6.s077 — INTRODUCTION TO DATA SCIENCE
EECS, MIT, Spring 2018

Lecture 1: Introduction

Instructors

Yury Polyanskiy

Devavrat Shah

John Tsitsiklis

TAs

Muhammad Jehangir Amjad

Eren Kizildag

Rishad Rahman

Today's agenda

- The scope of data science
  - versus "statistics," "machine learning," etc.

- The objectives:
  - modeling
  - "prediction"

  ("To model or not to model, that is the question")

- What can go wrong (= why this class)

- Course outline

- Probability background

Course mechanics

- The course Stellar site is the central point:
  http://stellar.mit.edu/S/course/6/sp18/6.S077/
- Join Piazza: https://piazza.com/class/jcll9afcjs56oe
- Read the "About this Course" handout
  - and other documents already posted on Stellar

- Lectures
- Recitations
- Problem sets (paper-pencil, plus easy computations)
- Computational projects
- Office hours
- Midterm & final

Prerequisites: 6.041A $+\epsilon$; Python

Our computational platform

- JupyterHub

- Python
  (including rich libraries)

- Follow the detailed instructions in the
  "Computation Platform Access and Setup" document
- Quick demo in this Friday's recitation

- Before Friday, start with "Problem Set 0"
  - to get familiar with platform
  - to be able to discuss sticky points at Friday's recitation
  - not due/graded

- Problem Set 1, due Thursday, Feb. 15

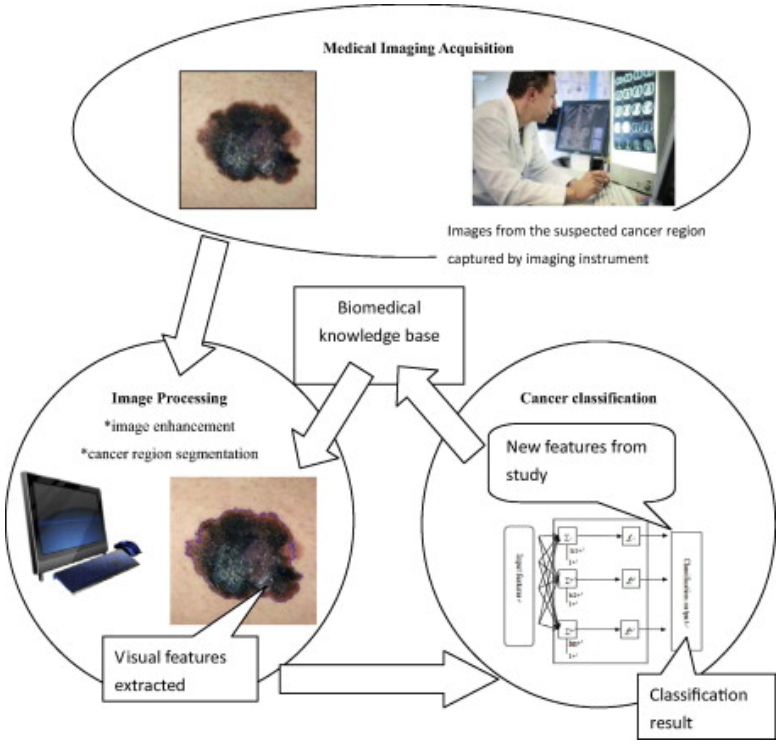
What is "Data Science"?

- 'a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven)' [attributed to Jim Gray]
- "data-driven science"
- scientific methods, processes, and systems to extract knowledge or insights from data in various forms

