Big (or small) data and William & Mary EXTREEMS-QED program

Junping Shi

College of William and Mary

Math 410

January 20, 2016

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: Sarah Day, 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

What is Big Data

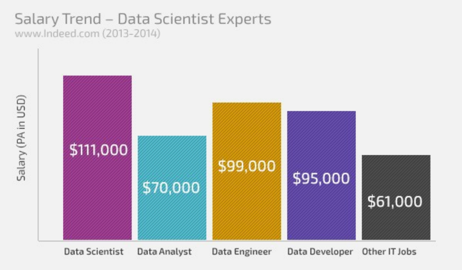
Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is
doing it, so everyone claims they
are doing it...

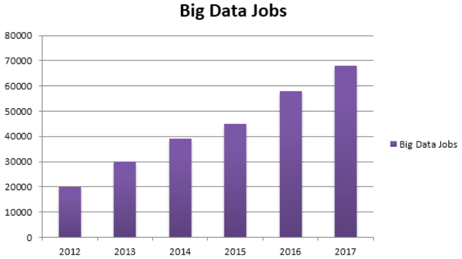
(Dan Ariely)

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)

Big Data Job Market





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 ?

What to learn in data science?

Data Science master degree program in University of Virginia
https://dsi.virginia.edu/academics

The Master of Science in Data Science (MSDS) is an 11-month professional masters program, designed to meet the increasingly data-intensive needs of industry and government. The program starts near the beginning of July and ends the next year in mid-May. Core program courses will be taught by faculty from the Departments of Computer Science, Statistics, and Systems and Information Engineering.

Three key features of this program are (a) an integrated curriculum and data experience; (b) the compressed duration; and (c) a cohort experience. To achieve these, the curriculum is tightly prescribed with about 80% common to all students. The curriculum is integrated across courses, with several large complicated data sets woven across courses to increase program cohesion. The compressed duration is designed to minimize the time from start to finish, and the cohort experience will allow students to consistently work together in teams. At the conclusion of the program the students will address an important data science challenge with their capstone experience. Students will commence this work with a proposal describing their objectives. In conducting this final exercise students will be guided, mentored, and eventually evaluated by faculty members from the different disciplines.

Data science courses in University of Virginia

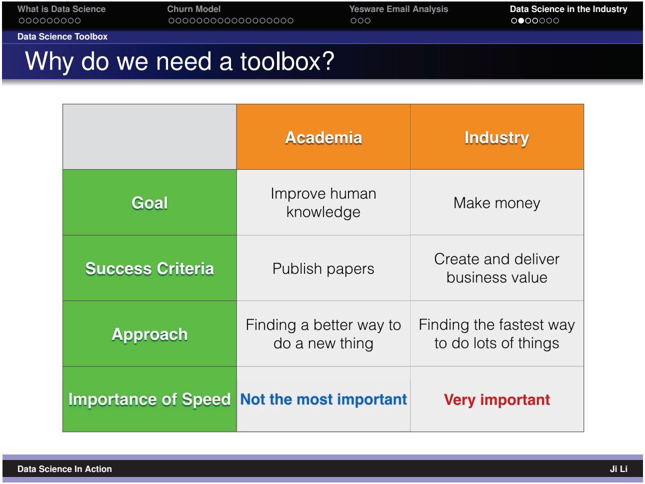
Courses: `https://dsi.virginia.edu/curriculum`

1. Summer (July-August):
   CS 5010: Programming and Systems for Data Science;
   STAT 6430: Statistical Computing for Data Science
2. Fall (September-December):
   STAT 6021: Linear Models for Data Science;
   CS 5012: Foundations of Computer Science;
   SYS 6018: Data Mining;
   DS 6001: Practice and Application of Data Science;
   DS 6002a: Ethics of Big Data;
   DS 6003a: Capstone Project;
3. Spring (January-May):
   SYS 6016: Machine Learning;
   DS 6002b: Ethics of Big Data;
   DS 6003b: Capstone Project;
   Elective 1 and 2.

Elective courses: CS 6501: Special Topics in Computer Science; CE 6400: Traffic Operations; STAT 6130: Applied Multivariate Statistics; STAT 5390: Exploratory Data Analysis; SYS 6001: Introduction to Systems Engineering; SYS 6003: Optimization I; SYS 6005: Stochastic Systems I; CS 6750: Database Systems; STAT 5170: Applied Time Series; STAT 5340: Bootstrap and Other Resampling Methods; MATH 5110: Stochastic Processes.

