



Data 102: Data, Inference, and Decisions

Lecture 1

Michael Jordan

University of California, Berkeley

DS 102 team this semester



Jacob
Steinhardt



Mihaela
Curmei



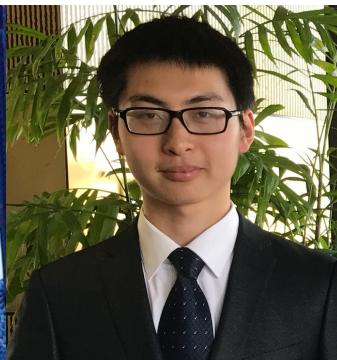
Jake
Soloff



Yimeng
Wang



Clara Wong-
Fannjiang



Banghua
Zhu

Building on work by Moritz Hardt, Fernando Perez, and the Fall 2019 and Spring 2020 teams.

Announcements

All class discussion on Piazza. Please be respectful and community-oriented.

TAs and faculty will not answer questions by email. Available via Piazza/Labs/Office hours.

Enrollment cap of 250 is firm. Instructors cannot change anything about that.

There is no course textbook, and lectures are of particular importance. (Watch them twice!)

Email ds-enrollments@berkeley.edu for enrollment related questions.

Data Science: A Personal Perspective

- It arose both in science and in technology, over many decades
 - in science as the “fourth paradigm” or “data-intensive science”
 - in technology as new business models based on data flows and data analysis
 - in my view, it’s the latter which is the more strikingly era-defining phenomenon
- Historical points of reference: the development of chemical engineering, civil engineering, and electrical engineering
 - in each case, real-world use cases combined with basic capabilities led to general principles and eventually to academic disciplines
 - we’re currently in the early days of the development of a new, human-centric form of engineering
- Engineering means “real-world systems that work and deliver value to humans”
 - we have a lot of work to do to realize that promise in this new emerging field

Still Further Perspective

- How does “Data Science” relate to “Machine Learning” and to “Artificial Intelligence”?
- The phrase “Machine Learning” arose in the early 1980’s
 - the idea was that instead of programming computers, we would let them learn from experience, somewhat like humans and animals
 - the actual methods and concepts developed in the field are clearly related to, if not identical to, those of statistical inference and decision theory
- I think of Machine Learning as the engineering side of Statistics (treating “engineering” with reverence)
 - on the next slide, see my industry-centric history of Machine Learning

Machine Learning in Industry

- First Generation ('90-'00): the **backend**
 - e.g., fraud detection, search, supply-chain management
- Second Generation ('00-'10): the **human side**
 - e.g., recommendation systems, commerce, social media
- Third Generation ('10-now): **pattern recognition**
 - e.g., speech recognition, computer vision, translation

Machine Learning in Industry

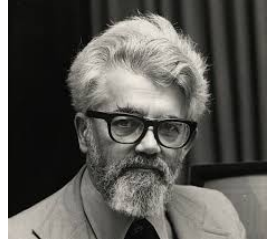
- First Generation ('90-'00): the **backend**
 - e.g., fraud detection, search, supply-chain management
- Second Generation ('00-'10): the **human side**
 - e.g., recommendation systems, commerce, social media
- Third Generation ('10-now): **pattern recognition**
 - e.g., speech recognition, computer vision, translation
- Fourth Generation (emerging): **markets**
 - not just one agent making a decision or sequence of decisions
 - but a huge interconnected web of data, agents, decisions
 - many new challenges!

Machine Learning in Industry

- First Generation ('90-'00): the **backend**
 - e.g., fraud detection, search, supply-chain management
- Second Generation ('00-'10): the **human side**
 - e.g., recommendation systems, commerce, social media
- Third Generation ('10-now): **pattern recognition**
 - e.g., speech recognition, computer vision, translation
- Fourth Generation (emerging): **markets**
 - not just one agent making a decision or sequence of decisions
 - but a huge interconnected web of data, agents, decisions
 - many new challenges!

- What about “AI”?

