Mathematical Foundations of Data Science

Thomas Strohmer
Department of Mathematics
University of California, Davis

Spring 2019

Course Objective

Experiments, observations, and numerical simulations in many areas of science nowadays generate massive amounts of data.

This rapid growth heralds an era of "data-centric science," which requires new paradigms addressing how data are acquired, processed, distributed, and analyzed.

This course covers mathematical concepts and algorithms (many of them very recent) that can deal with some of the challenges posed by Artificial Intelligence and Big Data.

Details about this course

This course is about mathematical methods for Data Science

Prerequisite:
Linear algebra and a basic knowledge in Probability and programming (preferably Python or Matlab).

What this class is not about:

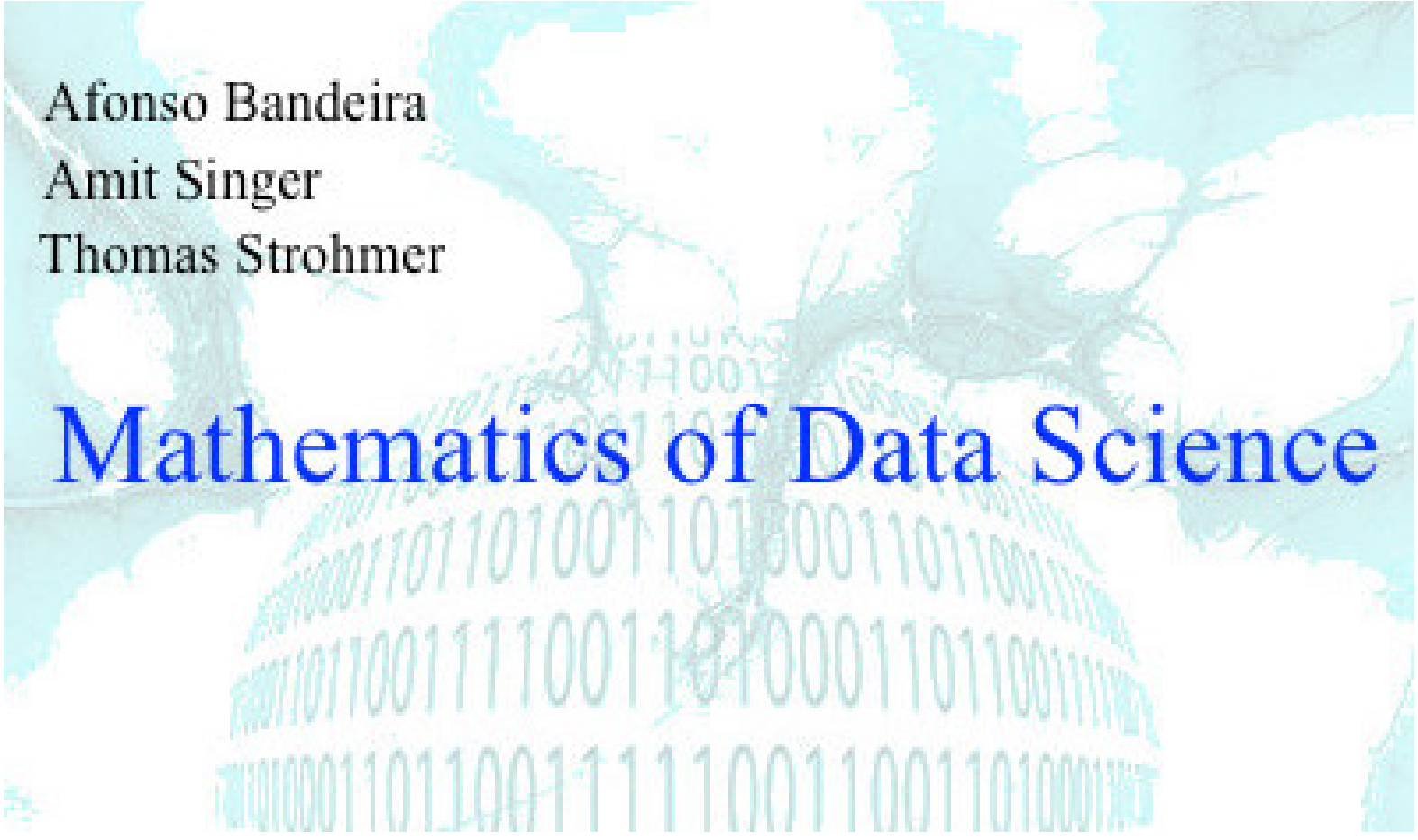
- Formal software development
- Database theory
- Specific applications
- Heuristic methods that lack mathematical foundations (well, except for deep learning ...)

Textbooks

There is no required textbook. The following books contains some material on these topics (but there is no need to buy these books)

- C. Bishop. Pattern Recognition and Machine Learning.
- F. Cucker, D. X. Zho. Learning Theory: an approximation theory viewpoint.
- S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing.
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction.
- M. Mahoney. Randomized Algorithms for Matrices and Data.
- G. Strang. Linear Algebra and Learning from Data.
- R.Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science.

Textbook in development



Notes from the book draft will be made available.

Grading Scheme

- 40% Homework: will be assigned about every other week. A subset of these problems will be graded.
- 60% Final Project

Final Project:

Write a 8-page (or so) report on one of the following topics:

- Describe how some of the methods you learned in this course will be used in your research.
- Find a practical application yourself (not copying from papers/books) using the methods you learned in this course; describe how to use them; include numerical demonstrations.
- Find an interesting data set and present a careful numerical comparison of existing algorithms related to one of the topics of this couse.
- I will post and discuss interesting problems for the Final Project during class.
- If in doubt, please ask me!

Teaching Assistant

Yang Li

TA office hours: M: 3pm-4pm, W: 4pm-5pm.

