UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

3250 Foundations of Data Science

Module 1: Introduction to Data Science

Course Plan

Module Titles
Current Focus: Module 1 – Introduction to Data Science
Module 2 – Introduction to Python
Module 3 – NumPy
Module 4 – Pandas
Module 5 – Data Collection and Cleaning
Module 6 – Descriptive Statistics and Visualization
Module 7 – Workshop (No Content)
Module 8 – Time Series
Module 9 – Introduction to Regression and Classification
Module 10 – Databases and SQL
Module 11 – Data Privacy and Security
Module 12 – Term Project Presentations (no content)

Learning Outcomes for this Module

- Outline the course logistics
- Discuss the history of Data Science
- Introduce applications of Predictive Modeling
- Identify the skills and knowledge a Data Scientist needs
- Review the job market
- Describe relevant certifications

Topics for this Module

- **1.1** Introductions and course overview
- **1.2** History of data science
- **1.3** Define predictive modeling and data mining
- **1.4** Examples of applications of predictive modeling
- **1.5** What it takes to become a data scientist
- **1.6** Job market overview
- **1.7** Homework

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 1 – Section 1

Introductions and Course Overview

Certified Analytics Professional

- Industry Certification
- Operated by INFORMS, the world's largest professional society for those in the field of analytics, operations research (O.R.), and the management sciences
- Requires experience doing analytics and a related degree (or equivalent additional experience)
- Code of ethics

Certificate in Data Science

- Understand the techniques and methods of predictive and Big Data analytics
- Learn how to use tools such as Python and Hadoop to tackle data analysis challenges
- Develop and use models and tools to solve business problems and mine data for fresh insights

Certificate in Data Science (cont'd)

What You'll Learn

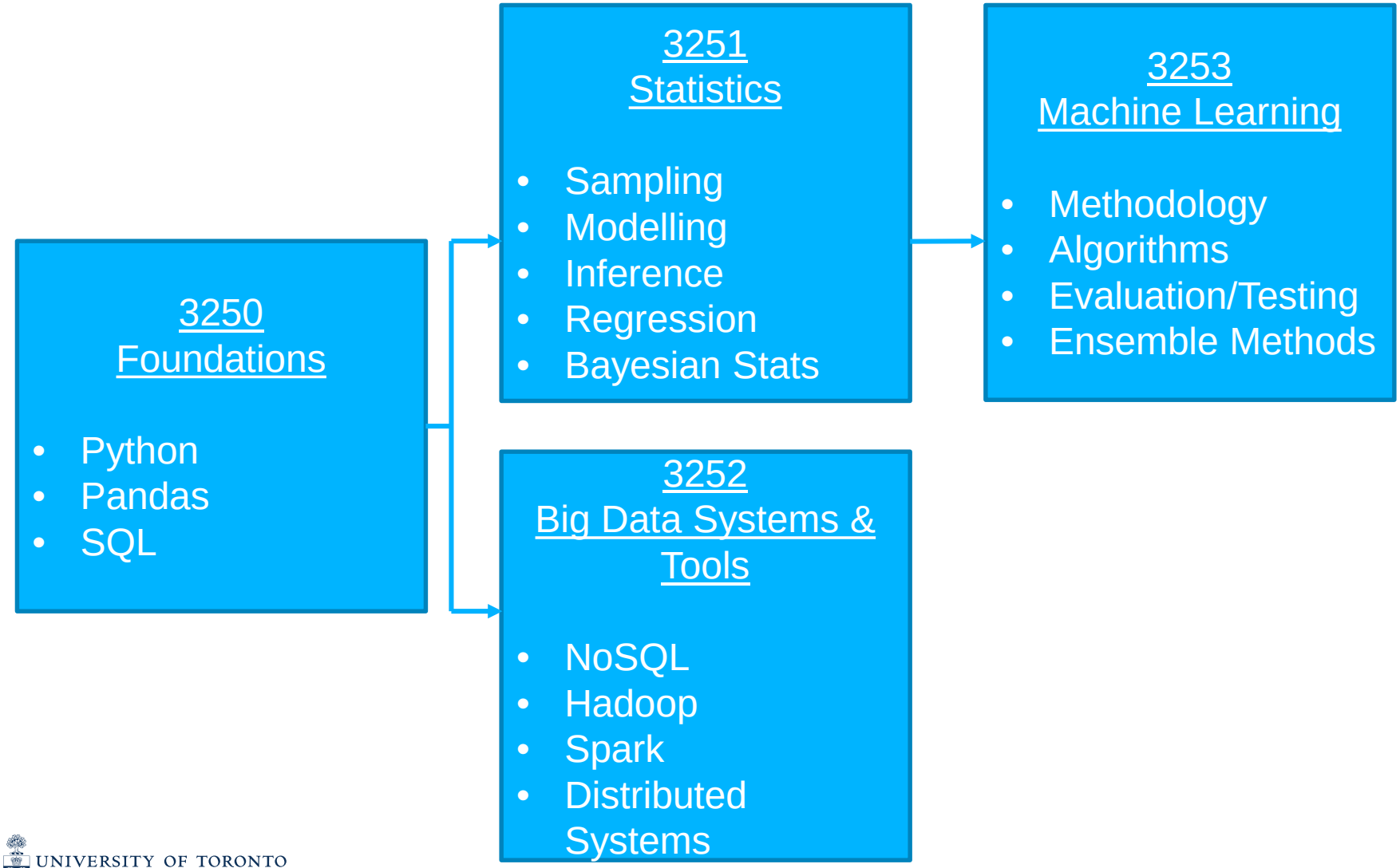
- Explore the evolution of data science and predictive analytics
- Know statistical concepts and techniques including regression, correlation and clustering
- Apply data management systems and technologies that reflect concern for security and privacy
- Adopt techniques and technologies including data mining, neural network mapping and machine learning
- Represent big data findings visually to aid decision-makers

Certificate in Data Science (cont'd)

Courses

- **SCS 3250 – Foundations of Data Science**
- SCS 3251 – Statistics for Data Science
- SCS 3252 – Big Data Management Systems & Tools
- SCS 3253 – Machine Learning

Certificate in Data Science (cont'd)



The CAP Domains

Coverage in this certificate program

	3250	3251	3252	3253
I. Business Problem (Question) Framing	✓	✓✓	✓	✓✓✓
II. Analytics Problem Framing	✓	✓✓✓	✓	✓✓✓
III. Data	✓✓	✓✓✓	✓✓	✓✓✓
IV. Methodology (Approach) Selection	✓	✓✓		✓✓✓
V. Model Building		✓✓✓	✓	✓✓✓
VI. Deployment			✓✓✓	✓
VII. Model Life Cycle Management			✓✓	✓✓✓

✓ = Introductory content
✓ ✓ = Substantial coverage
✓ ✓ ✓ = Major focus

About this Course and the Certificate Program

- Not a course in general Python programming
- But a course that introduces the use of Python in data analytics
- Subsequent courses in the certificate program
  - Teach the various disciplines of data science and Big Data technologies
  - Overall content more technical than the “Management of Enterprise Data Analytics” certificate program
  - Not all mathematics (i.e. analytical solutions) but also relying on the use of programming to understand data

How to Benefit the Most from this Course?

- Come to class
- Working with classmates is encouraged
- Use Quercus to share questions and insights (10% participation mark)
  - If you come across an interesting article on the subject matter, share with the class
  - If you have problems with the homework, post a question there
- Do the readings and homework

Quick Poll

Why would you like to become a Data Scientist?

