Generic Recipe for AI-related Problems

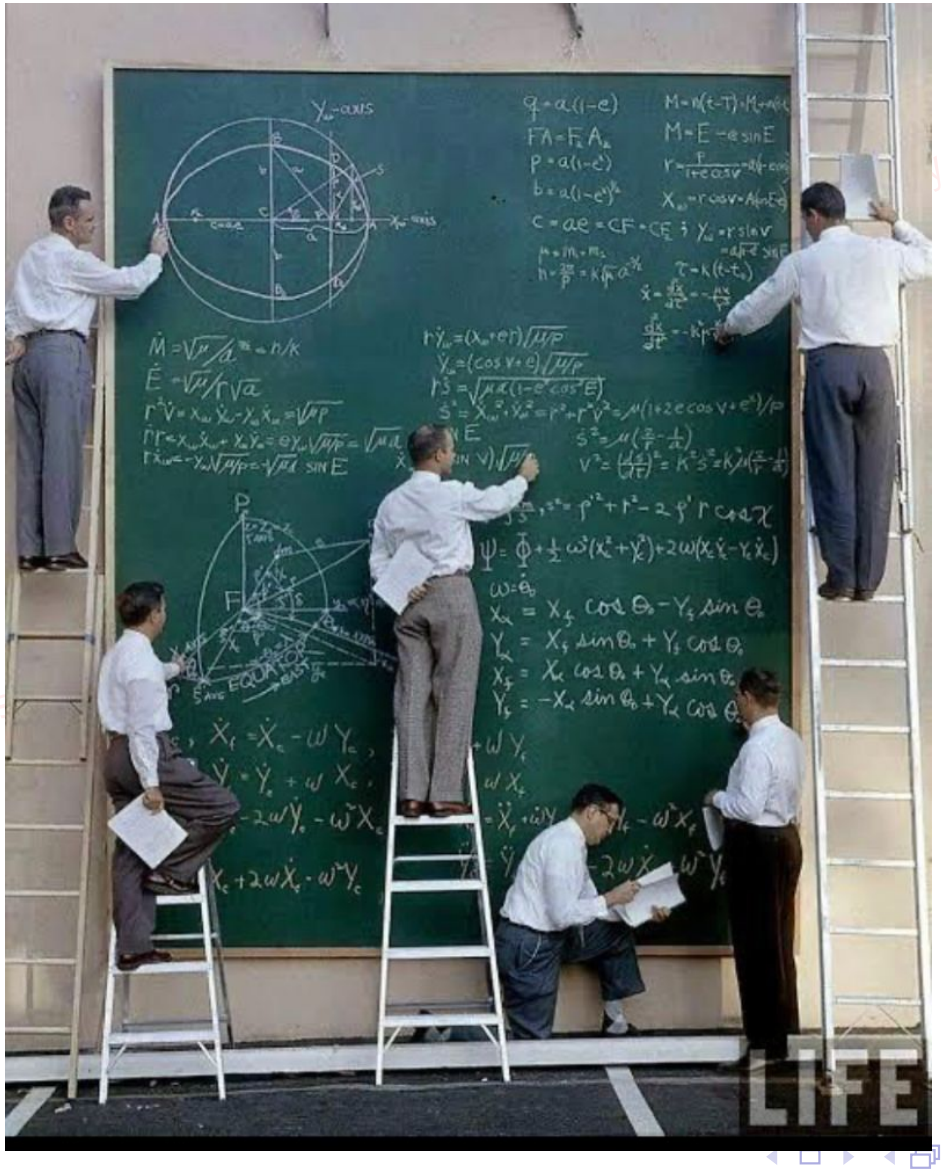
Ali Hirsa
Professor & Director of Financial Engineering
Director of Center for AI in Business Analytics & FinTech
Industrial Engineering & Operations Research (IEOR)
Data Science Institute (DSI)
Columbia University
&
Chief Scientific Officer, ASK2.ai
Managing Partner, Sauma Capital, LLC

AI Course Columbia+

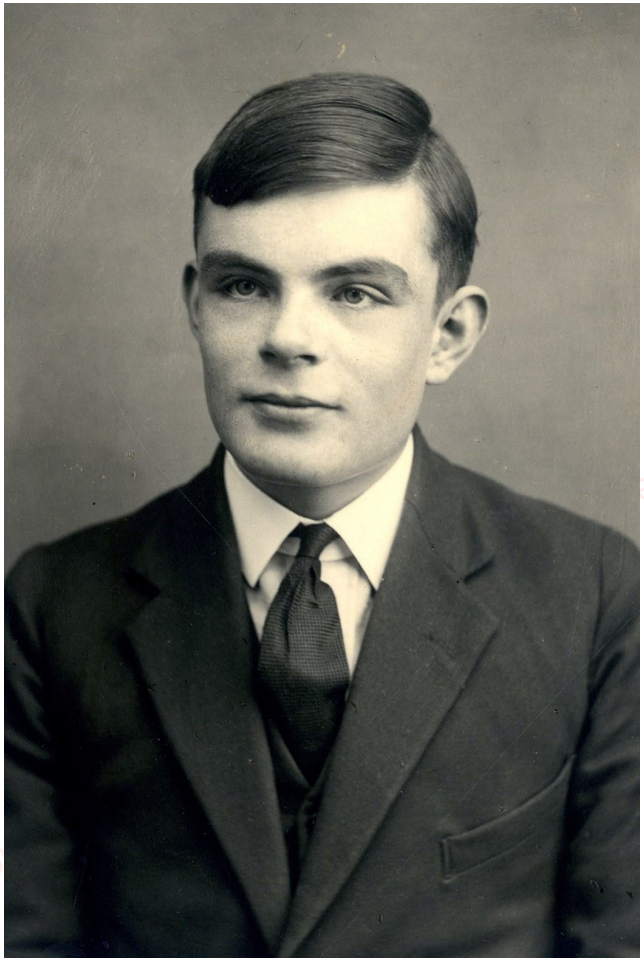
Agenda

- $\mathbb{AI}$ evolution & revolution
- recipe for becoming an $\mathbb{AI}$ chef
  - data
  - models & robustness
  - objective function
  - initial condition/starting point
  - optimization

Picture of the Day



Quote of the Day



A computer would deserve to be called intelligent if it could deceive a human into believing that it was human

`Alan Turing` (wrote in 1950 defining his now-famous `Turing Test`)

Number of pages in Google search?

