

# Gold standard data

Miroslav Batchkarov  
Teebly



# Slides

[https://github.com/mbatchkarov/dsf\\_london\\_2019](https://github.com/mbatchkarov/dsf_london_2019)

# Machine learning 101

Miro, £2.75 transaction in  
London, 09:51h, Friday, @Strong  
Coffee LTD

OK

User Miro, £1400 transaction in Milan,  
22:51, Monday, Versace store

Fraud

Aim: Teach a machine to do this

# Another example

"Hi doctor, I feel a sharp pain in my chest when I run for more than 30 seconds"



Specialty: cardiology  
Urgency: high

# Machine learning

**Data + Algorithms + Question**

# The data in data science

## Your options

- Get data online
  - Yeah, right!
- A month of unsupervised machine learning
- A week of gathering examples yourself

# This talk

How do we get good examples to train a machine?

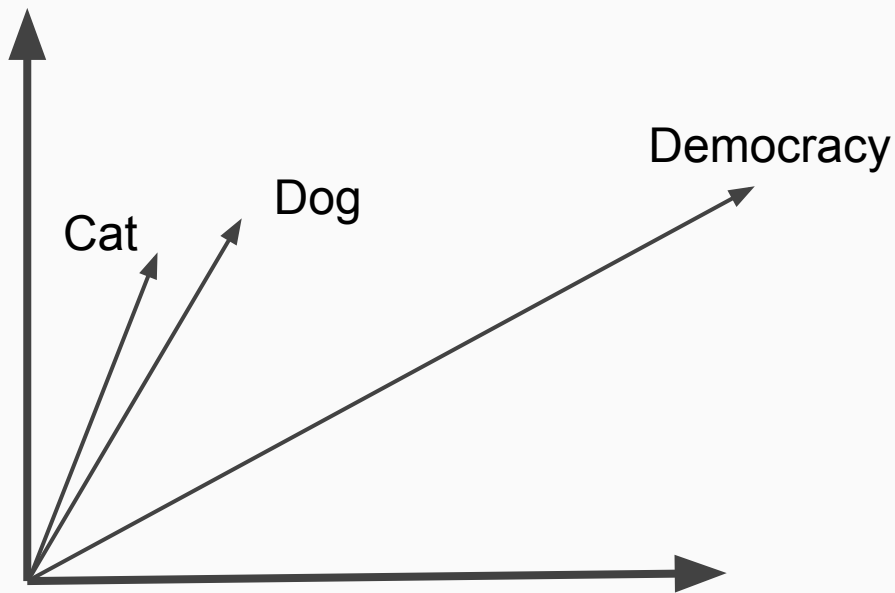
- Garbage in, garbage out
- But how does garbage get created?
- What can we do?

# Many ways to get data

- Automatically, from a system/ device
  - “Logs”
  - Need to gather and transmit
  - A whole different can of worms
- Manually, from humans, using a helper tool



# Case study 1: word embeddings

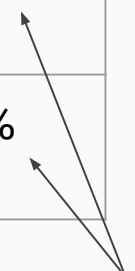


- Word = vector
- Distance = similarity
- Popular recently
- Easy to train
- Interesting semantic properties

