# DS 102: Data, Inference, and Decisions

Lecture 1

Michael Jordan
University of California, Berkeley

# Data Science, Qu’est-ce que C’est?

# Data Science, Qu’est-ce que C’est?

- You all hopefully have your own thoughts at this point…

# Data Science, Qu’est-ce que C’est?

- You all hopefully have your own thoughts at this point…
- Let’s not aim for a definition, but rather try to capture a zeitgeist

# Data Science: A Personal Perspective

- It arose both in science and in technology, over the past two 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

# Data Science: A Personal Perspective

- It arose both in science and in technology, over the past two 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

# Data Science: A Personal Perspective

- It arose both in science and in technology, over the past two 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
- Oddly, “engineering” is not a prized term compared to “science”
  - compare “social science” to “social engineering”
  - and even “computer science” vs. “computer engineering”

# Data Science: A Personal Perspective

- It arose both in science and in technology, over the past two 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
- Oddly, "engineering" is not a prized term compared to "science"
  - compare "social science" to "social engineering"
  - and even "computer science" vs. "computer engineering"
- But engineering really 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

# An Academic Perspective

- Blending the engineering perspective with the science perspective, and remembering the “human-centric” aspect, the scope is that of a College
- Existing Colleges at most universities, such as Arts & Humanities, Biology, Business, Engineering, and Science, already existed a hundred years ago
- The ensuing century has given rise to computer science, information theory, optimization, statistical inference, economics, etc---the data, information, and decision fields
- From an academic perspective, “Data Science” stands for the union of these fields, and the real-world phenomena that they focus on

# 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 (again, 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

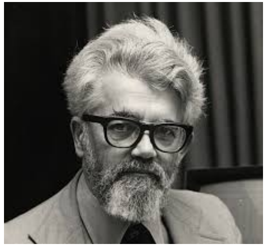
- First Generation (‘90-’00): the backend
  - e.g., fraud detection, search, supply-chain management
- Second Generation (‘00-’10): the human side
  - e.g., recommendation systems, commerce, social media
- Third Generation (‘10-now): pattern recognition
  - e.g., speech recognition, computer vision,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

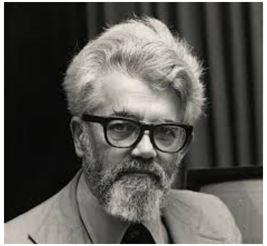
# Perspectives on AI*



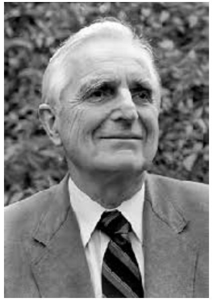
- The classical “human-imitative” aspiration

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

# Perspectives on AI*



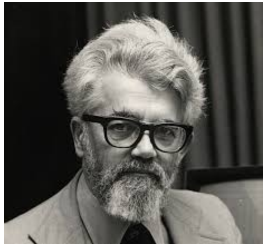
- The classical "human-imitative" aspiration



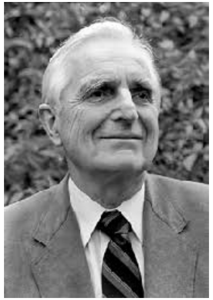
- The "intelligence augmentation" (IA) perspective

*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

# What Intelligent Systems Currently Exist?

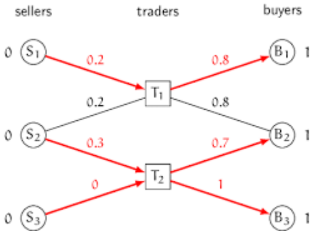
# What Intelligent Systems Currently Exist?

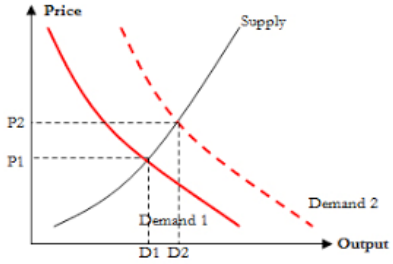
- Brains and Minds

# What Intelligent Systems Currently Exist?

- Brains and Minds

- Markets



# 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

# DS 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, and reinforcement learning
- We're trying to prepare you for the next wave, not the preceding wave!

