HUDK

5053:FEATURE

ENGINEERING

STUDIO

11/3/16 8:28 AM

Report Out

- Goal for week
- Problem/Barrier
- Discuss solution

Natural Language Processing

NLP

Analyses of language produced by humans (by computers)

- Treats language as a varied pool of information sources
- In order to:
  - Understand language (Cognitive Science)
  - Respond to the speaker appropriately (AI)

- Examples
  - Translation
  - Automated feedback (education, shopping)
  - Study linguistics, cognition, development, etc.

Methodological History

1930s

Understanding

Rule based

- Complex sets of rules (grammar/syntax)
- Chomsky

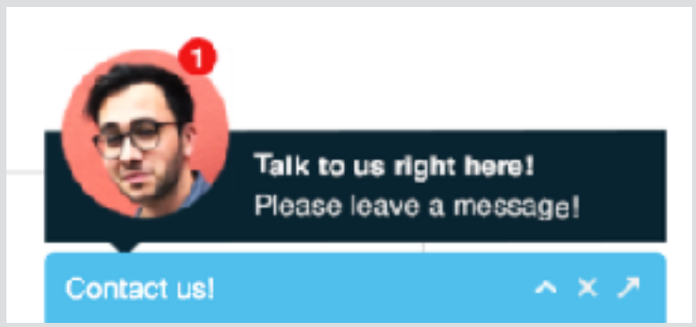
1980s

Processing

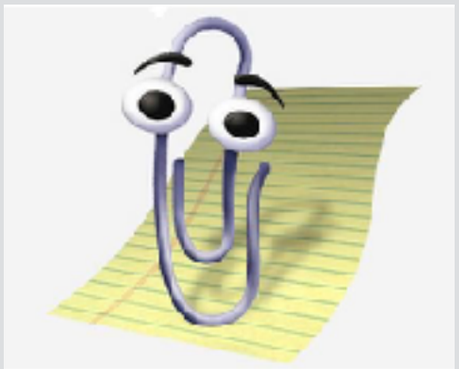
Statistical

- Infer rules from data
- IBM

Industry



Education

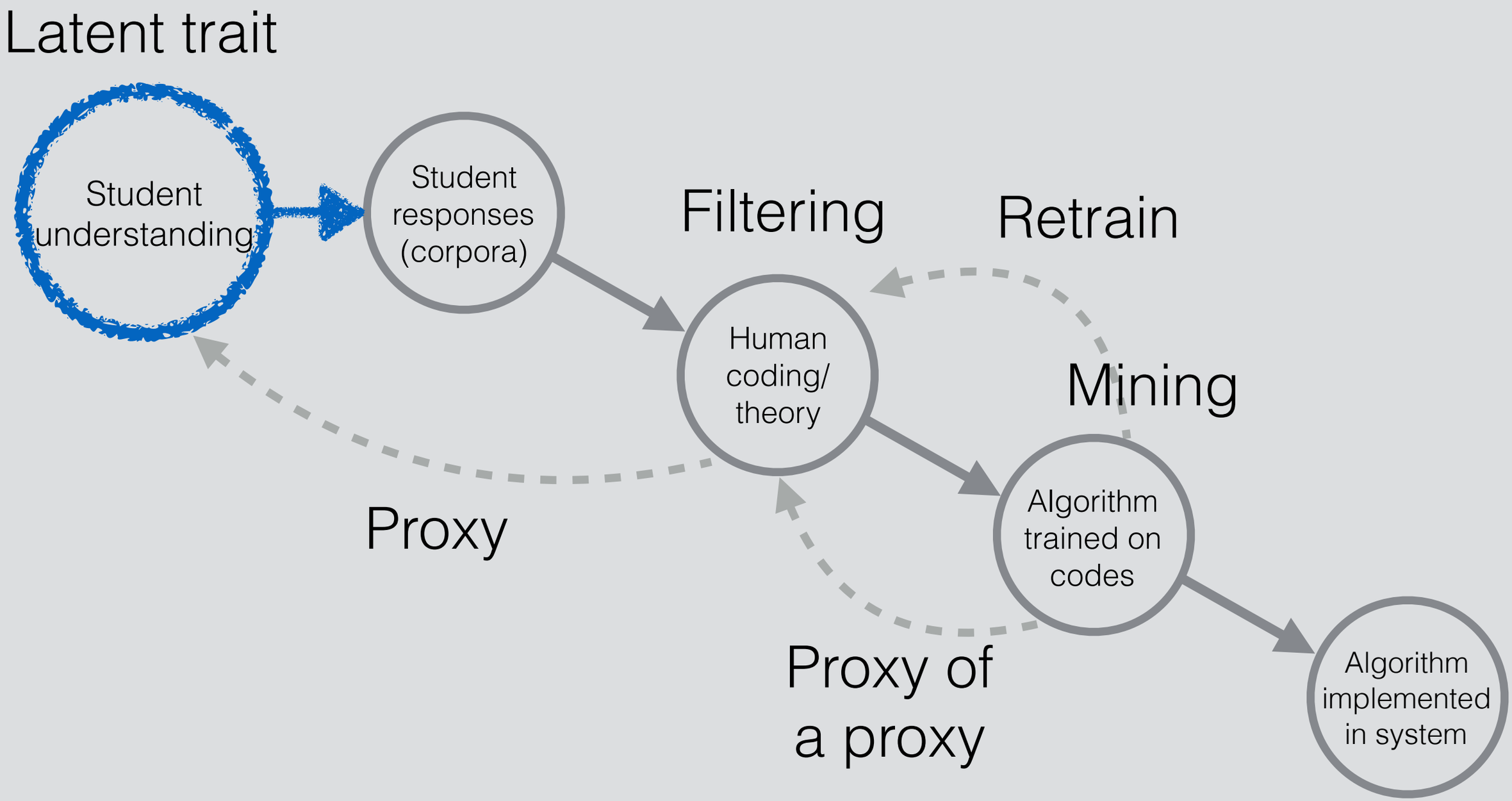


Essential Problem

- Heterogeneity
- We get rid of this by asking MCQ questions - but we also throw out a lot of information when we do that

- Collect more data and more complex data through written answers

Overall Method



Coding

Word counting

Google books Ngram Viewer

Types of Expressions

"I don't know…"
"I dunno…"

Stemming

Take the root of the word:
educate, education, educating

Tokenization (bag of words)

Chopping word/phrase into *tokens*

- Remove punctuation
- Find best number of letters to represent a word/meaning
- Consider all possible versions of word
- Stop word removal

↓

Features

Algorithms

Feature selection

- Not all tokens are useful, which ones can we scrap?

Feature extraction

- Extracting features from combining tokens