

DATA SCIENCE AND ARTIFICIAL INTELLIGENCE (AI)

Dr Marcos Barreto

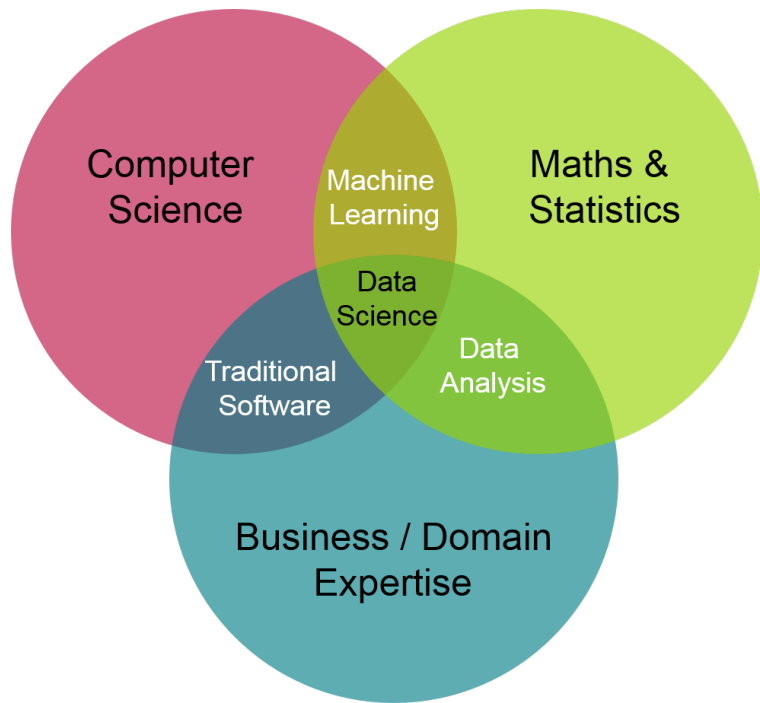
Department of Statistics
London School of Economics and Political Science (LSE)

w: marcosebarreto.github.io
e: m.e.barreto@lse.ac.uk

Bishop Douglass Catholic School
1st November 2023



WHAT IS DATA SCIENCE?

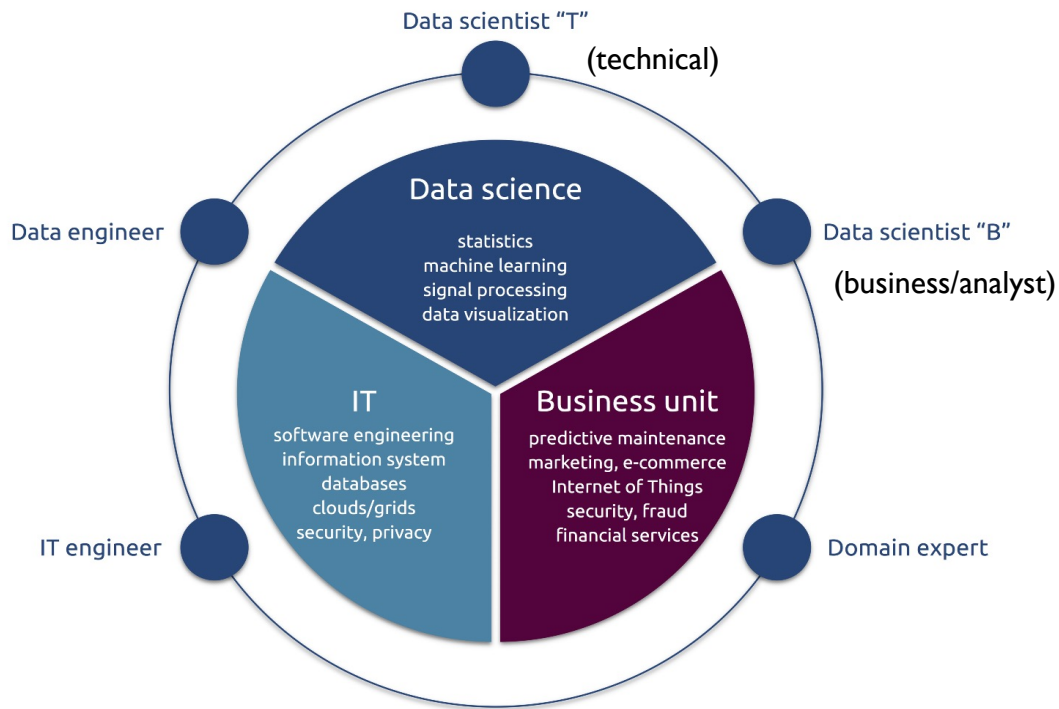


Initial (simplified?) definition
≅ 2010

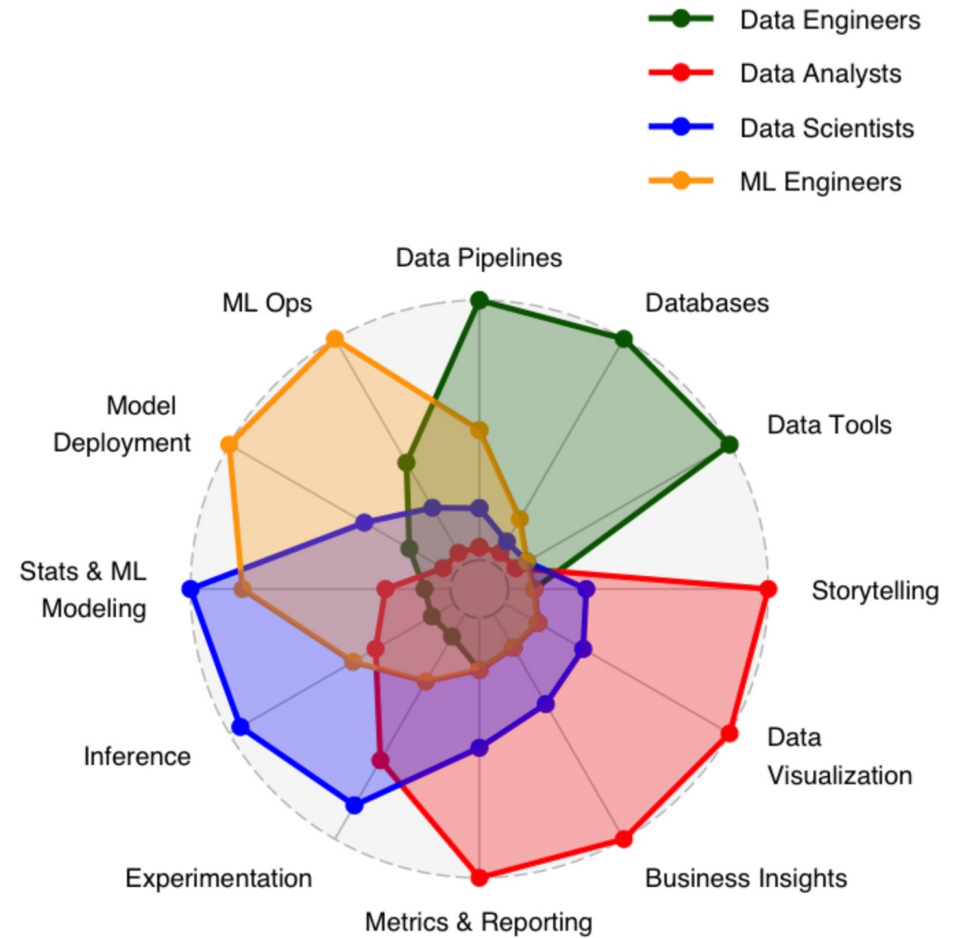


Recent definition
≅ 2020

DATA SCIENCE JOBS AND SKILLS



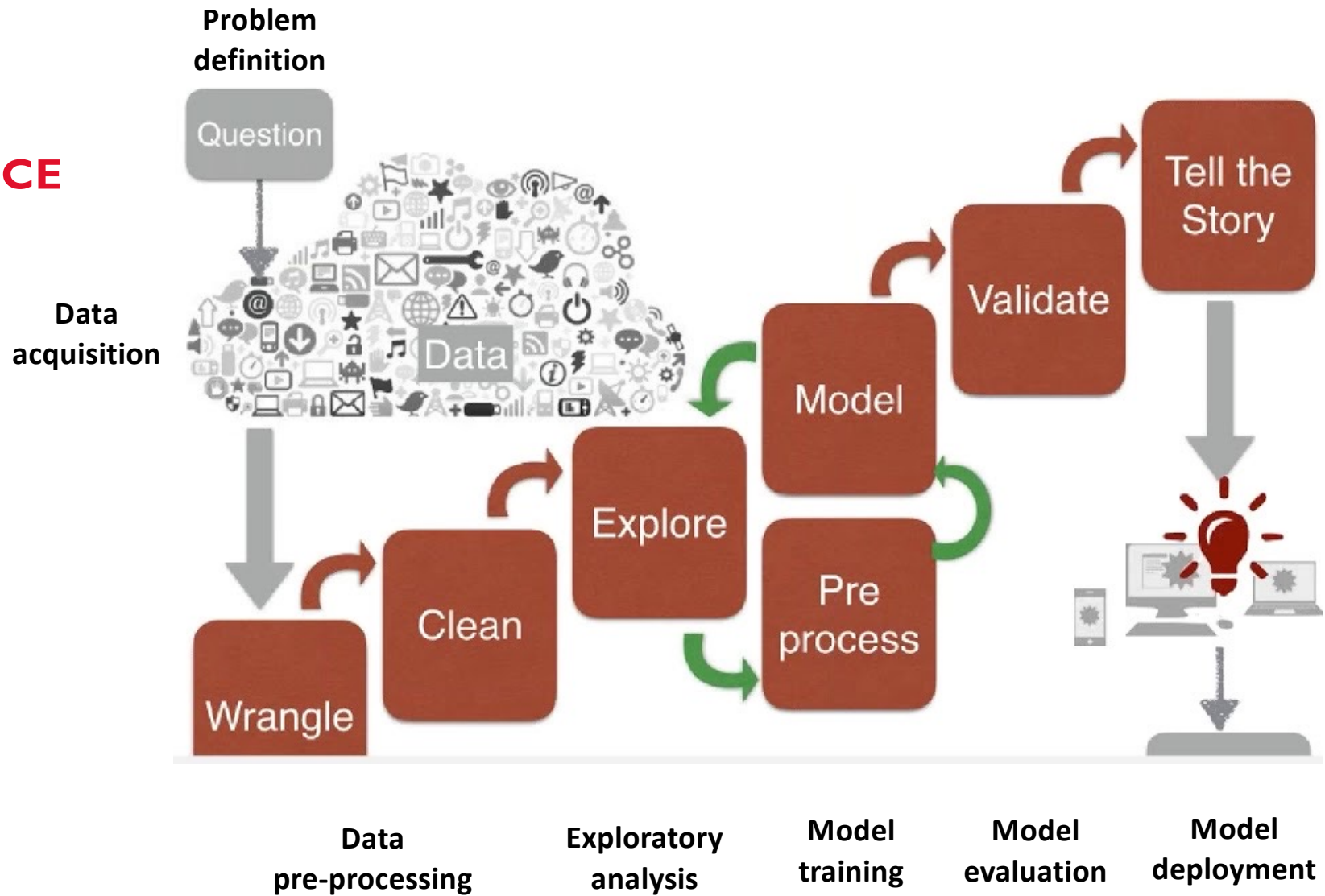
<https://towardsdatascience.com/the-data-science-ecosystem-industrial-edition-938582427466>



