Data Science

Session 1

tdi@cphbus

Intended Learning Today

What Do You Learn?

- ❑ The frame of the picture and the main characters in it
- ❑ What is the value of Data?
- ❑ Which are the stages in the workflow?
- ❑ What is an intelligent agent and its role?

What Do You Do?

- ❑ Work in teams
- ❑ Analyze use cases
- ❑ Define data categories
- ❑ Building a ML model
- ❑ Building an AI agent for games

Agenda

Data Science

Data Science Workflow

AI Agents

Hitting the road

Data Science

as a frame of a picture

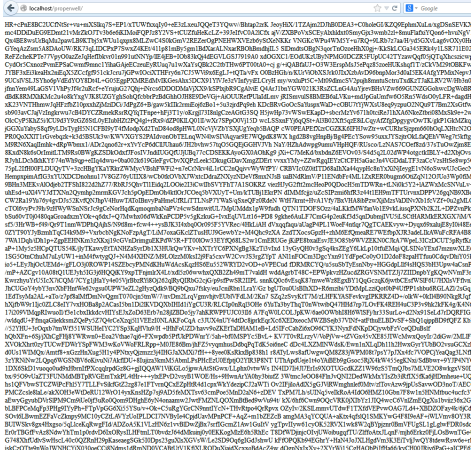
Data science is multifaceted and can be described as a science, as a research paradigm, as a research method, as a discipline, as a workflow, and as a profession.

Data science is an interdisciplinary field focused on extracting knowledge from typically large data sets and applying the knowledge and insights from that data to solve problems in a wide range of application domains.

Data science - Wikipedia

Data is Product

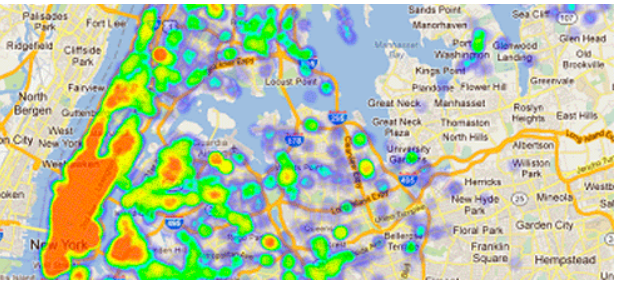
Typical Business Model



raw data

	Gender	Age	H1	F1	H2	F2	H3	F3	Height	Foot	Factor
1	Female	19.00	156.50	23.30	155.80	23.50	156.50	23.10	156.27	23.30	14.91
2	Male	21.00	165.50	25.50	165.30	24.50	166.50	25.60	165.77	25.20	15.20
3	Male	35.00	167.00	24.50	167.50	24.40	168.50	24.80	167.67	24.57	14.65
4	Female	19.00	158.50	24.60	158.50	23.50	158.70	23.50	158.57	23.87	15.05
5	Female	27.00	162.50	25.80	163.40	25.40	162.80	25.70	162.90	25.63	15.74
6	Female	19.00	159.50	23.50	158.00	23.40	158.00	23.60	158.50	23.50	14.83
7	Female	19.00	162.50	24.20	162.00	24.00	161.80	24.30	162.10	24.17	14.91
8	Female	20.00	165.40	23.90	165.00	23.80	166.00	23.80	165.47	23.83	14.40
9	Female	20.00	168.00	24.40	168.50	24.40	169.90	24.40	168.80	24.40	14.45
10	Female	22.00	159.20	24.70	159.50	24.50	160.00	24.50	159.57	24.57	15.40
11	Female	26.00	158.50	23.70	156.50	23.60	156.90	23.40	157.30	23.57	14.98
12	Female	19.00	162.00	24.30	161.00	24.60	160.90	24.40	161.30	24.43	15.15
13	Female	20.00	153.50	23.20	153.50	23.70	152.50	23.00	153.17	23.30	15.21
14	Female	21.00	158.00	24.90	157.80	24.60	158.50	24.90	158.10	24.80	15.69
15	Female	20.00	154.50	24.50	156.40	25.10	154.50	24.30	155.13	24.63	15.88
16	Female	20.00	163.00	23.00	161.50	24.00	163.50	23.50	162.67	23.50	14.45
17	Female	22.00	182.00	27.50	181.50	27.00	182.00	28.00	181.83	27.50	15.12
18	Male	27.00	175.00	27.50	178.50	27.50	176.50	27.50	176.67	27.50	15.57
19	Male	19.00	153.10	23.30	152.00	23.00	151.00	22.50	152.03	22.93	15.08
20	Female	22.00	170.00	25.00	172.50	25.60	170.50	25.50	171.00	25.37	14.83
21	Male	19.00	160.00	23.50	159.50	23.00	159.00	23.80	159.50	23.43	14.69
22	Female	20.00	155.50	23.20	154.40	23.00	153.50	22.50	154.47	22.90	14.83
23	Female	21.00	151.50	23.50	151.00	23.80	150.80	23.10	151.10	23.47	15.53
24	Female	21.00	158.00	23.50	158.50	24.20	157.80	23.50	158.10	23.73	15.01