A. Enjoy making sense of data

B. Good pay

C. Interesting work

D. Data Scientists are in high demand

E. All of the above

Module 1 – Section 2

History of Data Science

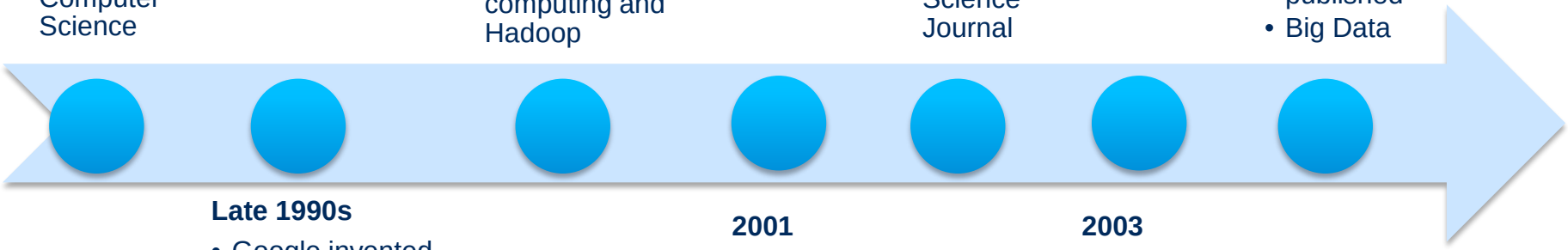
What is Data Science?

“Data Science” is a fairly new term, for a new profession that is trying to make sense of Big Data.

Collecting, storing, and making sense of Big Data (another fairly new term) is quickly becoming part of every business and everyone’s life.

A Brief History of Data Science

The term "Data Science" is attributed to William S. Cleveland who, in 2001, wrote "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.“



1960s-1970s

- Advances in Statistics and Computer Science

Late 1990s

- Google invented a new search engine combining math, statistics, data engineering and computation (which replaced AltaVista).

1998-2000

- Hard drives become cheap
- Dot-Com “boom”
- Cloud computing and Hadoop

2001

- Data Science term gets “coined”

2002

- CODATA Data Science Journal

2003

- Columbia University began publishing The Journal of Data Science

2010

- What is Data Science? Article is published
- Big Data

Evolution of Analytics

1.0 Traditional Analytics

- Primarily Descriptive and Reporting
- Internally sources, relatively small, structured data
- “Backroom” teams of analysts
- Internal Systems of Support

2.0 Big Data

- Complex, large, unstructured data sources
- Starting mid 2000s (the term Big Data was coined in 2010)
- Stored and processed rapidly, with new analytical and computational technologies like Hadoop
- “Data Scientists Emerge
- Online firms create data-based products and services

Analytics 3.0

- **What defines Analytics 3.0**
  - An environment and combines analytics 1.0 and 2.0 that yields insights with speed and impact
  - Analytics integral to running a business and becomes part of strategy and operations
  - Predictive and Prescriptive Models
  - Artificial Intelligence techniques
- Sources
  - Analytics 3.0 FAQ
  - Analytics 3.0

From Data Analysts to Data Scientists

Traditional Analysts

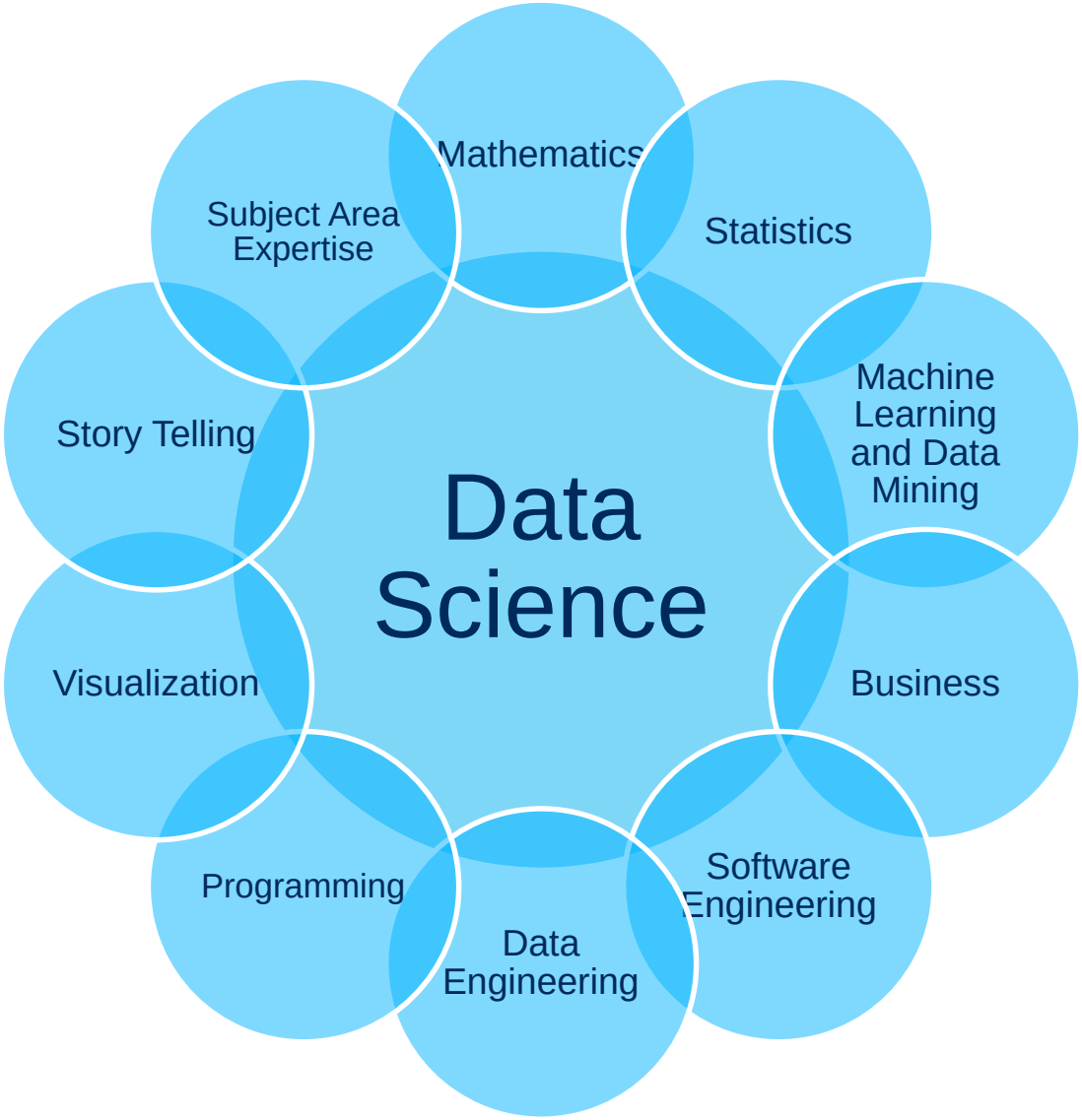
- Tend to use tools like SAS and SQL
- Use Relational DBMS

Data Scientists

- Tend to use tools like Python and R (often in addition of SQL and SAS)
- Use Hadoop environment as well as in-memory databases, and in-memory computing

IMPORTANT: Once you learn skills and tools in one environment you can easily transition to the other. The underlying skills are the same.

Data Science Is Multidisciplinary



Big Data

Big data is defined as a large volume of data (structured and unstructured) that “floods” a business on a day-to-day basis.

