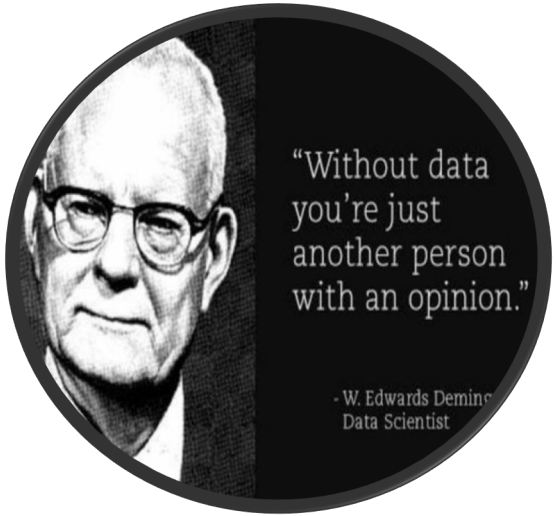
Objectives and considerations of (Big) data science education in Life Sciences domain

26 October 2018, Lukasz Grus



WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

glassdoor Jobs Company Reviews Salaries Interviews Know Your Worth Sign In Write Review For Employers

Job Title, Keywords, or Company | Jobs | Location | Search

50 Best Jobs in America

Awards
- Best Places to Work
- Top CEOs
- Best Places to Interview

Lists
- Best Jobs
- Best Cities for Jobs
- Highest Paying Jobs
- Oddball Interview Questions

Trends
- Overview

Job Trends
- Trends by Industry
- Trends by Location

Company Trends
- Trends by Industry
- Trends by Location
- Trends by Job Title

Employer Resources

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

United States | 2018 | 0 Shares

1 Data Scientist

4.8 / 5 Job Score
4.2 / 5 Job Satisfaction
$110,000 Median Base Salary
4,524 Job Openings
View Jobs

2 DevOps Engineer

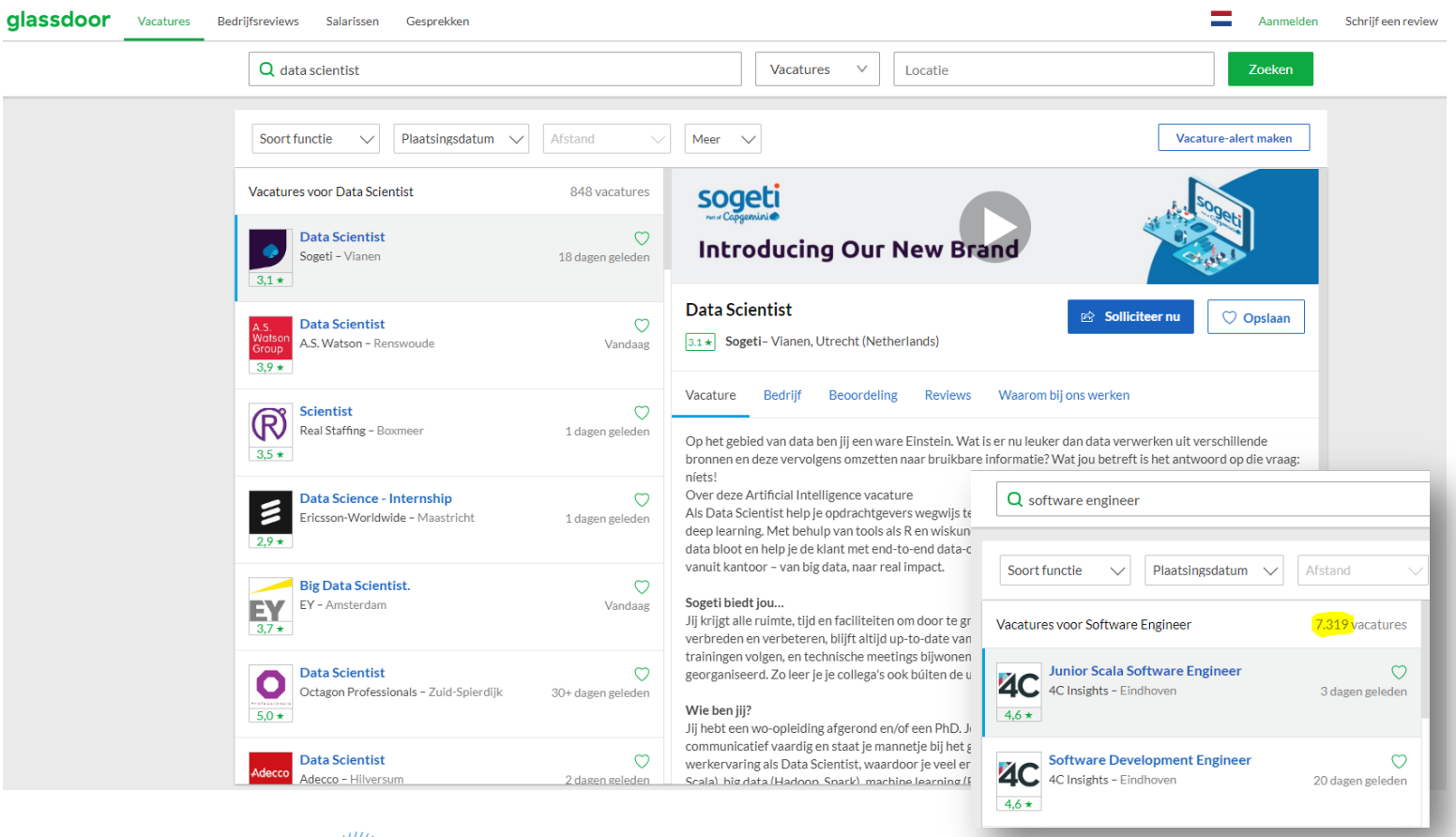
4.6 / 5 Job Score
4.0 / 5 Job Satisfaction
$105,000 Median Base Salary
3,369 Job Openings
View Jobs

3 Marketing Manager

4.6 / 5 Job Score
4.0 / 5 Job Satisfaction
$85,000 Median Base Salary
6,439 Job Openings

WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018



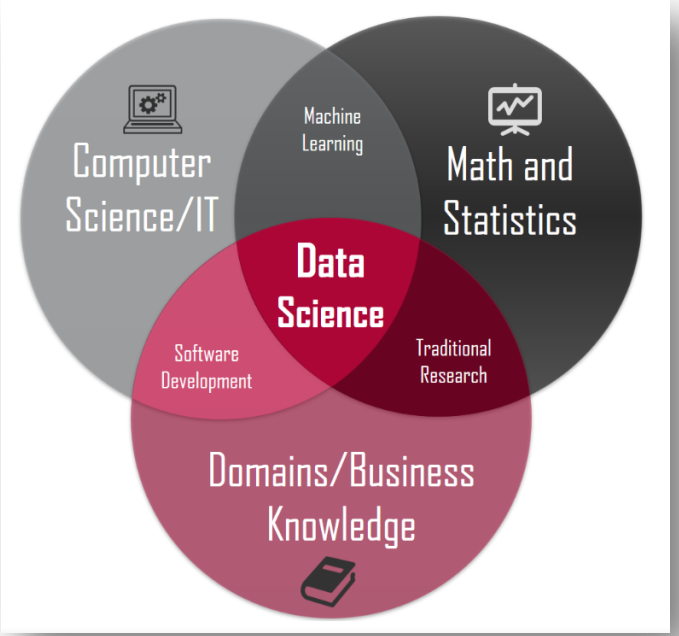
WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

What is data science ?