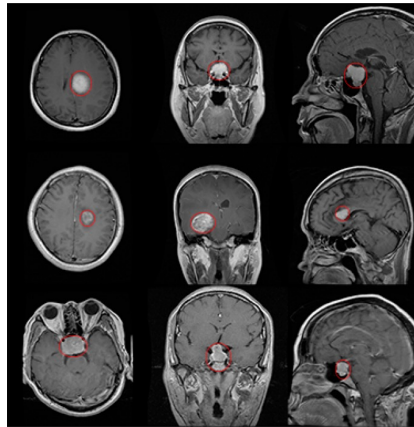
<https://www.datacaptains.com/blog/guide-to-data-roles>



DATA SCIENCE PIPELINE



DATA SCIENCE PIPELINE



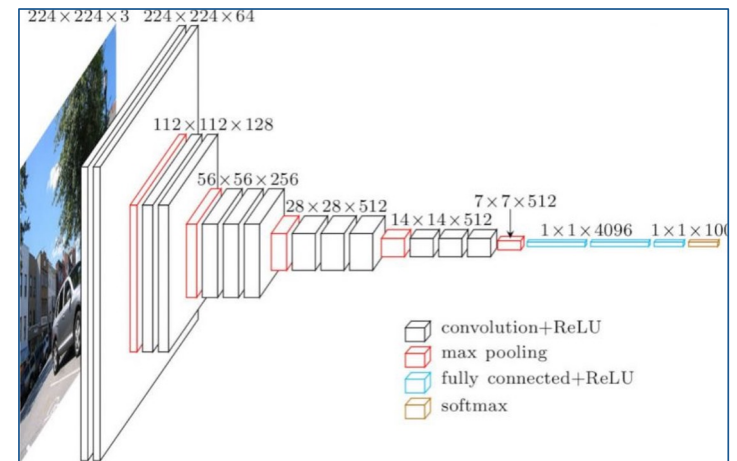
data = brain MRI

model = convolutional neural network

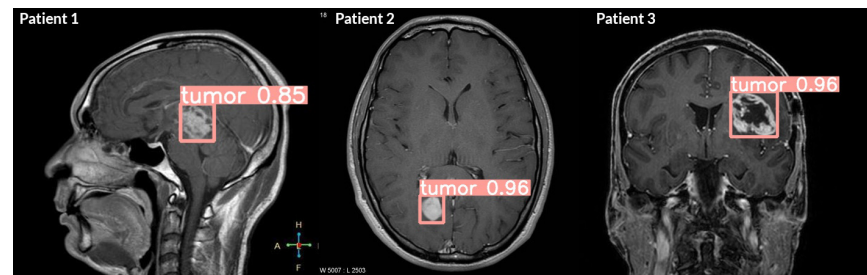
```

load_data()
split(train_data, test_data)
create_model()
train_model(train_data)
evaluate_model(test_data)
deploy_model(new_data)

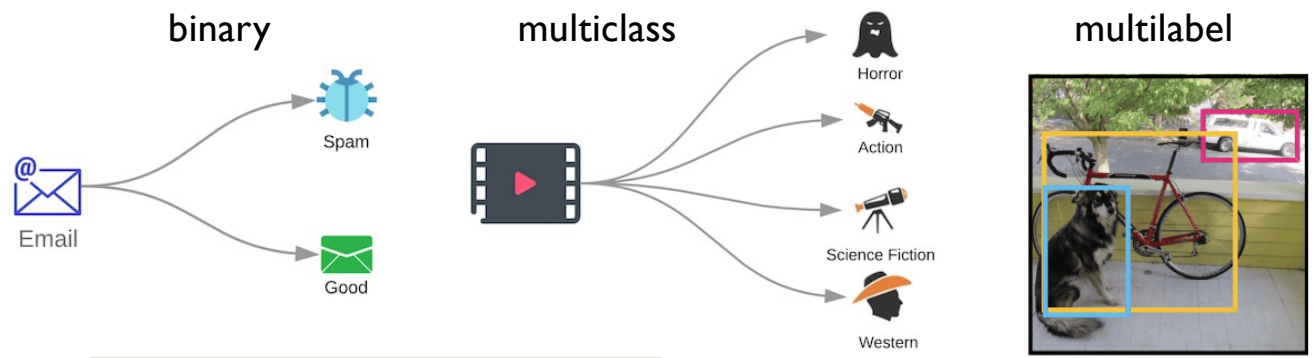
```



evaluation = tumor (yes/no)



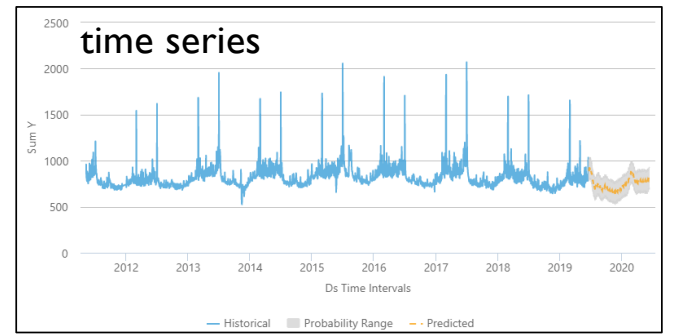
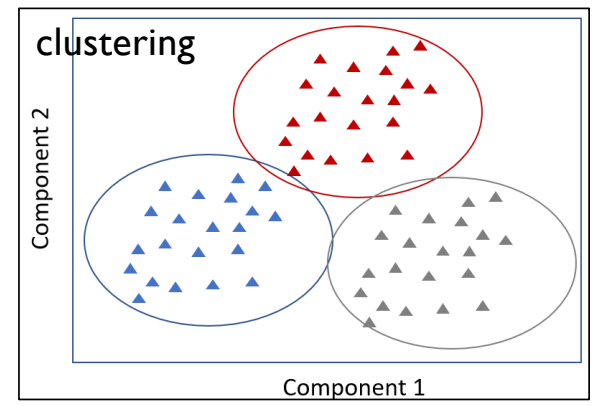
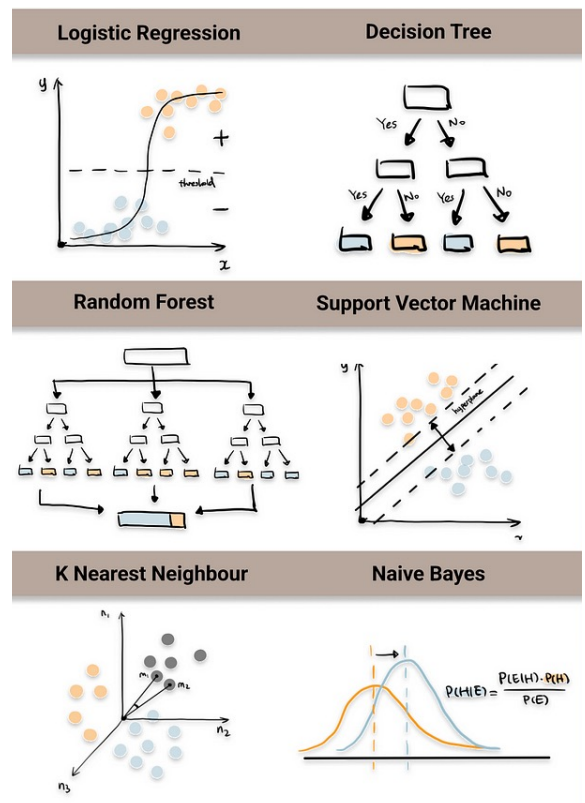
DATA SCIENCE PIPELINE



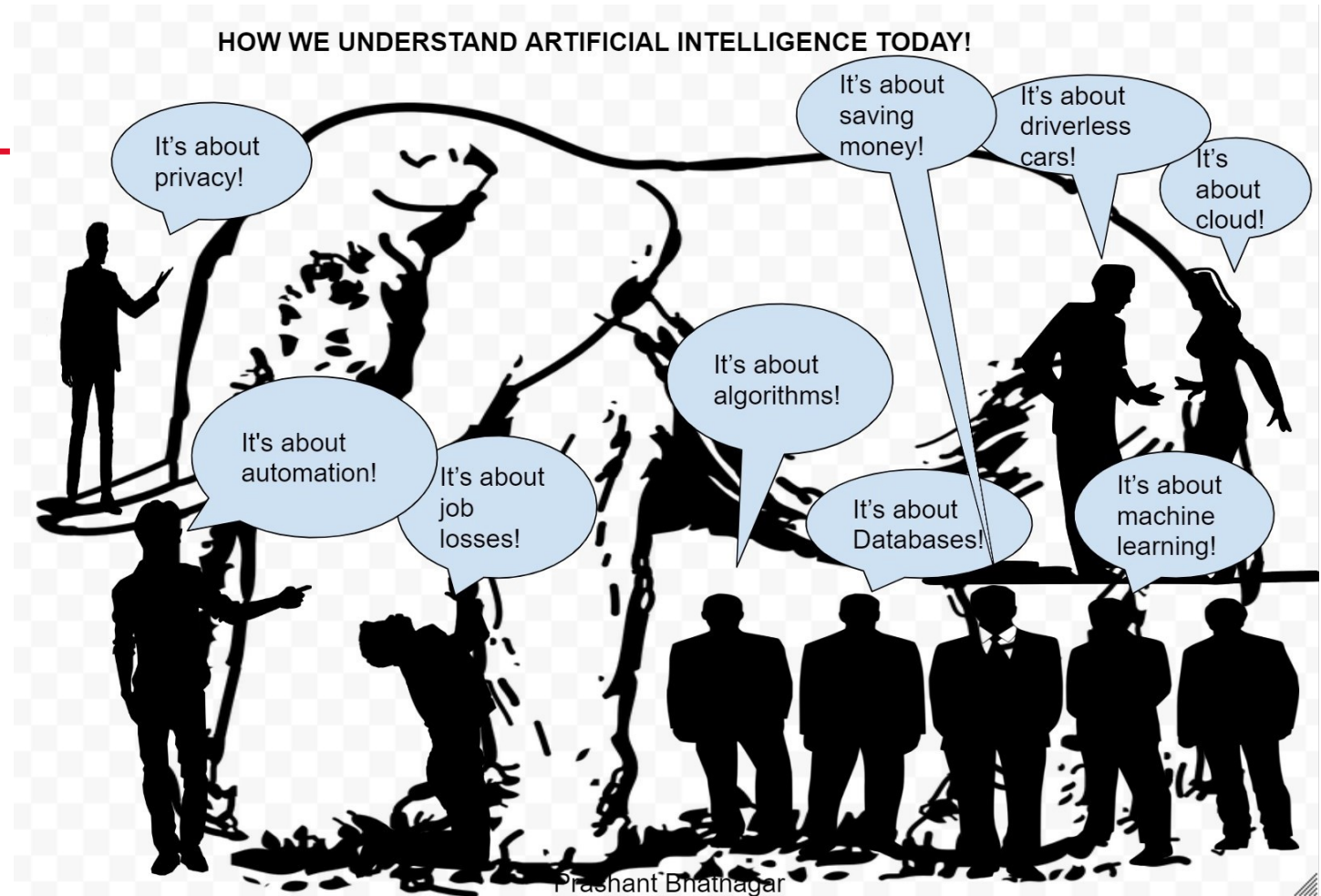
↓

```

load_data()
split(train_data, test_data)
create_model()
train_model(train_data)
evaluate_model(test_data)
deploy_model(new_data)
    
```



