Optimization for Data Science
Lecture 00: Introduction

Kimon Fountoulakis

School of Computer Science
University of Waterloo

05/09/2019

Outline

- Introduction to applications that use optimization models and algorithms.
- Course logistics.

Instructor

Assistant Professor

WATERLOO.AI
WATERLOO ARTIFICIAL INTELLIGENCE INSTITUTE

Kimon Fountoulakis

Scientific Computation Group
University of Waterloo

8 years of experience
in data science and numerical optimization

PhD in numerical optimization from the School of Mathematics at the University of Edinburgh

Postdoctoral researcher for 3 years at the Dep. of Statistics at the University of California Berkeley

What is optimization?

- In its simplest form optimization is about finding the minimum (or maximum) of a given function.

$$\text{minimize } f(x)$$

- Optimization (decision) variables

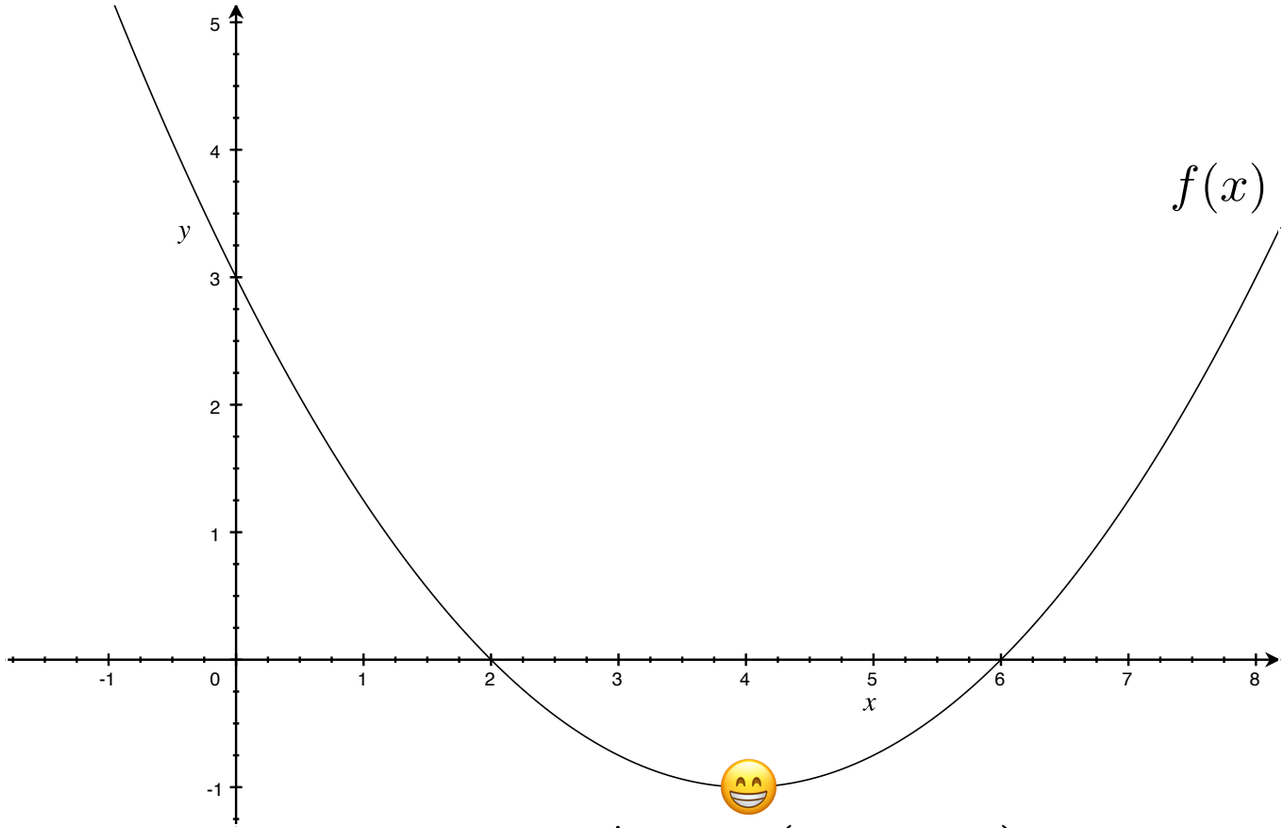
$$x = (x_1, x_2, \ldots, x_n)$$

- Objective function

$$f(x) : \mathbb{R}^n \to \mathbb{R}$$

Example

- In its simplest form optimization is about finding the minimum (😁) of a given function.



$$f(x) := (0.5x - 2)^2 - 1$$

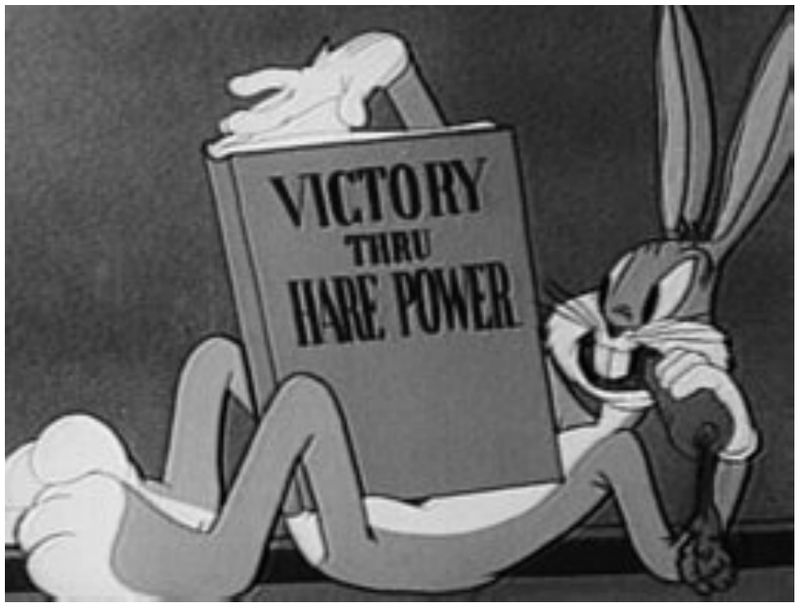
$$x^* = (4, -1)$$

Real Examples/Models in Data Science

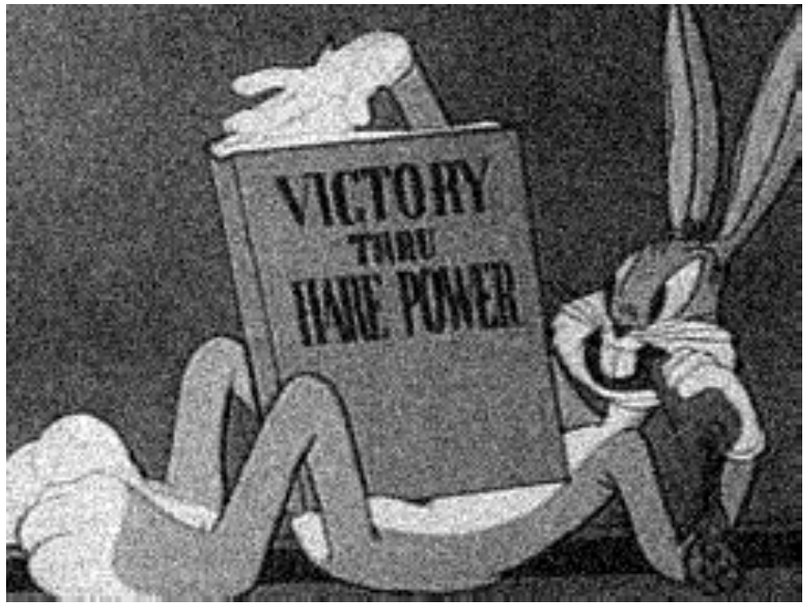
- Image denoising
- Signal processing
- Magnetic Resonance Image Reconstruction
- Single Pixel Camera
- Sound reconstruction
- Classification of data
- Face recognition
- Recomenders systems
- Clustering of facebook and twitter accounts
- Finding galaxies
- Finding segments in images
- Artificial Intelligence problems

Image Denoising

Original Image



Noisy Image



Denoised Image

Image Denoising

$$\text{minimize } \lambda\|\nabla x\|_1 + \|x - y\|_2^2$$

Regularizes the output

Measures distance to noisy input

We want “x” (denoised image) to be close to “y” (noisy image), but not identical.

Image Denoising: people

Stanley Osher

Physica D: Nonlinear Phenomena
Volume 60, Issues 1–4, 1 November 1992, Pages 259-268

Nonlinear total variation based noise removal algorithms ☆

Leonid I. Rudin [2], Stanley Osher, Emad Fatemi [2]

⊞ Show more

https://doi.org/10.1016/0167-2789(92)90242-F

https://www.sciencedirect.com/science/article/pii/016727899290242F

13500! citations on Google Scholar

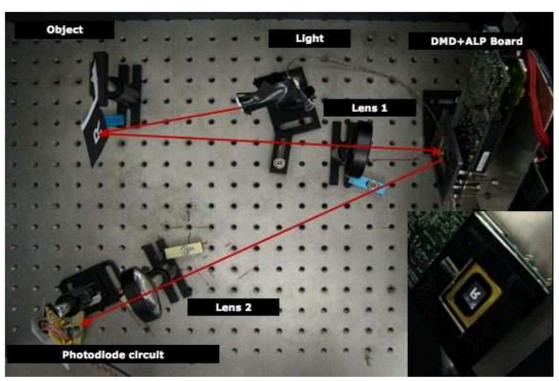
Image Reconstruction: single pixel camera

- Usual process of taking photos: given a camera with 10 megapixel sensor, we take 10 million samples to represent an image.

- Why would we take 10 million samples?? Most images have redundant information in them, i.e., blue sky background.

- Is it possible to take a small fraction of the samples and still be able to obtain a high resolution representation of an object?

