Machine Learning in Production

Goals and Measurement

Exploring Requirements...

Fundamentals of Engineering AI-Enabled Systems

Holistic system view: AI and non-AI components, pipelines, stakeholders, environment interactions, feedback loops

Requirements:
System and model goals
User requirements
Environment assumptions
Quality beyond accuracy
Measurement
Risk analysis
Planning for mistakes

Architecture + design:
Modeling tradeoffs
Deployment architecture
Data science pipelines
Telemetry, monitoring
Anticipating evolution
Big data processing
Human-AI design

Quality assurance:
Model testing
Data quality
QA automation
Testing in production
Infrastructure quality
Debugging

Operations:
Continuous deployment
Contin. experimentation
Configuration mgmt.
Monitoring
Versioning
Big data
DevOps, MLOps

Teams and process: Data science vs software eng. workflows, interdisciplinary teams, collaboration points, technical debt

Responsible AI Engineering

Provenance, versioning, reproducibility

Safety

Security and privacy

Fairness

Interpretability and explainability

Transparency and trust

Ethics, governance, regulation, compliance, organizational culture

Learning goals

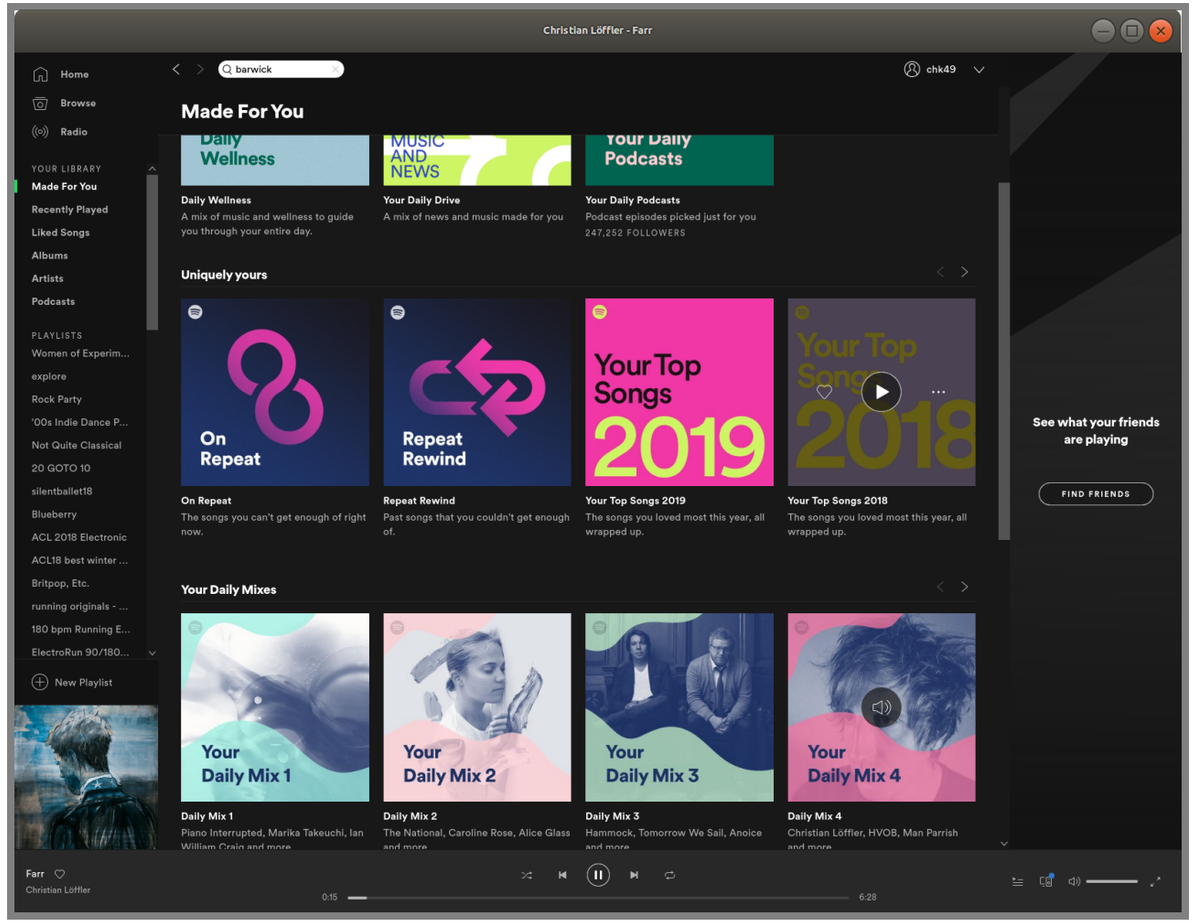
- Judge when to apply ML for a problem in a system
- Define system goals and map them to goals for ML components
- Understand the key concepts and risks of measurement

Readings

Required Readings: Hulten, Geoff. "Building Intelligent Systems: A Guide to Machine Learning Engineering" (2018), Chapters 2 (Knowing when to use IS) and 4 (Defining the IS's Goals)

Suggested complementary reading: Ajay Agrawal, Joshua Gans, Avi Goldfarb. "Prediction Machines: The Simple Economics of Artificial Intelligence" 2018

Today's Case Study: Spotify Personalized Playlists



When to use Machine Learning?

