Data science

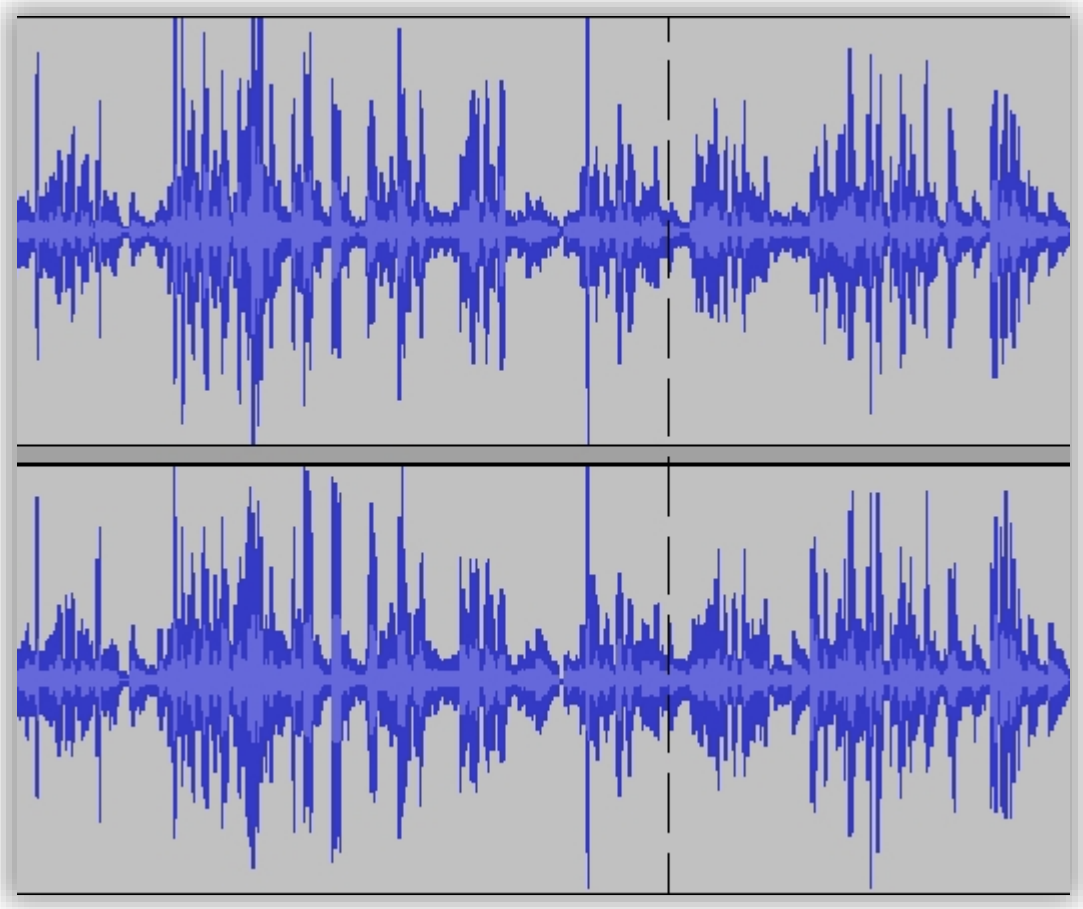
Chapter 1 – Introduction

2025-2026

What is data?

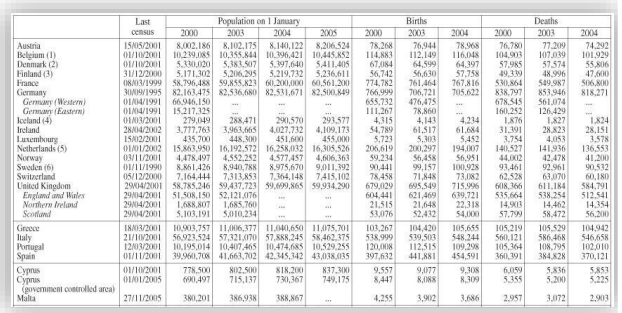
	Last census	Population on 1 January				Births			Deaths		
		2000	2003	2004	2005	2000	2003	2004	2000	2003	2004
Austria	15/05/2001	8,002,186	8,102,175	8,140,122	8,206,524	78,268	76,944	78,968	76,780	77,209	74,292
Belgium (1)	01/10/2001	10,239,085	10,355,844	10,396,421	10,445,852	114,883	112,149	116,048	104,903	107,039	101,929
Denmark (2)	01/10/2001	5,330,020	5,383,507	5,397,640	5,411,405	67,084	64,599	64,397	57,985	57,574	55,806
Finland (3)	31/12/2000	5,171,302	5,206,295	5,219,732	5,236,611	56,742	56,630	57,758	49,339	48,996	47,600
France	08/03/1999	58,796,488	59,855,823	60,200,000	60,561,200	774,782	761,464	767,816	530,864	549,987	506,800
Germany	30/09/1995	82,163,475	82,536,680	82,531,671	82,500,849	766,999	706,721	705,622	838,797	853,946	818,271
Germany (Western)	01/04/1991	66,946,150	...	...	...	655,732	476,475	...	678,545	561,074	...
Germany (Eastern)	01/04/1991	15,217,325	...	...	...	111,267	78,860	...	160,252	126,429	...
Iceland (4)	01/03/2001	279,049	288,471	290,570	293,577	4,315	4,143	4,234	1,876	1,827	1,824
Ireland	28/04/2002	3,777,763	3,963,665	4,027,732	4,109,173	54,789	61,517	61,684	31,391	28,823	28,151
Luxembourg	15/02/2001	435,700	448,300	451,600	455,000	5,723	5,303	5,452	3,754	4,053	3,578
Netherlands (5)	01/01/2002	15,863,950	16,192,572	16,258,032	16,305,526	206,619	200,297	194,007	140,527	141,936	136,553
Norway	03/11/2001	4,478,497	4,552,252	4,577,457	4,606,363	59,234	56,458	56,951	44,002	42,478	41,200
Sweden (6)	01/11/1990	8,861,426	8,940,788	8,975,670	9,011,392	90,441	99,157	100,928	93,461	92,961	90,532
Switzerland	05/12/2000	7,164,444	7,313,853	7,364,148	7,415,102	78,458	71,848	73,082	62,528	63,070	60,180
United Kingdom	29/04/2001	58,785,246	59,437,723	59,699,865	59,934,290	679,029	695,549	715,996	608,366	611,184	584,791
England and Wales	29/04/2001	51,508,150	52,121,076	...	...	604,441	621,469	639,721	535,664	538,254	512,541
Northern Ireland	29/04/2001	1,688,807	1,685,760	...	...	21,515	21,648	22,318	14,903	14,462	14,354
Scotland	29/04/2001	5,103,191	5,010,234	...	...	53,076	52,432	54,000	57,799	58,472	56,200
Greece	18/03/2001	10,903,757	11,006,377	11,040,650	11,075,701	103,267	104,420	105,655	105,219	105,529	104,942
Italy	21/10/2001	56,923,524	57,321,070	57,888,245	58,462,375	538,999	539,503	548,244	560,121	586,468	546,658
Portugal	12/03/2001	10,195,014	10,407,465	10,474,685	10,529,255	120,008	112,515	109,298	105,364	108,795	102,010
Spain	01/11/2001	39,960,708	41,663,702	42,345,342	43,038,035	397,632	441,881	454,591	360,391	384,828	370,121
Cyprus	01/10/2001	778,500	802,500	818,200	837,300	9,557	9,077	9,308	6,059	5,836	5,853
Cyprus (government controlled area)	01/01/2005	690,497	715,137	730,367	749,175	8,447	8,088	8,309	5,355	5,200	5,225
Malta	27/11/2005	380,201	386,938	388,867	...	4,255	3,902	3,686	2,957	3,072	2,903

What is data?

