FACULTY OF ENGINEERING
AND ARCHITECTURE

Big Data Algorithms

Dieter De Witte (main contact), Jefrey Lijfijt, and Tijl De Bie

GHENT
UNIVERSITY

Outline part 1

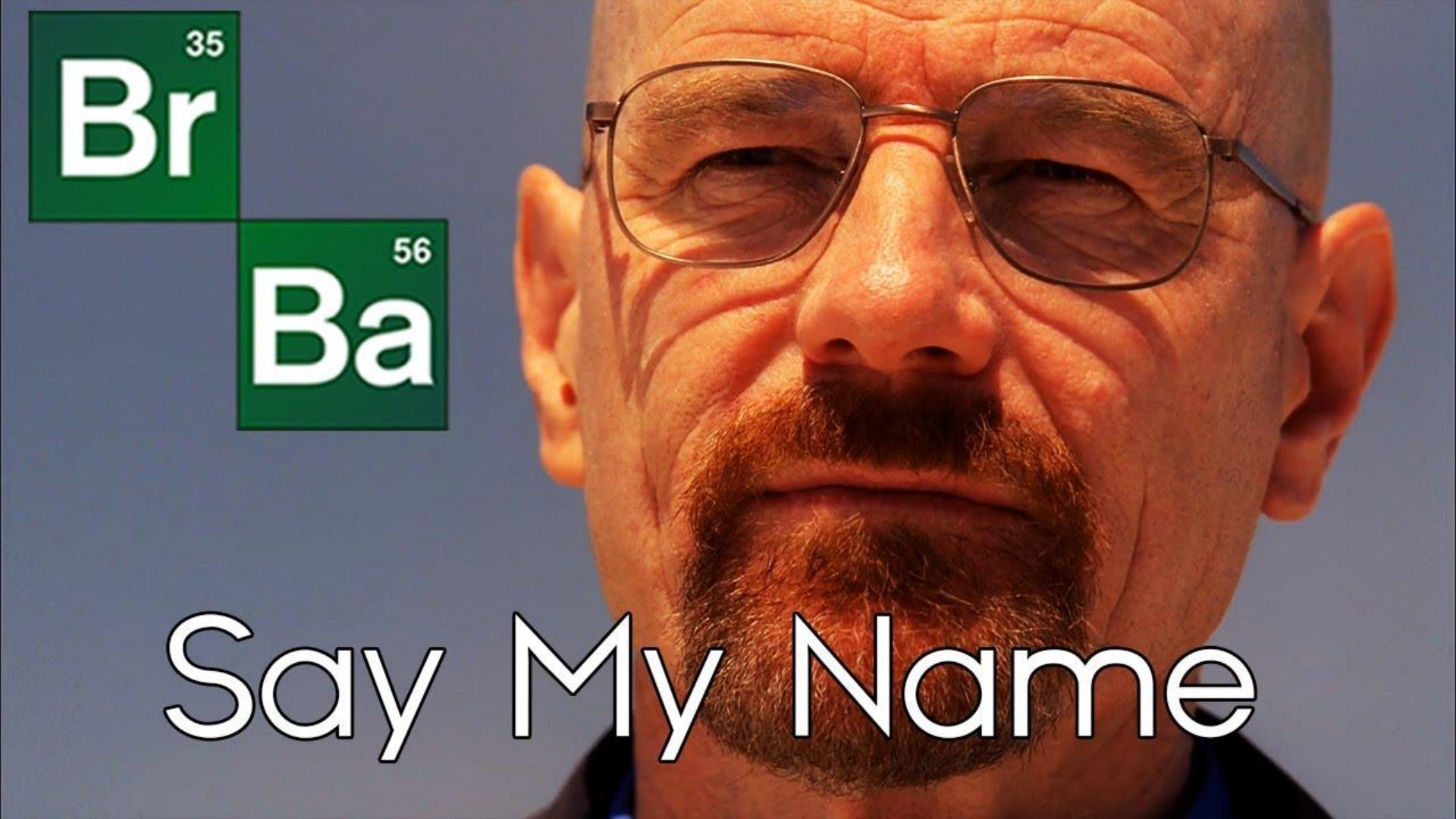
- Introducing the lecturers
- BDA Practicalities:
  - Theory
  - Exam + Project
- Inspirational Case

