What is Big Data

Junping Shi

College of William and Mary, USA

Math 410

January 22, 2014

EXTREEMS-QED

Expeditions in Training, Research, and Education for Mathematics and Statistics through Quantitative Explorations of Data (EXTREEMS-QED) program is an National Science Foundation educational program to support efforts to educate the next generation of mathematics and statistics undergraduate students to confront new challenges in computational and data-enabled science and engineering (CDS&E). EXTREEMS-QED projects will enhance the knowledge and skills of mathematics majors through training that incorporates computational tools for analysis of large data sets and for modeling and simulation of complex systems.

CDS&E: A New Discipline

CDS&E is now clearly recognizable as a distinct intellectual and technological discipline lying at the intersection of applied mathematics, statistics, computer science, core science and engineering disciplines. It is dedicated to the development and use of computational methods and data mining and management systems to enable scientific discovery and engineering innovation.

We regard CDS&E as explicitly recognizing the importance of data-enabled, data-intensive, and data centric science. CDS&E broadly interpreted now affects virtually every area of science and technology, revolutionizing the way science and engineering are done. Theory and experimentation have for centuries been regarded as two fundamental pillars of science. It is now widely recognized that computational and data-enabled science forms a critical third pillar.

W&M EXTREEMS-QED

Website:
http://www.wm.edu/as/mathematics/undergraduate_research/EXTREEMS-QED/index.php
News Story:
http://www.wm.edu/as/mathematics/news/EXTREEM.php

Grant: NSF DMS-1331021 (EXTREEMS-QED: Computational and Statistical theory and techniques in the study of large data sets) 2013-2018, $880K
http://www.nsf.gov/awardsearch/showAward?AWD_ID=1331021&HistoricalAwards=false

Principal Investigator: Junping Shi
Co-Principal Investigators: Tanujit Dey, Chi-Kwong Li, Gexin Yu

All W&M EXTREEMS-QED faculty members:
http://www.wm.edu/as/mathematics/undergraduate_research/EXTREEMS-QED/faculty/index.php

W&M EXTREEMS-QED courses

Spring 2014

MATH 352: Data Analysis (Tanujit Dey)
MATH 410-01/CSCI 688-02: Internet Algorithms & Econ (Anke van Zuylen)
MATH 410-07: Analysis of Big Data (Tanujit Dey, Junping Shi)
MATH 452/552: Mathematical Statistics (Tanujit Dey)
CSCI 618: Model/Applications Operation Research (Rex Kincaid)
CSCI 678: Analysis of Simulation Models (Larry Leemis)
CSCI 688-01: Combinatorial Optimization (Frans Schalekamp)

Fall 2014

Math 410: Matrices and Statistics (Chi-Kwong Li, Ross Iaci)

W&M EXTREEMS-QED summer research program

- Summer research program by teams of undergraduate students and faculty members starting summer 2014.
- 8-10 W&M undergrad students, and 2-4 undergrad students from Virginia State U, Hampton U, and Norfolk State U.
- Stipend $4000 and free summer on-campus housing
- Eligibility: math major, US citizen/permanent residents
- Application deadline: March 15, 2014.
- Research Program: May 26-July 18 (8 weeks)

