INTRODUCTION TO DATA SCIENCE

JOHN P DICKERSON

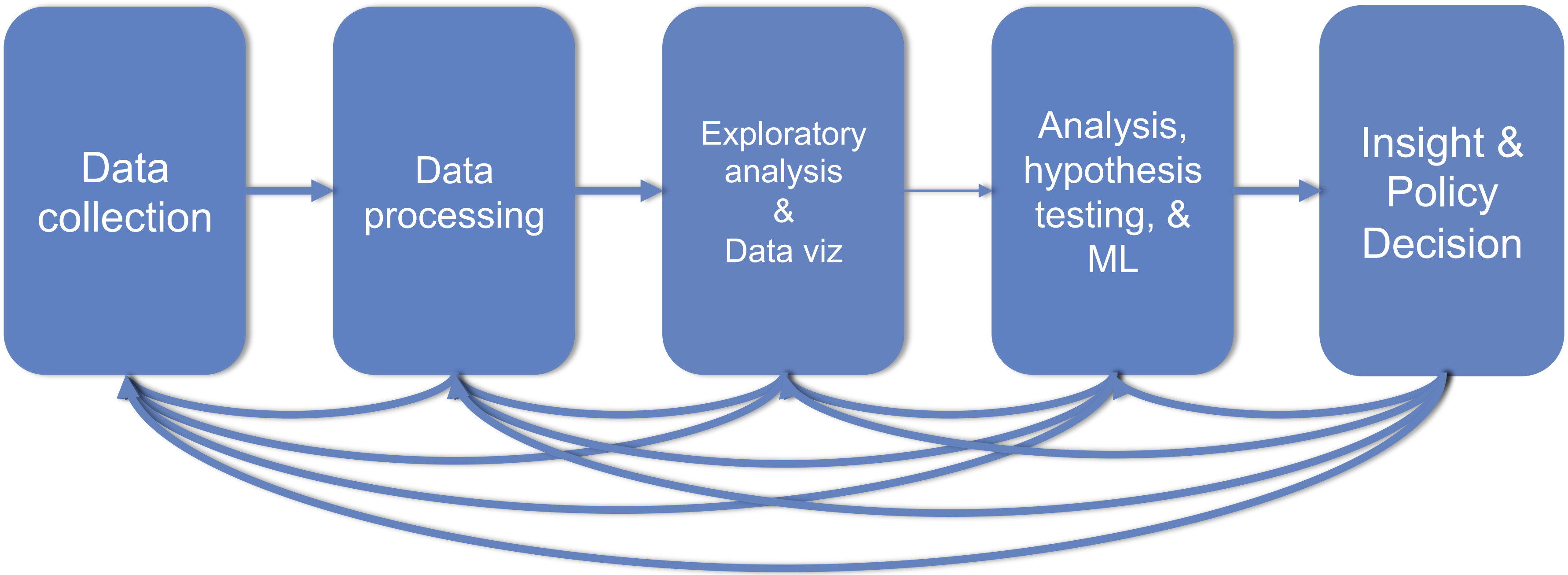
GUEST LECTURER *DU JOUR*

Lecture #4 – 2/5/2020

CMSC320
Tuesdays & Thursdays
5:00pm – 6:15pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

TODAY'S LECTURE



**BIG THANKS: Zico Kolter (CMU)
& Amol Deshpande (UMD)**

OUTLINE

Informed Consent

Reproducibility

~~p-value Hacking~~

Who owns the data?

Privacy & Anonymity

Debugging Data Science

Algorithmic fairness

Data validity/provenance

INFORMED CONSENT

Respect for persons -- cornerstone value for any conception of research ethics

Informed consent de facto way to "operationalize" that principle

- Integral component of medical research for many decades
- Applicable for any research where "personal information" is divulged or human experimentation performed
- Institutional Review Boards (IRBs) in charge of implementing

How it translates into the "big data" world?

- Largely ignored by most researchers

HISTORY

Systematic scientific experimentation on human subjects rare and isolated prior to the late 19th century

Some early directives in late 19th century and early 20th century

- Prussian directive in 1900: any medical intervention for any purpose other than diagnosis, healing, and immunisation must obtain “unambiguous consent” from patients after “proper explanation of the possible negative consequences” of the intervention

Nuremberg Code, drafted after conclusion of Nazi Doctors’ trials:

- established a universal ethical framework for clinical research
- “the voluntary consent of the human subject is absolutely essential” to ethical research
- Detailed specific guidelines on what to present to subjects (nature/ duration/purpose, how conducted, effects on health, etc)

HISTORY

Salgo v Leland Stanford etc. Board of Trustees (1957) … cited as establishing the legal doctrine of informed consent for medical practice and biomedical research in the United States

- plaintiff was awarded damages for not receiving full disclosure of facts

In 1953: NIH put the first IRB in place in its own hospital

- … voluntary agreement based on informed understanding shall be obtained from the patient
- … will be given an oral explanation in terms suited for his comprehension
- Only required a voluntary signed statement if the procedure involved “unusual hazard.”

HISTORY

A more detailed list of requirements emerged later

- 1) A fair explanation of the procedures to be followed, including an identification of those which are experimental;
- 2) A description of the attendant discomforts and risks;
- 3) A description of the benefits to be expected;
- 4) A disclosure of appropriate alternative procedures that would be advantageous for the subject;
- 5) An offer to answer any inquires concerning the procedures;
- 6) An instruction that the subject is free to withdraw his consent and to discontinue participation in the project or activity at any time

"Common Rule" – codification of "respect for persons, beneficence, and justice"

- Regulates use of human subjects in US today
- More elaborate treatment of all of these aspects

NON-MEDICAL RESEARCH

Unclear how the rules translate to other types of research

Identifying harm or potential risks difficult

Requirements and experiments change over the course of the study

The list of subjects itself evolving

CS has rarely had to deal with IRBs

- Although changing…

INDUSTRY RESEARCH

Less distinction between conventional or academic social scientific research and industry- or market-oriented research

Data fusion can lead to new insights and uses of data

Hard to translate the “informed consent” requirements to these settings

CASE STUDY: FACEBOOK EMOTIONAL EXPERIMENT

Facebook routinely does A/B testing to test out new features (e.g., layouts, features, fonts, etc)

In 2014: intentionally manipulated news feeds of 700k users

- Changed the number of positive and negative stories the users saw
- Measured how the users themselves posted after that

Hypothesis: Emotions spread over the social media

Huge outcry

Facebook claims it gets the "consent" from the user agreement

OKCUPID EXPERIMENTS

Experiment 1: Love is Blind

- Turned off photos for a day
- Activity went way down, but deeper conversations, better responses
- Deeper analysis at the link below

Experiment 2:

- Turned off text or not – kept picture
- Strong support for the hypothesis that the words don’t matter

Experiment 3: Power of Suggestion

- Told people opposite of what the algorithm suggested

https://theblog.okcupid.com/we-experiment-on-human-beings-5dd9fe280cd5

GDPR AND CONSENT

General Data Protection Regulation – new law in EU that recently went into play

Requires unambiguous consent

- data subjects are provided with a clear explanation of the processing to which they are consenting
- the consent mechanism is genuinely of a voluntary and "opt-in" nature
- data subjects are permitted to withdraw their consent easily
- the organisation does not rely on silence or inactivity to collect consent (e.g., pre-ticked boxes do not constitute valid consent);

OUTLINE

Informed Consent

Reproducibility

p-value Hacking

Who owns the data?

Privacy & Anonymity

Debugging Data Science

Algorithmic fairness

Data validity/provenance

THE REPRODUCIBILI… E

Noted by research community; in mu…

•
• Especially in preclinical research

Why animal research needs to improve

Many of the studies that use animals to model human diseases are too small and too prone to bias to be trusted, says **Malcolm Macleod**.

Beware the creeping cracks of bias

Evidence is mounting that research is riddled with systematic errors. Left unchecked, this could erode public trust, warns **Daniel Sarewitz**.

The Economist | World politics | Business & finance | Economics | Science & technology | Culture

Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition   Like 11k   Tweet 1,227

Believe it or not: how much can we rely on published data on potential drug targets?

Florian Prinz, Thomas Schlange and Khusru Asadullah

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

Raise standards for preclinical cancer research

C. Glenn Begley and **Lee M. Ellis** propose how methods, publications and incentives must change if patients are to benefit.

Drug targets slip-sliding away

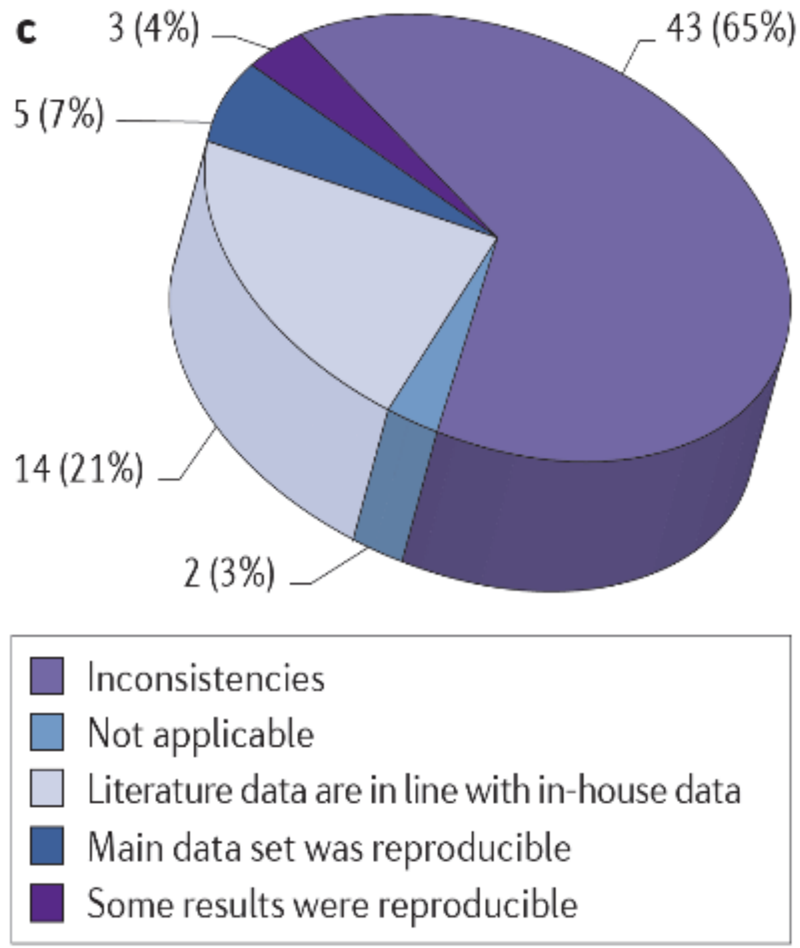
The starting point for many drug discovery programs is a published report on a new drug target. Assessing the reliability of such papers requires a nuanced view of the process of scientific discovery and publication.

Reforming Science: Methodological and Cultural Reforms

Believe it or not: how much can we rely on published data on potential drug targets?

