Data Science Applications & Use Cases

Instructor: Ekpe Okorafor

1. Accenture – Big Data Academy
2. Computer Science African University of Science & Technology

Objectives

Objectives

- **Understand Big Data Challenges**
- **What exactly is Data Science and what do Data Scientists do**
- **Data Science contrasted with other disciplines**
- **Case Study & Use Cases**

Outline

- **Big Data & Challenges**
- **What is Data Science**
- **Data Science & Academia**
- **Data Science & Others**
- **Case Studies**
- **Essential points**
- **Conclusion**

Data All Around

- **Lots of data is being collected and warehoused**
  - Scientific Experiments
  - Internet of Things
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network
  - ……many more!

Big Data

- Big Data are data sets so large or so complex that traditional methods of storing, accessing, and analyzing their breakdown are too expensive. However, there is a lot of potential value hidden in this data, so organizations are eager to harness it to drive innovation and competitive advantage.
- Big Data technologies and approaches are used to drive value out of data rich environments in ways that traditional analytics tools and methods cannot.
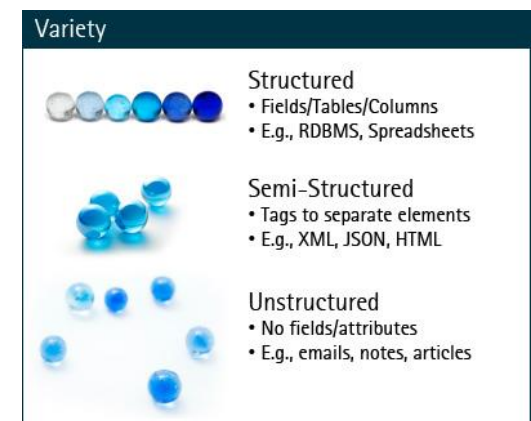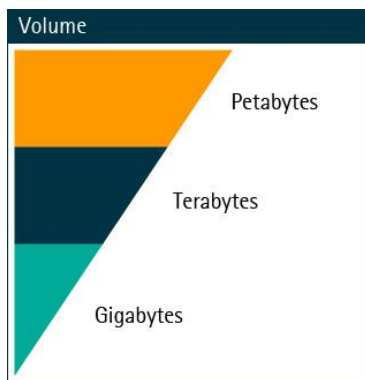
What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling
- Data Driven
  - Predictive Analytics
 - Deep Learning

Big Data & Data Science

- "… the sexy job in the next 10 years will be statisticians," Hal Varian, Google Chief Economist

- The U.S. will need 140,000-190,000 predictive analysts and 1.5 million managers/analysts by 2018. McKinsey Global Institute's June 2011

- New Data Science institutes being created or repurposed – NYU, Columbia, Washington, UCB,...

- New degree programs, courses, boot-camps:
  - e.g., at Berkeley: Stats, I-School, CS, Astronomy…
  - One proposal (elsewhere) for an MS in "Big Data Science"
  - Plans for Data Science Stream at AUST
  - RDA-CODATA School of Research Data Science

What is Data Science?

