澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Course Information

CISC7204: DATA SCIENCE & VISUALIZATION

Derek F. Wong

NLP²CT – *Natural Language Processing & Portuguese-Chinese Machine Translation Research Group*

derekfw@um.edu.mo

E11-4010 (Ext: 4478)

Office Hours: Thu – 16:00~17:30, Fri 11:00~12:30

Data Science & Visualization

Learning Outcomes

1. *State the development and principles of data analytics and data visualization*
2. *Identify different types of data and use appropriate analysis techniques best to explore them*
3. *Draw conclusions and formulate hypotheses from data presented graphically*
4. *Apply theories of data analytics and data visualization and competence in using software for data visualization and data analytics*
5. *Analyze, critique, and revise data visualizations*

Data Science & Visualization

Course Information

There is no single textbook. However, there are a couple of books that are particularly useful and we will reference:

- Vijay Kotu, and Bala Deshpande (2019). *Data Science: Concepts and Practice*. Elsevier.
- Kristen Sosulski (2019). *Data Visualization Made Simple*. Taylor & Francis.
- Kirthi Raman (2015). *Mastering Python Data Visualization*. Packt Publishing.
- Cathy O'Neil, and Rachel Schutt (2013). *Doing Data Science*. O'Reilly.
- Cole Knaflic (2015). *Storytelling with Data*. Wiley.
- Ryan Sleeper (2018). *Practical Tableau*. O'Reilly.

Course Website

- *http://ummoodle.umac.mo/*

Data Science & Visualization

Course Structure

Assignments

- Some *hands-on* exercises
- *No way* to really internalize *without doing* it

Mini Project

- Chance to explore a *special interest* at mid of term
- Implement and extend based on the *selected topic*

Classroom participation

Data Science & Visualization

Assessment

Evaluation Method

- Assignments 30% Exams 30% Projects 40%

Assignment Policies

- Late submissions
  - *Deduct 15% for 1 day late Deduct 30% for 2 days late*
  - *Deduct 50% for 3 days late Receive 0% for 4 or more days late*

Class Participation

- In-class assignments

Data Science & Visualization

Course Syllabus

We will be intermingling discussions of:

- *Basic Concepts*
  - E.g. *Data Science*, *Big Data*, *Analytical Life Cycle*, *Data Interpretation* & *Visualization*
- *Theoretical Foundations*
  - E.g. *Classification*, *Regression*, *Clustering*
- *Analytical Processes*
  - E.g. *Data Acquiring, Data Preprocessing, Model Selection*
- *Interpretation*
  - E.g. *Data Interpretation, Data Visualization*

Tableau & Python Programming

Software Tools

Instruction will be focused and directed based on the capabilities and features of visualization software:

- *Tableau Desktop Professional* (TFT License), Student License or Tableau Public
- *Python* Programming Language
- Microsoft Excel (Win 2007/Mac 2008 or Win 2010/Mac 2011 or Win 2013) - Optional

Noted: A full copy of Tableau Desktop is available to full-time students for free for a year, available from Tableau (https://www.tableau.com/academic/students).

Your Information & Expectation!

Survey of your background, knowledge and skills!

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Introduction

CISC7204: DATA SCIENCE & VISUALIZATION

Derek F. Wong

NLP²CT – *Natural Language Processing &*
Portuguese-Chinese Machine Translation Research Group

derekfw@um.edu.mo

E11-4010 (Ext: 4478)

Office Hours: Thu – 16:00~17:30, Fri 11:00~12:30

Natural Language Processing & Portuguese – Chinese Machine Translation Laboratory
自然語言處理與中葡機器翻譯實驗室

Content

- What is *Data Science*?
  - Differences between *Big Data* & *Data Analyst*
  - *Life Cycle* of Data Science Project?
- What is *Data Visualization*?
  - *Tools* for Data Visualization

A Warn Up Exercise

What is Data Science? Please define data science in your own words!

What is Data Science?

Definitions

- *"It's what a data-scientist does"*
- *"Machine learning/data mining/statistics"*
- *"Collecting, manipulating, and analyzing data in order to extracting value from it"*
- Wikipedia: *"Data Science is the extraction of knowledge from data, which is a continuation of the field of data mining and predictive analytics"*
- NIST Big Data Working Group: *"Data Science is the empirical synthesis of actionable knowledge from raw data through the complete data lifecycle process"*

What is Data Science?

A 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*, *interact* with *data* to *create data products*

Turn *data* into *data products*!!

(O'Reilly Radar report, 2011)

What's Driving Data Deluge?