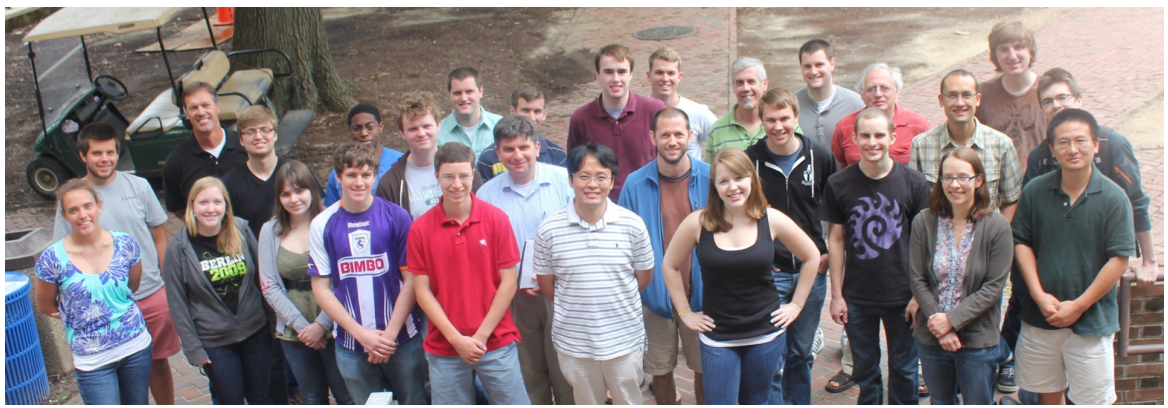
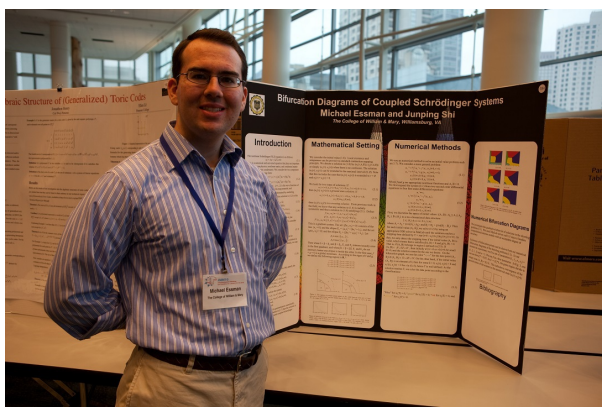
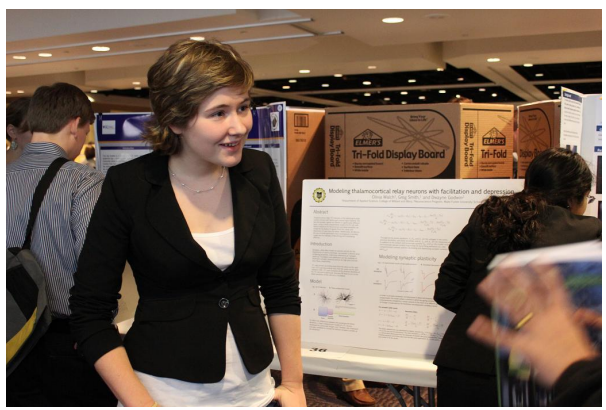


Photo: CSUMS program summer 2012

Other W&M EXTREEMS-QED activities

- Short courses in summer (topics in statistics, network theory)
- Lectures by external speakers (March 21, Don Brown, U Va)
- Visiting large data centers and computing facilities in summer (NSA, DOE Jefferson Lab, NASA research center)
- Participating research conferences to present your work (2015 Joint Math Meeting, San Antonio)





CSUMS in Joint Meeting. Left: Michael Essman, 2010, San Francisco; Right: Olivia Walch, 2011, New Orleans

What is Big Data

Everyone talks about it

1. (Mar 2012) Obama Administration unveils "Big data" initiative: announces $200 million in new R&D investment
2. New York Times (Feb 2012): *The Age of Big Data*
3. Harvard Business Review (Oct 2012): (i) *Big Data: The Management Revolution*; (ii) *Data Scientist: The Sexiest Job of the 21st Century*
4. Wall Street Journal (Mar 2013): *How Big Data Is Changing the Whole Equation for Business*
5. Wall Street Journal (Apr 2012): *Big Data's Big Problem: Little Talent* (nobody really knows how to do it)
6. Since 2012, new data science institutes (initiatives) have been established in MIT, UC Berkeley, Columbia U, Duke U, NYU, U Virginia and many other universities (everyone thinks everyone else is doing it, so everyone claims they are doing it)

What is big data: definitions

From `http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/`

"And yet ask a chief technology officer to define big data and he or she will will stare at the floor. Chances are, you will get as many definitions as the number of people you ask. And thats a problem for anyone attempting to buy or sell or use big data services–what exactly is on offer?"