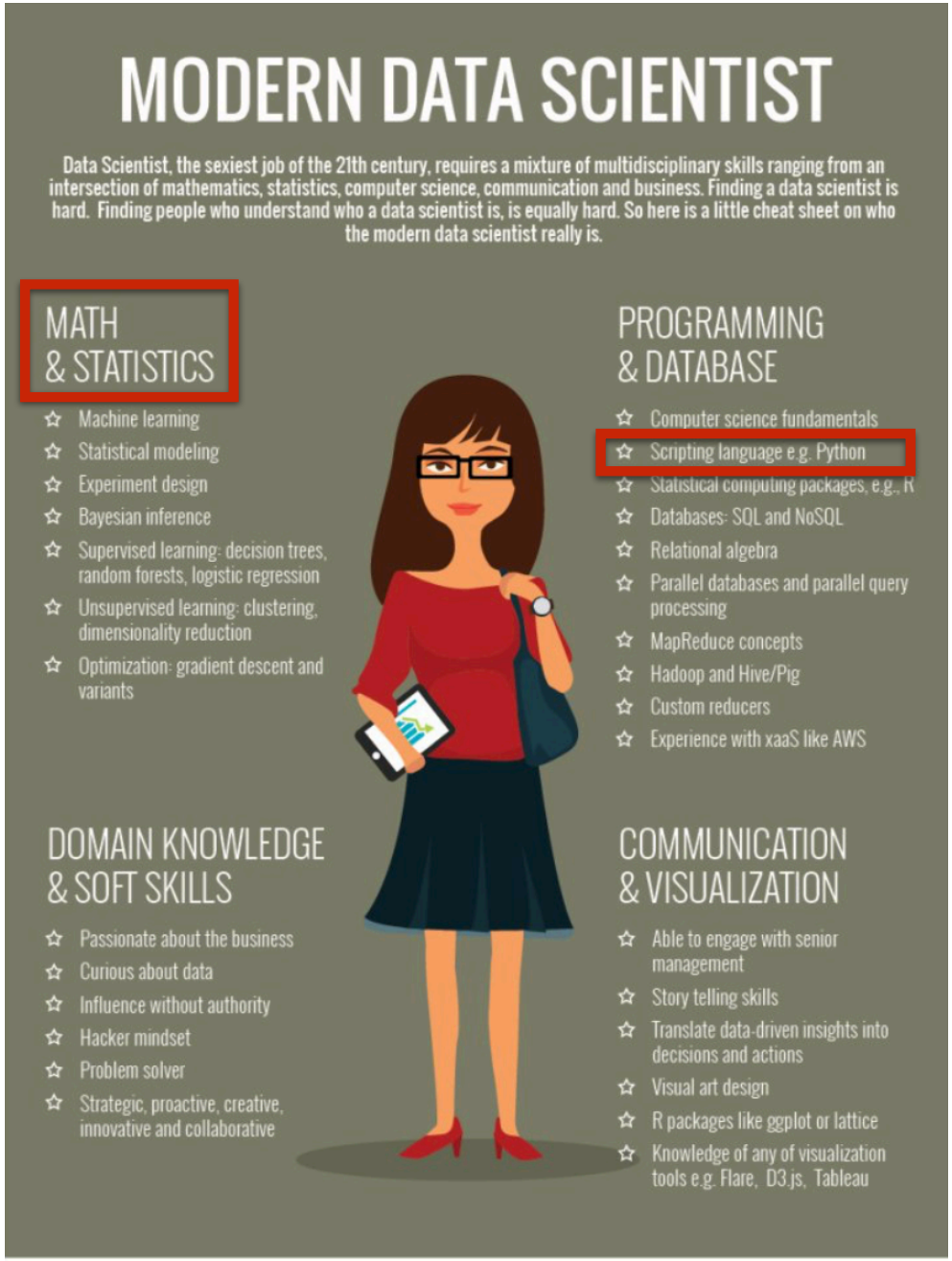
  https://en.wikipedia.org/wiki/Data_science

Data science is the science and art
of how to use data,
in order to do something useful or insightful.

