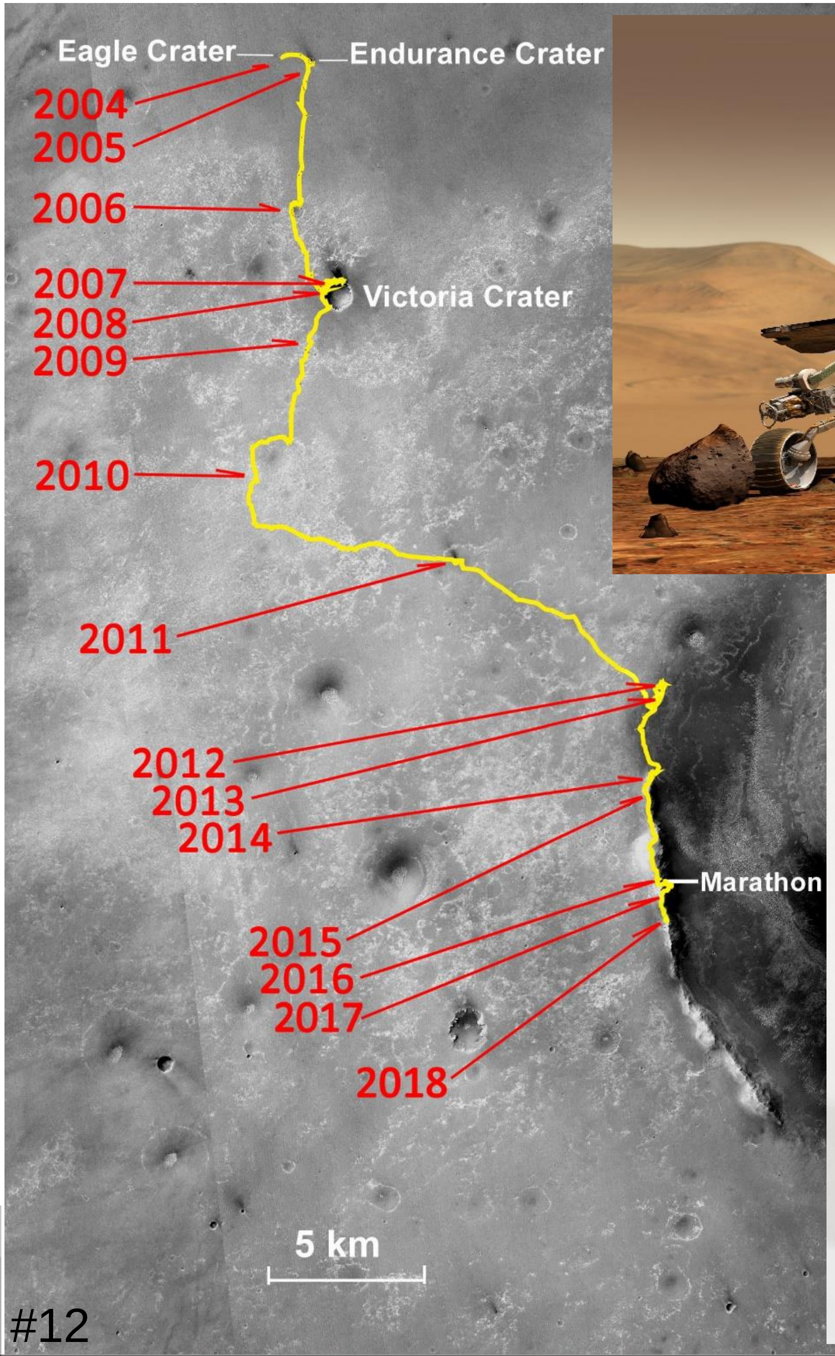
Math Dual e.V.

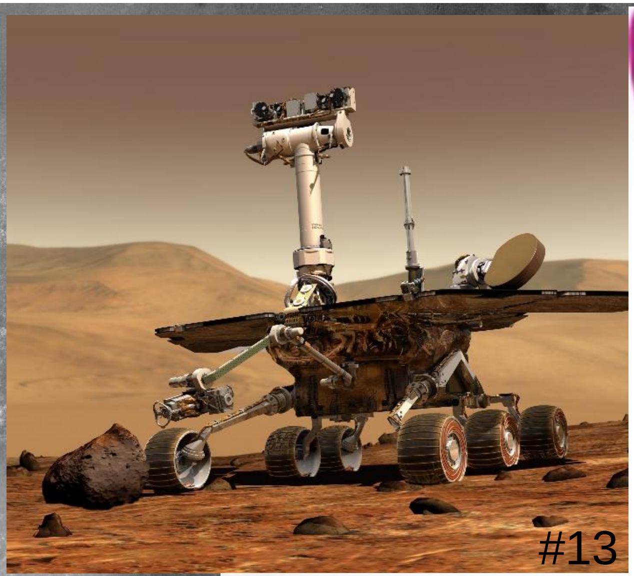
Data Science & Big Data

Björn Tings[1], Karl Kortum[1], James Imber[1], Dmitrii Murashkin[2,1]

[1]*Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)*

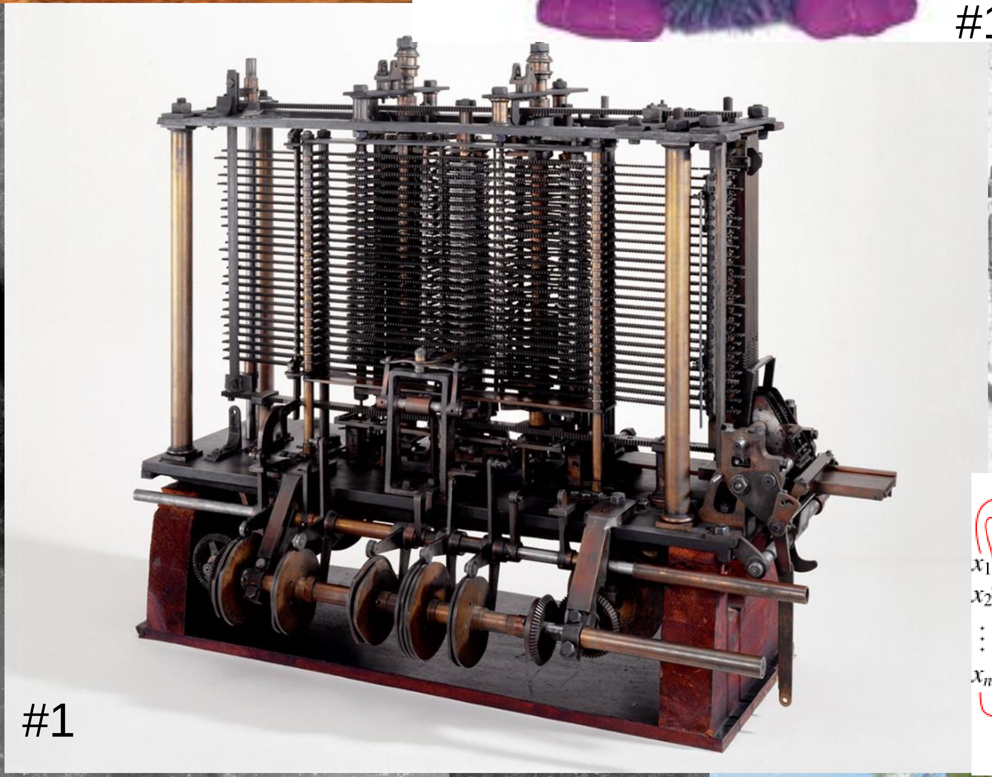
[2]*University of Bremen*

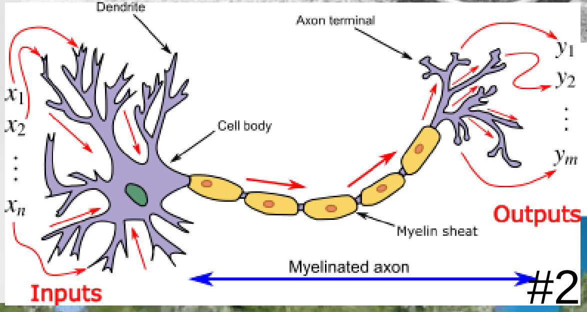

Eagle Crater
Endurance Crater
2004
2005
2006
2007
2008
2009
Victoria Crater
2010
2011
2012
2013
2014
Marathon
2015
2016
2017
2018
5 km
#12


#13

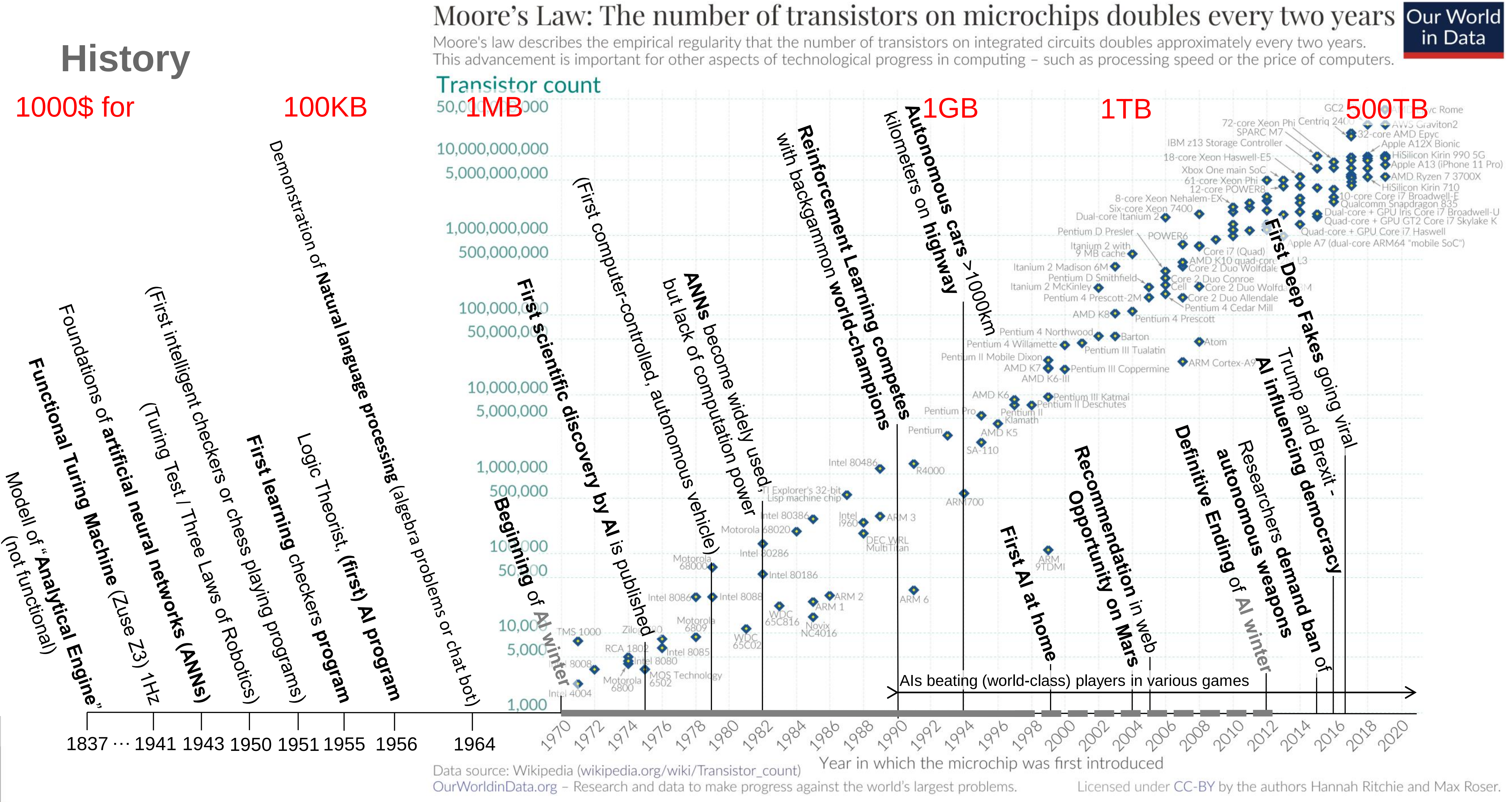

#11


#9

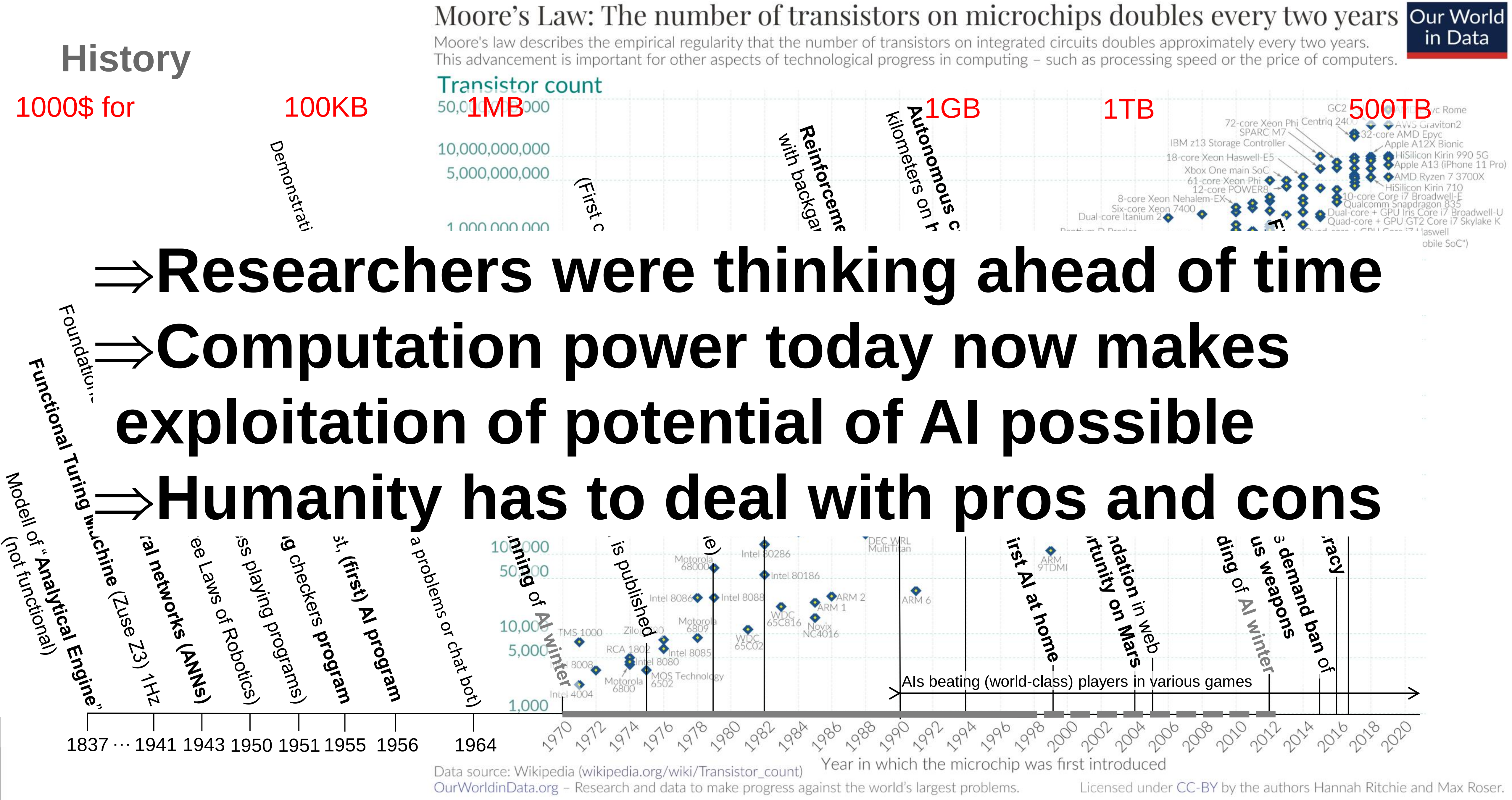

#1


Dendrite
Axon terminal
Cell body
x_1
x_2
x_n
y_1
y_2
y_m
Myelin sheat
Outputs
Myelinated axon
Inputs
#2