1. Gartner. In 2001, a Meta (now Gartner) report noted the increasing size of data, the increasing rate at which it is produced and the increasing range of formats and representations employed. This report predated the term " big data" but proposed a three-fold definition encompassing the "three Vs": Volume, Velocity and Variety. This idea has since become popular and sometimes includes a fourth V: Veracity, to cover questions of trust and uncertainty.
2. Oracle. Big data is the derivation of value from traditional relational database-driven business decision making, augmented with new sources of unstructured data.
3. Intel. Big data opportunities emerge in organizations generating a median of 300 terabytes of data a week. The most common forms of data analyzed in this way are business transactions stored in relational databases, followed by documents, e-mail, sensor data, blogs, and social media.

What is big data: definitions

From `http://www.technologyreview.com/view/519851/the-big-data-conundrum-how-to-define-it/`

4. Microsoft. Big data is the term increasingly used to describe the process of applying serious computing power–the latest in machine learning and artificial intelligence–to seriously massive and often highly complex sets of information.
5. The Method for an Integrated Knowledge Environment open-source project. The MIKE project argues that big data is not a function of the size of a data set but its complexity. Consequently, it is the high degree of permutations and interactions within a data set that defines big data.
6. The National Institute of Standards and Technology. NIST argues that big data is data which "exceed(s) the capacity or capability of current or conventional methods and systems." In other words, the notion of "big" is relative to the current standard of computation.
7. Their definition: Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning.
7. Your definition ?

How big is the data

1000=kilobyte(KB); 1000^2=megabyte(MB)
1000^3=gigabyte(GB); 1000^4=terabyte(TB)
1000^5=petabyte(PB); 1000^6=exabyte(EB)
1000^7=zettabyte(ZB); 1000^8=yottabyte(YB).

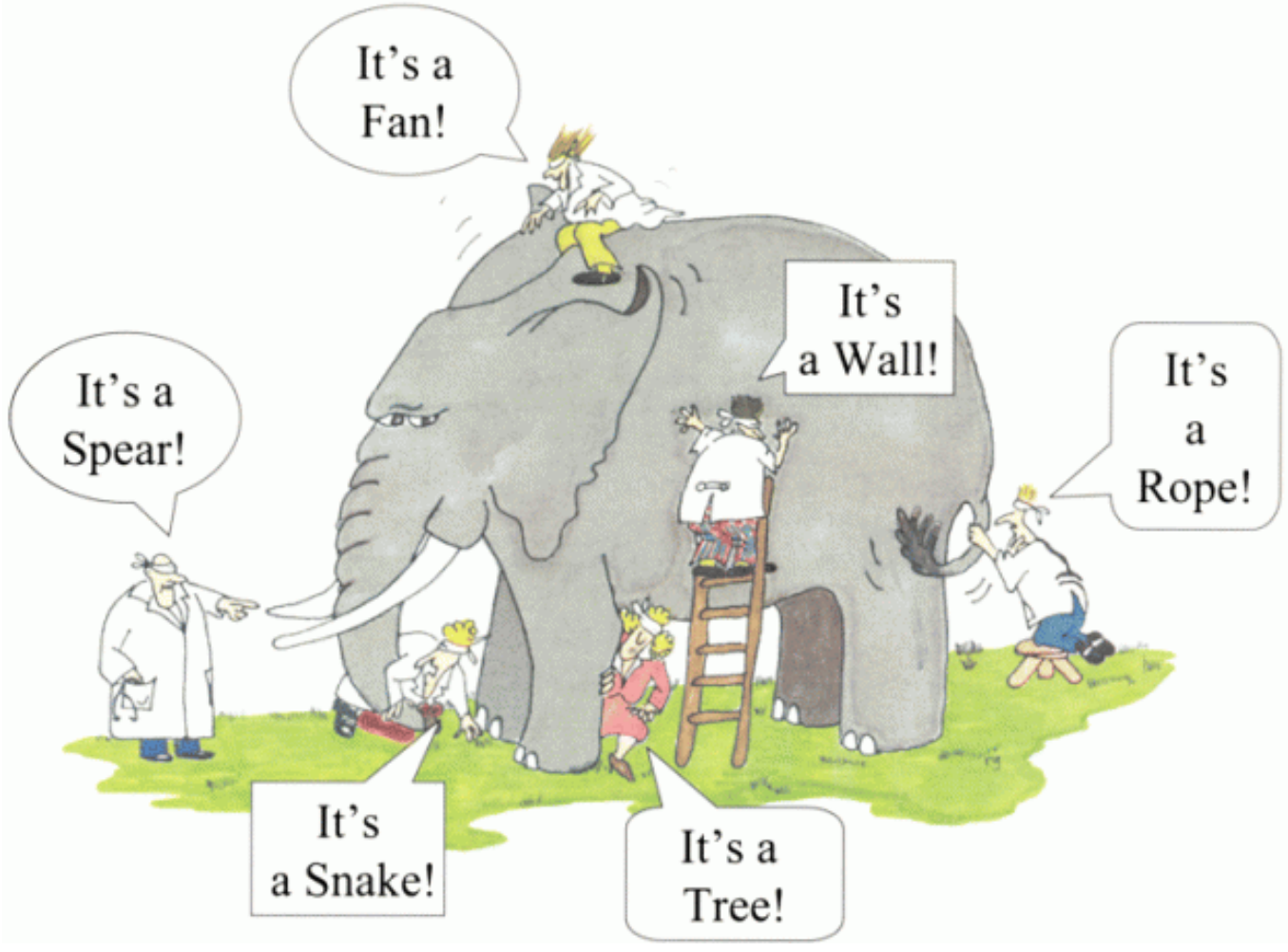
- As of 2012, every day 2.5 exabytes (2.5×10^{18}) of data were created, with 2012 total about 2.7 zettabytes
- Facebook generates 10 terabytes data daily, and Twitter generates 7 terabytes daily
- Wal-Mart handles more than 1 million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes.