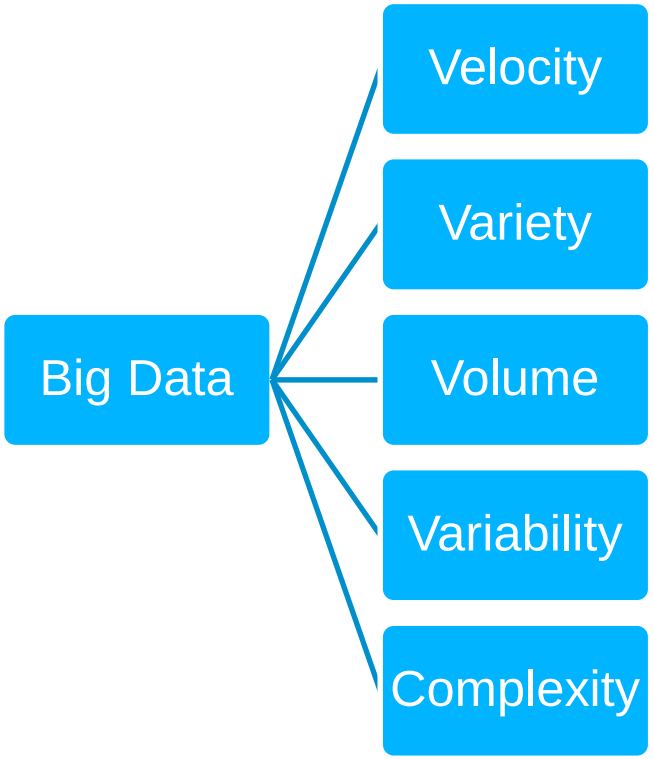
“Data is growing faster than ever before and by the year 2020, about 1.7 MB of new information will be created every second for every human being on the planet” (Marr, 2015)

Source

The Three+ Vs of Big Data

- "**Big Data**" is a relatively new term, however collecting, storing and analysing data is centuries old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of Big Data as the three Vs: **Velocity, Variety and Volume**.
  What is Big Data

- SAS Institute also considers **Variability and Complexity**
- Some also include **Veracity**



Questions:

Where would you find Big Data?

Can you provide an example of Big Data?

Module 1 – Section 3

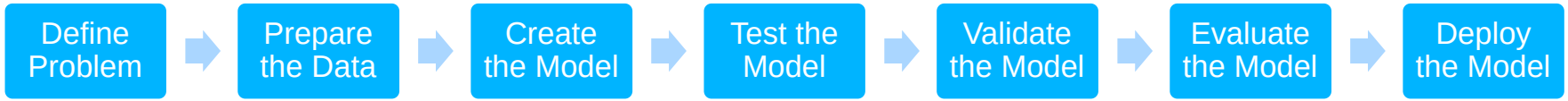
Defining Predictive Modeling and Data Mining

Predictive Modelling

Predictive modeling is a process used in analytics to create a statistical model of future behaviour.

Predictive analytics is the area of data science concerned with forecasting probabilities and trends.

The business process of Predictive Modelling often consists of the following steps:



What is Data Mining?

- Data Mining is defined as examining data to uncover patterns in the data to generate new information

- Data Mining is comprised of:
  - Massive data collection
  - Powerful multiprocessor computers
  - Data mining algorithms

Data Mining and Predictive Analytics

- Both branches are grounded in a huge amount of mathematical theory dating back several decades.

- Predictive analytics and data mining both apply complex mathematics to data in order to solve business problems. However, when we talk about data mining, we are usually referring to an analytic toolset that automatically searches for useful patterns in large data sets.

- **Data mining is often one stage in developing a predictive model.**

Examples of Predictive Modelling Techniques

- **Decision Trees**
  - Classification and Regression Trees (CART), CHAID, C4.5, C5.0, etc.
  - Random Forests (work by constructing many decision trees)
  - Boosted Trees

- **Regression**
  - OLS, GLM (Logistic Regression is special case of GLM, where other include Poisson, Gamma and Multinomial regression), MARS (multivariate adaptive regression splines), Semi-parametric regression

- **Neural Network**
- **Support Vector Machines**
- **k-Nearest Neighbour** algorithm (k-NN) is a non-parametric method used for classification and regression
- **Naive Bayes classifiers** are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features
- **k-means** algorithm is a distance-based clustering algorithm that partitions the data into a predetermined number of clusters

Module 1 – Section 4

Applications of Predictive Modeling - Examples

Data Science - Every Day

- How is data science impacting our daily lives:
  - Do you shop online?
  - Do you receive coupons and offers by email/mail?
  - How many credit cards do you have?
  - Why you received an offer to go watch a movie this weekend?
  - How does Facebook always “know” what ads you would like to see?
  - Do you watch Netflix, and follow their “recommendations”?

Social Media and the Data “Explosion”

Banking and Finance

- Data Science Applications include:
  - Customer acquisition (acquire new credit card customers, investors, traders, etc.)
  - Churn models (prevent customer attrition)
  - Risk models (to assess who is qualified for a mortgage, credit line, etc.)
  - Next best product model
  - Customer Satisfaction
  - Drive revenue, reduce cost

Data Science in Healthcare

Product Recommendation Engine

amazon.ca

Recommendations for you in Shoes

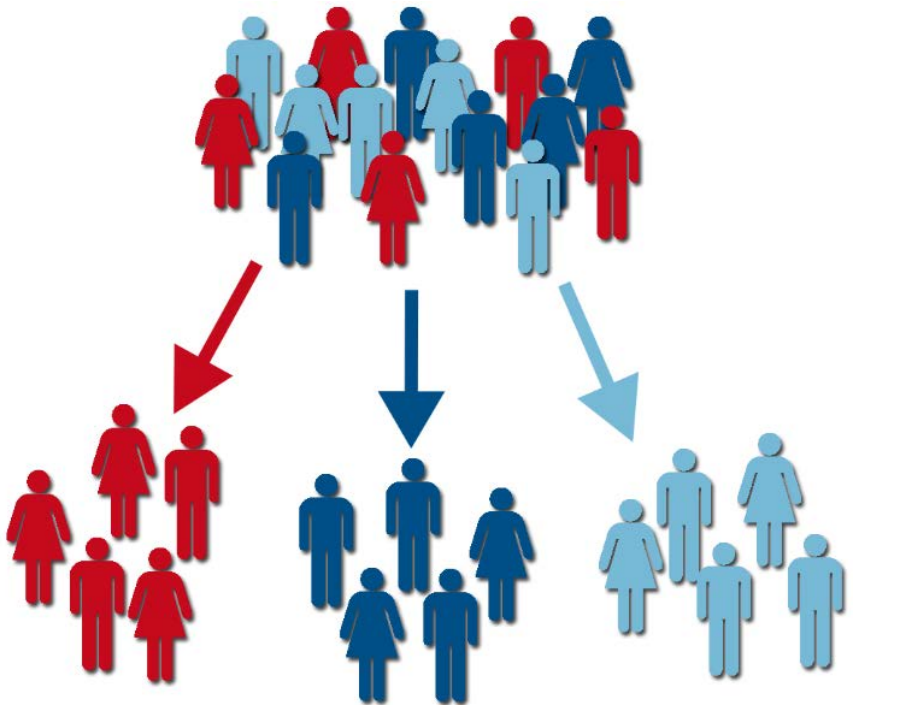
Recommendations for you in Clothing

Example: Customer Segmentation

Why Segment?

