Text Mining reliable methods

Antske Fokkens & Pia Sommerauer
antske.fokkens@vu.nl pia.sommerauer@vu.nl
CBS 1 July 2019

Planning

- Session 1:
  - Brief recap: basics of text mining
  - Diving deeper in evaluation
- Session 2:
  - Recap machine learning & neural networks
  - Word Embeddings

Planning

- Session 3:
    - Latest advances in Natural Language Processing
        - Feature engineering vs implicit representation
- Session 4:
    - Knowing your language model: Analyzing neural networks

Text Mining

- Minimal Requirements:

1. Well-defined task

2. Representative evaluation data

Text Mining

- What do you want to know?
- What information do you have?
- What technologies can be used to extract what you have from what you want to know?

Approaches

- Rule-based with resources
- Supervised machine learning
- Semi-supervised machine learning
- Unsupervised machine learning

Approaches

- Rule-based with resources
- **Supervised machine learning**
- **Semi-supervised machine learning**
- Unsupervised machine learning

(semi-)supervised machine learning

- overall question:

  what **information** is **relevant** and **how** to get this in the model?

- explicit feature representation

- feature learning

Evaluation

- standard evaluation in NLP (minimal requirement)

- this is not enough….

Annotations

- Make information that is *implicitly* present **explicit**

- Assign meaning to signals and can be stored or queried

Annotation Methods

- Manual expert annotations
- Crowd annotations
- Machine annotations

Methodology

- Reproducibility: possibility of repeating the study
- What can be done by which approach?
  - what goes wrong?
  - how does this influence the overall conclusions?

Expert annotations

- Annotators are trained for the specific task
- Training phase (with possible revisions)
- Annotators work independently

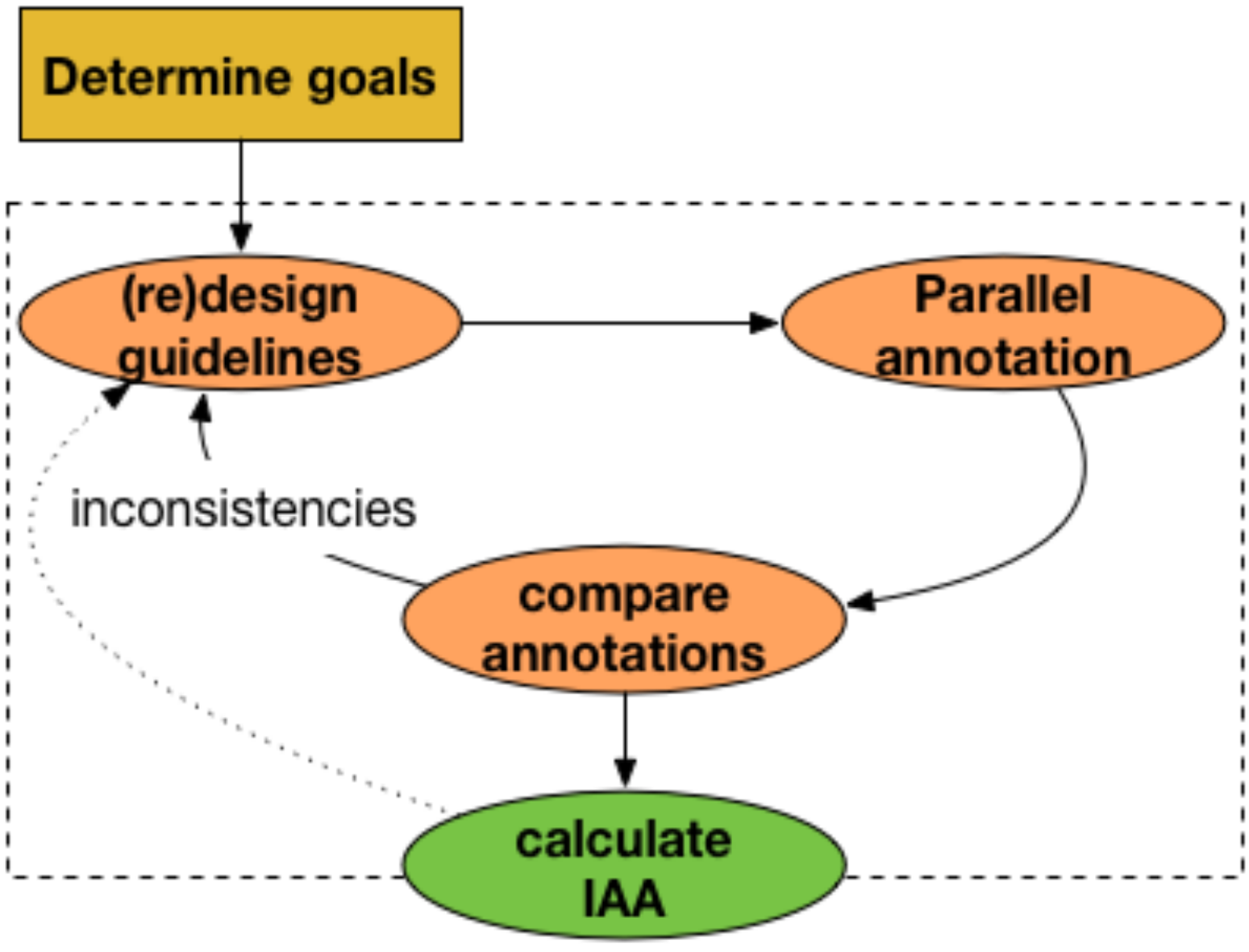
Coen was a person, very **modest** in life style, of **good character**, **not a drunkard**, **not haughty**, a **very proficient council** and **well-educated in bookkeeping and business.**