processing/refinement



product/information

monetizing

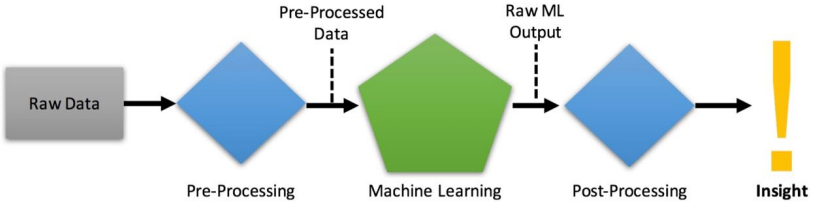
See also Data Marketplaces https://research.aimultiple.com/data-marketplace/

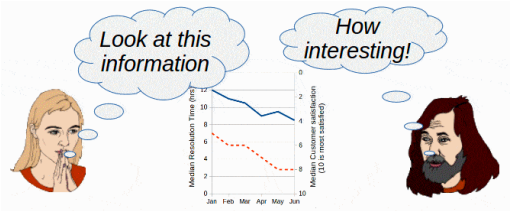
Data is Activator

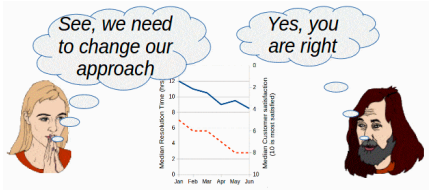
Every new insight drives new questions, new data, more analytics, and new insights!

