Principles of AI Engineering

Chapter 4: Goals

Prof. Dr. Steffen Herbold

Credit:

Based on contents from Christian Kästner (https://github.com/ckaestne/seai)

Contents

When to use machine learning

ML as universal solution

https://www.facebook.com/ProgrammersCreateLife/photos/a.241809332534619/4417862111595966/

But is it really always the best tool?

When not to use ML

- Clear specification available
  - Implement the specification directly! Learning adds risk.
- Simple heuristics are good enough
  - No need to spend the additional effort
- Cost of building and maintaining the ML system outweighs the benefits
  - ML components are complex and hard to maintain, simpler solutions or human effort may be cheaper
- Correctness is of utmost importance
  - That is still the issue with ML for safety critical systems!
- ML is only used for the hype
  - Marketing should not affect the system design

Consider non-ML baselines

- How far can simple heuristics get you?
- What are the costs and benefits of a semi-automated approach with human supervision?
- What would the system look like without the ML features?

When to use ML

- Big problems
  - Many inputs
  - Massive scale

- Open-ended problems
  - No single final solution
  - No fixed specification
  - Incremental improvements and growth over time

- Time-changing problems
  - Adapting to constant changes
  - Learning with and from the users

- Intrinsically hard problems
  - Unclear rules
  - Heuristics perform poorly

Live exercise: ML or not?

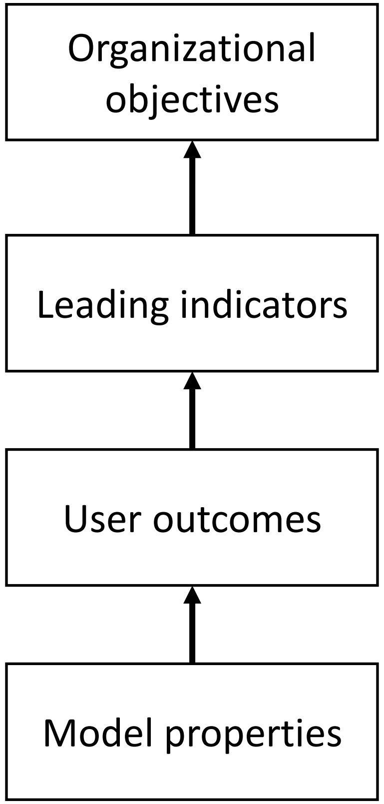
- Recommending products in a huge webshop
- Recommending products in a small webshop
- Filtering hate speech or profanity in public forums
- Credit card fraud detection
- Controlling water use in a washing machine

Additional consideration for ML

- Check if partial solutions are acceptable
  - Requires that mistakes are acceptable or can be mitigated
- Data for continuous improvement should be available
- Predictions can have an influence on the system objectives
  - Ensure that they contribute to the organizations objectives
- Cost effectiveness also affects ML model choice
  - Should use a ML approach that is clearly cheaper and has a better cost/benefit ratio than non-ML approaches

System goals

Layers of success measures



Innate/overall goals of the organization

Measures correlating with future success, from the business perspective

How well the system is serving its users, from the users’ perspective

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

Organizational Objectives

- Businesses
  - Current revenue, profit
  - Future revenue, profit
  - Reduce business risks

- Non-profits
  - Quality of life (e.g., lives saved, animal welfare increased, higher convenience in daily life)
  - Public policy goals (e.g., social justice improved, CO_2 reduced, catastrophes averted)

- Research
  - Knowledge gained
  - (any of the above, depending on type of research)

Implication: Accurate models are often not themselves the goals!

