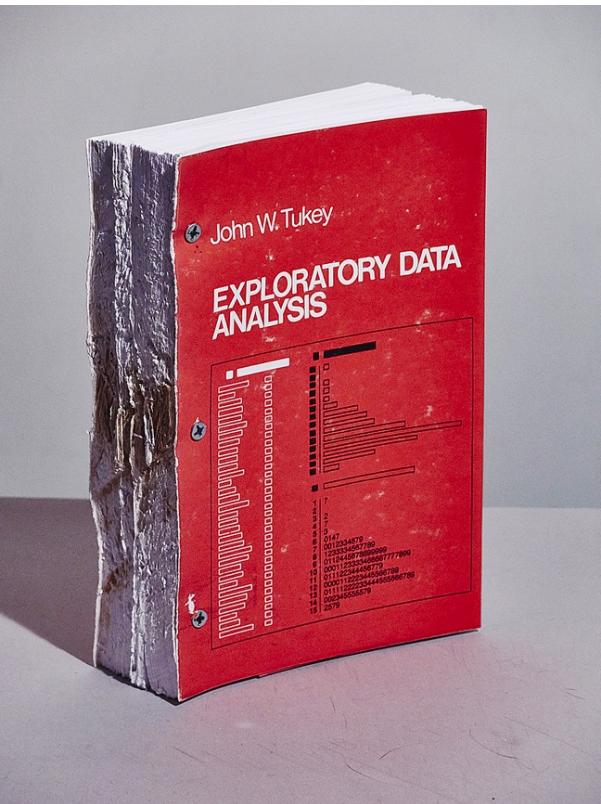




Your options in Data Science

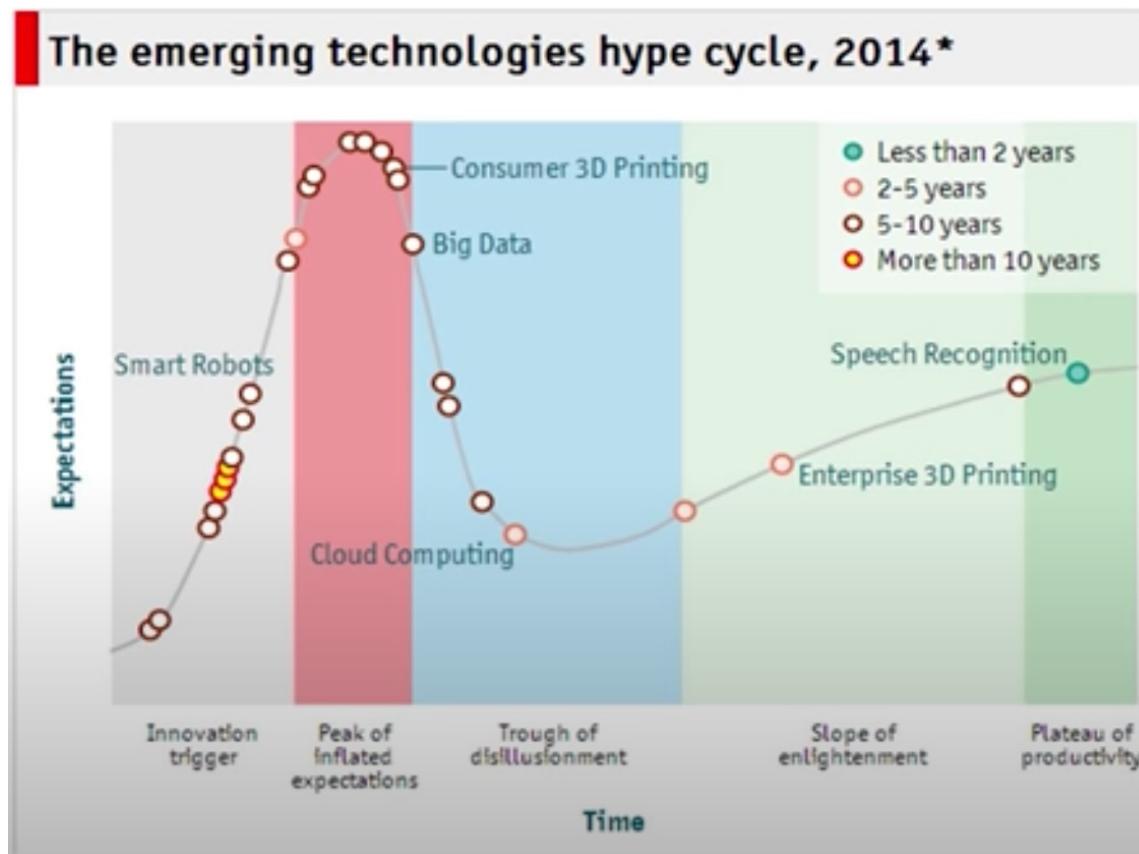
MAXIMOVSAKAYA
ANASTASIA

A little about history



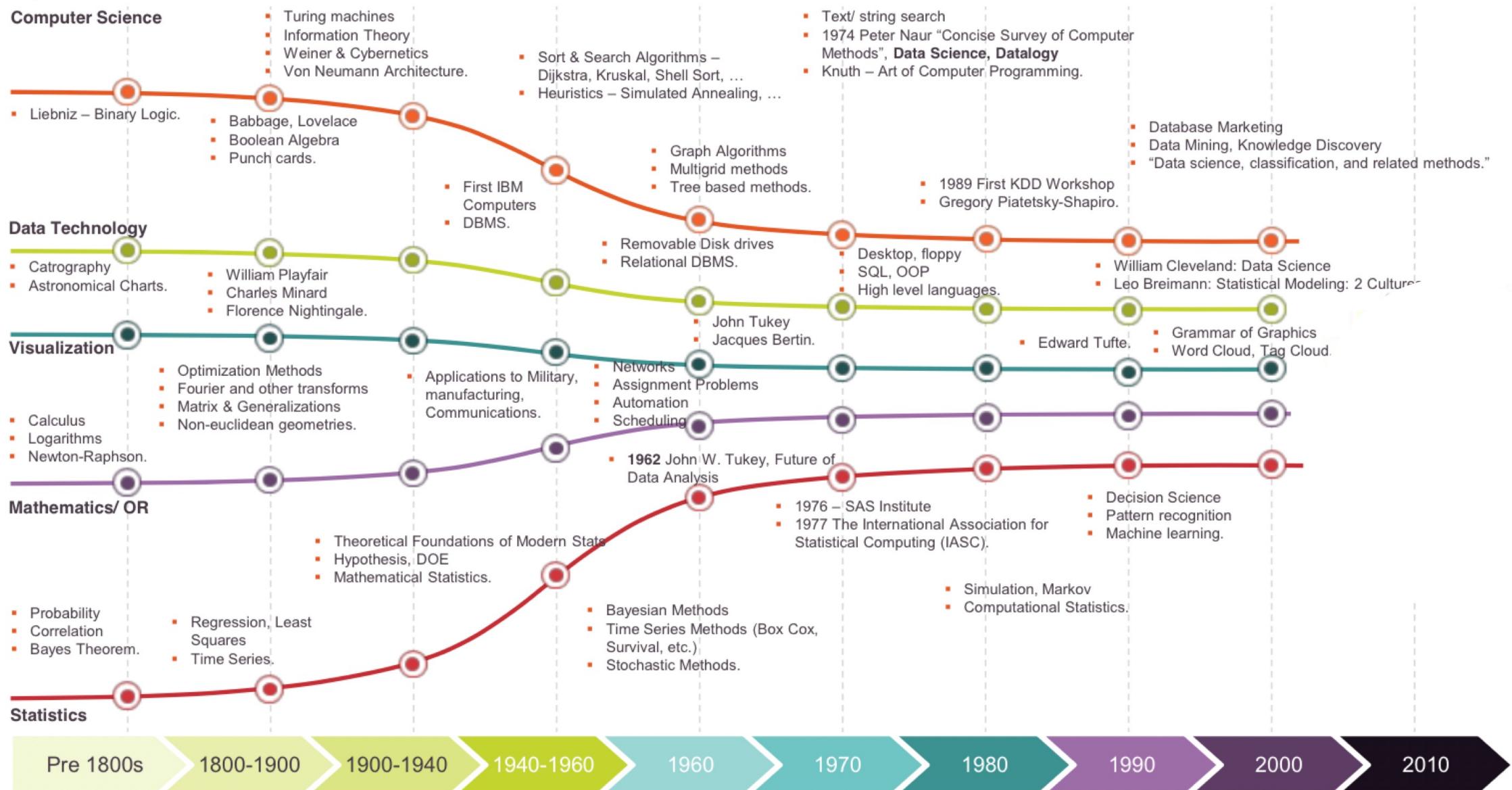
In the 1960s and 1970s, the American mathematician John W. Tukey published a series of papers where he suggested making hypotheses based on data.

