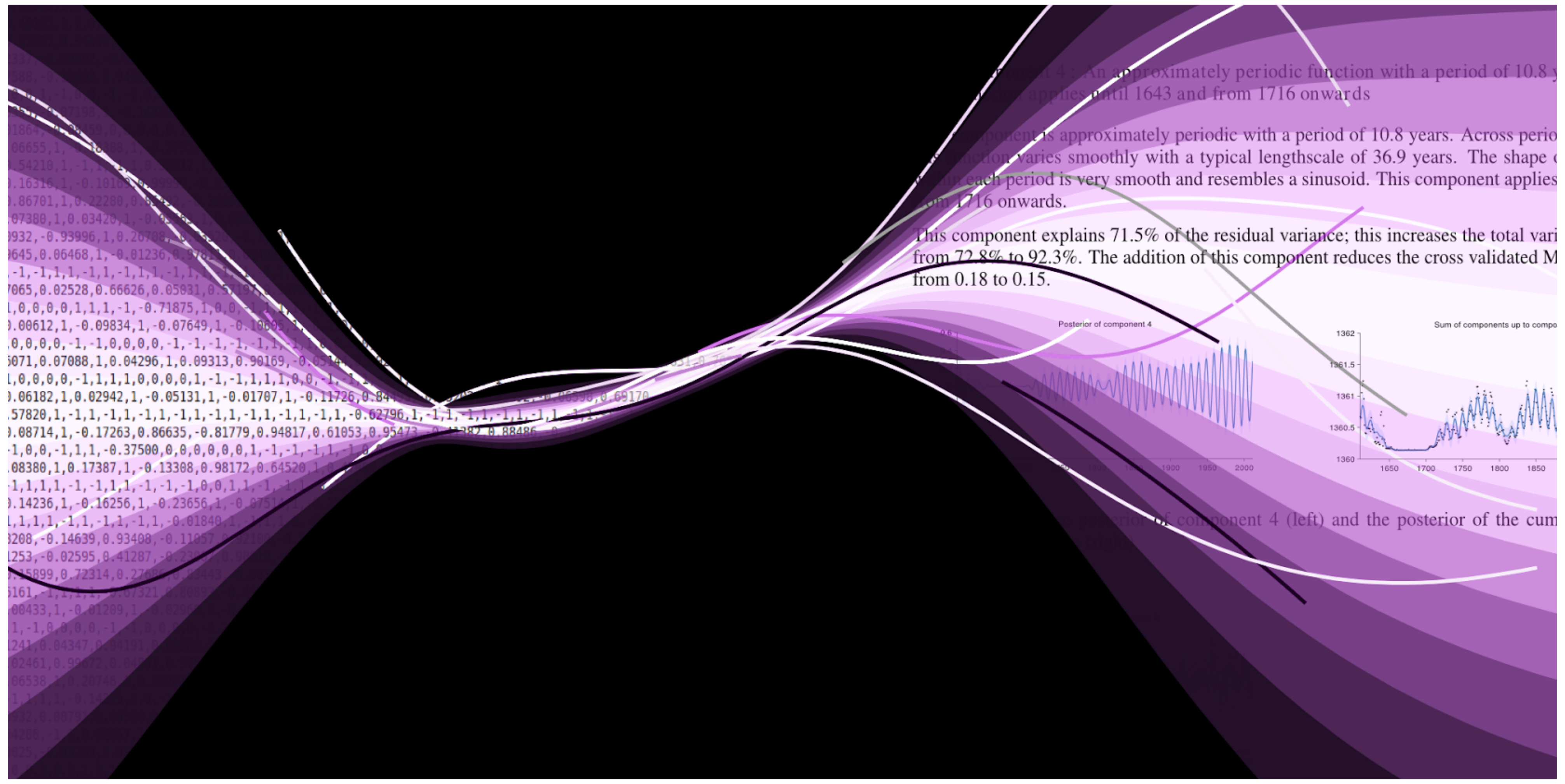
JSC370 and JSC470: Data Science II and III