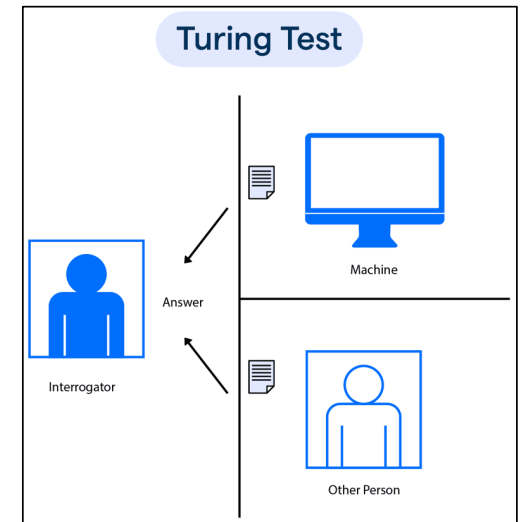
WHAT IS ARTIFICIAL INTELLIGENCE?



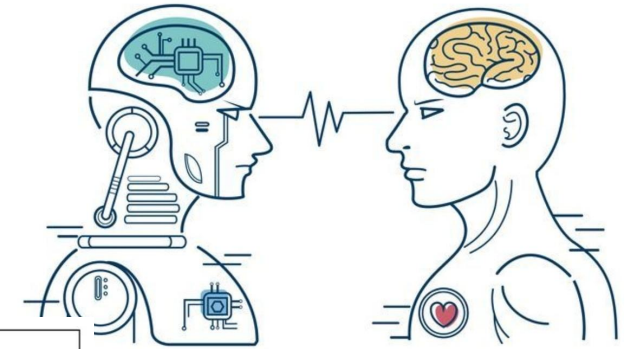
<https://medium.com/swlh/artificial-intelligence-the-story-of-blind-men-and-an-elephant-5781812d5554>

WHAT IS ARTIFICIAL INTELLIGENCE?

- ✓ No standard definition(!), but some consensus over time...
- ✓ **Alan Turing, 1950**
 - “A computer would deserve to be called intelligent if it could deceive a human into believing that it was human”
- ✓ **John McCarthy, 1955**
 - “the science and engineering of making intelligent machines”
- ✓ **Nils J. Nilsson, 2010**
 - “activity devoted to making machines intelligent, and intelligence is that quality that enables an entity to function appropriately and with foresight in its environment”
- ✓ **Andrew Ng, 2023**
 - “the ability of machines to perform tasks that would normally require human intelligence”



WHAT IS ARTIFICIAL INTELLIGENCE?

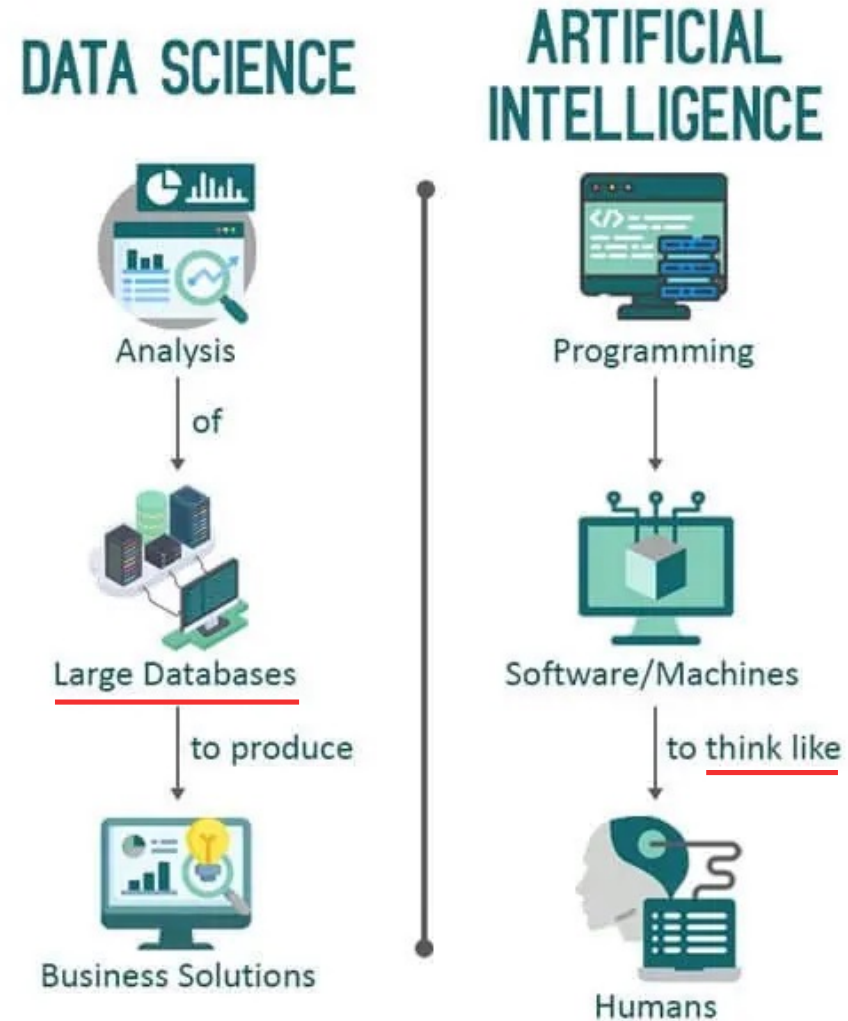


<https://skywell.software/blog/artificial-intelligence-vs-human-intelligence/>

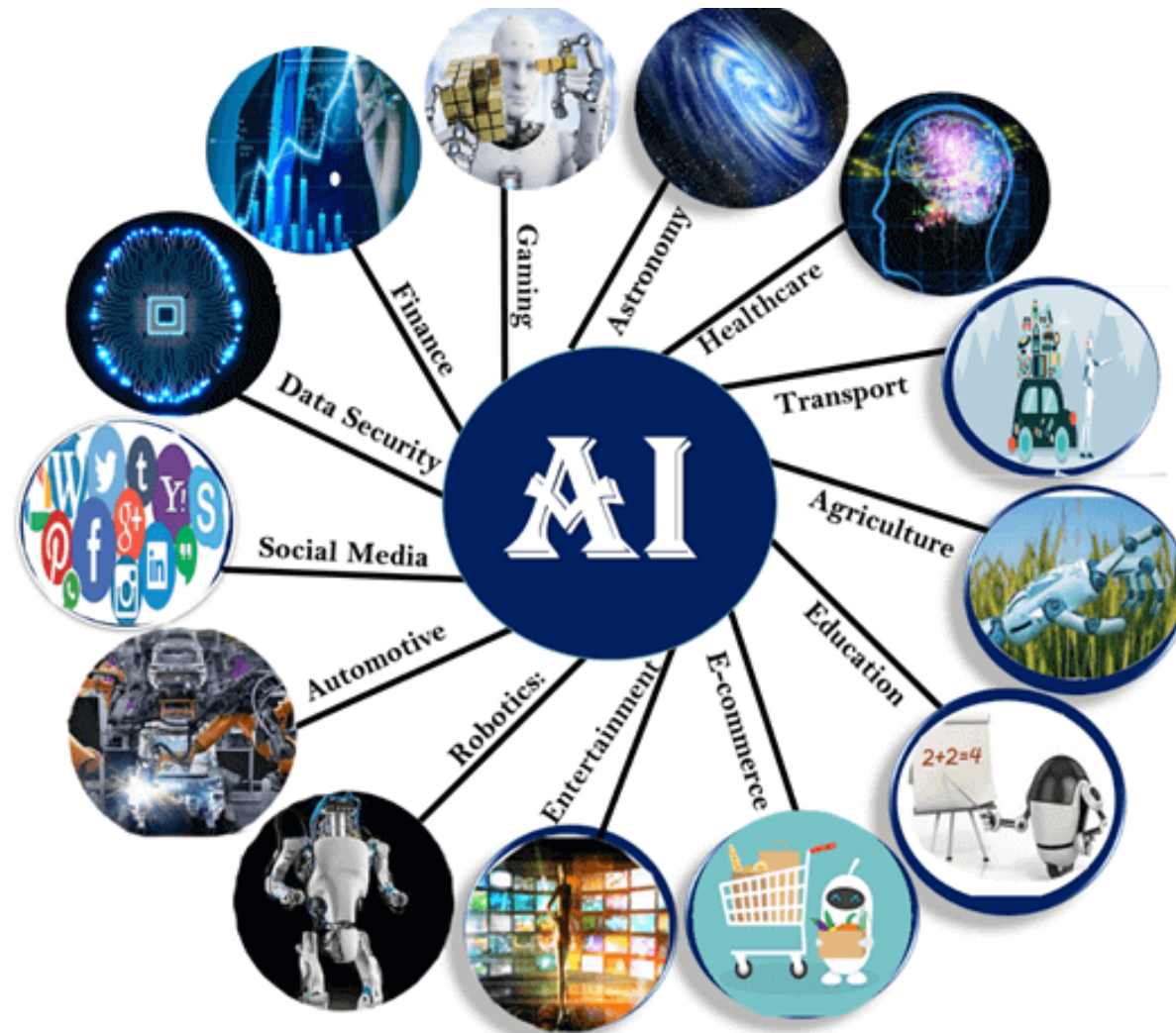
	Humans	Computers (AI)
Data	Sight, Sound, Smell, Touch	Audio, Video, Text, Numbers
Data collection assets	Eyes, Ears, Nose, Skin	Sensors (IoT), Cameras, Cookies, Text Inputs, Usage data
Data storage	Human brain (Hardware) Knowledge (Software)	Memory Devices (Hardware) Database (Software)
Data Processing	Human brain (Hardware) Thoughts, Education, Experiences (Software)	Computer Chip (Hardware), Programs Algorithms (Software)
Action and Presentation	Actions Thoughts, Speech, Emotions, Facial expressions	Data visualisation tools, Charts, Graphs, Automated actions by machines, Predictive Analytics, Suggestions, Recommendations

