A Mathematical Introduction to Data Science

Lecture 0: Introduction

Yuan Yao

Peking University

2015.4.30.

國立交通大學

National Chiao Tung University

Short Course Information

- Instructor: 姚远
- Email: yuany@math.pku.edu.cn
- Base Course Website: http://www.math.pku.edu.cn/teachers/yaoy/Fall2014
- Ebanshu public lectures: http://www.ebanshu.com/
- Time & Venue:
  - Seminar 1, A Dynamic Approach to Sparse Recovery, May 1, NCTS
  - Lecture 1, 2, 3: May 5, 6, 7; 3-5pm, 107 Hung-Ching Bldg, NCU
  - Seminar 2, Applied Hodge Theory, May 8, NCTS

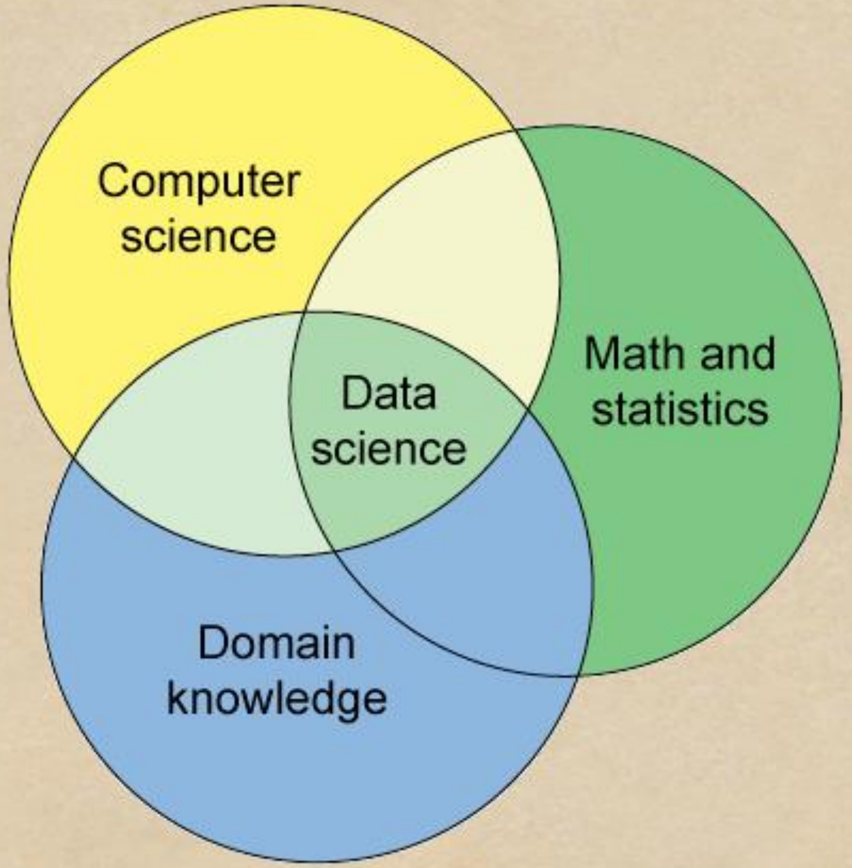
What's Data Science

- ***Data science*** is the study of the generalizable extraction of knowledge from data.
- Reference: wikipedia
- Dhar, V. (2013). "Data science and prediction". *Communications of the ACM* **56** (12): 64.

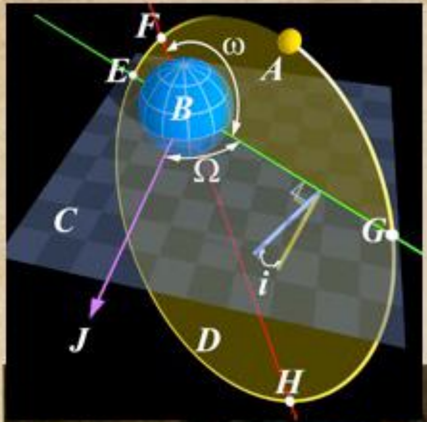
Where the term from...

- 1960, Peter Naur, Computer Scientist
- 1972, John W. Tukey, Mathematician
  - It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: intellectual adventure, demanding calls upon insight, and a need to find out "how things really are" by investigation and the confrontation of insights with experience. (Tukey's definition of 'Data Science'?)
- 1997, C. F. Jeff Wu (吴建福), Statistician
  - Statistics = Data Science?