Entities

SENTIMENT

The Pangeram of Bantam, a **Muslim priest**, and so **an utter enemy of Christians** was at the head of government while **the King** was a minor.

Expert annotation



Named Entities

Coen

The **Pangeram** of **Bantam** /
The **Pangeram of Bantam** /
The Pangeram of Bantam

Prof. **Westerdijk** /
Prof. Westerdijk

Doctor **Watson** /
Doctor Watson

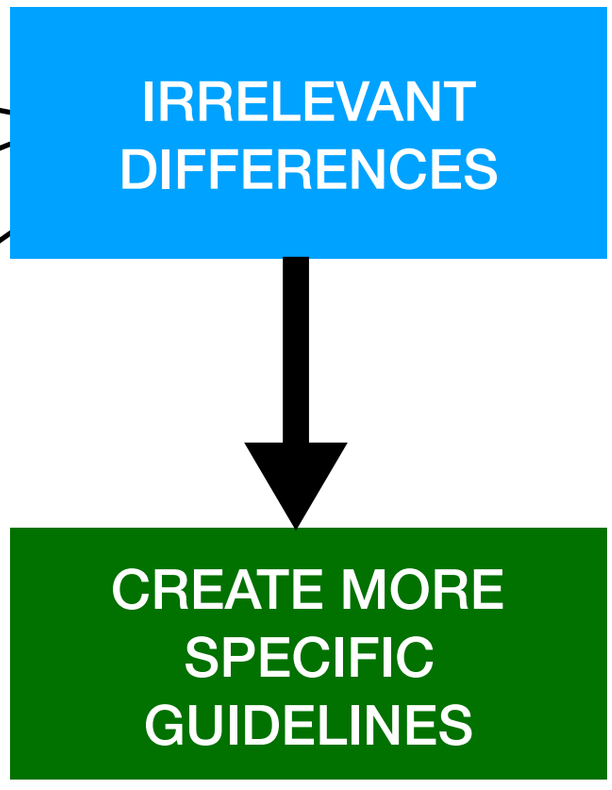
Named Entities

Coen

The **Pangeram** of **Bantam** /
The **Pangeram of Bantam** /
The Pangeram of Bantam