It follows that ML usually only indirectly influences the organizational objectives → Hard to quantify

Leading indicators

- Key factors related to organizational objectives

- Examples
  - Customer sentiment: do they like the product?
  - Customer engagement: how often do they use the product?
  - Time spent using product
  - Changes in customer base (growth, steady, decline)
  - Changes in reviews and ratings
  - …

- Often indirect proxy measures

- Can be misleading
  - Example: more users does not automatically mean higher profits

User outcomes

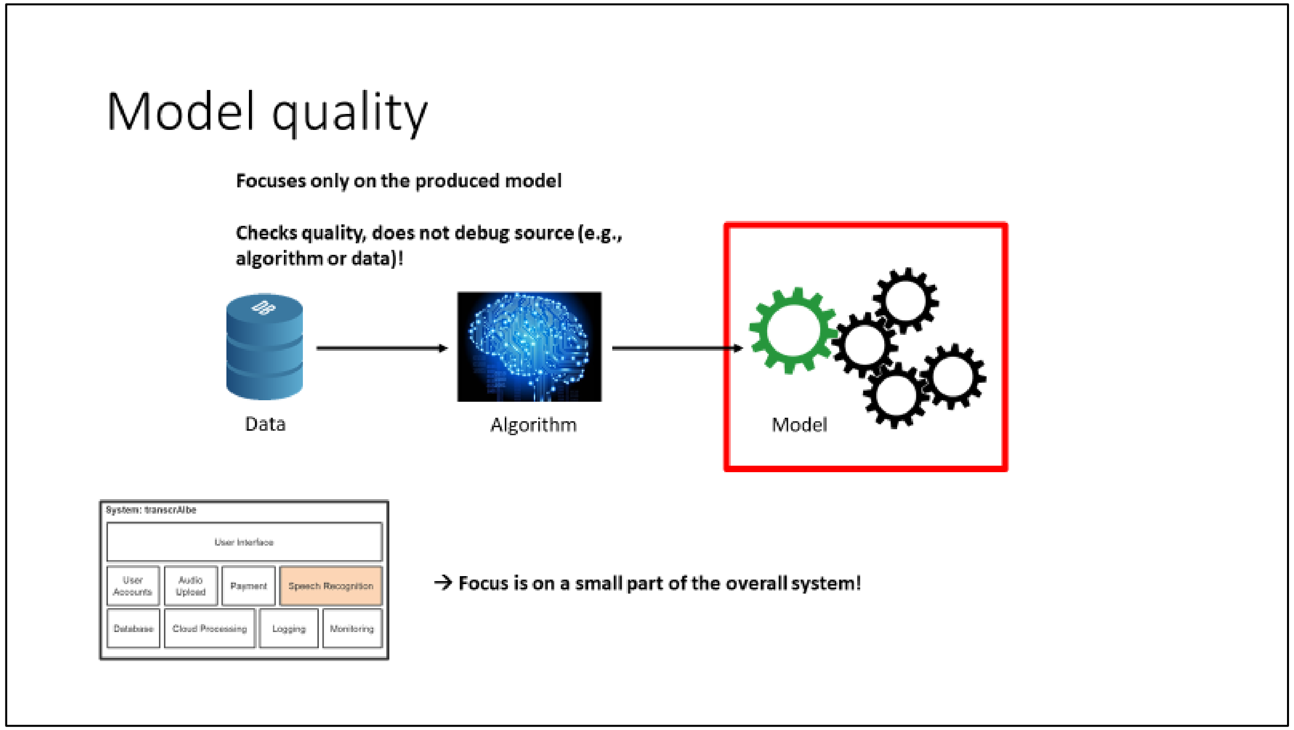
- Measure how the system is serving the users
- Examples
  - Users choose recommended items
  - Users make better decisions
  - Users save time
  - Users achieve their goals
  - …
- Easier to measure than leading indicators
  - Can often be automated
- Only indirect relation to organizational objectives

Model properties

- Directly related to model quality

- Examples
    - Accuracy
    - Rate and kinds of mistakes
    - User interactions
    - Inference time
    - Training costs
    - …

- No direct link to organizational objectives
    - Only indirect through user outcomes



Live exercise

- Consider a movie stream service
- One of your customer promises is to suggest good movies
- What are relevant …
  - organizational objectives
  - leading indicators
  - user outcomes
  - Model properties

Measurement

Defining measurements

Measurement is the empirical, objective assignment of numbers, according to a derived rule 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: finding the value of intangibles in business)

Everything is measurable

- If we care about something, it must be detectable!
  - Quality, risk, security, public image, …
  - Detection may not be easy!

