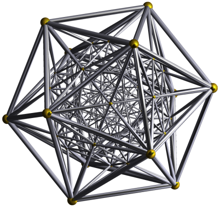
Computational Principles for High-dim Data Analysis

(Lecture One)

Yi Ma

EECS Department, UC Berkeley

August 26, 2021



Instructors of EECS 208, Fall 2021

1. **Prof. Yi Ma** (main instructor): yima@eecs.berkeley.edu
2. **Prof. Jiantao Jiao** (co-instructor): jiantao@eecs.berkeley.edu
3. **Simon (Yuexiang) Zhai** (head GSI): simonzhai@berkeley.edu
4. **Xiangyu Yue** (GSI): xyyue@berkeley.edu

Office hours are all posted on the course websites. Office hours are all to be held virtually **via Zoom** (for now).

Main Textbook

High-Dimensional Data Analysis with Low-Dimensional Models
Principles, Computation, and Applications

John Wright and Yi Ma
Cambridge University Press, 2022.

Pre-production Copy from Website: https://book-wright-ma.github.io

- Github:
  https://book-wright-ma.github.io/Book-WM-20210422.pdf
- Dropbox:
  https://www.dropbox.com/s/0jm9qhm0t5v21gq/Book-WM-20210422.pdf?dl=0

Course Websites

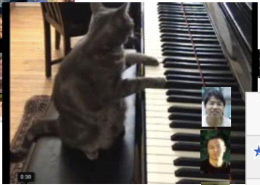
- **Course Website**:
  https://pages.github.berkeley.edu/UCB-EECS208/course_site/
  Course information, detailed schedules, and resources etc.
- **Piazza**:
  https://piazza.com/berkeley/fall2021/eecs208/
  Interactive functions, announcements, Q&A, discussions, and team work etc.

Grading Policy

- **Participation**: 10%.
- **Homeworks**: 50%.
  - 4-5 Written Homeworks (principles).
  - 4-5 Programming Homeworks (practices).
- **Final Project**: 40%.
  - Midterm: 5min pitch of ideas; 2-3 pages of proposal.
  - Final: 15min presentation; 8-10 pages of final report (conference paper style).

A New Paradigm for Modern Data Science

Principles $\xleftrightarrow{\textbf{Computation}}$ **Applications**



Images

⇓➤ ***1M pixels***

Compression
De-noising
Super-resolution
Recognition...

Videos

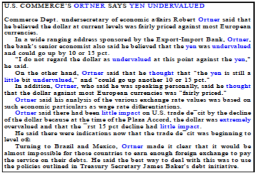
⇓➤ ***1B voxels***

Streaming
Tracking
Stabilization...

User data

⇓➤ ***1B users***

Clustering
Classification
Collaborative filtering...



U.S. COMMERCE'S ORTNER SAYS YEN UNDERVALUED

Commerce Dept. undersecretary of economic a®airs Robert Ortner said that he believed the dollar at current levels was fairly priced against most European currencies.

In a wide ranging address sponsored by the Export-Import Bank, Ortner, the bank's senior economist also said he believed that the yen was undervalued and could go up by 10 or 15 pct.

"I do not regard the dollar as undervalued at this point against the yen," he said.

On the other hand, Ortner said that he thought that "the yen is still a little bit undervalued," and "could go up another 10 or 15 pct."

In addition, Ortner, who said he was speaking personally, said he thought that the dollar against most European currencies was "fairly priced."

Ortner said his analysis of the various exchange rate values was based on such economic particulars as wage rate di®erentiations.

Ortner said there had been little impact on U.S. trade de¯cit by the decline of the dollar because at the time of the Plaza Accord, the dollar was extremely overvalued and that the ¯rst 15 pct decline had little impact.

He said there were indications now that the trade de¯cit was beginning to level o®.

Turning to Brazil and Mexico, Ortner made it clear that it would be almost impossible for those countries to earn enough foreign exchange to pay the service on their debts. He said the best way to deal with this was to use the policies outlined in Treasury Secretary James Baker's debt initiative.