An approximately periodic function with a period of 10.8 y
applies until 1643 and from 1716 onwards
is approximately periodic with a period of 10.8 years. Across perio
varies smoothly with a typical lengthscale of 36.9 years. The shape o
period is very smooth and resembles a sinusoid. This component applies
1716 onwards.
This component explains 71.5% of the residual variance; this increases the total vari
from 72.8% to 92.3%. The addition of this component reduces the cross validated M
from 0.18 to 0.15.
Posterior of component 4
1900
1950
2000
Sum of components up to compo
1362
1361.5
1361
1360.5
1360
1650
1700
1750
1800
1850
component 4 (left) and the posterior of the cum

David Duvenaud
January 2021

Lecture 2

Lecture Topics (Tentative)

- Confounding, censoring, causality, and the do-calculus
- Latent variable models + not-missing-at-random data
- Decision theory and Goodhart's law
- Natural Language Processing
- Using large off-the-shelf models
- Outlier detection
- Time Series Models + Validation
- Reproducibility and version control for data

Assignment 1

- Released tomorrow
- Build a recommender system, in context
- Focus on:
    - Connections to larger business and operational questions
    - Dealing with confounding

Parts of a good analysis

- What value you could add?
- How to measure success?
- Look at the data.
- Look for complementary sources of data.
- Brainstorm an everything model.
- Propose a model staircase (series of more sophisticated models)
- Fit the models + do sanity checks
- Report evidence it works, expected value added, conditions for accurate use