- A single pixel camera does this!



https://www.ece.rice.edu/~duarte/images/csCamera-SPMag-web.pdf

Image Reconstruction: single pixel camera

Image Reconstruction: single pixel camera

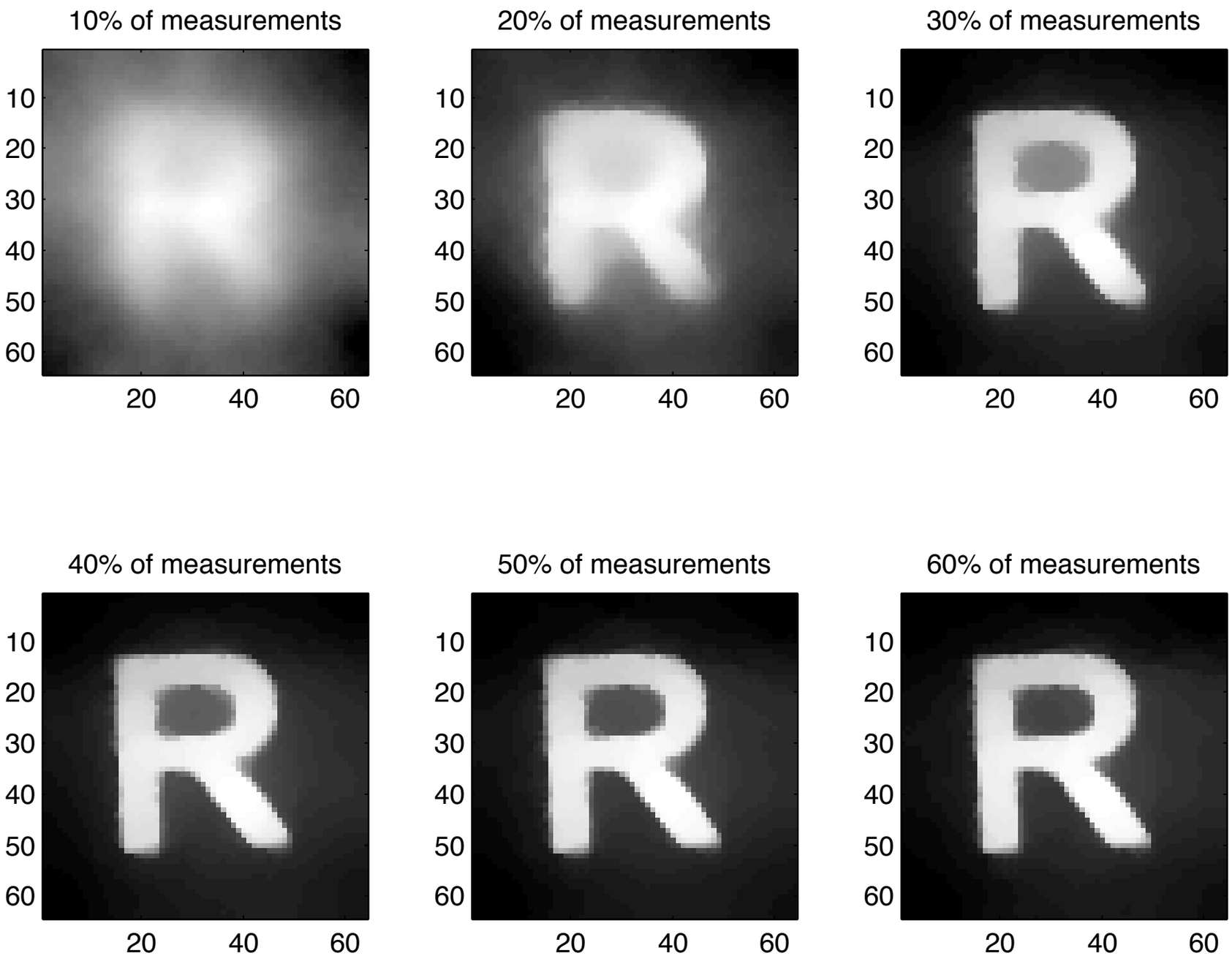


Image Reconstruction: single pixel camera

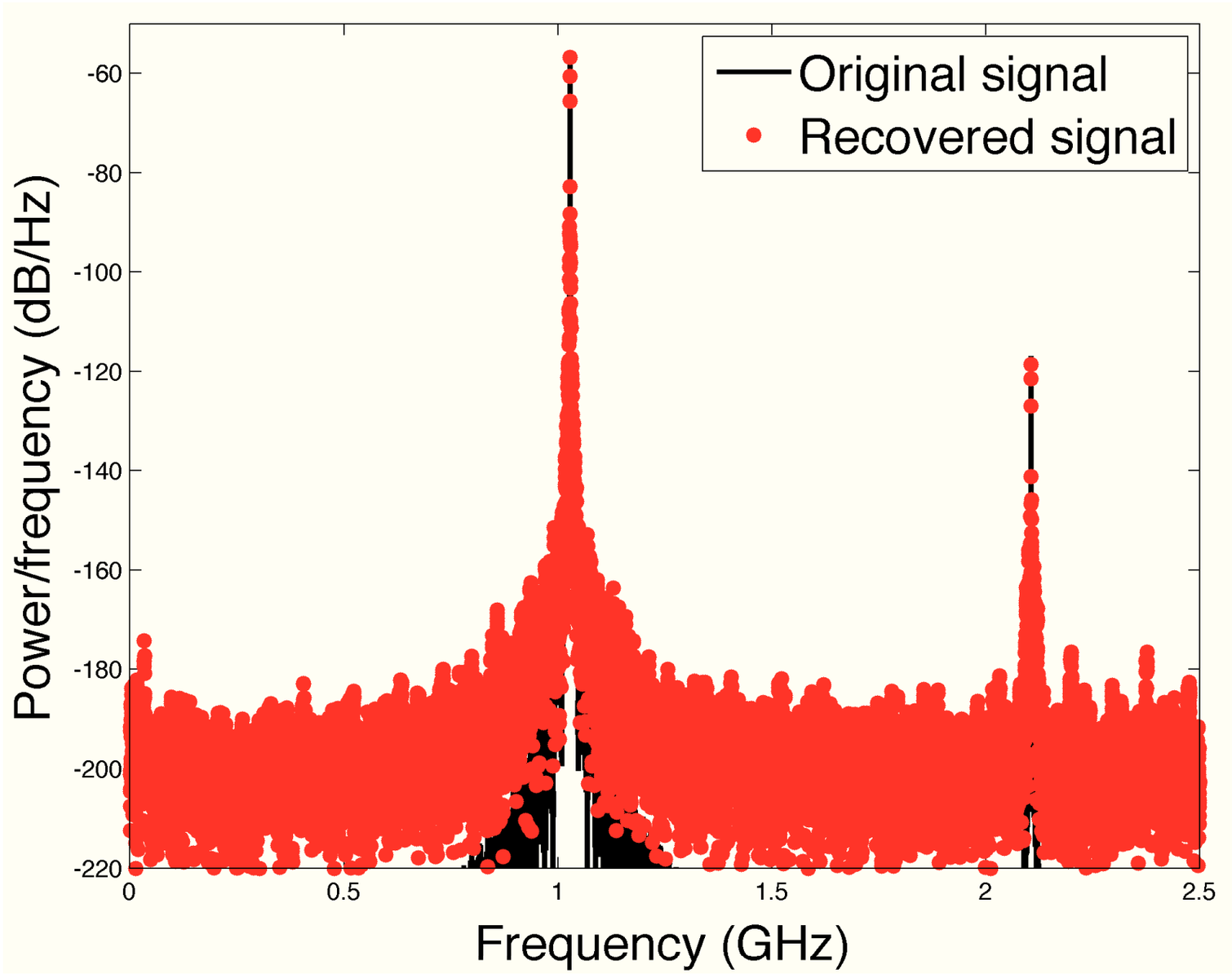
$$\text{minimize } \lambda\|Dx\|_1 + \frac{1}{2}\|Ax - b\|_2$$

Sparsity operator

Decision variables: the reconstructed image

Input under-sampled data

Signal Processing: Reconstruction of two radio-frequency radar tones



We only use a fraction of samples of the original signal

Sound reconstruction

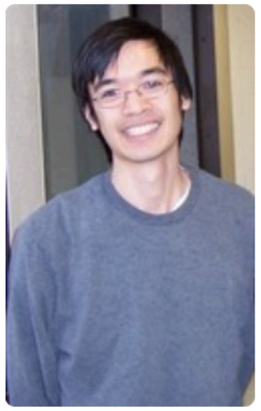
Noisy Handel Hallelujah

Denoised Handel Hallelujah

Signal reconstruction: people

Terence Tao

math.ucla.edu/~tao



Terence Chi-Shen Tao is an Australian-American mathematician who has worked in various areas of mathematics. He currently focuses on harmonic analysis, partial differential equations, algebraic combinatorics, arithmetic combinatorics, geometric combinatorics, compressed sensing and analytic number theory., he holds the James and Carol Collins chair in mathematics at the University of California, Los Angeles. Tao was a co-recipient of the 2006 Fields Medal and the 2014 Breakthrough Prize in Mathematics. **Wikipedia**

Born: 17, 1975, Adelaide, South Australia

Residence: Los Angeles, California

Nationality: Australia, United States

David Donoho

David Leigh Donoho is a professor of statistics at Stanford University, where he is also the Anne T. and Robert M. Bass Professor in the Humanities and Sciences. **Wikipedia**

Born: Mar 5, 1957, Los Angeles, CA, United States

Nationality: American

Fields: Mathematics

Emmanuel Candès

Emmanuel Jean Candès is a professor of mathematics, statistics, and electrical engineering at Stanford University, where he is also the Barnum-Simons Chair in Mathematics and Statistics. **Wikipedia**