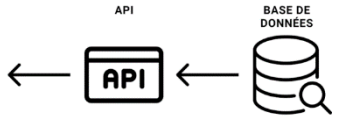


What is data?

What is data?

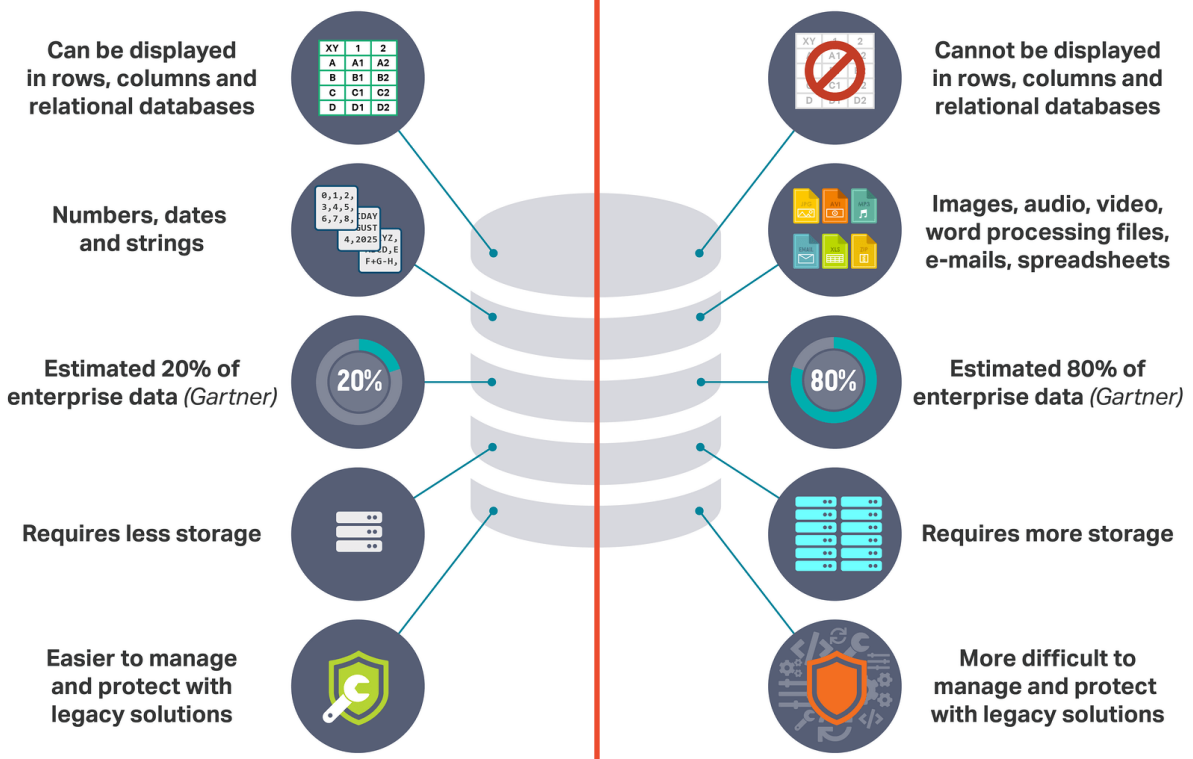


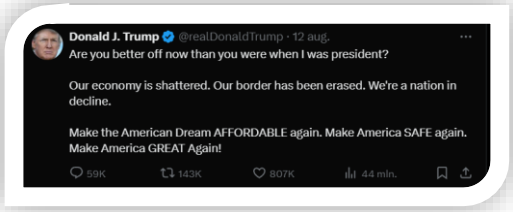
	Last census	Population on 1 January				Births				Deaths			
		2000	2003	2004	2005	2000	2003	2004	2000	2003	2004
Austria	15/05/2001	8,002,186	8,102,175	8,140,122	8,206,524	78,268	76,944	78,968	76,780	77,209	74,292
Belgium (1)	01/10/2001	10,239,085	10,355,844	10,396,421	10,445,852	114,883	112,149	116,048	104,903	107,039	101,929
Denmark (2)	01/10/2001	5,330,020	5,383,507	5,397,640	5,411,405	67,084	64,599	64,397	57,985	57,574	55,806
Finland (3)	31/12/2000	5,171,302	5,206,295	5,219,732	5,236,611	56,742	56,630	57,758	49,339	48,996	47,600
France	08/03/1999	58,796,488	59,855,823	60,200,000	60,561,200	774,782	761,464	767,816	530,864	549,987	506,800
Germany	30/09/1995	82,163,475	82,536,680	82,531,671	82,500,849	766,999	706,721	705,622	838,797	853,946	818,271
Germany (Western)	01/04/1991	66,946,150	...	...	...	655,732	676,475	...	678,545	561,074	...
Germany (Eastern)	01/04/1991	15,217,325	...	...	...	111,267	78,860	...	160,252	126,429	...
Iceland (4)	01/03/2001	279,049	288,471	290,570	293,577	4,315	4,143	4,234	1,876	1,827	1,824
Ireland	28/04/2002	3,777,763	3,963,665	4,027,732	4,109,173	54,789	61,517	61,684	31,391	28,823	28,151
Luxembourg	15/02/2001	435,700	448,300	451,600	455,000	5,723	5,303	5,452	3,754	4,053	3,578
Netherlands (5)	01/01/2002	15,863,950	16,192,572	16,258,032	16,305,526	206,619	200,297	194,007	140,527	141,936	136,553
Norway	03/11/2001	4,478,497	4,552,252	4,577,457	4,606,363	59,234	56,458	56,951	44,002	42,478	41,200
Sweden (6)	01/11/1990	8,861,426	8,940,788	8,975,670	9,011,392	90,441	99,157	100,928	93,461	92,961	90,532
Switzerland	05/12/2000	7,164,444	7,313,853	7,364,148	7,415,102	78,458	71,848	73,082	62,528	63,070	60,180
United Kingdom	29/04/2001	58,785,246	59,437,723	59,699,865	59,934,290	679,029	695,549	715,996	608,366	611,184	584,791
England and Wales	29/04/2001	51,508,150	52,121,076	...	...	604,441	621,469	639,721	535,664	538,254	512,541
Northern Ireland	29/04/2001	1,688,807	1,685,760	...	...	21,515	21,648	22,318	14,903	14,462	14,354
Scotland	29/04/2001	5,103,191	5,010,234	...	...	53,076	52,432	54,000	57,799	58,472	56,200
Greece	18/03/2001	10,903,757	11,006,377	11,040,650	11,075,701	103,267	104,420	105,655	105,219	105,529	104,942
Italy	21/10/2001	56,923,524	57,321,070	57,888,245	58,462,375	538,999	539,503	548,244	560,121	586,468	546,658
Portugal	12/03/2001	10,195,014	10,407,465	10,474,685	10,529,255	120,008	112,515	109,298	105,364	108,795	102,010
Spain	01/11/2001	39,960,708	41,663,702	42,345,342	43,038,035	397,632	441,881	454,591	360,391	384,828	370,121
Cyprus	01/10/2001	778,500	802,500	818,200	837,300	9,557	9,077	9,308	6,059	5,836	5,853
Cyprus (government controlled area)	01/01/2005	690,497	715,137	730,367	749,175	8,447	8,088	8,309	5,355	5,200	5,225
Malta	27/11/2005	380,201	386,938	388,867	...	4,255	3,902	3,686	2,957	3,072	2,903

