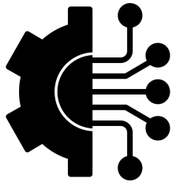
INTRODUCTION TO

Machine Learning and Data Mining



Hoai Thuan TRAN
Gia Dinh Univerrsity

Who is real ?

Why ML & DM ?

- "The most important general-purpose technology of our era is artificial intelligence, particularly **machine learning**" – Harvard Business Review

  https://hbr.org/2017/07/the-business-of-artificial-intelligence
- A huge demand on Data Science
- "Data scientist: the sexiest job of the 21st century" – Harvard Business Review.

  http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/
- "The Age of Big Data" – The New York Times

  http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-theworld.html?pagewanted=all&_r=0

Data Analyst
San Francisco Bay Area
Posted 18 days ago

Data Analyst
Greater New York City Area
Posted 25 days ago

Statistical Analyst - Data…
Greater New York City Area
Posted 9 hours ago

Data Analyst
Greater New York City
Posted 15 days ago

DATA SCIENTIST
Greater New York City
Posted 25 days ago

Data Scientist
Greater New York City
Posted 14 days ago

Marketing Analytics Associate
Greater New York City Area
Posted 24 days ago

Financial Data Analyst
Greater New York City Area
Posted 20 days ago

Data Analyst…
Greater New York City Area
Posted 13 days ago

Home Profile Network Jobs Interests

Data Analyst
Amazon - Newark, NJ
Posted 24 days ago
Apply on company website Save

Senior Data Analyst - Big Data, Meta Product
TripAdvisor - Newton, MA
Posted 12 days ago
Apply now Save

Home Profile Network Jobs Interests

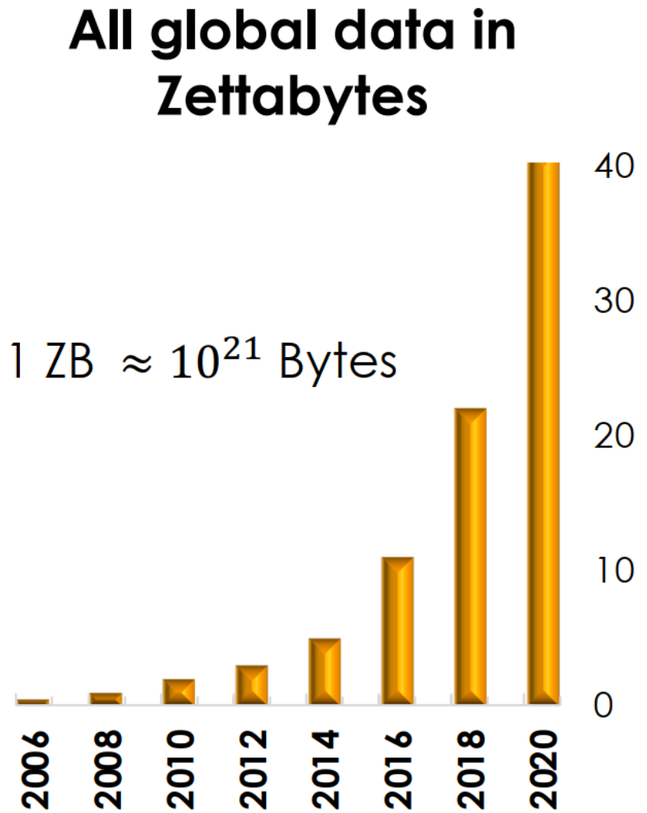
Data Analyst
Apple - Daly City - California -US
Posted 18 days ago
Apply on company website Save

Why ML & DM ?

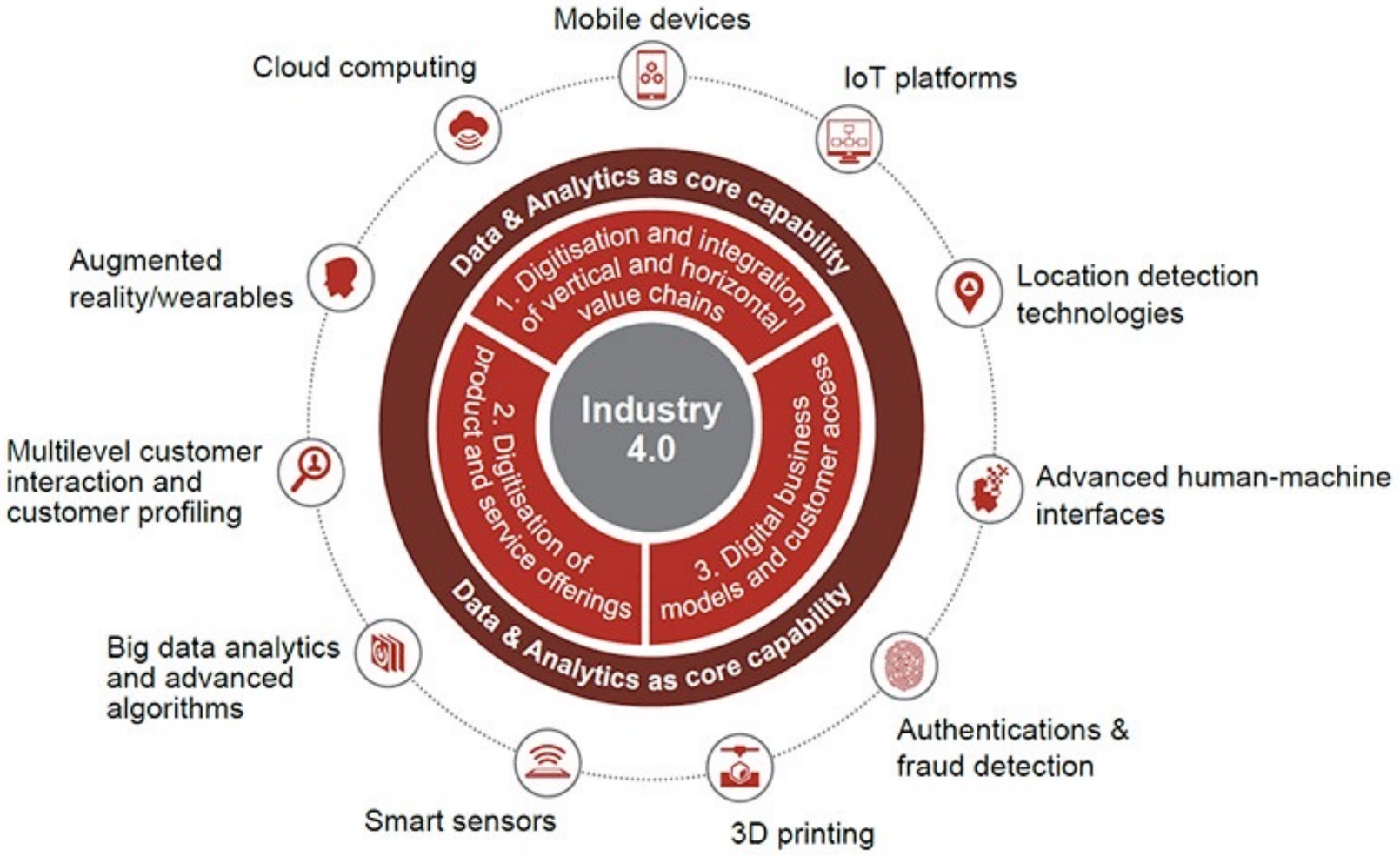
- Data mining, inference, prediction.
- ML & DM provides an efficient way to make intelligent systems/services.
- ML provides vital methods

  and a foundation for Big Data.