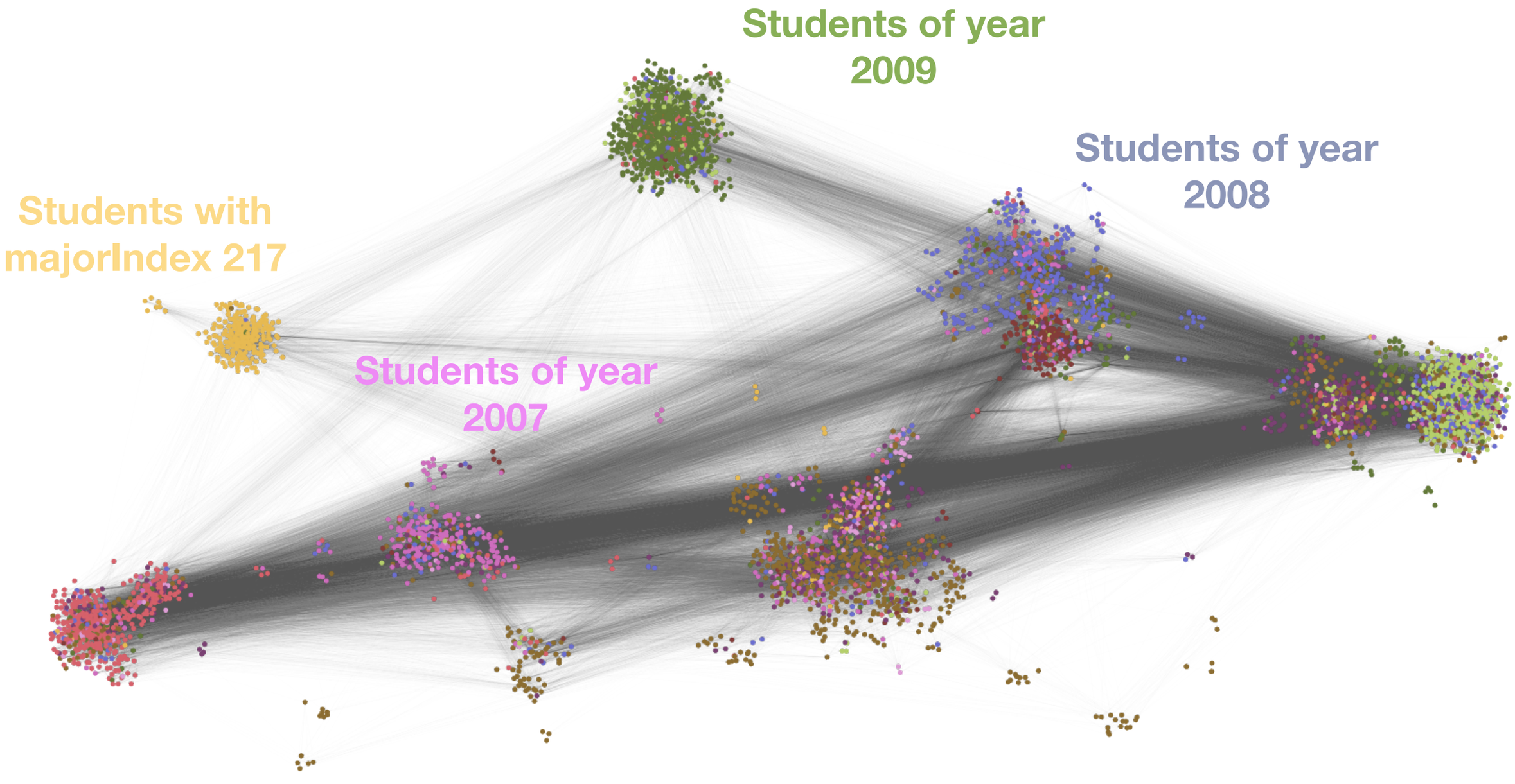
Born: 27, 1970, Paris, France

Nationality: French

Fields: Statistician, Mathematician

Multiple awards for their work on compressed sensing

Community detection in Facebook



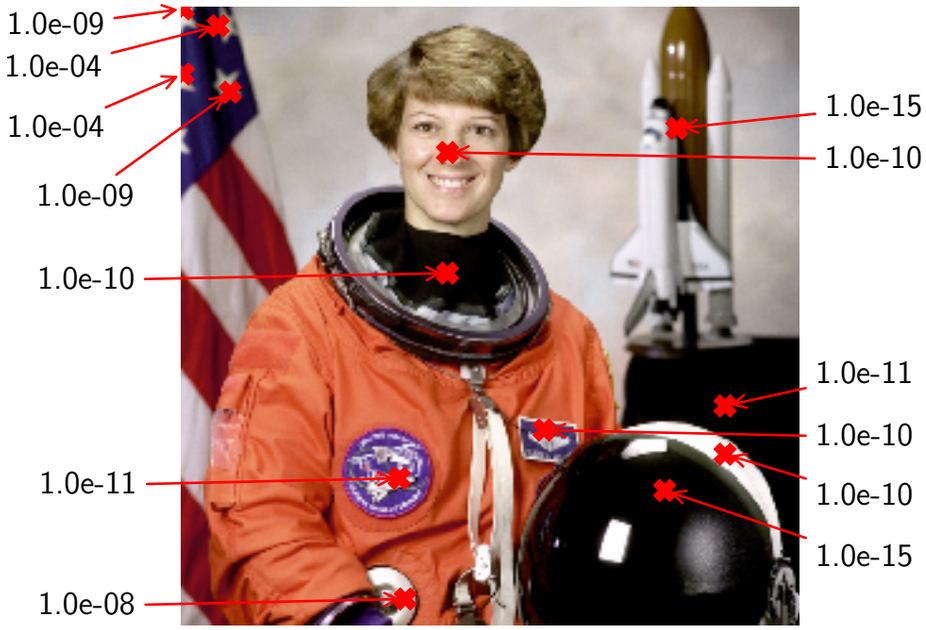
Data: Facebook Johns Hopkins, A. L. Traud, P. J. Mucha and M. A. Porter, Physica A, 391(16), 2012

Finding communities/ clusters

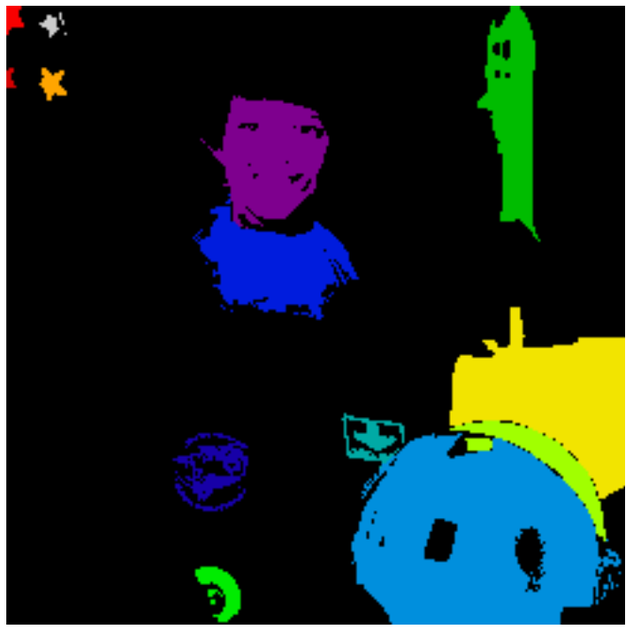
$$\text{minimize } \underbrace{\frac{1}{2}x^T Qx - \alpha x^T s}_{f(x)} + \underbrace{\rho\alpha\|Dx\|_1}_{g(x)}$$

- $f(x)$ represents the PageRank model
- $g(x)$ represents the penalty on large probabilities

Finding Segments in Images



Input image and target clusters



Output segments

Boundaries of output segments

Finding Segments in Images

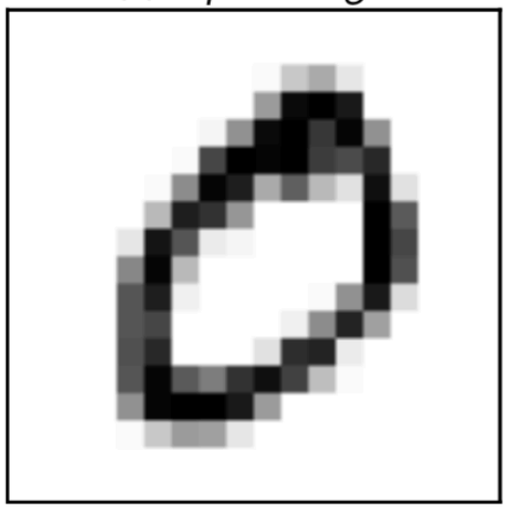
- Convert an image into a graph by considering the pixels as nodes and edges among nodes represent the pixel similarity, i.e., color and distance similarity.

- Regions of an image with similar structure should be represented as clusters of nodes in the graph with high internal connectivity and low external connectivity.

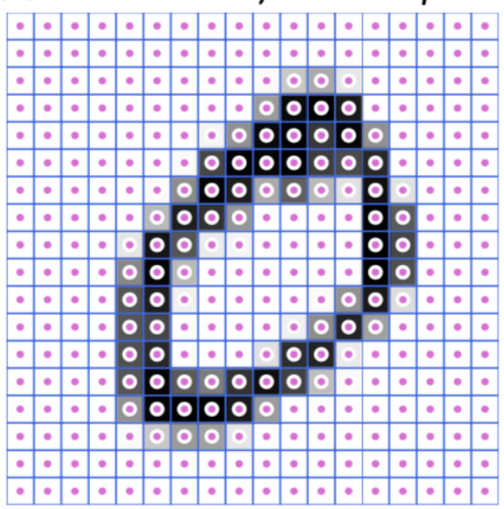
- Find clusters in the graph (segments in image) using clustering algorithms such as PageRank.