Goal and challenges of Data Science

Goal: The goal is to turn data into information

Challenges: Capture, curation, time-limitations, storage, search, sharing, transfer, analysis, and visualization of the data.

Data can be massive, non-static, multi-modal, incomplete, noisy, non-random, unstructured, dynamic, streaming, ...

“Data is the new (crude) oil for the economy!”

“Data is the new (crude) oil for the economy!”

You are not Google’s customer.

“Data is the new (crude) oil for the economy!”

You are not Google’s customer.

You are Google’s commodity (crude oil)

Big Data Everywhere!

Lots of data is being collected and warehoused

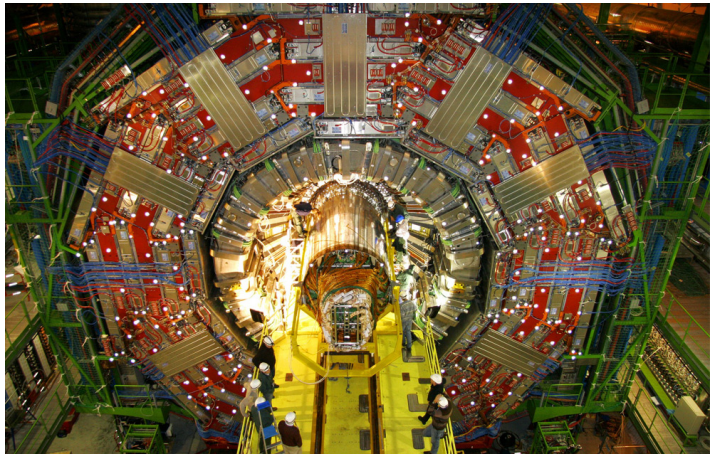
- Web data (often user-provided)
- e-commerce, purchases at stores
- Medical data, health care
- Bank/Credit Card transactions
- Social Network
- Traffic, GPS, ...
- Scientific experiments
- ...

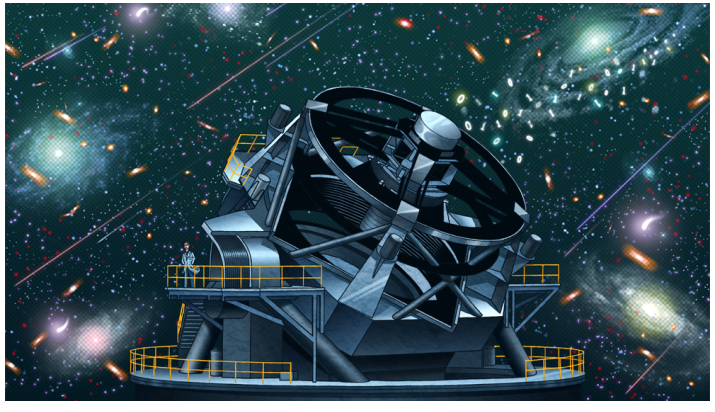
How much data?

- YouTube contains 120 million videos and 72 hours of video uploaded every minute.
- Google processes 3.5 billion requests per day
- There is currently an estimate of 3.8 trillion photographs, 10% of them taken in the last year.
- Facebook has about 140 billion images with about 300 million new images a day.
- 2.5PB are flowing through Walmart's databases
- NYSE collects 1 TB each day.

How much data?

- CERN's Large Hydron Collider generates 15 PB a year
- The BRAIN initiatives produce terabytes of data a day
- The Large Synoptic Survey Telescope in Chile will collect 30TB per night. Headed by Tony Tyson from UC Davis





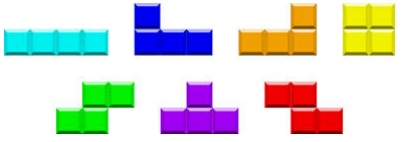
How much data?

Governments (USA, China, Russia, UK, Israel, Germany, ...) collect ??? PB /day

How much data?

Governments (USA, China, Russia, UK, Israel, Germany, ...) collect ??? PB /day

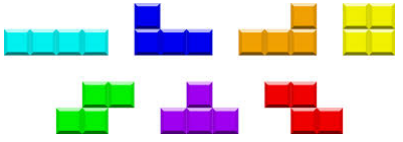
The CIA (via In-Q-Tel) was an early investor in Facebook



How much data?

Governments (USA, China, Russia, UK, Israel, Germany, ...) collect ??? PB /day

The CIA (via In-Q-Tel) was an early investor in Facebook



Somewhere in Nevada is an 8-Football field large storage area that collects all the emails sent in the USA.

More Data ...

Experts now predict that 40 zettabytes of data will be in existence by 2020.

Big Data does not just mean massive amounts of data
Big Data also means complex data

- Heterogeneous data
- Incomplete data
- Unstructured/semi-structured Data
- Graph Data
- Social Network, Semantic Web
- Streaming Data

Big Data is not new

- Seismic data acquisition and processing
- Census
- Wall Street hedge funds (e.g. Renaissance Technologies)
- Governments
- Banks, Insurances
- Scientific Research

