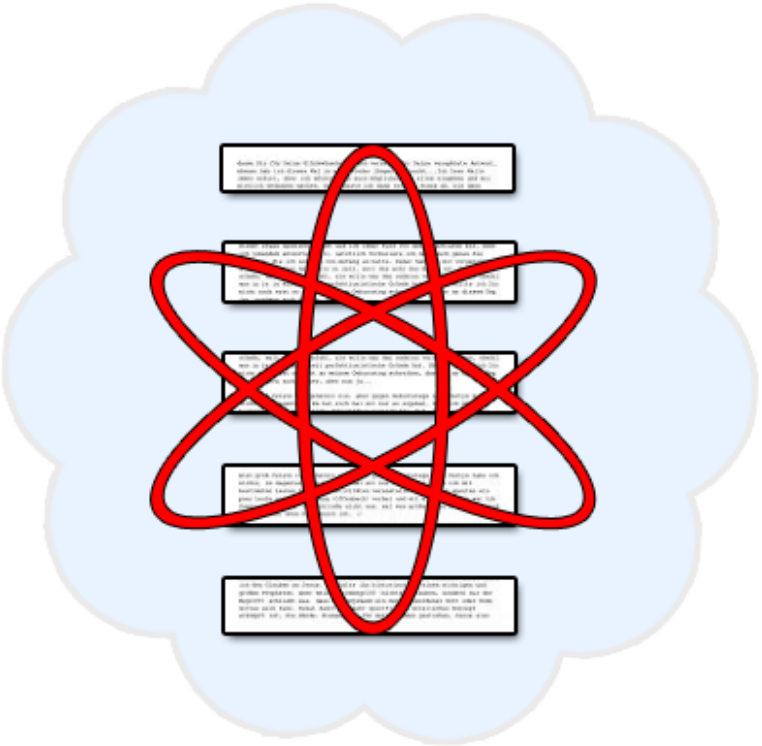
Intrinsic Plagiarism Detection

Register & Genre

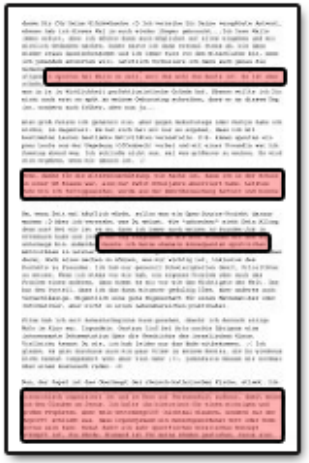
Oren Halvani

TECHNISCHE UNIVERSITÄT DARMSTADT

Before



After



Outline

TECHNISCHE UNIVERSITÄT DARMSTADT

- Motivation
- External / Internal Plagiarism Detection
- Text layers
- Features / Feature Extraction
- Similarity Functions
- Putting It All Together
- Take-Home-Messages / References

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Thanks to the internet the amount of information nowadays has grown drastically during the last two decades

> *"The Amount of Digital Information Reached 281 Exabytes (281 Billion Gigabytes)."* [1]

- A huge part of these information is represented through text, e.g. e-Mails, websites, office documents, e-Books, etc.

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Due to [8] current world population estimate: 6,852,472,823
- From this it follows: many different authors have been involved in the generation of these textual information ☺
- Sometimes an author A "borrows" some information (text) from an author B
- Furthermore A may not mention this action…
- This act refers to the term of **plagiarism**

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- What is plagiarism exactly?

> *"**Definition:** Plagiarism is the act, intentional or otherwise, of copying or borrowing words or ideas without properly acknowledging the original source."* [3]

- Since manual detection ist very time-consuming, research focuses on automatic plagiarism detection

- Majority of existing approaches refer to External Plagiarism Detection (EPD)

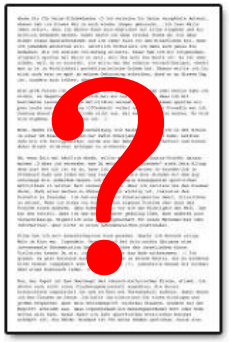
- Rough example…

External Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

Reference corpora (sample documents)

