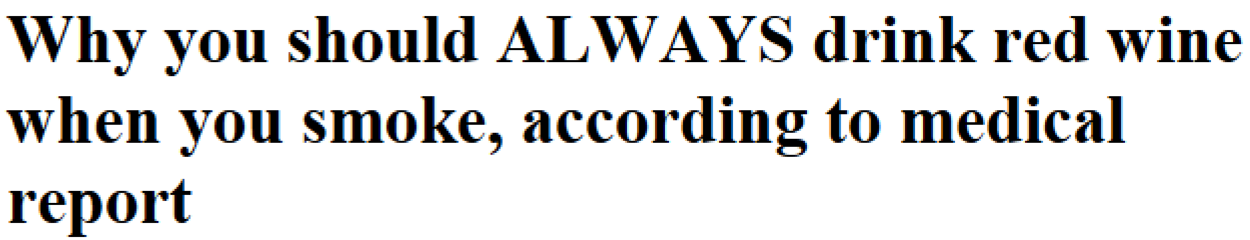
Webis Talks Flash 2017
12.01.2017

Clickbait Detection
GPU Based Text-Analytics
This is Offensive Language
Essay Scoring
Argument Unit Segmentation
Web Resource for Argument Mining
Vandalism in Wikipedia
Argumentation Strategies
2017 ACM DEBS Grand Challenge
A Large-scale Analysis of the Mnemonic Password Advice

Clickbait Detection

by Kristof Komlossy, Putrasandy Dana Riawan Sebastian Schuster



Why you should ALWAYS drink red wine when you smoke, according to medical report

This Teen Blew Everyone Out Of The Water With Her Insane
Prom Entrance

Definition

- A headline is considered as clickbait if its main purpose is to make the user click on the link by unnecessarily omitting crucial information, creating a curiosity gap.
- A clickbait is also present if there is an exaggerating headline which has nothing to do with the content.

Reasons

- Increase of clickbait over the last few years

- Not much existing reasearch (yet)

Problems

- Lowers the perception and standards of journalism
- Trains reader to skim over articles

Solution

- Journalists could abandon practice
- Readers could stop rewarding clickbait

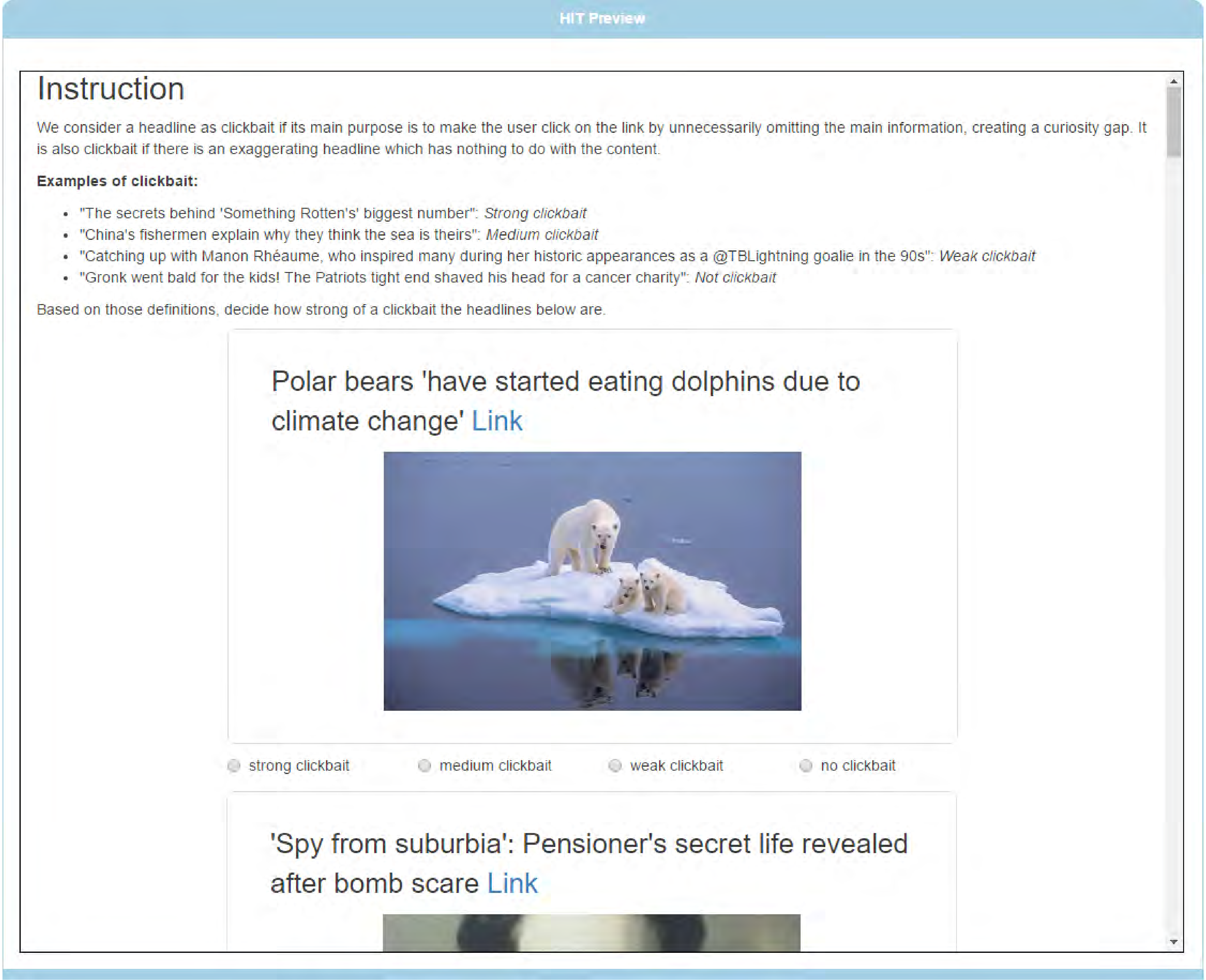
Solution

- ~~Journalists could abandon practice~~
- ~~Readers could stop rewarding clickbait~~
- give users a tool to detect clickbait content (maybe hide it)

Process

- Searched for concise **clickbait definition**
- **Archived tweets**[more than 250,000] and related **media** of news publishers

  → **Scored** after **degree of clickbait** by Amazon Mechanical Turk workers
- Use certain **features of the data** to train the classifiers
- Write browser plugin



Instruction

We consider a headline as clickbait if its main purpose is to make the user click on the link by unnecessarily omitting the main information, creating a curiosity gap. It is also clickbait if there is an exaggerating headline which has nothing to do with the content.

Examples of clickbait:

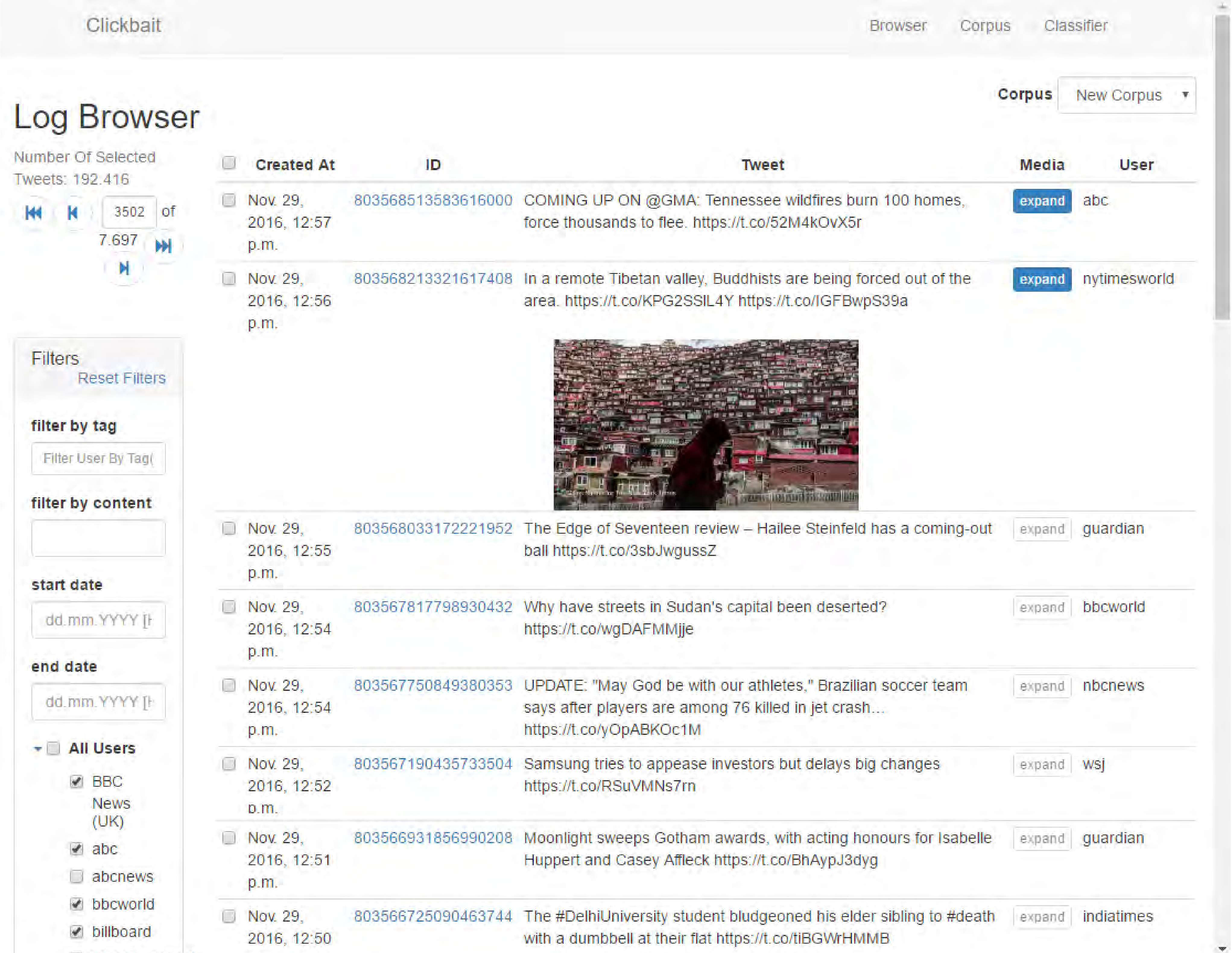
- "The secrets behind 'Something Rotten's' biggest number": *Strong clickbait*
- "China's fishermen explain why they think the sea is theirs": *Medium clickbait*
- "Catching up with Manon Rhéaume, who inspired many during her historic appearances as a @TBLightning goalie in the 90s": *Weak clickbait*
- "Gronk went bald for the kids! The Patriots tight end shaved his head for a cancer charity": *Not clickbait*

Based on those definitions, decide how strong of a clickbait the headlines below are.

Polar bears 'have started eating dolphins due to climate change' Link

○ strong clickbait ○ medium clickbait ○ weak clickbait ○ no clickbait

'Spy from suburbia': Pensioner's secret life revealed after bomb scare Link



Corpus: New Corpus ▾

Log Browser

Number Of Selected Tweets: 192.416

⏮ ◀ 3502 of 7.697 ⏭ ▶

Filters
Reset Filters

filter by tag: Filter User By Tag(

filter by content:

start date: dd.mm.YYYY [I

end date: dd.mm.YYYY [I

▾ □ All Users
☑ BBC News (UK)
☑ abc
□ abcnews
☑ bbcworld
☑ billboard

□	Created At	ID	Tweet	Media	User
□	Nov. 29, 2016, 12:57 p.m.	803568513583616000	COMING UP ON @GMA: Tennessee wildfires burn 100 homes, force thousands to flee. https://t.co/52M4kOvX5r	expand	abc
□	Nov. 29, 2016, 12:56 p.m.	803568213321617408	In a remote Tibetan valley, Buddhists are being forced out of the area. https://t.co/KPG2SSIL4Y https://t.co/IGFBwpS39a	expand	nytimesworld
□	Nov. 29, 2016, 12:55 p.m.	803568033172221952	The Edge of Seventeen review – Hailee Steinfeld has a coming-out ball https://t.co/3sbJwgussZ	expand	guardian
□	Nov. 29, 2016, 12:54 p.m.	803567817798930432	Why have streets in Sudan's capital been deserted? https://t.co/wgDAFMMjje	expand	bbcworld
□	Nov. 29, 2016, 12:54 p.m.	803567750849380353	UPDATE: "May God be with our athletes," Brazilian soccer team says after players are among 76 killed in jet crash… https://t.co/yOpABKOc1M	expand	nbcnews
□	Nov. 29, 2016, 12:52 p.m.	803567190435733504	Samsung tries to appease investors but delays big changes https://t.co/RSuVMNs7rn	expand	wsj
□	Nov. 29, 2016, 12:51 p.m.	803566931856990208	Moonlight sweeps Gotham awards, with acting honours for Isabelle Huppert and Casey Affleck https://t.co/BhAypJ3dyg	expand	guardian
□	Nov. 29, 2016, 12:50	803566725090463744	The #DelhiUniversity student bludgeoned his elder sibling to #death with a dumbbell at their flat https://t.co/tiBGWrHMMB	expand	indiatimes