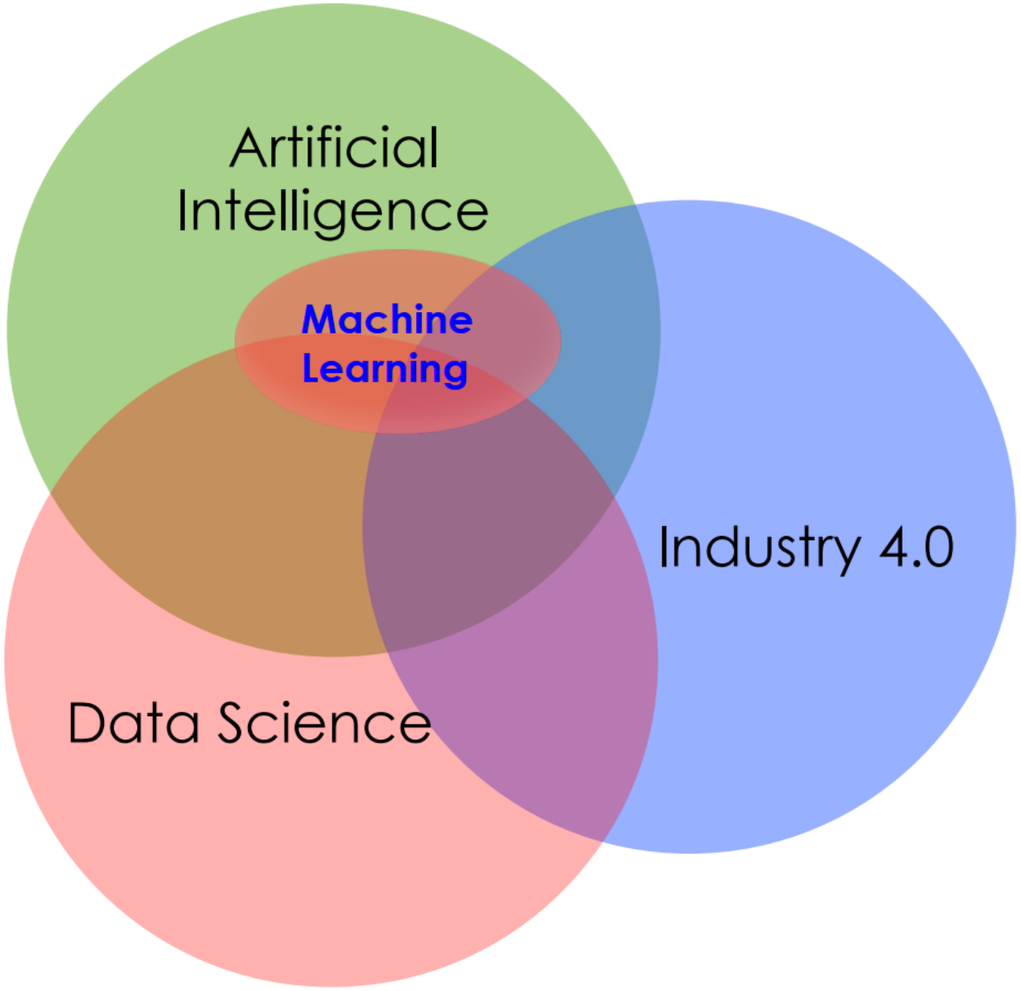
Each day:
230M tweets,
2.7B comments to FB,
86400 hours of video
to YouTube



Industry 4.0


Mobile devices
Cloud computing
IoT platforms
Augmented reality/wearables
Location detection technologies
Multilevel customer interaction and customer profiling
Advanced human-machine interfaces
Big data analytics and advanced algorithms
Authentications & fraud detection
Smart sensors
3D printing
Data & Analytics as core capability
1. Digitisation and integration of vertical and horizontal value chains
2. Digitisation of product and service offerings
3. Digital business models and customer access
Industry 4.0
Data & Analytics as core capability

AI & DS & Industry 4.0

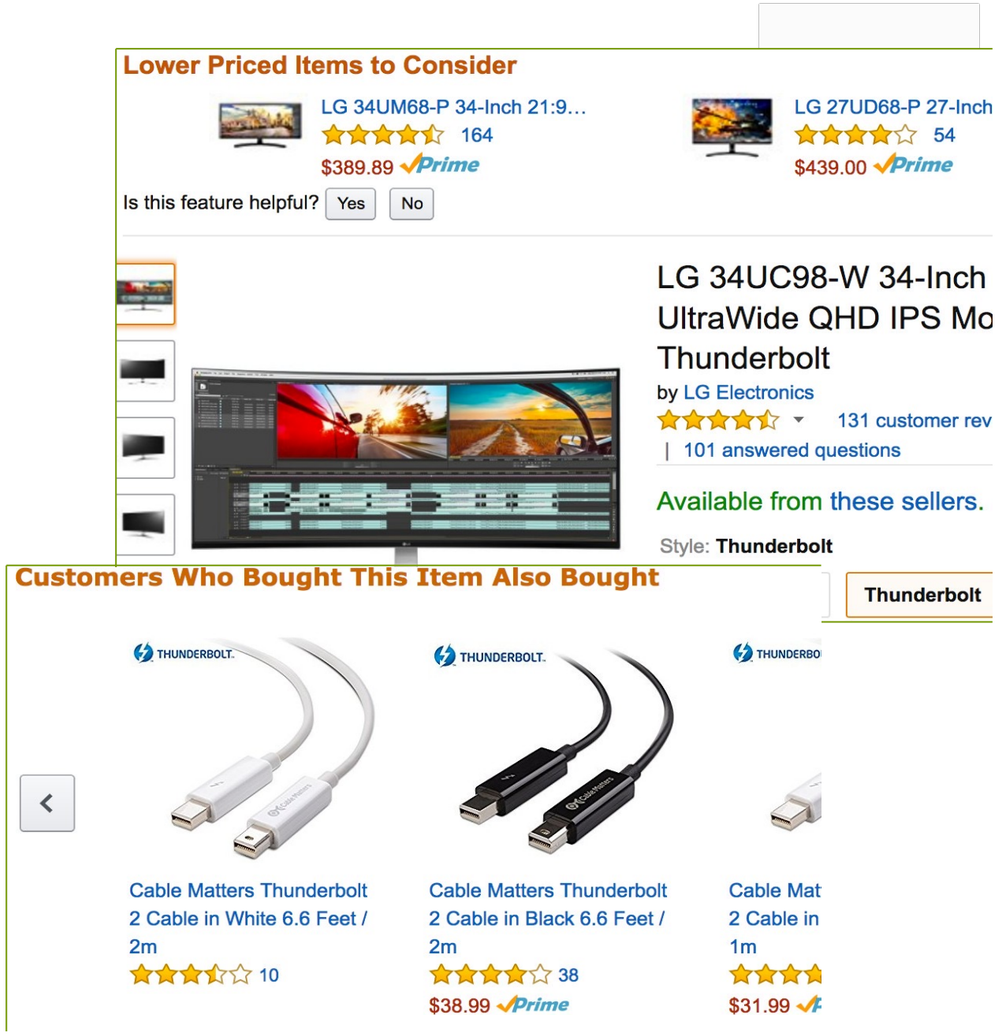


Some Successes: IBM's Watson (2011)

IBM's Watson Supercomputer Destroys Humans in Jeopardy (2011)

Some Successes: Amazon's secret (2012)