X



External Plagiarism Detector

X is / isn't plagiarized

External Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- Limitation: in Real-World-scenarios reference material isn't always available
- In such cases we wish at least to figure out if a document has been written by only ONE author
- How can we make it possible, if a given document is simultaneously

  → the reference corpus itself ?
- The discipline which occupied with this non-trivial challenge is the so-called:

Intrinsic Plagiarism Detection (IPD)

Intrinsic Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- IPD is a relatively young discipline, in comparison to EPD
- Massively influenced by the following pioneer in this field
- Prof. Dr. Benno Stein,
  Chair of the Web-Technology & Information Systems Group
  → Bauhaus-Universität Weimar

[4]

- Stein and his team published many fundamental methods & theories for IPD
- Furthermore they've developed a unique corpus for text plagiarism cases (PAN)

Intrinsic Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- What does the term intrinsic stand for? One possible definition:

> *"Metaphysically, an intrinsic property of an object is a property that the object has by virtue of itself, depending on no other thing…"* [9]

- In the context of IPD intrinsic stands for "the analysis from inside"

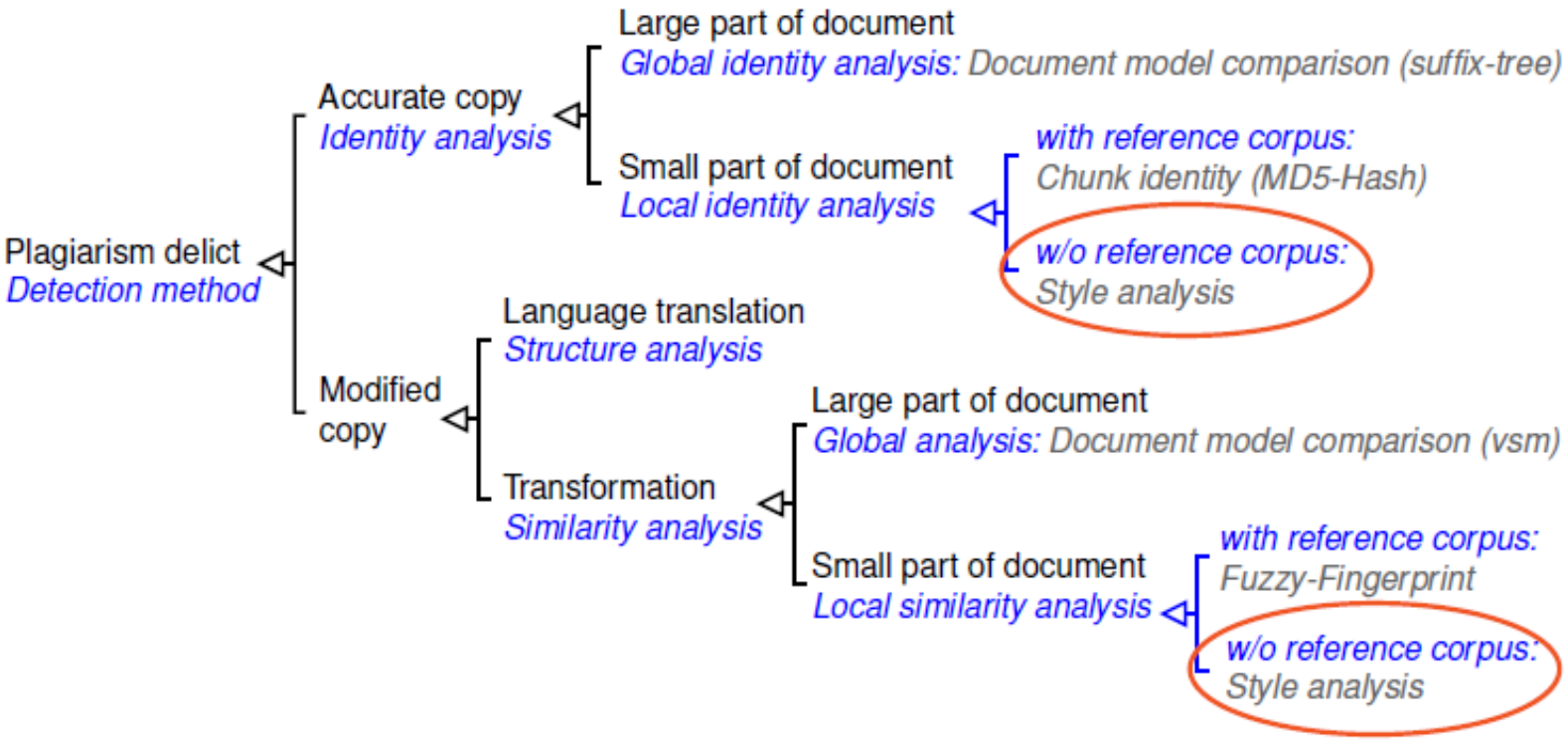
- Main purpose: assists to judge if a document has been plagiarized