"Big" means big volume, big velocity, and big variety.

Types of data

- Signal (time series) $f : \mathbb{R} \to \mathbb{R}$, or $f : \mathbb{Z} \to \mathbb{R}$
- Image: $f : [0,1] \times [0,1] \to \mathbb{R}$ (black-white) of $f : [0,1] \times [0,1] \to \mathbb{R}^3$ (colored)
- Spreadsheet: $N \times M$ matrix $A = (a_{ij})_{N \times M}$
- Higher dimensional spreadsheet: matroid $A = (a_{ijkl})_{N \times M \times P \times Q}$ (4-dimensional)
- Social media data: a graph with vertices (people), nodes (connected or not) and spreadsheet (personal info)
- Spatial data (ecological observation, astronomy observation): $f : \mathbb{R} \times X \to D$ ($\mathbb{R}$=time, X=space, D=data collected at that location and time)

Mathematics and Data



We are all like blind men feeling an elephant, and big data is that elephant.

Compressed sensing (linear algebra)

Compressed sensing (also known as compressive sensing, compressive sampling, or sparse sampling) is a signal processing technique for efficiently acquiring and reconstructing a signal, by finding solutions to underdetermined linear systems. An underdetermined system of linear equations has more unknowns than equations and generally has an infinite number of solutions. In order to choose a solution to such a system, one must impose extra constraints or beliefs (such as smoothness) as appropriate.

1 Undersample

A camera or other device captures only a small, randomly chosen fraction of the pixels that normally comprise a particular image. This saves time and space.

2 Fill in the dots

An algorithm called l_1 minimization starts by arbitrarily picking one of the effectively infinite number of ways to fill in all the missing pixels.

3 Add shapes

The algorithm then begins to modify the picture in stages by laying colored shapes over the randomly selected image. The goal is to seek what's called sparsity, a measure of image simplicity.

4 Add smaller shapes

The algorithm inserts the smallest number of shapes, of the simplest kind, that match the original pixels. If it sees four adjacent green pixels, it may add a green rectangle there.