Finding Segments in Images

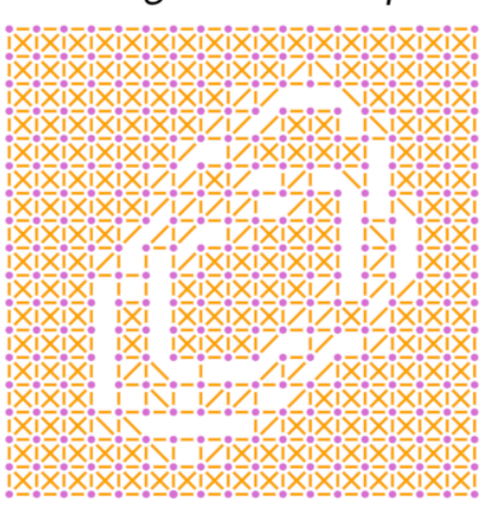
(a) Input Image

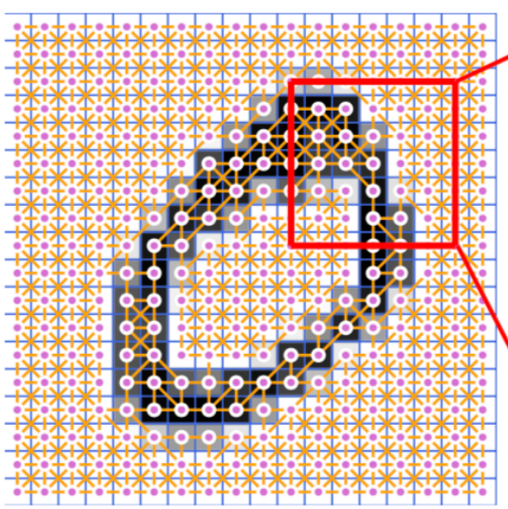


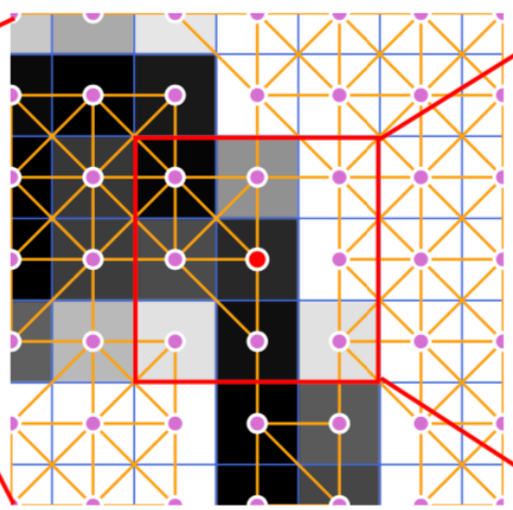
(b) Add nodes for each pixel

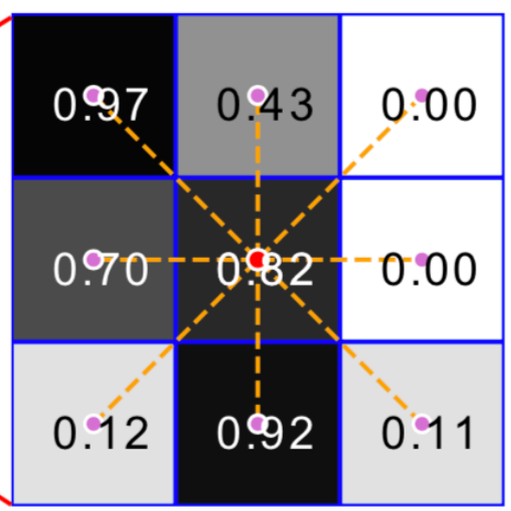


(c) Add edges between pixels









Finding Segments in Images

(d) Target cluster

(e) Input

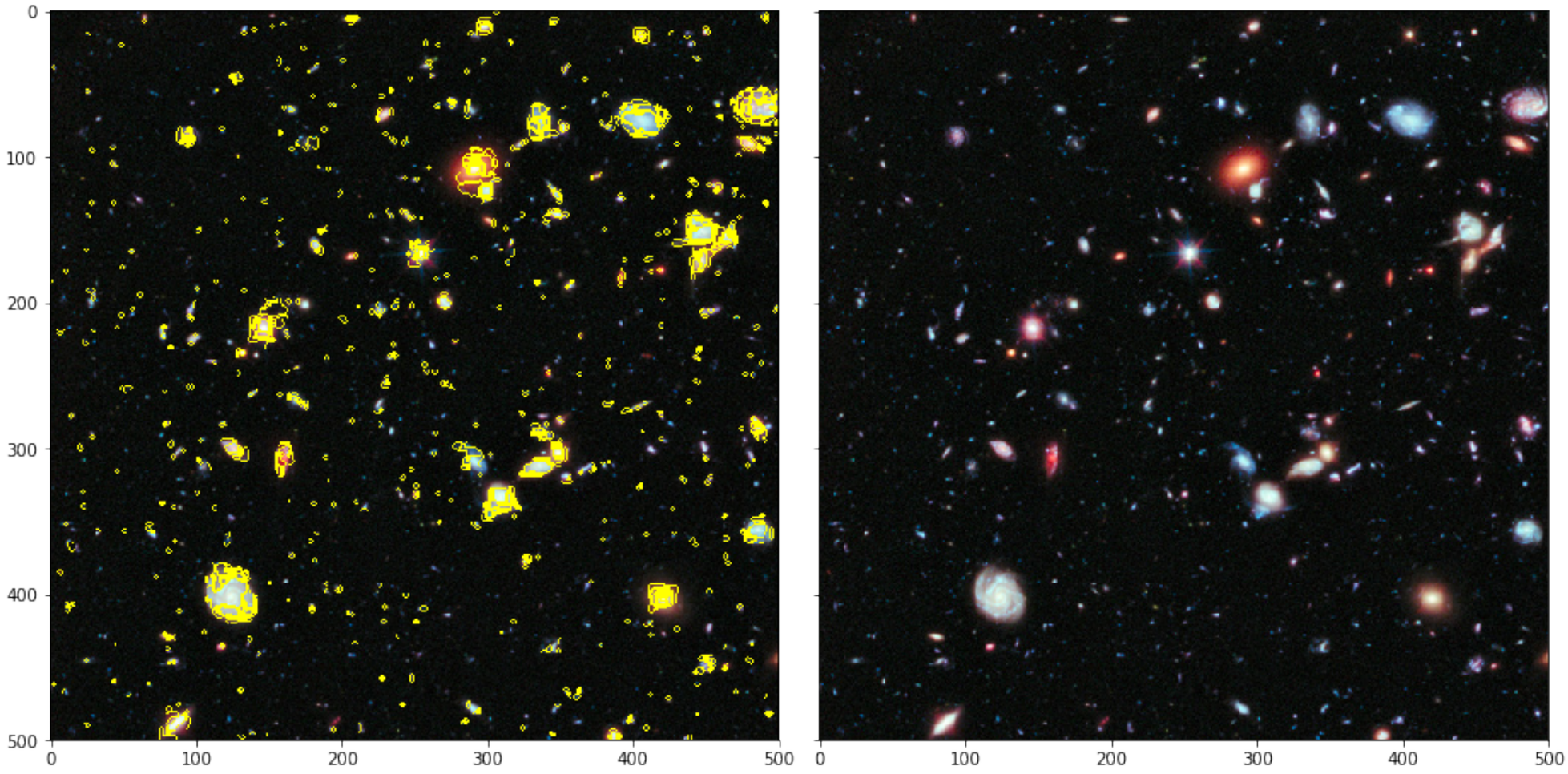
(f) SimpleLocal

(g) Target cluster

(h) Input

(i) SimpleLocal

Finding Galaxies



Clustering: people

PageRank

From Wikipedia, the free encyclopedia

"Google search algorithm" redirects here. For other search algorithms used by Google, see Google Penguin, Google Panda, and Google Hummingbird.

PageRank (**PR**) is an algorithm used by Google Search to rank web pages in their search engine results. PageRank was named after Larry Page,[1] one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:

Larry Page

Lawrence Edward Page is an American computer scientist and Internet entrepreneur who co-founded Google with Sergey Brin. **Wikipedia**

Born: Lawrence Edward Page, 26, 1973, East Lansing, Michigan, U.S.

Residence: Palo Alto, California, U.S.

Citizenship: United States

PageRank was invented for ranking web pages, but it is also very efficient as a clustering algorithm.

Clustering: people

- I have spent about 3 years working on clustering related problems.

- Currently we have the algorithm with the best worst-case guarantees and running time for the problem of local graph clustering.

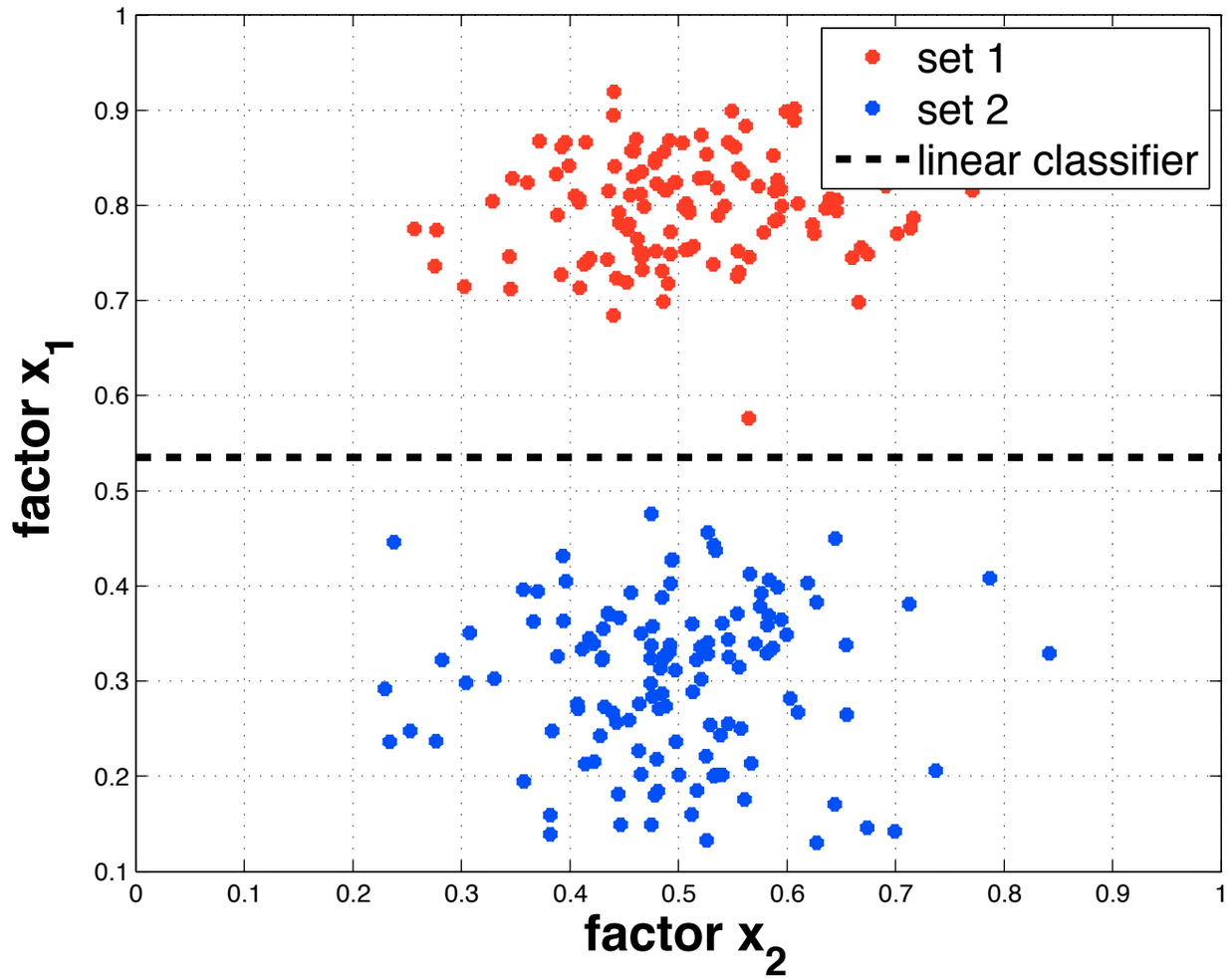
Capacity releasing diffusion for speed and locality
D Wang, K Fountoulakis, M Henzinger, MW Mahoney, S Rao
Proceedings of the 34th International Conference on Machine Learning-Volume …