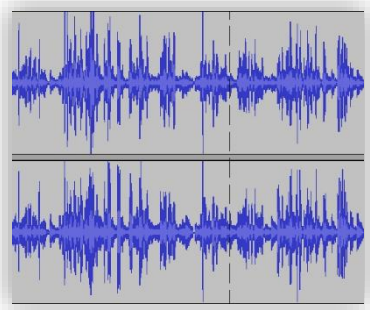



```
[
  {
    "id": 1,
    "name": "Leanne Graham
    "username": "Bret",
    "email": "Sincere@apri
    "address": {
      "street": "Kulas Light",
      "suite": "Apt. 556",
      "city": "Gwenborough",
      "zipcode": "92998-3874",
      "geo": {
        "lat": "-37.3159",
        "lng": "81.1496"
      }
    },
    "phone": "1-770-736-8031 x56442",
    "website": "hildegard.org",
    "company": {
      "name": "Romaguera-Crona",
      "catchPhrase": "Multi-layered client-server neural-net",
      "bs": "harness real-time e-markets"
    }
  },
  {
```


Structured Data vs Unstructured Data



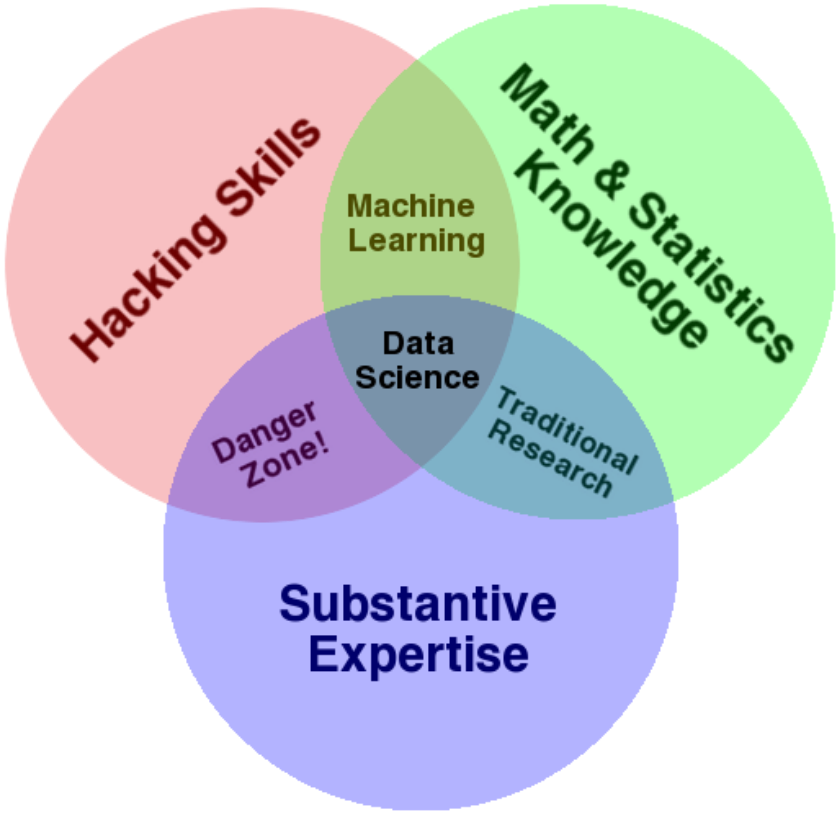




Focus in this course

Deep learning

What is data science?

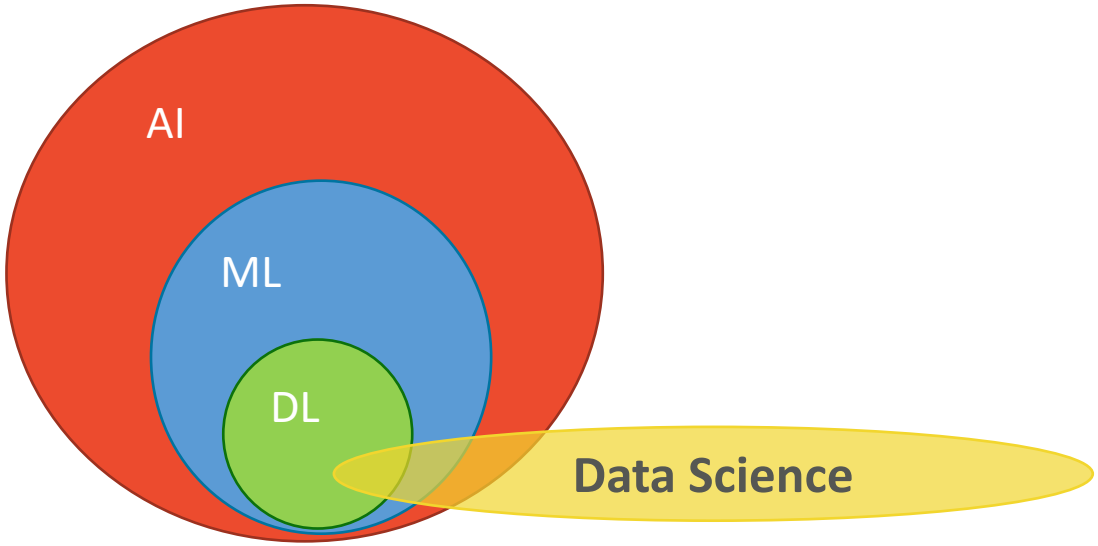


Data science combines math and statistics, specialized programming skills with specific subject matter expertise to uncover actionable insights hidden in an organization's data.
(Definition IBM)