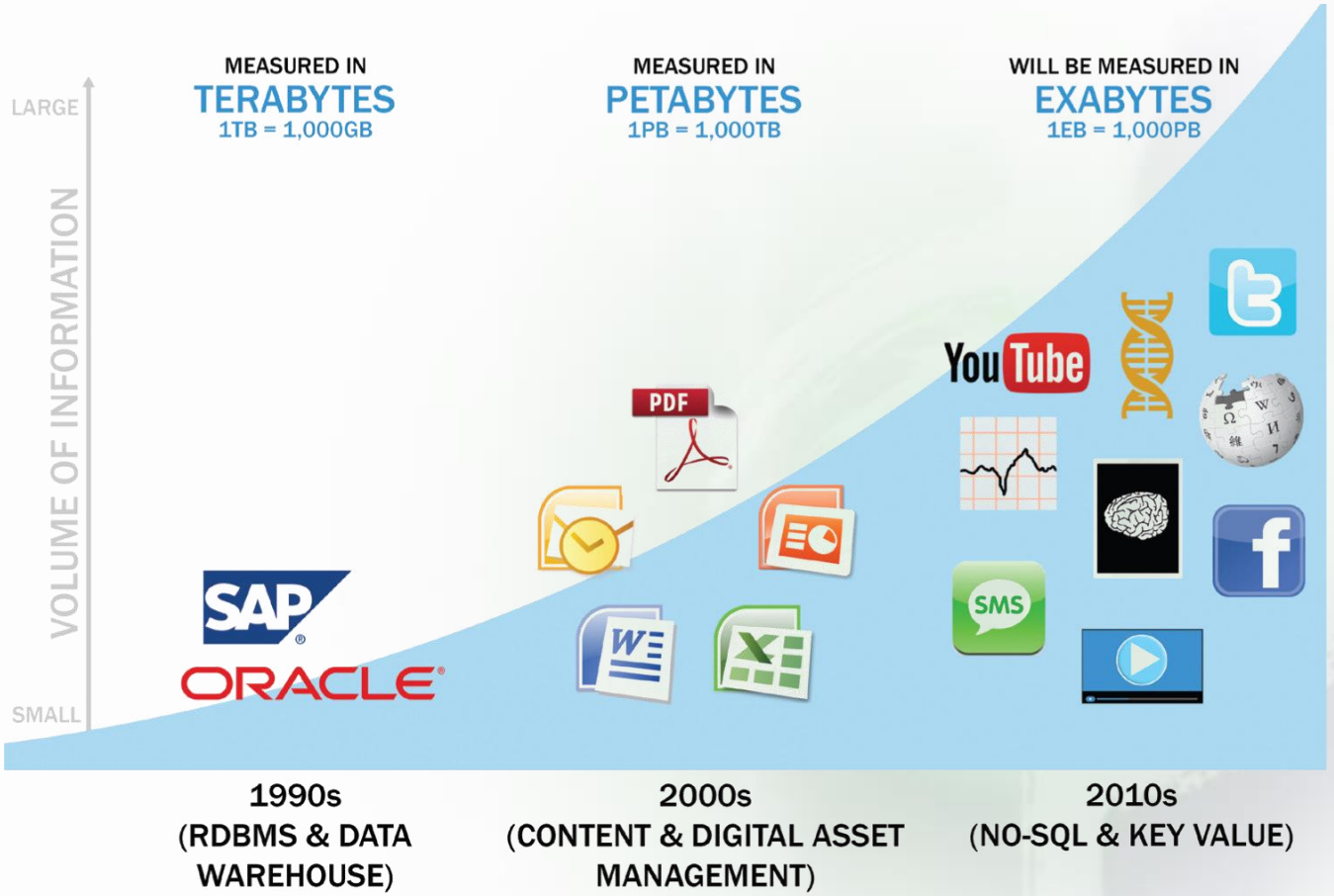
Big Data's 4 Vs

- Volume
  - *Scale of data*
- Velocity
  - *Analysis of streaming data*
- Variety
  - *Different forms of data*
- Veracity
  - *Uncertainty of data*