# Existing IT Business Models

- Many modern IT companies collect data as part of providing a service on a platform
  - often the value provided by these services is limited
  - so the monetization comes from advertising
  - i.e., many companies are in fact creating markets based on data and learning algorithms, but these markets only link the IT company and the advertisers
- Humans are treated as a product, not as a player in a market
  - the results (ads) are not based on the utility (happiness) of the providers of the data, and does not pay them for their data

# Existing IT Business Models

- Many modern IT companies collect data as part of providing a service on a platform
  - often the value provided by these services is limited
  - so the monetization comes from advertising
  - i.e., many companies are in fact creating markets based on data and learning algorithms, but these markets only link the IT company and the advertisers
- Humans are treated as a product, not as a player in a market
  - the results (ads) are not based on the utility (happiness) of the providers of the data, and does not pay them for their data
- This is broken---humans should be able to participate fully in a market in which their data are being used

# Example: Music in the Data Age

- More people are making music than ever before, placing it on sites such as SoundCloud
- More people are listening to music than ever before
- But there is no economic value being exchanged between producers and consumers
- And, not surprisingly, most people who make music cannot do it as their full-time job
  - i.e., human happiness is being left on the table

# Example: Music in the Data Age

- More people are making music than ever before, placing it on sites such as SoundCloud
- More people are listening to music than ever before
- But there is no economic value being exchanged between producers and consumers
- And, not surprisingly, most people who make music cannot do it as their full-time job
  - i.e., human happiness is being left on the table

- There do exist companies who make money off of this; they stream data from SoundCloud to listeners, and they make their money … from advertising! ☹

# The Alternative: Create a Market

- Use data to provide a dashboard to musicians, letting them learn where their audience is
- The musician can give shows where they have an audience
- And they can make offers to their fans

# The Alternative: Create a Market

- Use data science to provide a dashboard to musicians, letting them learn where their audience is
- The musician can give shows where they have an audience
- And they can make offers to their fans
- I.e., consumers and producers become linked, and value flows: a market is created
  - the company that creates this market profits simply by taking a cut from the transactions

# The Alternative: Create a Market

- Use data science to provide a dashboard to musicians, letting them learn where their audience is
- The musician can give shows where they have an audience
- And they can make offers to their fans
- I.e., consumers and producers become linked, and value flows: a market is created
  - the company that creates this market profits simply by taking a cut from the transactions

- In the US, the company *United Masters* is doing precisely this; see www.unitedmasters.com

# Data Science Meets Culture

# Social Consequences

- By creating a market based on the data flows, new jobs can be created!
- So here's a way that Data Science or AI can be a job creator, and not a job killer
- This can be done in a wide range of other domains, not just music
  - entertainment
  - information services
  - personal services

# 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

# The Basic Two-by-Two Table

| Reality \ Decision | 0 | 1 |
|---|---|---|
| 0 | | |
| 1 | | |

# The Basic Two-by-Two Table

| | | Decision | |
|---|---|---|---|
| | | 0 | 1 |
| Reality | 0 | TN | FP |
| | 1 | FN | TP |

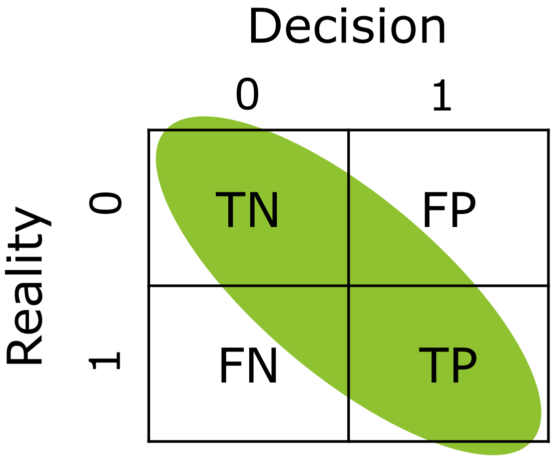
TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

# The Basic Two-by-Two Table




TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

# The Basic Two-by-Two Table

| Reality \ Decision | 0 | 1 |
| --- | --- | --- |
| 0 | TN | FP |
| 1 | FN | TP |

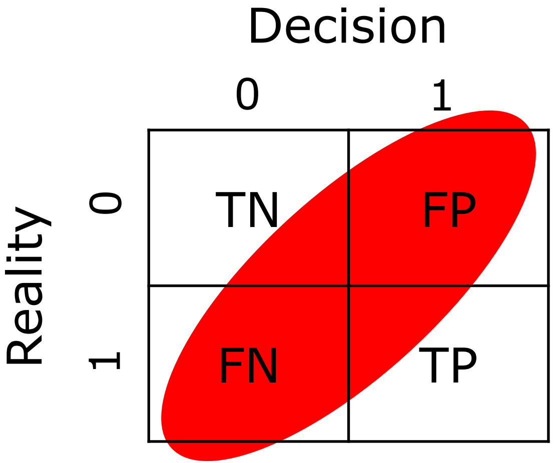
TN = True Negative

FP = False Positive

FN = False Negative

FP = True Positive

# The Basic Two-by-Two Table




TN = True Negative

FP = False Positive

FN = False Negative

FP = 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

# 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

- 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

| Reality \ Decision | 0 | 1 |
|---|---|---|
| 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

| Reality | | 0 | 1 |
|---|---|---|---|
| | 0 | $n_{00}$ | $n_{01}$ |
| | 1 | $n_{10}$ | $n_{11}$ |