Gartner Hype Cycle 2014

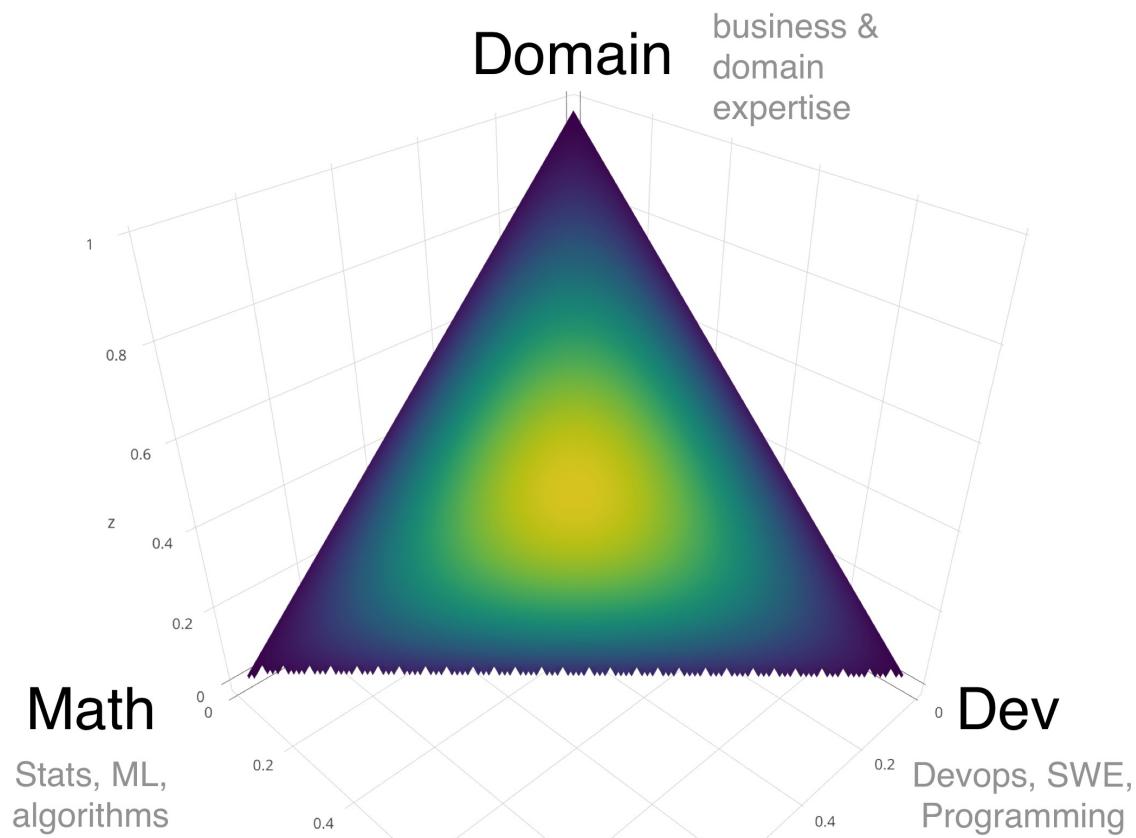


In 2015, Gartner excluded the Big Data area. Why?

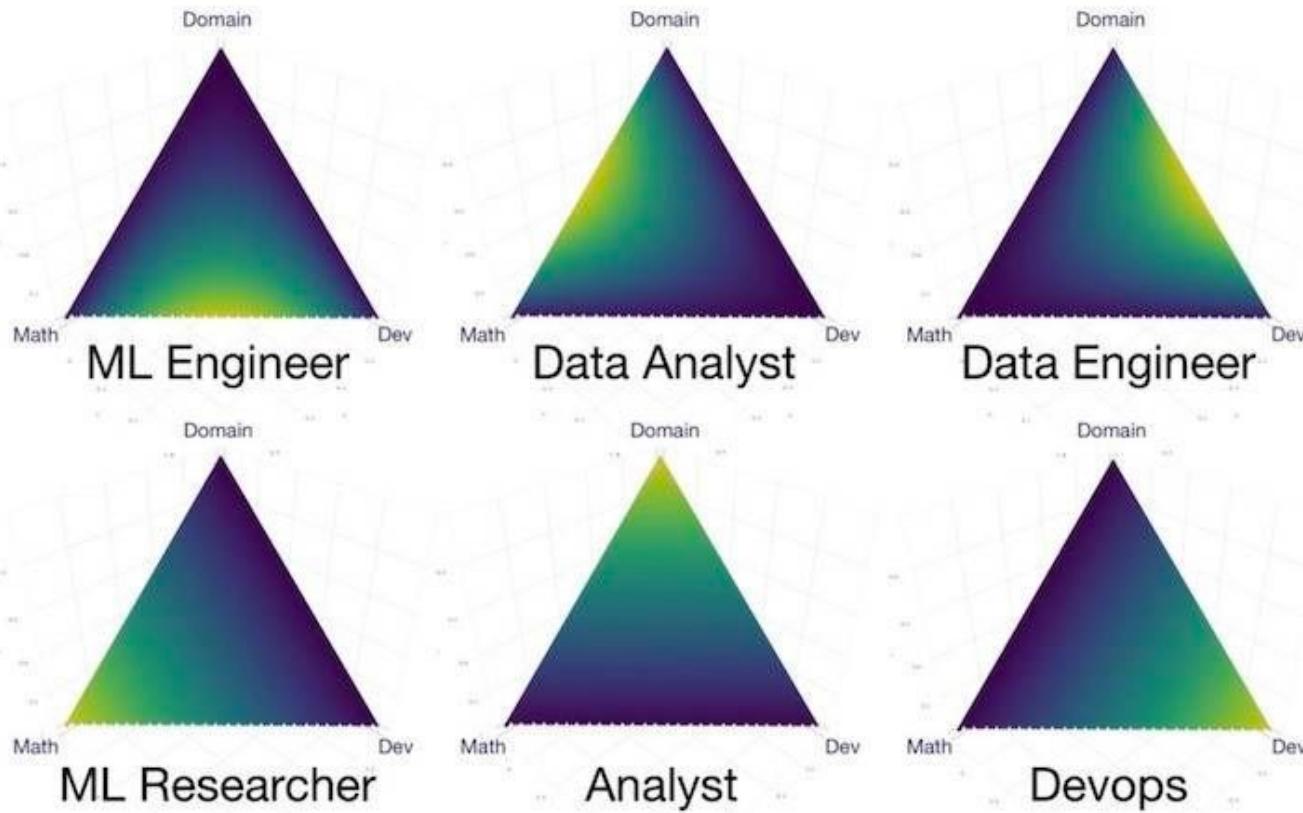
Do data scientists really
exist?