- Customers may differ in:
  - What they want to buy
  - Amount willing to pay
  - Quantity they buy
  - Time, place, frequency of purchase
  - Personal taste (likes and dislikes)
    - Media, telephone plan, newspapers, magazines, movies, social media

Example: Customer Segmentation (cont'd)



Premium

$$$

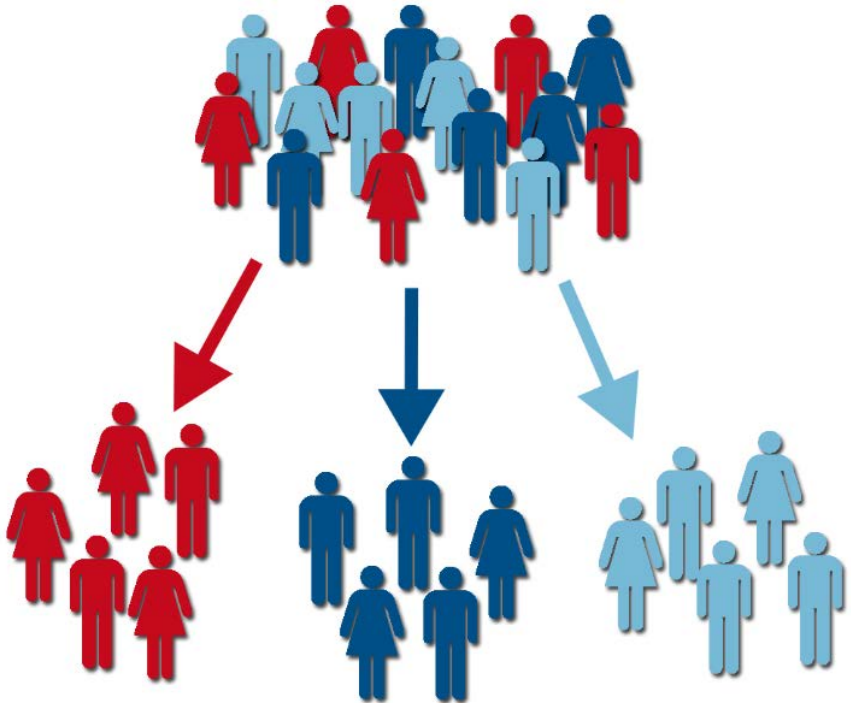
Valuable

$$

Potential

$

Customize Marketing Strategy for Each Customer Segment



Premium

Say "Thank you" through personalized communication

Valuable

Maintain and Grow

Potential

Grow these into Valuable customers through offers/promotions

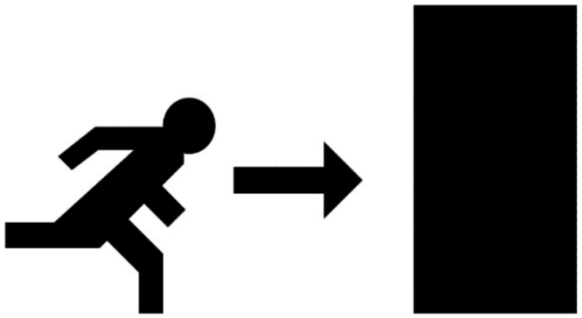
Example: Customer Churn (Attrition)

Customer **churn** or **attrition**, is defined as the number of customers who discontinue a service or employees who leave a company during a specified time period.

Why do customers leave?

Better price? Better service? Convenient location? Etc.

Data Scientists may build a predictive model to flag early signs of customer churn, to help business develop strategy to prevent churn.



Example: Fraud Detection

Market Basket Analysis

Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

Bought Milk and Eggs → Bought Oil

Business strategy could include:

a. Offer coupon on **Eggs** with a purchase of **Milk**
b. Place **Milk** and **Eggs** close on the shelf
c. Place **Oil** near **Milk** and **Eggs**
d. Place **Oil** far from **Milk** and **Eggs** (to force customer "shop through the store")

Quick Poll

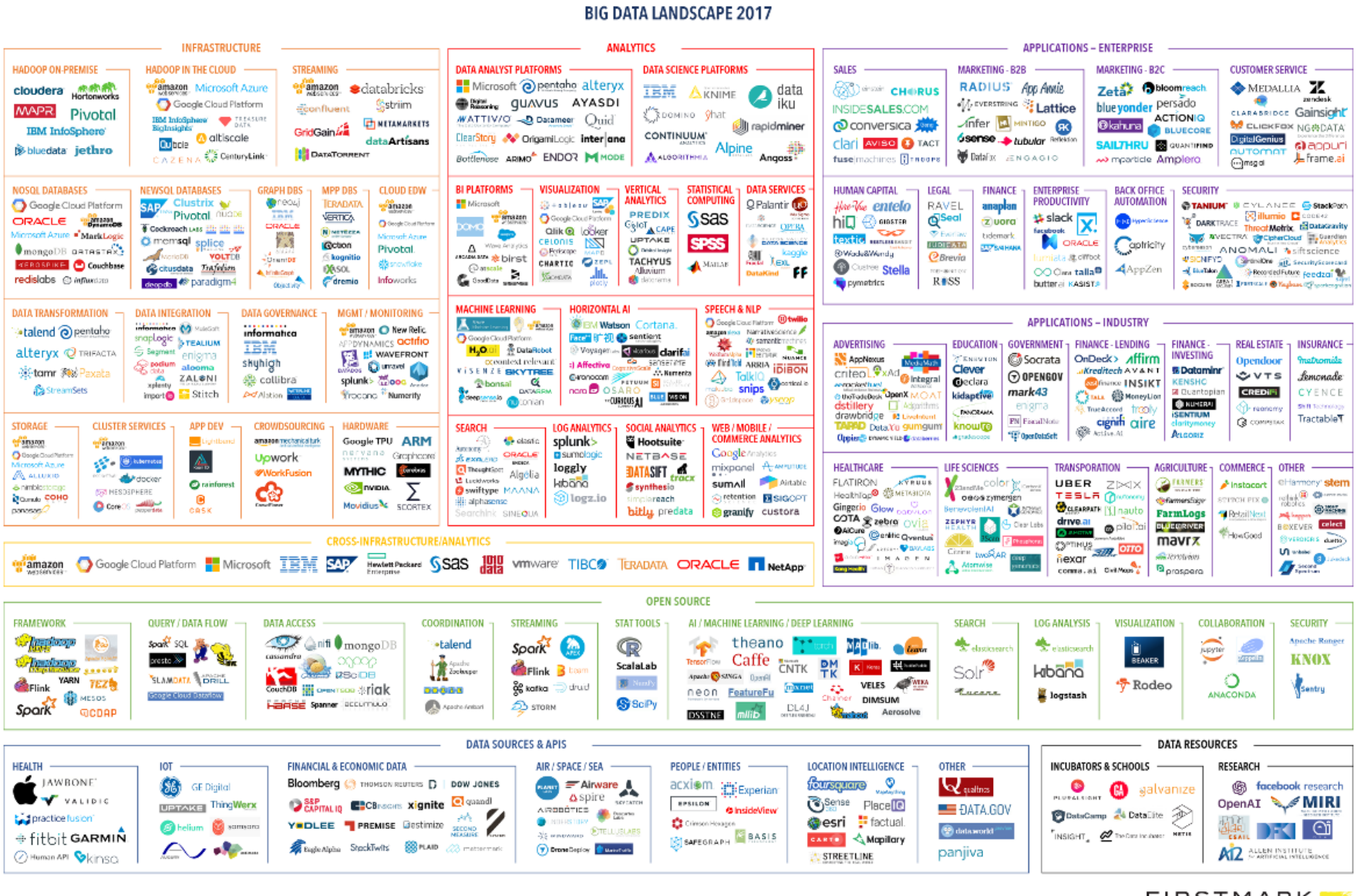
Does your organization have a Big Data road map?