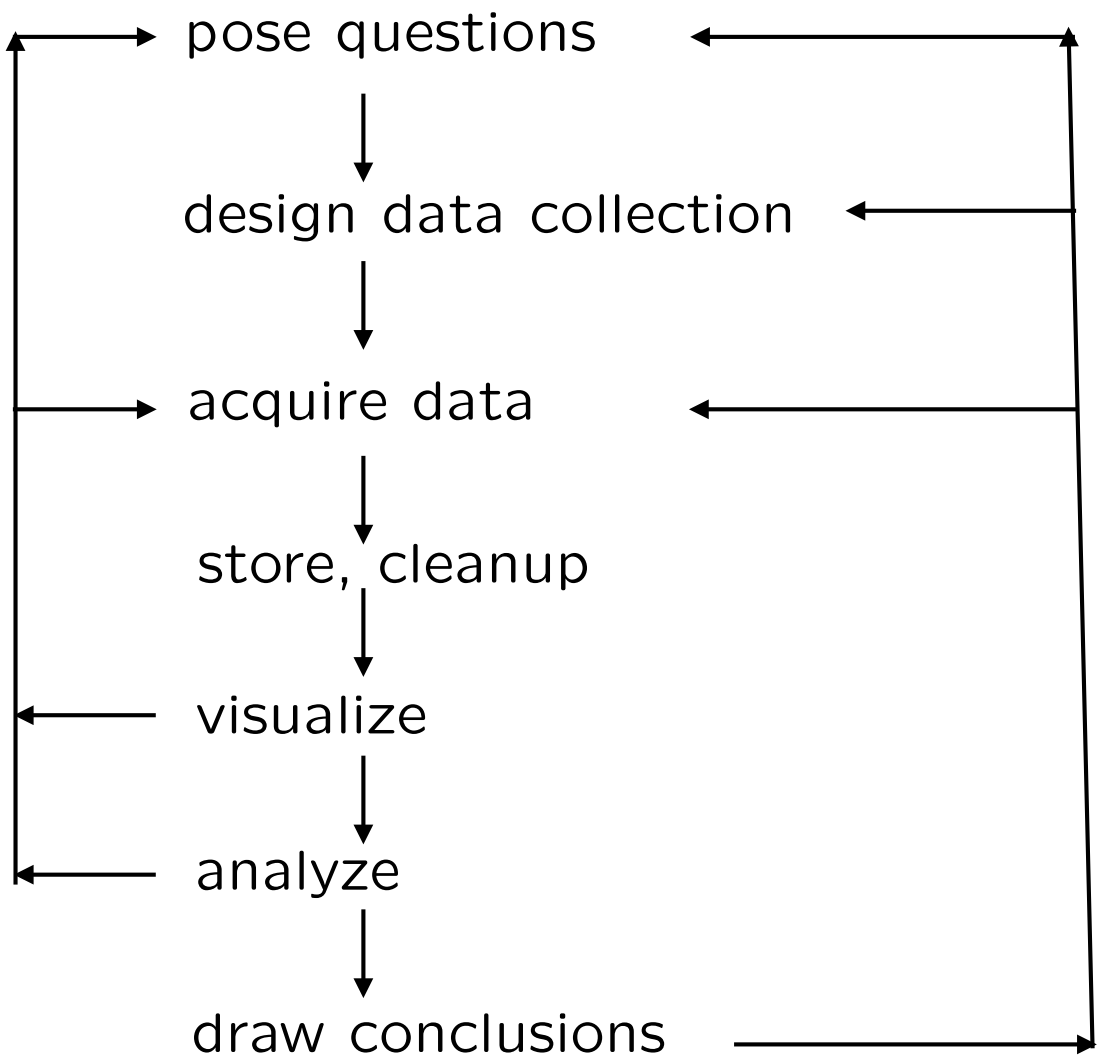
body of knowledge,
methods, and tools

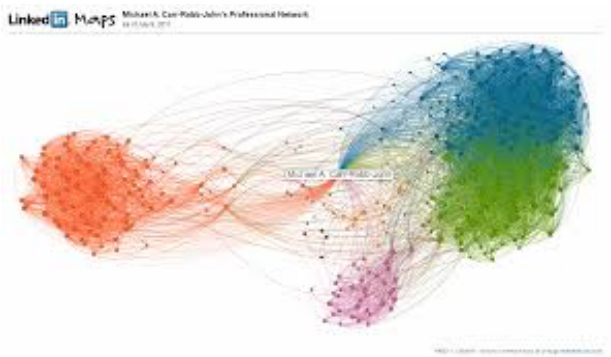


Non-ideological alternative: "whatever data scientists do"