DATA SCIENCE AND ARTIFICIAL INTELLIGENCE

✓ Summary



AI APPLICATIONS



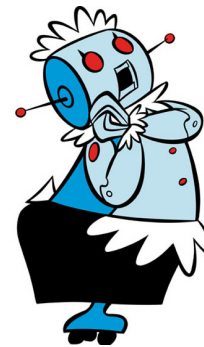
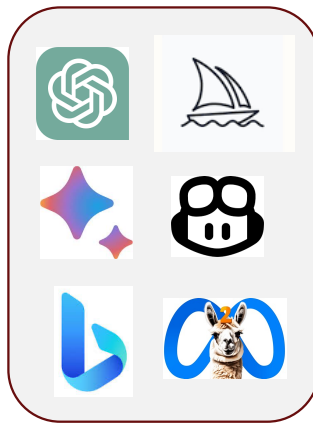
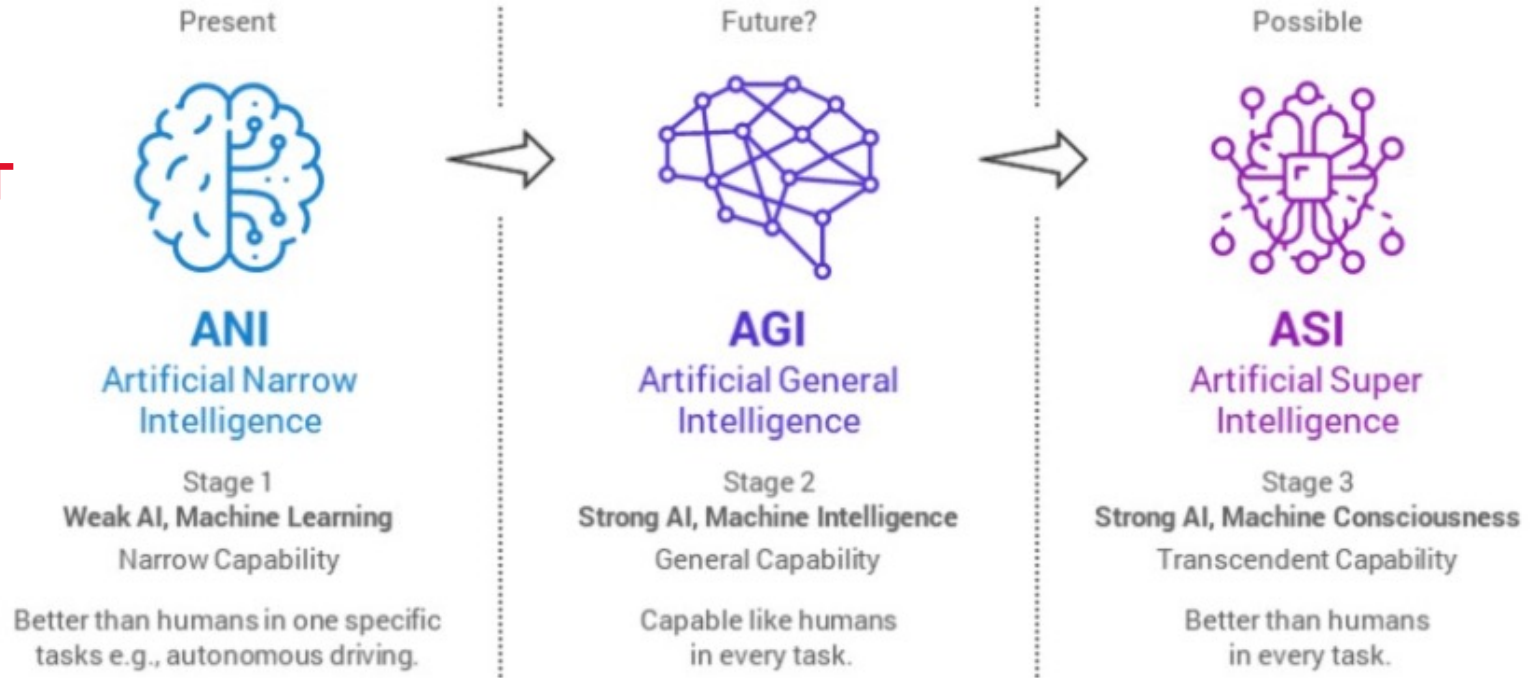
AI ALPHABET SOUP

- ✓ Artificial narrow intelligence (ANI)
- ✓ Artificial general intelligence (AGI)
- ✓ Artificial super-intelligence (ASI)



AI ALPHABET SOUP

<https://www.linkedin.com/pulse/agi-simplified-rajesh-dangi/>



Rosey / The Jetsons



HOW ABOUT GENERATIVE AI?

Defining Generative AI

To understand generative artificial intelligence (GenAI), we first need to understand how the technology builds from each of the AI subcategories listed below.

Expert System AI

Programmers teach AI exactly how to solve specific problems by providing precise instructions and steps.

Artificial Intelligence

The theory and methods to build machines that think and act like humans.

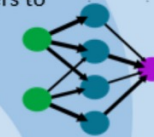


Machine Learning

The ability for computers to learn from experience or data without human programming.

Deep Learning

Mimics the human brain using artificial neural networks such as **transformers** to allow computers to perform complex tasks.



Generative AI

Generates new text, audio, images, video or code based on content it has been **pre-trained** on.



ChatGPT Midjourney Bard

HOW CHATGPT WORKS?

- ✓ ChatGPT is a “**text-to-text**” model
 - ✓ it produces a “reasonable continuation” of whatever text it’s got so far
 - ✓ the “reasonable continuation” is determined by a ranked list of words that might follow, together with “probabilities”

✓ Example

The best thing about AI is its ability to



input sentence
 (“prompt”)

word	probability
learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%



Out[=]= { The best thing about AI is its ability to,
 The best thing about AI is its ability to learn,
 The best thing about AI is its ability to learn from,
 The best thing about AI is its ability to learn from experience,
 The best thing about AI is its ability to learn from experience.,
 The best thing about AI is its ability to learn from experience. It,
 The best thing about AI is its ability to learn from experience. It's,
 The best thing about AI is its ability to learn from experience. It's not }

HOW CHATGPT WORKS?

ChatGPT 2 (2019, “zero temperature”)

The best thing about AI is its ability to learn from experience. It's not just a matter of learning from experience, it's learning from the world around you. The AI is a very good example of this. It's a very good example of how to use AI to improve your life. It's a very good example of how to use AI to improve your life. The AI is a very good example of how to use AI to improve your life. It's a very good example of how to use AI to

ChatGPT 3 (2020, “zero temperature”)

The best thing about AI is its ability to automate processes and make decisions quickly and accurately. AI can be used to automate mundane tasks, such as data entry, and can also be used to make complex decisions, such as predicting customer behavior or analyzing large datasets. AI can also be used to improve customer service, as it can quickly and accurately respond to customer inquiries. AI can also be used to improve the accuracy of medical diagnoses and to automate the process of drug discovery.

Temperature = 0..1, control for “randomness”

ChatGPT 2 (2019, temperature=0.8)

The best thing about AI is its ability to learn. I've always liked the
The best thing about AI is its ability to really come into your world and just
The best thing about AI is its ability to examine human behavior and the way it
The best thing about AI is its ability to do a great job of teaching us
The best thing about AI is its ability to create real tasks, but you can

ChatGPT 3 (2020, temperature=0.8)

The best thing about AI is its ability to learn and develop over time, allowing it to continually improve its performance and be more efficient at tasks. AI can also be used to automate mundane tasks, allowing humans to focus on more important tasks. AI can also be used to make decisions and provide insights that would otherwise be impossible for humans to figure out.

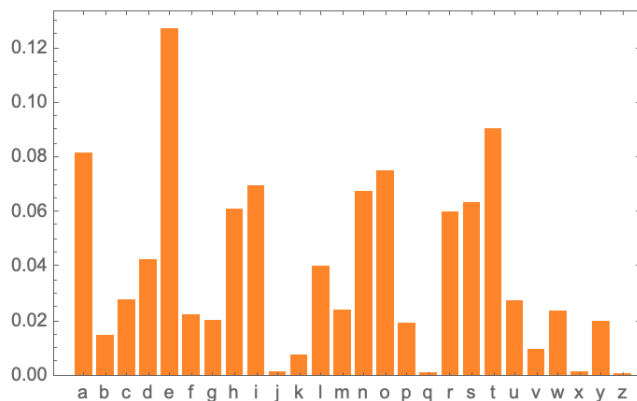
HOW CHATGPT WORKS?

- ✓ We can count letters in several documents and generate frequencies

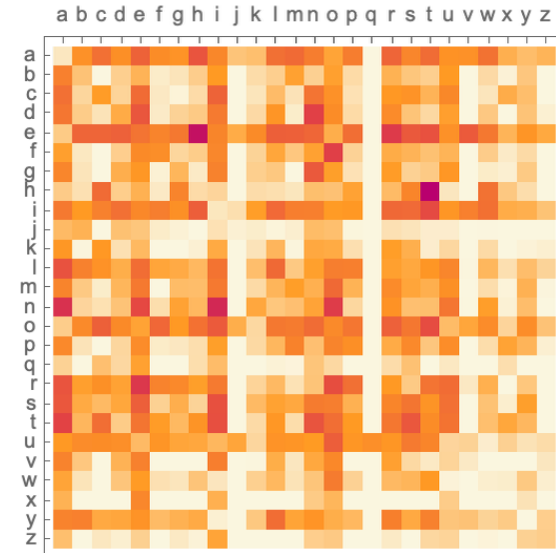
```
In[*]:= English LANGUAGE [ character frequencies ]
```

```
Out[*]:= { e → 12.7%, t → 9.06%, a → 8.17%, o → 7.51%, i → 6.97%, n → 6.75%,
           s → 6.33%, h → 6.09%, r → 5.99%, d → 4.25%, l → 4.03%, c → 2.78%, u → 2.76%,
           m → 2.41%, w → 2.36%, f → 2.23%, g → 2.02%, y → 1.97%, p → 1.93%, b → 1.49%,
           v → 0.978%, k → 0.772%, j → 0.153%, x → 0.150%, q → 0.0950%, z → 0.0740% }
```

Probabilities for letters (unigram)



Where do the probabilities come from?



Pair of letters (bigram)
 1st letter: across
 2nd letter: down



on inguman men ise forerenoft weat iofobato buc ous corew ousesetiv
 falle tinouco ryefo ra the ecederi pasuthrgr cuconom tra tesla wil tat pere thi

cat through shipping variety is made the aid emergency can the
 cat for the book flip was generally decided to design of
 cat at safety to contain the vicinity coupled between electric public
 cat throughout in a confirmation procedure and two were difficult music
 cat on the theory an already from a representation before a

(n-gram)

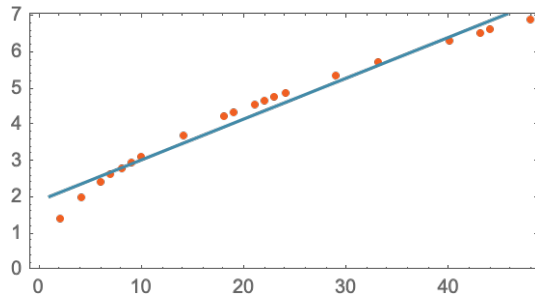


Why do we need a model and
which one to choose?