A. Yes

B. No

C. Don’t Know

D. What is Big Data??


BIG DATA LANDSCAPE 2017
INFRASTRUCTURE
HADOOP ON-PREMISE
HADOOP IN THE CLOUD
STREAMING
NOSQL DATABASES
NEWSQL DATABASES
GRAPH DBS
MPP DBS
CLOUD EDW
DATA TRANSFORMATION
DATA INTEGRATION
DATA GOVERNANCE
MGMT / MONITORING
STORAGE
CLUSTER SERVICES
APP DEV
CROWDSOURCING
HARDWARE
ANALYTICS
DATA ANALYST PLATFORMS
DATA SCIENCE PLATFORMS
BI PLATFORMS
VISUALIZATION
VERTICAL ANALYTICS
STATISTICAL COMPUTING
DATA SERVICES
MACHINE LEARNING
HORIZONTAL AI
SPEECH & NLP
SEARCH
LOG ANALYTICS
SOCIAL ANALYTICS
WEB / MOBILE / COMMERCE ANALYTICS
CROSS-INFRASTRUCTURE/ANALYTICS
APPLICATIONS – ENTERPRISE
SALES
MARKETING - B2B
MARKETING - B2C
CUSTOMER SERVICE
HUMAN CAPITAL
LEGAL
FINANCE
ENTERPRISE PRODUCTIVITY
BACK OFFICE AUTOMATION
SECURITY
APPLICATIONS – INDUSTRY
ADVERTISING
EDUCATION
GOVERNMENT
FINANCE - LENDING
FINANCE - INVESTING
REAL ESTATE
INSURANCE
HEALTHCARE
LIFE SCIENCES
TRANSPORTATION
AGRICULTURE
COMMERCE
OTHER
OPEN SOURCE
FRAMEWORK
QUERY / DATA FLOW
DATA ACCESS
COORDINATION
STREAMING
STAT TOOLS
AI / MACHINE LEARNING / DEEP LEARNING
SEARCH
LOG ANALYSIS
VISUALIZATION
COLLABORATION
SECURITY
DATA SOURCES & APIS
HEALTH
IOT
FINANCIAL & ECONOMIC DATA
AIR / SPACE / SEA
PEOPLE / ENTITIES
LOCATION INTELLIGENCE
OTHER
DATA RESOURCES
INCUBATORS & SCHOOLS
RESEARCH
Last updated 4/5/2017
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark (@firstmarkcap)
mattturck.com/bigdata2017
FIRSTMARK
EARLY STAGE VENTURE CAPITAL

Module 1 – Section 5

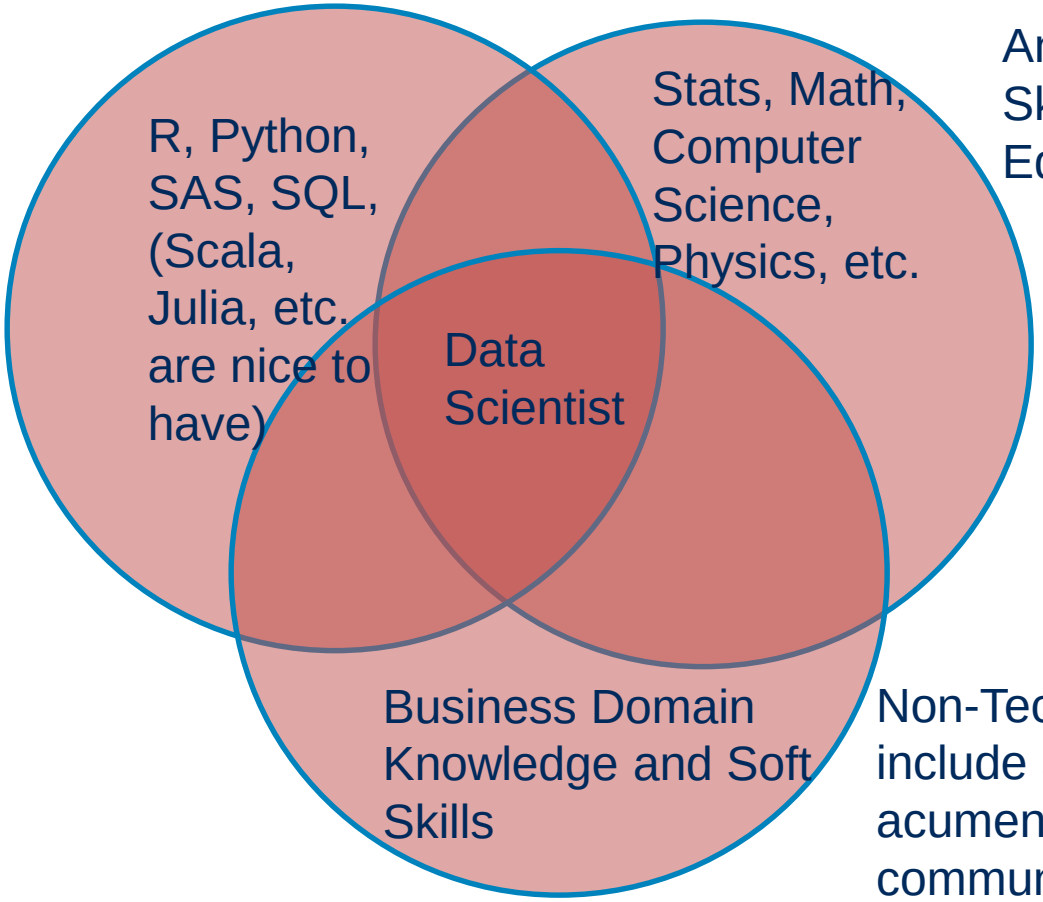
Becoming a Data Scientist

How to Become a Data Scientist

- To be come a data scientist, one would need to have
  - background in statistics, math and programming
  - soft skills (communication, scientific curiosity)
  - business understanding—and gut instinct
  - strong technical skills (databases and coding)

