GOALS AND SUCCESS MEASURES FOR AI-ENABLED SYSTEMS

Christian Kaestner

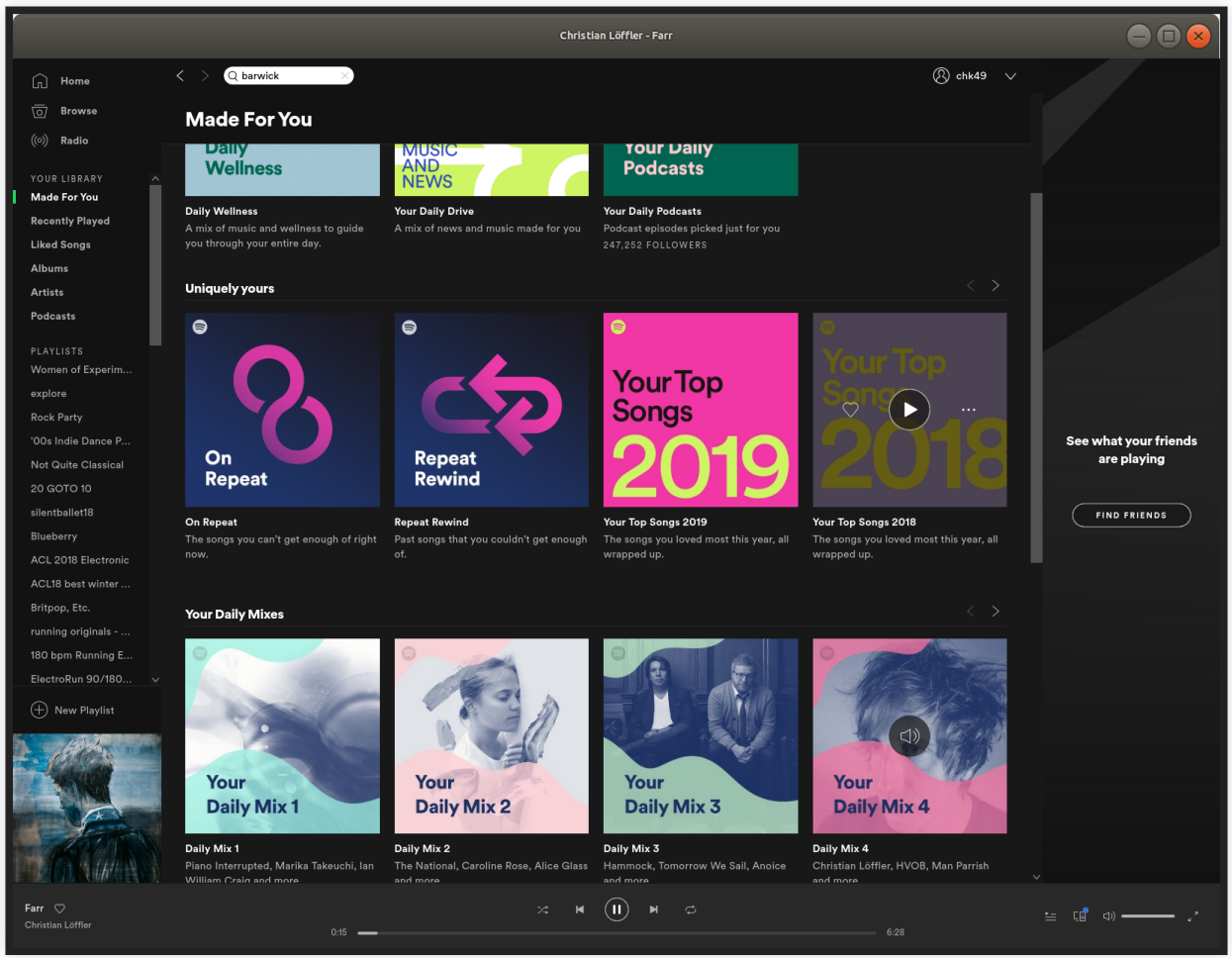
Required Readings: Hulten, Geoff. "Building Intelligent Systems: A Guide to Machine Learning Engineering." (2018), Chapters 2 (Knowing when to use IS), 4 (Defining the IS's Goals) and 15 (Intelligent Telemetry)

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

LEARNING GOALS

- Judge when to apply AI for a problem in a system
- Define system goals and map them to goals for the AI component
- Design and implement suitable measures and corresponding telemetry

TODAY'S CASE STUDY: SPOTIFY PERSONALIZED PLAYLISTS



WHEN TO USE MACHINE LEARNING?

WHEN NOT TO USE MACHINE LEARNING?

- If clear specifications are available
- Simple heuristics are *good enough*
- Cost of building and maintaining the system outweighs the benefits (see technical debt paper)
- Correctness is of utmost importance
- Only use ML for the hype, to attract funding

Examples?

Speaker notes

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

- **Discuss Examples**
    - Ranking apps, recommending products
    - Filtering spam or malicious advertisement
    - Creating subtitles for conference videos
    - Summarizing soccer games
    - Controlling a washing machine

WHEN TO USE MACHINE LEARNING

- Big problems: many inputs, massive scale
- Open-ended problems: no single solution, incremental improvements, continue to grow
- Time-changing problems: adapting to constant change, learn with users
- Intrinsically hard problems: unclear rules, heuristics perform poorly

Examples?