When to use Machine Learning?

When not to use Machine Learning?

Clear specifications are available

Simple heuristics are *good enough*

Cost of building and maintaining the ML system outweighs its benefits (see the technical debt paper)

Correctness is of utmost importance

ML is used only for the hype (e.g., to attract funding)

Examples of these?

Speaker notes

Heuristics: Filtering out profanity in languages

Tasks that are done infrequently or once in a while

Accounting systems, inventory tracking, physics simulations, safety railguards, fly-by-wire

Consider Non-ML Baselines

Consider simple heuristics -- how far can you get?

Consider semi-manual approaches -- cost and benefit?

Consider the system without that feature

Examples:

- Recommending products on Amazon
- Filtering comments with profanity on public forums
- Credit card fraud detection
- Controlling a washing machine

When to use Machine Learning

Big problems: Many inputs, massive scale

Open-ended problems: No single "final" solution; incremental improvements and growth over time

Time-changing problems: Adapting to constant changes, learning with users

Intrinsically hard problems: Unclear rules, heuristics perform poorly

Examples?

Additional Considerations for ML

Partial solution is acceptable: Mistakes are acceptable or mitigable

Data for continuous improvement is available

Predictions can have an influence on system objectives: Does it actually contribute to organizational objectives?

Cost effective: Cheaper than other approaches, or benefits clearly outweigh costs

Examples?

see Hulten, Chapter 2

Spotify: Use of ML?

Big problem? Open ended? Time changing? Hard? Partial solution acceptable? Data continuously available? Influence objectives? Cost effective?

Recidivism: Use of ML?

Big problem? Open ended? Time changing? Hard? Partial solution acceptable? Data continuously available? Influence objectives? Cost effective?

The Business View

Ajay Agrawal, Joshua Gans, Avi Goldfarb. “Prediction Machines: The Simple Economics of Artificial Intelligence” 2018

AI as Prediction Machines

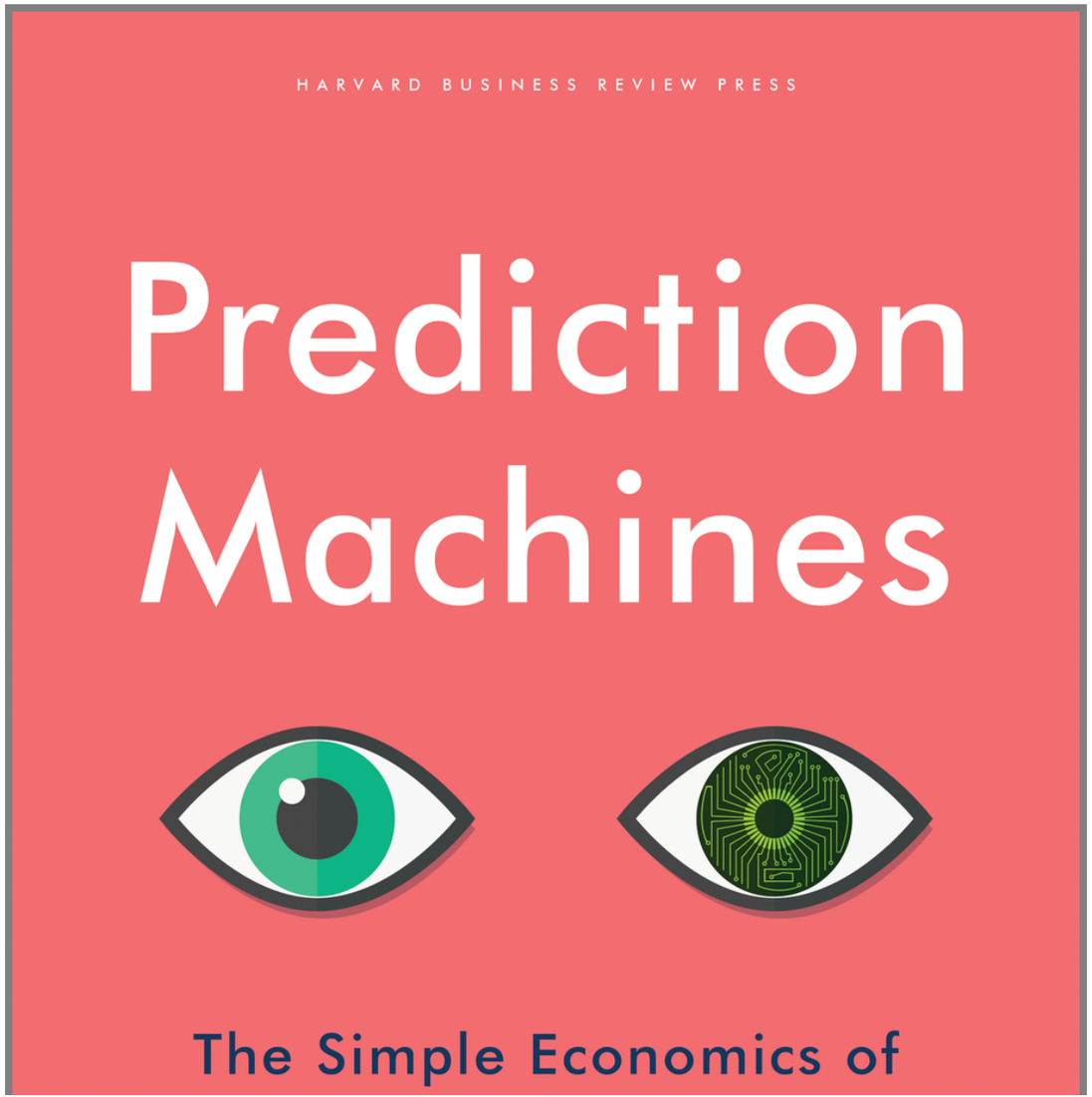
AI: Higher accuracy predictions at much lower cost