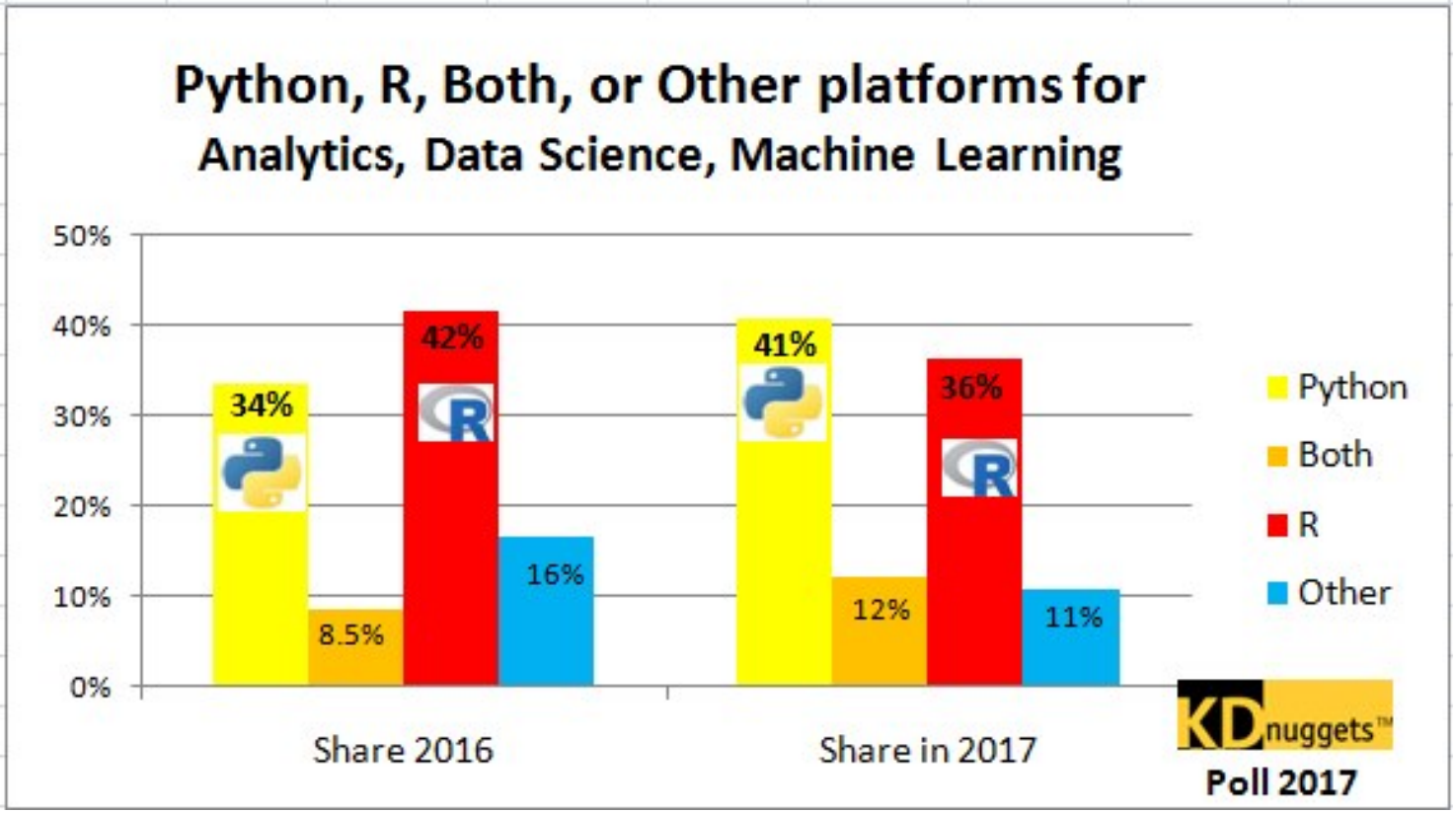
- Formal education
  - though these days a Masters or PhD isn't a requirement in Data Science; one could supplement a bachelors degree with experience and relevant certifications
  - (*Masters in Information and Data Science* MIDS at UC Berkeley costs ~ $60,000!)
- Certifications and non-degree programs (such as continuing education)
- Python, R and/or SAS, SQL
- Strong background in analytics

Skills Required



Non-Technical Skills include a strong business acumen, solid communication skills, and ability to tell a “story”

Python #1 for Data Science



Non Technical Skills

- **Intellectual curiosity** –This is a key skill, as one needs to think about the problem critically and ask the right questions to be able to formulate and eventually answer the business problem at hand.

- **Business acumen** – one needs a good understanding of the industry they are working in, and have a grasp of problems the company is trying to solve.

- **Communication skills** – a data scientist must be able to clearly and fluently translate their findings to a non-technical team (Marketing or Sales departments); as well as be able to communicate with the business to understand objectives and business problem.

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 1 – Section 6

Data Science: Job Market Overview

Demand for Data Science

- The statistics listed below represent this significant and growing demand for data scientists.
  - #16 Highest Paying Job in Demand
  - 3,433 Number of Job Openings
  - $105,395 Average Base Salary
  - #1 Best Job in America for 2016

- Sources: 25 Best Jobs in America and 25 Highest Paying Jobs in America for 2016

Data Scientists are “Sexy”

- The Harvard Business Review, a noted authority on “things that are sexy,” has declared “Data Scientist“ to be the sexiest career of the 21st century, publishing an article titled:

  “Data Scientist: The Sexiest Job of the 21st Century “

(Thomas H. Davenport, D.J. Patil October 2012 Issue)

Source

Data Scientists are #1 in the US

For 2016, **Glassdoor** has identified the 25 Best Jobs in America (based on highest overall Glassdoor Job Score, determined by combining three key factors – number of job openings, salary and career opportunities rating). Glassdoor rankings

In # 1 spot: Data Scientist

Job Openings (1,736) in the US

Median Base Salary ($116,840) in the US

Data Scientists are in Demand

- Forbes Published an Article **“The 10 Toughest Jobs to Fill in 2016”,** with Data Scientist in the top 10

Source

- In another article, **“Where Big Data Jobs Will Be In 2015”,** published in 2014, Columbus states:

  “Demand for big data expertise across a range of occupations saw significant growth over the last twelve months”

Source

Job Titles for Data Scientists

Salary Overview - USA

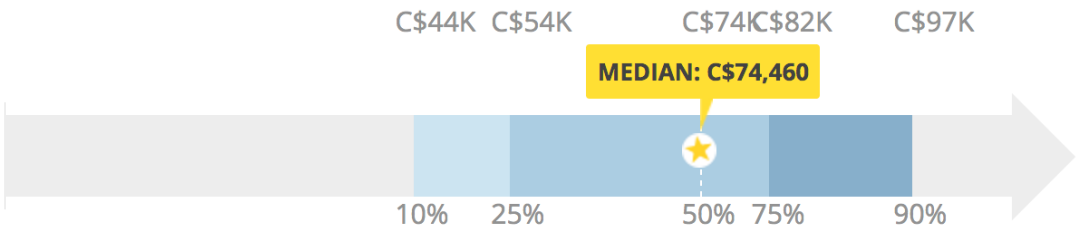
“The average data scientist today earns $123,000 a year, according to Indeed.com” (2016, USA)

“Why Data Scientists Get Paid So Much”
Data Scientist Salaries

Data Scientist Salary - Canada

Data Scientist, IT Salary (Canada)

The average salary for a Data Scientist, IT is C$74,461 per year. Most people with this job move on to other positions after 10 years in this career. A skill in Big Data Analytics is associated with high pay for this job.



Add this chart to your site: 640px 300px

Show Hourly Rate

+ experience | job

National Salary Data (?)	
Salary	C$43,781 - C$97,112
Bonus	C$0.00 - C$8,000
Profit Sharing	C$2,500
Total Pay (?)	**C$40,749 - C$100,688**

C$0 C$50K C$100K C$150K

© Payscale, Inc. @ www.payscale.com

Country: Canada | Currency: CAD | Updated: 28 Oct 2016 | Individuals Reporting: 68