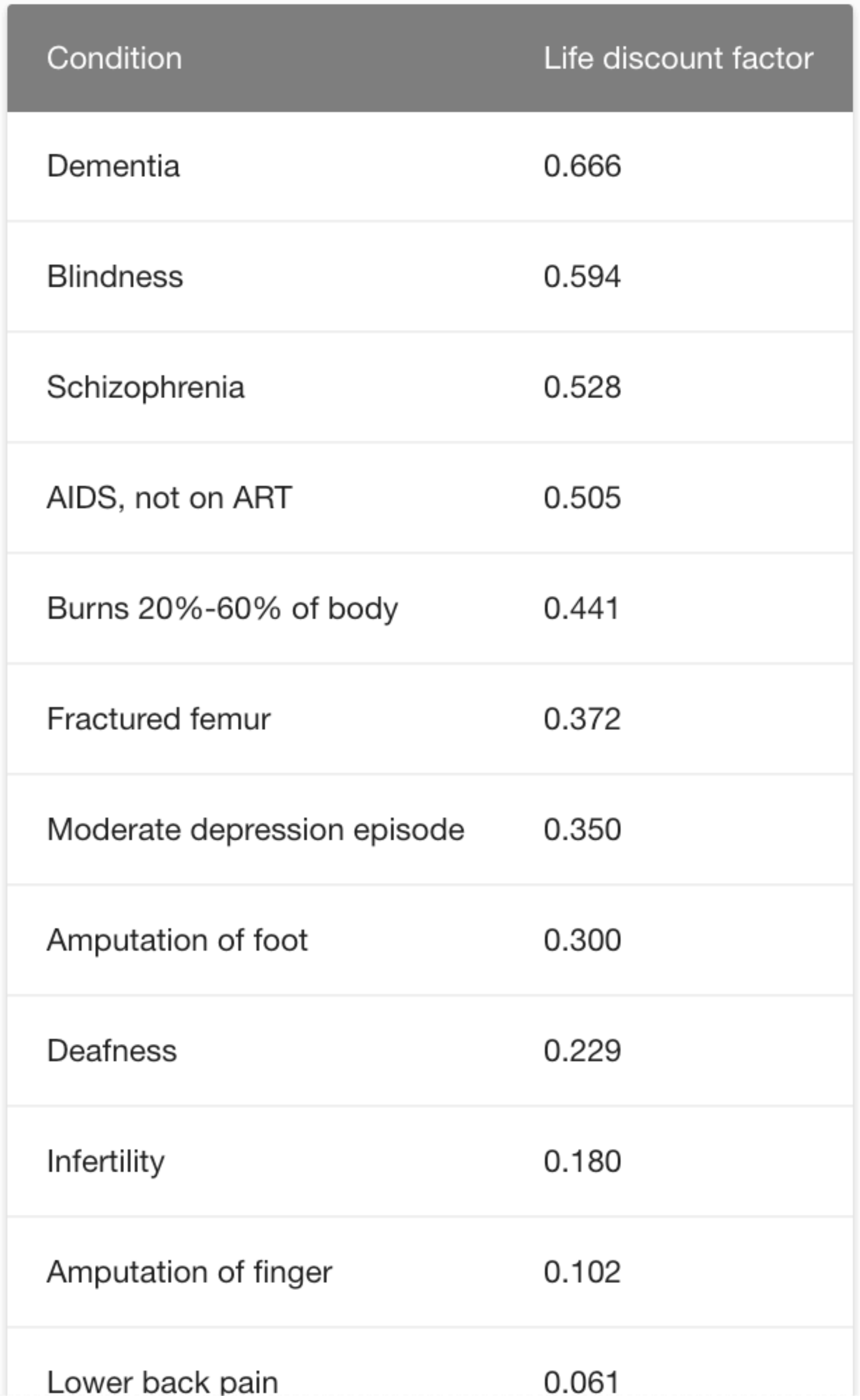
Parts of a good analysis: Value Add

- Find out what kind of value you can add, in principle
  - Make recommendations
  - automate decisions, audit decisions, aid decisions
  - Blue-sky research, provide context for proposals

- What would happen if you did nothing?

- What actions might you change based on info / analysis? Only worth doing any thinking / info gathering if you expect it will change actions.

- Get order-of-magnitude estimates of impact with made-up numbers



Condition	Life discount factor
Dementia	0.666