	5/1/2021	9/7/2021	1/15/2022	9/6/2022	5/20/2023	8/29/2023
AI Artificial Intelligence Machine Learning Deep Learning Natural Language Processing						
Big Data Data Science Data Analytics Business Analytics Financial Technology						
OpenAI ChatGPT Language Modeling Large Language Models						

Number of pages in Google search?

	5/1/2021	9/7/2021	1/15/2022	9/6/2022	5/20/2023	8/29/2023
AI	3,160,000,000					
Artificial Intelligence	753,000,000					
Machine Learning	2,230,000,000					
Deep Learning	2,130,000,000					
Natural Language Processing	964,000,000					
Big Data	5,400,000,000					
Data Science	3,060,000,000					
Data Analytics	1,740,000,000					
Business Analytics	-					
Financial Technology	-					
OpenAI	-					
ChatGPT	-					
Language Modeling	-					
Large Language Models	-					

Number of pages in Google search?

	5/1/2021	9/7/2021	1/15/2022	9/6/2022	5/20/2023	8/29/2023
AI	3,160,000,000	3,200,000,000				
Artificial Intelligence	753,000,000	890,000,000				
Machine Learning	2,230,000,000	2,090,000,000				
Deep Learning	2,130,000,000	2,420,000,000				
Natural Language Processing	964,000,000	-				
Big Data	5,400,000,000	6,320,000,000				
Data Science	3,060,000,000	2,620,000,000				
Data Analytics	1,740,000,000	1,140,000,000				
Business Analytics	-	1,100,000,000				
Financial Technology	-	79,400,000				
OpenAI	-	-				
ChatGPT	-	-				
Language Modeling	-	-				
Large Language Models	-	-				

Number of pages in Google search?

	5/1/2021	9/7/2021	1/15/2022	9/6/2022	5/20/2023	8/29/2023
AI	3,160,000,000	3,200,000,000	3,740,000,000			
Artificial Intelligence	753,000,000	890,000,000	829,000,000			
Machine Learning	2,230,000,000	2,090,000,000	2,190,000,000			
Deep Learning	2,130,000,000	2,420,000,000	2,230,000,000			
Natural Language Processing	964,000,000	-	1,110,000,000			
Big Data	5,400,000,000	6,320,000,000	6,140,000,000			
Data Science	3,060,000,000	2,620,000,000	2,780,000,000			
Data Analytics	1,740,000,000	1,140,000,000	1,140,000,000			
Business Analytics	-	1,100,000,000	3,320,000,000			
Financial Technology	-	79,400,000	1,840,000,000			
OpenAI	-	-	-			
ChatGPT	-	-	-			
Language Modeling	-	-	-			
Large Language Models	-	-	-			

Number of pages in Google search?

	5/1/2021	9/7/2021	1/15/2022	9/6/2022	5/20/2023	8/29/2023
AI	3,160,000,000	3,200,000,000	3,740,000,000	8,910,000,000		
Artificial Intelligence	753,000,000	890,000,000	829,000,000	1,360,000,000		
Machine Learning	2,230,000,000	2,090,000,000	2,190,000,000	2,240,000,000		
Deep Learning	2,130,000,000	2,420,000,000	2,230,000,000	1,600,000,000		
Natural Language Processing	964,000,000	-	1,110,000,000	667,000,000		
Big Data	5,400,000,000	6,320,000,000	6,140,000,000	8,040,000,000		
Data Science	3,060,000,000	2,620,000,000	2,780,000,000	5,470,000,000		
Data Analytics	1,740,000,000	1,140,000,000	1,140,000,000	2,130,000,000		
Business Analytics	-	1,100,000,000	3,320,000,000	3,320,000,000		
Financial Technology	-	79,400,000	1,840,000,000	2,390,000,000		
OpenAI	-	-	-	-		
ChatGPT	-	-	-	-		
Language Modeling	-	-	-	-		
Large Language Models	-	-	-	-		

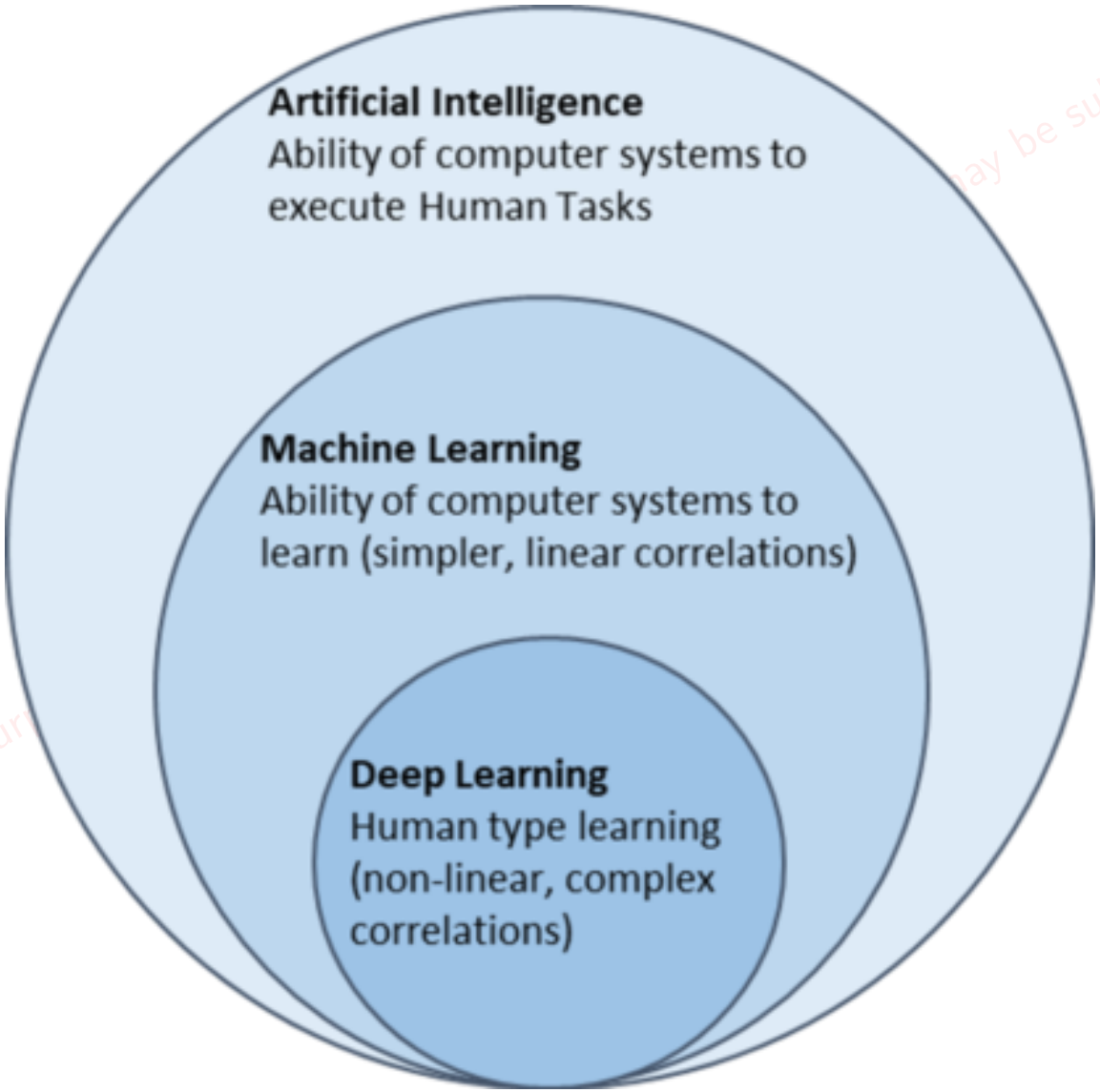
Number of pages in Google search?

