Learning Composite Events: Visual Exploration of Temporal Event Sequences through Volume and Variety

Andreas Mathisen
PhD student

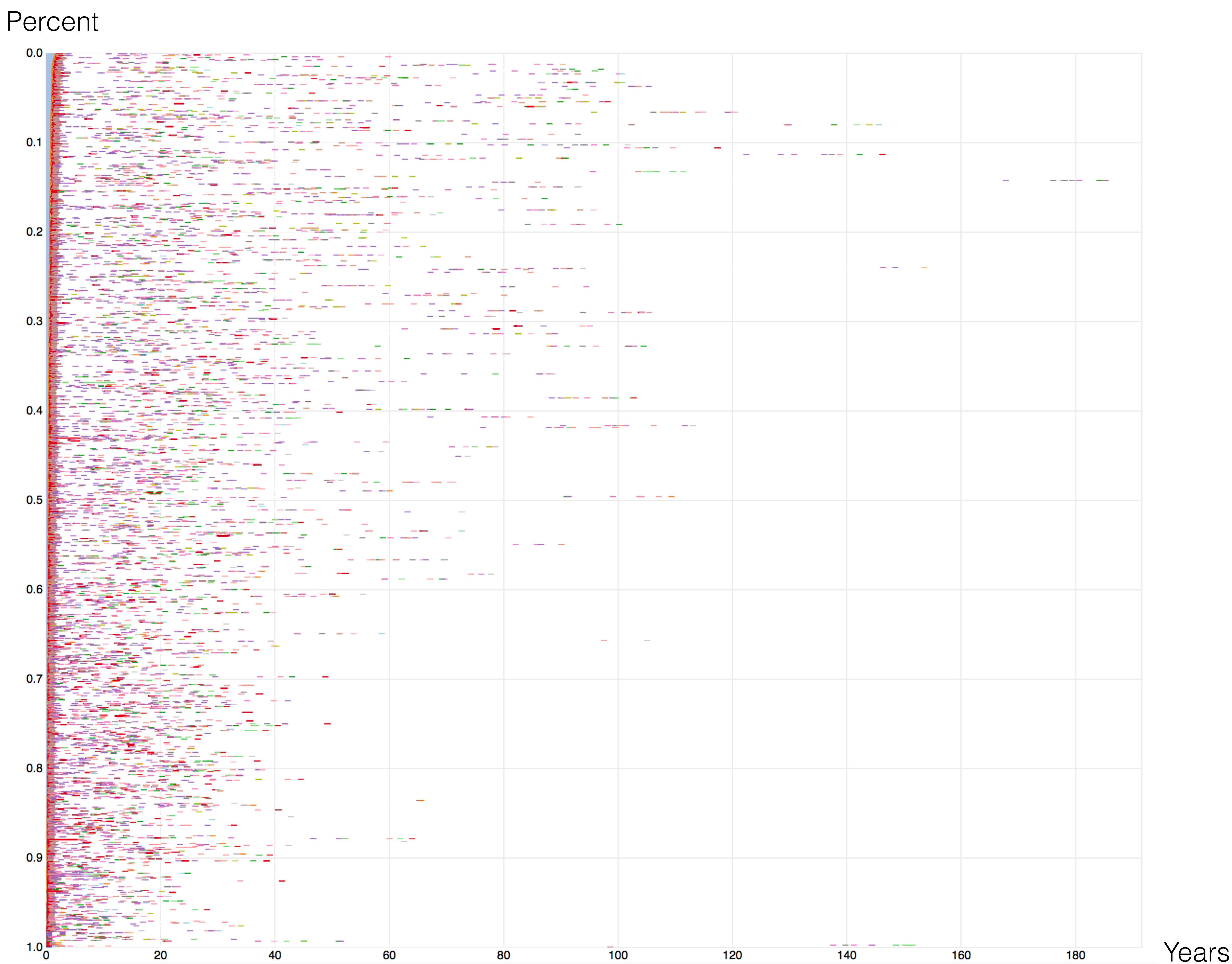
AARHUS UNIVERSITY

Volume and Variety

Temporal event sequence: a series of timestamped events (and potential attributes), which together form a sequence (or record).

Volume: number and length of event sequences.

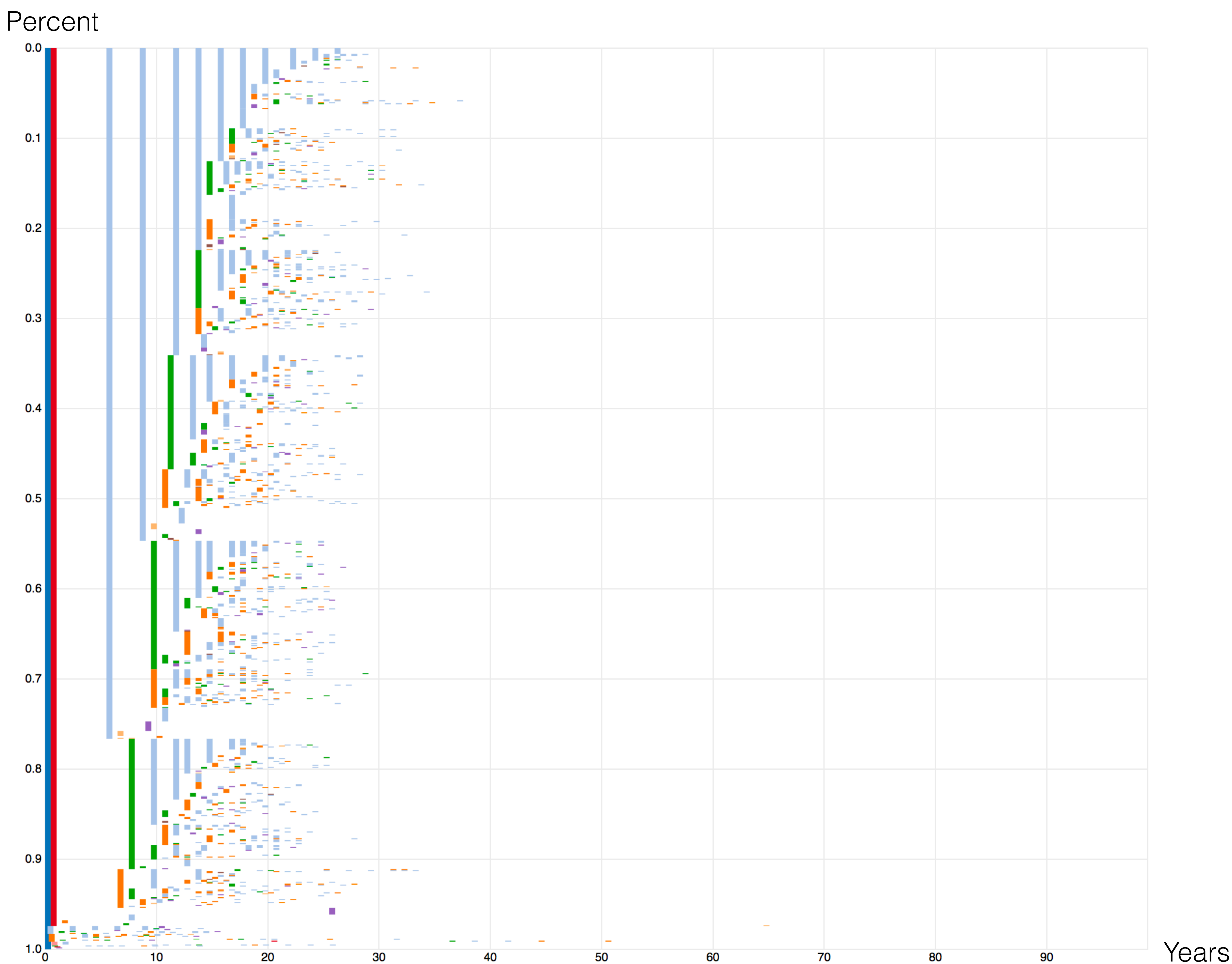
Variety: number of event types (and event attributes) and variation in time.


Percent
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
20
40
60
80
100
120
140
160
180
Years

Composite Events

Goal: reduce variety to allow for aggregation of sequences that would otherwise be unique.

Idea: learn composite events based on time segments.

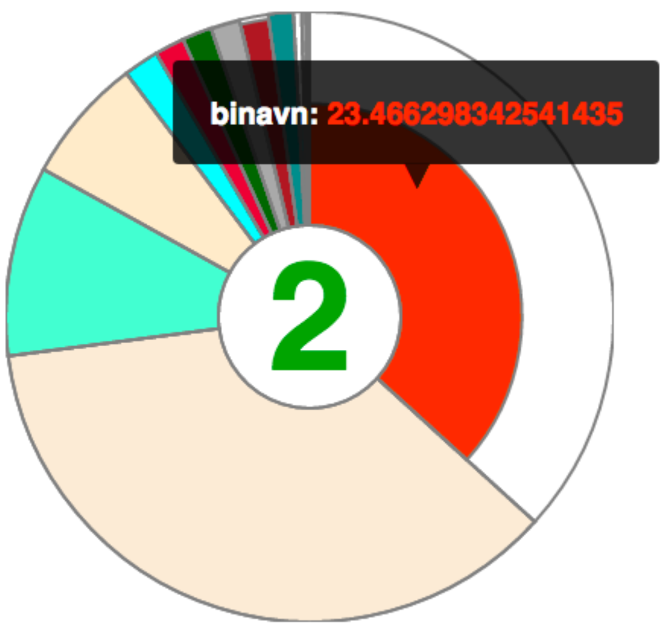


Percent
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
10
20
30
40
50
60
70
80
90
Years

Challenges

Explain to users what the new high-level events mean

Find suitable parameters (number of clusters, time segment sizes)