Sources

http://www.dailymail.co.uk/health/article-3939616/Why-drink-red-wine-smoke-according-medical-report.html

https://www.buzzfeed.com/stephaniemcneal/people-are-losing-their-minds-over-this-teens-insanely-over

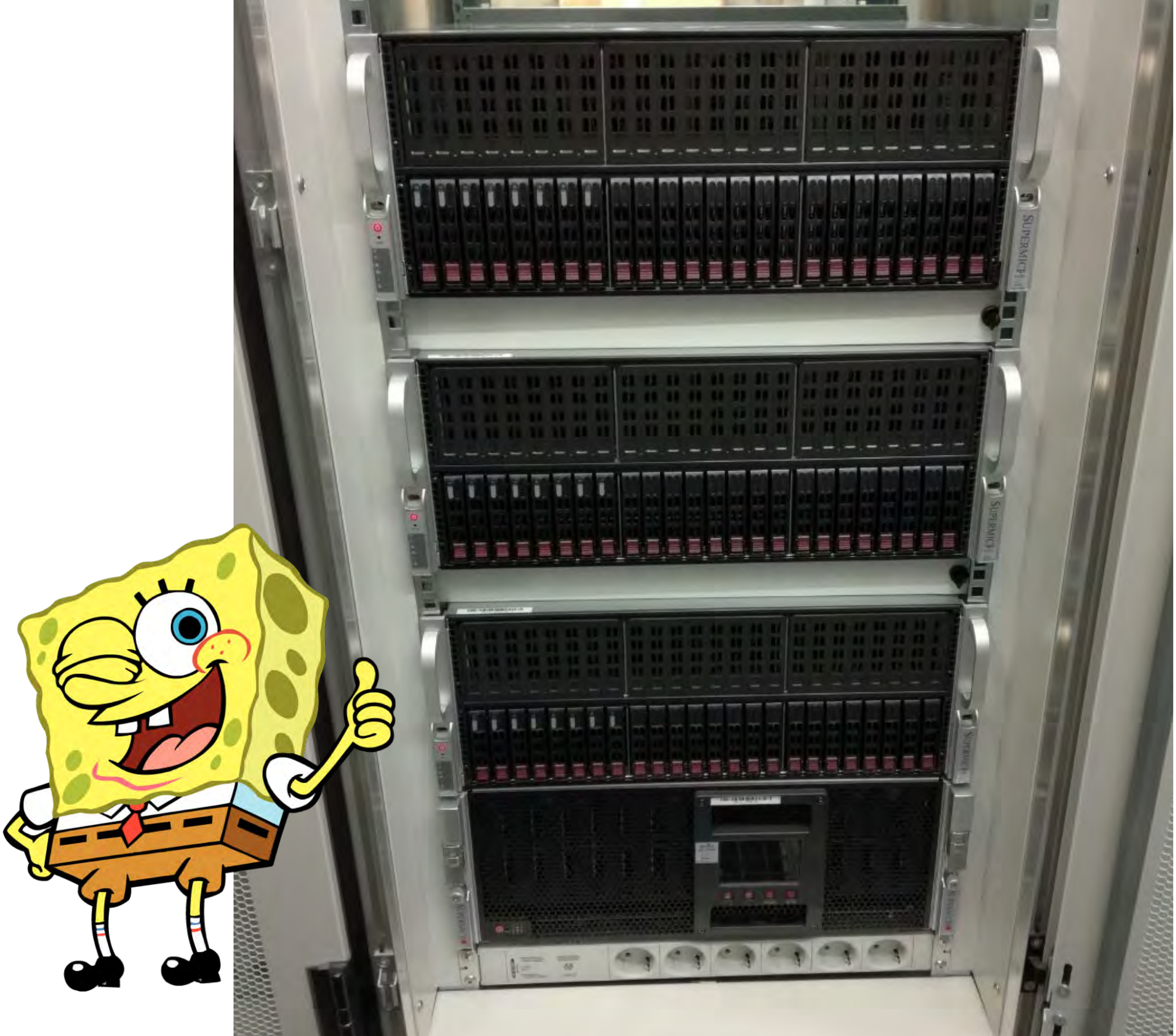
GPU Based Text-Analytics

André Karge
andre.karge@uni-weimar.de

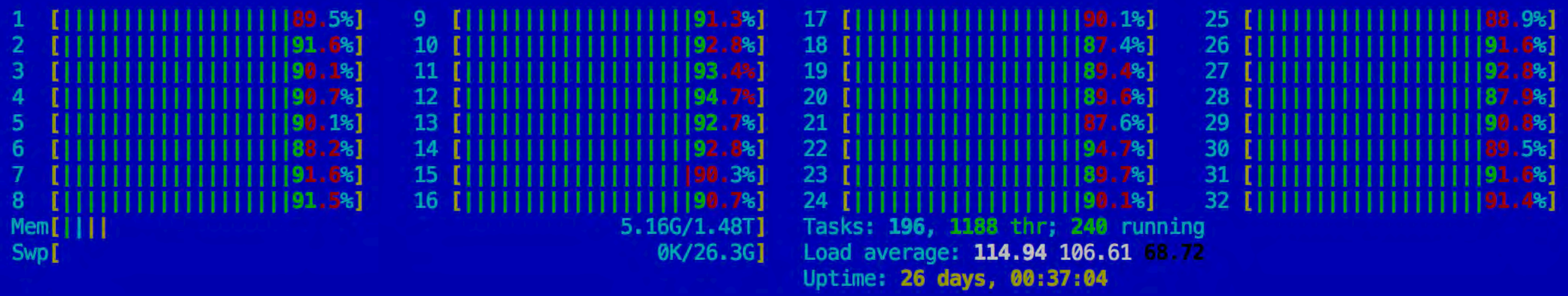
Benedikt S. Vogler
benedikt.vogler@uni-weimar.de

Setup

USV

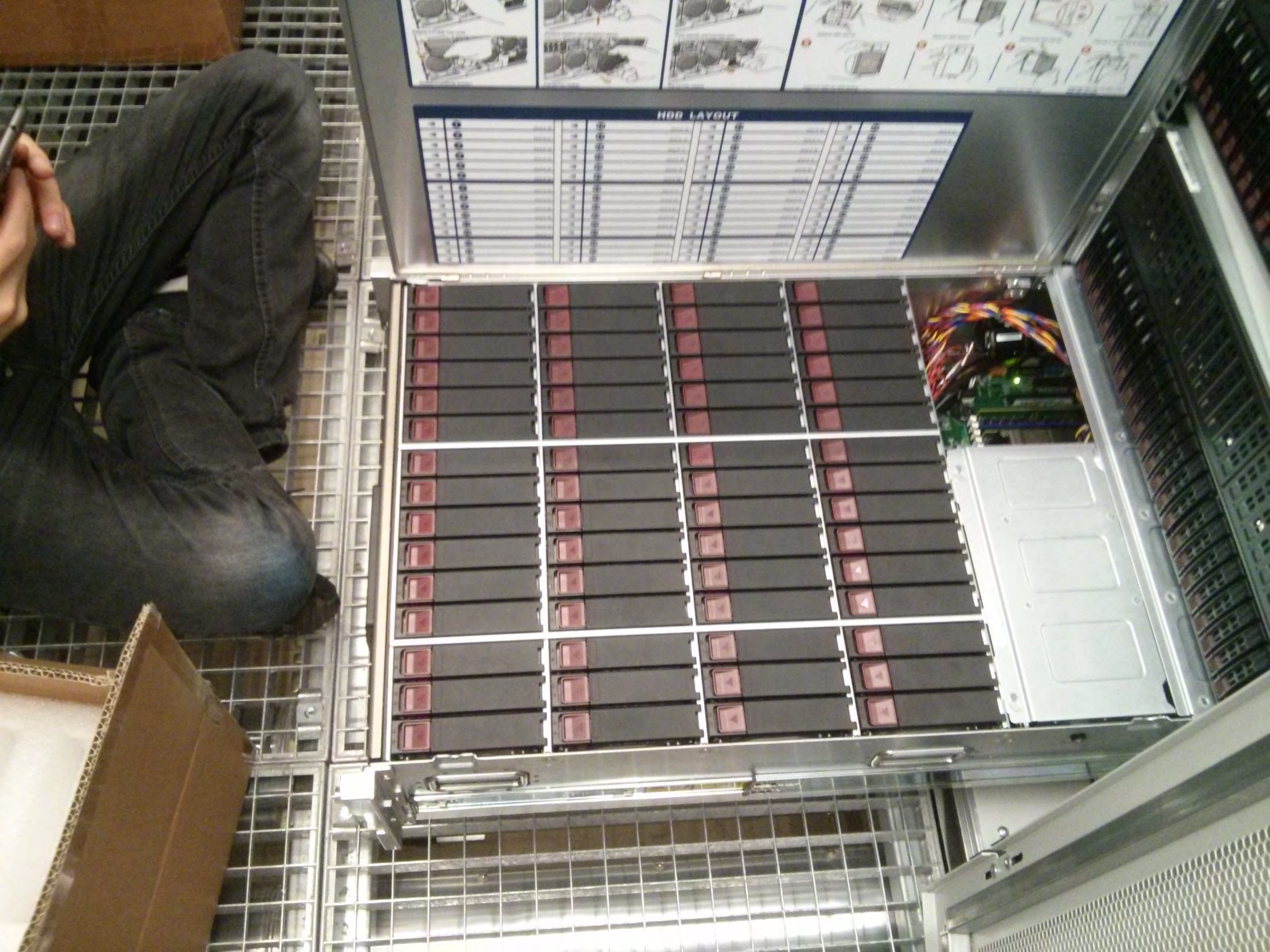

SUPERMICRO

Three times



```
1  [|||||||||||||||||||89.5%]   9  [|||||||||||||||||||91.3%]   17 [|||||||||||||||||||90.1%]   25 [|||||||||||||||||||88.9%]
2  [|||||||||||||||||||91.6%]   10 [|||||||||||||||||||92.8%]   18 [|||||||||||||||||||87.4%]   26 [|||||||||||||||||||91.6%]
3  [|||||||||||||||||||90.1%]   11 [|||||||||||||||||||93.4%]   19 [|||||||||||||||||||89.4%]   27 [|||||||||||||||||||92.8%]
4  [|||||||||||||||||||90.7%]   12 [|||||||||||||||||||94.7%]   20 [|||||||||||||||||||89.6%]   28 [|||||||||||||||||||87.9%]
5  [|||||||||||||||||||90.1%]   13 [|||||||||||||||||||92.7%]   21 [|||||||||||||||||||87.6%]   29 [|||||||||||||||||||90.8%]
6  [|||||||||||||||||||88.2%]   14 [|||||||||||||||||||92.8%]   22 [|||||||||||||||||||94.7%]   30 [|||||||||||||||||||89.5%]
7  [|||||||||||||||||||91.6%]   15 [|||||||||||||||||||90.3%]   23 [|||||||||||||||||||89.7%]   31 [|||||||||||||||||||91.6%]
8  [|||||||||||||||||||91.5%]   16 [|||||||||||||||||||90.7%]   24 [|||||||||||||||||||90.1%]   32 [|||||||||||||||||||91.4%]
Mem[||||                                         5.16G/1.48T]   Tasks: 196, 1188 thr; 240 running
Swp[                                                0K/26.3G]   Load average: 114.94 106.61 68.72
                                                                Uptime: 26 days, 00:37:04
```