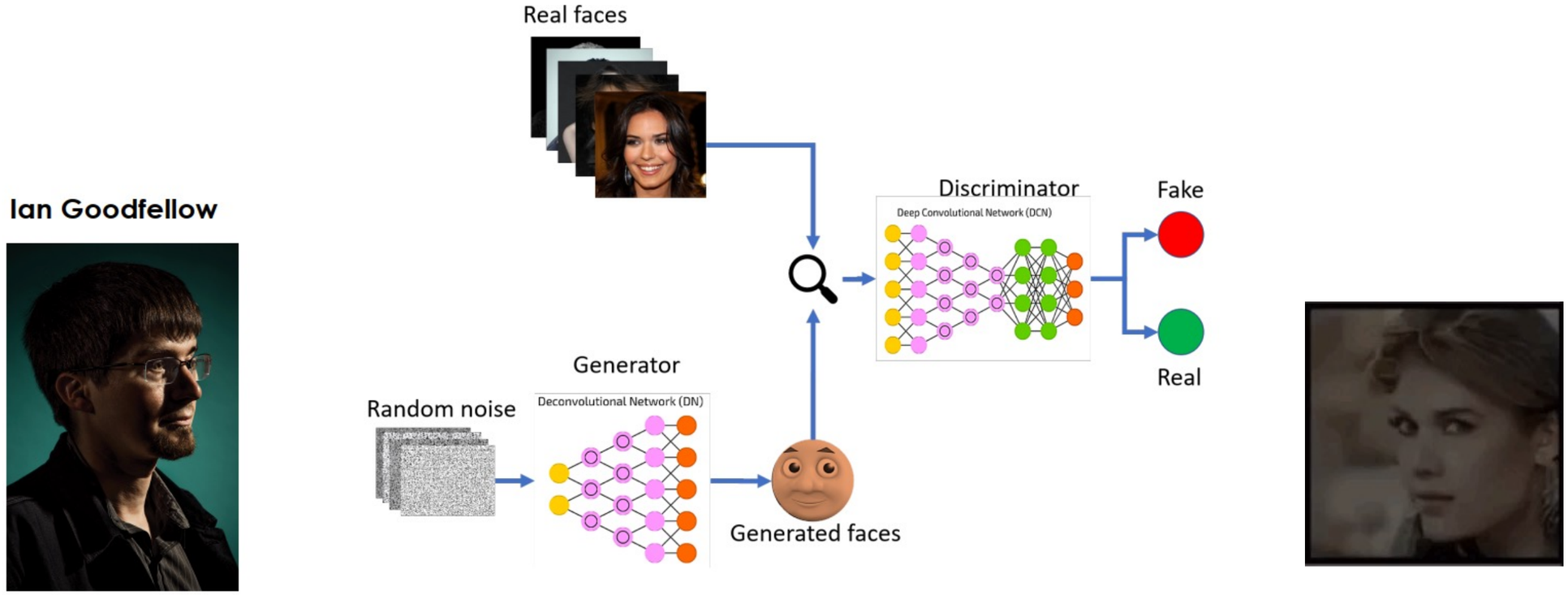
"The company reported a **29% sales increase** to $12.83 billion during its second fiscal quarter, up from $9.9 billion during the same time last year." – Fortune, July 30, 2012



Some Successes: GAN (2014)

- Imagination

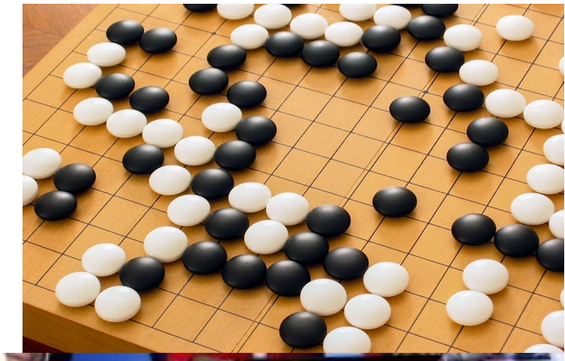
$$\min_G \max_D \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D\left(G(z)\right)\right)\right]$$



Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "**Generative adversarial nets**." In NIPS, pp. 2672-2680. 2014

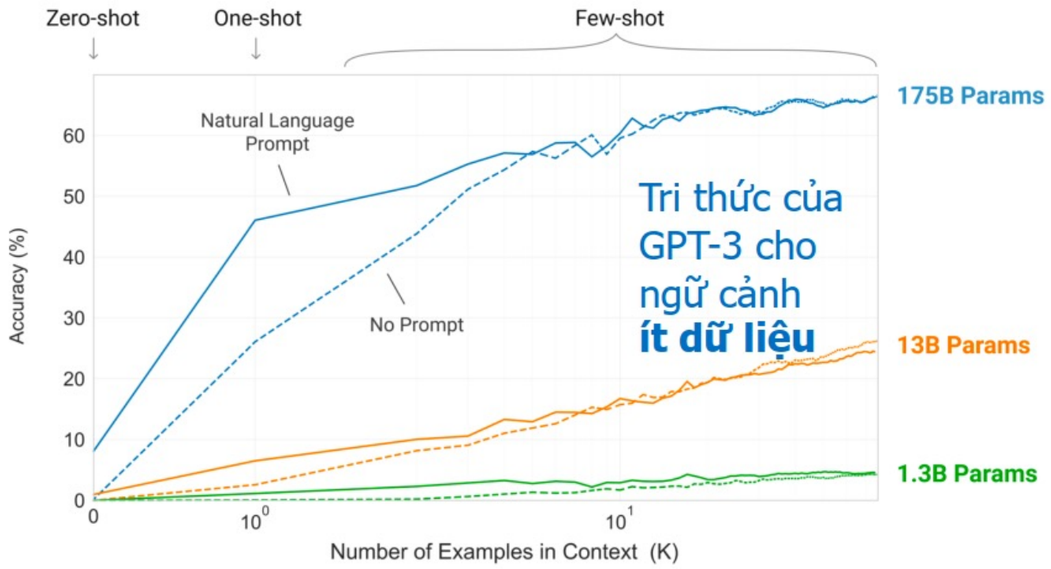
Some Successes: AlphaGo (2016)

- AlphaGo of Google DeepMind the world champion at Go, 3/2016.
  - Go is a 2500 year-old game.
  - Go is one of the most complex games.
- AlphaGo learns from 30 millions human moves, and plays itself to find new moves.
- It beat Lee Sedol (World champion).



Some Successes: Chat GPT -3.0

- Create writing capabilities for computers
  - A huge model trained from huge data.
  - It can be used in many problems with little data.



Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." NeurIPS (2020). Best Paper Award.

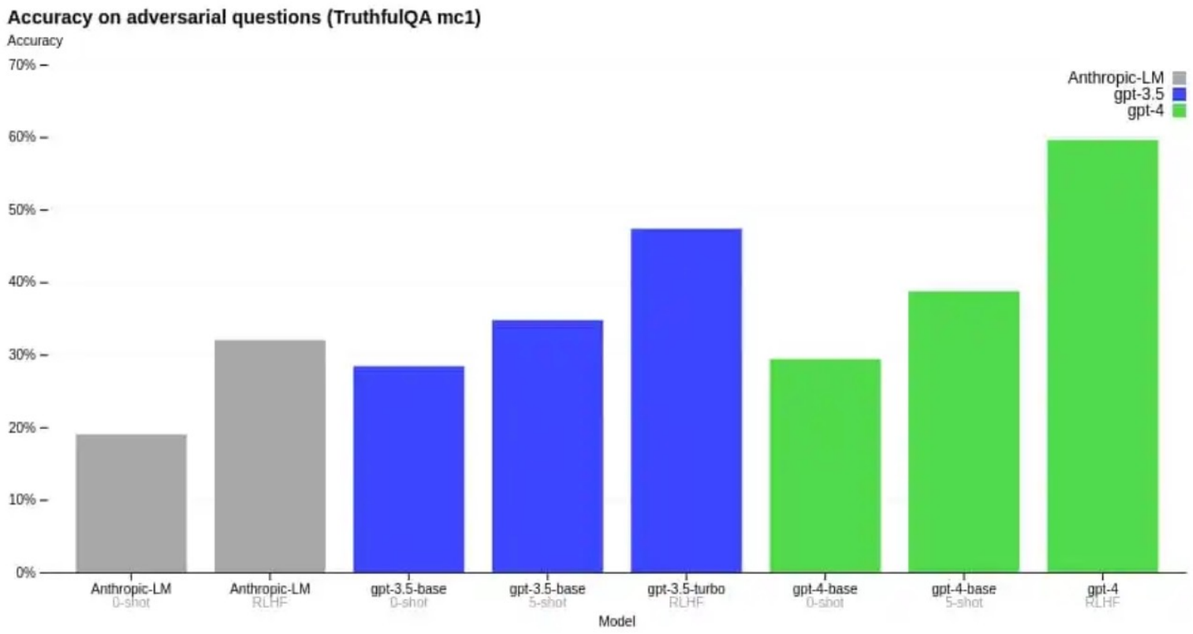
I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

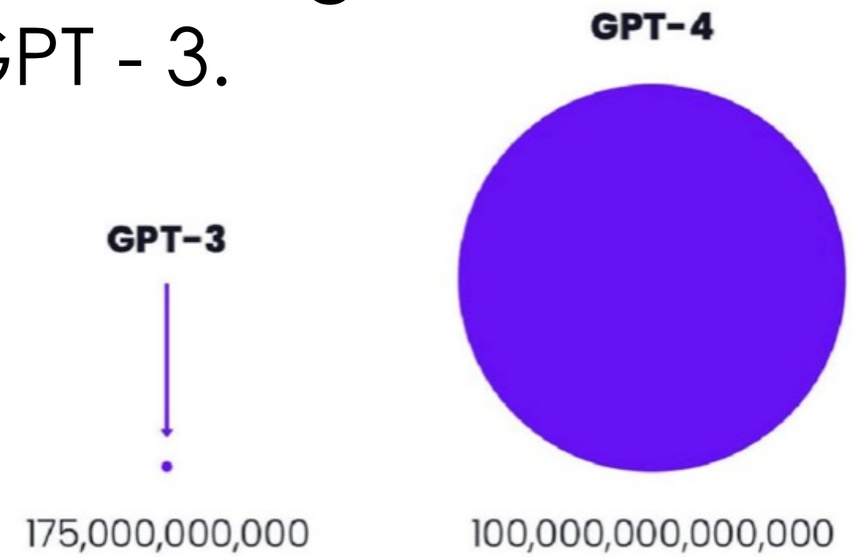
The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

	Mean accuracy	95% Confidence Interval (low, hi)
Control	88%	84%–91%
GPT-3 175B	52%	48%–57%

Some Successes: Chat GPT – 4.0

- GPT - 4 can analyze the contents of an image to find relevant information.
- GPT - 4 can handle larger text inputs, processing 25,000 words – eight times more than GPT - 3.