History


Moore's Law: The number of transistors on microchips doubles every two years
Our World in Data
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years.
This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.
Transistor count
1000$ for
100KB
1MB
1GB
1TB
500TB
50,000,000,000
10,000,000,000
5,000,000,000
1,000,000,000
500,000,000
100,000,000
50,000,000
10,000,000
5,000,000
1,000,000
500,000
100,000
50,000
10,000
5,000
1,000
Modell of "Analytical Engine" (not functional)
Functional Turing Machine (Zuse Z3) 1Hz
Foundations of artificial neural networks (ANNs)
(Turing Test / Three Laws of Robotics)
(First intelligent checkers or chess playing programs)
First learning checkers program
Logic Theorist, (first) AI program
Demonstration of Natural language processing (algebra problems or chat bot)
Beginning of AI winter
First scientific discovery by AI is published
(First computer-controlled, autonomous vehicle)
ANNs become widely used, but lack of computation power
Reinforcement Learning competes with backgammon world-champions
Autonomous cars >1000km kilometers on highway
First AI at home
Recommendation in web
Opportunity on Mars
Definitive Ending of AI winter
Researchers demand ban of autonomous weapons
AI influencing democracy
Trump and Brexit -
First Deep Fakes going viral
AIs beating (world-class) players in various games
1837 ··· 1941 1943 1950 1951 1955 1956 1964
1970 1972 1974 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016 2018 2020
Year in which the microchip was first introduced
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems.
Licensed under CC-BY by the authors Hannah Ritchie and Max Roser.

History



⇒Researchers were thinking ahead of time

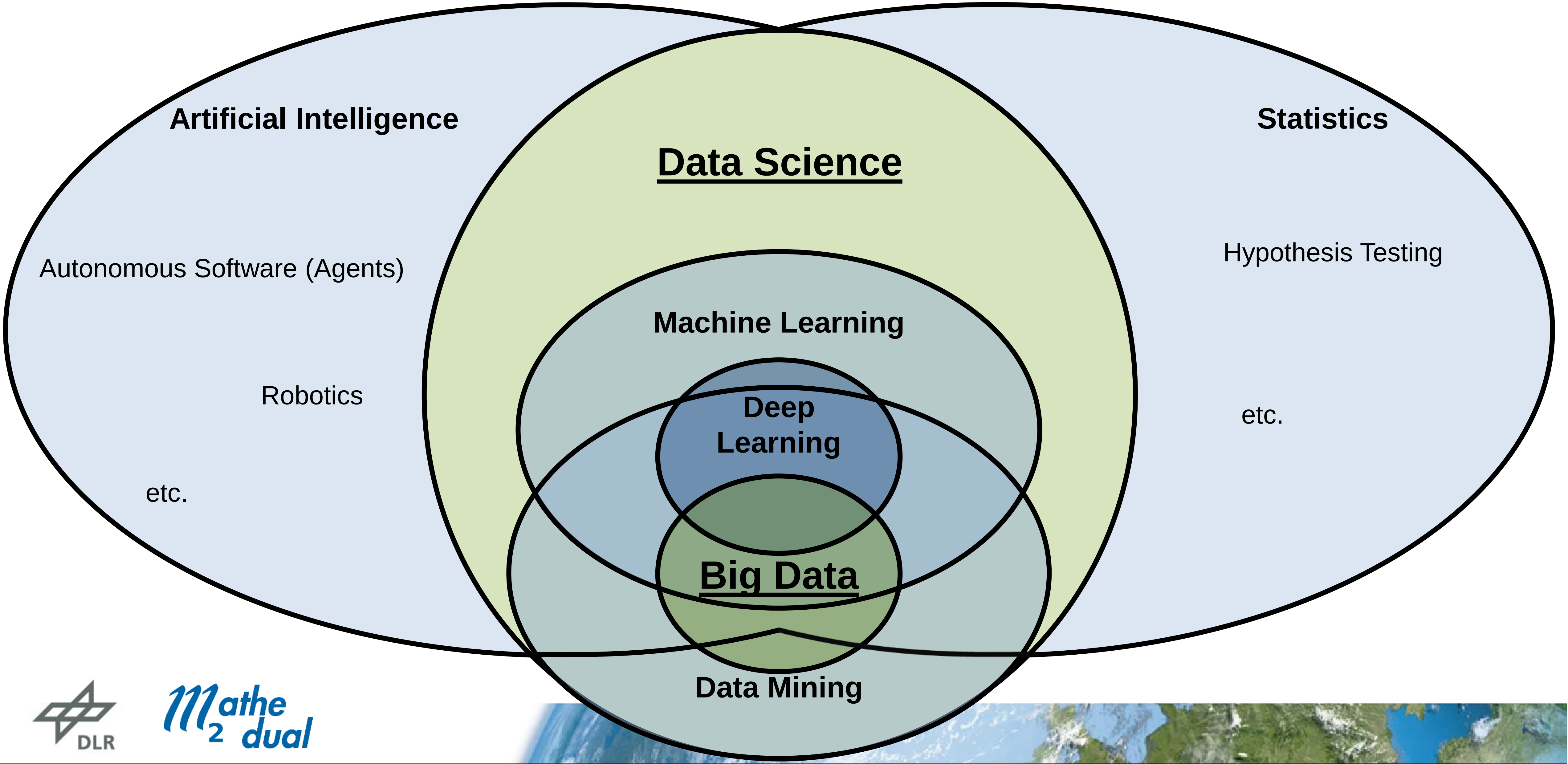
⇒Computation power today now makes exploitation of potential of AI possible

⇒Humanity has to deal with pros and cons

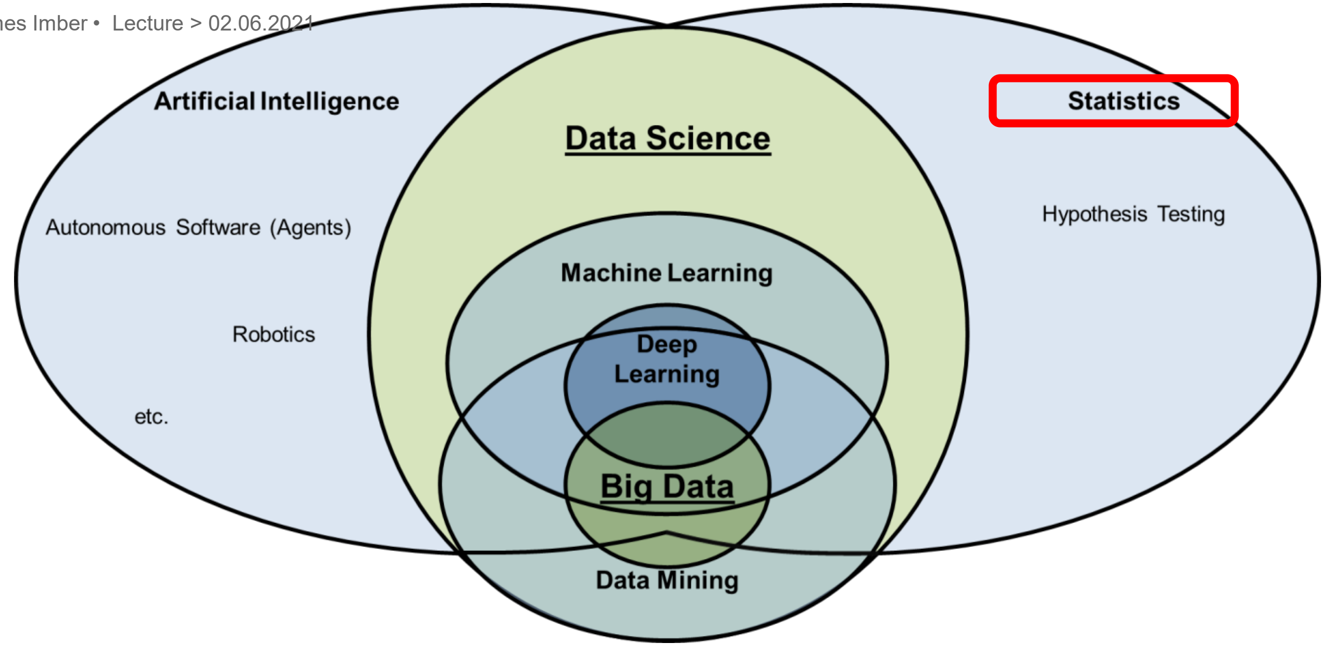
References

1. Science Museum London / Science and Society Picture Library - Babbage's Analytical Engine, 1834-1871. Uploaded by Mrjohncummings, Link: https://de.wikipedia.org/wiki/Analytical_Engine
2. This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. Attribution: Egm4313.s12 at English Wikipedia, Link: https://en.wikipedia.org/wiki/Artificial_neural_network
3. This file is licensed under the Creative Commons Attribution-Share Alike 3.0 Unported license. Author Glosser.ca, Link: https://en.wikipedia.org/wiki/Artificial_neural_network
4. This file is licensed under the Creative Commons Attribution-Share Alike 2.5 Generic license. Publish by= User:Stephane8888 Self made (GIF and problem). Link: https://en.wikipedia.org/wiki/Draughts
5. Tech - A Complete History of Artificial Intelligence by Rebecca Reynoso, https://learn.g2.com/history-of-artificial-intelligence
6. Author: Pete Birkinshaw from Manchester, UK, This file is licensed under the Creative Commons Attribution 2.0 Generic license. https://en.wikipedia.org/wiki/Punched_card
7. CC BY-SA 3.0, Copyright 2003 by Daniel P. B. Smith. Licensed under to the terms of the Wikipedia copyright. https://en.wikipedia.org/wiki/Magnetic_tape
8. CC BY-SA 3.0, Evan-Amos - Own work, https://en.wikipedia.org/wiki/Hard_disk_drive
9. cyberneticzoo.com - a history of cybernetic animals and early robots: A young Hans Moravec with the Stanford Cart c1977. http://cyberneticzoo.com/cyberneticanimals/1960-stanford-cart-american/
10. CC BY-SA 3.0, Ernst D. Dickmanns - Ernst D. Dickmanns, https://en.wikipedia.org/wiki/VaMP
11. Fair use, Source (WP:NFCC#4), https://en.wikipedia.org/wiki/Furby
12. This file is licensed under the Creative Commons Attribution-Share Alike 4.0 International license. Author: James919, https://en.wikipedia.org/wiki/Opportunity_(rover)
13. Source: http://photojournal.jpl.nasa.gov/catalog/PIA04413 (image link), Author: NASA/JPL/Cornell University, Maas Digital LLC
14. Face2Face: Real-time Face Capture and Reenactment of RGB Videos, Justus Thies[1]; Michael Zollhöfer[2]; Marc Stamminger[1]; Christian Theobalt[2]; Matthias Nießner[3]; [1]University of Erlangen-Nuremberg; [2]Max-Planck-Institute for Informatics; [3]Stanford University