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

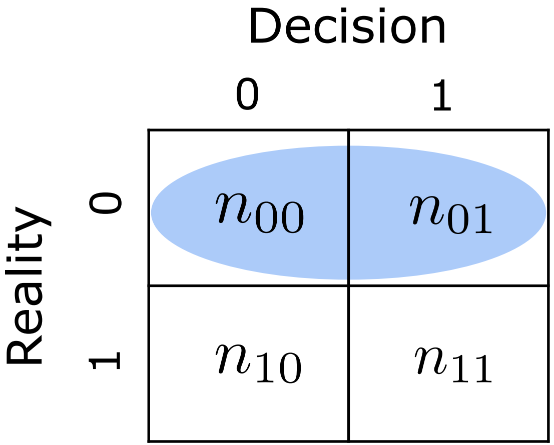
# Some Row-Wise Rates

| Reality \ Decision | 0 | 1 |
|---|---|---|
| 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




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

# Some Row-Wise Rates

| Reality \ Decision | 0 | 1 |
| --- | --- | --- |
| 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 $P(\text{Decision} = 1 \,|\, \text{Reality} = 1)$
- 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 later
  - specificity = 1 – Type I error rate
  - sensitivity = 1 – Type II error rate = power

# Towards Inference

- We'd like to have 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 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

# 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

| Reality \ Decision | 0 | 1 |
|---|---|---|
| 0 | $n_{00}$ | $n_{01}$ |
| 1 | $n_{10}$ | $n_{11}$ |

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

# Some Column-Wise Rates

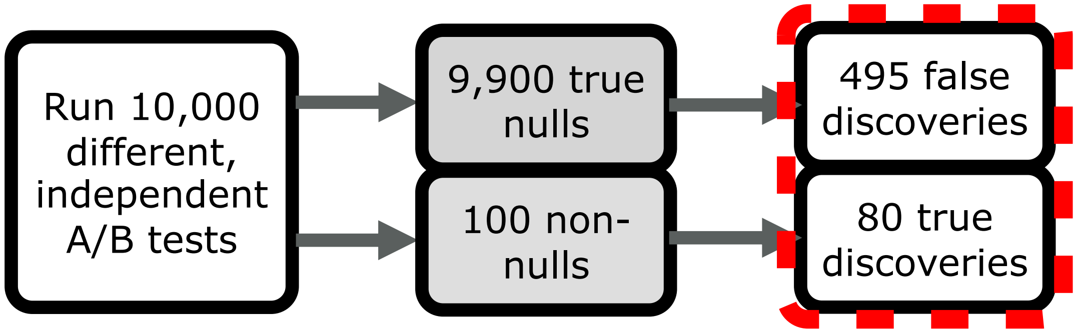
| Reality \ Decision | 0 | 1 |
|---|---|---|
| 0 | $n_{00}$ | $n_{01}$ |
| 1 | $n_{10}$ | $n_{11}$ |

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

# 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

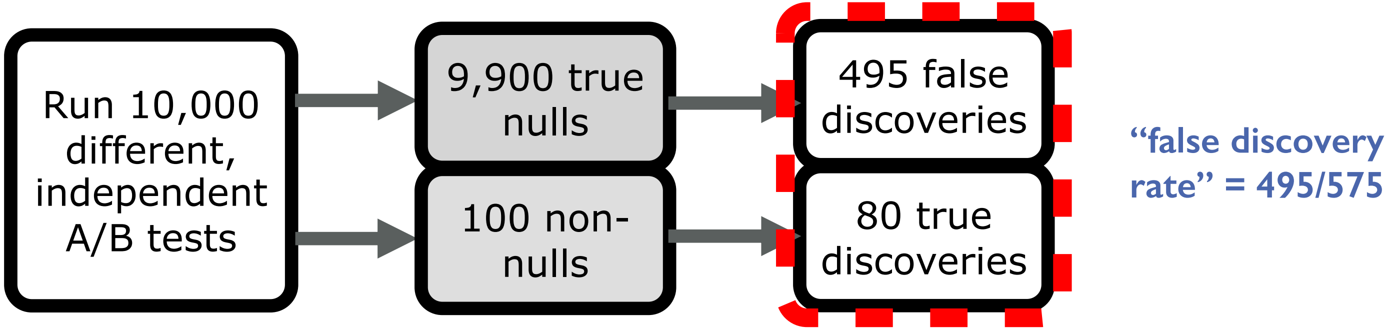
Type I error rate (per test) = 0.05




"false discovery rate" = 495/575

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…