A Short Introduction to Data Science, Big Data Analytics and Statistical Learning

Yu-Tzung Chang and Hsuan-Wei Lee

College of Social Sciences, National Taiwan University

2018.09.13

Data science as a field

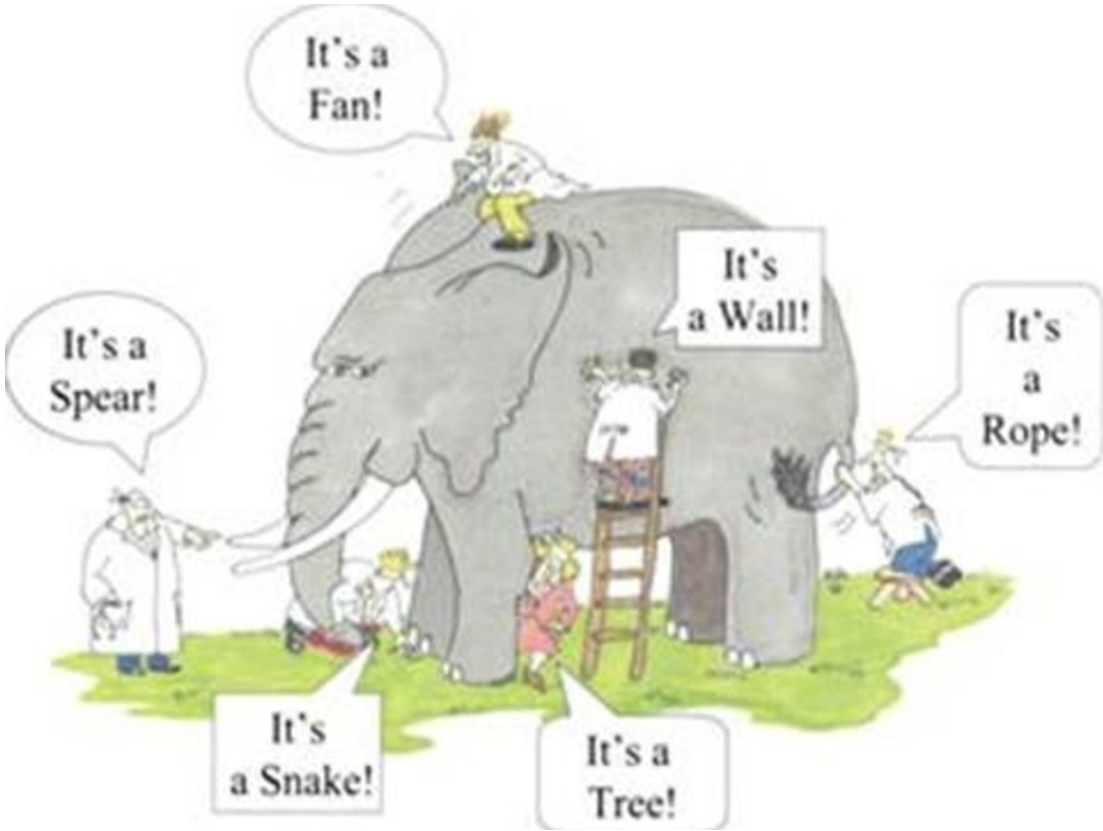
Abundance of data

- Volume of data roughly doubles in every three years – thank to the technology
  - computers
  - large scale storage devices
  - communication and sensor technologies
- Broad impacts
  - physical, biological, medical sciences
  - all branches of engineering
  - economics, social science, banking and all kinds of commerce
  - government, policy, election
  - environmental science, ecology, outer space
  - sports, recreation

What is big data analytics?

Wikipedia:
Big data is a term used to refer to the study and applications of data sets that are so big and complex that traditional data-processing application software are inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating, information privacy and data source.
There are a number of concepts associated with big data: originally there were 3 concepts volume, variety, velocity. Other concepts later attributed with big data are veracity (i.e., how much noise is in the data) and value.

What is data science?