Clarification of terms

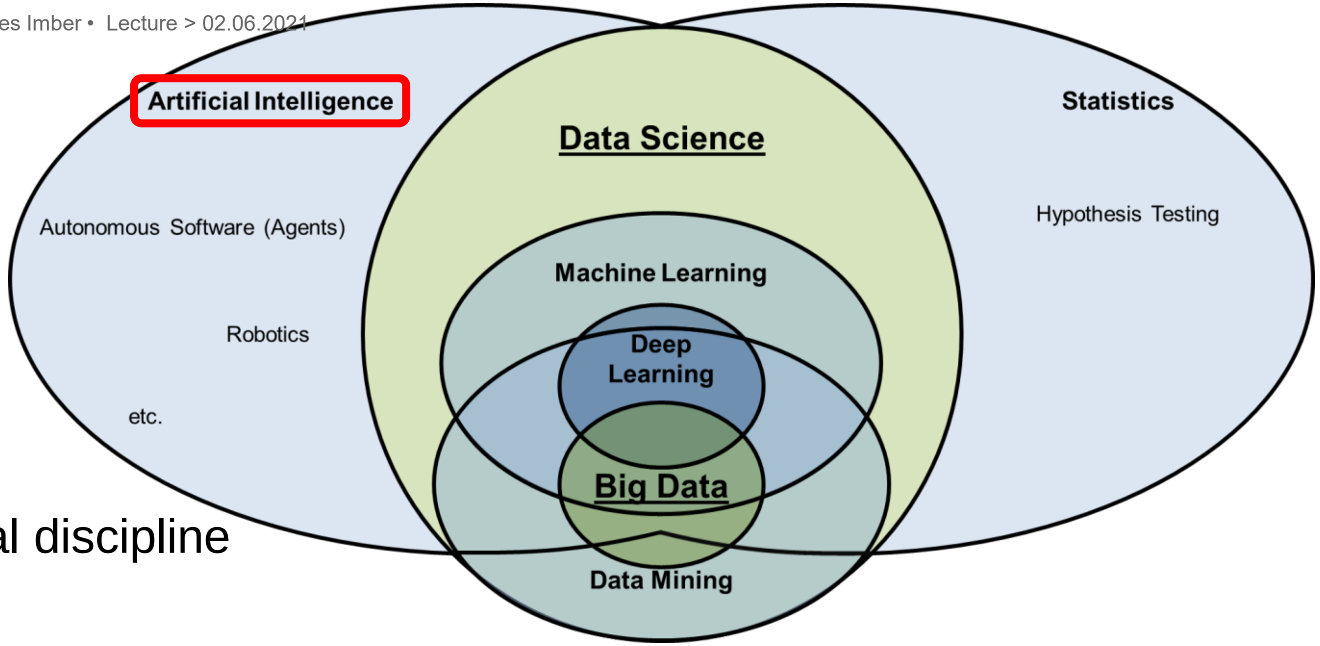


Data Science & Big Data



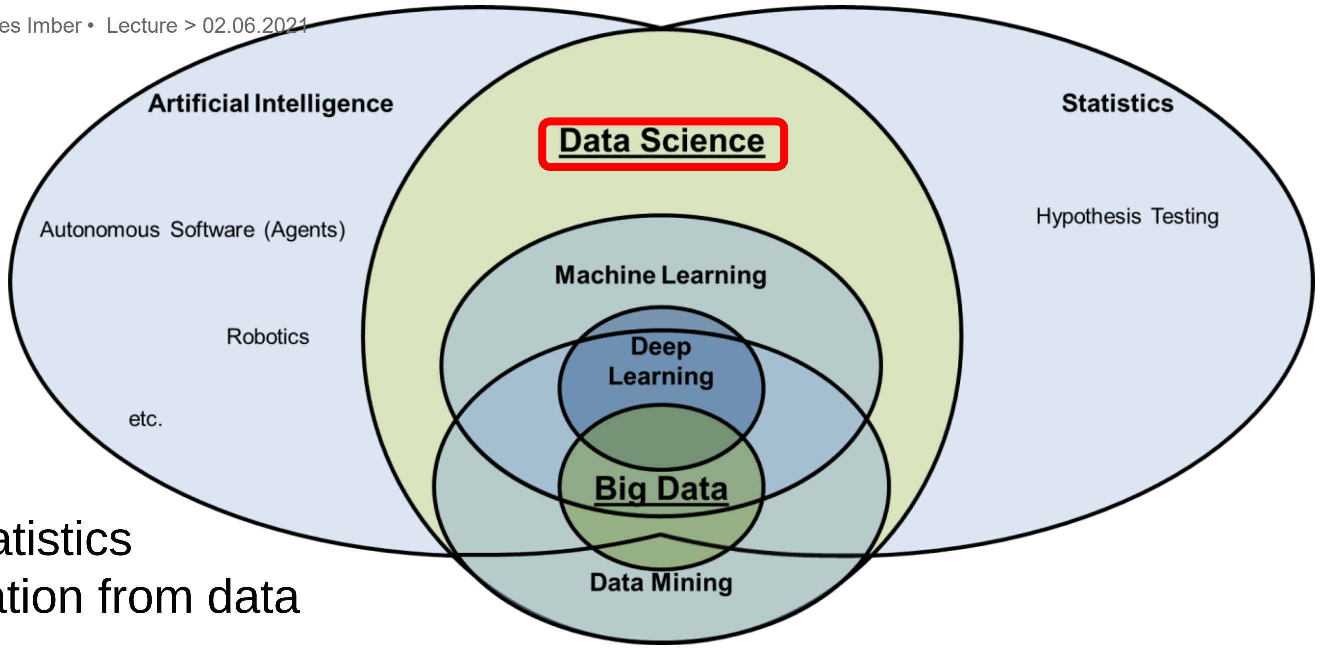
- **Statistics:**
  - Field of mathematics, which deals with the description and understanding of empirical data:
    - Collection, organization, analysis, interpretation, and presentation of data
    - This is also part of data science
  - Pure mathematical statistics without data science:
    - Proving of hypotheses by mathematical methods like analysis or algebra

Data Science & Big Data



- **Artificial Intelligence:**
  - Term was founded in 1955 as a new academical discipline
  - IBM defines AI as follows:
  "Artificial Intelligence enables computers and machines to mimic the perception, learning, problem-solving, and decision-making capabilities of the human mind."
  The level of intelligence does not matter for a machine to be intelligent.
  - **But: AI effect**: After a new capability is achieved by an AI's, it is often not considered as intelligent anymore. The capability becomes self-evident and is not surprising anymore.
  Larry Tesler: "AI is whatever hasn't been done yet."
  - Tasks not achievable by humans are normally part of the subfield of Data Science.

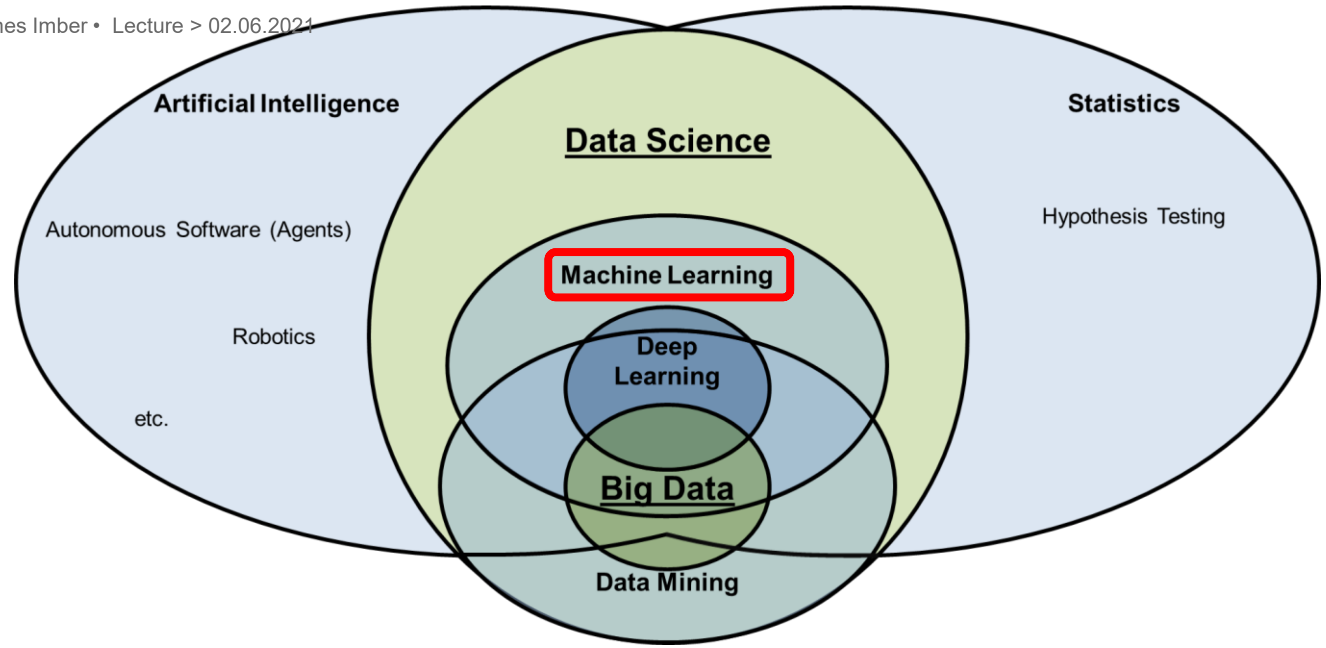
Data Science & Big Data



- **Data Science:**
  - Intersection field of Artificial Intelligence and Statistics
  - Combines methods for the extraction of information from data
    1. Input (Collection & Organization)
       - ➢Includes generation, preparation, storage, pre-processing of data, etc.
    2. Processing (Analysis)
       - ➢Includes modelling, simplification, augmentation of data, etc.
    3. Output (Interpretation & Presentation)
       - ➢Includes displaying of statistical properties, representation of models, charting of results, etc.
  - Additionally includes topics for handling of data, mainly IT-related not covered by AI or Statistics

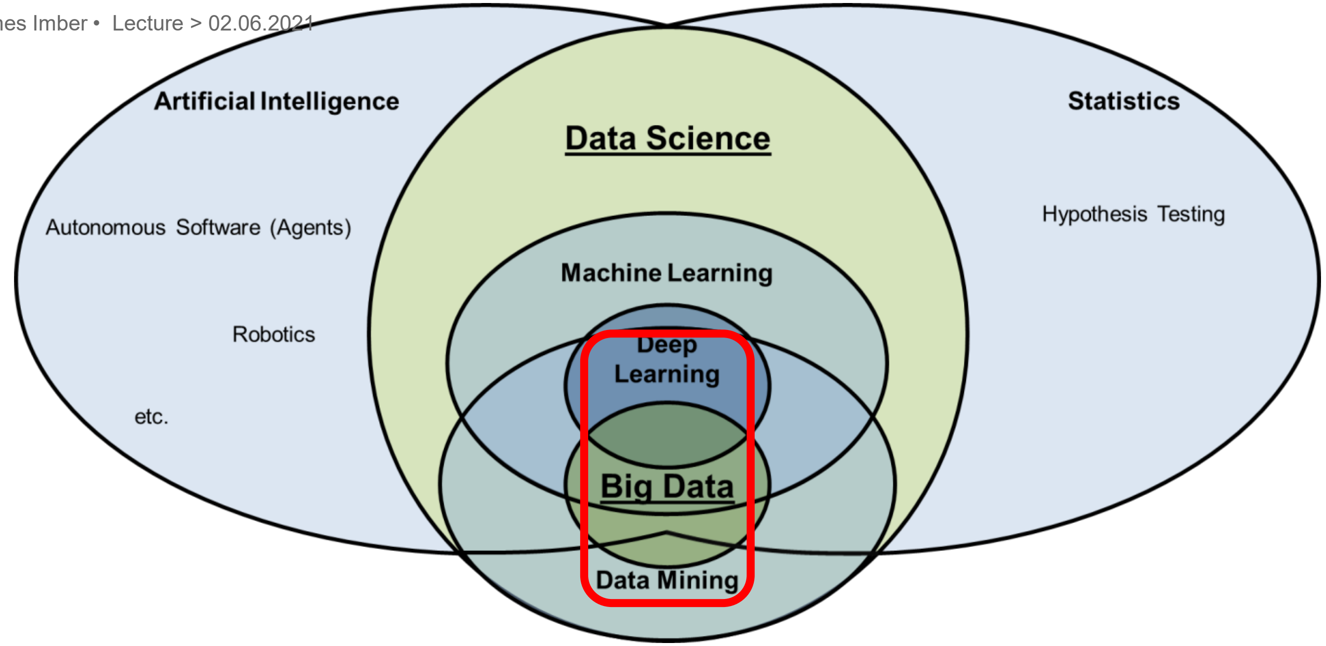
  - Machine Learning, Deep Learning, Data Mining and Big Data are subfields
  - Discrimination of subfields is not distinctively possible, subfields are overlapping

Data Science & Big Data



- **Machine Learning**
  - Complex statistical algorithms, which are learning from so called “training” data

Data Science & Big Data



- **Buzzwords:**
  - **Deep Learning**
    - Subfield of Machine Learning comprising even more complex statistical algorithms
    - Require huge amounts of computation power for learning
  - **Data Mining**
    - ~~Collection/Production of data from different sources~~
    - Analysis of data to extract additional information (input and output is data)
  - **Big Data**
    - Subfield of Data Mining, dealing with massive amounts of data
    - Optimizations for processing and storage of large datasets
    - Consideration of data privacy

Machine Learning

Supervised Learning

Data is labelled (every input has a given desired output).

Example Tasks:
- Classification
- Prediction/Regression

Unsupervised Learning

Data needs no labels.

Example Tasks:
- Dimensionality Reduction
- Clustering
- Association

Reinforcement Learning

Feedback is generated from the task at runtime.

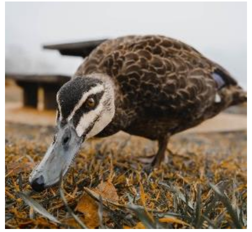
Example Tasks:
- Self-taught AI

Supervised Learning Example: Classification, Llama or Duck

Task: Given an Image of either a Llama or a duck, return the animal's correct species.

Training:

Input:

( ,  ,  ,  , ... ,  , ).

Output:

(0 , 1 , 0 , 0 , ... , 1 , 1),

With 0 = llama, 1 = duck.

Supervised Learning Example: Classification, Facial Recognition

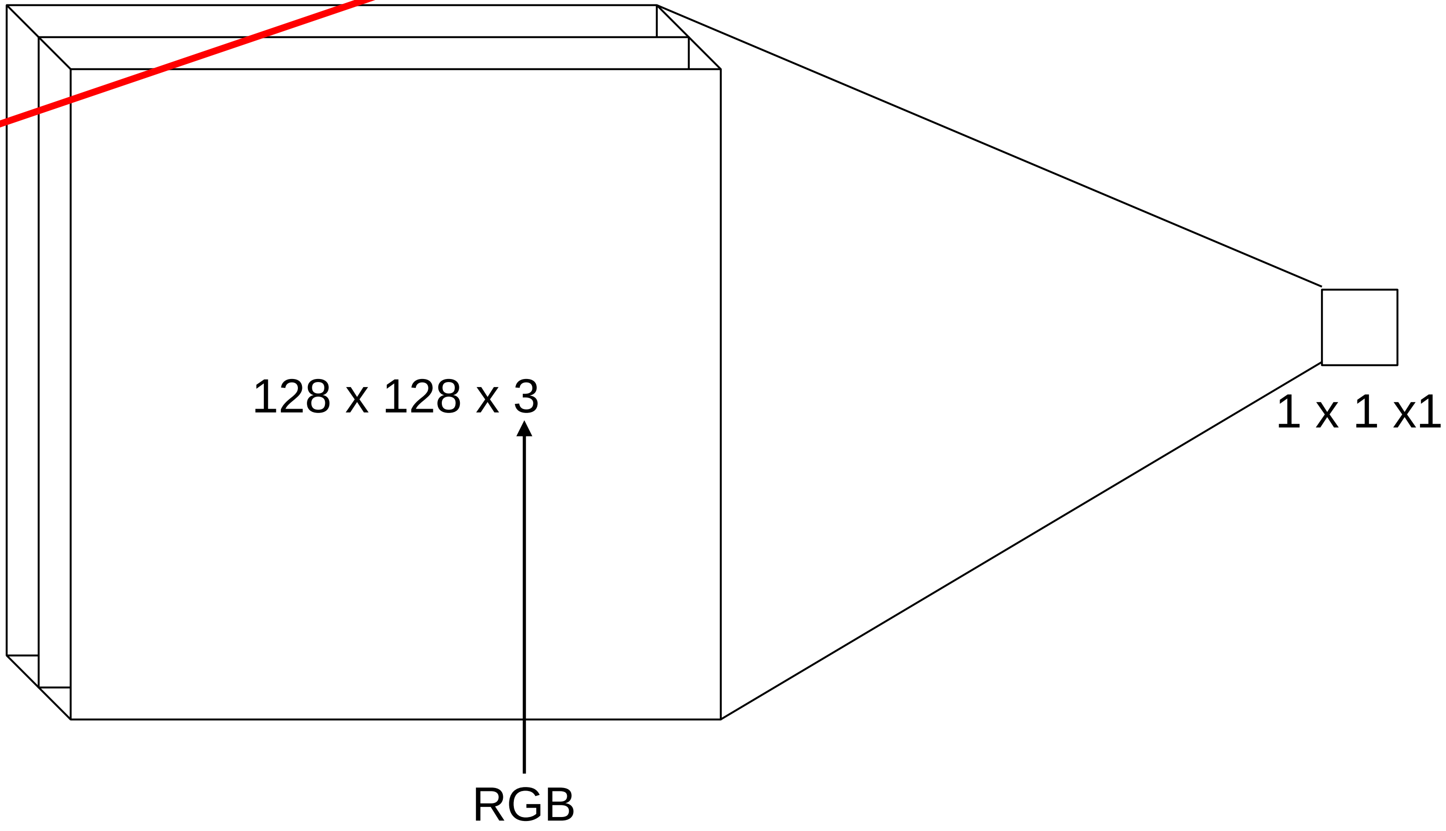
Task: Face ID, e.g. recognise whether or not **Bob or Frank** is in a given image.

