# GOALS AND SUCCESS MEASURES FOR AI-ENABLED SYSTEMS

Eunsuk Kang

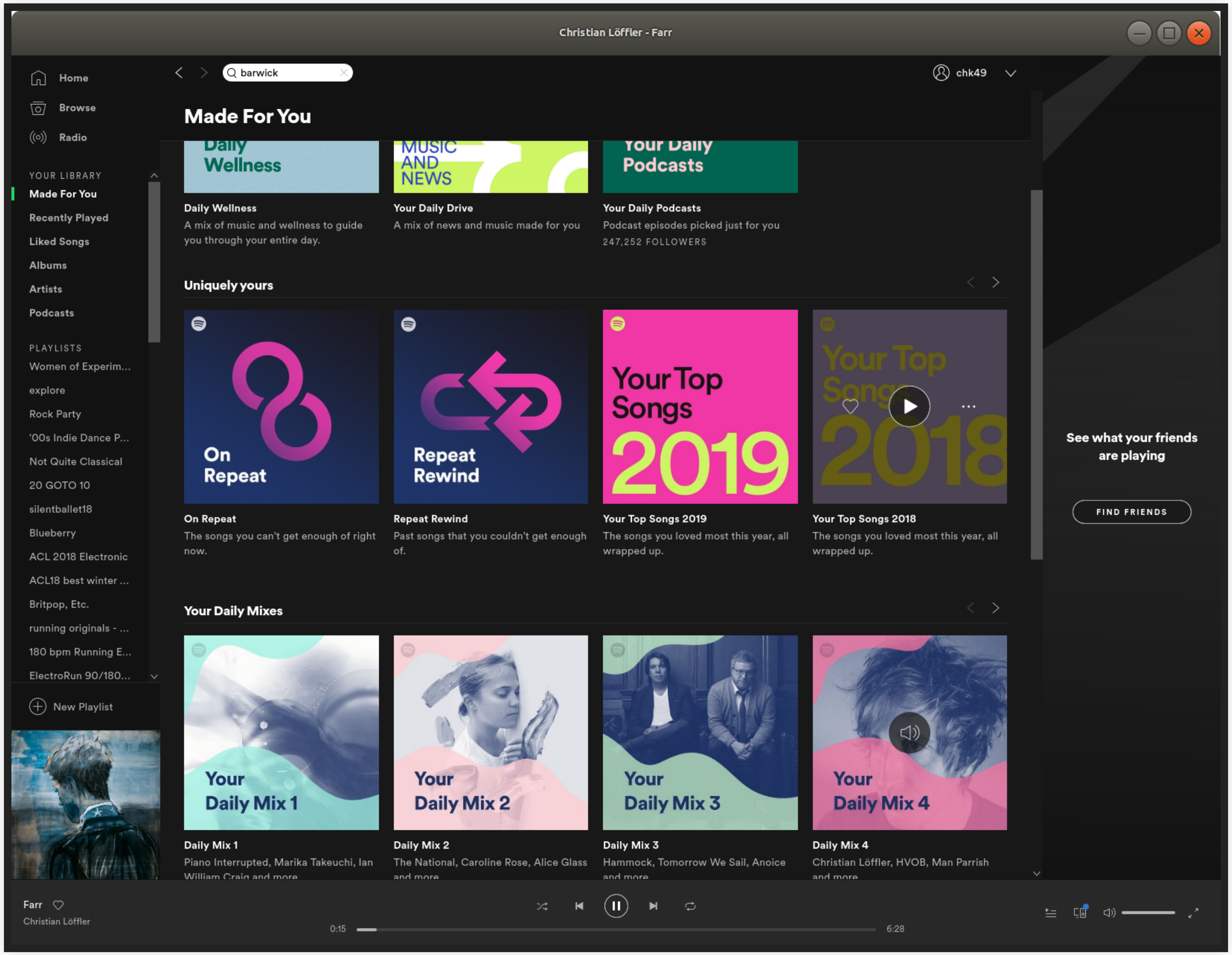
Required Readings: Hulten, Geoff. "Building Intelligent Systems: A Guide to Machine Learning Engineering" (2018), Chapters 2 (Knowing when to use IS) and 4 (Defining the IS's Goals)