Perspectives on AI*

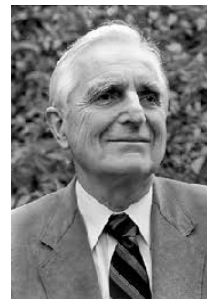
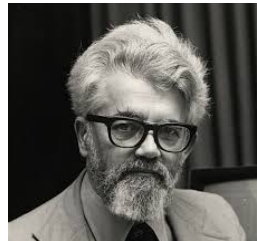


- The classical “human-imitative” aspiration

*M. I. Jordan, Artificial Intelligence: The Revolution Hasn't Happened Yet, *Medium*, 2019

Perspectives on AI*

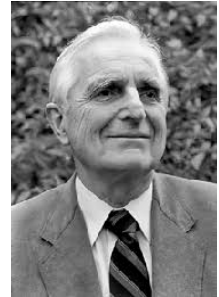
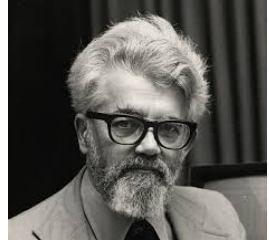
- The classical “human-imitative” aspiration
- The “intelligence augmentation” (IA) perspective



*M. I. Jordan, Artificial Intelligence: The Revolution Hasn't Happened Yet, *Medium*, 2019

Perspectives on AI*

- The classical “human-imitative” aspiration
- The “intelligence augmentation” (IA) perspective
- The “intelligent infrastructure” (II) perspective



*M. I. Jordan, Artificial Intelligence: The Revolution Hasn't Happened Yet, *Medium*, 2019

Pattern Recognition

- The third generation of Machine Learning has focused on supervised learning (aka, **classification** and **regression**)
 - labeled training data are used to train huge neural networks, via some form of gradient descent
 - this has traditionally been called **pattern recognition**
- This has been a major success, yielding human-level performance in speech recognition and computer vision
 - and has yielded super-human-level performance in some tasks
- Pattern recognition has become a commodity
 - companies are springing up worldwide to hire humans to provide labels for all kinds of data, transferring some aspects of human pattern recognition skill to computers

Decision Making

- Is pattern recognition (or classification/regression) all there is?
- The overall goal of a learning system is generally to make a decision of some kind
- Is decision making merely a matter of setting an appropriate threshold on the output of a neural network, if the training data is good enough?
- Let's do a thought experiment

A Visit to the Doctor's Office

- Consider a medical checkup in the not-too-distant future, where the doctor measures thousands of physiological variables and even obtains your genome
- This massive data vector is then input to a massive neural network, which has been trained to predict disease
- Suppose that one of the outputs has been trained to predict kidney failure, with a value over 0.7 suggesting an imminent failure
- Your value is 0.701
- The neural network has “decided” that you’re in trouble—what do you actually do?

A Visit to the Doctor's Office

- You will probably want to engage in a dialog, hopefully with a human but perhaps with the machine
- You will want to know:
 - what are the error bars on that 0.701?
 - what kind of uncertainty is being captured by those error bars?
 - what is the provenance of the data; i.e., what subset of humans was it taken from, on what measuring devices, how long ago, and under what conditions?
 - given this provenance, how relevant is that prediction of 0.701 to me?
- You will want to ask things like:
 - are you aware of certain facts about my history, my family, etc?
 - what if I were to exercise more, eat better, etc?
 - what are my treatment options, what are their costs, etc?
 - can I get a second opinion?

Decisions and Context

- Real-world decisions with consequences
 - counterfactuals, provenance, relevance, causal inference, dialog
- Sets of decisions across a network
 - false-discovery rate (instead of sensitivity/specificity/accuracy)
- Sets of decisions across a network over time
 - streaming, asynchronous decisions
- Decisions when there is scarcity and competition
 - need for an economic perspective
- Decisions which affect future data and future decisions
 - need for a dynamical-systems, control-theoretic perspective
- Decisions when there are consequences for others
 - need for an ethical perspective



“[T]echnologies are developed and used within a particular social, economic, and political context. They arise out of a social structure, they are grafted on to it, and they may reinforce it or destroy it, often in ways that are neither foreseen nor foreseeable.”

Ursula Franklin, 1989



“[C]ontext is not a passive medium but a dynamic counterpart. The responses of people, individually, and collectively, and the responses of nature are often underrated in the formulation of plans and predictions.”

Ursula Franklin, 1989

Data 102

- Unusually, this course will focus more on decision making and less on pattern recognition
- Decision-oriented topics that we'll cover include **false-discovery rate control**, **bandit algorithms**, **causal inference**, **experimental design**, **matching markets**, and **reinforcement learning**
- This set of topics spills over from statistics into economics, game theory, and control theory
- It will include discussion of incentives and unwanted outcomes

Early example of dynamics in decision making



In 1696, England's King William III seeks to tax wealth, but how to know one's wealth?