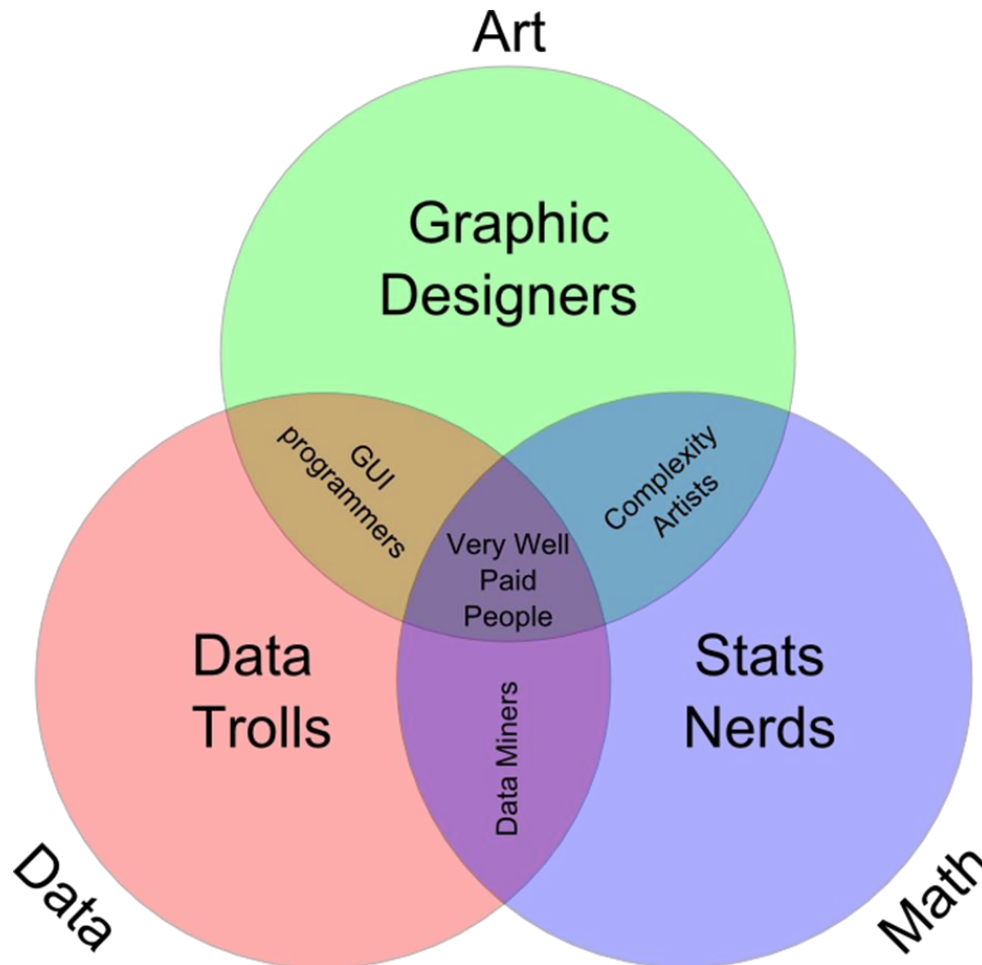- Some definitions link computational, statistical, and substantive expertise.
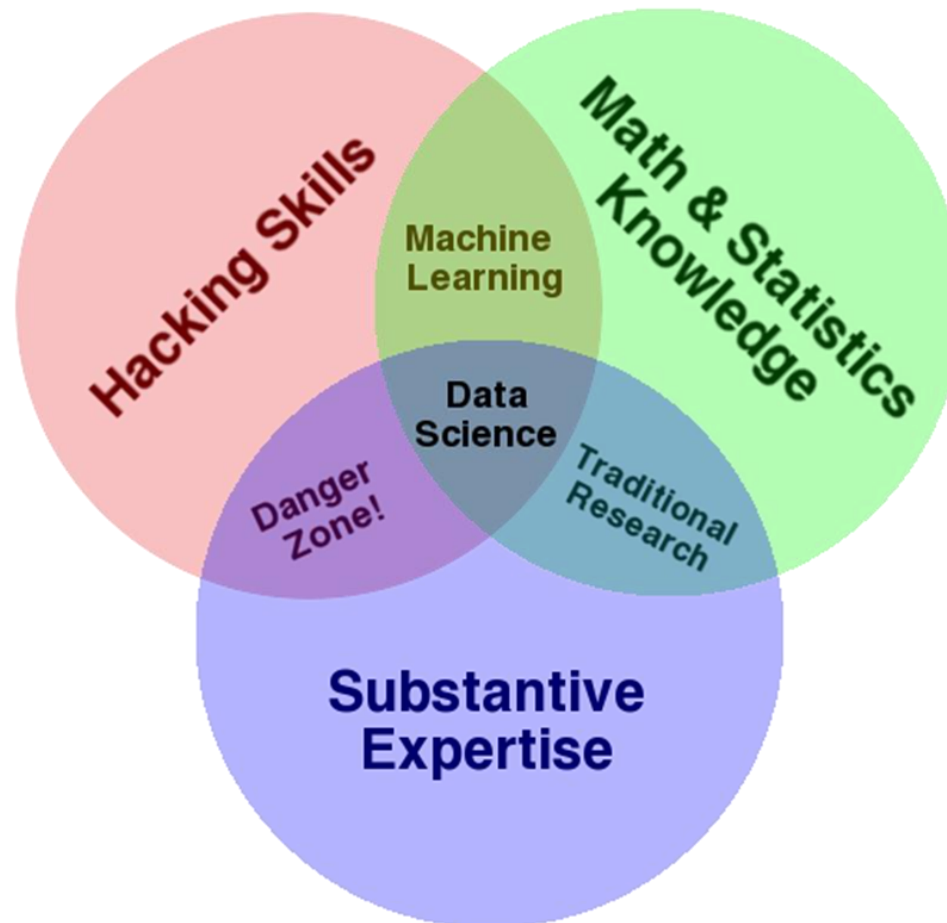
What is Data Science?