The comparison between GPT-3 and GPT-4 based on the number of parameters used in their architecture.

Machine Learning & Data Mining

- Machine Learning – ML

To build computer systems that can improve themselves by learning from data.

(Xây dựng những hệ thống mà có khả năng tự cải thiện bản thân bằng cách học từ dữ liệu.)

- Some venues:

NeurIPS, ICML, IJCAI, AAAI, ICLR, ACML, ECML

- Data Mining – DM

To find new and useful knowledge from datasets.

(Tìm ra/Khai phá những tri thức mới và hữu dụng từ các tập dữ liệu lớn.)

- Some venues:

KDD, PKDD, PAKDD, ICDM, CIKM

Data

Structured – relational (table-like)

	A	B	C	D	E	F	G
1	Country	Region	Population	Under15	Over60	Fertil	LifeExp
2	Zimbabwe	Africa	13724	40.24	5.68	3.64	54
3	Zambia	Africa	14075	46.73	3.95	5.77	55
4	Yemen	Eastern M	23852	40.72	4.54	4.35	64
5	Viet Nam	Western P	90796	22.87	9.32	1.79	75
6	Venezuela (Bo	Americas	29955	28.84	9.17	2.44	75
7	Vanuatu	Western P	247	37.37	6.02	3.46	72
8	Uzbekistan	Europe	28541	28.9	6.38	2.38	68
9	Uruguay	Americas	3395	22.05	18.59	2.07	77

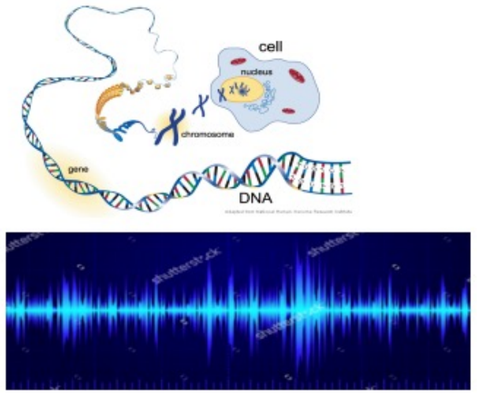
Un-structured


```
{
    "code": "1473a6fd39d1d8fa48654aac9d8cc2754232
    "title": "[Updating] Câu chuyện xuyên mưa về
    "url": "http://techtalk.vn/updating-cau-chuye
    "labels": "techtalk/Cong nghe",
    "content": "Vào chiều tối ngày 09/12/2016 vừa
    "image_url": "",
    "date": "2016-12-10T03:51:10Z"
}
```