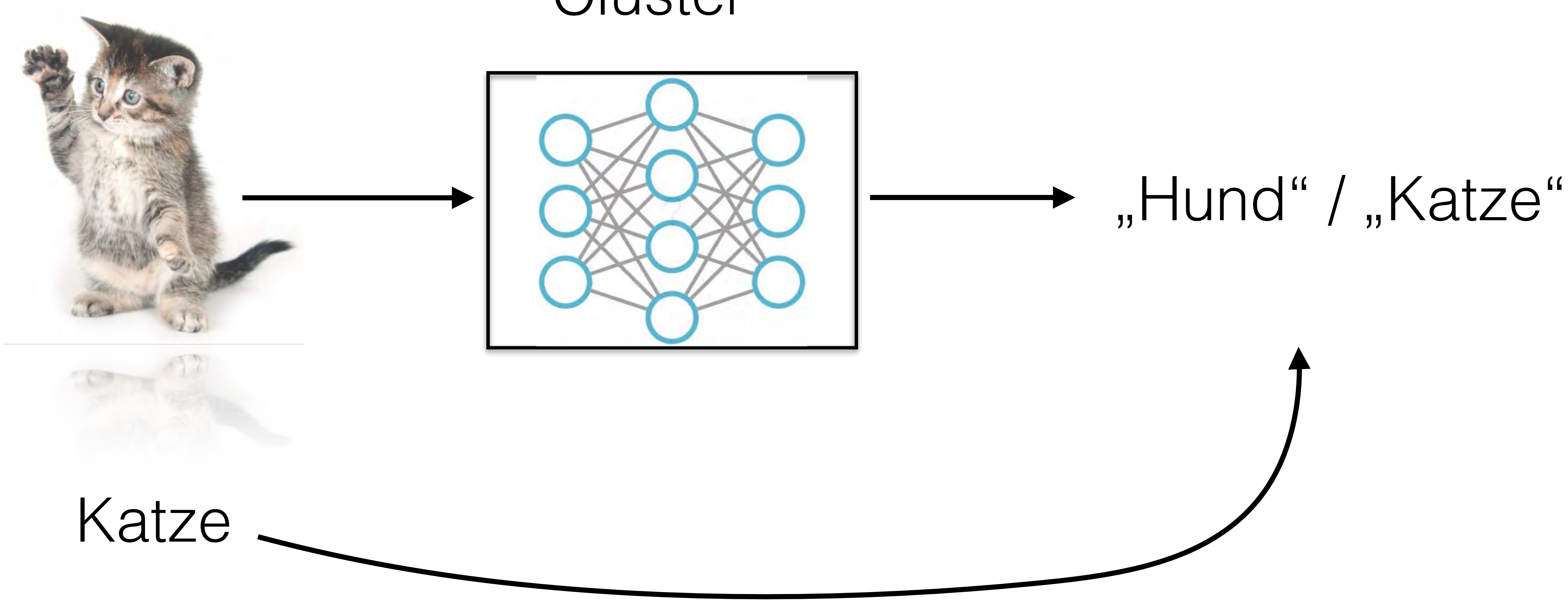
- 1,5 TB RAM
- 8x GeForce GTX 1080 (8113MiB)
- 2 Intel Xeon CPU:
- SSDs


HDD LAYOUT

Steps

- Install Software
- Health Checks
- Train own Deep Learning Networks
- Improve

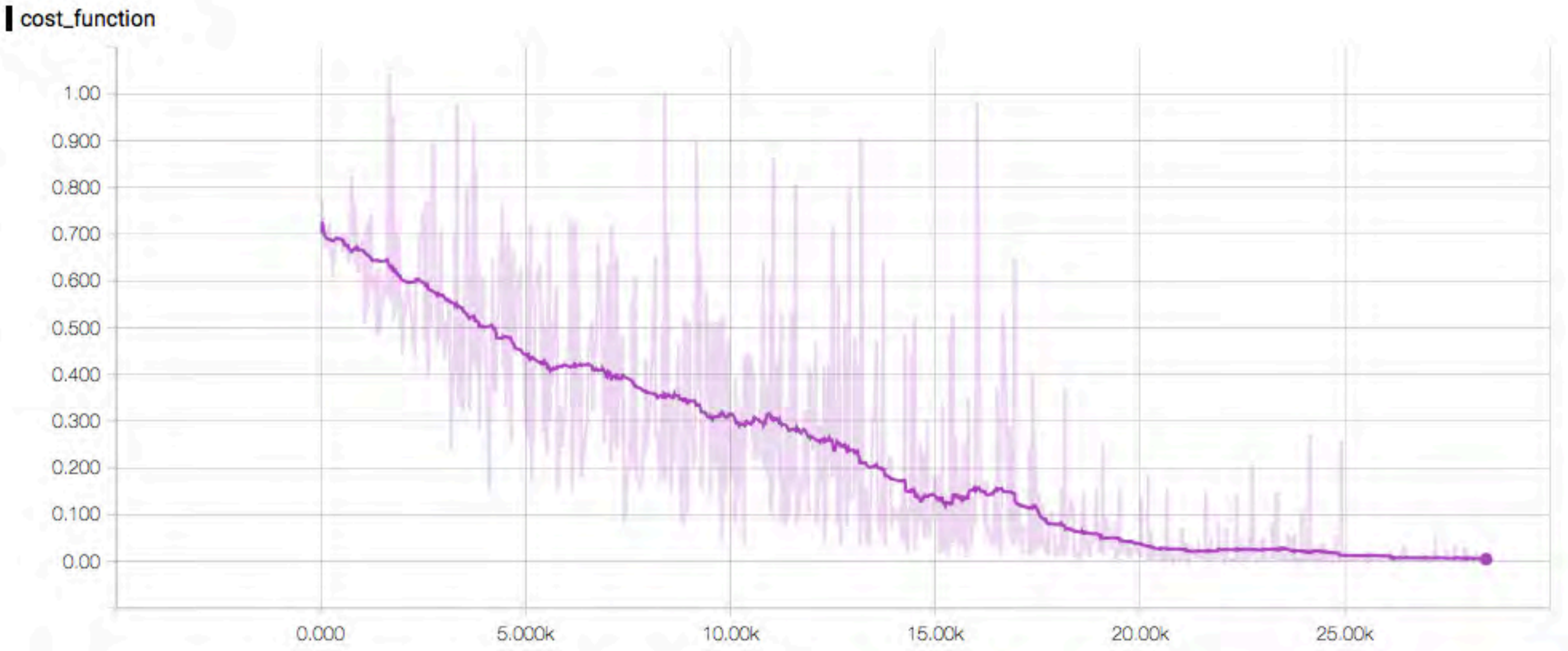
Supervised Learning



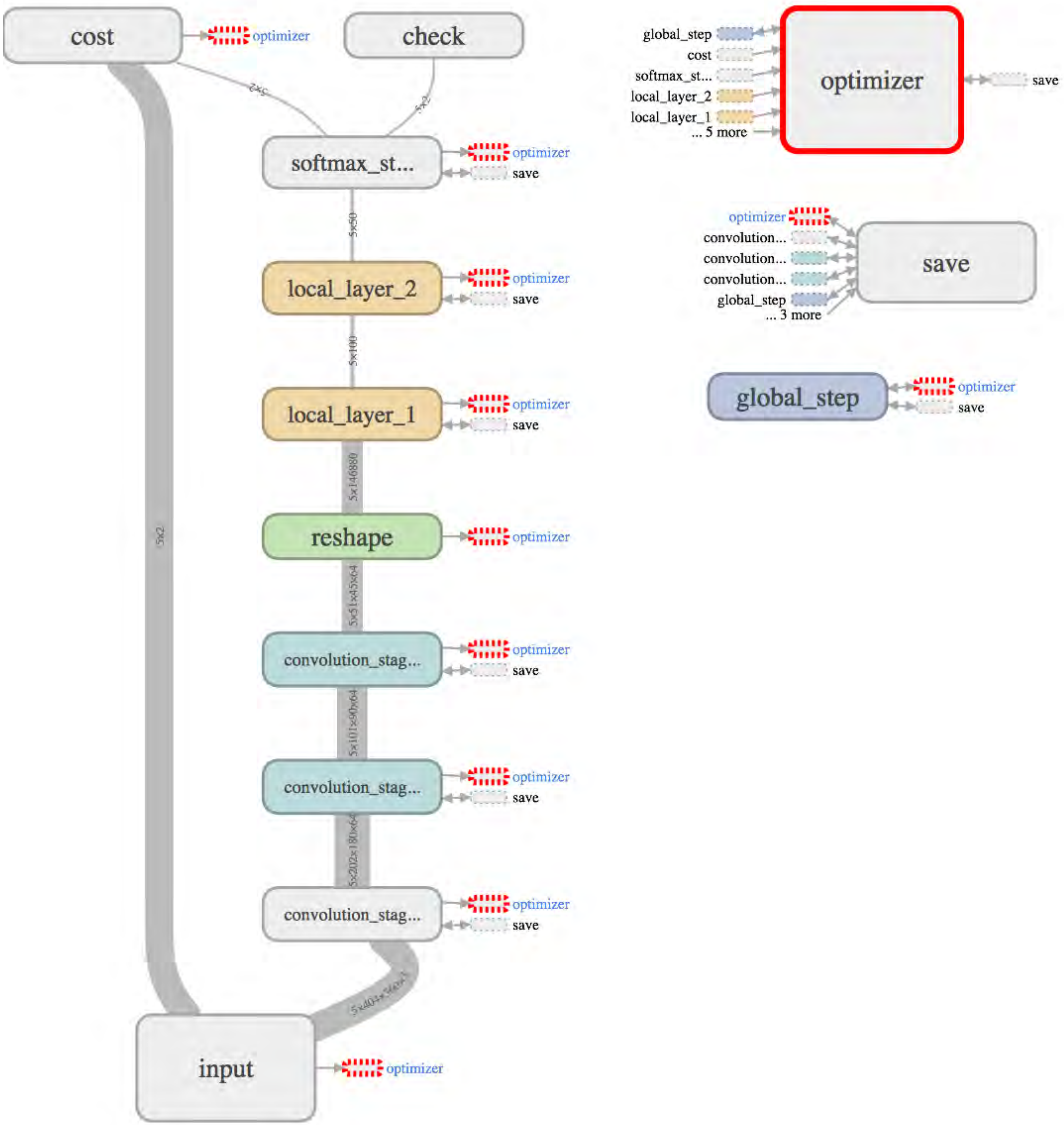
Cluster

„Hund“ / „Katze“

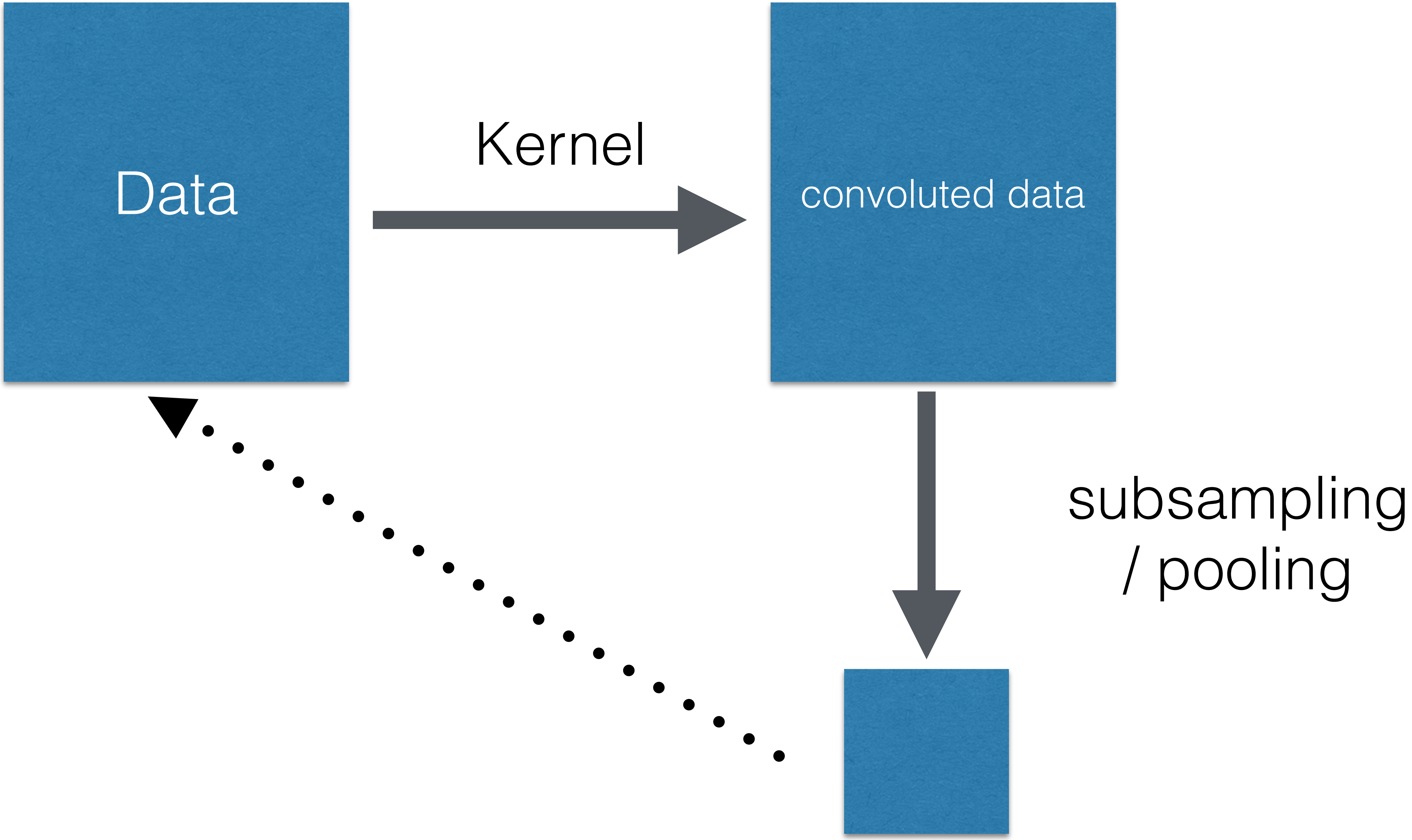
Katze



geglättet Kosten über Epochen


cost
optimizer
check
5x2
5x2
softmax_st...
optimizer
save
5x50
local_layer_2
optimizer
save
5x100
local_layer_1
optimizer
save
5x146880
reshape
optimizer
5x51x45x64
convolution_stag...
optimizer
save
5x101x90x64
convolution_stag...
optimizer
save
5x202x180x64
convolution_stag...
optimizer
save
5x404x360x3
5x2
input
optimizer
global_step
cost
softmax_st...
local_layer_2
local_layer_1
... 5 more
optimizer
save
optimizer
convolution...
convolution...
convolution...
global_step
... 3 more
save
global_step
optimizer
save