Career tracks in DS

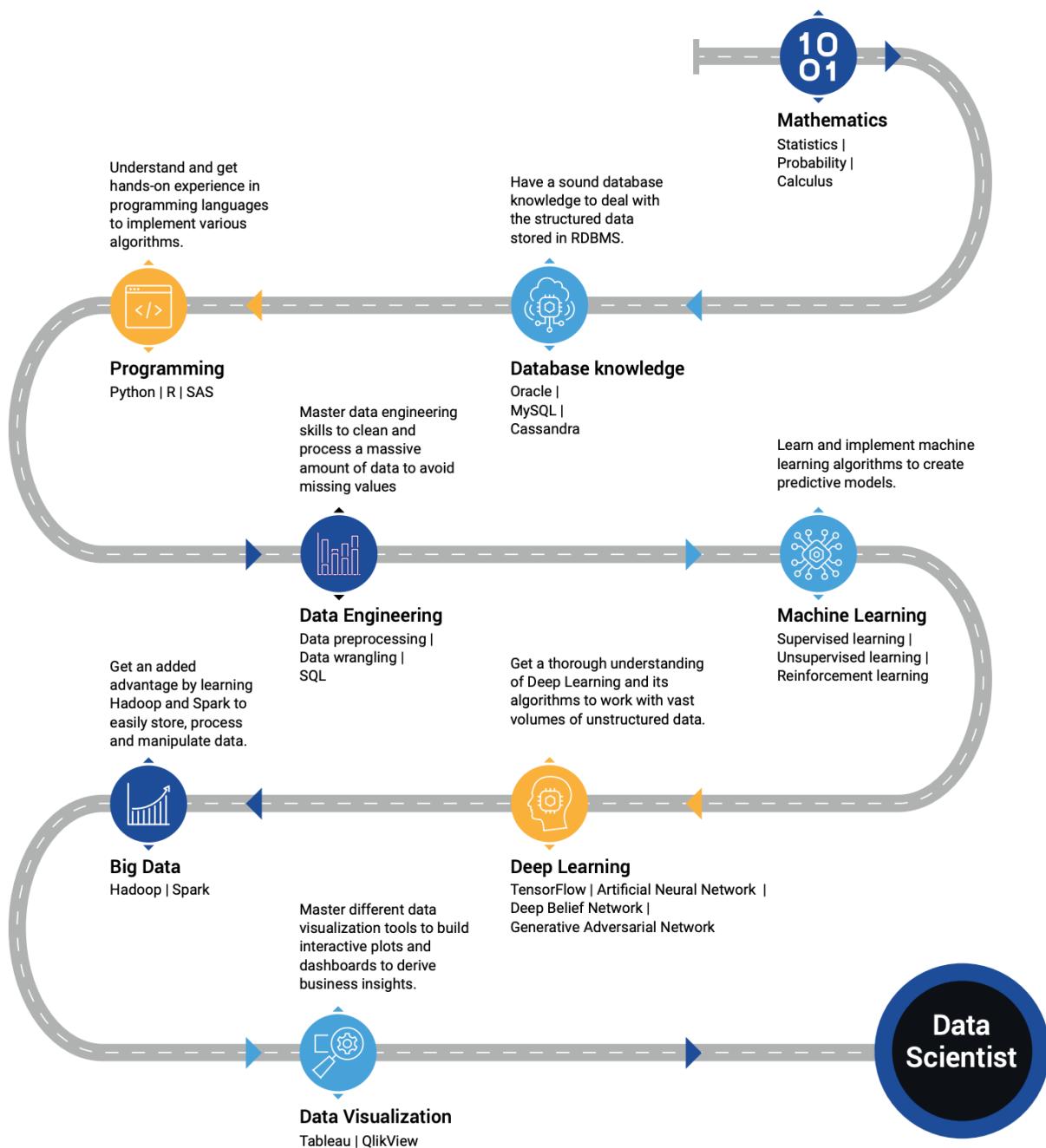


Career tracks in DS



Top skills

1. Data mining
2. Data visualization
3. Python
4. Probability Theory and Statistics
5. Machine learning
6. Deep learning
7. SQL
8. Docker and containerization
9. Cloud services
10. Databases



Terminology and interviews

ML system design intro

1. Problem statement
2. KPI and limitations
3. Baseline (heuristics or a simple model)
4. Defining the task
5. Data
6. Modelling
7. Metrics and validation
8. Problems
9. Deployment

Case study

- The client is a restaurant chain that wants to open a new spot
- Several accommodation options
- You need to choose the one that will bring the most profit

Case study

- x - object, what we are making a prediction for (restaurant location)
- \mathbb{X} - object space (all possible restaurant locations)
- y - target variable, what we predict (restaurant profit)
- \mathbb{Y} - decision space, all response values (all real numbers)

Training sample

- We collect many objects with a known value of the target variable
- Existing restaurants, their indicative description and profit

Features

- An object is an abstract entity, and computers work with numbers
- Feature - the numerical characteristic of the object

Types of features:

- Numeric
- Binary (0/1)
- Categorical
- Features with a complex internal structure

Features

- An object is an abstract entity, and computers work with numbers.
 - Features - the numerical characteristic of the object
-
- What features should we collect for this task?

Possible answers

- Location
- Number of cars passing by per day
- Distance to the nearest competitor
- Average cost per square meter of housing nearby
- The average age of residents of the nearest quarters
- and so on

How to encode location data

- We have coordinates - longitude and latitude
- It is not very likely that there is a relationship between these two numbers and the target variable - profit
- What should we do?

How to encode location data

- We have coordinates - longitude and latitude
 - We generate features based on this data (feature generation)
-
- What examples of features can you think of?