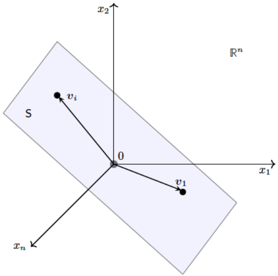
Web data

⇓➤ ***100B webpages***

Indexing
Ranking
Search...

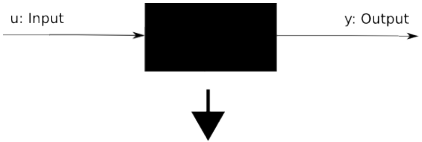
Pursue Low-dim Structures in High-dim Data

Introduction: A Universal Task



"Entities should not be multiplied without necessity."
– William of Ockham, *Law of Parsimony*

System Identification: Linear Systems or RNNs



$$x_{k+1} = Ax_k + Bu_k$$
$$y_k = Cx_k + Du_k$$

Problem: determine the system (A, B, C, D) from the input and output sequences

$$\{u_0, u_1, u_2, \ldots\}, \quad \{y_0, y_1, y_2, \ldots\}.$$

Fact: If the dimension of $\boldsymbol{x} \in \mathbb{R}^n$, then

$$\text{rank}(\boldsymbol{Y}\mathbf{U}^{\perp}) \leq n. \tag{1}$$

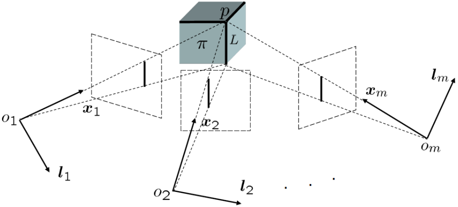
Variants: Recursive Neural Networks (RNNs):

$$\begin{cases} \boldsymbol{x}(t+1) & = & \sigma_{\boldsymbol{x}}\big(\boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{b}\big), \\ \boldsymbol{y}(t) & = & \sigma_{\boldsymbol{y}}\big(\boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{d}\big), \end{cases} \tag{2}$$

Visual Patterns and Correlations

If we view pixels of an $n \times n$ image as entries of a matrix M, then

$$\text{rank}(M) = d \ll n. \qquad (3)$$



Fact:
Let M be the multiview matrix associated with corresponding features (points, lines, planes, symmetric structures), we have

$$\text{rank}(M) \leq 1 \text{ or } 2. \qquad (4)$$

An Invitation to 3D Vision, Ma, Soatto, Kosecka, and Sastry, Springer, 2004.

Signal Acquisition and Processing

Fact: Sample band-limited signals with Nyquist frequency: $f = 2 \cdot \frac{\Omega}{2\pi}$.

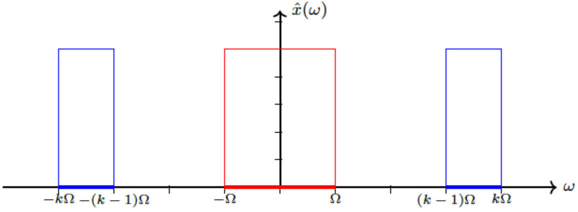
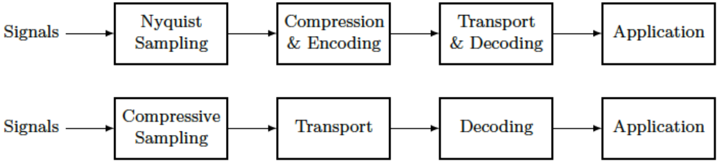


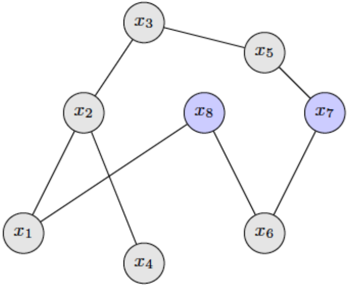
Figure: Comparing Classical Signal Processing and Compressive Sensing Pipelines



