

Data Science Care Team

```
sudo dnf update
```

2022-04-11, Rick de Klerk

"DATA IS THE NEW OIL"

Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.

From the beginning of recorded time until 2003, we created

5 exabytes of data.

(5 billion gigabytes)

In 2011 the same amount was created every two days.

By 2013, it's expected that the time will shrink to 10 minutes.

Every hour, we create enough Internet traffic to fill

7 billion DVDs.

Side by side, that's that's seven times the height of Everest!

There are nearly as many bits of information in the digital universe as there are stars in our actual universe.

As of August 2012, there were just over

4 million articles in the English Wikipedia.

There are **133 million** BLOGS on the web.

Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, **high-frequency traders**, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.

These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave

5 milliseconds

from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.

With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.

This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).

How they save 5 milliseconds

The depth of the Atlantic Ocean varies.

The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.



60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.

10% of all photos ever taken were taken in 2011.

80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 64% of citizens are smartphone users.)

English is the dominant language of the web. But by 2014 it will be **Chinese**, if its current rate of increase continues.

Top languages used on the web (May 2011):



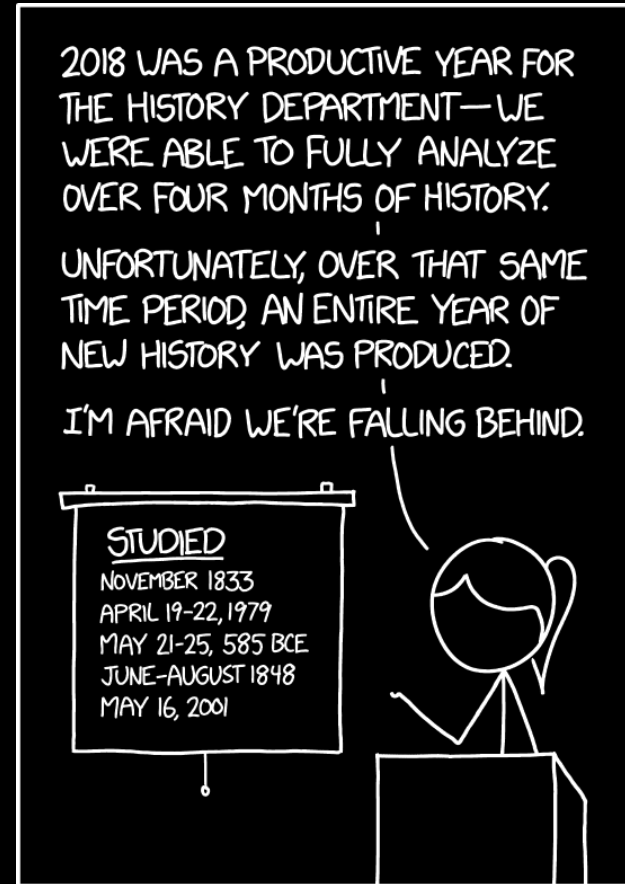
247 billion EMAILS are sent every day. (Up to 80% are spam.)

50% of 8-year-old kids in the U.S. are given access to a smartphone.



Relevantie voor onderzoek(ers)