AARHUS
UNIVERSITY

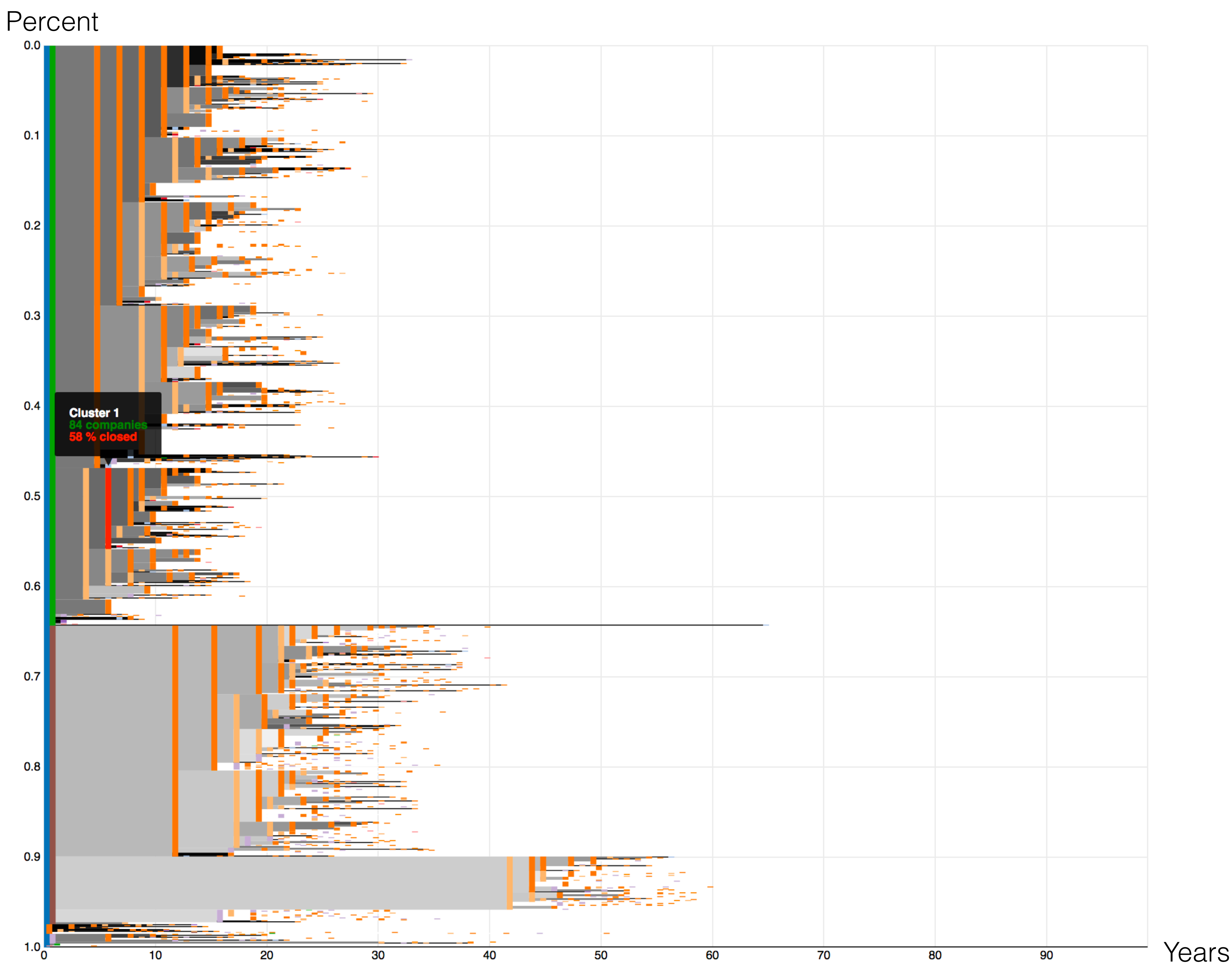
Sequence Outcomes

Sequence outcomes can be used for two things:

1. Give meaning to the composite event sequences.

2. Guide the automatic search for number of clusters and time segment sizes.

An example of an outcome is whether a company went bankrupt



Percent
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Cluster 1
84 companies
58 % closed
0
10
20
30
40
50
60
70
80
90
Years

Computing Flood Risk Based on Sea-Level Forecast

Yujin Shin

PhD Student
MADALGO, Aarhus University

DABAI

MADALGO CENTER FOR MASSIVE DATA ALGORITHMICS

AARHUS UNIVERSITY

Motivation

- Storm surge
  - Cyclone Xaver (DMI: Bodil, 4th December 2013)


05. DEC. 2014 KL. 06.54 | OPDATERET 05. DEC. 2014 KL. 06.56

Stormfloden efter Bodil sender stadig regninger til Stormrådet

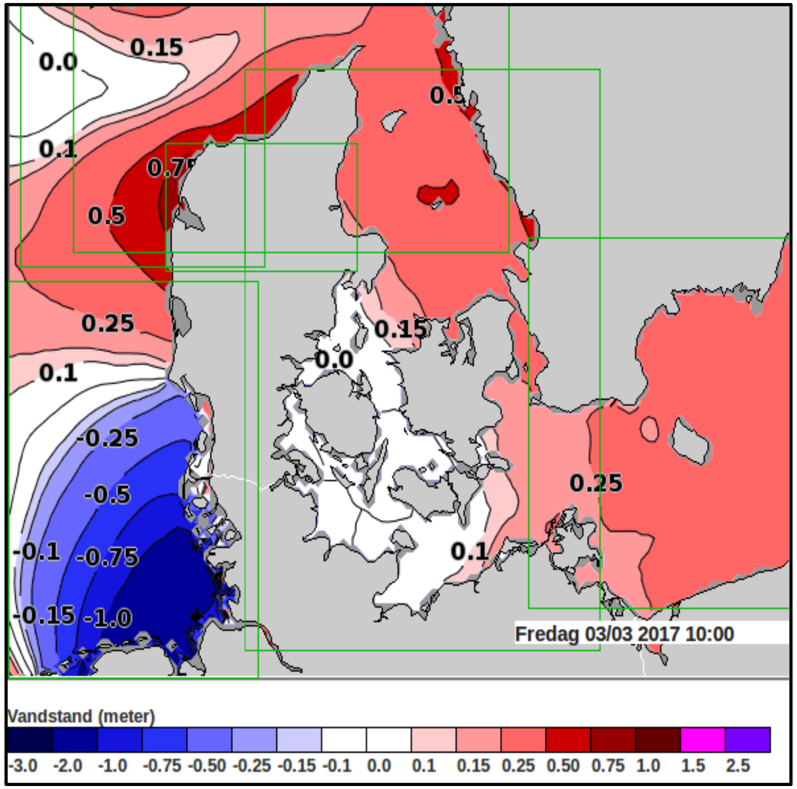
630 mio. kr. er indtil nu blevet udbetalt til danskere, der var udsat for stormen Bodil.

I alt har der været 2.973 sager til Stormrådet efter Bodil og på årsdagen mangler man at afslutte omkring 90 af dem. (Foto: Katrine Emilie Andersen © Scanpix)

https://www.dr.dk/nyheder/regionale/sjaelland/stormfloden-efter-bodil-sender-stadig-regninger-til-stormraadet
http://www.bt.dk/content/item/477962

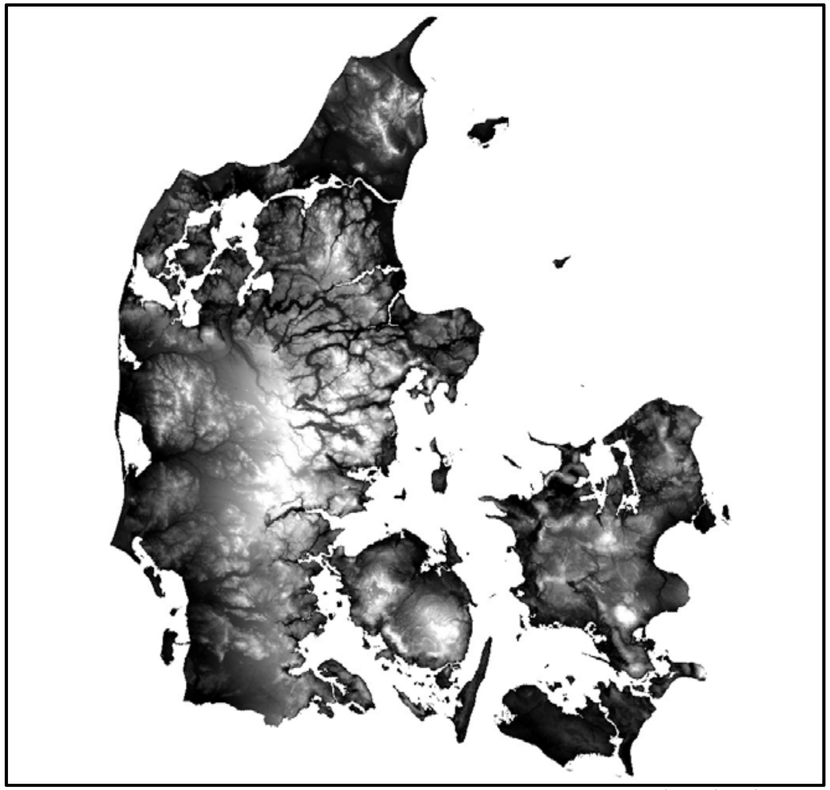
Flood Risk Computation

Sea-level forecast



The Danish Meteorological Institute

Terrain model



The Danish Geodata Agency

Sea- level forecasts of next 5 days
Updated every 6 hours

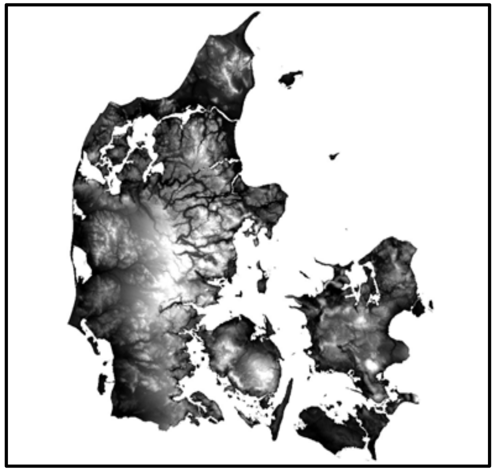
Mesured in every 0.4 meter

Issues

1. Handle massive terrain data
   - Reading and writing take more than 4 hours
2. Sea level is not the same everywhere
   - Existing algorithm is designed for uniform sea-level rise
3. Different resolutions
   - Terrain and sea-level data are measured by using different methods