Graphical Models in Machine Learning

$\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. Let $\boldsymbol{\Theta} \equiv \boldsymbol{\Sigma}^{-1}$. Then:

$$\textbf{Fact:} \quad \theta_{ij} = 0 \quad \text{iff} \quad x_i \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-i,-j}. \tag{5}$$



$$\boldsymbol{\Theta} = \begin{bmatrix} * & * & 0 & 0 & 0 & 0 & 0 & * \\ * & * & * & * & 0 & 0 & 0 & 0 \\ 0 & * & * & 0 & * & 0 & 0 & 0 \\ 0 & * & 0 & * & 0 & 0 & 0 & 0 \\ 0 & 0 & * & 0 & * & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & * & * & * & 0 \\ * & 0 & 0 & 0 & 0 & * & 0 & * \end{bmatrix}$$

Figure: A graphical model for dependency among random variables.

Graphical Model Identification

Let $\boldsymbol{x}$ be partitioned into **observed** and **hidden** $\boldsymbol{x} = (\boldsymbol{x}_o, \boldsymbol{x}_h)$. Its covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_o & \boldsymbol{\Sigma}_{o,h} \\ \boldsymbol{\Sigma}_{o,h}^* & \boldsymbol{\Sigma}_h \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{\Theta}_o & \boldsymbol{\Theta}_{o,h} \\ \boldsymbol{\Theta}_{o,h}^* & \boldsymbol{\Theta}_h \end{bmatrix}^{-1} \in \mathbb{R}^{n \times n}. \tag{6}$$

From linear algebra (Schur complement):

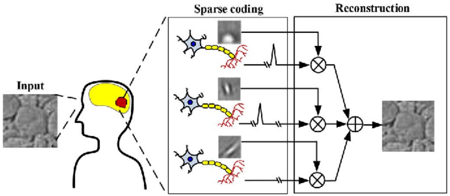
$$\textbf{Fact:} \quad \boldsymbol{\Sigma}_o^{-1} = \boldsymbol{\Theta}_o - \boldsymbol{\Theta}_{o,h} \boldsymbol{\Theta}_h^{-1} \boldsymbol{\Theta}_{o,h}^* \quad \in \mathbb{R}^{n_o \times n_o}. \tag{7}$$

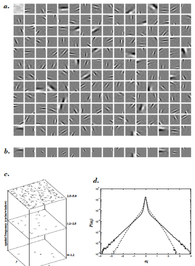
Hence to infer $\boldsymbol{\Theta}$ from the observable $\boldsymbol{\Sigma}_o$, we need to solve a problem of **sparse plus low-rank** decomposition:

$$\underbrace{\boldsymbol{\Sigma}_o^{-1}}_{\text{observed}} = \underbrace{\boldsymbol{S}}_{\text{sparse}} + \underbrace{\boldsymbol{L}}_{\text{low rank}} \quad \in \mathbb{R}^{n_o \times n_o}. \tag{8}$$

History: Nature and Neuroscience

Dogma for natural vision [Barlow 1972]: *"... to represent the input as completely as possible by activity in as few neurons as possible."*





Find sparse $\{x_i\}$ such that

$$\boldsymbol{y} = \sum_{i=1}^{n} x_i \boldsymbol{a}_i + \boldsymbol{\epsilon} \quad \in \mathbb{R}^m, \qquad (9)$$

[Nature, Olshausen and Field 1996.]

History: Signal Processing

Model y as a **linear function** of variables $a_1, \ldots, a_n$:

$$y = f(\boldsymbol{a}) = \boldsymbol{a}^* \boldsymbol{x} = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n, \tag{10}$$

from measurements

$$y_i = \boldsymbol{a}_i^* \boldsymbol{x} + \epsilon_i, \quad i = 1, 2, \ldots, m, \tag{11}$$

where ϵ_i is possible measurement noise or error.