GitHub

Software

kfoynt / **LocalGraphClustering**

Classification



Classification

- Given data points $(a_i, b_i)\ \forall i = 1, \ldots, n$
- For example “a” is a d-dimensional vector (d=2 in our previous example) and “b” is 1 or -1. These are data that are classified into two groups.
- We want to “train” a function such as

$$f(a; x) = x^T a$$

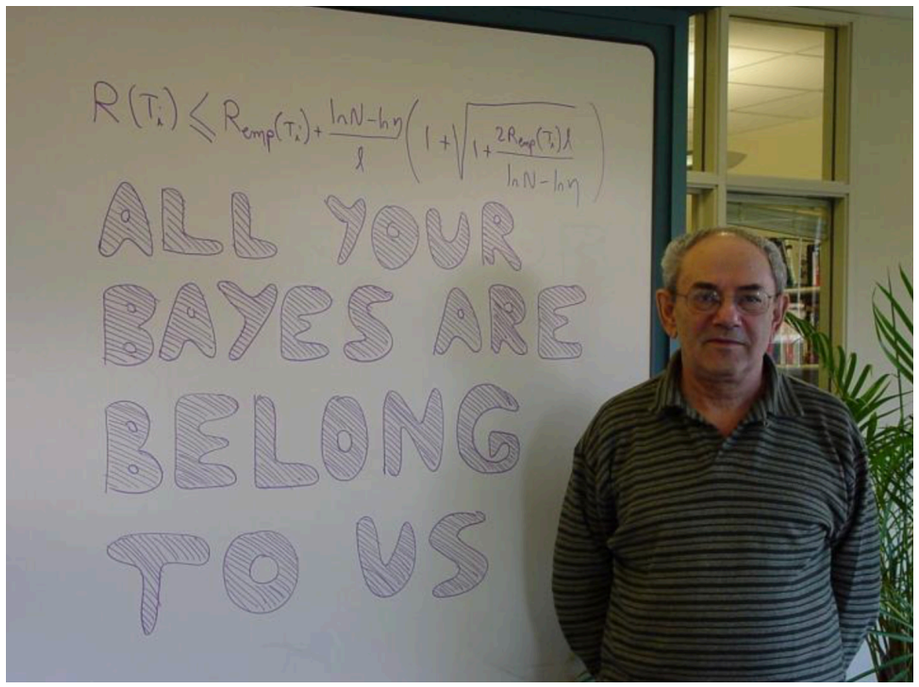
- Meaning that we want to find the value of each component in vector “x”, such that when we get a new data point we can use this function to predict the class of the new data point.

Classification

- Training is often achieved by logistic regression, which is formulated as an optimization problem.

$$\text{minimize } \|x\|_1 + \sum_{i=1}^{m} \log(1 + e^{b_i x^T a_i})$$

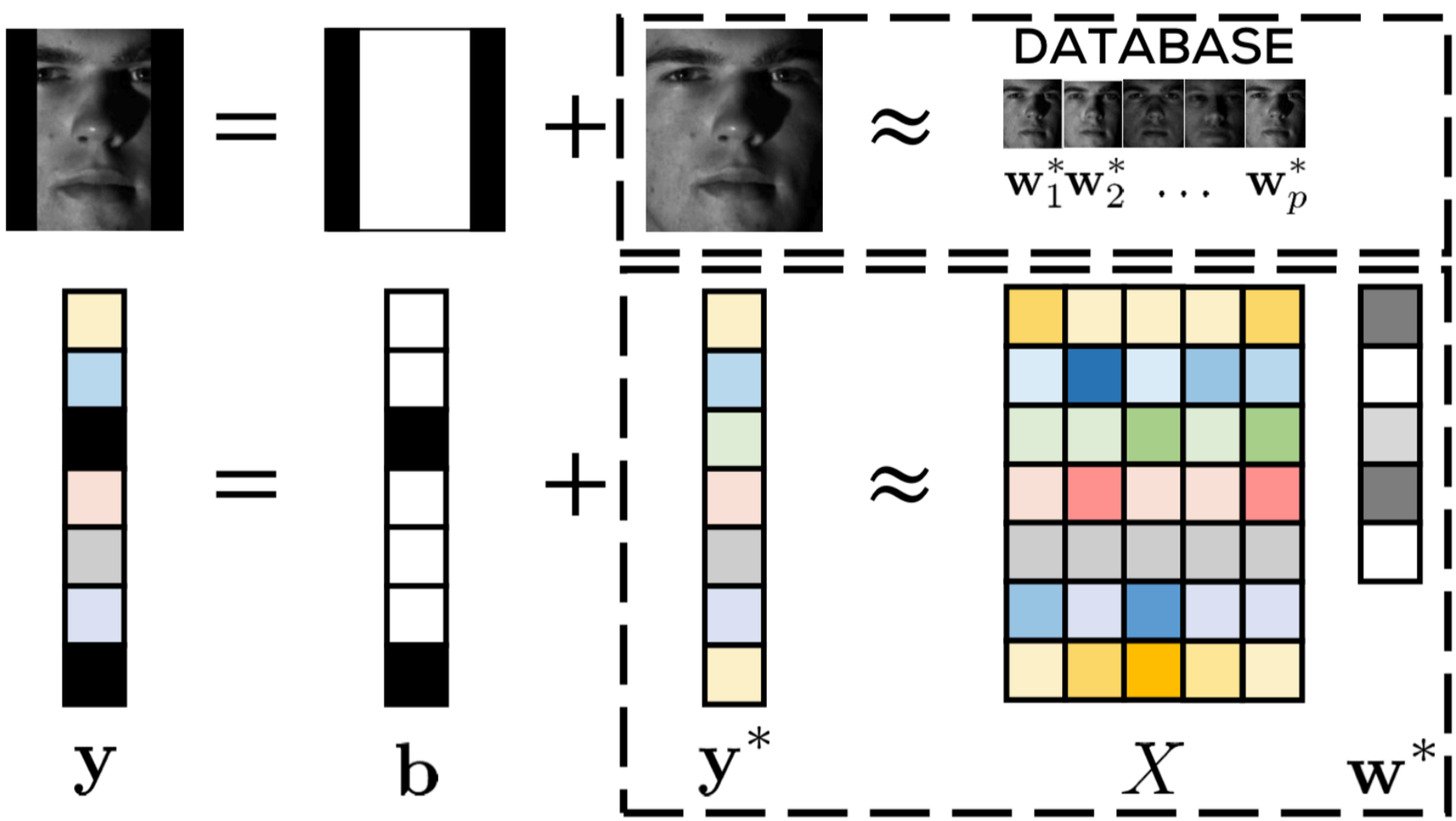
- Or the Support Vector Machine model by Vladimir Vapnik



Cited by		VIEW ALL
	All	Since 2014
Citations	184988	74993
h-index	77	54
i10-index	199	143

Biometrics: Face Reconstruction

- Reconstruct faces that have been randomly or maliciously corrupted.



Source: https://arxiv.org/pdf/1712.07897.pdf

Biometrics: Face Reconstruction

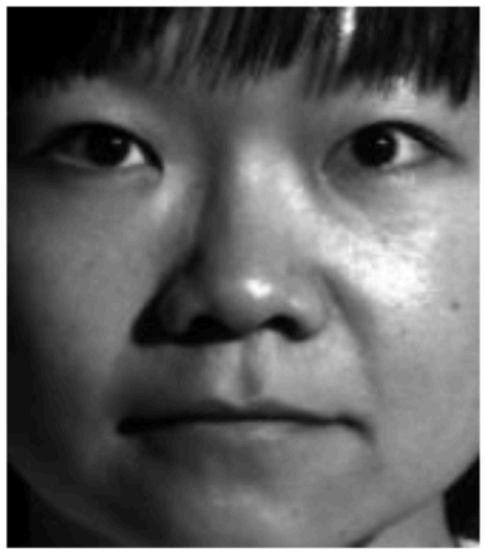
- Reconstruct faces that have been randomly or maliciously corrupted.
- This problem is formulated as an optimization problem

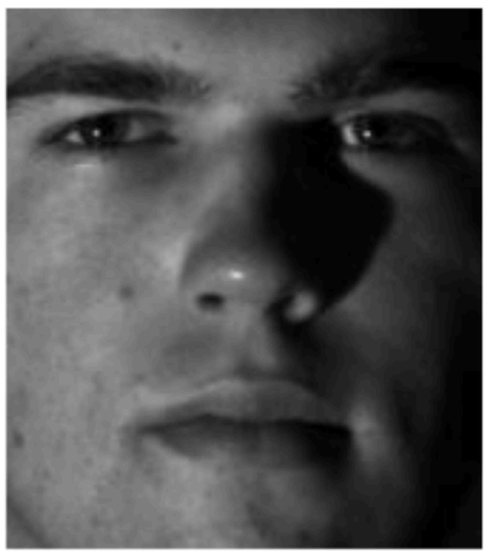
$$\min_{\substack{\mathbf{w}\in\mathbb{R}^p,\mathbf{b}\in\mathbb{R}^n \\ \|\mathbf{b}\|_0\leq k}} \|\mathbf{y} - X\mathbf{w} - \mathbf{b}\|_2^2$$