see Hulten, Chapter 2

WHEN TO USE MACHINE LEARNING

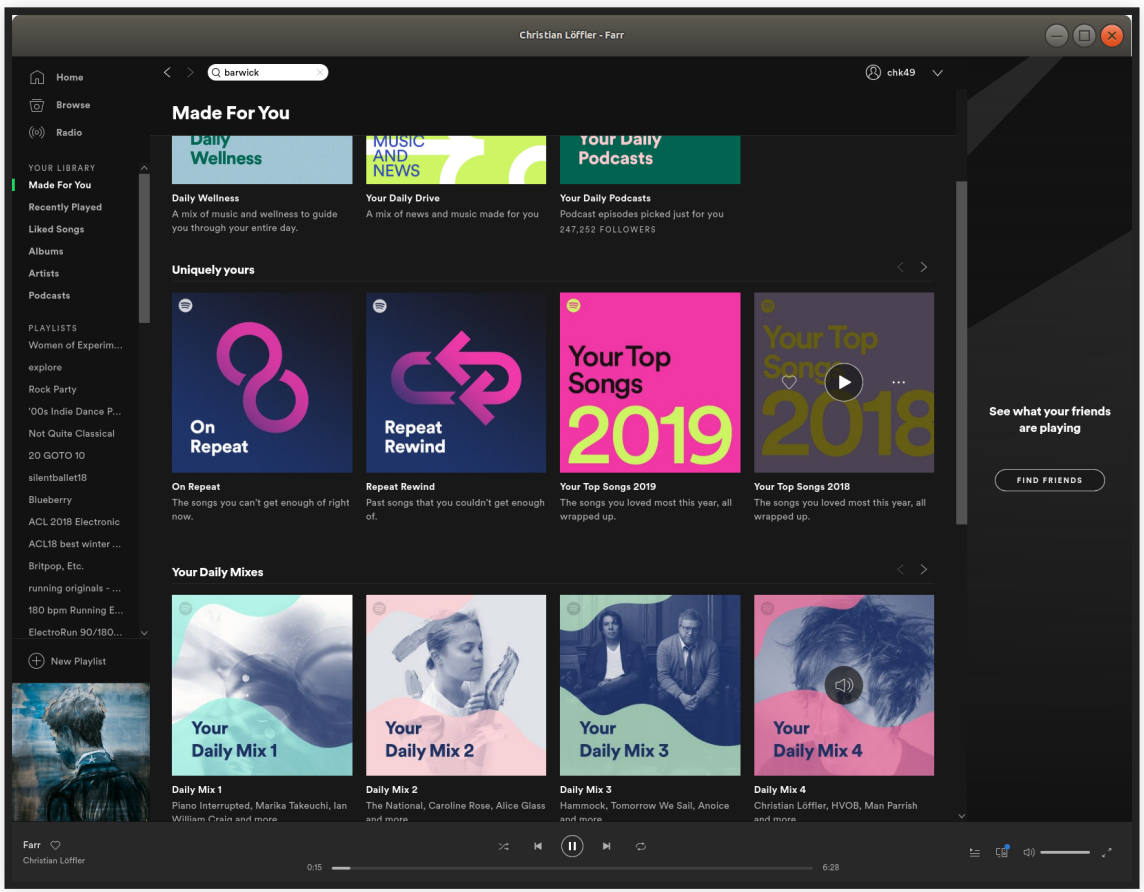
- Partial system is viable and interesting: mistakes are acceptable or mitigatable, benefits outweigh costs
- Data for continuous improvement is available: telemetry design
- Predictions can have an influence on system objectives: systems act, recommendations, ideally measurable influence
- Cost effective: cheaper than other approaches, meaningful benefits

Examples?

see Hulten, Chapter 2

DISCUSSION: SPOTIFY

Big problem? Open ended? Time changing? Hard? Partial system viable? 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

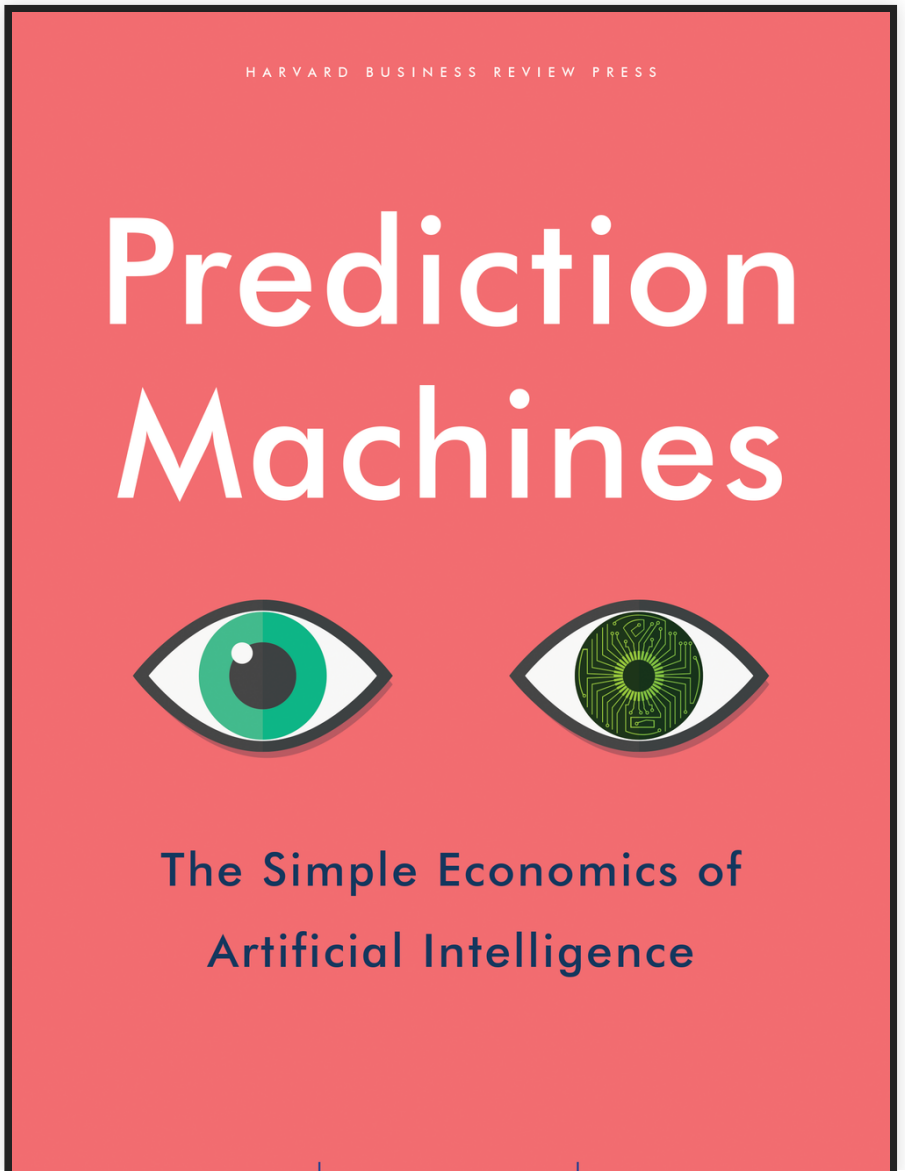
AI: Higher accuracy predictions at much much lower cost