Convolution



„Neural networks are engineering, they are not applied mathematics […]“

–S. Matthews, review of machine learning book

This is offensive language!! Stop attacking me

Negin Yaghoubisharif, Masoud Allahyari, Patrick Saad, Christian Paffrath, Philipp Rudloff

Supervisor: Khalid Al-Khatib

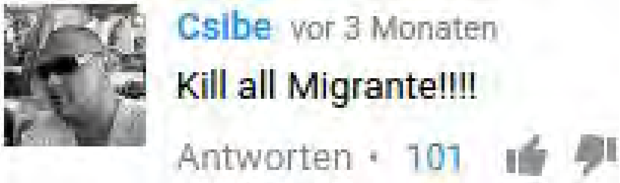
Introduction

Donald J. Trump
@realDonaldTrump
@caaataclysm Little Jon Stewart(?) Is a pussy, he would be hopeless in a debate with me!
07:22 am - 11 Mai 2013
(https://twitter.com/realDonaldTrump/status/333089735570493440)

This is offensive language!! Stop attacking me

Introduction

Donald J. Trump
@realDonaldTrump
@caaataclysm Little Jon Stewart(?) Is a pussy, he would be hopeless in a debate with me!
07:22 am - 11 Mai 2013
(https://twitter.com/realDonaldTrump/status/333089735570493440)



Csibe vor 3 Monaten
Kill all Migrante!!!!
Antworten · 101

The New York Times

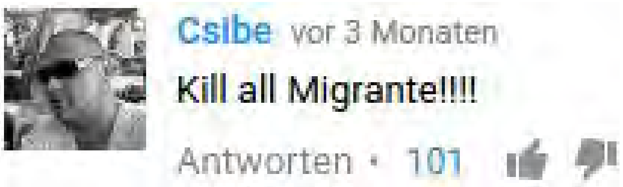
The Washington Post

YouTube

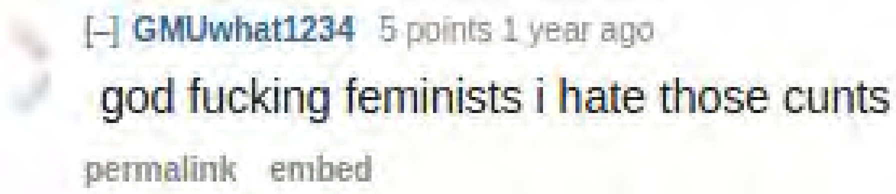
theguardian

Introduction

Donald J. Trump
@realDonaldTrump
@caaataclysm Little Jon Stewart(?) Is a pussy, he would be hopeless in a debate with me!
07:22 am - 11 Mai 2013
(https://twitter.com/realDonaldTrump/status/333089735570493440)



Csibe vor 3 Monaten
Kill all Migrante!!!!
Antworten • 101



[–] GMUwhat1234 5 points 1 year ago
god fucking feminists i hate those cunts
permalink embed

Introduction

Donald J. Trump
@
@
wo
07
(ht
)

CP
@realCPPP
I heard someone say “All russians are drunk bastards” on the bus today.
11:50 am - 8 January 2017
(made up example 1)

permalink embed

Introduction

Donald J. Trump
@
@
wo
07
(ht
)

CP
@realCPPP
I heard someone say “All russians are drunk bastards” on the bus today.
11:50 am - 8 January 2017
(made up example 1)

PD
@realPhifeDog
Those sick dogs!
25:61 am - 13 January 2017
(made up example 2)

permalink embed

Overview

Goal

- **Annotated dataset** for offensive language studies
- **Classifier** with good prediction accuracy, even for new datasets

Overview

Goal

- **Annotated dataset** for offensive language studies
- **Classifier** with good prediction accuracy, even for new datasets

Input

- **Unannotated dataset:** e.g. Twitter tweets

Overview

Goal

- **Annotated dataset** for offensive language studies
- **Classifier** with good prediction accuracy, even for new datasets

Input

- **Unannotated dataset:** e.g. Twitter tweets

Process

- Data annotation: **crowdsourced** instead of expert
- Machine learning method: **active** instead of passive **learning**
- Classification: **Offensive, Other**

Process

Crowdsourcing

- Amazon Mechanical Turk (AMT)
- Why
  - Annotate a large amount of data (HITs)
    - Scalable
  - Cost

amazonmechanical turk
beta
Artificial Artificial Intelligence

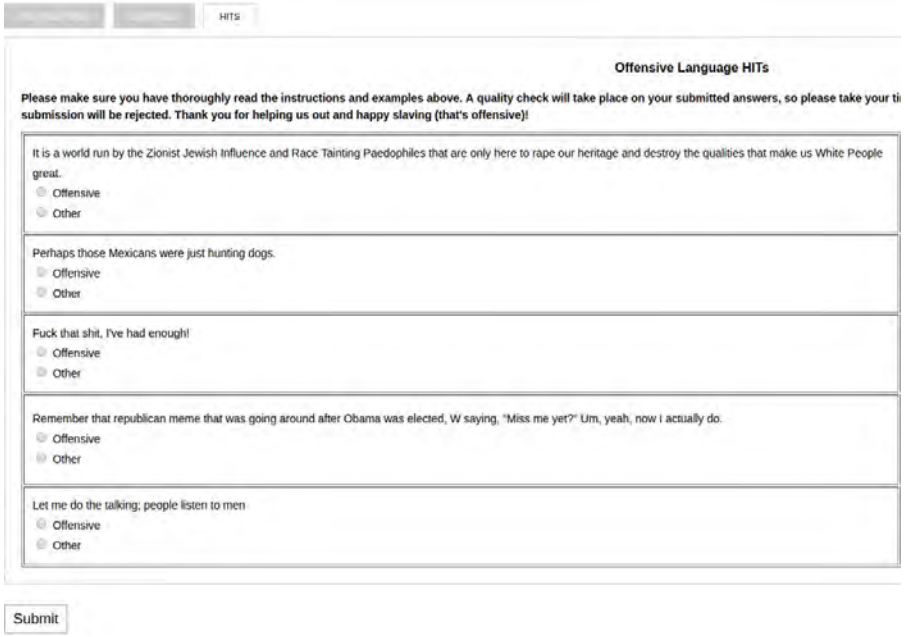
Process

Crowdsourcing

- Amazon Mechanical Turk (AMT)
- Why
  - Annotate a large amount of data (HITs)
    - Scalable
  - Cost

Example with Twitter dataset

- Send tweets for annotation as *Human Intelligence Tasks* (HITs)
- Receive annotated HITs



Offensive Language HITs

Please make sure you have thoroughly read the instructions and examples above. A quality check will take place on your submitted answers, so please take your ti submission will be rejected. Thank you for helping us out and happy slaving (that's offensive)!

It is a world run by the Zionist Jewish Influence and Race Tainting Paedophiles that are only here to rape our heritage and destroy the qualities that make us White People great.
○ Offensive
○ Other

Perhaps those Mexicans were just hunting dogs.
○ Offensive
○ Other

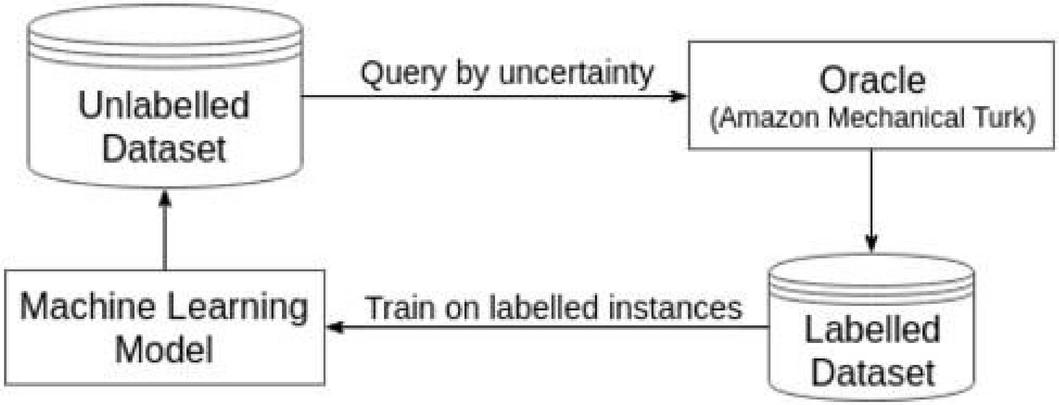
Fuck that shit, I've had enough!
○ Offensive
○ Other

Remember that republican meme that was going around after Obama was elected, W saying, "Miss me yet?" Um, yeah, now I actually do.
○ Offensive
○ Other

Let me do the talking; people listen to men
○ Offensive
○ Other

Process

Active Learning



Applications

- Browser plugin to hide/mark offensive content
- Automatically flag content for moderation
  - Wordpress
  - All sites that have user-generated content
- Use degree of offensiveness as quality measurement for posts
  - Numerical value that shows an offensiveness score for every post

Essay Scoring

Patrick Saad, Yamen Ajjour, Henning Wachsmuth, Khalid Al-Khatib

WS15-16 project | HIWI in 2016

HIWI supervisors: Khalid Al-Khatib, Henning Wachsmuth

Essay Scoring

The goal is to ...

improve the efficiency of large-scale essay scoring

> International Corpus of Learner English (ICLE) (Granger et al., 2009)

> 10 prompt topics

Essay Scoring

What kind of score

Score between 1 to 4 regarding a specific **dimension**

> Argument Strength *(Persing et al., 2015)*

> Thesis Clarity (*Persing et al., 2013)*

more examples

> Organization

> Coherence

Essay Scoring

Technical

What we used

> Apache UIMA
> JAVA 8
> Tomcat 8

Features

> POS (1-5)
> Token N-grams (1-3)
> Sentiment Analysis
> Argumentative Discourse Units
> Document Foundness

Results

	Persing et al.	Patrick et Yamen
Argument Strength	0.244	**0.2251**
Thesis Clarity	0.369	**0.437**

Essay Scoring

Online demo

http://webis16.medien.uni-weimar.de/essay-scoring

Bauhaus-Universität Weimar

Media Faculty

Webis

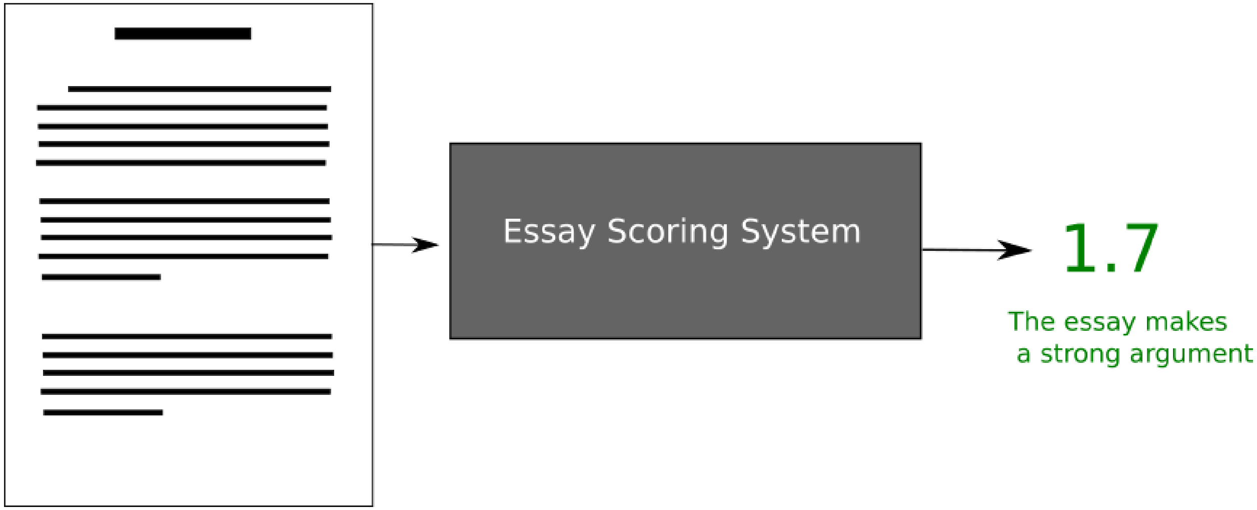
Argument Unit Segmentation

By: Yamen Ajjour (HIWI)