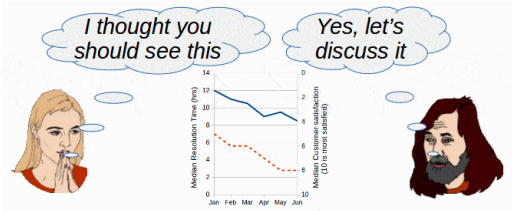
Data is Asset

- Informing
- Persuading
- Engaging



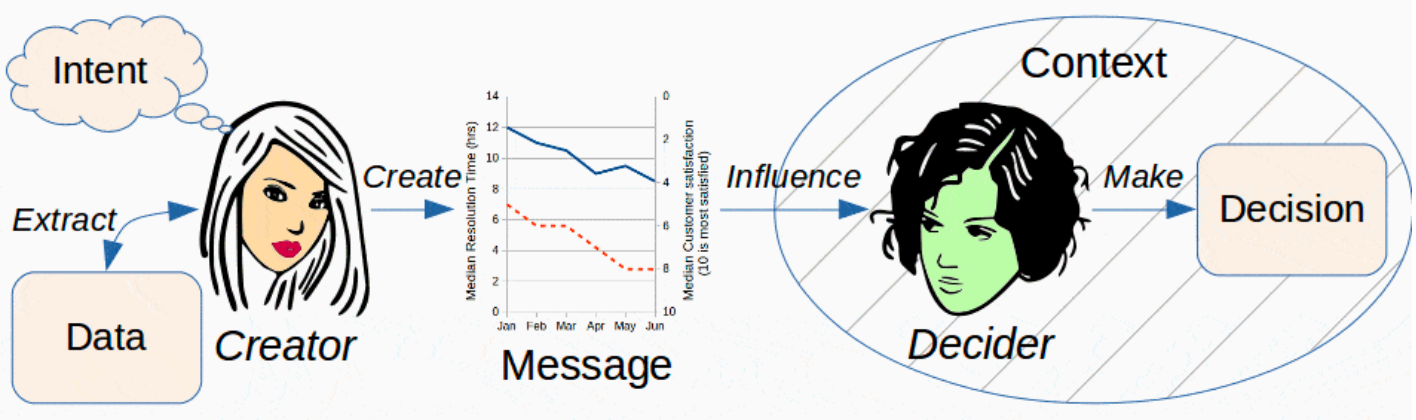




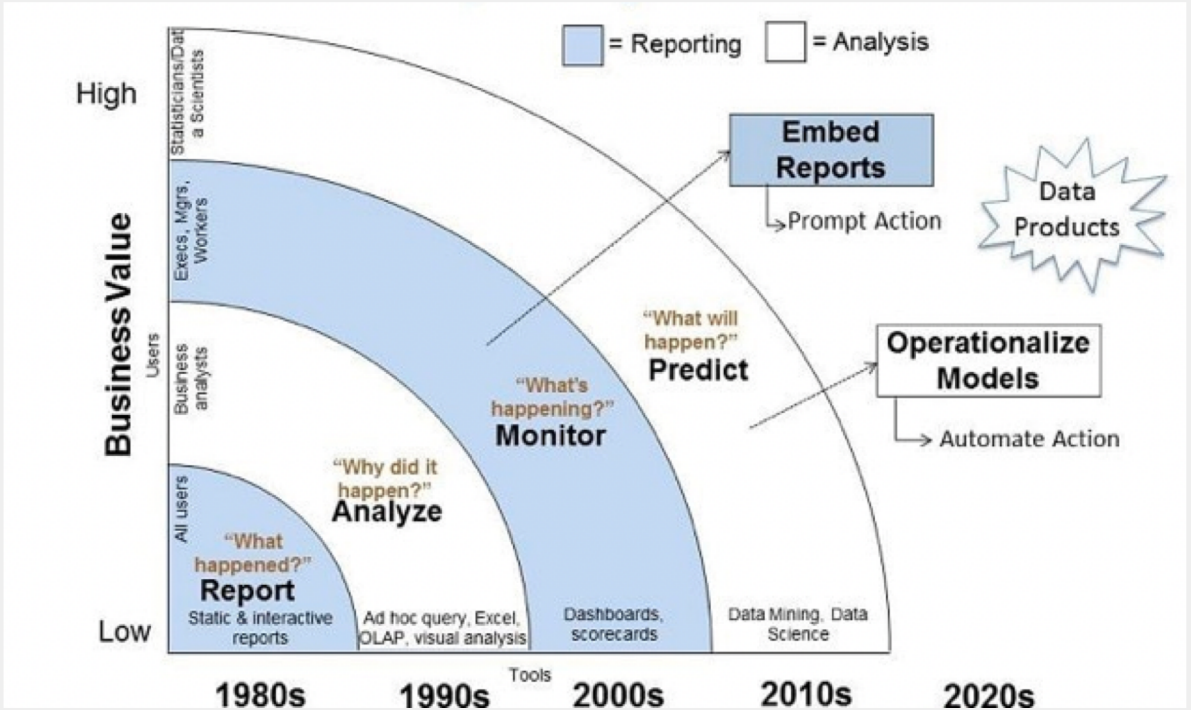


https://www.3cs.ch/information-visualization-data-visualization/

Every Company is Data Company



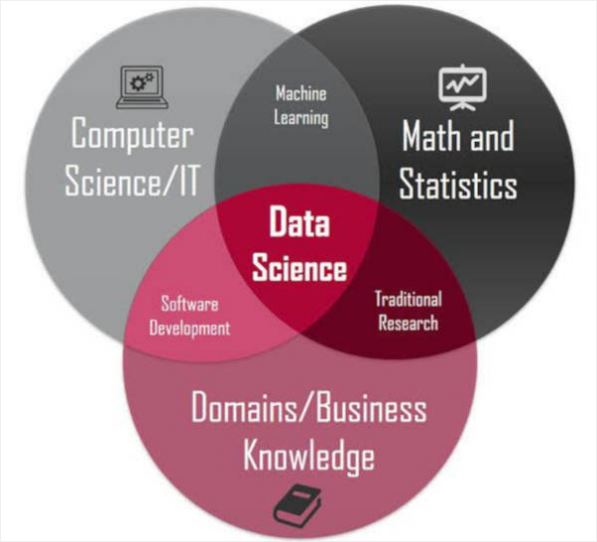
https://www.3cs.ch/information-visualization-data-visualization/



https://polestarllp.com/business-intelligence-trends-2020

Evolution

Inter-disciplinary knowledge



- What comes first?
- Where is the driving force?