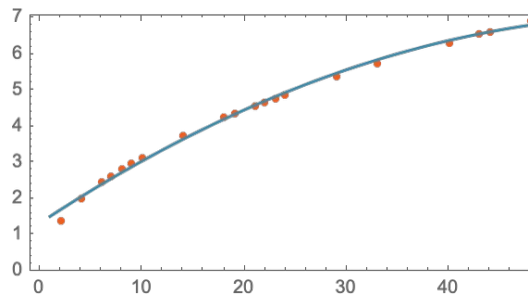
HOW CHATGPT WORKS?

- ✓ We need a model to estimate *n-gram* probabilities from a few hundred billion words
- ✓ Nearly 40,000 common words in English => possible 2-grams: 1.6 billion; for 3-grams: 60 trillion.

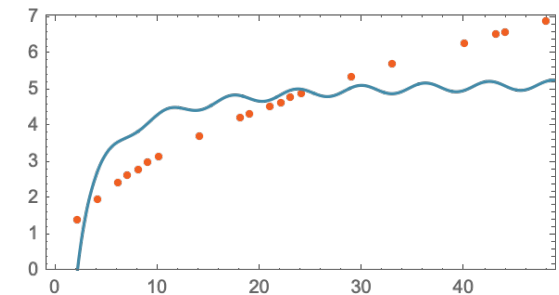
$$y = mx + c$$



$$a + bx + cx^2$$



$$a + b/x + c \sin(x)$$



Challenge #1: we need a **model of human-like language!**

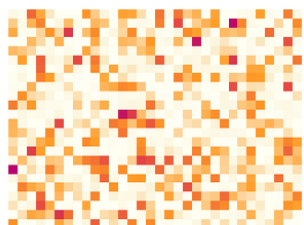
Challenge #2: we need to train **lots of parameters**

(large language model, *voilà!*)

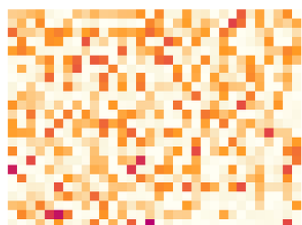
Version	Parameters (Unit: billion)
GPT - 1 - launched in 2015	1.1
GPT - 2 - launched in 2019	17.5
GPT - 3 - launched in 2020	1750
GPT - 4 - launched in 2023	10000

HOW CHATGPT WORKS?

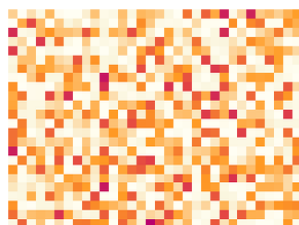
- ✓ Embedding = numerical representation of text
- ✓ Each word is assigned a number and “nearby words” are represented by “nearby numbers”



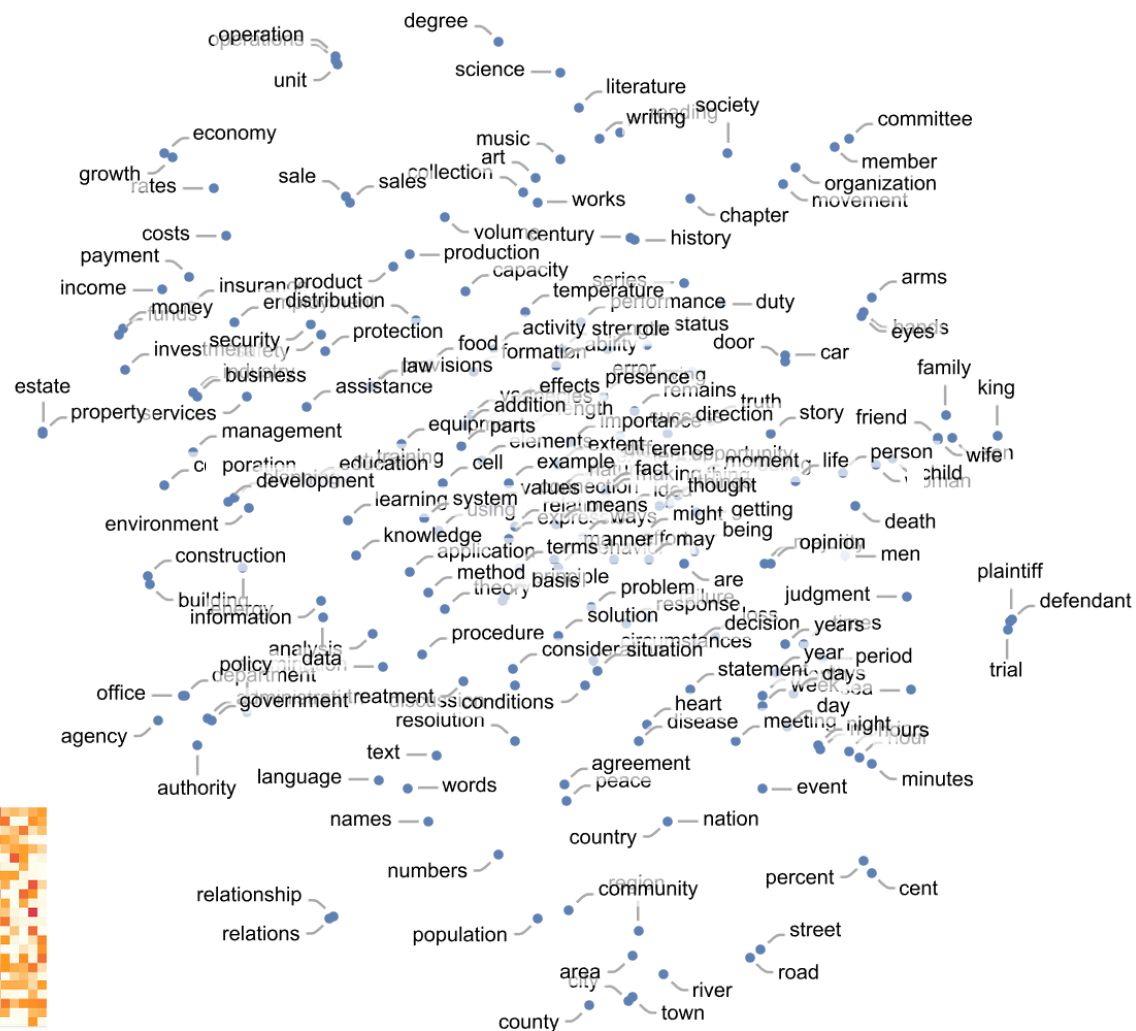
cat



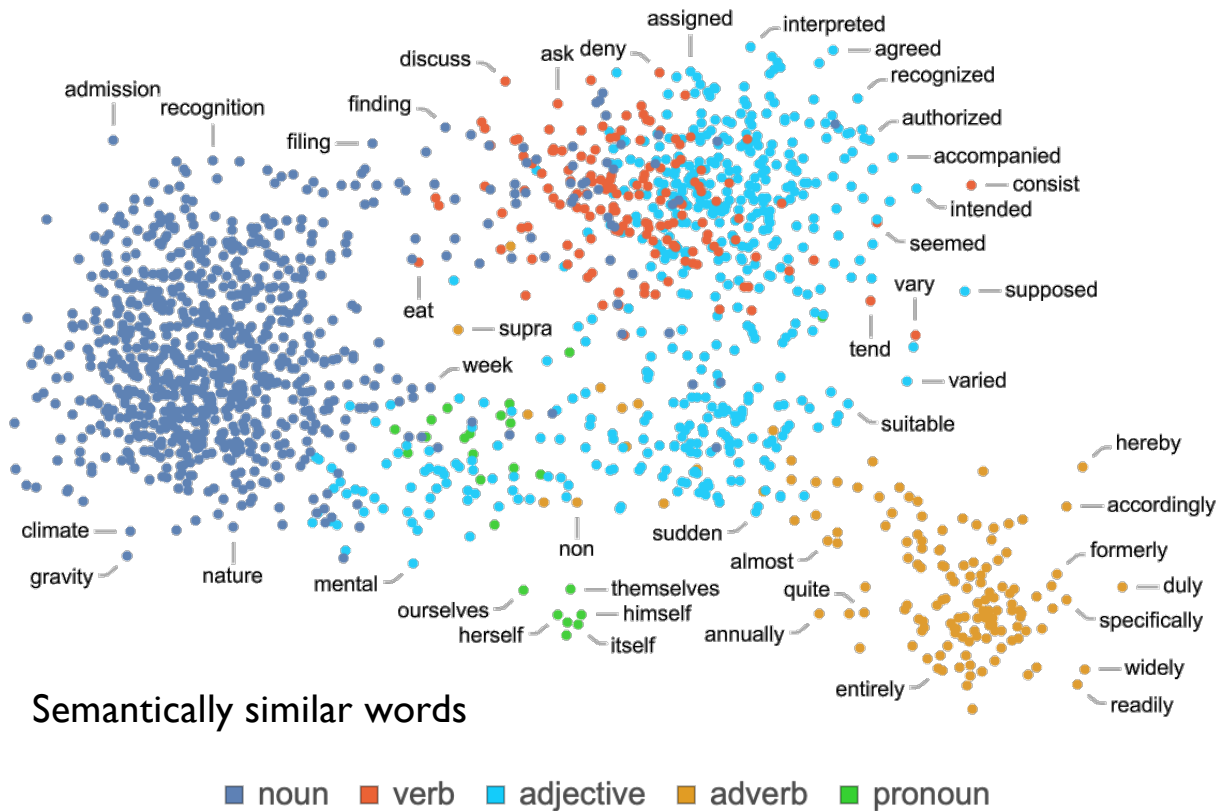
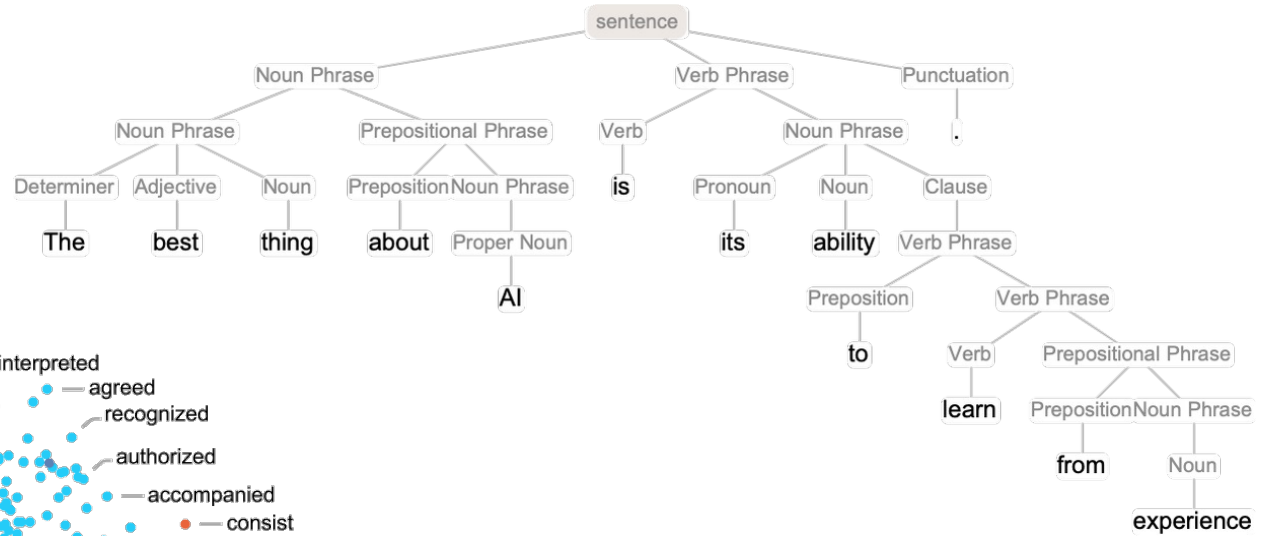
dog