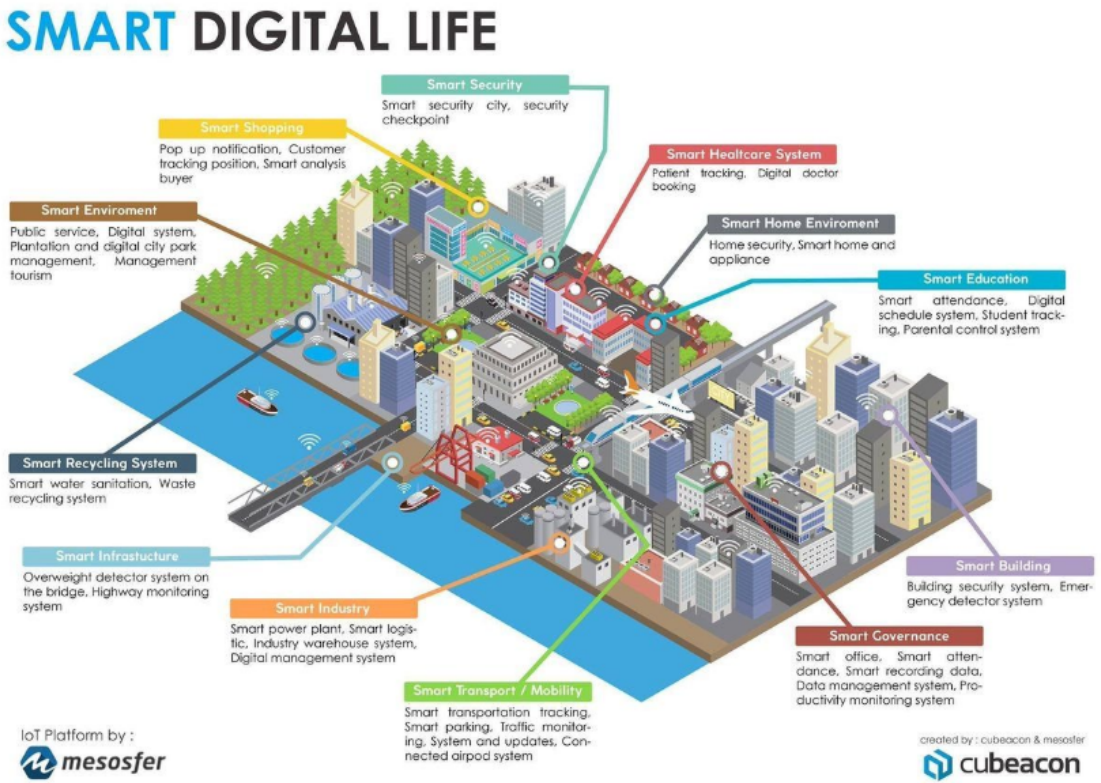
Data science is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing.

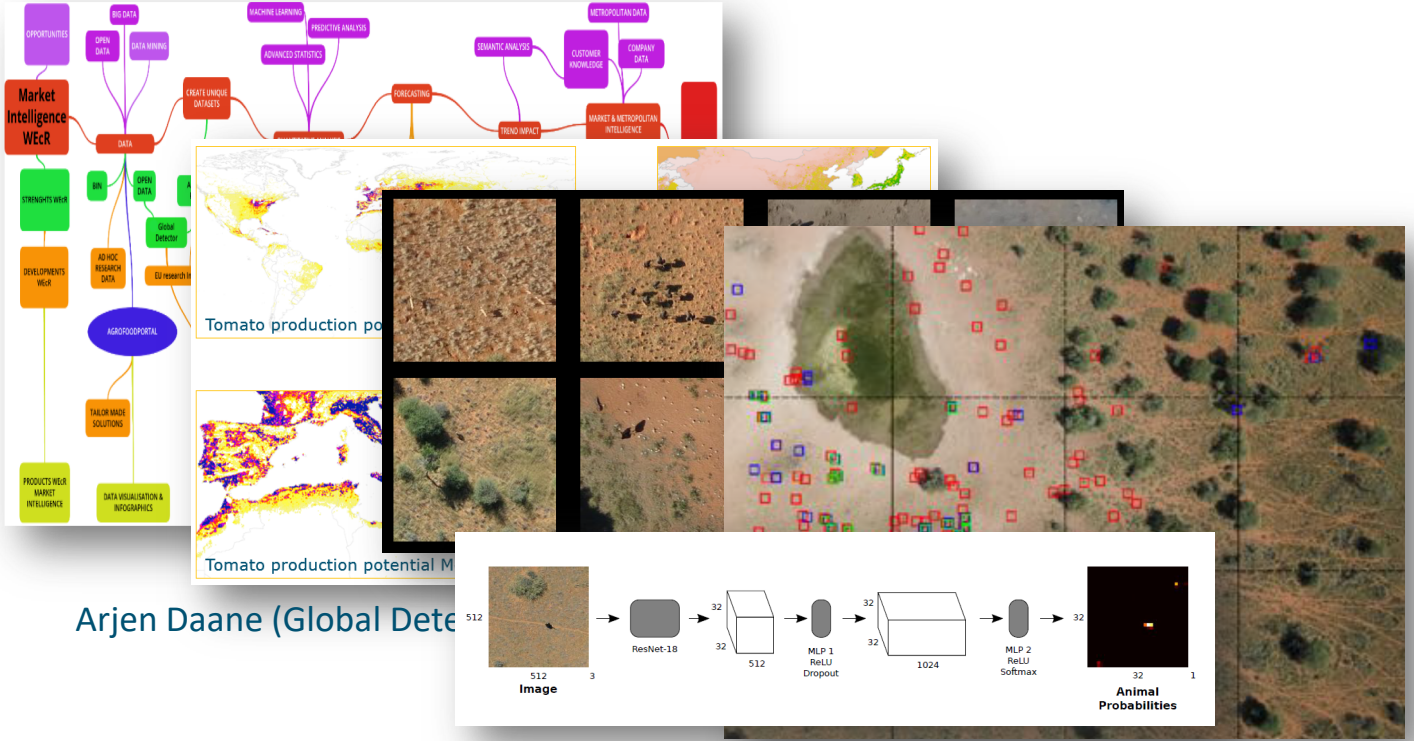
Data science [...] incorporates principles, techniques, and methods from many disciplines and domains including **(advanced) data management, data cleansing, analytics, visualization, engineering, etics, policy**, in the context of Big Data*



* Adapted from National Institute of Standards and Technology (NIST) Big Data Working Group (2015)

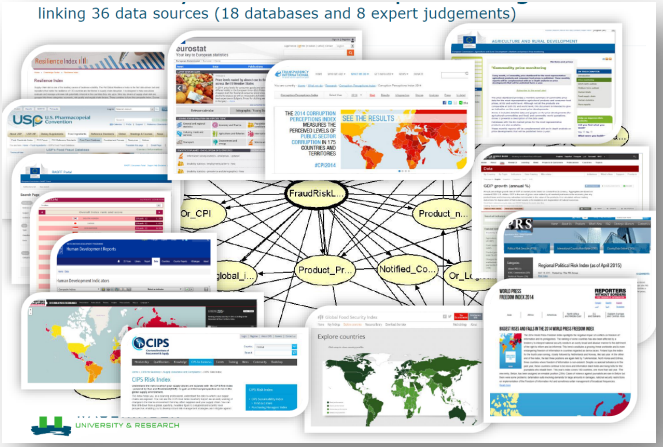
Why data science?