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, reduced cost of search with the internet)




AJAY AGRAWAL
JOSHUA GANS
AVI GOLDFARB

Speaker notes

May use new, cheaper predictions for traditional tasks -> inventory and demand forcast; 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 distinct 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/

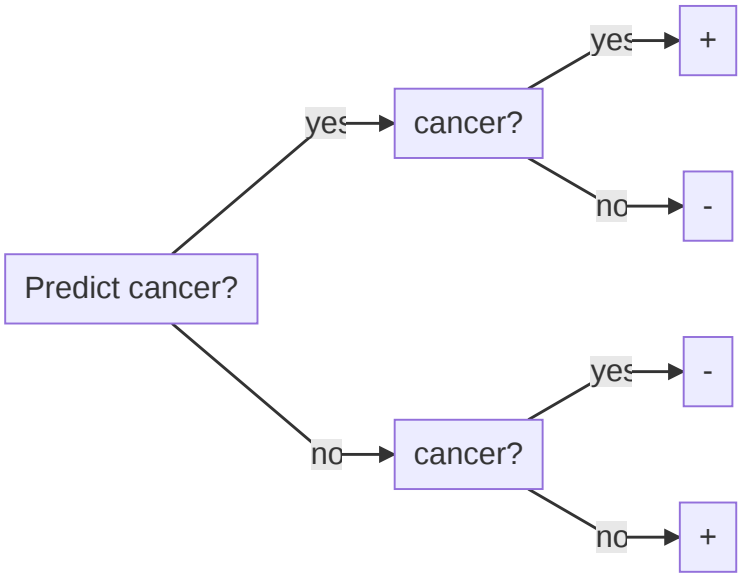
PREDICTIONS VS JUDGEMENT

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



Determine value function from value of each outcome and probability of each outcome

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

AUTOMATION IN CONTROLLED ENVIRONMENTS

Speaker notes

Source https://pixabay.com/photos/truck-giant-weight-mine-minerals-5095088/

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

The AI Canvas

What task/decision are you examining?

Briefly describe the task being analyzed.

Prediction

Identify the key uncertainty that you would like to resolve.

Judgment

Determine the payoffs to being right versus being wrong. Consider both false positives and false negatives.

Action

What are the actions that can be chosen?

Outcome

Choose the measure of performance that you want to use to judge whether you are achieving your outcomes.

Training

What data do you need on past inputs, actions and outcomes in order to train your AI and generate better predictions?

Input

What data do you need to generate predictions once you have an AI algorithm trained?

Feedback

How can you use measured outcomes along with input data to generate improvements to your predictive algorithm?

How will this AI impact on the overall workflow?

Explain here how the AI for this task/decision will impact on related tasks in the overall workflow. Will it cause a staff replacement? Will it involve staff retraining or job redesign?

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

COST PER PREDICTION

What contributes to the average cost of a single prediction?

Examples: Credit card fraud detection, product recommendations on Amazon

COST PER PREDICTION

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

AI RISKS

- Discrimination and thus liability
- Creating false confidence when predictions are poor
- Risk of overall system failure, failure to adjust
- Leaking of intellectual property
- Vulnerable to attacks if learning data, inputs, or telemetry can be influenced

- Societal risks
    - Focus on few big players (economies of scale), monopolization, inequality
    - Prediction accuracy vs privacy

DISCUSSION: FEASIBLE ML-EXTENSIONS

Discuss in groups

Each group pick a popular open-source system (e.g., Firefox, Kubernetis, VS Code, WordPress, Gimp, Audacity)

Think of possible extensions with and without machine learning

Report back 1 extension that would benefit from ML and one that would probably not

10 min

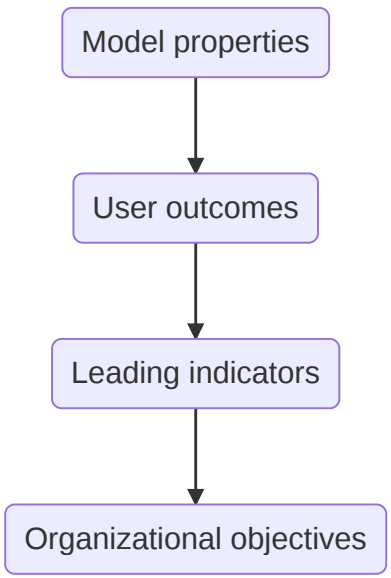
Guiding questions:

- ML Suitable: *Big problem? Open ended? Time changing? Hard? Partial system viable? Data continuously available? Influence objectives? Cost effective?*
- ML Profitable: *Prediction opportunities, cost per prediction, data costs, automation potential*

SYSTEM GOALS

LAYERS OF SUCCESS MEASURES

- Organizational objectives: Innate/overall goals of the organization
- Leading indicators: Measures correlating with future success, from the business' perspective
- 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



Some are easier to measure then others (telemetry), some are noisier than others, some have more lag