Issues

1. Handle massive terrain data *
   - Reading and writing take more than 4 hours
2. Sea level is not the same everywhere
   - Existing algorithm is designed for uniform sea-level rise
3. Different resolutions
   - Terrain and sea-level data are measured by using different methods



Input terrain

Almost 1 trillion cells! (420 GB)

Cannot lower the resolution

Cannot use parallelization

Output

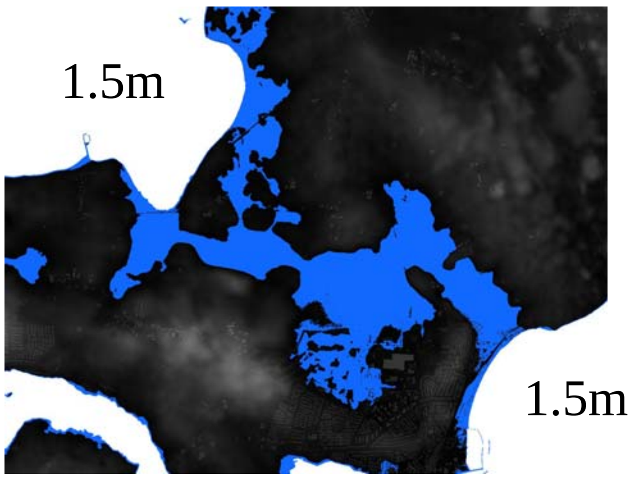
Write flooded water on each cell

As large as the input terrain

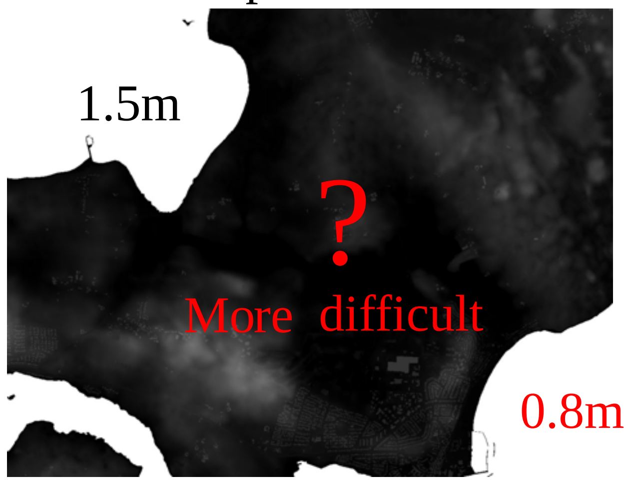
Issues

1. Handle massive terrain data
   - Reading and writing take more than 4 hours
2. Sea level is not the same everywhere *
   - Existing algorithm is designed for uniform sea-level rise
3. Different resolutions
   - Terrain and sea-level data are measured by using different methods

Existing algorithm*



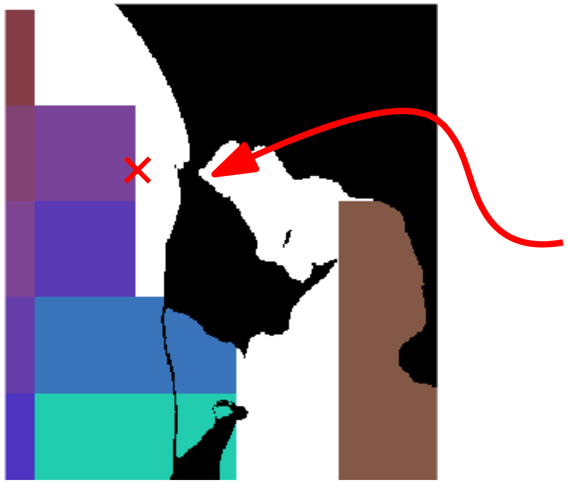
In practice



*A. Danner et. al., TerraStream: from Elevation Data to Watershed Hierarchies. 2007.

Issues

1. Handle massive terrain data
   - Reading and writing take more than 4 hours
2. Sea level is not the same everywhere
   - Existing algorithm is designed for uniform sea-level rise
3. Different resolutions
   - Terrain and sea-level data are measured by using different methods



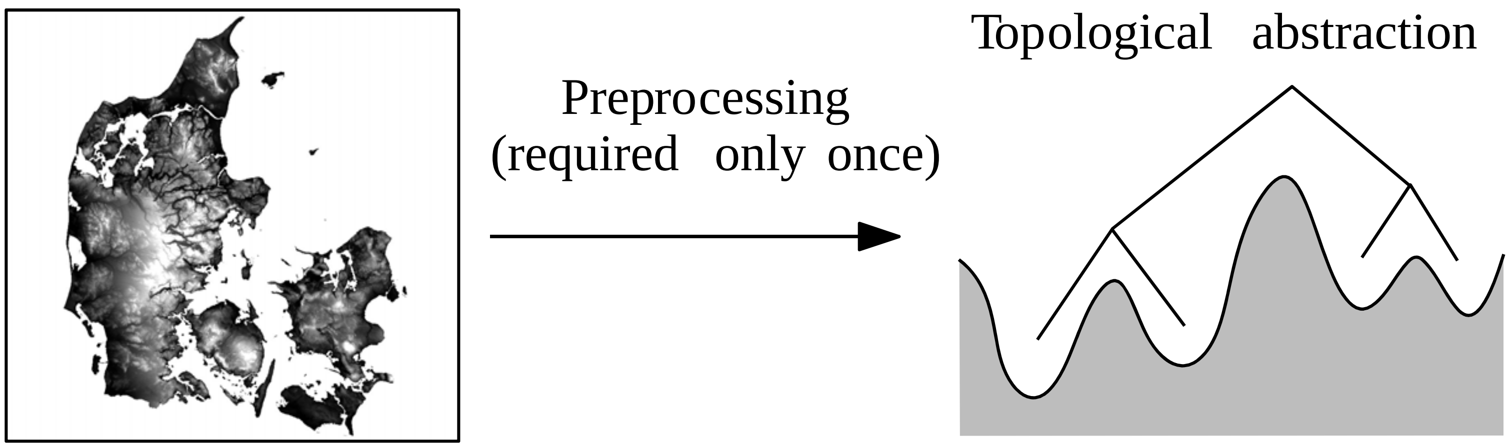
Cannot take the nearest cell
(sea water does not cross the terrain)

Solutions

1. Handle massive terrain data
   - Reading and writing take more than 4 hours
2. Sea level is not the same everywhere
   - Existing algorithm is designed for uniform sea-level rise
3. Different resolutions
   - Terrain and sea-level data are measured by using different methods

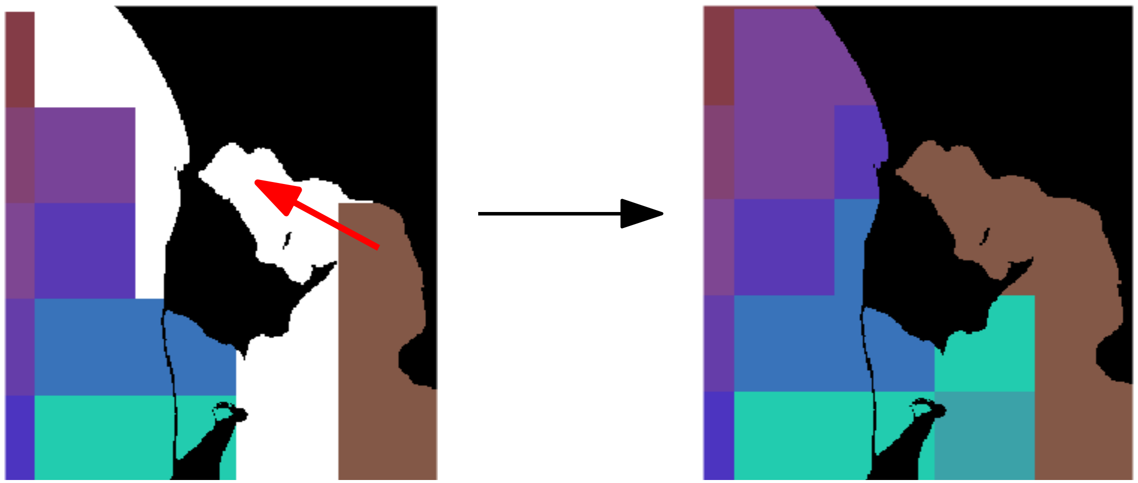
Solutions

1. Handle massive terrain data *
   - Reading and writing take more than 4 hours
2. Sea level is not the same everywhere *
   - Existing algorithm is designed for uniform sea-level rise
3. Different resolutions
   - Terrain and sea-level data are measured by using different methods



Solutions

1. Handle massive terrain data
   - Reading and writing take more than 4 hours
2. Sea level is not the same everywhere
   - Existing algorithm is designed for uniform sea-level rise
3. Different resolutions *
   - Terrain and sea-level data are measured by using different methods



Future Work

- Avoid writing all the result
  - Use more compact representation
- Validation
  - Compare with real-world event
- Integrate with DMI
  - Real world application

Future Work

- Avoid writing all the result
  - Use more compact representation
- Validation
  - Compare with real-world event
- Integrate with DMI
  - Real world application

Thank you.

Svend Christian Svendsen

Effective (semi-automatic) identification of hydrological corrections

DABAI

Who am I?

