Data Science Applications & Use Cases

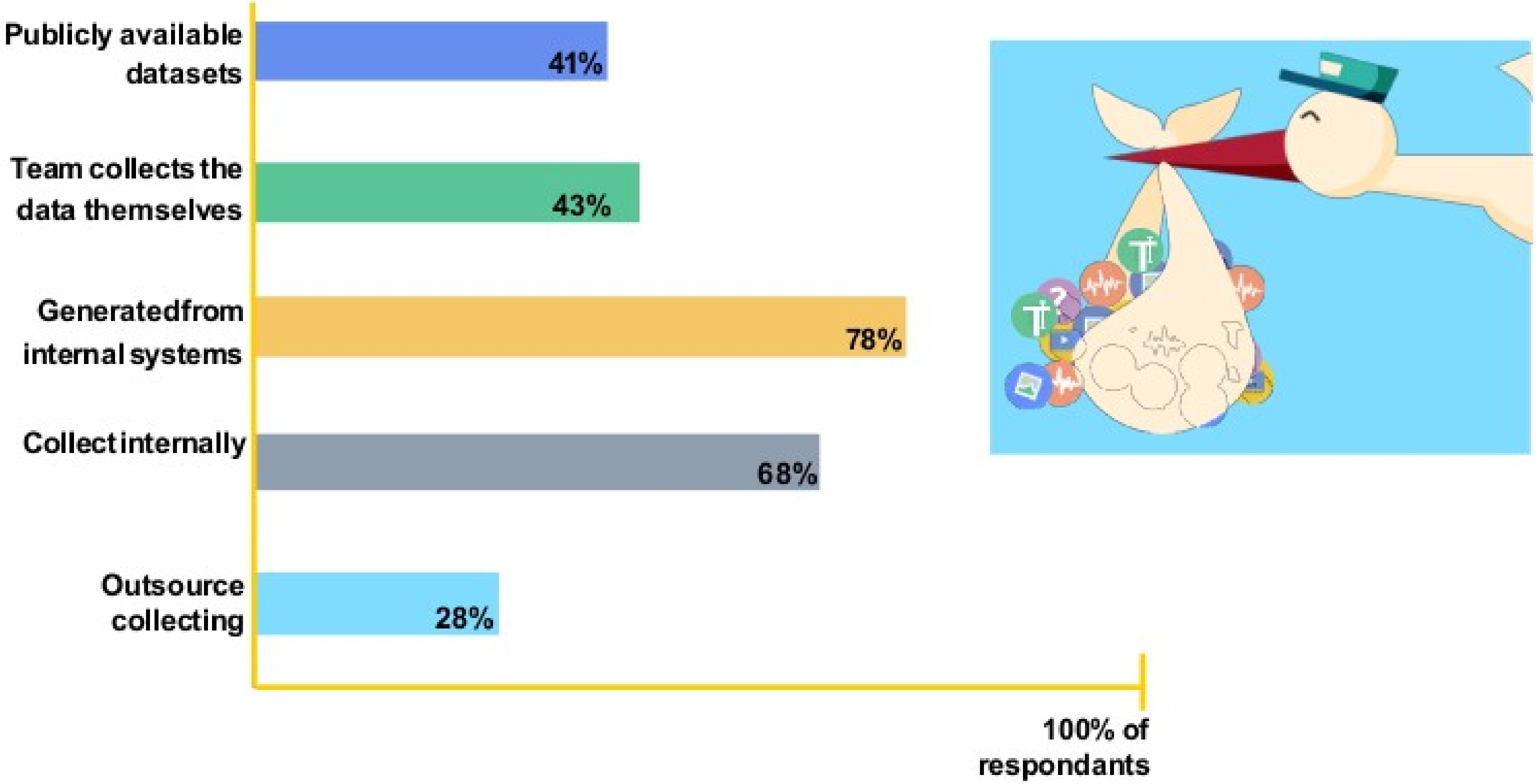
Instructor: Joseph.

Objectives

Objectives

- Understand Big Data Challenges
- What exactly is Data Science
- What do Data Scientists do
- Case Study & Use Cases

Where does data come from?

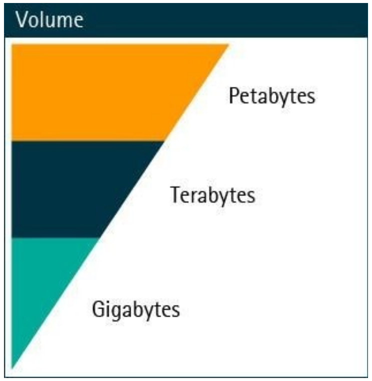


Data All Around

- **Lots of data is being collected and warehoused**
  - Scientific Experiments
  - Internet of Things
  - Health, Agriculture, Marketing
  - 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.