How to encode location data

- We have coordinates - longitude and latitude
- We generate features based on this data (feature generation)

Some examples:

- distance from the city center, from the nearest metro station
- if restaurants in different countries - time zones
- and so on

Modelling

- $a(x)$ - algorithm, model (a function that predicts the answer for any object)
- Maps \mathbb{X} to \mathbb{Y} (object space to decision space)

Metrics

- After the previous steps, we have a model that produces profit predictions for restaurants.
- Now it is necessary to evaluate the quality of the obtained predictions
- For this, metrics are used - measures of the quality of the algorithm on the sample
- Selected based on business requirements

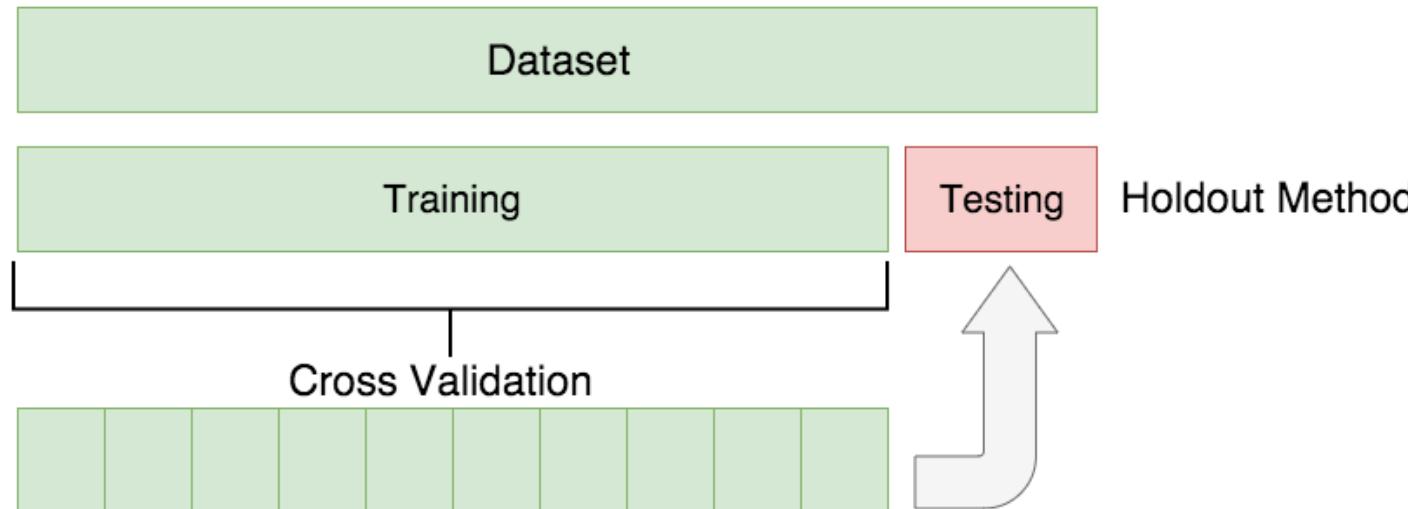
Example of a metric for regression

Mean Squared Error:

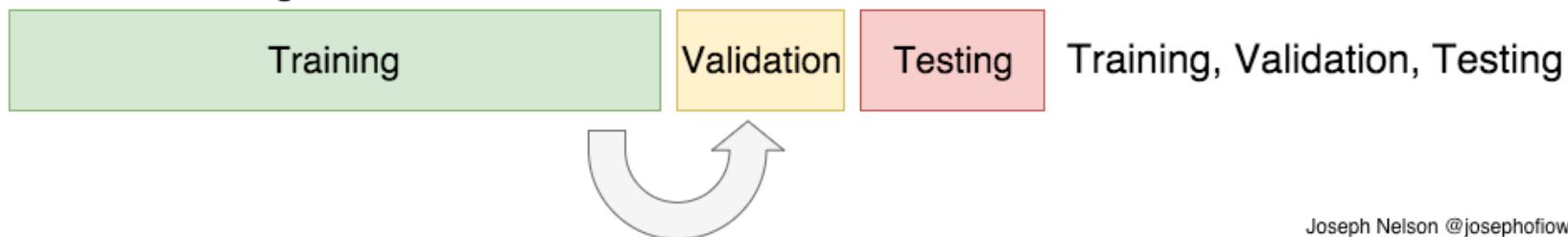
$$\text{MSE}(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

- The less the better

Validation



Data Permitting:

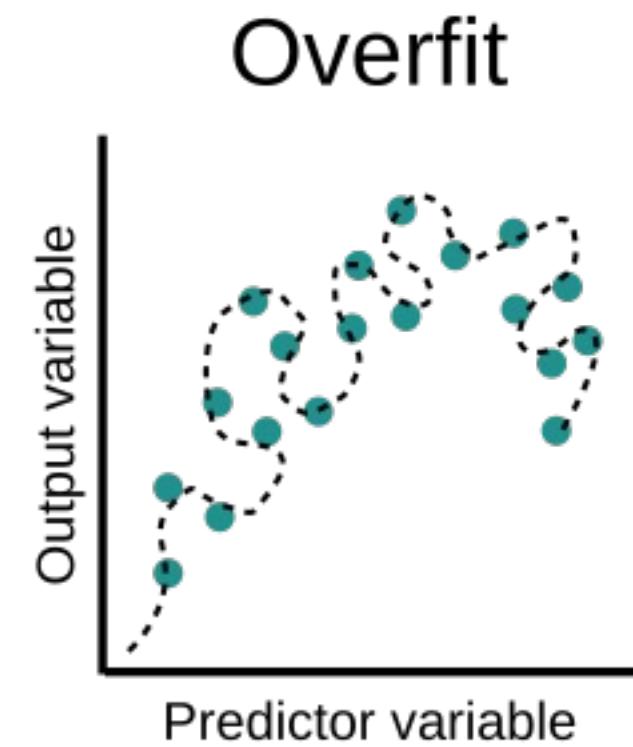
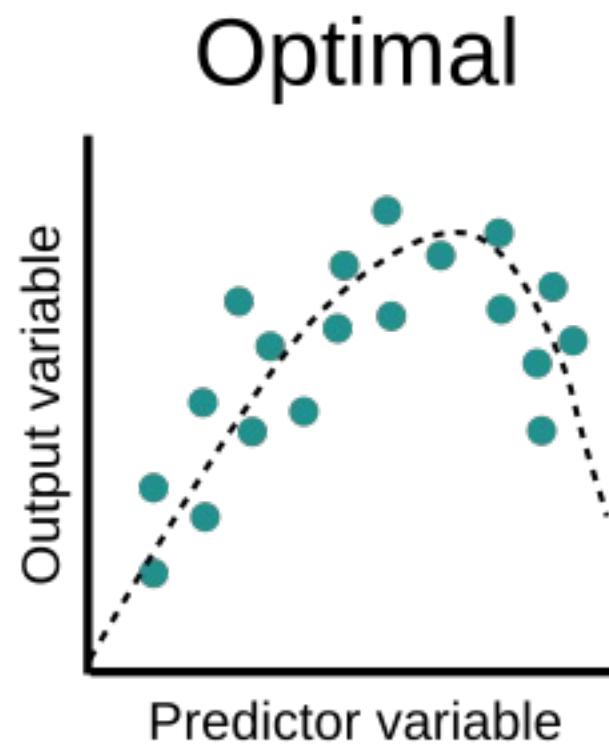
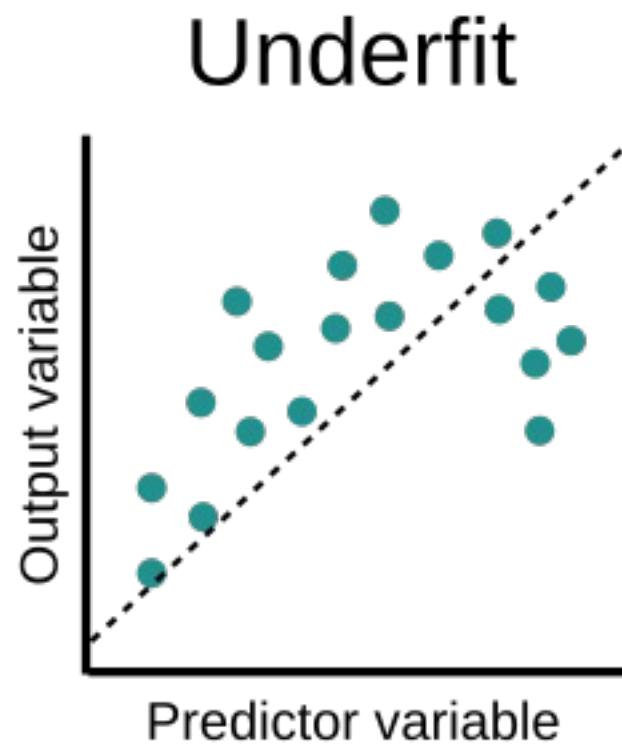