- Core concept: **feature extraction → outlier analysis → decision !**

Intrinsic Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- For which plagiarism delicts can we use IPD ?



Encircled parts refer to IPD !

[2]

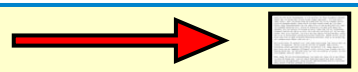
Intrinsic Plagiarism Detection

TECHNISCHE UNIVERSITÄT DARMSTADT

- IPD can only be achieved if features are suitable enough to detect breach of style within the text
- Key question: which features should be used?
- In order to answer this question we should first understand one important fact
- Text typically consists of several layers…

Text layers (small overview)

TECHNISCHE UNIVERSITÄT DARMSTADT

Grapheme Layer:	This is a text
Symbol Layer:	T h i s i s a t e x t
Constituents Layer:	(This (is (a (text))))
Token Layer:	[This] [is] [a] [text]
Phoneme Layer:	/ ðɪs ɪz ə tɛkst /
Part-Of-Speech Layer:	This_**DT** is_**VBZ** a_**DT** text_**NN**
Semantics Layer:	

Features for IPD

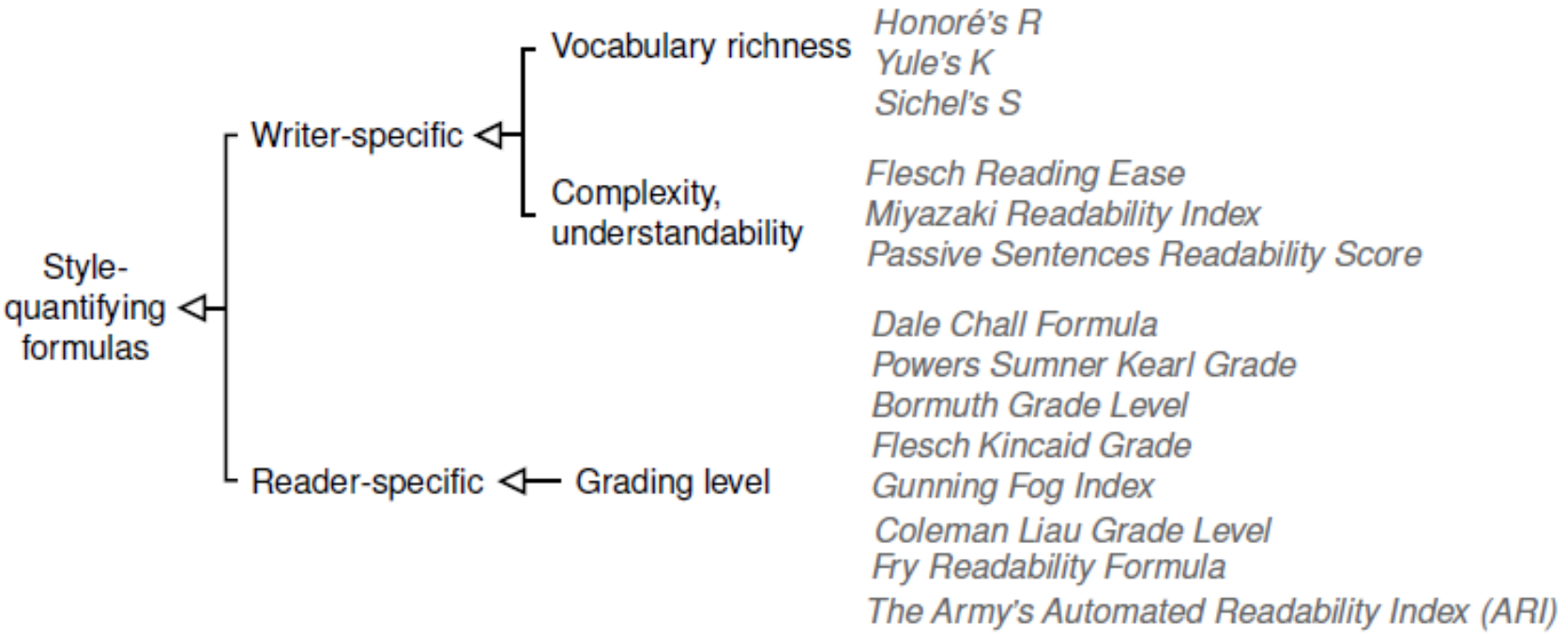
TECHNISCHE UNIVERSITÄT DARMSTADT

- Features should be extracted from any meaningful layer
- Statistically spoken: one/very few layers are not sufficient for the task
- Some possible features:

Symbol Layer	**Token Layer**	**POS Layer**	**POS Phrase Layer**	**Semantics Layer**
Ratio of vowels per word: (e, a, u, o, i, y)	Ratio of content-words/ all words	Ratio of nouns, verbs, adverbs adjectives and all words	Ratio of NP‘s, VP‘s, PP‘s and all sentences	Ratio of synonym usage and the corresponding synset

Features for IPD

- Other (more reliable) features…

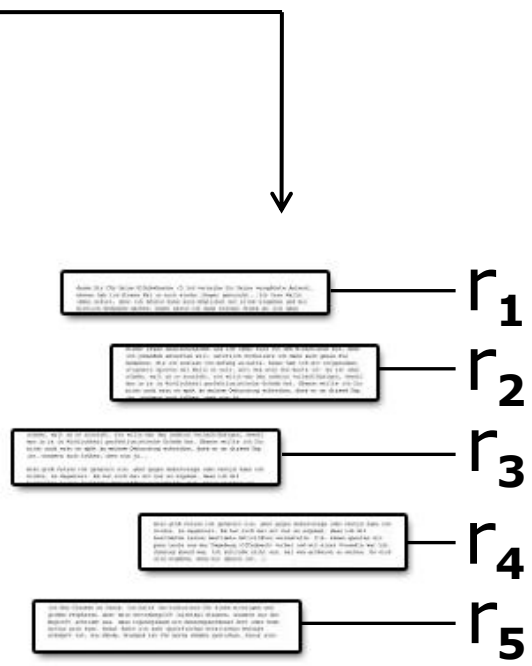


[2]