Prinz, Schlange and Asadullah
Bayer HealthCare

Nature Reviews Drug Discovery
2011; 10:712-713



PERSPECTIVE

doi:10.1038/nature11556

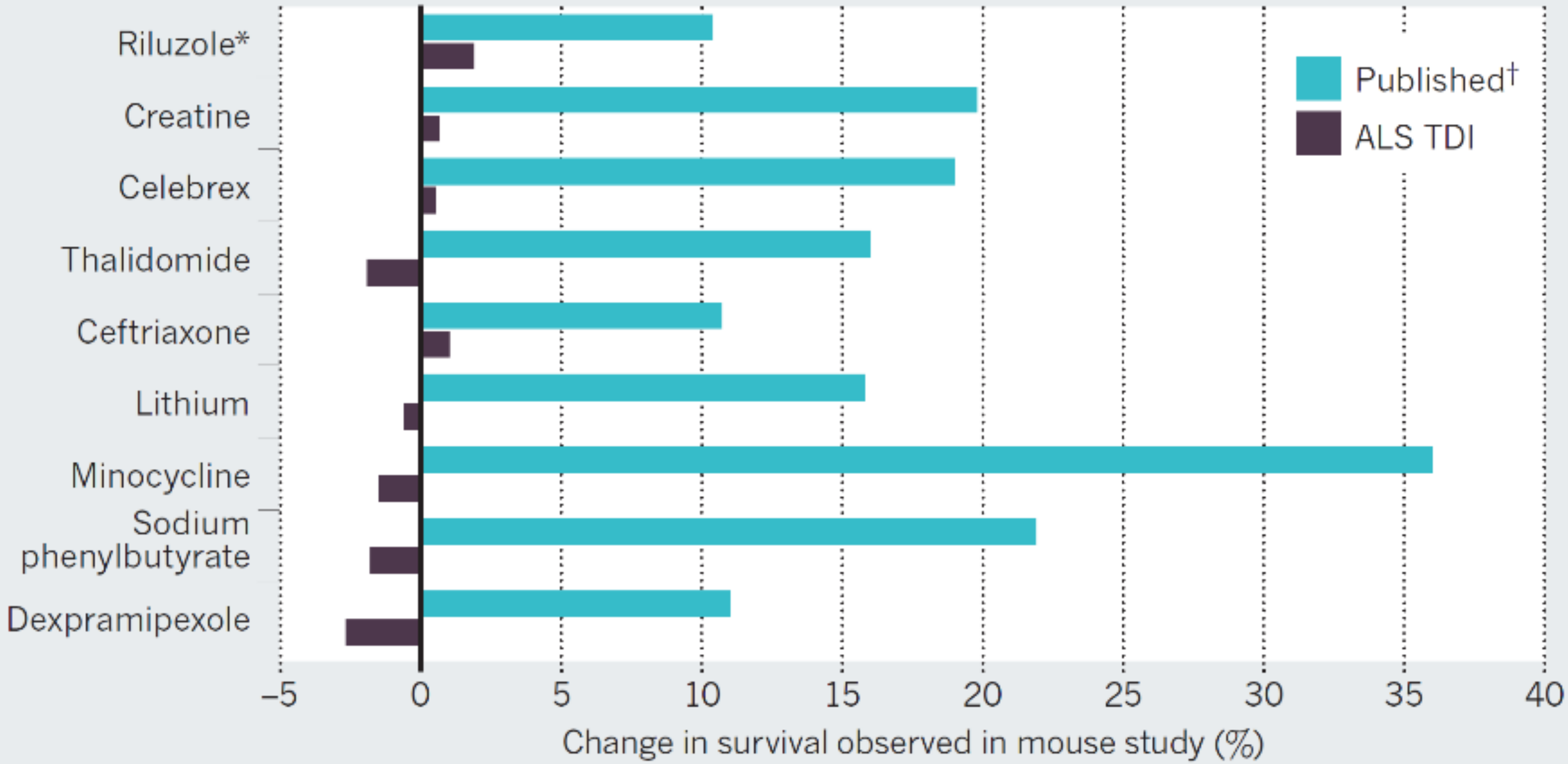
A call for transparent reporting to optimize the predictive value of preclinical research

Story C. Landis[1], Susan G. Amara[2], Khusru Asadullah[3], Chris P. Austin[4], Robi Blumenstein[5], Eileen W. Bradley[6], Ronald G. Crystal[7], Robert B. Darnell[8], Robert J. Ferrante[9], Howard Fillit[10], Robert Finkelstein[1], Marc Fisher[11], Howard E. Gendelman[12], Robert M. Golub[13], John L. Goudreau[14], Robert A. Gross[15], Amelie K. Gubitz[1], Sharon E. Hesterlee[16], David W. Howells[17], John Huguenard[18], Katrina Kelner[19], Walter Koroshetz[1], Dimitri Krainc[20], Stanley E. Lazic[21], Michael S. Levine[22], Malcolm R. Macleod[23], John M. McCall[24], Richard T. Moxley III[25], Kalyani Narasimhan[26], Linda J. Noble[27], Steve Perrin[28], John D. Porter[1], Oswald Steward[29], Ellis Unger[30], Ursula Utz[1] & Shai D. Silberberg[1]

The US National Institute of Neurological Disorders and Stroke convened major stakeholders in June 2012 to discuss how to improve the methodological reporting of animal studies in grant applications and publications. The main workshop recommendation is that at a minimum studies should report on sample-size estimation, whether and how animals were randomized, whether investigators were blind to the treatment, and the handling of data. We recognize that achieving a meaningful improvement in the quality of reporting will require a concerted effort by investigators, reviewers, funding agencies and journal editors. Requiring better reporting of animal studies will raise awareness of the importance of rigorous study design to accelerate scientific progress.

DUE DILIGENCE, OVERDUE

Results of rigorous animal tests by the Amyotrophic Lateral Sclerosis Therapy Development Institute (ALS TDI) are less promising than those published. All these compounds have disappointed in human testing.



*Although riluzole is the only drug currently approved by the US Food and Drug Administration for ALS, our work showed no survival benefit.
†References for published studies can be found in supplementary information at go.nature.com/hf4jf6.

Perrin, *Nature* 2014; 507: 423-425

CHALLENGES TO RIGOR AND TRANSPARENCY IN REPORTING SCIENCE

Science often viewed as self-correcting

- Immune from reproducibility problems?
- Principle remains true over the long-term

In the short- and medium-term, interrelated factors can short-circuit self-correction

- Leads to reproducibility problem
- Loss of time, money, careers, public confidence

FACTORS THAT "SHORT CIRCUIT" SELF-CORRECTION

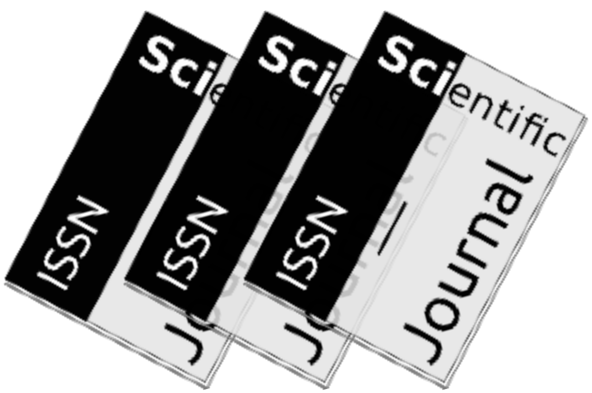
Current "hyper-competitive" environment fueled, in part, by:

- Historically low funding rates **$$**
- Grant review and promotion decisions depend too much on "high profile" publications

FACTORS THAT "SHORT CIRCUIT" SELF-CORRECTION

Publication practices:

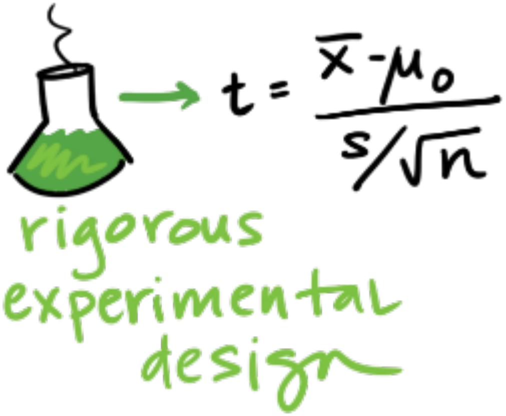
- Difficulty in publishing negative findings
- Overemphasis on the "exciting, big picture" finding sometimes results in publications leaving out necessary details of experiments



FACTORS THAT "SHORT CIRCUIT" SELF-CORRECTION

Poor training

- Inadequate experimental design
- Inappropriate use of statistics ("p-hacking")
- Incomplete reporting of resources used and/or unexpected variability in resources



REPRODUCIBILITY

Extremely important aspect of data analysis

- "Starting from the same raw data, can we reproduce your analysis and obtain the same results?"

Using libraries helps:

- Since you don't reimplement everything, reduce programmer error
- Large user bases serve as "watchdog" for quality and correctness

Standard practices help:

- Version control: git, git, git, …, git, svn, cvs, hg, Dropbox
- Unit testing: unittest (Python), RUnit (R), testthat
- Share and publish: github, gitlab

Slides adapted from Hector Corrado Bravo

PRACTICAL TIPS

Many tasks can be organized in modular manner:

- Data acquisition:
  - Get data, put it in usable format (many ‘join’ operations), clean it up – Anaconda lab from Tuesday!
- Algorithm/tool development:
  - If new analysis tools are required
- Computational analysis:
  - Use tools to analyze data
- Communication of results:
  - Prepare summaries of experimental results, plots, publication, upload processed data to repositories

Usually a single language or tool does not handle all of these equally well – **choose the best tool for the job!**

PRACTICAL TIPS

Modularity requires organization and careful thought

In Data Science, we wear two hats:

- Algorithm/tool developer
- **Experimentalist**: we don't get trained to think this way enough!

It helps two consciously separate these two jobs

THINK LIKE AN EXPERIMENTALIST

Plan your experiment

Gather your raw data

Gather your tools

Execute experiment

Analyze

Communicate

THINK LIKE AN EXPERIMENTALIST

Let this guide your organization. One potential structure for organizing a project:

```
project/
| data/
| | processing_scripts
| | raw/
| | proc/
| tools/
| | src/
| | bin/
| exps
| | pipeline_scripts
| | results/
| | analysis_scripts
| | figures/
```

THINK LIKE AN EXPERIMENTALIST

Keep a lab notebook!

Literate programming tools are making this easier for computational projects:

- http://en.wikipedia.org/wiki/Literate_programming
- https://ipython.org/
- http://rmarkdown.rstudio.com/
- http://jupyter.org/

THINK LIKE AN EXPERIMENTALIST

Separate experiment from analysis from communication

- Store results of computations, write separate scripts to analyze results and make plots/tables

Aim for reproducibility

- There are serious consequences for not being careful
  - Publication retraction
  - Worse: http://videolectures.net/cancerbioinformatics2010_baggerly_irrh/
- Lots of tools available to help, use them! Be proactive: learn about them on your own!

OUTLINE

Informed Consent

Reproducibility

p-value Hacking

Who owns the data?

Privacy & Anonymity

Debugging Data Science

Algorithmic fairness

Data validity/provenance

DATA OWNERSHIP

Consider your "biography"

- About you, but is it yours?
- No, the authors owns the copyright – not much you can do

If someone takes your photo, they own it

- Limits on taking photos in private areas
- Can't use the photo in certain ways, e.g., as implied endorsement or implied libel

Intellectual Property Basics:

- Copyright vs Patent vs Trade Secret
- Derivative works

DATA OWNERSHIP

Data Collection and Curation takes a lot of effort, and whoever does this usually owns the data “asset”

Crowdsourced data typically belongs to the facilitator

- Rotten tomatoes, yelp, etc.

What about personal data though?

- e.g., videos of you walking around a store, etc?
- Written contracts in some cases, but not always

New regulations likely to come up allowing customers to have more control over what happens with their data (e.g., GDPR)

OUTLINE

Informed Consent

Reproducibility

p-value Hacking

Who owns the data?