- PhD Student at DABAI since 2016
- Supervised by Prof. Lars Arge
- Research interest in I/O efficient Algorithms

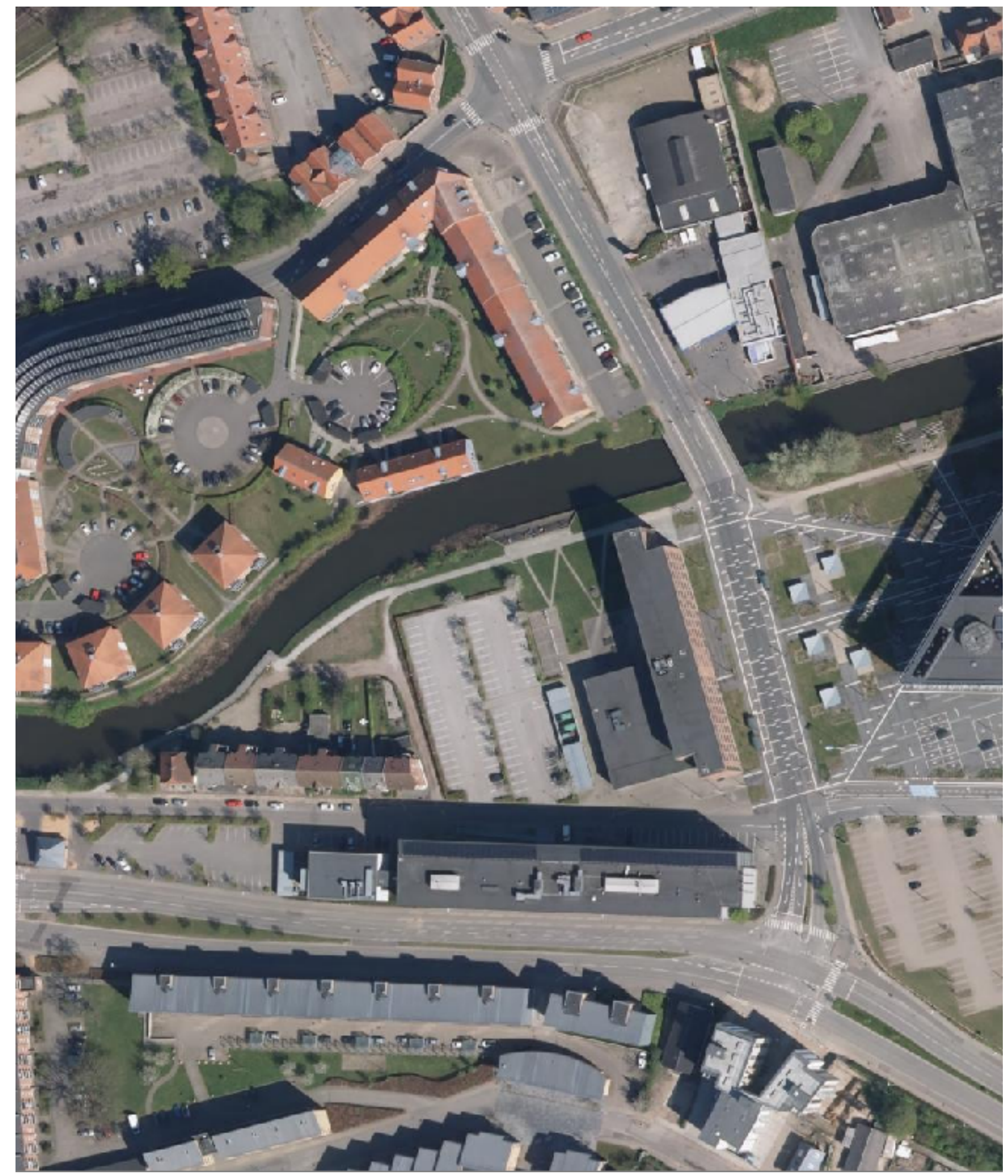
DABAI

Correction Identification

- Condition terrain data
  - Removal of bridges
  - Inclusion of culverts

Current Solutions

- Traditionally done manually and with local input
- Expensive and time consuming
- Error prone

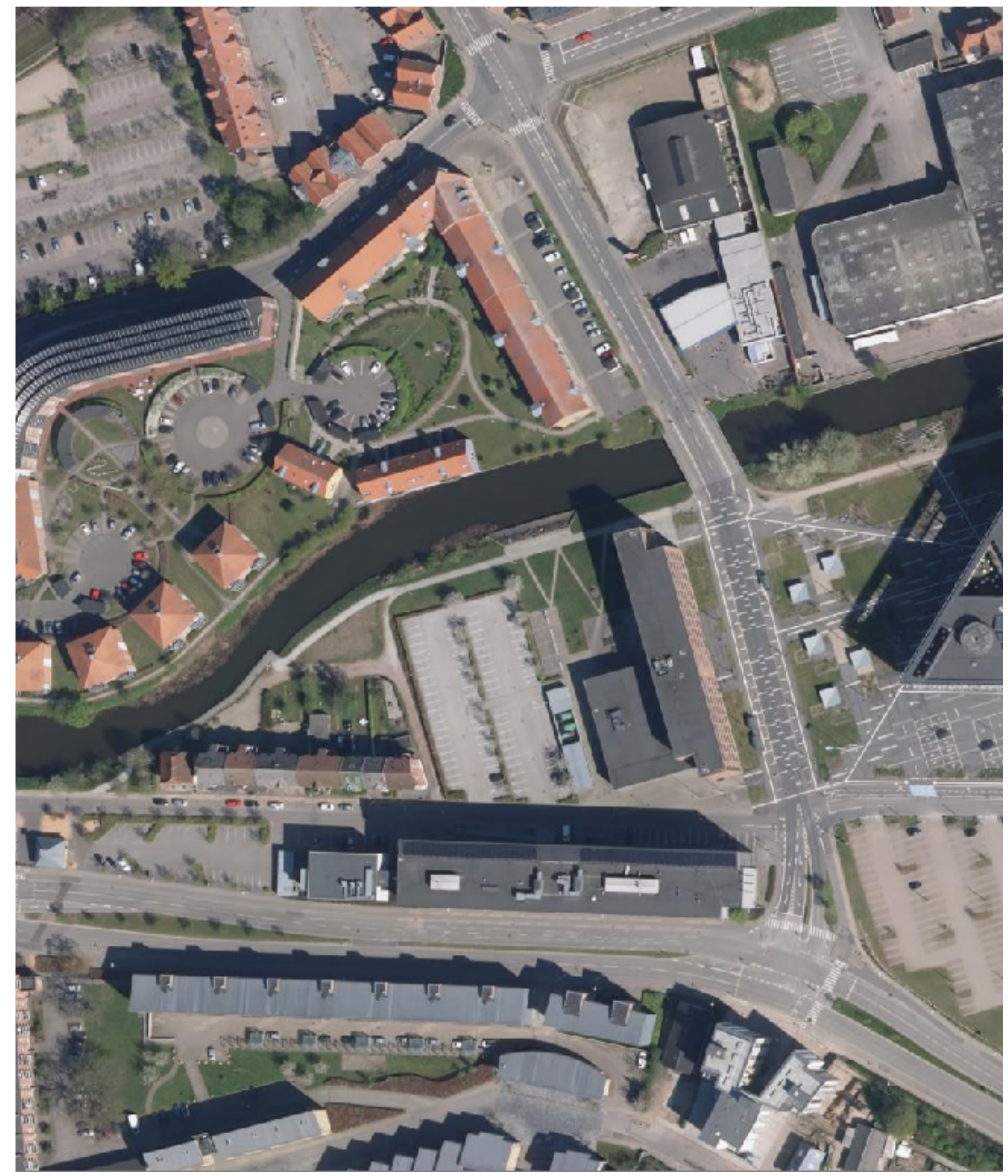


Current Solutions

- Use road and river data to burn river lines into data

- Alignment issues

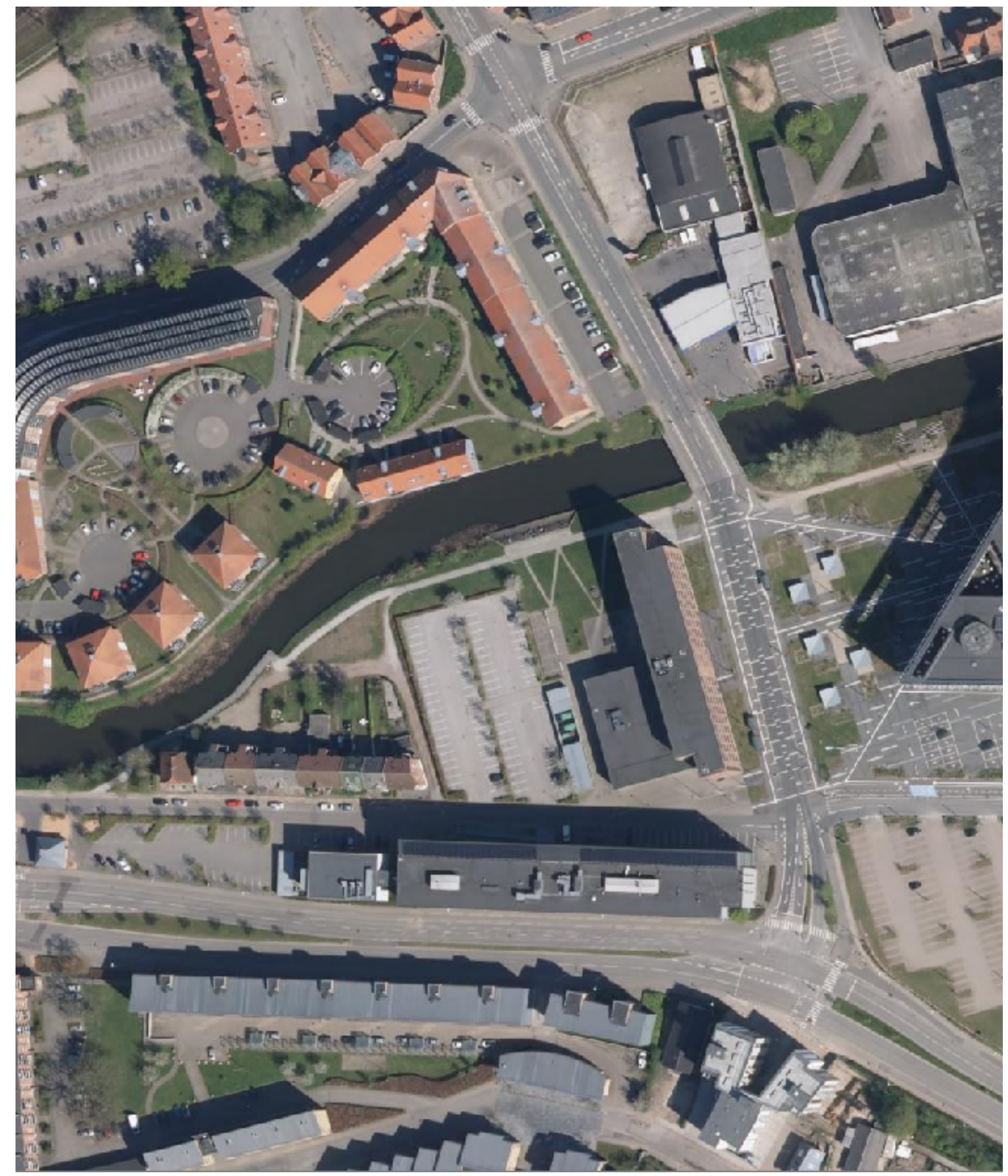
- Missing small streams and drainage pipes