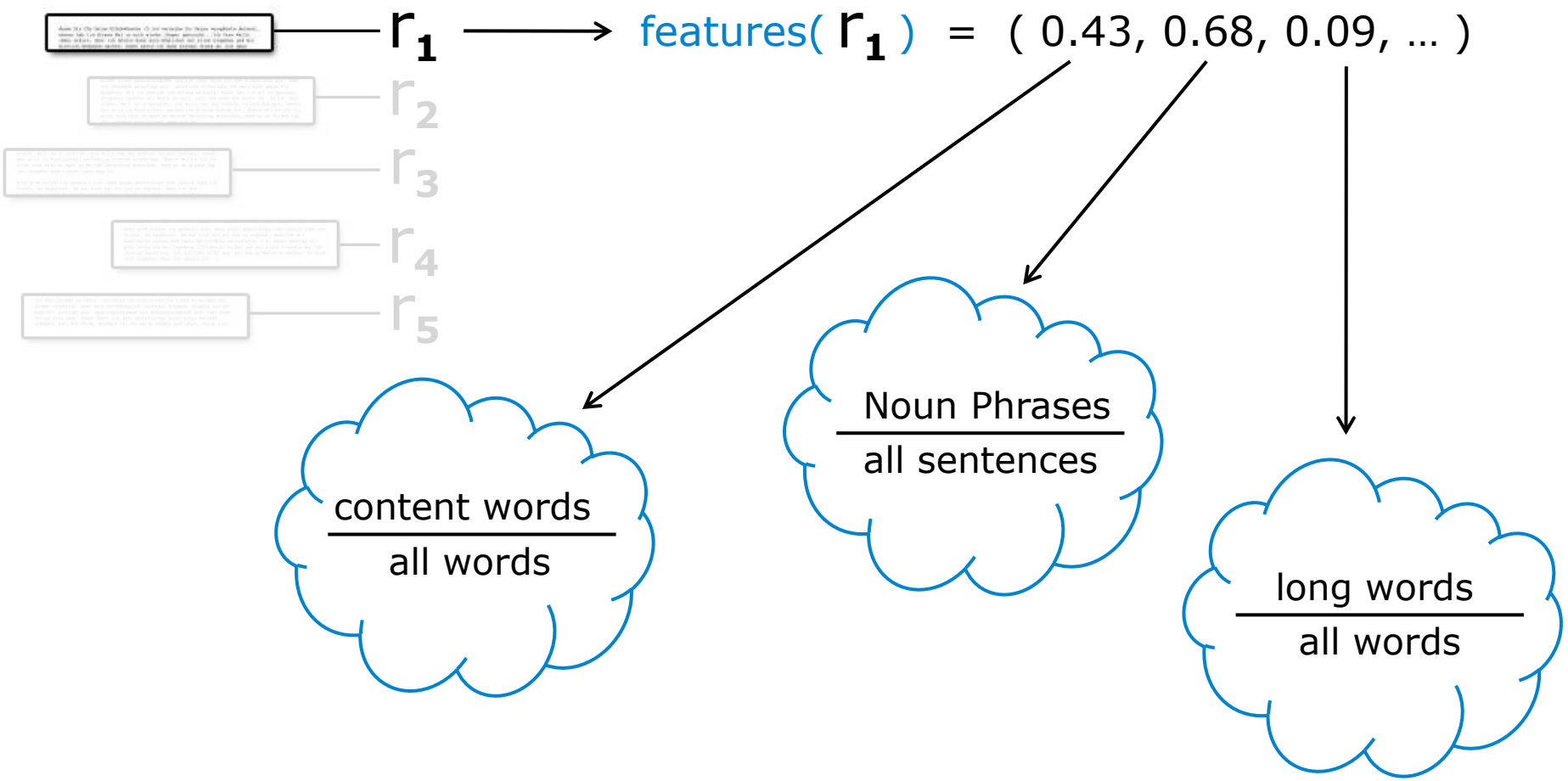
Feature Extraction

TECHNISCHE UNIVERSITÄT DARMSTADT

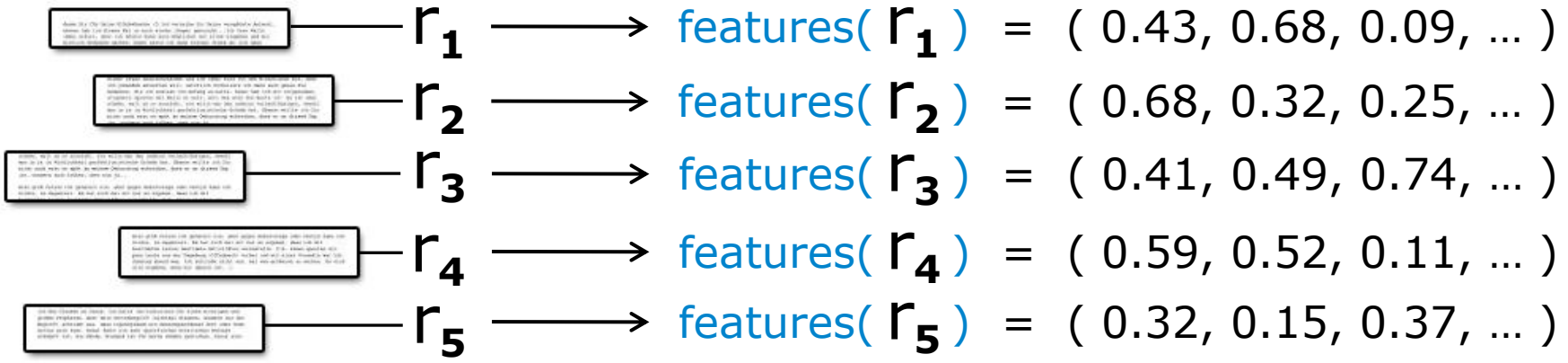
- Once we know which features are promising for the task, the next step would be to "get" them
- Solution: split document into non-overlapping regions
- Apply feature extraction features(r_i) on each region
- Example…