Privacy & Anonymity

Algorithmic fairness

Data validity/provenance

PRIVACY

First concern that comes to mind

- How to avoid the harms that can occur due to data being collected, linked, analyzed, and propagated?
- Reasonable rules ?
- Tradeoffs?

No option to exit

- In the past, could get a fresh start by moving to a new place, waiting till the past fades
- big data is universal and never forgets
- Data science results in major asymmetries in knowledge

WAYBACK MACHINES

Archives pages on the web (https://archive.org/web/ - 300 billion pages saved over time)

- almost everything that is accessible
- should be retained forever

If you have an unflattering page written about you, it will survive for ever in the archive (even if the original is removed)

RIGHT TO BE FORGOTTEN

Laws are often written to clear a person's record Law in EU and Argentina since 2006 after some years.

impacts search engines (not removed completely, but hard to find)

Collection vs Use

- Privacy usually harmed upon use of data
- Sometimes collection without use may be okay
- Survenillance:
    - By the time you know what you need, it is too late to go back and get it

WHY PRIVACY?

Data subjects have inherent right and expectation of privacy

"Privacy" is a complex concept

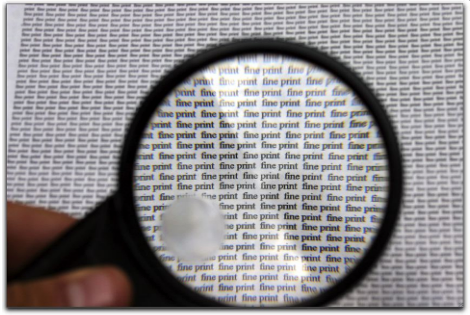
- What exactly does "privacy" mean? When does it apply?
- Could there exist societies without a concept of privacy?

Concretely: at collection "small print" outlines privacy rules

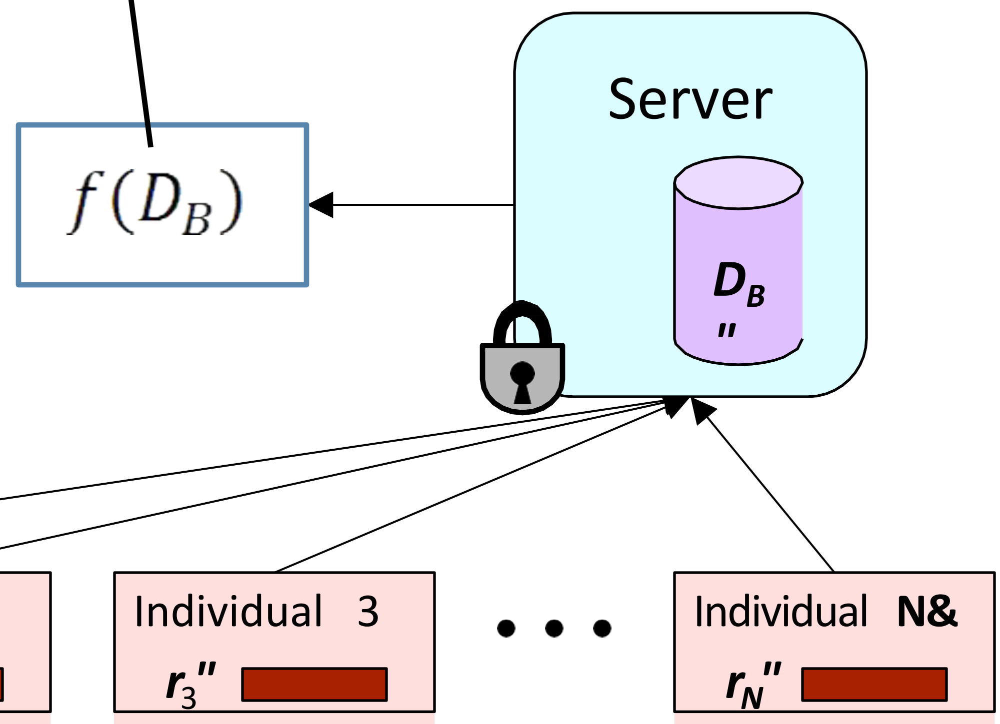
- Most companies have adopted a privacy policy
- E.g. AT&T privacy policy att.com/gen/privacy-policy?pid=2506

Significant legal framework relating to privacy

- UN Declaration of Human Rights, US Constitution
- HIPAA, Video Privacy Protection, Data Protection Acts



RELEASE THE DATA "ANONYMOUSLY" OR RELEASE A MODEL?



WHY ANONYMIZE?

For Data Sharing

- Give real(istic) data to others to study without compromising privacy of individuals in the data
- Allows third-parties to try new analysis and mining techniques not thought of by the data owner

For Data Retention and Usage

- Various requirements prevent companies from retaining customer information indefinitely
- E.g. Google progressively anonymizes IP addresses in search logs
- Internal sharing across departments (e.g. billing [?] marketing)

WHY ANONYMIZE?

2.1. Definitions in the EU Legal Context

Directive 95/46/EC refers to anonymisation in Recital 26 to exclude anonymised data from the scope of data protection legislation:

> *"Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible;".*[1]

Releasing data is bad?

What if we ensure our names and other identifiers are never released?

CASE STUDY: US CENSUS

Raw data: information about every US household

- Who, where; age, gender, racial, income and educational data

Why released: determine representation, planning

How anonymized: aggregated to geographic areas (Zip code)

- Broken down by various combinations of dimensions
- Released in full after 72 years

Attacks: no reports of successful deanonymization

- Recent attempts by FBI to access raw data rebuffed

Consequences: greater understanding of US population

- Affects representation, funding of civil projects
- Rich source of data for future historians and genealogists

CASE STUDY: NETFLIX PRIZE

Raw data: 100M dated ratings from 480K users to 18K movies

Why released: improve predicting ratings of unlabeled examples

How anonymized: exact details not described by Netflix

- All direct customer information removed
- Only subset of full data; dates modified; some ratings deleted,
- Movie title and year published in full

Attacks: dataset is claimed vulnerable [Narayanan Shmatikov 08]

- Attack links data to IMDB where same users also rated movies
- Find matches based on similar ratings or dates in both

Consequences: rich source of user data for researchers

- unclear if attacks are a threat—no lawsuits or apologies yet

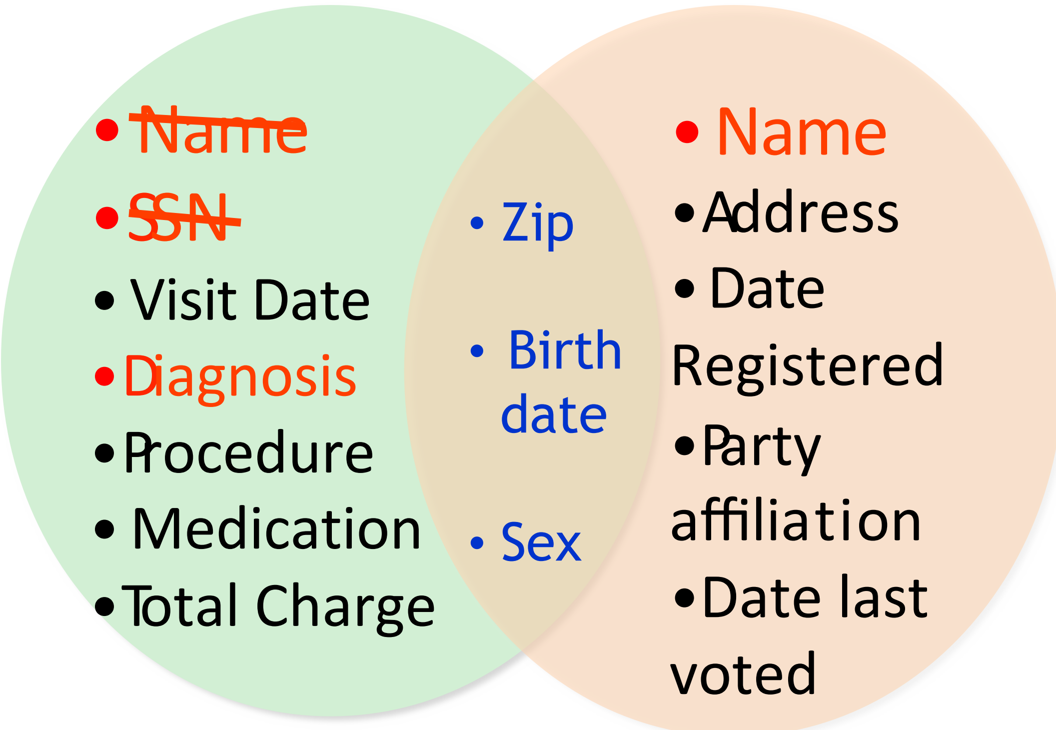
THE MASSACHUSETTS GOVERNOR PRIVACY BREACH [SWEENEY IJUFKS 2002]

- ~~Name~~
- ~~SSN~~
- Visit Date
- Diagnosis
- Procedure
- Medication
- Total Charge

- Zip
- Birth date
- Sex

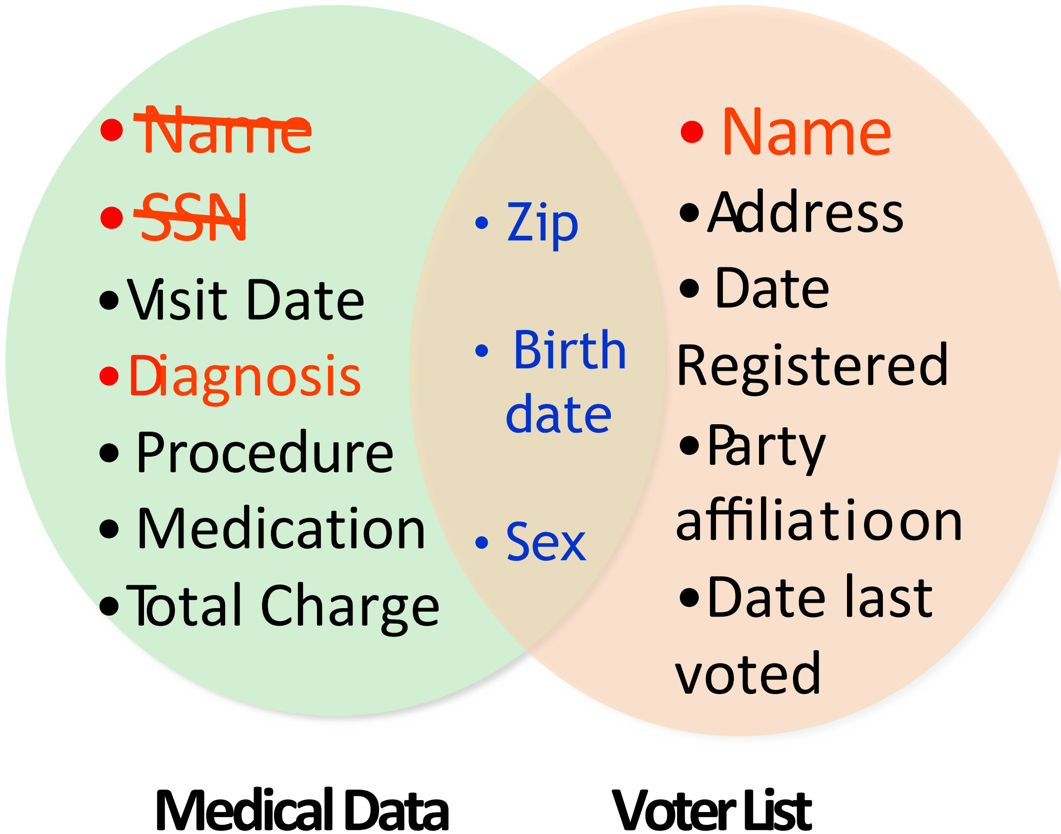
Medical Data

THE MASSACHUSETTS GOVERNOR PRIVACY BREACH [SWEENEY IJUFKS 2002]



