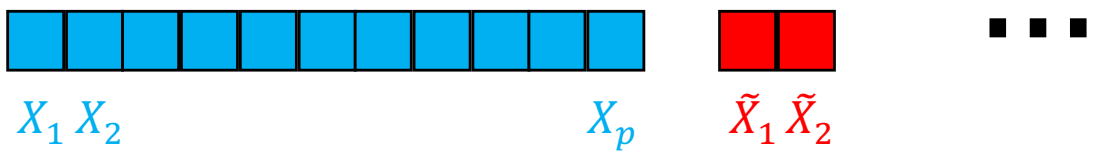


Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

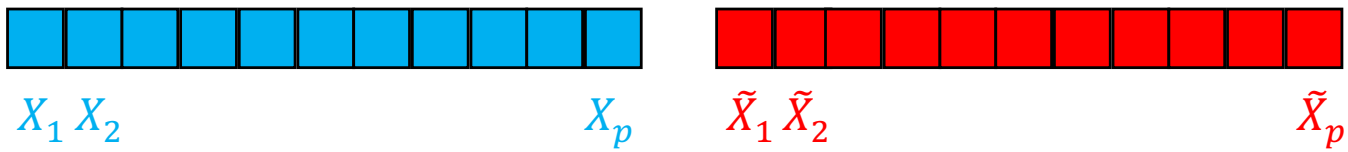


Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

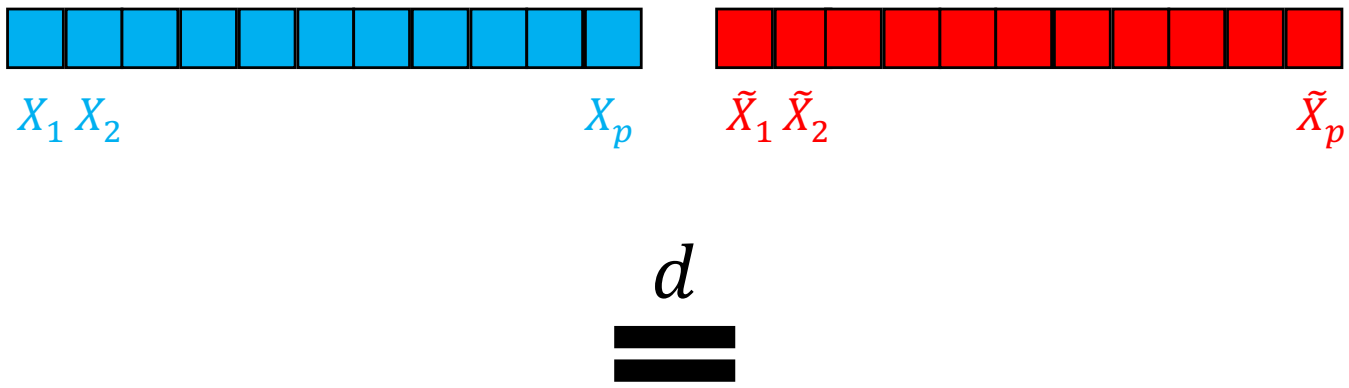


Sequential conditional exchangeable pairs (SCEP)

Bates, C., Janson & Wang, '19

Leverages ideas from:

- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling

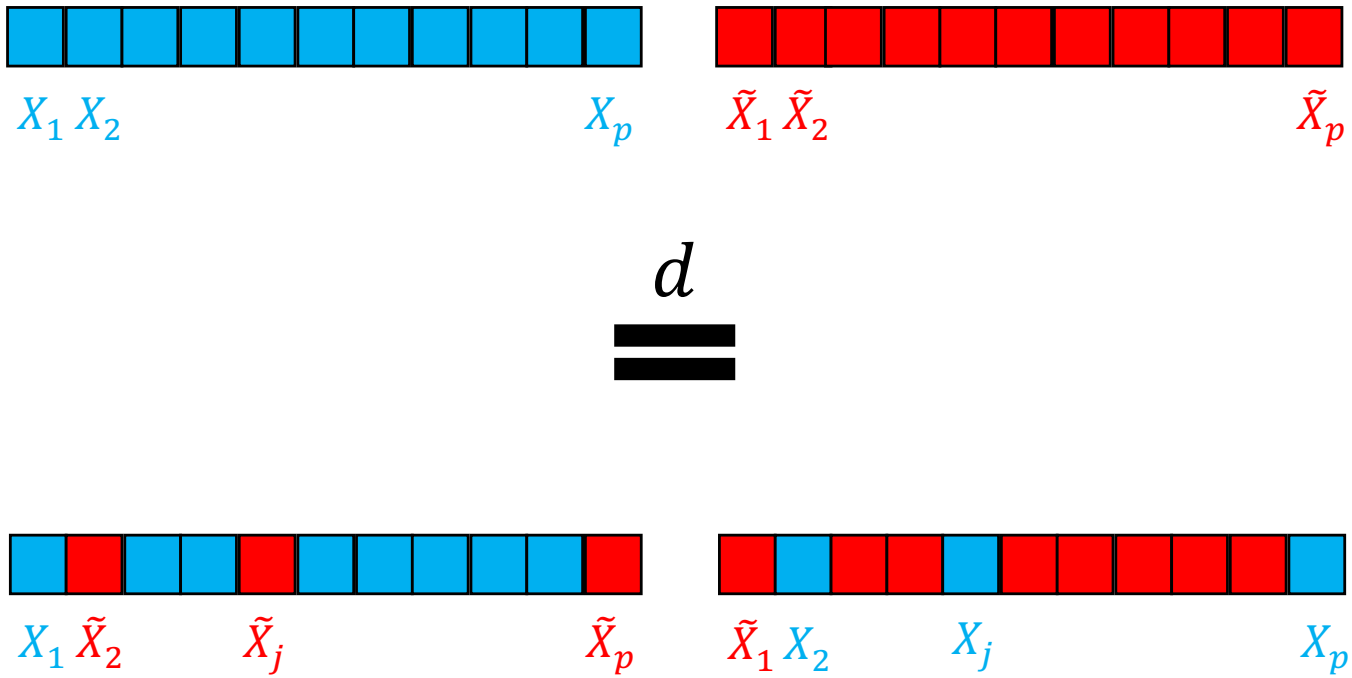


Sequential conditional exchangeable pairs (SCEP)