Prof. **Westerdijk** /
Prof. Westerdijk

Doctor **Watson** /
Doctor Watson

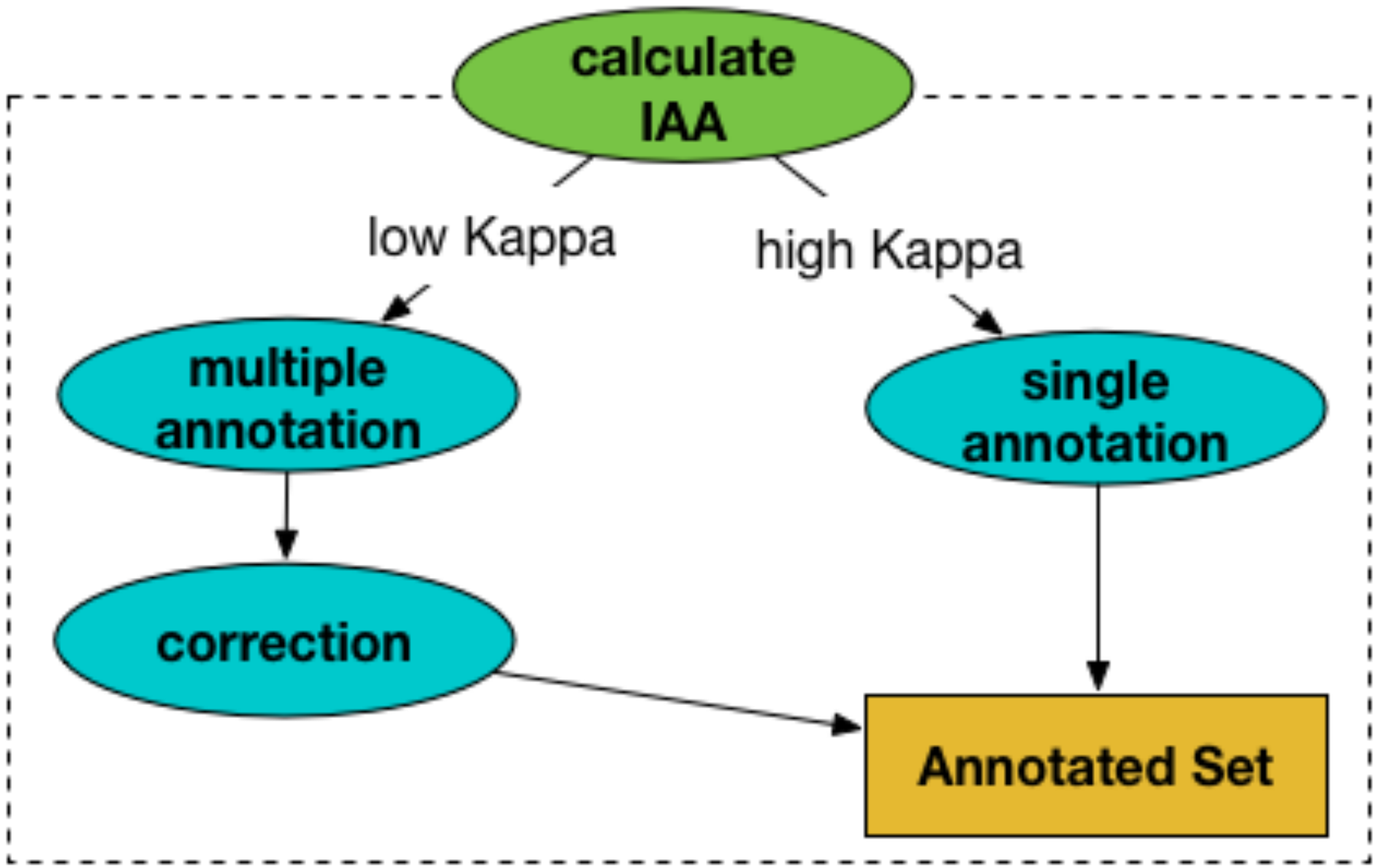


Cohen’s Kappa

$$\text{Kappa} = \frac{\text{Observed\_Agreement (Pa)} - \text{ChanceAgreement (Pe)}}{1 - \text{ChanceAgreement (Pe)}}$$

κ	**Interpretation**
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

Expert Annotations



Expert Annotations

- Method: good IAA -> good guidelines
- Usage:
  - `gold data' for machine learning
  - directly applied to large corpora to address social science/humanities research question

Draw-backs

- Relatively expensive, when used
  - as general research method (full corpus annotation)
  - to create training data
- Does `gold' really exist?

Motivations for the crowd:

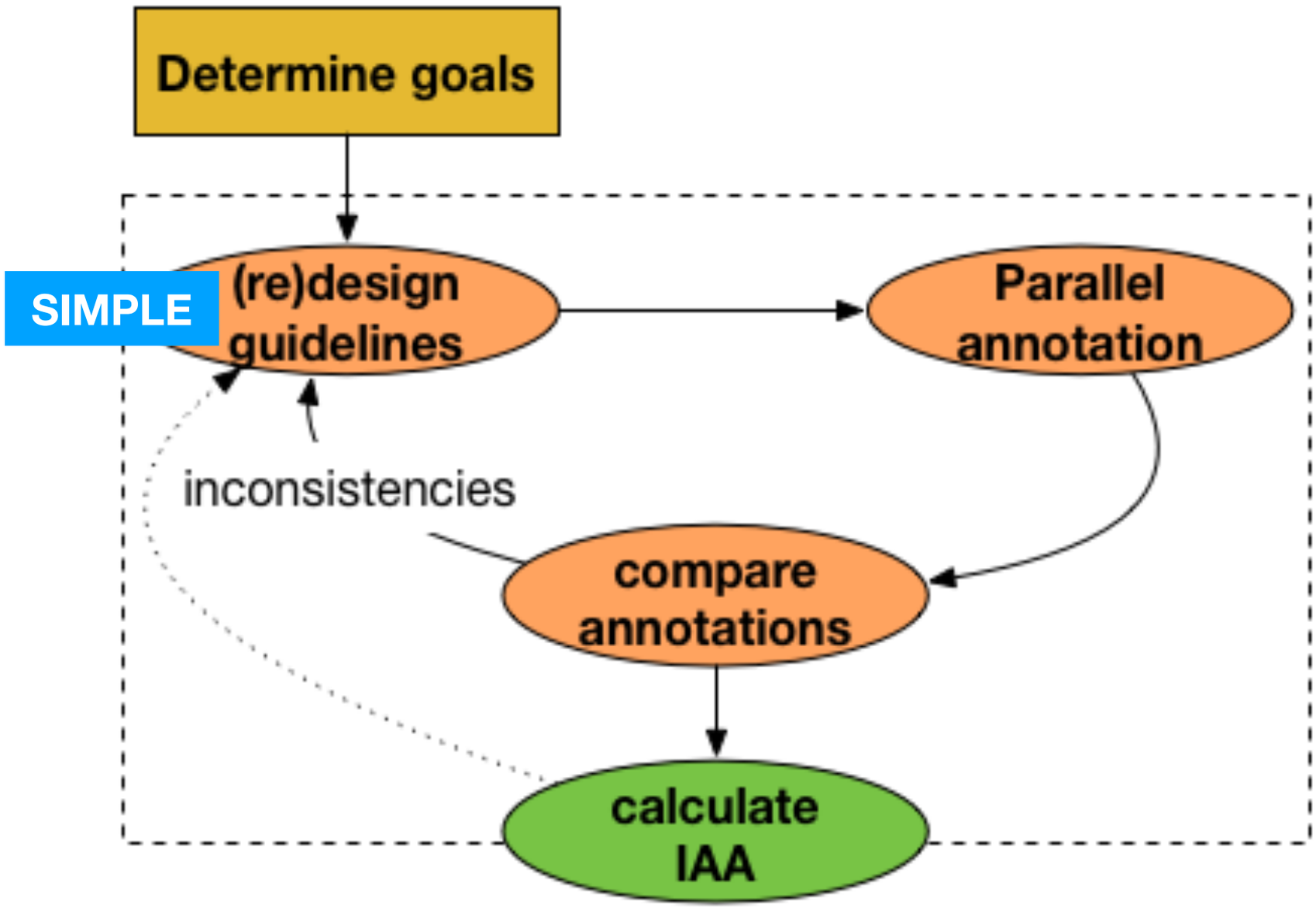
- Much cheaper than experts

- The average of multiple laymen is worth more than that of one expert

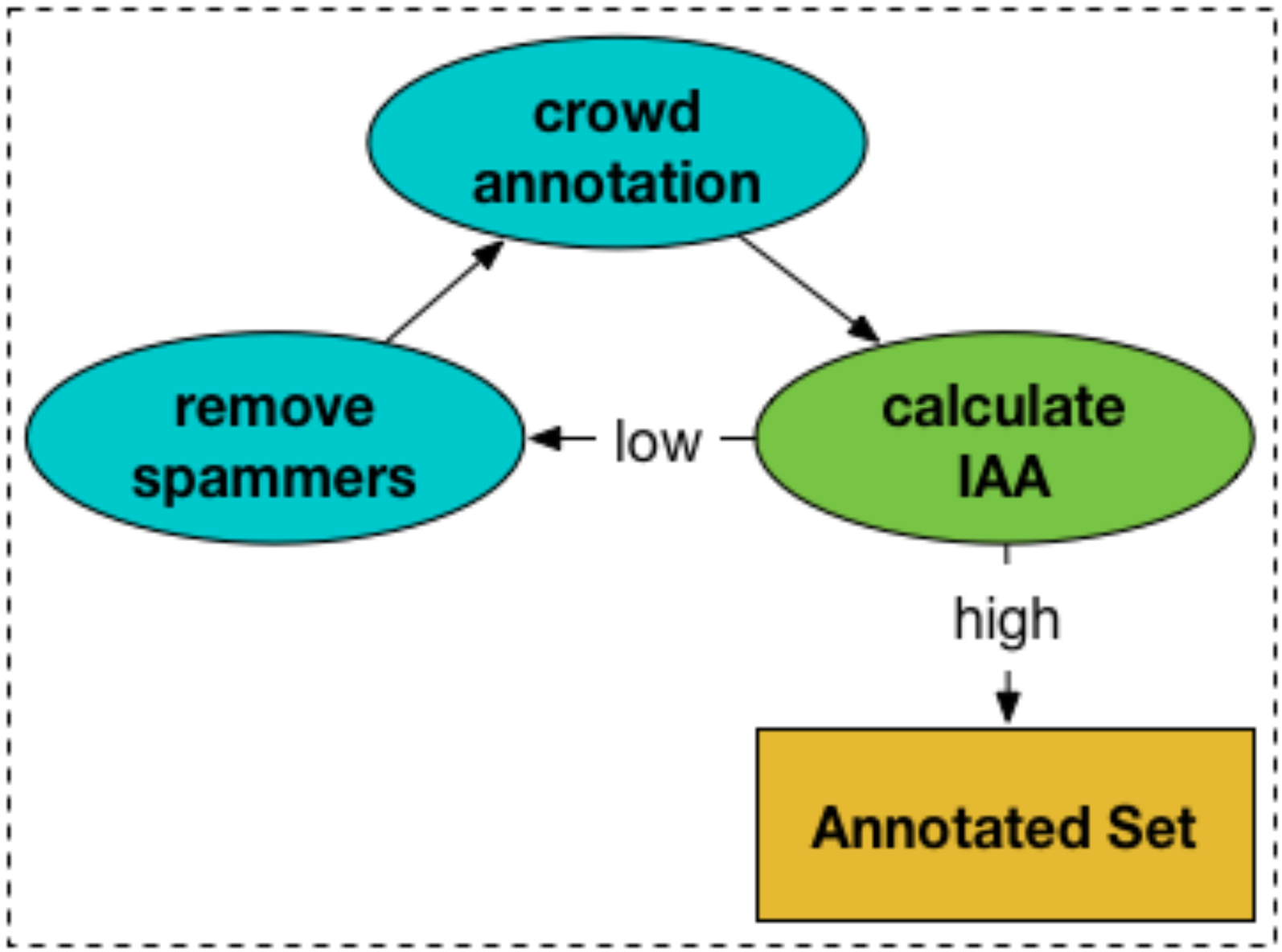
Sir Francis Galton

- Asked 787 people to guess the weight of an ox
- Nobody guessed right
- The collective guess (average) was almost perfect