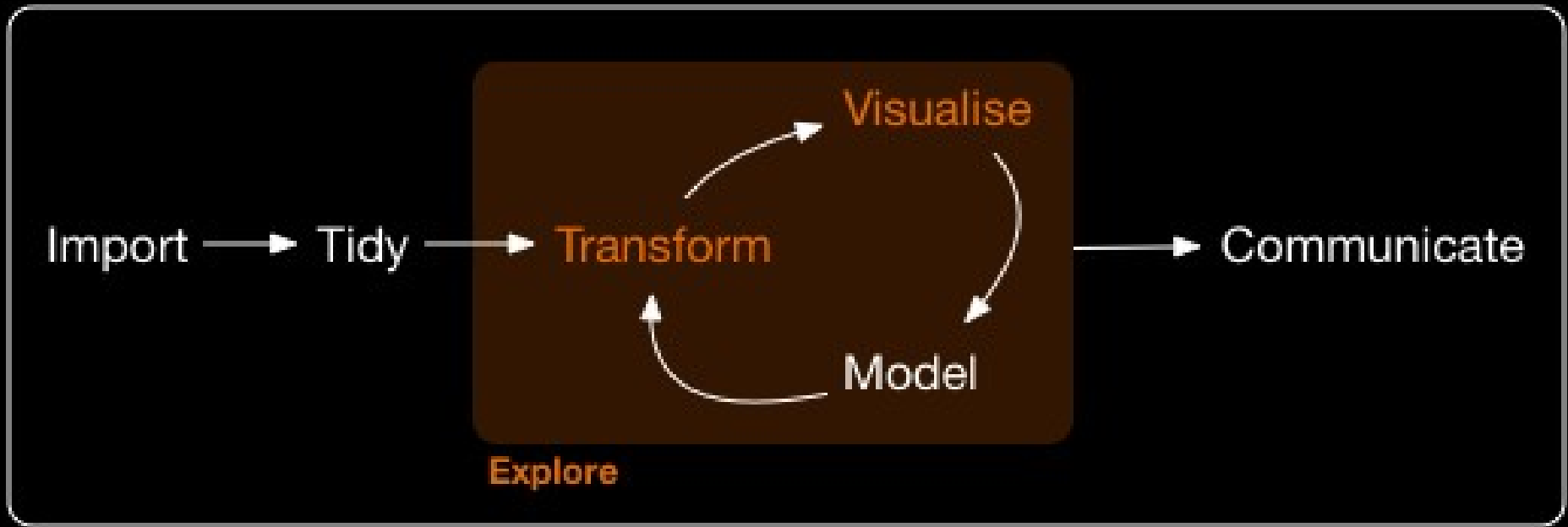
- Data stortvloed
- Open Science
- Oplossingen
- Digital skills



“Data science is the process
by which data becomes
understanding, knowledge
and insight”