	5/1/2021	9/7/2021	1/15/2022	9/6/2022	5/20/2023	8/29/2023
AI	3,160,000,000	3,200,000,000	3,740,000,000	8,910,000,000	10,430,000,000	
Artificial Intelligence	753,000,000	890,000,000	829,000,000	1,360,000,000	1,090,000,000	
Machine Learning	2,230,000,000	2,090,000,000	2,190,000,000	2,240,000,000	2,030,000,000	
Deep Learning	2,130,000,000	2,420,000,000	2,230,000,000	1,600,000,000	1,770,000,000	
Natural Language Processing	964,000,000	-	1,110,000,000	667,000,000	610,000,000	
Big Data	5,400,000,000	6,320,000,000	6,140,000,000	8,040,000,000	7,620,000,000	
Data Science	3,060,000,000	2,620,000,000	2,780,000,000	5,470,000,000	4,300,000,000	
Data Analytics	1,740,000,000	1,140,000,000	1,140,000,000	2,130,000,000	2,190,000,000	
Business Analytics	-	1,100,000,000	3,320,000,000	3,320,000,000	1,570,000,000	
Financial Technology	-	79,400,000	1,840,000,000	2,390,000,000	2,730,000,000	
OpenAI	-	-	-	-	268,000,000	
ChatGPT	-	-	-	-	796,000,000	
Language Modeling	-	-	-	-	1,770,000,000	
Large Language Models	-	-	-	-	2,830,000,000	

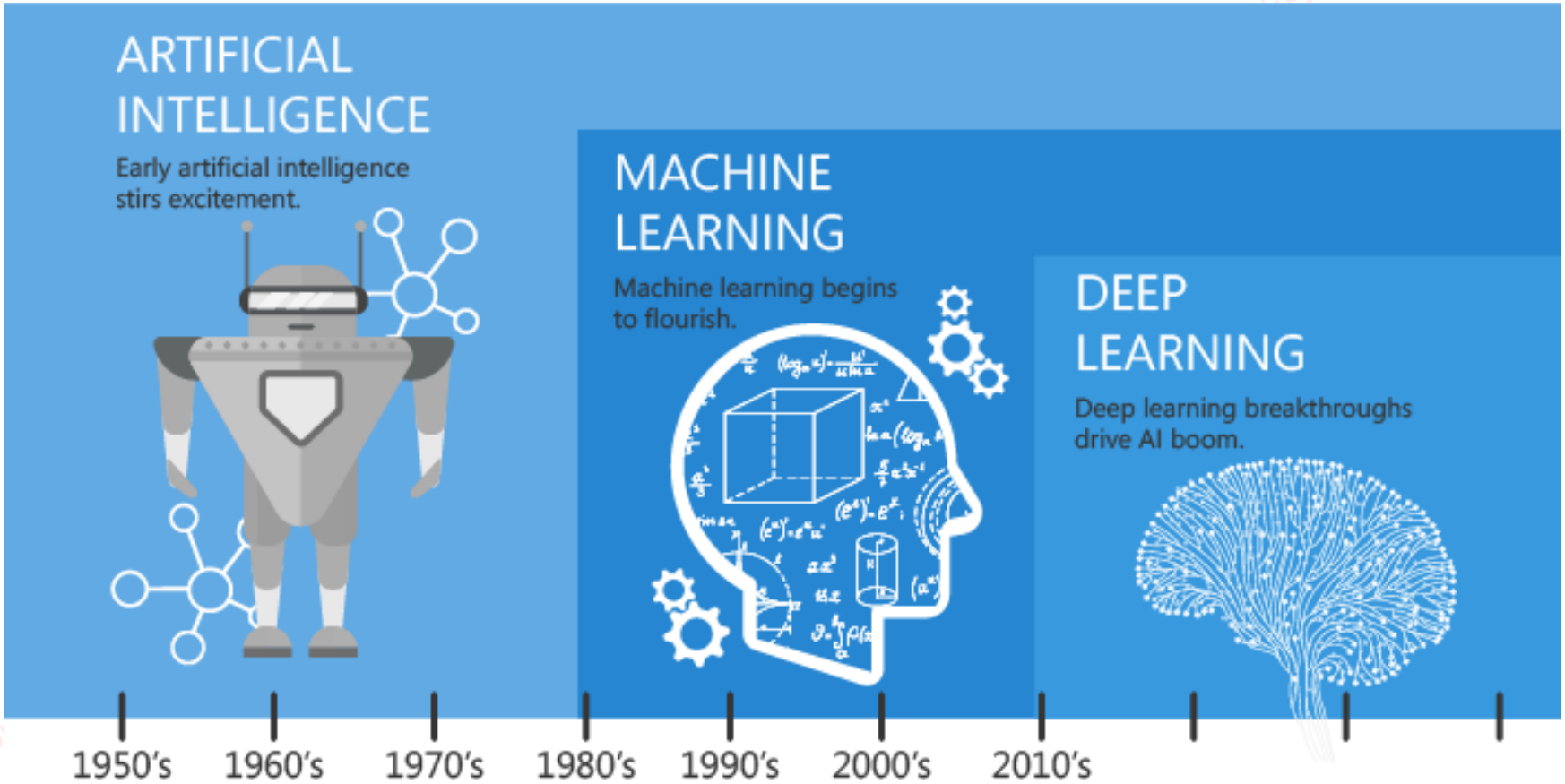
Number of pages in Google search?