Medical Data

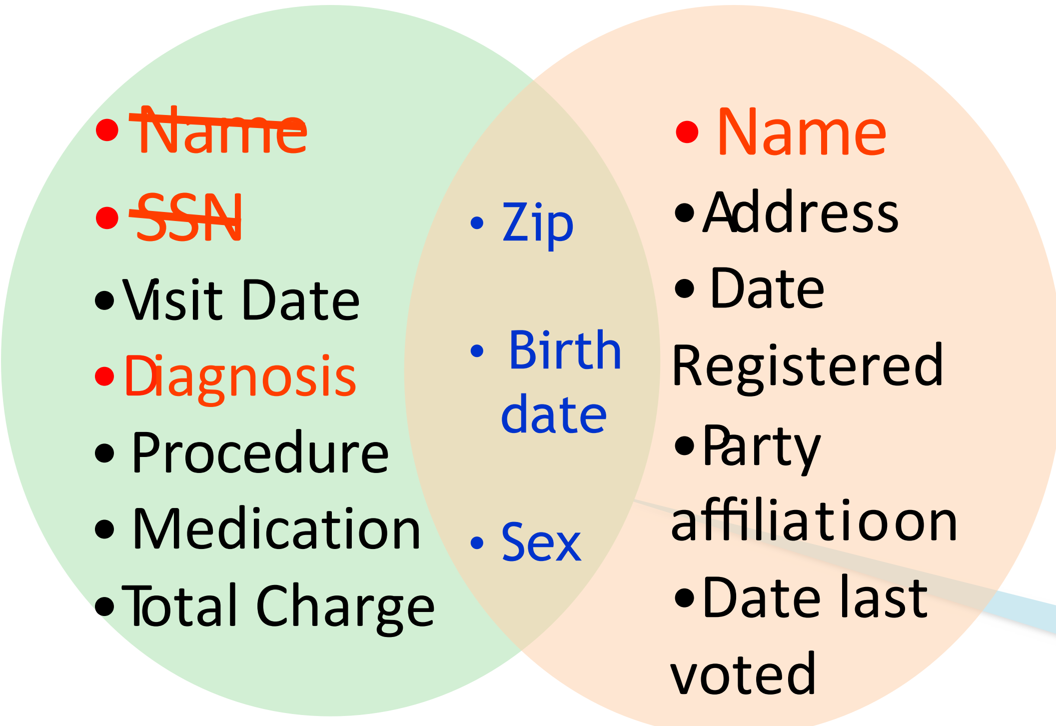
THE MASSACHUSETTS GOVERNOR PRIVACY BREACH [SWEENEY IJUFKS 2002]



- Governor of MA **uniquely identified** using ZipCode, Birth Date, and Sex.

Name linked to Diagnosis

THE MASSACHUSETTS GOVERNOR PRIVACY BREACH [SWEENEY IJUFKS 2002]



Medical Data **Voter List**

- 87 % of US population **uniquely identified** using ZipCode, Birth Date, and Sex.

Quasi-Identifiers

AOL DATA PUBLISHING FIASCO …

AOL “anonymously” released a list of 21 million web search queries.

Ashwin222	Uefa cup
Ashwin222	Uefa champions league
Ashwin222	Champions league final
Ashwin222	Champions league final 2007
Pankaj156	exchangeability
Pankaj156	Proof of deFinitti s theorem
Cox12345	Zombie games
Cox12345	Warcraft
Cox12345	Beatles anthology
Cox12345	Ubuntu breeze
Ashwin222	Grammy 2008 nominees
Ashwin222	Amy Winehouse rehab

AOL DATA PUBLISHING FIASCO …

AOL “anonymously” released a list of 21 million web search queries.

UserIDs were replaced by random numbers …

865712345	Uefa cup
865712345	Uefa champions league
865712345	Champions league final
865712345	Champions league final 2007
236712909	exchangeability
236712909	Proof of deFinitti s theorem
112765410	Zombie games
112765410	Warcraft
112765410	Beatles anthology
112765410	Ubuntu breeze
865712345	Grammy 2008 nominees
865712345	Amy Winehouse rehab

Privacy Breach

[NYTimes 2006]

A Face Is Exposed for AOL Searcher No. 4417749

By MICHAEL BARBARO and TOM ZELLER Jr.
Published: August 9, 2006

SIGN IN TO E-
THIS

CASE STUDY: AOL SEARCH DATA

Raw data: 20M search queries for 650K users from 2006

Why released: allow researchers to understand search patterns

How anonymized: user identifiers removed

- All searches from same user linked by an arbitrary identifier

Attacks: many successful attacks identified individual users

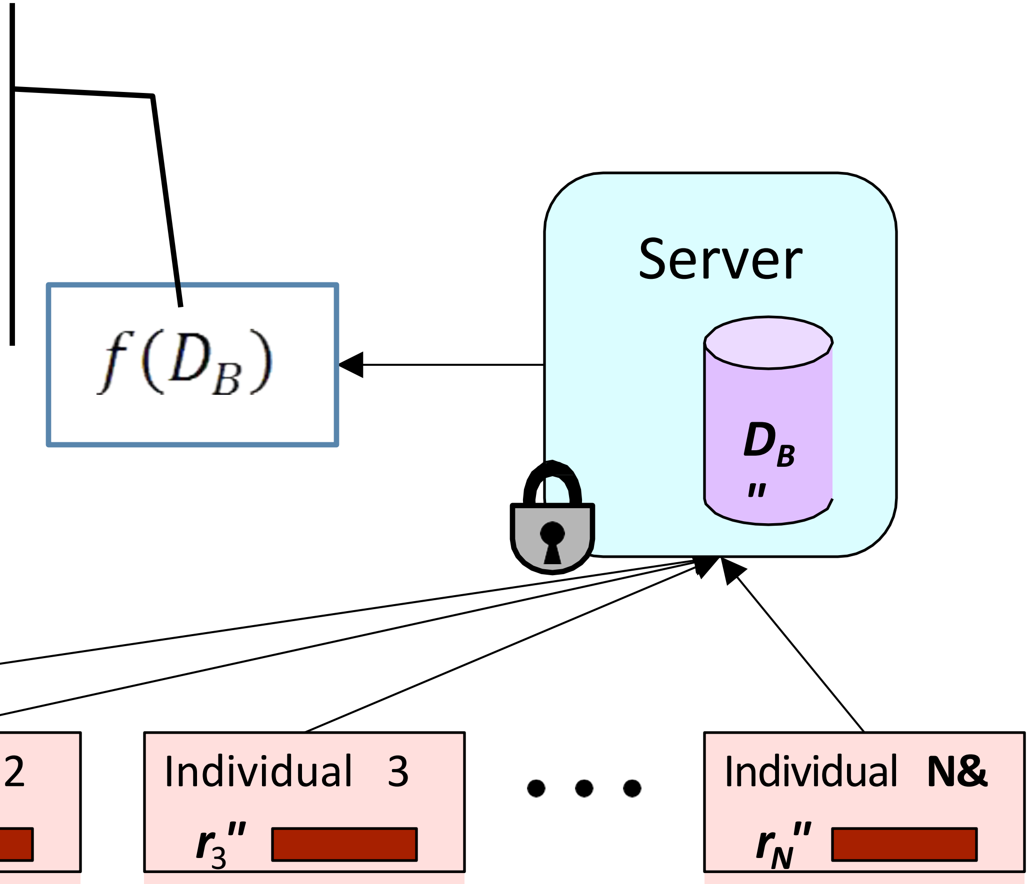
- Ego-surfers: people typed in their own names
- Zip codes and town names identify an area
- NY Times identified 4417749 as 62yr old GA widow [Barbaro Zeller 06]

Consequences: CTO resigned, two researchers fired

- Well-intentioned effort failed due to inadequate anonymization

CAN WE RELEASE A MODEL ALONE?

Release the data "~~anonymously~~" or release a model



RELEASING A MODEL CAN ALSO BE BAD

[Korolova JPC 2011]

Facebook profile

+

Online Data

- who live in the **United States**
- who live within 50 miles of **Staten Island, NY**
- between the ages of **23** and **27** inclusive
- who are **female**
- who are connected to **DogAnd PonyShow**
- in one of the categories: **Pop Culture**, **Science Fiction/Fantasy**, **Alternative**, **Rock**, **Classic Rock** or **iPhone**

Number of Impressions

+ Who are interested in **Men**	25
+ Who are interested in **Women**	0

Facebook's learning algorithm uses private information to predict match to ad

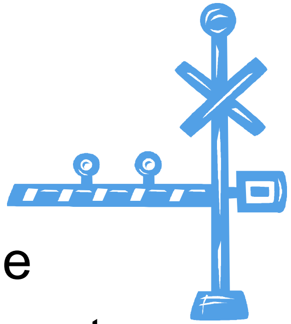
Model Inversion

[Frederickson et al., USENIX Security 2014]

- An attacker, given the model and some demographic information about a patient, can predict the patient's genetic markers.

We show, however, that warfarin models do pose a privacy risk (Section 3). To do so, we provide a general model inversion algorithm that is optimal in the sense that it minimizes the attacker's *expected misprediction rate* given the available information. We find that when one knows a target patient's background and stable dosage, their genetic markers are predicted with significantly better accuracy (up to 22% better) than guessing based on marginal distributions. In fact, *it does almost as well as regression models specifically trained to predict these markers (only ˜5% worse)*, suggesting that model inversion can be nearly as effective as learning in an "ideal" setting. Lastly, the inverted model performs measurably better for members of the training cohort than others (yielding an increased 4% accuracy) indicating a leak of information specifically about those patients.

MODELS OF ANONYMIZATION



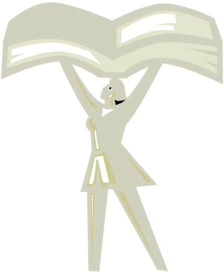
Interactive Model (akin to statistical databases)

- Data owner acts as "gatekeeper" to data
- Researchers pose queries in some agreed language
- Gatekeeper gives an (anonymized) answer, or refuses to answer

"Send me your code" model

- Data owner executes code on their system and reports result
- Cannot be sure that the code is not malicious

Offline, aka "publish and be damned" model



- Data owner somehow anonymizes data set
- Publishes the results to the world, and retires
- Seems to model most real releases

OBJECTIVES FOR ANONYMIZATION

Prevent (high confidence) inference of associations

- Prevent inference of salary for an individual in “census”
- Prevent inference of individual’s viewing history in “video”
- Prevent inference of individual’s search history in “search”
- All aim to prevent linking sensitive information to an individual

Prevent inference of presence of an individual in the data set

- Satisfying “presence” also satisfies “association” (not vice-versa)
- Presence in a data set can violate privacy (eg STD clinic patients)

Have to model what knowledge might be known to attacker

- Background knowledge: facts about the data set (X has salary Y)
- Domain knowledge: broad properties of data (illness Z rare in men)

UTILITY

Anonymization is meaningless if utility of data not considered

- The empty data set has perfect privacy, but no utility
- The original data has full utility, but no privacy

What is "utility"? Depends what the application is…

- For fixed query set, can look at max, average distortion
- Problem for publishing: want to support unknown applications!
- Need some way to quantify utility of alternate anonymizations

PRIVACY IS NOT ANONYMITY

- Bob's record is indistinguishable from records of other Cancer patients
  - We can infer Bob has Cancer !

