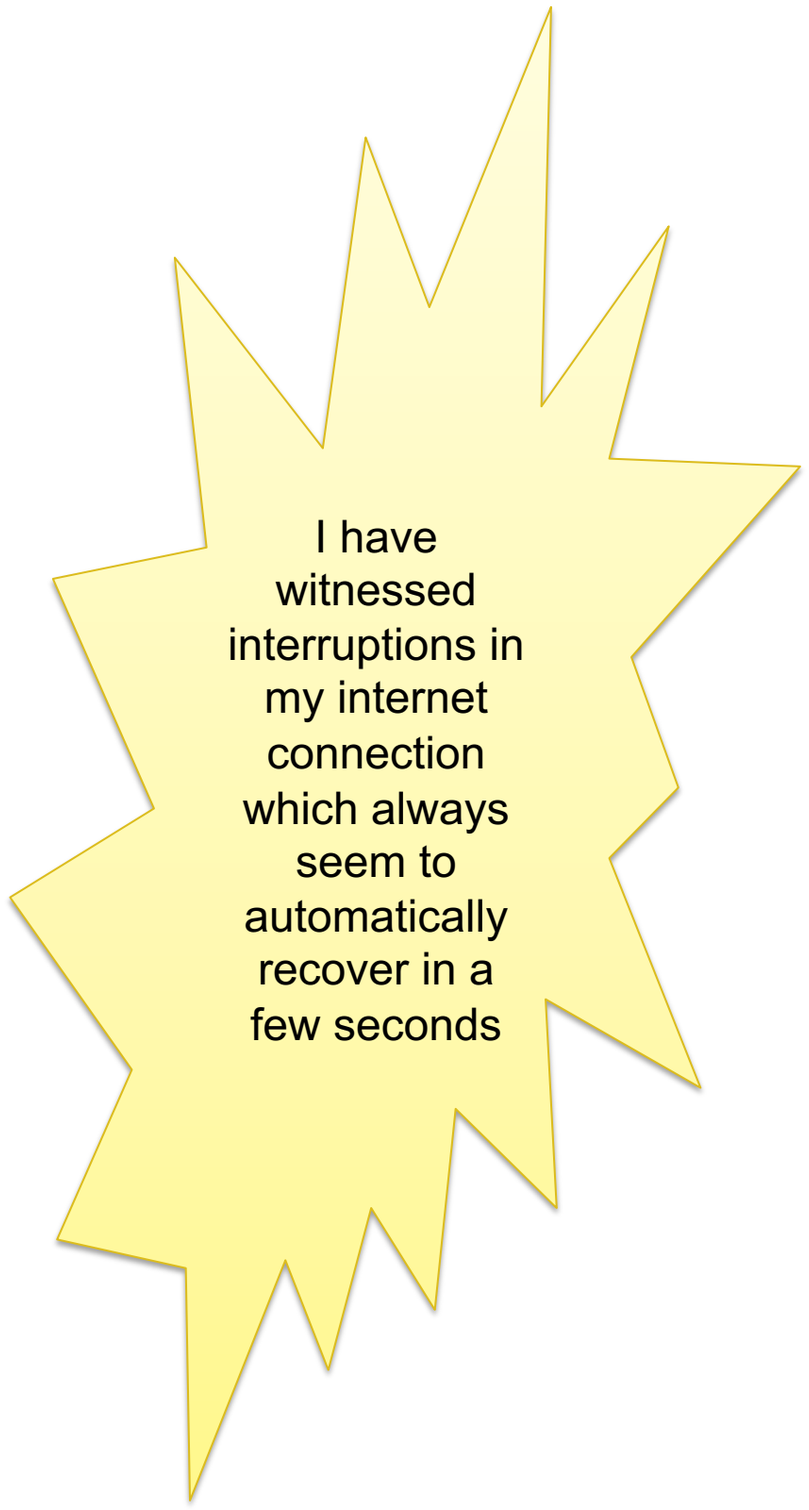
UNIVERSITY OF TWENTE.

Introduction

Master Course Data Science

2022/23 1B

Maurice van Keulen
Nov, 2022



[slide contributions: C. Seifert, M. v. Keulen]

Teaching Staff

Course Coordinators

Karin Oudshoorn
Assistant Professor
Coordinator Q1

Maurice van Keulen
Associate Professor
Coordinator Q2

Faizan Ahmed
Assistant Professor
Coordinator Q3

Other topic teachers

- Ellen-Wien Augustijn
- Nacir Bouali
- Faiza Bukhsh
- Rolf de By
- Elena Mocanu
- Estefania Talavera
- Brenda Voorthuis
- Shenghui Wang

TAs (Q2)

- Tifani Zata Lini
- Fleur Veenman
- Raef Kazi
- Radhika Kapoor
- Ajay Joseph
- Jan Menzel
- Rakshitha Vijay Kumar
- Ruben Hessels
- David Pröschold

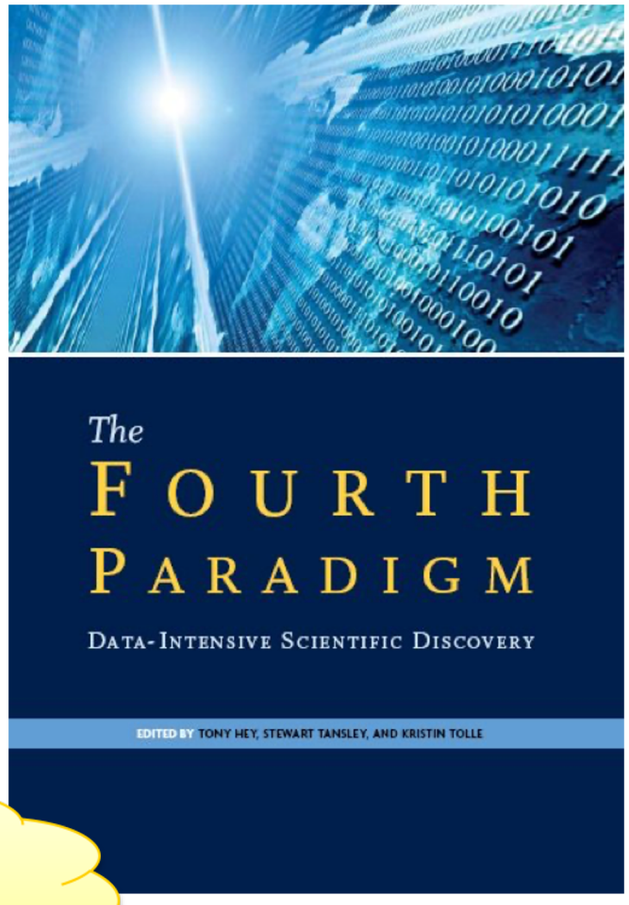
- Priyadharshini Shanmuganathan
- Lilian van Oosterhout
- Merit Fernhout
- Willem Vincent
- Carlota Trigo
- Sneha Borkar
- Gayathri Dhanapal
- Raj Ashokan
- Cindy Pistorius

Data Science

Revolution of Scientific Method

Paradigms of scientific method

1. Empiricism (observe the flying apple)
2. Mathematical modelling (exact formulas for the trajectory)
3. Simulation (people behaviour on concerts for crowd control measures)



Bio-Informatics professor:
" PhD of 4 years, 3 years devoted to 'data fiddling' "

A new paradigm: Data-intensive Scientific Discovery

4. Combining and analysing data in novel ways is capable of tackling research questions that could not be answered before

Data Science

PACS / EHR

Web

Publications

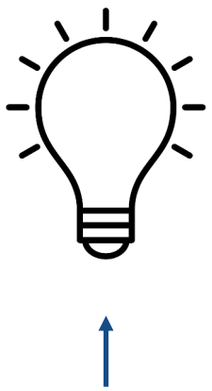
Linked Data

Gene Sequences

Sensor Data

Data

- Interdisciplinary
- Scientific and economic progress
- Exploration



- Big Data
- Visualization
- Data Mining (Deep Learning)

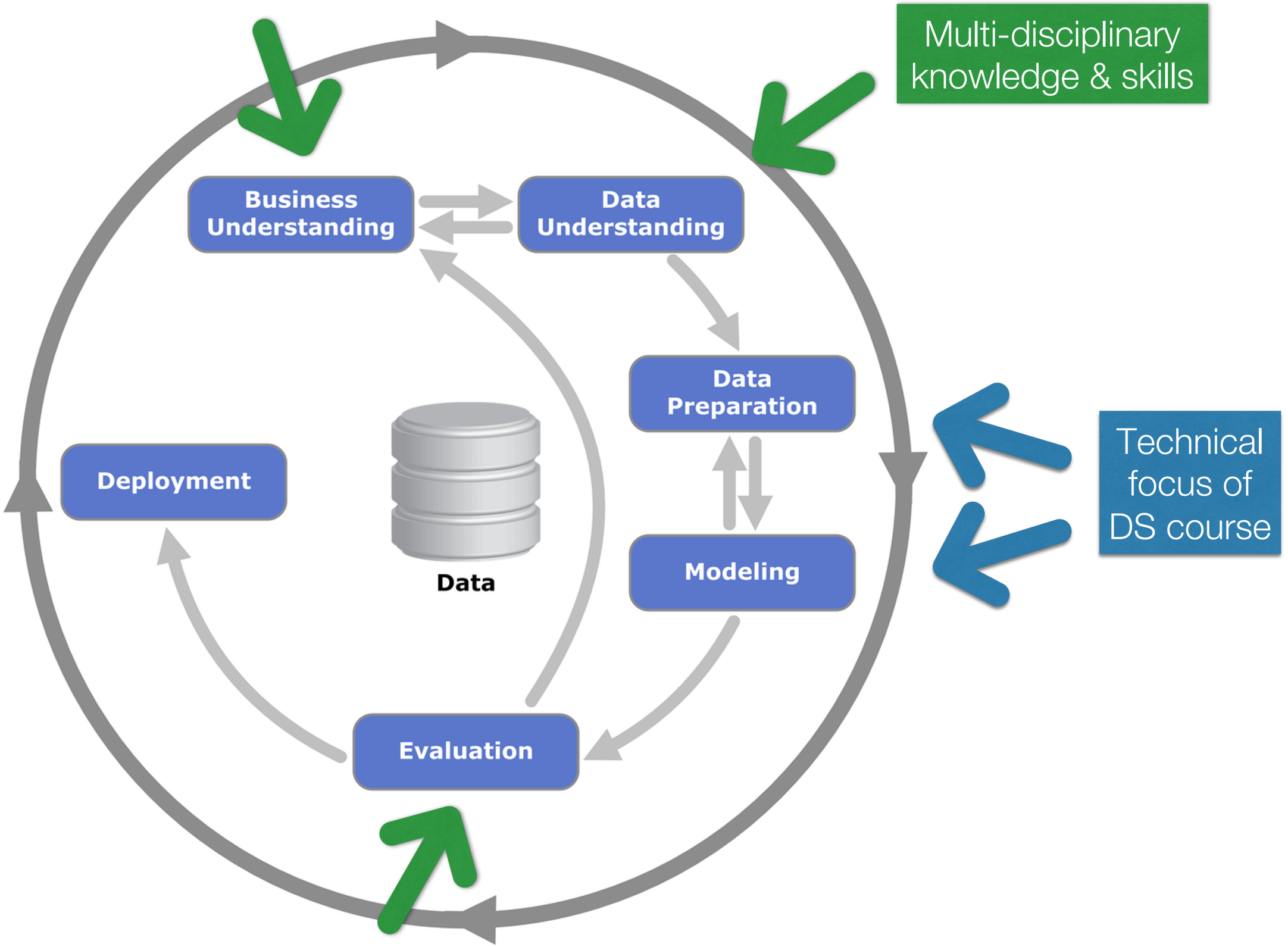
Big data

The 4 V's of big data

- Volume (big)
- Velocity (fast)
- Variety (diverse)
- Veracity (with quality issues)

- Not always all 4, usually a combination of V's

CRISP cycle


Multi-disciplinary knowledge & skills
Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data
Technical focus of DS course

Organisation

Topics and Projects

- 2x Topic (educational unit)
  - Purpose: Learn and practice, hands-on
  - Form: 1 Lecture + self-study assignments => sign off
  - Choose: 2 topics

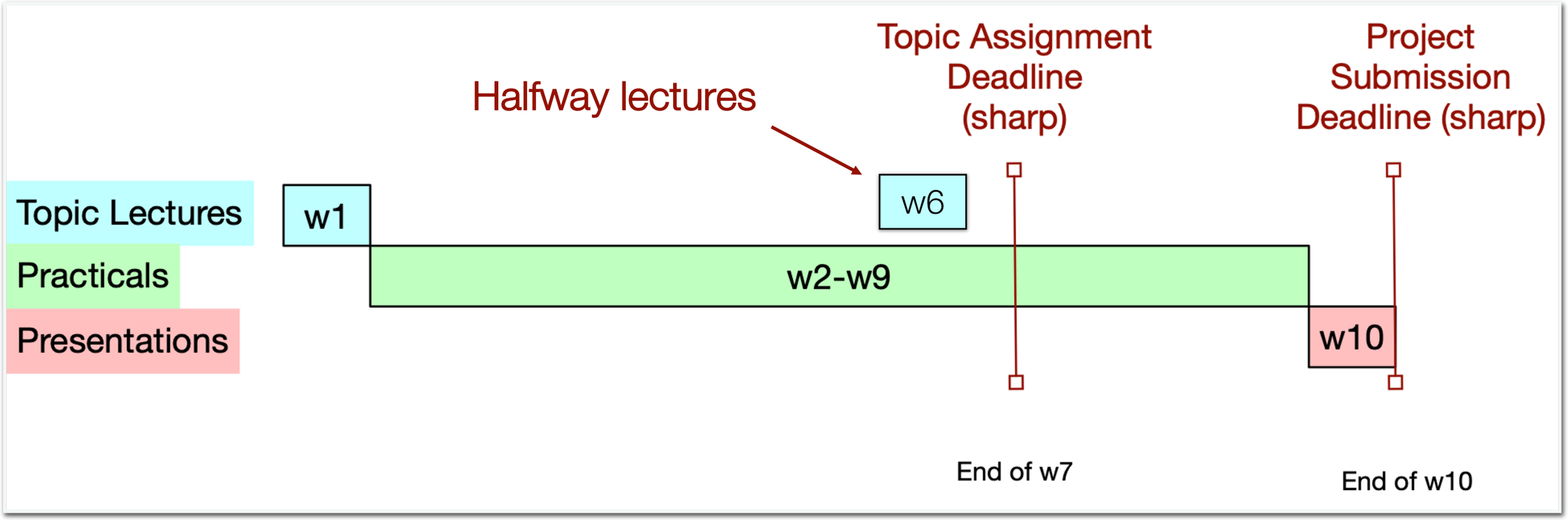
pass/fail

- 1x Project
  - Purpose: Deepening and assessment
  - Form: open project → report + presentation => grade
  - Choose: 1 project

grade for course
=
grade for project

- Requirement: Apply primary topic in project
  Optional: Also applying secondary topic in project is a plus