May use new, cheaper predictions for traditional tasks (**examples?**)

May now use predictions for new kinds of problems (**examples?**)

May now use more predictions than before

(Analogies: Reduced cost of light; internet reduced cost of search)



Speaker notes

May use new, cheaper predictions for traditional tasks -> inventory and demand forecast; May now use predictions for new kinds of problems -> navigation and translation

The economic lense

- predictions are critical input to decision making (not necessarily full automation)
- decreased price in predictions makes them more attractive for more tasks
- increases the value of data and data science experts
- decreases the value of human prediction and other substitutes
- decreased cost and increased accuracy in prediction can fundamentally change business strategies and transform organizations
  - e.g., a shop sending predicted products without asking
- use of (cheaper, more) predictions can be economic advantage

Predicting the Best Route

Speaker notes

Cab drivers in London invested 3 years to learn streets to predict the fasted route. Navigation tools get close or better at low cost per prediction. While drivers' skills don't degrade, they now compete with many others that use AI to enhance skills; human prediction no longer scarce commodity.

At the same time, the value of human judgement increases. Making more decisions with better inputs, specifying the objective.

Picture source: https://pixabay.com/photos/cab-oldtimer-taxi-car-city-london-203486/

Predictions vs Judgement

Predictions are an input to decision making under uncertainty

Making the decision requires judgement (determining relative payoffs of decisions and outcomes)

Judgement often left to humans ("value function engineering")

ML may learn to predict human judgment if enough data

Automation with predictions

- Automated predictions scale much better than human ones
- Automating prediction vs predict judgement
- Value from full and partial automation, even with humans still required
- Highest return with full automation
  - Tasks already mostly automated, except predictions (e.g. mining)
  - Increased speed through automation (e.g., autonomous driving)
  - Reduction in wait time (e.g., space exploration)
- Liability concerns may require human involvement

Automated decisions desirable but not necessary

The Cost and Value of Data

- (1) Data for training, (2) input data for decisions, (3) telemetry data for continued improving
- Collecting and storing data can be costly (direct and indirect costs, including reputation/privacy)
- Diminishing returns of data: at some point, even more data has limited benefits
- Return on investment: investment in data vs improvement in prediction accuracy
- May need constant access to data to update models

Where to use AI?

- Decompose tasks to identify the use of (or potential use of) predictions
- Estimate the benefit of better/cheaper predictions
- Specify exact prediction task: goals/objectives, data
- Seek automation opportunities, analyze effects on jobs (augmentation, automate steps, shift skills, see taxis)
- Focus on steps with highest return on investment

Cost Per Prediction

What contributes to the average cost of a single prediction?

Examples: Music recommendation, credit card fraud detection, product recommendations on Amazon