How about from industry point of view?

From `http://jxshix.people.wm.edu/Math410-2015/Data_Science_in_Action.pdf`
author: Ji Li (PhD, 2007, Brandeis Math; now works at FaceBook)



What is Data Science ○○○○○○○○○ Churn Model ○○○○○○○○○○○○○○○○○○○ Yesware Email Analysis ○○○ Data Science in the Industry ○●○○○○○

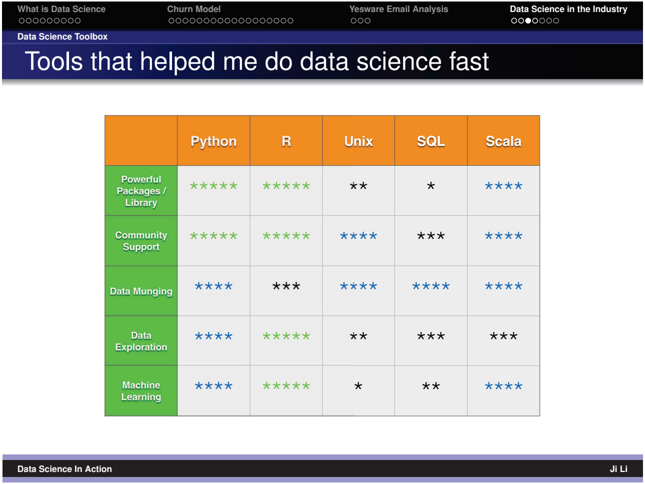
Data Science Toolbox

Why do we need a toolbox?

	Academia	Industry
Goal	Improve human knowledge	Make money
Success Criteria	Publish papers	Create and deliver business value
Approach	Finding a better way to do a new thing	Finding the fastest way to do lots of things
Importance of Speed	Not the most important	Very important

Data Science In Action Ji Li

How about from industry point of view?

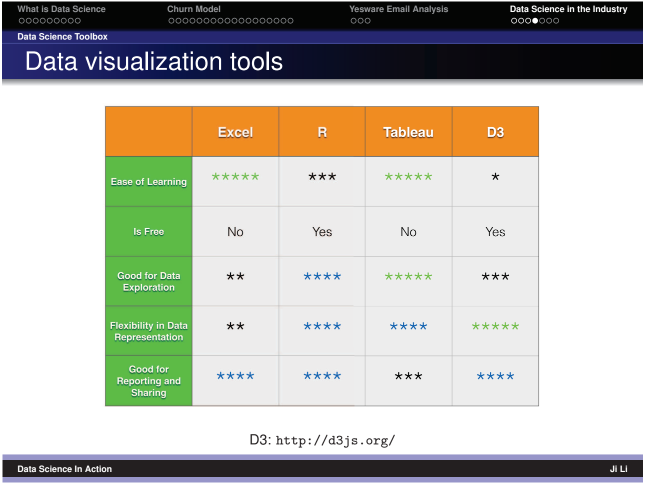


Data Science Toolbox

Tools that helped me do data science fast

	Python	R	Unix	SQL	Scala
Powerful Packages / Library	★★★★★	★★★★★	★★	★	★★★★
Community Support	★★★★★	★★★★★	★★★★	★★★	★★★★
Data Munging	★★★★	★★★	★★★★	★★★★	★★★★
Data Exploration	★★★★	★★★★★	★★	★★★	★★★
Machine Learning	★★★★	★★★★★	★	★★	★★★★

How about from industry point of view?



What is Data Science ○○○○○○○○○ Churn Model ○○○○○○○○○○○○○○○○○○ Yesware Email Analysis ○○○ Data Science in the Industry ○○○●○○○

Data Science Toolbox

Data visualization tools

	Excel	R	Tableau	D3
Ease of Learning	*****	***	*****	*
Is Free	No	Yes	No	Yes
Good for Data Exploration	**	****	*****	***
Flexibility in Data Representation	**	****	****	*****
Good for Reporting and Sharing	****	****	***	****

D3: http://d3js.org/

Data Science In Action Ji Li

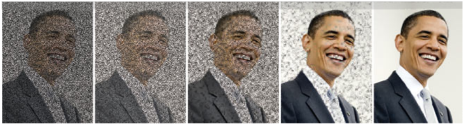
Mathematics and Data

It's a Fan!
It's a Spear!
It's a Wall!
It's a Rope!
It's a Snake!
It's a Tree!