- If something is detectable, then it must be quantifiable
  - Number of bugs, deviation from project plan, positive/negative statements on social media
  - Often only partial aspects

- If we can observe it, we can use this to define measures
  - … but the measures may be imprecise

Measurement terminology

- *Quantification* is turning observations into numbers
- *Metric* and *measure* refer to a method or standard format for measuring something
  - We use both terms synonymously, which is not always the case!
- *Operationalization* is identifying and implementing a method to measure some factors

Measurements

Software Engineering

- Which project to fund
- Need for more testing
- Need for more training
- Execution speed
- Code quality
- Importance of features
- Time and cost estimation
- …

Data Science

- Model accuracy
- Generalization
- Noise in data
- Fairness of models
- Robustness of models
- …

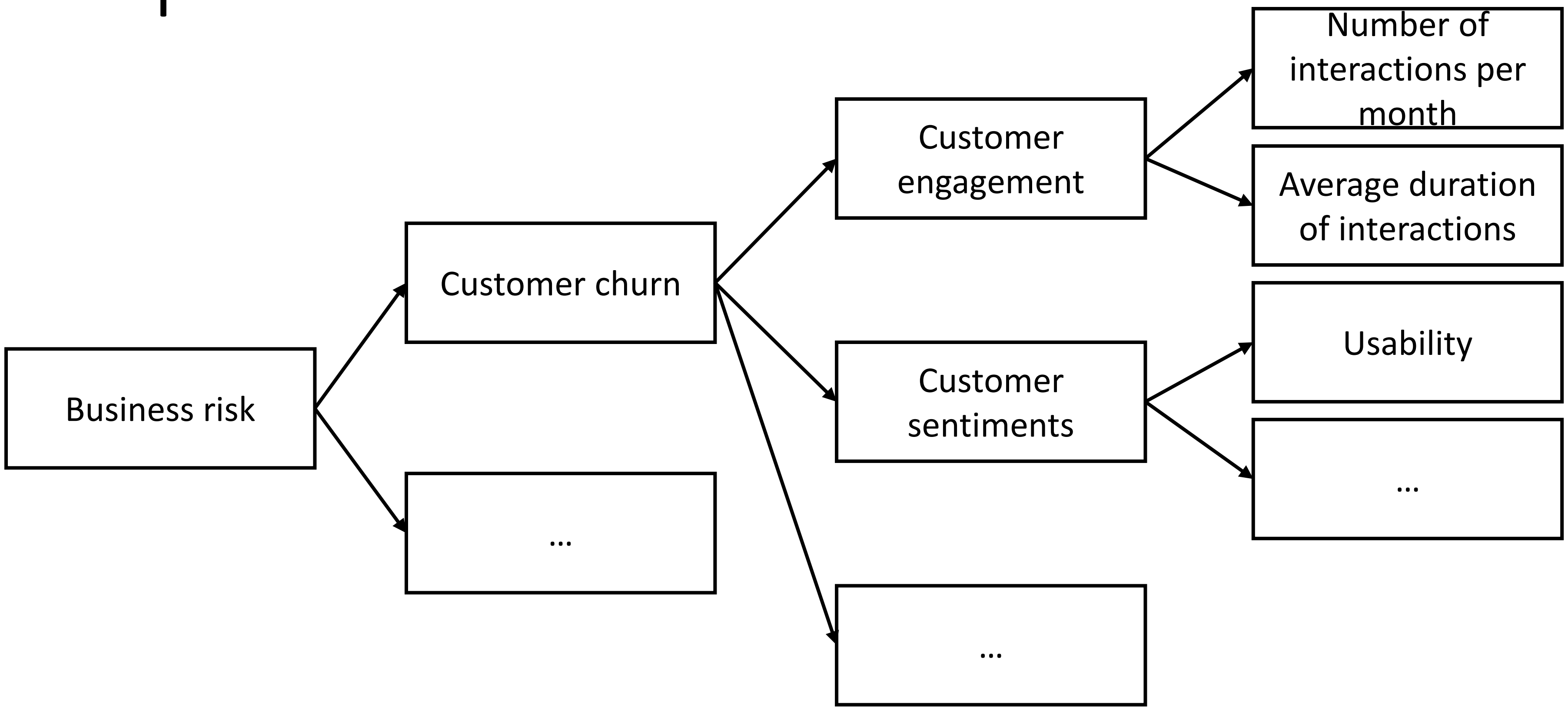
Measurement scales

	Scale	Property	Allowed Operations	Example
Categorial	Nominal	Classification or membership	$=, \neq$	Color as “black”, “white” and “blue”
Categorial	Ordinal	Comparison or levels	$=, \neq, >, <$	Size in “small”, “medium”, and “large”
Numeric	Interval	Differences or affinities	$=, \neq, >, <, +, -$	Dates, temperatures, discrete numeric values