Source

Salary Overview – Python

Python Programmer Big Data Salary

What Python Programmer Big Data — Job Title, Keywords

Where Toronto, ON — City, State or Zip

Add Comparison

☐ **Search Job Titles Only** View Salary

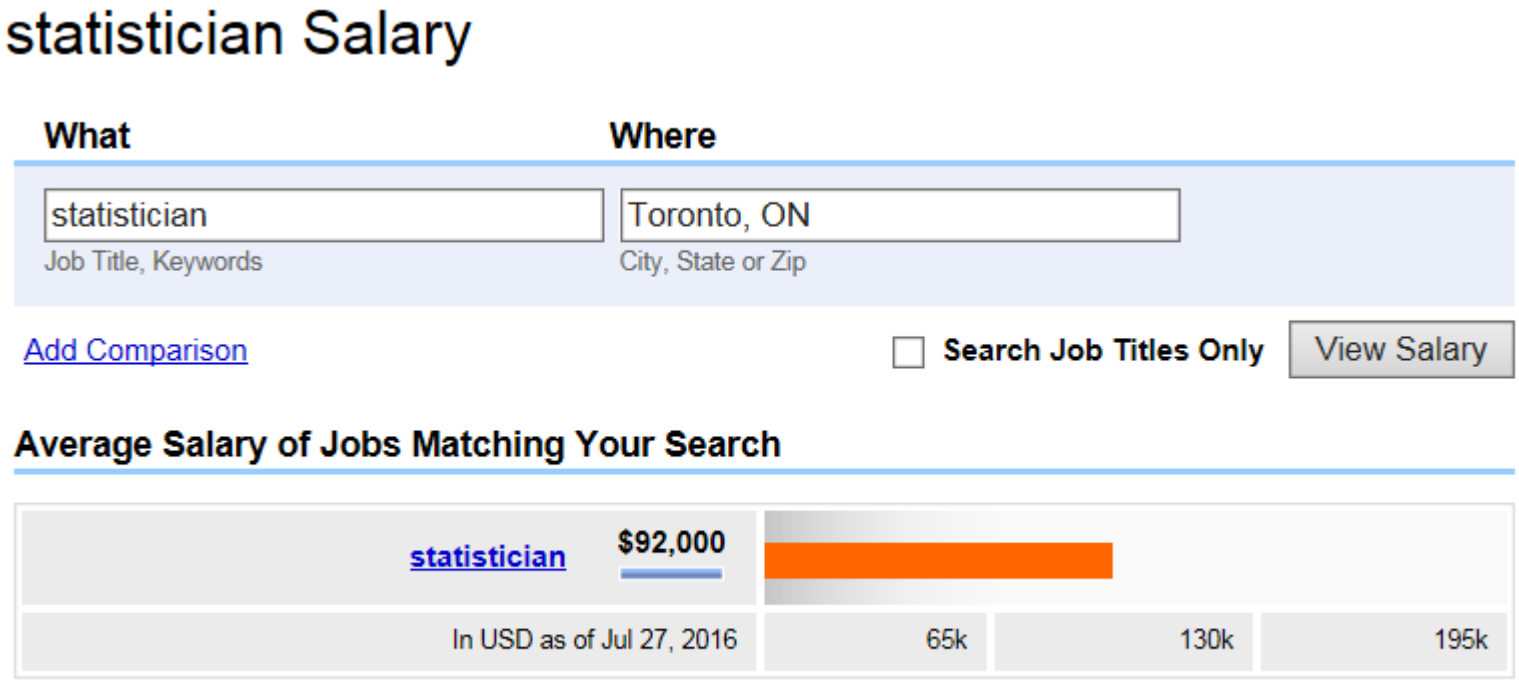
Average Salary of Jobs Matching Your Search

Python Programmer Big Data	$112,000			
In USD as of Jul 27, 2016		40k	80k	120k

Average Python Programmer Big Data salaries for job postings nationwide are 95% higher than average salaries for all job postings nationwide.

Source

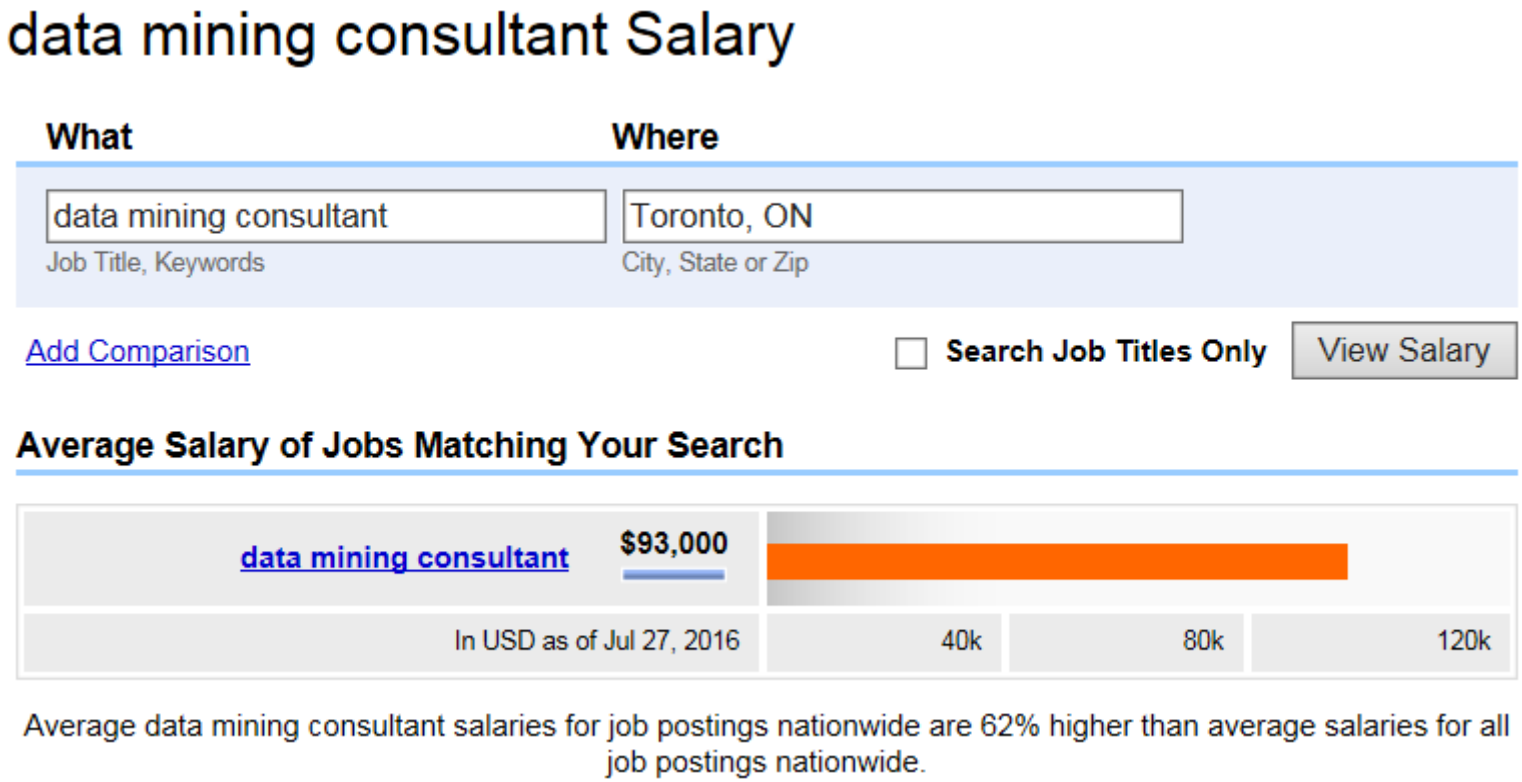
Salary Overview - Statistician



Average statistician salaries for job postings nationwide are 59% higher than average salaries for all job postings nationwide.

Source

Salary Overview – Data Mining



Source

What Determines Salary?