Source: https://arxiv.org/pdf/1712.07897.pdf

Face Reconstruction

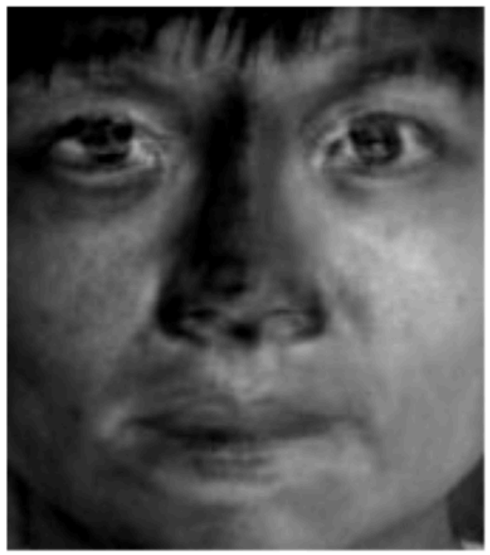
ORIGINAL

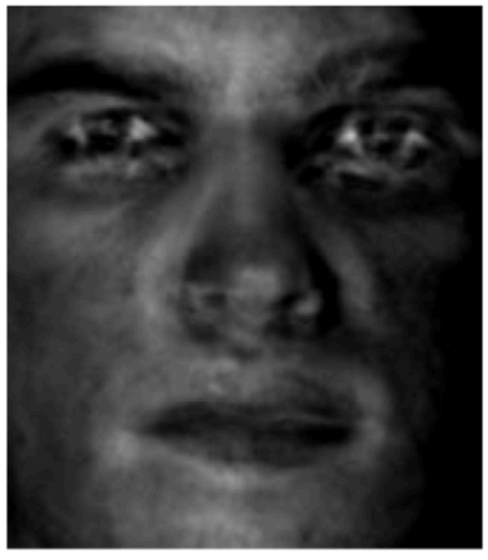




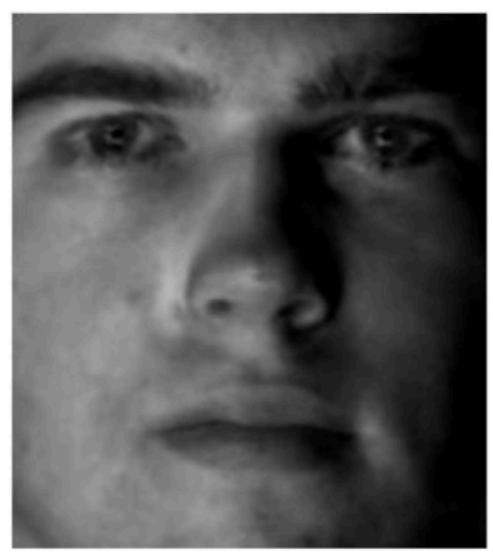
DISTORTED

OLS





AM-RR



Recommenders Systems

- Given a table of 10000 movies and 1 million viewers
- We assume that each viewer has watched some tiny fraction of the moves and rated them.

	Movies		
	10	6	
Viewer ID			8
	9		5
			2

Recommenders Systems

- Our objective is to find another table, which we use to predict unknown ratings by the viewers.

Movies

Viewer ID			
	10	6	
			8
	9		5
			2

Input

Movies

Viewer ID			
	10	6	**7**
	8	**5**	8
	9	**4**	5
	9	**6**	2

Output

Recommenders Systems

- We can use the predicted ratings to make recommendations

Movies

Viewer ID			
	10	6	
			8
	9		5
			2

Input

Movies

Viewer ID			
	10	6	**7**
	8	**5**	8
	9	**4**	5
	9	**6**	2

Output

Recommenders Systems

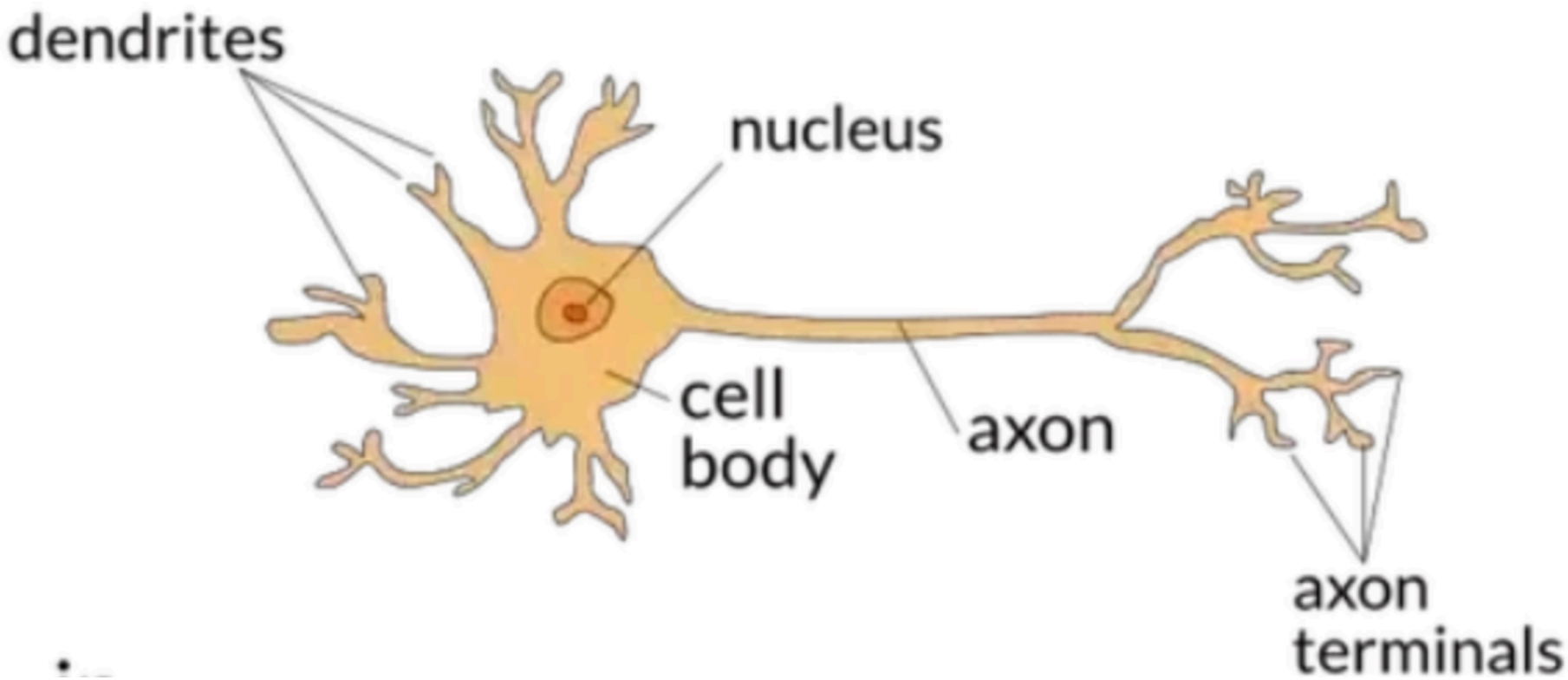
- This is formulated as an optimization problem

$$\text{minimize } \|X\|_*$$

$$\text{subject to: } \sum_{ij}(X_{ij} - A_{ij}) \leq \epsilon$$

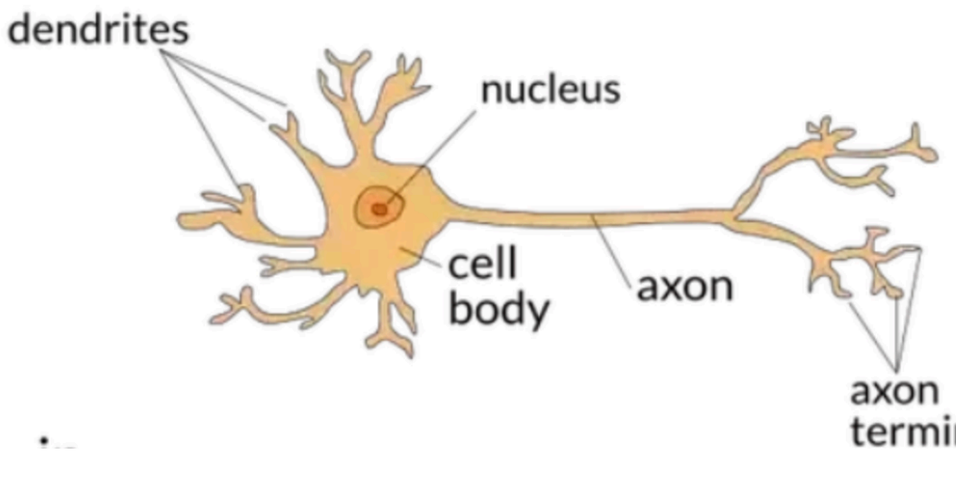
- “X” is the table that we are looking for
- “A” is the input table
- The above problems says that we want a table “X” that is close to table “A” and also has small norm.