What data scientists do: the workflow





Statistics versus machine learning – a carricature of a spectrum

Deep understanding
of simple models/methods

Theory does not
always explain success

Deep theorems
on complex models

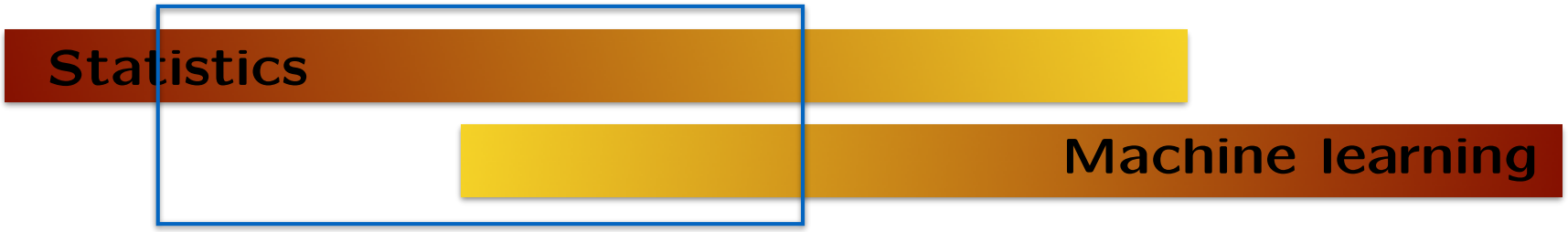
Fearless methods

Algorithmic emphasis

Cumulative art

Models	Structured	Unstructured or absent
Data	Few/moderate	Big
Q:	Does this treatment help?	Is it a cat or a dog?

Positioning of this class



Focus on:

- Fundamental concepts
- Key methods
  - not a laundry list
  - when do/don't they work
- What can go wrong

Need a "language": probability

A conceptual big picture — through a prototypical problem

Model?

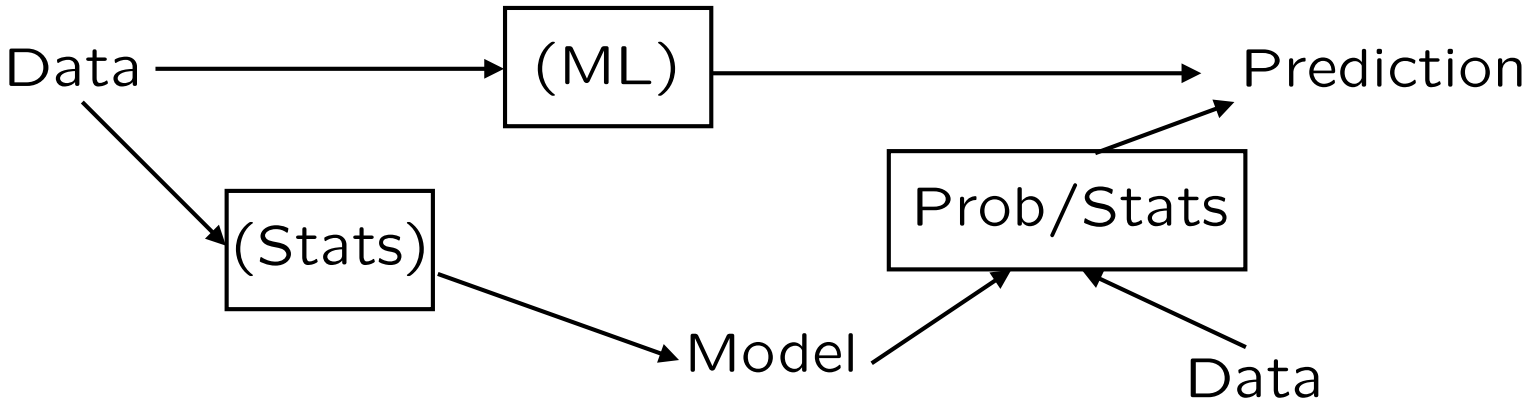
X_1	Y_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
X_n	Y_n
X	Y?