- **Experience** - people who more experience, get paid more
- **Managerial roles** - managers and directors in this field do get paid more
- **Academic achievement.**

  More degrees = more $
- **Company size –** start-ups may not be able to pay top $, however many start-ups love to hire data scientists

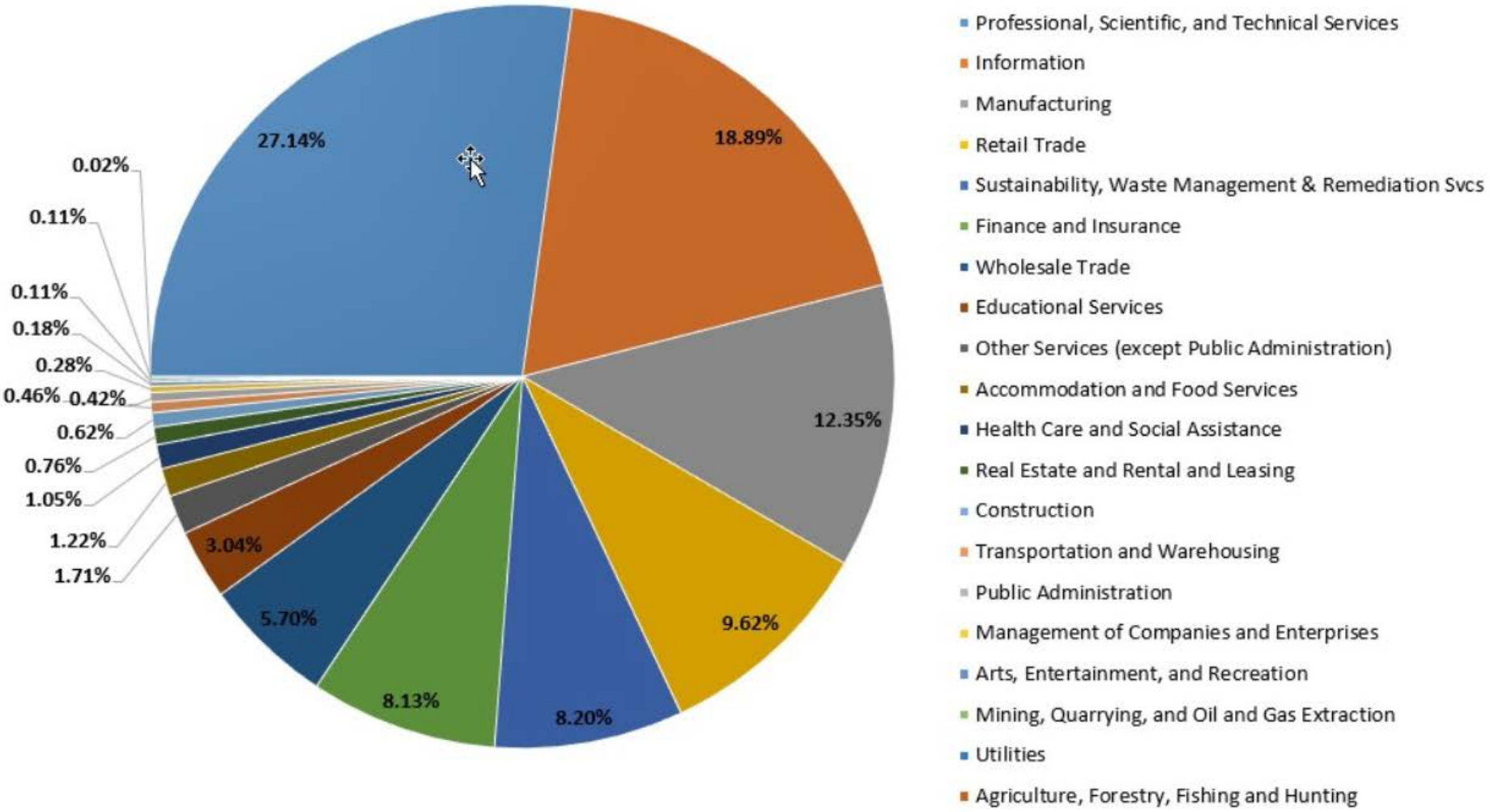
Who is Hiring Data Scientists?

Any company that has a great deal of “Big Data” would seek out a Data Scientist

- Banking and Finance
- Insurance
- Healthcare
- Biotechnology
- Pharmaceutical
- Retail
- Marketing
- Social Media
- Energy Sector
- Engineering
- Information Technology
- Telecommunication
- Media
- Transportation

Top 20 Industries Hiring Big Data Expertise

Source: Wanted Analytics, 2014

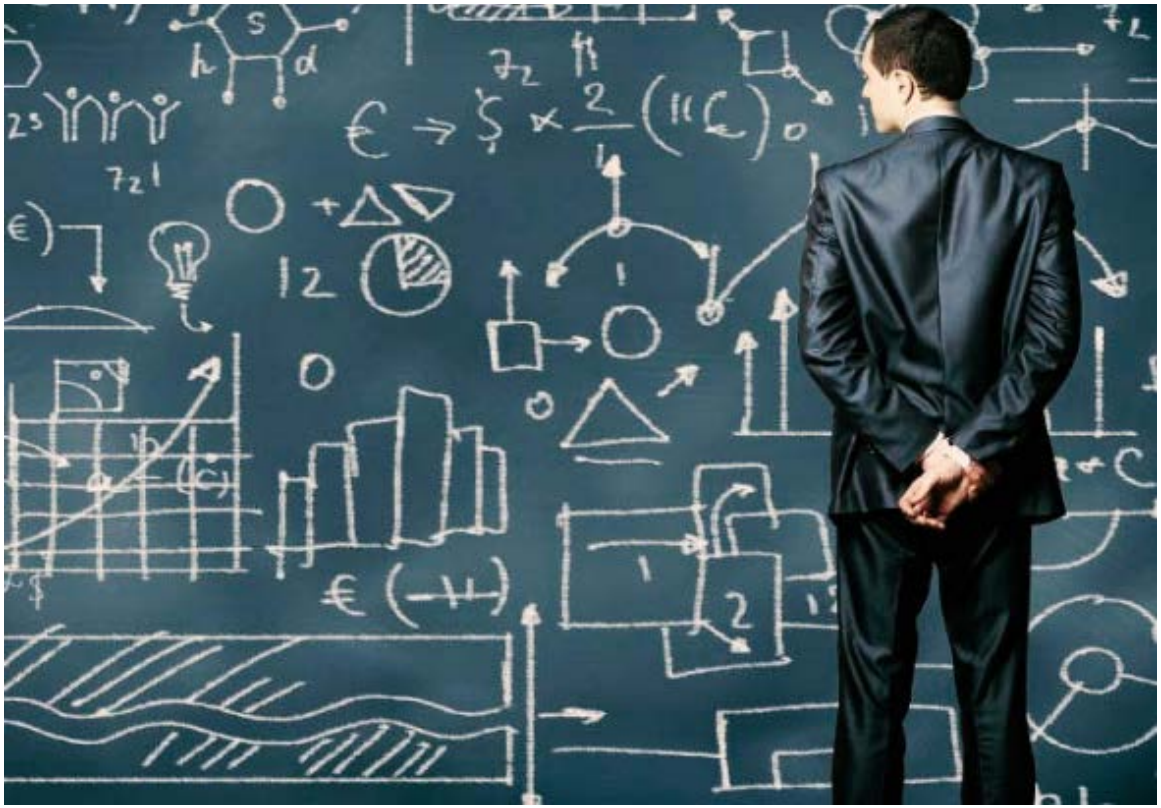


Source

Certifications

The following website provides an extensive overview of certifications:

www.kdnuggets.com



Where there is data… there is Data Science

Quick Poll

Would you like to be a Data Scientist?

A. Yes

B. No

C. Don’t Know

D. Still Thinking About It!

Summary

- Analytics have broad application in business and science
- Data Science brings together ideas from computer science, statistics and engineering to solve new problems
- Business skills (formulating questions, gathering information, building consensus) are essential to applying data science to solving business problems

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Module 1 – Section 7

Homework

Next Class

- In preparation:
  - If you are reading *Think Python*, continue with Ch. 8 – 14
  - Install Anaconda Python according to the instructions provided

- Introduction to Python
  - The core syntax of Python
  - Hands-on

Follow us on social

Join the conversation with us online:

facebook.com/uoftscs

@uoftscs

linkedin.com/company/university-of-toronto-school-of-continuing-studies

@uoftscs

Any questions?

UNIVERSITY OF TORONTO
SCHOOL OF CONTINUING STUDIES

Thank You

Thank you for choosing the University of Toronto School of Continuing Studies