Supervisor : Johannes Kiesel

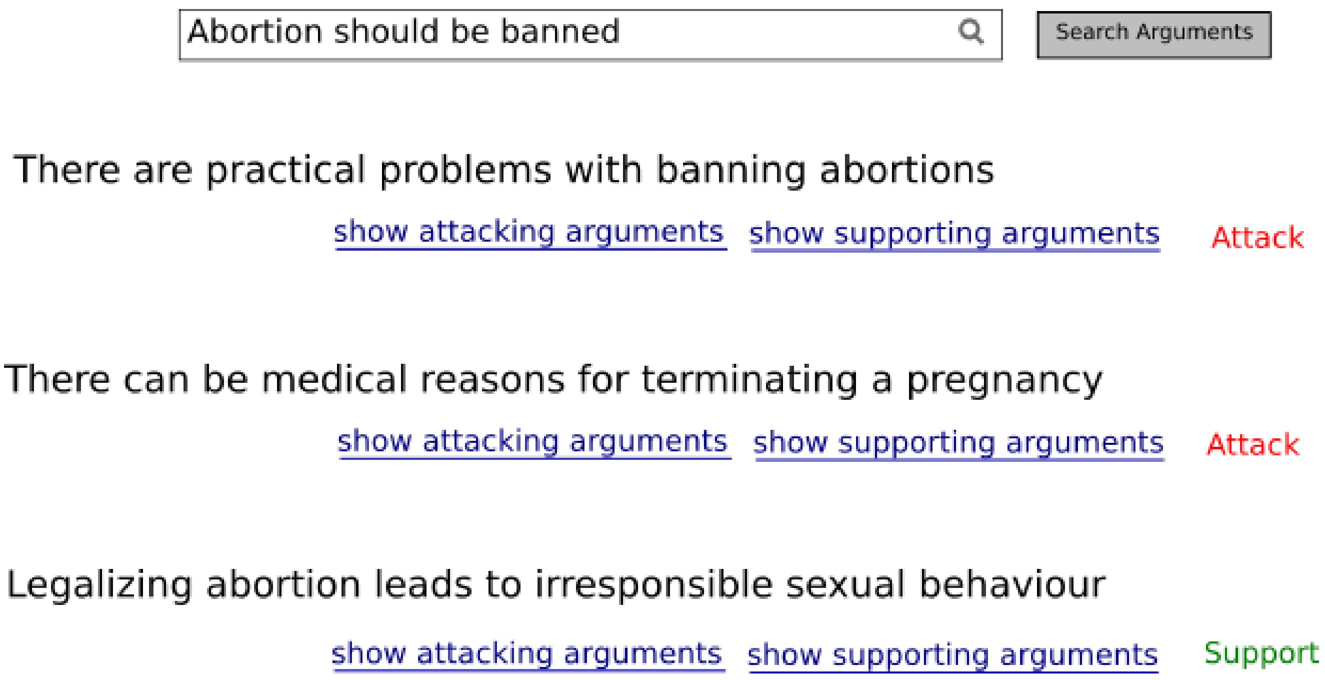
Argument Unit Segmentation
Motivation

Essay Scoring Systems



Argument Unit Segmentation
Motivation

Argument Search Engine



Abortion should be banned

There are practical problems with banning abortions

show attacking arguments show supporting arguments Attack

There can be medical reasons for terminating a pregnancy

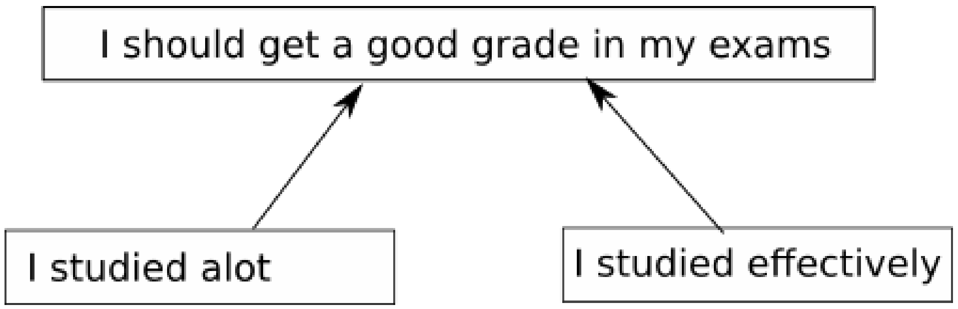
show attacking arguments show supporting arguments Attack

Legalizing abortion leads to irresponsible sexual behaviour

show attacking arguments show supporting arguments Support

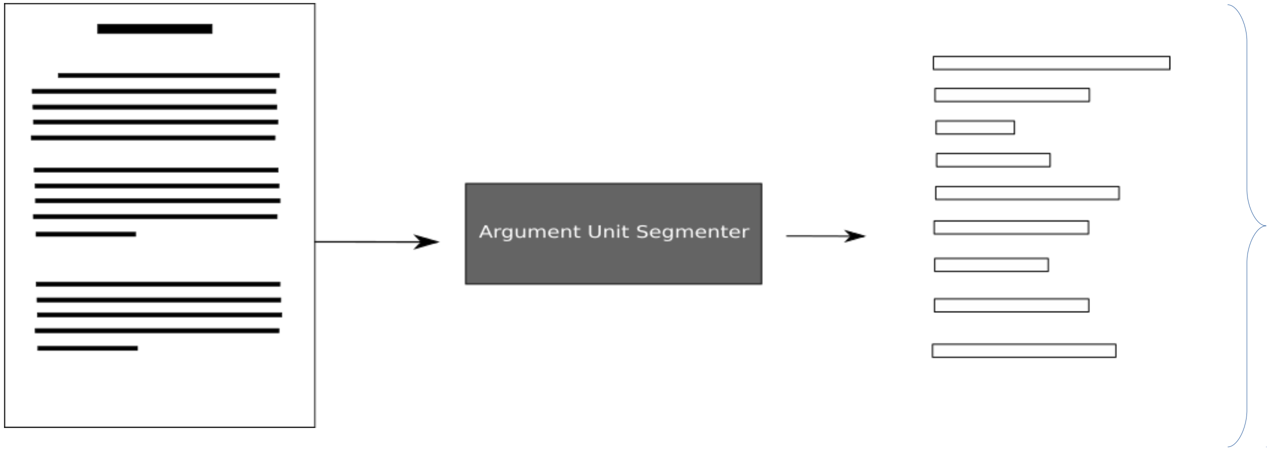
Argument Unit Segmentation
What is an argument ?

An argument is a set of argument units (a conclusion and set of premises)



Argument Unit Segmentation
Problem Definition

Given a plain text how to split it into text segments that cover argument units ?

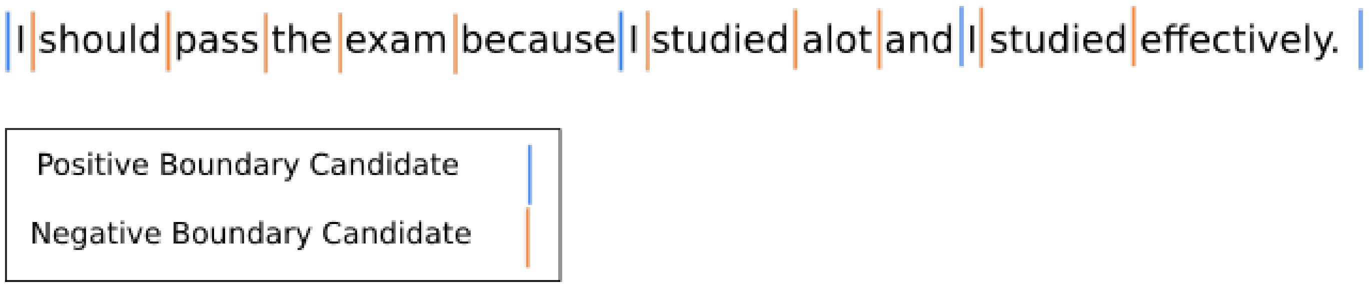


Argument Units Candidates

Argument Unit Segmentation

Approach

- **Conventional:** An argument unit spans a grammatical unit, such as clause or sentence
- **Training a classifier to distinguish the boundaries of an argument unit**
  - An input instance: Each two consecutive words is a boundary candidate
  - Features :
    - Is the boundary candidate at the start of a clause
    - Is the boundary candidate after a comma
    - Is the boundary candidate after a connective



I should pass the exam because I studied alot and I studied effectively.

Positive Boundary Candidate |

Negative Boundary Candidate |

Hiwi Work:
Web Resources for Argument Mining

by Kevin Lang and Jakob Herpel

Motivation

We want to develop a system to detect arguments.

To improve...

- Search Engines
- Dialog Systems
- Writing Assistant Applications

Solution: Machine Learning!

- Problem: require labeled dataset (manual labeling very difficult and expensive)
- Solution: automatic creation of labeled dataset for argument mining

How? Finding appropriate web resources and filter them automatically!

ClueWeb12

“**Classifying Arguments by Scheme**” by Feng and Hirst, 2011

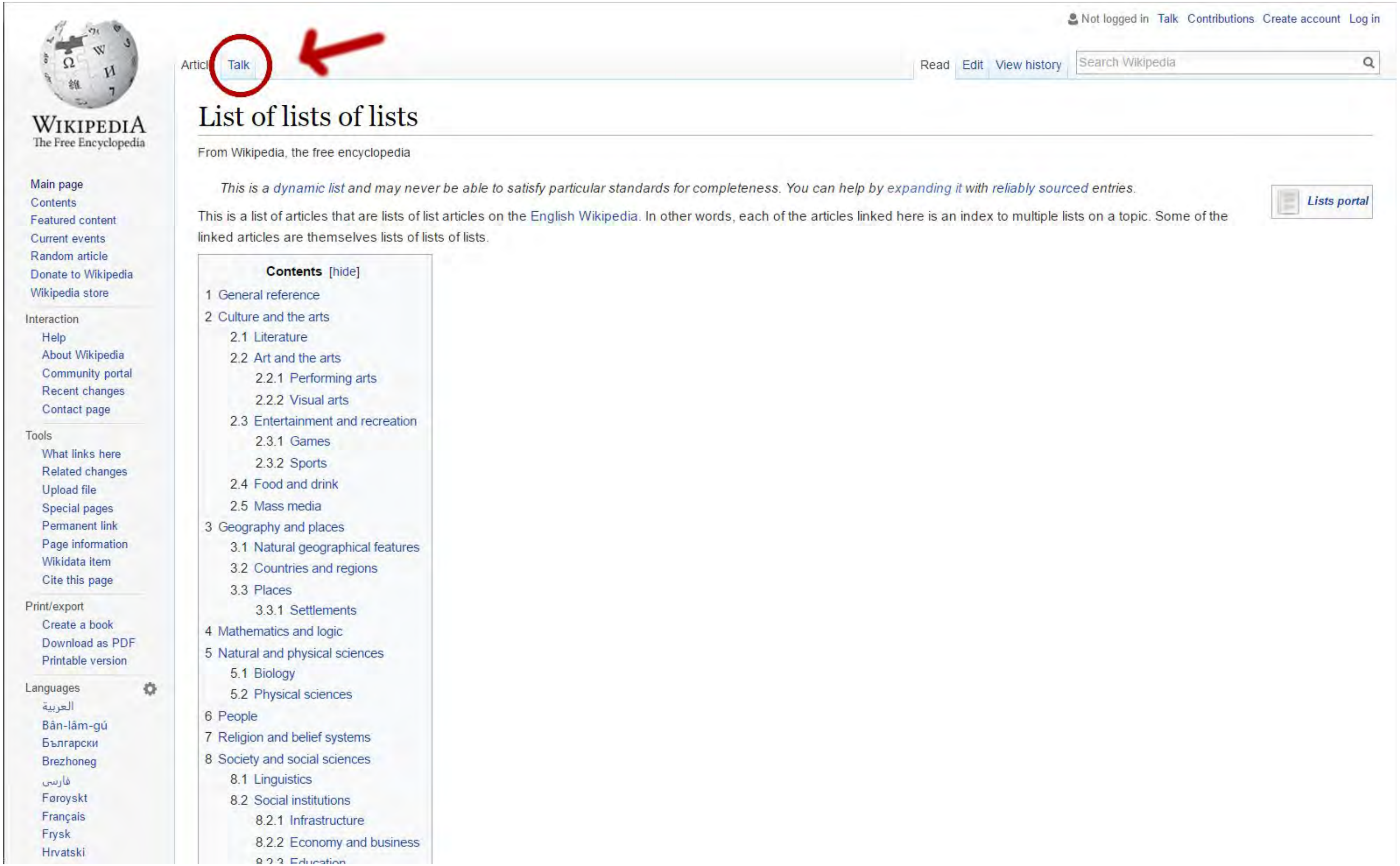
- *Argument from example*
- *Argument from cause to effect*
- *Practical reasoning*
- **Argument from consequences**
- *Argument from verbal classification*

Consequences in ClueWeb12

- How to extract consequences:
  - lexical indicators, sentiment analysis, relatedness to good/bad concepts