- Other definitions focus more on technical skills alone.

What is Data Science?

- An area that manages, manipulates, extracts, and interprets knowledge from tremendous amount of data

- Data science (DS) is a multidisciplinary field of study with goal to address the challenges in big data

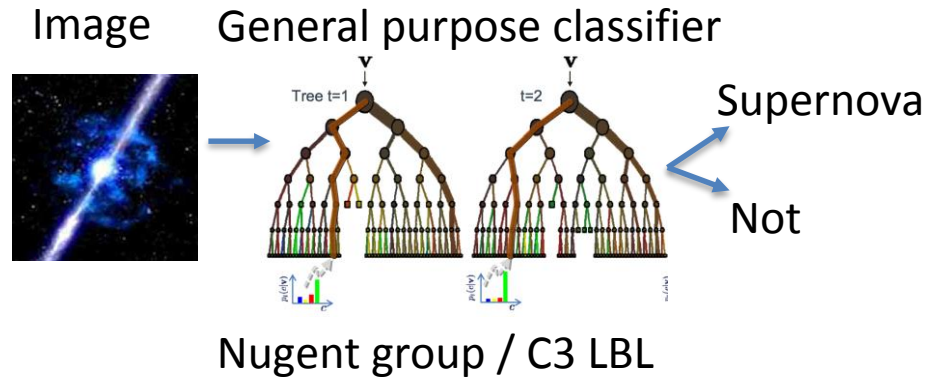- Data science principles apply to all data – big and small
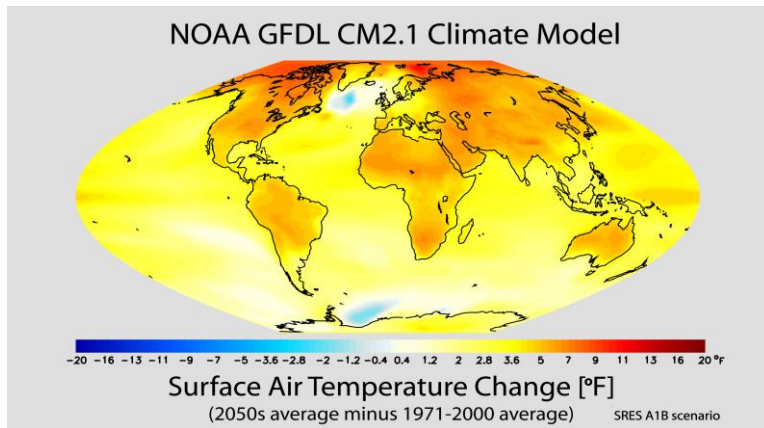
What is Data Science?

- Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries such as science, engineering, economics, politics, finance, and education
  - Computer Science
    - Pattern recognition, visualization, data warehousing, High performance computing, Databases, AI
  - Mathematics
    - Mathematical Modeling
  - Statistics
    - Statistical and Stochastic modeling, Probability.