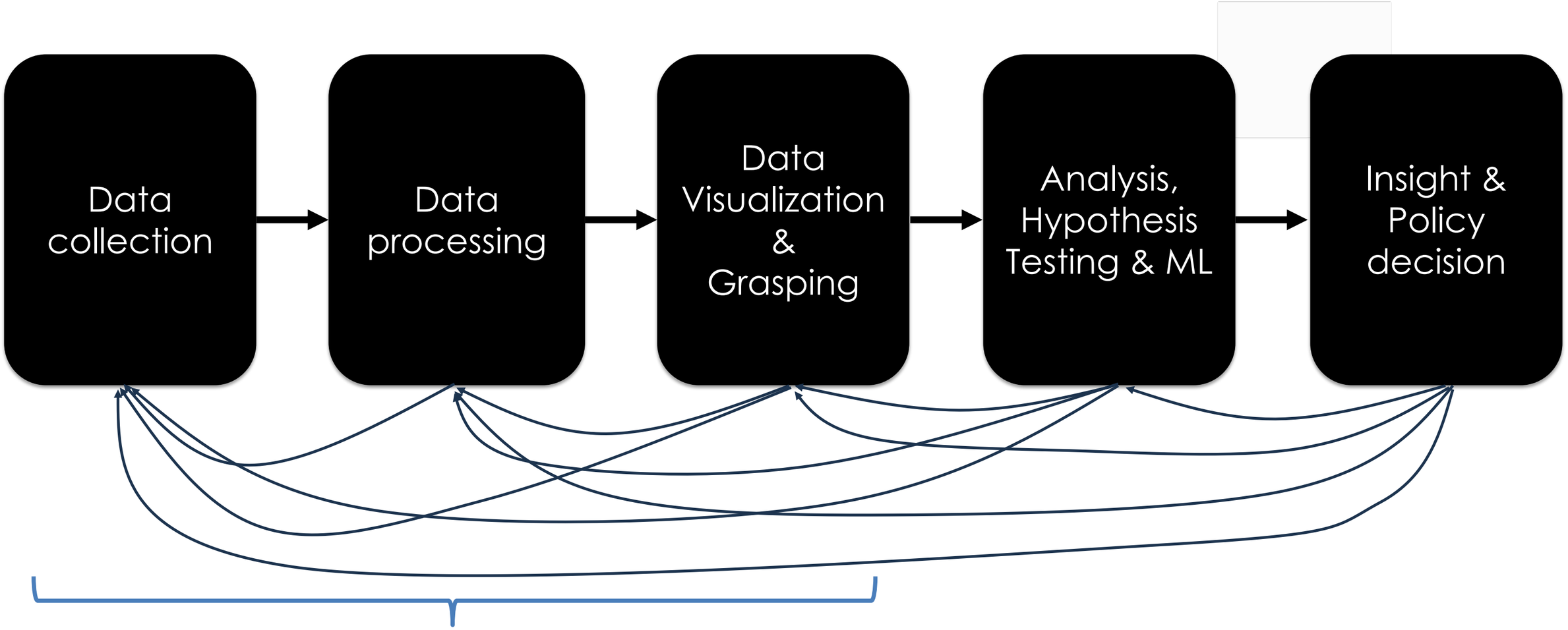

texts in websites, emails, articles, tweets

2D/3D images, videos + meta

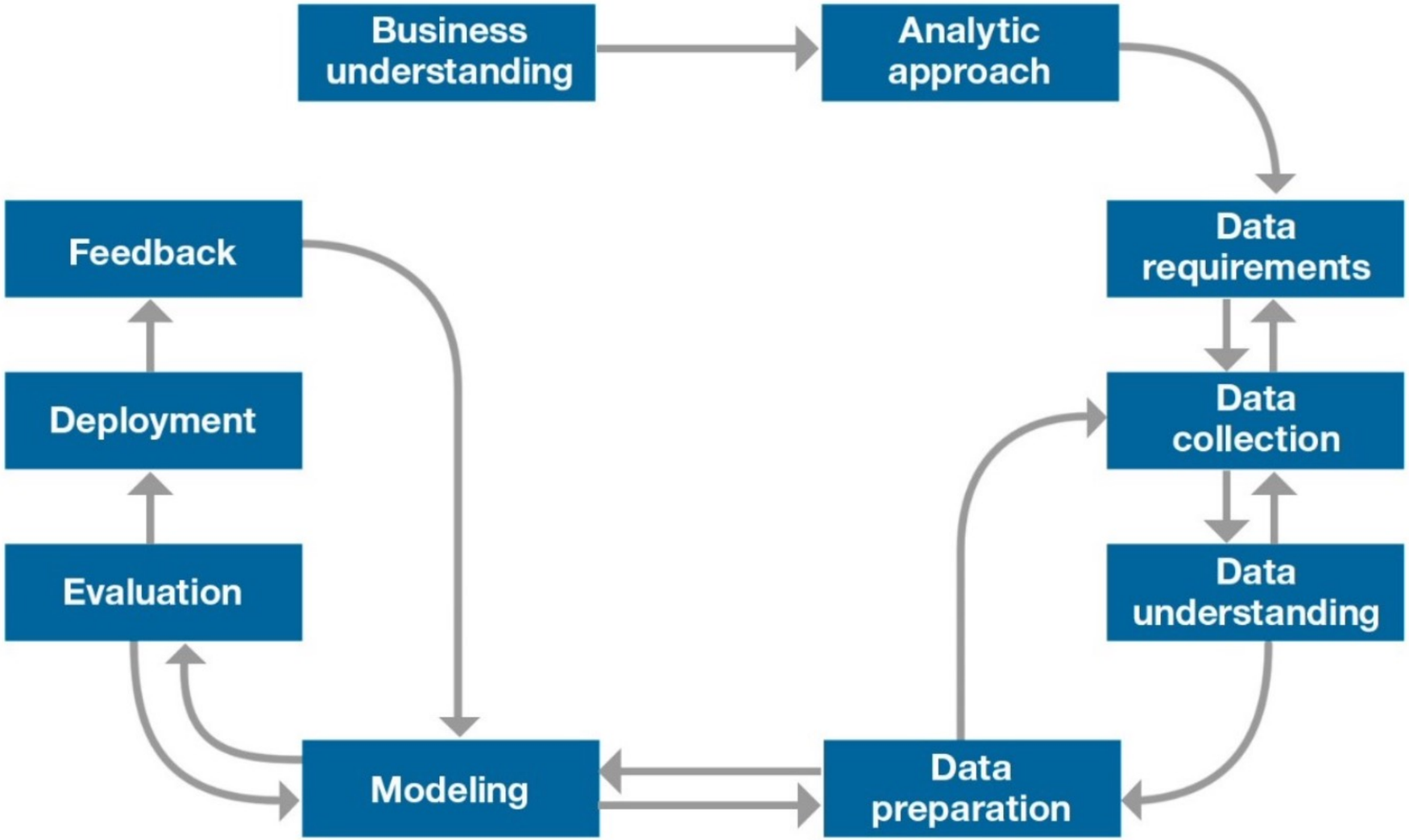
spectrograms, DNAs, …



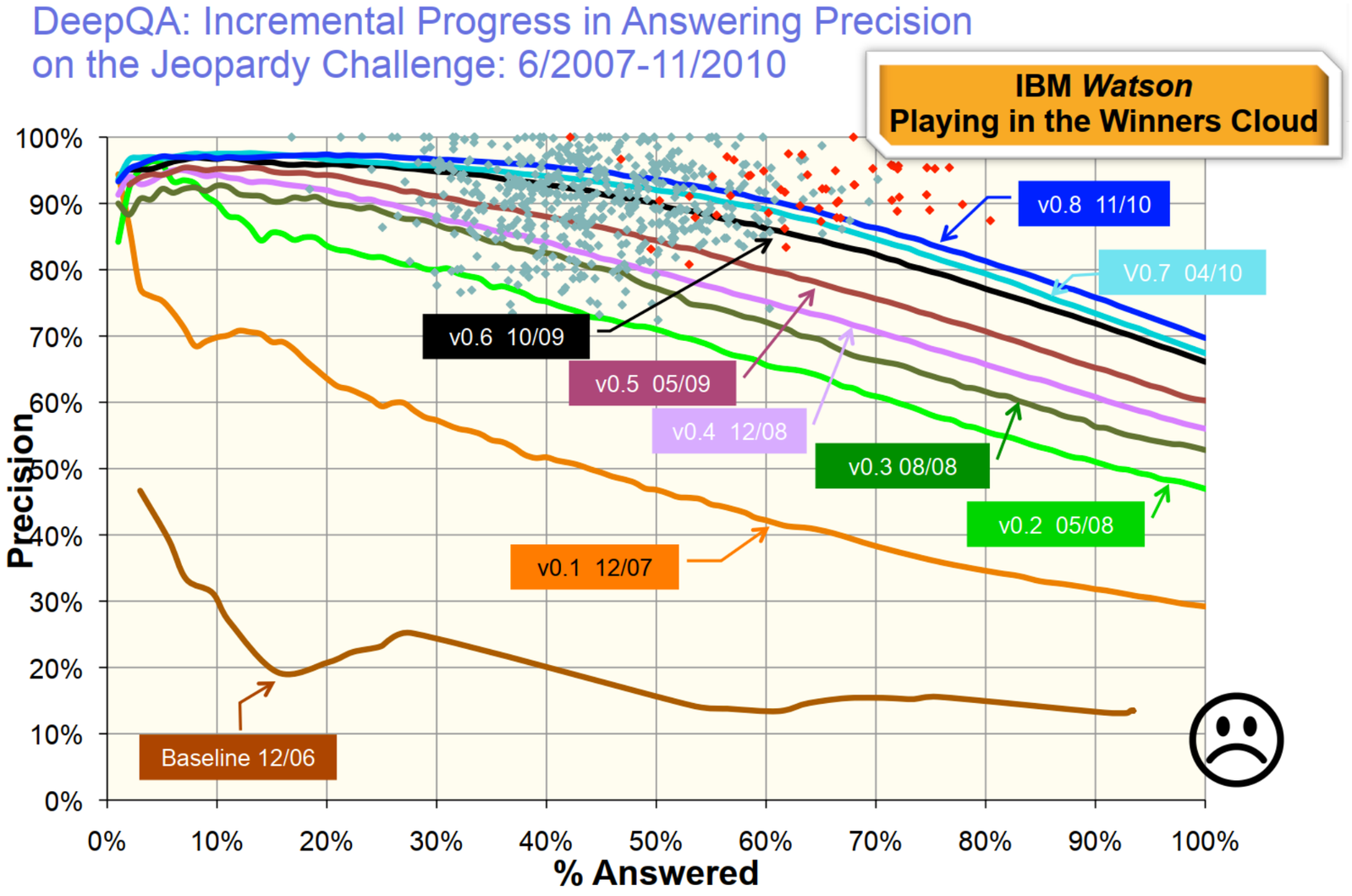
Methodology: insight-driven



Methodology: product-driven



Product development: experience



What is Machine Learning?

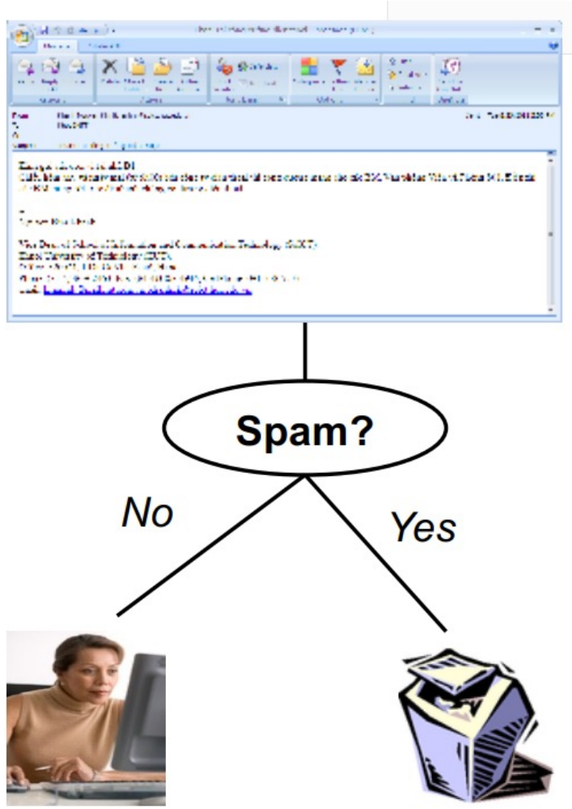
- Machine Learning (ML) is an active subfield of Artificial Intelligence
- ML seeks to answer the question [Mitchell, 2006]
  - How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?
- Some other views on ML:
  - Build systems that automatically improve their performance [Simon,1983].
  - Program computers to optimize a performance objective at some task, based on data and past experience [Alpaydin, 2020].