5 Achieve clarity

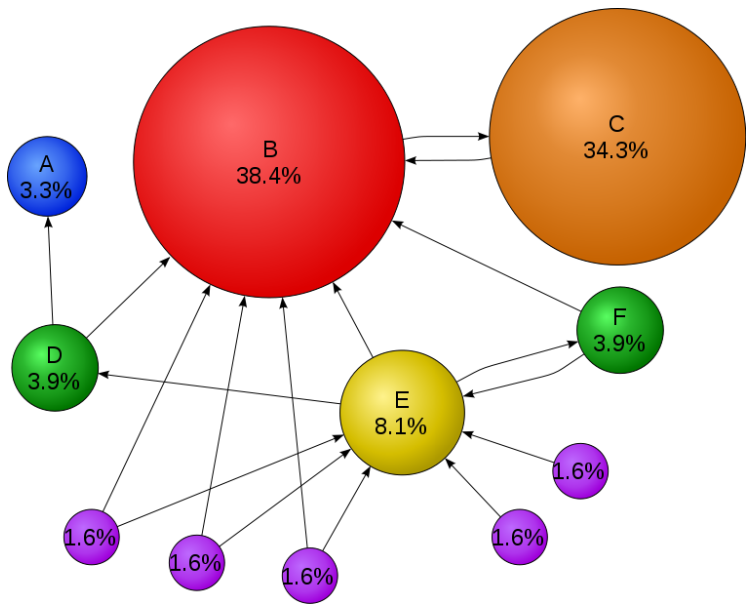
Iteration after iteration, the algorithm adds smaller and smaller shapes, always seeking sparsity. Eventually it creates an image that will almost certainly be a near-perfect facsimile of a hi-res one.

Photos: Obama: Corbis; Image Simulation: Jarvis Haupt/Robert Nowak

http://en.wikipedia.org/wiki/Compressed_sensing

Page Rank (linear algebra)

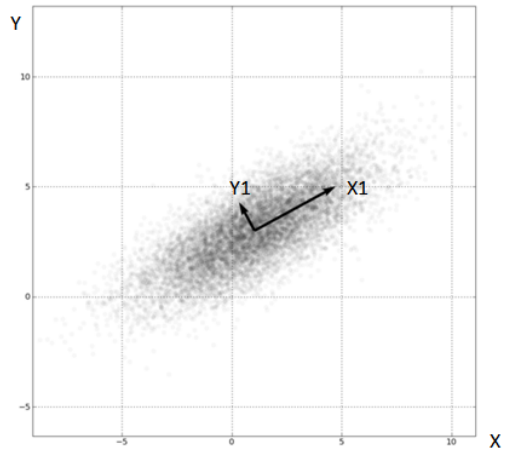
PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank is a way of measuring the importance of website pages. According to Google: "PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites."



http://en.wikipedia.org/wiki/PageRank

Principal component analysis (statistics)

Principal component analysis (PCA) is a statistical procedure that uses orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components.