Data Science Vs Analysis Vs Software Delivery

Component	Traditional Analysis	Traditional Software Delivery	Data Science
Tools	SAS, R, Excel, SQL, in-house tools	Java, source control, Linux, continuous integration, unit testing, bug reports and project management	R, Java, scientific Python libraries, Excel, SQL, Hadoop, Hive, Pig, Mahout and other machine learning libraries, github for source control and issue management
Analytical Methods	Regressions, classifications, measuring prediction accuracy and coverage/error, sampling	N/A	Classification, clustering, similarity detection, recommenders, unsupervised and supervised learning, small- and large-scale computations, measuring prediction accuracy and coverage/error
Team Structure	Statisticians, Mathematicians, Scientists	Developers, Project Managers, Systems Engineers	Mathematicians, Statisticians, Scientists, Developers, Systems Engineers
Time Frame	Either: • Usually on-going research and discovery within a team in the organization Or: • Specific project to determine answers	Regular software release cycle, continuous delivery, etc.	Either: • Discovery/learning phase leading to product development Or: • On-going research and product invention/improvement

Contrast: Scientific Computing

NOAA GFDL CM2.1 Climate Model
Surface Air Temperature Change [°F]
(2050s average minus 1971-2000 average)   SRES A1B scenario

Image   General purpose classifier
Supernova
Not
Nugent group / C3 LBL

Scientific Modeling	Data-Driven Approach
Physics-based models	General inference engine replaces model
Problem-Structured	Structure not related to problem
Mostly deterministic, precise	Statistical models handle true randomness, and **un-modeled complexity**.
Run on Supercomputer or High-end Computing Cluster	Run on cheaper computer Clusters (EC2)

Contrast: Machine Learning
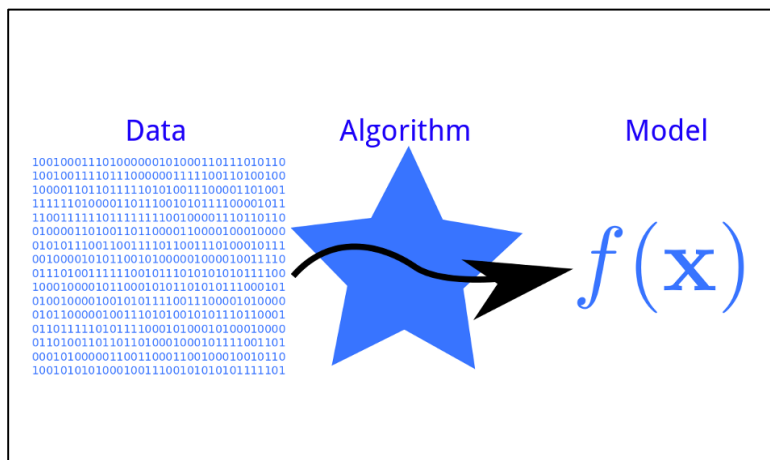
Machine Learning
Develop new (individual) models
Prove mathematical properties of models
Improve/validate on a few, relatively clean, small datasets
Publish a paper ☺

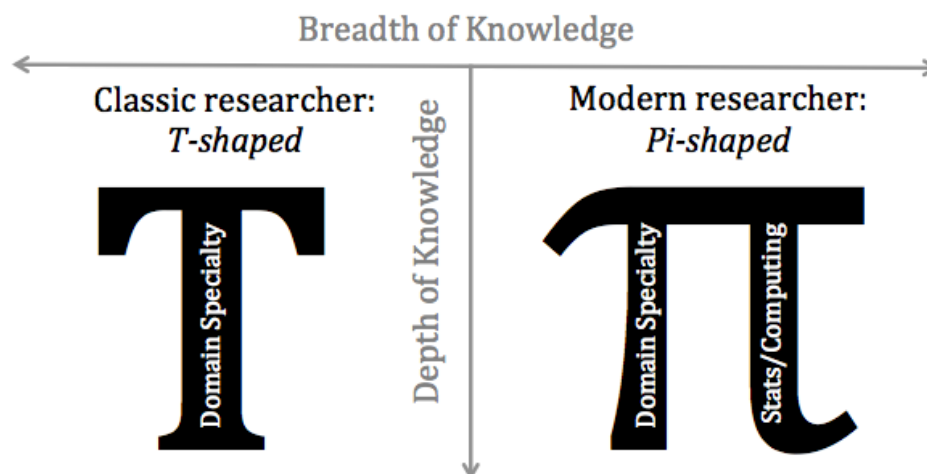| Data Science |
|---|
| Explore many models, build and tune hybrids |
| Understand empirical properties of models |
| Develop/use tools that can handle massive datasets |
| Take action! |