Cost Per Prediction

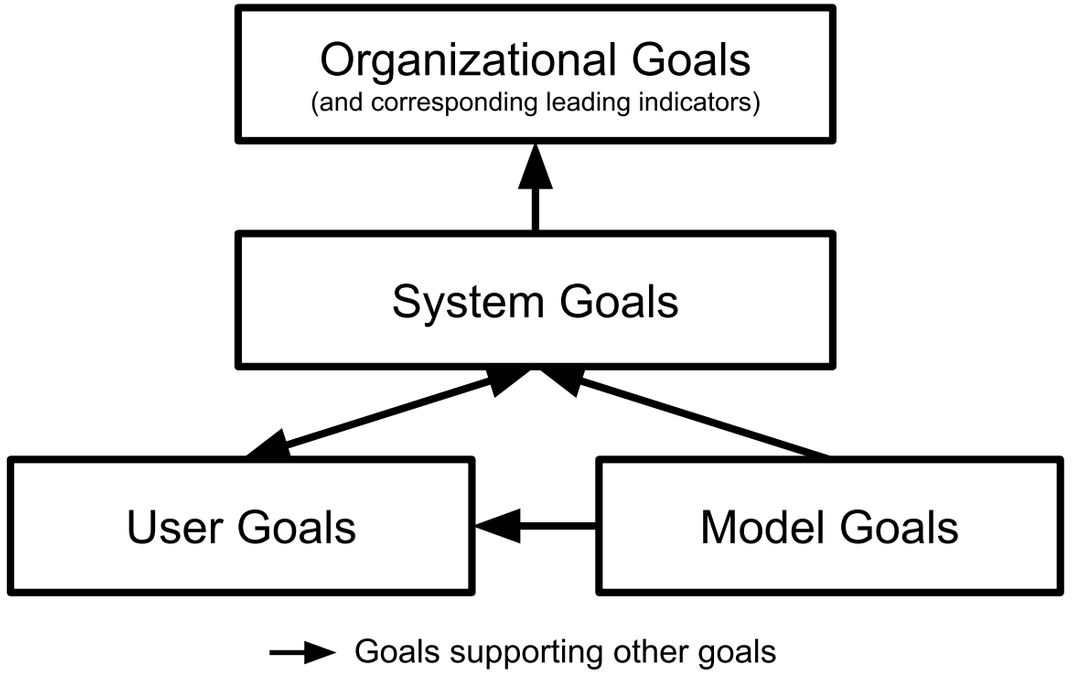
- Useful conceptual measure, factoring in all costs
  - Development cost
  - Data acquisition
  - Learning cost, retraining cost
  - Operating cost
  - Debugging and service cost
  - Possibly: Cost of deadling with incorrect prediction consequences (support, manual interventions, liability)
  - ...

Setting Goals

Layers of Success Measures

- **Organizational objectives:** Innate/overall goals of the organization
- **Leading indicators:** Measures correlating with future success, from the business perspective
- **System goals:** Goals of the software system/feature to be built
- **User outcomes:** How well the system is serving its users, from the user's perspective
- **Model properties:** Quality of the model used in a system, from the model's perspective

Ideally, these goals should be aligned with each other



Organizational Goals

Innate/overall goals of the organization

- Business
  - Current/future revenue, profit
  - Reduce business risks
- Non-Profits
  - Lives saved, animal welfare increased, CO2 reduced, fires averted
  - Social justice improved, well-being elevated, fairness improved
- Often not directly measurable from system output; slow indicators

Implication: Accurate ML models themselves are not the ultimate goal!

ML may only indirectly influence such organizational objectives; influence is often hard to quantify; lagging measures

Leading Indicators

Measures correlating with future success, from the business perspective

Examples:

- Customers sentiment: Do they like the product? (e.g., surveys, ratings)
- Customer engagement: How often do they use the product?
  - Regular use, time spent on site, messages posted
  - Growing user numbers, recommendations

Caveats

- Often indirect, proxy measures
- Can be misleading (e.g., more daily active users => higher profits?)

System/Feature Goals

Concrete outputs the system (or a feature of the system) should produce

Relates to system requirements

Examples:

- Detect cancer in radiology scans
- Provide and recommend music to stream
- Make personalized music recommendations
- Transcribe audio files
- Provide legal help with a self-service chatbot

User Goals

How well the system is serving its users, from the user's perspective

Examples:

- Users choosing recommended items and enjoying them
- Users making better decisions
- Users saving time thanks to the system
- Users achieving their goals

Easier and more granular to measure, but possibly only indirect relation to organization/system objectives