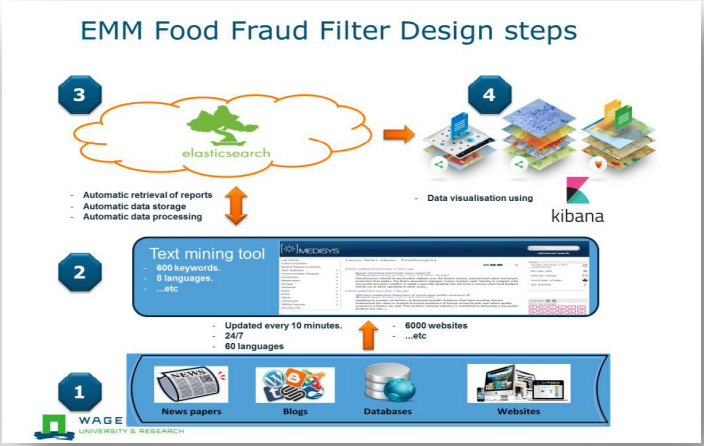


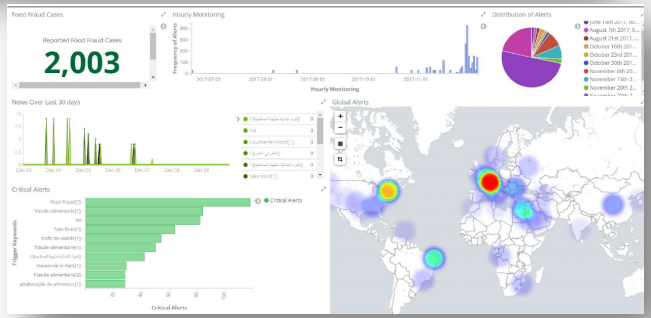


Arjen Daane (Global Dete

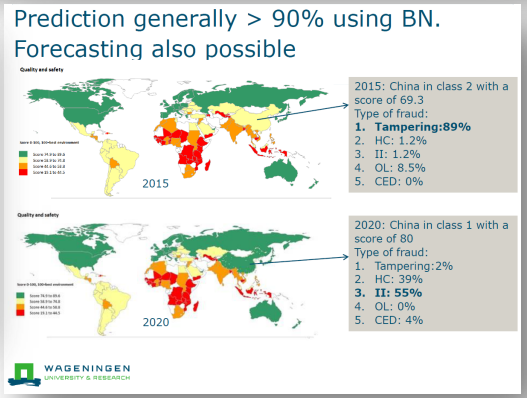
Devis Tuia, Benjamin Kellenberger (GRS)

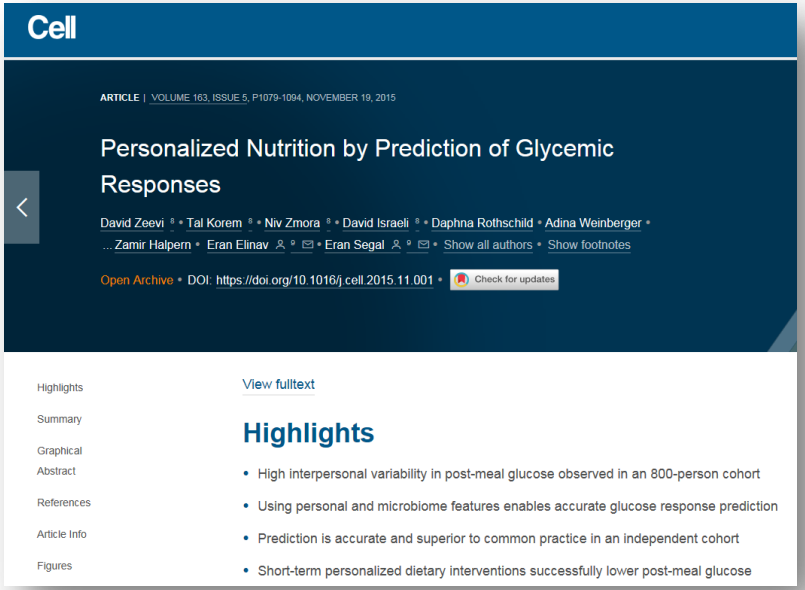






Hans Marvin, Yamine Bouzembrak (RIKILT)





Cell

ARTICLE | VOLUME 163, ISSUE 5, P1079-1094, NOVEMBER 19, 2015

Personalized Nutrition by Prediction of Glycemic Responses

David Zeevi [8] • Tal Korem [8] • Niv Zmora [8] • David Israeli [8] • Daphna Rothschild • Adina Weinberger • ... Zamir Halpern • Eran Elinav • Eran Segal • Show all authors • Show footnotes

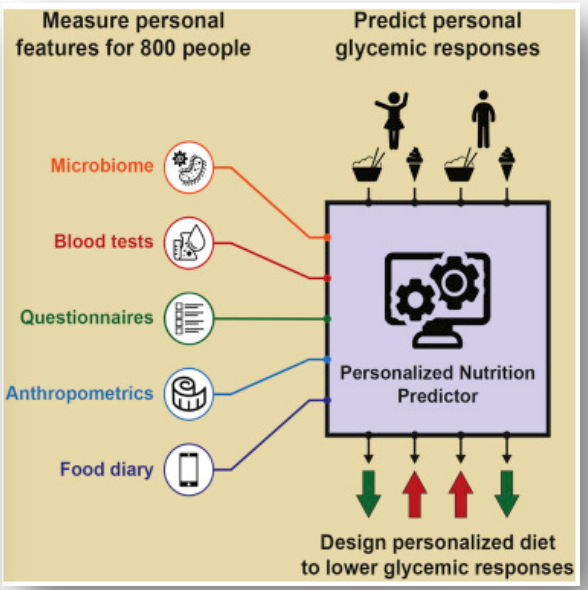
Open Archive • DOI: https://doi.org/10.1016/j.cell.2015.11.001 • Check for updates

Highlights
Summary
Graphical Abstract
References
Article Info
Figures

View fulltext

Highlights

- High interpersonal variability in post-meal glucose observed in an 800-person cohort
- Using personal and microbiome features enables accurate glucose response prediction
- Prediction is accurate and superior to common practice in an independent cohort
- Short-term personalized dietary interventions successfully lower post-meal glucose