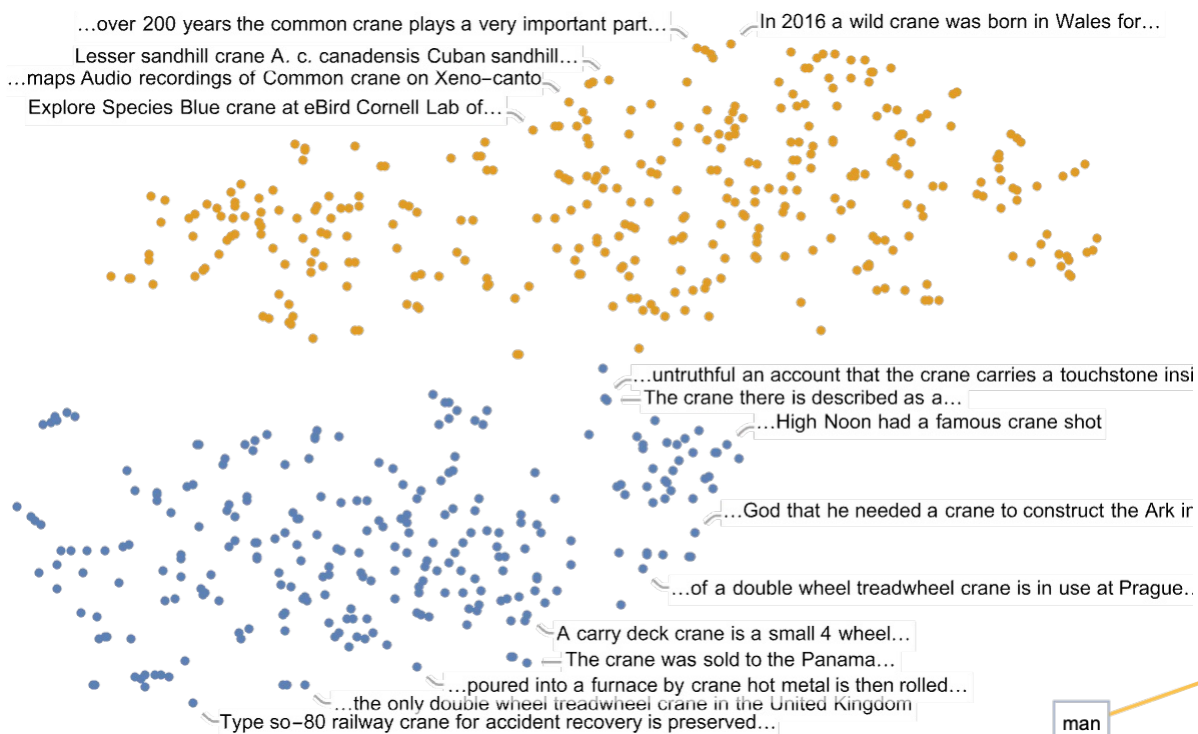
chair



HOW CHATGPT WORKS?



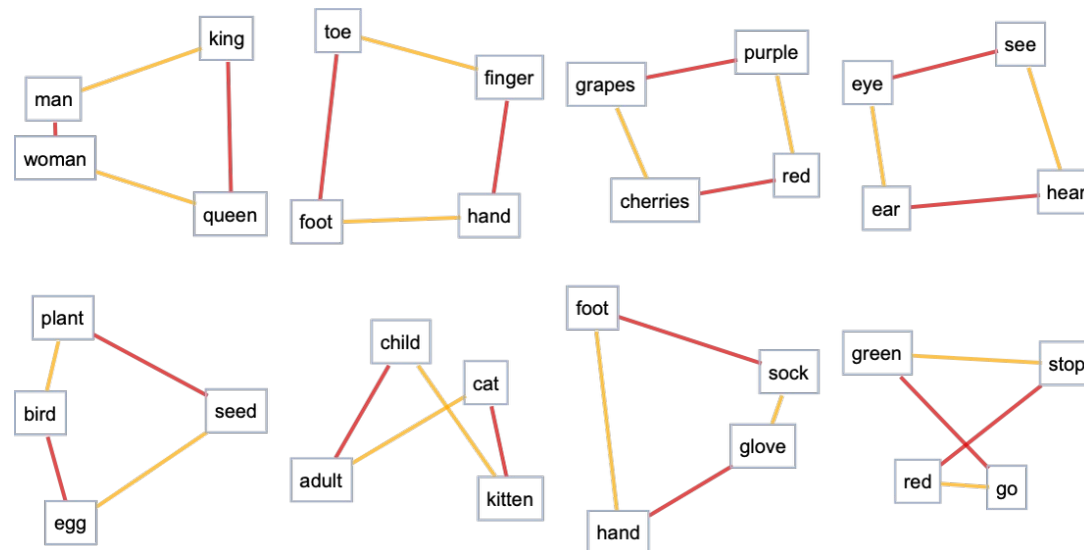
Modelling human text
embedding



1: we check for different meanings for a given word

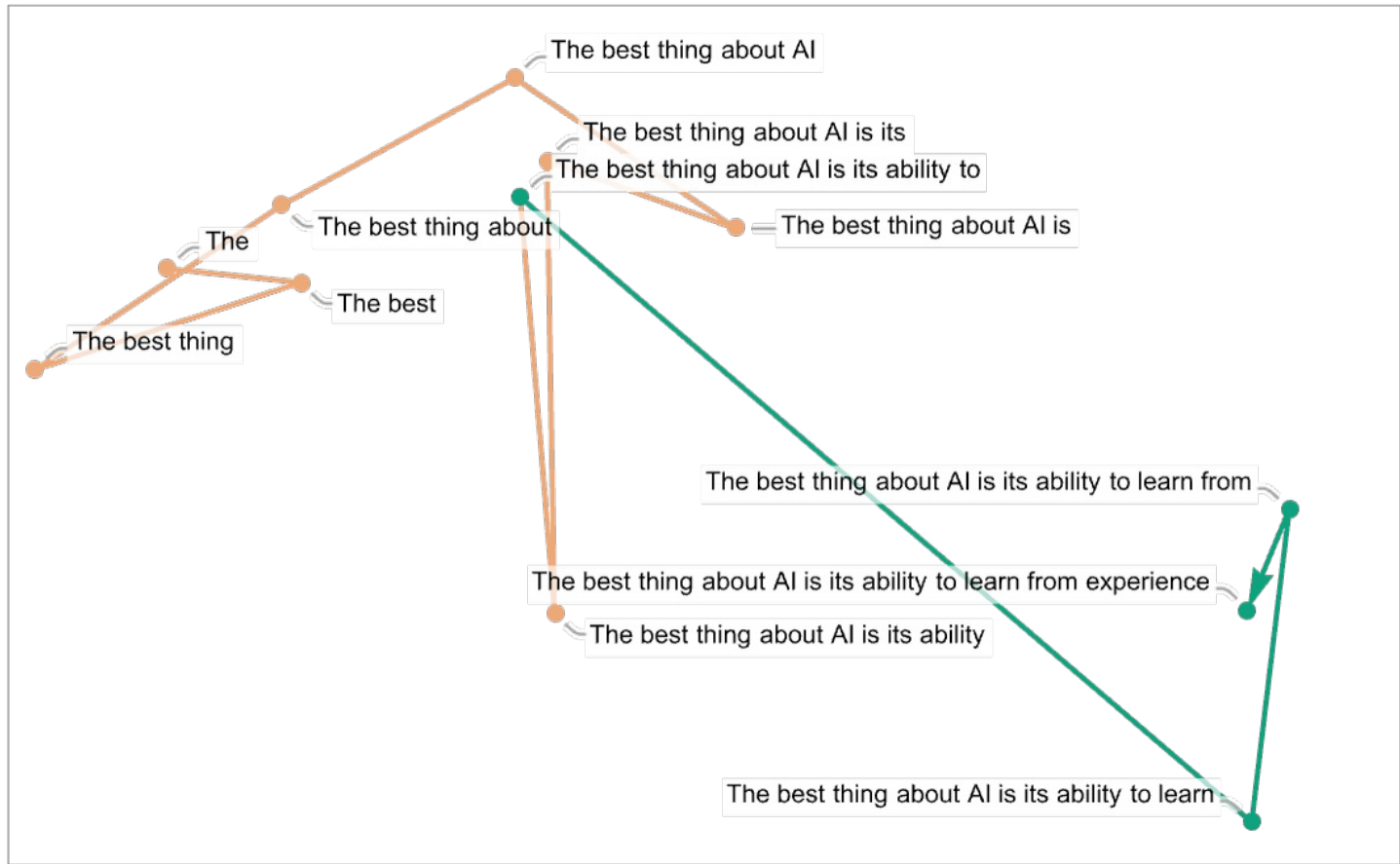
Example: "crane" (bird or machine?)