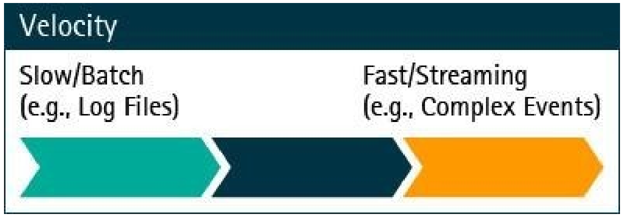


Volume

Petabytes

Terabytes

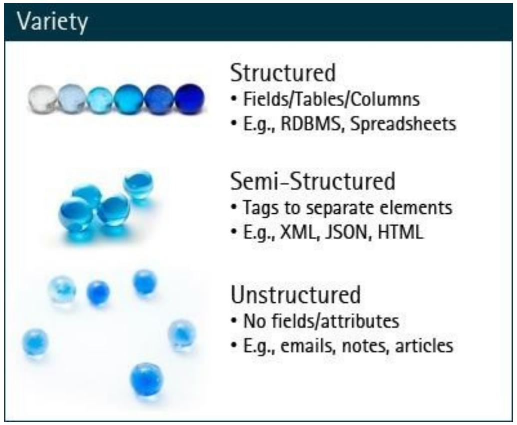
Gigabytes



Velocity

Slow/Batch
(e.g., Log Files)

Fast/Streaming
(e.g., Complex Events)



Variety

Structured
- Fields/Tables/Columns
- E.g., RDBMS, Spreadsheets

Semi-Structured
- Tags to separate elements
- E.g., XML, JSON, HTML

Unstructured
- No fields/attributes
- E.g., emails, notes, articles

What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching
- Knowledge discovery
  - Data Mining
  - Statistical Modeling
- Data Driven
  - Predictive Analytics
  - Machine Learning Models

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.

Definition…..

- **Data Science** is the science which uses computer science, statistics and machine learning, visualization and human-computer interactions to collect, clean, integrate, analyze, visualize, model, interact with data to create data products.

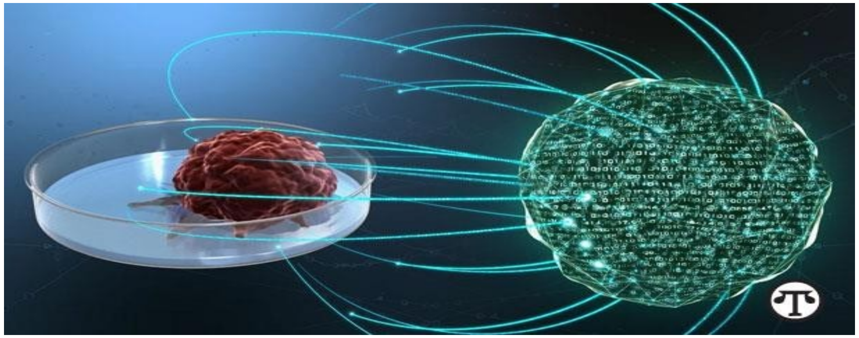
- * Data products are typically referred to in the business space, meaning any application of data that is of value to the business

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"