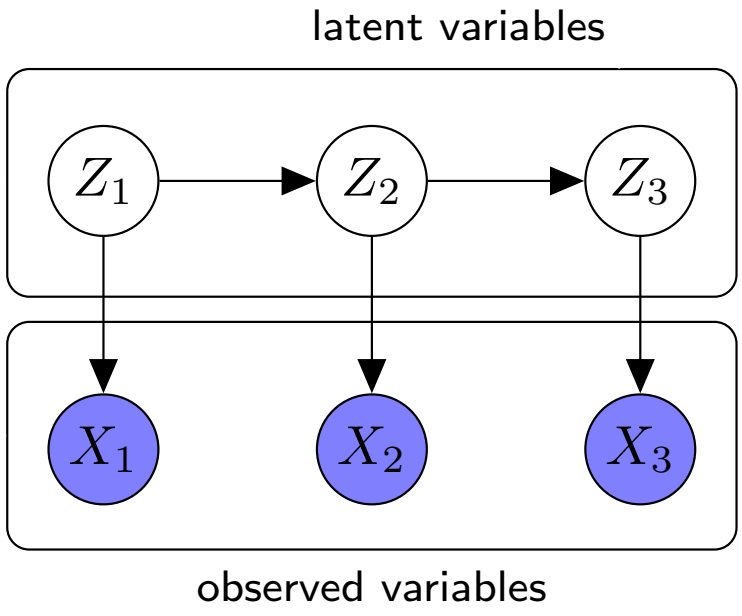
Bates, C., Janson & Wang, '19

Leverages ideas from:

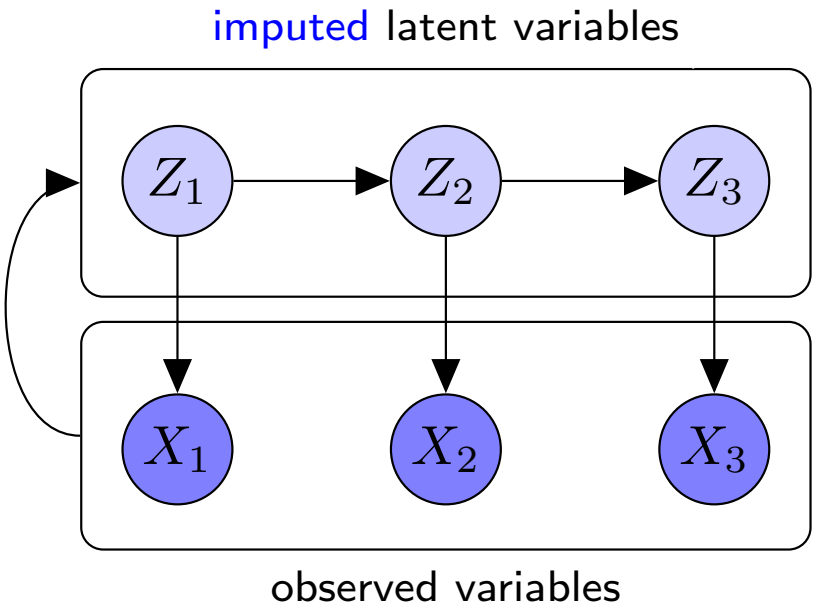
- MCMC: Metropolis-Hastings correction
- Importance sampling
- Graphical modeling



Example: hidden Markovs models (HMM)

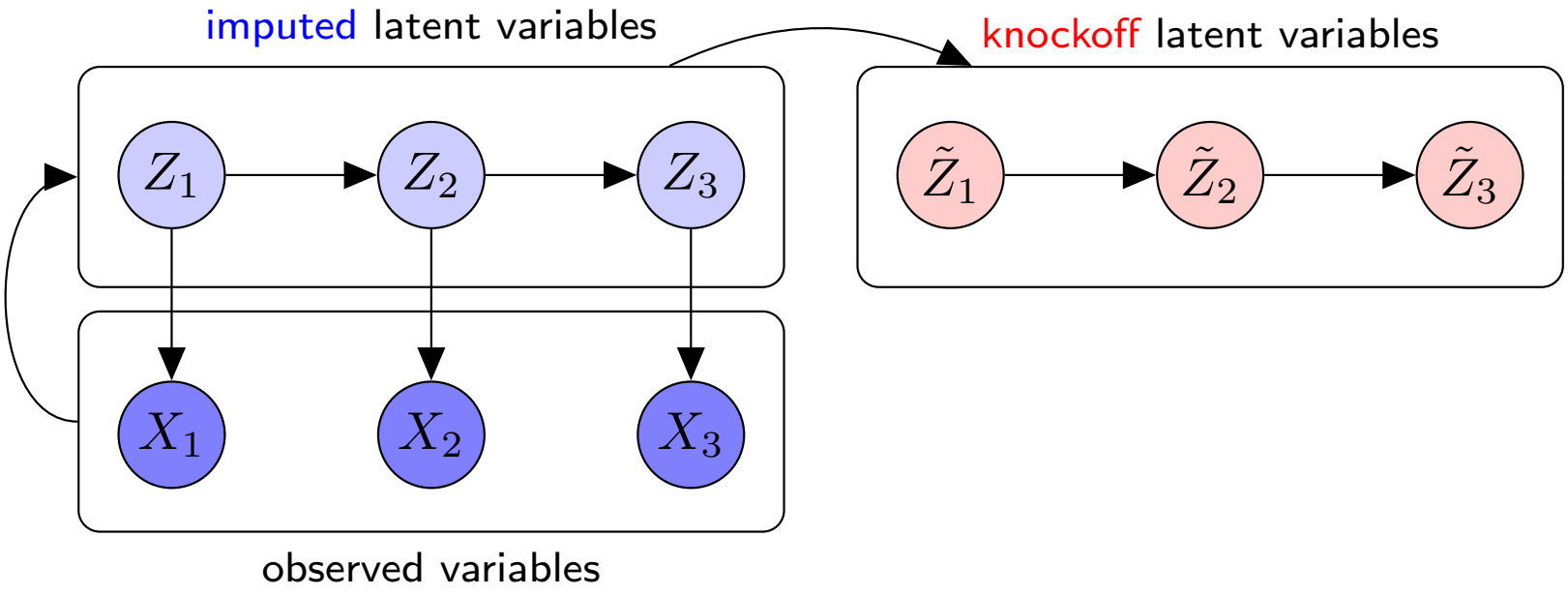


Example: hidden Markovs models (HMM)