Timeline



- Note: Topic assignments can be submitted later,
- BUT:
  - If topic assignment deadline is not met, we can’t guarantee grading in the grading period of this quarter (this is to balance workload for TAs and teachers)
  - You run into the danger in not being able to complete the project (empirical evidence from recent years)

Time Management

- 5 EC course = 5x 28 = 140 hours
- 5 Lectures: 1 Intro, 2 Topic, 1 Global, 1 Presentation = 10 hr
- 130 hr left for both topics and project
- 16 practical sessions = 32 hr supervised
- 130 – 32 = ±100 unsupervised hours
- Outside of scheduled ours, we expect you to invest approx. 100 hours on the practical assignments and the project!
- Guideline: Of the 8 weeks of practical sessions, a normal schedule has 5 weeks for assignments and 3 weeks for project

Prerequisites

Programming skills

- SEMI (SQL, Java)
- IENLP (Python)
- PDBDQ (any)
- TS (basic Python)
- DPV, DM (R or Python)
- DINT (Python)
- CV&IP (Python)
- **New: GIS** (Python)

Probability theory

- DM

Working with tools / software

- DPV visualisation, PM

Assignments have one primary programming language. But projects can be solved in any language.

Mandatory / Optional

Mandatory

- Submission ONLY through Canvas
- Topic assignments signed off for both topics
- Presentation of project
- Project deliverables submitted to Canvas

important to ensure proper grading

Optional

- Attendance at lectures and practical sessions (topic lectures and introduction recorded on video)
- Extra ‘side-path’ or ‘deepening’ on topic

Project assessment

Project assessed by two teachers:

- results are averaged
- if grade disagreement is bigger than 1 grade, there is a discussion between the two and they have to both agree on the grade

Assessment criteria (see Assessment Form on Canvas)

- [20%] Communication (presentation / report)
- [40%] Technical depth
  - Proper application of technology of main topic
  - Depth: going beyond what is taught in the topic(s)
  - Understanding
- [40%] Method
  - Proper interpretation of results
  - Interesting insights
  - Critical attitude towards source data and results
  - Clear goal, proper goal oriented methodology and priorities / choices with argumentation
  - Relevance of the steps taken

Learning goals

After completing the course, students:

- have **knowledge and understanding** of various data science skills for preparing different kinds of raw data and for analysing data,
- are able to properly **apply** these data science skills in a **real-world** project,
- are able to make proper **methodical decisions** in a real-world project: taking steps with relevancy and justified priority, showing a **critical attitude** towards data and results, and properly interpreting data science results,
- are able to properly **communicate** the results of a real-world project both orally as well as in written form,
- and have the ability and **attitude for continued learning** in data science techniques and methods.

Project do’s and don’ts

- Project = **Real-world** data & Real-world challenge
- **Show** that you can apply, communicate, and have the right attitudes (continued learning, critical attitude)

Do’s and don’ts

- Choose a relevant & interesting (for you) **focus**
- Attempt to **go beyond** what has been taught
  - Find & apply other techniques, methods
  - Use literature
- Other tools and programming languages allowed

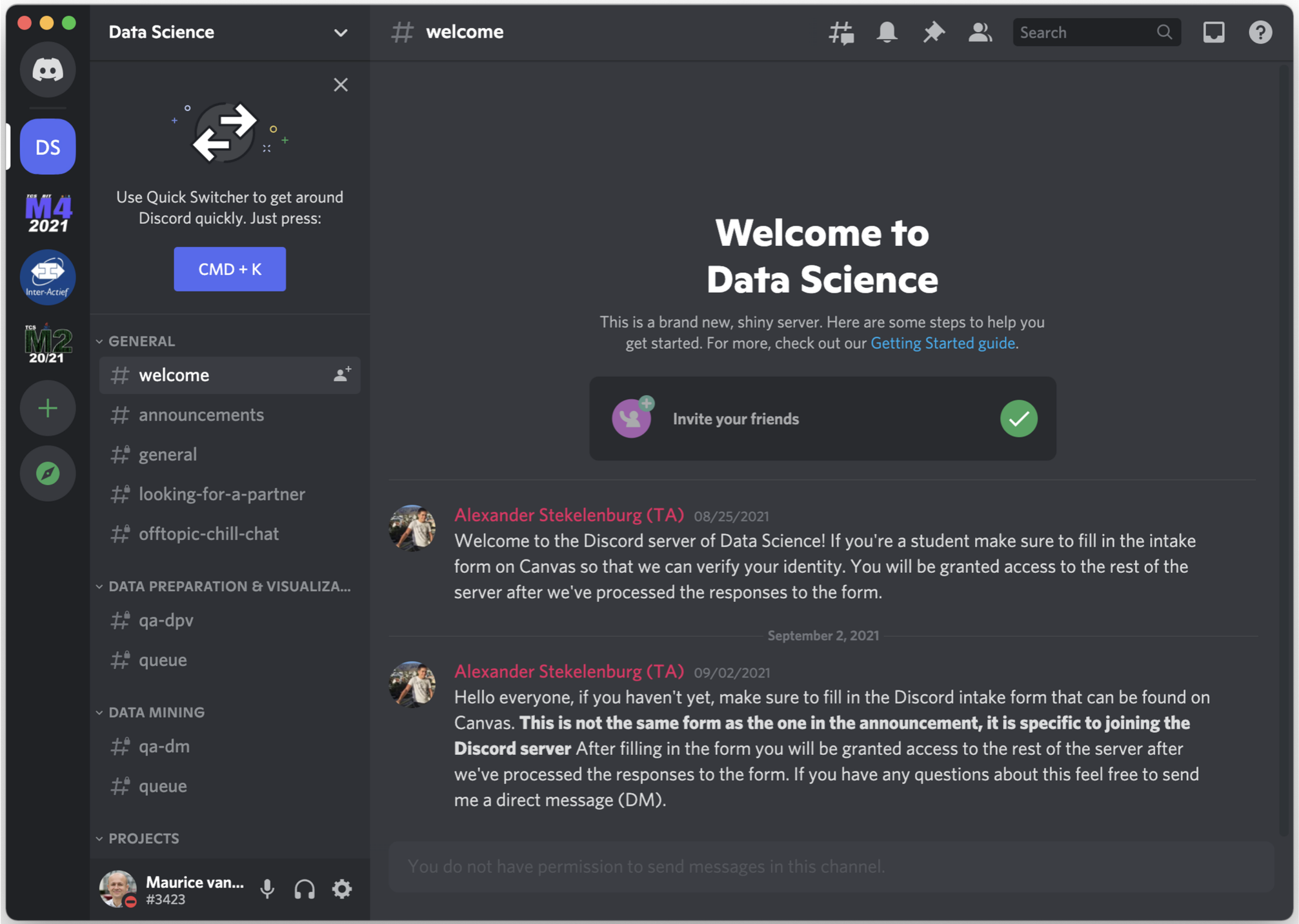
Supervision and Discussion

In practical sessions

- Supervision and guidance is ’mostly’ limited to practical sessions
- Supervision by main teachers: Maurice van Keulen, Karin Oudshoorn, Faizan Ahmed
  - and sporadically by other topic teachers
  - and teaching assistants (typically students who did Data Science before)
- At least one person present at all sessions
- First few sessions with more … we’ll see how many are needed

Discussion Tool: Discord

powerful team communication service



Supervision and Discussion

Discord during practicals & outside scheduled hours

- Sign-up: registration link (see Syllabus)
- Website + native tablet & smartphone apps
- While working, join “table” (on-campus & on-line)
- Question?: post in appropriate #qa-channel (topic / prj)
  Or during practical: post table+location in #queue
  then wait for TA/teacher to respond
  on-campus: raise hand
  on-line TA/teacher 'moves' to table
- Also outside scheduled hours … but slower responses
  Help each other by asking and answering questions
  (you learn a lot)

Hybrid (COVID) arrangements

Project rooms on-campus

Hybrid =
on-line is a must, on-campus is an enriching extra

- Practical sessions most important for face-to-face
  - **2 queues on Discord: fairness on-campus / on-line**
- Lectures: mostly on-line (broadcasted + recorded)
- Presentations: choose on-campus or on-line session

Topics

Topics

First quarter only DPV & DM offered

- Topics
  - Data Preparation & Visualization (DPV)
  - Data Mining (DM)
  - Information extraction and Natural Language Processing (IENLP)
  - Feature Extraction from Time Series data (TS)
  - Semi-structured data (SEMI)
  - Probabilistic Databases & Data Quality (PDBDQ)
  - Process Mining (PM)
  - Data Integration (DINT)
  - Computer Vision and Image Processing (CV&IP)
  - New: "Spatial data" (GIS)

Students of Industrial Engineering and Management [**IEM**], Health Science [**HS**],Communication Science (**COM**), and Business Administration (**BA**) are required to do topics **DPV** and **DM**

M-TM-MSS
Master Technical Medicine specialization MSS are required to do topics **TS** and **DM**

M-BIT
(Master Business Information Technology) are required to select 2 out of 3 topics **DPV, DM** and **PM** for the CORE of the program (if taken for first time)

Course “Data Science Additional Topics”: Do DS a scond time with 2 different topics & 1 different project

Topic 0 - Programming in R or Python

If you don't have (much) programming experience, then it is advised to also do the optional topic "zero"

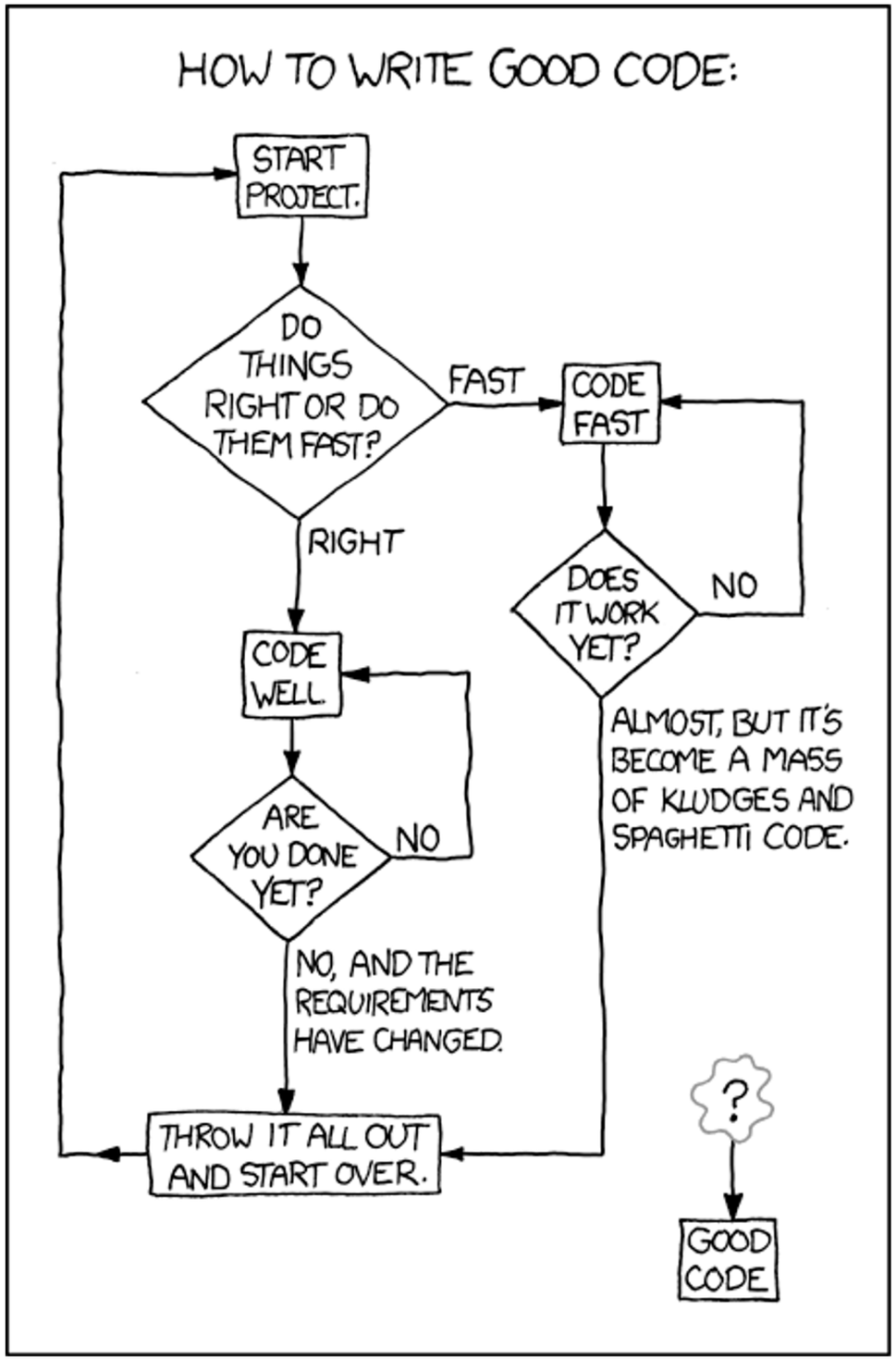
Main contents:

- The Programming Environment
- Variables and Control Flow
- Basic Data Structures
- Functions
- Libraries for Data Processing, Visualization and Manipulation
- How to solve errors / bugs
- Introduction to programming specifically geared towards programming for data science

If you do not join, then you have no right to complain about programming

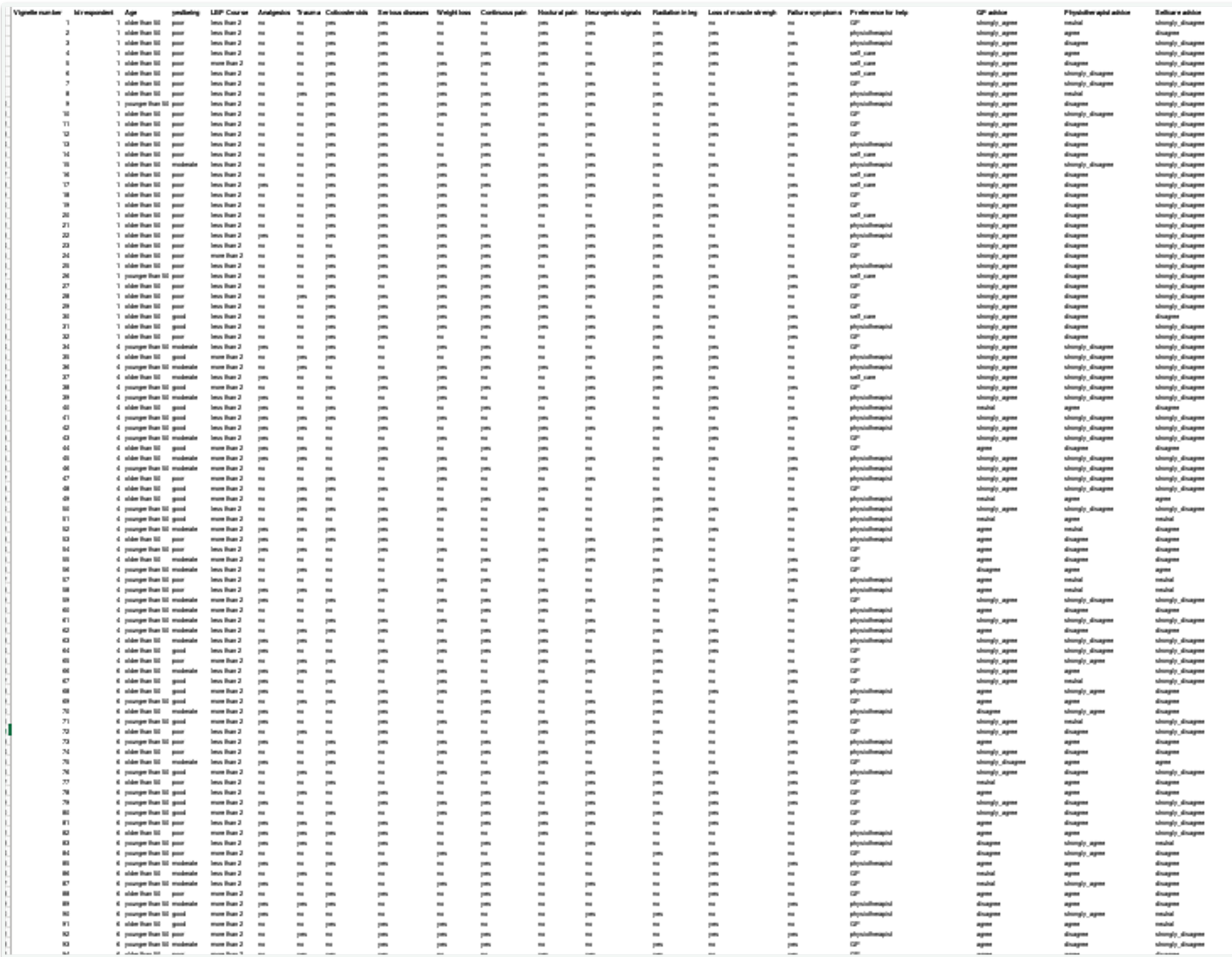
About Programming

- There are many ways to solve a problem (course guide gives hint with ONE solution in mind)
- Programming one line might require 3 hours of thinking
- First conceptualise, then write code (getting the idea on paper first, helps)
- Check all intermediate steps (light-weight debugging)



Topic Data Prep. and Vis. (DPV)

- Just by “looking at” data we can’t see anything

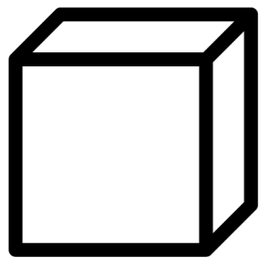


Topic Data Prep. and Vis. (DPV)

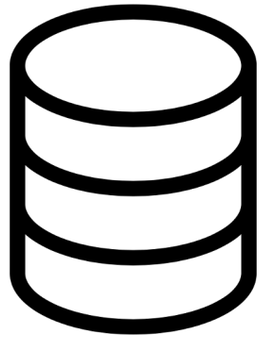
sources

transform

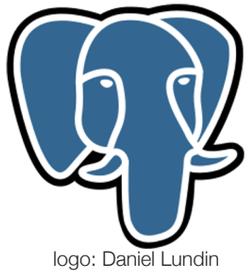
data cube



stored in

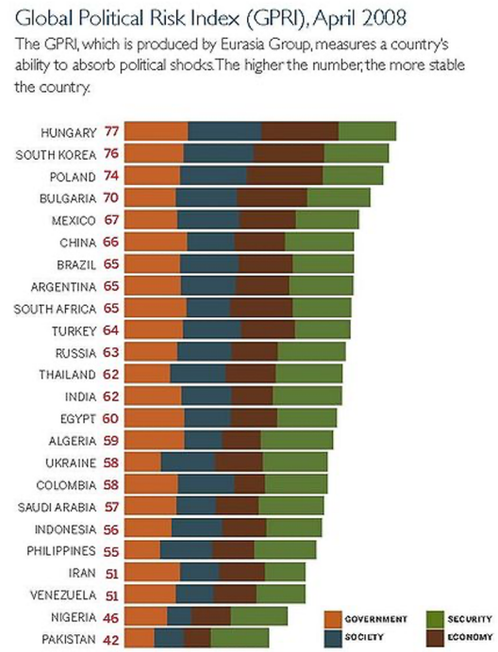


DBMS



logo: Daniel Lundin
wikimedia commons

accesses



via Wikimedia commons
https://commons.wikimedia.org/wiki/File:GPRI.JPG

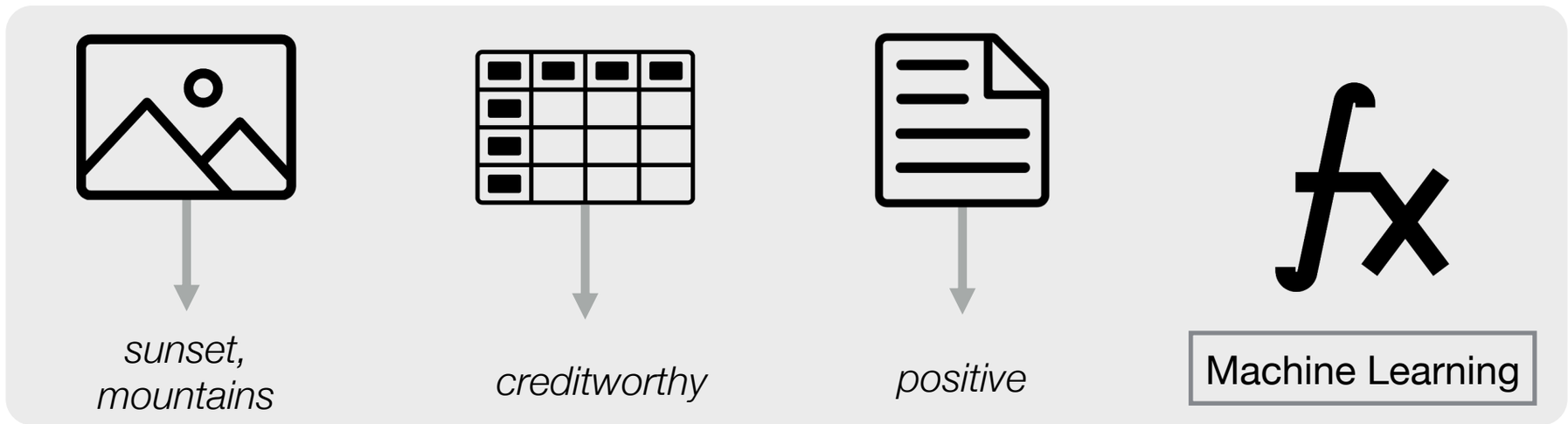
Design method for DPV

The proper way to come from data to insights
(based on insights from business intelligence)

1. Formulate “Questions to data”
2. Imagine visualisations / reports
3. Design star schema(s) for cube(s)
   by analysing (1) and (2) for fact(s) and dimensions
4. Create (empty) database with schema
5. Fill database by transforming sources
6. Realise visualisations / reports by connecting to the database

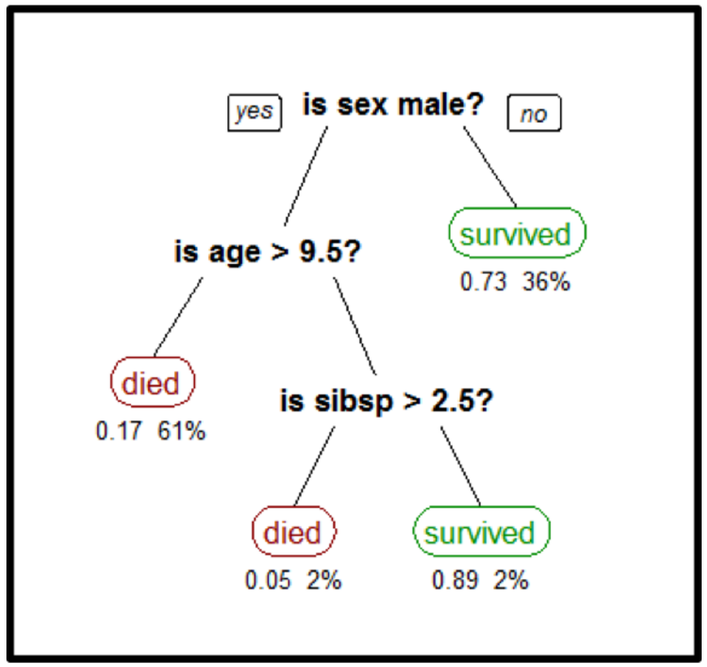
Topic Data Mining (DM)

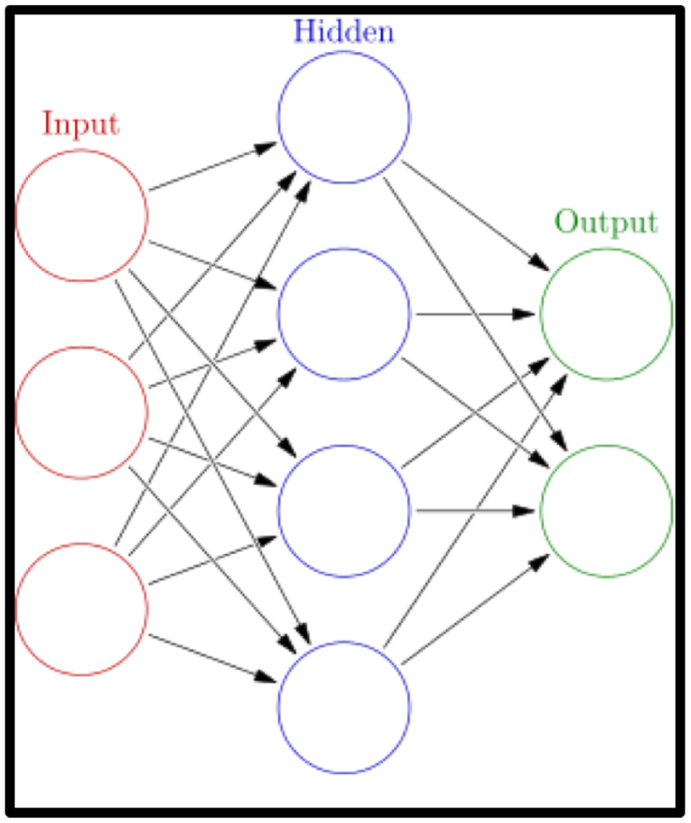
- Techniques to automatically extract knowledge from data (by hand is simply not feasible anymore.
- Supervised techniques: learn a target function by examples
- Unsupervised techniques: find “obvious” patterns