Explore

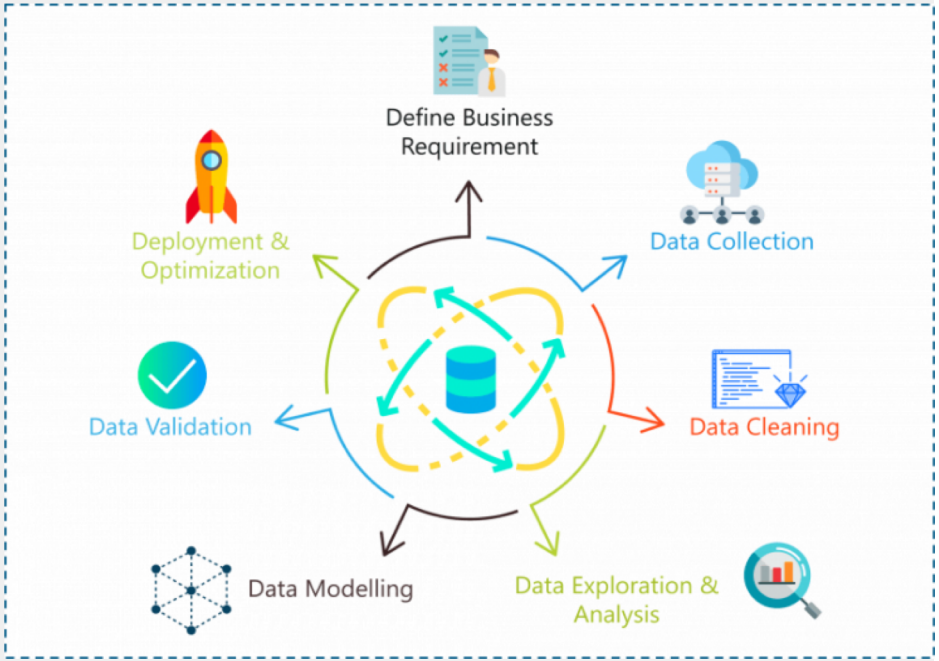
Find arguments for and against

- [ ] Go to https://github.com/metrica-sports/sample-data
  - explore the content of some data files
  - define as many use cases related to this data as you can for 10 minutes
- [ ] Work in pairs
- [ ] Share your experience and ideas with the others

Data Science Workflow

the process, the life cycle

Data Science Workflow

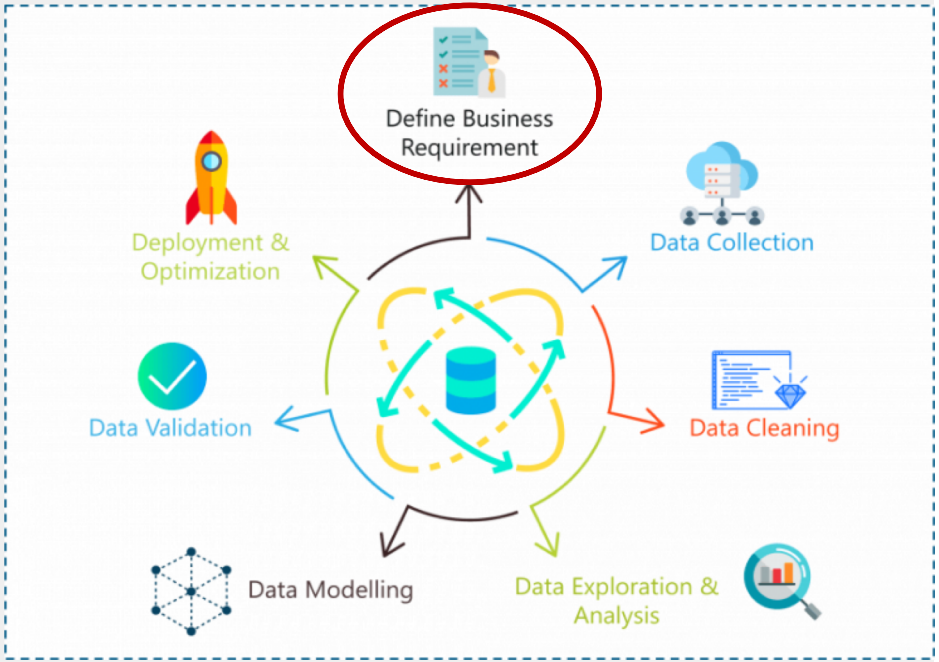


https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Data Science Workflow - 1

Define research goal
- what?
- why?
- how?

Start a project
- apply project management
  methodology
  team
  time
  resources
  risks



https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Example

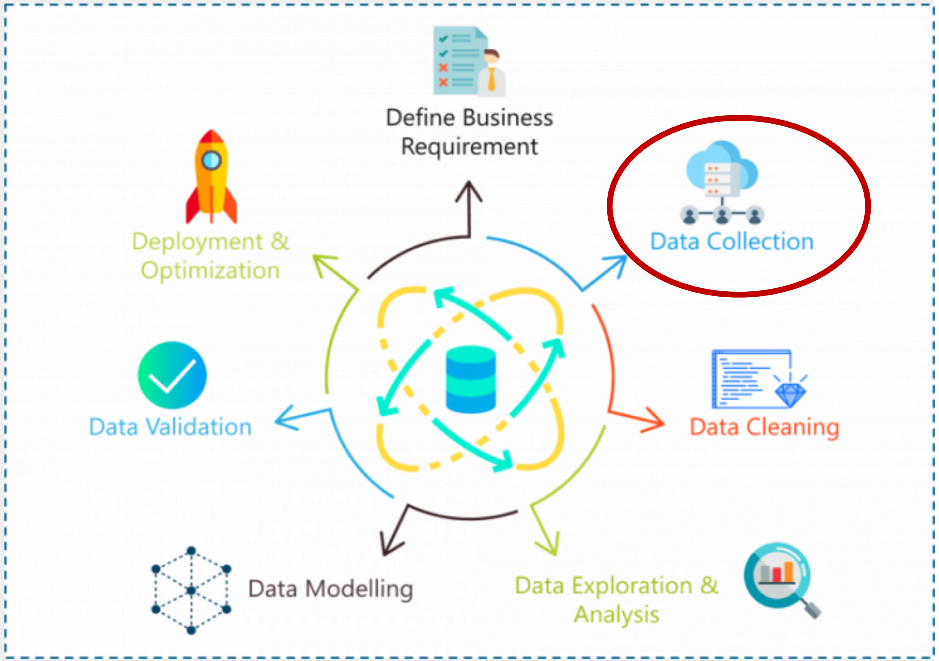
- First, go to https://github.com/datsoftlyngby/soft2022spring-DS/tree/main/Data/mercedesbenz-greener-manufacturing
- open the file `test.csv`
- What do you see?