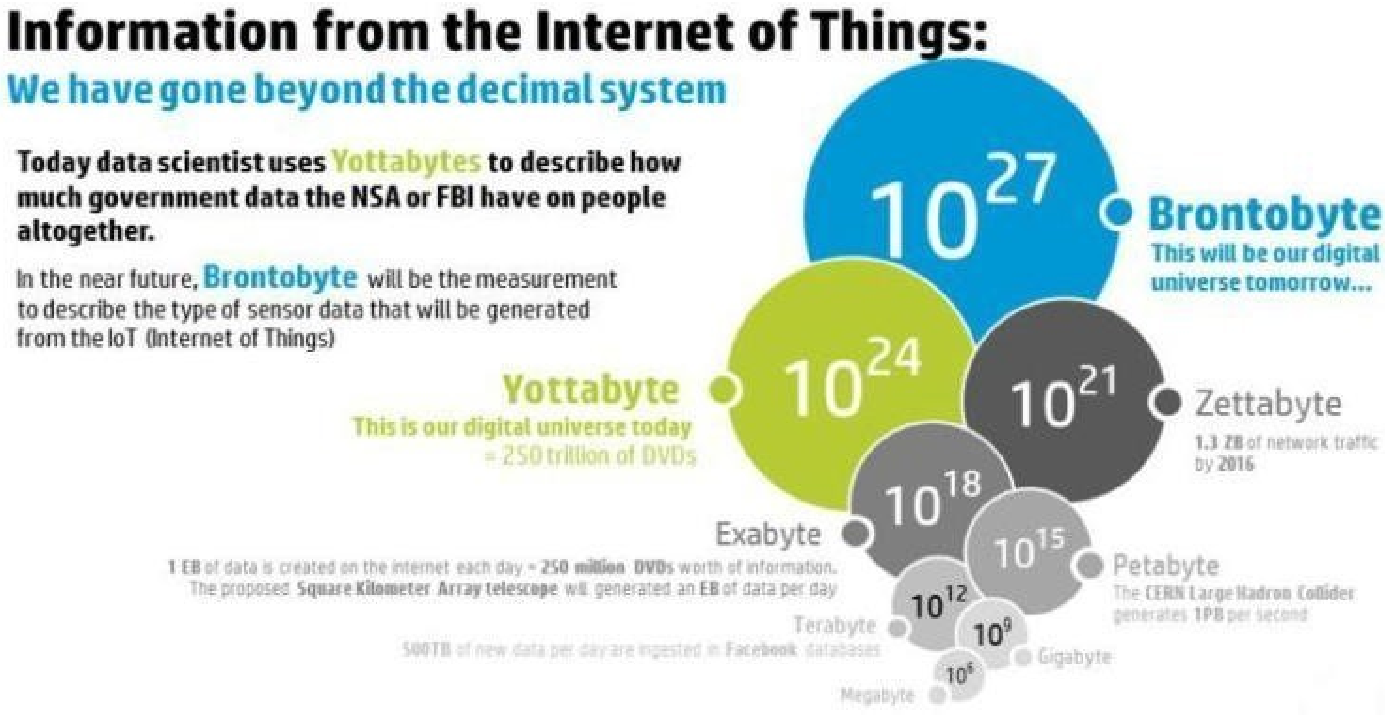
Data Science: Case Study
Health Research

- 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
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

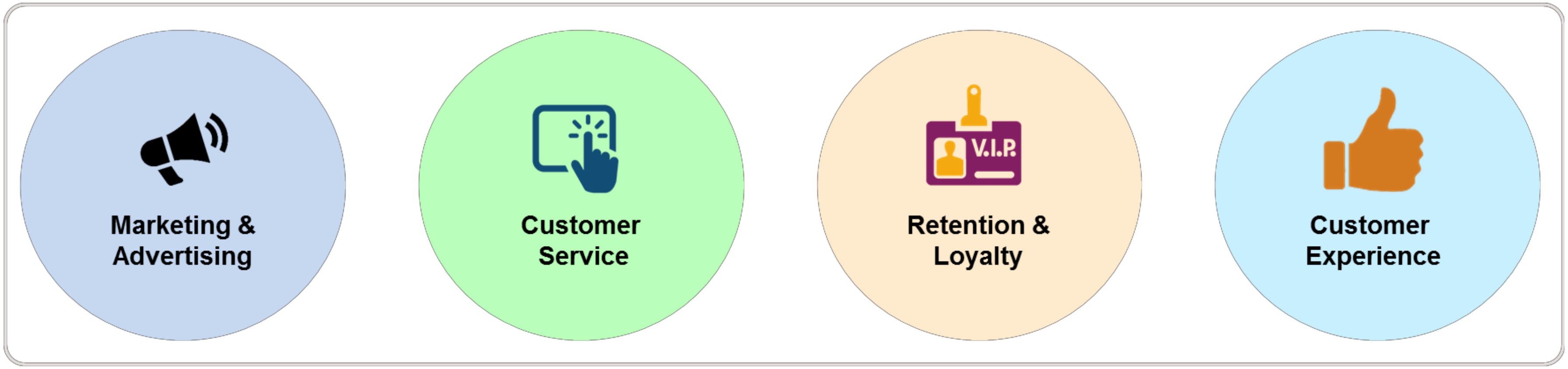
Today data scientist uses Yottabytes to describe how much government data the NSA or FBI have on people altogether.

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.

Data Science: Case Study
Customer Analytics



Marketing & Advertising

Customer Service

Retention & Loyalty

Customer Experience

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.