A learning machine

- We say that a machine learns if the system reliably improves its performance **P** at task **T**, following experience **E**.
- A *learning problem* can be described as a triple (**P**, **T**, **E**).
- ML is close to and intersects with many areas:
  - Computer Science,
  - Statistics, Probability,
  - Optimization,
  - Psychology, Neuroscience,
  - Computer Vision,
  - Economics, Biology, Bioinformatics, …

Some real examples (1)

- Spam filtering for emails
  - **T**: filter/predict all emails that are spam.
  - **P**: the accuracy of prediction, that is the percentage of emails that are correctly
  - **E**: set of old emails, each with a label of spam/normal.



Some real examples (2)

- Image tagging
  - **T**: give some words that describe the meaning of a picture.
  - **P**: ?
  - **E**: set of pictures, each has been labelled with a set of words.

FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

What does a machine learn?

- A mapping (function):

$$f : x \mapsto y$$

  - x: observations (data), past experience
  - y: prediction, new knowledge, new experience,…

- A model (mô hình):
  - Data are often supposed to follow or be generated from an unknown model.
  - Learning a model means learning the parameters of that model.

Where does a machine learn from?

- Learn from a set of training examples (training set):

$$\{ \{x_1, x_2, \ldots, x_N\}; \{y_1, y_2, \ldots, y_M\} \}$$

  - x_i is an observation (quan sát) of x in the past.
  - y_j is an observation of y in the past, often called label or response or output.

- After learning:
  - We obtain a model, new knowledge, or new experience f.
  - We can use that model/function to do **prediction** or **inference** for future observations, e.g.,

$$y = f(x)$$

Two basic learning problems

- Supervised learning: learn a function y = f(x) from a given training set $\{x_1, x_2, \ldots, x_N, y_1, y_2, \ldots, y_N\}$ so that $y_i \cong f(x_i)$ for every i.
  - **Classification**, **categorization** : if y only belongs to a discrete set, for example {spam, normal}.
  - **Regression**: if y is a real number.
- Unsupervised learning: learn a function y = f(x) from a given training set $\{x_1, x_2, \ldots, x_N\}$.
  - y can be a data cluster, a hidden structure or a trend
- Other: semi-supervised learning, reinforcement learning,…

Supervised learning: Classification

- Multiclass classification: when the output y is one of the pre-defined labels $\{c_1, c_2, \ldots, c_L\}$
  (mỗi đầu ra chỉ thuộc 1 lớp, mỗi quan sát x chỉ có 1 nhãn)
  - Spam filtering: y in {spam, normal}
  - Financial risk estimation: y in {high, normal, no}
- Multilabel classification: when the output y is a subset of labels
  (mỗi đầu ra là một tập nhỏ các lớp; mỗi quan sát x có thể có nhiều nhãn)
  - Image tagging: y = {birds, nest, tree}
  - Sentiment analysis

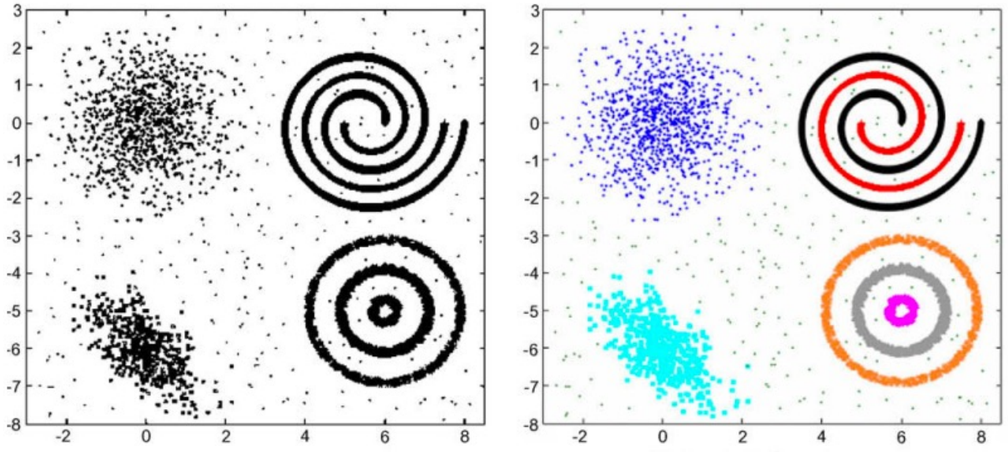
BIRDS NEST TREE
BRANCH LEAVES

Supervised learning: Regression

- Prediction of stock indices

Unsupervised learning: examples (1)

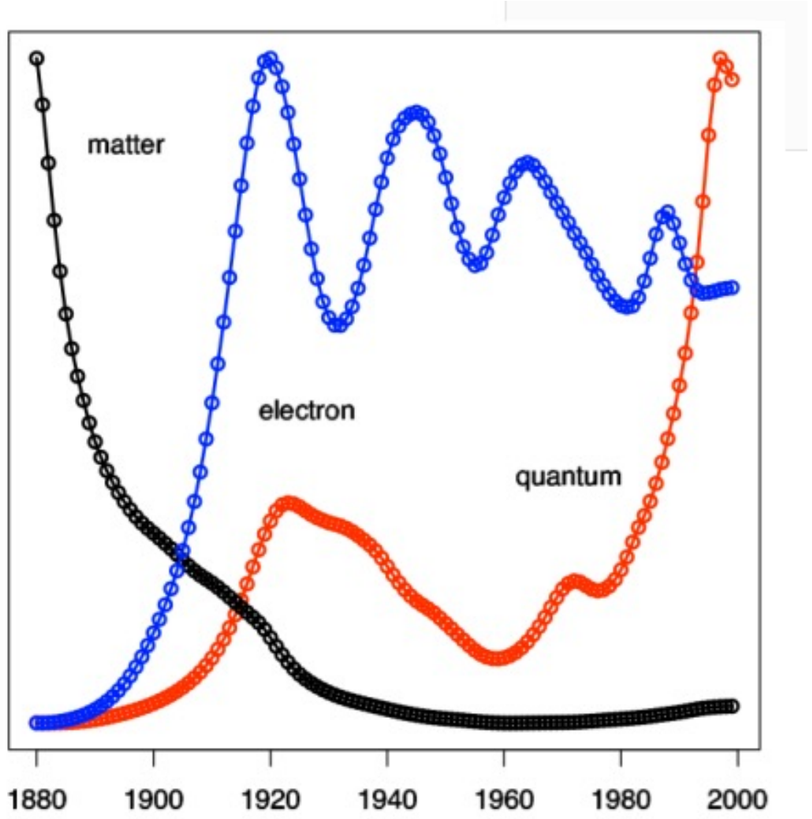
- Clustering data into clusters
  - Discover the data groups/clusters



- Community detection
  - Detect communities in online social networks

Unsupervised learning: examples (2)