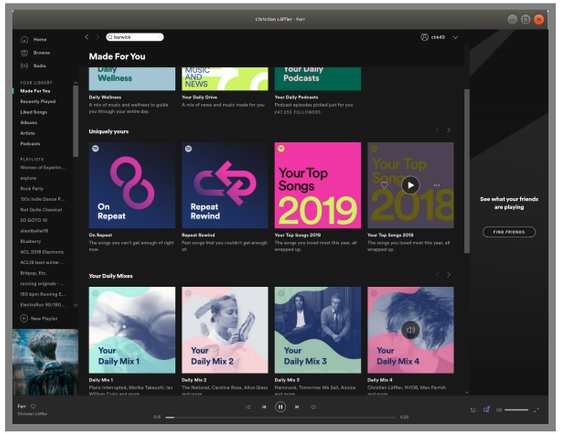
Model Goals

Quality of the model used in a system, from the model's perspective

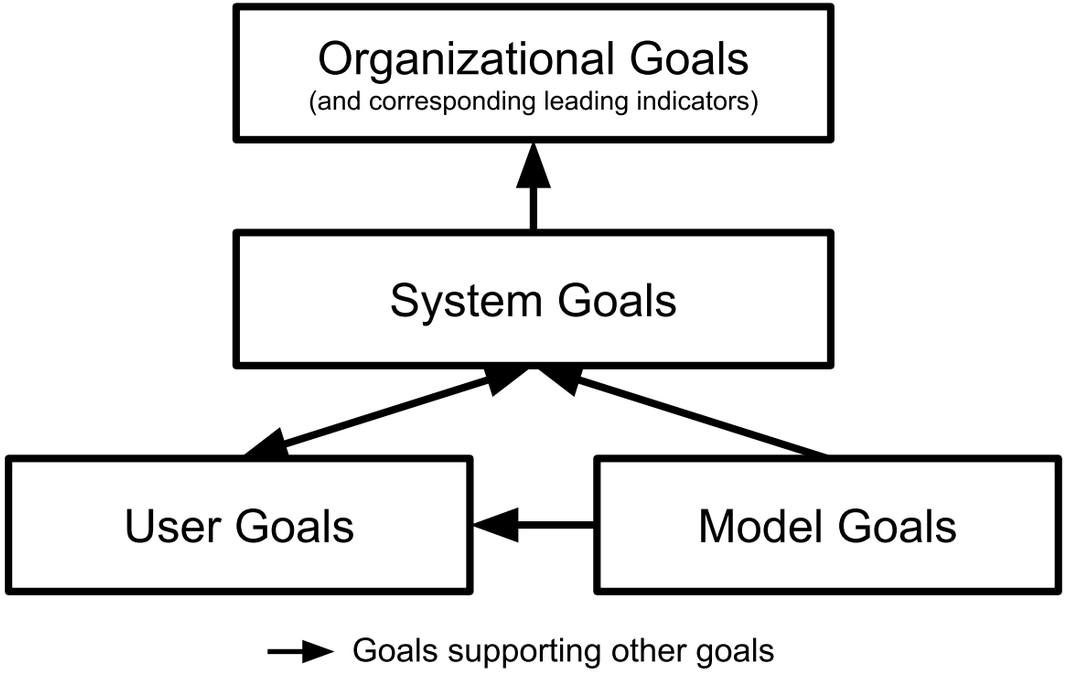
- Model accuracy
- Rate and kinds of mistakes
- Successful user interactions
- Inference time
- Training cost

Often not directly linked to organizational/system/user goals

Success Measures in the Spotify Scenario?



Organizational goals? Leading indicators? System goals? User goals? Model goals?



Speaker notes

Accuracy of song predictions does not necessarily lead to increased user engagement (e.g., if the UI is terrible)

Breakout: Automating Admission Decisions

What are different types of goals behind automating admissions decisions to a Master's program?

As a group post answer to `#lecture` tagging all group members using template:

Measurement

What is Measurement?

Measurement is the empirical, objective assignment of numbers, according to a rule derived from a model or theory, to attributes of objects or events with the intent of describing them. – Craner, Bond, "Software Engineering Metrics: What Do They Measure and How Do We Know?"

A quantitatively expressed reduction of uncertainty based on one or more observations. – Hubbard, "How to Measure Anything …"

Everything is Measurable

1. If X is something we care about, then X, by definition, must be detectable.
    - How could we care about things like “quality,” “risk,” “security,” or “public image” if these things were totally undetectable, directly or indirectly?
    - If we have reason to care about some unknown quantity, it is because we think it corresponds to desirable or undesirable results in some way.
2. If X is detectable, then it must be detectable in some amount.
    - If you can observe a thing at all, you can observe more of it or less of it
3. If we can observe it in some amount, then it must be measurable.

But: Not every measure is precise, not every measure is cost effective

Douglas Hubbard, “How to Measure Anything: finding the value of intangibles in business" 2014

On Terminology

- **Quantification** is turning observations into numbers
- **Metric** and **measure** refer a method or standard format for measuring something (e.g., number of mistakes per hour)
  - Metric and measure synonymous for our purposes (some distinguish metrics as derived from multiple measures, or metrics to be standardizes measures)
- **Operationalization** is identifying and implementing a method to measure some factor (e.g., identifying mistakes from telemetry log file)

Measurement in Software Engineering

- Which project to fund?
- Need more system testing?
- Need more training?
- Fast enough? Secure enough?
- Code quality sufficient?
- Which features to focus on?
- Developer bonus?
- Time and cost estimation?

Measurement in Data Science

- Which model is more accurate?
- Does my model generalize or overfit?
- How noisy is my training data?
- Is my model fair?
- Is my model robust?