Some quotes from the Internet

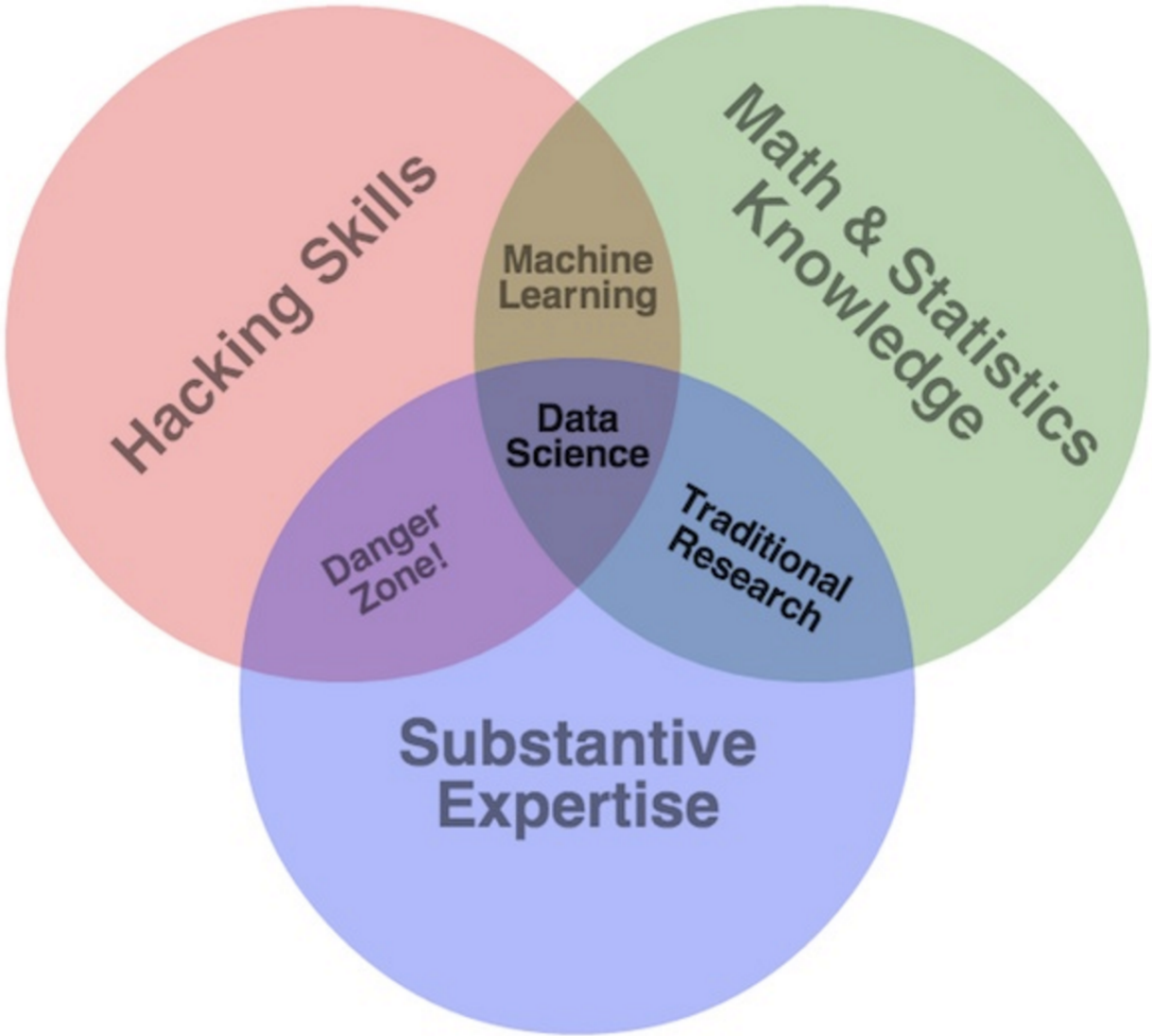
- "A data scientist is a data analyst who lives in California."
- "A data scientist is a statistician who lives in San Fransisco."
- "Data science is statistics on a Mac."
- "A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician."