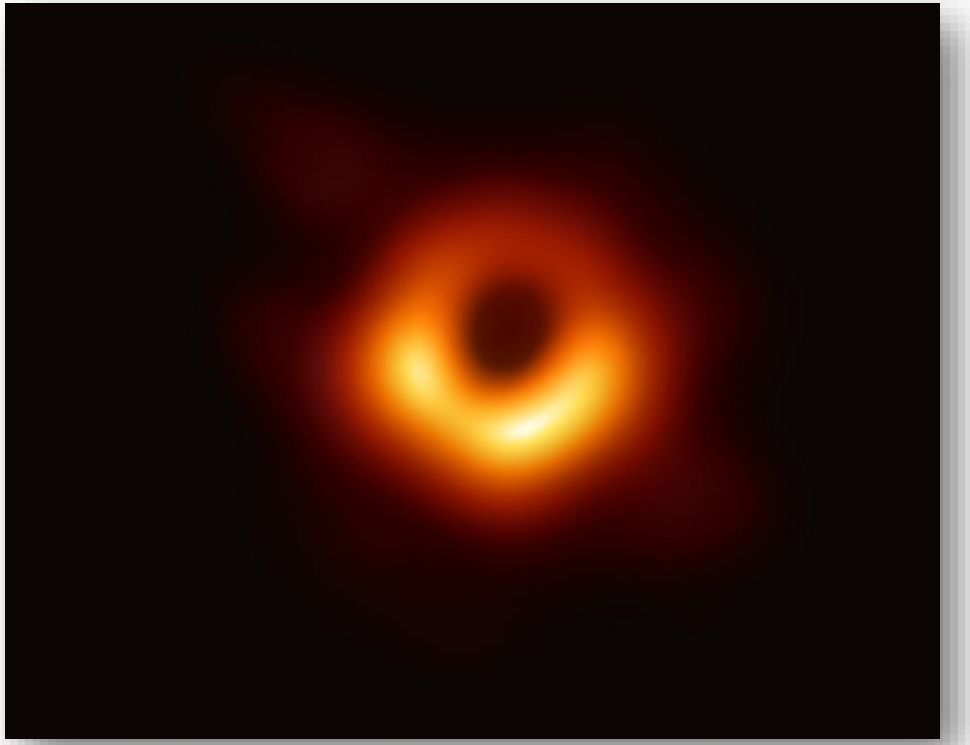
The 4 V's of Big Data

Volume: Scale of Data

Big data exceeds the *storage capacity* of *conventional* databases



Behind the 1st Black Hole Image



2019@Event Horizon Telescope

Data collection began in 2017

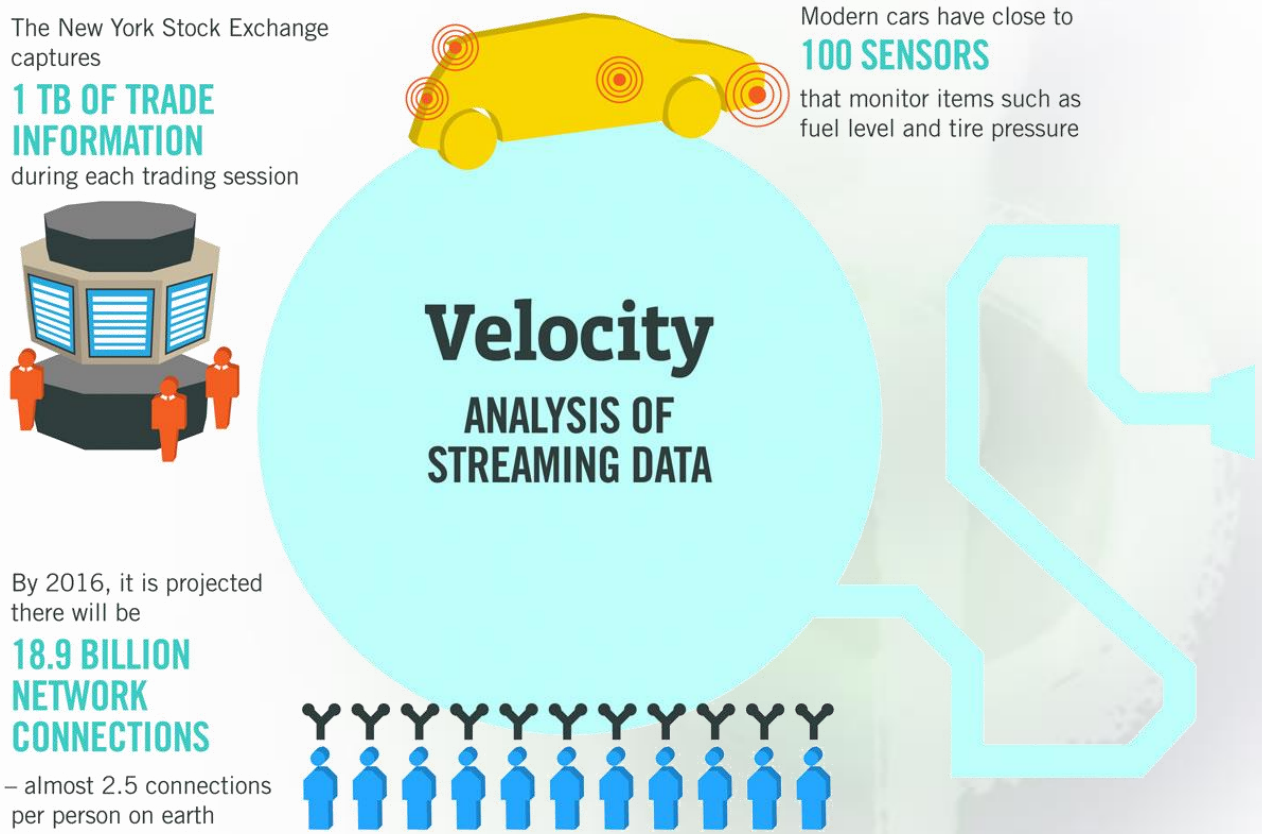
2019@ExtremeTech

"At the end of that, we had five petabytes of data recorded… it amounts to more than half a ton of hard drives. Five petabytes is a lot of