Blindness	0.594
Schizophrenia	0.528
AIDS, not on ART	0.505
Burns 20%-60% of body	0.441
Fractured femur	0.372
Moderate depression episode	0.350
Amputation of foot	0.300
Deafness	0.229
Infertility	0.180
Amputation of finger	0.102
Lower back pain	0.061

Source: WHO

Designing Sensible Metrics

- Brainstorm ways to measure success.

- What would success look like, big picture? E.g. person is healthy, company makes money + provides value

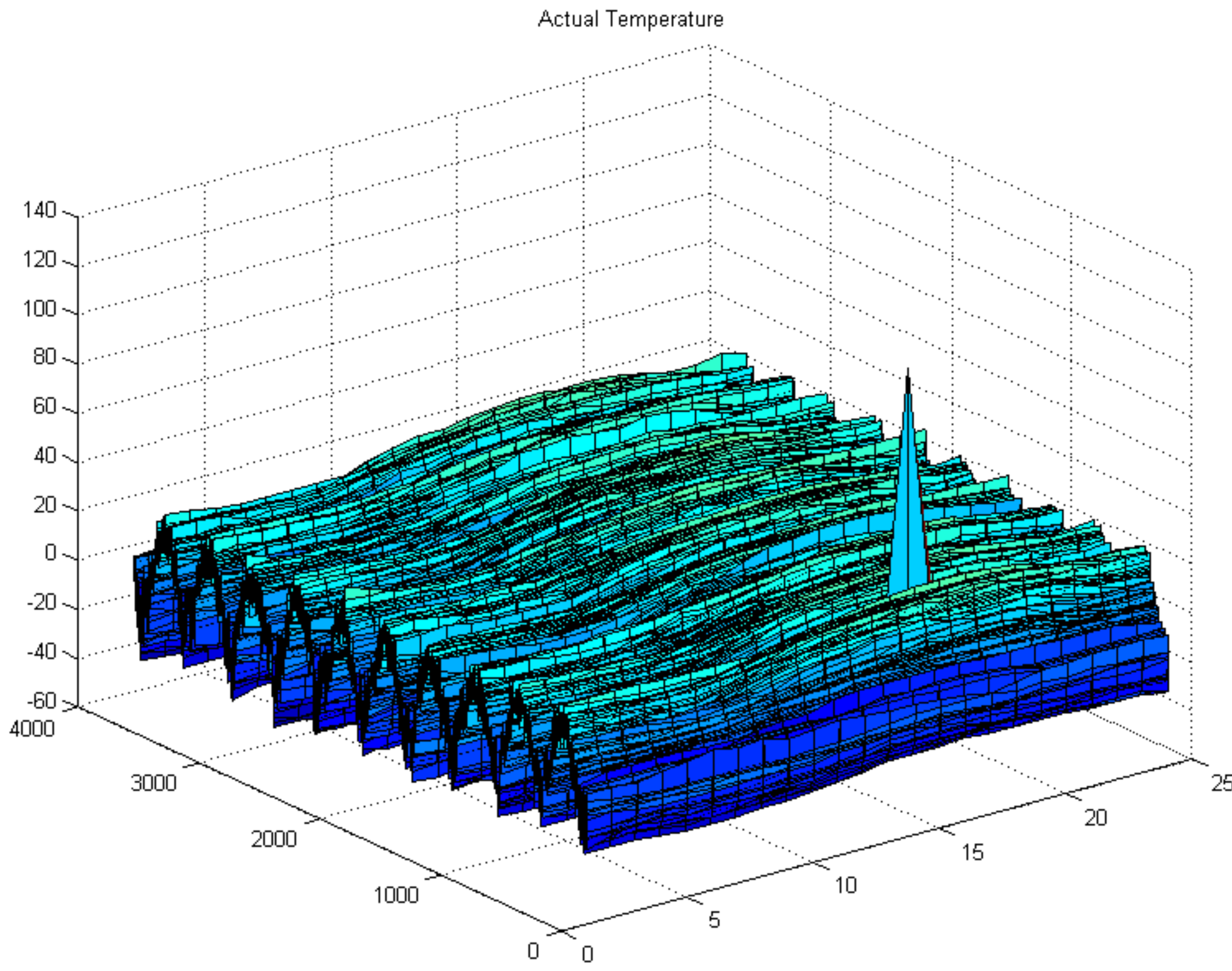
- What are some correlates / preconditions of that success. E.g. accuracy at diagnosis.