Introduces tax based on **number of windows**

Idea spreads to France, Spain, Scotland

People adapt



One row of houses in Edinburgh featured no bedroom windows at all.

Tax revenues fell

Goodhart's law

“Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.” - Charles Goodhart, 1975

Related:

Lucas critique 1976 in macroeconomics

Campbell's law 1979 in social sciences

Learning invites gaming

- Correlation is all you need for prediction
- Typically lots of features
- Features often easy to change

Behavior Revealed in Mobile Phone Usage

Predicts Credit Repayment

Daniel Björkegren¹ and Darrell Grissen²

Number of outgoing calls
Text response rate
Average airtime balance
Entropy of GPS coordinates

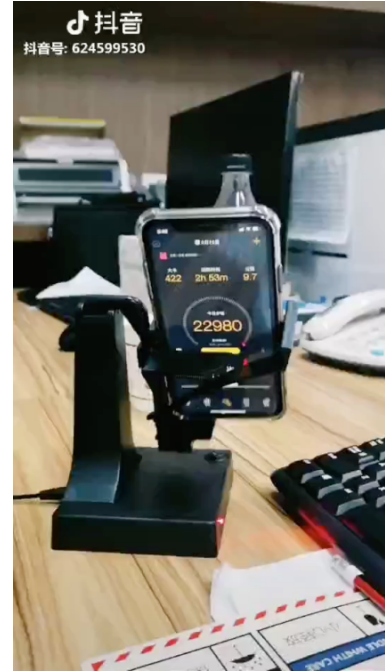
What behavior do our decisions *incentivize*?

Get moving.

Start a healthy habit in the new year and **set some activity goals** to get moving. Want to know the fastest way to get in shape? Walk more. **Studies** have shown that low-intensity cardiovascular exercise increases fat burn. It also produces endorphins, natural mood elevators that make you feel good! Whether you enjoy a retro speed-walk with your friends or are training for the NYC Marathon, we recognize that a little healthy competition is good! **Download the Oscar app to your phone and sync Apple Health or Google Fit to track your steps and earn \$1 a day in Amazon® Gift Card rewards when you meet your step goals.**



But there are two ways of going about it



Basics of Decision Making

- We'll start by considering the most simple of decision-making formulations
- Let's suppose that **Reality** is in one of two states, which we denote as 0 or 1
- We don't observe this state, but we do obtain **Data** that is drawn from a distribution that depends whether the state is 0 or 1
- We make a **Decision** based on the Data, which we denote as 0 or 1
- We can think of the Decision as our best guess as to the state of Reality or, more generally, as an action we think is best given our guess of the state of Reality
- Example: **COVID-19 testing**

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0		
	1		

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

TN = True Negative

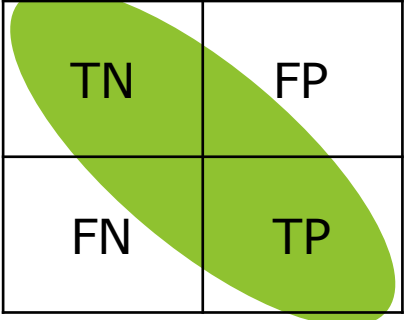
FP = False Positive

FN = False Negative

TP = True Positive

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

A 2x2 confusion matrix diagram. The vertical axis is labeled 'Reality' with values 0 and 1. The horizontal axis is labeled 'Decision' with values 0 and 1. The four quadrants are labeled: top-left is 'TN', top-right is 'FP', bottom-left is 'FN', and bottom-right is 'TP'. A green diagonal highlight covers the TN and TP cells.

TN = True Negative

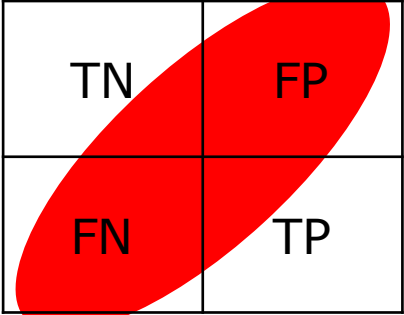
FP = False Positive

FN = False Negative

TP = True Positive

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

A 2x2 confusion matrix diagram. The vertical axis is labeled 'Reality' with values 0 and 1. The horizontal axis is labeled 'Decision' with values 0 and 1. The four quadrants are labeled: Top-Left (Reality 0, Decision 0) is 'TN'; Top-Right (Reality 0, Decision 1) is 'FP'; Bottom-Left (Reality 1, Decision 0) is 'FN'; Bottom-Right (Reality 1, Decision 1) is 'TP'. A red oval highlights the diagonal elements, TN and TP.