DABAI

Feature Extraction

- Terrain model
- Orthophotos
- Flow Accumulation
- Flash-Flood Mapping



DABAI

Training

- Select techniques with success in similar areas
- Apply machine learning algorithms to detect hydrological corrections
- Generalise to detection of hydrological corrections on new data

DABAI

Big Data Challenges in the City of Copenhagen

Casper Hansen

Department of Computer Science
University of Copenhagen

DTU, Mar 29, 2017

Planned projects

Ongoing

- Predicting future service need for citizens receiving home care (danish: hjemmehjælp)

Future

- (Social) Case worker assistance
- Social fraud detection
- Early intervention in the child and youth area

Predicting future service needs in home care

Data

- Daily log of received services (medicine help, laundry, cooking, rehabilitation, etc.)
- Journal data on each citizen ("Michael fell down the stairs and is feeling unwell")
- Hospitalization info (duration and admitted hospital units)

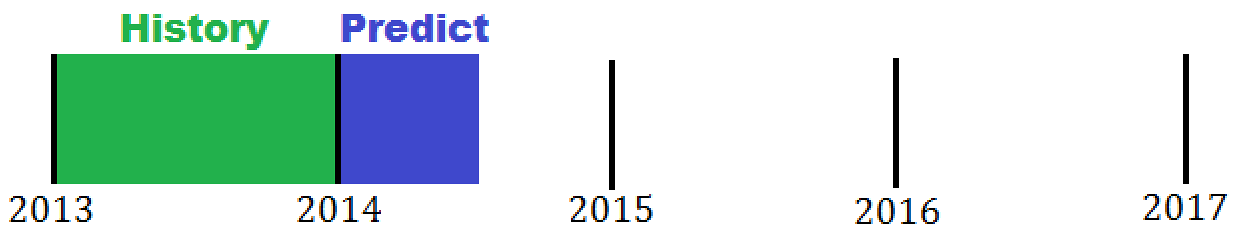
Data processing

- Aggregate historical past using one-hot encoding for categorical variables
- NLP for journal data (sentiment + top N keywords)

Predicting future service needs in home care

Initial basic approach

- Predict if a citizen will need an increased/not increased number of *<u>hours of help</u>*
- Predict if a citizen has a high risk of being *<u>hospitalized</u>* (or re-hospitalized)
- Time series prediction. Using X months historical data, predict the target variable in the following Y months?



Initial basic results with a random forest:

- Using just the daily log data with a binary variable of increased/not increased number of hours, yields an average 77% accuracy on both labels.

Similarity among quizzes

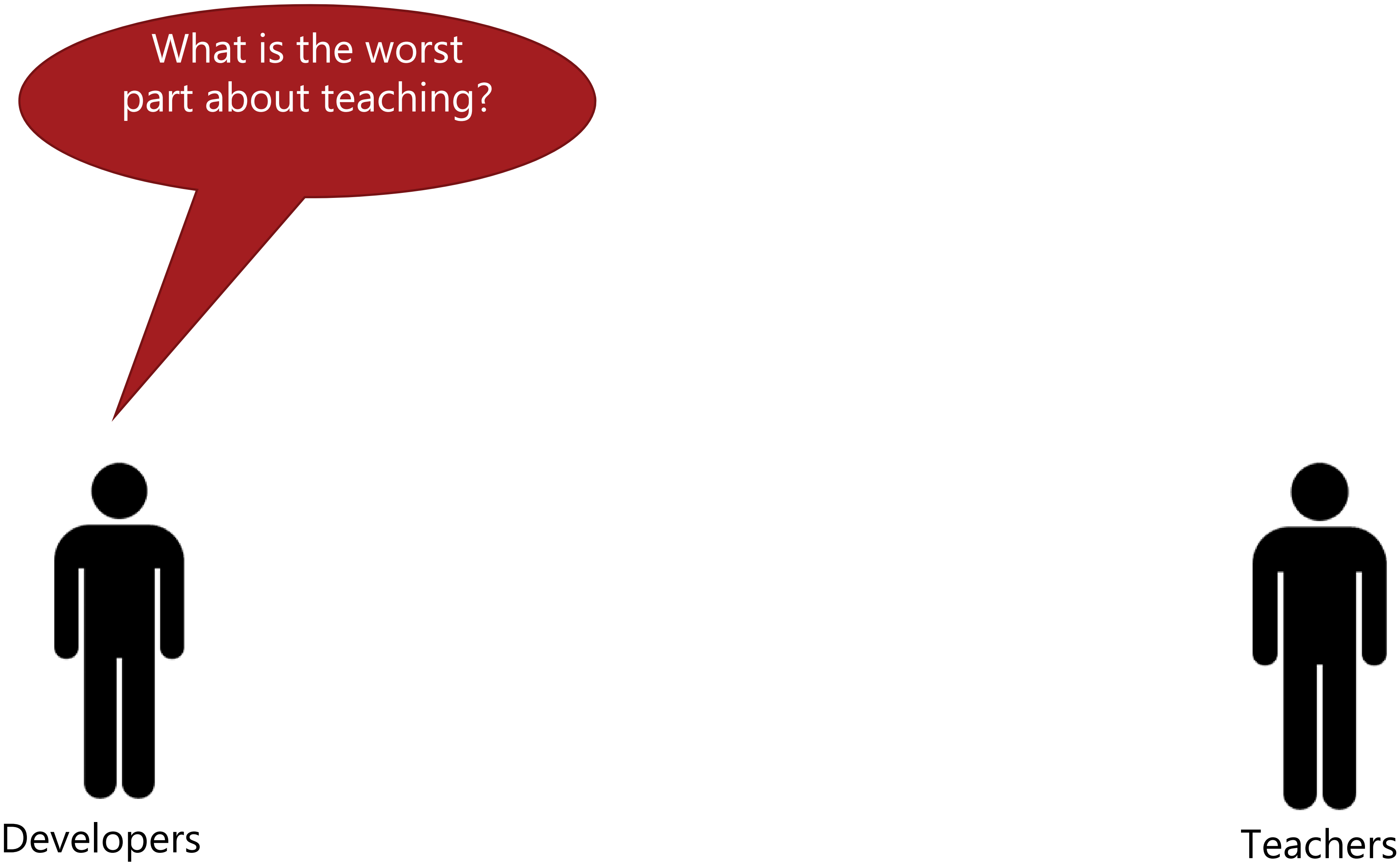
DABAI

DABAI education

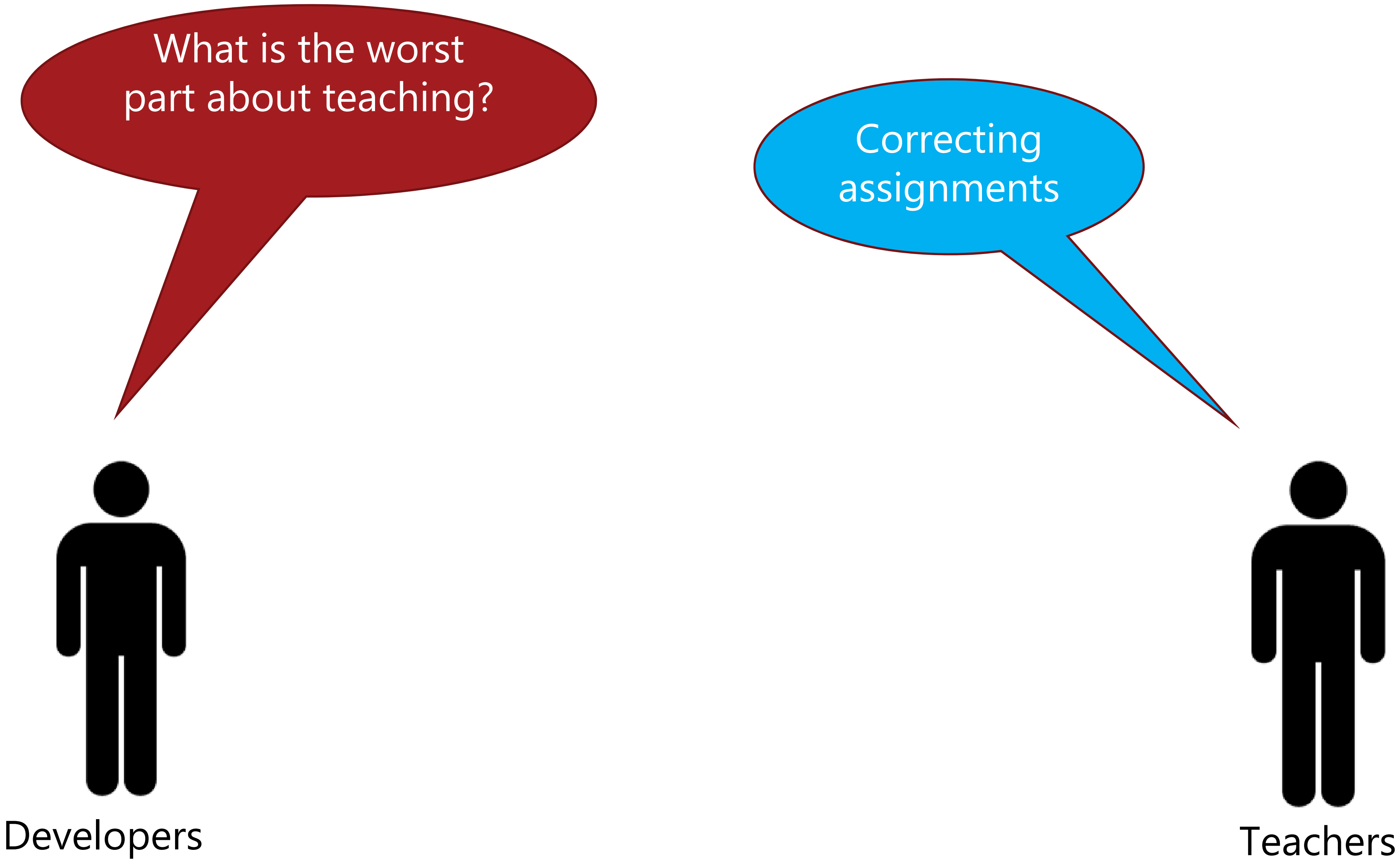
Niklas Hjuler
PhD student
DIKU
DTU 29-03-2017