Input:




128

128



Output:

1

2

0

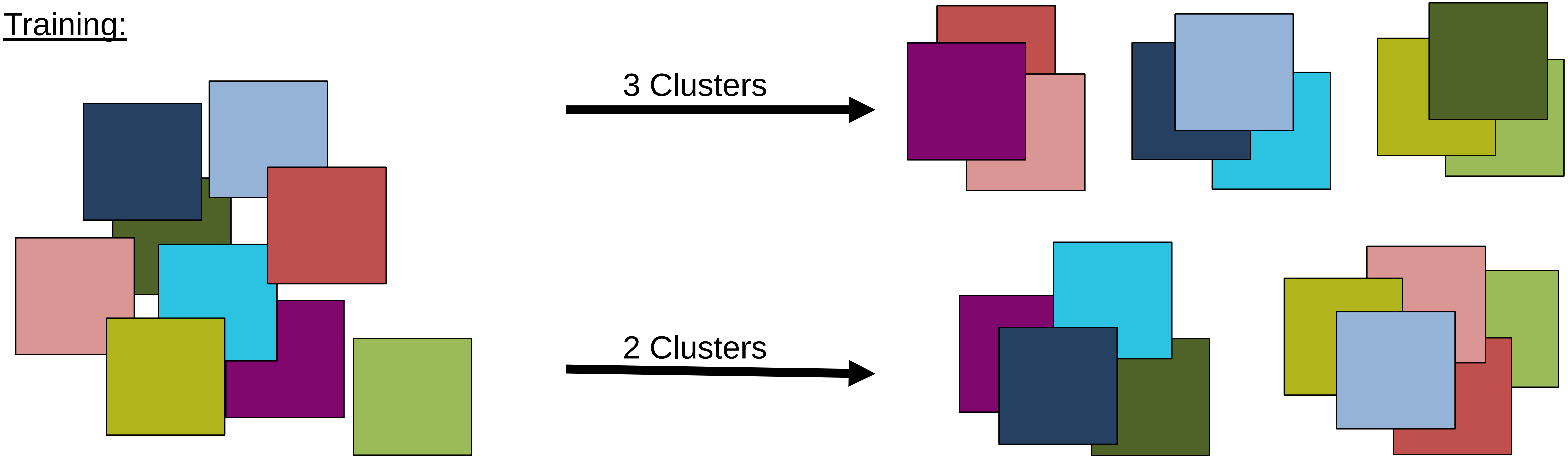
1

with 2 = Frank,
1 = Bob,
0 = not Bob

Unsupervised Learning Example: Clustering

Task: Given a collection of colour swatches, sort them into two/three different groups.

Training:

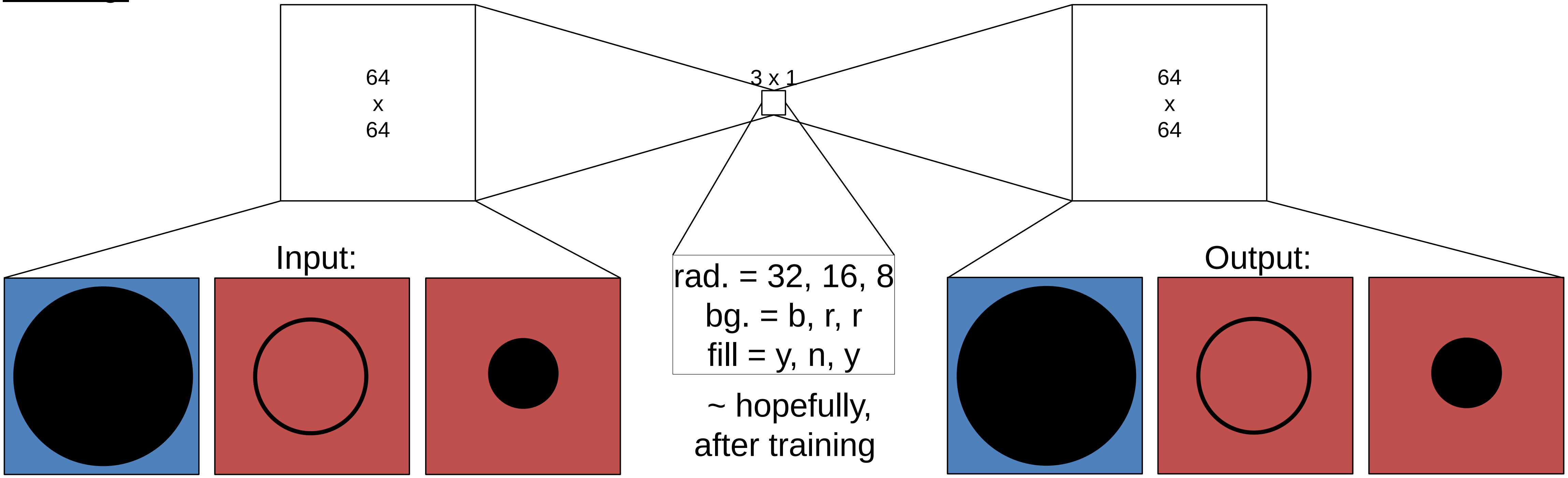


Note: One needs to introduce some metric of similarity. Here for example Euclidian distance between the RGB vectors could be used. The resulting model can associate any new colour to any of the groups immediately.

Unsupervised Learning Example: Dimensionality Reduction

Task: Given an image of a circle find a three-dimensional parametrisation of the Image.

Training:



IMPORTANT: The radii, background colour and fill parameters are NOT given as training data. The method learns this (or an equivalent) parametrisation itself. As input == output, no labels are required.

Reinforcement Learning Example Videos: Cartpole

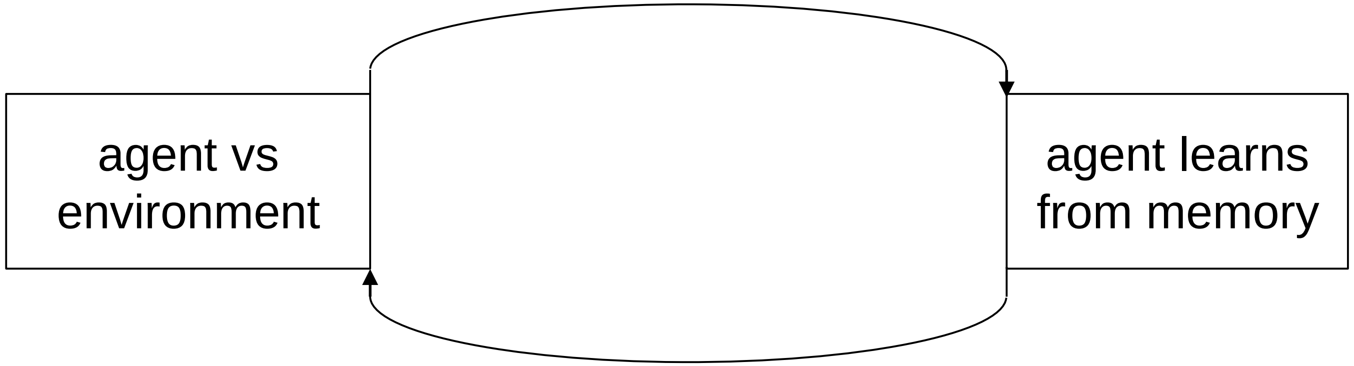
Source: https://www.youtube.com/watch?v=XiigTGKZfks , Author: PilcoLearner

Reinforcement Learning Example: Cartpole

Task: Train an AI cart, that can balance a pole on its end by moving back and forth.

Training: Iterative Process. Agents attempt to balance the stick (select action with best predicted quality), getting rewarded depending on how well they did. Then they learn to predict the quality of an action, in essence trying to predict the action that will give the highest reward. In this case the higher the pole is, the better the reward will be, causing the agent to try and keep the pole up as long as possible.

Agent attempts to balance the pole, selecting actions with best predicted quality. Memorises feedback and states.



Agent has some memory of the feedback given and then learns to predict quality using feedback (Supervised Learning).

IMPORTANT: Feedback/reward is generated ‘automatically’ by the environment, thus no labels are needed.

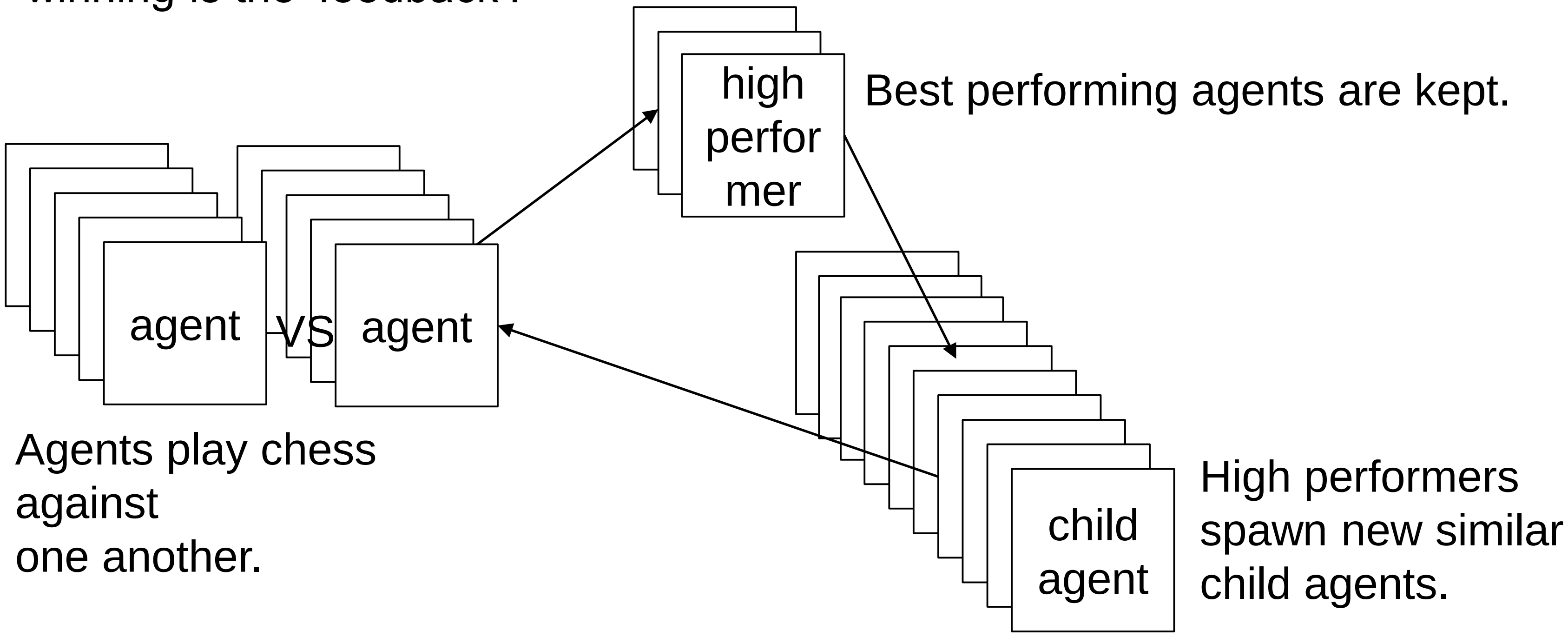
Reinforcement Learning Example Videos: Hide & Seek