- What are relevant UI / rollout details that determine correlation of metrics with value added? E.g. Accuracy@top K where K is the number of items shown.

- Any "known good" decisions, predictions, to calibrate or learn from?

Look at your data

- Plot a few of each type of data you have
- Histogram anything that can be binned (ratings, number of ratings)
- Sanity check data
  - E.g. find a movie you've seen, check its ratings + description
- Look at extreme examples
  - movies + users with most and fewest ratings
- Check for time-dependence (e.g. average rating over time)
- Check for missing / incomplete data (users, movies with no ratings)



Look at data collection pipeline

- Find out as much as you can about where it came from.
- What did the UI look like when the data was created? What were the options, the context, the stakes, distractions, instructions.
- Did anything change during the data collection process? Ask about feedback loops.
- Write all this down and who told it to you.
- Find sanity checks - Lizardman's constant is 4%

Find Complementary Data

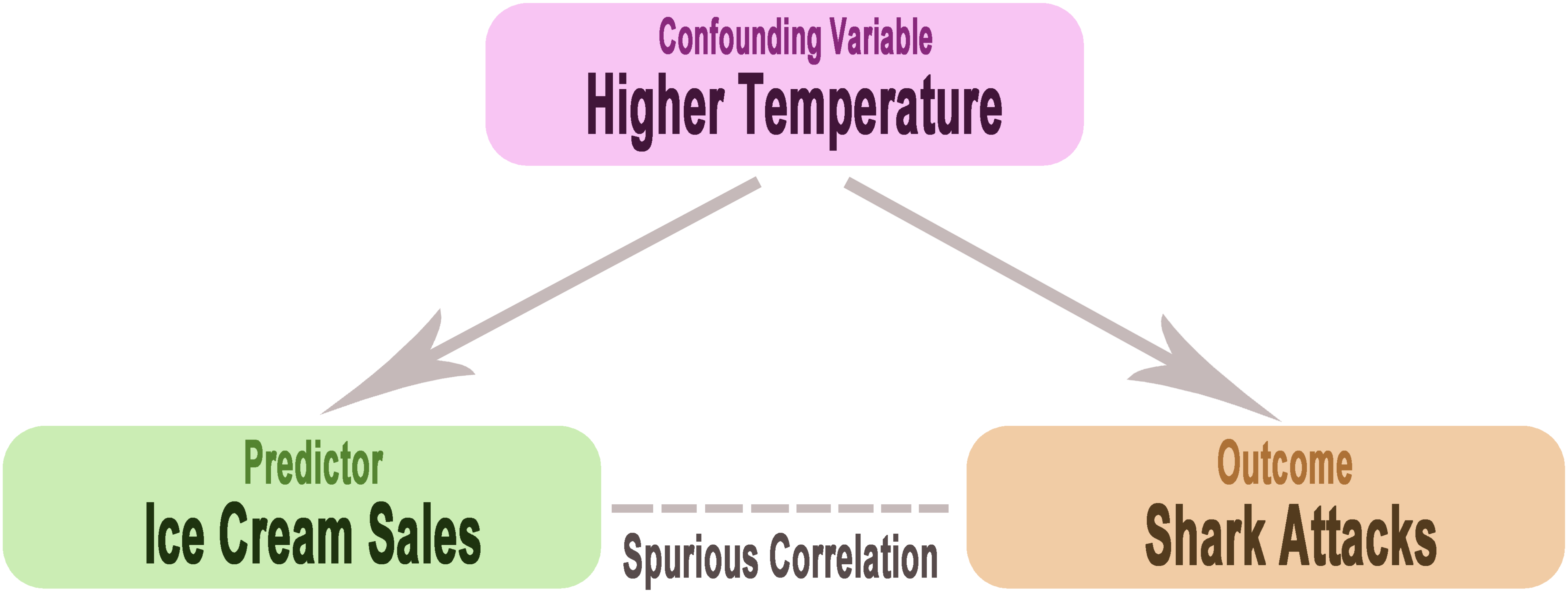
- Ask what data, in principle, could help identify / route around some of the relevant confounders (e.g. RCTs)
  - Similar analyses done elsewhere.
  - E.g. for movies: Scrape IMBD.

- Consider global variables to condition on, e.g. weekends, holidays, price changes, selection increasing, seasons changing, country of users, language, age-related laws

Sketch a kitchen-sink model

- Write down a somewhat-realistic all-encompassing model.

- Try to brainstorm all possible unobserved confounders.
  - E.g. wealth, mood, relationships, health, genes, treatments, accidents, interference by others, foul play, corruption, deliberate misinformation, sabotage, ignorance, misunderstandings, illiteracy, weather, power outages...