Feature Extraction



Feature Extraction



- Computing feature vectors is not enough - how should we detect outliers?
- Simple approach: Compute the document **mean** vector, which is the mean of all the feature vectors.
- Once we have the mean vector, we can start to measure similarities…

Similarity Functions

TECHNISCHE UNIVERSITÄT DARMSTADT

- Loose definition:

  - *Numerical measure of how alike two data objects are*
  - *Is higher when objects are more alike*
  - *Often falls in the range [0 ; 1]* [7]

- Many similarity functions can be used in IPD

  - Overlap Coefficient
  - Cosine Similarity
  - Bray–Curtis dissimilarity
  - Jaccard Index
  - Minkowski distance

Similarity Functions

TECHNISCHE UNIVERSITÄT DARMSTADT

- Example:

- Overlap Coefficient, defined by:

$$over(X, Y) = \frac{\sum_{i=0}^{n} min(x_i, y_i)}{min\left(\sum_{i=0}^{n} x_i, \sum_{i=0}^{n} y_i\right)}$$

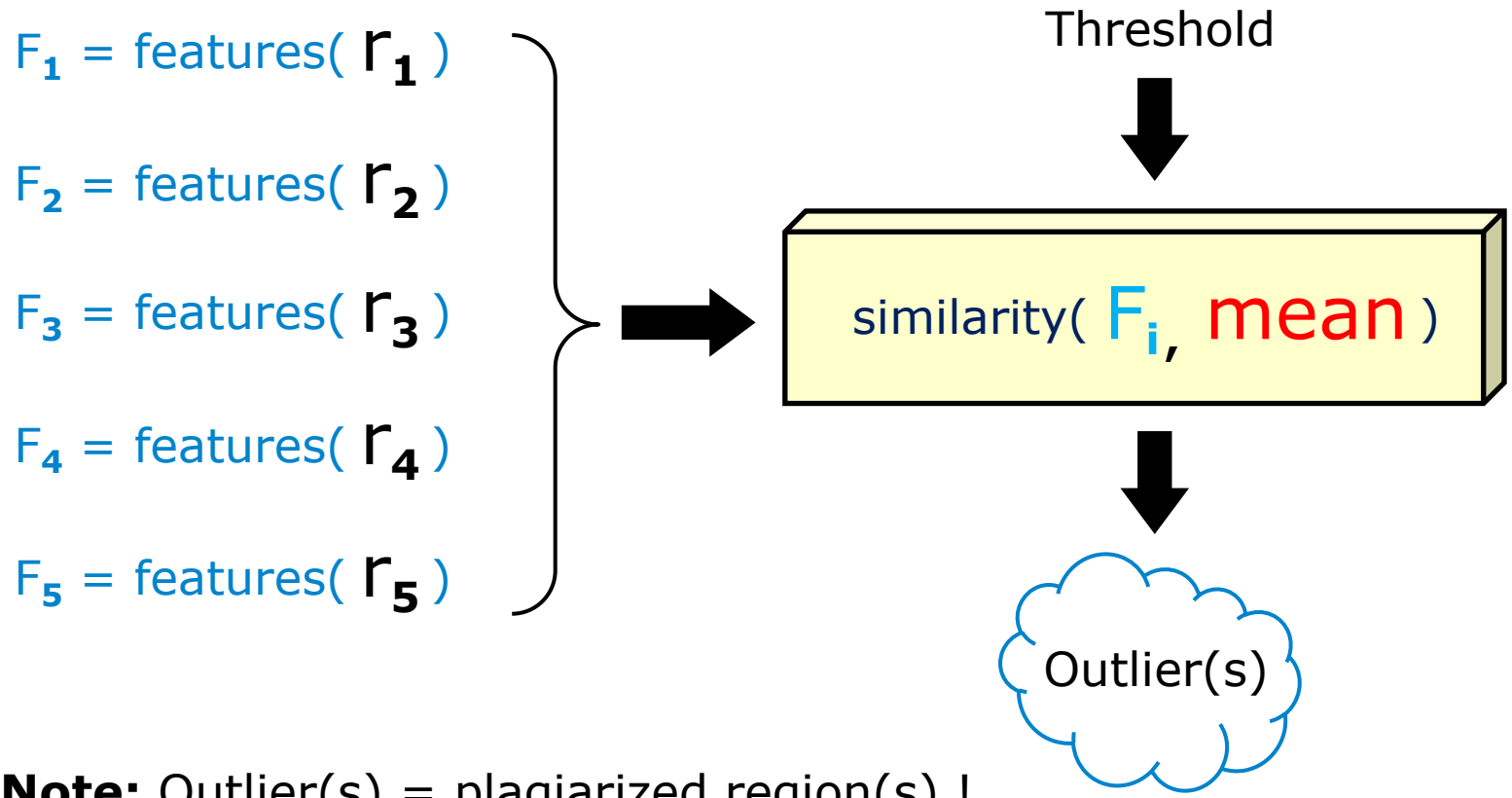
- Assume the following vectors are given:

  - X = (2, 2, 2)
  - Y = (7, 1, 0)

$$over(X, Y) = \frac{min(2,7) + min(2,1) + min(2,0)}{min(2+2+2, 7+1+0)}$$

$$over(X, Y) = \frac{2 + 1 + 0}{min(6, 8)} = \frac{3}{6} = \frac{1}{2}$$

Putting It All Together...



Note: Outlier(s) = plagiarized region(s) !

Take-Home-Messages

TECHNISCHE UNIVERSITÄT DARMSTADT

Intrinsic Plagiarism Detection:

…is an intensively discussed research discipline (especially since the last decade), but in comparision to EPD, comparatively a poorly investigated area.

…involves alot of complex techniques (in this talk: focus on the "basic idea")

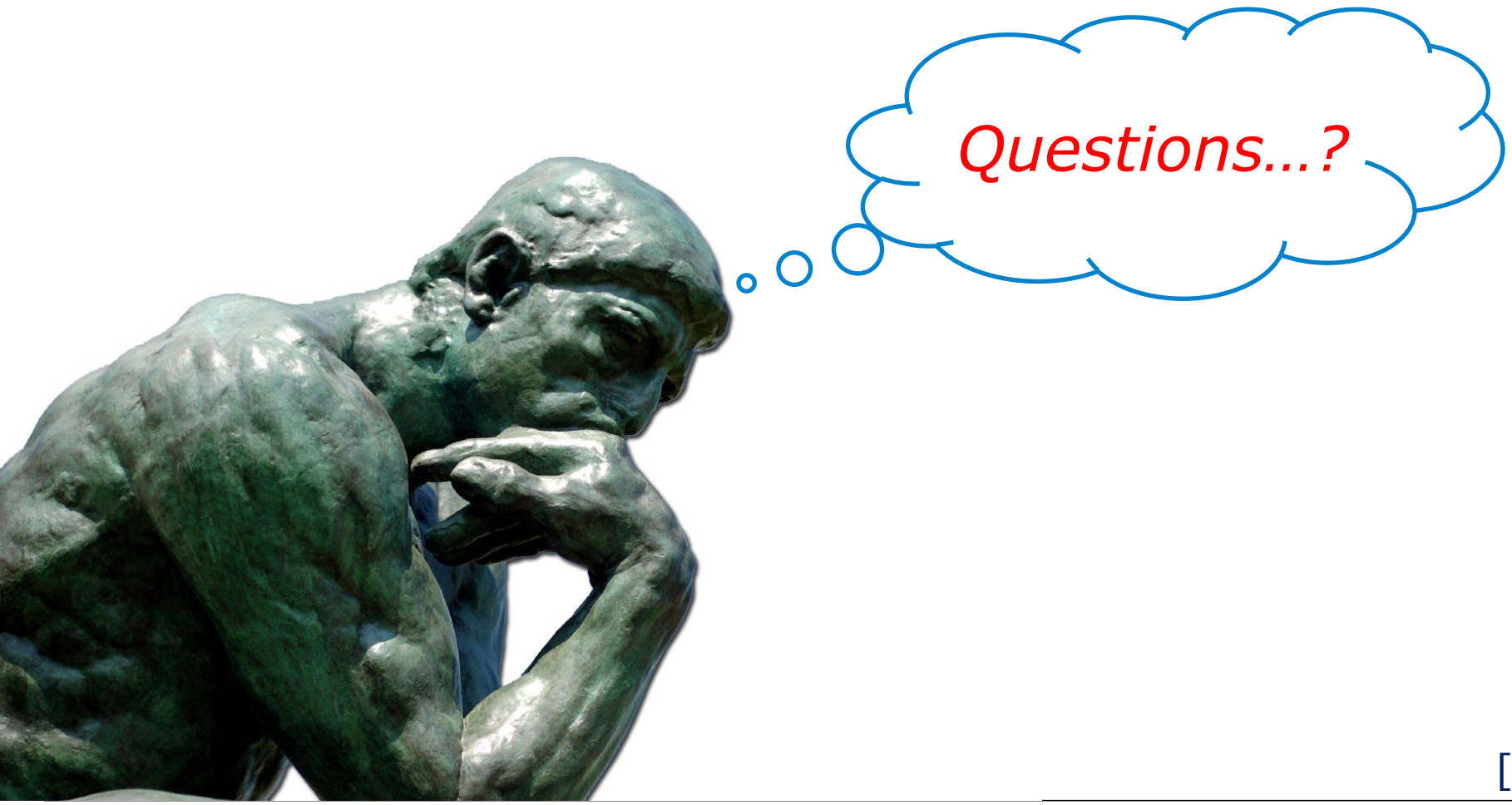
…requires a lot of creativity (and patience)

☺

Thanks for your attention !

TECHNISCHE UNIVERSITÄT DARMSTADT

[6]



[5]

References

[1] "**The Amount of Digital Information**",

www.infoniac.com/hi-tech/amount-digital-information-reached-281-exabytes.html

[2] "**Plagiarism Detection without Reference Collections**",

Sven Meyer zu Eissen, Benno Stein, and Marion Kulig,
Decker and Lenz (Eds.): Advances in Data Analysis
Selected Papers from the 30th Annual Conference of the German Classification Society (GfKl)
Berlin, ISBN 978-3-540-70980-0, pp. 359-366, c Springer 2007.

http://www.springerlink.com/content/w87475j2m5413220/