ORGANIZATIONAL OBJECTIVES

Innate/overall goals of the organization

- Business
  - current revenue, profit
  - future revenue, profit
  - reduce risk
- Non-Profits
  - Lives saved, animal welfare increased
  - CO_2 reduced, fires averted
  - Social justice improved, well-being elevated, fairness improved

accurate models themselves are not a goal

AI may only very indirectly influence such organizational objectives, influence hard to quantify, lagging measures

BREAKING DOWN PROCESSES

Break overall goals along processes

- Break workflow into tasks
- Identify decisions in tasks
- Evaluate benefit of AI for prediction or automation
- Evaluate the influence of improving some tasks on process

Maintain mapping from task-specific goals to system goals

LEADING INDICATORS

Measures correlating with future success, from the business' perspective

- Customers sentiment, customers liking the products (e.g., through surveys, rating)
- Customer engagement, regular use, time spent on site, messages posted
- Growing user numbers, recommendations

indirect proxy measures, lagging, bias

can be misleading (more daily active users => higher profits?)

USER OUTCOMES

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

- Users receive meaningful recommendations, enjoying content
- Users making better decisions
- Users saving time due to system
- Users achieving their goals

easier and more granular to measure, but only indirect relation to organization objectives

MODEL PROPERTIES

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

not directly linked to business goals

LAYERING OF SUCCESS MEASURES

Example: Amazon shopping recommendations

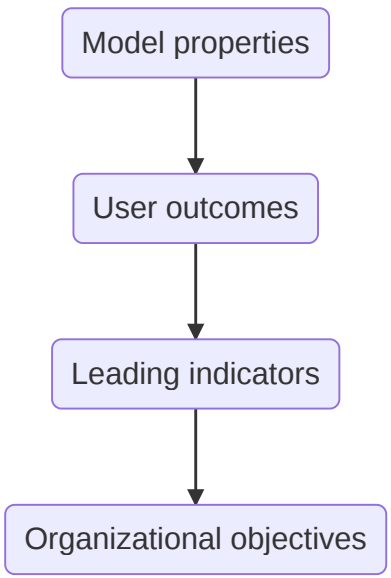
Closely watch model properties for degradation; optimize accuracy; ongoing

Weekly review user outcomes, e.g., sales, reviews returns

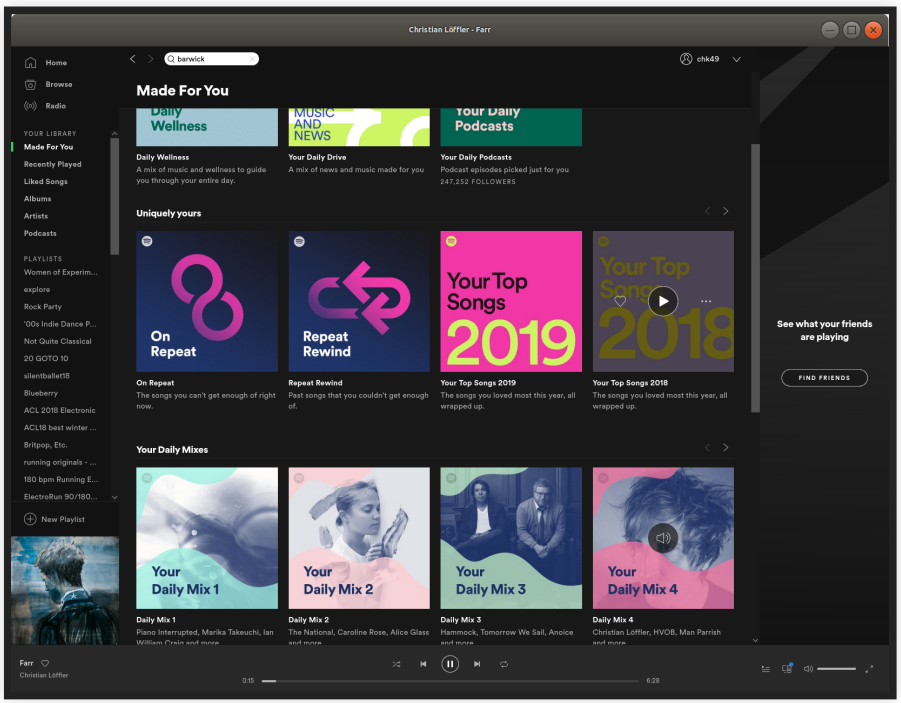
Monthly review trends of leading indicators, e.g., shopper loyalty

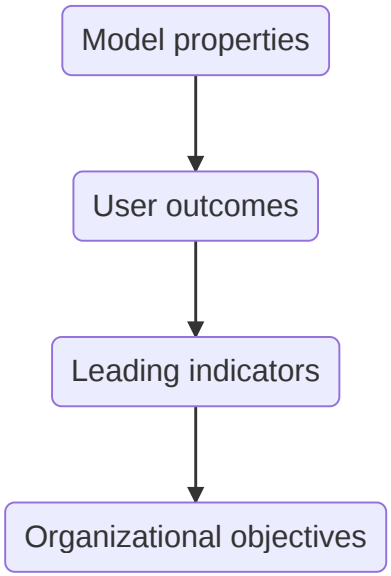
Quarterly ensure look at organizational objectives

Telemetry?



SUCCESS MEASURES IN THE SPOTIFY SCENARIO?





EXERCISE: AUTOMATING ADMISSION DECISIONS TO MASTER'S PROGRAM

Discuss in groups, breakout rooms

What are the *goals* behind automating admissions decisions?

Organizational objectives, leading indicators, user outcomes, model properties?