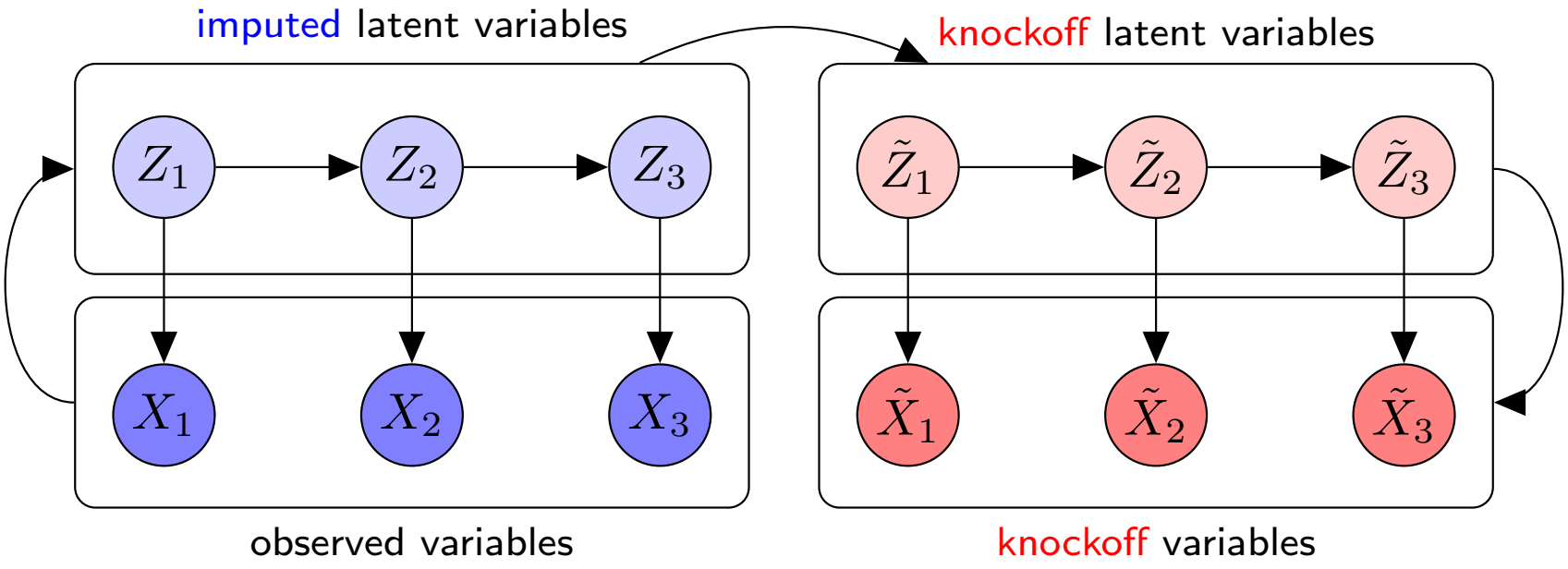
- Sample $Z \sim p(Z \mid X)$ (variation on Viterbi's algorithm)

Example: hidden Markovs models (HMM)



- Sample $Z \sim p(Z \mid X)$ (variation on Viterbi's algorithm)
- Sample $\tilde{Z}_j \sim p(Z_j \mid Z_{-j}, \tilde{Z}_{1:(j-1)})$ for $j = 1, \ldots, p$

Example: hidden Markovs models (HMM)



- Sample $Z \sim p(Z \mid X)$ (variation on Viterbi's algorithm)
- Sample $\tilde{Z}_j \sim p(Z_j \mid Z_{-j}, \tilde{Z}_{1:(j-1)})$ for $j = 1, \ldots, p$
- Sample $\tilde{X} \sim p(X \mid Z = \tilde{z})$ from emission probs.

Application to genetic data

C., Sabatti and Sesia ('17)

Haplotypes and genotypes well modeled by HMMs

Scheet ('06), Marchini ('07, '11), Li ('10), Browning ('10)

Application to genetic data

C., Sabatti and Sesia ('17)

Haplotypes and genotypes well modeled by HMMs

Scheet ('06), Marchini ('07, '11), Li ('10), Browning ('10)

Wellcome Trust Case Control Consortium

$\approx 5,000$ subjects and 400,000 SNPs

Response: Crohn's disease (CD)

Northern Finland 1966 Birth Cohort

$\approx 4,700$ subjects and 330,000 SNPs

Response: lipid levels

Dataset	Number of discoveries	
	Original study	Knockoffs (average)
CD	9	22.8
HDL	5	8
LDL	6	9.8

Nominal FDR level at 10%

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Franke et al. '10	WTCCC '07
100%	rs11209026 (2)	1	67.31–67.42	yes	yes
99%	rs6431654 (20)	2	233.94–234.11	yes	yes
98%	rs6688532 (33)	1	169.4–169.65		yes
97%	rs17234657 (1)	5	40.44–40.44	yes	yes
95%	rs11805303 (16)	1	67.31–67.46	yes	yes
91%	rs7095491 (18)	10	101.26–101.32	yes	yes
91%	rs3135503 (16)	16	49.28–49.36	yes	yes
81%	rs7768538 (1145)	6	25.19–32.91	yes	yes
80%	rs6601764 (1)	10	3.85–3.85		yes
75%	rs7655059 (5)	4	89.5–89.53		
73%	rs6500315 (4)	16	49.03–49.07	yes	yes
72%	rs2738758 (5)	20	61.71–61.82	yes	
70%	rs7726744 (46)	5	40.35–40.71	yes	yes
68%	rs11627513 (7)	14	96.61–96.63		
66%	rs4246045 (46)	5	150.07–150.41	yes	yes
62%	rs9783122 (234)	10	106.43–107.61		
61%	rs6825958 (3)	4	55.73–55.77		

Table: SNP clusters found to be important for CD over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
100%	rs1532085 (4)	15	58.68–58.7	yes	yes
100%	rs7499892 (1)	16	57.01–57.01	yes	yes
100%	rs1800961 (1)	20	43.04–43.04	yes	
99%	rs1532624 (2)	16	56.99–57.01	yes	yes
95%	rs255049 (142)	16	66.41–69.41	yes	yes

Table: SNP clusters found to be important for HDL over 100 repetitions of knockoffs.