- “New Information” principle
  - Privacy is breached if releasing D (or f(D)) allows an adversary to learn sufficient new information.
  - *New Information = distance(adversary's prior belief, adversary's posterior belief after seeing D)*
  - *New Information* can't be 0 if the output D or f(D) should be useful.

PRIVACY DEFINITIONS

- Many privacy definitions
  - L-diversity, T-closeness, M-invariance, **ε- Differential privacy,** E- Privacy, …
- Definitions differs in
  - What information is considered sensitive
    - Specific attribute (disease) vs all possible properties of an individual
  - What is the adversary's prior
    - All values are equally likely vs Adversary knows everything about all but one individuals
  - How is new information measured
    - Information theoretic measures
    - Pointwise absolute distance
    - Pointwise relative distance

NO FREE LUNCH

- Why can't we have a single definition for privacy?
  - For every adversarial prior and every property about an individual, new information is bounded by some constant.
- No Free Lunch Theorem: For every algorithm that outputs a D with even a sliver of utility, there is some adversary with a prior such that privacy is not guaranteed.

RANDOMIZED RESPONSE MODEL

- N respondents asked a sensitive “yes/no” question.
- Surveyor wants to compute fraction π who answer “yes”.
- Respondents don't trust the surveyor.
- What should the respondents do?

RANDOMIZED RESPONSE MODEL

- Flip a coin
  - heads with probability p, and
  - tails with probability 1-p ($p > ½$)

- Answer question according to the following table:

	True Answer = Yes	True Answer = No
Heads	Yes	No
Tails	No	Yes

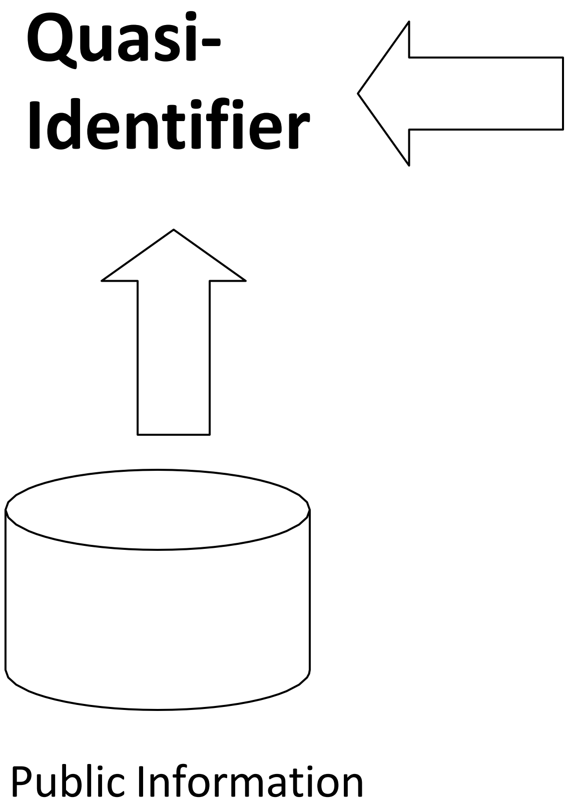
SAMPLE MICRODATA

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer
265-04-1275	14853	55	Russian	Heart
574-22-0242	14850	47	American	Viral
388-32-1539	14850	59	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer

REMOVING SSN ...

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

LINKAGE ATTACKS



Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

K-ANONYMITY

[Samarati et al, PODS 1998]

- Generalize, modify, or distort quasi-identifier values so that no individual is uniquely identifiable from a group of *k*
- In SQL, table T is **k-anonymous** if each

```
SELECT COUNT(*)
FROM T
GROUP BY Quasi-Identifier
```

  is **≥ k**

- Parameter k indicates the “degree” of anonymity

EXAMPLE: GENERALIZATION (COARSENING)

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053			
13053			
13068			
13068			

Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Equivalence Class: Group of k-anonymous records that share the same value for Quasi-identifier attribtutes

K-ANONYMITY THROUGH MICROAGGREGATION

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

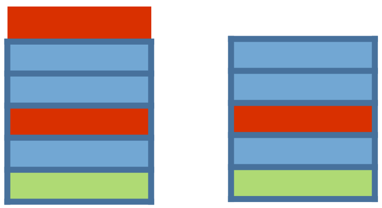
Zip	Age	Nationality	Disease
4 tuples Zip code = 130** 23 < Age < 29 Average(age) = 25			2 Heart and 2 Flu
4 tuples Zip = 1485* 47 < Age < 59 Average(age) = 53			1 Cancer, 1 Heart and 2 Flu
4 tuples Zip = 130** 31 < Age < 37 Avergae(age) = 34			All Cancer patients

DIFFERENTIAL PRIVACY

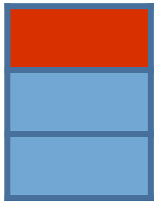
[Dwork ICALP 2006]

For every pair of inputs that differ in one row

For every output …



D_1 *D_2*



O

Adversary should not be able to distinguish between any D_1 and D_2 based on any O

$$\log\left(\frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]}\right) < \varepsilon \quad (\varepsilon > 0)$$

DIFFERENTIAL PRIVACY

- **Typically achieved by adding controlled noise (e.g., Laplace Mechanism)**

- **Some adoption in the wild:**
  - US Census Bureau
  - Google, Apple, and some others have used this for collecting data

- **Issues:**
  - Effectiveness in general still unclear

OUTLINE

Informed Consent

Reproducibility

p-value Hacking

Who owns the data?

Privacy & Anonymity

Debugging Data Science

Algorithmic fairness

Other Issues

Data Science in Industry

Traditional debugging

Traditional debugging of programs is relatively straightforward

You have some desired input/output pairs

You have a mental model (or maybe something more formal) of how each step in the algorithm “should” work

You trace through the execution of the program (either through a debugger or with print statement), to see where the state diverges from your mental model (or to discover your mental model is wrong)

Data science debugging

You have some desired input/output pairs

Your mental model is that an ML algorithm should work because … math? ... magic?

What can you trace through to see why it may not be working? Not very useful to step through an implementation of logistic regression…

Debugging data science vs. machine learning

Many of the topics here overlap with material on “debugging machine learning”

We are indeed going to focus largely on debugging data science prediction tasks (debugging web scraping, etc, is much more like traditional debugging)

But,

The first step of data science debugging

Step 1: determine if your problem is impossible

There are plenty of tasks that would be really nice to be able to predict, and absolutely no evidence that there the necessary signals to predict them (see e.g., predicting stock market from Twitter)

But, hope springs eternal, and it’s hard to prove a negative…

A good proxy for impossibility

Step 1: ~~determine if your problem is impossible~~ see if *you* can solve your problem manually

Create an interface where you play the role of the prediction algorithm, you need to make the predictions of the outputs given the available inputs

To do this, you'll need to provide some intuitive way of visualizing what a complete set of input features looks like: tabular data for a few features, raw images, raw text, etc

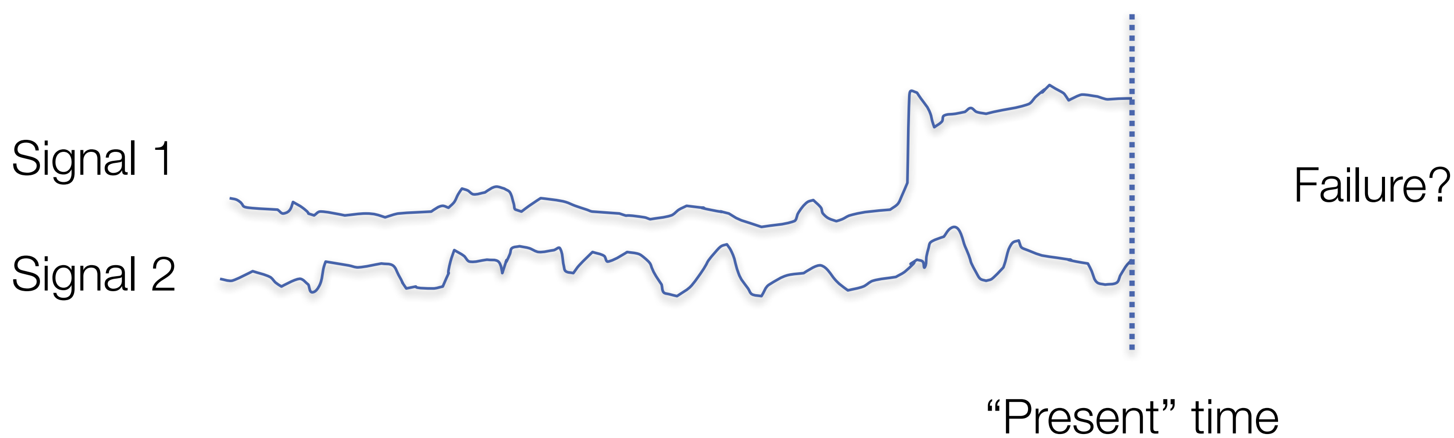
Just like a machine learning algorithm, you can refer to training data (where you know the labels), but you can't peak at the answer on your test/validation set

An example: predictive maintenance

An example task: you run a large factory and what to predict whether any given machine will fail within the next 90 days

You’re given signals monitoring the state of this device

Your interface: visualize the signals (but not whether there was a failure or not), and see if you can identify whether or not a machine is about to fail?



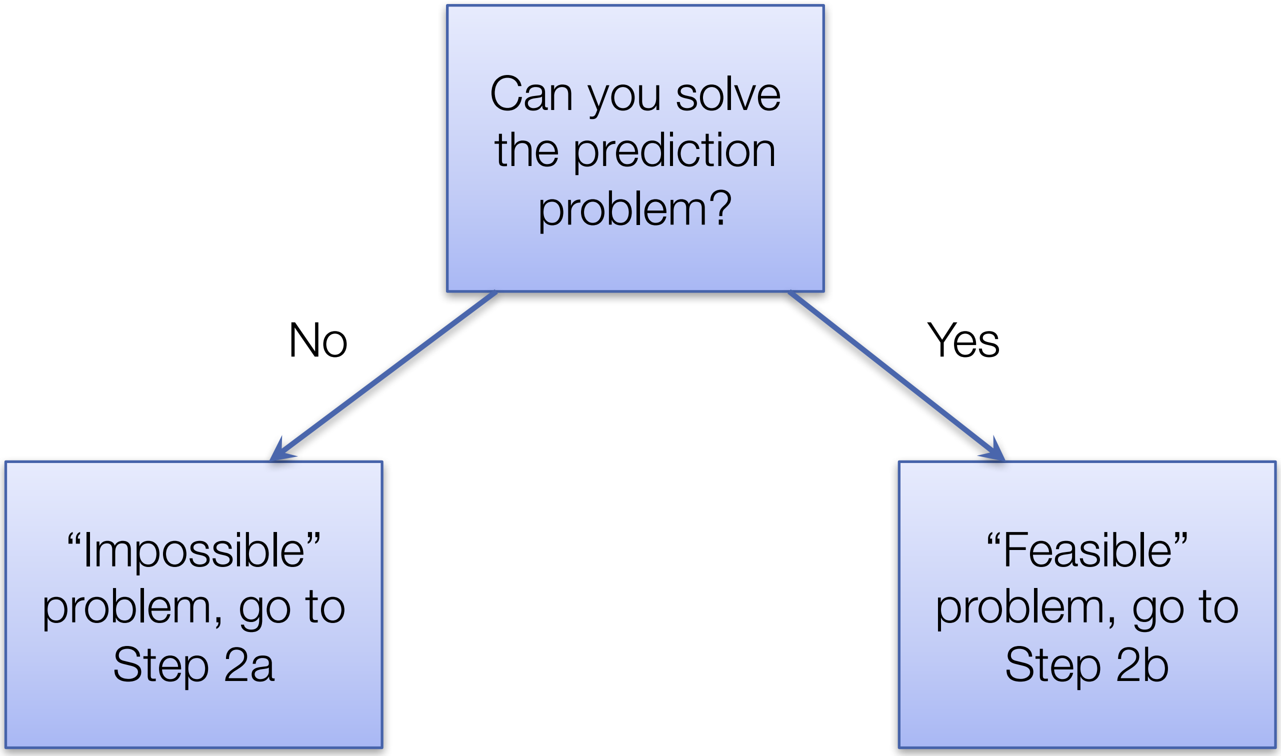
What about "superhuman" machine learning

It's a common misconception that machine learning will *outperform* human experts on most tasks

In reality, the benefit from machine learning often doesn't come from superhuman performance in most cases, it comes from the ability to scale up expert-level performance extremely quickly

If you can't make good predictions, neither will a machine learning algorithm (at least the first time through, and probably always)

Decision diagram



Dealing with "impossible" problems

So you've built a tool to manually classify examples, run through many cases (or had a domain expert run through them), and you get poor performance

What do you do?