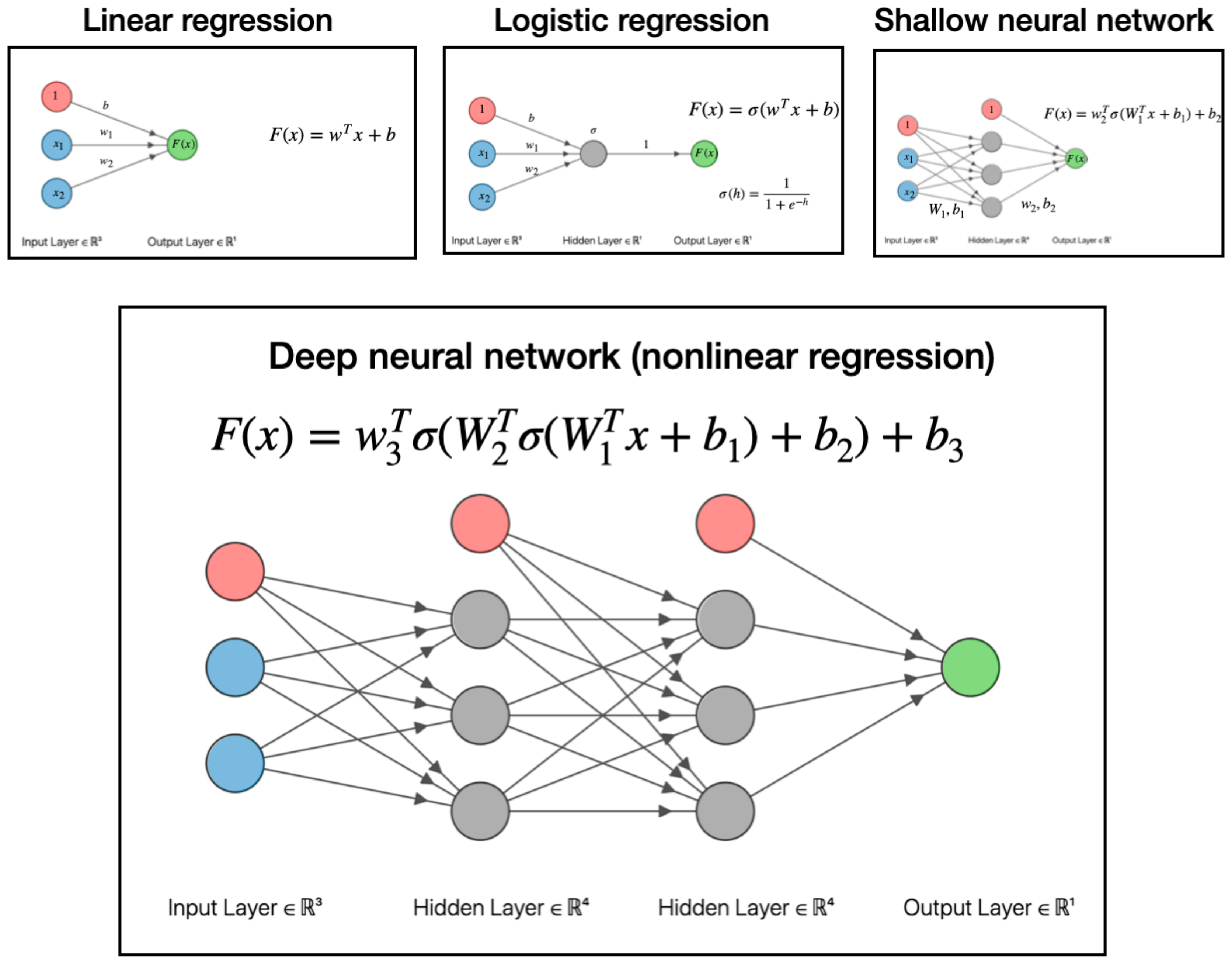


Credit: Vivekananda Das

Propose a Model Staircase

- Write down a series of more and more sophisticated models, starting with stupidly simple ones. E.g. start with everything i.i.d.
  - Prioritize the parts of the model that are most related to the outcomes you want to measure / predict.
  - Can make up coarse-grained versions of variables, e.g. movie content -> overall movie quality.

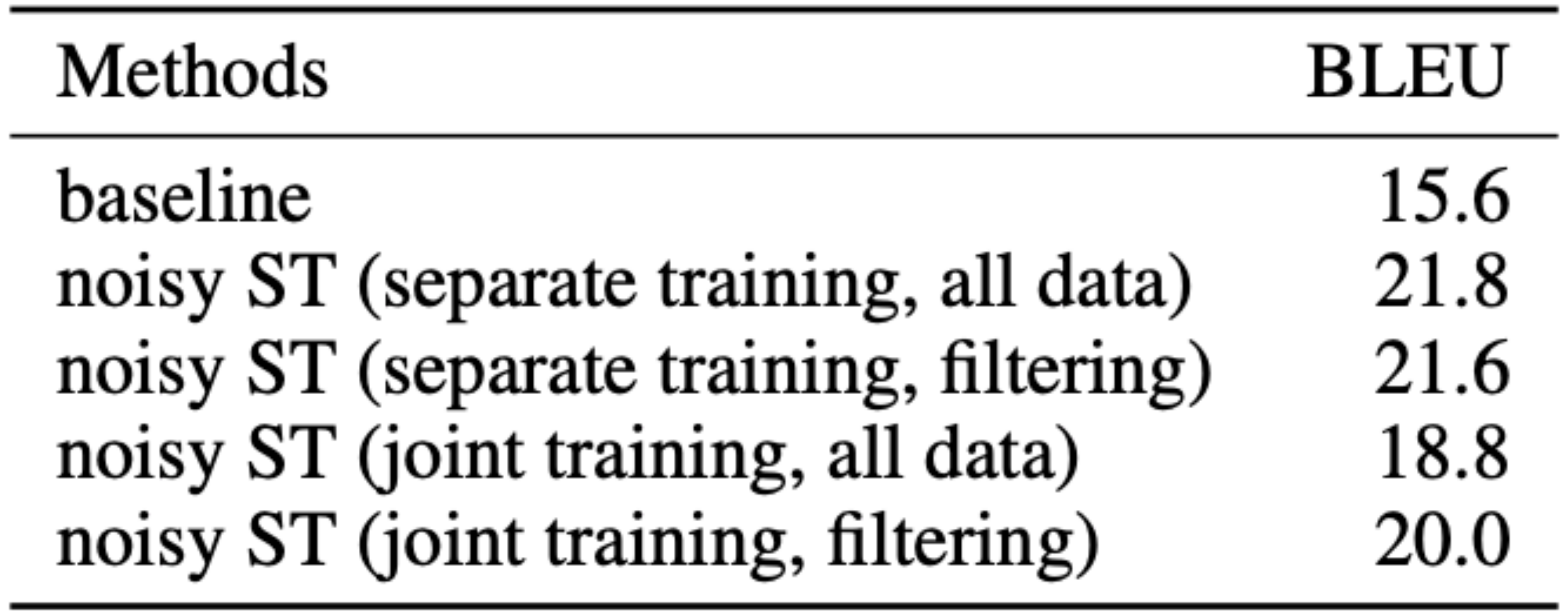
- Every piece of complexity must be justified + checked.
  - A simple, convincing analysis will be used.
  - A complicated, hard-to-verify analysis will usually be ignored.
  - Can check e.g. latent variables by showing they correspond to known aspects of the system.