THE LECTURERS…

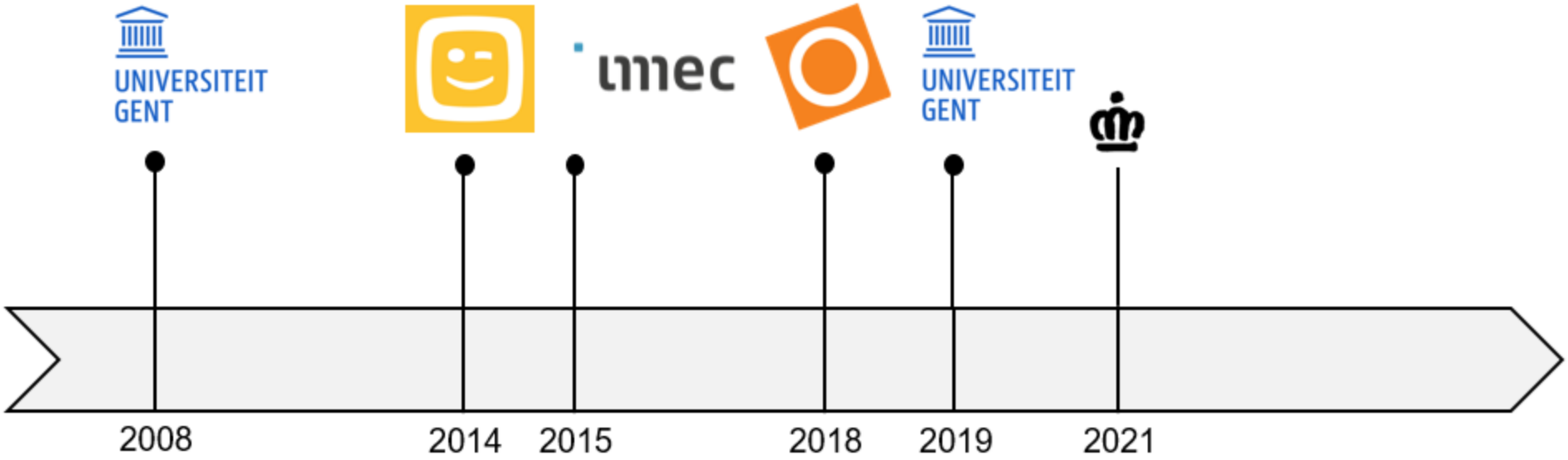
Introducing the lecturers: Dieter De Witte

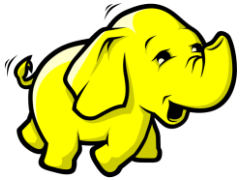
- **Personal:** FamilyOFive
- **Habitat**: Tielt, West-Flanders

- **Email**: dieter.dewitte@ugent.be


35
Br
56
Ba
Say My Name

CV


UNIVERSITEIT GENT
imec
UNIVERSITEIT GENT
2008
2014
2015
2018
2019
2021

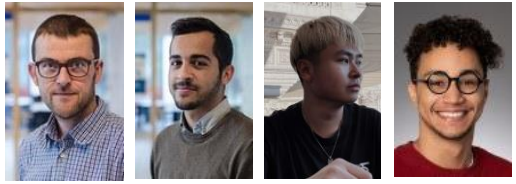



blazegraph™

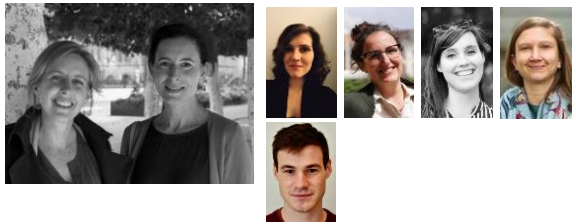

DMP
ONLINE

2 institutions, 2* #friends

Team G:



Team B:



Topics:

FAIR Data Pub
Provenance
CV: Enrichment
CV: Visual IR

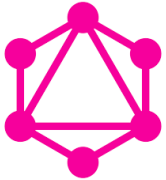
Heritage

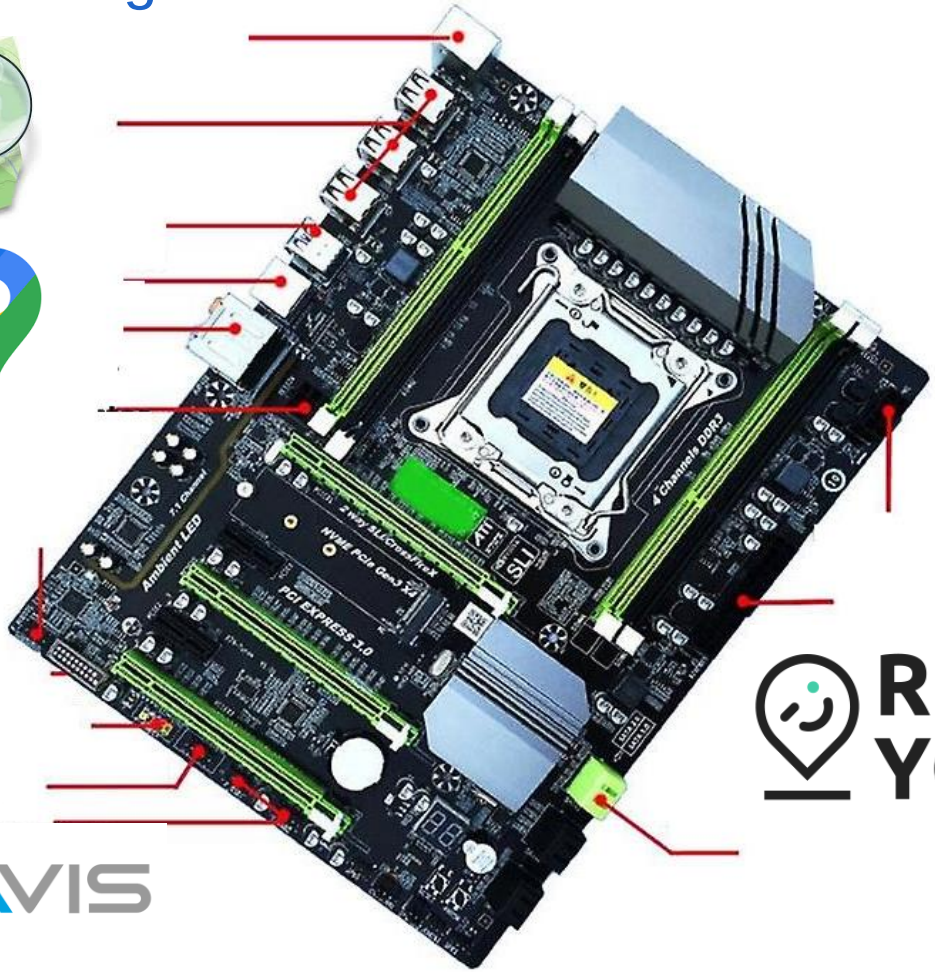
Media

Mental healthcare

Education

IMPACT with AI on heritage & mental healthcare





ROUTE YOU

LAVIS

Collectie van de Gentenaar

Technical Focus: Enrich & Interact

A valuable tool

FAME: Video face browser

centralized

privacy

Meemoo is a nonprofit organization funded by the Flemish government and focused on preserving Flemish history. Nico Verplancke, General Manager, explains: “We work together with about 160 media organizations in Flanders, including public broadcasters and regional broadcasters, as well as performing art centers and other heritage institutions, to digitize historical content. This can include everything from theatre performances to images of a strike to footage of a speech made in the 1940s. It’s a very diverse set of material.”

It’s also a very large set of material. The meemoo archive already contains about 19 PB of data, and that number grows by approximately 2 PB each year. Meemoo makes the material available to the citizens of Flanders for a variety of uses, collaborating with cultural, heritage and media organizations to carry out projects designed to help preserve and celebrate the region’s rich history. One of meemoo’s flagship projects is The Archive for Education.

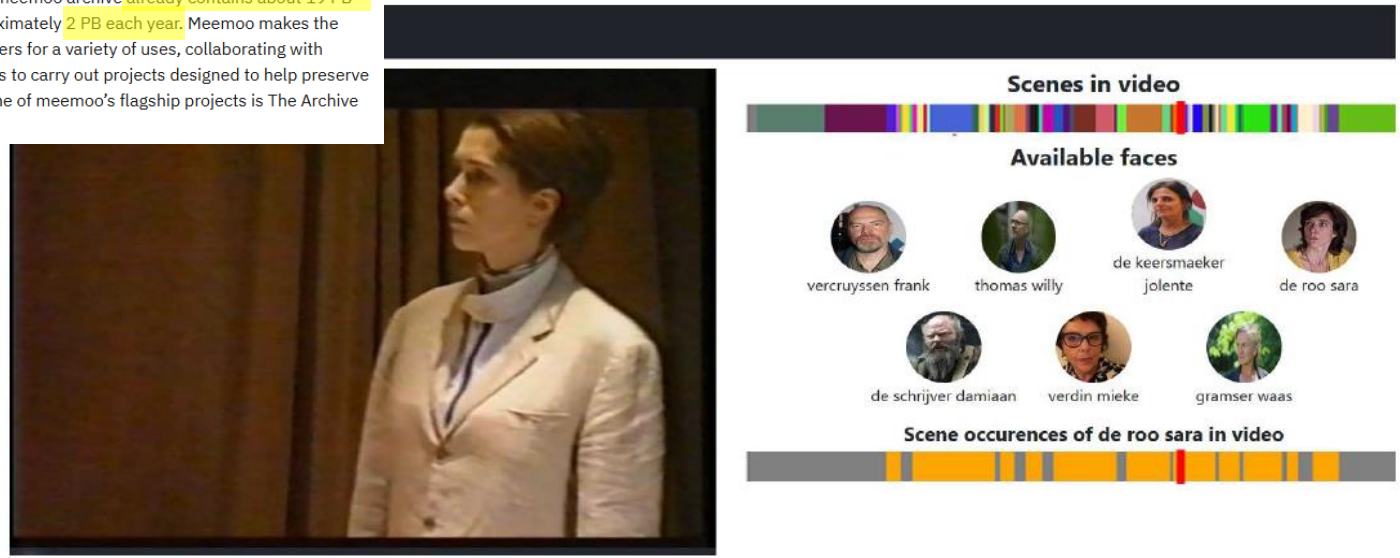


Figure 1 - Timeline based FAME video browser. For each actor, all shots he/she appears in are highlighted on the timeline. By clicking on the timeline, the video browser starts playing the video at that particular location.