Joseph Nelson @josephofiowa

Training the model

- Training is a search for the optimal algorithm in terms of quality
- Having a training sample and a quality functional (metric), we select an appropriate algorithm from a certain family of algorithms
- Important criteria: quality that meets business requirements, sustainability of results, in some cases - interpretability

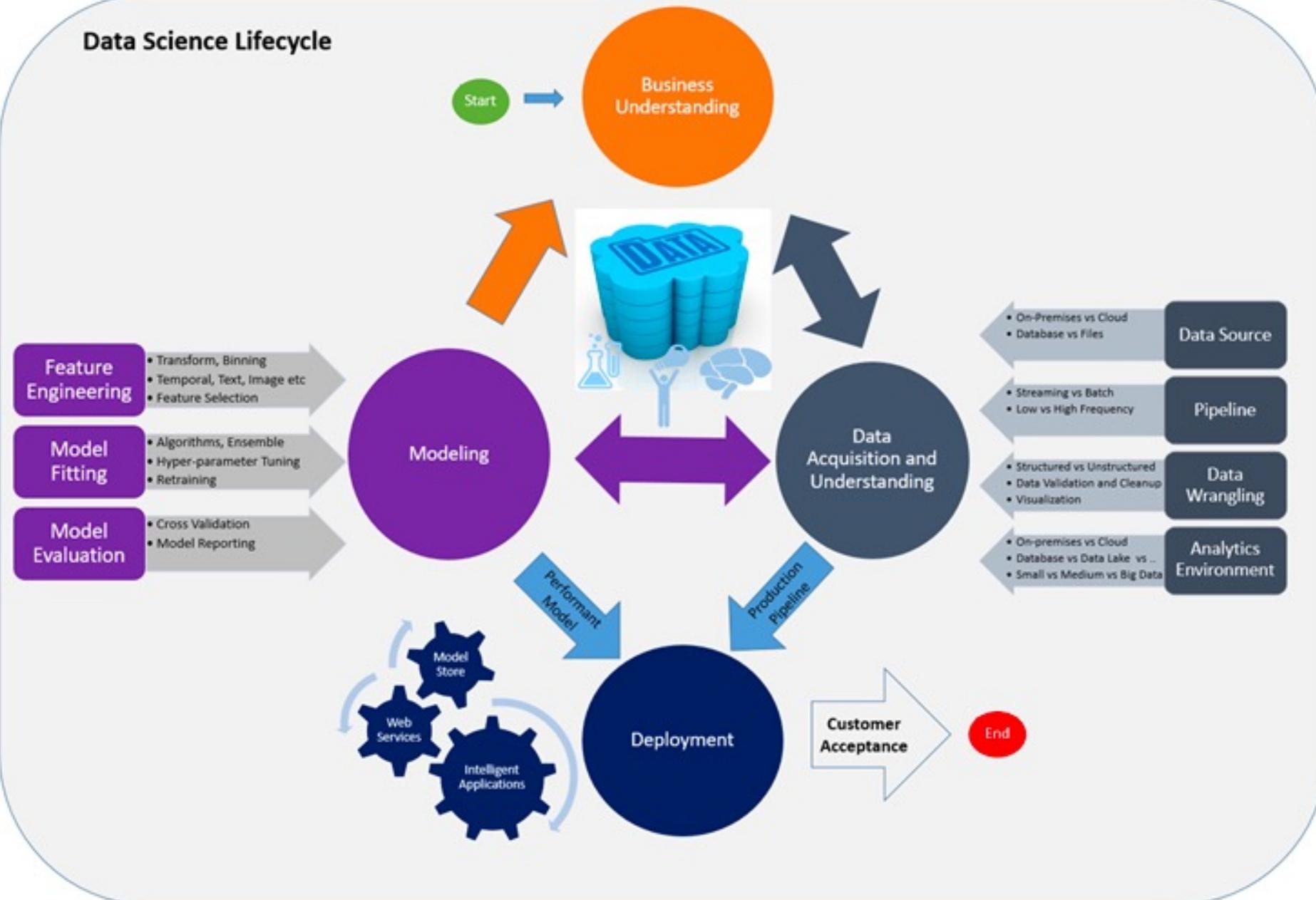
Overfitting and underfitting

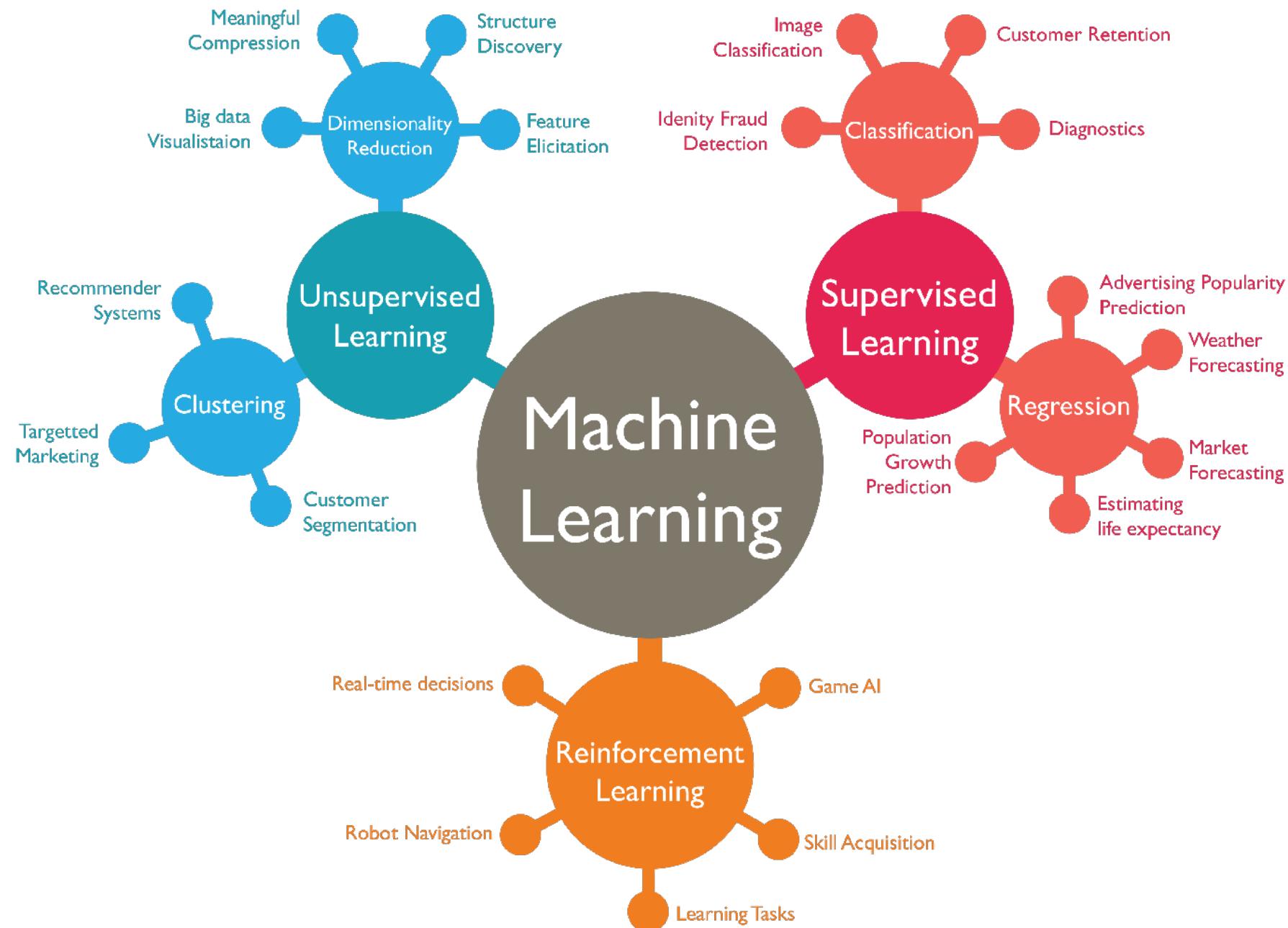


Data science project lifecycle

1. Definition of the task and target metric
2. Data collection
3. Data preprocessing and EDA (exploratory data analysis)
4. Generation of additional features
5. Selection and validation of models
6. As a rule, return to point 2 or 3

Data Science Lifecycle





Regression

- Regression is a class of supervised learning problems, when it is necessary to predict the target variable for a certain set of object features.