TN = True Negative

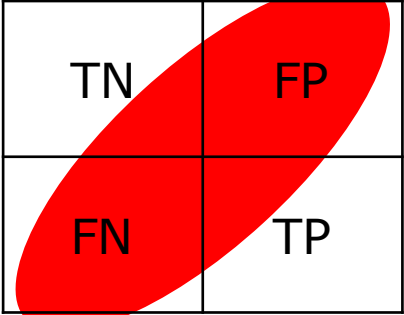
FP = False Positive

FN = False Negative

TP = True Positive

The Basic Two-by-Two Table

		Decision	
		0	1
Reality	0	TN	FP
	1	FN	TP

A 2x2 confusion matrix with a red diagonal highlight. The matrix is labeled with 'Reality' on the y-axis and 'Decision' on the x-axis. The y-axis has values 0 and 1, and the x-axis has values 0 and 1. The cells contain 'TN' (True Negative) at (0,0), 'FP' (False Positive) at (0,1), 'FN' (False Negative) at (1,0), and 'TP' (True Positive) at (1,1). A red oval highlights the diagonal cells (TN and TP).

TN = True Negative

FP = False Positive

FN = False Negative

TP = True Positive

Rough goal: lots of green outcomes, few red outcomes!

Examples: How Serious are FP and FN (and How Desirable are TP and TN)?

- Medical: 0 = no disease, 1 = disease
- Commerce: 0 = no fraud, 1 = fraud
- Physics: 0 = no Higgs boson, 1 = Higgs boson
- Social network: 0 = no link, 1 = link
- Self-driving car: 0 = no pedestrian, 1 = pedestrian
- Search: 0 = not relevant, 1 = relevant
- Oil-Well Drilling: 0 = no oil, 1 = oil

Examples: How Serious are FP and FN (and How Desirable are TP and TN)?

- Medical: 0 = no disease, 1 = disease
- Commerce: 0 = no fraud, 1 = fraud
- Physics: 0 = no Higgs boson, 1 = Higgs boson
- Social network: 0 = no link, 1 = link
- Self-driving car: 0 = no pedestrian, 1 = pedestrian
- Search: 0 = not relevant, 1 = relevant
- Oil-Well Drilling: 0 = no oil, 1 = oil

- In real-world domains, there are many, many complications that arise

Towards a Statistical Framework

- Although the two-by-two table is useful conceptually, it's not clear how to make use of it in a real problem, because we don't know Reality
- We need to move towards a statistical framework, where we consider not just one decision, but a **set of related decisions**

Towards a Statistical Framework

- Let's now imagine that we not only make a decision, but we build a **decision-making algorithm**
- We want to evaluate the algorithm not just on one problem, but on a set of related problems

Towards a Statistical Framework

- Let's now imagine that we not only make a decision, but we build a **decision-making algorithm**
- We want to evaluate the algorithm not just on one problem, but on a set of related problems
- Concretely, we may have a collection of hypothesis-testing problems, where we repeatedly decide whether to accept the null or accept the alternative
- Or we may have a set of classification decisions, where we repeatedly classify data points into one of two classes

Towards a Statistical Framework

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$N = n_{00} + n_{01} + n_{10} + n_{11}$$

Towards a Statistical Framework

- Our language will start to involve **rates** and **probabilities**
- Indeed, the variables n_{00} , n_{01} , n_{10} , and n_{11} are **random variables**
- In just what sense they are random will need to be made clear (e.g., is the state of Reality random, is the Decision random, is N random?)