	5/1/2021	9/7/2021	1/15/2022	9/6/2022	5/20/2023	8/29/2023
AI	3,160,000,000	3,200,000,000	3,740,000,000	8,910,000,000	10,430,000,000	19,470,000,000
Artificial Intelligence	753,000,000	890,000,000	829,000,000	1,360,000,000	1,090,000,000	1,570,000,000
Machine Learning	2,230,000,000	2,090,000,000	2,190,000,000	2,240,000,000	2,030,000,000	3,300,000,000
Deep Learning	2,130,000,000	2,420,000,000	2,230,000,000	1,600,000,000	1,770,000,000	4,730,000,000
Natural Language Processing	964,000,000	-	1,110,000,000	667,000,000	610,000,000	777,000,000
Big Data	5,400,000,000	6,320,000,000	6,140,000,000	8,040,000,000	7,620,000,000	8,280,000,000
Data Science	3,060,000,000	2,620,000,000	2,780,000,000	5,470,000,000	4,300,000,000	5,360,000,000
Data Analytics	1,740,000,000	1,140,000,000	1,140,000,000	2,130,000,000	2,190,000,000	3,110,000,000
Business Analytics	-	1,100,000,000	3,320,000,000	3,320,000,000	1,570,000,000	2,200,000,000
Financial Technology	-	79,400,000	1,840,000,000	2,390,000,000	2,730,000,000	2,880,000,000
OpenAI	-	-	-	-	268,000,000	271,000,000
ChatGPT	-	-	-	-	796,000,000	1,130,000,000
Language Modeling	-	-	-	-	1,770,000,000	7,270,000,000
Large Language Models	-	-	-	-	2,830,000,000	3,960,000,000

AI vs. ML vs. DL (Venn Diagram)

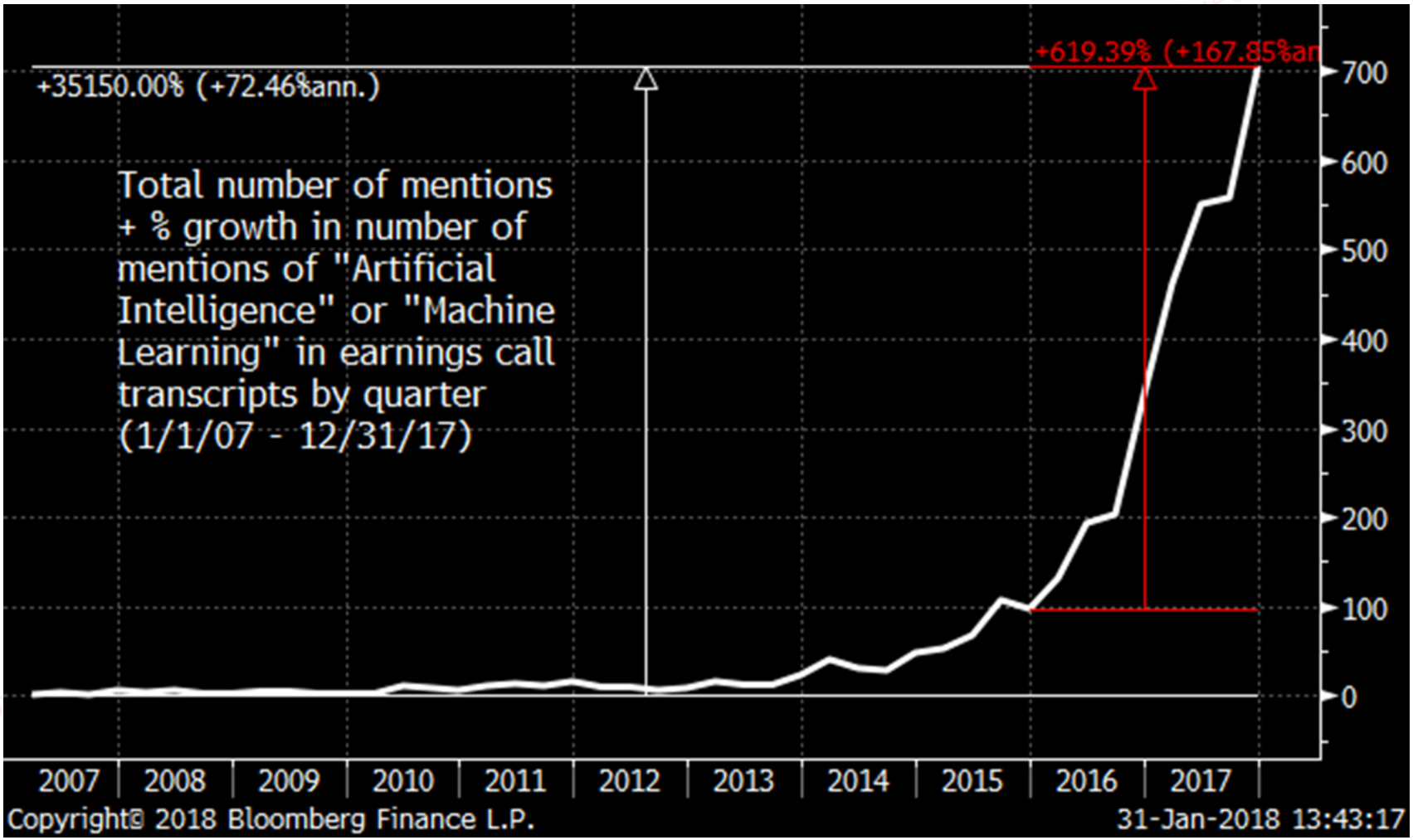


AI vs. ML vs. DL (Time Evolution)



Since an early flush of optimism in the 1950's, smaller subsets of artificial intelligence - first machine learning, then deep learning, a subset of machine learning - have created ever larger disruptions.