UNIVERSITY OF COPENHAGEN

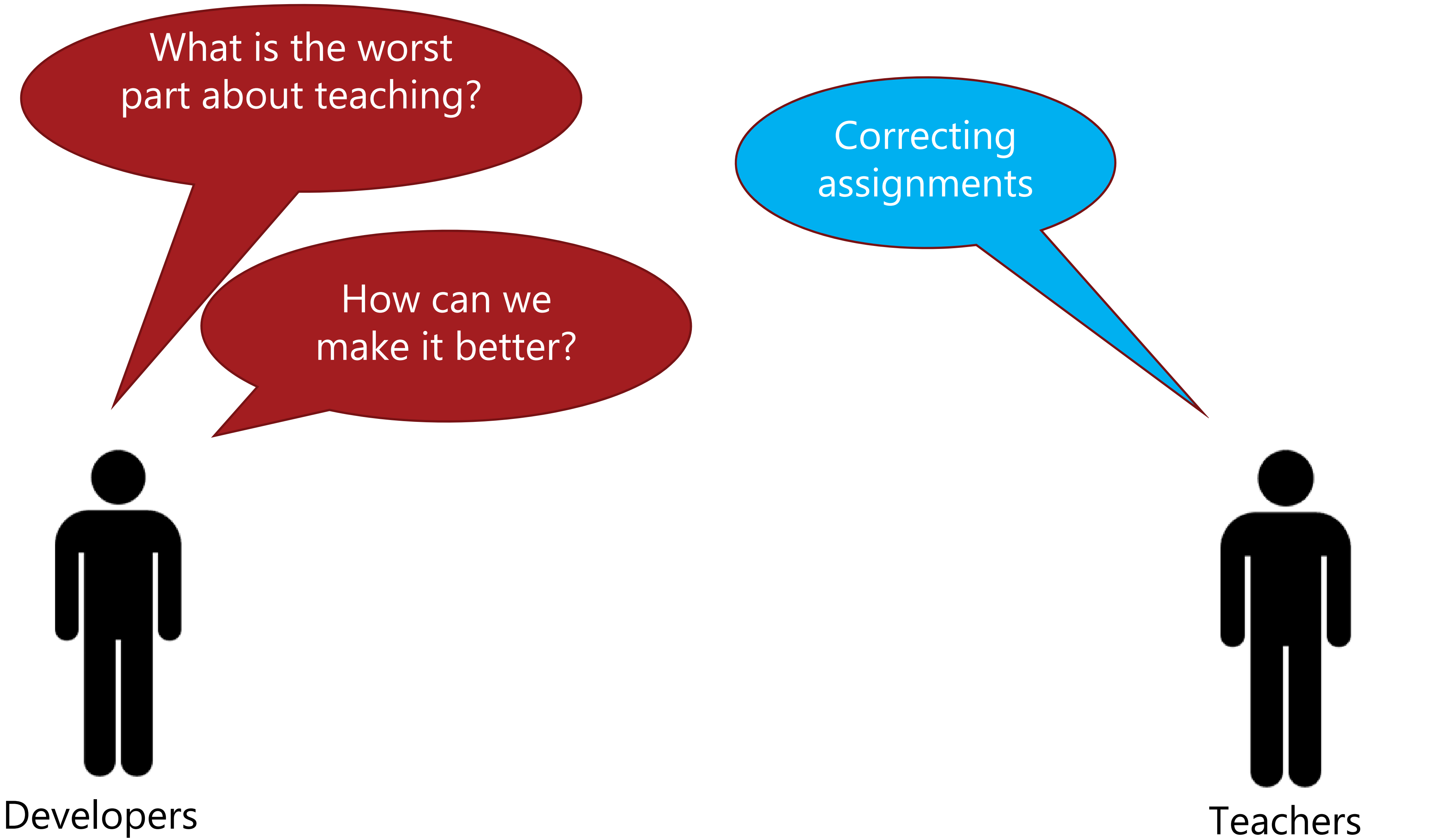
Introduction



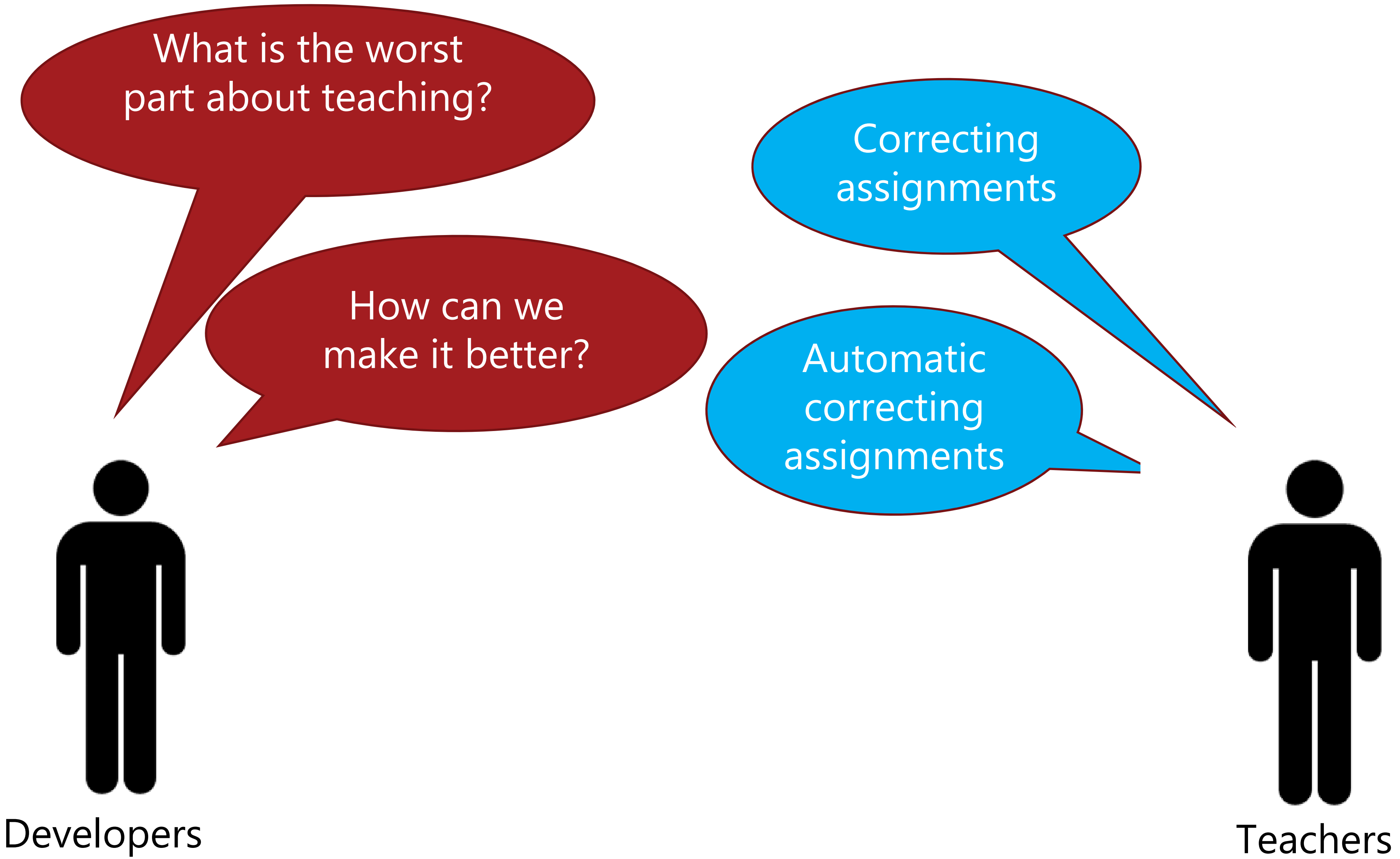
Introduction



Introduction



Introduction



Motivation

Motivation

- Teacher don’t want to spend too much time on correcting assignments.

Motivation

- Teacher don't want to spend too much time on correcting assignments.