WAGENINGEN UNIVERSITY & RESEARCH

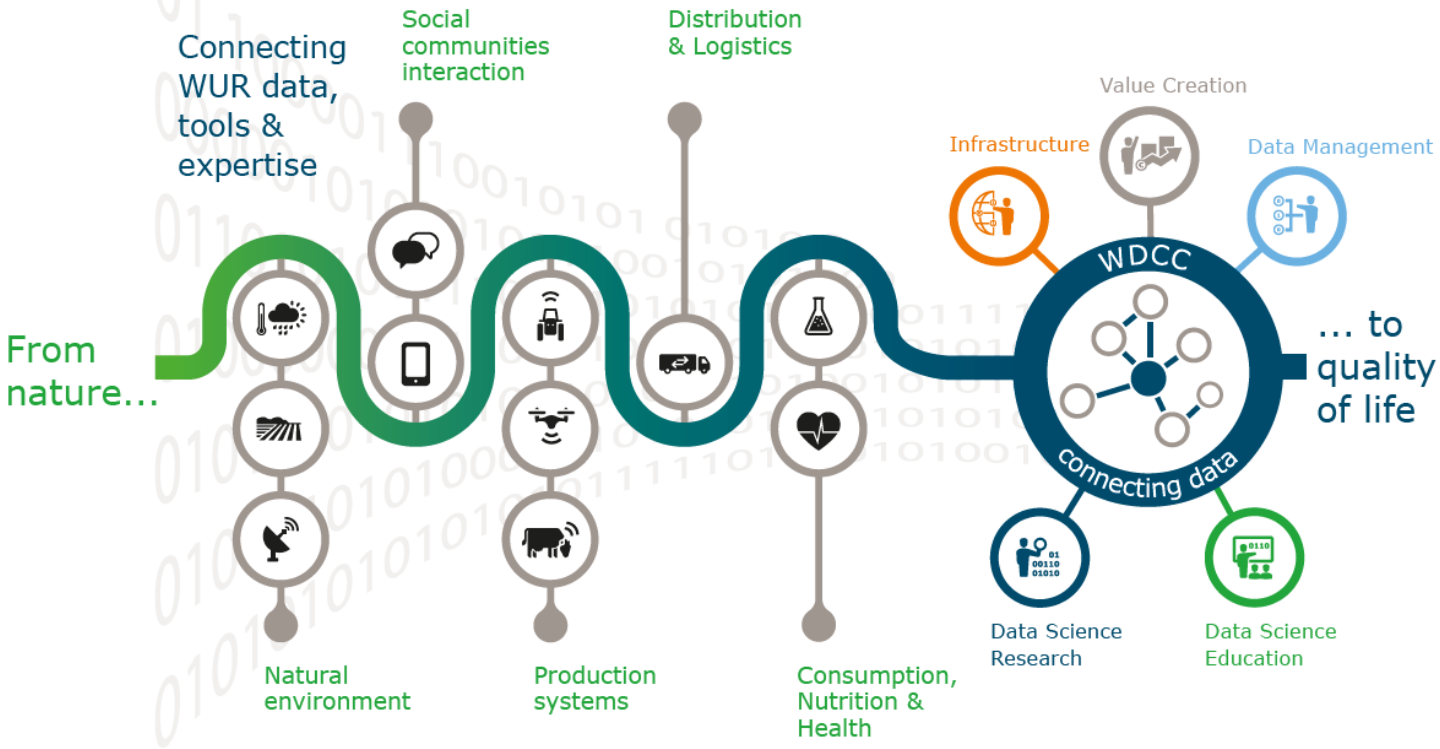
100years 1918 — 2018

Cell. 2015 Nov 19;163(5):1079-1094. doi: 10.1016/j.cell.2015.11.001.

Personalized Nutrition by Prediction of Glycemic Responses.

Zeevi D[1], Korem T[1], Zmora N[2], Israeli D[3], Rothschild D[1], Weinberger A[1], Ben-Yacov O[1], Lador D[1], Avnit-Sagi T[1], Lotan-Pompan M[1], Suez J[4], Mahdi JA[4], Matot E[1], Malka G[1], Kosower N[1], Rein M[1], Zilberman-Schapira G[4], Dohnalová L[4], Pevsner-Fischer M[4], Bikovsky R[1], Halpern Z[5], Elinav E[6], Segal E[7].

Wageningen Data Competence Centre



dr. WJJ (Willem Jan) Knibbe
Head of WDCC

drs.ing. J (Jacquelijn) Ringersma
Coordinator Data Management

dr.ir. L (Lukasz) Grus
Coordinator Education

dr.ir. CWPM (Chantal) Hukkelhoven
Coordinator Research

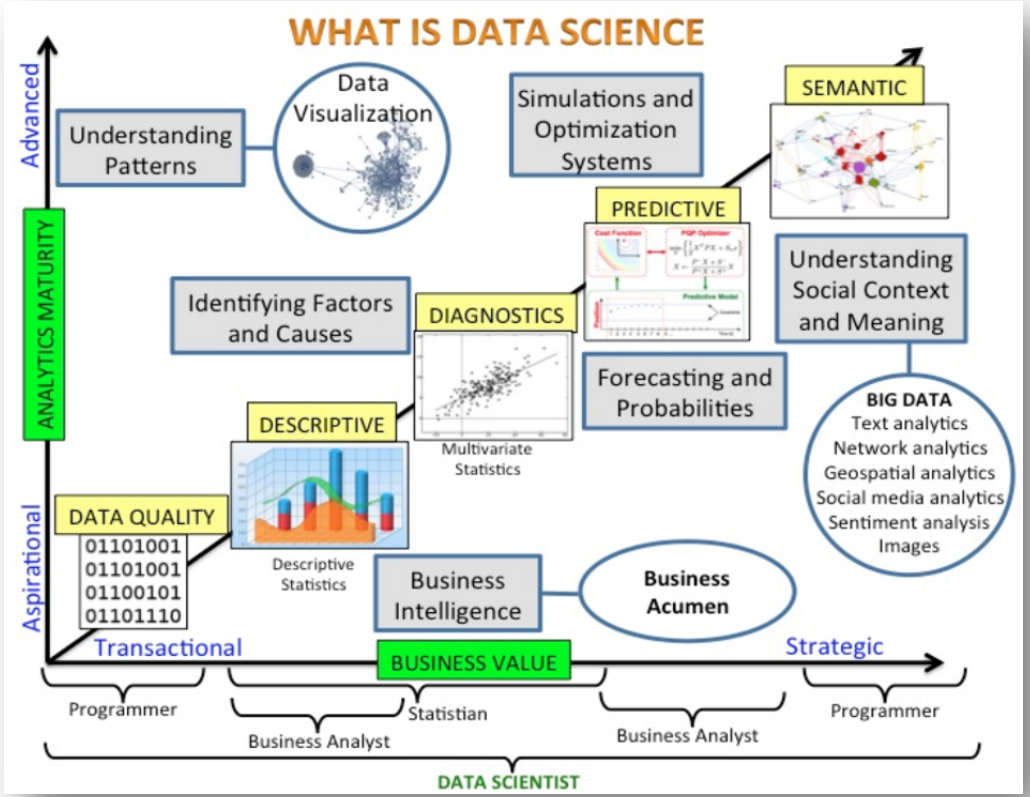
E (Erik) van den Bergh PhD
Coordinator Infrastructure

ir. HHM (Hilda) van der Pol-Sturm
Data Librarian

Considerations

(in relation to education)

Consideration 1 – Data science field is very broad



Vincent Granville, Data Science Central (2017)

Profile Data scientist - competences (adapted from EDISON project)