- Hadley Wickham, Chief Scientist at RStudio

Data Science proces



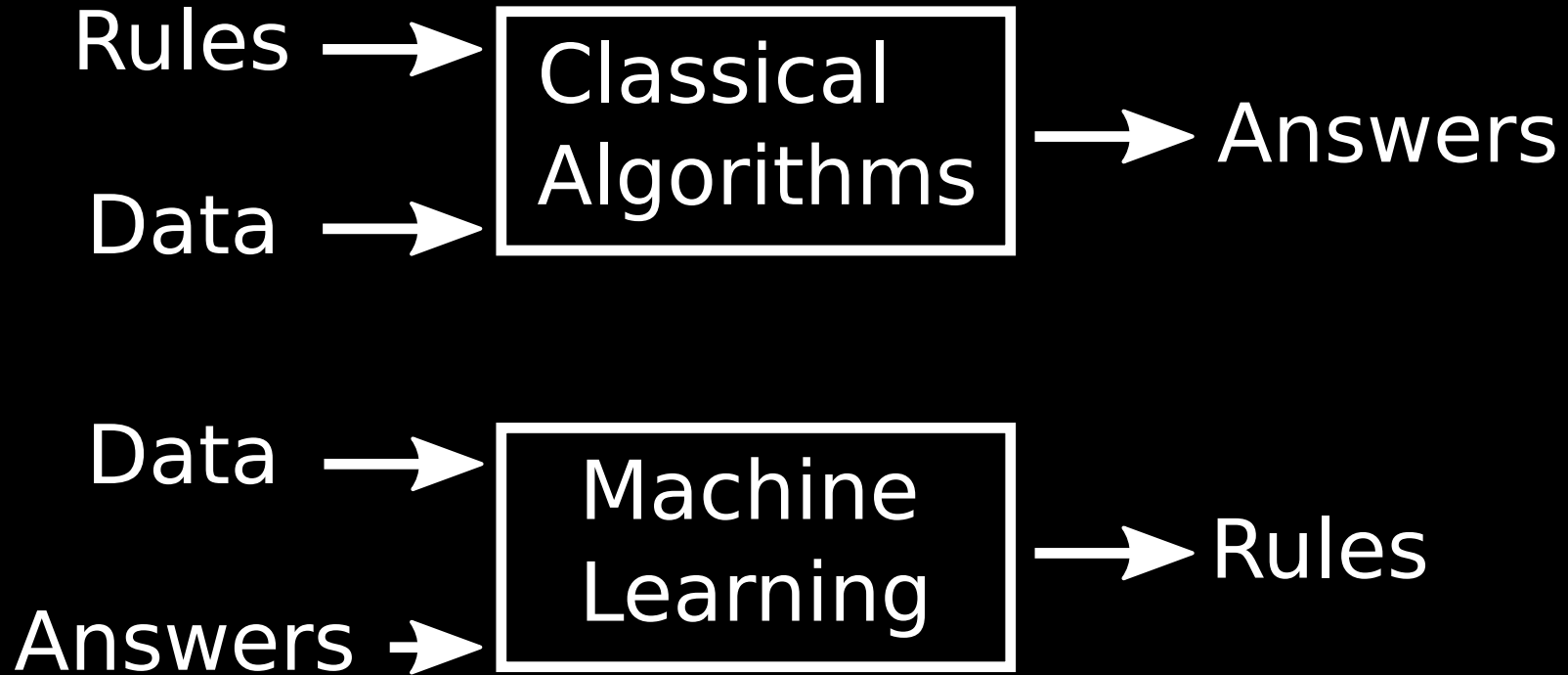
Program

Machine learning

Machine learning



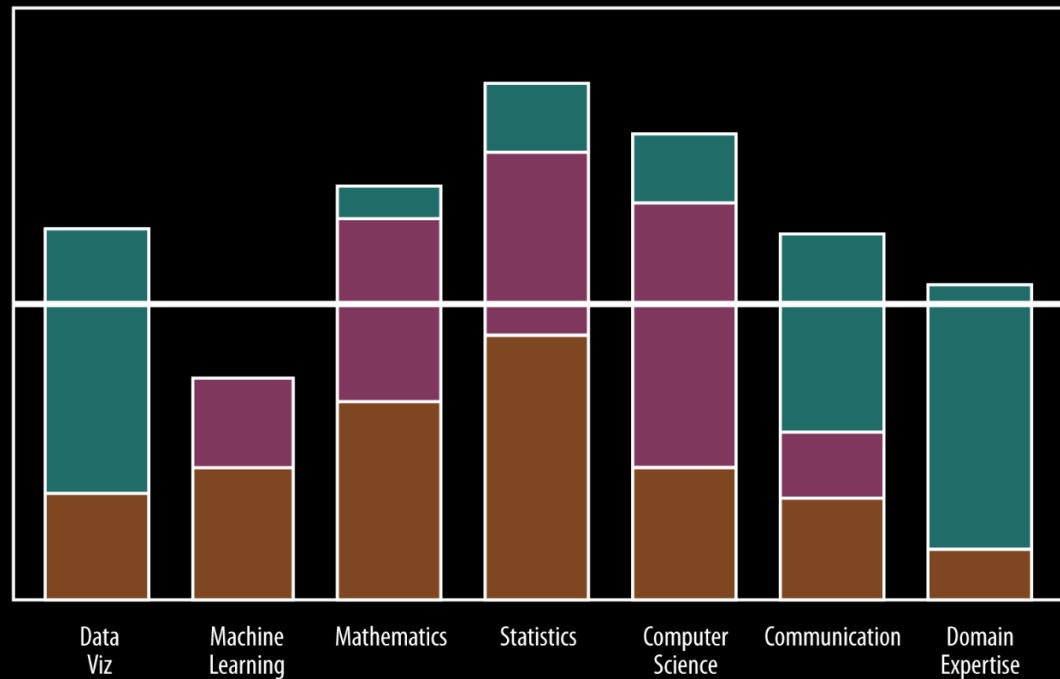
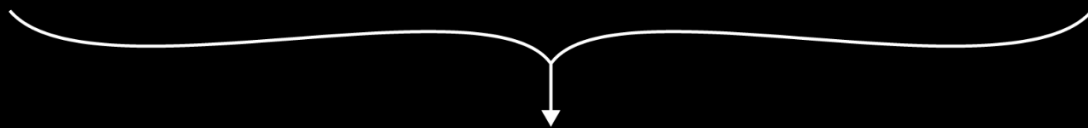
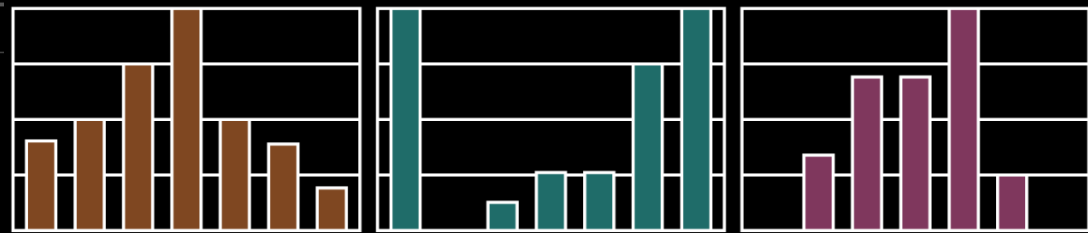
Machine learning



Doing data science

No one person can be the perfect data scientist, so we need teams.