- We use:
  - regular expressions, sentiment analysis tools, word2vec

Wiki Talks



List of lists of lists

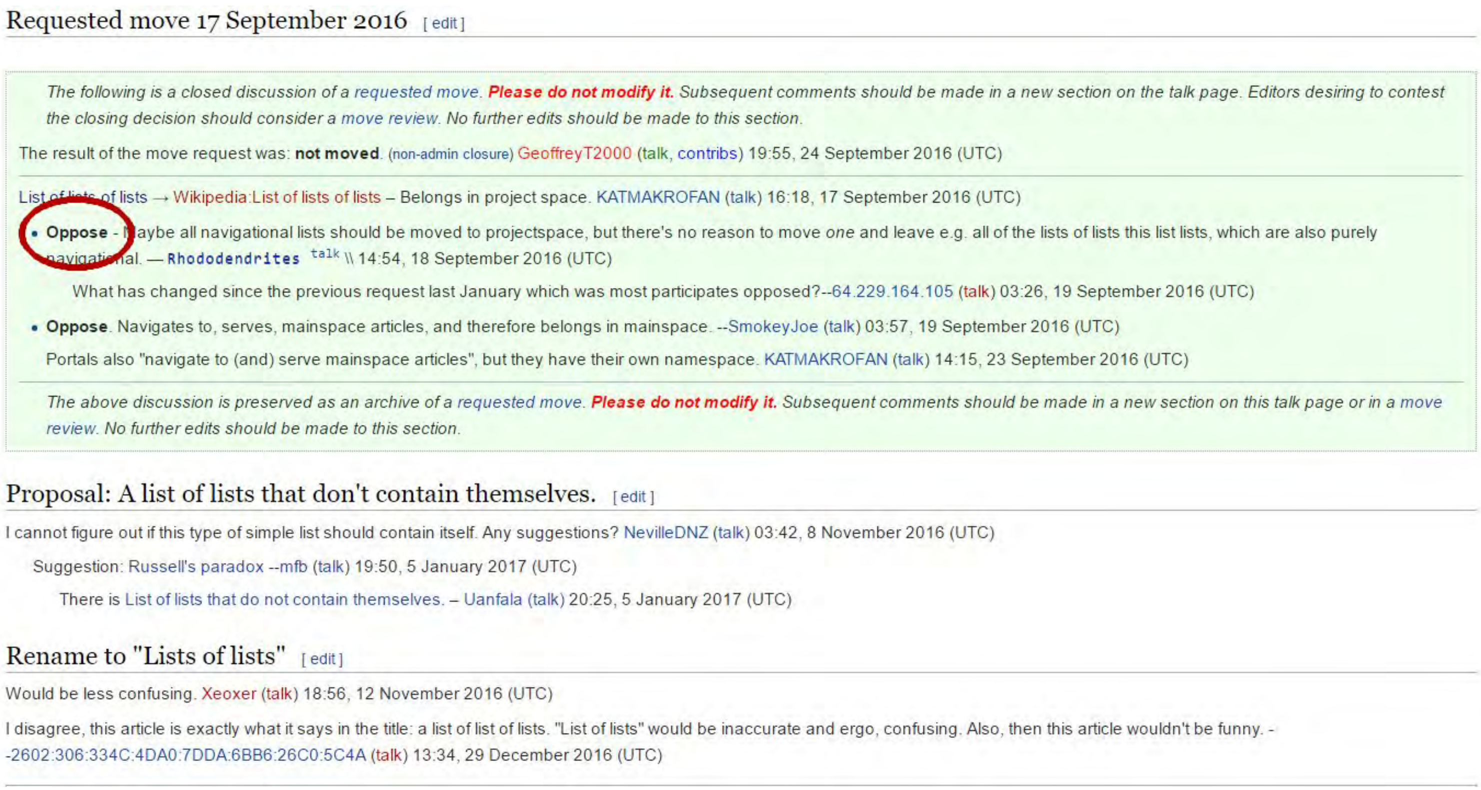
From Wikipedia, the free encyclopedia

This is a dynamic list and may never be able to satisfy particular standards for completeness. You can help by expanding it with reliably sourced entries.

This is a list of articles that are lists of list articles on the English Wikipedia. In other words, each of the articles linked here is an index to multiple lists on a topic. Some of the linked articles are themselves lists of lists of lists.

Contents [hide]

Comment Tags



Requested move 17 September 2016 [edit]

*The following is a closed discussion of a requested move. **Please do not modify it.** Subsequent comments should be made in a new section on the talk page. Editors desiring to contest the closing decision should consider a move review. No further edits should be made to this section.*

The result of the move request was: **not moved**. (non-admin closure) GeoffreyT2000 (talk, contribs) 19:55, 24 September 2016 (UTC)

List of lists of lists → Wikipedia:List of lists of lists – Belongs in project space. KATMAKROFAN (talk) 16:18, 17 September 2016 (UTC)

- **Oppose** - Maybe all navigational lists should be moved to projectspace, but there's no reason to move *one* and leave e.g. all of the lists of lists this list lists, which are also purely navigational. — Rhododendrites talk \\ 14:54, 18 September 2016 (UTC)

  What has changed since the previous request last January which was most participates opposed?--64.229.164.105 (talk) 03:26, 19 September 2016 (UTC)

- **Oppose**. Navigates to, serves, mainspace articles, and therefore belongs in mainspace. --SmokeyJoe (talk) 03:57, 19 September 2016 (UTC)

  Portals also "navigate to (and) serve mainspace articles", but they have their own namespace. KATMAKROFAN (talk) 14:15, 23 September 2016 (UTC)

*The above discussion is preserved as an archive of a requested move. **Please do not modify it.** Subsequent comments should be made in a new section on this talk page or in a move review. No further edits should be made to this section.*

Proposal: A list of lists that don't contain themselves. [edit]

I cannot figure out if this type of simple list should contain itself. Any suggestions? NevilleDNZ (talk) 03:42, 8 November 2016 (UTC)

Suggestion: Russell's paradox --mfb (talk) 19:50, 5 January 2017 (UTC)

There is List of lists that do not contain themselves. – Uanfala (talk) 20:25, 5 January 2017 (UTC)

Rename to "Lists of lists" [edit]

Would be less confusing. Xeoxer (talk) 18:56, 12 November 2016 (UTC)

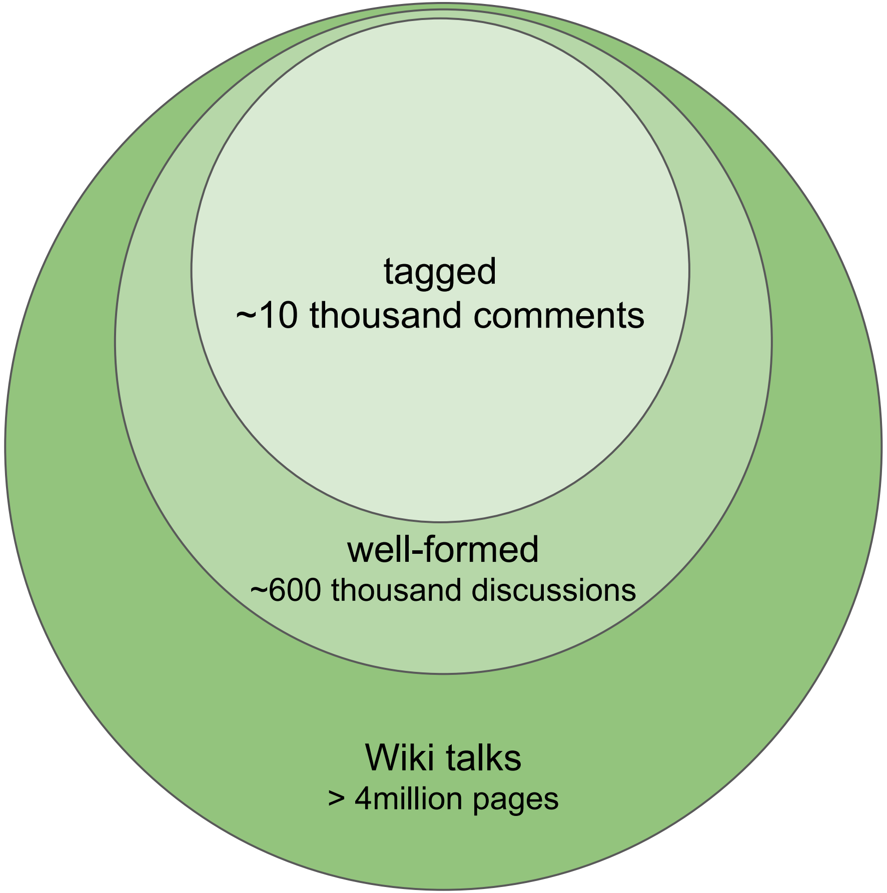
I disagree, this article is exactly what it says in the title: a list of list of lists. "List of lists" would be inaccurate and ergo, confusing. Also, then this article wouldn't be funny. --2602:306:334C:4DA0:7DDA:6BB6:26C0:5C4A (talk) 13:34, 29 December 2016 (UTC)

How:

- parsing wiki discussions with user's metadata (e.g., support and opposed tags)

Challenges:

- ambiguous discussion hierarchies



Conclusion

- Exploited two web resources:
  (a) clueweb12 and (b) wiki talks
  for argument mining.
- Investigated how to extract argument consequences from (a) and argument discourse structures from (b).

Future Works

- Understand segments and parts of sentence where to find consequences
- Fix hierarchy problem by generating more heuristic rules to fix the structure of ill-formed discussions

Thanks for your Attention!

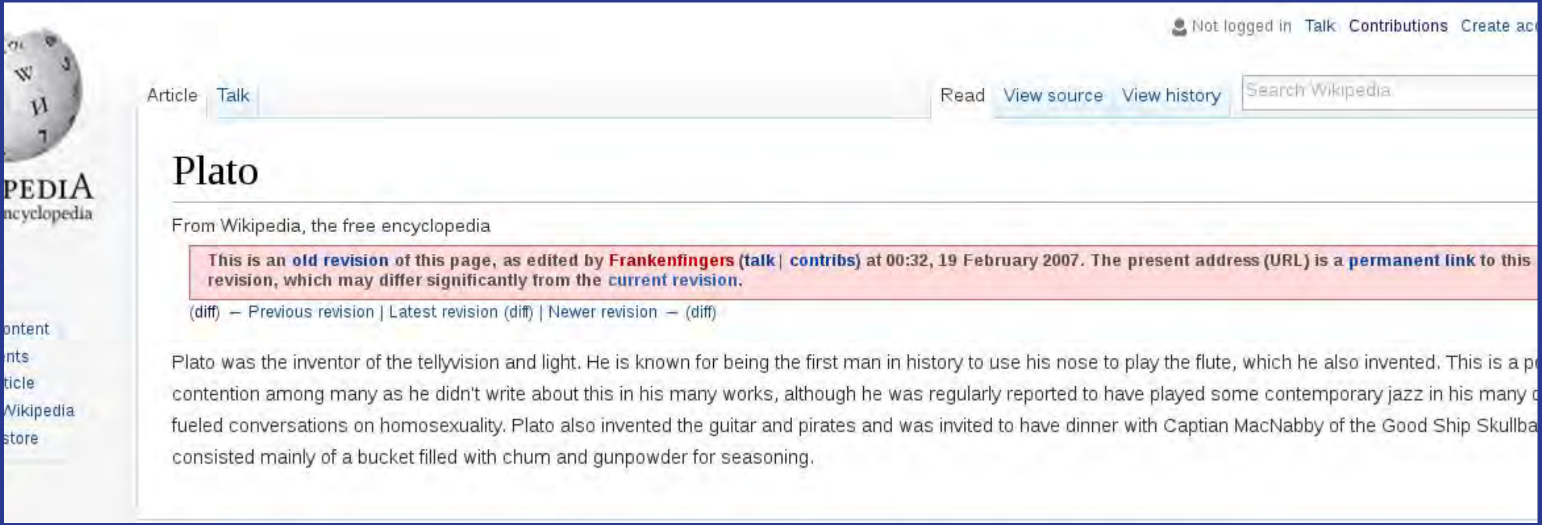
Vandalism In Wikipedia

Milad Alshomary

Bauhaus University

Vandalism In Wikipedia

- Vandalism is defined by Wikipedia as : “Any malicious edit which attempts to reverse the main goal of the project of Wikipedia.”



Plato

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by Frankenfingers (talk | contribs) at 00:32, 19 February 2007. The present address (URL) is a permanent link to this revision, which may differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

Plato was the inventor of the tellyvision and light. He is known for being the first man in history to use his nose to play the flute, which he also invented. This is a p contention among many as he didn't write about this in his many works, although he was regularly reported to have played some contemporary jazz in his many c fueled conversations on homosexuality. Plato also invented the guitar and pirates and was invited to have dinner with Captian MacNabby of the Good Ship Skullba consisted mainly of a bucket filled with chum and gunpowder for seasoning.

Vandalism In Wikipedia

- Solutions against vandalism:
  - Reviewing edits and contributions by users.
  - Automated tools to prevent vandalism.
  - **What about mining the vandalism and trying to understand the pattern behind it?**

Vandalism In Wikipedia

- A research to perform spatio-temporal analysis of vandalism on Wikipedia.