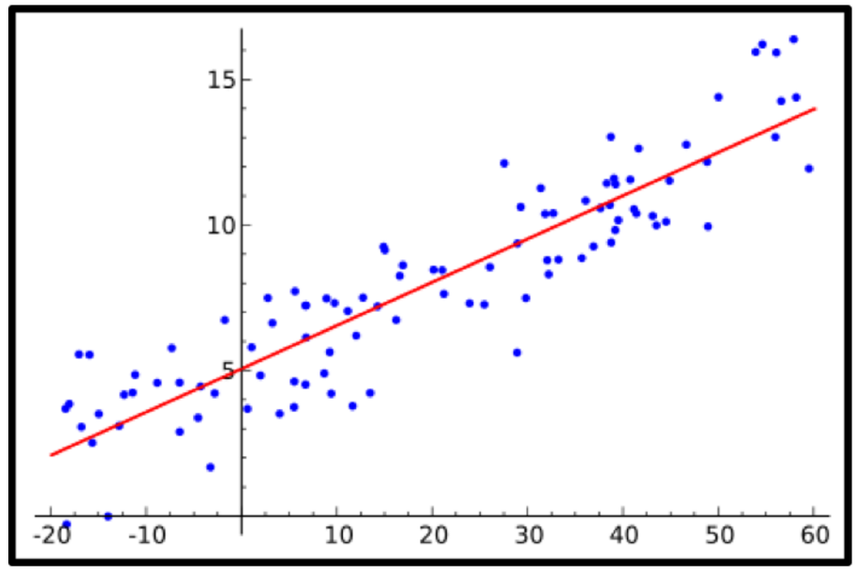


Data Mining Model (supervised)

- For decision tree mining, model = decision tree
- For deep learning, model = neural network with weights on connections
- For regression, model = (linear) function







Topic IENLP (Information Extraction and Natural Language Processing)

Fact

Relation

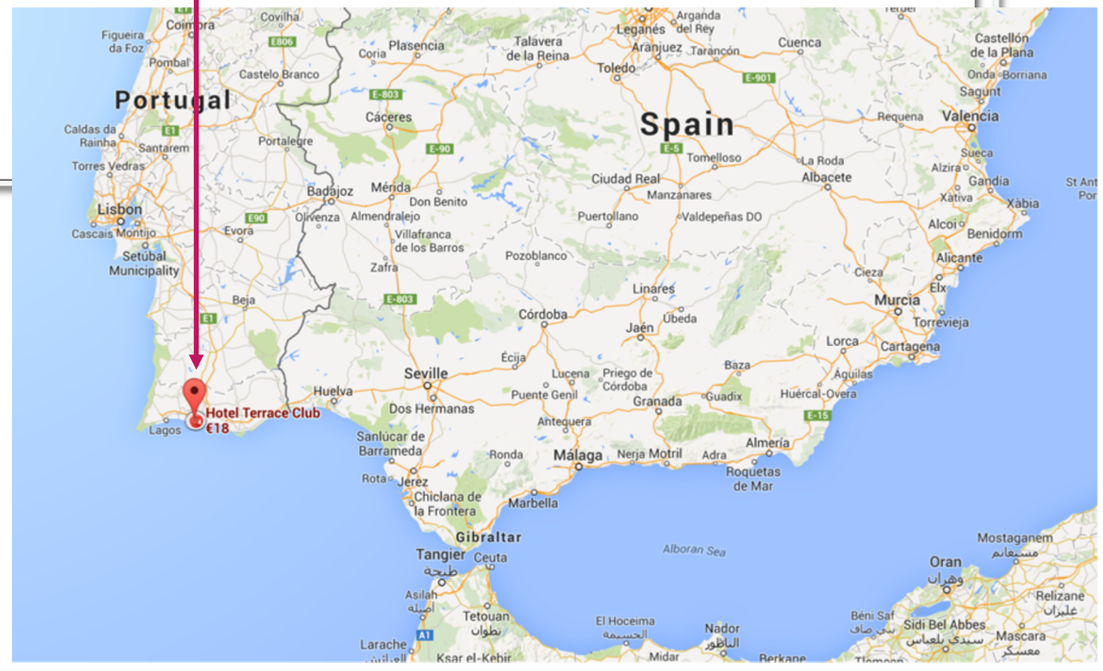
2-room apartment 55 m2: living/dining room with 1 sofa bed and satellite- TV, exit to the balcony. 1 room with 2 beds (90 cm, length 190 cm). Open kitchen (4 hotplates, freezer). Bath/bidet/WC. Electric heating. Balcony 8 m2. Facilities: telephone, safe (extra). Terrace Club: Holiday complex, 3 storeys, built in 1995 2.5 km from the centre of Armacao de Pera, in a quiet position. For shared use: garden, swimming pool (25 x 12 m, 01.04.-30.09.), paddling pool, children's playground. In the house: reception, restaurant. Laundry (extra). Linen change weekly. Room cleaning 4 times per week. Public parking on the road. Railway station "Alcantarilha" 10 km. Please note: There are more similar properties for rent in this same residence. Reception is open 16 hours (0800-2400 hrs). Lounge and reading room, games room. Daily entertainment for adults and children. Bar-swimming pool open in summer. Restaurant with Take Away service. Breakfast buffet, lunch and dinner(to be paid for separately, on site). Trips arranged, entrance to water parks. Car hire. Electric cafetiere to be requested in adavance. Beach football pitch. IMPORTANT: access to the internet in the computer room (extra). The closest beach (350 m) is the "Sehora da Rocha", Playa de Armacao de Pera 2.5 km. Please note: the urbanisation comprises of eight 4 storey buildings, no lift, with a total of 185 apartments. Bus station in Armacao de Pera 4 km.

Potential tags:

- LOCATION
- TIME
- PERSON
- ORGANIZATION
- MONEY
- PERCENT
- DATE

Named Entity Types

Named Entity Disambiguation



Topic IENLP (Information Extraction and Natural Language Processing)

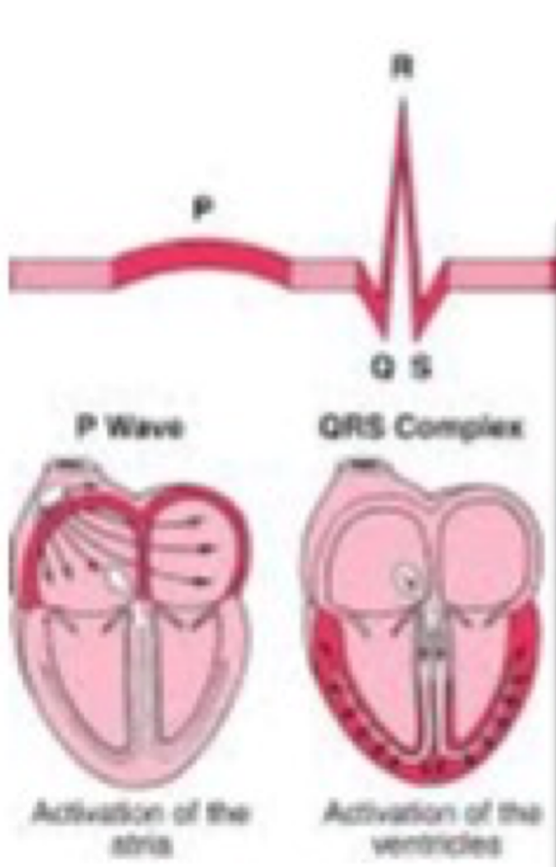
Skills

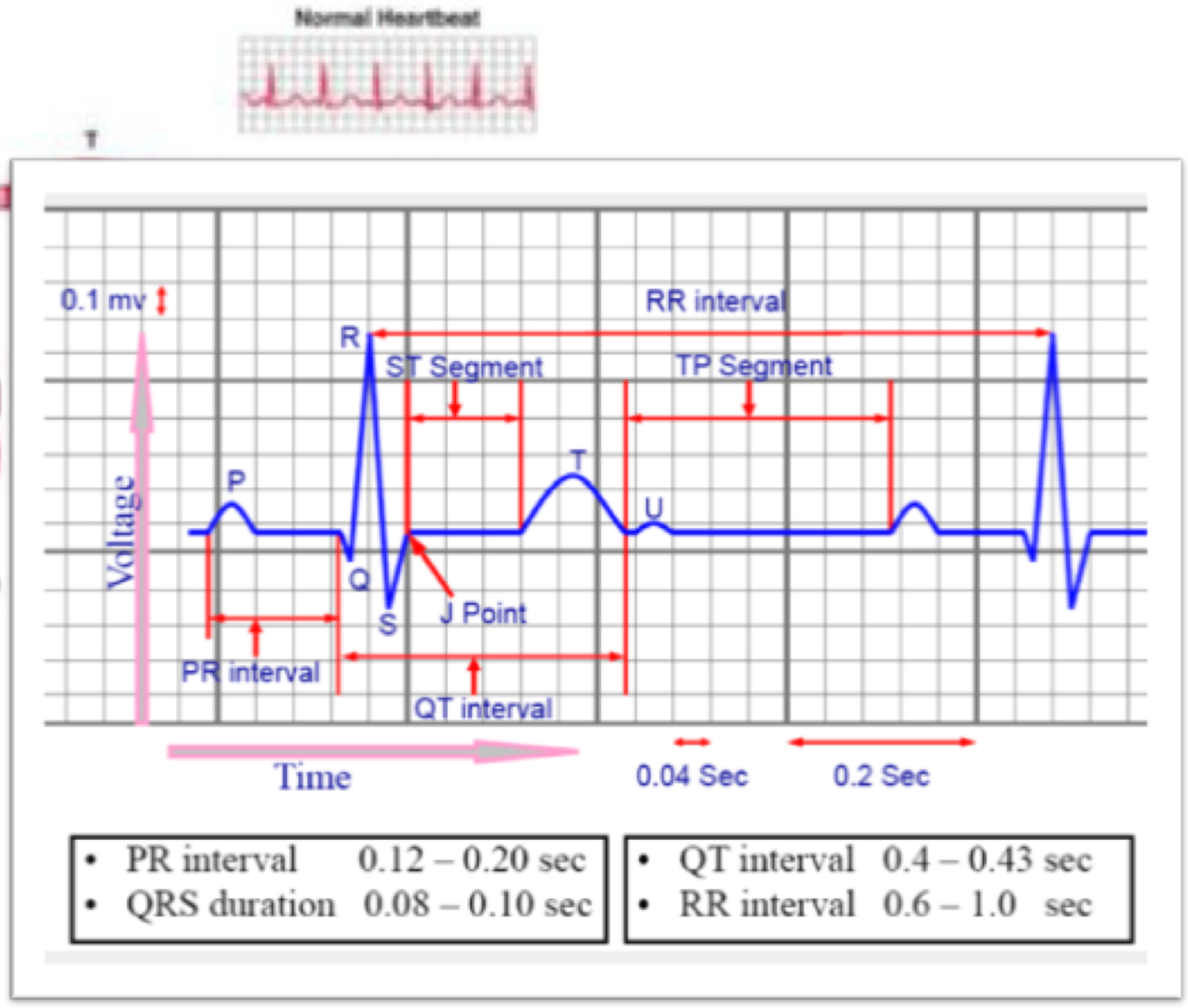
- Basic text processing: Normalization, regular expressions
- Text classification: Naive Bayes, evaluation
- Information extraction: Part-of-Speech tagging, named entity recognition

Data sets

- Hamlet
- Ada Lovelace Wikipedia page
- SMS
- Forum posts from a Single Forum
- Firefox Forum posts

Topic TS (Feature Extraction from Time-Series Data)





• PR interval	0.12 – 0.20 sec	• QT interval	0.4 – 0.43 sec
• QRS duration	0.08 – 0.10 sec	• RR interval	0.6 – 1.0 sec

Topic TS (Feature Extraction from Time-Series Data)

- Sensor data is like crude oil …
- It needs many stages of refinement to obtain features usable for data mining!

Topic TS (Feature Extraction from Time-Series Data)

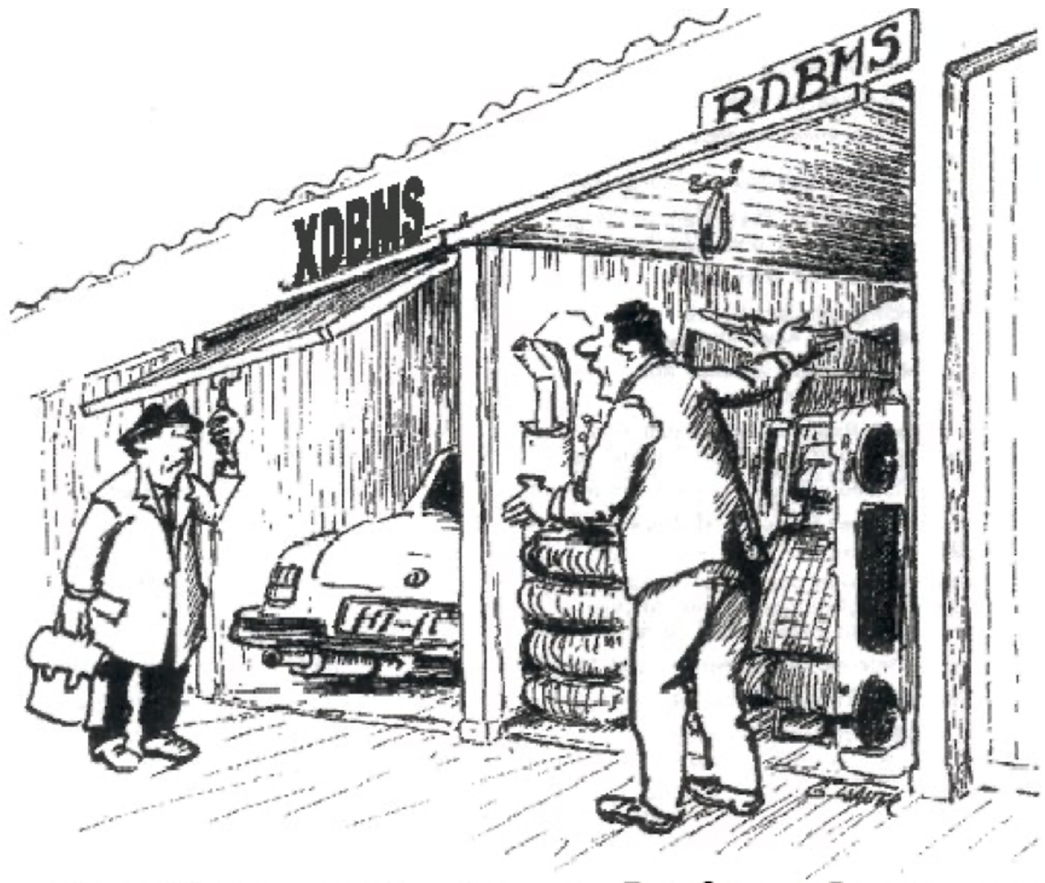
What this topic is about

- (Learning Python programming enough to do sensor data manipulation)
- Preprocessing: e.g., filtering (noise) / segmentation
- Inferring high-level events: e.g., peaks
- Transformation: e.g., frequency domain, time warping