Source: https://www.youtube.com/watch?v=kopoLzvh5jY, Author: OpenAi

Reinforcement Learning Example: Chess AI, Genetic Algorithm

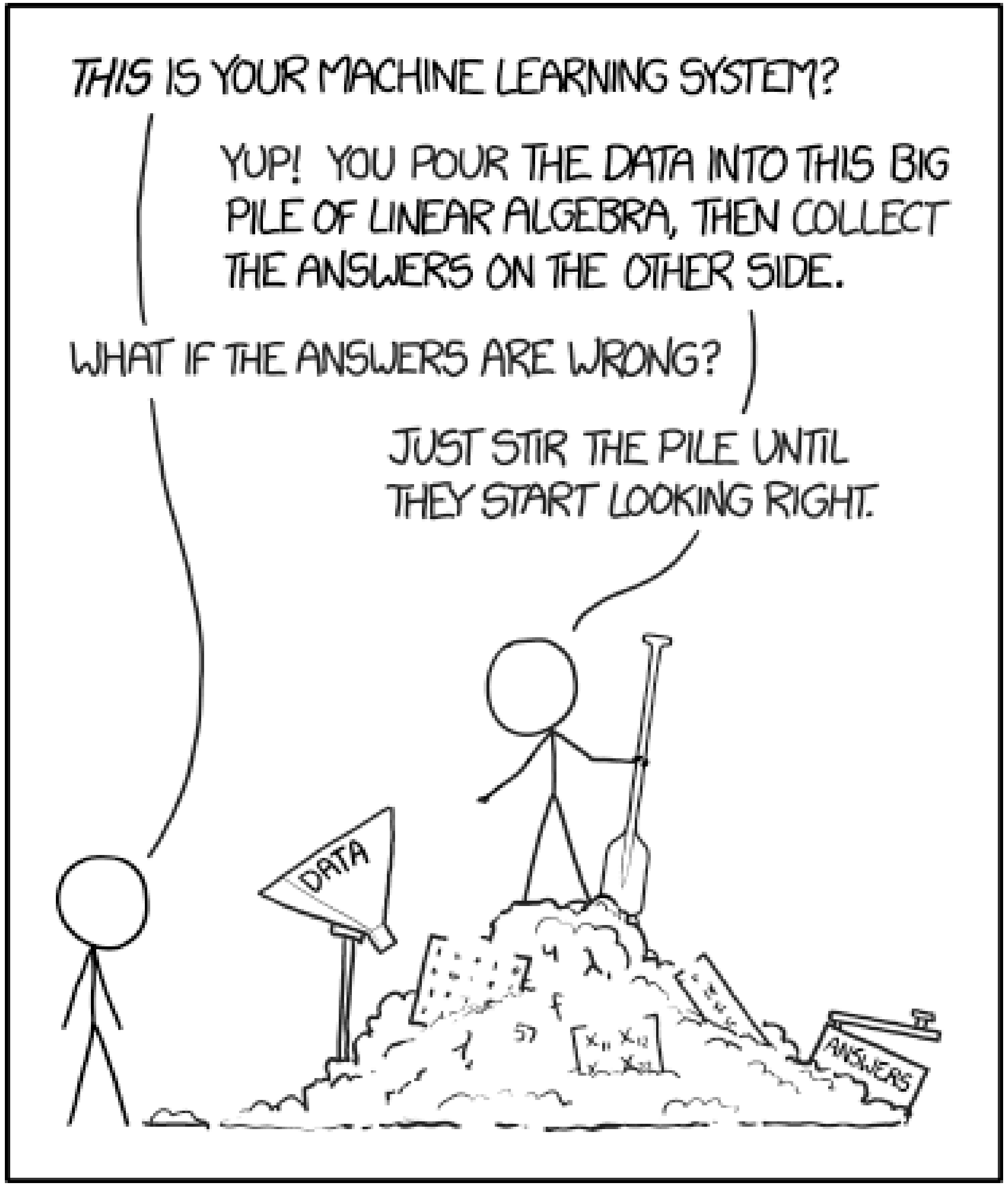
<u>Task:</u> Train a working chess AI, that learns by only playing against itself.

<u>Training:</u> Iterative Process. Agents play vs other agents - this is the ‘environment’. Winning agents move on to the next iteration - winning is the ‘feedback’.



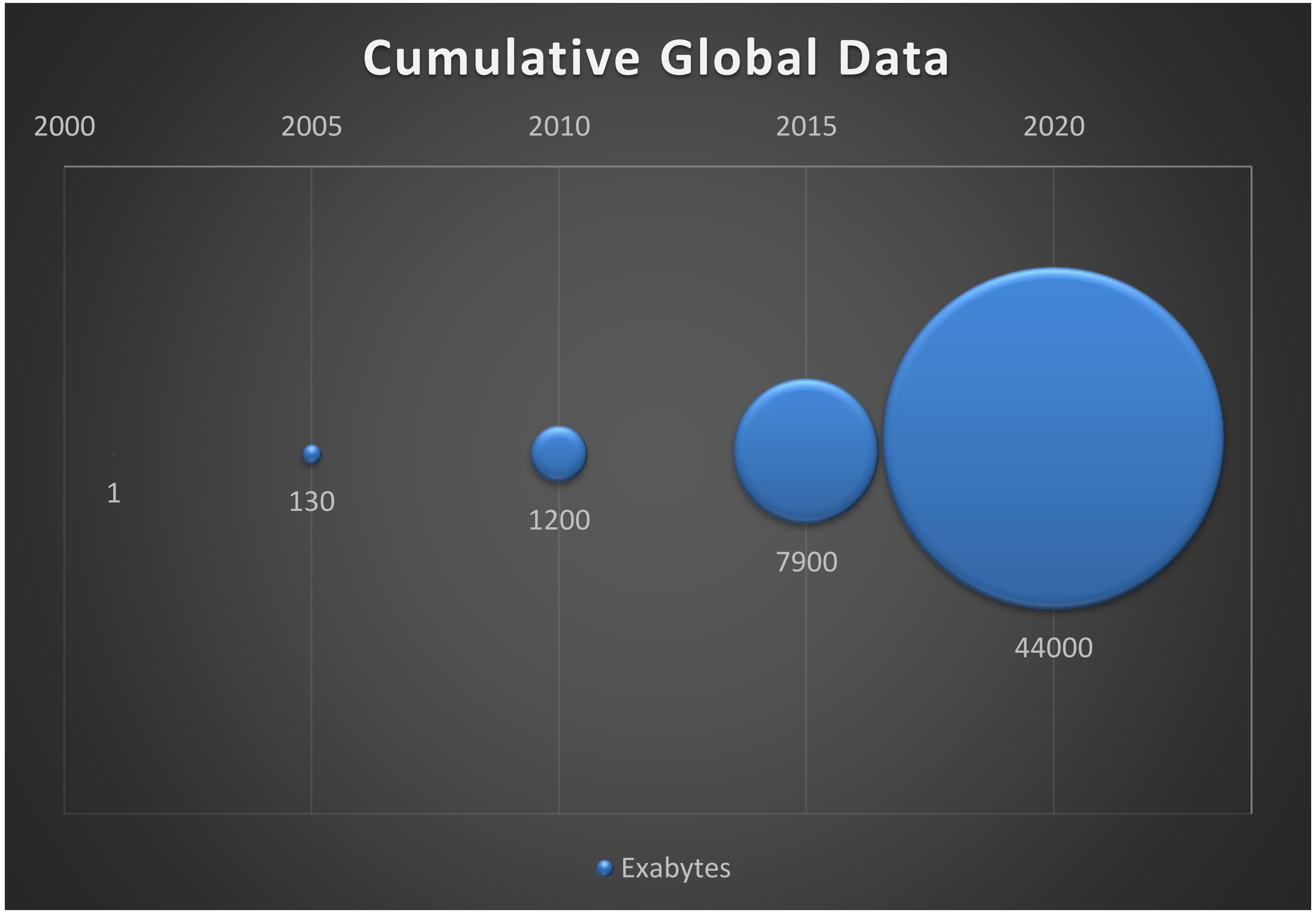
IMPORTANT: Feedback is generated ‘automatically’ by the environment, thus no labels are needed.

Break – Questions?



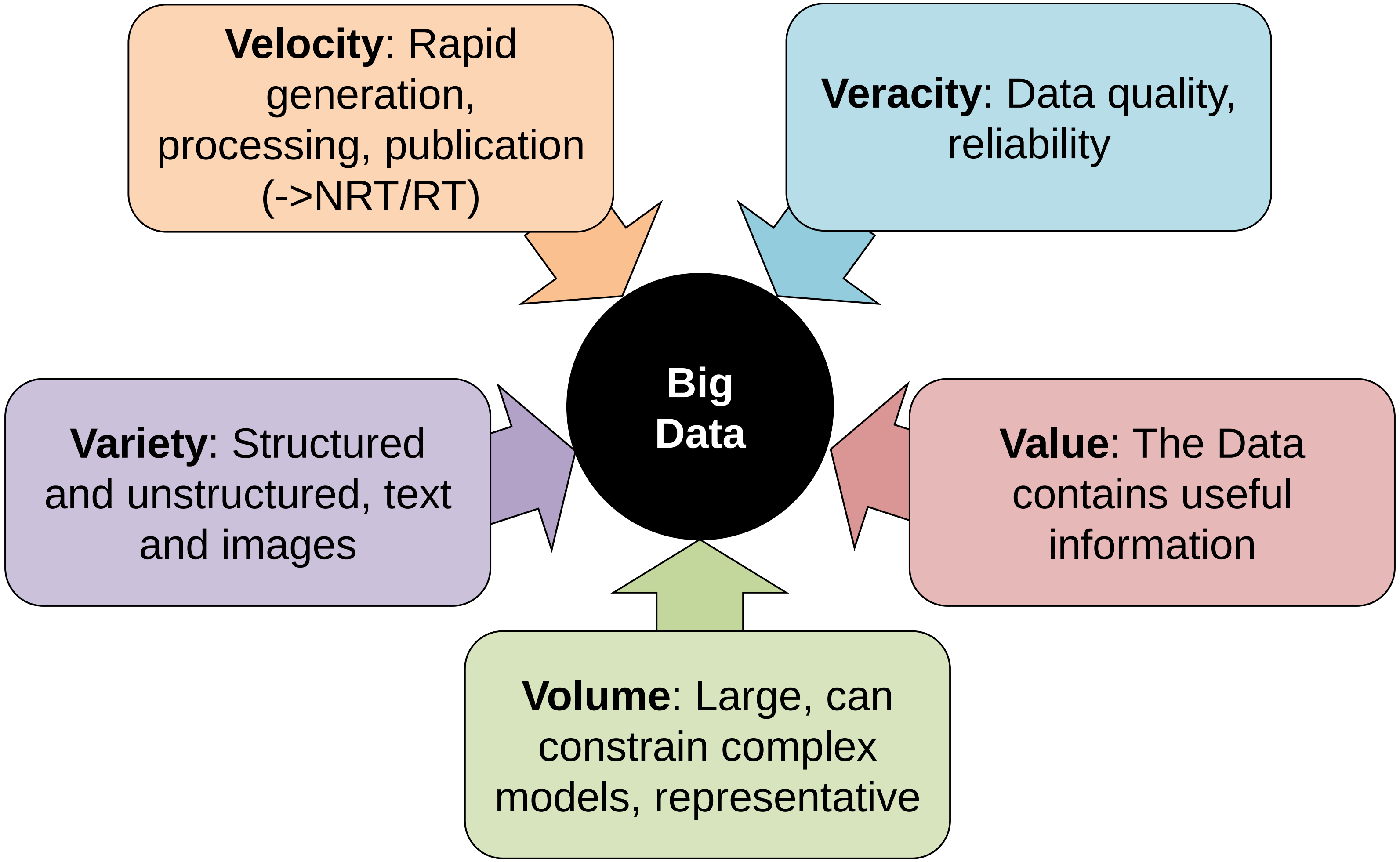
Big Data

- Exponential growth of data – ExaBytes per day! (10^{18}Bytes)
- Very little data will ever be viewed by a human
- Automated processes for data ingestion, reduction, analysis
- Not just large, but complex
- Big data is **not** in databases

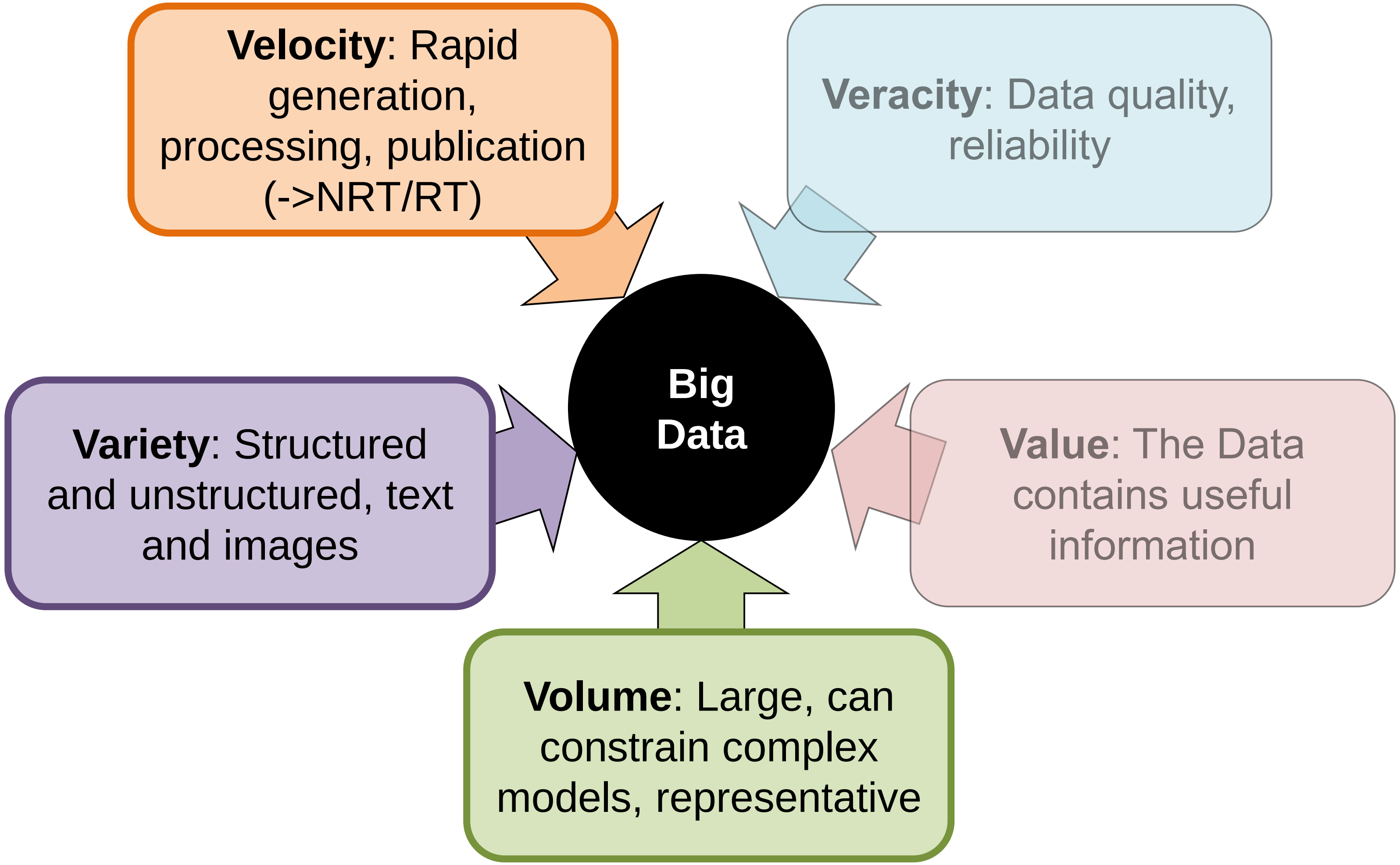


Source: CSC, IDC

Big Data – The Five V`s



Big Data – The Five V`s



Big Data - Volume (Data Storage)