Report back in 10 min

EXCURSION: 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

- 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.
- 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
- 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

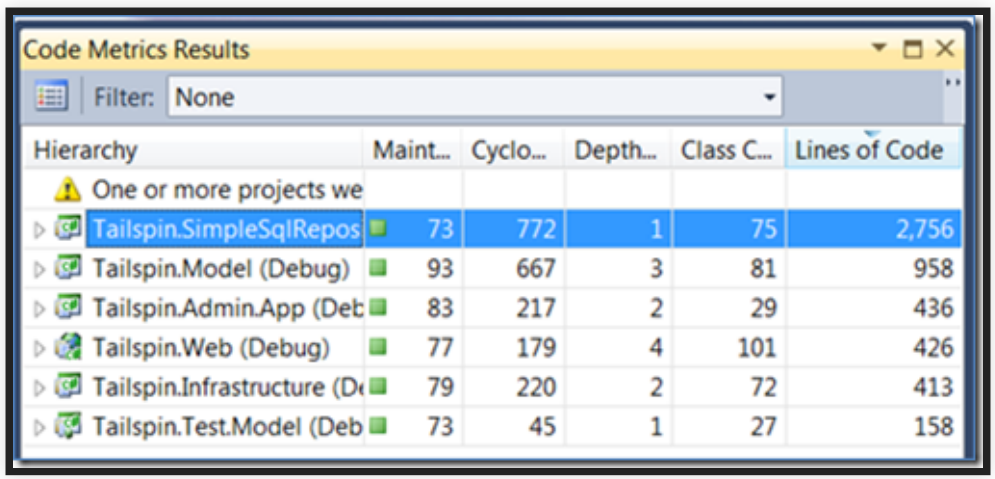
Douglas Hubbard, “How to Measure Anything: finding the value of intangibles in business" 2014

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 synonmous 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)

THE MAINTAINABILITY INDEX

$max(0, (171 - 5.2ln(HV) - 0.23CC - 16.2ln(LOC) * 100/171)) < 10$: low maintainability, ≤ 20 high maintainability



Further reading: Arie van Deursen, Think Twice Before Using the "Maintainability Index", Blog Post, 2014

Speaker notes

The maintainability index is a machine learning classifier in the modern sense. The function was derived through linear regression from training data. Thresholds were defined at Microsoft manually.

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? Predictions reliable?

MEASUREMENT IN DATA SCIENCE

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

MEASUREMENT SCALES

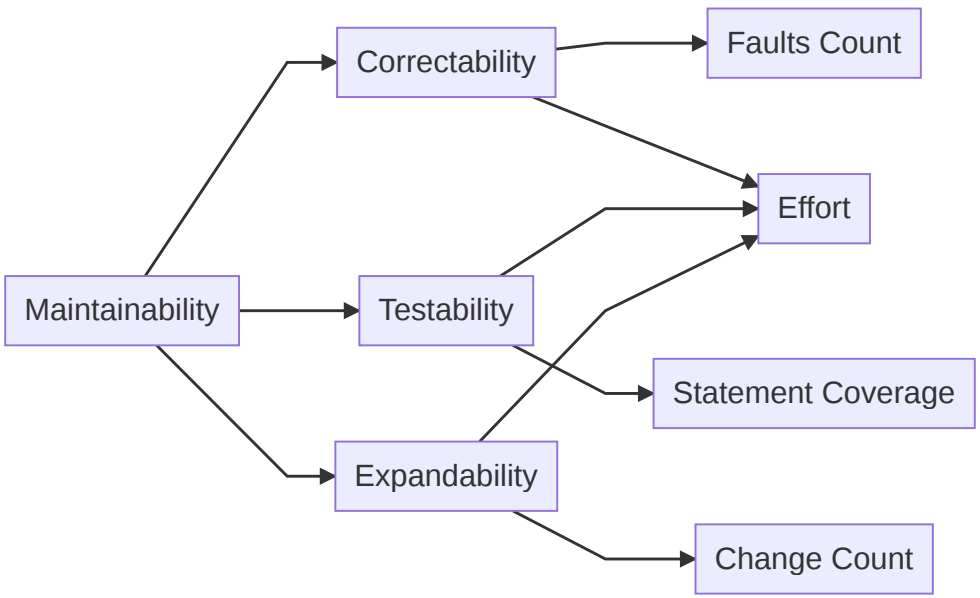
- Scale: The type of data being measured; dictates what sorts of analysis/arithmetic is legitimate or meaningful.
- Nominal: Categories ($=$, $\neq$, frequency, mode, ...)
    - e.g., biological species, film genre, nationality
- Ordinal: Order, but no meaningful magnitude ($<$, $>$, median, rank correlation, ...)
    - 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)

Aside: Understanding scales of features is also useful for encoding or selecting learning strategies in ML

DECOMPOSITION OF MEASURES

Often higher-level measures are composed from lower level measures

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



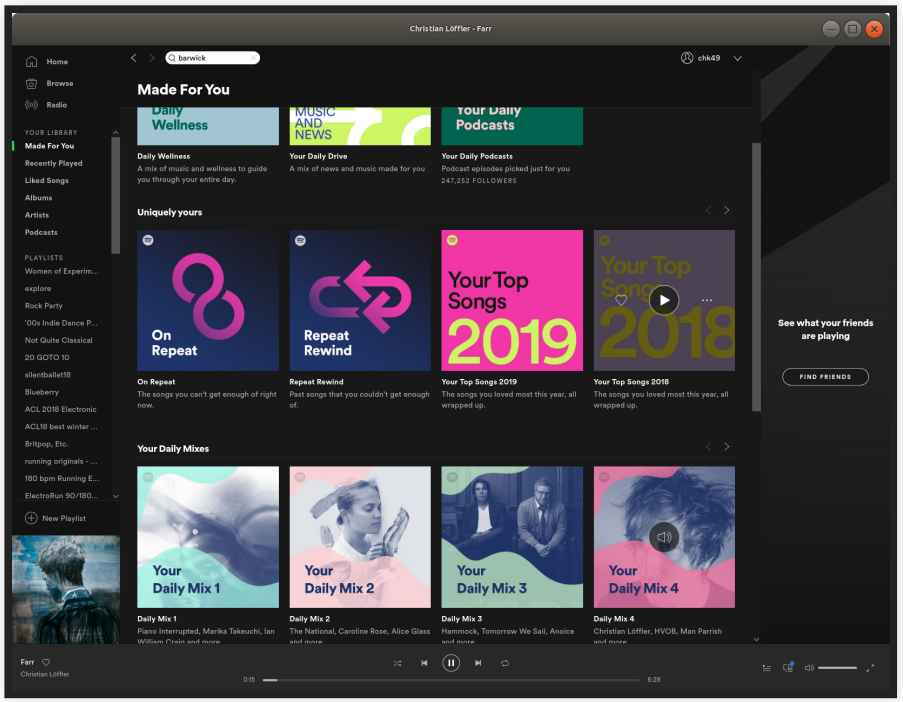
For design strategy, see Goal-Question-Metric approach