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

Compressed sensing (linear algebra, partial differential equations)

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.

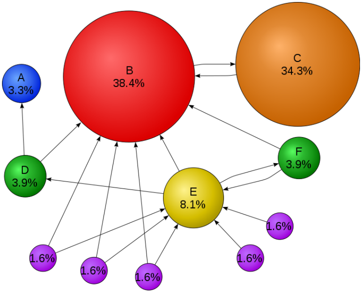
5 **Achieve clarity**

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

Page Rank (linear algebra)

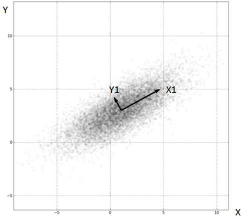
PageRank is an algorithm used by Google Search to rank websites in their search engine results. PageRank is a way of measuring the importance of website pages. According to Google: "PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the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

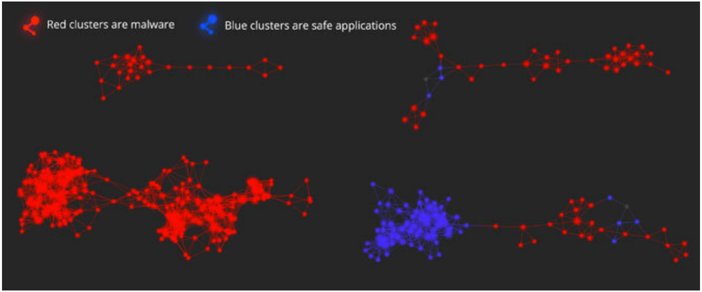


Variables X and Y appear to be correlated. They are transformed by PCA into variables X1 and Y1 which are now uncorrelated in the X1-Y1 space. We can see that X1 accounts for a larger amount of variance in the data (more spread) than Y1. Thus X1 is the first principal component. Y1 is the second principal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

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

Topological data analysis (topology)

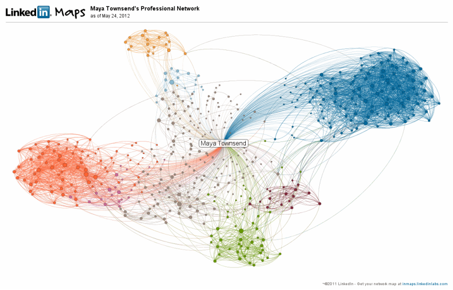
Topological data analysis (TDA) is a new area of study aimed at having applications in areas such as data mining and computer vision. The main problems are: (1) how one infers high-dimensional structure from low-dimensional representations; and (2) how one assembles discrete points into global structure. 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.


LinkedIn Maps
Maya Townsend's Professional Network
as of May 24, 2012
Maya Townsend
©2011 LinkedIn - Get your network map at inmaps.linkedinlabs.com

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

William and Mary EXTREEMS-QED

1. Undergraduate curriculum development
    1. Freshman, sophomore level introductory courses
    2. New advanced courses in complex networks, graphical models, matrix and graph theory techniques in statistics
2. Summer research program
    1. For each summer, 4-6 research teams will be formed with each team working on a data-enabled topic
    2. Each team consists of 1-2 faculty advisers and 2-3 undergraduate students, with total of 8 faculty advisers and 12 research students per year
    3. 1-2 faculty advisers and 2-4 students will be from partner HBCU institutes (Virginia St. U, Hampton U, Norfolk St. U)
3. Faculty professional development and outreach activities

Faculty advisers

1. Principal Investigator: Junping Shi (PDE and Math Biology)
2. Co-Principal Investigators: Sarah Day (Computational Topology), Chi-Kwong Li (Matrix Analysis), Gexin Yu (Graph Theory)
3. Senior Personnel: John Delos (Physics), Ross Iaci (Statistics), Rex Kincaid (Operation Research), Larry Leemis (Operation Research), Rom Lipcius (Marine Science), Anh Ninh (Big data analytics and Optimization), Mainak Patel (Math Biology), George Rublein (Math Education), Margaret Saha (Biology), Leah Shaw (Adaptive Network and Epidemics), Greg Smith (Applied Sci), Guan-Nan Wang (Statistics), Anke van Zuylen (Optimization)