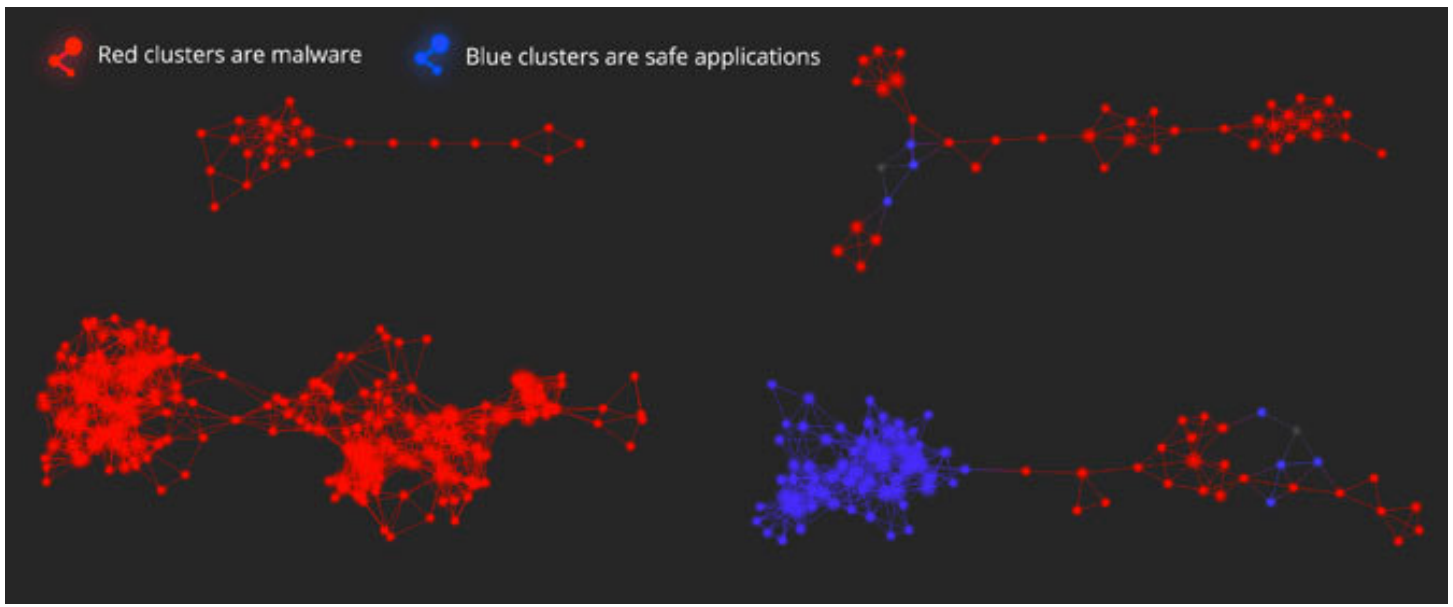


Variables X and Y appear to be correlated. They are transformed by PCA into variables X1 and Y1 which are now uncorrelated in the X1-Y1 space. We can see that X1 accounts for a larger amount of variance in the data (more spread) than Y1. Thus X1 is the first principal component. Y1 is the second principal component.

http://en.wikipedia.org/wiki/Principal_component_analysis

Topological data analysis (topology)

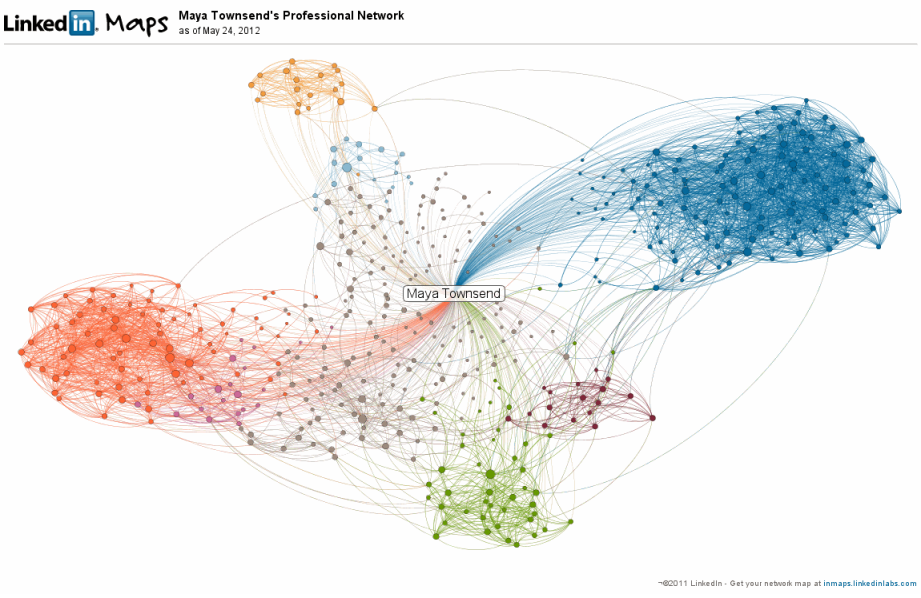
Topological data analysis (TDA) is a new area of study aimed at having applications in areas such as data mining and computer vision. The main problems are: (1) how one infers high-dimensional structure from low-dimensional representations; and (2) how one assembles discrete points into global structure. The human brain can easily extract global structure from representations in a strictly lower dimension, i.e. we infer a 3D environment from a 2D image from each eye. The inference of global structure also occurs when converting discrete data into continuous images, e.g. dot-matrix printers and televisions communicate images via arrays of discrete points.