Some Row-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{sensitivity} = \frac{n_{11}}{n_{10} + n_{11}}$$

Some Row-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{sensitivity} = \frac{n_{11}}{n_{10} + n_{11}}$$

aka, "true positive rate"
or "recall" or "power"

Some Row-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{specificity} = \frac{n_{00}}{n_{00} + n_{01}}$$

Some Row-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{specificity} = \frac{n_{00}}{n_{00} + n_{01}}$$

aka, "true negative rate"
or "selectivity"

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
 - e.g., sensitivity approximates $P(\text{Decision} = 1 \mid \text{Reality} = 1)$

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
 - e.g., sensitivity approximates $P(\text{Decision} = 1 \mid \text{Reality} = 1)$
- As such, they are not dependent on the **prevalence** (i.e., the probabilities of the two states of Reality in the population)

Comments on the Row-Wise Rates

- They can be thought of as estimates of conditional probabilities
 - e.g., sensitivity approximates $P(\text{Decision} = 1 \mid \text{Reality} = 1)$
- As such, they are not dependent on the **prevalence** (i.e., the probabilities of the two states of Reality in the population)
- They are the kinds of quantities that are the focus of Neyman-Pearson inferential theory, which we'll review in a moment
 - specificity = $1 - \text{Type I error rate}$
 - sensitivity = $1 - \text{Type II error rate} = \text{power}$

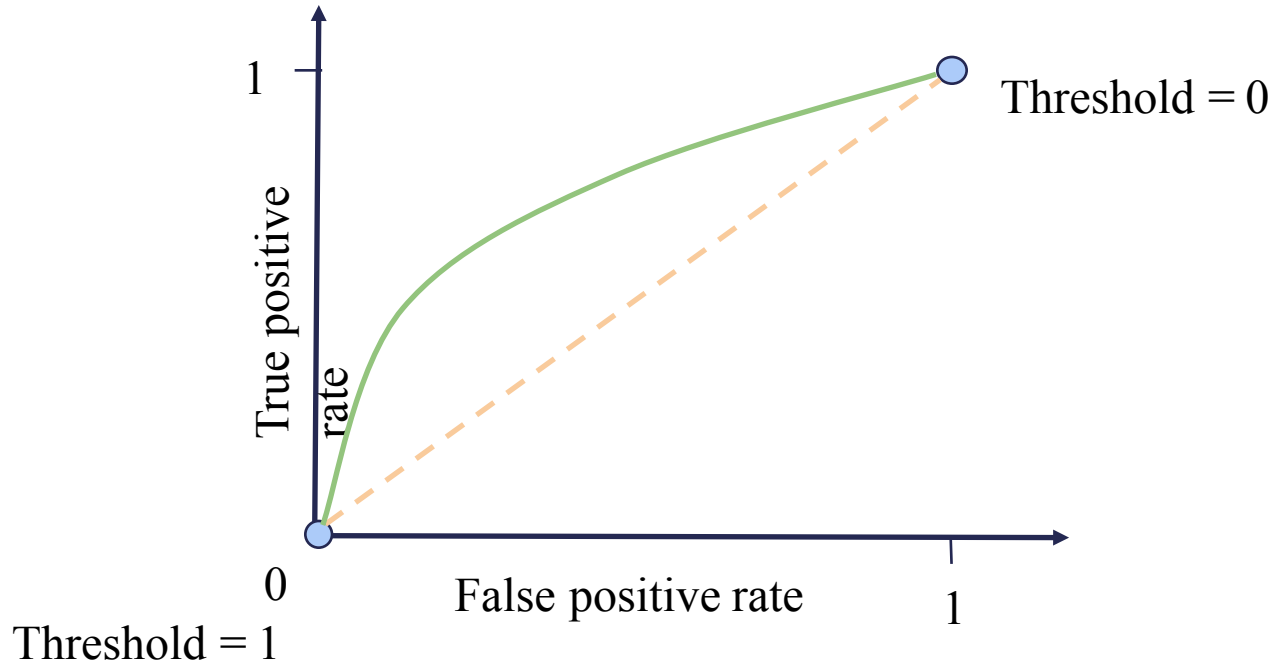
Towards Inference

- We'd like to have high sensitivity and high specificity
 - but in general there is a tradeoff (see whiteboard drawings)
 - we have to figure out how to manage the tradeoff

Towards Inference

- We'd like to have high sensitivity and high specificity
 - but in general there is a tradeoff (see whiteboard drawings)
 - we have to figure out how to manage the tradeoff
- Neyman and Pearson (1932) formulated this problem as a **constrained optimization problem**:
 - maximize the sensitivity while constraining the specificity to be more than some fixed number (e.g., .95)
 - i.e., maximize the power while constraining the false-positive rate to be less than some fixed number (e.g., .05)
 - we're neglecting the distinction between rates and probabilities here; we'll be more clear on this later

A Tradeoff Curve (the “ROC curve”)



The Neyman-Pearson Formulation (1932)