Selection frequency	SNP (cluster size)	Chr.	Position range (Mb)	Confirmed in Willer et al. '13	Found in Sabatti et al. '09
99%	rs4844614 (34)	1	207.3–207.88		yes
97%	rs646776 (5)	1	109.8–109.82	yes	yes
97%	rs2228671 (2)	19	11.2–11.21	yes	yes
94%	rs157580 (4)	19	45.4–45.41	yes	yes
92%	rs557435 (21)	1	55.52–55.72	yes	
80%	rs10198175 (1)	2	21.13–21.13	yes	yes
76%	rs10953541 (58)	7	106.48–107.3		
62%	rs6575501 (1)	14	95.64–95.64		

Table: SNP clusters found to be important for LDL over 100 repetitions of knockoffs.

Deep knockoffs

Romano, Sesia & C. '18

P_X unknown? Repurpose deep generative models

- KnockoffsGAN

  Jordon, Yoon & van der Schaar '19

- Knockoffs via VAE

  Liu & Zheng '18

Deep knockoffs: overall view

Deep neural net f_θ

Draw knockoff $\tilde{X}^i$ for ith example

$$\tilde{X}^i = f_\theta\left(X^i, Z\right) \;\; Z \sim N(0, \mathbf{I})$$

$$(\mathbf{X}, \tilde{\mathbf{X}}) \overset{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})_{\text{Swap}(j)} \;\; ?$$

$$J_\theta(\mathbf{X}, \tilde{\mathbf{X}})$$

Measures distance to exchangeability

SGD

$$\theta \leftarrow \theta - \mu \nabla_\theta J_\theta(\mathbf{X}, \tilde{\mathbf{X}})$$

$$J_\theta(\mathbf{X}, \tilde{\mathbf{X}}) = \sum_{j=1}^{p} \mathcal{D}\left((\mathbf{X}, \tilde{\mathbf{X}}), (\mathbf{X}, \tilde{\mathbf{X}})_{\text{Swap}(j)}\right) + \delta \sum_{j=1}^{p} \left(\mathbf{X}_j^T \tilde{\mathbf{X}}_j\right)^2$$

Maximum Mean Discrepancy (MMD) [Gretton et al. ('12)]

- Classic two-sample problem: given **U** and **V**, test whether $P_U = P_V$
- Discrepancy measure ($\mathcal{H}$ is RKHS)

$$\mathcal{D}_\phi = \|\mathbb{E}_U[\phi(U)] - \mathbb{E}_V[\phi(V)]\|_{\mathcal{H}}^2$$

- E.g. $U, V \in \mathbb{R}$

$\phi(U) = U$ ➡ Distance between **means**

$\phi(U) = (U, U^2)$ ➡ Error in **first two moments**

- How to compare higher-order moments?

MMD & the 'kernel-trick' [Gretton et al. ('12)]

$$\mathcal{D}_\phi = \|\mathbb{E}_U[\phi(U)] - \mathbb{E}_V[\phi(V)]\|^2_{\mathcal{H}}$$

Expand quadratic and replace inner products with kernel operations

$$\mathrm{MMD}(P_U, P_V) = \mathbb{E}_{UU'}[\kappa(U, U')] - 2\mathbb{E}_{UV}[\kappa(U, V)] + \mathbb{E}_{VV'}[\kappa(V, V')]$$

Characteristic kernel, e.g. Gaussian, implies $\mathrm{MMD} = 0$ iff $P_U = P_V$

MMD & the 'kernel-trick' [Gretton et al. ('12)]

$$\mathcal{D}_\phi = \|\mathbb{E}_U[\phi(U)] - \mathbb{E}_V[\phi(V)]\|^2_{\mathcal{H}}$$

Expand quadratic and replace inner products with kernel operations

$$\mathrm{MMD}(P_U, P_V) = \mathbb{E}_{UU'}[\kappa(U, U')] - 2\mathbb{E}_{UV}[\kappa(U, V)] + \mathbb{E}_{VV'}[\kappa(V, V')]$$

Characteristic kernel, e.g. Gaussian, implies $\mathrm{MMD} = 0$ iff $P_U = P_V$

Unbiased estimate

$$\widehat{\mathrm{MMD}}(\mathbf{U}, \mathbf{V}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \kappa(U^i, U^j) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \kappa(U^i, V^j) + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \kappa(V^i, V^j)$$

Optimization: stochastic gradient descent

Generate knockoffs

Evaluate $\tilde{X}^i = f_{\theta_t}(X^i, Z^i)$ for each example. The network f_{θ_t} is fixed