- Connects well with Topic DM
- … and projects TIMETABLES, AF and EEG

Topic SEMI: semi-structured Data

- Structured data (aka tabular data)
  - Spread sheets / database tables
- Other less-structured formats for data publishing and exchange
  - XML
  - JSON
  - RDF



Topic SEMI: semi-structured Data

RDF http://en.wikipedia.org/wiki/Amsterdam

```
@prefix dbpedia
<http://dbpedia.org/resource/>.
@prefix
dbterm  <http://dbpedia.org/property/>.

dbpedia:Amsterdam
  dbterm:officialName "Amsterdam" ;
  dbterm:longd "4" ;
  dbterm:longm "53" ;
  dbterm:longs "32" ;
  dbterm:website
<http://www.amsterdam.nl> ;
  dbterm:populationUrban "1364422" ;
  dbterm:areaTotalKm "219" ;
  ...
dbpedia:ABN_AMRO
  dbterm:location dbpedia:Amsterdam ;
  ...
```

Amsterdam

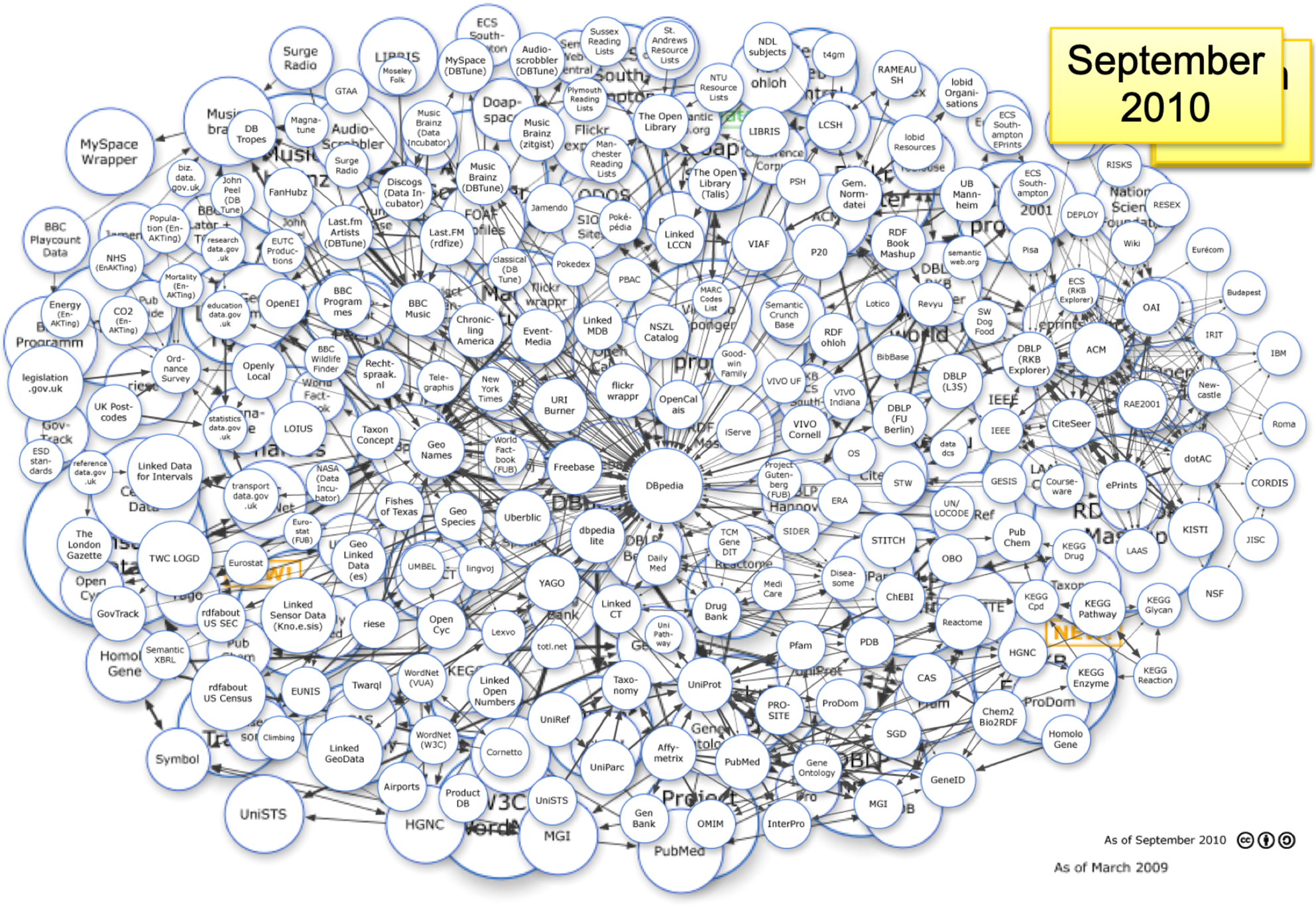
— Municipality / City —

Coordinates: 52°22′23″N 4°53′32″E

Country	Netherlands
Province	North Holland
COROP	Amsterdam
Boroughs	Boroughs
Government	
- Mayor	Eberhard van der Laan (PvdA)
- Aldermen	Carolien Gjlkiels Hans Gerson Maarten van Poelgeest Freek Ossel Marijke Vos
- Secretary	Henk de Jong
Area[1][2]	
- Municipality / City	219 km² (84.6 sq mi)
- Land	166 km² (64.1 sq mi)
- Water	53 km² (20.5 sq mi)
- Urban	1,003 km² (387.3 sq mi)
- Metro	1,815 km² (700.8 sq mi)
Elevation[3]	2 m (7 ft)
Population (June 2009)[4][5]	
- Municipality / City	762,057
- Density	4,459/km² (11,548.8/sq mi)
- Urban	1,364,422

Topic SEMI: semi-structured Data

Linked Open Data



Topic PDBDQ: Probabilistic Data Bases and Data Quality

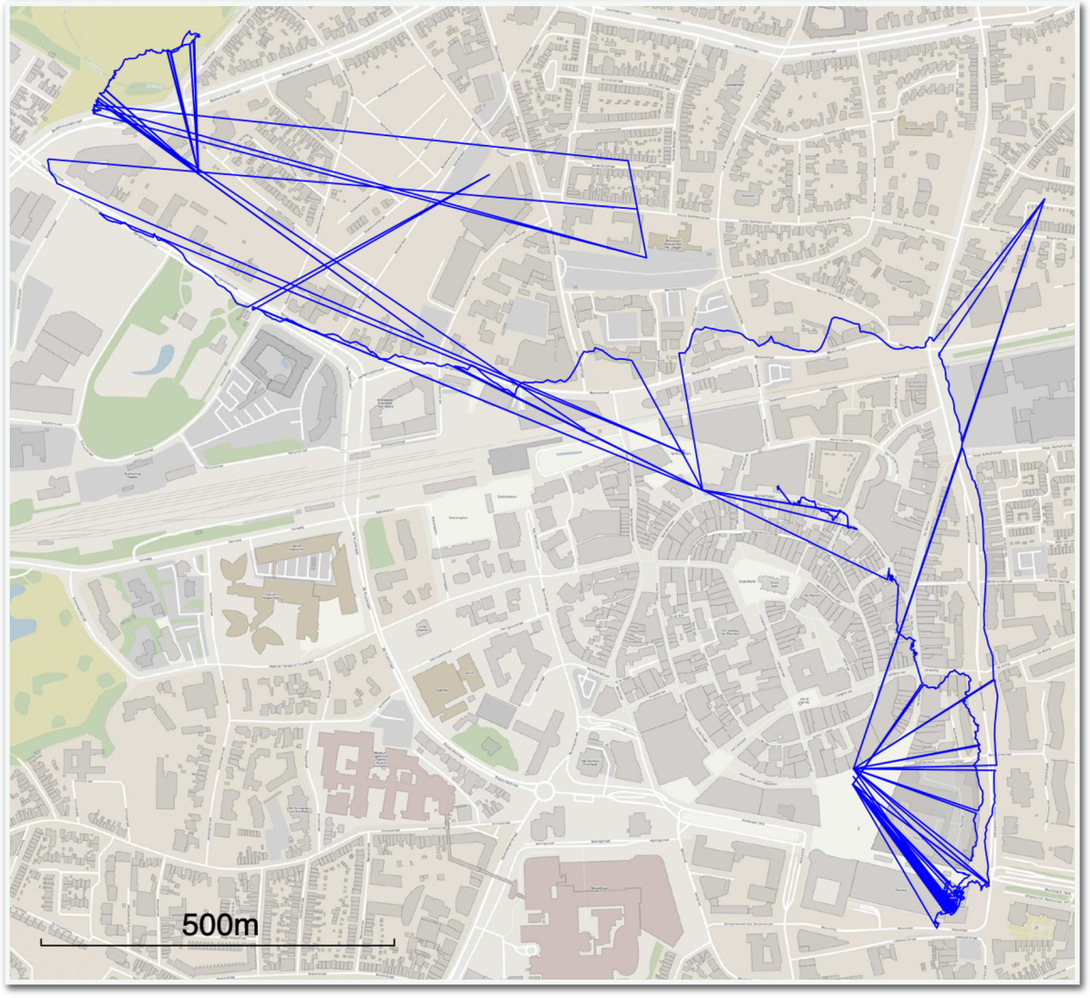
Before “real” DS starts (analysis) much needs to be done!

- Primary activities e-scientist: data preparation (extract, transform, clean), integration, curation
- Proves harder than originally thought
- Consuming most of an e-scientist’s time

- Bioinformatics: “fiddling with the data” may often consume more than half of the time of a PhD project
- Dirty data costs US businesses billions of dollars annually; cleaning accounts for 30% - 80% of the development time in a data warehouse project
- Gartner: Poor data quality primary reason for 40% of all business initiatives failing to achieve their targeted benefits

Topic PDBDQ: Probabilistic Data Bases and Data Quality

- What are these strange “attractors” in these GPS-traces?
- How to clean?

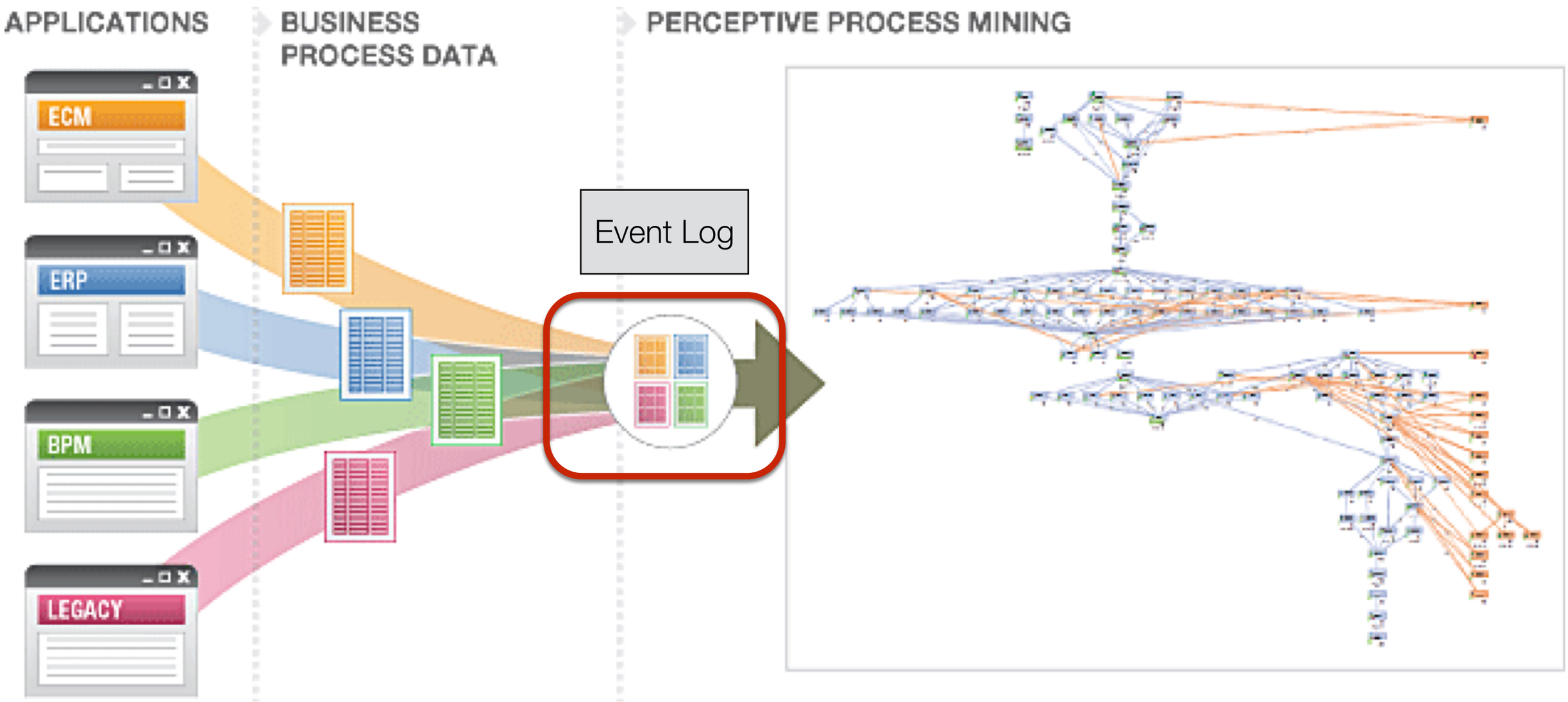


- Nitty gritty details may remain undiscovered rendering results invalid
- Excel corrupts data with automatic format conversions: gene names (1SEP) converted to dates, Riken identifiers to floating point numbers in micro array data