Big Data Tasks

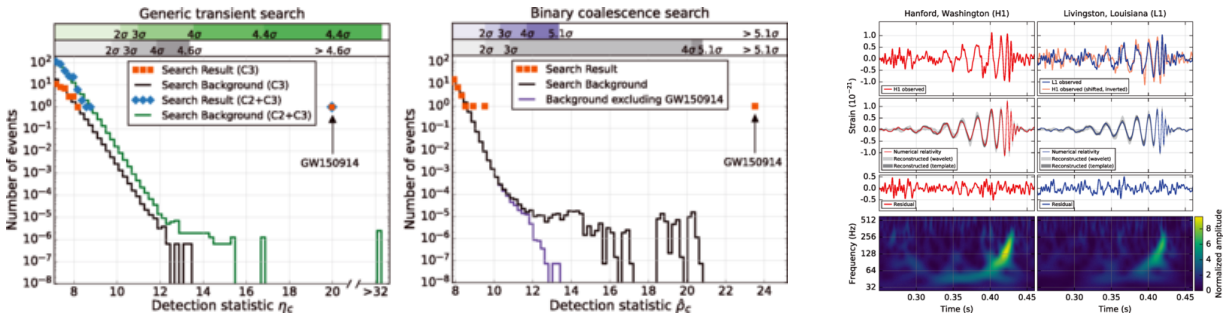
- Discovery of useful, possibly unexpected, patterns in data
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Finding outliers (security threat, credit card theft, ...)
- Clustering
- Classification
- Object recognition
- Visualization, dimension reduction
- “Data cleaning”: denoising, smoothing, grouping, ...
- Association Rule Mining (Costumers who buy X often buy Y, Costumer 123 likes product p10)
- Collaborative filtering: users collaborate in filtering information to find information of interest (Amazon, Netflix)

LIGO and Gravitational Waves

LIGO: Laser Interferometer Gravitational-Wave Observatory

Gravitational waves: predicted by Einstein's Theory of General Relativity.
Indirectly observed in the 1980s (Nobel Prize in 1993)
Directly observed for the first time by LIGO in 2016.

What means "directly observed"?
Need to use carefully designed algorithms to detect a certain pattern in massive amounts of very noisy data



Meta Data Analysis

The idea is 100 years old (see Karl Pearson), but its full potential will be unleashed only now.

Example:
In a recent analysis researchers developed a framework for comparing classiffers common in Machine Learning (Boosted decision trees, Random Forests, SVM, KNN, PAM and DLDA) based on a standard series of datasets.

Result: A simple (but mathematically rigorous) method gave better classification results across the data sets than the "glamorous" methods.

The dawning Age of Big Data will make it not just possible but very common (and perhaps necessary?) to validate methods via such meta data analyses. [David Donoho, Stanford].

Big Data Startups

Crunchbase records more than 3000 Startups and
Angellist more than 4000 Startups in Big Data and AI.

Two random examples (out of 1000+?) of Bay area startups:

- helm.ai: (Menlo Park) Self-driving cars
- 23andMe (Mountain View): Genomics

Some startups by mathematicians:

- PredPol: Crime prediction (Andrea Bertozzi)
- ThetaRay: Threat detection (R. Coifman, Amir Averbuch)
- Ayasdi: Topological data analysis (Gunnar Carlsson)

Many Data Initiatives Nationwide

Campus-wide initiatives at NYU, Columbia, Michigan, Harvard, MIT, Berkeley, ...

New Master's Degree programs in Data Science, for example at Berkeley, NYU, Stanford, UC Davis, ...

New Alan Turing Institute for Data Sciences in UK

For a long list across the world see
`http://data-science-university-programs.silk.co`

Topic Overview (tentative)

- Basic goals of AI and Machine Learning
- Curses and blessings of dimensionality, Surprises in high dimensions
- Singular Value Decomposition, Principal Component Analysis
- Data Clustering: k-means, graph Laplacian
- Classification: Some basics on Deep Learning
- Linear dimension reduction, random projections
- Nonlinear dimension reduction, diffusion maps, manifold learning, intrinsic geometry of data
- Compressive sensing and sparsity
- Matrix completion and low-rank modeling
- Randomized algorithms

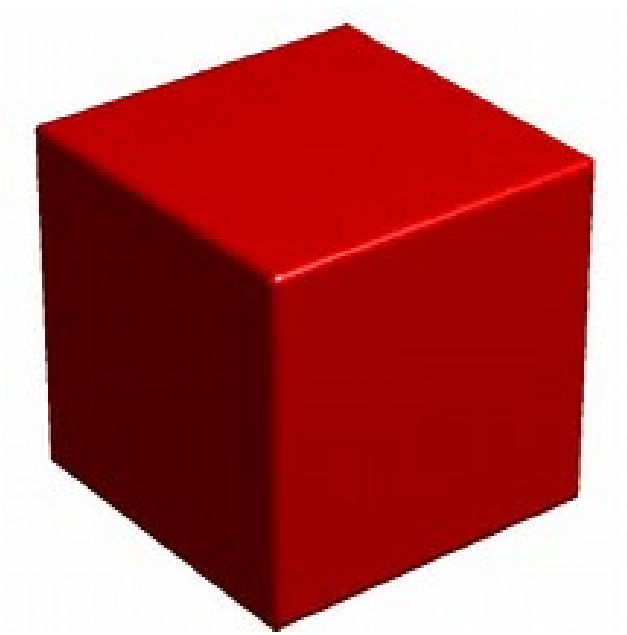
High-dim. probability; Curses and blessings

Things in high dimension can behave very differently than in low dimension.

High-dim. probability; Curses and blessings

Things in high dimension can behave very differently than in low dimension.