SPECIFYING METRICS

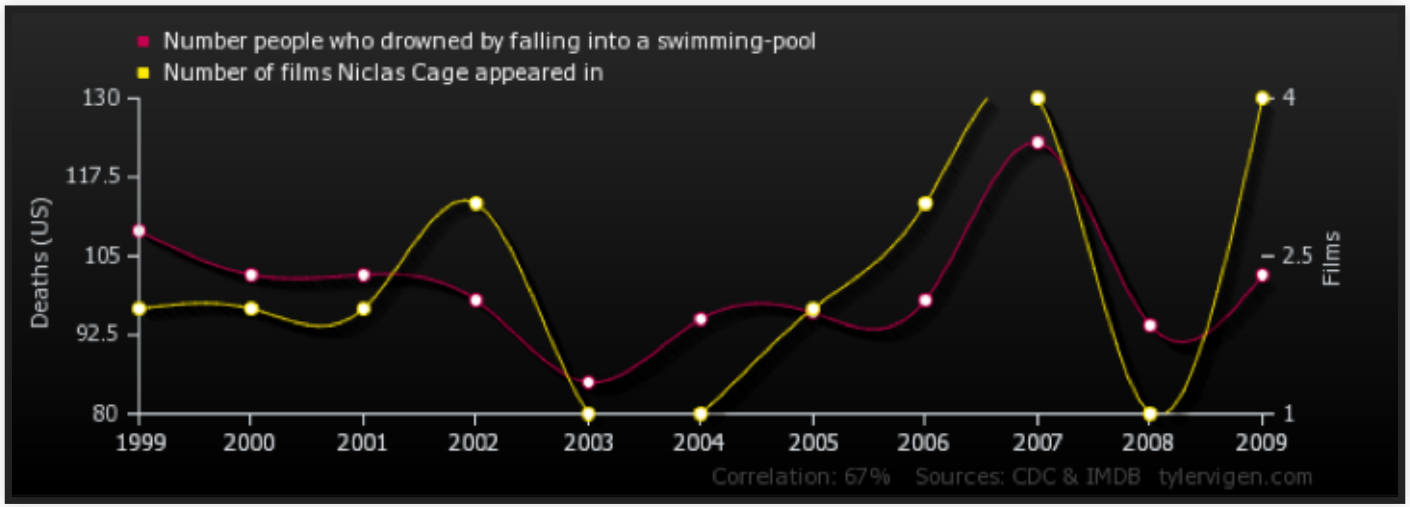
- Always be precise about metrics
  - "measure accuracy" -> "evaluate accuracy with MAPE"
  - "evaluate test quality" -> "measure branch coverage with Jacoco"
  - "measure execution time" -> "average and 90%-quantile response time for REST-API x under normal load"
  - "assess developer skills" -> "measure average lines of code produced per day and number of bugs reported on code produced by that developer"
  - "measure customer happyness" -> "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

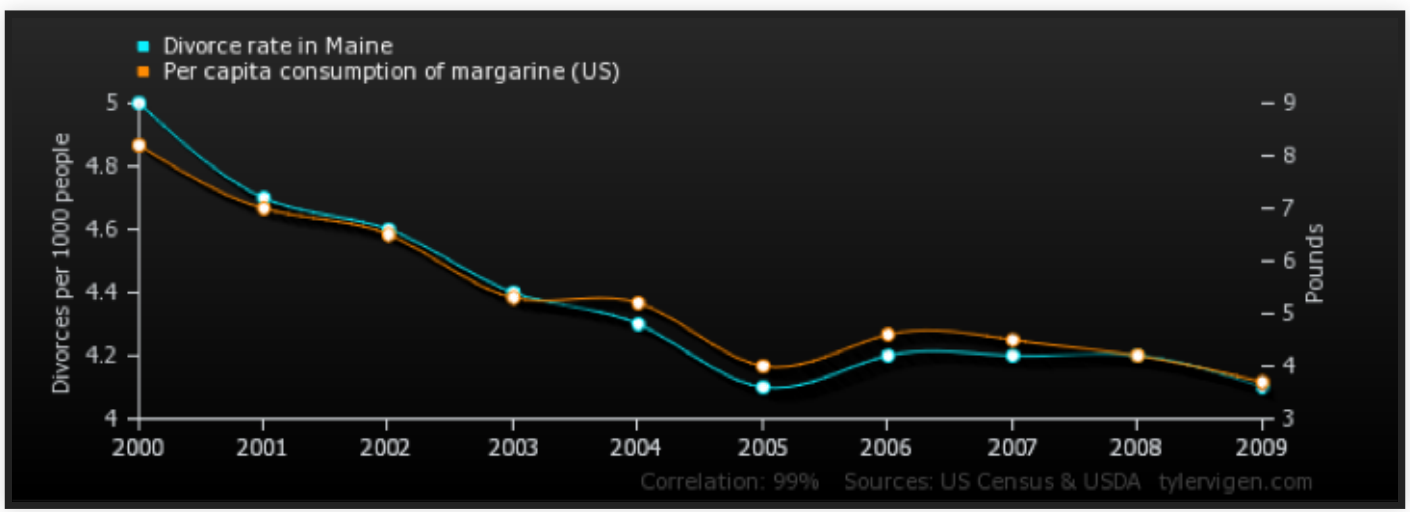
EXERCISE: SPECIFIC METRICS FOR SPOTIFY GOALS?