- The task of regression is to find dependencies between defining variables and the defined variable if it is a continuous number. For example, to determine the cost of a house by its area.
- The target value is any real number.
- The task of linear regression is to find such a relationship if it is linear.

Classification

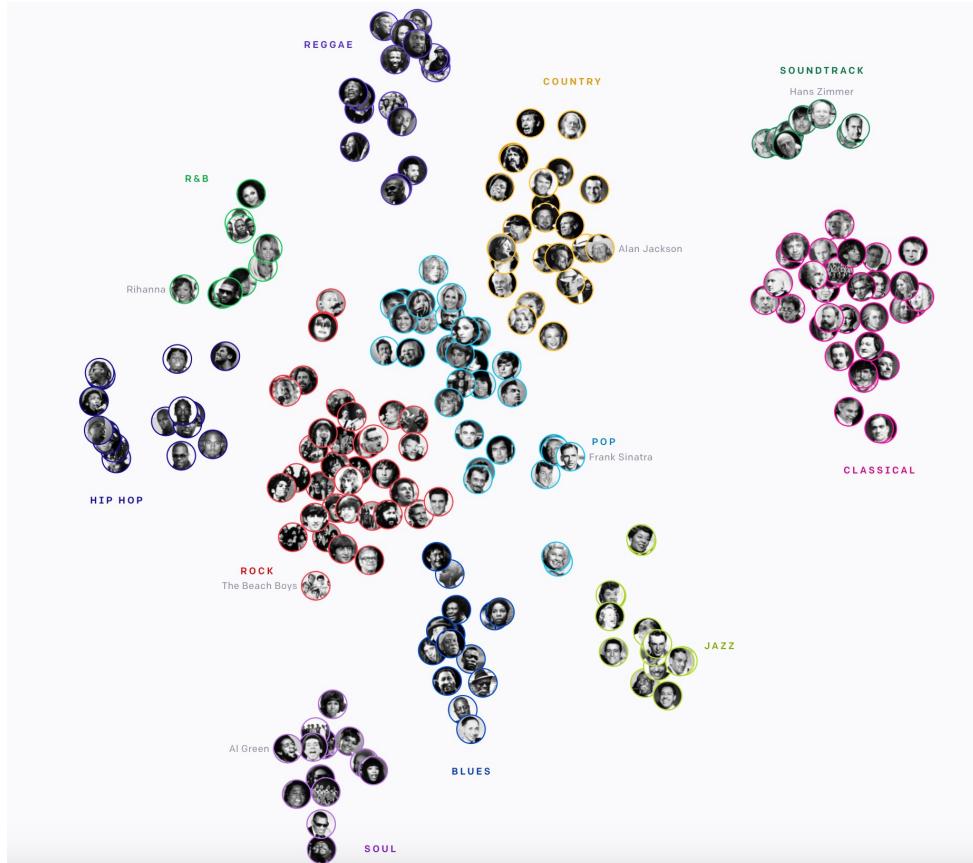
- Classification is a class of supervised learning tasks, when, according to a certain set of features of an object, it is necessary to predict its belonging to a certain class.
- The task of classification is to find dependencies between defining variables and the defined variable, if it belongs to a fixed set of values. For example, to determine whether a person will return a loan or not (binary classification) or to determine the scientific field of an article (multi-class classification).
- The target value is a class (one from a fixed set).
- There is also a multiclass classification with overlapping classes.