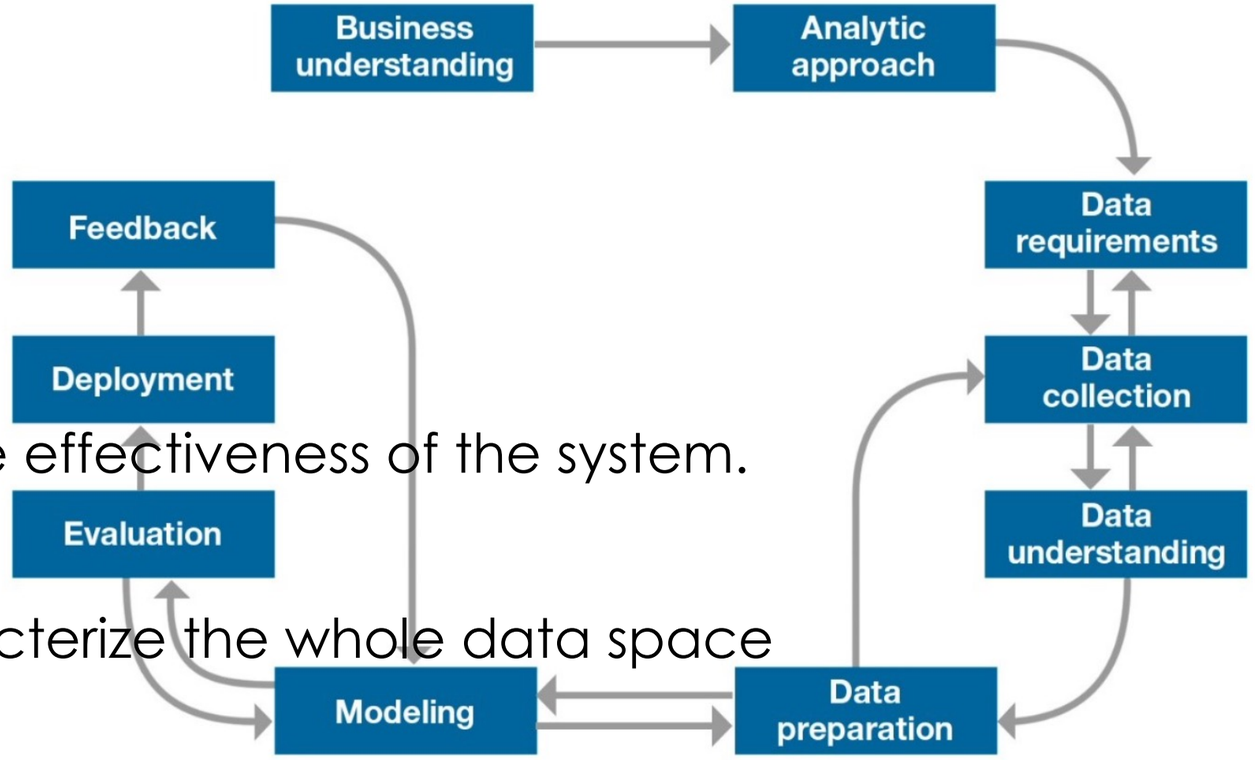
- Trends detection
  - Discover the trends, demands, future needs of online users



Design a learning system

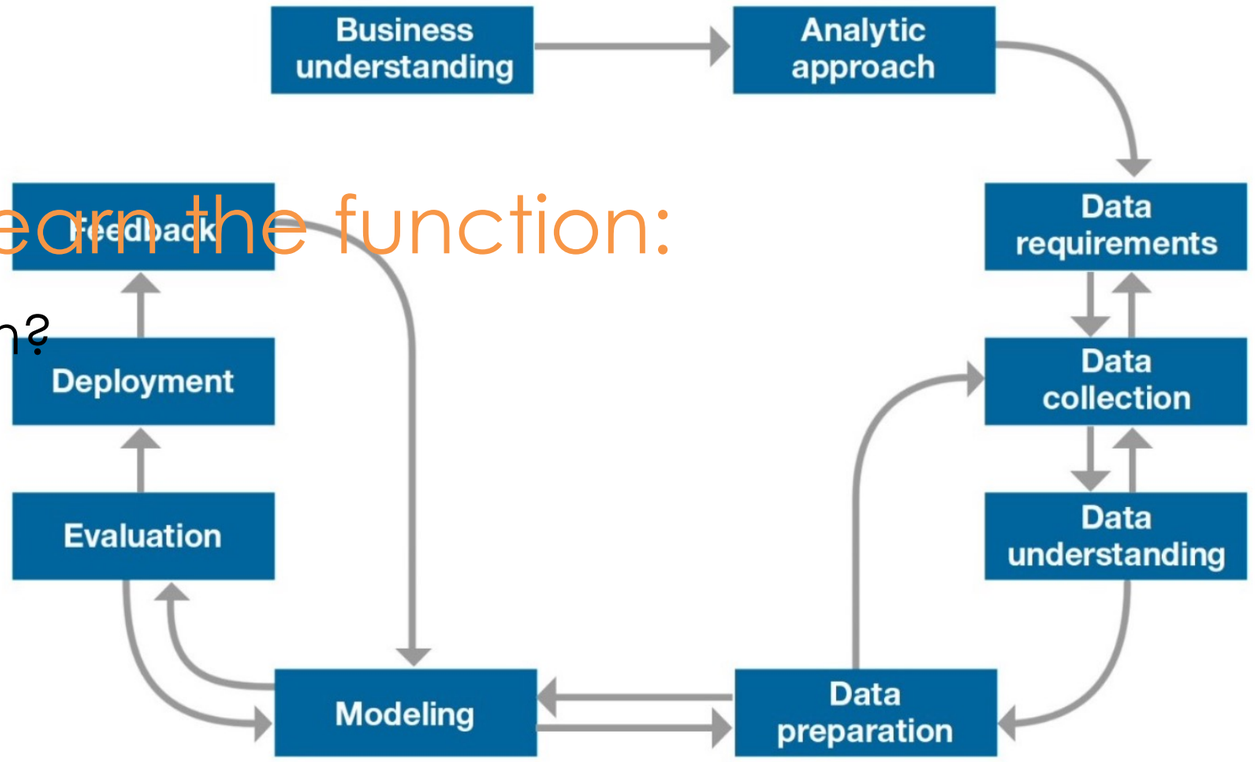
Some issues should be carefully considered when designing a learning system.

- Determine the type of the function to be learned
  - F: X → {0,1}
  - F: X → set of labels/tags
  - F: X → R
- Select a training set:
  - The training set plays the key role in the effectiveness of the system.
  - Do the observations have any label?
  - The training observations should characterize the whole data space ⟹ good for future predictions.



Design a learning system

- Select a representation for the function: (model)
  - Linear?
  - A neural network?
  - A decision tree? …
- Select a good algorithm to learn the function:
  - Ordinary least square? Ridge regression?
  - Back-propagation?
  - D3?



ML: some issues

- **Learning algorithm**
  - Under what conditions the chosen algorithm will (asymtotically) converge?
  - For a given application/domain and a given objective function, what algorithm performs best?
- **No-free-lunch theorem** [Wolpert and Macready, 1997]

If an algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems.

No algorithm can beat another on all domains.

ML: some issues

- **Training data**
  - How many observations are enough for learning?
  - Whether or not does the size of the training set affect performance of an ML system?
  - What is the effect of the disrupted or noisy observations?

ML: some issues

- **Learnability**
  - The goodness/limit of the learning algorithm?
  - What is the generalization of the system?

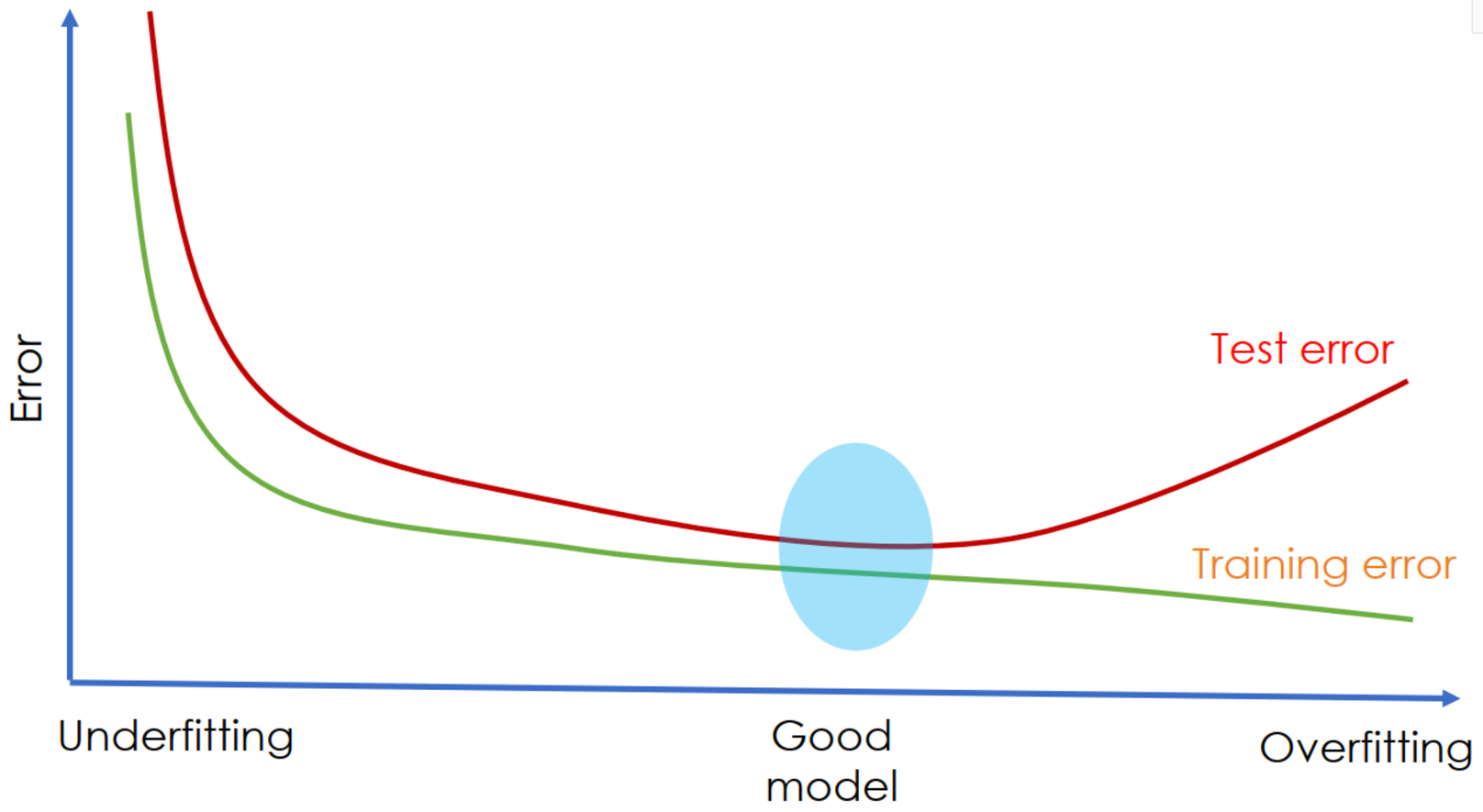
  - Predict well new observations, not only the training data.