AI Revolution[1]



[1] source: Bloomberg

Impact of Big Data & Computational Power

- the success of `Machine Learning`, `Deep Learning`, and `LLMs` is more due to:

  (a) availability of data (`big data`)

  &

  (b) availability of more powerful computational engines like GPUs and more recently TPUs and NPUs

  &

  (c) and quantum computing in early future

Floating point operations per second (FLOPs)

just to give us some perspective

- 1952 – Estimating the entropy of English[2]: kiloFLOPs 10^3
- 1979 – Speech recognition: gigaFLOPs 10^9
- 1994 – Machine translation: teraFLOPs 10^{12}
- 2020 – Language understanding: yottaFLOPs 10^{24}
- 2025 – xxxGPT: quettaFLOPs 10^{30}

[2] the average number of bits per letter of the text required to translate the language into binary bits

Recipe for becoming an AI chef

- from the simplest problems such as regressions and binary classification to more advanced cases in machine learning, deep learning, or LLMs, we follow a generic recipe

- here are the steps in that recipe
  1. specification of dataset
  2. choice of an AI model/architecture
  3. choice of an objective/cost/loss function
  4. choice of an initial parameter set for the model
  5. choice of an optimization routine
  6. training & validation
  7. testing
  8. robustness against adversarial attacks

Key: Understanding your Data (1 of 2)

- AI & AI-based models are data driven

Without data, you're just another person with an opinion

William Edwards Deming

Key: Understanding your Data (2 of 2)

- should know your data, and always ask the following questions:
  1. stationary or non-stationary
     - unconditional joint probability distribution does not change when shifted in time
  2. parametric or non-parametric
     - distribution-free, do not rely on assumptions that data are drawn from a given parametric family (with a limited number of parameters)
  3. Markovian vs. non-Markovian
     - memoryless property, the conditional probability of a future state is only dependent on the present state (and is independent of any prior state)

Few Questions

Q: are markets `Markovian`[3]?

$$P(X_{t+1} \in \mathcal{A} | X_t, X_{t-1}, \ldots, X_{t-\ell}) \stackrel{?}{=} P(X_{t+1} \in \mathcal{A} | X_t)$$

[3] ask yourself about your choice for X_t

Few Questions

Q: are markets `Markovian`[3]?

$$P(X_{t+1} \in \mathcal{A} | X_t, X_{t-1}, \ldots, X_{t-\ell}) \stackrel{?}{=} P(X_{t+1} \in \mathcal{A} | X_t)$$

Q: a knee-jerk reaction (what is relevant?)

$$P(X_{t+1} \in \mathcal{A} | X_t, X_{t-1}, \ldots, X_{t-\tau}, X_{t-\tau-1} \ldots, X_{t-\ell})$$

[3] ask yourself about your choice for X_t

Few Questions

Q: are markets `Markovian`[3]?

$$P(X_{t+1} \in \mathcal{A} | X_t, X_{t-1}, \ldots, X_{t-\ell}) \stackrel{?}{=} P(X_{t+1} \in \mathcal{A} | X_t)$$

Q: a knee-jerk reaction (what is relevant?)

$$P(X_{t+1} \in \mathcal{A} | \boxed{X_t, X_{t-1}, \ldots, X_{t-\tau}}, X_{t-\tau-1} \ldots, X_{t-\ell})$$

Q: regime-switch (is anything relevant anymore?)

$$P(X_{t+1} \in \mathcal{A} | \boxed{X_t, X_{t-1}, \ldots, X_{t-\ell}})$$

`learning with rejection!`

[3] ask yourself about your choice for X_t

1. Specification of dataset (1 of 3)

i. `type` of dataset:

(a) tick data, time series, etc (e.g. signal extraction/data mining)

(b) market data & holding data (asset management)

(c) alternative data (e.g. data collected by `IoT`)

(d) categorical data

(e) text (e.g. classification)

(f) images (e.g. computer vision)

g) sound (e.g. voice recognition)

(h) video (e.g. motion detection)

& etc

1. Specification of dataset (2 of 3)

ii. collecting, generating, cleaning, (pre- & post-) processing, and/or choosing/picking (`relevant`) dataset

(a) collecting & cleaning data for training

(b) generating more data for training (real data are scarce)

(c) generating labels (supervised learning)

(d) proper data classification

(e) imputation for missing data

& more

Who wants clean data?

Who wants to clean the data?

1. Specification of dataset (3 of 3)

iii. dividing it into:

(a) `training` set

(b) `validation` set (to avoid over-fitting)

(c) `test` set

Quote of Overfitting[4]

With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

John von Neumann

[4] With enough parameters, i.e. a complex model, one can fit any data set.

Data & $\mathbb{AI}$

- data: main factor given models are data-driven

Q: working with data, is it $\mathbb{AI}$?

- ▷ cleaning
- ▷ pre- & post-processing
- ▷ classification
- ▷ curating
- ▷ generating & replicating
- ▷ imputation

& more

2. Choice of a model (1 of 2)

(a) classification models

- logistic regression
- decision trees
- random forest
- naive Bayes

(b) dimensionality reduction models

- `PCA` unsupervised technique used primarily for dimensionality reduction
- robust rolling `PCA` (`R2-PCA`)
- kernel `PCA`
- `ICA`
- autoencoder

(c) clustering methods used in unsupervised learning

- K-means
- robust rolling K-means (`R2K-means`)
- density-based spatial clustering of applications with noise (`DBSCAN`)
- `Gaussian` mixture model

2. Choice of a model (2 of 2)

(d) solving equations (explicit/implicit replication)
- mapping input data to labels via `FNNs` supervised
- using a neural network as a solution unsupervised

(e) image classification
- `CNNs`

(f) sequence analysis and `NLP` (sentiment analysis & more)
- `RNNs`
- `LSTMs`
- `GRUs`

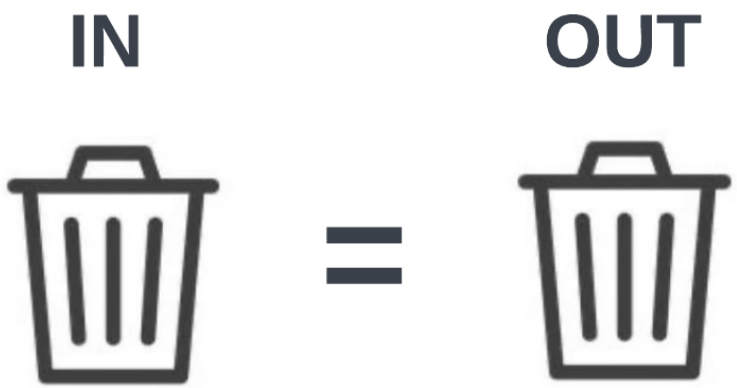
(g) `LLMs` (sentiment analysis, mathematical reasoning, & more)
- transformers

(h) sampling models (simulating/generating data preserving stylized facts)
- `MCMC` parametric
- `GANs` non-parametric