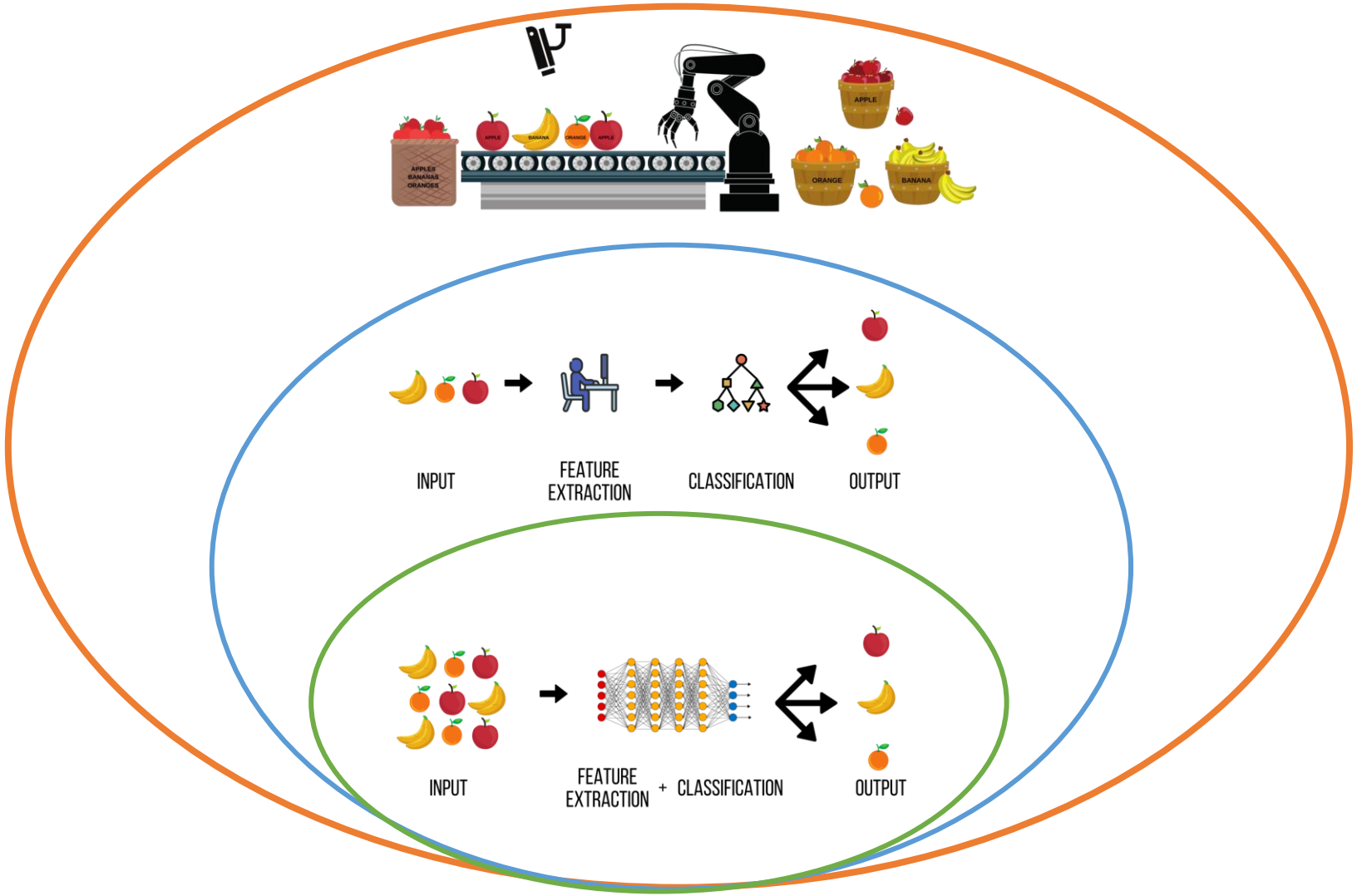
Buzzword Bingo: AI – ML – DL

- **AI** - Artificial intelligence: techniques to mimic human intelligence.
- **ML** - Machine learning: subset of AI to have machines doing a specific task better and better by gaining experience.
- **DL** - Deep learning: subset of ML that uses artificial neural networks with many layers.

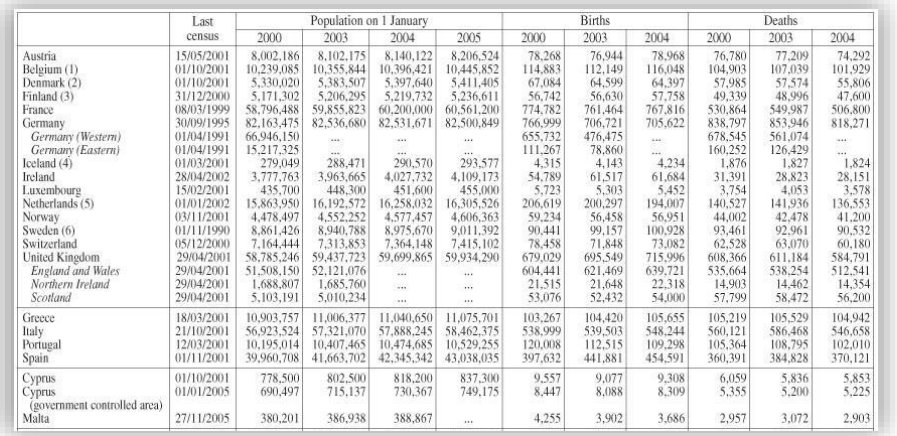
First step to do AI/ML/DL well is data science. You need to understand your data.



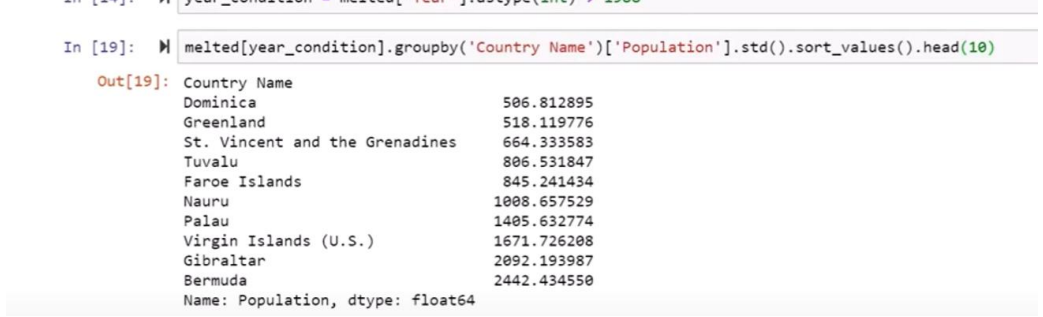
Buzzword Bingo: AI – ML – DL



Data science or machine learning?