  - Avoid overfitting.

Overfitting

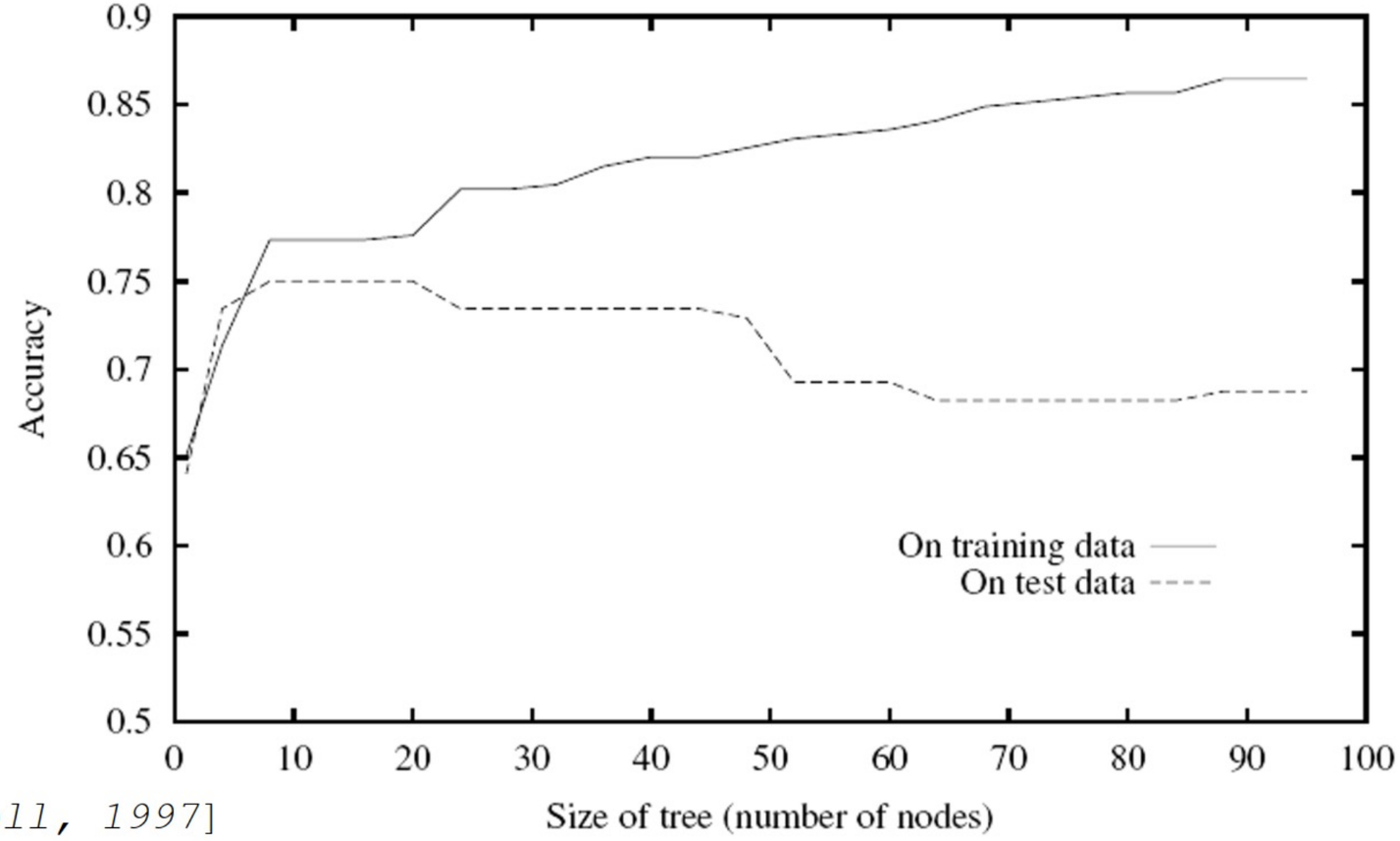
- Function h is called overfitting [Mitchell, 1997] if there exists another function g such that:
  - g might be worse than h for the training data, but
  - g is better than h for future data.
- A learning algorithm is said to overfit relative to another one if it is more accurate in fitting known data, but less accurate in predicting unseen data.
- Overfitting is caused by many factors:
  - The trained function/model is too complex or have too much parameters.
  - Noises or errors are present in the training data.
  - The training size is too small, not characterizing the whole data space.

Overfitting



Overfitting: Example

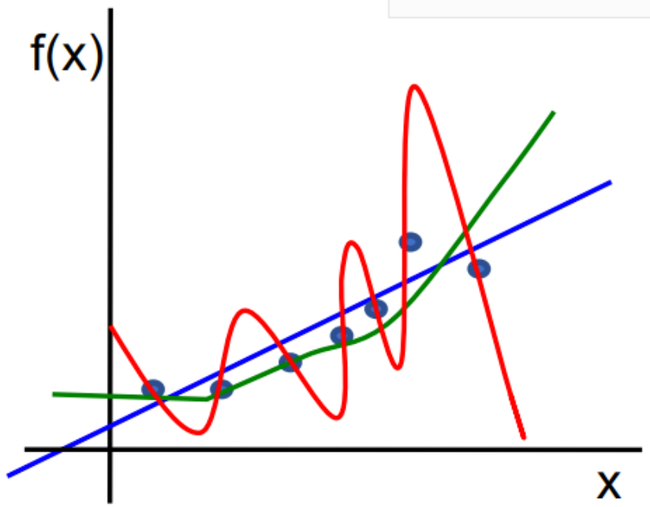
- Increasing the size of a decision tree can degrade prediction on unseen data, even though increasing the accuracy for the training data.

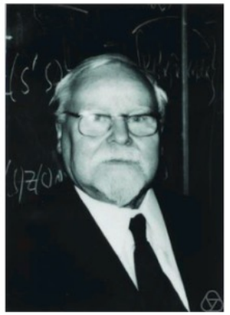


[*Mitchell, 1997*]

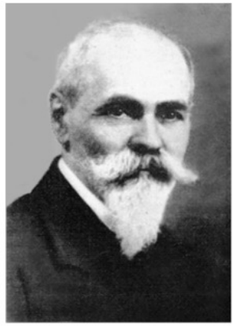
Overfitting: Regularization

- Among many functions, which one can generalize best from the given training data?
  - Generalization is the main target of ML
  - Predict unseen data well.
- Regularization: a popular choice





Tikhonov, smoothing an ill-posed problem



Zaremba, model complexity minimization

Bayes: priors over parameters

Andrew Ng: need no maths, but it prevents overfitting!