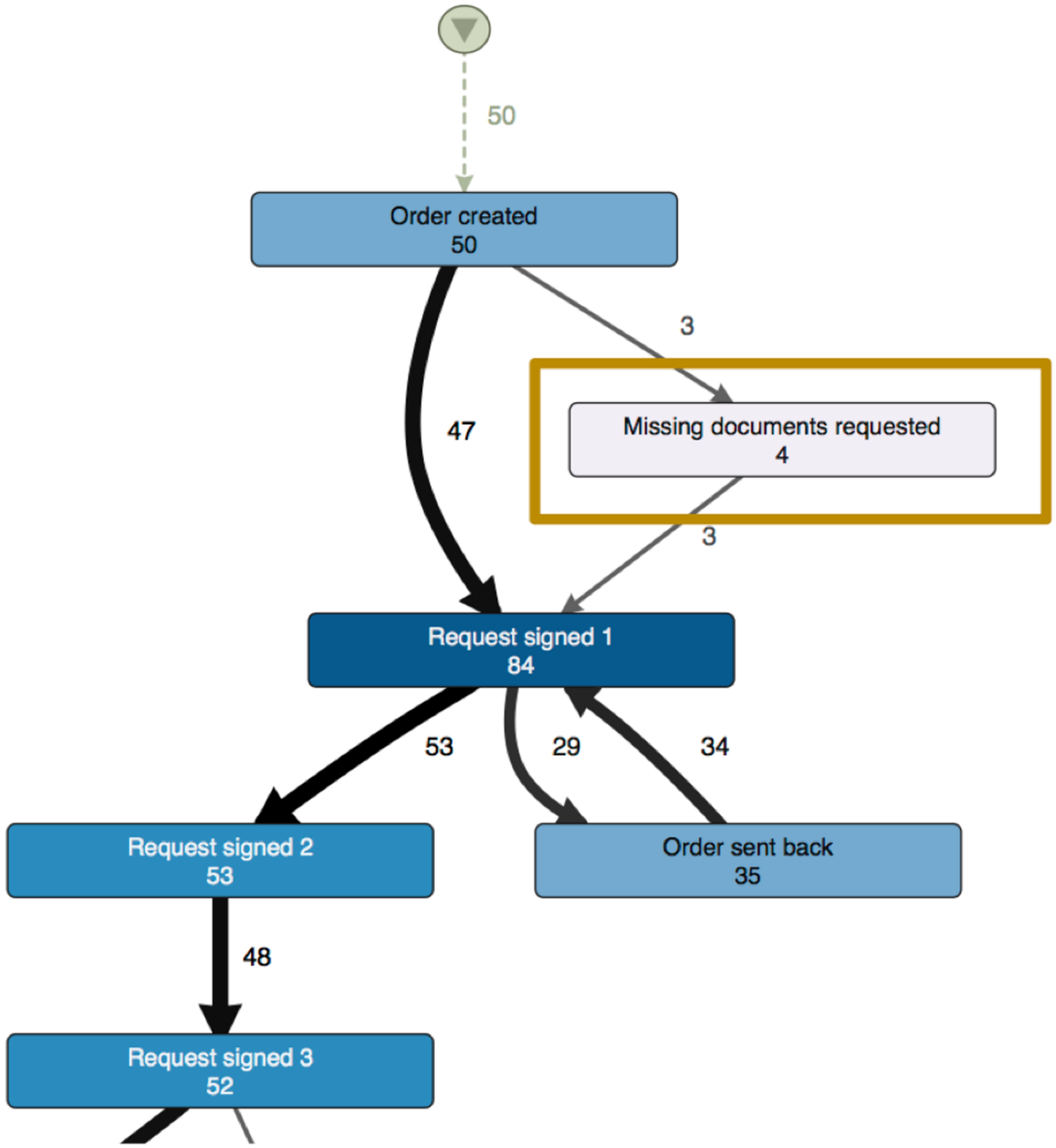
Topic PDBDQ: Probabilistic Data Bases and Data Quality

id	name	address	postcode	age	
1	John Smith	Street 1	7599 AA	50	**1,0**
2	Jan Smith	Street 1	7599 AA	48	**1,0**
3	Toddler Smith	Street 1	7599 AA	6	**1,0**
4	Teenager Smith	Street 1	7599 AA	12	**1,0**
5	Uncle Smith	Street 3	7599 AA	29	**0,2**
		Lane 5	7588 BB	31	**0,8**
6	Baby Smith	Lane 5	7588 BB	1	**0.7 ?**
	GrandPa Smith				**0,9**

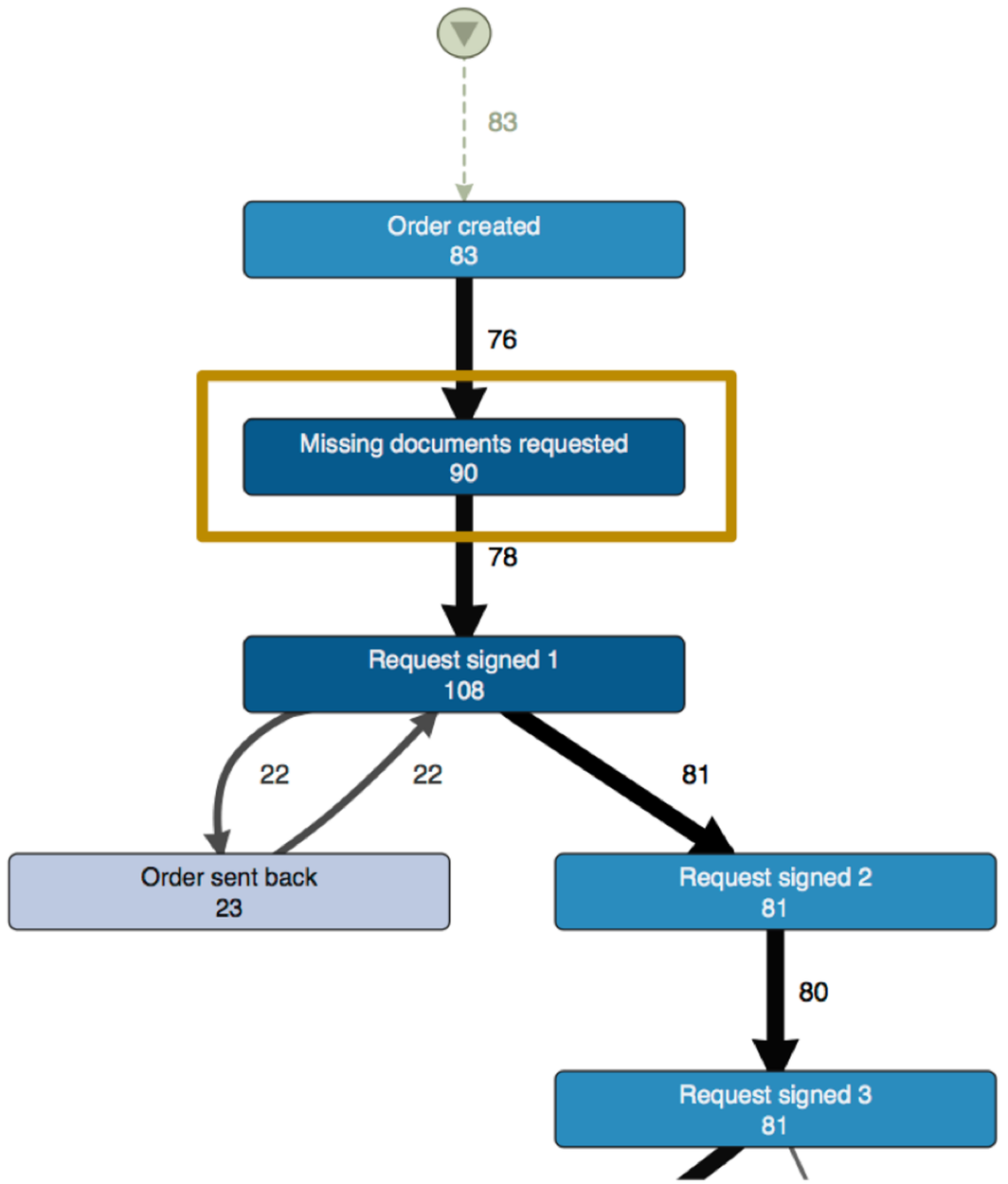
Topic PM: Process Mining



Topic PM: Process Mining



(a) Via **Callcenter**



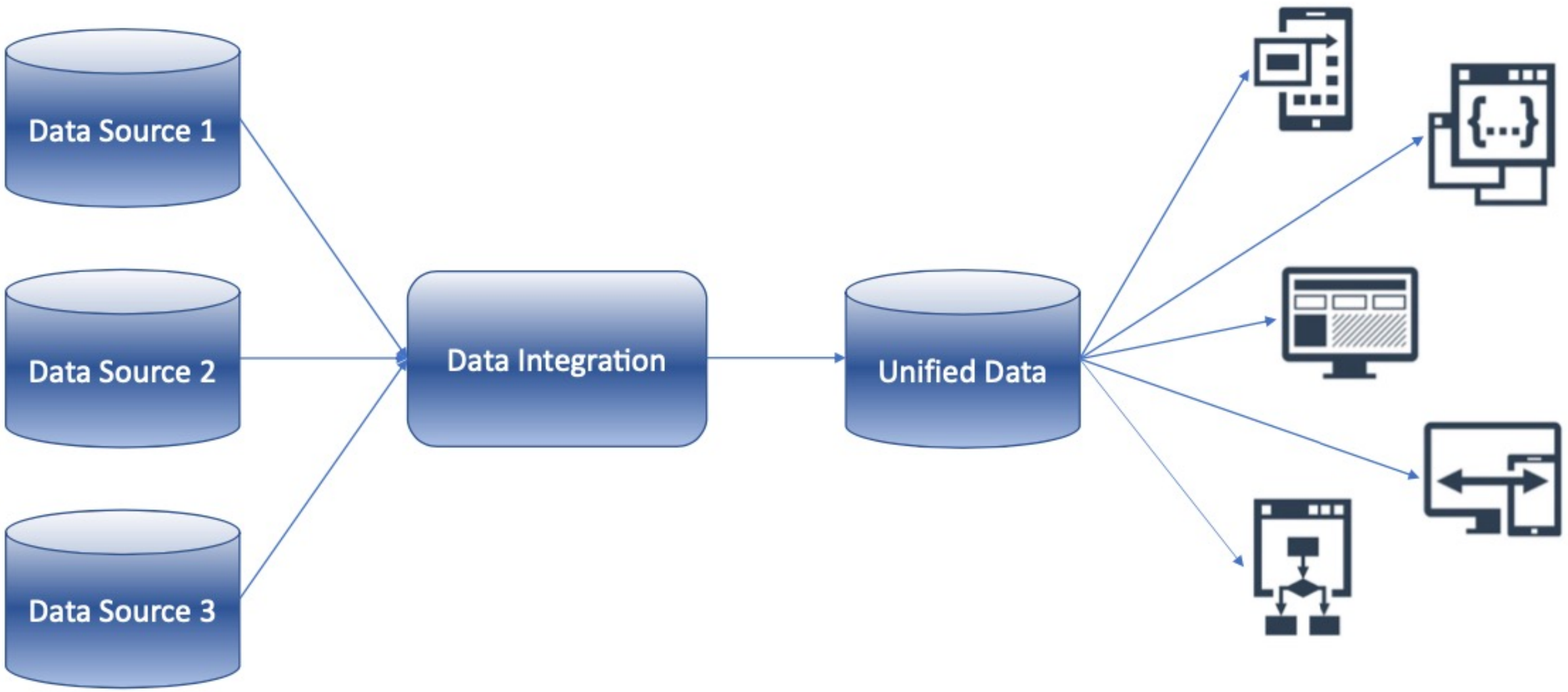
(b) Via **Internet**

Topic PM: Process Mining

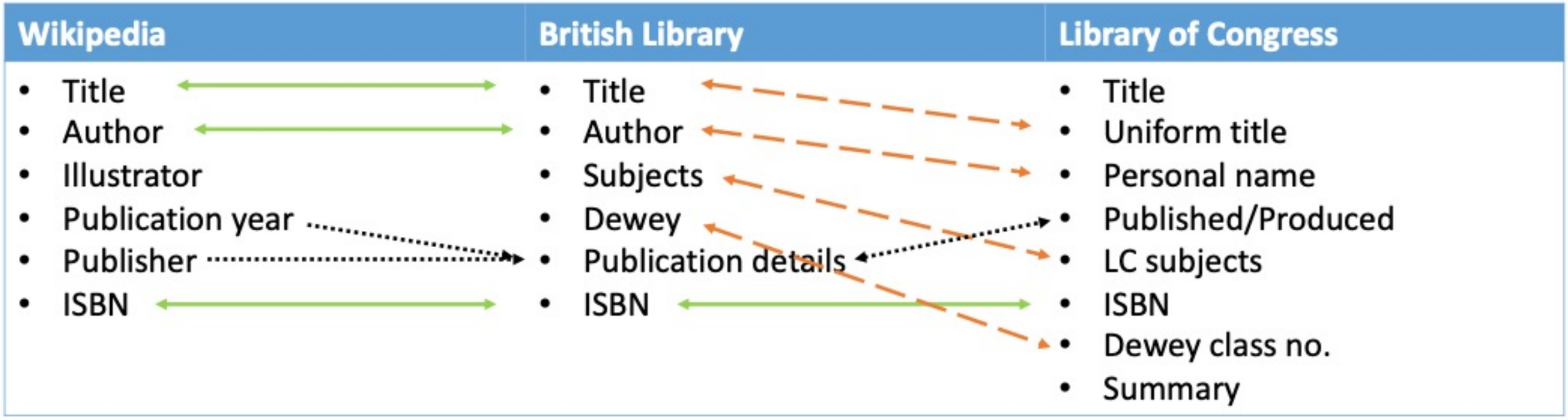
Most important concepts and skills
for applying and understanding Process Mining