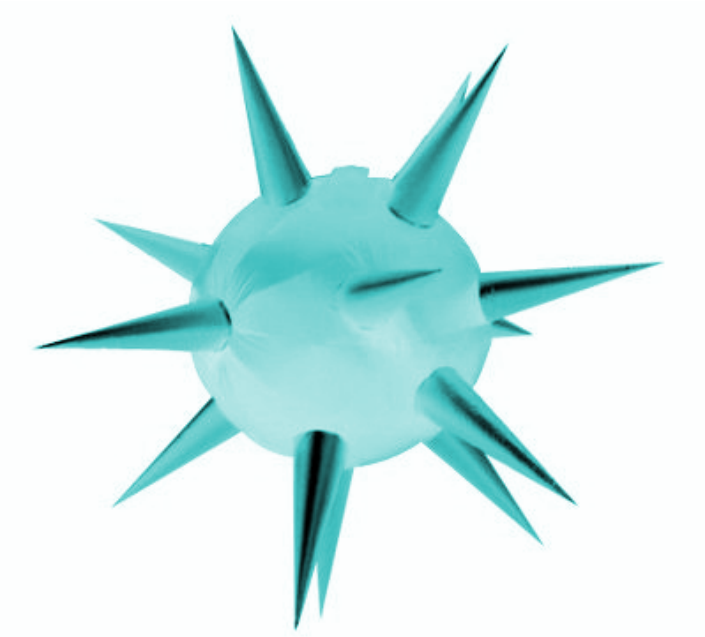
A cube in high dimensions looks like this:



High-dim. probability; Curses and blessings

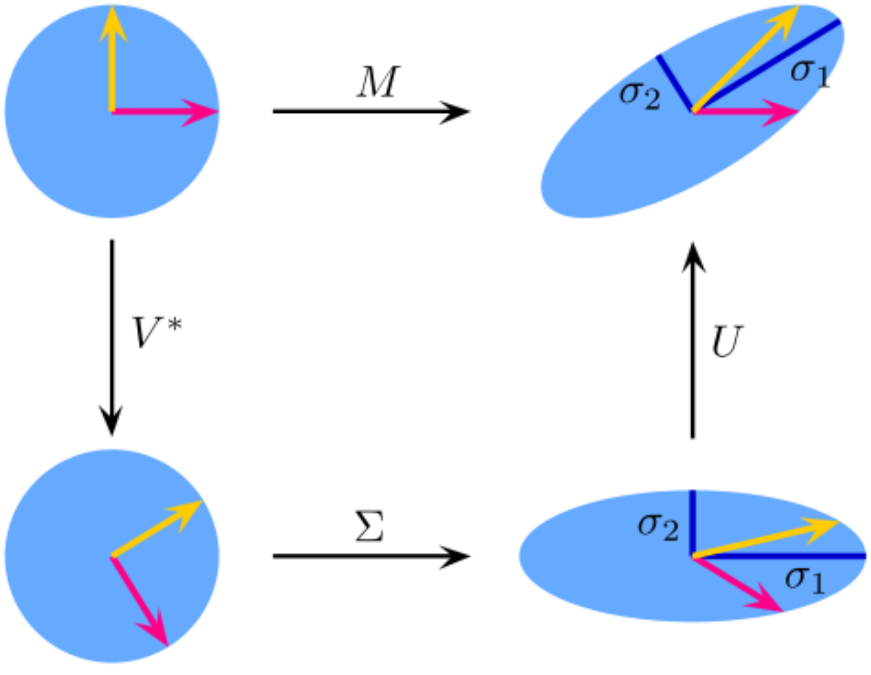
Things in high dimension can behave very differently than in low dimension.

But a cube in high dimensions also looks like this:



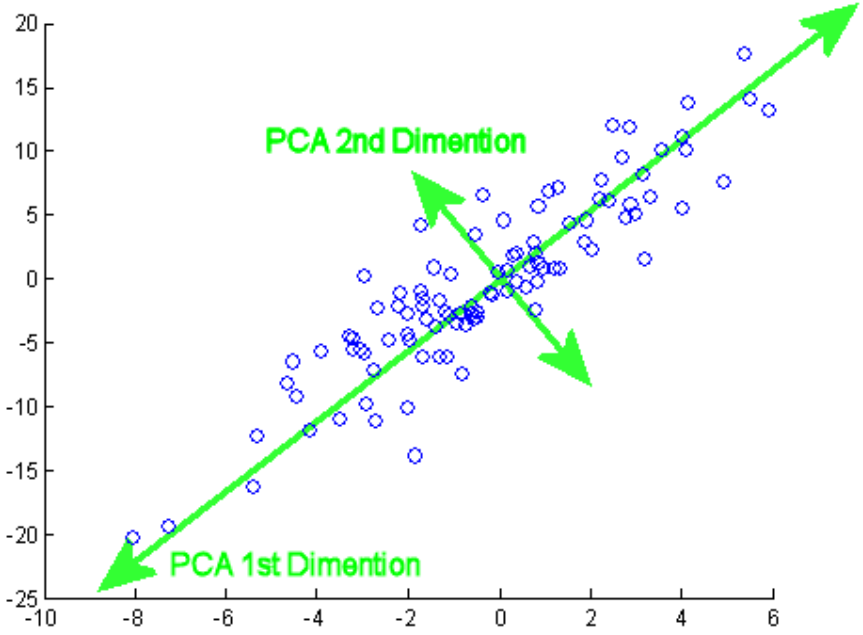
SVD and PCA

Singular Value Decomposition



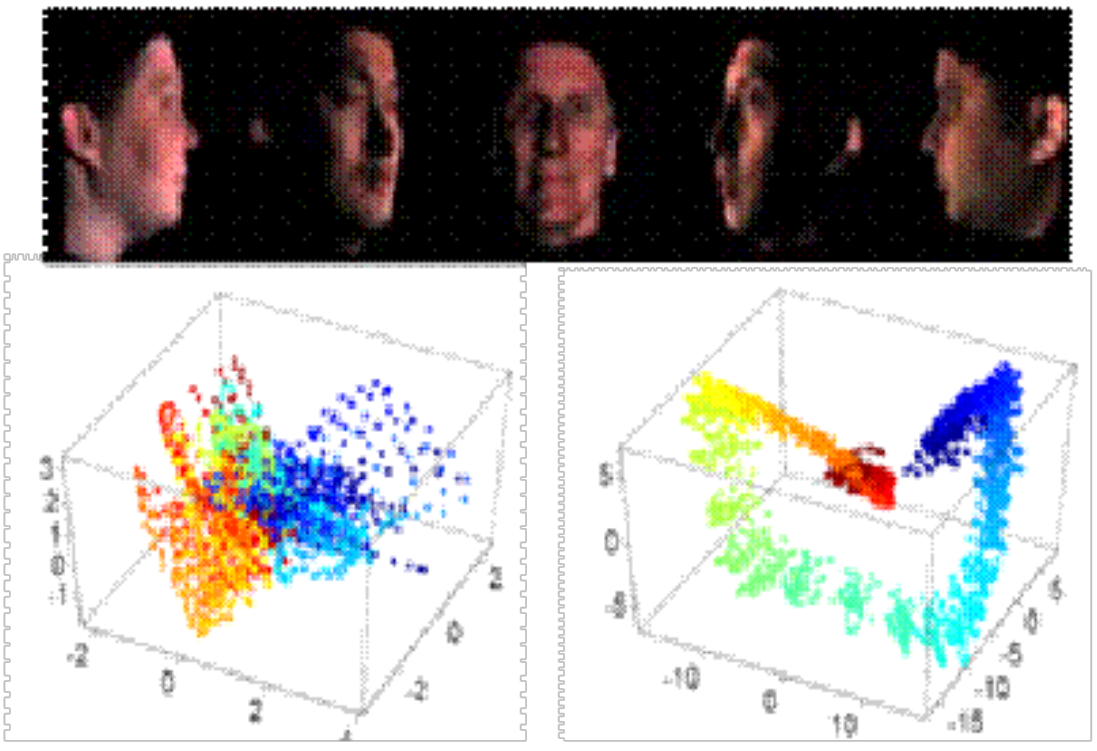
$$M = U \cdot \Sigma \cdot V^*$$

Principal Component Analysis



Dimension reduction

Linear dimension reduction and random projections

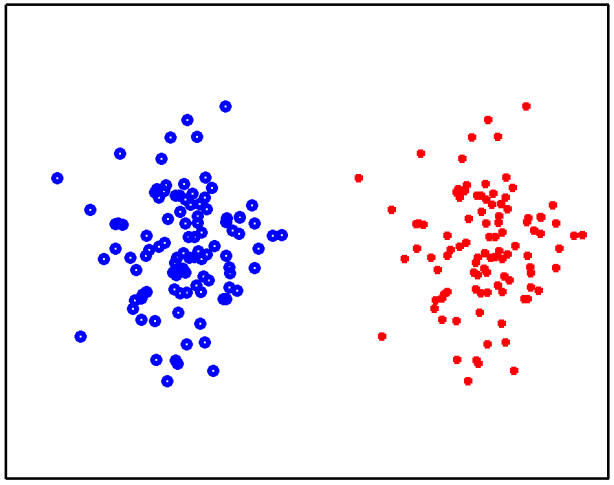


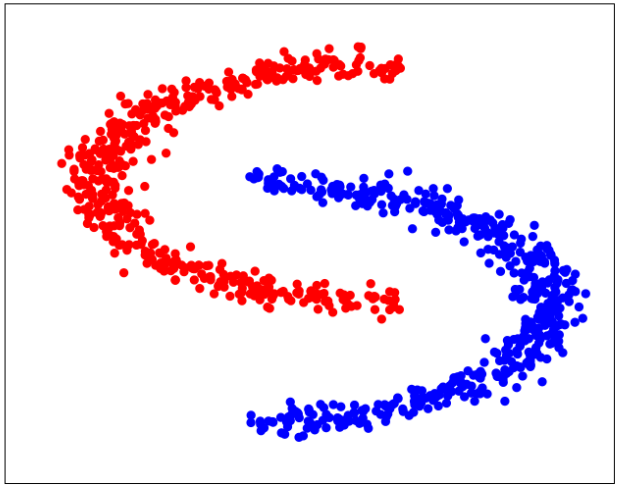
Johnson-Lindenstrauss projections

Clustering

A basic task in data analysis is clustering:

k-means: advantages and limitations





Graph Laplacian, spectral clustering

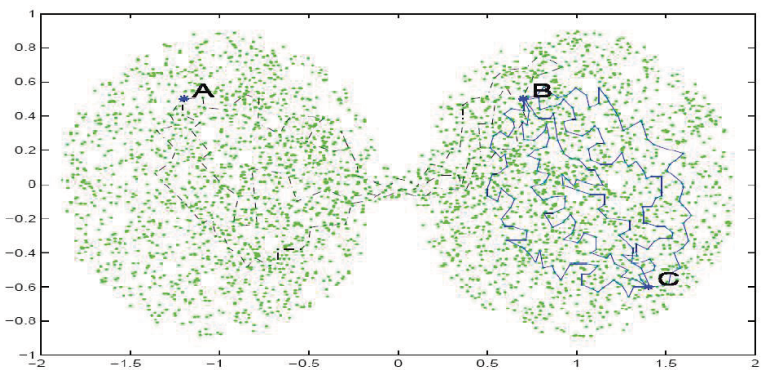
Diffusion maps

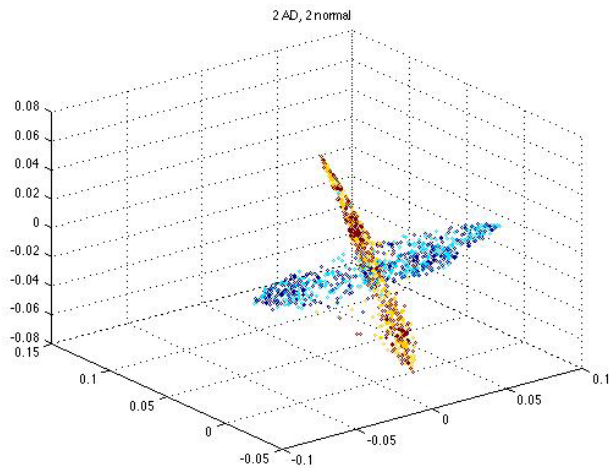
What is a diffusion map?