data. It's equivalent to 5,000 years of MP3 files, or according to one study I read, the entire selfie collection over a lifetime for 40,000 people." – by Dan Marrone, University of Arizona.

The 4 V's of Big Data

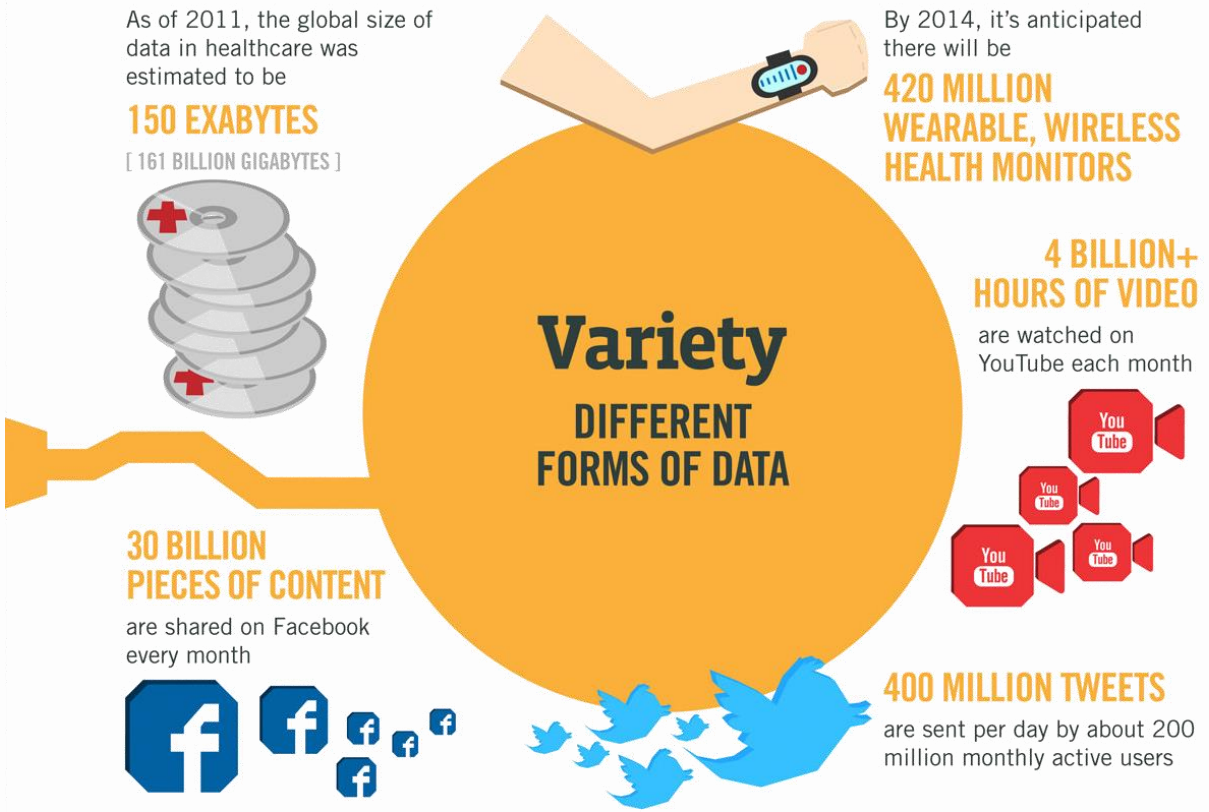
Velocity: Explosion of Data

Data velocity is *accelerating*. Streams of *tweets*, *Facebook* entries, *financial* information, etc., are being generated by *more users* at an ever increasing pace