- Turn the problem into a **constrained optimization problem**:
 - maximize the power while constraining the false-positive rate to be under some fixed number (e.g., .05)
- A very fruitful idea, and sometimes the right idea, but not to be viewed as written in stone

Some Column-Wise Rates

		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{false omission rate} = \frac{n_{10}}{n_{00} + n_{10}}$$

Some Column-Wise Rates

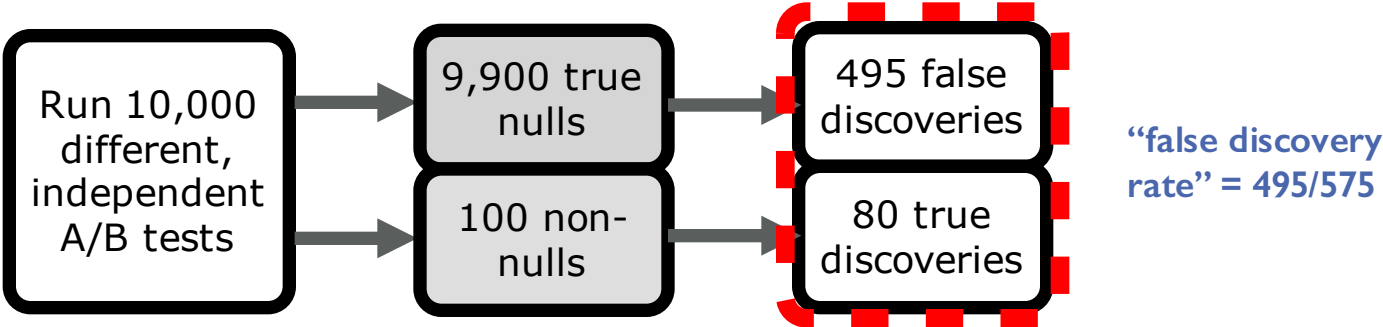
		Decision	
		0	1
Reality	0	n_{00}	n_{01}
	1	n_{10}	n_{11}

$$\text{false discovery rate} = \frac{n_{01}}{n_{01} + n_{11}}$$

Comments on the Column-Wise Rates

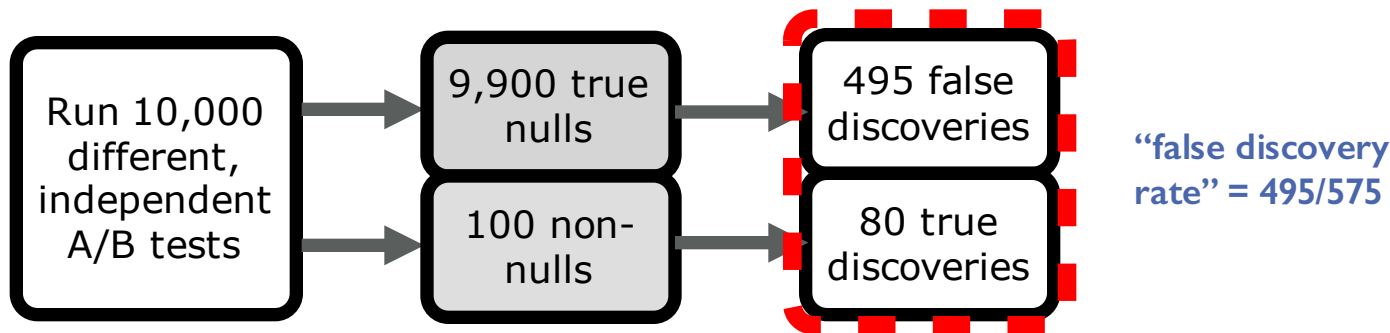
- They can be thought of as estimates of conditional probabilities
 - e.g., false discovery rate approximates $P(\text{Reality} = 0 \mid \text{Decision} = 1)$
- They **are** dependent on the **prevalence** (i.e., the probabilities of the two states of Reality in the population), via Bayes' Theorem
 - as such, they are more Bayesian
- This is arguably a good thing, as we'll see on the next slide

Type I error rate (per test) = 0.05



Power (per test) = 0.80

Type I error rate (per test) = 0.05



Power (per test) = 0.80

(NB: We're again not being rigorous at this point; FDR is actually an **expectation** of this proportion. We'll do it right anon.)

Back to Inference

- Can we develop general frameworks that allow us to control column-wise quantities like the false-discovery rate (FDR)?
 - in a similar way as Neyman-Pearson controls the false-positive rate
- To be continued...