Johannes Kiesel

Martin Potthast

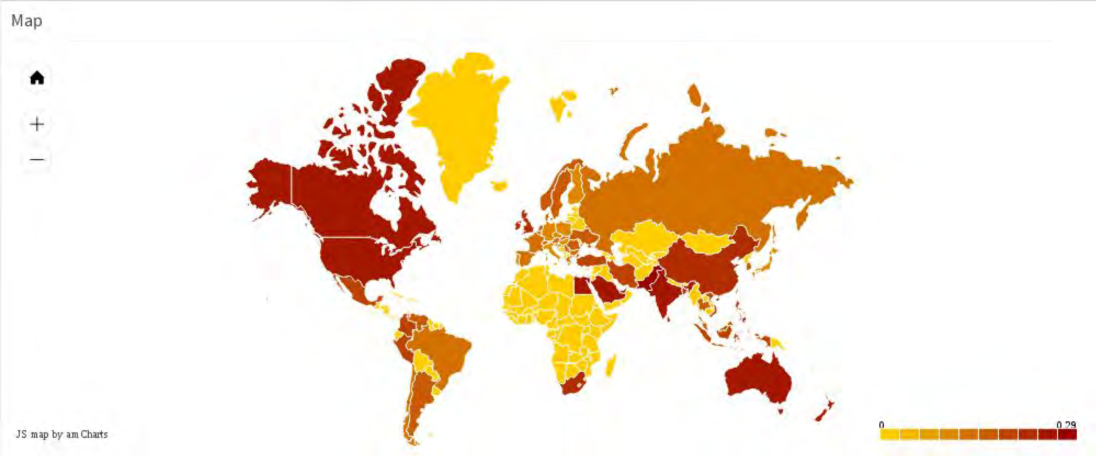
Matthias Hagen

Benno Stein

Vandalism In Wikipedia

Interactive web interface to communicate the results:

How the spatial distribution of vandalism ratio looks like?

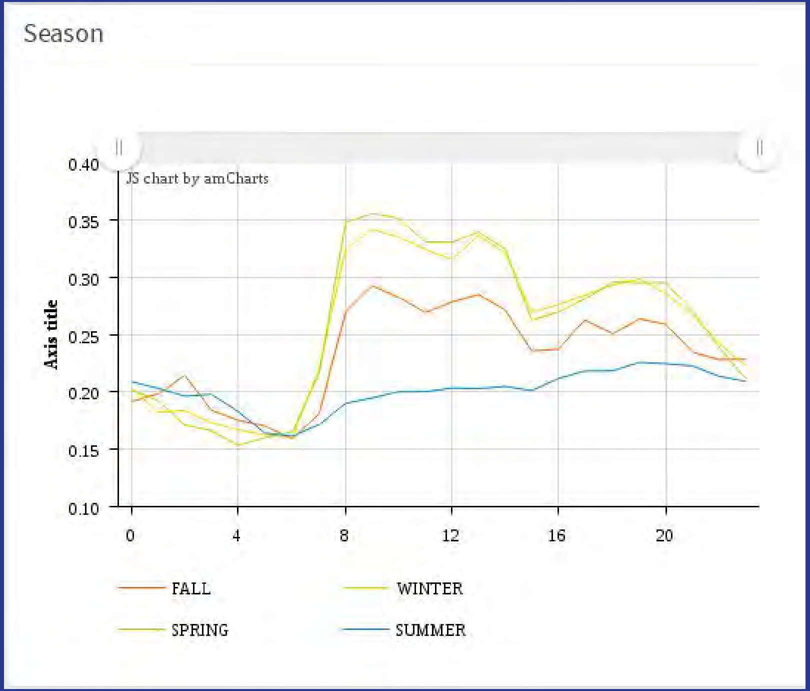


Heat map of vandalism ratio

Vandalism In Wikipedia

Interactive web interface to communicate the results:

How the vandalism ratio changes over hour of day in each season in USA?

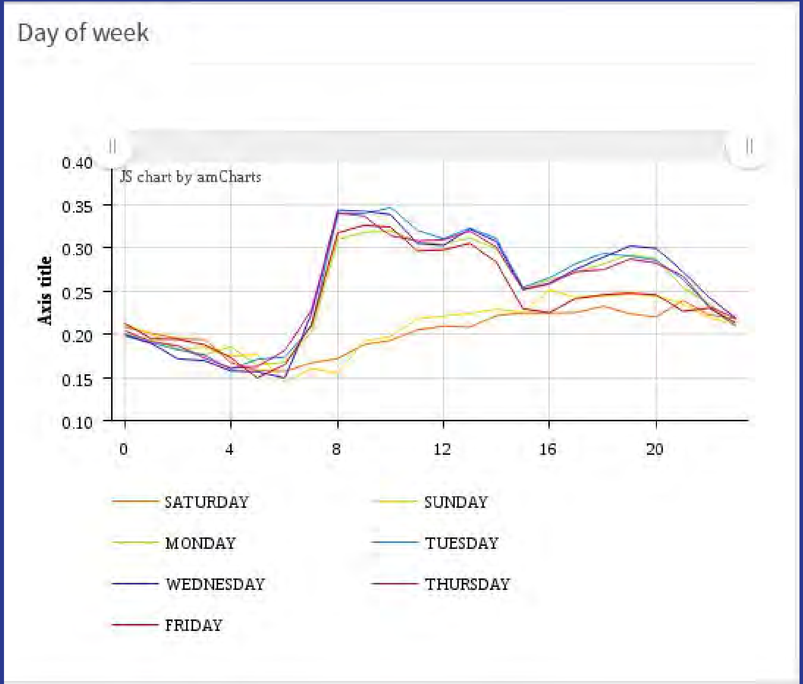


Hour of the day to ratio of vandalism (By season)

Vandalism In Wikipedia

Interactive web interface to communicate the results:

How the vandalism ratio changes over hour of day for each week day in USA?

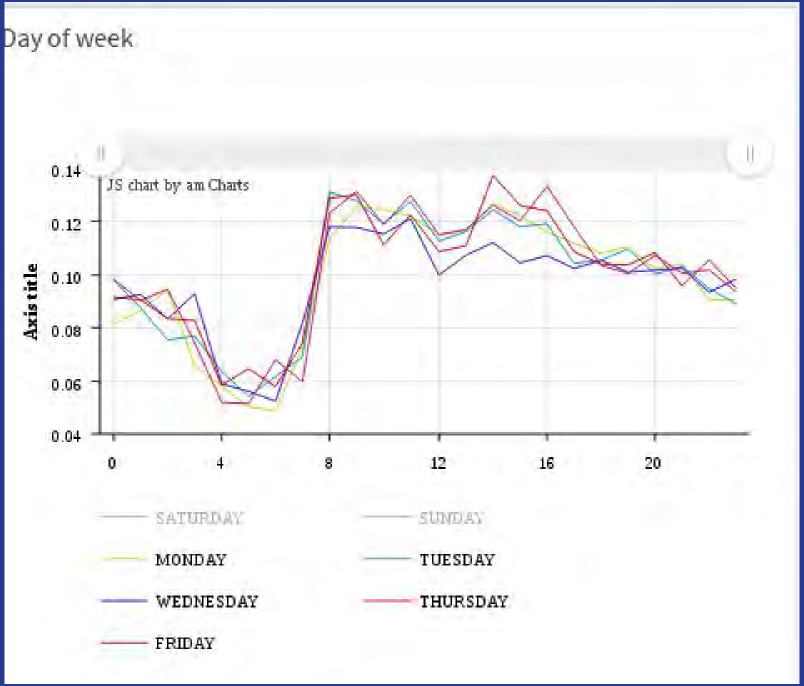


Hour of the day to ratio of vandalism (By day of the week)

Vandalism In Wikipedia

Interactive web interface to communicate the results:

How the vandalism ratio changes over hour of day for each week day in France?



Hour of the day to ratio of vandalism (By day of the week) In France

Vandalism In Wikipedia

- Research Results :

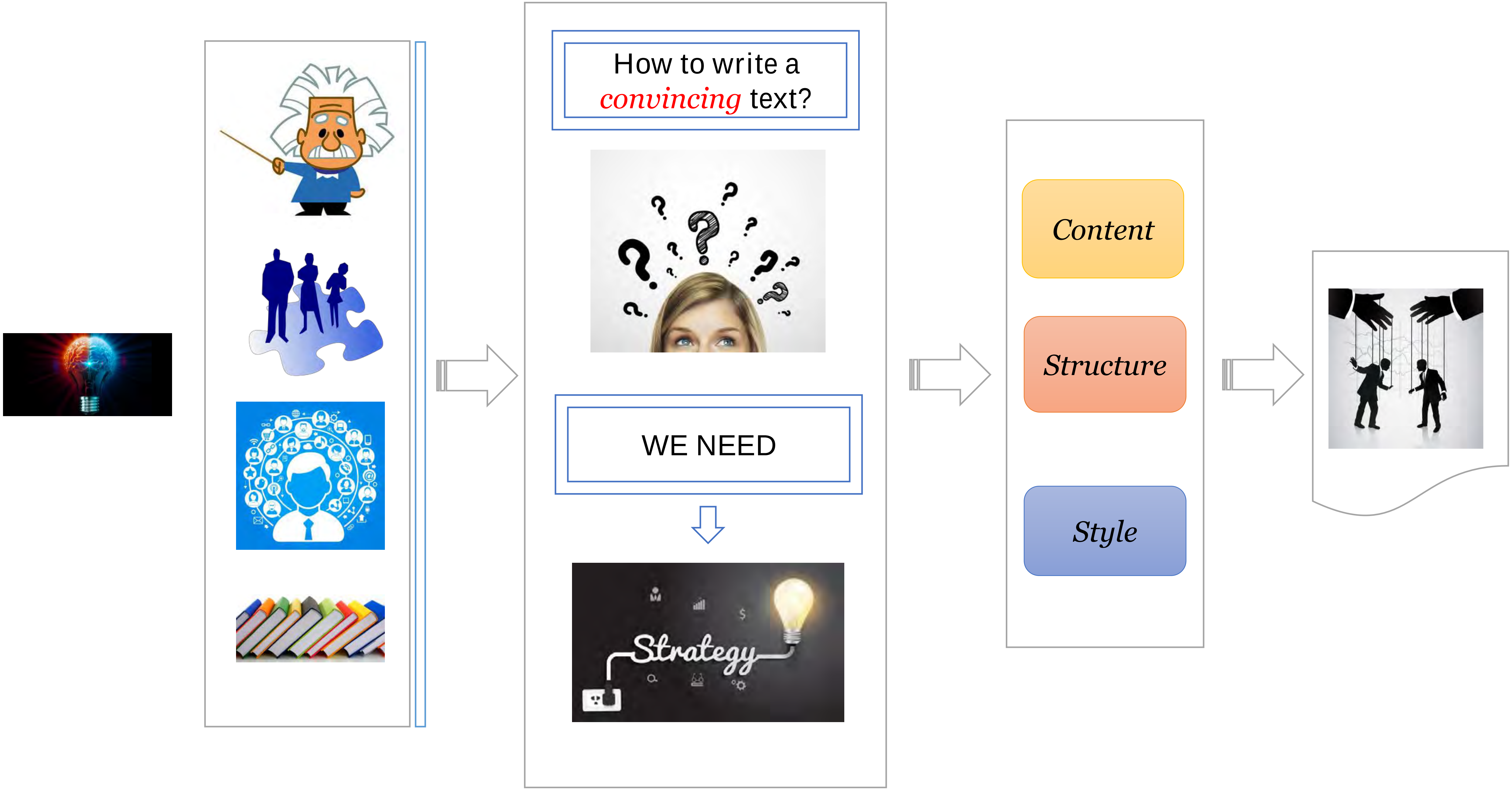
Vandalism is related to work/study habits!!