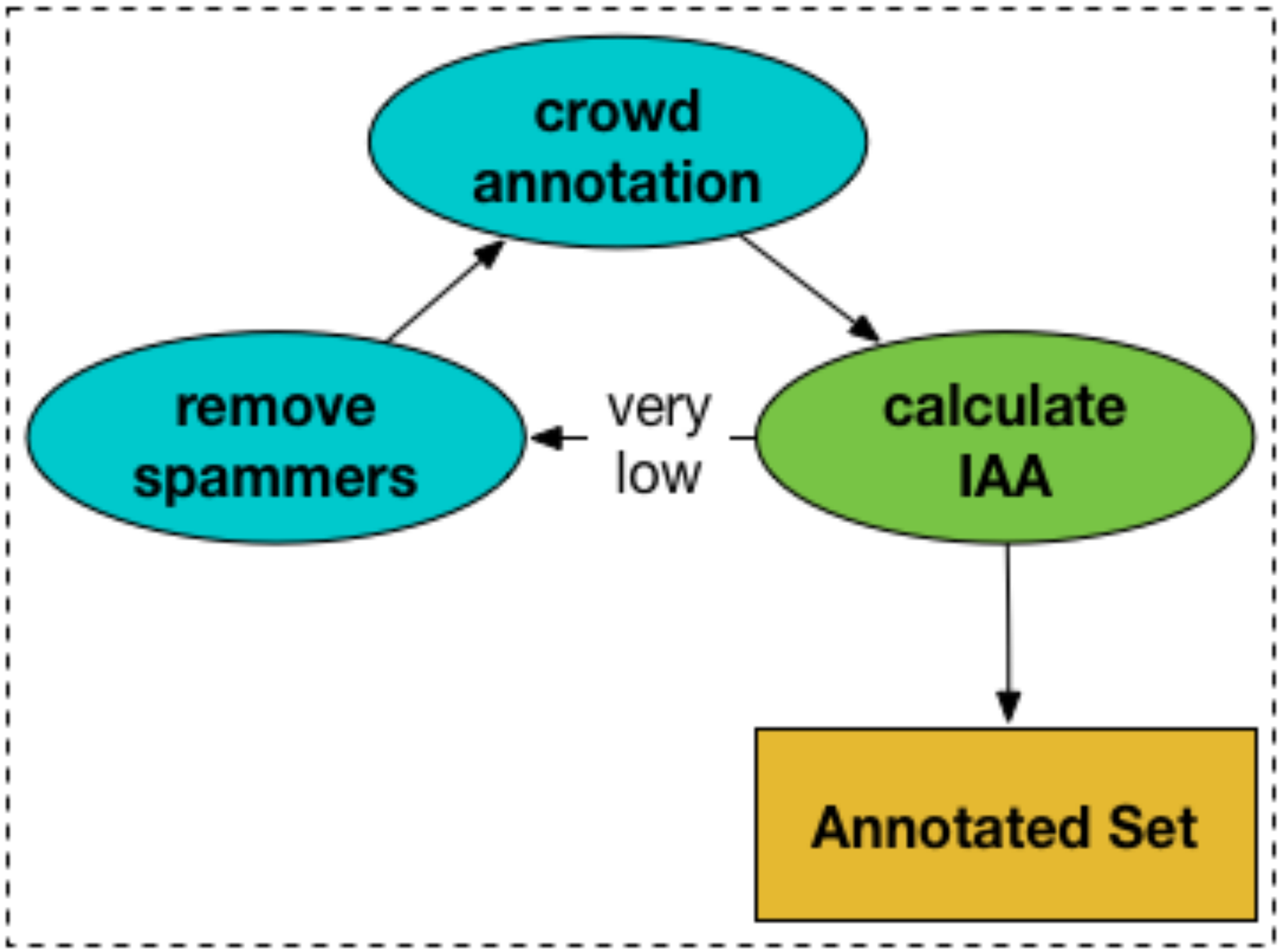
Crowd for Gold



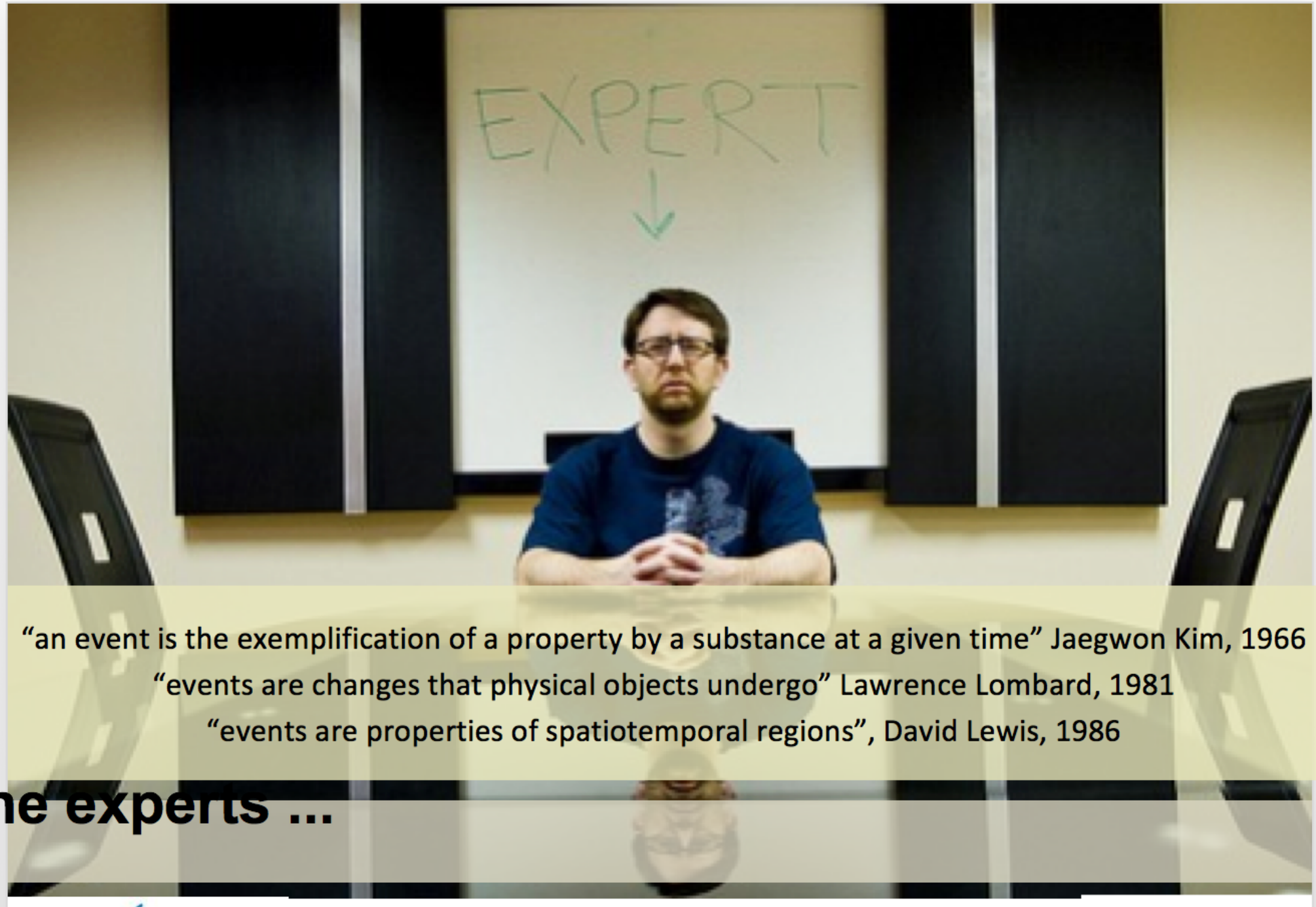
Crowd for Gold



CrowdTruth


crowd annotation
remove spammers
very low
calculate IAA
Annotated Set

If you ask the experts ...



many things such as: An **observable occurrence, phenomenon** or an **extraordinary occurrence**."

"an event is an **incident** that's very **important** or **monumental**"

"A **planned** public or social get together or **occasion**."

An event is something occurring **at a specific time** and/or **date** to **celebrate** or **recognize** a **particular occurrence**."

thing like a **function is held**. you could tell if something is an event if there **people gathering** for a **purpose**."

If you ask the crowd ...

`expert' annotations

- A token or sequence of tokens that refers to a (hypothetical) happening or state that can be bound to a time

- Typically expressed by verbs, but not copula, modals or auxiliaries

- Nominalizations

- Adjectives and attributes when occurring in a predictive construction

Top Israeli officials **SENT** strong new **SIGNALS** Sunday that Israel wants to withdraw from southern Lebanon, ...

CrowdFlower

does not refer to an event → it seems to refer to an inference or communicated feeling more than specific event.

refers to an event → a group of people did something specific at a specific point in time.

refers to an event → the actors in question (top Israeli officials) performed an action during a specified time (Sunday).

refers to an event → it refers to what the israelis did on sunday, a specific time.

That 1978 resolution calls for Israel's unconditional **WITHDRAWAL** from the self-declared security zone it occupies in south Lebanon, ...