2: we build a feature space by placing "words nearby in meaning" close to each other

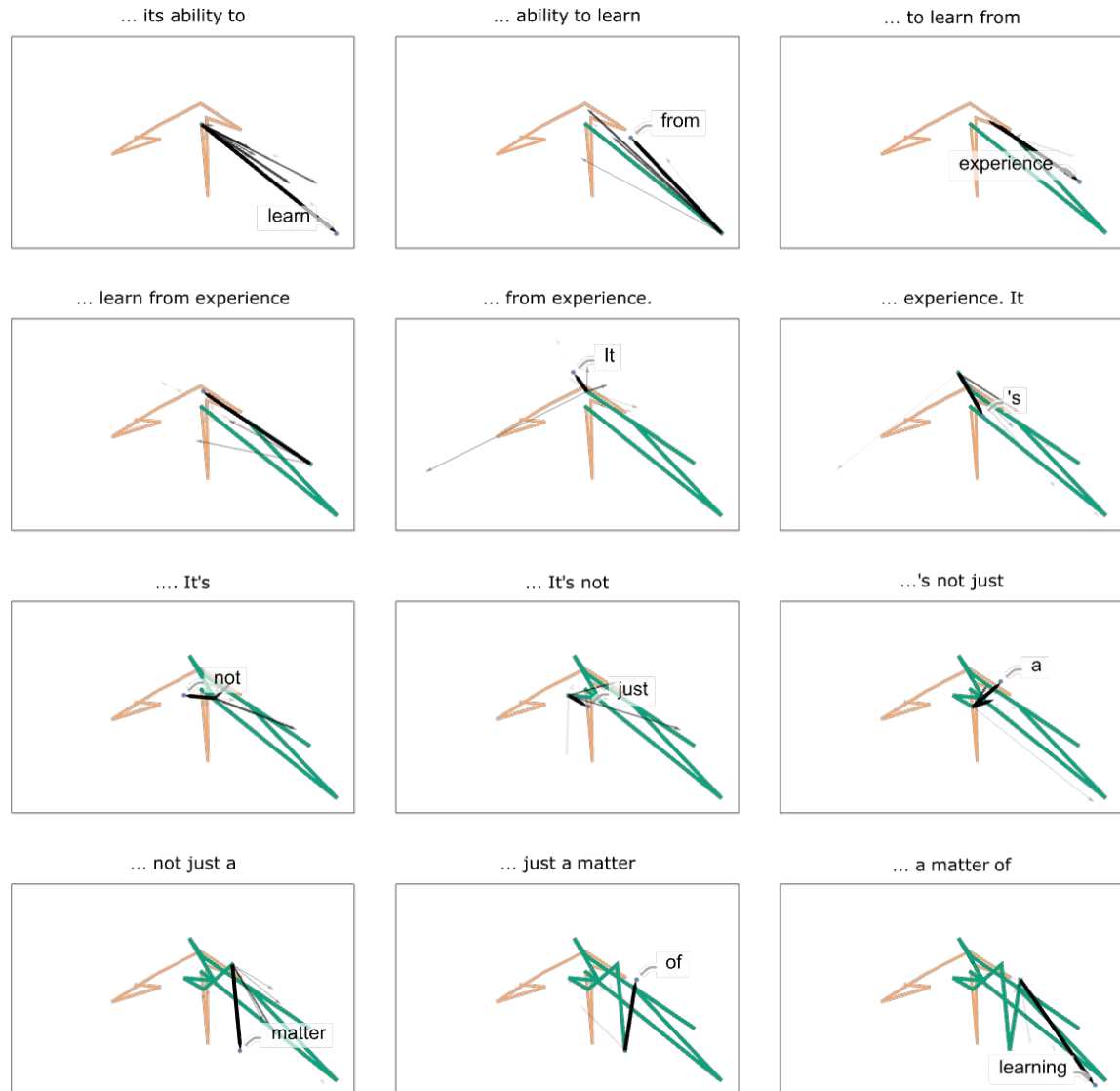


“Navigating” the feature space

HOW CHATGPT WORKS?



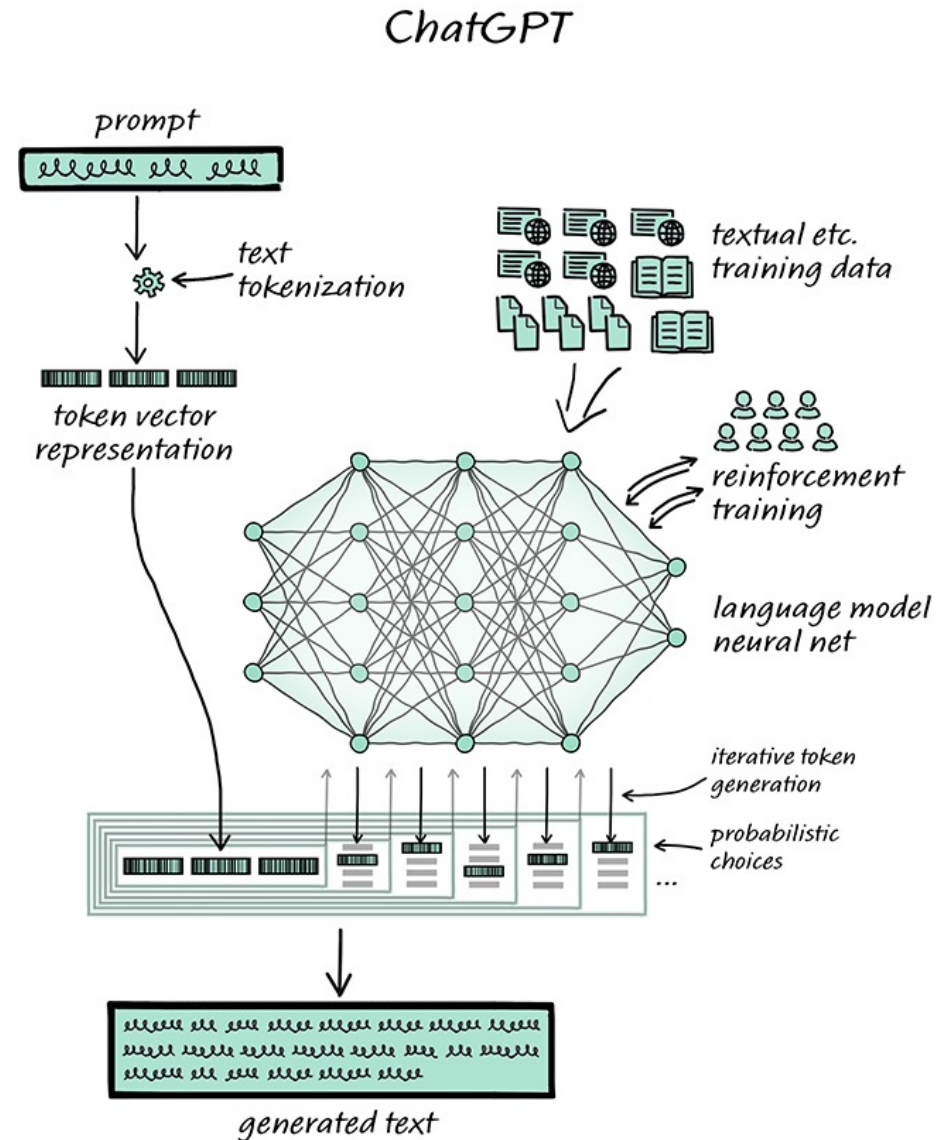
HOW CHATGPT WORKS?



“Navigating”
the feature space

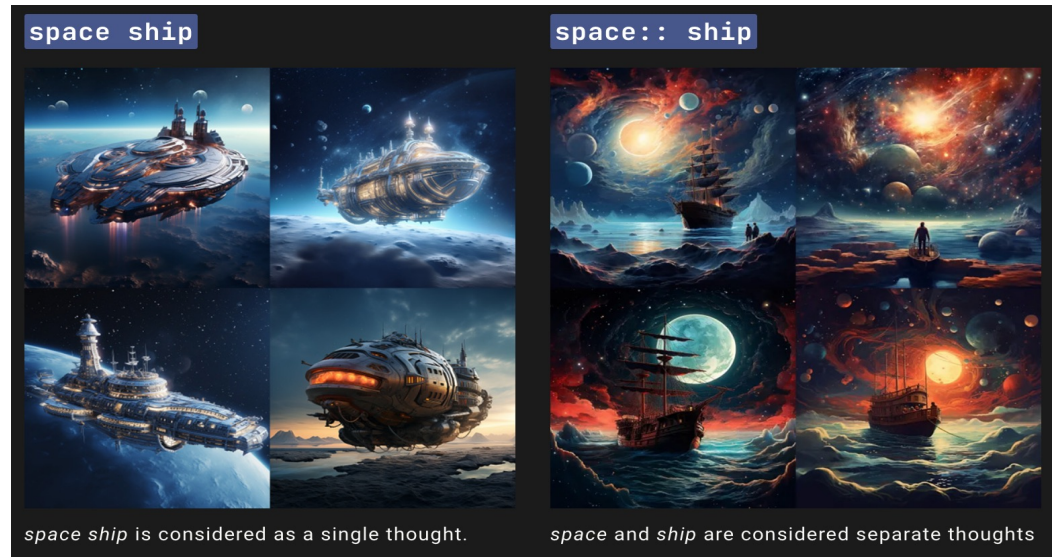
HOW CHATGPT WORKS?

- ✓ Input = prompt
- ✓ Text (token) representation = embedding
- ✓ Large language model
- ✓ Training data (books, Web pages, code etc)
- ✓ Reinforcement learning = user interaction
- ✓ Output = generated text



DO GENERATIVE AI TOOLS PERFORM WELL?

- ✓ it really depends on the model and the prompt structure
- ✓ Example: **text-to-image**



DO GENERATIVE AI TOOLS PERFORM WELL?



← → ↻ 🏠 textsynth.com/text_to_image.html

📧 Gmail 📁 LSE 📁 CIDACS 📁 UFBA 📁 UCL_IHI 📁 Tools 📁 Courses 📁 Books 📁 Journals 📁 Publications

[restrictions](#) associated with this model.

Translation

Text to Image

Chat

Number of images: Size: Timesteps: Guidance scale:

Select an example

A horse riding an astronaut

Generate

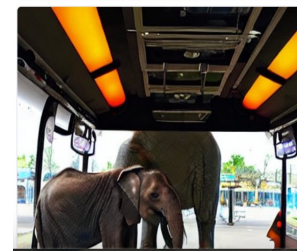
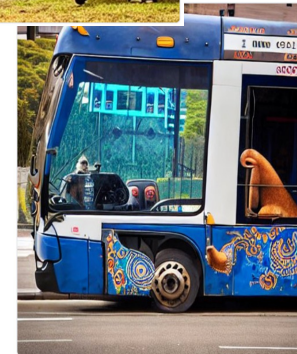
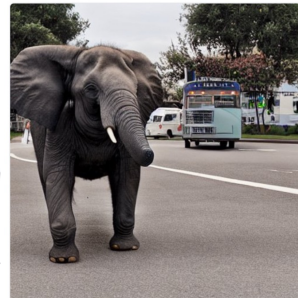
Stable Diffusion v1-5 Demo

Stable Diffusion v1-5 is the latest version of the state of the art text-to-image model.

For faster generation you can try [text to image tool](#) at Runway.

an elephant playing guitar within a double deck bus in london

Generate image



DO GENERATIVE AI TOOLS PERFORM WELL?