Evaluate loss

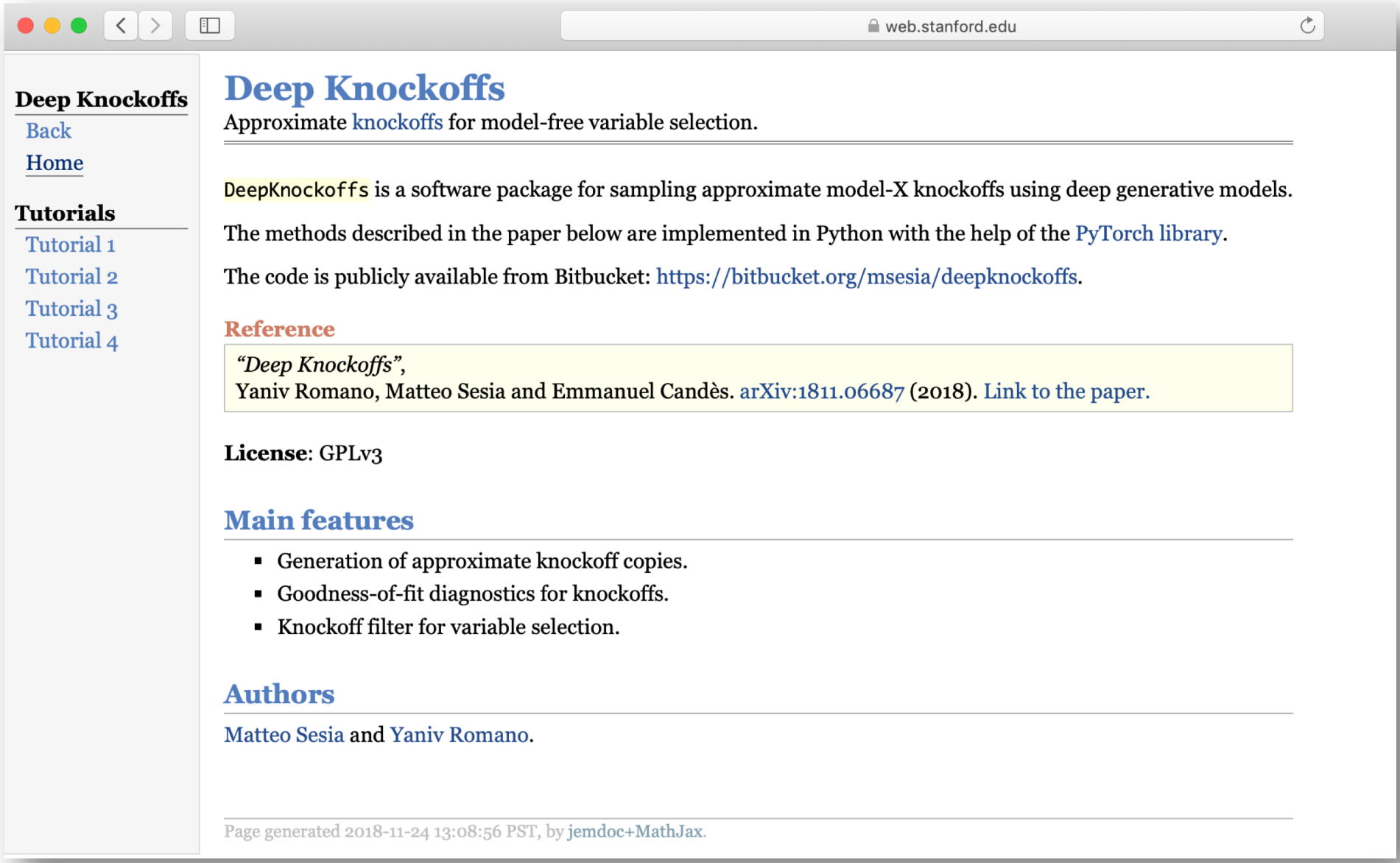
$$J_{\theta_t}(\mathbf{X}_t, \tilde{\mathbf{X}}_t) = \sum_{j=1}^{p} \widehat{\mathrm{MMD}}\left((\mathbf{X}, \tilde{\mathbf{X}}), (\mathbf{X}, \tilde{\mathbf{X}})_{\mathrm{Swap}(j)}\right) + \delta \sum_{j=1}^{p} \left(\mathbf{X}_j^T \tilde{\mathbf{X}}_j\right)^2$$

Update parameters

$$\theta_{t+1} \leftarrow \theta_t - \mu \nabla_{\theta_t} J_{\theta_t}(\mathbf{X}_t, \tilde{\mathbf{X}}_t)$$

- Mini-batch SGD
- Random swaps
- Evaluate MMD on disjoint subsets of samples

Software tools



Deep Knockoffs

Approximate knockoffs for model-free variable selection.

Numerical experiments I (training)

Notebook written by Matteo Sesia and Yaniv Romano

Stanford University, Department of Statistics

Last updated on: November 19, 2018

The purpose of this notebook is to allow the numerical experiments described in the paper to be reproduced easily. Running this code may take a few hours on a graphical graphical processing unit.

Load the required libraries

In [1]:
```
import numpy as np
from DeepKnockoffs import KnockoffMachine
from DeepKnockoffs import GaussianKnockoffs
import data
import parameters
```

Data generating m

We model $X \in \mathbb{R}^p$ as a multiv
correlation parameter for this c

In [2]:
```
# Number of features
p = 100

# Load the built-in mu
# The currently availa
# - gaussian : Multiva
# - gmm      : Gaussia
# - mstudent : Multiva
# - sparse   : Multiva
model = "mstudent"
distribution_params = p

# Initialize the data g
```