Credit: Joshua Goings

Propose a Model Staircase

- Biggest engineering failures: a giant, do-everything, all-or-nothing fancy model.
- Quicker feedback. Things only ever work when we have lots of feedback.
- Can give any-time results.
- Simpler models usually faster to fit, require less data, more interpretable, easier to maintain.
- Can justify complexity of each step with performance gain.
- A.K.A. Ablation studies



Methods	BLEU
baseline	15.6
noisy ST (separate training, all data)	21.8
noisy ST (separate training, filtering)	21.6
noisy ST (joint training, all data)	18.8
noisy ST (joint training, filtering)	20.0

Table 7: Ablation analysis on WMT100K dataset.

https://openreview.net/pdf?id=SJgdnAVKDH

Example Model Staircase 1

- Ratings are iid
- Ratings are iid for each movie
- Ratings depend on known movie features
- Ratings depends on known movie features + known user features
- Ratings depend on known movie features + known + latent user features
- Ratings depend on known + latent movie features + known + latent user features

Example Model Staircase 2

- Treatment effect is iid
- Mean treatment effect is linear in age
- Mean treatment effect is linear in age + binned by gender
- Mean treatment effect is nonlinear in age + binned by gender

Models vs Decisions

- In principle, can divide analysis into
  1) Model p(result | action)
  2) Choose action = argmax_action E_(result|action)[value of result]

- Aka Bayesian decision theory. Can also incorporate multiple rounds of actions + observation. Optimal exploration / exploitation strategy is not mysterious, just intractable.

Example Decision Procedures

- Propose movie with highest predicted rating for that user
- Propose movie with most uncertain rating for that user
- Combinations of those two
- Run a complicated decision tree
- Propose movie which would most reduce uncertainty about all users on average, weighted by their expected number of watches.