Numeric	Ratio	Magnitudes or amounts	$=, \neq, >, <, +, -, \cdot, /$	Size in cm, duration in seconds, continuous numeric values

Not only relevant for features for ML, but for any measurement!

S. S. Stevens: On the Theory of Scales of Measurement, Science, 103(2684):677-680

Decomposition of measures



Higher-level measure often composed from lower level measures → Clear trace from specific low-level measurements to high-level metrics

Use, e.g., Goal-Question-Metric approach to define the measures

Specifying metrics

measure accuracy		evaluate accuracy with MAPE
evalute test quality		measure branch coverage of Java code with Jacoco
	VS	
measure execution time		average and 90%-quantile response time of REST-API under normal load
measure customer happiness		report response rate and average customer rating on survey shown to 2% of all customers (randomly selected)

Independent party should be able to set up infrastructure and measure outcomes

Live exercise

- What are measures you could define for the movie recommendation service goals?

Risk of measurements

Measurement validity

- Construct validity
  - Are we measuring what we intent to measure?
  - Does the abstract concept match the specific scale/measurement?

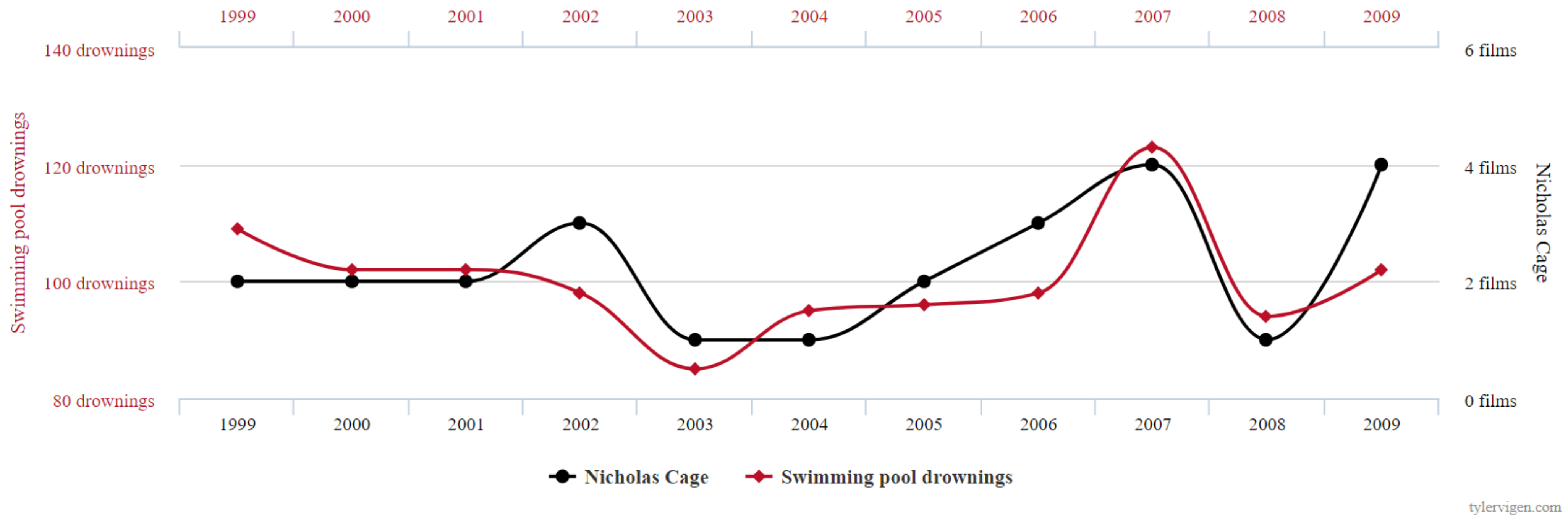
  - Example:
    - What is concept IQ actually measuring? Is the scale meaningful?
    - Are the questions in a usability assessment suitable?