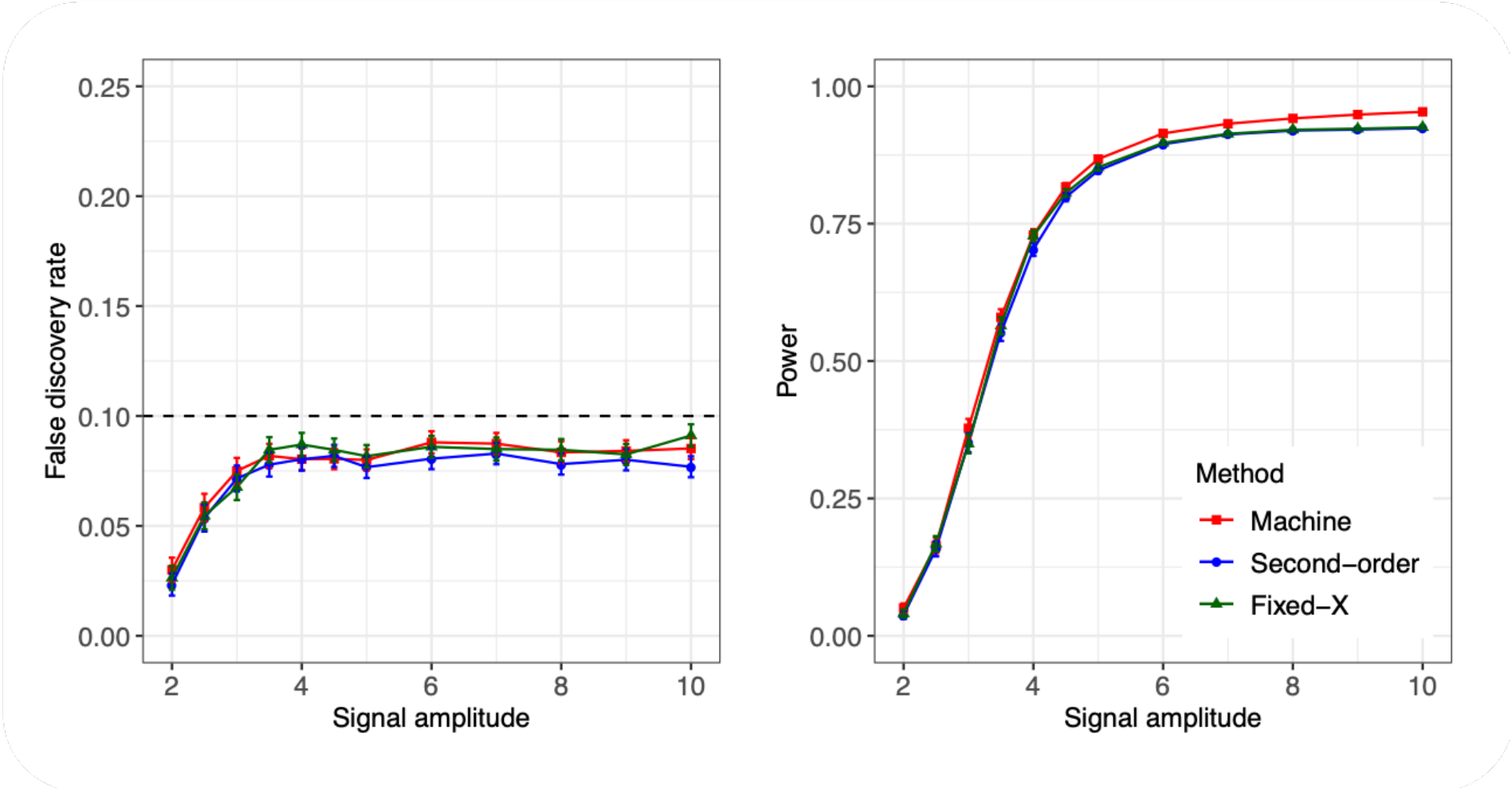
In [8]:
```
# Train the machine
print("Fitting the knockoff machine...")
machine.train(X_train)
```

```
Fitting the knockoff machine...
[   1/ 100], Loss: 0.1876, MMD: 0.1726, Cov: 1.289, Decorr: 0.299
[   2/ 100], Loss: 0.1454, MMD: 0.1366, Cov: 0.927, Decorr: 0.441
[   3/ 100], Loss: 0.1332, MMD: 0.1265, Cov: 0.786, Decorr: 0.502
[   4/ 100], Loss: 0.1294, MMD: 0.1236, Cov: 0.730, Decorr: 0.536
[   5/ 100], Loss: 0.1272, MMD: 0.1220, Cov: 0.702, Decorr: 0.555
[   6/ 100], Loss: 0.1260, MMD: 0.1212, Cov: 0.671, Decorr: 0.564
[   7/ 100], Loss: 0.1253, MMD: 0.1207, Cov: 0.694, Decorr: 0.567
[   8/ 100], Loss: 0.1247, MMD: 0.1204, Cov: 0.680, Decorr: 0.569
[   9/ 100], Loss: 0.1238, MMD: 0.1198, Cov: 0.668, Decorr: 0.569
[  10/ 100], Loss: 0.1233, MMD: 0.1195, Cov: 0.650, Decorr: 0.572
[  11/ 100], Loss: 0.1230, MMD: 0.1194, Cov: 0.651, Decorr: 0.570
[  12/ 100], Loss: 0.1225, MMD: 0.1191, Cov: 0.647, Decorr: 0.572
[  13/ 100], Loss: 0.1224, MMD: 0.1190, Cov: 0.623, Decorr: 0.570
[  14/ 100], Loss: 0.1221, MMD: 0.1189, Cov: 0.633, Decorr: 0.567
```

HIV Drug Resistance

- Detect mutations in HIV associated with drug resistance (to protease inhibitors)
- Y: log-fold-increase of lab-tested drug resistance
- X_j: presence or absence of mutation $\#j$
- $n = 1431$, $p = 150$