Measurement Scales

- Scale: Type of data being measured; dictates what analysis/arithmetic is meaningful
- Nominal: Categories ($=$, $\neq$, frequency, mode, ...)
  - e.g., biological species, film genre, nationality
- Ordinal: Order, but no meaningful magnitude ($<$, $>$, median, rank correlation, ...)
  - Difference between two values is not meaningful
  - Even if numbers are used, they do not represent magnitude!
  - e.g., weather severity, complexity classes in algorithms
- Interval: Order, magnitude, but no definition of zero ($+$, $-$, mean, variance, ...)
  - 0 is an arbitrary point; does not represent absence of quantity
  - Ratio between values are not meaningful
  - e.g., temperature (C or F)
- Ratio: Order, magnitude, and zero ($*$, $/$, log, $\sqrt{}$, geometric mean)
  - e.g., mass, length, temperature (Kelvin)
- Understand scales of features and use an appropriate encoding for learning algorithms!
  - e.g., One-hot encoding for nominal features

Composing/Decomposing Measures

Often higher-level measures are composed from lower level measures

Clear trace from specific low-level measurements to high-level metric

Stating Measures Precisely

- Always be precise about metrics
  - "measure accuracy" -> "evaluate accuracy with MAPE"
  - "evaluate test quality" -> "measure branch coverage with Jacoco"
  - "measure execution time" -> "average and 90%-quantile response time for REST-API x under normal load"
  - "assess developer skills" -> "measure average lines of code produced per day and number of bugs reported on code produced by that developer"
  - "measure customer happiness" -> "report response rate and average customer rating on survey shown to 2% of all customers (randomly selected)"
- Ideally: An independent party should be able to independently set up infrastructure to measure outcomes

Measuring

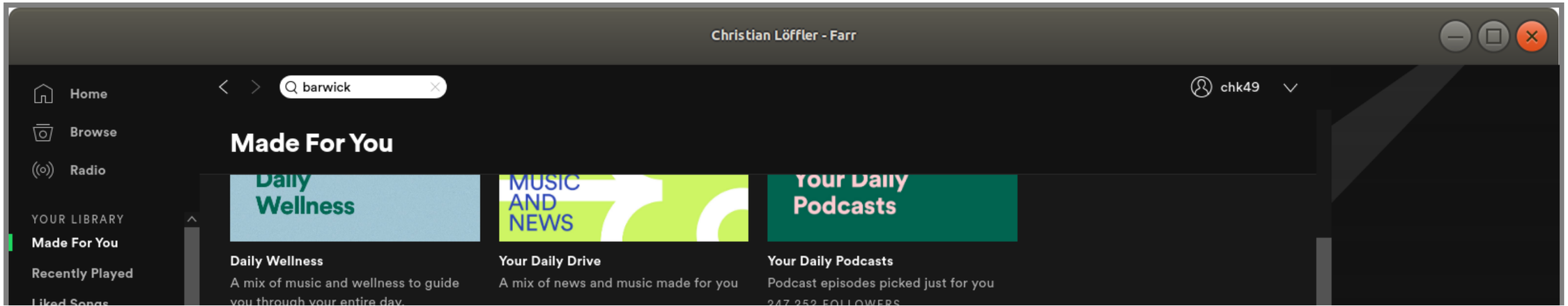
Three ingredients:

1. **Measure:** What do we try to capture?
2. **Data collection:** What data is collected and how?
3. **Operationalization:** How is the measure computed from the data?

Example: Measuring for Spotify Goals?

- Organization goals?
- Leading indicators?
- System goals?
- User goals?
- Model goals?

Identify measure, data collection, and operationalization...



Risks with Measurements

Bad statistics: A basic misunderstanding of measurement theory and what is being measured.

Bad decisions: The incorrect use of measurement data, leading to unintended side effects.

Bad incentives: Disregard for the human factors, or how the cultural change of taking measurements will affect people.

Measurement Validity

Construct validity: Are we measuring what we intended to measure?

- Does the abstract concept match the specific scale/measurement used?
- e.g., IQ: What is it actually measuring?
- Other examples: Pain, language proficiency, personality...