Data Federation

Meta Belgica

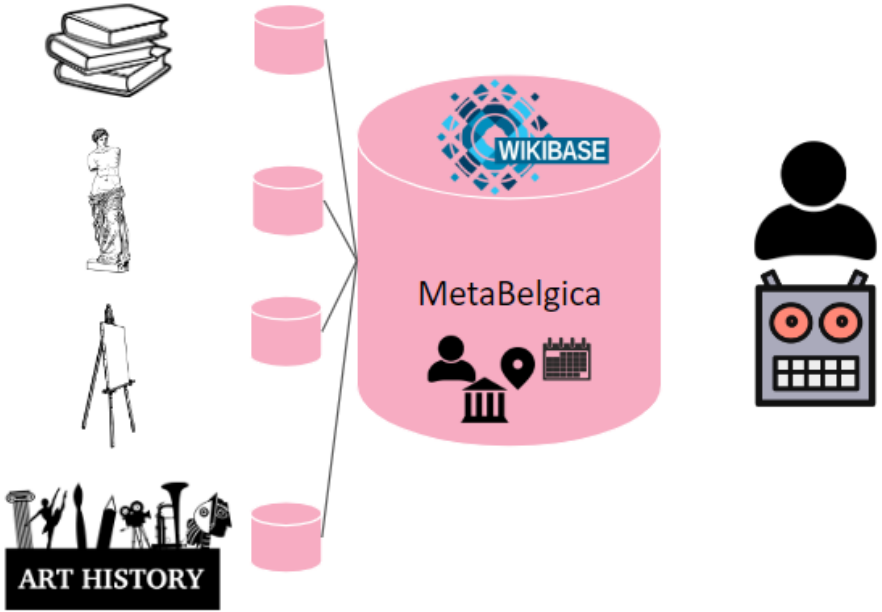
decentralized

KBR

IRPA KIK Royal Institute for Cultural Heritage

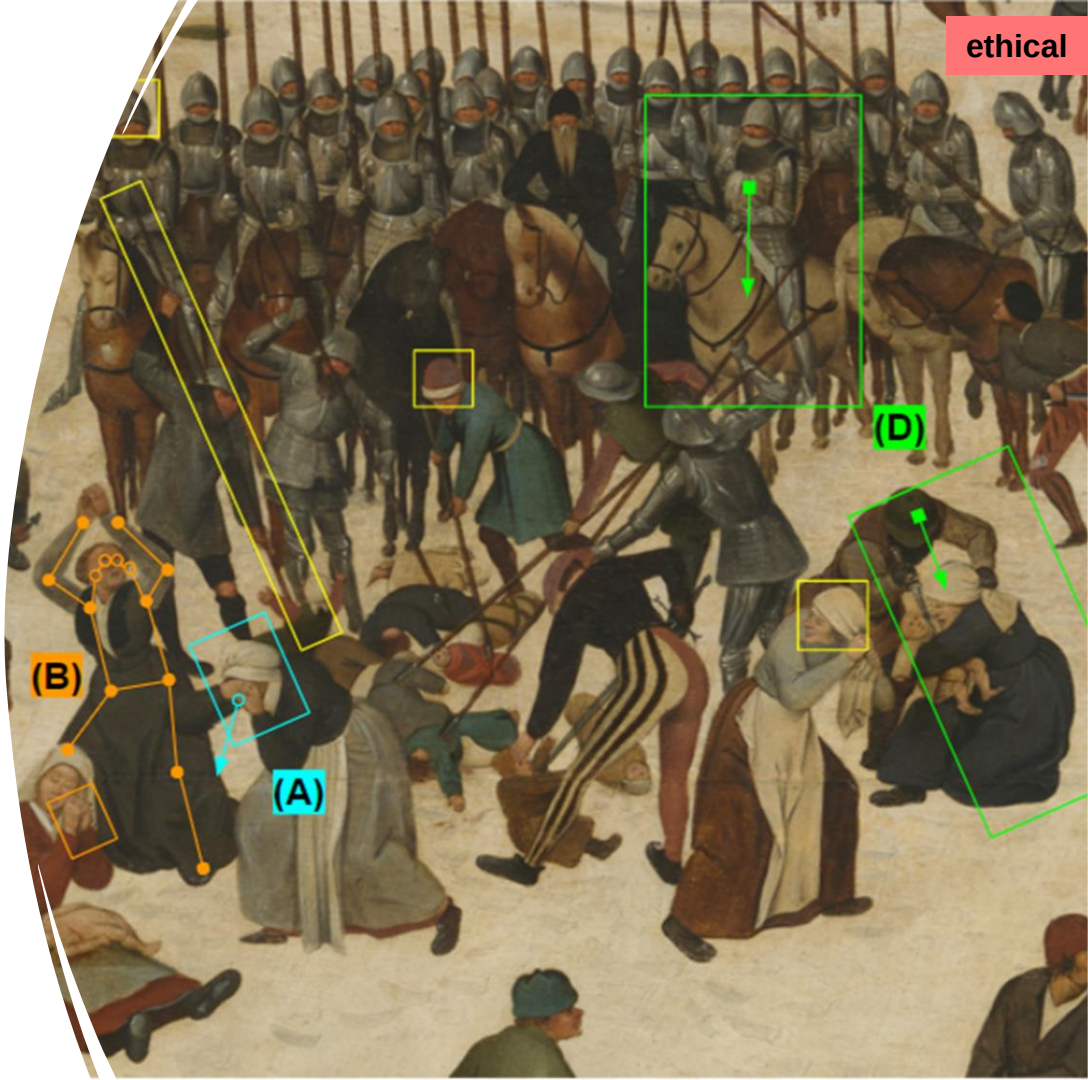
Royal Museums of Fine Arts of Belgium

ROYAL MUSEUMS OF ART AND HISTORY
KONINKLIJKE MUSEA VOOR KUNST EN GESCHIEDENIS
MUSÉES ROYAUX D'ART ET D'HISTOIRE



The **museum** as a safespace for sensitive AI research

FRIDA project works with citizen scientists to try and measure power structures, representation of minorities, etc. through time via the Royal Museum collection that spans 7 centuries.



KunstContact: "Networks of art to connect people"

similarity

graphs

"orchestra"