# Evaluating embeddings

## Human similarity score

Cat	Dog	80%
Cat	Democracy	20%

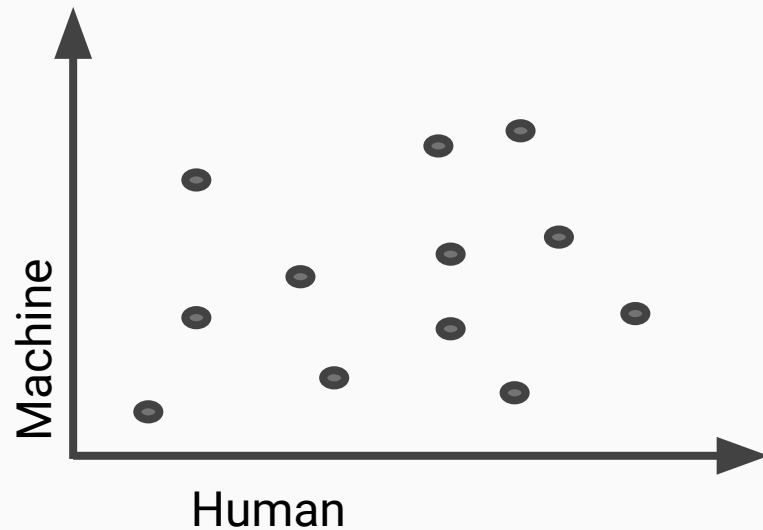
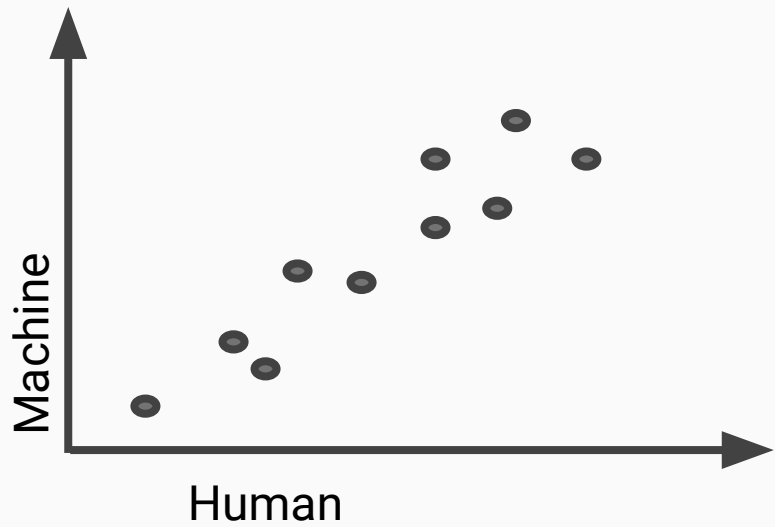


## Machine similarity score

Cat	Dog	76%
Cat	Democracy	58%

We need humans to give us these numbers!

# Agreement as correlation



# Question: is the task clearly defined?

- What makes words similar?
  - Topical relatedness
  - Physical appearance (for concrete nouns)
- Humans need clear unambiguous instructions
  - Corner cases hard to pin down, e.g.
  - Antonyms
- Results
  - Similarity of “Tiger” and “Cat” ranges from 50% to 90% in a popular academic data set

# Question: is the task easy?

- Some tasks are inherently subjective, even with clear instructions
  - Cat vs Tiger similarity (show of hands)
  - Written sarcastic comments, out of context

# What can we do?

- Write down annotator guidelines early
  - Even if you are the one doing the annotation
- Have them proof-read and interpreted by others
- Simplify task
  - Will a coarser-grained annotation schema do?
  - Business case

# Coarse vs fine-grained models



# Question: do you have quality controls?

Many ways to do it -- pick one

- Inter-annotator agreement
- Agreement with a known gold standard
- Random perturbation
- Detect (and remove) random clickers
  - Stackoverflow does a great job



# Stackoverflow edit dialog

Always needs help needing calling or individual page super after sox controller overriding?

I read, that **it's** always **best practice** to call `super.method()`, if you override a method.

However, overriding the `perform()` method of an `UIStoryboardSegue` and calling `super.perform()` throws an exception. If i'm not calling `super.perform()`, **it's** working fine.

Why?

I read, that **matters whats is its getting some observable create chart show** always **knowing that some fields valued 389 14807751365845** to **practice** call `super.method()`, if you **just using A jquery list removes markers** override a method.

However, overriding the `perform()` method **will helps the programming concepts entities cart tpl inites** of an `UIStoryboardSegue` and calling `super.perform()` throws an exception. If i'm not calling **templates 32 int cvalue2 std string lio netty netty common** `super.perform()`, working **but I observer throwed** fine.

Why?

# Stackoverflow edit dialog

Congratulations!