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

# 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

# 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 system outweighs the benefits (see technical debt paper)
- Correctness is of utmost importance
- ML is used only for the hype, to attract funding

**Examples?**

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

- **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

# ADDITIONAL CONSIDERATIONS FOR ML

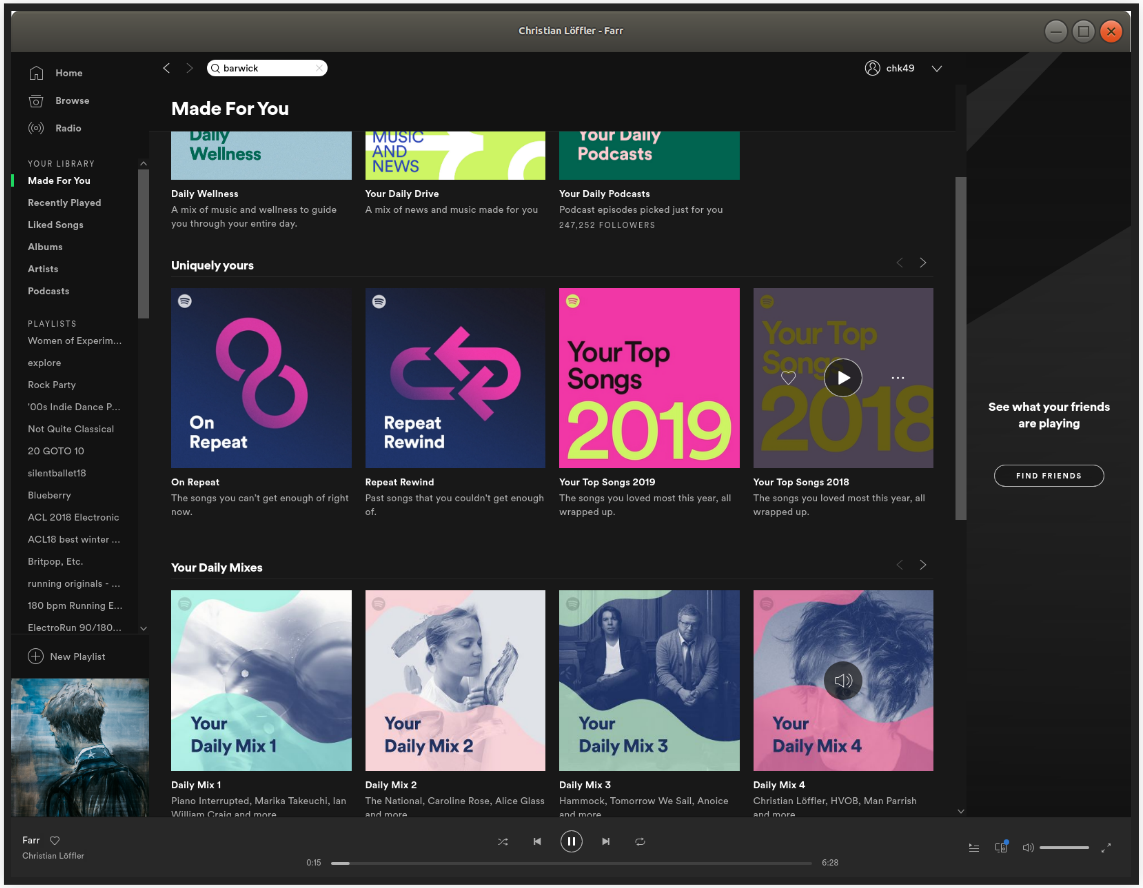
- Partial solution is acceptable: Mistakes are acceptable or mitigable
- Data for continuous improvement is available
- Predictions can have an influence on system objectives: Does it actually contribute to organizational objectives?
- Cost effective: Cheaper than other approaches, or benefits clearly 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?