- Data distribution networks → Code distribution networks (containerization)
  - It used to be that data was moved to desired processor running the code
  - Now the code is moved to the data storage

- Data Storage requirements
  - **Scalability** (quickly and flexibly)
  - **Redundancy** (resilient and persistent)
  - **Accessibility** (query time independent of scale and location agnostic)

Global Energy Usage
By Data Centres:

205 TWh in 2018*

Total German Energy Production in 2018:
592.3 TWh**

Power usage effectiveness:

$$PUE = \frac{Total\ Power}{IT\ Equipment\ Power}$$

1.2 ⟷ **3.0**

Carbon usage effectiveness:

$$CUE = PUE \times \frac{CO_2\ emissions(kg)}{Energy(KWh)}$$

Big Data – Velocity

- Data is being produced at unprecedented rates

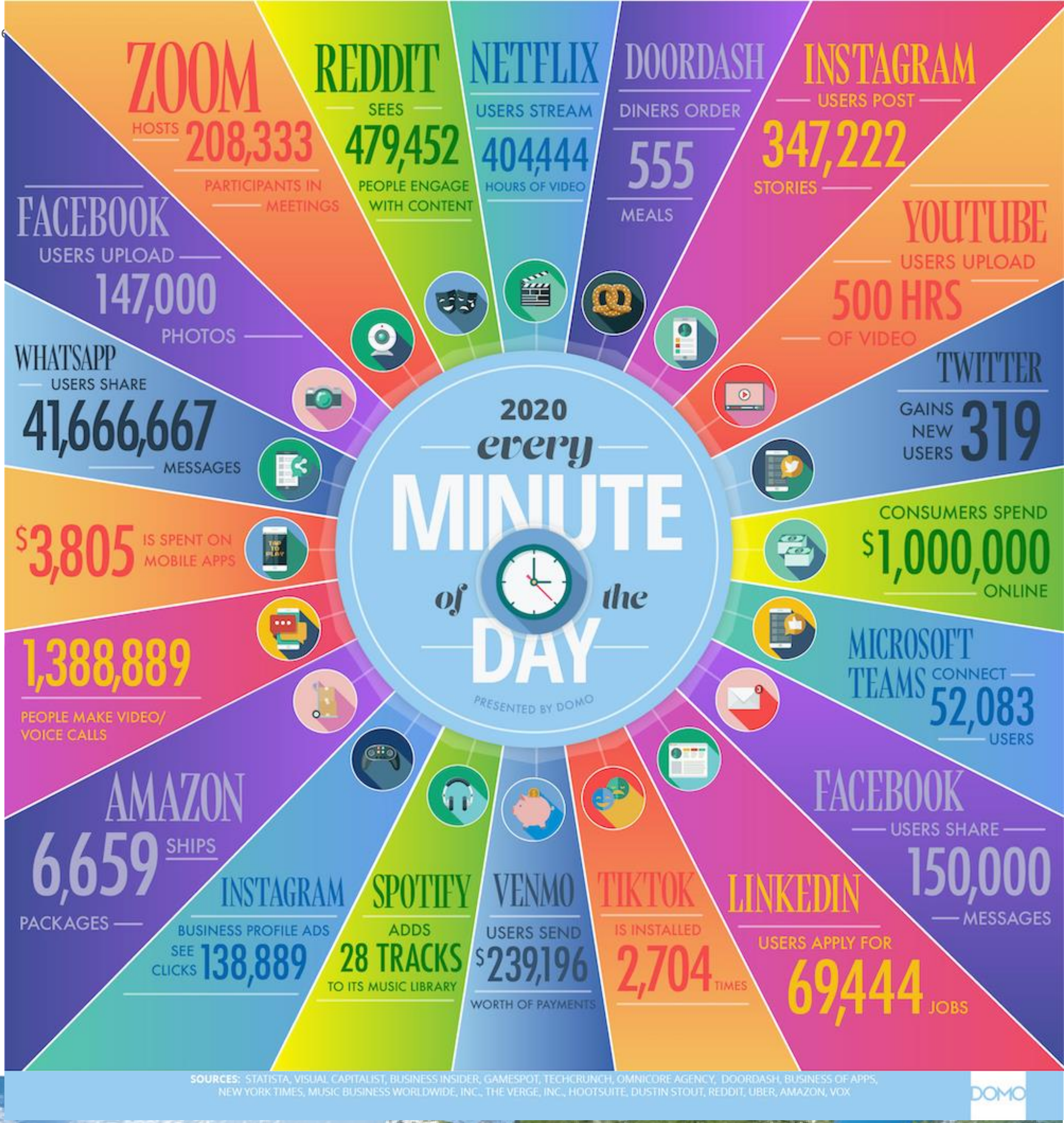
- Data can be **large** and **time sensitive**
  - trending topics from tweets
  - fraud detection in financial transactions
- Data age/relevancy based storage/access
  - Hot, warm, cold

- Uninterrupted data ingestion (Feed the beast!)
- Bound to high speed **response**

  decision making

  action

➢ Data needs to be processed as fast as it is generated



Big Data – Variety

- Text, numbers, images, audio, video, meta-data …
- Structured vs. Semi-structured vs. Unstructured
- Missing entries/sparse data
- How can such diverse datasets be used together effectively? – Data fusion

Bremen

From Wikipedia, the free encyclopedia

Coordinates: 53°5′N 8°48′E

This article is about the German city. For the German state consisting of Bremen and Bremerhaven, see Bremen (state). For other uses, see Bremen (disambiguation).

The **City Municipality of Bremen** (/ˈbreɪmən/, also US: /ˈbrɛmən/;[3][4][5] German: *Stadtgemeinde Bremen*, IPA: [ˈʃtatgəˌmaɪndə ˈbʁeːmən] listen); Low German also: *Breem* or *Bräm*) is the capital of the German state Free Hanseatic City of Bremen (also called just "Bremen" for short), a two-city-state consisting of the cities of Bremen and Bremerhaven. With around 570,000 inhabitants, the Hanseatic city is the 11th largest city of Germany as well as the second largest city in Northern Germany after Hamburg.

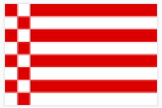
Bremen is the largest city on the River Weser, the longest river flowing entirely in Germany, lying some 60 km (37 mi) upstream from its mouth into the North Sea, and is surrounded by the state of Lower Saxony. A commercial and industrial city, Bremen is, together with Oldenburg and Bremerhaven, part of the Bremen/Oldenburg Metropolitan Region, with 2.5 million people. Bremen is contiguous with the Lower Saxon towns of Delmenhorst, Stuhr, Achim, Weyhe, Schwanewede and Lilienthal. There is an exclave of Bremen in Bremerhaven, the "Citybremian Overseas Port Area Bremerhaven" (*Stadtbremisches Überseehafengebiet Bremerhaven*). Bremen is the fourth largest city in the Low German dialect area after Hamburg, Dortmund and Essen.

Bremen's port, together with the port of Bremerhaven at the mouth of the Weser, is the second largest port in Germany after the Port of Hamburg. The airport of Bremen (*Flughafen Bremen "Hans Koschnick"*) lies in the southern borough of Neustadt-Neuenland and is Germany's 12th busiest airport.

Bremen is a major cultural and economic hub of Northern Germany. The city is home to dozens of historical galleries and museums, ranging from historical sculptures to major art museums, such as the Bremen Overseas Museum (*Übersee-Museum Bremen*).[6] The Bremen City Hall and the Bremen Roland are UNESCO World Heritage Sites. Bremen is well known through the Brothers Grimm's fairy tale "Town Musicians of Bremen" (*Die Bremer Stadtmusikanten*), and there is a statue dedicated to it in front of the city hall.

Bremen

Clockwise from top: Bremer Marktplatz, Bremen Hauptbahnhof, the Werdersee and the Town Musicians statue



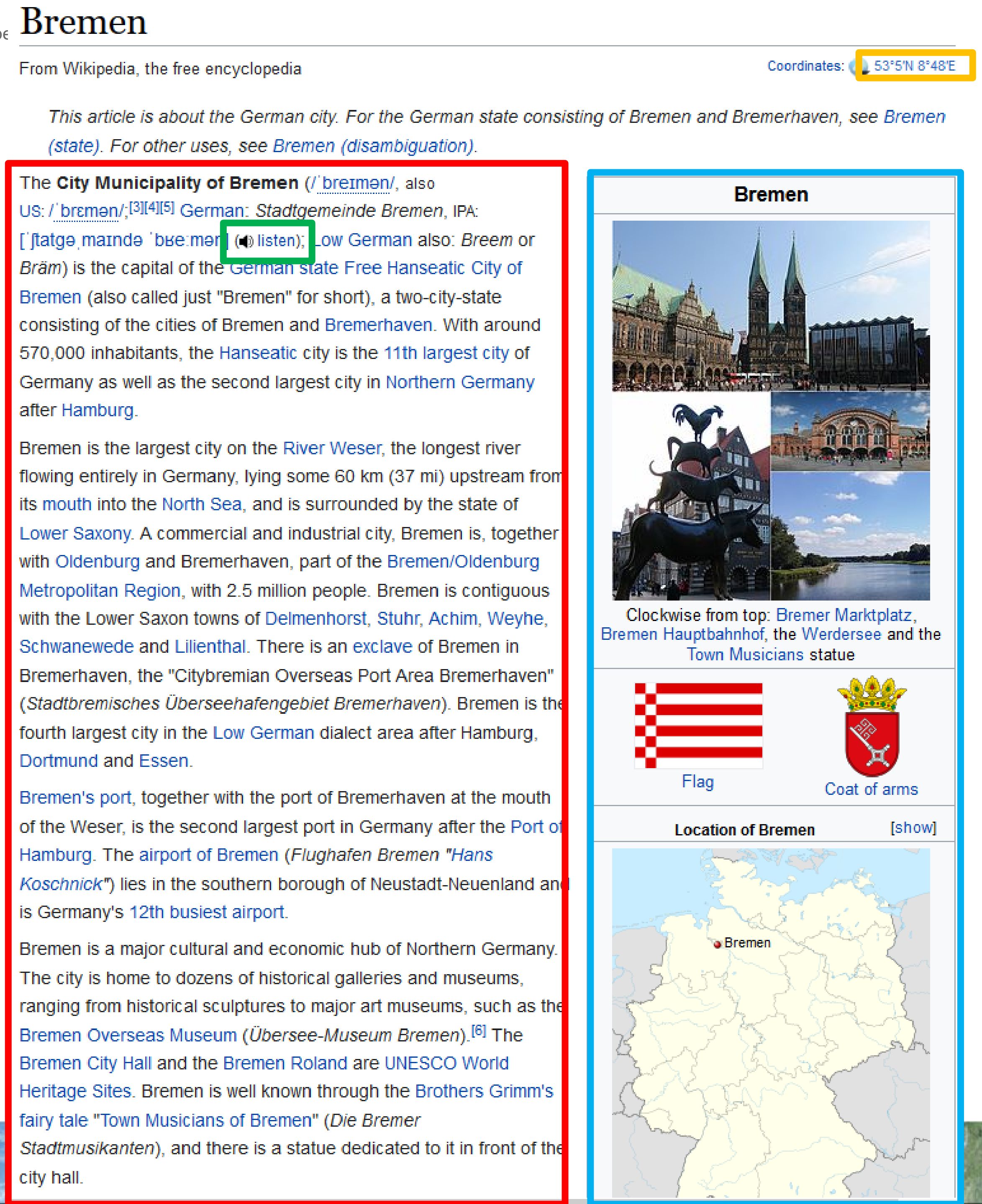
Flag

Coat of arms

Location of Bremen [show]

Big Data – Variety

- **Text**, numbers, **images**, **audio**, video, **meta-data** …
- Structured vs. Semi-structured vs. Unstructured
- Missing entries/sparse data
- How can such diverse datasets be used together effectively? – Data fusion



Bremen

From Wikipedia, the free encyclopedia

Coordinates: 53°5′N 8°48′E

This article is about the German city. For the German state consisting of Bremen and Bremerhaven, see Bremen (state). For other uses, see Bremen (disambiguation).

The **City Municipality of Bremen** (/ˈbreɪmən/, also US: /ˈbrɛmən/;[3][4][5] German: *Stadtgemeinde Bremen*, IPA: [ˈʃtatgəˌmaɪndə ˈbʁeːmən] (listen); Low German also: *Breem* or *Bräm*) is the capital of the German state Free Hanseatic City of Bremen (also called just "Bremen" for short), a two-city-state consisting of the cities of Bremen and Bremerhaven. With around 570,000 inhabitants, the Hanseatic city is the 11th largest city of Germany as well as the second largest city in Northern Germany after Hamburg.

Bremen is the largest city on the River Weser, the longest river flowing entirely in Germany, lying some 60 km (37 mi) upstream from its mouth into the North Sea, and is surrounded by the state of Lower Saxony. A commercial and industrial city, Bremen is, together with Oldenburg and Bremerhaven, part of the Bremen/Oldenburg Metropolitan Region, with 2.5 million people. Bremen is contiguous with the Lower Saxon towns of Delmenhorst, Stuhr, Achim, Weyhe, Schwanewede and Lilienthal. There is an exclave of Bremen in Bremerhaven, the "Citybremian Overseas Port Area Bremerhaven" (*Stadtbremisches Überseehafengebiet Bremerhaven*). Bremen is the fourth largest city in the Low German dialect area after Hamburg, Dortmund and Essen.

Bremen's port, together with the port of Bremerhaven at the mouth of the Weser, is the second largest port in Germany after the Port of Hamburg. The airport of Bremen (*Flughafen Bremen "Hans Koschnick"*) lies in the southern borough of Neustadt-Neuenland and is Germany's 12th busiest airport.

Bremen is a major cultural and economic hub of Northern Germany. The city is home to dozens of historical galleries and museums, ranging from historical sculptures to major art museums, such as the Bremen Overseas Museum (*Übersee-Museum Bremen*).[6] The Bremen City Hall and the Bremen Roland are UNESCO World Heritage Sites. Bremen is well known through the Brothers Grimm's fairy tale "Town Musicians of Bremen" (*Die Bremer Stadtmusikanten*), and there is a statue dedicated to it in front of the city hall.