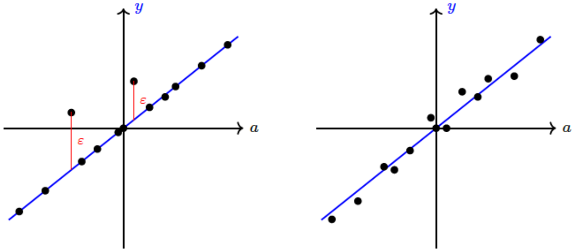


Figure: Left: $\epsilon \sim \frac{1}{2b} \exp\left(-\frac{|\epsilon|}{b}\right)$; Right: $\epsilon \sim \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right)$

History: Error Correction and Denoising ($m > n$)

Least Absolute Deviations [Roger Joseph Boscovich, 1750]:

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1 = \sum_{i=1}^{m} |y_i - \boldsymbol{a}_i^* \boldsymbol{x}|, \quad \epsilon \sim \frac{1}{2b} \exp\left(-\frac{|\epsilon|}{b}\right). \tag{12}$$

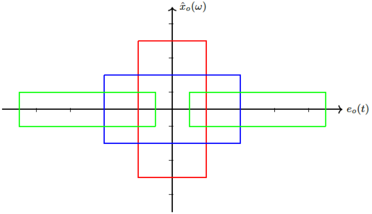
Least Squares [Legendre in 1805 and Gauss in 1809]:

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 = \sum_{i=1}^{m} (y_i - \boldsymbol{a}_i^* \boldsymbol{x})^2, \quad \epsilon \sim \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{13}$$

Error Correction [Ben Logan 1960]:

$$y(t) = x(t) + e(t): \quad \min \|x - y\|_1 \quad \text{subject to} \quad x \in \mathcal{B}_1(\Omega). \tag{14}$$

Logan's Phenomenon: $|T| \times \Omega < \frac{\pi}{2}$.



More Recent History: Linear Regression ($m < n$)

Best Subset Selection [Hocking, Leslie, and Beale 1967]:

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{x}\|_0 \le k, \tag{15}$$

Stepwise Regression ($\mathcal{P}_I$ projection on $\boldsymbol{A}_I$) [Efroymson 1966]:

$$i_k = \arg\min_{i \notin I_k} \|\boldsymbol{y} - \mathcal{P}_{I_k \cup \{i\}}(\boldsymbol{y})\|_2^2, \quad I_{k+1} = I_k \cup \{i_k\}. \tag{16}$$

Lasso Regression [Tibshirani 1996]:

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{x}\|_1 \le k. \tag{17}$$

Basis Pursuit [Chen, Donoho, and Saunders 1998] :

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}. \tag{18}$$

History: Principal Component Analysis (PCA)

[Pearson 1901, Hotelling 1933]

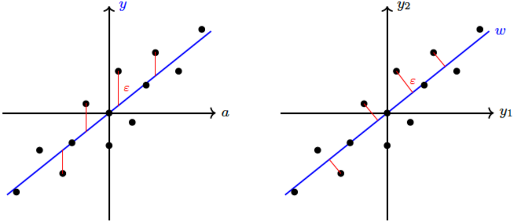


Figure: Left: regression; Right: principal component analysis.

A high-dim random vector $\boldsymbol{y}$ is approximated by the $d < m$ components as:

$$\boldsymbol{y} = \boldsymbol{u}_1 w_1 + \boldsymbol{u}_2 w_2 + \cdots + \boldsymbol{u}_d w_d + \boldsymbol{\epsilon} \doteq \boldsymbol{U}\boldsymbol{w} + \boldsymbol{\epsilon} \quad \in \mathbb{R}^m, \tag{19}$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d] \in \mathbb{R}^{m \times d}$, $\boldsymbol{w} = [w_1, w_2, \ldots, w_d]^* \in \mathbb{R}^d$, and the variance of the residual $\boldsymbol{\epsilon} \in \mathbb{R}^m$ is minimized:

$$\min \mathbb{E}\left[\|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{w}\|_2^2\right]. \tag{20}$$