& more

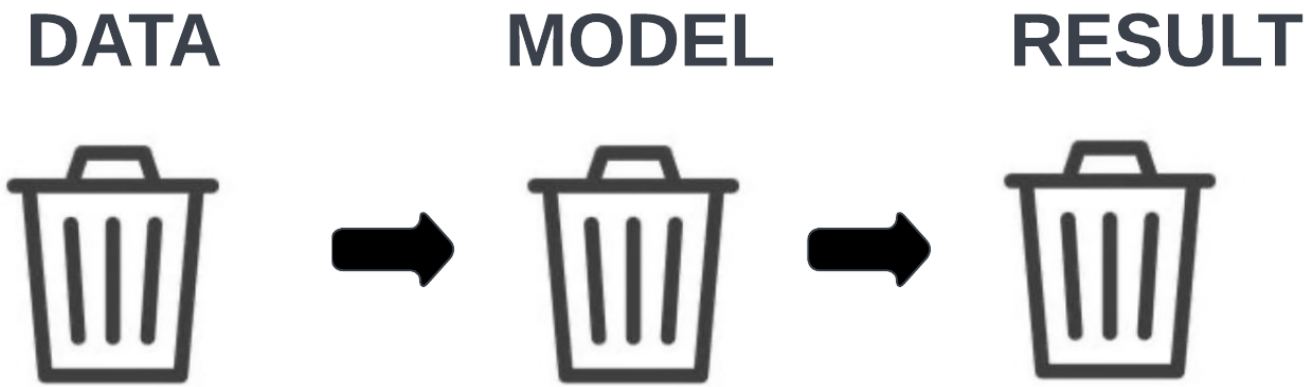
Data & Model (1 of 4)

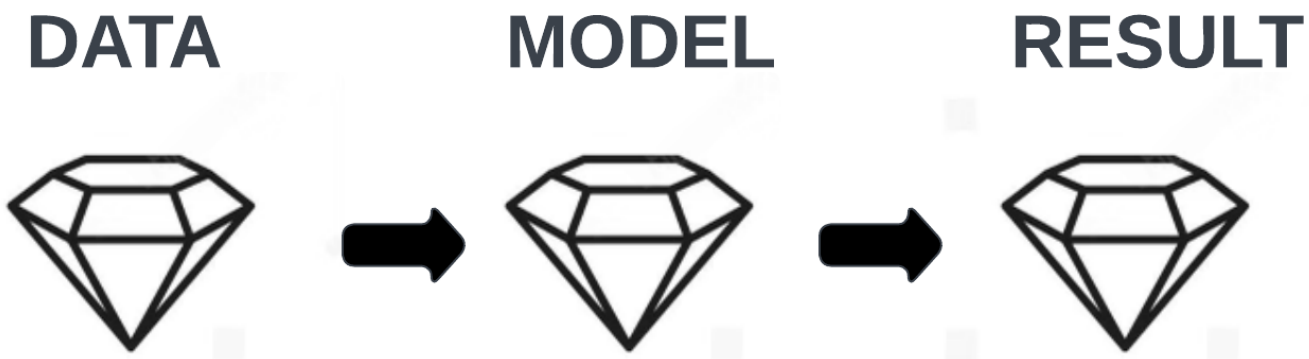
- `Garbage-in Garbage-out` (cannot learn from garbage)



Data & Model (2 of 4)

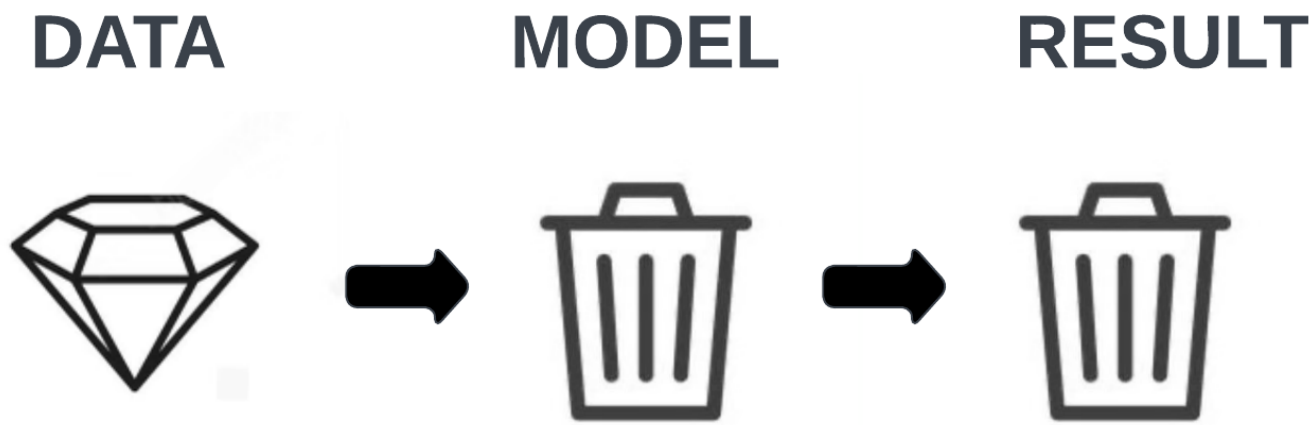
- it is obvious

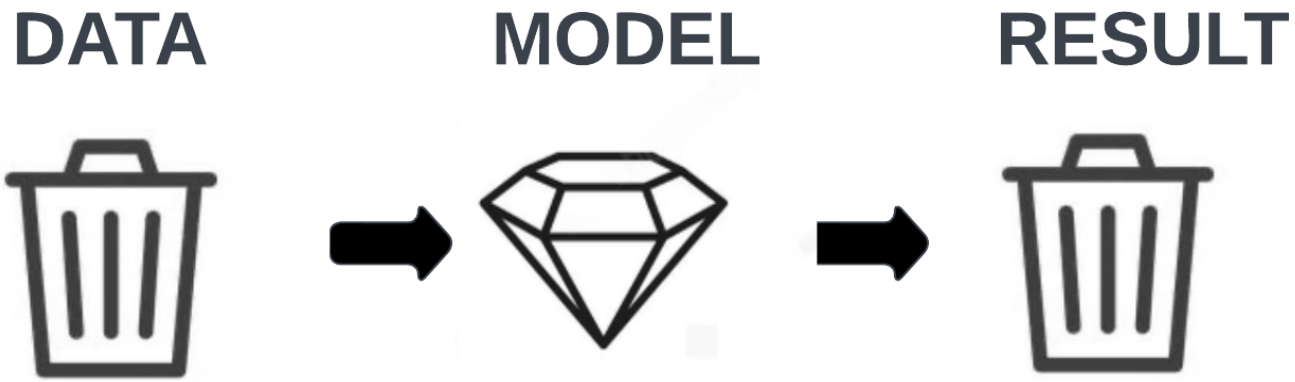




Data & Model (3 of 4)

- should be obvious, is it?