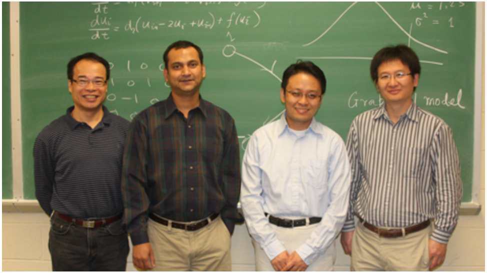


Photo: William and Mary EXTREEMS-QED PI and co-PIs: Li, Dey, Shi, Yu

William and Mary EXTREEMS-QED courses

Spring 2016

MATH 352: Data Analysis (Daniel McGibney)
MATH 410-02: Data Science: theory and applications
(Gexin Yu, Anh Ninh, Guannan Wang)
MATH 451: Probability & Statistics (Hyunchul Park)
MATH 452/552: Mathematical Statistics (Ross Iaci)
math 410-03/CSCI 618: Model/Applications Operational Research (Rex Kincaid)
Math 410-04/CSCI 688-01: Optimization Machine Learning (Anh Ninh)
CSCI 688-02: Econ Aspects of Internet (Anke van Zuylen)
Math 459-01/CSCI 688-03: Statistical Data Mining (Guan-Nan Wang)

Spring 2014 Math 410: Big Data Analysis

1. Format: one-credit course with one lecture per week by faculty members or external speakers.
2. Purpose: to introduce students to big data analysis, data science and possible undergraduate research projects (20 students in the class, and 5 entered 2014 summer EQ program).
3. Requirement: students take notes for each lecture and write a summary as assignments.
4. Sample topics: videos of 12 talks at
   `https://www.youtube.com/channel/UC2BiuU8gzH7eF7lxwj-VMIg`
   Albert Decatur (AidData) *Data from a Stone: Aid Transparency, PDF Ghettos, and Data Mining*
   Leah Shaw (Applied Science) *Adaptive Social Networks*
   John Delos (Physics) *Saving Infants in a Heartbeat!*
   Margaret Saha (Biology) *Big Data from RNA-Seq Experiments*
   Greg Smith (Applied Science) *Visual and Virtual Data: Using Simulation to Manage Your Expectations*
   Ersin Ancel (NASA Langley) *Data Mining at NASA Langley Research Center*
   Rom Lipcius (Marine Science) *Using Big Data for Marine Species to Achieve Conservation Goals*

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 (international students can apply for Charles center scholarship by March 18, 2016)
- 2016 Application deadline: March 25, 2016.
- 2015 Research Program: May 31-July 22, 2016 (8 weeks)

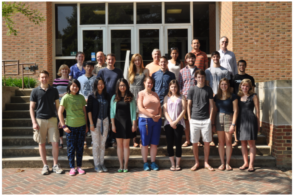


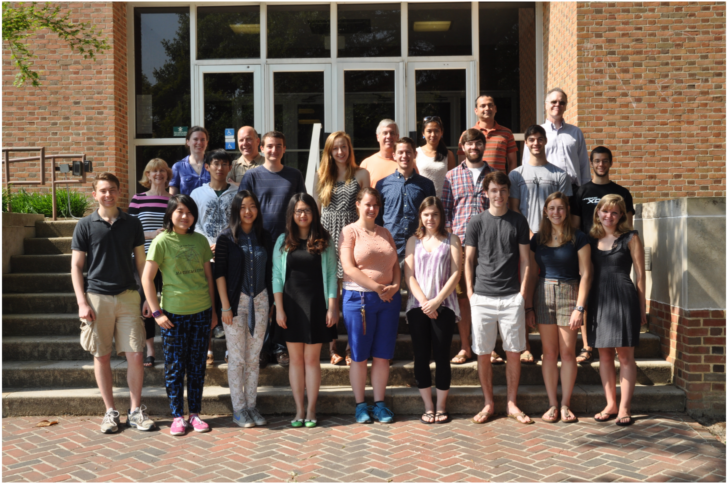
Photo: EXTREEMS-QED summer research program 2014