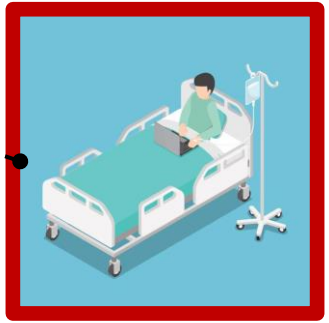
Marc Chagall,
Blue violinist

Vincent Van Gogh, *viooltjes*

Alfred Stevens,
The Violinist

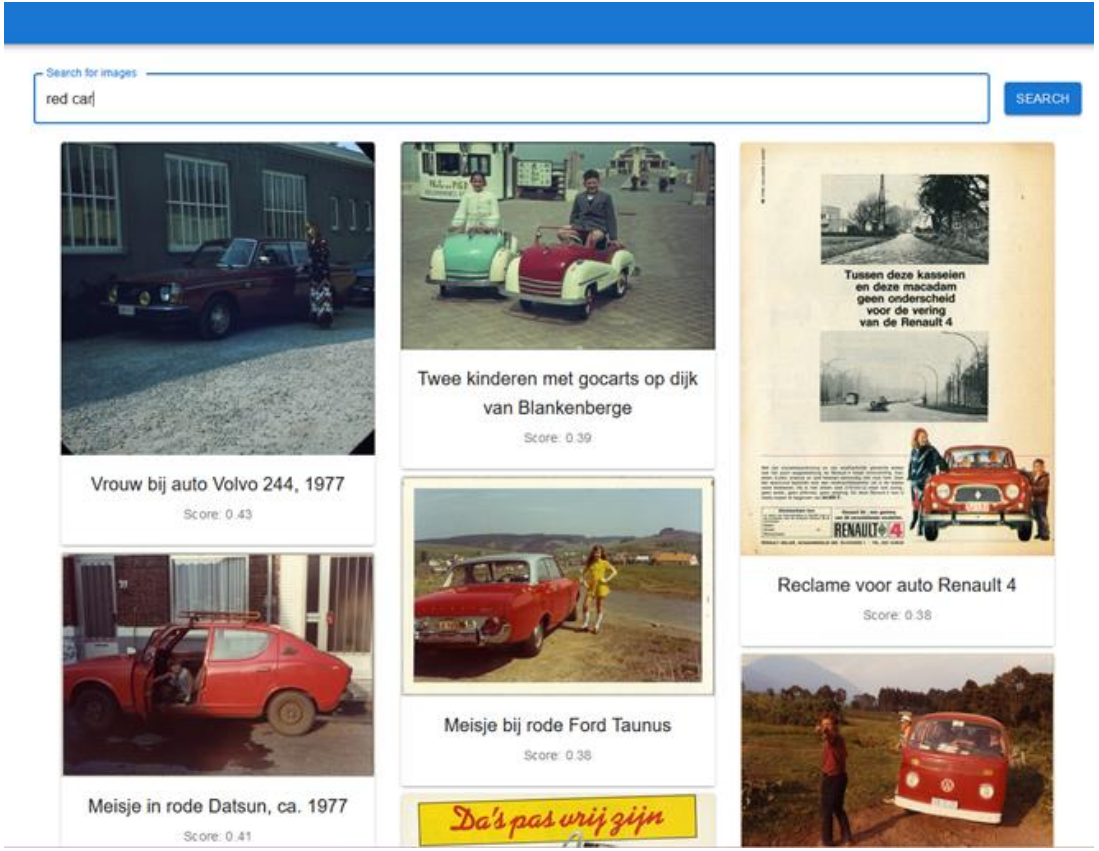
Jan Davidsz.
de Heem

"garden"



Searching big collections without metadata!

similarity



similiarity

(text query, multimodal embedding

embeddings

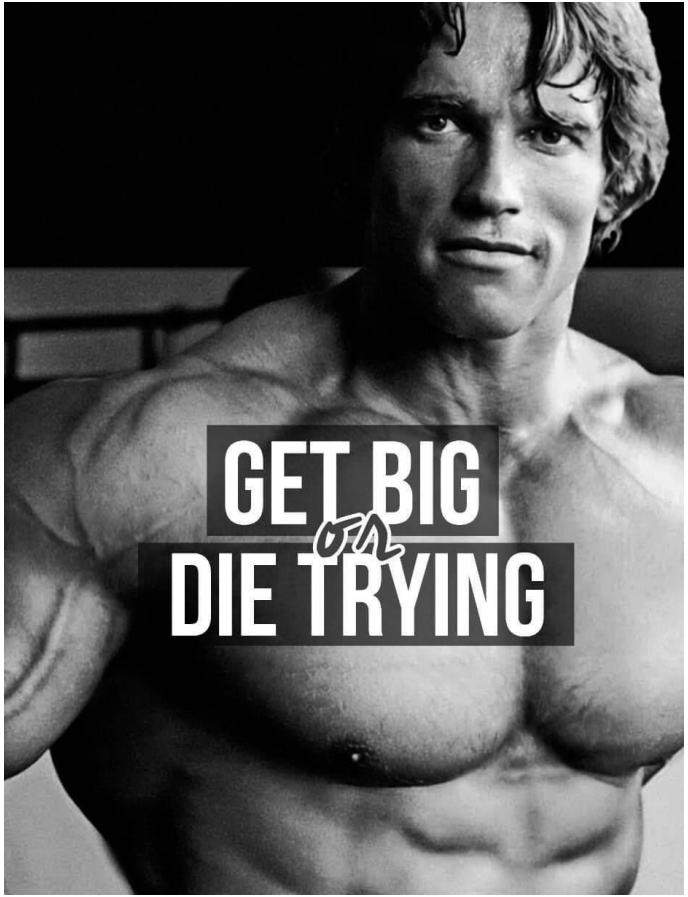
are stored in a vector database

=> indexing!?

THEORY

“Intelligence is not what we know, it’s what we do when we don’t know.”
Jean Piaget

Big Data is 'Tough' Data!!!



- 'Big' Data has always been a misfit ~ marketing term
- Big Data is data that is **'tough' to handle** and requires a nonstandard approach (!= SQL)
- In this course tough means:
  1. **Large**: central or decentralized
  2. **Ethics** & Privacy
  3. Connected data (graphs)
  4. Tough to **navigate** or **query**

Course agenda

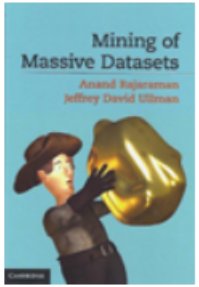
w1	11/02	Intro - meet the lecturers	DJT
w2	18/02	Data Mining	J1
w3	25/02	Data Mining / Decentralized	J/D
w4	04/03	**Free**	
w5	11/03	Ethical & Privacy-preserving AI	T1
w6	18/03	**Free**	
w7	25/03	**Free**	
w8	01/04	Graph Algorithms	J2
	EASTER		
w9	22/04	Graph Embeddings	T2
w10	29/04	Sketching & Hashing	D2
w11	06/05	**Free (Deadline project video)**	
w12	13/05	**PROJECT**	DJT