Real X with simulated Y: FDR and Power

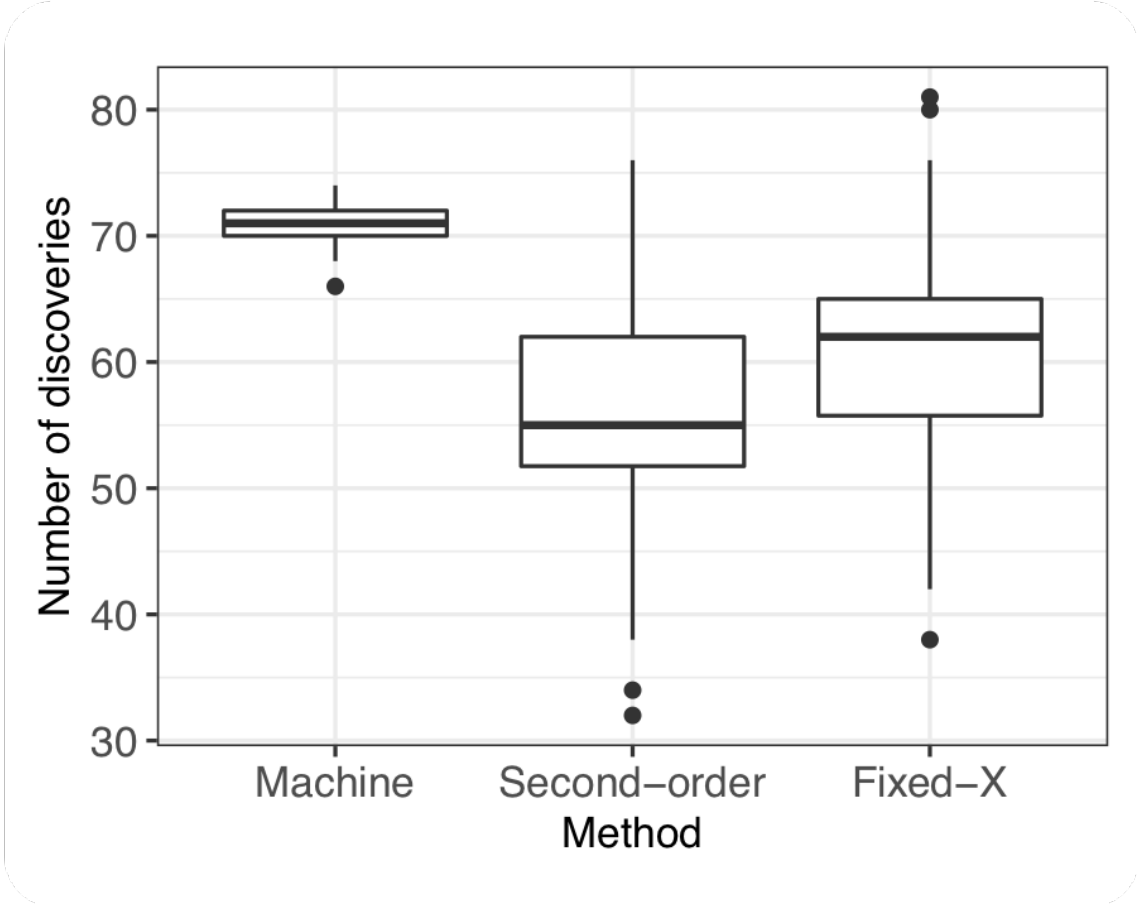


X fixed

Resample Y

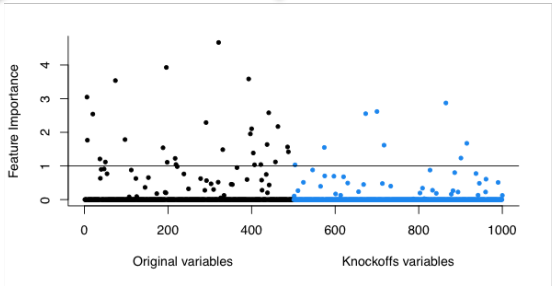
Resample $\tilde{X}$

Real data example



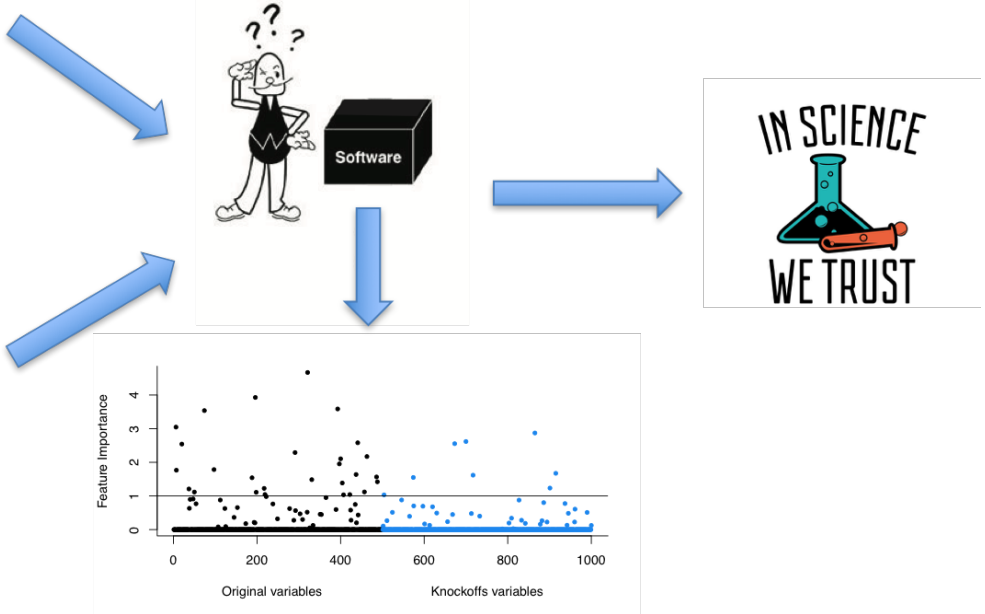
Method selects variables that mostly correspond to real (replicable) effects

Summary and challenges



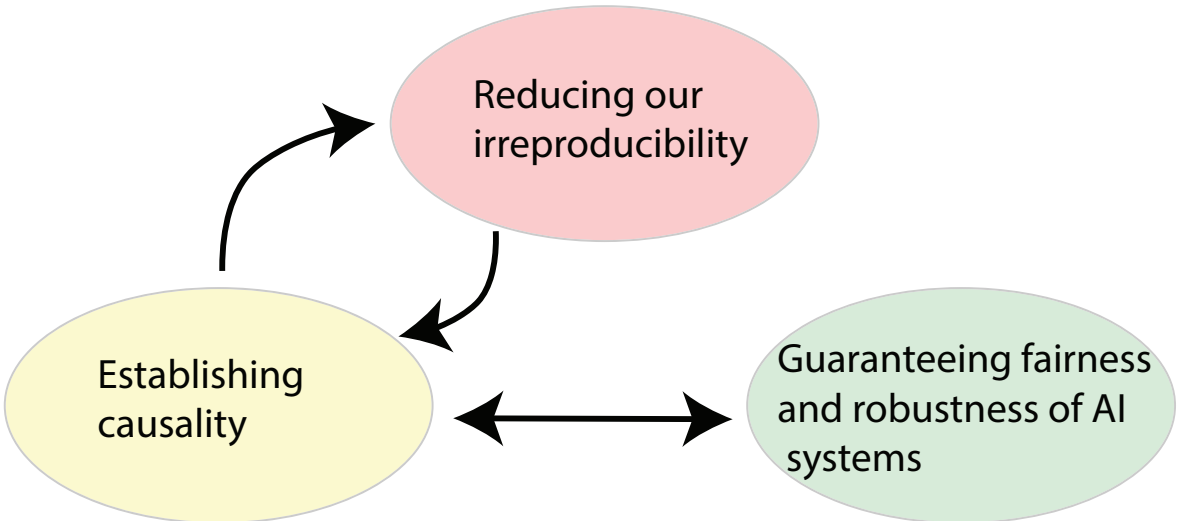
'Wrapper' around black-box algorithm rigorously addresses reproducibility issue