Data & Model (4 of 4)

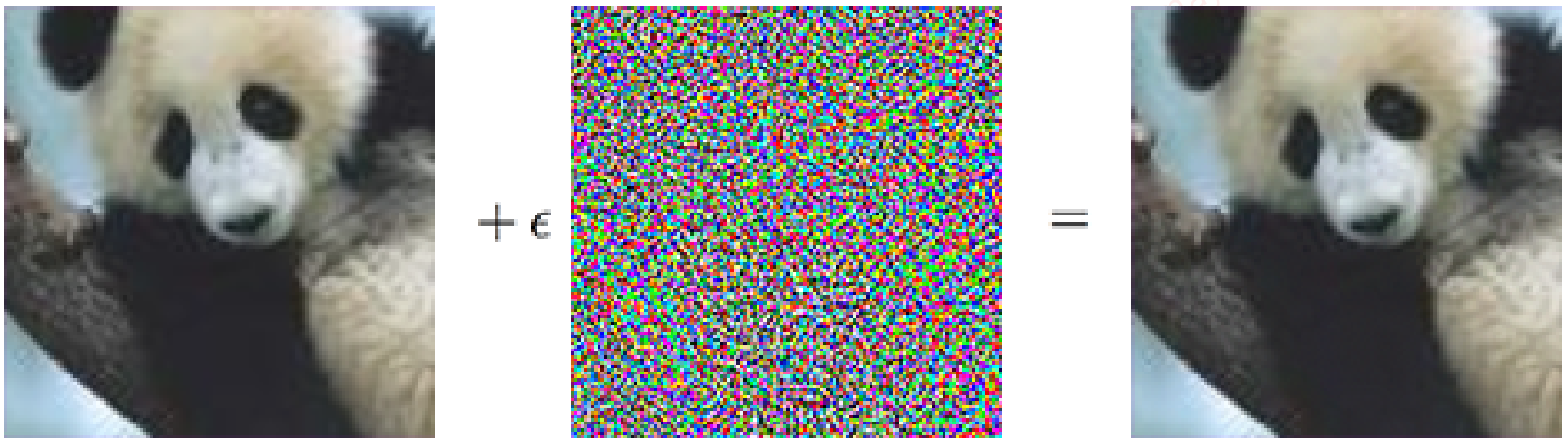
- try it on various simulated data and compare
- check *robustness* by adding noise to original data
- consistency between in-sample vs. out-sample

Robustness and Explainability of AI

few questions on AI-based models:

- Q: can it be understood by humans or simply a **black box**?
- Q: if the model does not do well, do we know why?
- Q: can you understand or follow model prediction/decision?
- Q: when do we know a trained model has failed?
- Q: how robust is the model against the adversarial attack?

Attacking Deep Learning w/ Adversarial Examples[5]



"panda"
57.7% confidence

"gibbon"
99.3% confidence

[5] classic example

Testing against Adversarial Attacks

intentionally left blank

3. Choice of objective/cost/loss function (1 of 2)

- mean squared error (L_2 loss)

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

- absolute error (L_1 Loss)

$$\frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

- cross entropy loss

$$-\frac{1}{n}\sum_{j=1}^{n} y_j \log(\hat{y}_j)$$

3. Choice of objective/cost/loss function (2 of 2)

intentionally left blank

Regularization (exogenous vs endogenous)

- add a penalty term to the objective function, but why?

$$\mathcal{L}(\Theta) + \lambda R(\Theta)$$

- for example

$$\mathcal{L}(\Theta) + \lambda \|\Theta\|_2^2$$

4. Choice of initial parameter set (1 of 7)

Q: how important is it to start from a good starting point?

A: in many cases it could be very crucial

- will demonstrate this during visual introduction to optimization

4. Choice of initial parameter set (2 of 7)

Q: how to find a good starting point?

A: there are lots of literature about the topic

- it depends on the problem and the model
- here are few of them:
  - zero initialization – have most zero but few, show know the problem and the model really well
  - random initialization
  - He et al. initialization[6]
  - Xavier initialization – tries to initialize weights with a smaller value such that neurons will not start training in saturation

[6] *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification* by He et al (2015)

4. Choice of initial parameter set (3 of 7)

Q: how does the loss surface look like?

A: impossible to know due to high dimensionality

- a naive way to visualize

$$\Theta = \alpha\Theta_1 + (1 - \alpha)\Theta_2$$

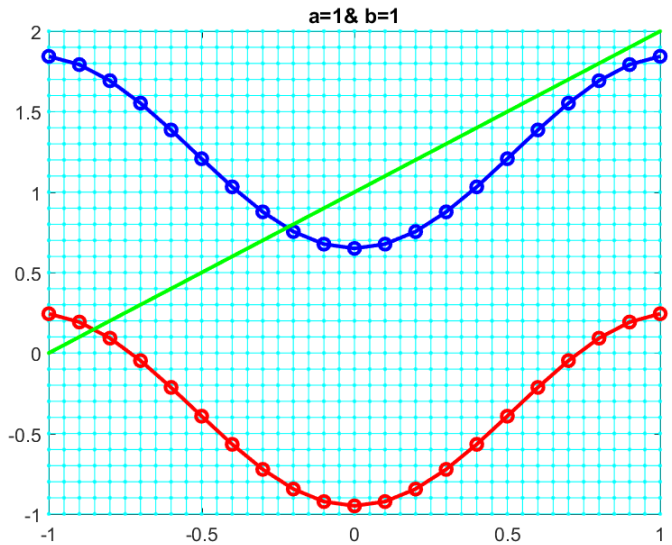
where Θ_1 and Θ_2 are two arbitrary parameter sets

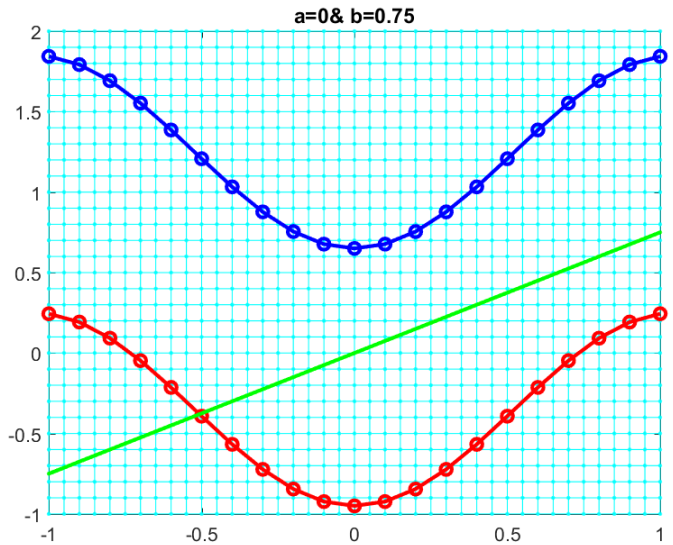
- could be viewed as a function of α as opposed to Θ