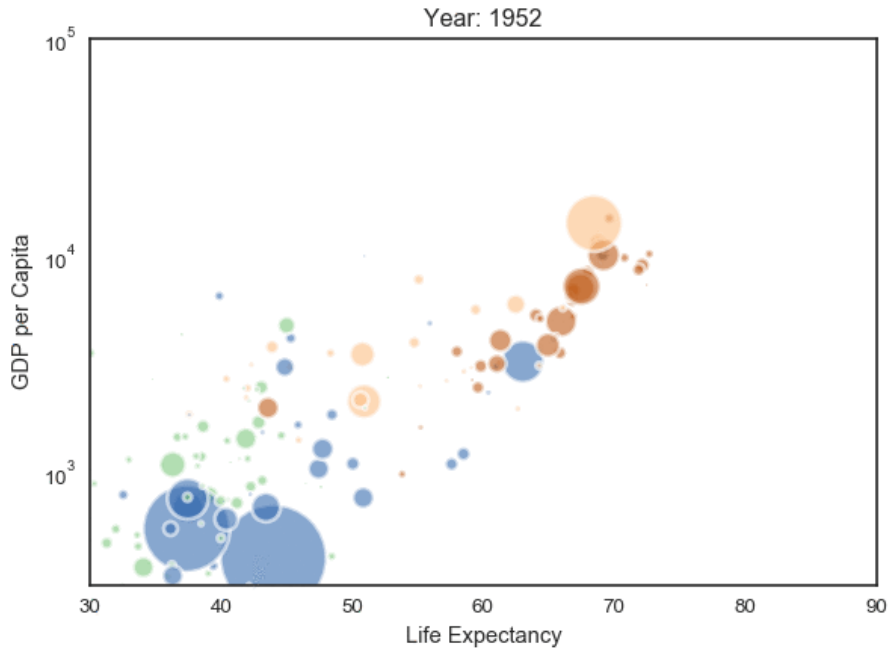
	Last census	Population on 1 January				Births			Deaths		
		2000	2003	2004	2005	2000	2003	2004	2000	2003	2004
Austria	15/05/2001	8,002,186	8,102,175	8,140,122	8,206,524	78,268	76,944	78,968	76,780	77,209	74,292
Belgium (1)	01/10/2001	10,239,085	10,355,844	10,396,421	10,445,852	114,883	112,149	116,048	104,903	107,039	101,929
Denmark (2)	01/10/2001	5,330,020	5,383,507	5,397,640	5,411,405	67,084	64,599	64,397	57,985	57,574	55,806
Finland (3)	31/12/2000	5,171,302	5,206,295	5,219,732	5,236,611	56,742	56,630	57,758	49,339	48,996	47,600
France	08/03/1999	58,796,488	59,855,823	60,200,000	60,561,200	774,782	761,464	767,816	530,864	549,987	506,800
Germany	30/09/1995	82,163,475	82,536,680	82,531,671	82,500,849	766,999	706,721	705,622	838,797	853,946	818,271
Germany (Western)	01/04/1991	66,946,150	...	...	...	655,732	476,475	...	678,545	561,074	...
Germany (Eastern)	01/04/1991	15,217,325	...	...	...	111,267	78,860	...	160,252	126,429	...
Iceland (4)	01/03/2001	279,049	288,471	290,570	293,577	4,315	4,143	4,234	1,876	1,827	1,824
Ireland	28/04/2002	3,777,763	3,963,665	4,027,732	4,109,173	54,789	61,517	61,684	31,391	28,823	28,151
Luxembourg	15/02/2001	435,700	448,300	451,600	455,000	5,723	5,303	5,452	3,754	4,053	3,578
Netherlands (5)	01/01/2002	15,863,950	16,192,572	16,258,032	16,305,526	206,619	200,297	194,007	140,527	141,936	136,553
Norway	03/11/2001	4,478,497	4,552,252	4,577,457	4,606,363	59,234	56,458	56,951	44,002	42,478	41,200
Sweden (6)	01/11/1990	8,861,426	8,940,788	8,975,670	9,011,392	90,441	99,157	100,928	93,461	92,961	90,532
Switzerland	05/12/2000	7,164,444	7,313,853	7,364,148	7,415,102	78,458	71,848	73,082	62,528	63,070	60,180
United Kingdom	29/04/2001	58,785,246	59,437,723	59,699,865	59,934,290	679,029	695,549	715,996	608,366	611,184	584,791
England and Wales	29/04/2001	51,508,150	52,121,076	...	...	604,441	621,469	639,721	535,664	538,254	512,541
Northern Ireland	29/04/2001	1,688,807	1,685,760	...	...	21,515	21,648	22,318	14,903	14,462	14,354
Scotland	29/04/2001	5,103,191	5,010,234	...	...	53,076	52,432	54,000	57,799	58,472	56,200
Greece	18/03/2001	10,903,757	11,006,377	11,040,650	11,075,701	103,267	104,420	105,655	105,219	105,529	104,942
Italy	21/10/2001	56,923,524	57,321,070	57,888,245	58,462,375	538,999	539,503	548,244	560,121	586,468	546,658
Portugal	12/03/2001	10,195,014	10,407,465	10,474,685	10,529,255	120,008	112,515	109,298	105,364	108,795	102,010
Spain	01/11/2001	39,960,708	41,663,702	42,345,342	43,038,035	397,632	441,881	454,591	360,391	384,828	370,121
Cyprus	01/10/2001	778,500	802,500	818,200	837,300	9,557	9,077	9,308	6,059	5,836	5,853
Cyprus (government controlled area)	01/01/2005	690,497	715,137	730,367	749,175	8,447	8,088	8,309	5,355	5,200	5,225
Malta	27/11/2005	380,201	386,938	388,867	...	4,255	3,902	3,686	2,957	3,072	2,903



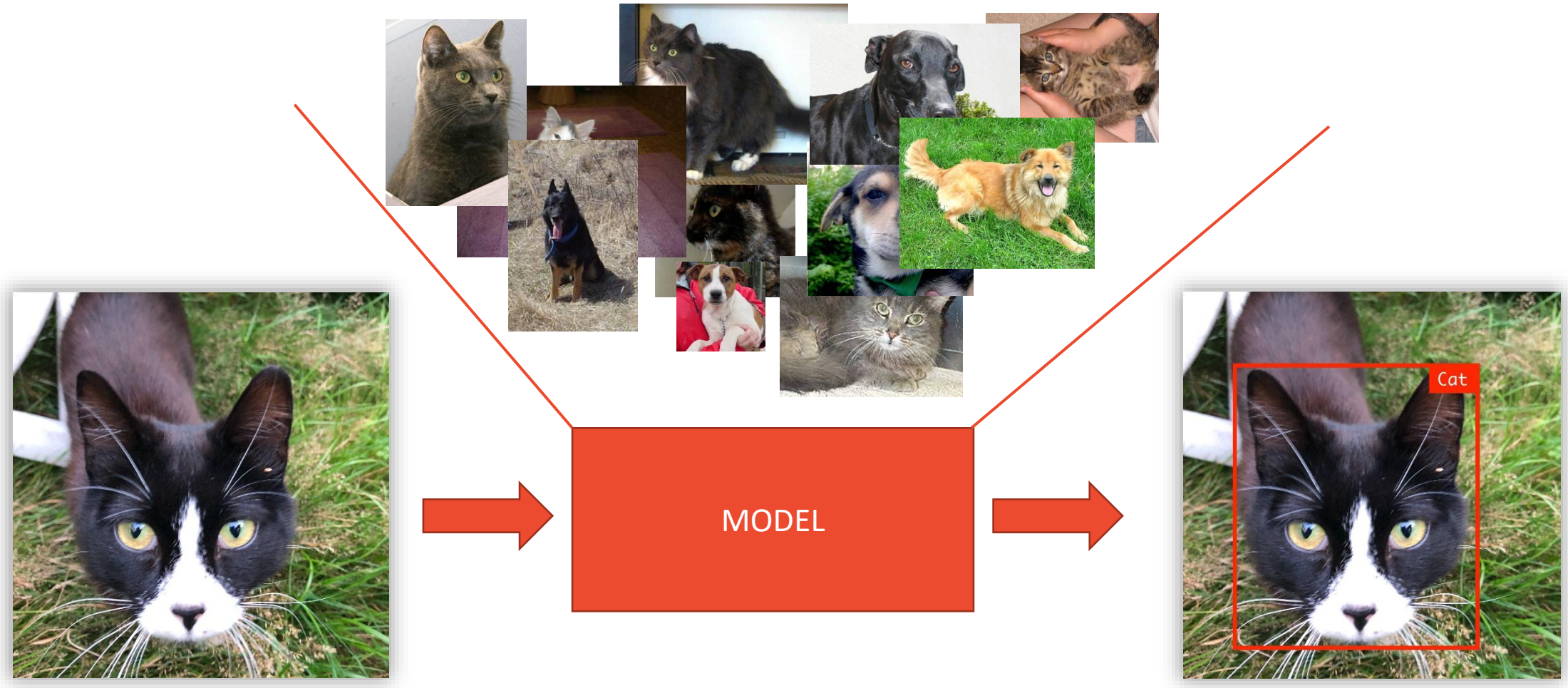
Data science or machine learning?