- Petri Nets: the theoretical foundation of process mining
- Concepts: event log, causal trace, Alpha algorithm
- Using the ProM tool for process discovery
- Answering analytical questions for a discovered process
- Using the ProM tool for process conformance checking

Topic DINT: Data Integration

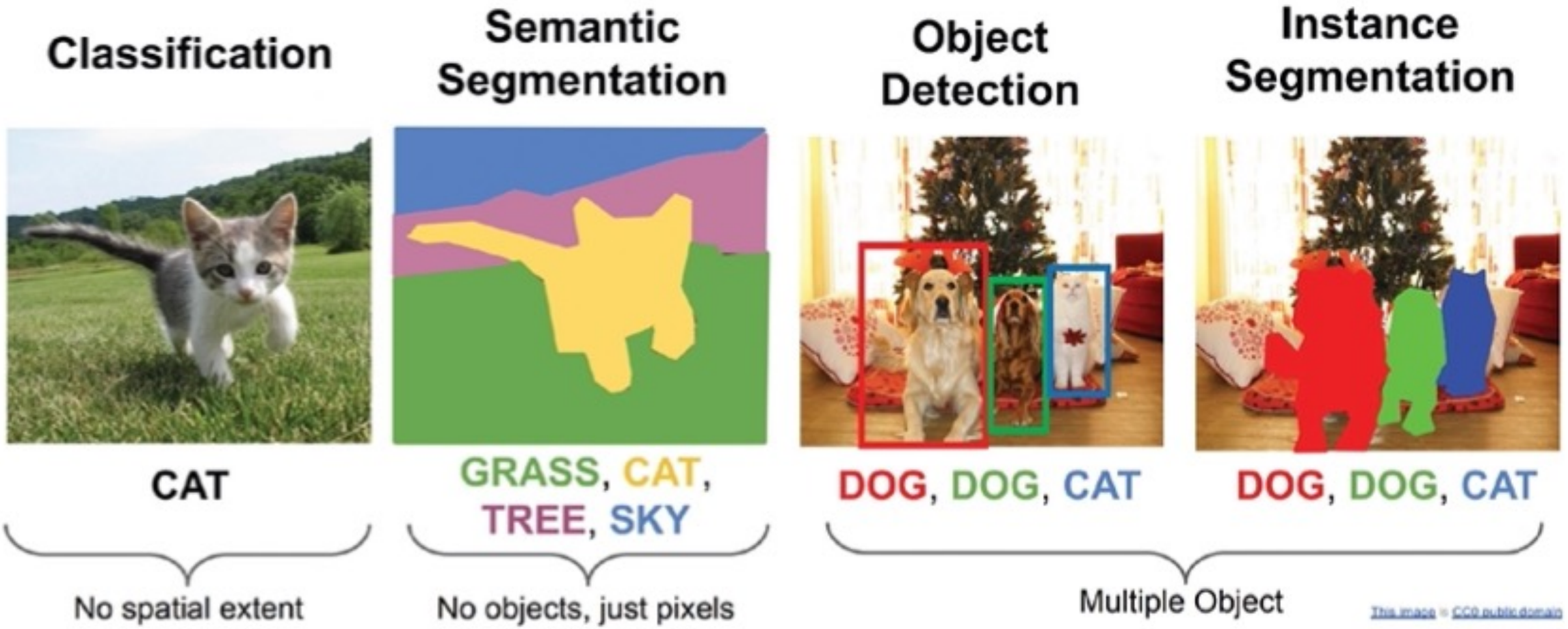


Topic DINT: Data Integration



	Wikipedia	British Library	Library of Congress
Title	Harry Potter and the Philosopher's Stone	Harry Potter and the Philosopher's Stone	Harry Potter and the sorcerer's stone
ISBN	0-7475-3269-9	0747574472 (pbk)	9781338299144 (paperback), 133829914X (paperback)
Subject		Juvenile fiction, fantasy fiction	Wizards—Fiction, Magic-—Fiction, Schools-Fiction
Dewey		823.914	823/.914 [Fic]
Publishing status	• Publication year: 1997 • Publisher: Bloomsbury (UK)	• Publication details: London : Bloomsbury, 1997 2004 printing.	• Published/Produced: New York, NY : Scholastic Inc., [2018]

Topic CV&IP: Computer Vision & Information Processing



Applications

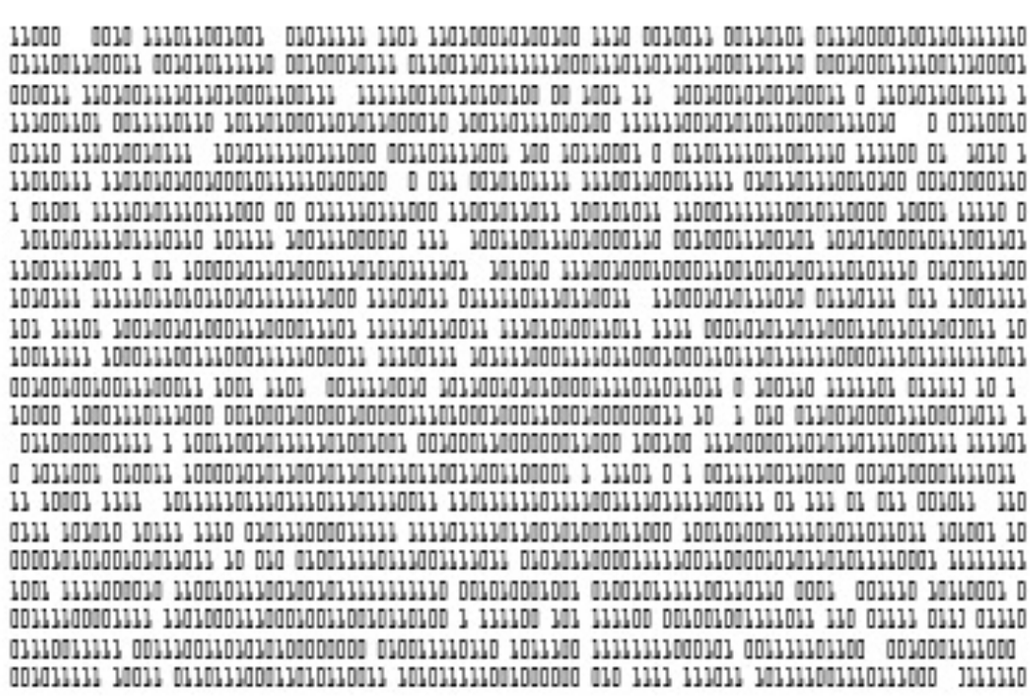
- Surveillance & security, medical imaging & health, media & entertainment, mobile computer vision, cultural inheritance, assistive driving, etc.

Topic CV&IP: Computer Vision & Information Processing

The problem

What we see

What the computer sees



Topic GIS: Handling spatial data”

NEW

Under development
Not truly available yet

Spatial data

- Geospatial vector data
  For example, map data like Google Maps & Open Street Map
- Geospatial raster data
  For example, satellite images, laser altimetry, etc.
- Thematic data with locations (coordinates/addresses/polygons)

Common tasks

- Segmentation, object detection & classification from raster data
- Combining ‘data with locations’ with maps & analyze / predict
- Visualization by means of thematic overlays or Google Earth

Projects

Project List

1. Decision Support for University Timetables
2. Business Intelligence
3. Analyse and Predict Transportations
4. Predicting neurological outcomes (EEG)
5. Text Classification (Conference Recommender), Twitter NER
6. Detection of Atrial fibriliation episodes
7. Linking Open Cultural Data
8. Probabilistic Music Data Base
9. Who should treat Low Back Pain?
10. Predicting surgical case durations for a Thorax centre
11. Web Harvesting for Smart Application
12. What can we learn from online healthcare communities
13. How do inventions evolve?
14. Discover and analyse processes

Slight updates each quarter. List on Canvas.

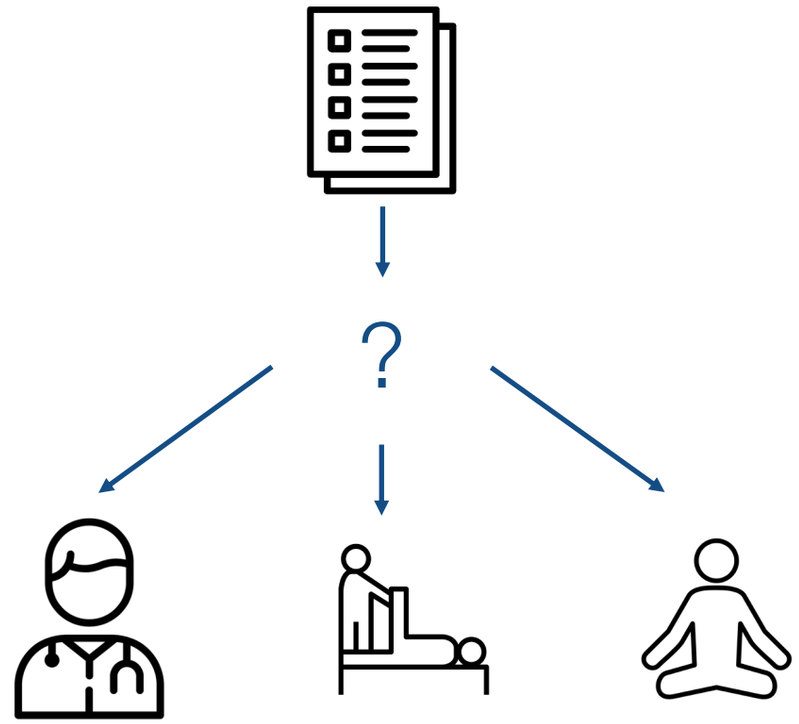
Project RA (Referral Advice)

Data set

- 1288 fictive patient cases (questionnaire) on low back pain that were judged by healthcare professionals on referral advices on a 5-point scale

Challenge

- Predict referral to GP, physiotherapist, self-care
- To what extent can one reduce the number of questions in the questionnaire



Project TIMETABLES

Data set

- Several years of data from MyTimeTable for UT & Saxion

Challenge

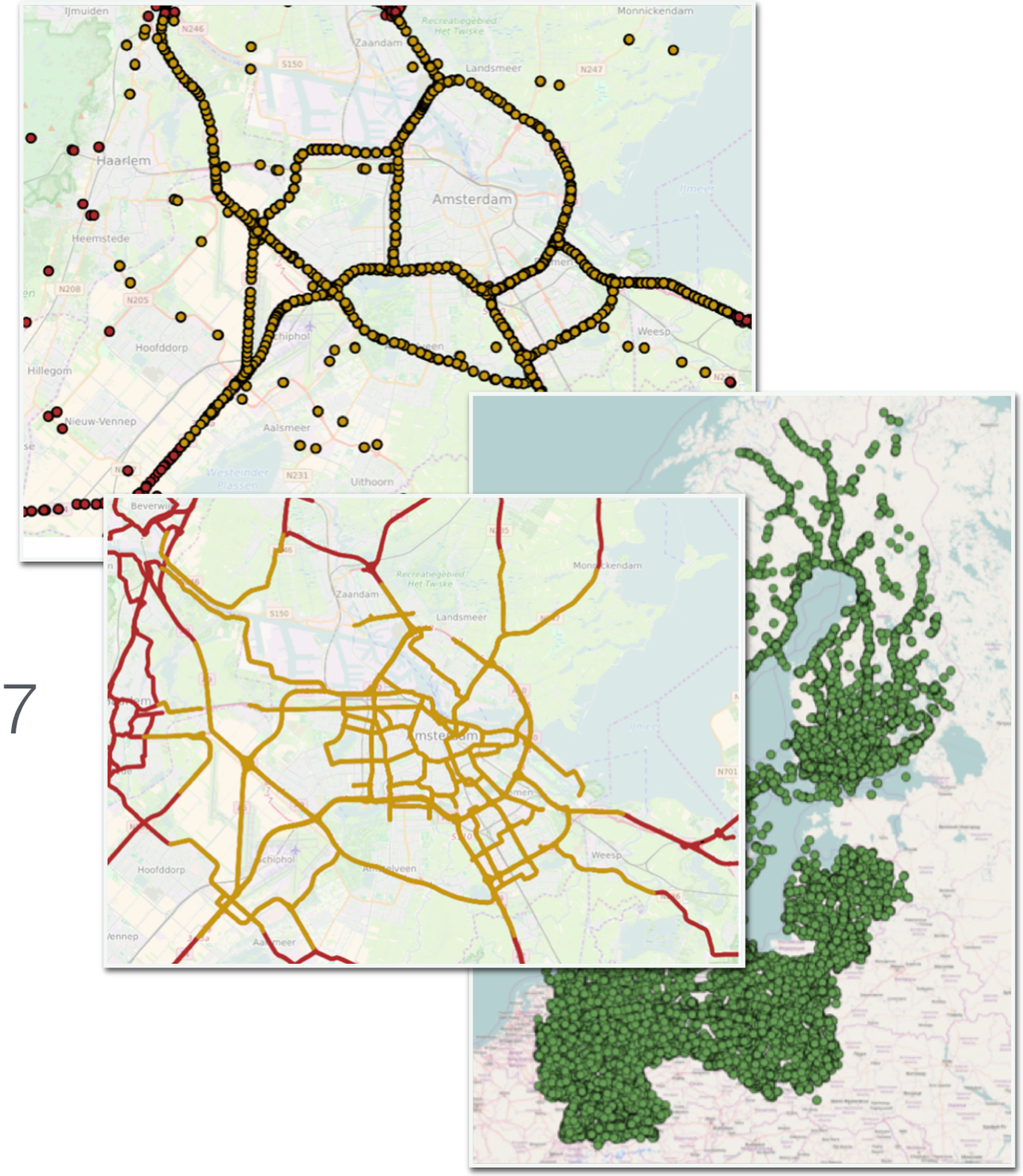
- Check KPIs, explore problems & trends

- Capacity: over programs, buildings, weekdays, periods
- ‘Travel’ distances
- Room occupation: too small/large, facilities