Physical Neurons

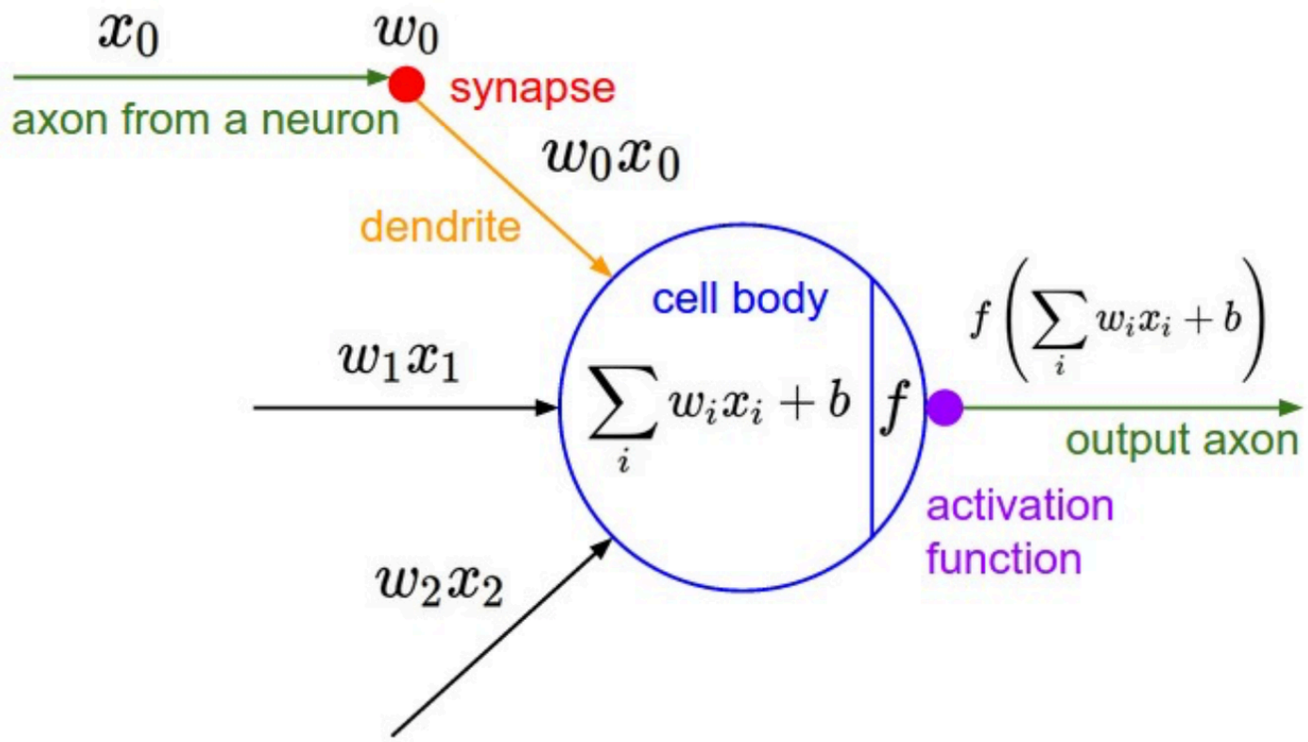


source

Artificial Neurons- Perceptrons



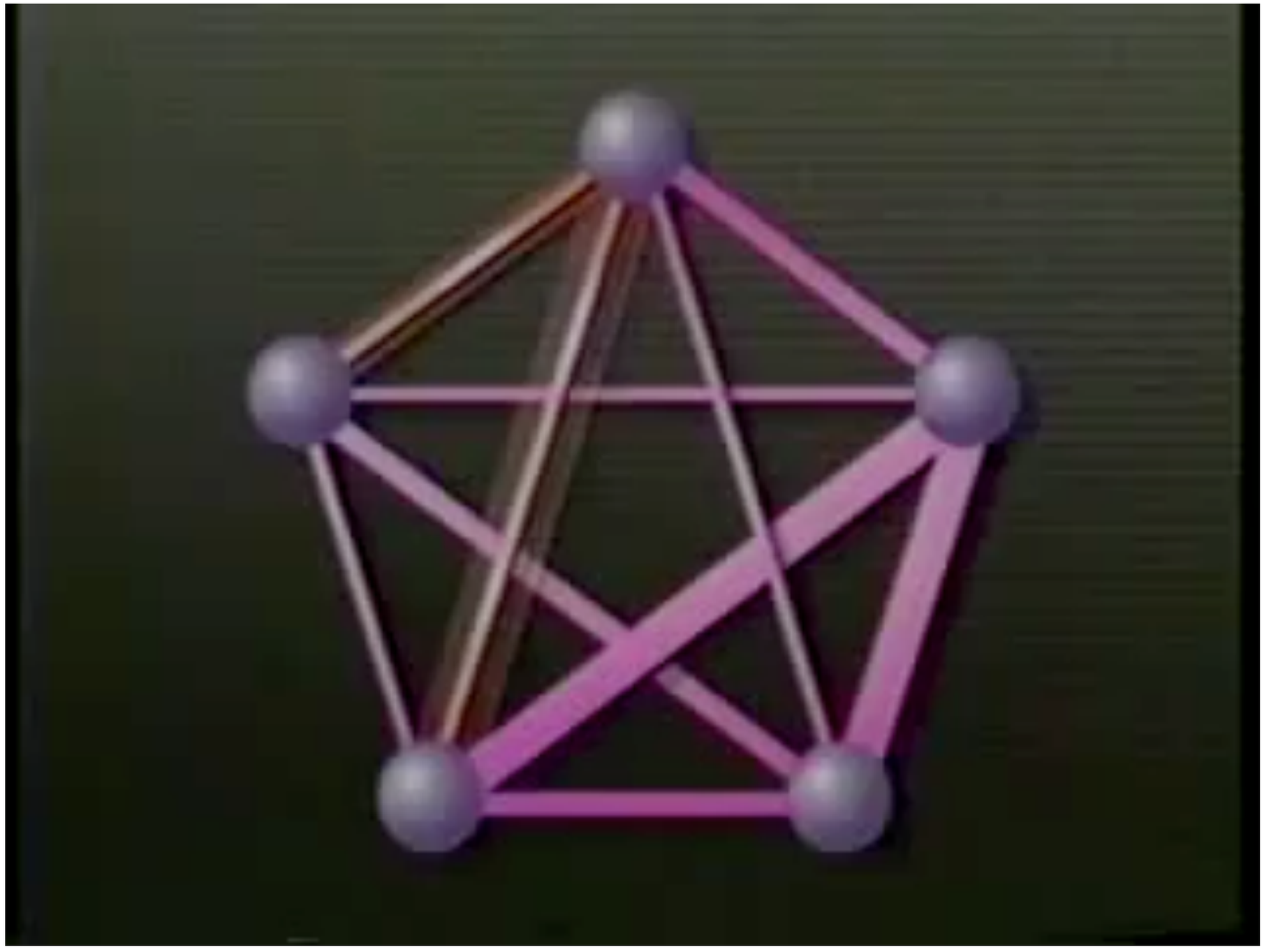
Physical representation



Artificial representation

source

Artificial Neural Networks in the 50s



The perceptron used to be a custom-build hardware used in the US navy

The first hype

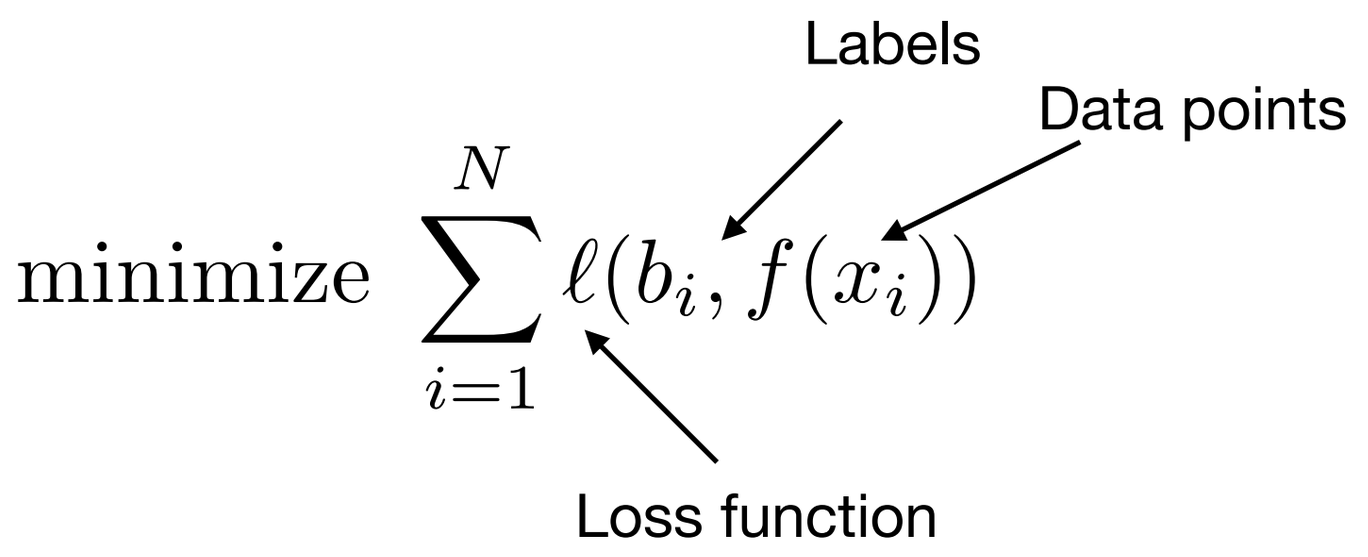
More on history of NN

- Since the 60s there have been multiple lows and highs in the history of artificial neural networks.

http://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning/

Artificial Neural Networks

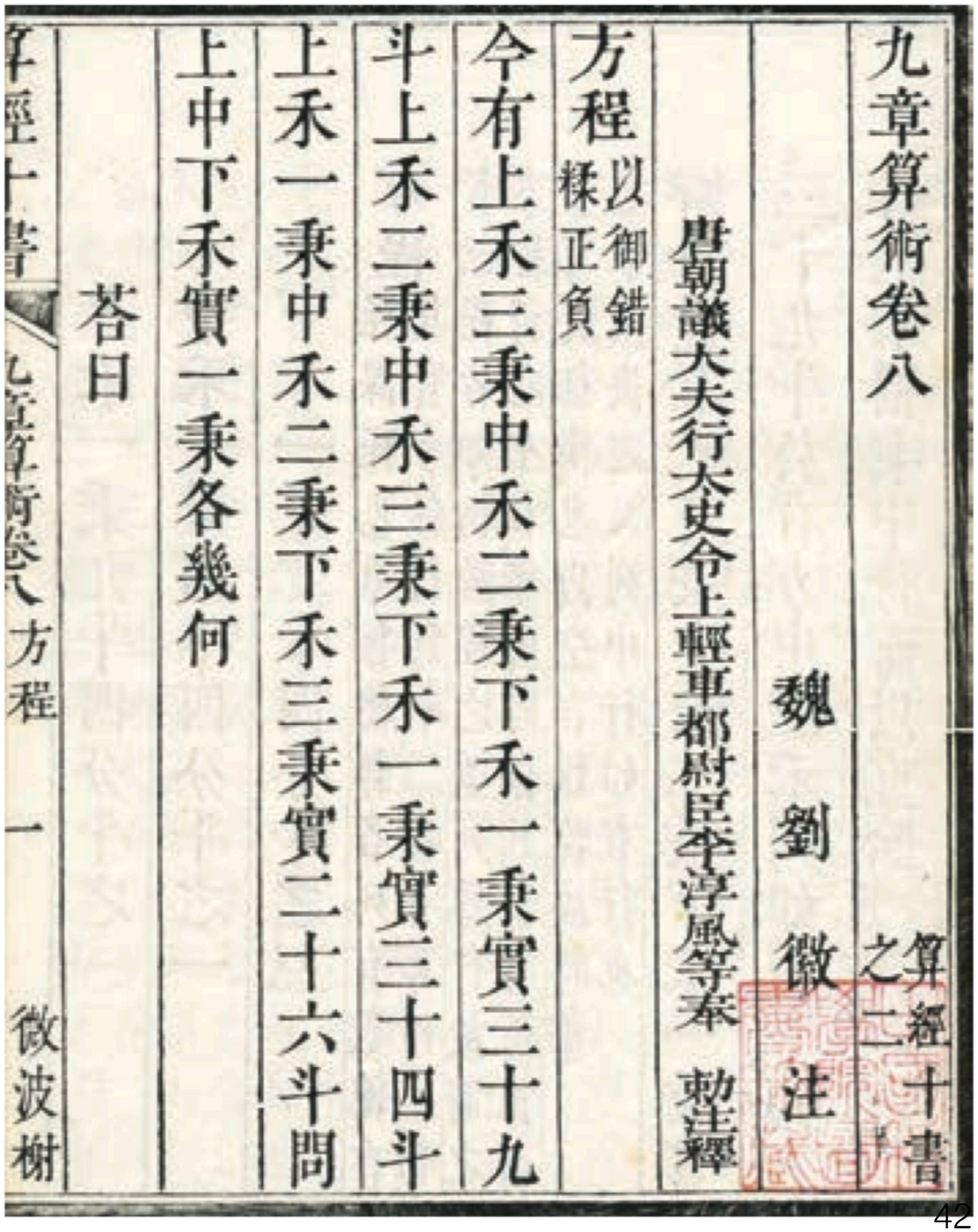
- Most supervised neural network techniques are formulated as finding a local minimum of a non-convex objective function.