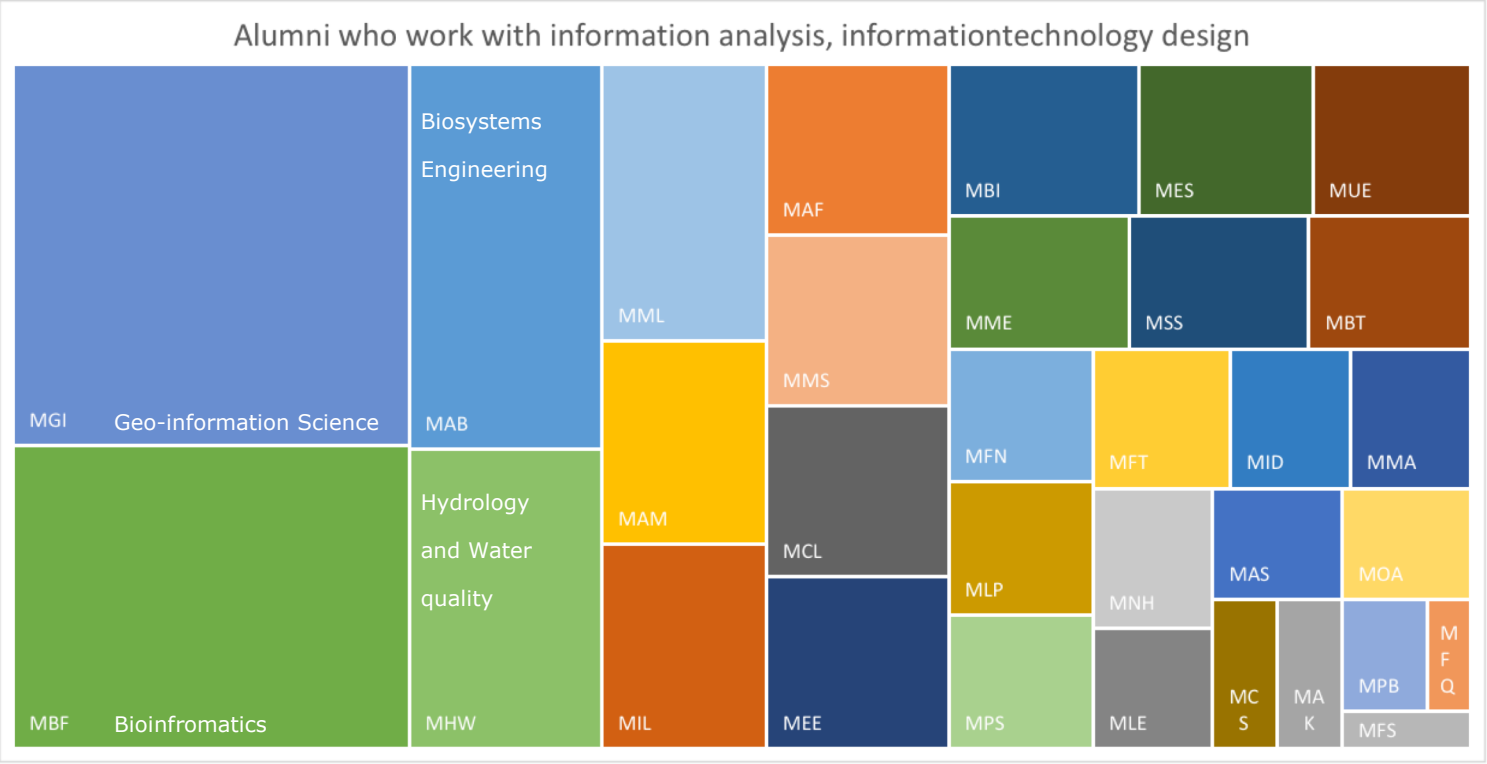
EDISON
building the data science profession

Data Analytics	Data Science Engineering	Data Management	Research Methods and Project Management	Domain related Competences
Use data analysis and statistical techniques on data *to deliver insights into research problem*	**Use engineering principles (software design and development) and modern computer technologies (programming) to research, design, implement new data analytics applications.**	**Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing**	**Create new understandings and capabilities by using the scientific method**	**Use domain knowledge to develop relevant data analytics applications**
Use techniques such as **Machine learning, Data Mining, Prescriptive and Predictive analytics**, for complex data analysis through the whole data cycle.	Use engineering principles (general and software) to research, **design, develop and implement instruments and applications** for data collection, storage, analysis and visualisation	Develop and implement data strategy, in particular, in a form of data management policy and Data Management Plan (DMP)	Create new understandings by using the research methods	**Analyse information needs, assess exisitng data and suggest/identify new data** required for specific context
Apply **statistics, time series analysis, optimization, simulation**, to deploy models for analysis and prediction	**Develop and apply computational solutions** to domain related problems using data analytics platforms, with the special focus on Big Data technologies and cloud based data analytics platforms	Develop and implement **data models**, define **metadata** using common standards and practices	Direct systematic study toward understanding of the observable facts, and discovers new methods	Operationalise fuzzy concepts to enable key performance indicators measurement to validate the research results or business analysis, identify and assess potential challenges
Identify, extract, and pull together available and pertinent **heterogeneous data, including modern data sources such as social media data, open data, governmental data**	**Develop and prototype specialised data analysis applicaions**, tools and supporting infrastructures for data driven scientific workflow.	**Integrate heterogeneous data** from multiple source and provide them for further analysis and use	Analyse domain related available data to **identify research questions and formulate sound hypothesis**	Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends
Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	**Develop, deploy and operate large scale data storage and processing solutions** using different distributed and cloud based platforms for storing data (e.g. Data Lakes, Hadoop, Hbase, Cassandra, MongoDB, Accumulo, DynamoDB, others)	Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)		Analyse opportunity and suggest use of historical data available in the study field or organization for creating new knowledge or optimization
Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making	**Consistently apply data security** mechanisms and controls at each stage of the data processing, including data anonymisation, privacy and IPR protection.	Ensure **data quality**, **accessibility**, **interoperability, compliance to standards, and publication** (data curation), **comply with FAIR principles**	Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	Analyse customer relations data to optimise/improve interacting with the specific user groups or in the specific business sectors
Visualise results of data analysis, design dashboard and use storytelling methods	Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	Develop and manage/supervise policies on **data protection, privacy, IPR and ethical issues** in data management	Develop and guide **data driven** projects, including project planning, experiment design, data collection and handling	Analyse multiple data sources for marketing purposes; identify effective marketing actions

Consideration 2

- **All** students share **interest and skills** in **Life Sciences**
- **Not all** students share **interest and skills** in **(Big) Data Science**

Consideration 2



Consideration 3

Know **Data science**

Is **Data scientist**

Consideration 3

Consideration 4

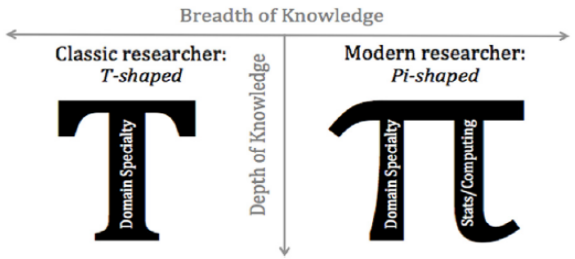


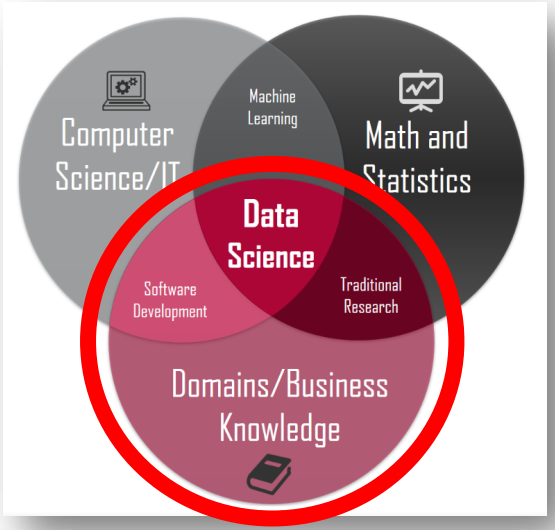
Fig. 1. Representations of T vs. Pi-shaped education.