Manifold learning

Intrinsic geometry of data

Nonlinear dimension reduction

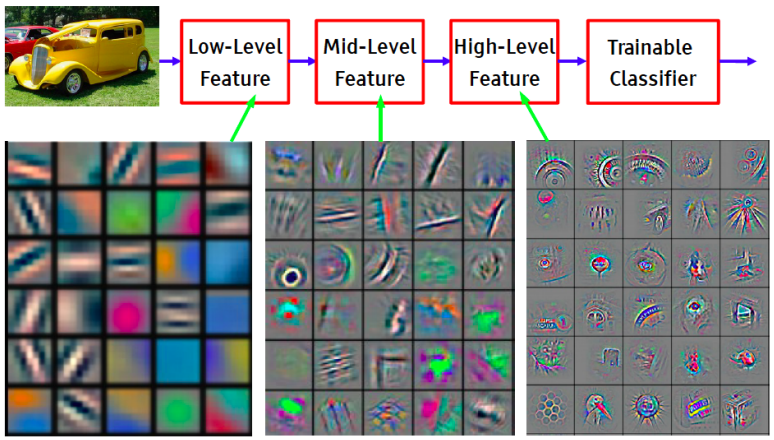




Deep Learning and Classification

Deep Learning: neural network with more than one layer

Deep networks achieve state-of-the-art results in several complex object recognition tasks



They learn a huge network of filter banks and non-linearities on large datasets

Heuristic method, a lot of trial-and-error

Almost no mathematical theory (yet)

Data Science, Machine Learning, and Artificial Intelligence

Artificial Intelligence

1950: Alan Turing's "Computing Machinery and Intelligence"

1955: McCarthy, Minsky, Rochester, and Shannon:

> *"We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."*

Artificial Intelligence

Definition:
Capability of a machine to imitate intelligent human behavior

Artificial Intelligence in the classical sense has by far not been realized (if ever), has not had any practical impact, and is therefore not studied in this course.

What is studied in this course?

Artificial Intelligence

Definition:
Capability of a machine to imitate intelligent human behavior

Artificial Intelligence in the classical sense has by far not been realized (if ever), has not had any practical impact, and is therefore not studied in this course.

What is studied in this course?
Augmented Intelligence (AI) or Intelligence Augmentation (IA)

There is no "I" in AI (yet)

Augmentated Intelligence is about empowering humans with tools that make them more capable, while traditional AI has been about removing humans fully from the loop.

Current AI algorithms are surprisingly powerful and at the same time surprisingly stupid. They are definitely not intelligent ...

Throughout this course:

Artificial Intelligence = Augmentated Intelligence

AI has had enormous impact in the last few years
Many tools that make AI useful in practice are fairly recent

What made recent breakthrough in AI possible?
Combination of Big Data with advances in machine learning, fast algorithms, and computer power

Machine Learning

Machine Learning is a subfield within Artificial Intelligence that builds algorithms, which allow computers to learn to perform tasks from data instead of being explicitly programmed.

Supervised Learning: uses a known dataset (the training dataset) to make predictions. The training dataset includes input data and labeled responses. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset.
A test dataset is often used to validate the model.

Unsupervised Learning: is used to draw inferences from datasets consisting of input data without labeled responses.

Semi-supervised Learning: We have a small amount of labeled data and a large amount of unlabeled data.

Machine Learning

Main tasks of supervised learning:

- Regression
- Classification

Main tasks of unsupervised learning:

- Clustering
- Density estimation
- Dimension reduction

Regression

Learning the relationship between independent variables (predictors) and dependent variables.
Used for prediction and forecasting.

Estimation of housing price (say, based on location, number of bedrooms, ...), prediction of stock price, weather forecast, ...

Methods:

- Linear regression (least squares, ...)
- Kernel regression
- Gaussian process regression

Clustering

Finding natural groupings of data and a label associated with each of these groupings.

Marketing (consumer groups), Netflix, Amazon, text mining, image segmentation, ...

Methods:

- k-means
- Spectral clustering
- Hierarchical clustering

Dimension reduction

Reducing the number of variables under consideration

Data visualization, faster processing of data, reducing storage,

Methods:

- Principal component analysis
- Manifold learning
- Random projections
- Compressive sensing

Density estimation

Construction of an estimate of an unobservable underlying probability density function based on observed data. Finding likelihood or frequency of objects.

Finance (risk estimation), medical diagnostics, outlier detetion

Methods:

- Histograms
- Kernel density estimation
- Mixture of Gaussians

Classification

Organizing data into gategories, predicting a category of a data.

Does a person have a certain illness or not?
Classifying an image according to the objects in the image
Anomaly detection: detecting if a transaction is a fraud or not
Spam filtering, News vs Fake News

Methods:

- Support vector machines
- Random forests
- Deep Learning

What AI can or cannot do

- Play a decent game of table tennis?
- Play a decent game of Jeopardy?
- Drive safely along a well-mapped road?
- Buy a week's worth of groceries on the web?
- Buy a week's worth of groceries at Whole Foods?
- Discover and prove a new mathematical theorem?
- Converse successfully with another person for an hour?
- Perform a surgical operation?
- Put away the dishes and fold the laundry?
- Translate spoken Chinese into spoken English in real time?
- Write an intentionally funny story?