Project TRANSPORT

Data sets

- Flows/speeds, travel times, status of A'dam area
- BE-mobile: trip-data of A10 west area
- Outbound Call Detail Records of Estonians in 7 other countries (roaming data)



Challenges

- many :) (see description)

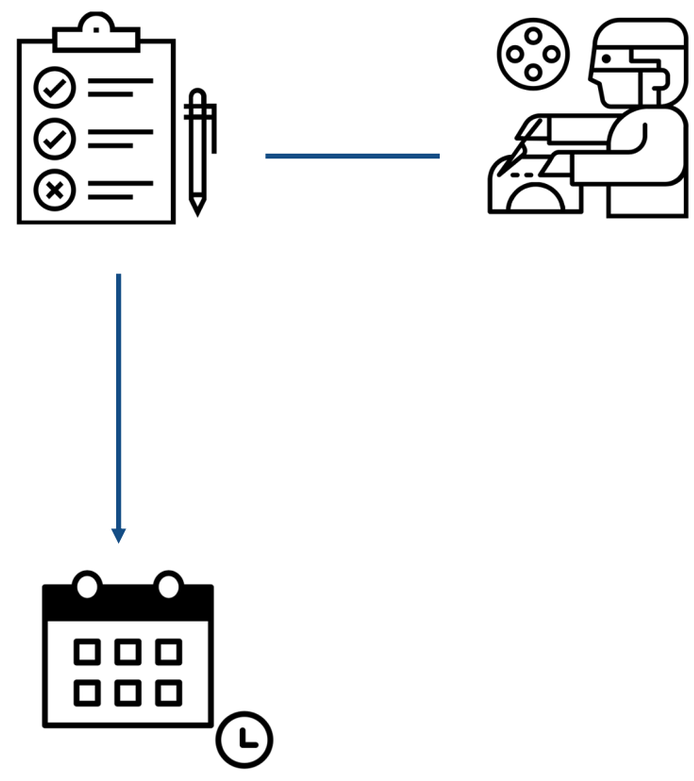
Project PSCD (Surgical Cases)

Data set

- Data on 4087 surgical cases performed from Jan 2013 to Jan 2016 at Thorax Centrum Twente

Challenge

- identify patterns in surgical case durations
- derive prediction models surgical case
- decreasing overtime while maintaining OR-utilisation
- finding possible causes for overtime / undertime



MOCHA: Public Priorities

- Descriptive, cross-sectional, quantitative questionnaire
  - representative sample of the general public
  - United Kingdom, the Netherlands, Germany, Spain and Poland
  - eliciting preferences wrt children's primary care + measure experiences with quality of currently provided care
- 2641 rows/respondents, 305 columns
- Challenge: clean it up, good quality cube, visualise and summarise

COVID: Predicting Mortality

Two datasets

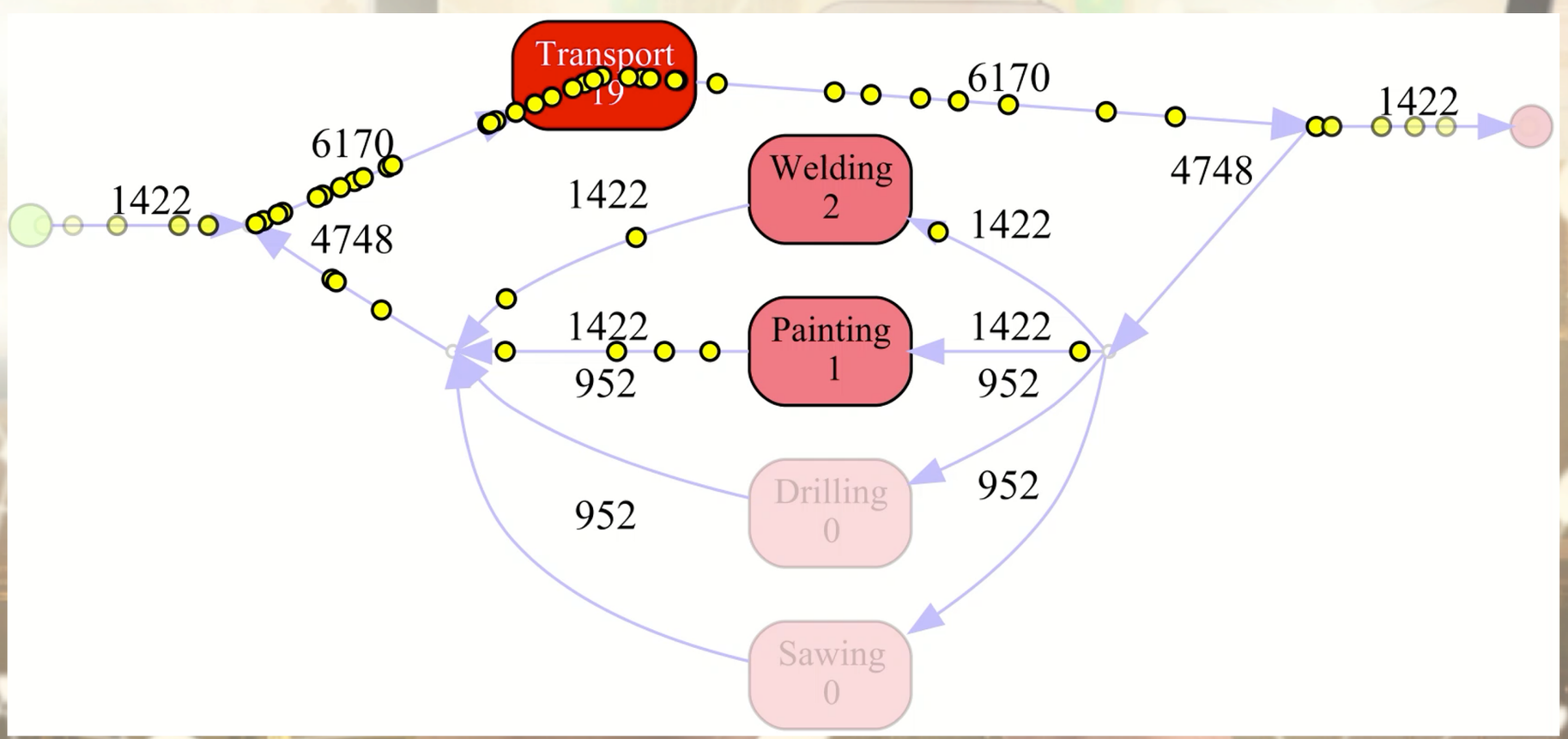
- training set 375 patients, Wuhan, 10 Jan-18 Feb, biomarkers measured, 174 died
- Test set 110 patients, Wuhan, 19 Feb-24 Feb, limited biomarkers, 13 died.

Paper suggests that the mortality of an individual COVID-19 case can be predicted with more than 90% accuracy and over 10 days in advance based on just three biomarkers: lactic dehydrogenase (LDH), lymphocyte, and high-sensitivity C-reactive protein (hs-CRP)

Project PDA (Process Discovery and Analysis)

Analyze discrepancies between specifications and observed reality

Discover & analyze business processes



Project GEO

NEW

Analysis of Geospatial Images

- Glyphosate is a controversial chemical that is used in agriculture to kill weeds
- May cause decline of insect life, affect animals in the soil and contaminate drinking water potentially causing cancer
- Policies in EU in place to reduce, but still legal to use it

Challenge

- Develop a workflow that can be used to detect from satellite images agricultural fields on which glyphosate has been applied
- Data: from Google Earth Engine (GEE) catalog and/or Sentinel multispectral images (with 10 m spatial resolution)

Summary

Next Tasks

1. Form a group (2 students) for topics + project
2. Choose 2 topics and 1 project
3. Enroll group on Canvas
4. Forms: (a) Register in Discord + (b) Questionnaire
5. Attend the topic lectures
   optionally the Topic 0-R or Topic 0-Python lecture
6. Sign in into Discord and read & answer & discuss

What if..?

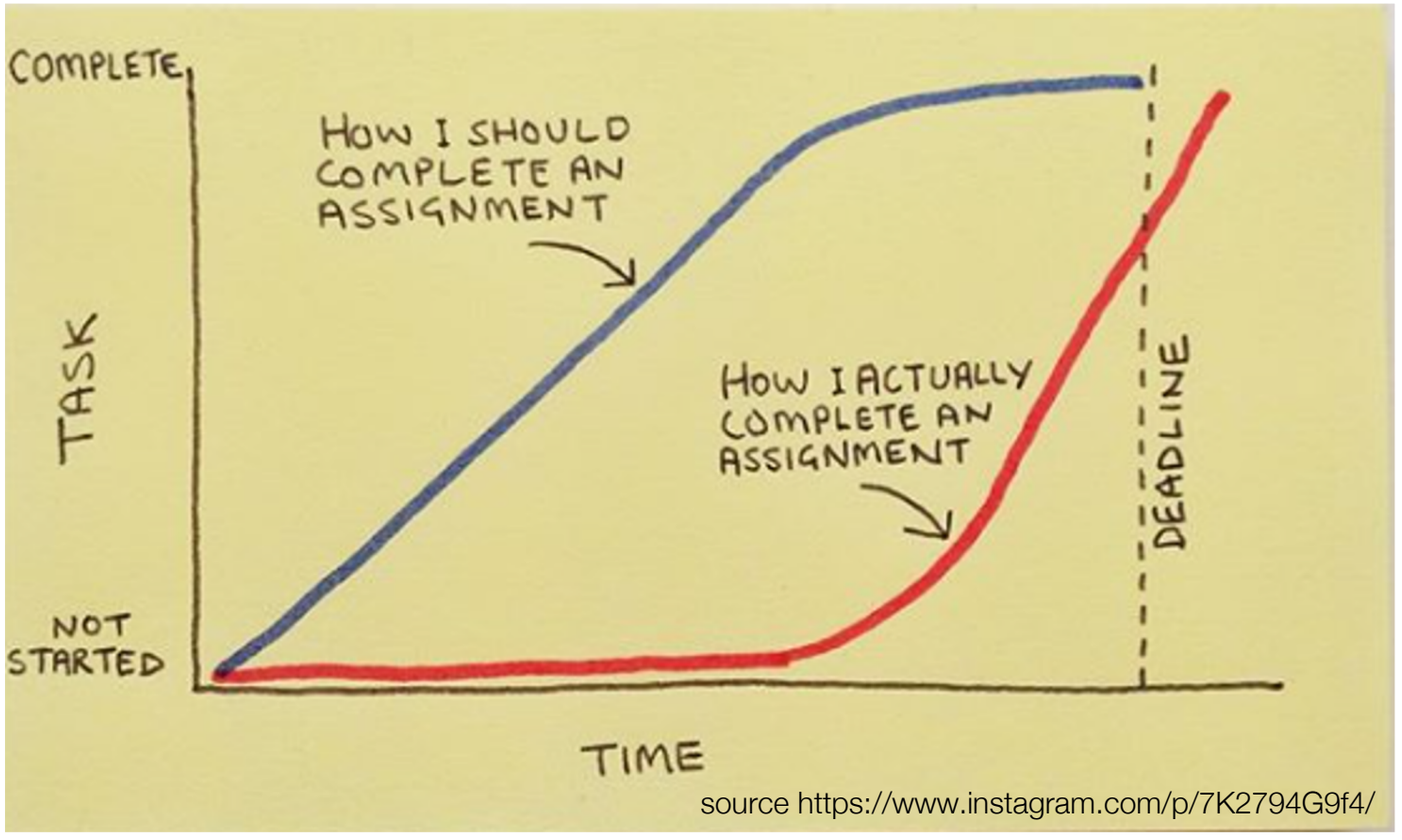
- There may be things we didn’t think of
- There may be things that are still rough and unclear
- There may be things that change at the last moment

1. Canvas!
   - Keep a close watch on Canvas/Discord announcements
   - New versions of course guide, material, and software
2. Let yourself be heard!
   - Suggestions or anything too hard, too unclear, unfair
   - Approach main teacher and assistants
   - Discord (allows private messages, too)

Course in a nutshell

- work in pairs, sign up for a group on canvas
- do 2 topics, they have to be signed off
- do 1 project, this is graded
- submit everything to canvas to get it checked/graded
- do 1 presentation of your project
- course communication via Canvas & Discord (sign up)
- read the **FAQ**

Questions?



source https://www.instagram.com/p/7K2794G9f4/