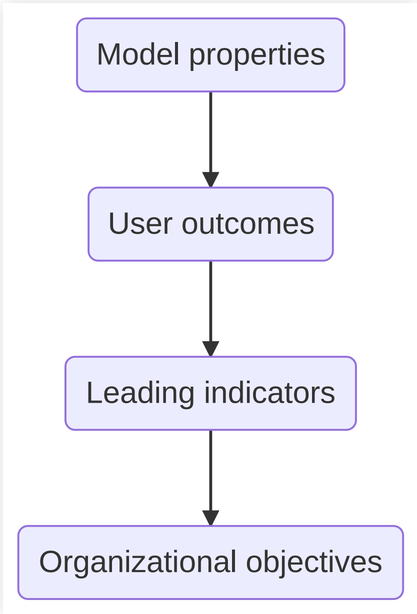
*Big problem? Open ended? Time changing? Hard? Partial solution acceptable? Data continuously available? Influence objectives? Cost effective?*

# 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

**Ideally, these goals should be aligned with each other**



# ORGANIZATIONAL OBJECTIVES

*Innate/overall goals of the organization*

- Business
  - Current revenue, profit
  - Future revenue, profit
  - Reduce business risks
- Non-Profits
  - Lives saved, animal welfare increased
  - $CO_2$ reduced, fires averted
  - Social justice improved, well-being elevated, fairness improved
- Often not directly measurable from system output; slow indicators