- By doing data manipulations, you can build animated visualisations
- Data visualisation is a separate course in the 2nd term.

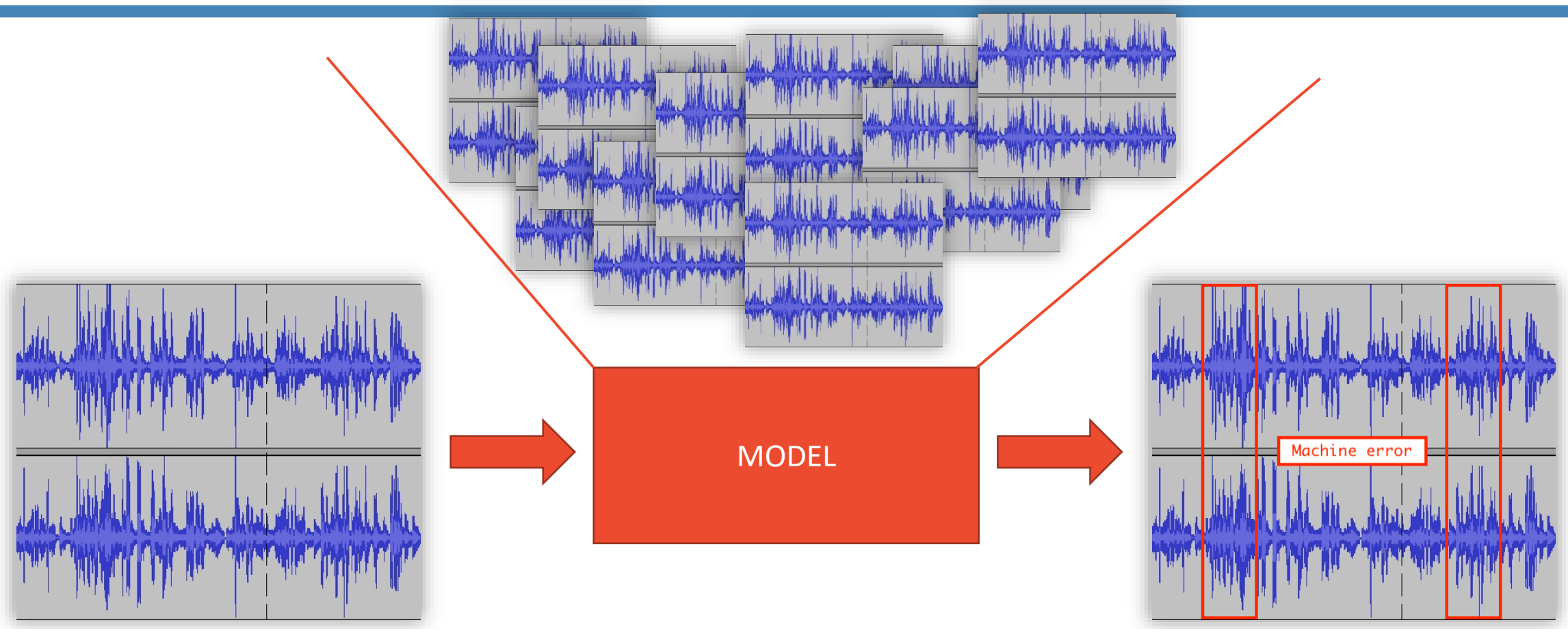
	Last census	Population on 1 January				Births			Deaths		
		2000	2003	2004	2005	2000	2003	2004	2000	2003	2004
Austria	15/05/2001	8,002,186	8,102,175	8,140,122	8,206,524	78,268	76,944	78,968	76,780	77,209	74,292
Belgium (1)	01/10/2001	10,239,085	10,355,844	10,396,421	10,445,852	114,883	112,149	116,048	104,903	107,039	101,929
Denmark (2)	01/10/2001	5,330,020	5,383,507	5,397,640	5,411,405	67,084	64,599	64,397	57,985	57,574	55,806
Finland (3)	31/12/2000	5,171,302	5,206,295	5,219,732	5,236,611	56,742	56,630	57,758	49,339	48,996	47,600
France	08/03/1999	58,796,488	59,855,823	60,200,000	60,561,200	774,782	761,464	767,816	530,864	549,987	506,800
Germany	30/09/1995	82,163,475	82,536,680	82,531,671	82,500,849	766,999	706,721	705,622	838,797	853,946	818,271
Germany (Western)	01/04/1991	66,946,150	...	...	...	655,732	476,475	...	678,545	561,074	...
Germany (Eastern)	01/04/1991	15,217,325	...	...	...	111,267	78,860	...	160,252	126,429	...
Iceland (4)	01/03/2001	279,049	288,471	290,570	293,577	4,315	4,143	4,234	1,876	1,827	1,824
Ireland	28/04/2002	3,777,763	3,963,665	4,027,732	4,109,173	54,789	61,517	61,684	31,391	28,823	28,151
Luxembourg	15/02/2001	435,700	448,300	451,600	455,000	5,723	5,303	5,452	3,754	4,053	3,578
Netherlands (5)	01/01/2002	15,863,950	16,192,572	16,258,032	16,305,526	206,619	200,297	194,007	140,527	141,936	136,553
Norway	03/11/2001	4,478,497	4,552,252	4,577,457	4,606,363	59,234	56,458	56,951	44,002	42,478	41,200
Sweden (6)	01/11/1990	8,861,426	8,940,788	8,975,670	9,011,392	90,441	99,157	100,928	93,461	92,961	90,532
Switzerland	05/12/2000	7,164,444	7,313,853	7,364,148	7,415,102	78,458	71,848	73,082	62,528	63,070	60,180
United Kingdom	29/04/2001	58,785,246	59,437,723	59,699,865	59,934,290	679,029	695,549	715,996	608,366	611,184	584,791
England and Wales	29/04/2001	51,508,150	52,121,076	...	...	604,441	621,469	639,721	535,664	538,254	512,541
Northern Ireland	29/04/2001	1,688,807	1,685,760	...	...	21,515	21,648	22,318	14,903	14,462	14,354
Scotland	29/04/2001	5,103,191	5,010,234	...	...	53,076	52,432	54,000	57,799	58,472	56,200
Greece	18/03/2001	10,903,757	11,006,377	11,040,650	11,075,701	103,267	104,420	105,655	105,219	105,529	104,942
Italy	21/10/2001	56,923,524	57,321,070	57,888,245	58,462,375	538,999	539,503	548,244	560,121	586,468	546,658
Portugal	12/03/2001	10,195,014	10,407,465	10,474,685	10,529,255	120,008	112,515	109,298	105,364	108,795	102,010
Spain	01/11/2001	39,960,708	41,663,702	42,345,342	43,038,035	397,632	441,881	454,591	360,391	384,828	370,121
Cyprus	01/10/2001	778,500	802,500	818,200	837,300	9,557	9,077	9,308	6,059	5,836	5,853
Cyprus (government controlled area)	01/01/2005	690,497	715,137	730,367	749,175	8,447	8,088	8,309	5,355	5,200	5,225
Malta	27/11/2005	380,201	386,938	388,867	...	4,255	3,902	3,686	2,957	3,072	2,903