X: symptoms

Y: disease

- **"Predict"** Y, based on X

- **Understand**
  - build a model, a theory, a narrative, a mechanism
- **Act/change**



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

What can go wrong? Selection bias

– No Cambridge resident ever rode an empty bus;

therefore we are pretty sure that Cambridge buses are almost never empty :)

What can go wrong? "Exploring the data"

0 1 1 1 1 0 0 0 1 1 0 0 1 0 1 0 1 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 0 1 0 1 1

"unremarkable"

0 0 0 1 0 0 0 0 0 0 "wow! how unusual"

CULTURE

Chicago Bears' Winning Streak Advances to 14 ... Coin Flips

Jay Smith | September 29, 2017 4:53 pm

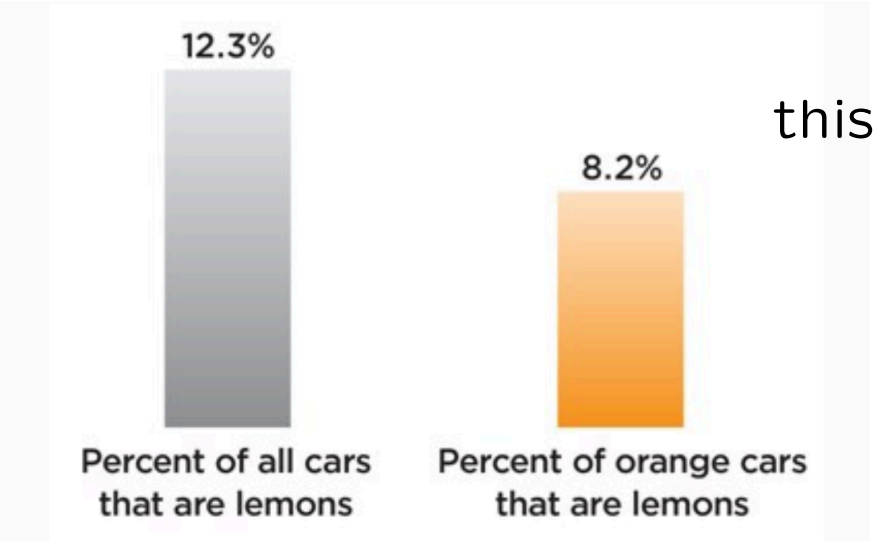
"fishing expedition" $\longrightarrow$ "see" structure in randomness

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

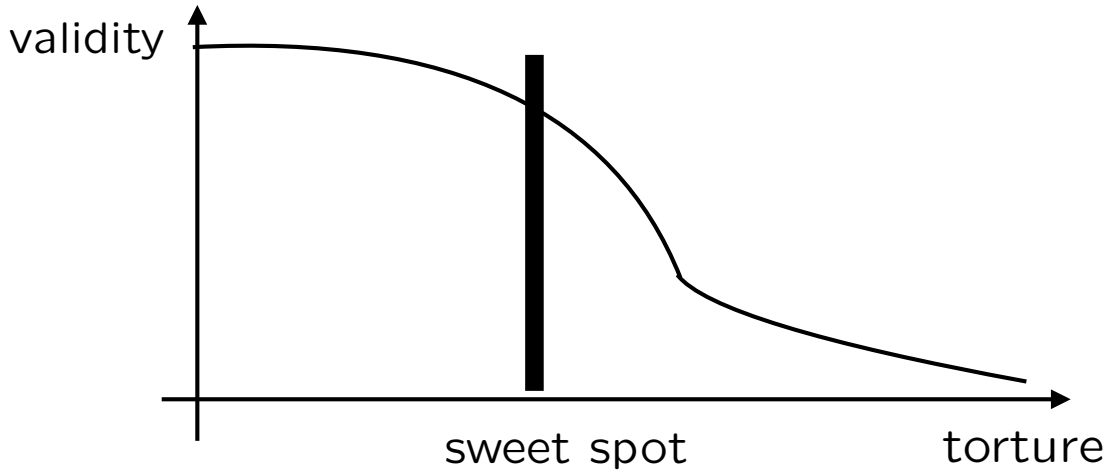
What can go wrong?