http://en.wikipedia.org/wiki/Principal_component_analysis

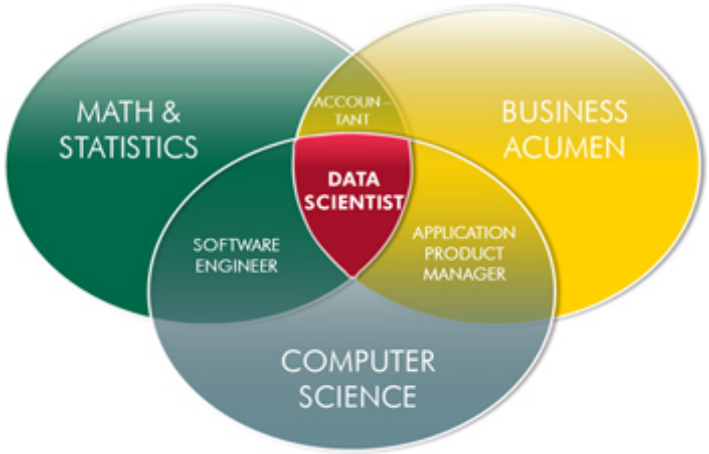
Network Science (graph theory?)

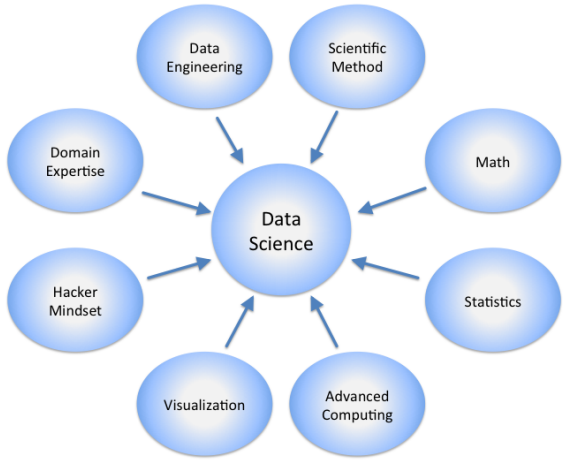
Network science is an interdisciplinary academic field which studies complex networks such as telecommunication networks, computer networks, biological networks, cognitive and semantic networks, and social networks. The field draws on theories and methods including graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science, inferential modeling from statistics, and social structure from sociology.



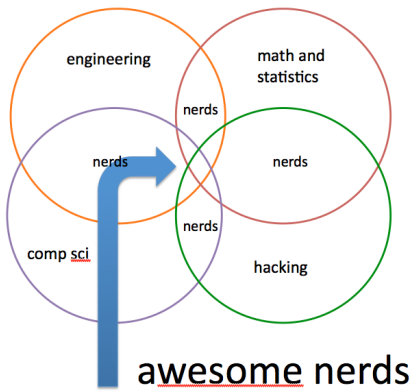
http://en.wikipedia.org/wiki/Network_science

Data Science and Data Scientists





Data scientists?



Conclusion

- Big Data Analysis (Data Science) is a hot topic in the world of business and science, and Data Scientists are in high demand for the near future.
- Big data is not a well-defined disciplinary yet. It requires knowledge in statistics, computer science, engineering, and applied (and maybe even classical) mathematics.
- In W&M EXTREEMS-QED program, we hope to provide students a solid foundation in mathematics, statistics and computer science, and students can learn about frontier of this newly emerging science.