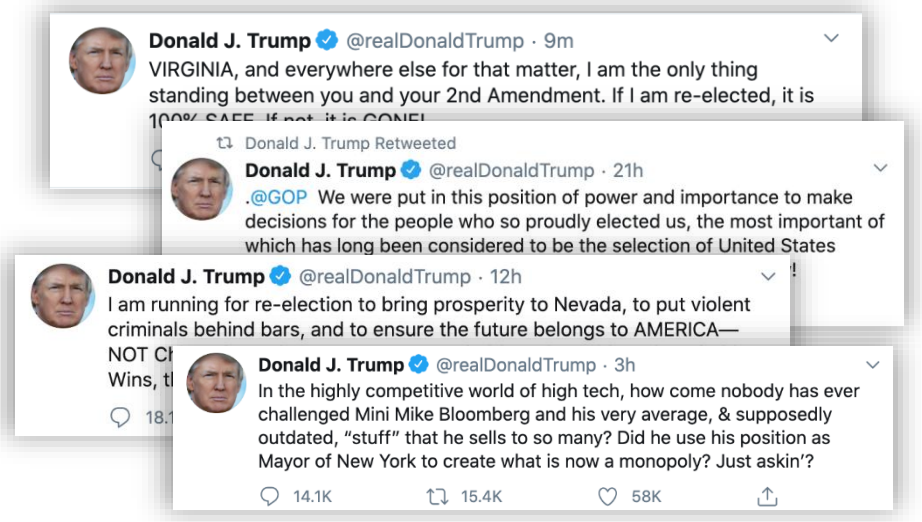
Data science or machine learning?



Data science or machine learning?



Data science or machine learning?

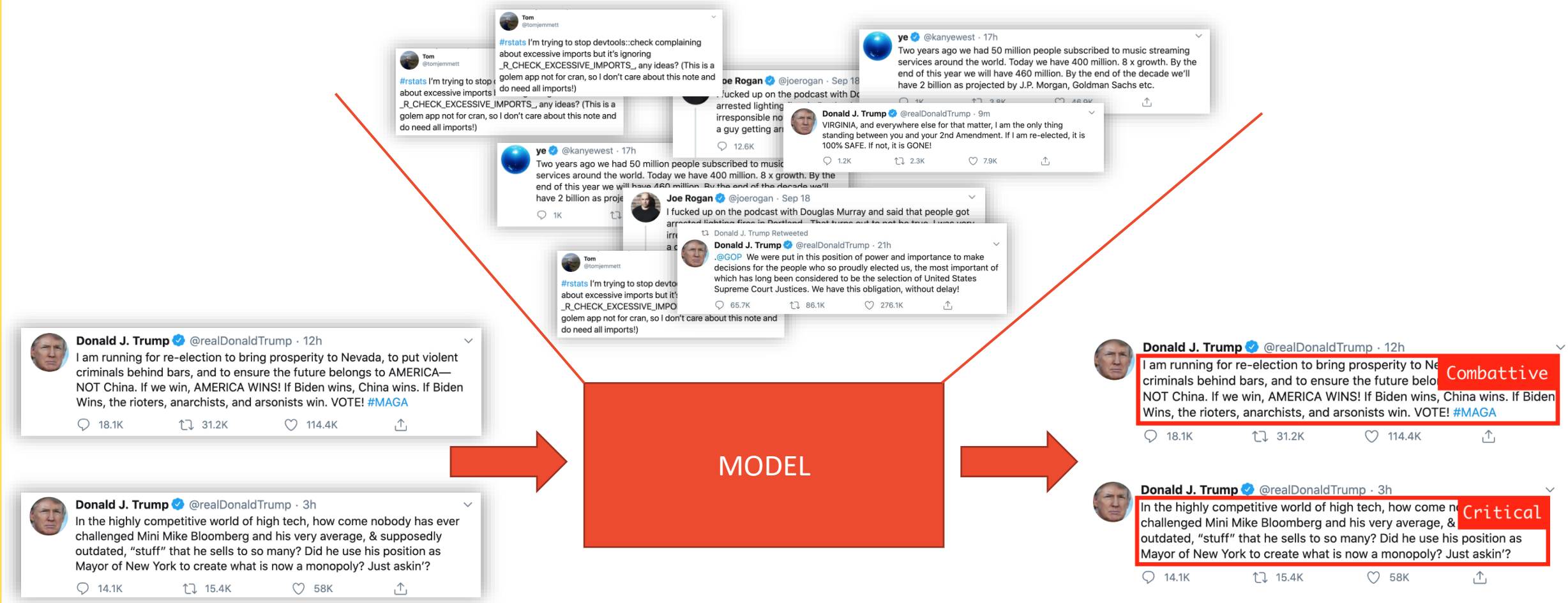


"Trump's tweets in 2020 contain on average 270.4 characters."

"Trump's tweets in 2020 contain on average 2.4 words that are unnecessarily capitalized"