What AI can or cannot do

- Play a decent game of table tennis?
- Play a decent game of Jeopardy?
- Drive safely along a well-mapped road?
- Buy a week's worth of groceries on the web?
- Buy a week's worth of groceries at Whole Foods?
- Discover and prove a new mathematical theorem?
- Converse successfully with another person for an hour?
- Perform a surgical operation?
- Put away the dishes and fold the laundry?
- Translate spoken Chinese into spoken English in real time?
- Write an intentionally funny story?

Green = Yes, Red = No, Blue=?

Applications of AI

DeepMind's AlphaGo beats a world class Go player

Image classification: Google, Facebook (DeepFace), ...

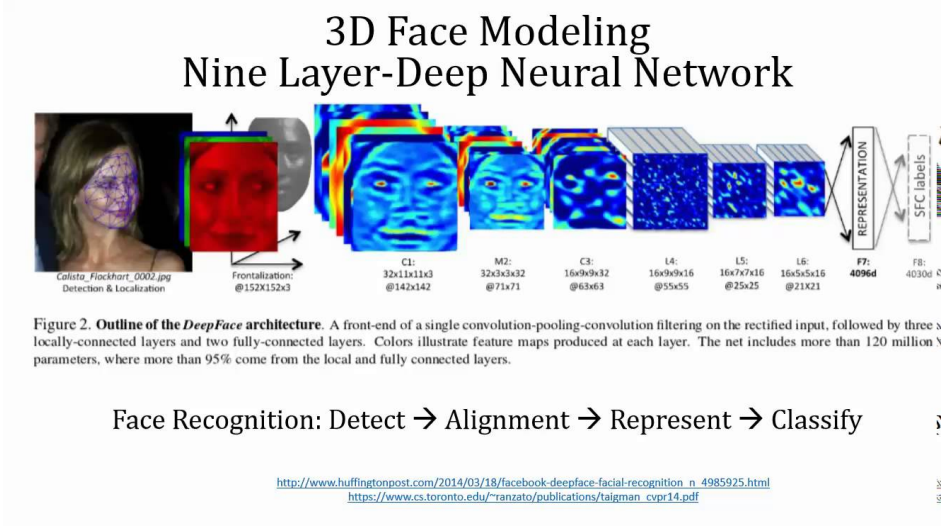


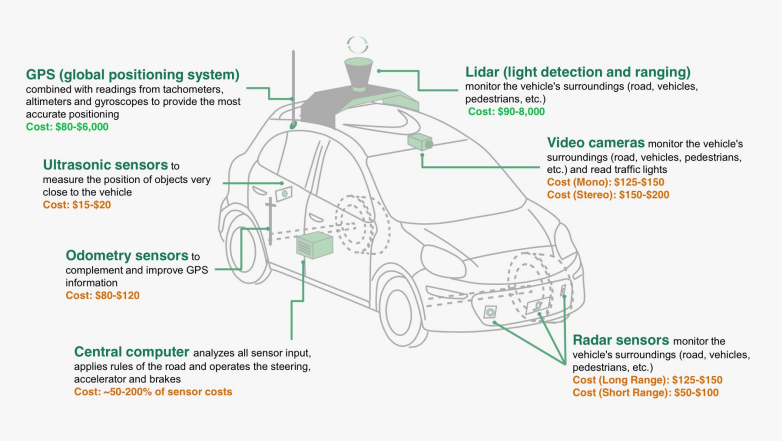
Figure 2. **Outline of the *DeepFace* architecture**. A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

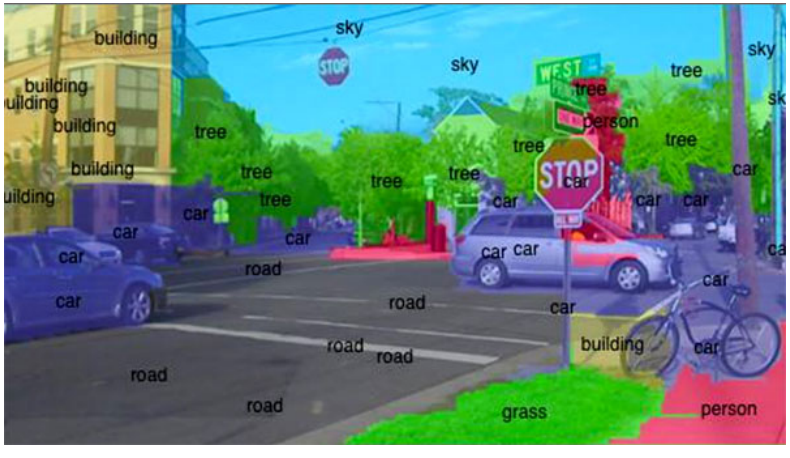
Face Recognition: Detect → Alignment → Represent → Classify

http://www.huffingtonpost.com/2014/03/18/facebook-deepface-facial-recognition_n_4985925.html
https://www.cs.toronto.edu/~ranzato/publications/taigman_cvpr14.pdf

Applications of AI

Self-driving cars





Uber Ends Self-Driving Car Test in San Francisco

Uber Suspends Tests of Self-Driving Vehicles After Arizona Crash

Applications of AI

Self-flying airplanes: easier than self-driving cars.

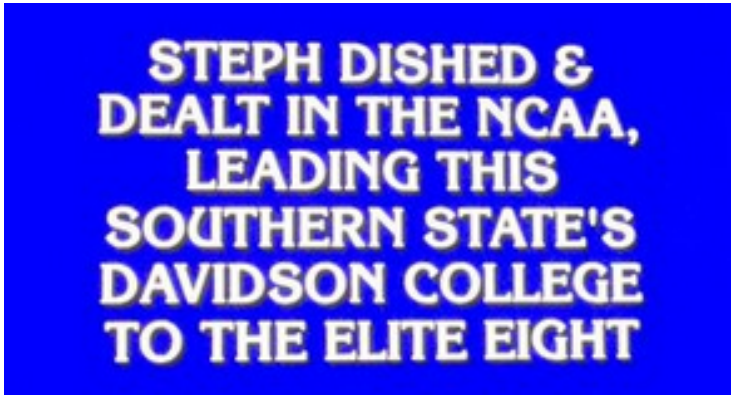
Airplanes use autopilots for many years

Airbus plans a self-flying car to be in operation by 2018

Applications of AI

1997 chess victory of IBM's Deep Blue over Garry Kasparov

2011 Jeopardy! victory of IBM's Watson system over two human contestants. Watson needed to be able to extract semantic meaning from the questions.