- But some information is "lost" when the teacher is no longer correcting the assignments

Motivation

- To solve the loss of information developers show teacher some statistics about his students

Grade Items Print

Grade Item	Points	Weight Achieved	Grade
Exam 1	70 / 100	17.5 / 25	70 %
Exam 2	85 / 100	21.25 / 25	85 %
Exam 3	90 / 100	22.5 / 25	90 %
Quizzes		24 / 25	
Quiz 1	50 / 50	5 / 5	100 %
Quiz 2	50 / 50	5 / 5	100 %
Quiz 3	45 / 50	4.5 / 5	90 %
Quiz 4	45 / 50	4.5 / 5	90 %
Quiz 5	50 / 50	5 / 5	100 %

Motivation

- To solve the loss of information developers show teacher some statistics about his students
- This statistics is only based on the students own data

Grade Items Print

Grade Item	Points	Weight Achieved	Grade
Exam 1	70 / 100	17.5 / 25	70 %
Exam 2	85 / 100	21.25 / 25	85 %
Exam 3	90 / 100	22.5 / 25	90 %
Quizzes		24 / 25	
Quiz 1	50 / 50	5 / 5	100 %
Quiz 2	50 / 50	5 / 5	100 %
Quiz 3	45 / 50	4.5 / 5	90 %
Quiz 4	45 / 50	4.5 / 5	90 %
Quiz 5	50 / 50	5 / 5	100 %

DABAI

Motivation

- To solve the loss of information developers show teacher some statistics about his students
- This statistics is only based on the students own data
- And not the data of all students (i.e. it makes no prediction)

Grade Items Print

Grade Item	Points	Weight Achieved	Grade
Exam 1	70 / 100	17.5 / 25	70 %
Exam 2	85 / 100	21.25 / 25	85 %
Exam 3	90 / 100	22.5 / 25	90 %
Quizzes		24 / 25	
Quiz 1	50 / 50	5 / 5	100 %
Quiz 2	50 / 50	5 / 5	100 %
Quiz 3	45 / 50	4.5 / 5	90 %
Quiz 4	45 / 50	4.5 / 5	90 %
Quiz 5	50 / 50	5 / 5	100 %

Description - Similarity among quizzes

DABAI

- When all the data is gathered in one place, we can use the data of other students as well.

Description - Similarity among quizzes

- When all the data is gathered in one place, we can use the data of other students as well.
- Which means we calculate how similar different quizzes are. By similar we mean the same students are good/bad.

Description - Similarity among quizzes

- When all the data is gathered in one place, we can use the data of other students as well.
- Which means we calculate how similar different quizzes are. By similar we mean the same students are good/bad.

1. Find similar/dissimilar quizzes

Description - Similarity among quizzes

DABAI

- When all the data is gathered in one place, we can use the data of other students as well.
- Which means we calculate how similar different quizzes are. By similar we mean the same students are good/bad.

1. Find similar/dissimilar quizzes
2. Finding the error source(s) of a quiz

Description - Similarity among quizzes

DABAI

- When all the data is gathered in one place, we can use the data of other students as well.
- Which means we calculate how similar different quizzes are. By similar we mean the same students are good/bad.

1. Find similar/dissimilar quizzes
2. Finding the error source(s) of a quiz
3. Predict what the student have troubles with

Description - Similarity among quizzes

- When all the data is gathered in one place, we can use the data of other students as well.
- Which means we calculate how similar different quizzes are. By similar we mean the same students are good/bad.

1. Find similar/dissimilar quizzes
2. Finding the error source(s) of a quiz
3. Predict what the student have troubles with
4. Cover the curriculum in a few number of quizzes

Pearson correlation as similarity

- Pearson correlation between quizzes: $$\frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

Pearson correlation as similarity

- Pearson correlation between quizzes: $$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

```
Retskrivningsprøve - Fra Andrea til Medina   mest typiske fejltype er   Nutids-r 4.1     med score 0.63121
Retskrivningsprøve - Musik   mest typiske fejltype er   -ene eller -ende 8.1     med score 0.54787
Retskrivningsprøve - Børne-tv    mest typiske fejltype er   -ene eller -ende 8.1     med score 0.645663
Retskrivningsprøve - Menneskerobotten    mest typiske fejltype er   Ordforveksling nutid/datid 6.3   med score 0.799154
Retskrivningsprøve - Madvarer    mest typiske fejltype er   Sammensatte ord 7.3  med score 0.762077
Retskrivningsprøve - Fodbold     mest typiske fejltype er   -ene eller -ende 8.1     med score 0.558926
Retskrivningsprøve - Retsmedicineren     mest typiske fejltype er   -ene eller -ende 7.1     med score 0.655344
Retskrivningsprøve - Mærkværdige oplevelser  mest typiske fejltype er   Udsagnsord 6.1   med score 0.616944
Retskrivningsprøve - Logget ud   mest typiske fejltype er   -ene eller -ende 7.1     med score 0.663364
Retskrivningsprøve - Talemåder   mest typiske fejltype er   Nutids-r 8.2     med score 0.688781
Retskrivningsprøve - Kabeltyve   mest typiske fejltype er   -ene eller -ende 8.1     med score 0.723588
Retskrivningsprøve - Hollywood   mest typiske fejltype er   Nutids-r 6.1     med score 0.624064
Retskrivningsprøve - Internettet     mest typiske fejltype er   -ene eller -ende 8.1     med score 0.694216
Retskrivningsprøve - Høj fart    mest typiske fejltype er   Nutids-r 6.1     med score 0.694645
Retskrivningsprøve - Som det burde være  mest typiske fejltype er   -ene eller -ende 5.1 med score 0.52998
Retskrivningsprøve - Det allersidste klip    mest typiske fejltype er   -ene eller -ende 8.1     med score 0.668234
Retskrivningsprøve - Den biologiske antropolog   mest typiske fejltype er   Udsagnsord - bøjningsformer 6.1  med score 0.738142
Retskrivningsprøve - Den syriske aktivist    mest typiske fejltype er   Konsonantfordobling 7.1  med score 0.564097
Retskrivningsprøve - Hotel 45    mest typiske fejltype er   Navneord i flertal 8.1   med score 0.694262
Retskrivningsprøve - Vandafvisende myg   mest typiske fejltype er   Konsonantfordobling 7.1  med score 0.7181
Retskrivningsprøve - Landevejsriddere    mest typiske fejltype er   -ene eller -ende 8.1     med score 0.621853
```

DABAI

Pearson correlation as similarity

- Pearson correlation between quizzes: $$\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$
- Initial conclusion: Nutids-r and ene/ende often has the highest correlation with writing tests.

```
Retskrivningsprøve - Fra Andrea til Medina   mest typiske fejltype er   Nutids-r 4.1     med score 0.63121
Retskrivningsprøve - Musik   mest typiske fejltype er   -ene eller -ende 8.1     med score 0.54787
Retskrivningsprøve - Børne-tv    mest typiske fejltype er   -ene eller -ende 8.1     med score 0.645663
Retskrivningsprøve - Menneskerobotten    mest typiske fejltype er   Ordforveksling nutid/datid 6.3   med score 0.799154
Retskrivningsprøve - Madvarer    mest typiske fejltype er   Sammensatte ord 7.3  med score 0.762077
Retskrivningsprøve - Fodbold     mest typiske fejltype er   -ene eller -ende 8.1     med score 0.558926
Retskrivningsprøve - Retsmedicineren     mest typiske fejltype er   -ene eller -ende 7.1     med score 0.655344
Retskrivningsprøve - Mærkværdige oplevelser  mest typiske fejltype er   Udsagnsord 6.1   med score 0.616944
Retskrivningsprøve - Logget ud   mest typiske fejltype er   -ene eller -ende 7.1     med score 0.663364
Retskrivningsprøve - Talemåder   mest typiske fejltype er   Nutids-r 8.2     med score 0.688781
Retskrivningsprøve - Kabeltyve   mest typiske fejltype er   -ene eller -ende 8.1     med score 0.723588
Retskrivningsprøve - Hollywood   mest typiske fejltype er   Nutids-r 6.1     med score 0.624064
Retskrivningsprøve - Internettet     mest typiske fejltype er   -ene eller -ende 8.1     med score 0.694216
Retskrivningsprøve - Høj fart    mest typiske fejltype er   Nutids-r 6.1     med score 0.694645
Retskrivningsprøve - Som det burde være  mest typiske fejltype er   -ene eller -ende 5.1 med score 0.52998
Retskrivningsprøve - Det allersidste klip    mest typiske fejltype er   -ene eller -ende 8.1     med score 0.668234
Retskrivningsprøve - Den biologiske antropolog   mest typiske fejltype er   Udsagnsord - bøjningsformer 6.1  med score 0.738142
Retskrivningsprøve - Den syriske aktivist    mest typiske fejltype er   Konsonantfordobling 7.1  med score 0.564097
Retskrivningsprøve - Hotel 45    mest typiske fejltype er   Navneord i flertal 8.1   med score 0.694262
Retskrivningsprøve - Vandafvisende myg   mest typiske fejltype er   Konsonantfordobling 7.1  med score 0.7181
Retskrivningsprøve - Landevejsriddere    mest typiske fejltype er   -ene eller -ende 8.1     med score 0.621853
```

Thank you

DABAI

- Identify error source (Done)
- Find similar/dissimilar quiz (Done)
- Predicting where the student have problems. (Future work)
- Cover the curriculum as good as possible in a fixed number of quizzes (Future work)

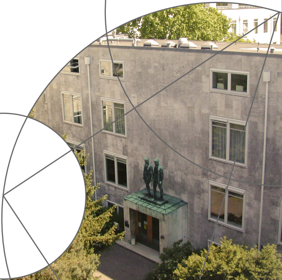
Faculty of Science

Detecting ghost-writing in high school assignments

Stephan Sloth Lorenzen

ph.d. student
Department of Computer Science, University of Copenhagen

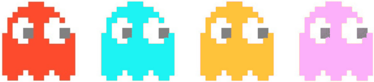
DABAI EDU



March 29th, 2017

Motivation

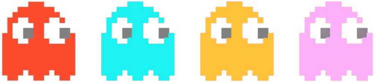
Large problem in academics (secondary education and universities):
Students hire teachers, professionals, etc. to write their assignments - this is known as *ghost-writing*.



[1] http://www.thebestschools.org/resources/ghostwriting-business-trade-standards-practices-secrets/

Motivation

Large problem in academics (secondary education and universities):
Students hire teachers, professionals, etc. to write their assignments - this is known as *ghost-writing*.