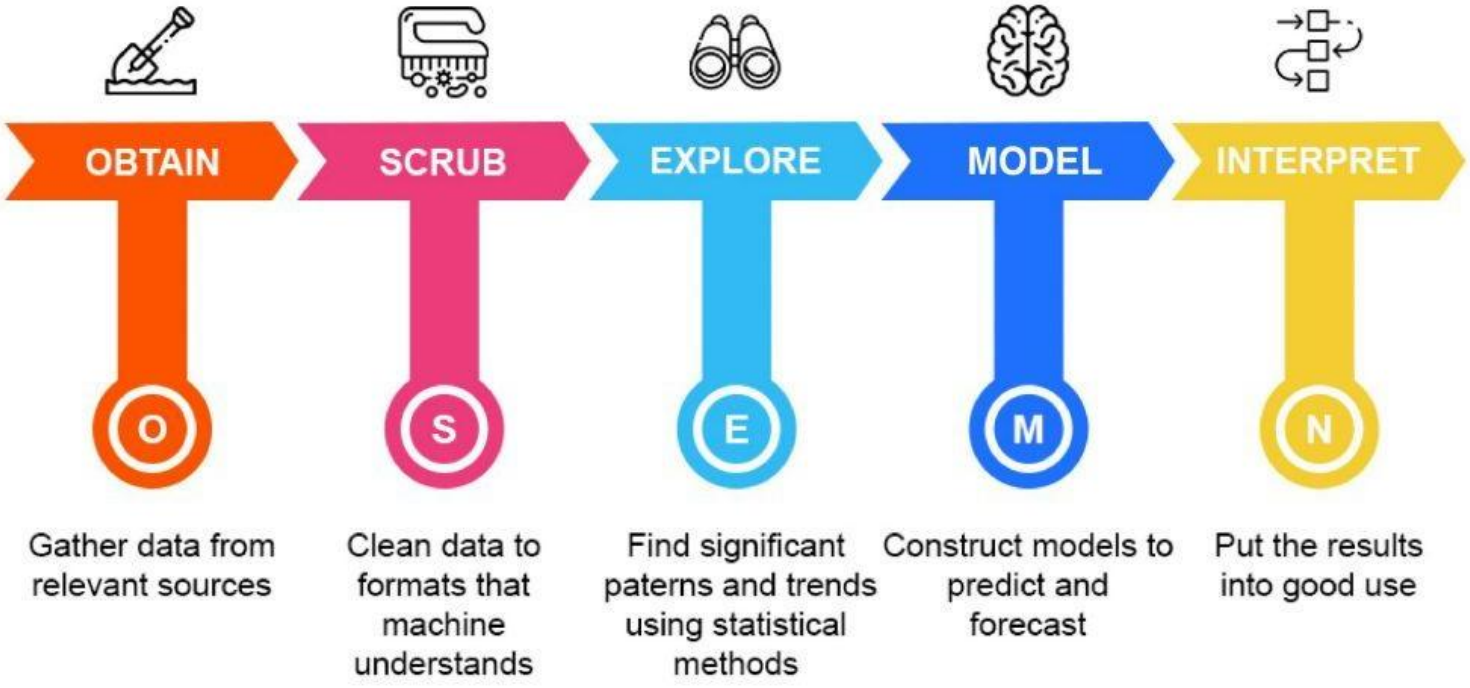
pandas

Data science or machine learning?



The data science process

The data science life cycle involves various roles, tools, and processes, which enables analysts to gain actionable insights. Typically, a data science project undergoes the following stages:



Step 1: Obtain

- Query databases
- Scrape websites
- Read Excel-files
- Call API’s
- Read CSV’s
- …

- Get and combine the data from anywhere you can
- But: only use properly governed data
  - More is not better when more means less reliable

Step 2: Scrub

= Data wrangling

- Are all dates recognized as dates?
- Are all numbers recognized as numbers?
- Is the information of different files correctly merged?
- Does every datapoint translate to a single piece of information?
  - National Insurance Number: date of birth and gender
  - Street name and number: Split? Merge?
- Are there missing values? How will you deal with them?

Step 3: Explore 

= EDA (Explorative data analysis)

- Try to make sense of all available data
- Create some graphs, examine the distribution of all variables
- Are there correlations between the data? Should there be?
  - High blood pressure in relation to height and weight
  - Do not assume causation!
- If needed, go back to “Scrub”
  - Identify and label categorical data
  - Merge or split datapoints
  - Deal with outliers
- Also: talk to the domain experts

Step 4: Model data

- Create models that can explain or predict based upon the data you have imported
  - Classify using logistic regression
  - Predict using linear regression
  - Cluster using k-means or hierarchical clustering

- We don’t cover this step in this course.

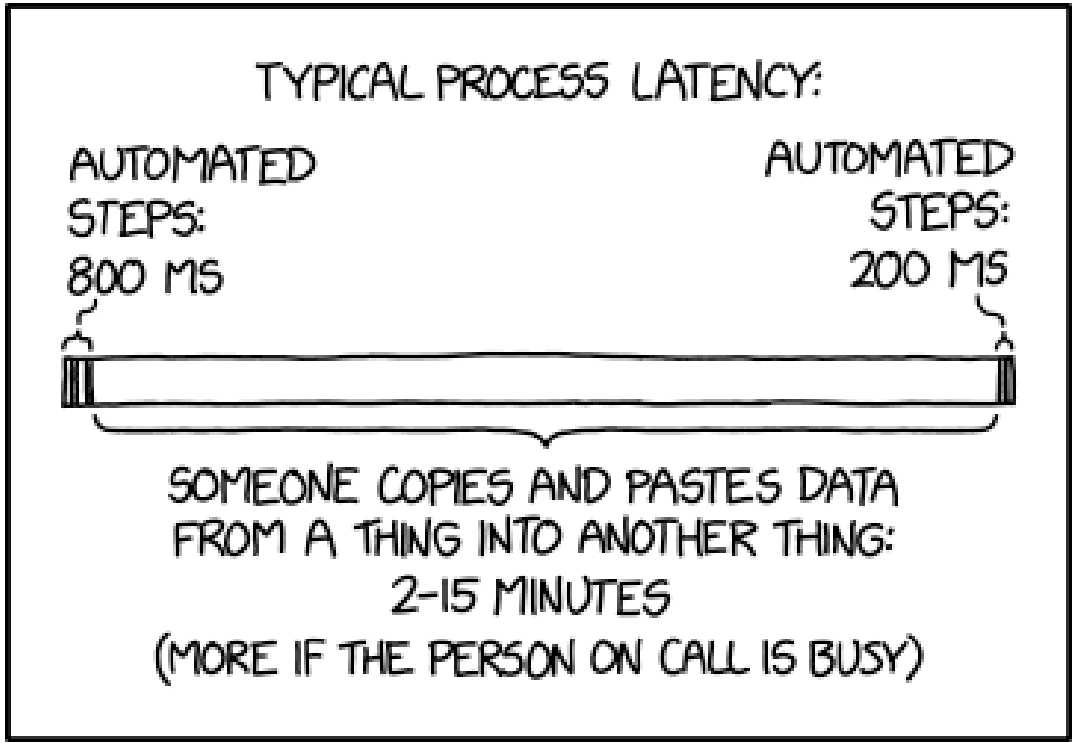
Step 5: Interpret the data

- The ultimate goal of every data science-project
- Test how the models we made generalize
- Explain our findings to a non-domain specialist
- Answer the questions we posed in the beginning
  - Do heavier cars always consume more gas?
  - In which season are there more insect-bites?
- Tell a story using the data you were given.

Summing up

OSEMN	Description	Technologies used
Obtain	Gathering data from databases, the web, open data platforms like Kaggle, …	SQL, Web APIs, **Web Scraping** via Python libraries like Beautiful Soup, …
Scrub	Organising and tidying up the data, removing what is no longer needed, replacing what is missing and standardising the format across all the data collected.	**Regex, Python Pandas library**
Explore	Inspecting and analyzing the data with descriptive statistics (min-max-mean-median-correlations-distributions-…) and basic visualisations.	Python libraries like **NumPy & Pandas** Visualisation libraries like **MatPlotLib** & Seaborn Visualisation tools like Qlik, Tableau, PowerBI
Model	Setting up models to forecast future values or classify and group values.	Python libraries like SciKit
Interpret	Presenting the data to an audience. Delivering the results in such a way business questions are responded, and actionable insights are provided.	Python libraries like Matplotlib & **Seaborn** Javascript libraries like **d3js** Dashboarding tools like Qlik, Tableau, PowerBI

What we want to avoid



Which language?

- R
  - Statistical language, preferred by researchers and mathematicians
  - Originally the first and only language to do data science in
- Python
  - General purpose programming language, preferred by computer scientists
  - Evolved into a perfect alternative to R
- Julia
  - Programming language made for data science
  - Fixes problems that you have with Python when dealing with ginormous amounts of data

Which IDE?

- VSCode and Jupyter notebooks
  - See separate presentation for installation and setup