Data Science is highly interdisciplinary



http://www.ibm.com/developerworks/jp/opensource/library/os-datascience/figure1.png



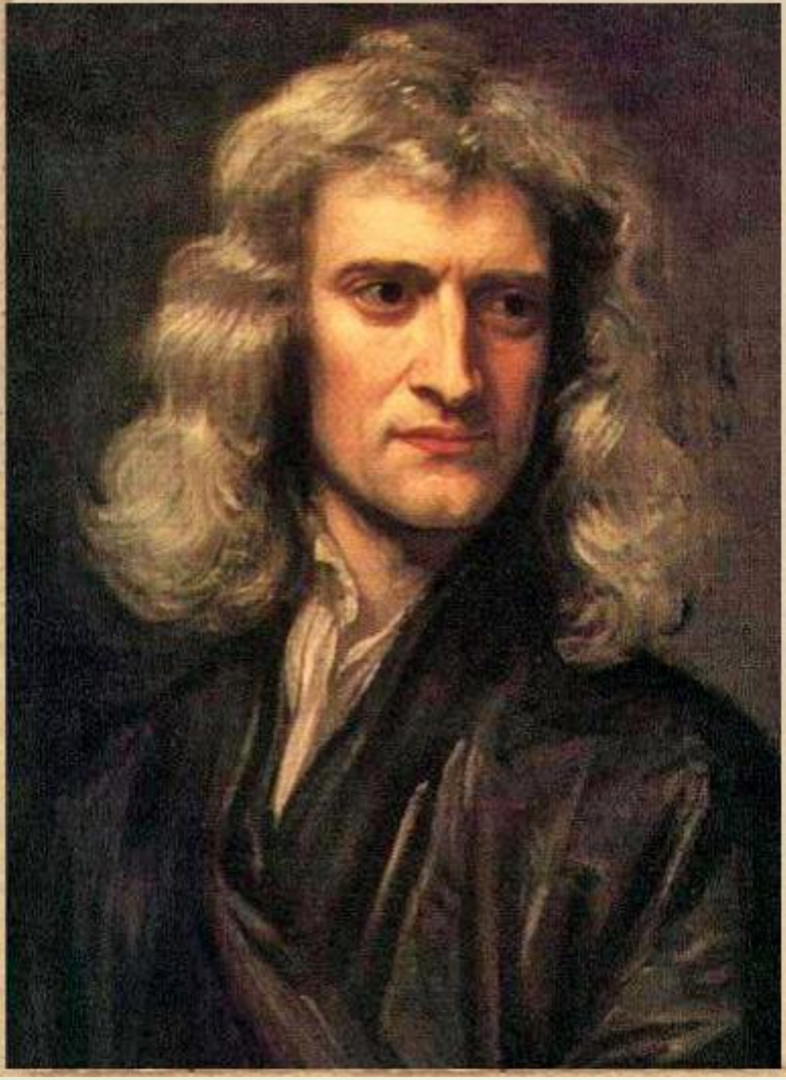
Johannes Kepler 1618

- 3 laws of planetary motion

Planet	Period (yr)	Average Distance (au)	T^2/R^3 (yr^2/au^3)
Mercury	0.241	0.39	0.98
Venus	.615	0.72	1.01
Earth	1.00	1.00	1.00
Mars	1.88	1.52	1.01
Jupiter	11.8	5.20	0.99
Saturn	29.5	9.54	1.00
Uranus	84.0	19.18	1.00
Neptune	165	30.06	1.00
Pluto	248	39.44	1.00

(**NOTE**: The average distance value is given in astronomical units where 1 a.u. is equal to the distance from the earth to the sun - 1.4957×10^{11} m. The orbital period is given in units of earth-years where 1 earth year is the time required for the earth to orbit the sun - 3.156×10^{7} seconds.)