Bremen

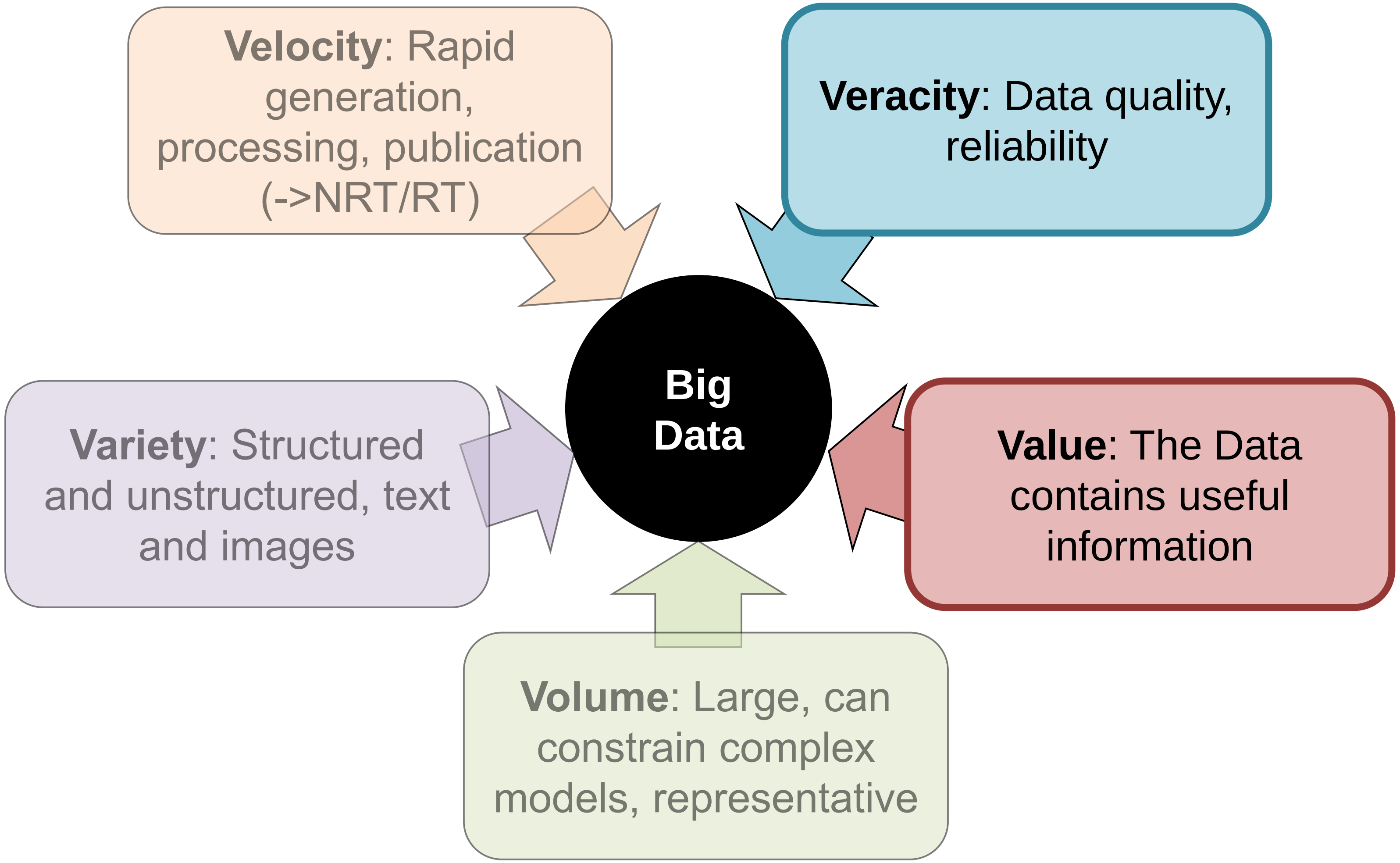
Clockwise from top: Bremer Marktplatz, Bremen Hauptbahnhof, the Werdersee and the Town Musicians statue

Flag

Coat of arms

Location of Bremen [show]

Big Data – The Five V`s

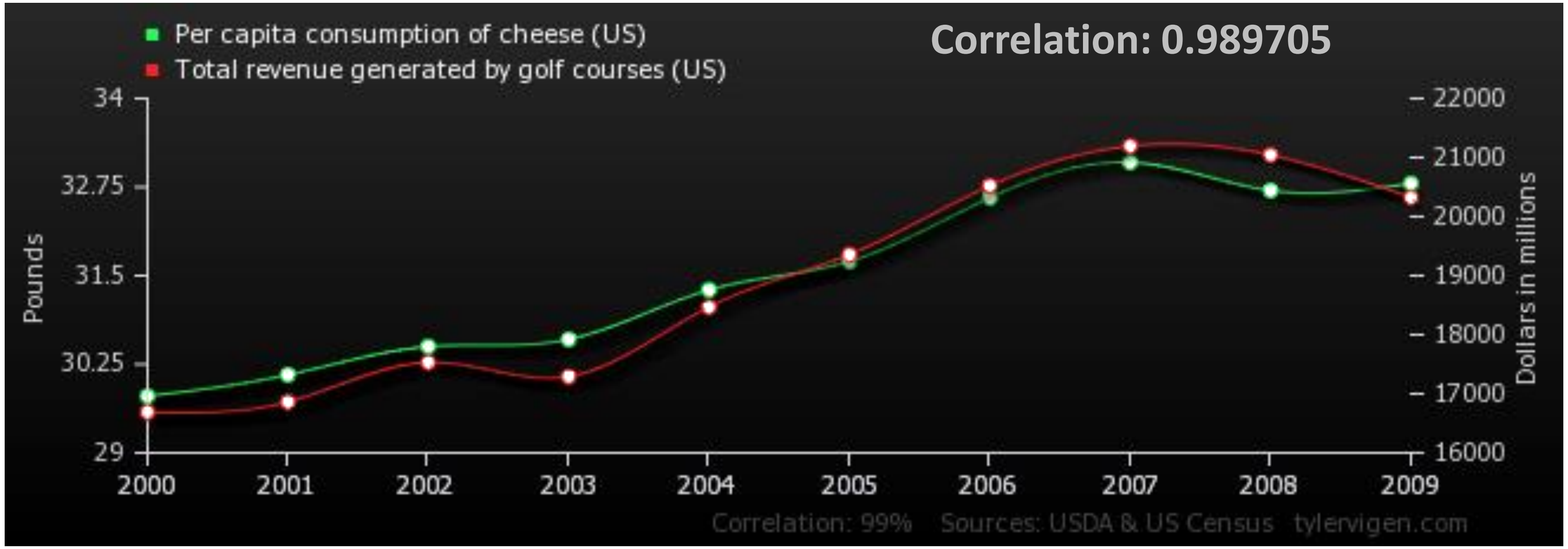


Big Data – Veracity (Data Quality)

- Always important, but critical for automated, empirical analysis (without theoretical underpinning)

- Completeness/sampling

- Bias
  - Confirmation bias
  - Selection bias
  - Outliers/Anomalies
  - "Look elsewhere"
  - …



The larger the dataset, the more likely for correlations to appear!

- Often easier to correct bias during the data taking than account for it in the analysis

Big Data – Value

- Is the data actually useful?
- What kind of analyses can be performed?
- Is there a market for the results?
- How can the value of data be quantified?

It may be that these questions are only properly answered after the fact

- Valuable features:
  - High statistics, Completeness, Real-time processing
- Large amounts of data are being wasted
  - **60%** of companies reported more than half of data unused

European Commission forecasts value of the data economy in EU27 to be **€829 billion** by 2025

Big Data – Privacy

- All records are now digital and data is a commodity
  - Medical data
  - Financial data
  - etc.
- New risks and vulnerabilities emerge regularly

In a 2020 survey: 49% of businesses had experienced a data breach, 26% had been breached in the past year.

- How can privacy be maintained?
  - Data level
    - Anonymization/tokenization
    - Randomization
    - Encryption
  - System level
    - Uniform procedures and regular testing
    - Security patching
    - Distributed data (no one can see everything at once)

86% of breaches are financially motivated
22% of breaches involve Phishing

New solutions will also come from **Big Data**!

Big Data – Summary

- The successful use of Big Data is not about implementing a particular technology, but a series of technologies implemented in a pipeline backed-up by processes and institutional culture

References and Attributions

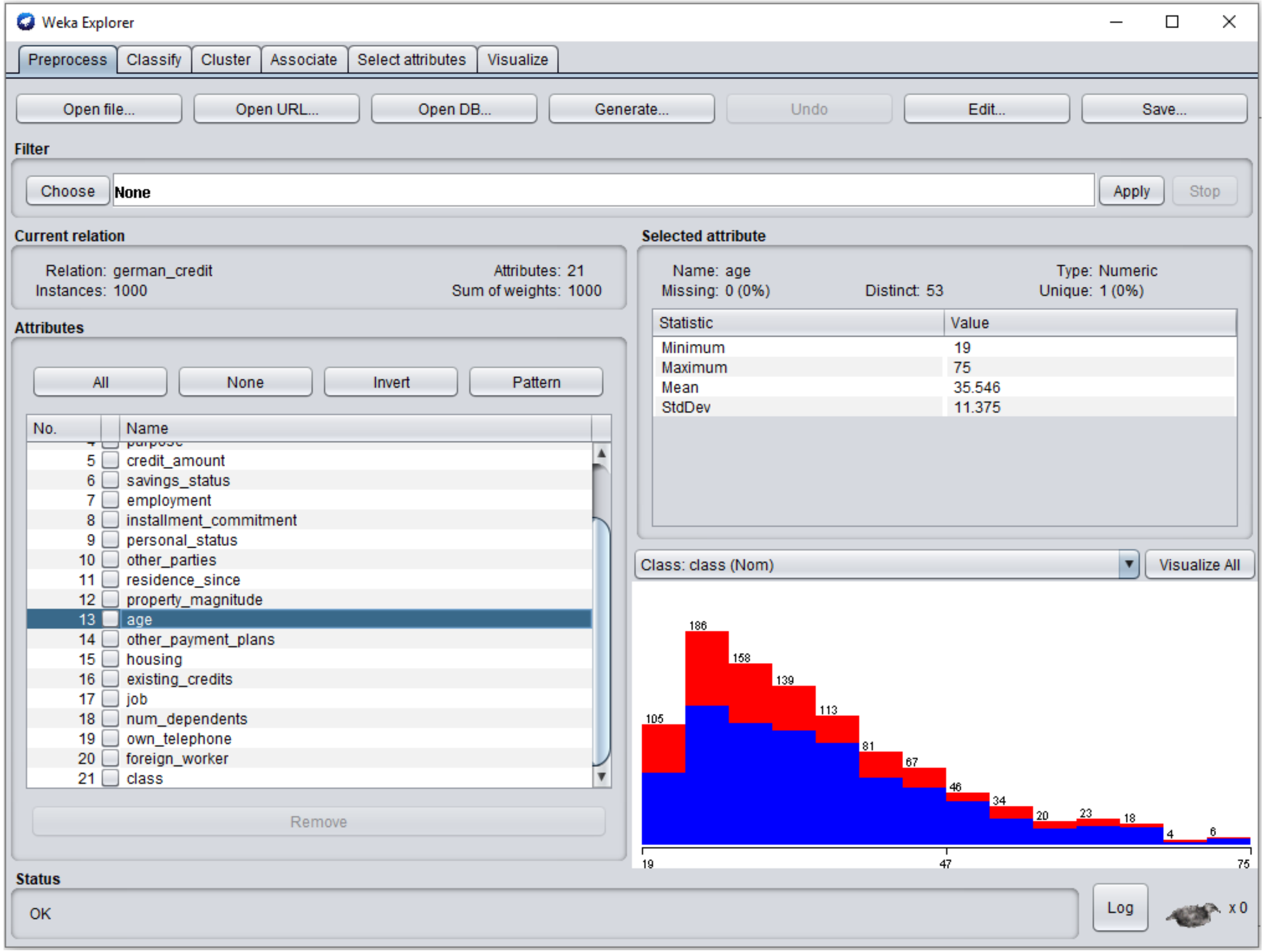
- Recalibrating global data center energy-use estimates, Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, Jonathan Koomey *Science*28 Feb 2020 : 984-986 https://science.sciencemag.org/content/367/6481/984
- BUNDESNETZAGENTUR Monitoring report 2019 - Key findings and summary https://www.bundesnetzagentur.de/SharedDocs/Downloads/EN/Areas/ElectricityGas/CollectionCompanySpecificData/Monitoring/KernaussagenEng_MB2019.pdf
- Domo “Data Never Sleeps” 8.0. Image courtesy of Domo. Full image at https://www.domo.com/learn/data-never-sleeps-8
- Bremen https://en.wikipedia.org/wiki/Bremen reproduced under Creative Commons Attribution-NonCommercial 3.0 License
- “Machine Learning” https://xkcd.com/1838/ reproduced under Creative Commons Attribution-NonCommercial 2.5 License
- Spurious Correlations, https://tylervigen.com/view_correlation?id=341 reproduced under Creative Commons Attribution-NonCommercial 4.0 License
- European Data Strategy, European Commission, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy
- The Changing Face of Data Security: 2020 Thales Data Threat Report Global Edition, Thales, https://cpl.thalesgroup.com/sites/default/files/content/research_reports_white_papers/field_document/2020-04/2020-data-threat-report.pdf
- 2020 Data Breach Investigations Report, Verizon, https://enterprise.verizon.com/resources/reports/dbir/

Data Science Tools

- GUI interfaces (no programming skills needed):
  - Weka (demonstrated at the end if enough time)
  - RapidMiner (commercial, but free for academic purpose)
- Languages
  - Python with scikit-learn
  - R
  - IDL (commercial)
  - MATLAB (commercial)
  - C++ with ROOT /by CERN
  - Julia
- Deep learning
  - TensorFlow + Keras (C++, Python, other languages supported unofficially) /by Google
  - Torch (Lua, Python) /by Facebook

- Also: Neural network playground: https://playground.tensorflow.org

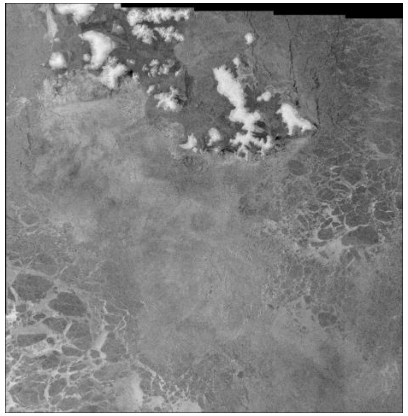
Demonstration of Weka



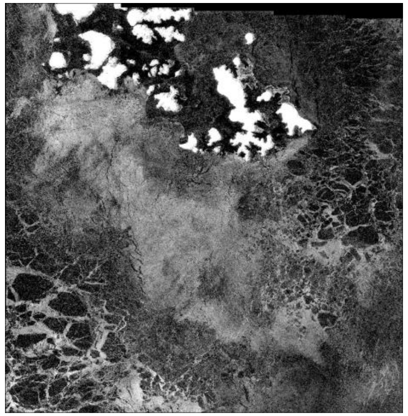
Example: Sea Ice mapping

Sentinel-1 Synthetic Aperture Radar (SAR)