Vandalism may be a form of stress-relief

Argumentation Strategies

Khalid Al-Khatib, Henning Wachsmuth, Benno Stein

Bauhaus-Universität Weimar



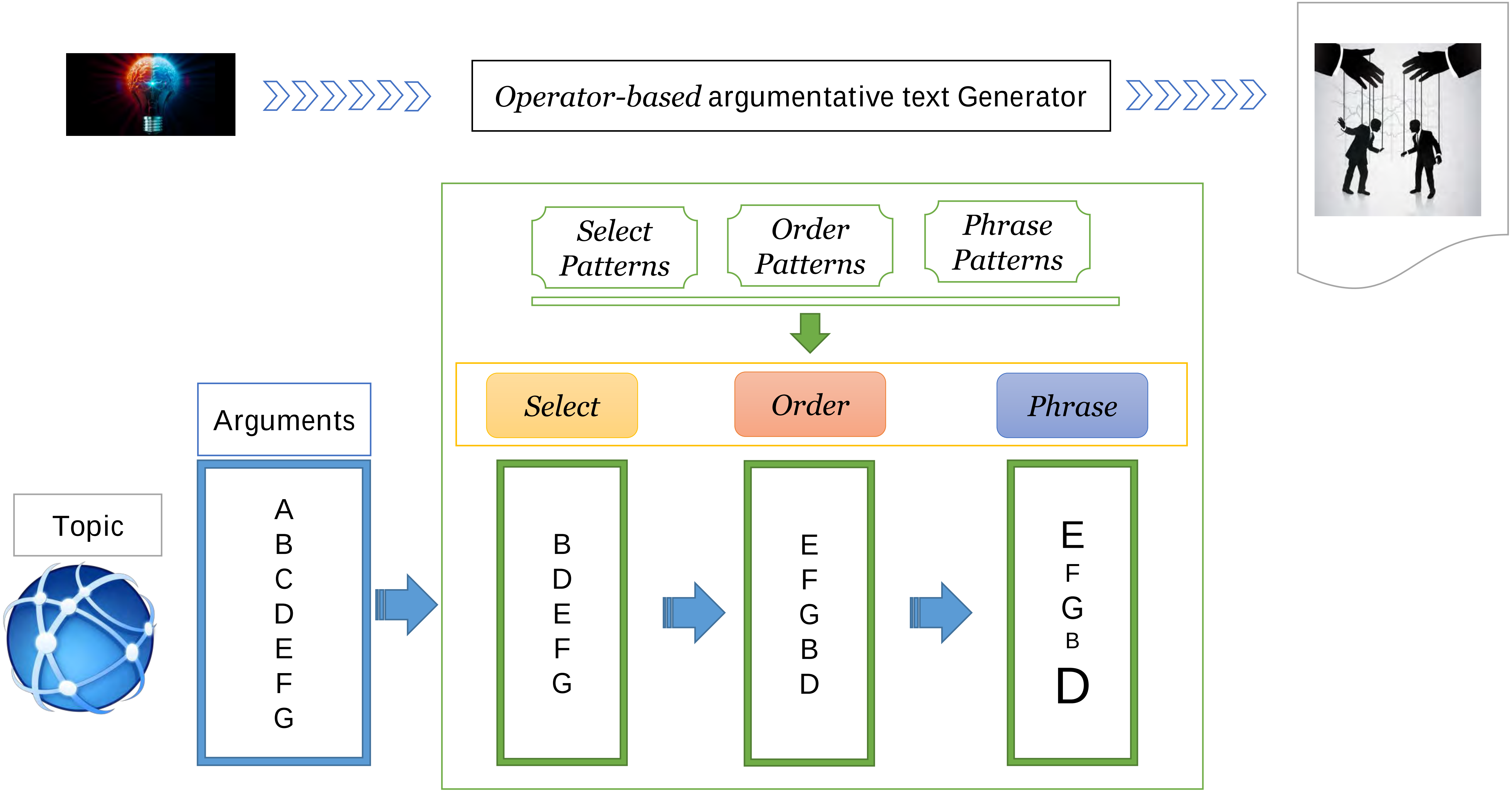
How to write a *convincing* text?

WE NEED

Content

Structure

Style



Operator-based argumentative text Generator

Select Patterns

Order Patterns

Phrase Patterns

Select

Order

Phrase

Topic

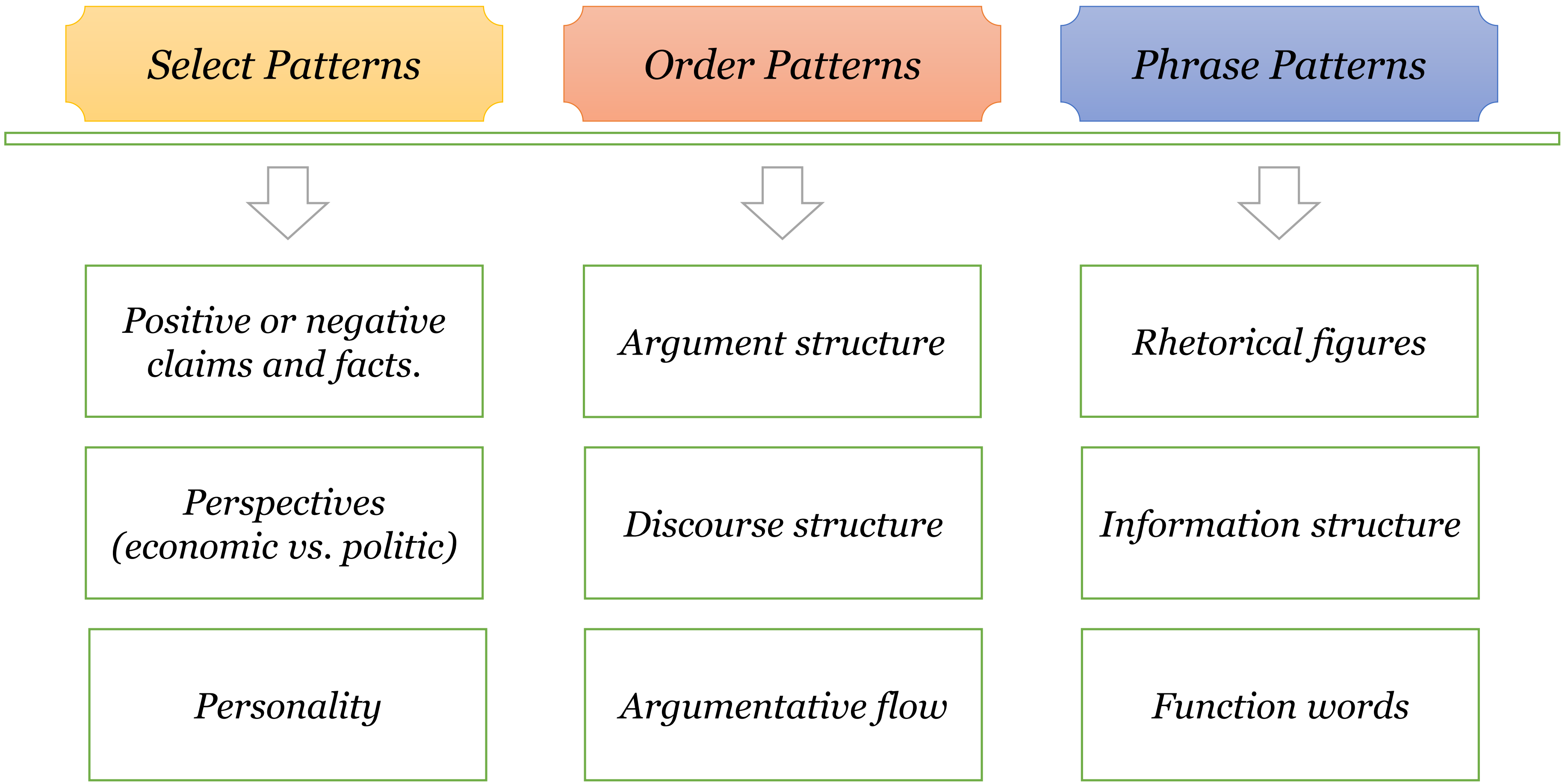
Arguments

A
B
C
D
E
F
G

B
D
E
F
G

E
F
G
B
D

E
F
G
B
D



Select Patterns	*Order Patterns*	*Phrase Patterns*
Positive or negative claims and facts.	*Argument structure*	*Rhetorical figures*
Perspectives (economic vs. politic)	*Discourse structure*	*Information structure*
Personality	*Argumentative flow*	*Function words*

2017 ACM DEBS Grand Challenge

12. Januar 2017

Overview

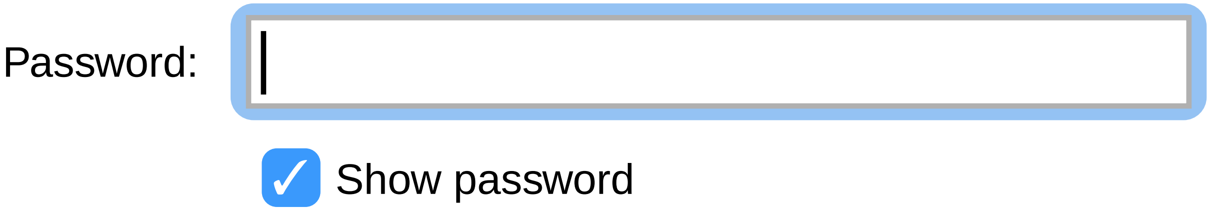
- goal: detecting anomalies in the behaviour of a manufacturing machine using sensor data
- restrictions:
  - two different types of machine: injection molding machine and assembly machine
  - sensor data delivered at a certain time
  - following procedure for detecting the anomalies:
    1. **clustering all the data of one sensor and one machine in the last W time units (where number of cluster is given)**
    2. **modelling the transitions from one cluster centre to another as a Markov process**
    3. **detecting anomalies**
  - input: sensor data of each machine represented as RDF tuples containing information on type, number of clusters per dimension and machine instance, current sensor data, timestamp
  - output: anomaly or not in a certain dimension and in a certain machine in RDF

A Large-scale Analysis of the Mnemonic Password Advice

Johannes Kiesel, Benno Stein, Stefan Lucks
Bauhaus-Universität Weimar
www.webis.de

Weimar, January 12th 2017

Problem: You need a new password



Password:

☑ Show password

Problem: You need a new password

Method: Standard

Let‘s just pick...

Password: Password123

☑ Show password

Problem: You need a new password

Method: Dictionary

➜

⚁ ⚃ ⚂ ⚄ ⚅	embalm
⚁ ⚅ ⚃ ⚅ ⚀	fuss
⚅ ⚄ ⚀ ⚄ ⚄	yogi
⚂ ⚅ ⚀ ⚅ ⚅	layup

Password: embalm fuss yogi layup

☑ Show password

Problem: You need a new password

Method: Mnemonic

➜

W**hen **I** **w**alked **to** **t**he **g**rocery **s**tore**,
***t**here **w**ere **c**amels **f**lying **o**verhead*

Password: WIw2tgs,twcfo |

☑ Show password

What method is best?

Example: passwords with on average 9 characters
(an overly simple comparison)

Method	Standard[1]	Dictionary[2]	Mnemonic[3]	Max[4]	
Like rolling a dice with	10^8	10^7	**?**	10^{17}	faces

[1] J. Bonneau, “The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords”, 2012. Average length unknown, but current minimum of the service is 8.
[2] Two diceware words, average length of 9.5
[3] Minimum length of 8, average length of 9.4
[4] Uniform distribution over 95 characters, length of 9

What method is best?

Example: passwords with on average 9 characters
(an overly simple comparison)

Method	Standard[1]	Dictionary[2]	Mnemonic[3]	Max[4]	
Like rolling a dice with	10^8	10^7	10^{11}	10^{17}	faces

[1] J. Bonneau, “The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords”, 2012. Average length unknown, but current minimum of the service is 8.

[2] Two diceware words, average length of 9.5

[3] Minimum length of 8, average length of 9.4

[4] Uniform distribution over 95 characters, length of 9

What method is best?

Example: passwords with on average 9 characters
(an overly simple comparison)

Method	Standard[1]	Dictionary[2]	Mnemonic[3]	Max[4]	
Like rolling a dice with	10^8	10^7	10^{11}	10^{17}	faces

Our approach:

[1] J. Bonneau, "The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords", 2012. Average length unknown, but current minimum of the service is 8.
[2] Two diceware words, average length of 9.5
[3] Minimum length of 8, average length of 9.4
[4] Uniform distribution over 95 characters, length of 9

What method is best?

Example: passwords with on average 9 characters
(an overly simple comparison)

Method	Standard[1]	Dictionary[2]	Mnemonic[3]	Max[4]	
Like rolling a dice with	10^8	10^7	10^{11}	10^{17}	faces

Our approach:

≠

↓ ↓

WIw2tgs,twcfo ≈ Pdnatraia7dowyrti

[1] J. Bonneau, “The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords”, 2012. Average length unknown, but current minimum of the service is 8.

[2] Two diceware words, average length of 9.5

[3] Minimum length of 8, average length of 9.4

[4] Uniform distribution over 95 characters, length of 9

Our research

The data

730,000,000	Web pages	(ClueWeb12, 27.3 TB)
3,400,000,000	Sentences	(aitools4-aq-web-page-content-extraction)
25,000,000,000	Passwords	(18 password generation rules)
6,400	Language models	(5.1 TB)
970	Strength estimates	(4 metrics)

Selected results for mnemonic passwords:

- Use lowercase letters only, but one character more
- Sentence complexity does not play a big role
- Password strength is far worse than for purely random passwords

Our research

The data

730,000,000	Web pages	(ClueWeb12, 27.3 TB)
3,400,000,000	Sentences	(aitools4-aq-web-page-content-extraction)
25,000,000,000	Passwords	(18 password generation rules)
6,400	Language models	(5.1 TB)
970	Strength estimates	(4 metrics)

Selected results for mnemonic passwords:

- Use lowercase letters only, but one character more
- Sentence complexity does not play a big role
- Password strength is far worse than for purely random passwords

Thank you for your attention