does not refer to an event → it is not a particular movement that has or is going on but a request that the country of Israel remove their forces from the zone they occupy.

does not refer to an event → the sentence is speaking of a demand for a withdrawal that had not yet occurred.

refers to an event → Because it is describing a historical issue concerning the resolution of 1978

CrowdFlower

Crowd

- Method: new annotations -> comparable results
  - most workers understand task in similar way
  - filtering of `spammers’
- Usage:
  - Gold (though often verification with experts required)
  - Training data
  - Annotations with subjective components

Draw-backs

- Limited control on workers
- Limited possibilities of instructing annotators
- Simplification of task

Creating Good Evaluation Data

- Representative (similar to what it will be used for)
- Awareness of potential bias: what will systems use to learn?
    - possibly this will only become clear after models have been created

Text Mining for down-stream research

- Apply both *intrinsic* and *extrinsic* evaluation

- How good is the output of the model?

- How suitable is it for a specific task?

Example

- Biography Portal of the Netherlands: descriptions of Dutch citizens and inhabitants

- Goal: Detect and disambiguate locations in the Biography Portal of the Netherlands

- Approach: prefer location in the Netherlands over locations elsewhere

Research questions

- Where are the officials working in The Hague originally from?
- Which city/region produced the most painters/poets?
- How often did officials from the Dutch colonies visit the Netherlands?