"If all color cars were similar,
this event would have chance $< 1\%$ to occur"

"So, we are pretty sure that
orange cars tend to be better"

- If you torture the data long enough, it will confess

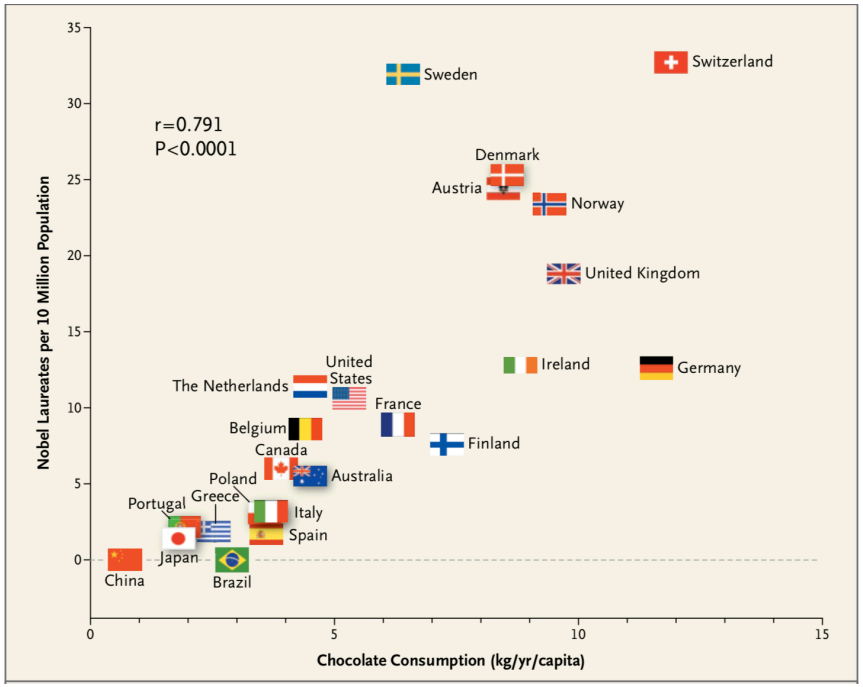


What can go wrong? Predicting ≪ Modeling

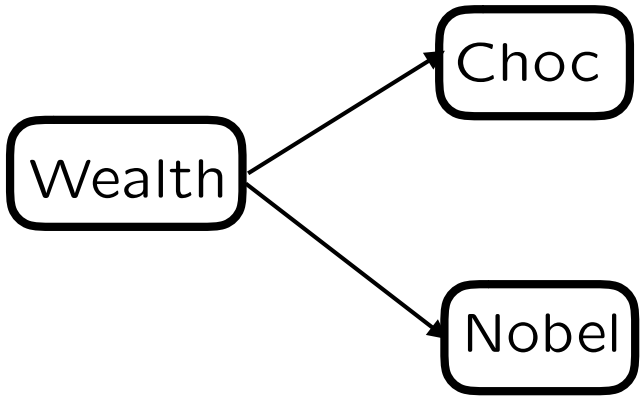
BUSINESS INSIDER

There's A Shocking Connection Between Eating More Chocolate And Winning The Nobel Prize

JOE WEISENTHAL
APR. 20, 2014, 11:10 AM



The NEW ENGLAND JOURNAL *of* MEDICINE



Need to infer a model/mechanism

Can we? How?

This is serious business

Why most published research findings are false 5452 2005
JPA Ioannidis
PLoS medicine 2 (8), e124

Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

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

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

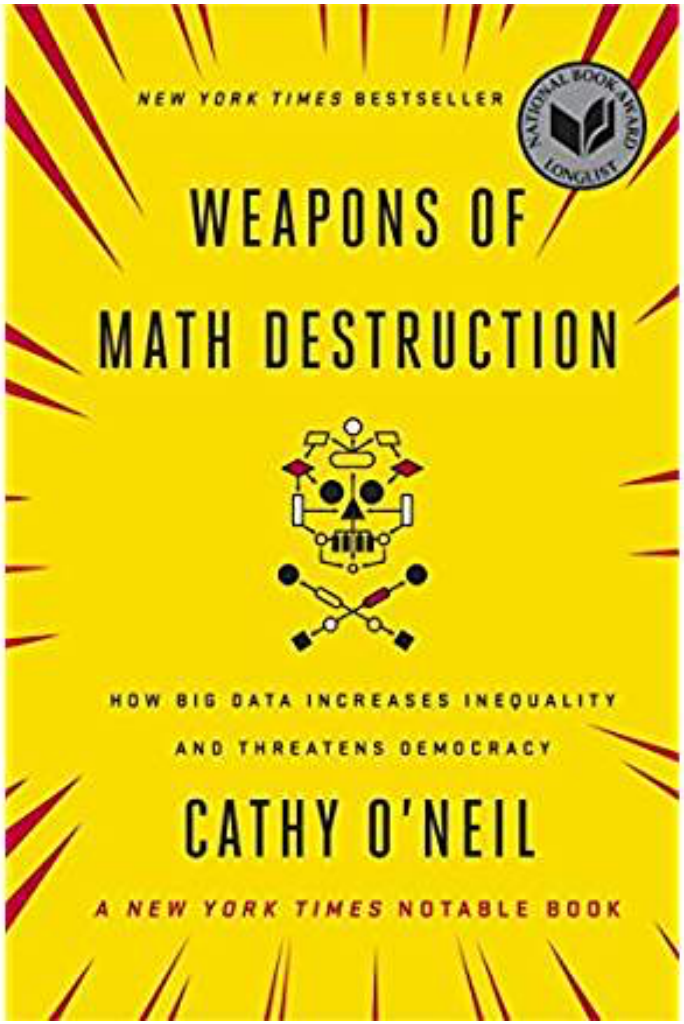
Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread notion that medical research articles