**Accurate models themselves are not the ultimate goal!**

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

# LEADING INDICATORS

*Measures correlating with future success, from the business perspective*

- 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

**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**

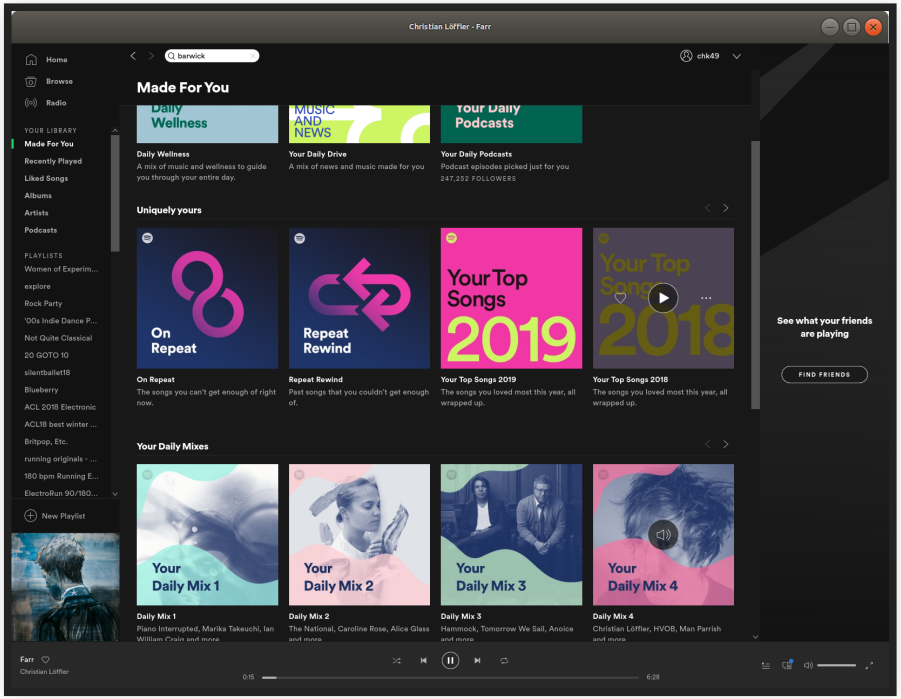
# MODEL PROPERTIES

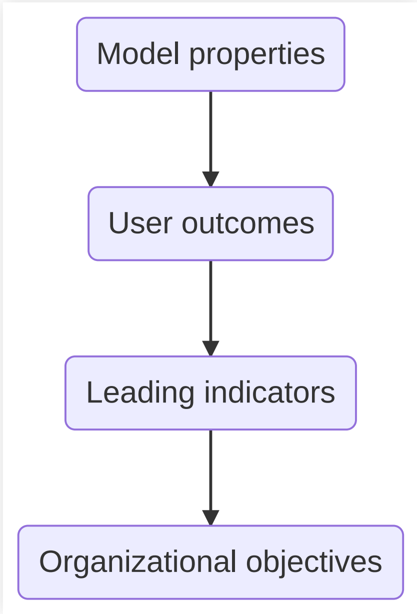
*Quality of the model used in a system, from the model's perspective*

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

**Not directly linked to business goals**

# SUCCESS MEASURES IN THE SPOTIFY SCENARIO?




Organizational objectives? Leading indicators? User outcomes? Model properties?

# 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

# 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

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

# 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.

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

# 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

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

# 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.

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

# 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 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? Predictions reliable?