Course notes

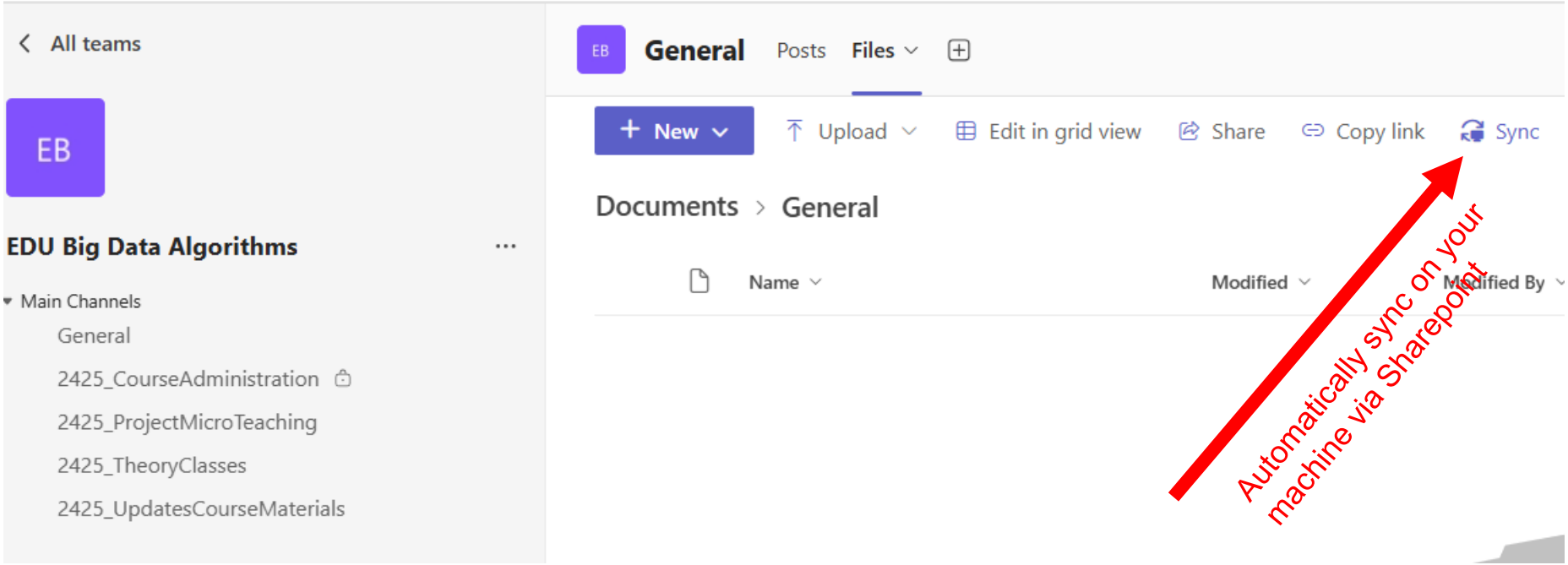
- mmds.org
- annotated slides

The 3rd edition of the book

Chapter	Title	Book
	Preface and Table of Contents	PDF
Chapter 1	Data Mining	PDF
Chapter 2	Map-Reduce and the New Software Stack	PDF
Chapter 3	Finding Similar Items	PDF
Chapter 4	Mining Data Streams	PDF
Chapter 5	Link Analysis	PDF
Chapter 6	Frequent Itemsets	PDF
Chapter 7	Clustering	PDF
Chapter 8	Advertising on the Web	PDF
Chapter 9	Recommendation Systems	PDF
Chapter 10	Mining Social-Network Graphs	PDF
Chapter 11	Dimensionality Reduction	PDF
Chapter 12	Large-Scale Machine Learning	PDF



One location for all up-to-date course files



> UpdatesCourseMat. => every update to slides,... will be mentioned here
> Other two channels are for Q&A and also for discussion between students
Note: Ufora doesn't sync so this is a means to make sure everyone has the latest version of the course without monitoring Ufora all the time.

Homework

- To get a bit of a feel where BDA can be applied prepare one use case with: (we will take some time to discuss these next week)

1. What major Belgian companies have Big Data clusters?
2. What use cases only work when learning is decentralized?
3. The Belgian company Ontoforce uses federated querying in its biomedial engine, can you explain how?
4. How do you remove bias from a CV? For example how can you guarantee that gender has no impact on the hiring?
5. How does LinkedIn decide whos content to show on your timeline?
6. How to recommend jobs in a way that it re-balances the job market (Yoosof Mashayekhi)
7. How does Spotify find similar songs in a huge collection instantaneously while you are navigating?
8. Apple counts emoji's with sketching!?

Exam & Project

Exam marks

- Theory: 10/20
  - Oral exam Live (!)
  - Closed book
  - Example questions (see later)

- Microteaching project: 10/20
  - 50% Microteaching video
  - 50% Hands-on (papers-with-code approach)

- No Labs!

Project

- form groups of 3 (use teams channel + ask for help if needed)

1. Within the scope of the course look for *'related'* topics, research papers that extend the course
2. submit a half-page abstract with a proposition of what you'd like to work on before 15/3 (topic + coding)
3. lecturers will approve / reject / give small suggestions for course changes.