- External validity
  - Generalization of the findings to context and environments, other than the one studied

  - Example:
    - Do the results of a usability assessment on a sample generalize to the target population?

Bad constructs lead to invalid measurements and conclusions

Correlation vs. causation



Did Nicholos Cage sign a contract when he read about drownings!?
Did people really jump into pools because of Nicholas Cage movies!?

Spurious correlation!

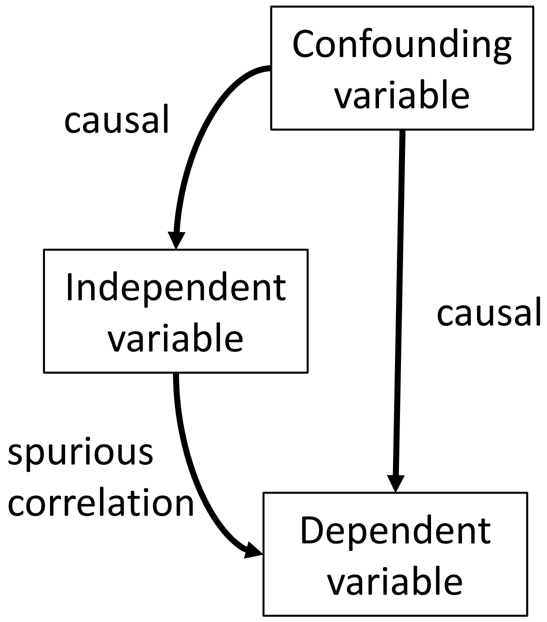
https://www.tylervigen.com/spurious-correlations

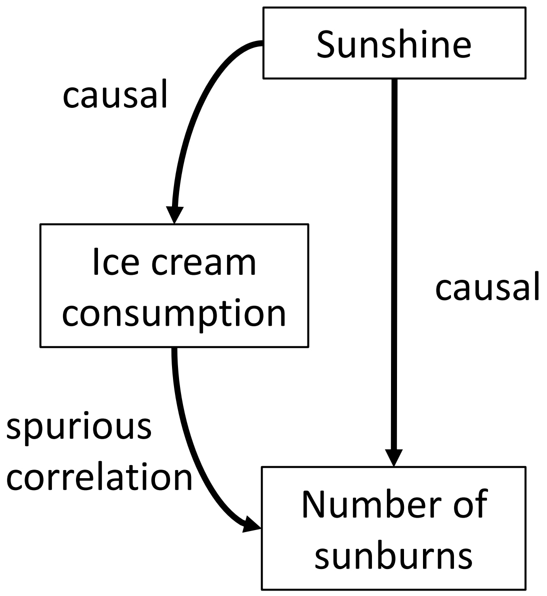
ML (usually) learns correlations!

- ML exploits correlations between inputs (features) and output to build a model
  - Notable exceptions: Bayesian networks, some symbolic AI methods
- Be careful about interpretation & intervention based on correlations
  - Does a positive correlation between exercise and skin cancer mean, we should exercise less to reduce our chance of skin cancer?
- To establish causality you need to
  - develop a theory (X causes Y) based on domain knowledge and independent data
  - indentify relevant variables suitable to measure the predictions of the theory
  - design a controlled experiment with a suitable construct that shows the predicted correlations

That is why checking model quality is important (and difficult)!