- Next, go to https://www.kaggle.com/c/mercedes-benz-greener-manufacturing
- find the explanation
- Does the business context make it more clear?

Data Science Workflow - 2

Collect data from
- internal sources
  - files
  - databases
  - text documents
- external sources
  - web pages
  - public repositories
  - other projects and applications

Aggregate the collected data



https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Programming Examples

- Create code to read
  - CSV file
  - PDF file
  - scanned document
  - SQL database
  - JSON objects
  - image
- Create code to
  - scrape a web page
  - request data from API

- https://www.dst.dk/da/
- https://www.imdb.com/chart/top
- https://dawadocs.dataforsyningen.dk/dok/api

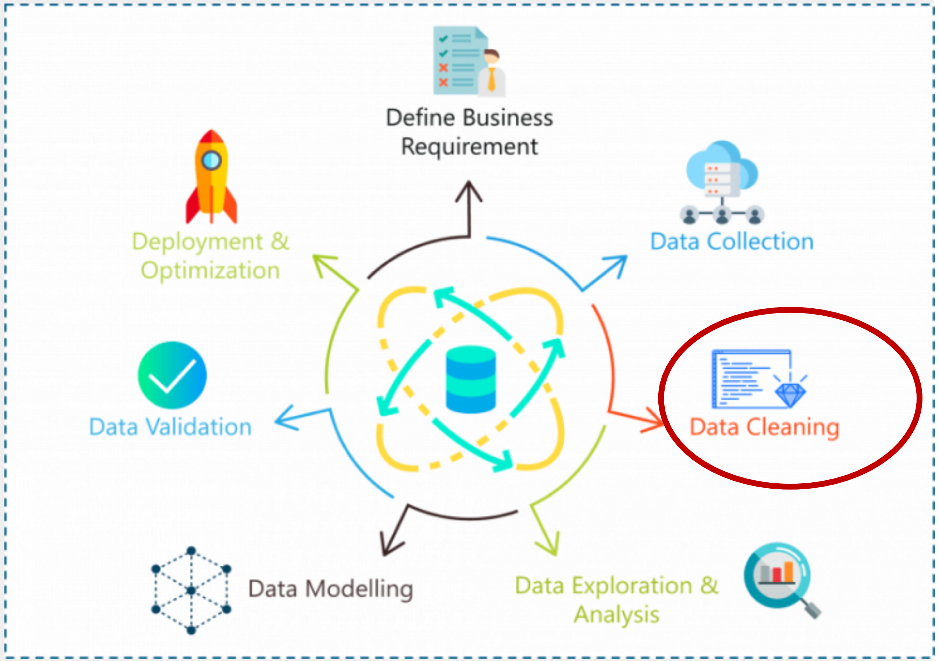
Data Science Workflow - 3

- Store the data into internal data structures

(like data frames, lists, arrays)

- Identify the dimensions and measurements
- Determine the types of the attributes
- Clean