This introductory course in data science is built on three interrelated perspectives: inferential thinking, computational thinking, and real-world relevance. Given data arising from some real-world phenomenon, how does one analyze that data so as to understand that phenomenon? How does one collect data to answer questions that one is interested in? Inferential thinking refers to an ability to connect data to underlying phenomena and to the ability to think critically about the conclusions that are drawn from data analysis. Computational thinking refers to the ability to conceive of the abstractions and processes that allow inferential procedures to be embodied in computer programs, and to ensure that such programs are scalable, robust and understandable. In addition to teaching critical concepts and skills in computer programming and statistical inference, the course will involve the hands-on analysis of a variety of real-world datasets, including economic data, document collections, geographical data and social networks, and it will delve into social and legal issues surrounding data analysis, including issues of privacy and data ownership.

Fig. 2. Syllabus of the undergraduate course Foundations of Data Science, Berkeley University (2015 edition).

Consideration 5

- Strong (Life Sciences) domain knowledge is an opportunity



Data Analytics	Data Science Engineering	Data Management	Research Methods and Project Management	Domain related Competences
Use data analysis and statistical techniques on data *to deliver insights into research problem*	**Use engineering principles (software design and development) and modern computer technologies (programming) to research, design, implement new data analytics applications.**	**Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing**	**Create new understandings and capabilities by using the scientific method**	**Use domain knowledge to develop relevant data analytics applications**
Use techniques such as **Machine learning, Data Mining, Prescriptive and Predictive analytics**, for complex data analysis through the whole data cycle.	Use engineering principles (general and software) to research, design, develop and implement new instruments and applications for data collection, storage, analysis and visualization	Develop and implement data strategy, in particular, in a form of data management policy and Data Management Plan (DMP)	Create new understandings by using the research methods	**Analyse information needs, assess exisitng data and suggest/identify new data** required for specific context
Apply **statistics, time series analysis, optimization, simulation**, to deploy models for analysis and prediction	Develop and apply computational solutions to domain related problems using data analytics platforms, with the special focus on Big Data technologies and cloud based data analytics platforms	Develop and implement **data models**, define **metadata** using common standards and practices	Direct systematic study toward understanding of the observable facts, and discovers new methods	Operationalise fuzzy concepts to enable key performance indicators measurement to validate the research results or business analysis, identify and assess potential challenges
Identify, extract, and pull together available and pertinent **heterogeneous data, including modern data sources such as social media data, open data, governmental data**	Develop and prototype specialised data analysis applicaions, tools and supporting infrastructures for data driven scientific workflow.	**Integrate heterogeneous data** from multiple source and provide them for further analysis and use	Analyse domain related available data to **identify research questions and formulate sound hypothesis**	Deliver business focused analysis using appropriate BA/BI methods and tools, identify business impact from trends; make business case as a result of organisational data analysis and identified trends
Understand and use different performance and accuracy metrics for model validation in analytics projects, hypothesis testing, and information retrieval	Develop, deploy and operate large scale data storage and processing solutions using different distributed and cloud based platforms for storing data (e.g. Data Lakes, Hadoop, Hbase, Cassandra, MongoDB, Accumulo, DynamoDB, others)	Maintain historical information on data handling, including reference to published data and corresponding data sources (data provenance)		Analyse opportunity and suggest use of historical data available in the study field or organization for creating new knowledge or optimization
Develop required data analytics for organizational tasks, integrate data analytics and processing applications into organization workflow and business processes to enable agile decision making	Consistently apply data security mechanisms and controls at each stage of the data processing, including data annonymisation, privacy and IPR protection.	Ensure **data quality, accessibility, interoperability, compliance to standards, and publication** (data curation), **comply with FAIR principles**	Design experiments which include data collection (passive and active) for hypothesis testing and problem solving	Analyse customer relations data to optimise/improve interacting with the specific user groups or in the specific business sectors
Visualise results of data analysis, design dashboard and use storytelling methods	Design, build, operate relational and non-relational databases (SQL and NoSQL), integrate them with the modern Data Warehouse solutions, ensure effective ETL (Extract, Transform, Load), OLTP, OLAP processes for large datasets	Develop and manage/supervise policies on **data protection, privacy, IPR and ethical issues** in data management	Develop and guide **data driven** projects, including project planning, experiment design, data collection and handling	Analyse multiple data sources for marketing purposes; identify effective marketing actions

Consideration 6

- Lots of bottom-up (research) and very specific DS education initiatives.
- Need for generic Data Science education offerings

SSB-53306 Life Sciences Information - Integrating, combining organizing and analyzing heterogeneous Biological information

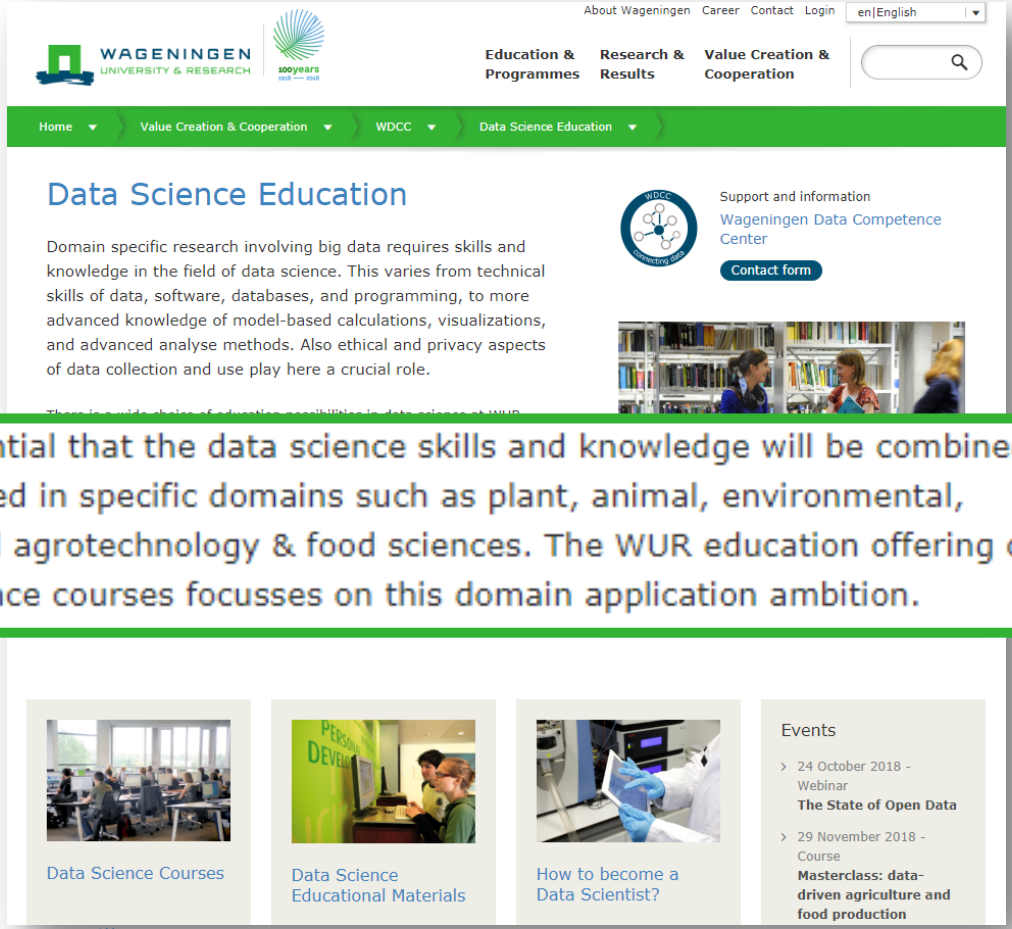
Learning outcomes:
After successful completion of this course students are expected to be able to:
- create a FAIR Data Management plan for complex biological experiments;
- design and perform advanced queries of (online) biological databases using SPARQL;
- design a custom-made ontology for different kinds of biological data;
- design a statistical experimental plan to address a complex biological question;
- extract relevant biological information using inferential approaches using R;
- apply data reduction techniques to analyse and investigate large biological data set.