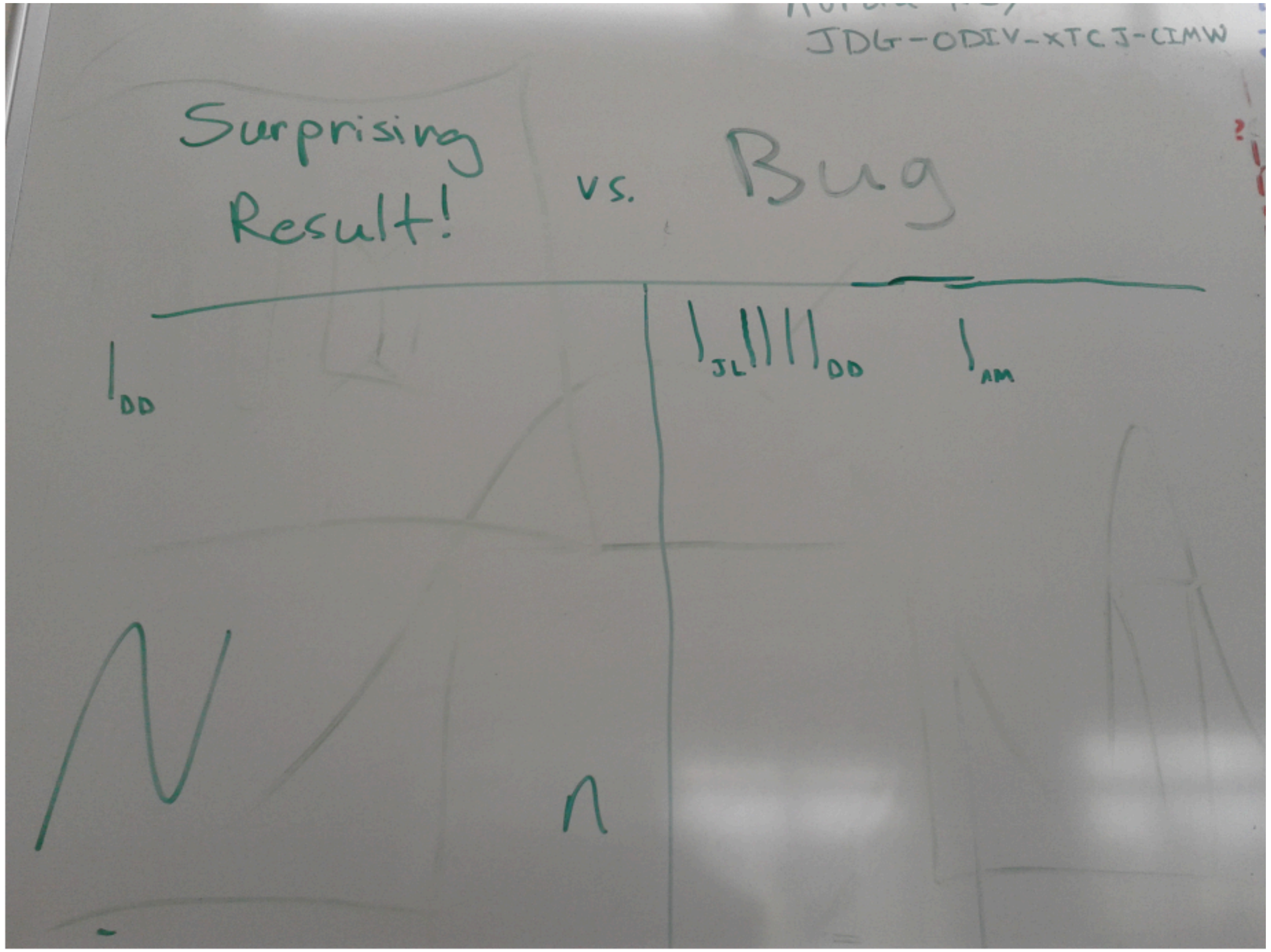
Parts of a good analysis: Fit the model

- Actually code up and fit these models (machine learning / stats courses)

- Might need to come up with cheaper / differentiable version of desired losses. E.g. log-likelihood instead of accuracy.

- Include sanity checks (for your own sake, to convince others, and for monitoring after deployment).

- Antipattern: Looking at only one number and making up a story about why it goes up or down.



Parts of a good analysis

- Report accuracy. Spell out expected value added, in dollars or life-years if possible.

- Report conditions necessary for continued accuracy once in use. E.g. if we change the UI, need to do it only for some users and first and record who used which one. If we change the clinical test we need to label that. Or E.g. If population who uses treatment changes, accuracy can go down.

- Suggest ways to improve the pipeline
  - Data collection (UI, incentives, controls, annotation)
  - Improving evaluation (anticipate Goodhart's law + gaming metrics)

Parts of a good analysis

- What value you could add?
- How to measure success?
- Look at the data.
- Look for complementary sources of data.
- Brainstorm an everything model.
- Propose a model staircase (series of more sophisticated models)
- Fit the models + do sanity checks
- Report evidence it works, expected value added, conditions for accurate use

Questions?