- ✓ it really depends on the model and the prompt structure
- ✓ Example: **image-to-image**
 - ✓ *Given the original image, generate variations*





GENIAL: GENerative AI Tools as a Catalyst for Learning

A Collaborative Focus Group

<https://lse-dsi.github.io/genial/>

EXAMPLE (DATABASES)

```

There are six tables in my database - department, dependent, dept_locations, employee, project, works_on. This is the code I have used to create these tables - CREATE TABLE "department" (
  "Dname" TEXT NOT NULL,
  "Dnumber" INTEGER NOT NULL,
  "Mgr_ssn" INTEGER,
  "Mgr_start_datge" TEXT,
  PRIMARY KEY("Dnumber")
); CREATE TABLE "dependent" (
  "Essn" INTEGER NOT NULL,
  "Dependent_name" TEXT NOT NULL,
  "Sex" TEXT,
  "Bdate" TEXT,
  "Relationship" TEXT,
  PRIMARY KEY("Essn","Dependent_name")
); CREATE TABLE "dept_locations" (
  "Dnumber" INTEGER,
  "Dlocation" TEXT,
  PRIMARY KEY("Dnumber","Dlocation")
); CREATE TABLE "employee" (
  "Fname" TEXT,
  "Lname" TEXT,
  "Job_id" INTEGER NOT NULL,
  "Hire_date" TEXT,
  "Salary" INTEGER,
  "Comm" INTEGER,
  PRIMARY KEY("Fname","Lname")
); CREATE TABLE "works_on" (
  "Essn" INTEGER NOT NULL,
  "Pno" INTEGER NOT NULL,
  "Hours" TEXT,
  PRIMARY KEY("Pno","Essn")
);

```

...

```

); CREATE TABLE "works_on" (
  "Essn" INTEGER NOT NULL,
  "Pno" INTEGER NOT NULL,
  "Hours" TEXT,
  PRIMARY KEY("Pno","Essn")
); How do I find the sum of the salaries of all employees, the maximum salary, the minimum salary, and the average salary.

```

← contextual prompt

→ task prompt

← task prompt

OUTPUT: AGGREGATION FUNCTIONS



To calculate the sum, maximum, minimum, and average salaries of all employees, you can use SQL aggregate functions. Here's the SQL query that you can use:

```

sql Copy code
SELECT
  SUM(Salary) AS TotalSalary,
  MAX(Salary) AS MaxSalary,
  MIN(Salary) AS MinSalary,
  AVG(Salary) AS AvgSalary
FROM
  employee;

```

This query selects the sum, maximum, minimum, and average of the "Salary" column from the "employee" table.

Simply execute this query in your database management tool or platform, and it will return the desired results.

OUTPUT: PATTERN MATCHING



); Retrieve all dependents born in the 70s (1970-1979)

To retrieve all dependents born in the 70s (1970-1979) from your database, you can use the following SQL query:

```

sql Copy code
SELECT Dependent_name, Bdate
FROM dependent
WHERE Bdate BETWEEN '1970-01-01' AND '1979-12-31';

```

This query selects the 'Dependent_name' and 'Bdate' columns from the 'dependent' table, where the birthdate ('Bdate') falls between January 1, 1970, and December 31, 1979. This will give you a list of dependents born in the 1970s along with their birthdates.

I am getting no output

I apologize for the confusion. It seems that the date format might not match the one stored in your database. If the date format is different, you'll need to adjust the query accordingly. If your date format is in the format 'YYYY-MM-DD', the query I provided should work.

✓ Example: **text-to-text**

EXAMPLE (DATABASES)

```
There are six tables in my database - department, dependent, dept_locations, employee, project, works_on. This is the code I have used to create these tables - CREATE TABLE "department" ( "Dname" TEXT NOT NULL, "Dnumber" INTEGER NOT NULL, "Mgr_ssn" INTEGER, "Mgr_start_date" TEXT, PRIMARY KEY("Dnumber") ); CREATE TABLE "dependent" ( "Essn" INTEGER NOT NULL, "Dependent_name" TEXT NOT NULL, "Sex" TEXT, "Bdate" TEXT, "Relationship" TEXT, PRIMARY KEY("Essn","Dependent_name") ); CREATE TABLE "dept_locations" ( "Dnumber" INTEGER, "Dlocation" TEXT, PRIMARY KEY("Dnumber","Dlocation") ); CREATE TABLE "employee" ( "Fname" TEXT,
```

...

```
); CREATE TABLE "works_on" ( "Essn" INTEGER NOT NULL, "Pno" INTEGER NOT NULL, "Hours" TEXT, PRIMARY KEY("Pno","Essn") ); How do I find the sum of the salaries of all employees, the maximum salary, the minimum salary, and the average salary.
```

← contextual prompt

task prompt ←

OUTPUT: JOINING TABLES



```
); Match each row from the DEPARTMENT table with every row from the PROJECT table based on inner join condition. Sure, based on the provided table structures, here's how you can perform an inner join between the 'DEPARTMENT' and 'PROJECT' tables: sql Copy code SELECT * FROM department INNER JOIN project ON department.Dnumber = project.Dnum; This query will match each row from the 'DEPARTMENT' table with every row from the 'PROJECT' table based on the condition that the 'Dnumber' column from the 'DEPARTMENT' table matches the 'Dnum' column from the 'PROJECT' table.
```

EXAMPLE (DATABASES)

```
The two different code you have provided are giving me different answers - SELECT
D.Dnumber AS DepartmentNumber,
COUNT(E.Ssn) AS NumberOfEmployees
FROM
department D
JOIN
employee E ON D.Dnumber = E.Dno
GROUP BY
D.Dnumber
HAVING
COUNT(E.Ssn) > 5
AND SUM(CASE WHEN E.Salary > 40000 THEN 1 ELSE 0 END) > 0; and WITH
DepartmentEmployeeStats AS (
SELECT
D.Dnumber AS DepartmentNumber,
COUNT(E.Ssn) AS TotalEmployees,
SUM(CASE WHEN E.Salary > 40000 THEN 1 ELSE 0 END) AS HighSalaryEmployees
FROM
department D
JOIN
employee E ON D.Dnumber = E.Dno
GROUP BY
D.Dnumber
HAVING
COUNT(E.Ssn) > 5
AND SUM(CASE WHEN E.Salary > 40000 THEN 1 ELSE 0 END) > 0;
)
```

←
contextual
prompt



OUTPUT: EXPLANATION



```
I apologize for the confusion. The two queries are conceptually similar, but there might be differences in the interpretation of the requirements. If they are giving you different answers, let's focus on the query using the common table expression (WITH clause) since it's more organized. Here's that query again for clarity:
```

```
sql Copy code
WITH DepartmentEmployeeStats AS (
SELECT
D.Dnumber AS DepartmentNumber,
COUNT(E.Ssn) AS TotalEmployees,
SUM(CASE WHEN E.Salary > 40000 THEN 1 ELSE 0 END) AS HighSalaryEmployees
FROM
department D
JOIN
employee E ON D.Dnumber = E.Dno
GROUP BY
D.Dnumber
HAVING
COUNT(E.Ssn) > 5
AND SUM(CASE WHEN E.Salary > 40000 THEN 1 ELSE 0 END) > 0;
)
SELECT
DepartmentNumber,
HighSalaryEmployees AS NumberOfEmployeesExceeding40000
FROM
```

Please ensure that your data and requirements are consistent with this query. If you are getting unexpected results, there might be issues with the data or the interpretation of the requirements. Double-check your data and the logic you want to implement in your query to make sure they align correctly. If you're still facing issues, please provide more context or specific details about your data and requirements, and I'll be happy to assist you further.



GENIAL: GENerative AI Tools as a Catalyst for Learning

A Collaborative Focus Group

<https://lse-dsi.github.io/genial/>

EXAMPLE: GITHUB COPILOT

LSE100_Copilot_demo.ipynb U

LSE100_Copilot_demo.ipynb > M Step 1: Read the Data > # Let's look at the data for London

+ Code + Markdown ▶ Run All ↺ Restart ☰ Clear All Outputs [33] Variables ☰ Outline ...

```
# Let's look at the data for London but focus on 2023

london_data_2023 = london_data[london_data['Date'] >= '2023-01-01']
london_data_2023.head()
```

[14] ✓ 0.0s

	Date	RegionName	AveragePrice
67170	2023-01-01	London	535789.0
67171	2023-02-01	London	529715.0
67172	2023-03-01	London	521321.0
67173	2023-04-01	London	531282.0
67174	2023-05-01	London	525327.0

```
# Plot the average price for London since 1965

import matplotlib.pyplot as plt

plt.plot(london_data['Date'], london_data['AveragePrice'])
plt.show()
```

Python

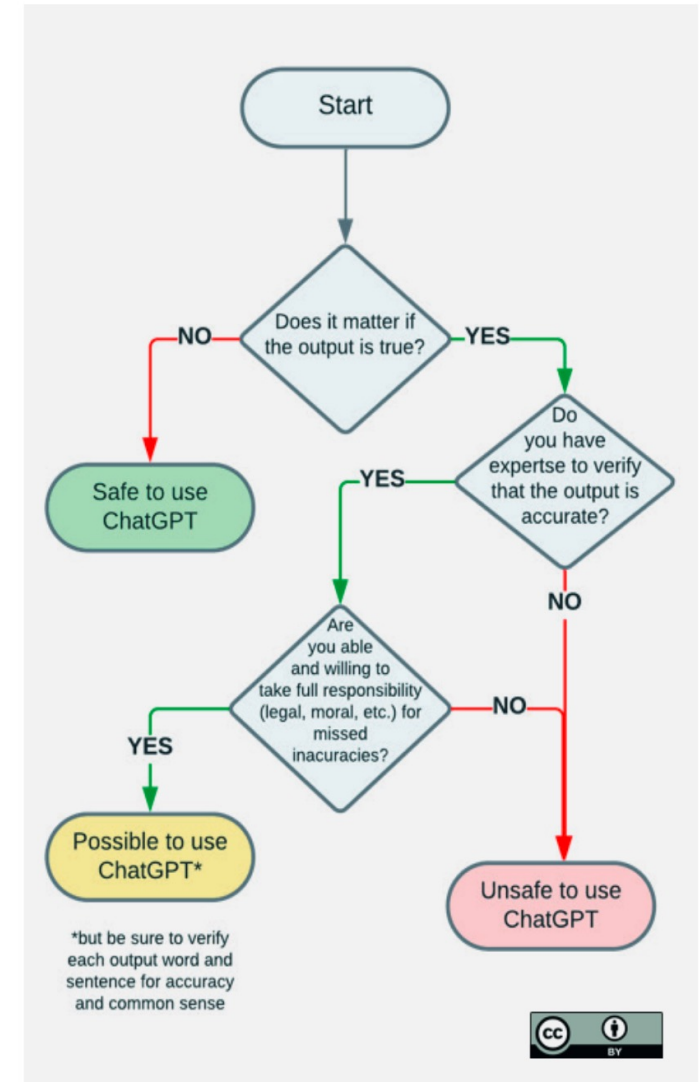
Image from LSE Moodle course:
Generative AI – Developing your AI literacy

SHOULD WE USE GENERATIVE AITOOOLS?

- ✓ **Embrace**
 - ✓ “just another tool”, such as Google, spell-checkers etc
 - ✓ great learning opportunity
 - ✓ boost productivity
 - ✓ prepare for the job market

- ✓ **Resist**
 - ✓ Not ethical
 - ✓ Not reliable
 - ✓ Not transparent
 - ✓ Not fair
 - ✓ Overreliance x essential skills (e.g., coding)

Figure 1: When is it safe to use ChatGPT?⁵



DATA SCIENCE AND AI

Dr Marcos Barreto

Department of Statistics / LSE

w: [marcosebarreto.github.io](https://github.com/marcosebarreto)

e: m.e.barreto@lse.ac.uk

THANK YOU!