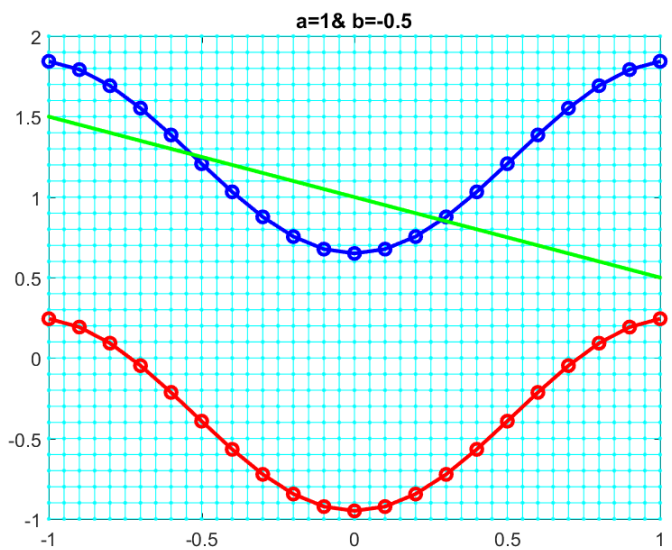
4. Choice of initial parameter set (4 of 7)

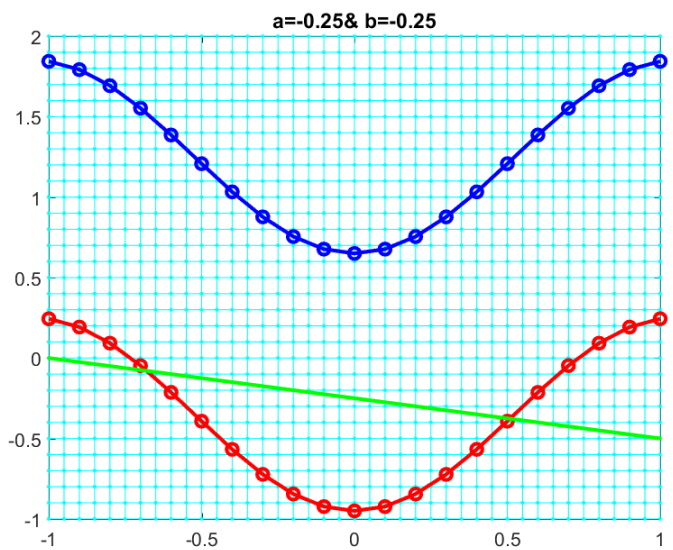
intentionally left blank

4. Choice of an initial parameter set (5 of 7)

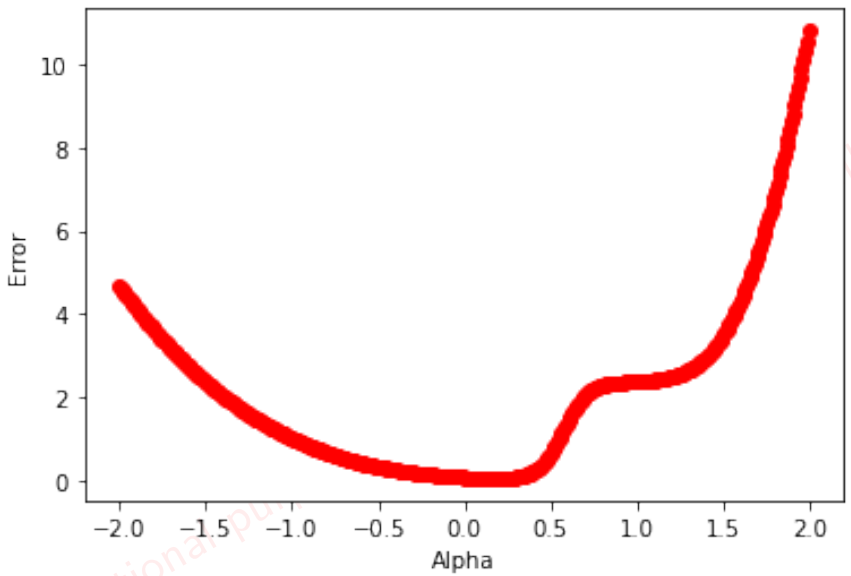


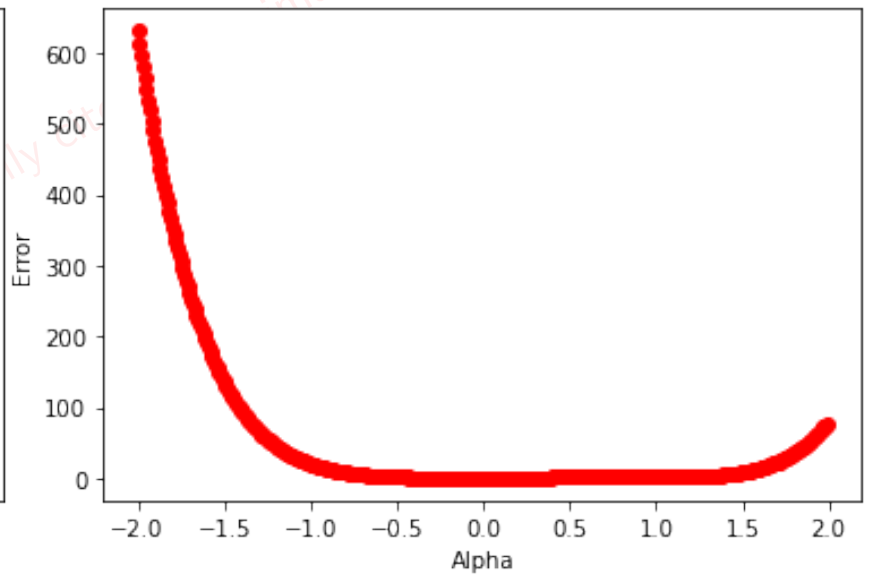






4. Choice of initial parameter set (6 of 7)





4. Choice of initial parameter set (7 of 7)

intentionally left blank

5. Optimization routines

there are three major ones:

- brute-force (grid) search
- gradient-free routines
- gradient-based routines

will discuss in depth in visual introduction to optimization

Machine Learning versus Optimization

- `optimization` is the central task of `machine learning`
- learning is an optimization problem