You do *not* try to throw more, bigger, badder, machine learning algorithms at the problem

Instead you need to change the problem by: 1) changing the input (i.e., the features), 2) changing the output (i.e., the problem definition)

Changing the input (i.e., adding features)

The fact that we can always add more features is what makes these problems “impossible” (with quotes) instead of impossible (no quotes)

You can always hold out hope that you just one data source away from finding the “magical” feature that will make your problem easy

But you probably aren’t… adding more data is good, but:

1. Do spot checks (visually) to see if this new features can help *you* differentiate between what you were previously unable to predict

2. Get advice from domain experts, see what sorts of data source they use in practice (if people are already solving the problem)

Changing the output (i.e., changing the problem)

Just make the problem easier! (well, still need to preserve the character of the data science problem)

A very useful procedure: instead of trying to predict the future, try to predict what an expert would predict given the features you have available

E.g., for predictive maintenance this shifts the question from: “would this machine fail?” to “would an expert choose to do maintenance on this machine?”

With this strategy we already have an existence proof that it’s feasible

Changing the output #2

Move from a question of getting “good” prediction to a question of characterizing the uncertainty of your predicts

Seems like a cop-out, but many tasks are *inherently* stochastic, the best you can do is try to quantify the likely uncertainty in output given the input

E.g.: if 10% of all machines fail within 90 days, it can still be really valuable to predict if whether a machine will fail with 30% probability

Dealing with feasible problems

Good news! Your prediction problem seems to be solvable (because you can solve it)

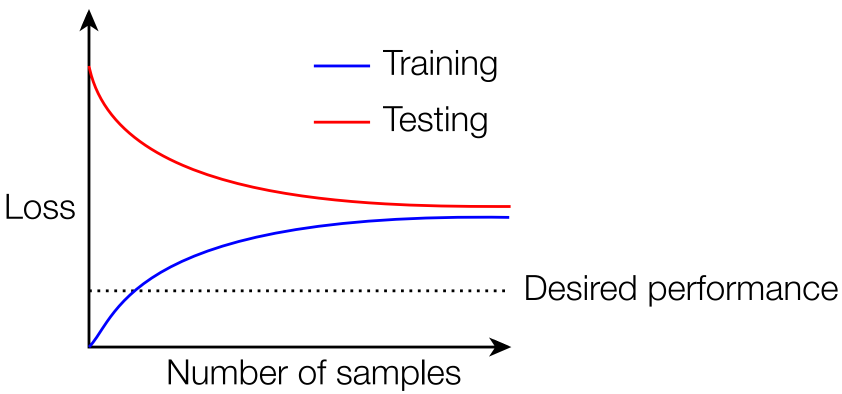
You run your machine learning algorithm, and find that it doesn't work (performs worse than you do)

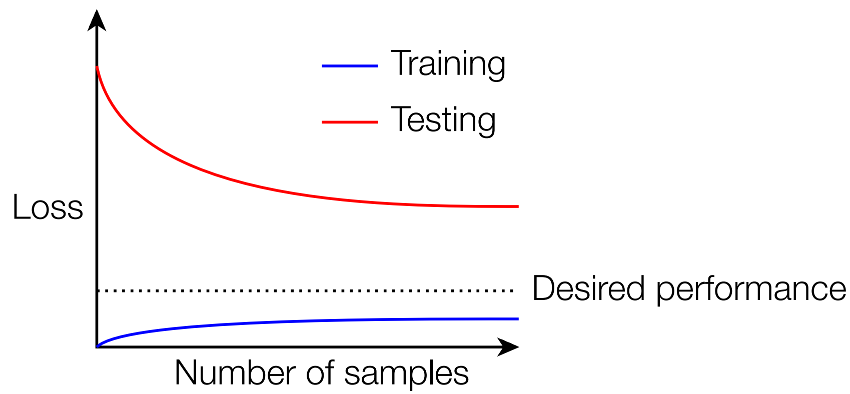
Again, you can try just throwing more algorithms, data, features, etc, at the problem, but this is unlikely to succeed

Instead you want to build diagnostics that can check what the problem may be

Characterizing bias vs. variance

Consider the training and testing loss of your algorithm (often plotting over different numbers of samples), to determine if you problem is one of high bias or high variance



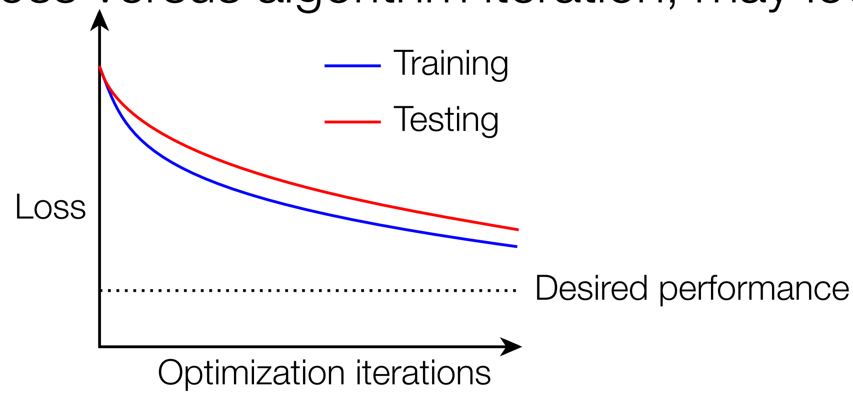


For high bias, add features based upon your own intuition of how you solved the problem

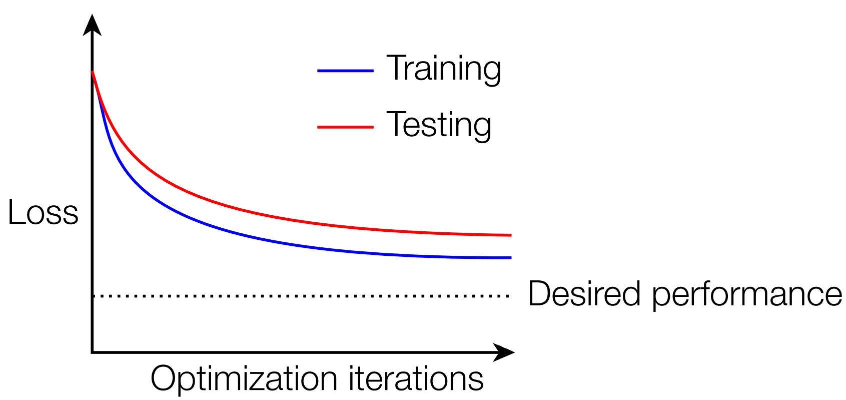
For high variance, add data or remove features (keeping features based upon your intuition)

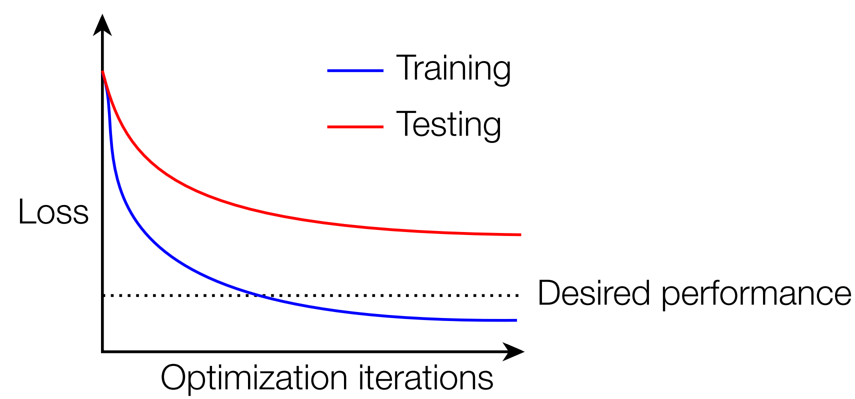
Characterizing optimization performance

It is a much less common problem, but you may want to look at training/testing loss versus algorithm iteration, may look like this:



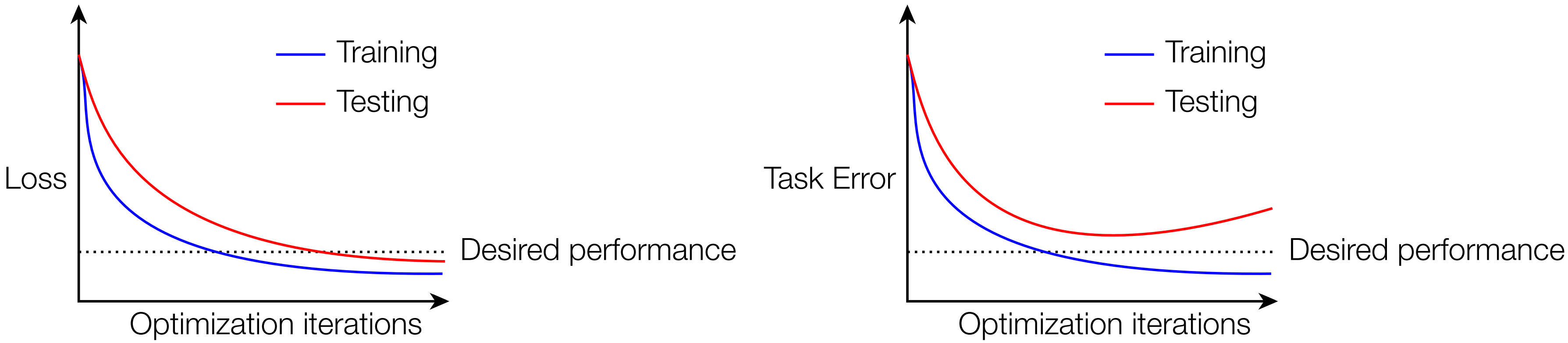
But it probably looks like this:





Consider loss vs. task error

Remember that machine learning algorithms try to minimize some loss, which may be different from the task error you actually want to optimize



This is common when dealing e.g. with imbalanced data sets for which cost of different classifications is very different

THE DREAM

You run your ML algorithm(s) and it works well (?!)

Still: be skeptical …

Very easy to accidentally let your ML algorithm cheat:

- Peaking (train/test bleedover)
- Including output as an input feature explicitly
- Including output as an input feature implicitly

Try to solve the problem by hand;

Try to interpret the ML algorithm / output

Continue being skeptical. Always be skeptical.

OUTLINE

Informed Consent

Reproducibility

p-value Hacking

Who owns the data?