INF-22803 Intro to Data Structures and Algorithms for Health

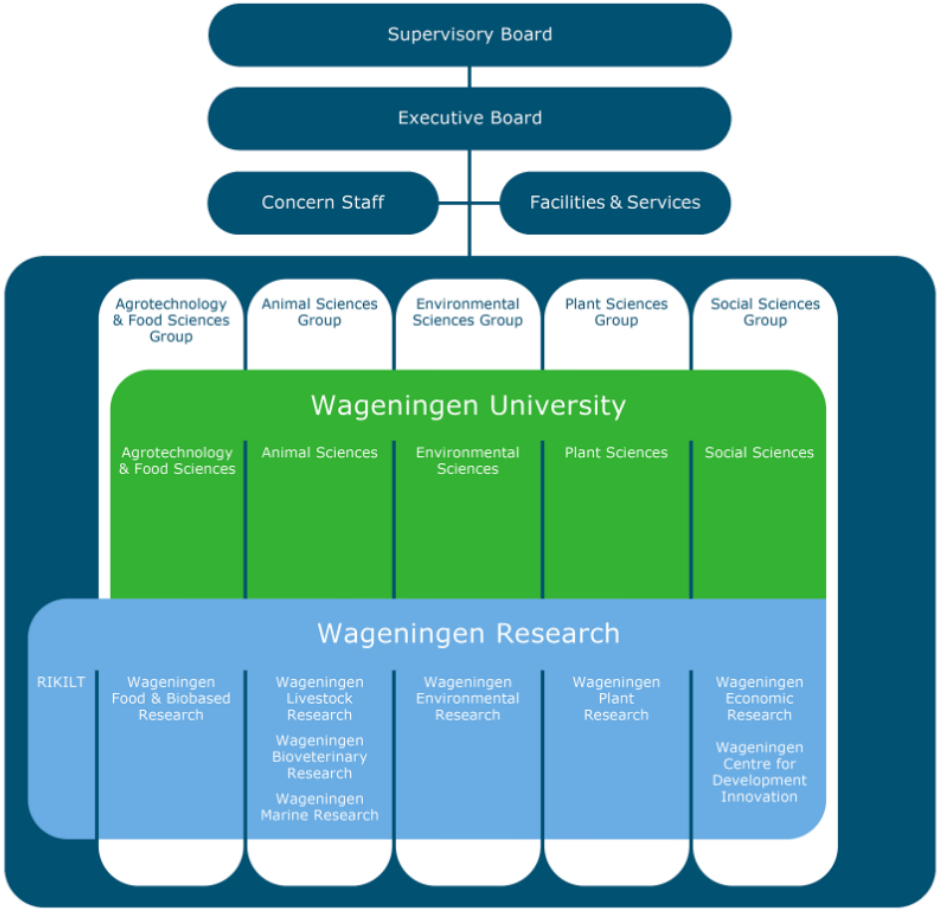
Learning outcomes:
After successful completion of this course students are expected to be able to:
- show insight in basic computer science terminology and concepts;
- identify several approaches to problem solving;
- explain key concepts of data structures and algorithms;
- compare and select suitable basic data structures for health related data;
- implement a proper data structure for health related data;
- analyse health related data problems using selected algorithms;
- solve health related data problems computationally;

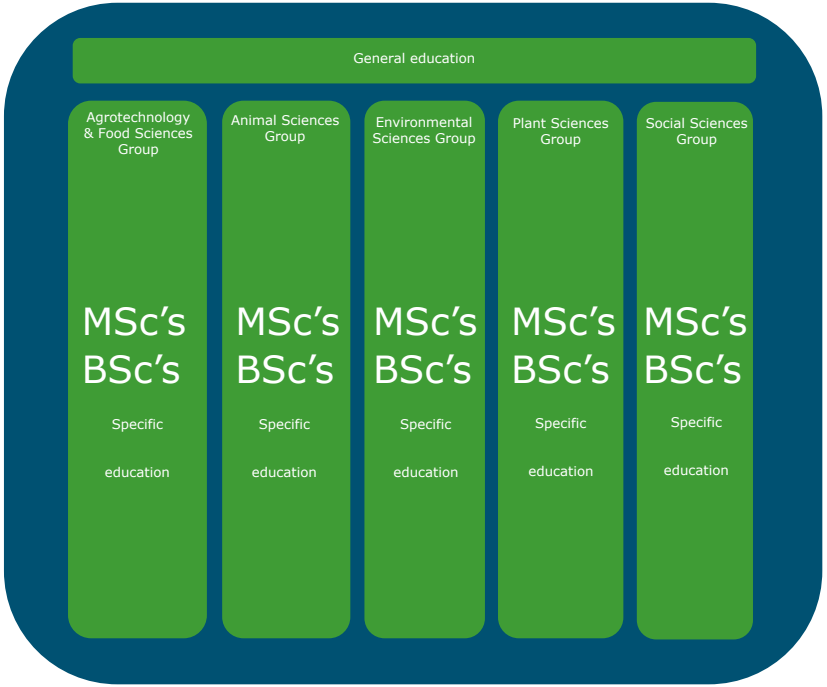
INF-33306 Linked Data

Learning outcomes:
After successful completion of this course students are expected to be able to:
- explain the research data life-cycle;
- explain linked data technologies (Internet, Semantic Web, ontologies, graph databases) and standards (as URI, XML, RDF, OWL, SPARQL) and their use in forming a Web of Data;
- use Linked Data technologies for retrieving information in the Semantic Web (i.e use of SPARQL);
- use of existing vocabularies for annotating research data and endpoints for finding data from the Life Sciences domain ;
- develop own ontologies and demonstrate the use of reasoners;
- debate why linked data is important and contemporary views on data stewardship, FAIR principles, data provenance, curation, ownership, open data and their application in the Life Sciences.



It is essential that the data science skills and knowledge will be combined and applied in specific domains such as plant, animal, environmental, social and agrotechnology & food sciences. The WUR education offering of data science courses focusses on this domain application ambition.