Summary and challenges

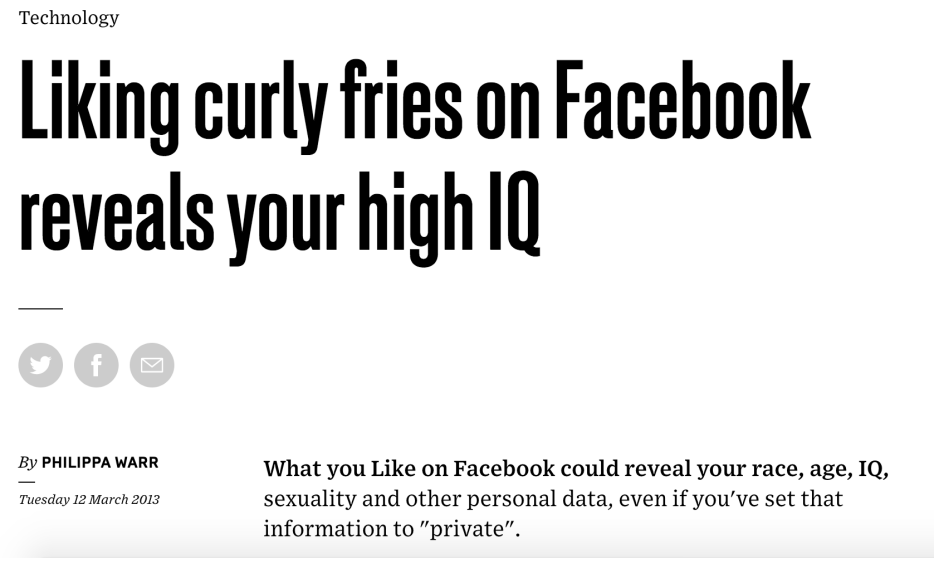


'Wrapper' around black-box algorithm rigorously addresses reproducibility issue

Other important things to think about



This is not just about not being wrong (irreproducibility)



Technology

Liking curly fries on Facebook reveals your high IQ

By **PHILIPPA WARR**

Tuesday 12 March 2013

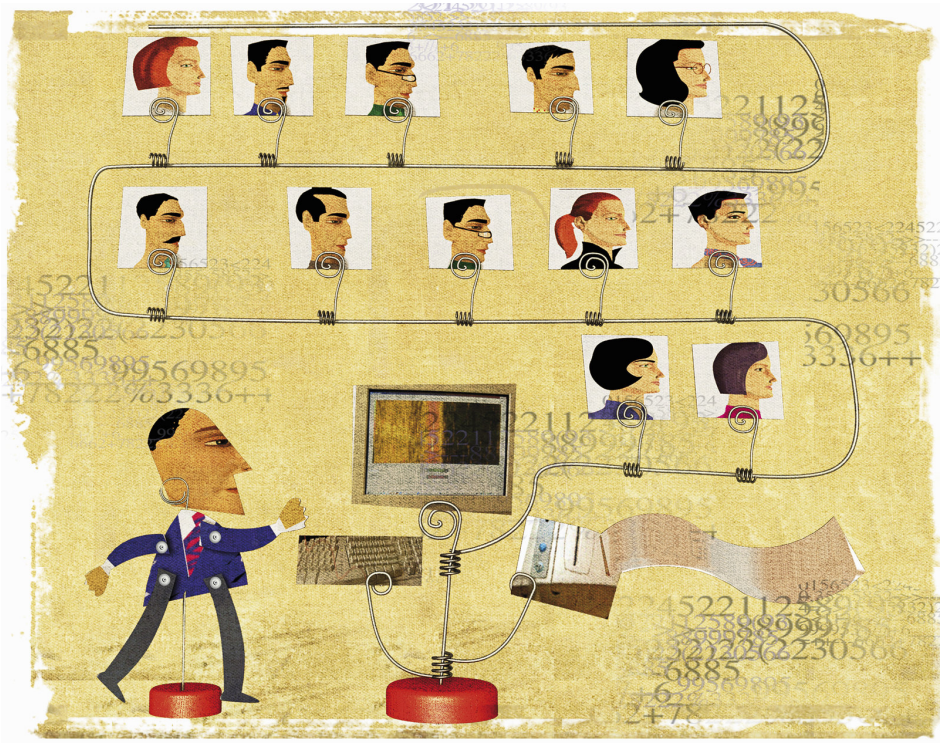
What you Like on Facebook could reveal your race, age, IQ, sexuality and other personal data, even if you've set that information to "private".

Robustness?

Would want predictions to be valid in different samples collected in different circumstances

"Constant conjunction" is a property of causal effects (Hume)

Fairness: can computer programs be racist and sexist?



Guido Rosa/Getty Images/Ikon Images

Blind application of machine learning runs risk of amplifying biases and prejudices

Identifying variables $\rightsquigarrow$ chance to scrutinize model built from one sample:

- Do we believe these variables are "structurally" important, or are they just reflecting a spurious association in this sample?
- Are we learning something about the world or reifying our prejudices?