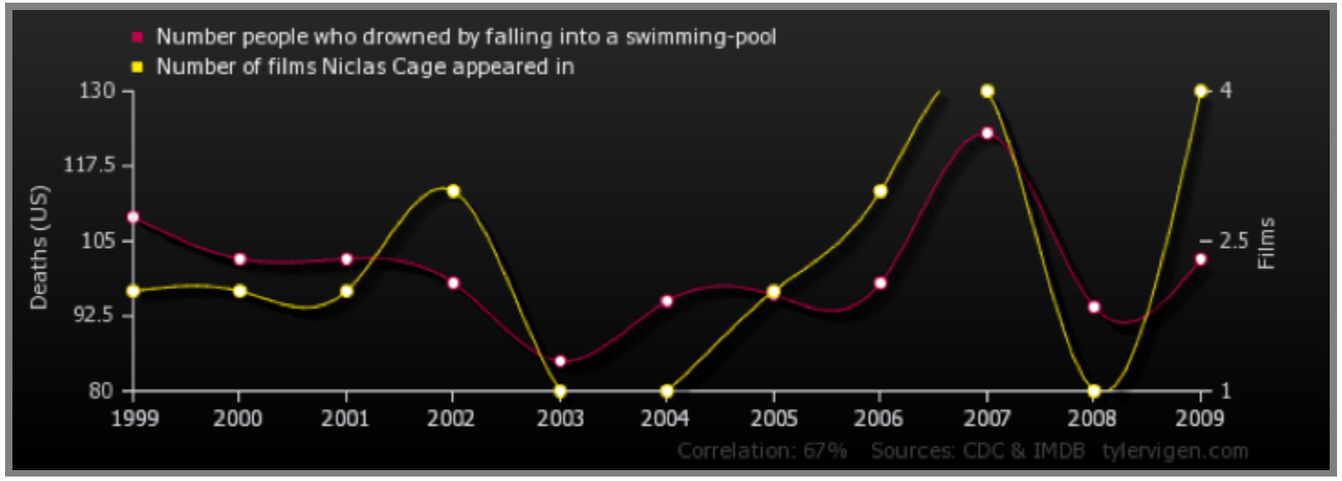
Predictive validity: The extent to which the measurement can be used to explain some other characteristic of the entity being measured

- e.g., Higher SAT scores => higher academic excellence?

External validity: Concerns the generalization of the findings to contexts and environments, other than the one studied

- e.g., Drug effectiveness on a test group: Does it hold over the general public?

Correlation vs Causation



Correlation vs Causation

In general, ML learns correlation, not causation

- (exception: Bayesian networks, certain symbolic AI methods)
- For more details: See causal inference

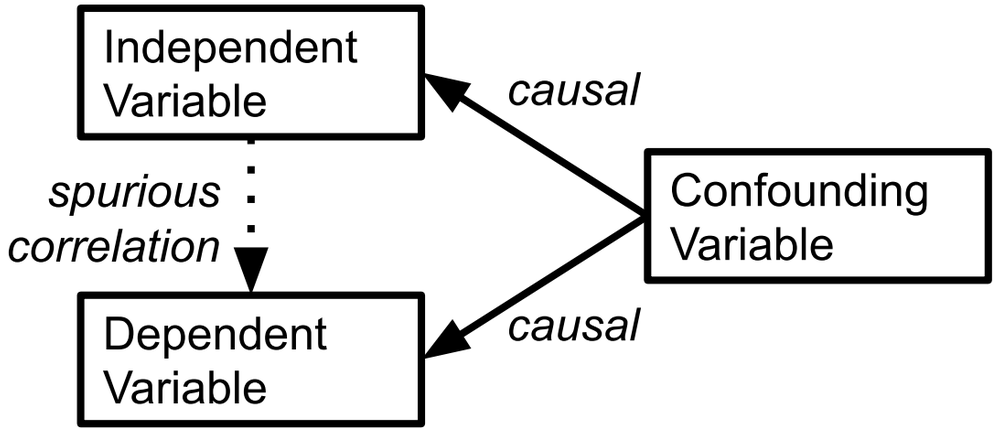
Be careful about interpretation & intervention based on correlations

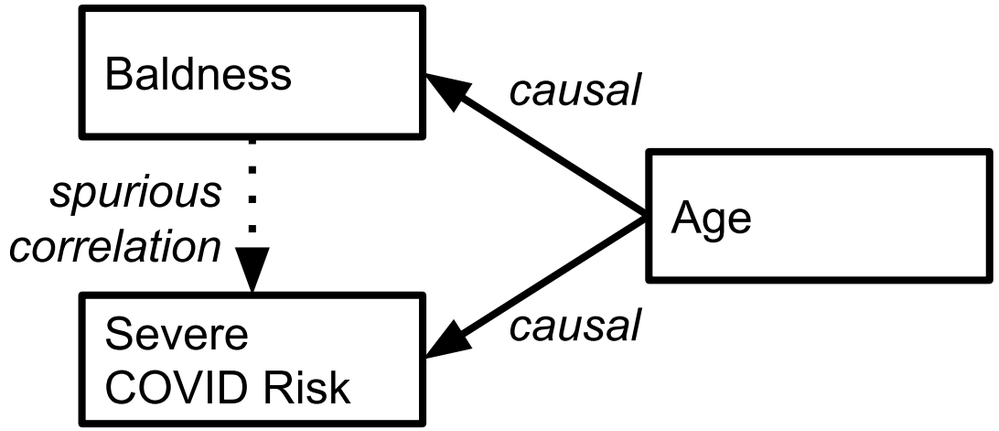
- e.g., positive correlation between exercise and skin cancer
- Exercise less => reduce chance of skin cancer?