# 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

# 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

# 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

# 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)

# 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)

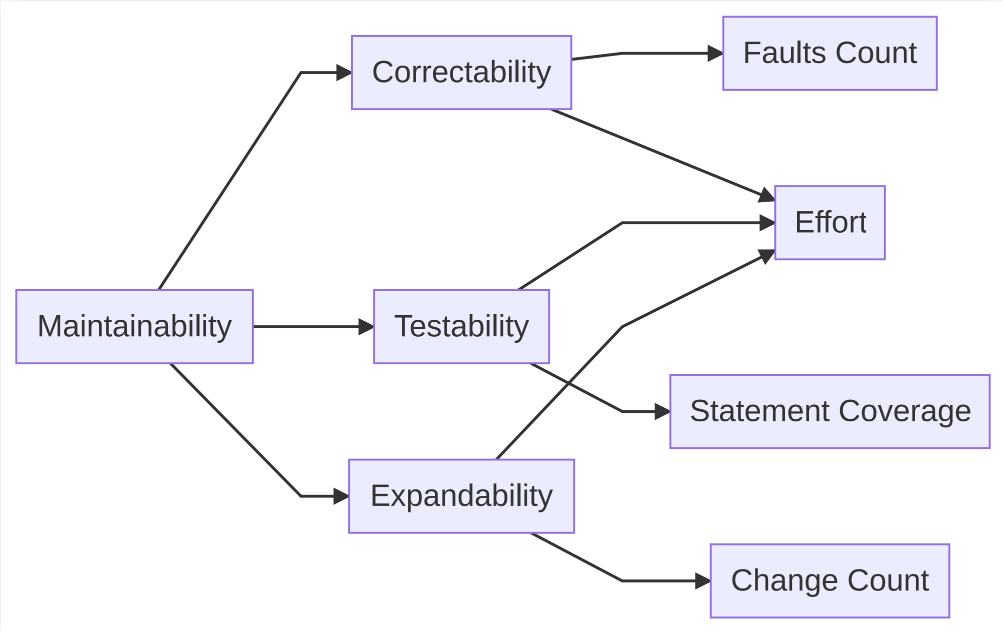
# MEASUREMENT SCALES

- 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

# DECOMPOSITION OF MEASURES

Often higher-level measures are composed from lower level measures

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



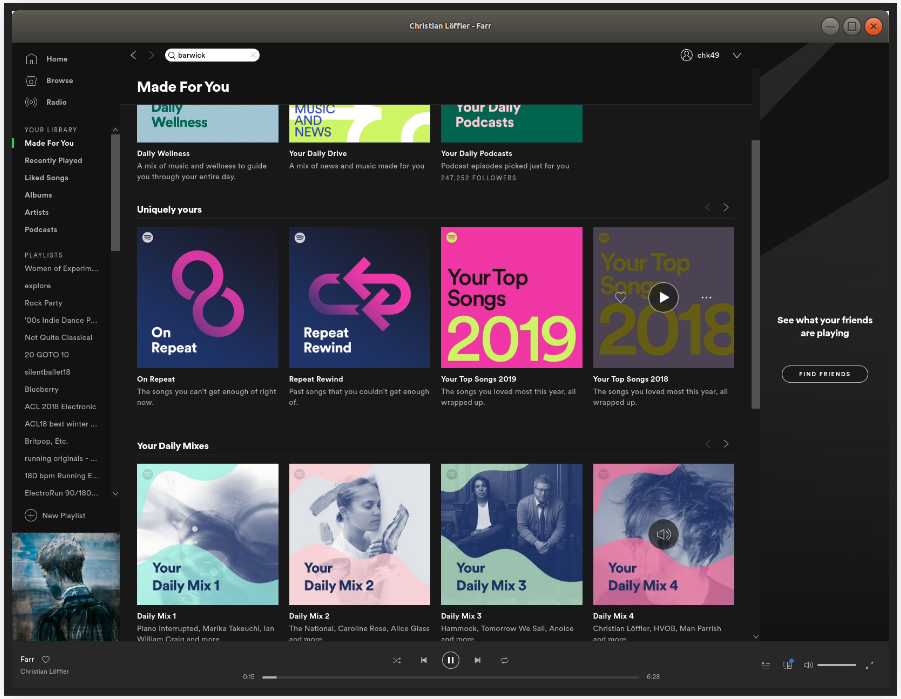

For design strategy, see Goal-Question-Metric approach

# 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 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

# EXERCISE: SPECIFIC METRICS FOR SPOTIFY GOALS?

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



# RISKS WITH MEASUREMENTS

# RISKS WITH MEASUREMENTS

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

# 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.

# 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

# 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...

# MEASUREMENT VALIDITY

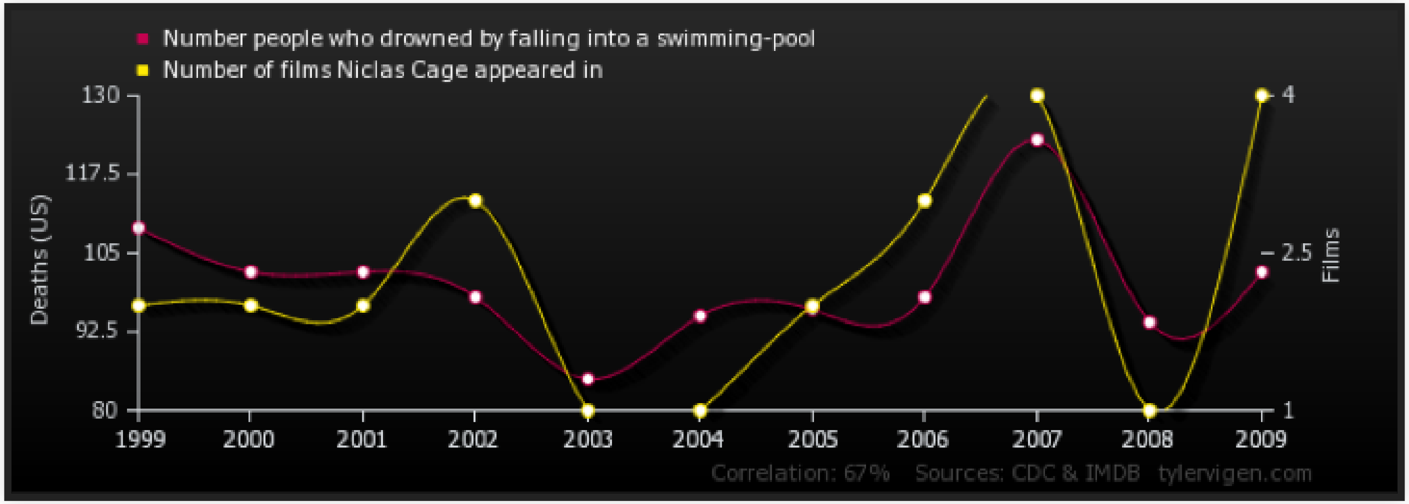
- Construct: Are we measuring what we intended to measure?
    - Does the abstract concept match the specific scale/measurement used?
    - e.g., IQ: What is it actually measuring?
    - Other examples: Pain, language proficiency, personality...
- Predictive: The extent to which the measurement can be used to explain some other characteristic of the entity being measured
    - e.g., Higher SAT scores => higher academic excellence?