Data science and prediction

- "Data science is the study of the generalizable extraction of knowledge from data"*
- A key epistemic requirement for new knowledge is its ability to **predict** and not just **explain**
- Compare: the benefit of knowing things in advance vs. the cost of being wrong

[1]Dhar V. Data Science and Prediction, Communications of the ACM, Vol. 56 No.12, December 2013

Data science overview



[1] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Related disciplines

- Mathematical and statistical modeling
- Data assimilation
- Estimation theory
- Regression analysis
- Time series analysis
- System identification
- Adaptive optimization
- Inverse probelm

Goal of data science – prediction

- Need for prediction is all pervasive
- Understand the patterns of nature (e.g. natural science)
- Prevent damages (e.g. weather prediction, crime prevention, disease propagation)
- Enhance welfare (e.g. medical diagnosis, automation)
- Speculation (e.g. investment, hedge funds, election prediction)

What makes prediction so hard?

- Noise (physical versus social systems)
- Not knowing the right question to ask
- Not having the right/enough data
- Not having enough understanding of the patterns
- Believing in the modeling too much
- Chaotic theory

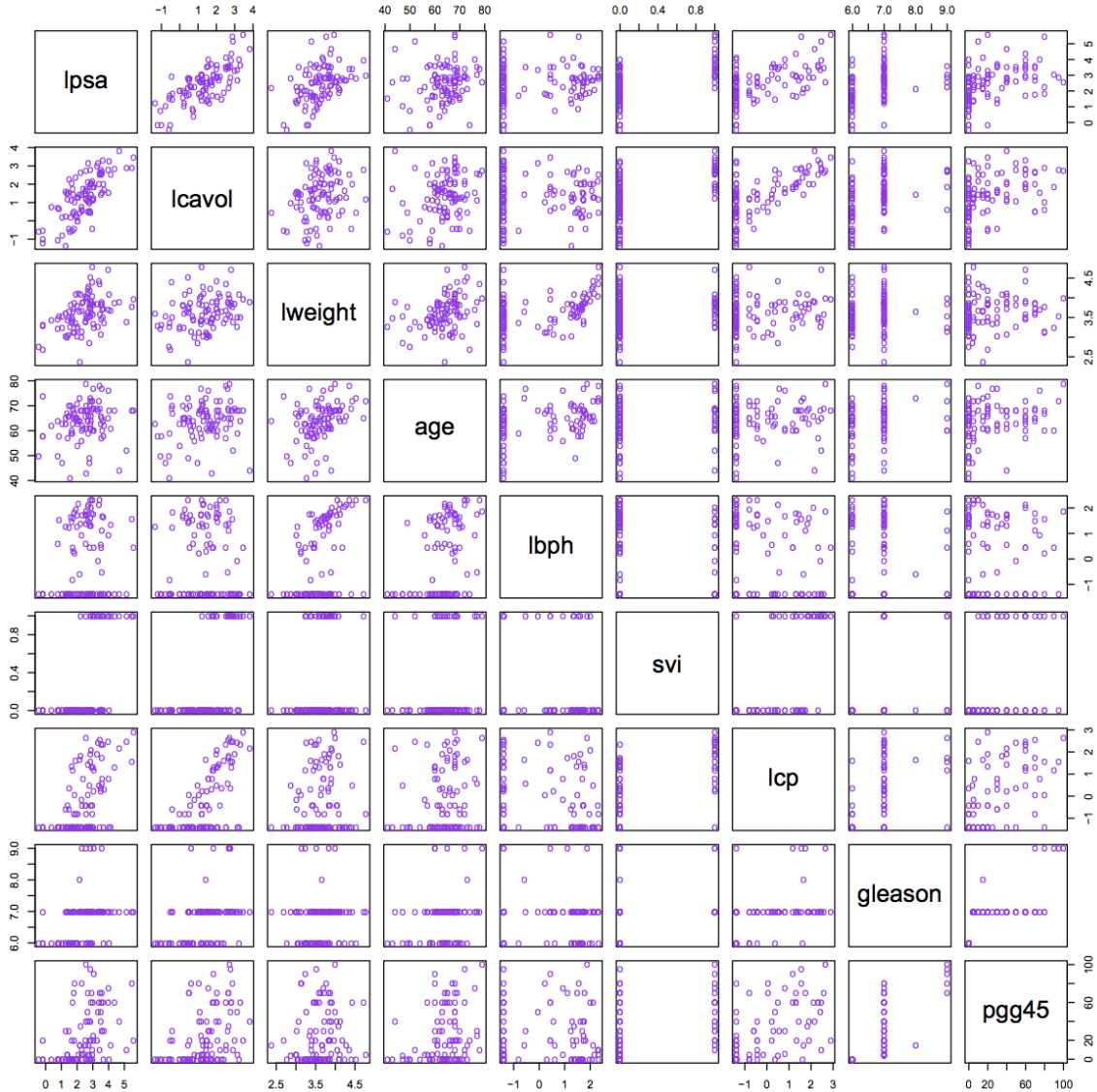
A mathematical and statistical approach

- A very vast discipline
- Key players: data, model, the process of fitting
- Four steps:
  - data mining:
    discover basis laws from the analysis of data
  - data assimilation:
    an inverse problem to estimate the unknowns
  - model selection:
    compare and evaluate models using some criteria
  - generate predictions:
    a direct problem to forecast

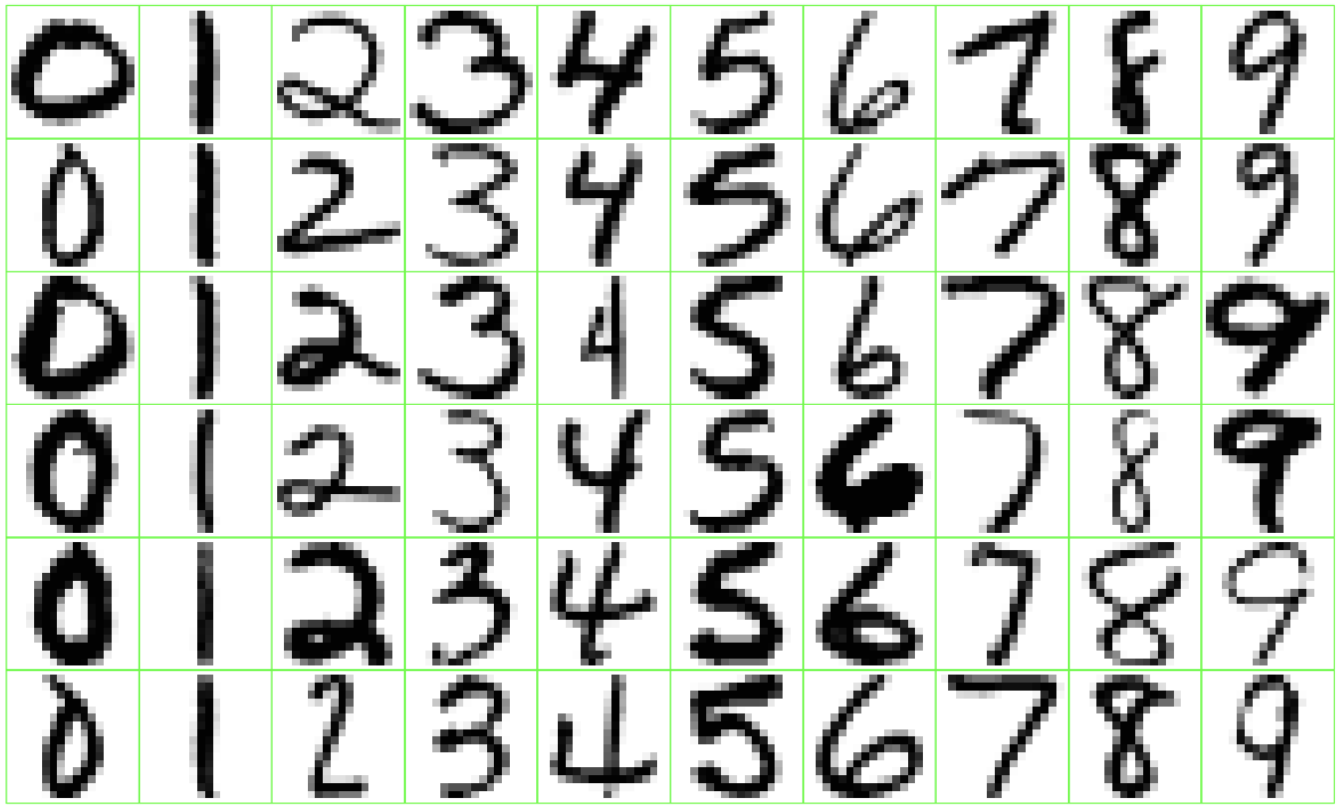
Statistical learning problems (email spam)