Contrast: Data Engineering

	Data Science	Data Engineering
Approach	Scientific (Exploration)	Engineering (Development)
Problems	Unbounded	Bounded
Path to Solution	Iterative, exploratory, nonlinear	Mostly linear
Education	More is better (PhD's common)	BS and/or self-trained
Presentation Skills	Important	Not as important
Research Experience	Important	Not as important
Programming Skills	Not as important	Important
Data Skills	Important	Important

Data Science & Academia

- In the words of Alex Szalay, these sorts of researchers must be "Pi-shaped" as opposed to the more traditional "T-shaped" researcher. In Szalay's view, a classic PhD program generates T-shaped researchers: scientists with wide-but-shallow general knowledge, but deep skill and expertise in one particular area. The new breed of scientific researchers, the data scientists, must be Pi-shaped: that is, they maintain the same wide breadth, but push deeper both in their own subject area and in the statistical or computational methods that help drive modern research:

Breadth of Knowledge

Classic researcher: *T-shaped*

Modern researcher: *Pi-shaped*

Depth of Knowledge

Domain Specialty

Domain Specialty

Stats/Computing

Data Science & Academia

- In a post by Jake Vanderplas in 2014 related to SciFoo discussion on: **Academia and Data Science**, the following questions below were discussed.

- I encourage you to develop your own thoughts on them and come up with your assessment

  - Where does Data Science fit within the current structure of the university & research institutions?

  - What is it that academic data scientists want from their career? How can academia offer that?

  - What drivers might shift academia toward recognizing & rewarding data scientists in domain fields?

  - Recognizing that graduates will go on to work in both academia and industry, how do we best prepare them for success in both worlds?

Data Science Applications

	Business	**Health Care**	**Urban Leaving**
Summary	From car design to insurance to pizza delivery, businesses are using data science to optimize their operations and better meet their customers' expectations.	Tomorrow's healthcare may look more efficient thanks to things like electronic health records. It also may look a lot more effective. Reduced readmissions, better care, and earlier detection are on the horizon.	For the first time in human history, more people live in cities than in suburban or rural areas. An emerging field called "urban informatics" combines data science with the unique challenges facing the world's growing cities
What is happening?	Two-Way Street for the Ford Focus Electric Car	Reducing Hospital Readmissions	Taking on Megacity Traffic
	Better Fraud Detection Boosts Customer Satisfaction	Better Point-of-Care Decisions	Fighting Crime with Data "predictive policing"
	E-Commerce Insights: Domino's Secret Sauce		
What is possible	Using Social Data to Select Successful Retail Locations .	Medical Exams by Bathroom Mirrors	Instrumenting cities

Contrast: Computational Sciences

- Is there a contrast between Data Science and Computational Science?

Data Science: Case Study
Cancer Research

- Cancer is an incredibly complex disease; a single tumor can have more than **100 billion cells**, and each cell can acquire mutations individually. The disease is always changing, evolving, and adapting.

- Employ the power of big data analytics and high-performance computing.

- Leverage sophisticated pattern and machine learning algorithms to identify patterns that are potentially linked to cancer

- Huge amount of data processing and recognition
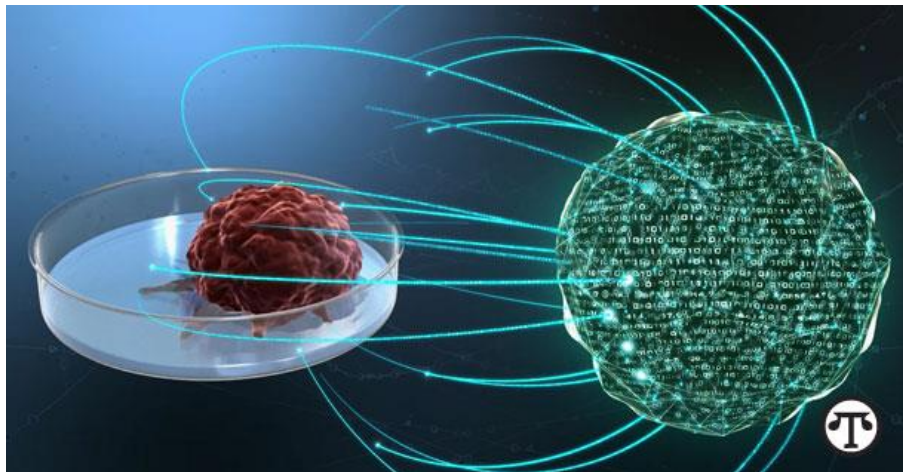
Data Science: Case Study
Health Care