missing values
outliers
wrong values
characters in strings



https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Programming Examples

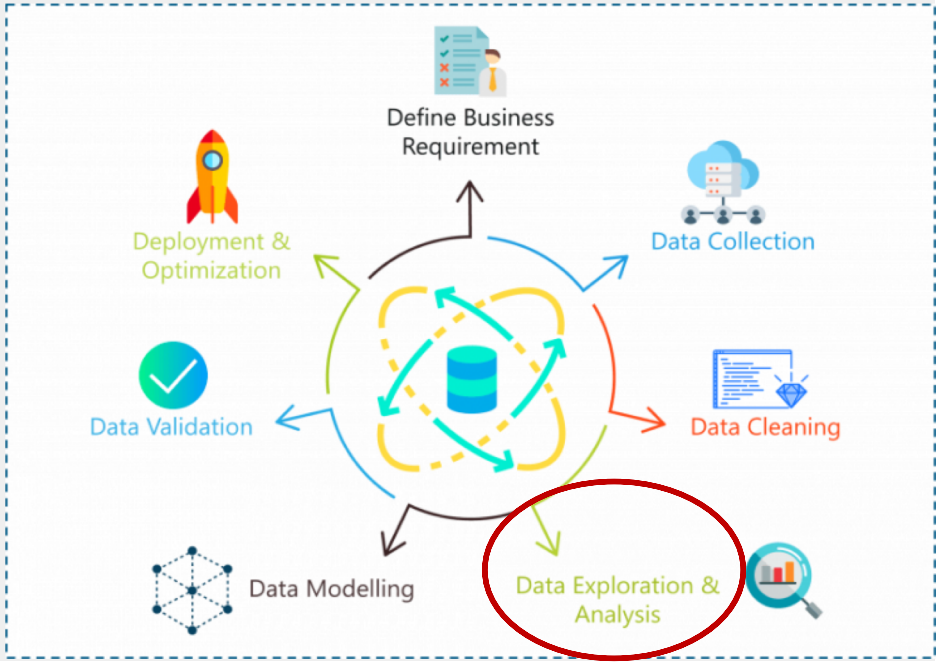
- Remove from text
  - punctuation
  - case
  - stop words
  - duplications
- Extend abbreviations
- Anonymise the sensitive data
  - replace it with fake values
  - hide or remove it
- Replace missing values with
  - average of the rest
  - mode of the rest
- Remove records with NULL values
- Transform data types and values

Data Science Workflow - 4

Exploratory data analysis for understanding, validation, and testing of the data

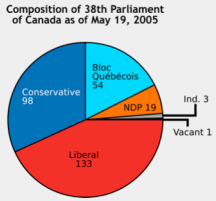
Get to know the data by applying measures from the Descriptive Statistics

- sample
- mean
- mode
- median
- range
- distribution



https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Programming Tasks



Suggest some methods of pre-processing

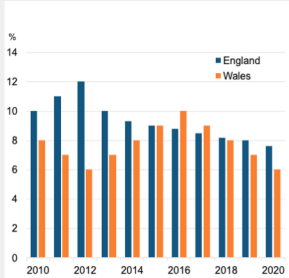
ID	X0	X1	X2	X3	X4	X5	X6	X8	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19	X20	X21	X22	X23	X24
1	az	v	n	f	d	t	a	w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	t	b	ai	a	d	b	g	y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	az	v	as	f	d	a	j	j	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
4	az	l	n	f	d	z	l	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	w	s	as	c	d	y	i	m	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
8	y	aa	ai	e	d	x	g	s	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
10	x	b	ae	d	d	x	d	y	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
11	f	s	ae	c	d	h	d	a	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
12	ap	l	s	c	d	h	j	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	o	v	as	f	d	g	f	v	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
15	ap	l	s	c	d	g	d	n	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

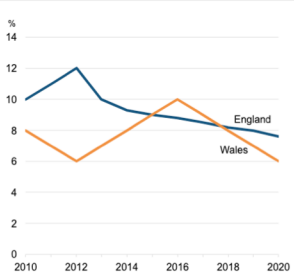
Data Engineering

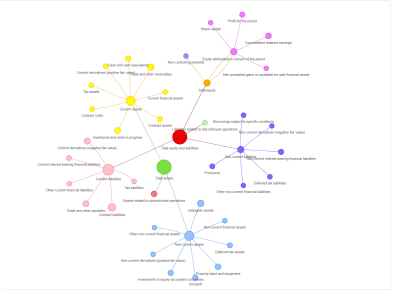
- feature importance
- dimensionality reduction