History: Low-Rank Matrix Approximation

[Eckart and Young 1936]

$$\textbf{A Matrix of Samples:} \quad \boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n] \quad \in \mathbb{R}^{m \times n}. \tag{21}$$

Matrix approximation by rank-1 factors (Beltrami and Jordan 1870's):

$$\boldsymbol{Y} = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^* + \sigma_2 \boldsymbol{u}_2 \boldsymbol{v}_2^* + \cdots + \sigma_d \boldsymbol{u}_d \boldsymbol{v}_d^* + \boldsymbol{E}, \tag{22}$$

Low-rank matrix approximation:

$$\boldsymbol{X}_\star = \arg\min_{\boldsymbol{X}} \|\boldsymbol{Y} - \boldsymbol{X}\|_2^2 \quad \text{subject to} \quad \text{rank}(\boldsymbol{X}) \leq d. \tag{23}$$

Solution via Singular Value Decomposition:

$$\boldsymbol{X}_\star = \boldsymbol{U}_d \boldsymbol{\Sigma}_d \boldsymbol{V}_d^*, \tag{24}$$

where $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ be the SVD of the matrix $\boldsymbol{Y} \in \mathbb{R}^{m \times n}$.

A Long and Rich History...

*A **long and rich history** of robust estimation with error correction and missing data imputation:*




R. J. Boscovich. *De calculo probailitatum que respondent diversis valoribus summe errorum post plures observationes …, before 1756*

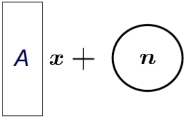
A. Legendre. *Nouvelles methodes pour la determination des orbites des cometes*, 1806

C. Gauss. *Theory of motion of heavenly bodies, 1809*

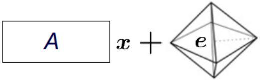
A. Beurling. *Sur les integrales de Fourier absolument convergentes et leur application a une transformation functionelle*, 1938

B. Logan. *Properties of High-Pass Signals*, 1965

⋮

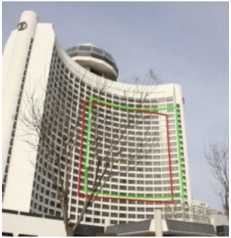


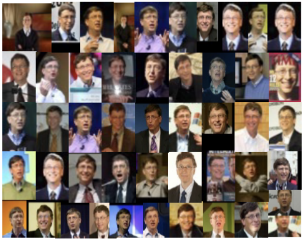
over-determined
+ dense, Gaussian



underdetermined
+ sparse, Laplacian

Why a Shift of Paradigm?





Real application data often contain ***missing observations***, ***corruptions***, *or subject to unknown* ***deformation or misalignment.***

Classical methods (e.g., PCA, least square regression) break down…

From Curses to Blessings of High-Dimensionality

For problems of identifying low-dimensional (e.g. sparse or low-rank) structures of massive data in high-dimensional spaces, we like to answer **two fundamental questions:**

1. Why many seemingly intractable high-dimensional problems can be solved efficiently **without suffering the curses of dimensionality**? (seemingly NP-hard, combinatorial, exponential, astronomical scale...)

2. What is the precise characterization of the required **data complexity and computational complexity** for certain guaranteed accuracy or probability of success?
   (number of samples needed, number of oracles computed...)

Our goal: develop a principled mathematical foundation to answer above two questions for the following problems:

- **Compressive Sensing** (Parsimony):

$$\min_{\boldsymbol{x}} \|\boldsymbol{x}\|_1 \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \tag{25}$$

- **Error Correction** (Robustness):

$$\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_1, \quad \text{with} \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}. \tag{26}$$

- **Deep Learning** (Nonlinearlity):

$$\begin{cases} \boldsymbol{z}_{\ell+1} &= \phi(\boldsymbol{A}^{\ell}\boldsymbol{z}_{\ell}), \quad \boldsymbol{z}_0 = \boldsymbol{x}, \quad \ell = 0, 1, \ldots, L-1, \\ \boldsymbol{y} &= \phi(\boldsymbol{C}\boldsymbol{z}_L), \end{cases} \tag{27}$$

  where $\phi(\cdot)$ is typically *sparsity-promoting* nonlinear activation.