It can be proven that most claimed research findings are false.

should be interpreted based only on

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability of claiming a relationship when none truly exists reflects the Type I error rate, α. Assuming that c relationships are being probed in the field, the expected values of the 2×2 table are given in Table 1. After a research finding has been claimed based on achieving formal statistical significance,

This is very serious business



- Bad risk models $\longrightarrow$ financial meltdown

Bad DNA forensics

The New York Times | https://nyti.ms/2x5M4gT

Traces of Crime: How New York's DNA Techniques Became Tainted

The city's medical examiner has been a pioneer in analyzing complex DNA samples. But two methods were recently discontinued, raising questions about thousands of cases.

By LAUREN KIRCHNER SEPT. 4, 2017

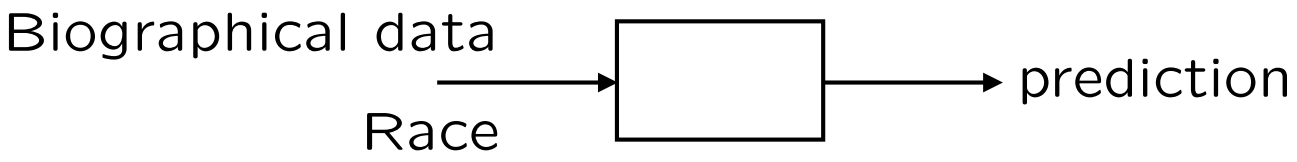
Misleading teacher evaluations

- Biased police deployments

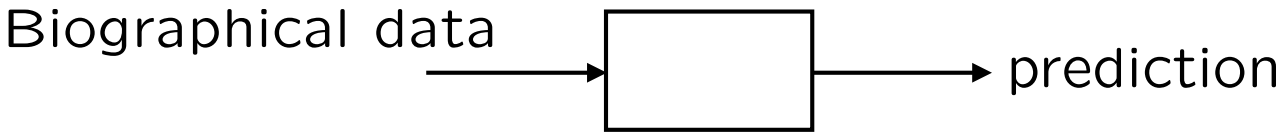
https://www.nytimes.com/2017/12/02/opinion/sunday/intelligent-policing-and-my-innocent-children.html

Can statistics become inadvertently racist?

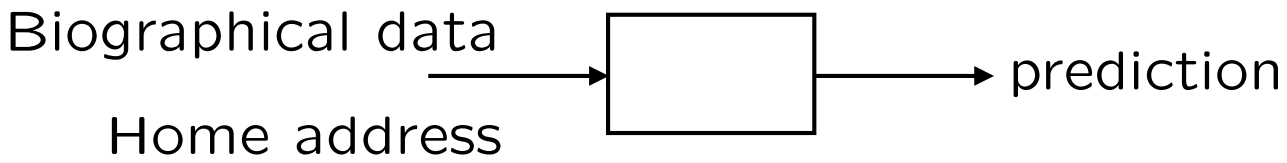
- Predict recidivism; or likelihood of loan default