- Technical skills
- People skills
- Science skills



Data-intensief onderzoek

1. Behoeften definiëren: digitale vaardigheden, kaders, en rollen
2. Training aanbieden
3. Gemeenschappen opbouwen
4. Beloningsstructuren
5. Bredere factoren

Digitale vaardigheden

- Informatie en data geletterdheid
- Communicatie en samenwerking
- Digitale content creation
- Veiligheid

“You can’t do data science with a GUI”

OECD Global Science Forum. 2020. Building digital workforce capacity and skills for data-intensive science.

Hadley Wickham. 2018. “You can’t do data science in a GUI”

DSCCT Producten

- Deze presentatie

Data Science Care Team

```
sudo dnf update
```

2022-04-11, Rick de Klerk

DSC T Producten

- Deze presentatie
- Workshop: Jupyter

The image displays two JupyterLab notebooks. The left notebook, titled 'Data.ipynb', shows the process of loading data from a CSV file using Pandas and a GeoJSON file. The right notebook, titled 'Lorenz.ipynb', shows the Lorenz attractor with interactive sliders for parameters sigma, beta, and rho, and a plot of the attractor.

Open a CSV file using Pandas

```
[4]: import pandas
df = pandas.read_csv('../data/iris.csv')
df.head(20)
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa
10	5.4	3.7	1.5	0.2	setosa
11	4.8	3.4	1.6	0.2	setosa
12	4.8	3.0	1.4	0.1	setosa
13	4.3	3.0	1.1	0.1	setosa
14	5.8	4.0	1.2	0.2	setosa
15	5.7	4.4	1.5	0.4	setosa
16	5.4	3.9	1.3	0.4	setosa
17	5.1	3.5	1.4	0.3	setosa
18	5.7	3.8	1.7	0.3	setosa
19	5.1	3.8	1.5	0.3	setosa

Read a GeoJSON file

```
[2]: import json
with open('../data/Museums_in_DC.geojson') as f:
    s = json.loads(f.read())
[3]: s
```

```
{'type': 'FeatureCollection',
 'features': [{'type': 'Feature',
 'properties': {'OBJECTID': 1,
 'ADDRESS': '716 MONROE STREET NE',
 'NAME': 'AMERICAN POETRY MUSEUM',
 'ADDRESS_ID': 309744,
 'LEGALNAME': 'HERITAGE US',
```

The Lorenz Differential Equations

Before we start, we import some preliminary libraries. We will also import (below) the accompanying `Lorenz.py` file, which contains the actual solver and plotting routine.

```
[1]: %matplotlib inline
from matplotlib import pyplot as plt
from ipywidgets import interactive, fixed
[2]: plt.style.use("dark_background")
```

We explore the Lorenz system of differential equations:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= \rho x - y - xz \\ \dot{z} &= -\beta z + xy\end{aligned}$$

Let's change (σ, β, ρ) with `ipywidgets` and examine the trajectories.