To establish causality:

- Develop a theory ("X causes Y") based on domain knowledge & indep. data
- Identify relevant variables
- Design a controlled experiment & show correlation
- Demonstrate ability to predict new cases

Confounding Variables





Confounding Variables

- To identify spurious correlations between X and Y:
  - Identify potential confounding variables
  - Control for those variables during measurement
    - Randomize, fix, or measure + account for during analysis
    - e.g., control for "smoke", check whether "drink coffee" => "pancreatic cancer"
- Other examples
  - Degree from top-ranked schools => higher salary
  - Age => credit card default rate
  - Exercise => skin cancer
  - and many more...

Streetlight effect

- A type of *observational bias*
- People tend to look for something where it's easiest to do so
  - Use cheap proxy metrics that only poorly correlate with goal
  - e.g., number of daily active users as a measure of projected revenue

Risks of Metrics as Incentives

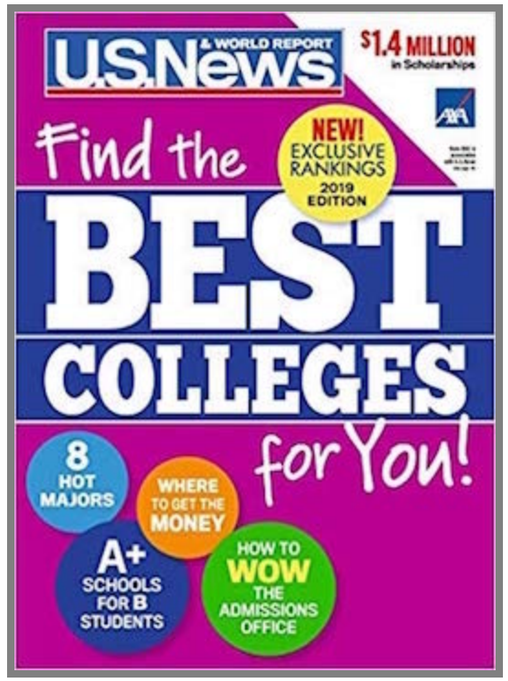
Metrics-driven incentives can:

- Extinguish intrinsic motivation
- Diminish performance
- Encourage cheating, shortcuts, and unethical behavior
- Become addictive
- Foster short-term thinking

Often, different stakeholders have different incentives

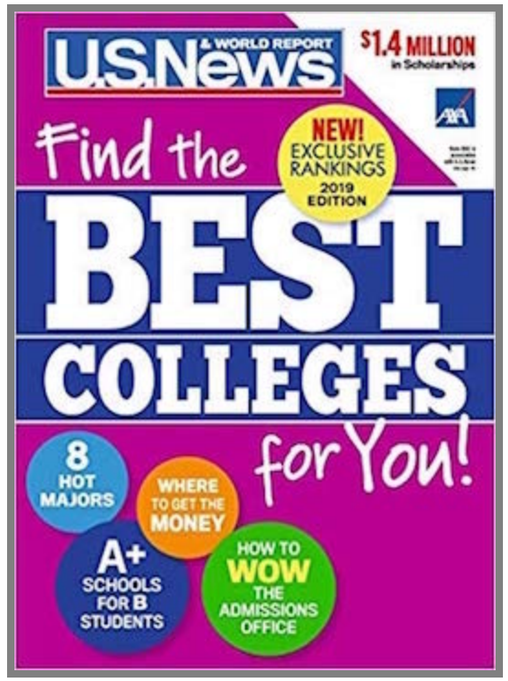
Make sure data scientists and software engineers share goals and success measures

Example: University Rankings



- Originally: Opinion-based polls, but complaints by schools on subjectivity
- Data-driven model: Rank colleges in terms of "educational excellence"
- Input: SAT scores, student-teacher ratios, acceptance rates, retention rates, campus facilities, alumni donations, etc.,

Example: University Rankings



- Can the ranking-based metric be misused or cause unintended side effects?

For more, see Weapons of Math Destruction by Cathy O'Neil

Speaker notes

- Example 1
  - Schools optimize metrics for higher ranking (add new classrooms, nicer facilities)
  - Tuition increases, but is not part of the model!
  - Higher ranked schools become more expensive
  - Advantage to students from wealthy families
- Example 2
  - A university founded in early 2010's
  - Math department ranked by US News as top 10 worldwide
  - Top international faculty paid $$ as a visitor; asked to add affiliation
  - Increase in publication citations => skyrocket ranking!

Successful Measurement Program

- Set solid measurement objectives and plans
- Make measurement part of the process
- Gain a thorough understanding of measurement
- Focus on cultural issues
- Create a safe environment to collect and report true data
- Cultivate a predisposition to change
- Develop a complementary suite of measures

Summary

Ask yourself: Do you really need ML? Identify the business case. Establish a non-ML solution as a baseline & consider cost vs benefit.

Identify and align your goals. Better ML models does not always lead to better business goals!

Consider risks of measurement. Are you really measuring what you want? Can your metric incentivize bad behaviors?