$f(x_i) = T_L \circ \cdots \circ T_1(x_i)$ f represents the neural network

$T_k(\cdot) := \sigma_k(W_k \cdot + y_k)$ σ_k is the activation function at layer k

History of Optimization: instances of optimization in ancient times



Problem 1, Chapter 8 of Jiu Zhang Suan Shu, Book title: The Nine Chapters on the Mathematical Art, 263 AD.

First written version of Gaussian Elimination was about solving a grains problem

History of Optimization: instances of optimization in ancient times

Problem I. There are three grades of grain: top, medium and low. Three sheaves of top-grade, two sheaves of medium-grade and one sheaf of low-grade are 39 *Dous*[4]. Two sheaves of top-grade, three sheaves of medium-grade and one sheaf of low-grade are 34 *Dous*. One sheaf of top-grade, two sheaves of medium-grade and three sheaves of low-grade are 26 *Dous*. How many *Dous* does one sheaf of top-grade, medium-grade and low-grade grain yield respectively?

- x: Dous (weight measurement) of 1 sheaf of top-grade
- y: Dous of 1 sheaf of medium-grade
- z: Dous of 1 sheaf of low-grade

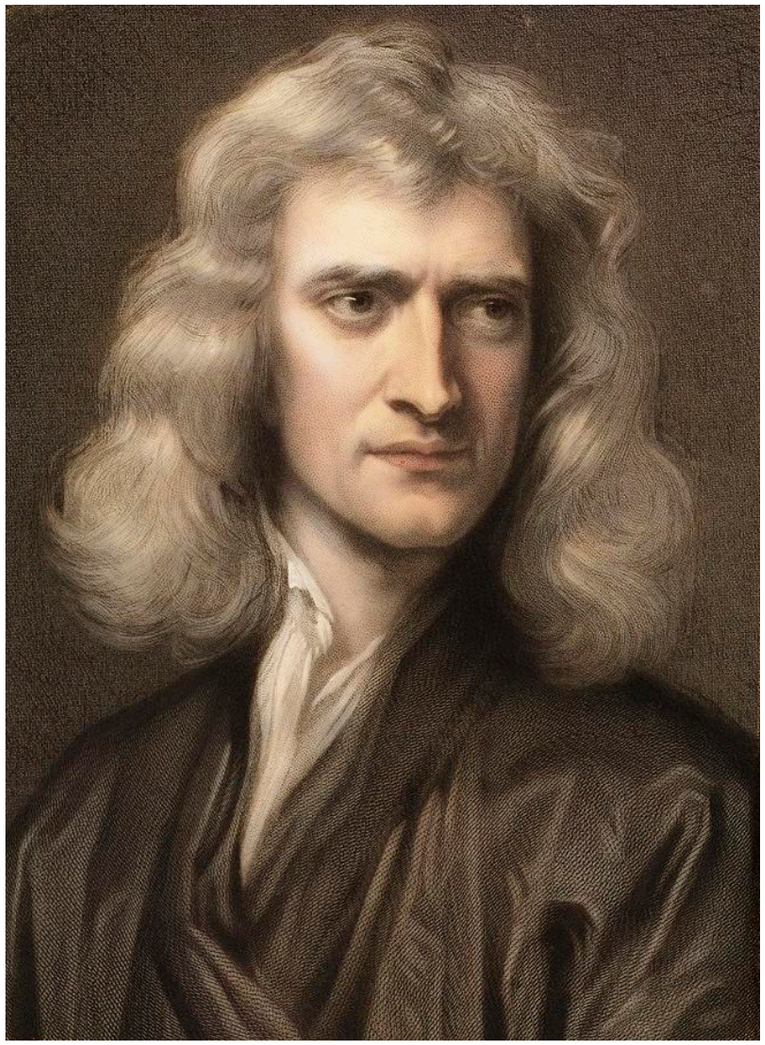
$$3x + 2y + z = 39$$

$$2x + 3y + z = 34$$

$$x + 2y + 3z = 26$$

What is “x”, “y” and “z”?

History of Optimization: instances of optimization methods in the middle ages and early modernity



Sir Isaac Newton (1642 - 1727)

Is Newton's method really Newton's method?

The origins of Newton's method are in the work of Jamshid al-Kashi (1380-1429), *The Key to Arithmetic*.

Newton improved the method of al-Kashi from linear to quadratic convergence (more on that in subsequent lectures).

History of Optimization: approaching today

- Combinatorial optimization. First written problems: T*he Bridges of Konigsberg* and the *Chinese Postman Problem.*

- Linear Programming: The Simplex algorithm by G. Dantzig in 1947, the first polynomial time algorithm in 1979 by L. G. Khachiyan.

- Continuous optimization (**this course**)

  - The gradient method was discovered by Louis Augustin Cauchy (1789 - 1857).

  - The optimality conditions for nonlinear optimization problems were first discussed by William Karush (1917 - 1997) in his Master's thesis

“Nothing in the world takes place without optimization, and there is no doubt that all aspects of the world that have a rational basis can be explained by optimization methods”

–Leonard Euler 1744

History of Optimization

- Optimization stories: book from the International Symposium of Mathematical Programming 2012.

Goals of the Course

- In this course we will mostly deal with unconstrained optimization problems.
- We will learn how to distinguish convex from nonconvex problems.
- We will learn how to find minima using iterative optimization algorithms.
- We will apply what we learn to popular real-world data science problems.

Goals of the Course

- Optimization methods that we will learn
  - Gradient descent
  - Accelerated gradient descent
  - Proximal gradient descent
  - Accelerated proximal gradient descent
  - Newton's method
  - Newton-based methods
  - Stochastic Gradient method
  - Coordinate descent methods

First-order methods

Good at finding low accuracy solutions very fast!

Second-order methods

Robust methods, can get high accuracy solutions faster than simple gradient methods

Randomized methods

Good for very large-scale problems

Tentative Schedule

Week 1	Sep 5	Introduction, Preliminaries
Week 2	Sep 10	Image Denoising, Convexity, First-order Optimality Conditions
Week 3	Sep 17	Gradient Descent, Convergence of Gradient Descent
Week 4	Sep 24	Total Variation Image Denoising
Week 5	Oct 1	Newton's method, Convergence of Newton's method
Week 6	Oct 8	Conjugate Gradients and Newton-CG
Week 7	Oct 15	Reading Week
Week 8	Oct 22	Compressed Sensing and Proximal Gradient Descent
Week 9	Oct 29	Local Graph Clustering and Coordinate Descent Method
Week 10	Nov 5	Logistic Regression and Randomized Optimization Methods
Week 12	Nov 17	Non-convex optimization, Recommender System and Face Reconstruction
Week 11	Nov 12	Optimization methods for non-convex optimization
Week 13	Nov 24	Neural networks and optimization methods for Neural Networks

Assignments

- Every time that we introduce a new application/problem. You will be asked to implement a solver for that problem.

- Also, you will be asked to compare algorithms that can solve the same problem.

Grades

- Each assignment has the same weight.
- Your final grade will simply be the average of the assignment grades.

Required mathematical background

- Basic linear algebra, i.e., vector and matrix operations, matrix decomposition, eigen- and svd-decompositions.

- Multivariate calculus

- Basic analysis, i.e., convergence and limits

- Basic probability, i.e., common distributions, means, and so on

Programming Languages

- Python and/or MATLAB

Optional texts

- S. Bubeck, *Convex Optimization*: Algorithms and Complexity, Foundations and Trends in Machine Learning, 2015. Preliminary version available on arxiv.org.

- P. Jain and P. Kar, *Nonconvex optimization for Machine Learning*, 2017. Preliminary version available on arxiv.org.

- L. Bottou and F. E. Curtis and J. Nocedal, *Optimization Methods for Large-Scale Machine Learning*, Preliminary version available on arxiv.org.

- S. Boyd and L. Vandenberghe, *Convex Optimization*, available on the web.

- J. Nocedal and S. J. Wright, *Numerical Optimization*. 2nd Edition.

Office hours

- Instructor: Kimon Fountoulakis, DC3611, Monday 3-4pm

- TA: Chufeng Hu, DC3594 Office #5, Monday 10-11am

Piazza

- Sign-up link: piazza.com/uwaterloo.ca/fall2019/cs794
- Class link: piazza.com/uwaterloo.ca/fall2019/cs794/home

Reproducibility

- All experiments demonstrated in these slides can be reproduced in Python, see the following repo

https://github.com/kfoynt/optimizationForDataScienceNotebook

- I will be uploading the code after each assignment. Currently, it should be empty.

More details

- Please read the outline PDF on Learn and piazza.