Privacy & Anonymity

Algorithmic fairness

Data validity/provenance

DATA SCIENCE LIFECYCLE: AN ALTERNATE VIEW

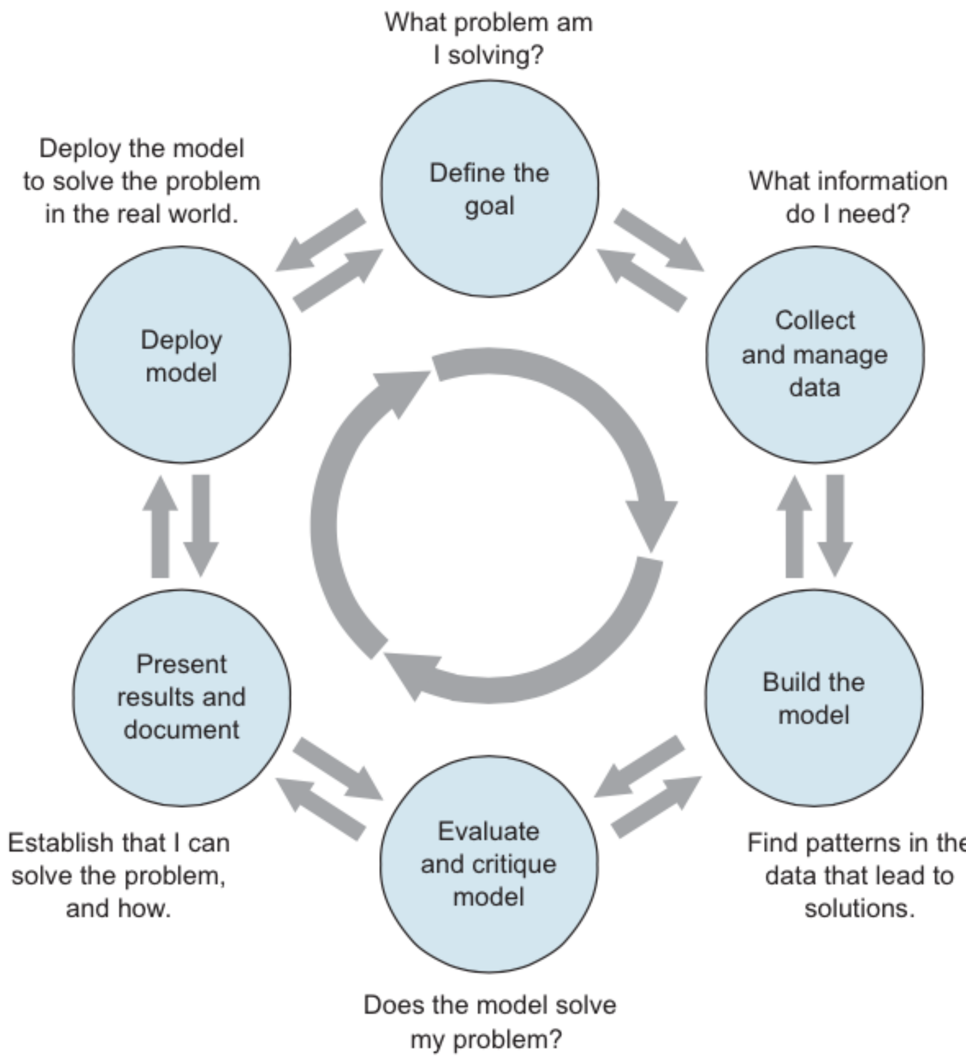


Figure 1.1 The lifecycle of a data science project: loops within loops

COMBATING BIAS

Fairness through blindness:

- Don’t let an algorithm look at protected attributes

Examples currently in use ??????????

- Race
- Gender
- Sexuality
- Disability
- Religion

Problems with this approach ?????????

COMBATING BIAS

“After all, as the former CPD [Chicago Police Department] computer experts point out, the algorithms in themselves are neutral. ‘This program had absolutely nothing to do with race… but multi-variable equations,’ argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound.”

COMBATING BIAS

If there is bias in the training data, the algorithm/ML technique will pick it up

- Especially social biases against minorities
- Even if the the protected attributes are not used

Sample sizes tend to vary drastically across groups

- Models for the groups with less representation are less accurate
- Hard to correct this, and so fundamentally unfair
- e.g., a classifier that performs no better than coin toss on a minority group, but does very well on a majority group

COMBATING BIAS

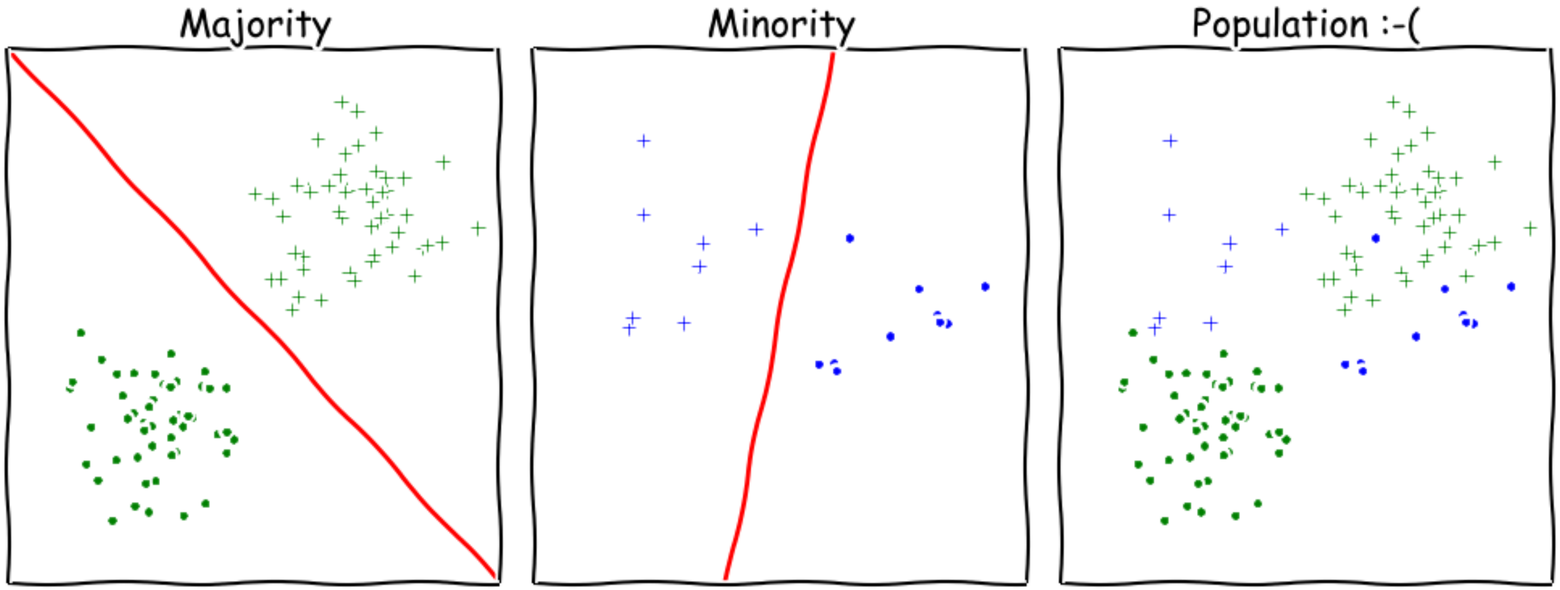
Cultural Differences

- Consider a social network that tried to classify user names into real and fake
- Diversity in names differs a lot – in some cases, short common names are ‘real’, in others long unique names are ‘real’

COMBATING BIAS

Undesired complexity

- Learning combinations of linear classifiers much harder than learning linear classifiers



COMBATING BIAS

Demographic parity:

- A decision must be independent of the protected attribute
- E.g., a loan application's acceptance rate is independent of an applicant's race (but can be depenedent on non-protected features like salary)

Formally: binary decision variable C, protected attribute A

- $P\{ C = 1 \mid A = 0 \} = P\{ C = 1 \mid A = 1 \}$

Membership in a protected class should have no correlation with the final decision.

- Problems ????????

Example from Moritz Hardt's blog

COMBATING BIAS

What if the decision isn't the thing that matters?

"Consider, for example, a luxury hotel chain that renders a promotion to a subset of wealthy whites (who are likely to visit the hotel) and a subset of less affluent blacks (who are unlikely to visit the hotel). The situation is obviously quite icky, but demographic parity is completely fine with it so long as the same fraction of people in each group see the promotion."

Demographic parity allows classifiers that select qualified candidates in the "majority" demographic and unqualified candidate in the "minority" demographic, within a protected attribute, so long as the expected percentages work out.

More: http://blog.mrtz.org/2016/09/06/approaching-fairness.html

Example from Moritz Hardt's blog

FATML

This stuff is really tricky (and really important).

- It's also not solved, even remotely, yet!

New community: Fairness, Accountability, and Transparency in Machine Learning (aka **FATML**)

"… policymakers, regulators, and advocates have expressed fears about the potentially discriminatory impact of machine learning, with many calling for further technical research into the dangers of inadvertently encoding bias into automated decisions."

F IS FOR FAIRNESS

In large data sets, there is always proportionally less data available about minorities.

Statistical patterns that hold for the majority may be invalid for a given minority group.

Fairness can be viewed as a measure of diversity in the combinatorial space of sensitive attributes, as opposed to the geometric space of features.

Thanks to: Faez Ahmed

A IS FOR ACCOUNTABILITY

Accountability of a mechanism implies an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms.

- Current accountability tools were developed to oversee human decision makers
- They often fail when applied to algorithms and mechanisms instead

Example, no established methods exist to judge the intent of a piece of software. Because automated decision systems can return potentially incorrect, unjustified or unfair results, additional approaches are needed to make such systems accountable and governable.

Thanks to: Faez Ahmed

T IS FOR TRANSPARENCY

Automated ML-based algorithms make many important decisions in life.

- Decision-making process is opaque, hard to audit

A transparent mechanism should be:

- understandable;
- more meaningful;
- more accessible; and
- more measurable.

Thanks to: Faez Ahmed

DATA COLLECTION

What data should (not) be collected

Who owns the data

Whose data can (not) be shared

What technology for collecting, storing, managing data

Whose data can (not) be traded

What data can (not) be merged

What to do with prejudicial data

Thanks to: Kaiser Fung

DATA MODELING

Data is biased (known/unknown)

- Invalid assumptions
- Confirmation bias

Publication bias

- WSDM 2017: https://arxiv.org/abs/1702.00502

Badly handling missing values

Thanks to: Kaiser Fung

DEPLOYMENT

Spurious correlation / over-generalization

Using "black-box" methods that cannot be explained

Using heuristics that are not well understood

Releasing untested code

Extrapolating

Not measuring lifecycle performance (concept drift in ML)

We will go over ways to counter this in the ML/stats/hypothesis testing portion of the course

Thanks to: Kaiser Fung

GUIDING PRINCIPLES

Start with clear user need and public benefit

Use data and tools which have minimum intrusion necessary

Create robust data science models

Be alert to public perceptions

Be as open and accountable as possible

Keep data secure

GOV.UK

Thanks to: UK cabinet office

SOME REFERENCES

Presentation on ethics and data analysis, Kaiser Fung @ Columbia Univ. http://andrewgelman.com/wp-content/uploads/2016/04/fung_ethics_v3.pdf

O'Neil, Weapons of math destruction. https://www.amazon.com/Weapons-Math-Destruction-Increases-Inequality/dp/0553418815

UK Cabinet Office, Data Science Ethical Framework. https://www.gov.uk/government/publications/data-science-ethical-framework

Derman, Modelers' Hippocratic Oath. http://www.iijournals.com/doi/pdfplus/10.3905/jod.2012.20.1.035

Nick D's MIT Tech Review Article. https://www.technologyreview.com/s/602933/how-to-hold-algorithms-accountable/

OUTLINE

Informed Consent

Reproducibility

p-value Hacking

Who owns the data?