Data Visualisation

- create diagrams and graphs



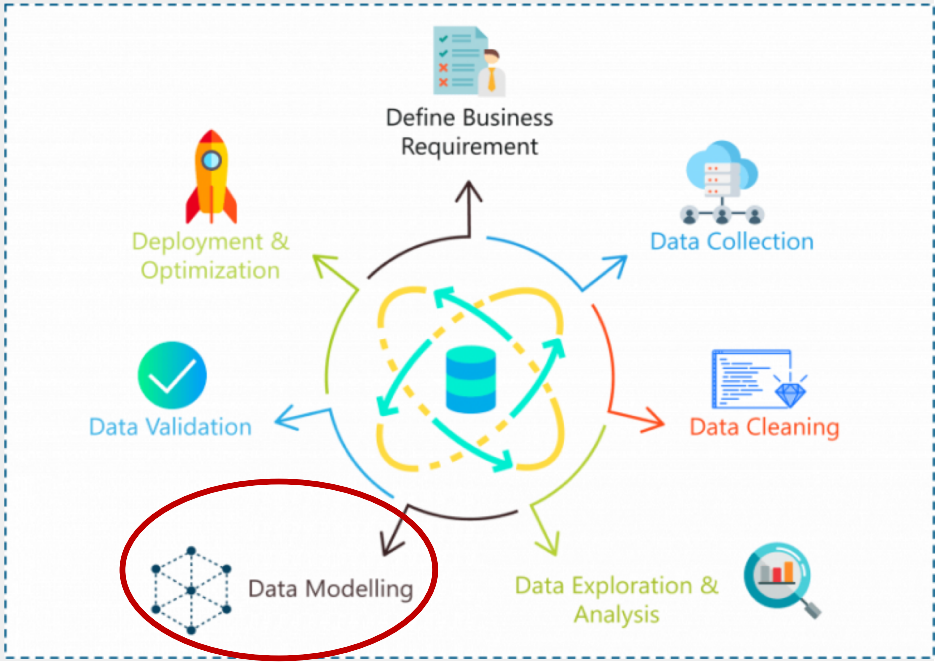




Data Science Workflow - 5

Select a machine learning model
- depending on the task and the data

Apply it for
- training
- testing
- evaluation

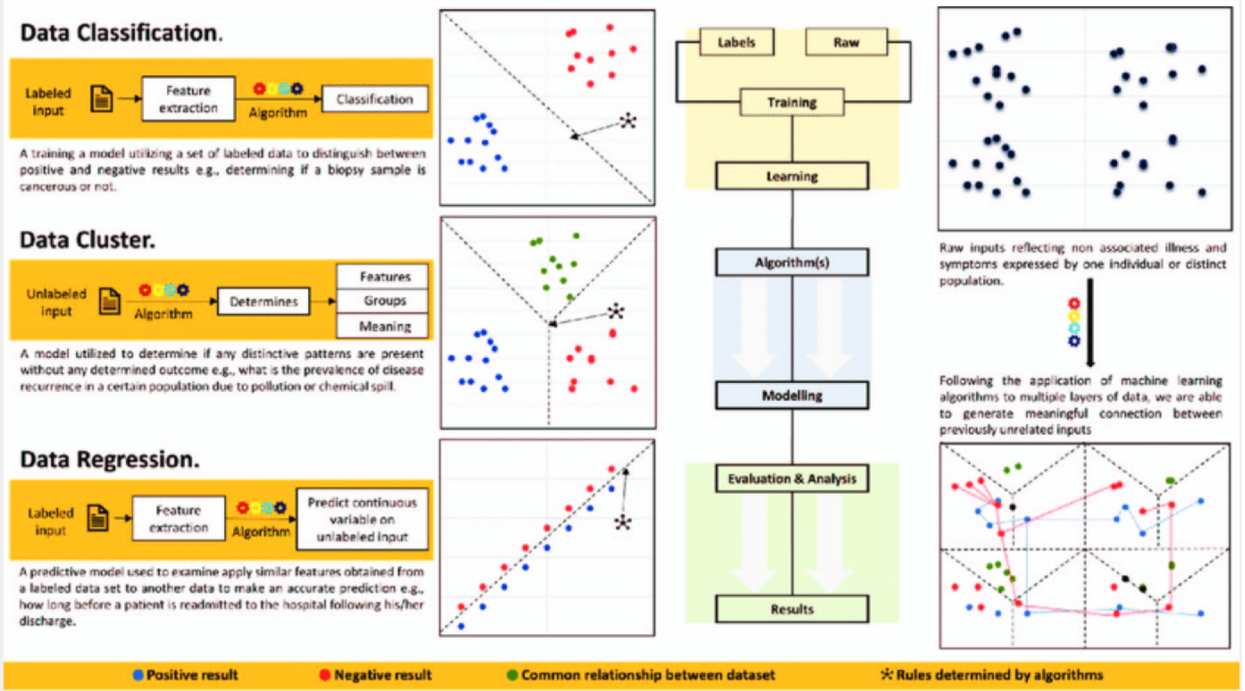


https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Data Modelling

- A data model is an abstract model that
  - organizes elements of data and
  - standardizes how they relate to one another and to the properties of real-world entities
- Determines the structure of data
- Specified in a data modelling notation