This was only a test, designed to make sure you were paying attention. **You passed.**

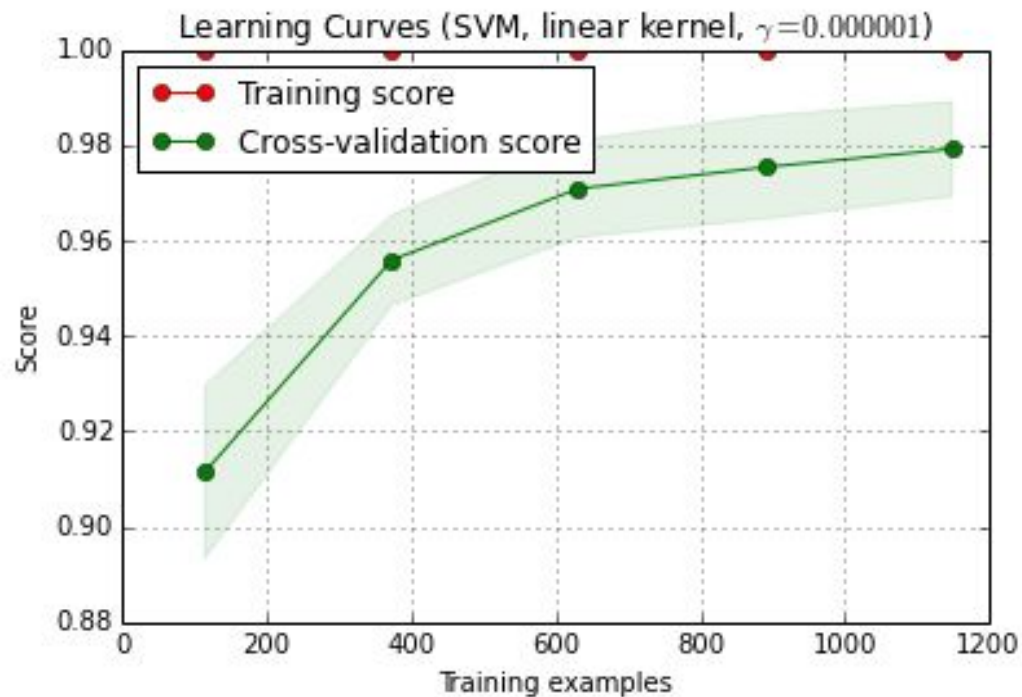
# Quality control (continued)

- Run checks constantly
- Watch out for systematic errors
  - Why? Are instructions clear?
  - Talk to annotators
  - Discard poor data

# Question: how much data?

- Costs can be high, even with crowdsourcing
  - Especially if data has to be discarded
- How much data does the machine need?
- Will more data help?
  - Learning curve