Privacy & Anonymity

Algorithmic fairness

Some other issues

Data Science in Industry

DATA VALIDITY/ PROVENANCE

Provenance: a history of how a data item or a dataset came to be

- Also called *lineage*

Crucial to reason about the validity of any results, or to do auditing

Lot of research over the years

- File system/OS-level provenance, data provenance, workflow provenance

Increasing interest in industry, but pretty nascent field

INTERPRETABILITY/ EXPLAINABILITY

Can you explain the results of an ML model?

Easy for decision trees (relatively), nearly impossible for deep learning

Can't use black box models in many domains

- e.g., health care, policy-making

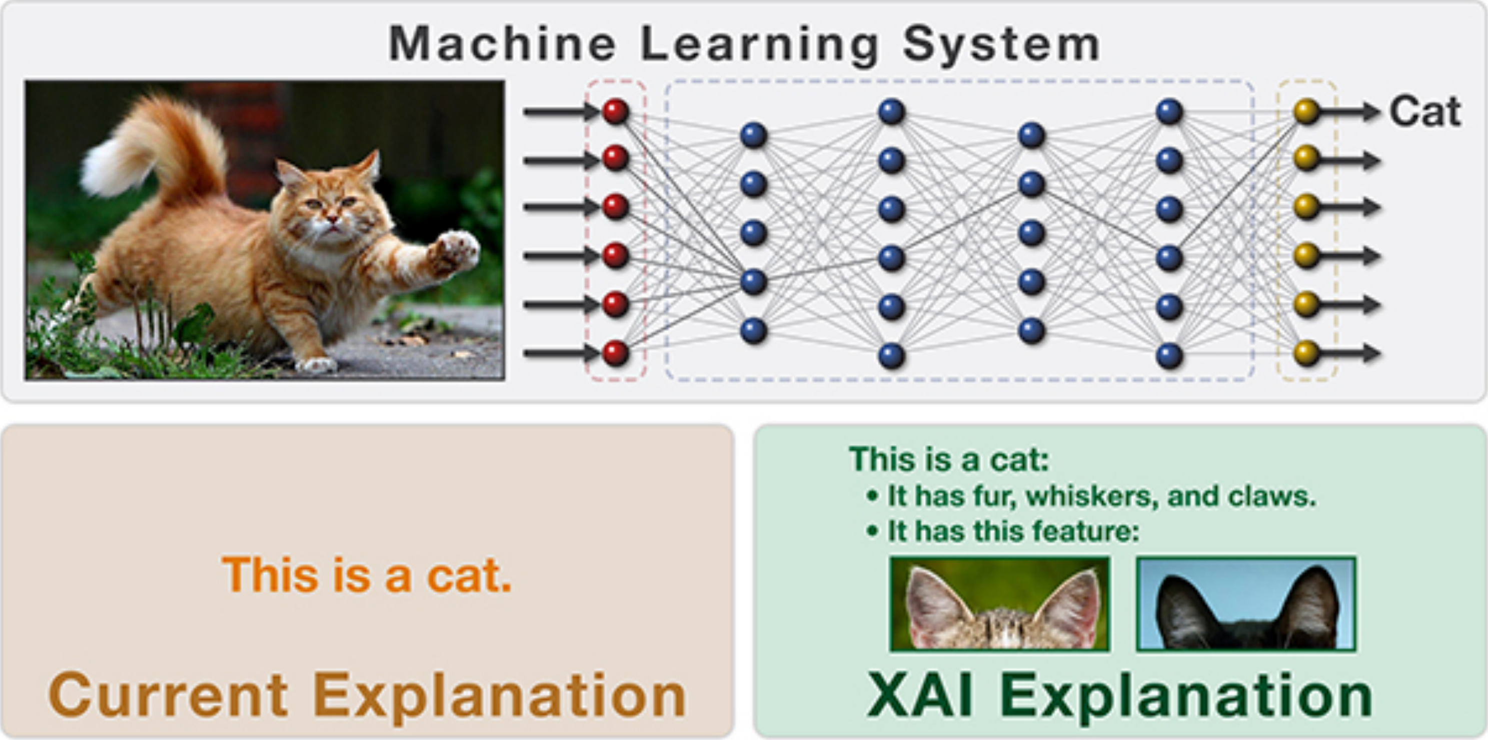
Several recent proposals on simpler models, but those tend to have high error rates

Other proposals on trying to interprete more complex models

- Evolving area…
- Big DARPA project: Explainable AI

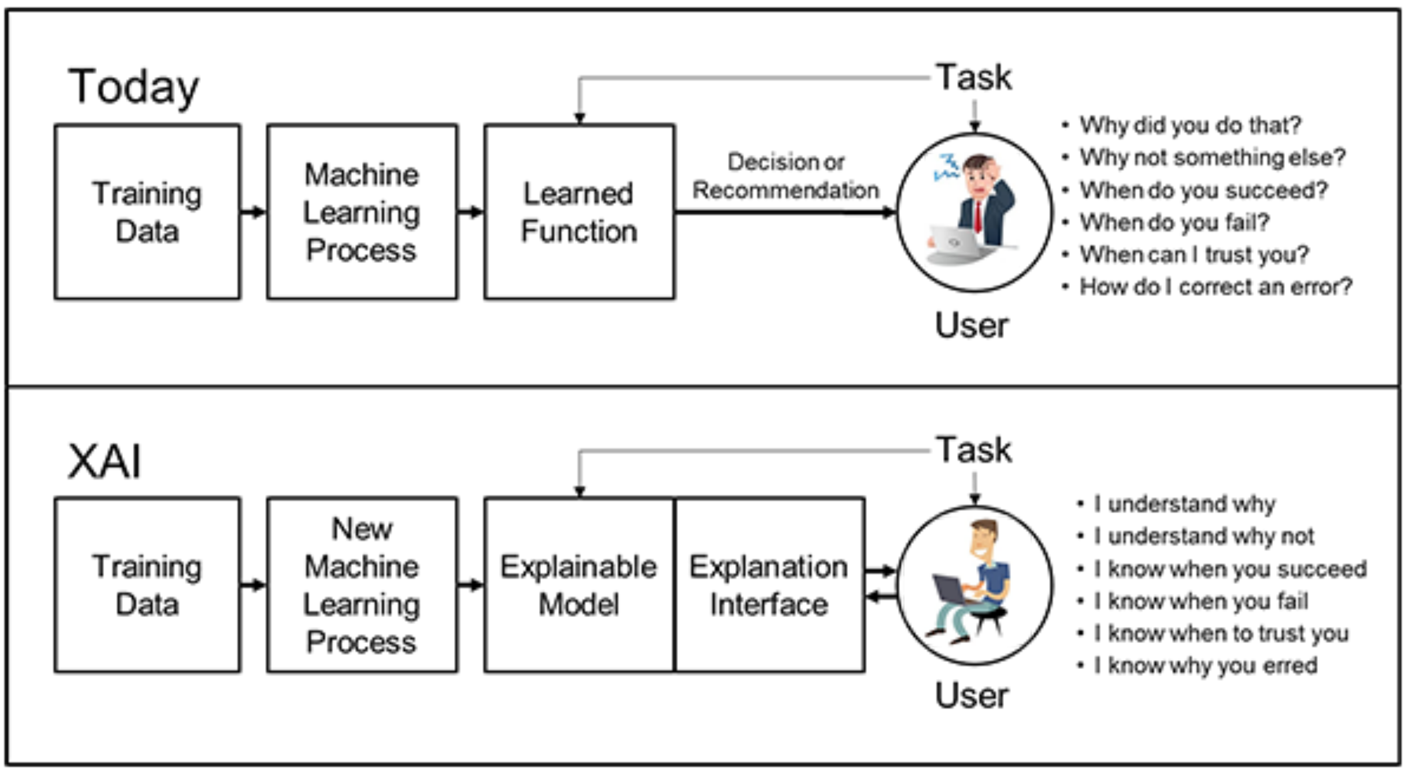
INTERPRETABILITY/ EXPLAINABILITY

From https://www.darpa.mil/program/explainable-artificial-intelligence



INTERPRETABILITY/ EXPLAINABILITY

From https://www.darpa.mil/program/explainable-artificial-intelligence



OUTLINE

Informed Consent

Reproducibility

p-value Hacking

Who owns the data?

Privacy & Anonymity

Algorithmic fairness

Some other issues

Data Science in Industry

WHAT IS A DATA SCIENTIST?

Many types of "data scientists" in industry …

- **Business analysts, renamed**
    - "… someone who analyzes an organization or business domain (real or hypothetical) and documents its business or processes or systems, assessing the business model or its integration with technology." – Wikipedia
- **Statisticians**
- **Machine learning engineer**
- **Backend tools developer**

Thanks to: Zico Kolter

KEY DIFFERENCES

Classical statistics vs machine learning approaches

- (Two are nearly mixed in most job calls you will see.)

Developing data science tools vs. doing data analysis

Working on a core business product vs more nebulous "identification of value" for the firm

FINDING A JOB

Make a personal website.

- Free hosting options: GitHub Pages, Google Sites
- Pay for your own URL (but not the hosting).
- Make a clean website, and make sure it renders on mobile:
    - Bootstrap: https://getbootstrap.com/
    - Foundation: http://foundation.zurb.com/

Highlight relevant coursework, open source projects, tangible work experience, etc

Highlight tools that you know (not just programming languages, but also frameworks like TensorFlow and general tech skills)

“REQUIREMENTS”

Data science job postings – and, honestly, CS postings in general – often have completely nonsense requirements

1. The group is filtering out some noise from the applicant pool
2. Somebody wrote the posting and went buzzword crazy

In most cases (unless the position is a team lead, pure R&D, or a very senior role) you can work around requirements:

- A good, simple website with good, clean projects can work wonders here …
- Reach out and speak directly with team members
- Alumni network, internship network, online forums

INTERVIEWING

We saw that there is no standard for being a "data scientist" – and there is also no standard interview style …

… **but, generally, you'll be asked about the five "chunks" we covered in this class, plus core CS stuff:**

- Software engineering questions
- Data collection and management questions (SQL, APIs, scraping, newer DB stuff like NoSQL, Graph DBs, etc)
- General "how would you approach …" EDA questions
- Machine learning questions ("general" best practices, but you should be able to describe DTs, RFs, SVM, basic neural nets, KNN, OLS, boosting, PCA, feature selection, clustering)
- Basic "best practices" for statistics, e.g., hypothesis testing

Take-home data analysis project (YMMV)

GRADUATE SCHOOL, ACADEMIA, R&D, …

Data science isn't really an academic discipline by itself, but it comes up everywhere within and without CS

- **Modern science is built on a "CS and Statistics stack"** …

Academic work in the area:

- **Outside of CS, using techniques from this class to help fundamental research in that field**
- **Within CS, fundamental research in:**
  - Machine learning
  - Statistics (non-pure theory)
  - Databases and data management
  - Incentives, game theory, mechanism design
- **Within CS, trying to automate data science (e.g., Google Cloud's Predictive Analytics, "Automatic Statistician,"** …**)**

CONCLUSIONS

Final project due in 2 weeks

Will send out a survey in a few days – please complete it

Sign up for remaining courses

Converting to MS