Supervisory Board
Executive Board
Concern Staff
Facilities & Services
Agrotechnology & Food Sciences Group
Animal Sciences Group
Environmental Sciences Group
Plant Sciences Group
Social Sciences Group
Wageningen University
Agrotechnology & Food Sciences
Animal Sciences
Environmental Sciences
Plant Sciences
Social Sciences
Wageningen Research
RIKILT
Wageningen Food & Biobased Research
Wageningen Livestock Research
Wageningen Bioveterinary Research
Wageningen Marine Research
Wageningen Environmental Research
Wageningen Plant Research
Wageningen Economic Research
Wageningen Centre for Development Innovation



WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018

Master's programmes

- Agroecology (European)
- Animal Sciences
- Aquaculture and Marine Resource Management
- Biobased Sciences
- Bioinformatics
- Biology
- Biosystems Engineering
- Biotechnology
- Climate Studies
- Communication, Health and Life Sciences
- Development and Rural Innovation
- Earth and Environment
- Environmental Sciences
- Food Quality Management
- Food Safety
- Food Studies (European)
- Food Technology
- Food Technology (online)
- Forest and Nature Conservation
- Geographical Information Management and Applications
- Geo-information Science
- International Development Studies
- International Land and Water Management
- Landscape Architecture and Planning
- Leisure, Tourism and Environment
- Management, Economics and Consumer Studies
- Metropolitan Analysis, Design, and Engineering
- Molecular Life Sciences
- Nutritional Epidemiology and Public Health (online)
- Nutrition and Health
- Organic Agriculture
- Plant Biotechnology
- Plant Breeding (online)
- Plant Sciences
- Statistical Sciences for the Life and Behavioural Sciences
- Urban Environmental Management
- Water Technology

Data science related courses

Information Technology (INF)
- Data Management
- Programming in Python
- Big Data
- Linked Data
- Applied Information Technology
- Software Engineering
- Computer Literacy
- Agent-Based Modelling of Complex Adaptive Systems
- Modelling and Simulation of Complex Socio-Technical Systems

Amsterdam Institute for Advanced Metropolitan Solutions (YMS)
- Metropolitan Data I
- Metropolitan Data II

Human nutrition (HNE)
- Applied Data Analysis

Systems and Synthetic Biology (SSB)
- Bioinformation Technology

Mathematical and Statistical Methods (MAT)
- Mathematics
- (Advanced) Statistics
- Advanced Statistics for Nutritionists
- R for Statistics

Library (LIB)
- Information Literacy
- Research Data Management (PhD / postdoc)

Farm Technology (FTE)
- Machine Learning
- Data Analysis Biosystems Engineering

Crop Systems Analysis (CSA)
- Ecological Modelling and Data Analysis in R

Bioinformatics (BIF)
- Advanced Bioinformatics
- Algorithms in Bioinformatics
- Biological Data Analysis and Visualisation
- Practical Computing for Biologists

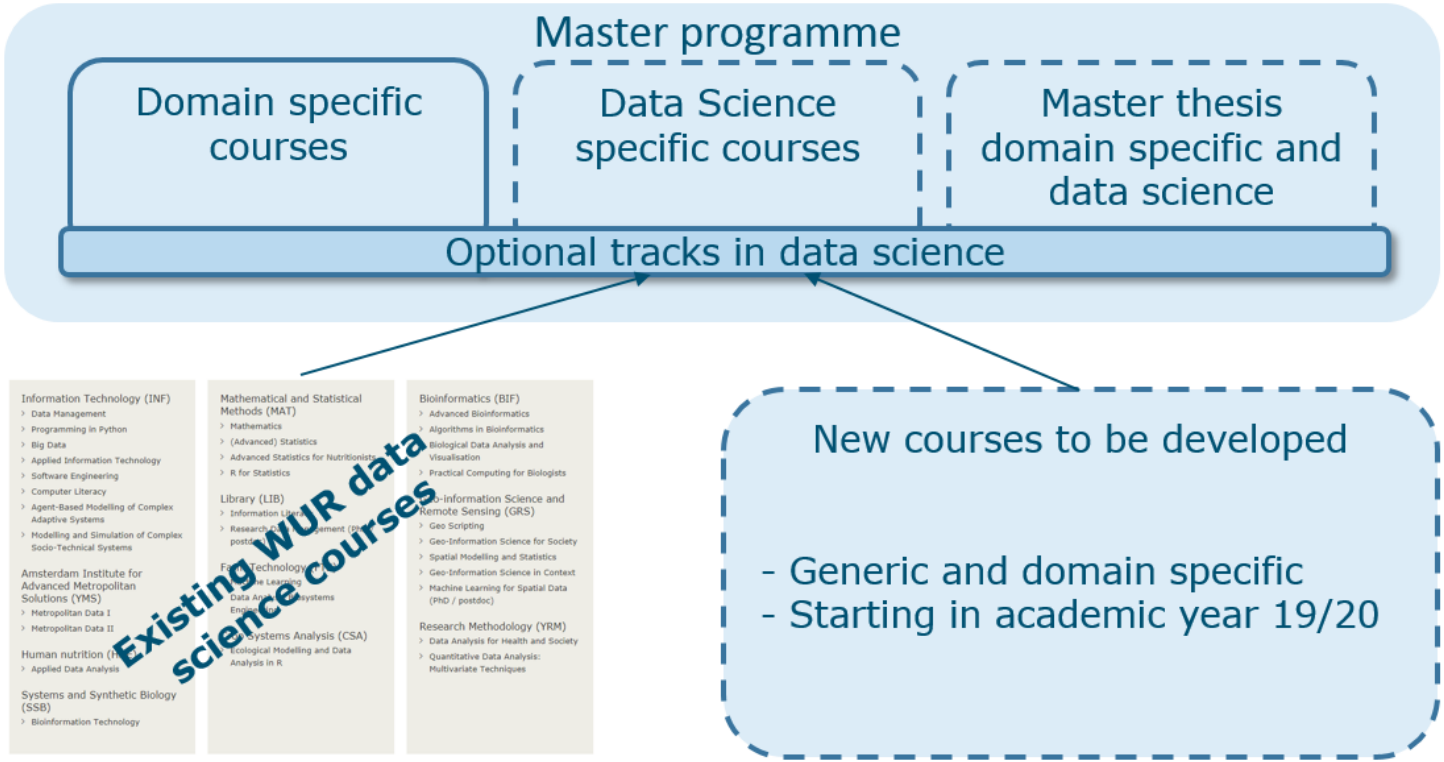
Geo-information Science and Remote Sensing (GRS)
- Geo Scripting
- Geo-Information Science for Society
- Spatial Modelling and Statistics
- Geo-Information Science in Context
- Machine Learning for Spatial Data (PhD / postdoc)

Research Methodology (YRM)
- Data Analysis for Health and Society
- Quantitative Data Analysis: Multivariate Techniques

Data Science minor

- Employ a **data lifecycle approach** to organize their data driven research;
- Identify appropriate **analysis techniques** when confronted with new datasets and questions;
- Design and **implement databases** for applications in the WUR domain;
- **Write** efficient and well-documented **computer programs**;
- Apply and evaluate relevant **visualization and quantitative data analysis methods**;
- Communicate research findings by **data storytelling**;
- Consider broader issues related to **data collection, processing and dissemination, including licensing and privacy issues**.

Data science **tracks** embedded in an application domains



Summary

- Data science is a broad domain;
- Strong domain knowledge as an opportunity for DS
- Not everyone must like it but most will probably need it;
- Dillema which knowledge level;
- T → ∏ education;
- Research groups recognize already the need for DS education
- € € €