Project

- **By 6/5** 23:59 submit a lecture video via Ufora (mp4 format)
  - a mini-lecture about your project of ~ 10 minutes
  - we will compile these into a playlist
  - everyone reviews all material until the next lecture
- **On 13/5** you give a 3 minute recap
  - your peers will prepare prepare one **question each**
- **On 13/5** you provide us with a small github repo
  - with a clear **readme.md** where we can review your **coding**

Micro-teaching

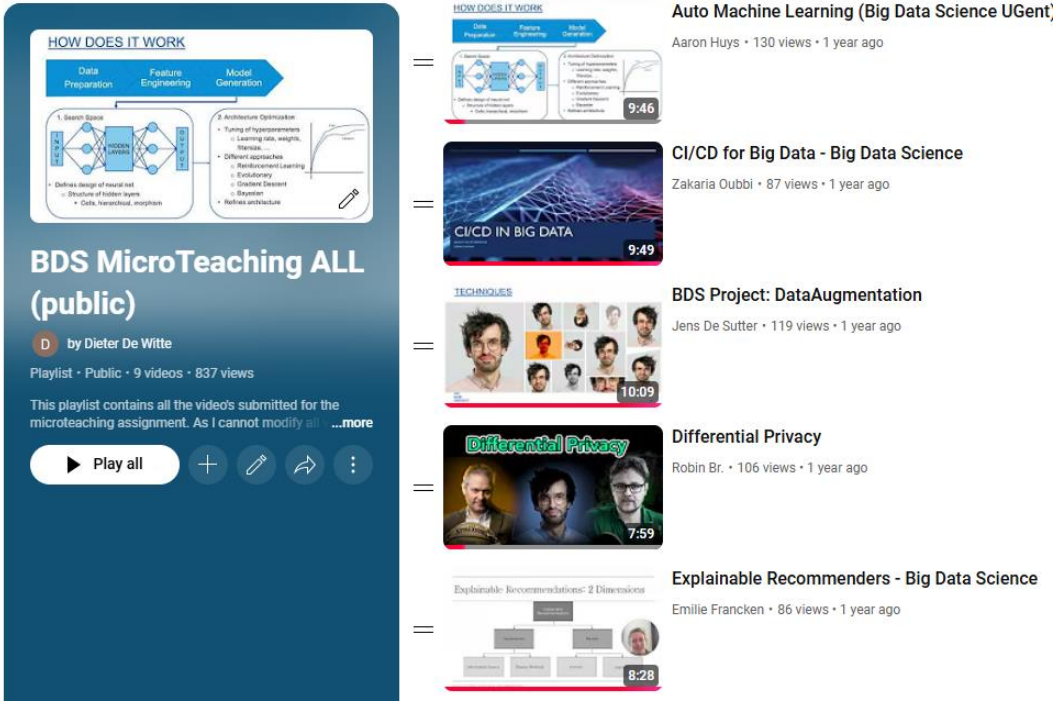
- The target audience of your lectures are your peers, take into account their background!

- The lecture video's /slides you create **will be part of the course material**s (**oral exam**!)

- Students / Lecturers will prepare questions for the last lecture on 13/05 (flipped classroom)

Similar project in course BDT with data engineering topics

https://youtube.com/playlist?list=PLFs4BmAq7GvHf0adWBp3TXEhz2MF1Czjk&si=IKaK6cKb33AMYr_7



Main difference?

> topics are more closely tied with the course

> focus is more on SotA

Coding? Paperswithcode philosophy

graph2vec: Learning Distributed Representations of Graphs

17 Jul 2017 · Annamalai Narayanan, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, Shantanu Jaiswal · Edit social preview

Recent works on representation learning for graph structured data predominantly focus on learning distributed representations of graph substructures such as nodes and subgraphs. However, many graph analytics tasks such as graph classification and clustering require representing entire graphs as fixed length feature vectors. While the aforementioned approaches are naturally unequipped to learn such representations, graph kernels remain as the most effective way of obtaining them. However, these graph kernels use handcrafted features (e.g., shortest paths, graphlets, etc.) and hence are hampered by problems such as poor generalization.

PDF Abstract

provide your own implementation, document and publish!

Code

Edit

benedekrozemberczki/karateclub	★ 2,190	
benedekrozemberczki/graph2vec	★ 908	TensorFlow
MLDroid/graph2vec_tf	★ 155	TensorFlow
paulmorio/geo2dr	★ 45	PyTorch
compnet/pang	★ 12	TensorFlow

See all 6 implementations

Tasks

Clustering
General Classification
Graph Classification
Graph Embedding
Graph Matching
Representation Learning

How much effort should I spend on this assignment?

- 3 SP ~ 90 hours
  - 50% of the marks ~ 45 hours pp x 3
- Gaps to work on this: no morning lectures + multiple weeks without lectures