Guarantees: High-Dim Geometry and Statistics

Minimal data that ensure tractable method for a correct solution?

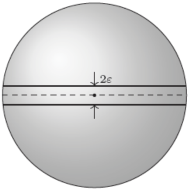
Phenomena against intuition from low-dim spaces:

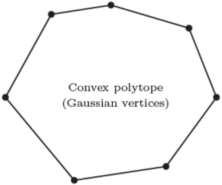
- **Measure Concentration** ($\epsilon \sim O(n^{-1/2})$)

$$\text{Area}\{\boldsymbol{x} \in \mathbb{S}^{n-1} : -\epsilon \leq x_n \leq \epsilon\} = 0.99 \cdot \text{Area}(\mathbb{S}^{n-1}), \tag{28}$$

- **Neighborly Polytopes** (vertices from a Gaussian matrix):

$$\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n] \in \mathbb{R}^{m \times n}.$$





Computational Cost: Scalable Optimization

Minimal computational cost that ensures an accurate solution?

Solutions defying conventional wisdoms:

- **Convex Optimization**: accelerated first order methods, augmented Lagrangian method, alternating minimization, etc.
- **Nonconvex Optimization**: symmetry, stochastic gradient descent, generalized power iteration, etc.

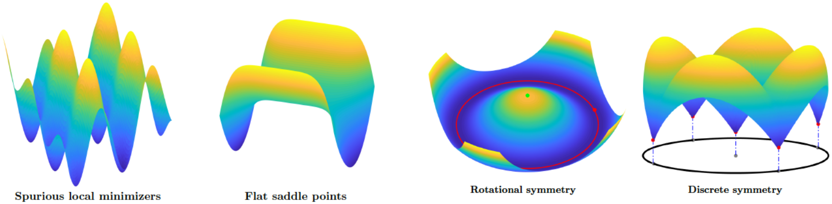
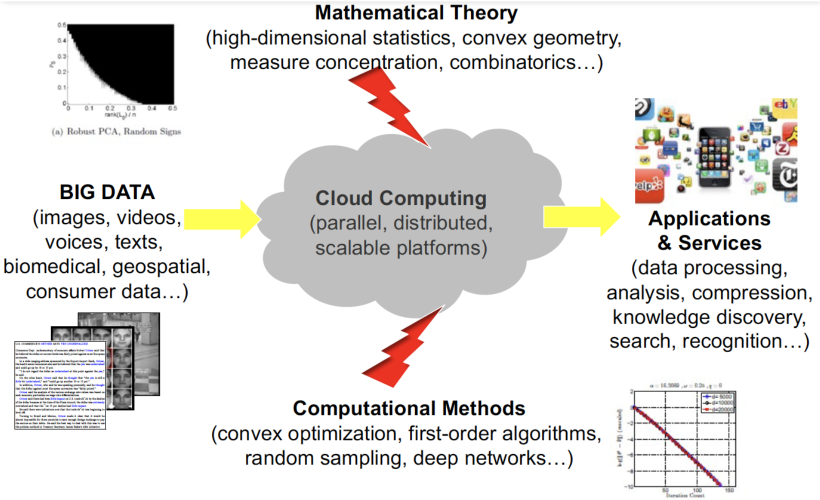


Figure: Left: conventional view. Right: actual landscape.

The first 20 years of the century

Figure: A **perfect storm** for unprecedented confluence and advancement in mathematics, computation, technology, and science.



Homework 0

- Reading Assignment I: Preface
- Reading Assignment II: Introduction (Chapter 1)
- Reading Assignment III: Linear Algebra (Appendix A)