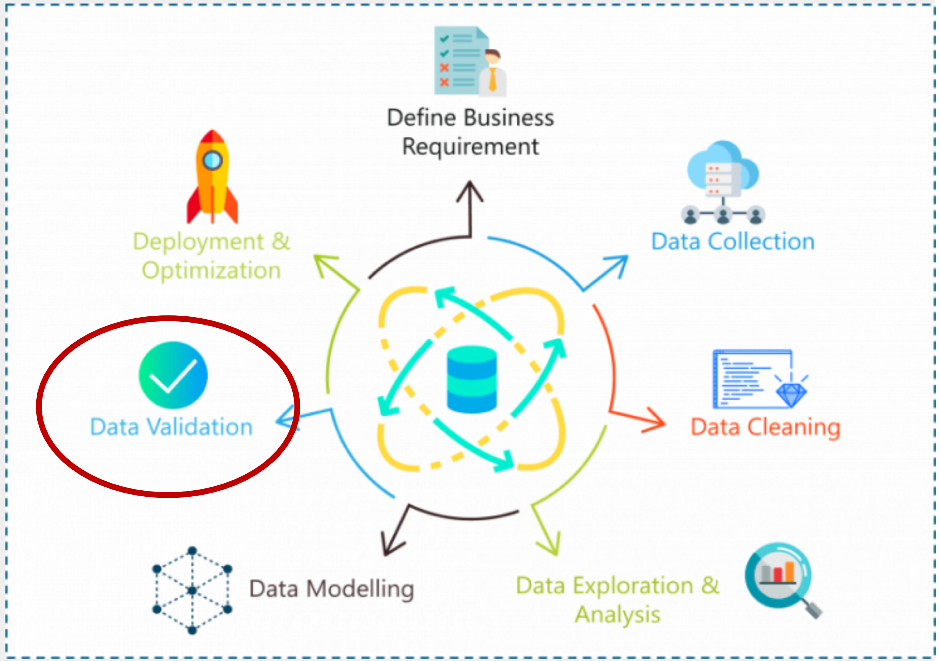
Predictive Analysis



researchgate.net

Data Science Workflow - 6

Applying the model for prediction or prescription of expected values and future behaviour
Reporting and testing



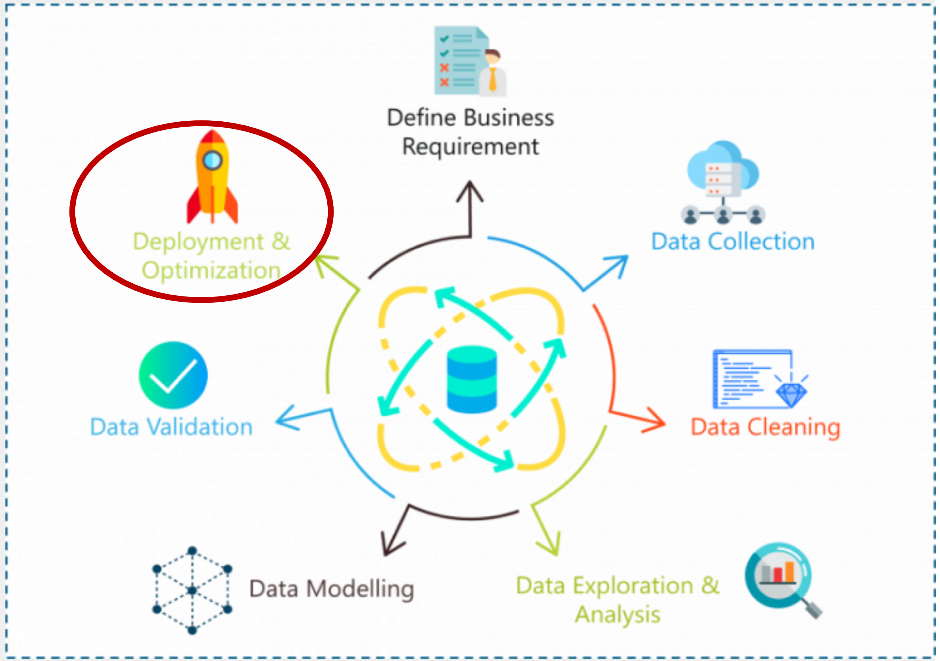
https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Programming Examples

- **Predictive analysis:** uncover trends
  - **Querying:** Asking the data specific questions, BI pulling the answers from the datasets
  - **Performance optimization:** Creating performance measures and customized dashboards

Data Science Workflow - 7

Deployment
Distribution
Dissemination
Visualisation
Collecting feedback
Performance and accuracy optimisation



https://datasciencepr.com/how-to-deliver-a-data-science-project-successfully-in-2020/

Programming Examples

- Creating application
- Enabling users to access it
- Reporting: Sharing data analysis to stakeholders in an appropriate format, so they can draw conclusions and make decisions
  - Data visualization: Turning data analysis into visual representations such as charts, graphs, histograms, and visual stories for business to more easily consume the data and the information of it
  - Interaction and usability: Creating approachable and understandable report, which can be used for self-service analytics

Data Science Use Cases

- IT Analytics
  - https://www.talend.com/customers/lenovo/
  - https://www.tableau.com/solutions/it-analytics
- Healthcare analytics
  - https://www.tableau.com/solutions/healthcare-analytics
  - https://www.tableau.com/covid-19-coronavirus-data-resources/healthcare-data-track
- Education
  - https://www.tableau.com/solutions/education-analytics
- https://experiments.withgoogle.com/collection/ai
- http://www.imaginarysoundscape.net
- http://places2.csail.mit.edu/demo.html
- https://dandelion.eu/semantic-text/sentiment-analysis-demo/?appid=us%3A333903271&exec=true

Books

- Data Science
- Machine Learning
- AI