2014 W&M EXTREEMS-QED summer research program

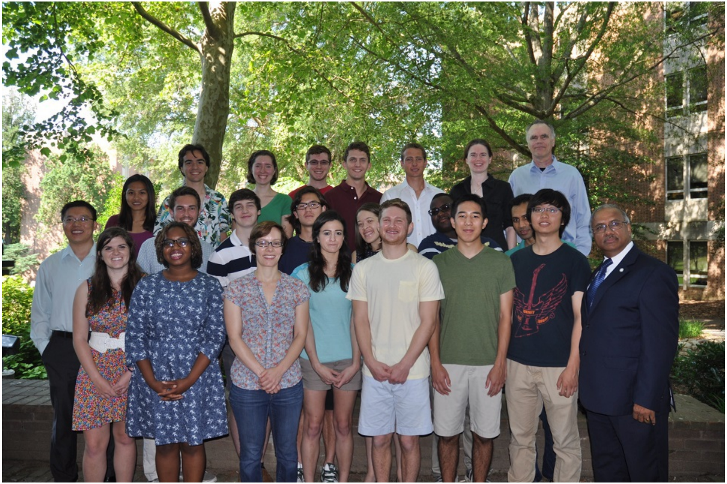
- Eric Berry: Decomposition of quantum gates (adviser: Li, Pelejo)
- Miranda Elliott: Automating data extraction in AidData (adviser: Kincaid)
- Aaron Finkle: High dimensional nonlinear modeling for analyzing robustness and plasticity in developing systems using transcriptome data (adviser: Saha, Vasiliu)
- Dean Katsaros: Decomposition of unitary gates (adviser: Li, Pelejo)
- Greg Kirwin: Adaptive social networks in online communities (adviser: Shaw)
- Devon Oberle: Text mining air transportation accident reports (adviser: Kincaid)
- Amanda Reeder (Norfolk State U): Heartbeats and Respiration (adviser: Delos, Lanz)
- Martin Salgado-Flores: Dynamics at continuous varying resolutions (adviser: Day)
- Robert Torrence: Identifying Message Boards as an Adaptive Network (adviser: Shaw)
- Fangyi Xu: study cell fate and differentiation in trichomes (adviser: Smith)

Other associated research students (not directly supported by EQ):
Mayee Chen (adviser: Shi), Corynne Dech (adviser: Shaw), Wade Hodson (adviser: Lipcius, Shaw, Shi), Haomiao Li (adviser: Shi), Catherine King (adviser: Day), Jing Yi Zhou (adviser: Lipcius, Shaw, Shi)
Topics: sustainable agriculture, pattern formation in intertidal zone, oyster population

2014 W&M EXTREEMS-QED summer research program students and faculty members



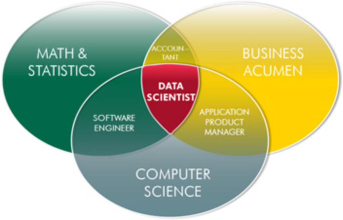
2015 W&M EXTREEMS-QED summer research program students and faculty members

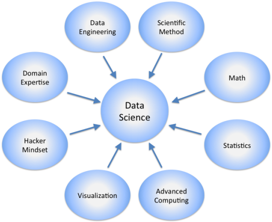


Summer activities

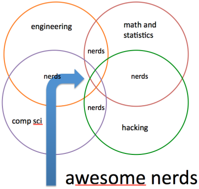
- Week 1-2: Short courses/tutorials (tutorials in `R`, `Matlab`, and statistics)(students have learned LaTeX in a sophomore course)
- Weekly student presentations (students make informal Beamer/Powerpoint presentation, and faculty members join discussions and help students to improve writing slides and presentation skills)
- Visiting large data centers, computing facilities and laboratories (2014)
  1. Week 2: W&M cyclone cluster (on campus)
  2. Week 3: DOE Thomas Jefferson Lab (40 miles away)
  3. Week 4: W&M Biology Lab on RNA sequencing (on campus)
  4. Week 5: NASA Langley Research Center Scientific Computing Group (30 miles away)
  5. Future Plan: Univ. Virginia Data Science Institute (120 miles away), NSA, NIST or other federal computing facilities (150 miles away)

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.
- We welcome students to participate in the summer research program in 2016 and beyond!