- Stanford Medicine, Google team up to harness power of data science for health care
- Stanford Medicine will use the power, security and scale of Google Cloud Platform to support precision health and more efficient patient care.
- Analyzing genetic data
- Focusing on precision health
- ***Data as the engine that drives research***

Data Science: Case Study
Elections

- The Obama campaigns in 2008 and 2012 are credited for their successful use of social media and data mining.

- Micro-targeting in 2012

  – http://www.theatlantic.com/politics/archive/2012/04/the-creepiness-factor-how-obama-and-romney-are-getting-to-know-you/255499/

  – http://www.mediabizbloggers.com/group-m/How-Data-and-Micro-Targeting-Won-the-2012-Election-for-Obama---Antony-Young-Mindshare-North-America.html

- Micro-profiles built from multiple sources accessed by aps, real-time updating data based on door-to-door visits, focused media buys, e-mails and Facebook messages highly targeted.

- 1 million people installed the Obama Facebook app that gave access to info on "friends".

Data Science: Case Study
Internet of Things (IoT)

- The Internet of Things is rapidly growing. It is predicted that more than 25 billion devices will be connected by 2020.

Information from the Internet of Things:
We have gone beyond the decimal system

Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

In the near future, **Brontobyte** will be the measurement to describe the type of sensor data that will be generated from the IoT (Internet of Things)

10^{27} **Brontobyte**
This will be our digital universe tomorrow...

10^{24} **Yottabyte**
This is our digital universe today
= 250 trillion of DVDs

10^{21} Zettabyte
1.3 ZB of network traffic by 2016

10^{18} Exabyte
1 EB of data is created on the internet each day = 250 million DVDs worth of information. The proposed Square Kilometer Array telescope will generated an EB of data per day

10^{15} Petabyte
The CERN Large Hadron Collider generates 1PB per second

10^{12} Terabyte
500TB of new data per day are ingested in Facebook databases

10^{9} Gigabyte

10^{6} Megabyte

- The Internet of Things (IOT) will soon produce a massive volume and variety of data at unprecedented velocity. If "Big Data" is the product of the IOT, "Data Science" is it's soul.

23

Data Science: Case Study
Customer Analytics

Marketing & Advertising

Leveraging customer data to move ever closer to the elusive goal of truly personalized marketing: the right offer, at the right time, in the right location and context, to the right person.

Customer Service

By capturing and analyzing the data from customer touch points within an organization, companies can identify customer pain points and issues proactively and update their customer service FAQs or other communications with existing customers.

Retention & Loyalty

Using customer data and analytics, these companies deploy and refine predictive models that help them retain customers with proactive approaches. Investments, in terms of offers and upgrades, can be made at the right time to increase the likelihood of retaining desirable customers.

Customer Experience

The experience that customers have with companies matters a great deal. Other recent research has highlighted the critical connection between experience and company financial performance.

Essential Points

- **Big Data has given rise to Data Science**

- **Data science is rooted in solid foundations of mathematics and statistics, computer science, and domain knowledge**

- **Sexy profession – Data Scientists ☺**

- **Not every thing with data or science is Data Science!**

- **The use cases for Data Science are compelling**

Conclusion

In this section you have learned

- What Big Data Challenges are

- What exactly is Data Science and what do Data Scientists do

- Data Science contrasted with other disciplines

- Case Study & Use Cases

Questions?

http://www.ign.com/articles/2015/12/16/star-wars-the-force-awakens-review