not allowed



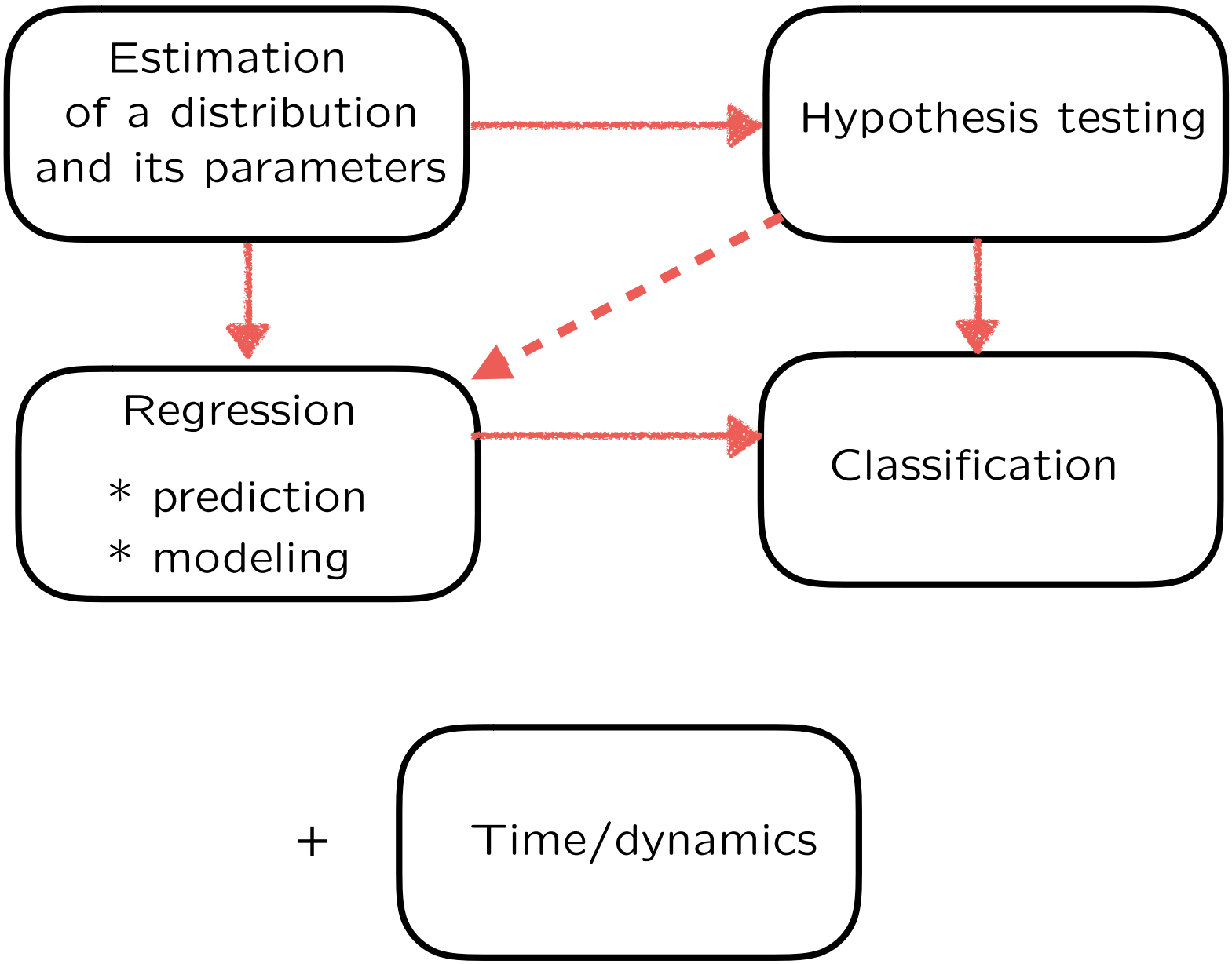
admissible



admissible,
but can replicate
the disallowed one

- Even formalizing the concept of a "race-blind algorithm" is subtle

Our class


Estimation
of a distribution
and its parameters
Hypothesis testing
Regression
* prediction
* modeling
Classification
+
Time/dynamics

For this week

- PSet 0 (become familiar with the platform!)

- Probability review!!!
  - read "probability background" document, on Stellar
  - identify gaps
  - fill any gaps!