- Organization objectives?
- Leading indicators?
- User outcomes?
- Model properties?
- What are their scales?



CORRELATION VS CAUSATION

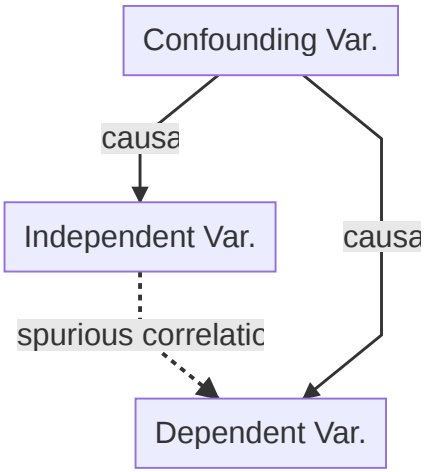


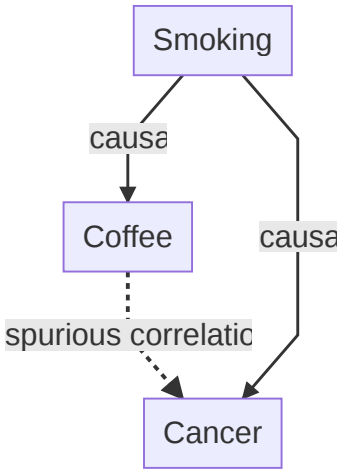


CORRELATION VS CAUSATION

- In general, ML learns correlation, not causation
  - (exception: Bayesian networks, certain symbolic AI methods)
- To establish causality:
  - Develop a theory ("X causes Y") based on domain knowledge & independent 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)
- Examples
  - Drink coffee => Pancreatic cancer?
  - Degree from high-ranked schools => Higher-paying jobs?

CHALLENGE: THE 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?

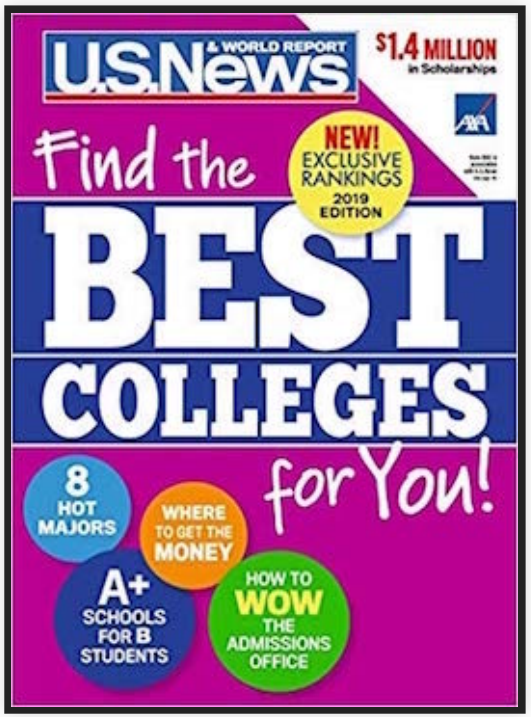

THIS IS WHERE YOU LOST YOUR WALLET?
NO, I LOST IT IN THE PARK. BUT THIS IS WHERE THE LIGHT IS.

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

ON INCENTIVES: UNIVERSITY RANKINGS



Speaker notes

- Originally: Opinion-based polls, but schools complained
- Data-driven model: Rank colleges in terms of "educational excellence"
- Input: SAT scores, student-teacher ratios, acceptance rates, retention rates, alumni donations, etc.,
- What is (not) being measured? Any streetlight effect?
- Is the measured data being used correctly?
- Are incentives for using these data good? Can they be misused?
- 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!

MEASUREMENT VALIDITY

- Construct: 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: 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 test group: Does it hold over the general public?