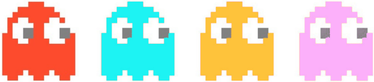


In universities in the U.S., 7% of students admit to cheating by handing in assignments written by others *[McCabe'05]*.

[1] http://www.thebestschools.org/resources/ghostwriting-business-trade-standards-practices-secrets/

Motivation

Large problem in academics (secondary education and universities):
Students hire teachers, professionals, etc. to write their assignments - this is known as *ghost-writing*.



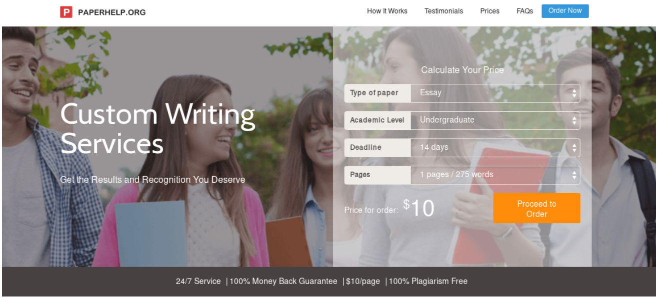
In universities in the U.S., 7% of students admit to cheating by handing in assignments written by others *[McCabe'05]*.

Ghost-writing has become an industry: more than 300 online services[1] providing ghost-writing for payment exist.

[1] http://www.thebestschools.org/resources/ghostwriting-business-trade-standards-practices-secrets/

Motivation


AFFORDABLE PAPERS
THE RIGHT WAY TO THE TOP
GET A FREE INQUIRY
PRICES
OUR WRITERS
TESTIMONIALS
FAQ
MANAGE YOUR ORDERS
GOING THE RIGHT WAY
FROM $ 9 page
Get your high-quality paper at affordable prices with the deadline you need
"PLAGIARISM-FREE" GUARANTEE
MONEYBACK GUARANTEE
ORDER A PAPER


PAPERHELP.ORG
How It Works
Testimonials
Prices
FAQs
Order Now
Custom Writing Services
Get the Results and Recognition You Deserve
Calculate Your Price
Type of paper
Essay
Academic Level
Undergraduate
Deadline
14 days
Pages
1 pages / 275 words
Price for order: $10
Proceed to Order
24/7 Service | 100% Money Back Guarantee | $10/page | 100% Plagiarism Free

Motivation: Danish Perspective

In Denmark, the problem has recently received more attention, as high school students hire ghost-writers to write their SRPs (large written third-year assignment).

Motivation: Danish Perspective

In Denmark, the problem has recently received more attention, as high school students hire ghost-writers to write their SRPs (large written third-year assignment).

Data available from MaCom:
MaCom is the company behind the learning platform **Lectio**:

Motivation: Danish Perspective

In Denmark, the problem has recently received more attention, as high school students hire ghost-writers to write their SRPs (large written third-year assignment).

Data available from MaCom:
MaCom is the company behind the learning platform **Lectio**:

- Lectio is used by 90% of Danish high schools.

Motivation: Danish Perspective

In Denmark, the problem has recently received more attention, as high school students hire ghost-writers to write their SRPs (large written third-year assignment).

Data available from MaCom:

MaCom is the company behind the learning platform **Lectio**:

- Lectio is used by 90% of Danish high schools.
- Covers more than 150,000 students.

Motivation: Danish Perspective

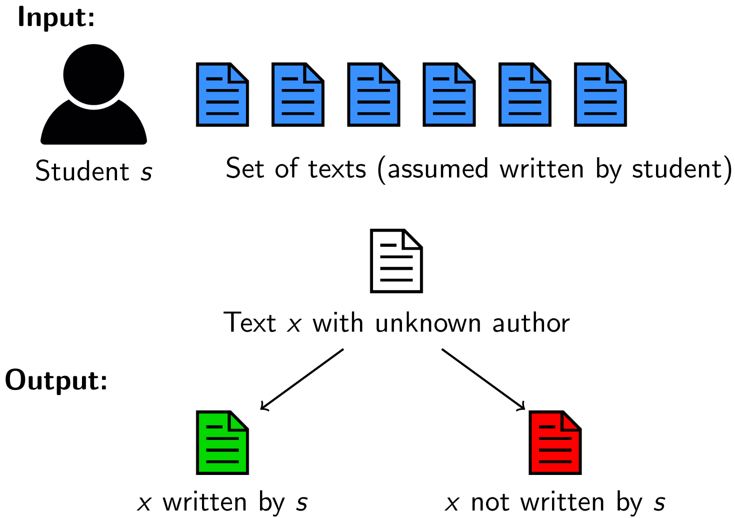
In Denmark, the problem has recently received more attention, as high school students hire ghost-writers to write their SRPs (large written third-year assignment).

Data available from MaCom:

MaCom is the company behind the learning platform **Lectio**:

- Lectio is used by 90% of Danish high schools.
- Covers more than 150,000 students.
- More than 15 million written assignments handed in.

The problem: Refuting authorship



Data

Data

Each assignment is given in plain text with a student id, subject. A feature vector is then constructed for each assignment.

Data

Each assignment is given in plain text with a student id, subject. A feature vector is then constructed for each assignment.

Textual features

- Average word length
- Average sentence length
- Ratio between number of commas and periods
- and more...

Data

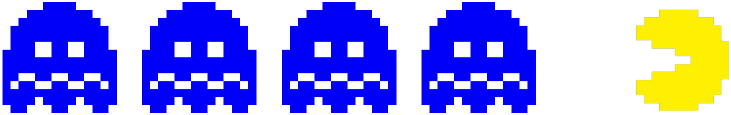
Each assignment is given in plain text with a student id, subject. A feature vector is then constructed for each assignment.

Textual features

- Average word length
- Average sentence length
- Ratio between number of commas and periods
- and more...

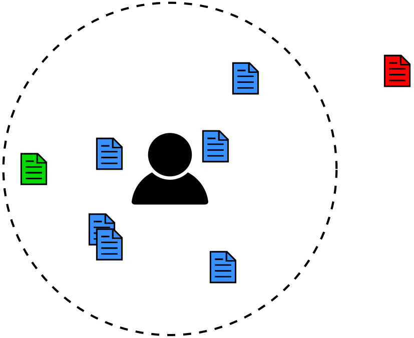
We assume that previous hand-ins are written by the given student.

Methods for detecting ghost-writers



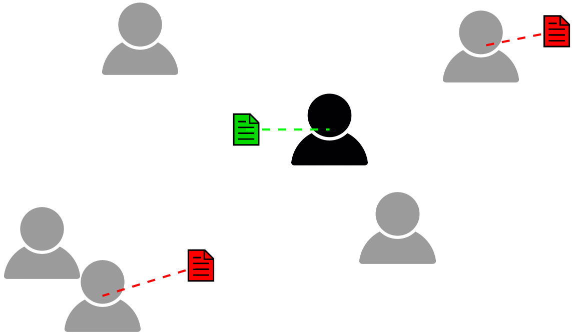
Method: distance based

Idea: construct profile vector for s based on earlier assignments (e.g. as the cluster center for assignments). Accept new assignment x if x is within an acceptable distance (compared to earlier assignments) from the profile, and reject x otherwise.



Method: distance based with imposters

Idea: construct profile vector for $s_0 = s$ and for several other students, $s_1, s_2, \ldots, s_m$. Accept new assignment x if x is closest to profile for s_0, and reject x otherwise.



Method: classification with imposters

Idea: train a classifier $\mathcal{C}$ based on the feature vectors for assignments handed in by $s_0 = s$ and several other students, $s_1, s_2, \ldots, s_m$. Accept new assignment x if $\mathcal{C}$ predicts that x is written by s_0, and reject x otherwise.

Method: classification with imposters

Idea: train a classifier $\mathcal{C}$ based on the feature vectors for assignments handed in by $s_0 = s$ and several other students, $s_1, s_2, \ldots, s_m$. Accept new assignment x if $\mathcal{C}$ predicts that x is written by s_0, and reject x otherwise.

Any multi-class classifier can be used. E.g.:

Method: classification with imposters

Idea: train a classifier $\mathcal{C}$ based on the feature vectors for assignments handed in by $s_0 = s$ and several other students, $s_1, s_2, \dots, s_m$. Accept new assignment x if $\mathcal{C}$ predicts that x is written by s_0, and reject x otherwise.

Any multi-class classifier can be used. E.g.:

- Support Vector Machines (SVM).

Method: classification with imposters

Idea: train a classifier $\mathcal{C}$ based on the feature vectors for assignments handed in by $s_0 = s$ and several other students, $s_1, s_2, ..., s_m$. Accept new assignment x if $\mathcal{C}$ predicts that x is written by s_0, and reject x otherwise.

Any multi-class classifier can be used. E.g.:

- Support Vector Machines (SVM).
- Random forests.

Method: classification with imposters

Idea: train a classifier $\mathcal{C}$ based on the feature vectors for assignments handed in by $s_0 = s$ and several other students, $s_1, s_2, ..., s_m$. Accept new assignment x if $\mathcal{C}$ predicts that x is written by s_0, and reject x otherwise.

Any multi-class classifier can be used. E.g.:

- Support Vector Machines (SVM).
- Random forests.

Experiments with MaCom data (using SVM) achieves accuracy of $\simeq 70\%$ *[Hansen, Lioma, Larsen, Alstrup 2014]*.

The goal

We wish to improve upon the previous results.

The goal

We wish to improve upon the previous results.
This is done by:

The goal

We wish to improve upon the previous results.

This is done by:

- Considering more textual features.

The goal

We wish to improve upon the previous results.

This is done by:

- Considering more textual features.
- Considering new methods/methods not used before in this domain (e.g. methods from *outlier detection*).

The goal

We wish to improve upon the previous results.

This is done by:

- Considering more textual features.
- Considering new methods/methods not used before in this domain (e.g. methods from *outlier detection*).

Future research: detecting student progress.

The writing style of a student may change over time; using the techniques discussed here, we can detect this change and thus the progress of the student.