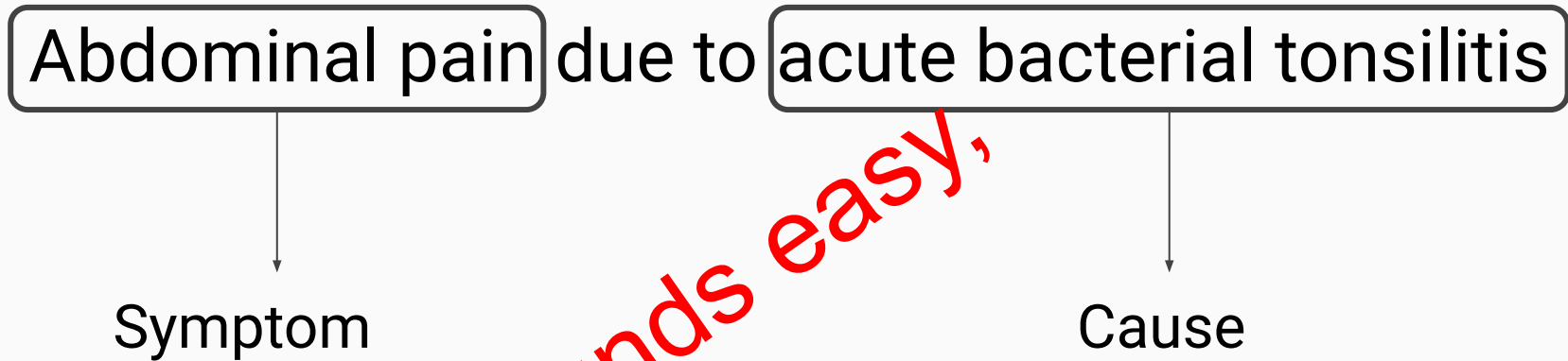
# Learning curve



# Ethics of crowdsourcing

- Annotators are people too
  - Pay fair wages- cheap is expensive!
  - Treat them with respect
- Reliable annotators are hard to find

## Case study 2: symptom recognition



Sounds easy,  
right?

# Challenges

Abd pian//acute b tons//n/a aleggies.

- Lots of terminology
  - Impossible to read for a layperson
- Extremely messy language
  - Can't even find word boundaries



# Question: do annotators need training?

- Task has to be done by a trained physician
- They prefer to practice medicine
- Typically not experts in linguistics/machine learning
- Can you explain the task?

Question: can you measure agreement?

Abdominal pain due to acute bacterial tonsilitis

Abdominal pain due to acute bacterial tonsilitis

Abdominal pain due to acute bacterial tonsilitis

# Question: can you measure agreement?

- No longer a simple correlation
- Need to consider
  - Partial overlap
  - Boundary mismatch
  - Missing units
  - Agreement by chance
- Artstein & Poesio (2008), Savkov (2016)

# Question: do you need specialist tooling?

- Technical requirements
  - Reliable and bug-free
  - Access control and data security
  - Can you use a cloud provider?
  - Continuous quality monitoring
  - Easy to write or install and maintain

# Tooling (continued)

- Non-technical requirements
  - User-friendly
  - Enables your annotators
  - Increases productivity
  - Bug-free

## Tooling (continued)

Crowdfunder, Mechanical Turk, BRAT, Prodigy, custom (?)

Diagram illustrating Named Entity Recognition (NER) on the sentence: "Por Viruca Atanes Madrid, 24 may (EFE)."

The entities are labeled as follows:

- PER** (Person): Viruca Atanes
- LOC** (Location): Madrid
- ORG** (Organization): (EFE)

The sentence is segmented into: Por, Viruca Atanes, Madrid, 24 may, (EFE).

# People issues

- Controversial claim
  - A single reliable annotator can contribute more to your project than a team of programmers
  - Poor annotation can kill your project faster than anything else

# People issues

- Build a good working relationship
- Helps to be in the same room
  - Identify misconceptions faster
  - Helps **you** learn about the problem
- Re-training if necessary
- Day-to-day problems



# Crowdsourcing

- All risks above, plus possibly
  - Poor command of English/German
  - Random clickers
  - Steep learning curve

# Case study 3: Address matching

1 April 2018

**CONFIDENTIAL**



Thierry Lehmann  
Ober-Beerbacher-Str. 27b  
Bettlach 2544  
CH

Dear Mr Lehmann

**PRINCIPLE STATEMENT OF TERMS AND CONDITIONS : FIXED  
TERM CONTRACT**

I am delighted to confirm your appointment as software developer with Finanz AG (the 'Company'). This document outlines the Terms and Conditions that apply to your contract, and other information which is relevant to your employment.

1. Subject to the terms and conditions outlined below, this contract will be

# Question: can you get data?

- Available but proprietary & expensive
- Questionable quality

# Data generation and augmentation

- No data? No problem. Just make it up
  - Images: rotation, translation, lighting
  - Spell checking: random transposition/replacement
  - Addresses: mix and match known fields

# Summary

- Get to know the problem domain
- Challenge your assumptions
- Do not be afraid to start from scratch
- Monitor quality continuously
- Be careful of crowdsourcing

# Thank you

Miroslav Batchkarov

@loglinear

@Teebly\_HQ