[3] "**Understanding Plagiarism**",

GENET 418/518 - Human genetics, Winter 2010 - Dr. Heather McDermid

www.biology.ualberta.ca/people/heather_mcdermid/genet418/UnderstandingPlagiarism.pdf

References

[4] "**Prof. Dr. Benno Stein**",

Picture taken from the Web Technology & Information Systems (Uni Weimar) website:

www.uni-weimar.de/cms/medien/webis/people/benno-stein.html

[5] "**Questions picture: photo of Le Penseur**",

A bronze sculpture made by Auguste Rodin, held in the Musée Rodin in Paris, France.

[6] "**Thank You**",

BRYAN DALTON / MISTAKE THE BEAUTIFUL. PRODUCES FREELANCE PHOTO-ILLUSTRATION, DESIGN AND ANIMATION IN PORTLAND, OR.

http://www.mistakethebeautiful.com/thankyou.html

References

[7] "**Introduction to Data Mining**",

Lecture Notes, by: Tan, Steinbach, Kumar

www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap2_data.pdf

[8] "**Current World Population and World Population Growth Since the Year One**",

www.geography.about.com/od/obtainpopulationdata/a/worldpopulation.htm

[9] "**Theodore Sider - Intrinsic Properties**",

Philosophical Studies 83 (1996): 1–27

www.tedsider.org/papers/intrinsic_properties.pdf

All external links have been accessed on: 30.10.2011 (some links updated…)