Research questions

- Where are the officials working in The Hague originally from?

OK

- Which city/region produced the most painters/poets?`

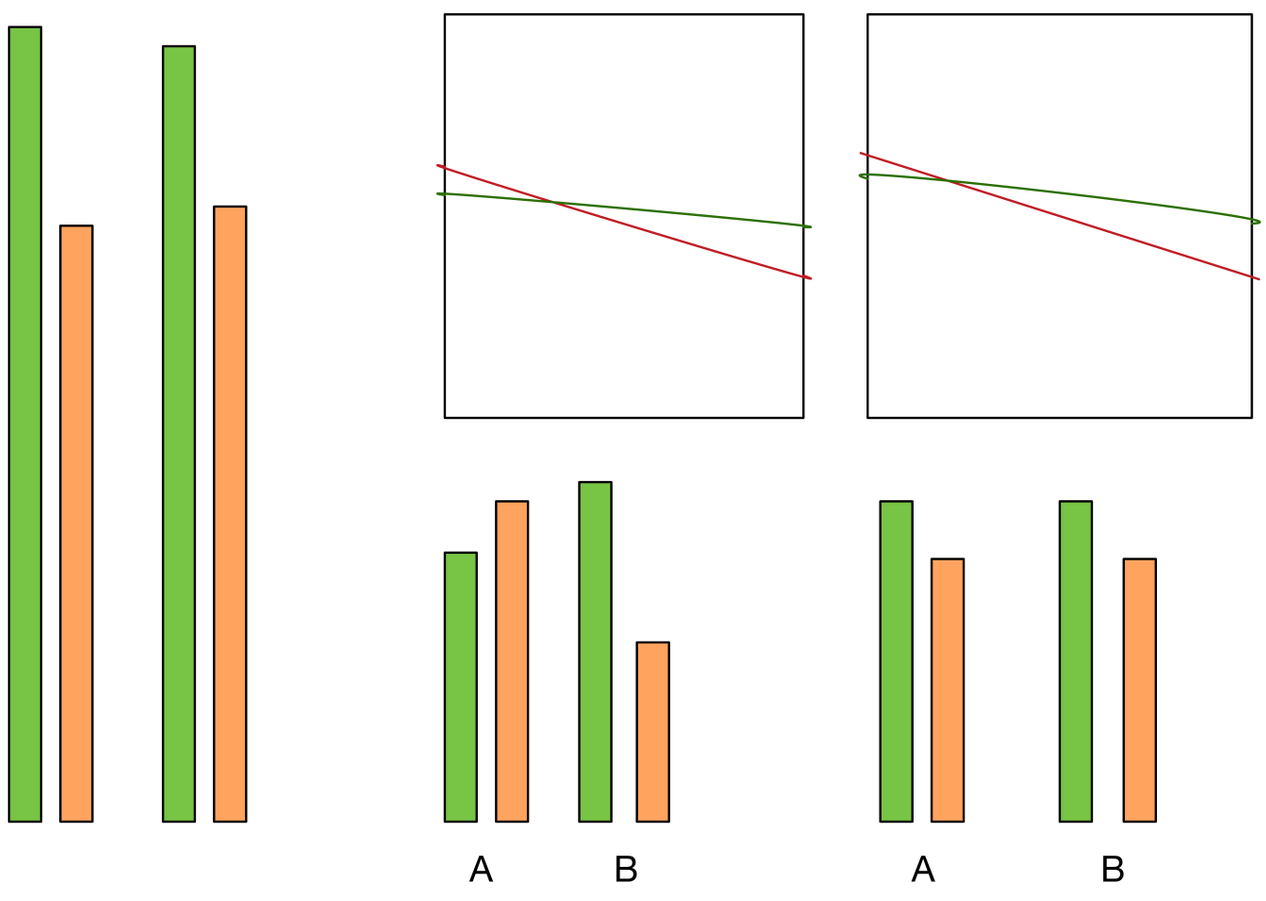
OK

- How often did officials from the Dutch colonies visit the Netherlands?

PROBLEM

Bias (sentiment mining)

- Is the overall trend the same?



Social Media Analysis

- Detect sentiment
- Detect offensive tweets
- Detect hate speech
- Detect racist remarks
- Detect cyberbullying

how to define the task?

how to obtain good data?

how to avoid biases?

Creating Good Training Data

- More is better:
    - it can be useful to create less quality data to have more volume
    - garbage in, garbage out: there are minimum requirements
- if your evaluation is good, you will at least know