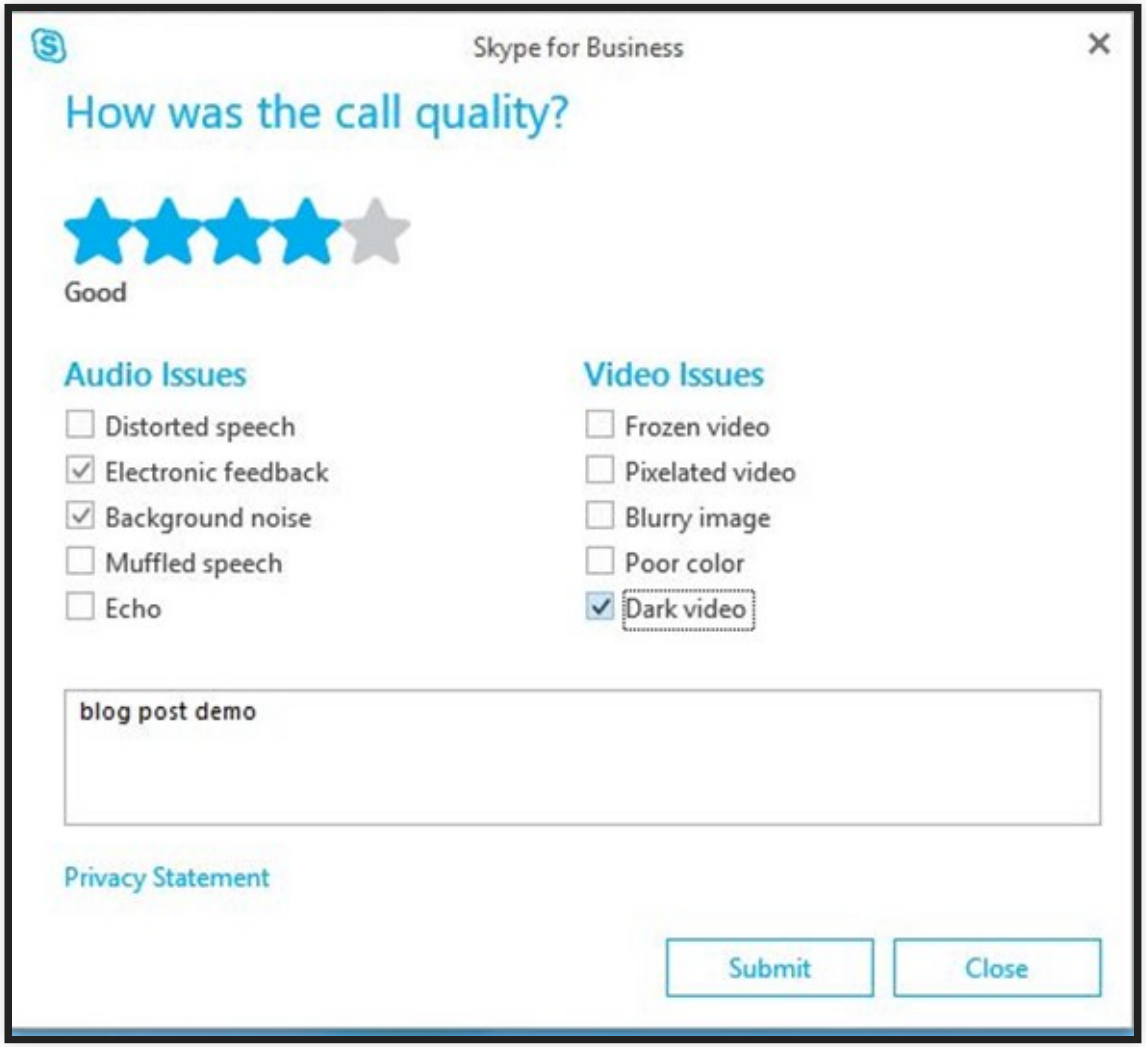
SUCCESSFUL MEASUREMENT PROGRAM

- 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

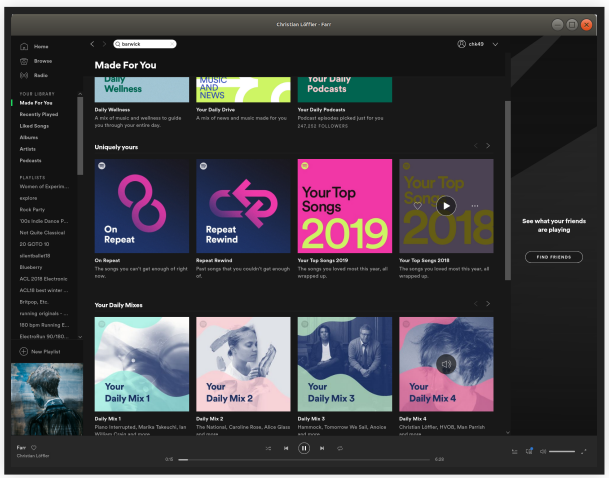
OUTLOOK: MEASUREMENT WITH TELEMETRY

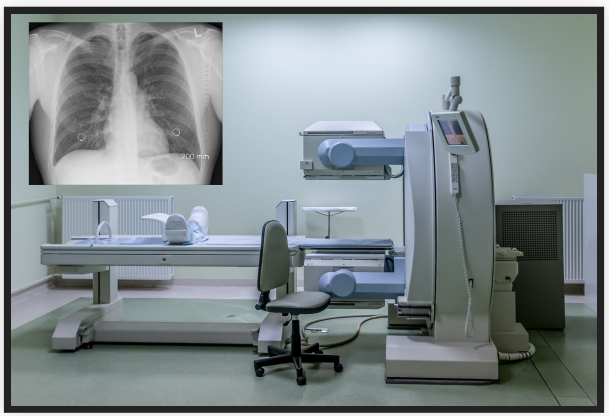
KEY CHALLENGE: TELEMETRY

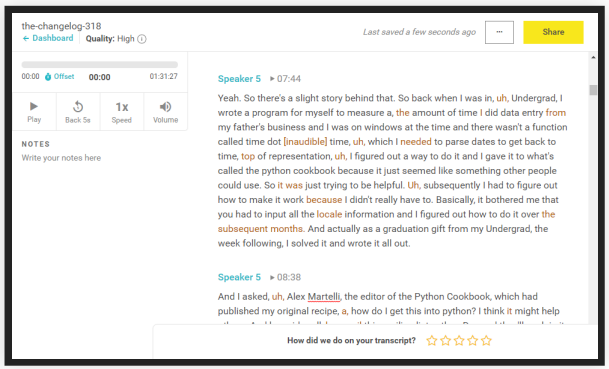
- Some goals can be quantified easily, others not so much
- Be creative in data sources (telemetry)
    - Existing logs and measures
    - Logging additional information
    - User surveys, feedback mechanisms
- Often only proxy measures, sometimes delayed, often only samples


Skype for Business
How was the call quality?
Good
Audio Issues
Distorted speech
Electronic feedback
Background noise
Muffled speech
Echo
Video Issues
Frozen video
Pixelated video
Blurry image
Poor color
Dark video
blog post demo
Privacy Statement
Submit
Close

HOW TO MEASURE SYSTEM SUCCESS?







SUMMARY

- Be deliberate about when to use AI/ML
- Understand the business case for AI (cheap predictions, automation, cost per prediction)
- Identify and break down system goals, define concrete measures
- Telemetry to quantify system goals
- Key concepts and challenges of measurement