Confounding variables

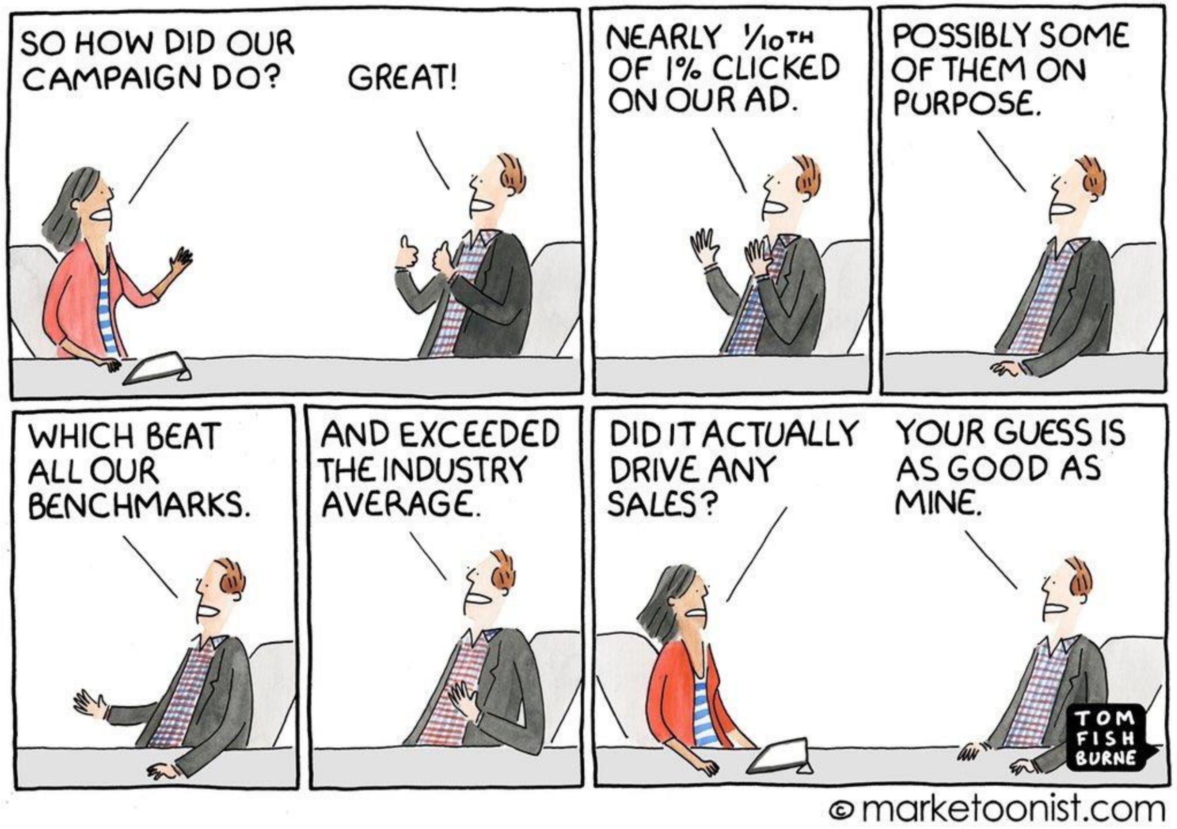




Controlling for confounding variables

- Identify confounding variables
- Control for those variables during measurement
  - Randomize, fix, or measure+account for during analysis
- Example
  - Want to study relation between coffee consumption and lung cancer
  - Use knowledge that coffee consumption is correlated with smoking → smoking as confounding variable
  - Ask study participants if they are smokers, consider this during analysis

Streetlight effect



Danger to avoid: focus on bad and easy to measure metrics in favor of good metrics

Goodhart‘s law

When a measure becomes a target, it ceases to be a good measure.

- Example
  - Number of visits is used as proxy for revenue
  - No problem: revenue is still the target and regularly considered to make decisions
  - Problem: the number of visits are increased, without checking if this is good or bad for the revenue

Questions?