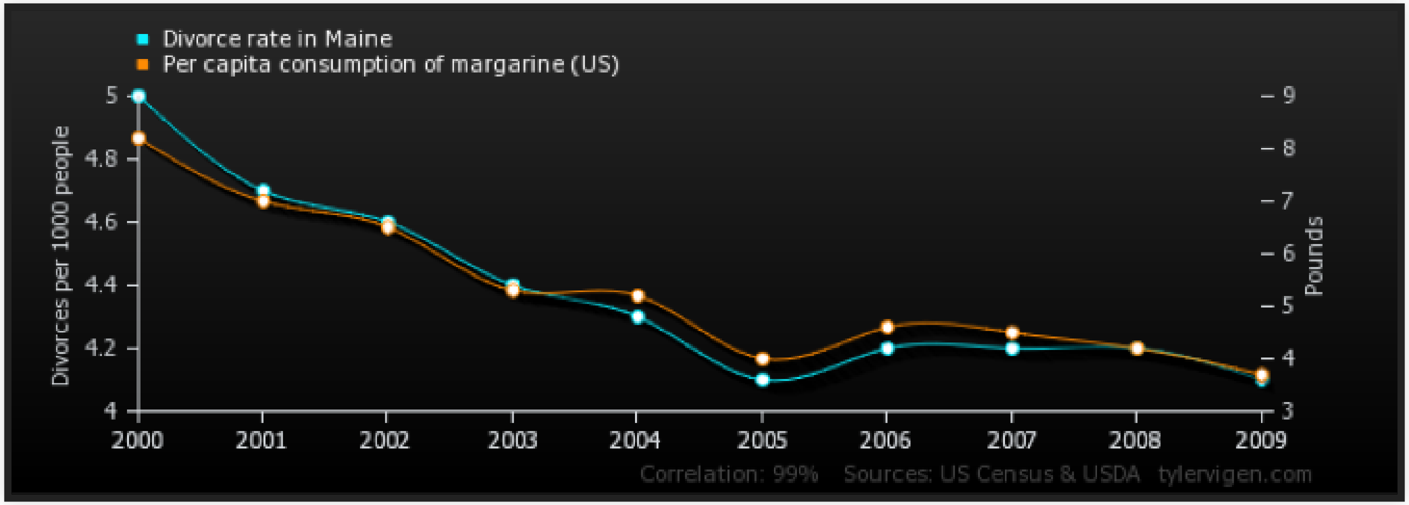
# MEASUREMENT VALIDITY

- 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?