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.

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.

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.

Companies learn your secrets, shopping patterns, and preferences.

…..many more..

- Fraud and Risk Detection.
  - Loan defaulters, Loan qualifications etc
- Internet Search.
- Targeted Advertising, Product recommendations
- Website Recommendations.
- Advanced Image Recognition.
- Speech Recognition.
- Airline Route Planning.
- Text Messaging – Chat bots
- Transport – Price Prediction
- Emails … Spam, Autocomplete features

The Data Science Pipeline

-
- The five key steps in Data Science:
- 1. **Acquisition**: Acquire data from a variety of sources, including RDBMS systems,
- NoSQL and document store, webscraping, Data Lakes, HDFS, etc.
- 2. **Exploration & Understanding:** Understanding the data that you will use and
- how it was collected, often requiring significant exploration
- 3. **Munging, Wrangling, & Manipulation**: Single most time-consuming and
- important step in the pipeline - data is rarely in the needed form for analysis
- 4. **Analysis & Modeling:** The fun part where the data scientist explores the
- statistical relationship between the variables in the data and uses a bag of
- machine learning tricks to cluster, categoralize, classify Predictive Models
- 5. **Communicate & Operationalize**: Give the data back in a compelling form and
- structure - one-off report, scalable web product, interactive data, ...

Conclusion

In this section you have learned

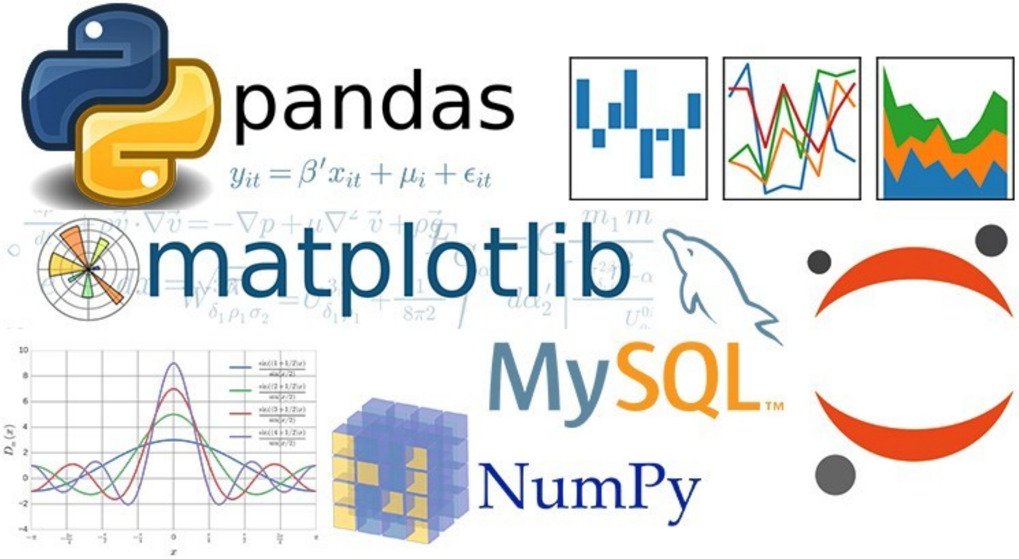
- **What exactly is Data Science and what do Data Scientists do**
- **Data Science contrasted with other disciplines**
- **Case Study & Use Cases**

Questions

Python Libraries for Data Science

- Python libraries.
- Standard Python libraries useful for Data Analysis with Python are:
- **Pandas** - for data munging and preparation
- **NumPy** - abundance of useful features for operations on n-arrays and matrices in Python
- **SciPy** - library of software for engineering and science
- **Matplotlib** - tailored for the generation of simple and powerful visualizations
- **Statsmodels** - data exploration by using methods of estimation of statistical models
- **scikit-learn** - concise & consistent interface to the common ML algorithms
- **Seaborn** - visualization of statistical models; based & highly dependent on Matplotlib
- **Bokeh** - is aimed at interactive visualizations; independent of Matplotlib - main focus is
- interactivity, with presentation via modern browsers in the style of Data-Driven Documents
- **Plotly** - web-based toolbox for building visualizations, exposing APIs to Python, etc.
- **Theano / TensorFlow / Keras**
- **NLTK** - Natural Language Toolkit - tasks of symbolic & statistical NL processing
- **Gensim / Scrapy**

DATA SCIENCE
python


pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
matplotlib
MySQL
NumPy