- year-round imaging, independent on weather conditions
- good coverage over the Arctic at 40 meter resolution and 400 km swath width
- White spots ono top are island, fractured area at bottom left are ice flows



co-polarization band

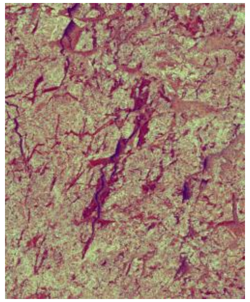


cross-polarization band

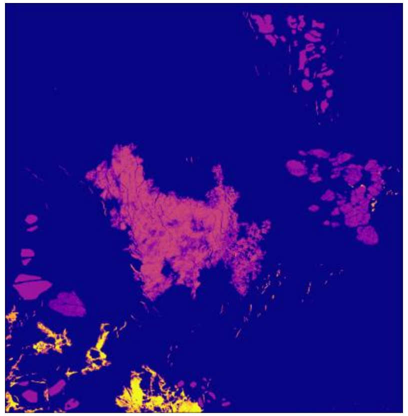
Classification:

- preprocessing
- texture descriptors calculation
- classifier training
- classifier prediction

Optical image

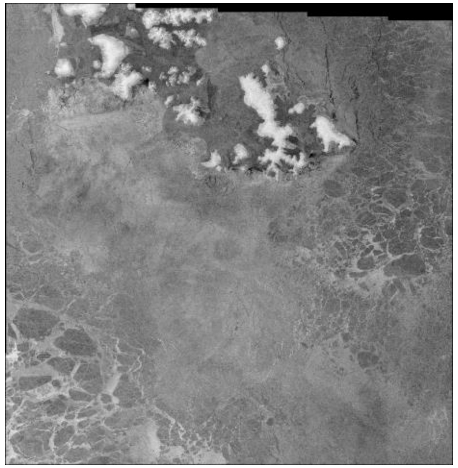


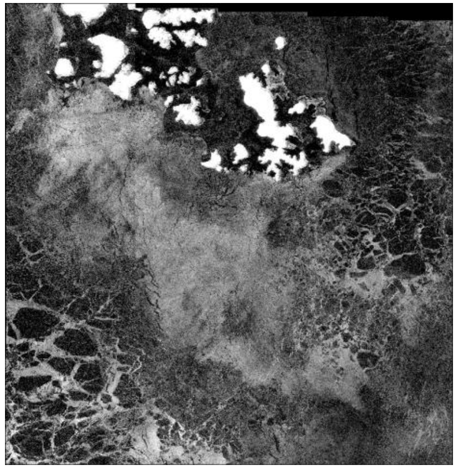
SAR image (pseudocolor)

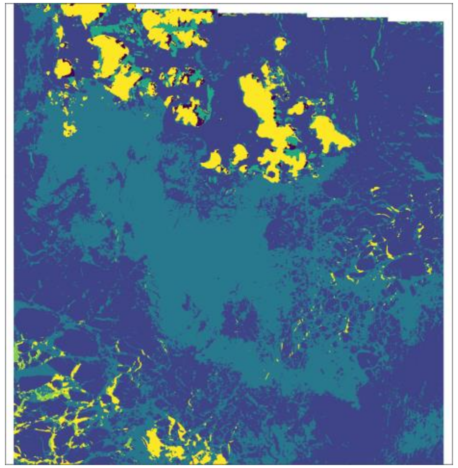


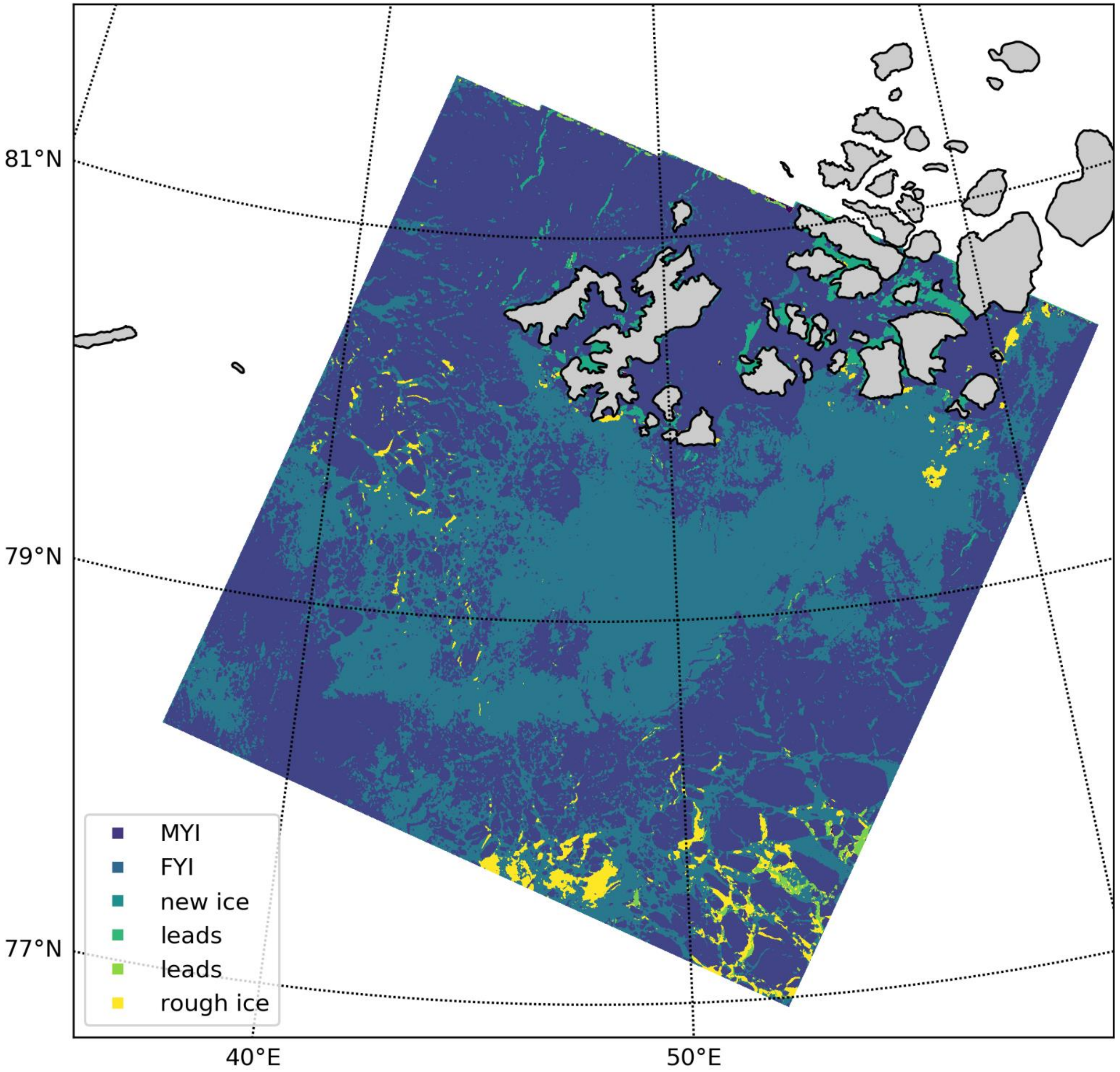
manual labels, unlabeled areas are blue

Example: semi-automatic classification


Sentinel-1 co-pol band

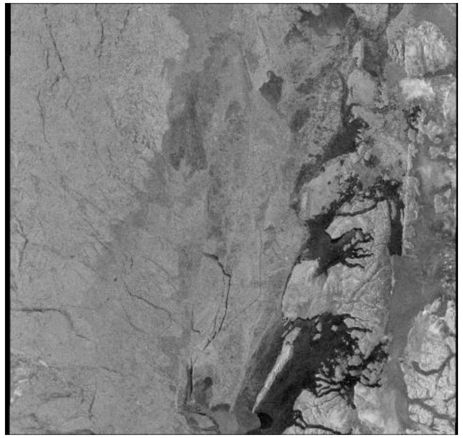

Sentinel-1 cross-pol band


Predicted labels

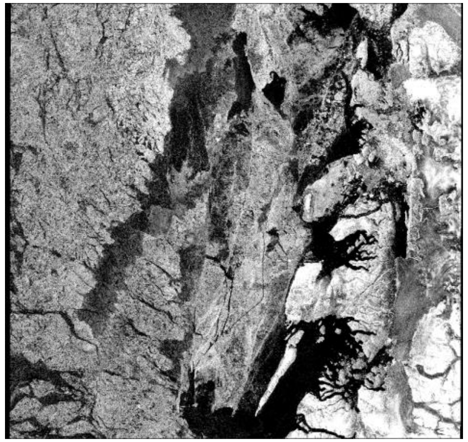


Franz Josef Land

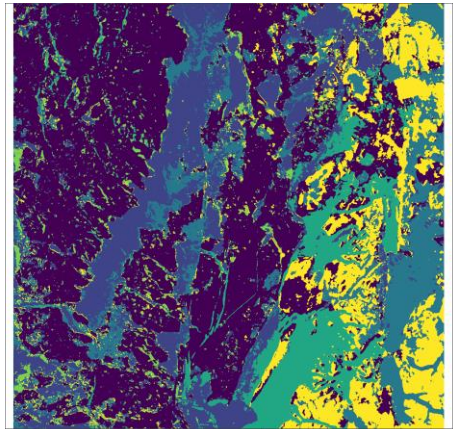
Example: automated classification



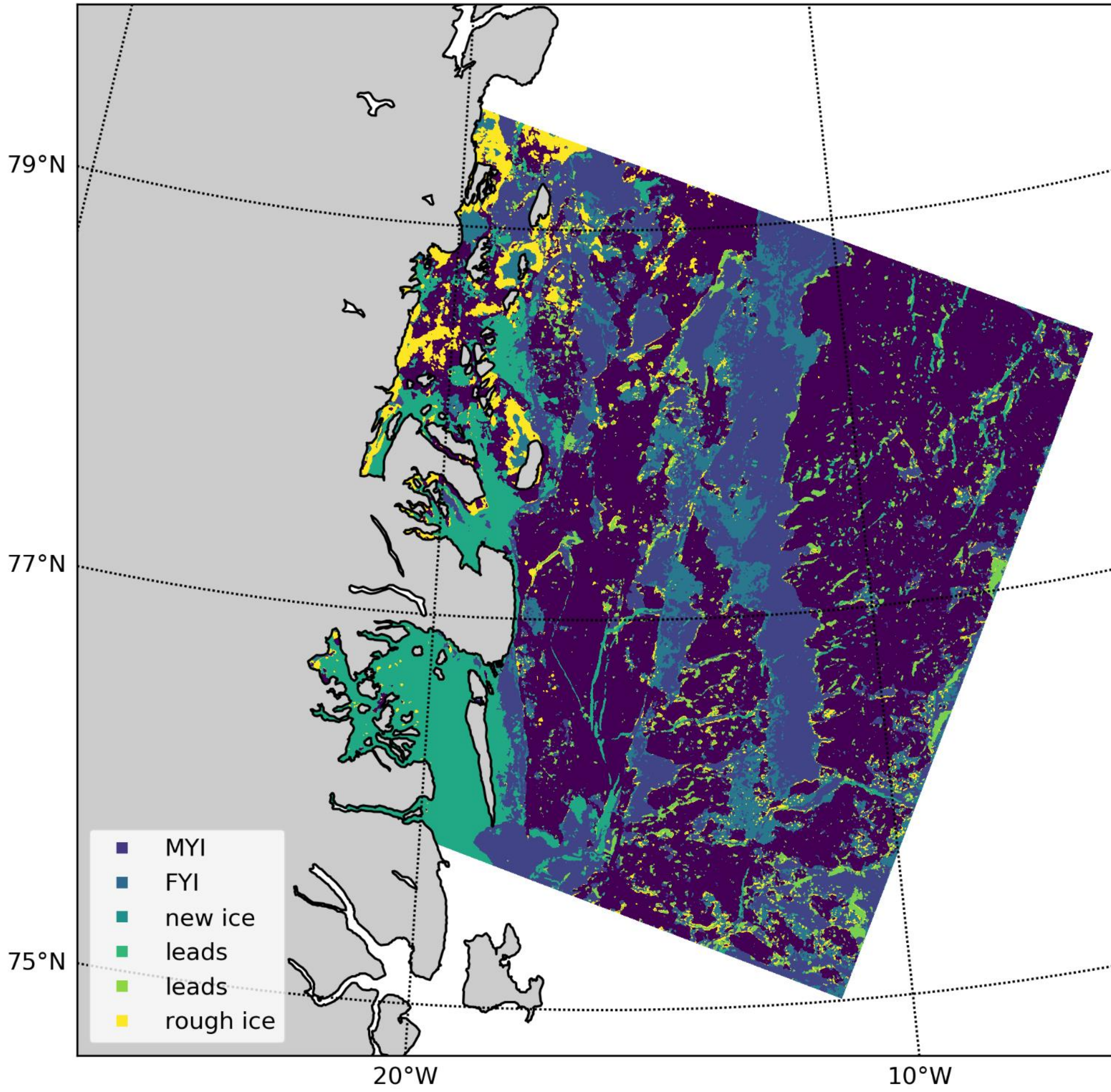
Sentinel-1 co-pol band



Sentinel-1 cross-pol band



Predicted labels



Greenland

Book recommendations

Basics about Machine Learning can be learned from the following books:

- "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani and Jerome Friedman (https://web.stanford.edu/~hastie/ElemStatLearn/)
- "An Introduction to Statistical Learning: With Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (https://ebooksbag.com/pdf-epub-an-introduction-to-statistical-learning-with-applications-in-r-download/)

Be aware that the field of data science is evolving quickly. Books can only introduce the field, but are hardly state-of-the-art.

Thank you!

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR)
German Aerospace Center
Remote Sensing Technology Institute | SAR Signal Processing |
Maritime Safety and Security Lab, Bremen
Am Fallturm 9 | 28359 Bremen

Björn Tings
bjoern.tings@dlr.de