```
[3]: from lorenz import solve_lorenz
w = interactive(solve_lorenz, sigma=(0.0, 50.0), rho=(0.0, 50.0))
w
```

sigma: 10.00
beta: 2.67
rho: 28.00

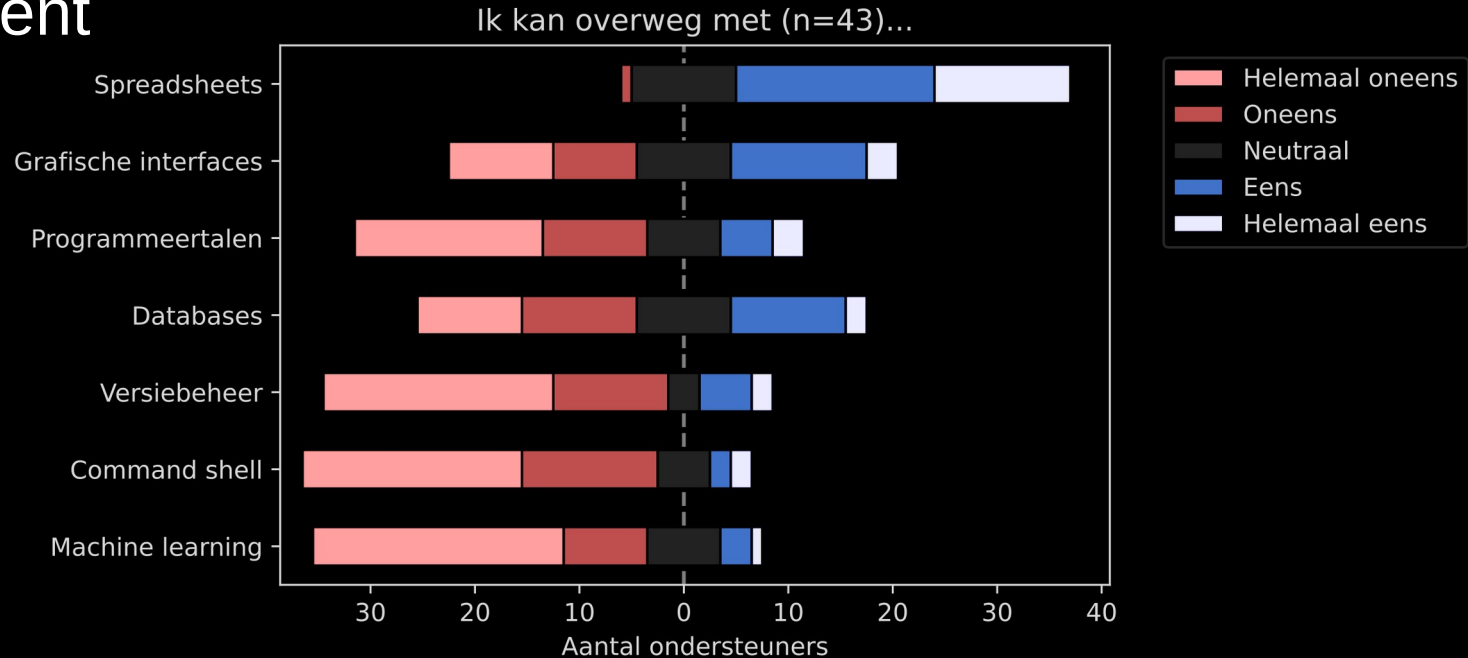
For the default set of parameters, we see the trajectories swirling around two points, called attractors.

The object returned by `interactive` is a `Widget` object and it has attributes that contain the current result and arguments:

```
[4]: t, x, y, z = w.result
```

DSCCT Producten

- Deze presentatie
- Workshop: Jupyter
- Needsassessment



DSCCT Producten

- Deze presentatie
- Workshop: Jupyter
- Needsassessment
- Workshop: Data Carpentry i.s.m. eScience Center



netherlands

eScience center

Producten

- Deze presentatie
- Workshop: Jupyter
- Needsassessment
- Workshop: Data Carpentry i.s.m. eScience Center
- Animatie Notebooks

Animatie notebooks

Producten

- Deze presentatie
- Workshop: Jupyter Notebooks
- Needsassessment
- Workshop: Data Carpentry i.s.m. eScience Center
- Animatie Notebooks
- Adviesdocument

Producten

- Deze presentatie
- Workshop: Jupyter Notebooks
- Needsassessment
- Workshop: Data Carpentry i.s.m. eScience Center
- Animatie Notebooks
- Adviesdocument
- Overig









Overig: FAIR file listing

- Integreren van FAIR in workflow
- Open file extensies
- Integratie in explorer, research drive, etc.

Navigation: < > Home Projects misc demo_files

Search: [Search Icon]

View: [Grid Icon] [Dropdown Arrow] [List Icon] [Close Icon]

	Name	Size	Permissions	Modified	Type
Recent					
★ Starred					
Home					
Desktop					
Documents					
Downloads					
Music					
Pictures					
Videos					
Trash					
Projects					
demo_files					
+ Other Locations					
	 new	0 bytes	-rw-rw-r--	7 Apr 2021	Text
	 new.docx	4.2 kB	-rw-rw-r--	2:45 PM	Document
	 new.md	0 bytes	-rw-rw-r--	7 Apr 2021	Text
	 new.ods	6.9 kB	-rw-rw-r--	6 Apr 2021	Spreadsheet
	 new.odt	8.0 kB	-rw-rw-r--	2:44 PM	Document
	 new.py	23 bytes	-rw-rw-r--	7 Apr 2021	Text
	 new.rst	0 bytes	-rw-rw-r--	7 Apr 2021	Text
	 new.txt	0 bytes	-rw-rw-r--	7 Apr 2021	Text

Vragen?

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.