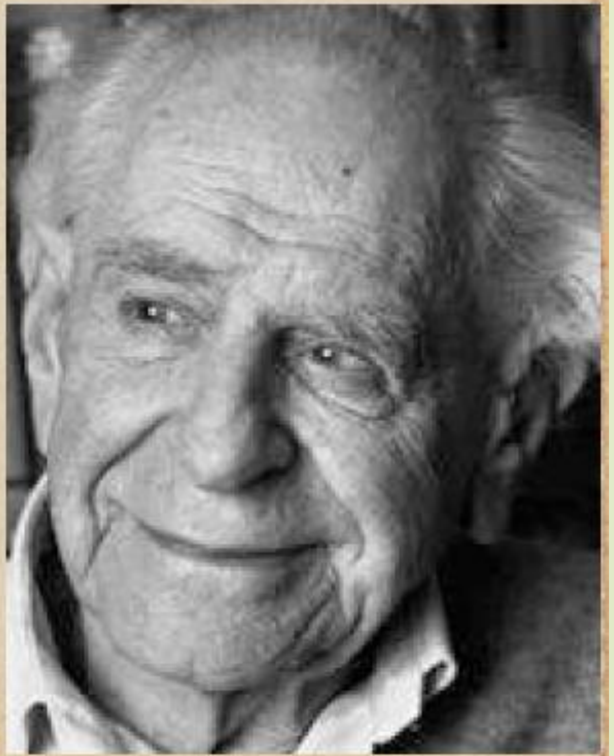
Issac Newton's learning



- Force f = m a
- Gratività f = G Mm/a
- => Kepler's law

Karl Popper 1950s

- Falsifiability = Science vs. Pseudoscience
- A theory in the empirical sciences can never be proven, but it can be falsified, meaning that it can and should be scrutinised by decisive experiments.



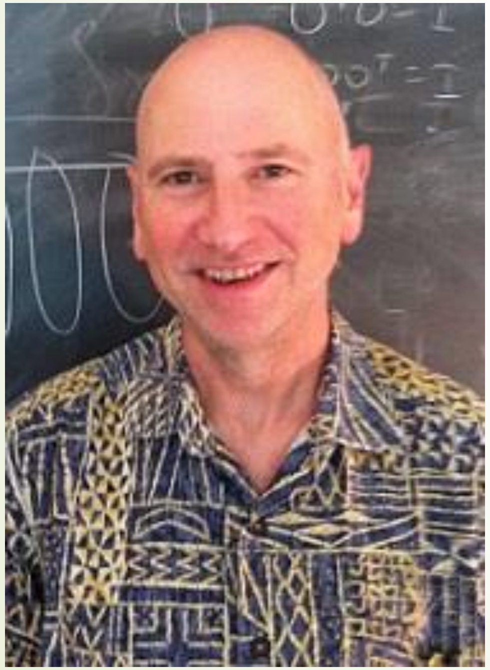
Occam's Razor

- The hypothesis has to be as simple as possible, but not simpler. (Einstein)

How to Quantify the Falsifiability and Occam's Razor

- Information Complexity for model uncertainty
  - V. Vapnik, Statistical Learning Theory, 1970s
  - Kearns and Vazirani, An Introduction to Computational Learning Theory
- Yet big data brings up computational challenges…

Larry Wasserman in his "All of Statistics"



Statistics/Data Mining Dictionary

Statisticians and computer scientists often use different language for the same thing. Here is a dictionary that the reader may want to return to throughout the course.

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from X
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains an unknown quantity with given frequency
directed acyclic graph	Bayes net	multivariate distribution with given conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update beliefs
frequentist inference	—	statistical methods with guaranteed frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

Data have structures, mathematically universal

- Sparsity structure lies in the core of high dimensional data analysis
  - Low dimensional vector spaces
  - Low rank matrices, tensors, etc.
- Lecture 1: Sample mean and Covariance (PCA/MDS): Fisher's Principle of Maximum Likelihood Estimate, yet things might go wrong --
  - Stein's phenomenon and shrinkage
  - Random matrix theory and failure of PCA
- Lecture 2: Generalized PCA/MDS
  - Random projections and compressed sensing
  - PCA/MDS with uncertainty
  - Nonlinear manifold learning
- Lecture 3: topological and geometric structures of data
  - From graphs to simplicial complexes

Seminar 1: Sparse Recovery or Variable Selection via Differential Equations

- LASSO path is biased
- Nonconvex optimization/regularization is hard to find global optimizer
- A simple dynamics of dual gradient descent of LASSO leads us: model selection consistency with debias, provably under the nearly same condition for LASSO.

Seminar 2: Applied Hodge Theory

- Data Analysis on simplicial complexes, extension from graphs
- Computer Vision: optical flow decomposition for object tracking
- Statistical Ranking: preference aggregation via crowdsourcing pairwise ranking data
- Game theory: every finite game admits a direct sum of potential game and harmonic (zero-sum) game

Acknowledgement

- Weinan E, PKU and Princeton
- Amit Singer, Princeton
- Bin Yu, UC Berkeley and PKU
- As well as numerous students in the past few years.