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

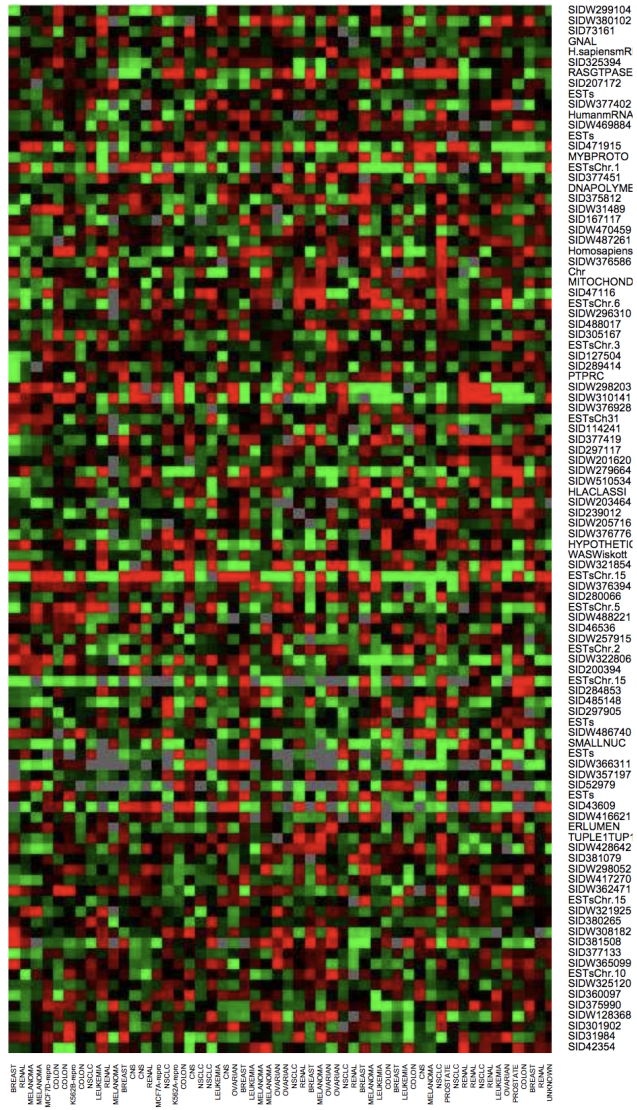
Statistical learning problems (prostate cancer)



Statistical learning problems (handwritten digit recognition)



Statistical learning problems (DNA expressions microarrays)



Different language for the same thing

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from $X \in \mathcal{X}$
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 unknown quantity with a prescribed frequency
directed acyclic graph	Bayes net	multivariate distribution with specified conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update subjective beliefs
frequentist inference	—	statistical methods for producing point estimates and confidence intervals with guarantees on frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

[1] All of Statistics, Larry A. Wasserman, Springer

Topics of this semester

- Linear regression
- Classification
- Resampling methods
- Linear model selection and regularization
- Moving beyond linearity
- Tree-based methods
- Unsupervised Learning