Clustering

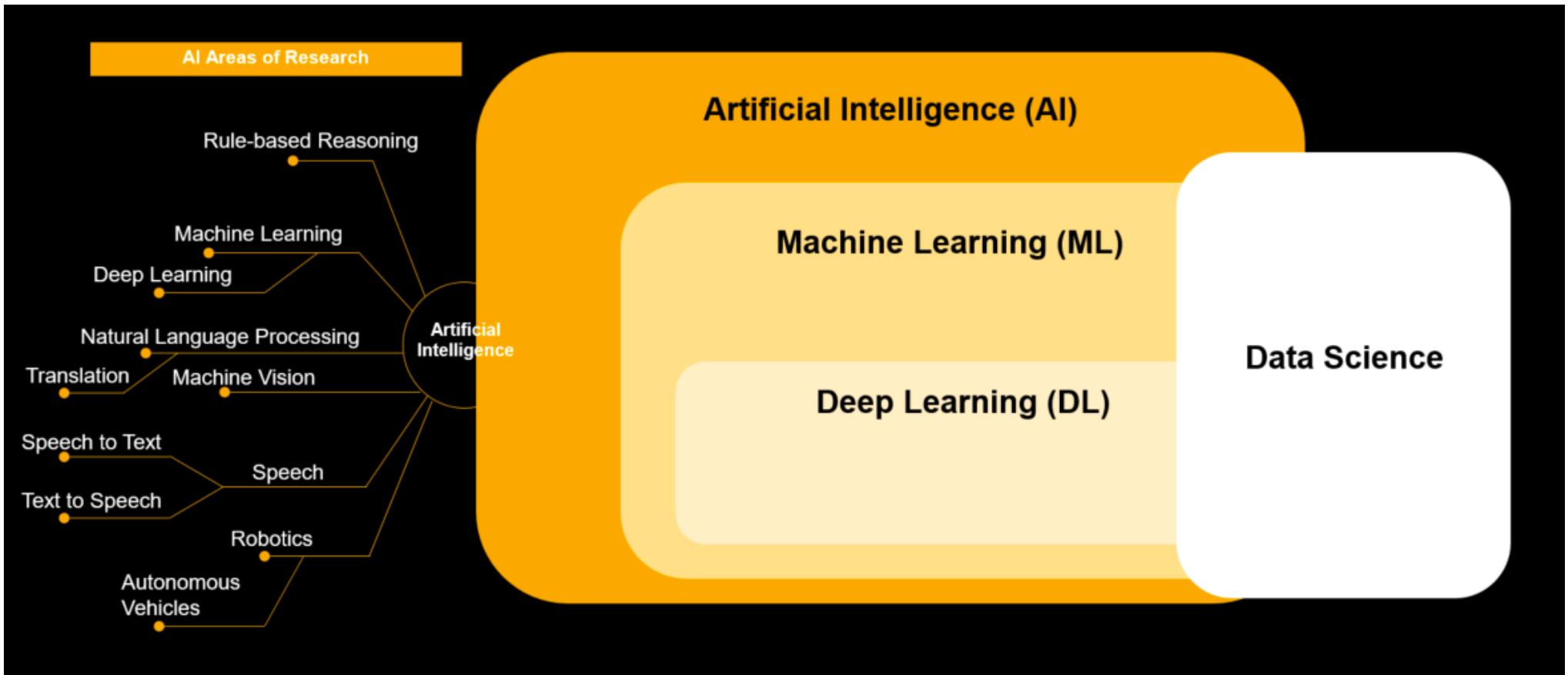
- Clustering is a type of unsupervised learning tasks, objects are divided into groups that have certain properties. Within the same cluster, objects must be similar
- Similarity is determined by a measure of distance
- Example: Euclidean Distance - Geometric Distance in Multidimensional Space

$$\rho(x, x') = \sqrt{\sum_i^N (x_i - x'_i)^2}$$

Example



<https://openai.com/blog/jukebox/>



Specializations

Recommender systems

- We predict what will be interesting to the user based on information about his profile
- Examples: Netflix, Amazon, Youtube

NLP - Natural Language Processing

➤ Working with natural language

NLP + ranking

Google search results for "how does netflix recommendation system work":

Search bar: how does netflix recommendation system work

Results count: Результатов: примерно 8 460 000 (0,47 сек.)

Thumbnail: A diagram showing a 3D grid of red cubes representing possible rows, with arrows pointing to a TV screen displaying the Netflix interface.

Text snippet:

The **recommendation system works** putting together data collected from different places. ... Every time you press play and spend some time watching a TV show or a movie, **Netflix** is collecting data that informs the **algorithm** and refreshes it. The more you watch the more up to date the **algorithm** is. 2 авр. 2018 г.

Source: uxplanet.org > netflix-binging-on-the-algorithm-a3a74... ▾

Link: [Netflix: Binging on the Algorithm | by Josefina Blattmann | UX ...](#)

Related queries (Похожие запросы):

- How do I fix my Netflix recommendations?
- What information system does Netflix use?
- How does Netflix know what I want to watch?

Bottom links: help.netflix.com > node ▾ Перевести эту страницу [How Netflix's Recommendation System Works](#)

Яндекс search results for "how does netflix recommendation system work":

Search bar: how does netflix recommendation system work

Results count: Нашлось 19 млн результатов

Result 1: **How Netflix's Recommendations System Works**

Source: help.netflix.com > en/node/100639 ▾

The **recommendations system** does not include demographic information (such as age or gender) as part of the decision making process. ... Below is a description of **how the system works** over time, and **how** these pieces of information influence what we present... Читать ещё >

Result 2: **How does the Netflix movie recommendation system work?**

Source: quora.com > How...Netflix...recommendation-system-work- ▾

More than 80 per cent of the TV shows people watch on **Netflix** are discovered through the platform's **recommendation system**. That means the majority of what you decide to watch on **Netflix** is the result of decisions made by a mysterious, black box of an algorithm. Intrigued? Here's **how it works**. **Netflix** uses machine learning and... Читать ещё >

Result 3: **How Netflix's Recommendation Engine Works? - Medium**

Source: medium.com > ...ind/how-netfixs-recommendation...works... ▾

How does Netflix come up with such precise genres for its 100 million-plus subscriber ... **Netflix's recommendation systems** have been developed by hundreds of engineers ... Whenever a user accesses **Netflix** services, the **recommendations system** estimates... Читать ещё >

Result 4: **How does Netflix's recommendation system work?**

Source: radiotimes.com > news...netflix...how-does-it-work/ ▾

The streaming service is **working** hard to match viewers to new shows they'll like - but if you feel like **Netflix** doesn't "get" you, you're not alone.

Result 5: **This is how Netflix's secret recommendation system works**

Source: wired.co.uk > article/how-do-netflixs...work-machine... ▾

This is how **Netflix**'s top-secret **recommendation system** works. ... **Netflix** uses machine learning and algorithms to help break viewers' preconceived notions and find ... To do this, it looks at nuanced threads within the content, rather than relying on broad genres to... Читать ещё >