# CORRELATION VS CAUSATION

https://www.tylervigen.com/spurious-correlations



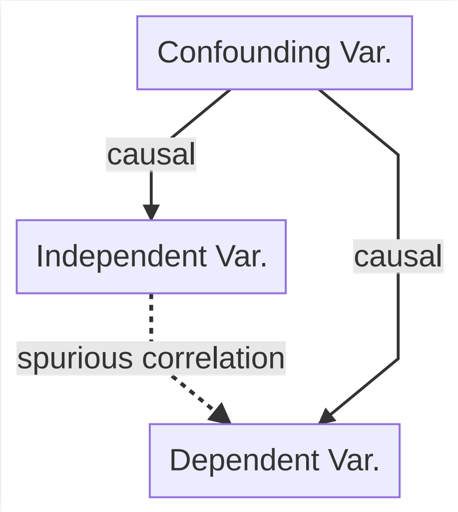


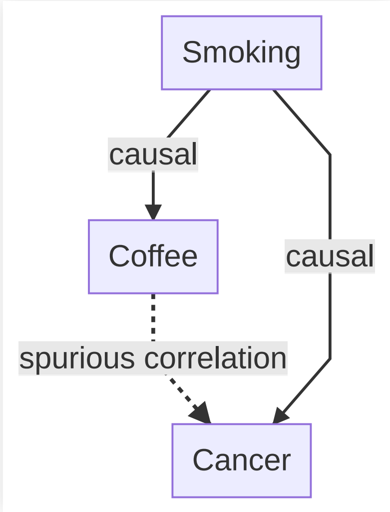
Divorce rate in Maine
Per capita consumption of margarine (US)
Divorces per 1000 people
Pounds
5
4.8
4.6
4.4
4.2
4
9
8
7
6
5
4
3
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Correlation: 99%
Sources: US Census & USDA
tylervigen.com

# 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 & 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
    - 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

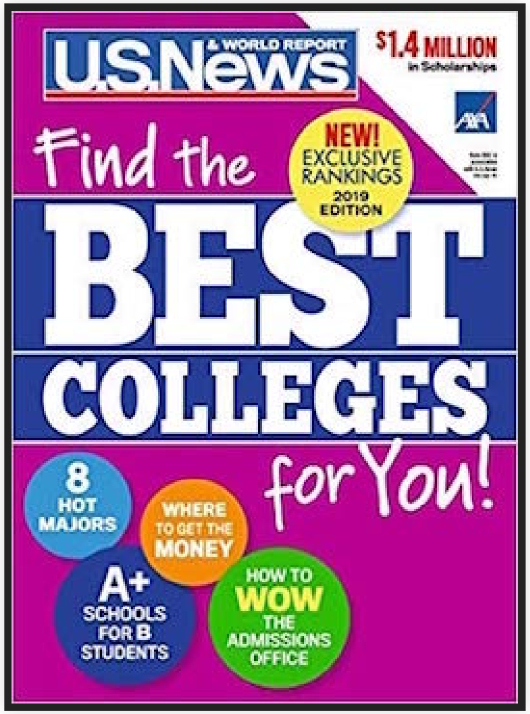
- A type of *observational bias*
- People tend to look for something where it's easiest to do so
  - Use cheap proxy metrics that only poorly correlate with goal
  - e.g., number of daily active users as a measure of projected revenue

# 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

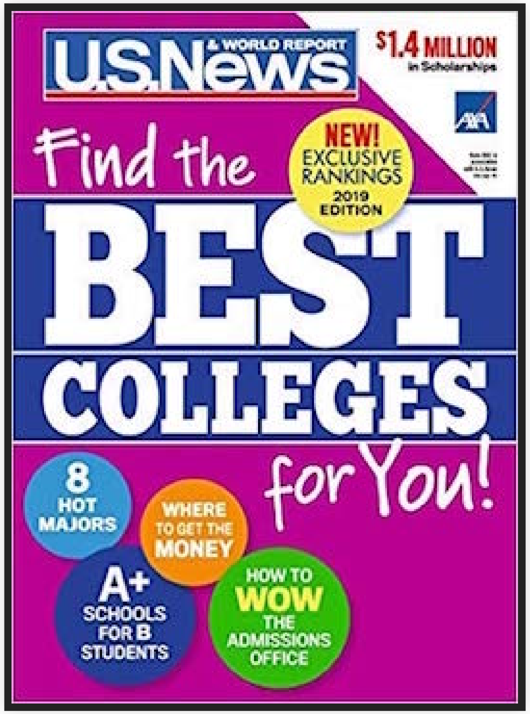
**Make sure data scientists and software engineers share goals and success measures**

# 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, alumni donations, etc.,

# EXAMPLE: UNIVERSITY RANKINGS




- Who are different stakeholders? What are their incentives? Can they 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

- Be deliberate about when to use AI/ML
- Identify and break down system goals, define concrete measures
- Key concepts and challenges of measurement