Speech recognition, automatic translation

Medical Diagnostics

- Machine Learning can help find patterns in large amounts of data to detect markers for diseases
- Assume we have a 3-dim. MRI of the brain:
  Doctors can look at 3-dim. MRIs only one slice at a time.
  They may miss patterns that can much better be detected by looking at the entire 3-dim. data simultaneously.
  Mathematical algorithms can easily analyze 3-dim. data
- The point is not to replace the doctor (as classical AI might attempt to do), but to assist the doctor with information that may be difficult to access - this is augmented intelligence.
- AI must be trustworthy. Will we develop trust as we interact with AI systems over time, as we have done with ATMs?

Applications of AI

MELANIE GONICK

Intelligent Machines

Machine learning is making pesto even more delicious

Applications of AI

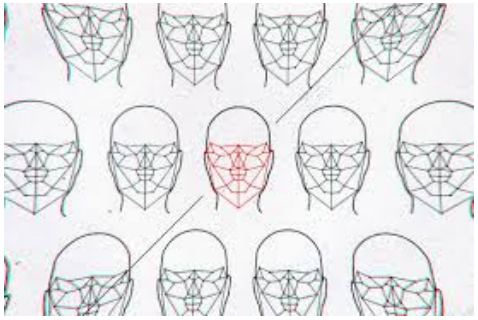
How can a computer learn concepts?

Currently AI heavily relies on vast amounts of training data.

Challenges of AI

Bias in AI algorithms:

April 4 2019: AI experts want Amazon to stop selling facial recognition tech to police.
Amazon's Rekognition program has much higher error rates when it is trying to recognize the gender of darker skinned women than lighter skinned man.



Bias comes from: people who design algorithms, collected data (sampling bias),

Self-driving cars: Trolley problem, Ethics of AI

Challenges of AI

Surprise:

What is difficult for humans is “easy” for AI
(playing chess, detecting patterns in complex data, ...)

What is easy for humans is very difficult for AI
(moving around, language, common sense reasoning, ...)

Consequences of AI

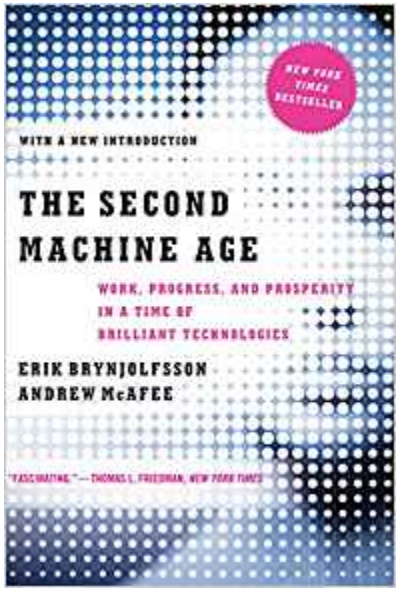
AI will lead to economic disruptions

Consequence: many people will lose their jobs, social turmoil

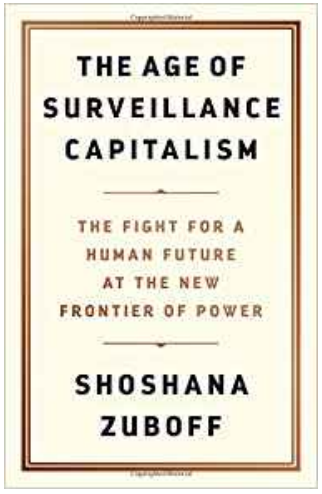
Inequalities in society will increase much further

Changes may be at least as forceful as during the Industrial Revolution

Steps to reduce massive negative impact of AI need to be taken already now before AI fully kicks in



Surveillance Capitalism



G-MAFIA: Google, Microsoft, Amazon, Facebook, IBM, Apple)
The G-MAFIA provides free services that billions of people cheerfully use, enabling the providers of those services to monitor the behaviour of those users in astonishing detail – often without their explicit consent.
Surveillance capitalism claims human experience as free raw material for translation into behavioural data, which is fed into AI algorithms and then fabricated into prediction products (advertising, ...).

Surveillance Capitalism

China recently introduced a Social Credit Score
"It allows the trustworthy to roam freely under heaven while making it hard for the discredited to take a single step."

"Utopian Big Data Bliss or Black Mirror on Steroids?"
[Forbes Magazine, Jan.21 2019]

Privacy-Preserving Machine Learning

Dangers: Surveillance capitalism, data breaches (Target, Facebook, Bank of America, ...), Fake News, Cambridge Analtyics, Insurance companies, ...

Challenge: Can we benefit from AI and Machine Learning, while minimizing surveillance, data intrusion, ...?
Can we design algorithms than enable us to do privacy-preserving machine learning?

Differential privacy: add noise to data for obfuscation, good idea, but limited usefulness

On-device machine learning: Run algorithms on the device, without sending them up and down the cloud.

And last but not least

Algorithms for AI and Big Data are powerful.

Use your power responsibly and carefully.

And last but not least

Algorithms for AI and Big Data are powerful.

Use your power responsibly and carefully.

Einstein: “Not everything that can be counted, counts.
And not everything that counts, can be counted.”