NLP

АНГЛИЙСКИЙ (ОПРЕДЕЛЕН АВТОМАТИЧЕСКИ)	РУССКИЙ	АНГЛИЙСКИЙ	РУССКИЙ	АНГЛИЙСКИЙ	УКРАИНСКИЙ
<p>The recommendation system works putting together data collected from different places. Recommended rows are tailored to your viewing habits. That's why you can tell when your little cousins have been using your account to watch a billion hours of Peppa Pig. In this case, algorithms are often used to facilitate machine learning. Systems like Netflix based on machine learning rewrite themselves as they learn from their own users. Every time you press play and spend some time watching a TV show or a movie, Netflix is collecting data that informs the algorithm and refreshes it. The more you watch the more up to date the algorithm is. </p>	<p>Система рекомендаций работает, собирая данные, собранные из разных мест. Рекомендуемые строки с учетом ваших привычек просмотра. Вот почему вы можете сказать, когда ваши маленькие двоюродные братья использовали вашу учетную запись, чтобы посмотреть миллиард часов Peppa Pig. В этом случае алгоритмы часто используются для облегчения машинного обучения. Такие системы, как Netflix, основанные на машинном обучении, переписывают себя, учась у своих пользователей. Каждый раз, когда вы нажимаете кнопку воспроизведения и проводите некоторое время за просмотром телепередачи или фильма, Netflix собирает данные, которые информируют алгоритм и обновляют его. Чем больше вы смотрите, тем более современным является алгоритм.</p>	<p>Sistema rekomendatsiy rabotayet, sobiraya dannyye, sobrannyye iz raznykh mest. Rekomenduyemye stroki s uchetom vashikh privychev prosmotra. Vot pochemu vy mozhete skazat', kogda vashi malen'kiye dvoyurodnyye brat'ya ispol'zovali vashu uchetnuyu</p>	<p>Развернуть</p>	<p>637/5000</p>	<p>Скачать</p>

Information retrieval

Frank Rosenblatt (July 11, 1928 – July 11, 1971) was an American psychologist notable in the field of artificial intelligence.

...

Frank Rosenblatt

Born	Frank Rosenblatt July 11, 1928 New Rochelle, New York, U.S.
Died	July 11, 1971 (aged 43) Chesapeake Bay
Alma mater	Cornell University
Known for	Perceptron

[Ещё 1 строка](#)

https://en.wikipedia.org/wiki/Frank_Rosenblatt

[Frank Rosenblatt - Wikipedia](#)

In the request I wrote a name, in response I got a card with key information

NLP - Natural Language Processing

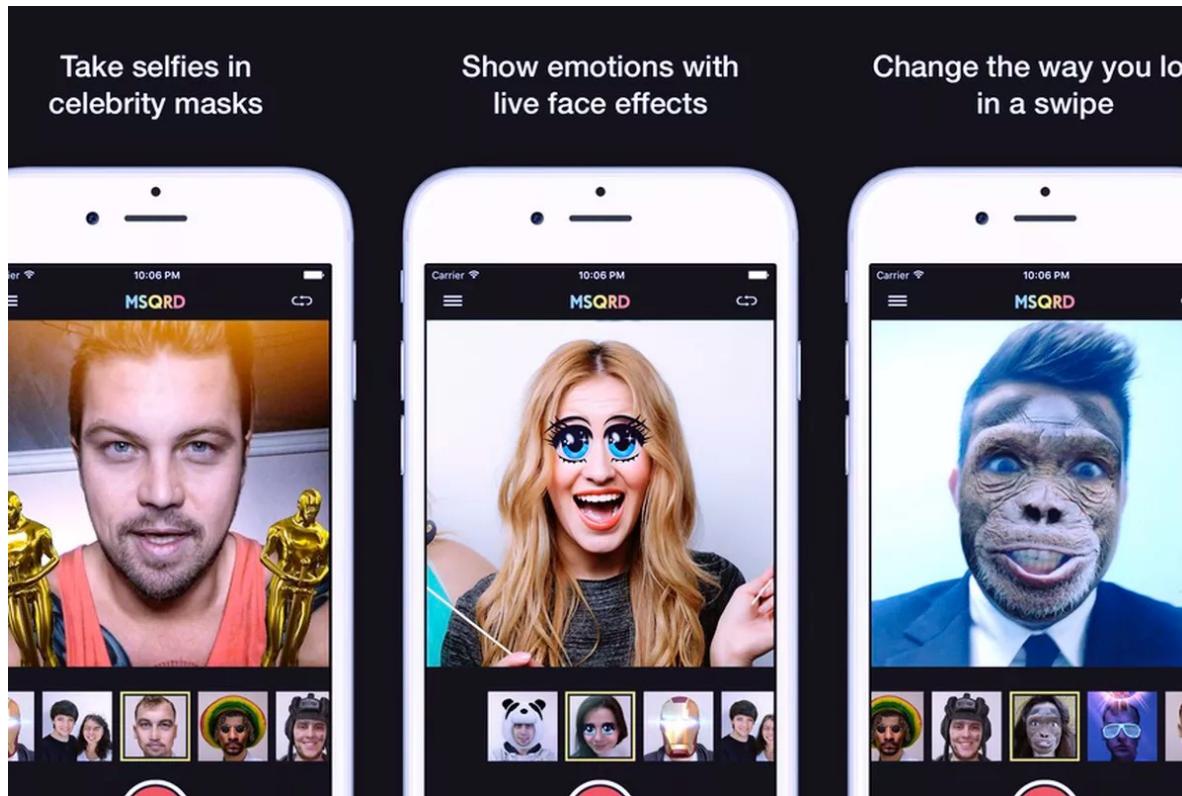
- Machine translation, chat bots, spam filter, search engine suggestions
- And in conjunction with speech recognition - Siri, Alexa

CV – computer vision

- We want to understand what is happening in the image
- Turing test for computer vision problems: answer any question about an image that a person can answer



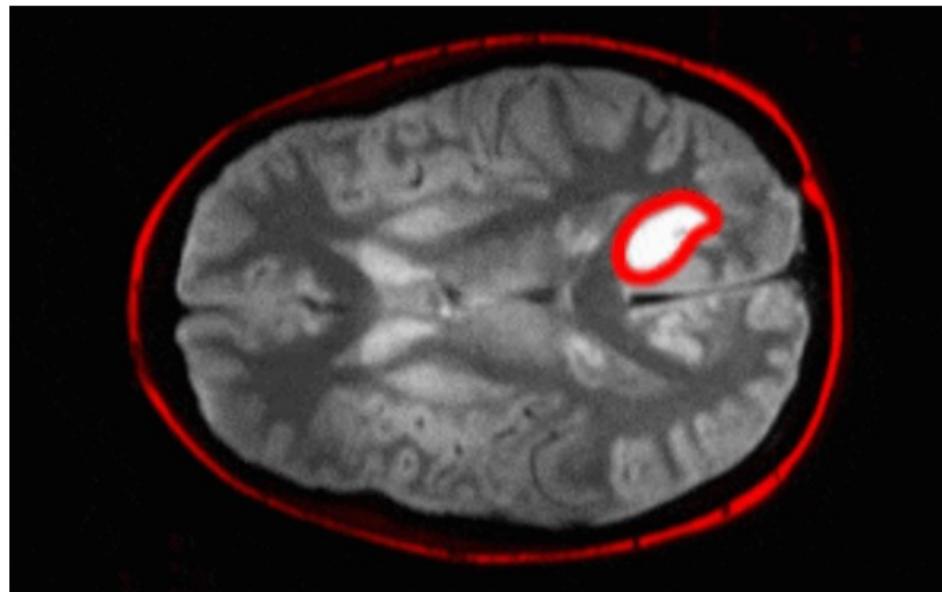
CV applications



CV applications



CV applications



Detection of a brain tumor