The 4 V's of Big Data

Variety: Different Forms of Data

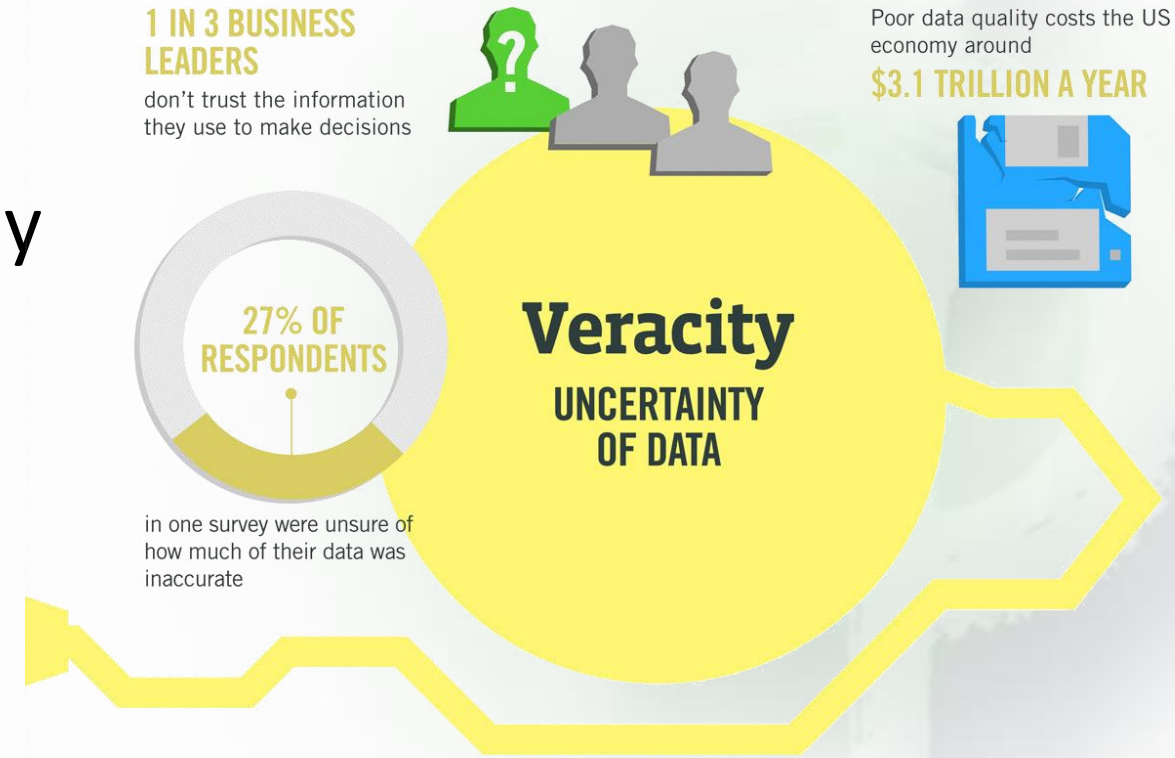


Data today comes from *many kinds of data sources*, and the level in which that data is *structured varies greatly* from data source to data source

The 4 V's of Big Data

Veracity: Uncertainty of Data

The *value* of almost anything and everything is directly proportional to the *quality of data*, and is *affected* by the way it is *entered*, *stored*, and *managed*



History of Data Science

Introduced knowledge worker

Excel became to tool knowledge worker

Industry & Business

1955s — 1965s — 1975s — 1985s — 1995s — 2000s — 2005s — 2010s — 2015s

Taylorism adopted across industrial revolution

Peter Drucker

Adopting production system in manufacturing

Interneting the world

Service industry transformed customer experience & improved with data

ML & AI became mainstream

Academia, Eng, Math, Comp.

1955s — 1965s — 1975s — 1985s

Computers & AI after World War II

Awareness importance of Visualizing data

HP - First desktop computer

John Tukey

Operating Systems, MS Windows

Boom - Company take their business online

Smartphones & on-demand economy era

(Kumar, 2018)

What Is Data Science?

- Data Science is a *blend of various tools*, *algorithms*, and *machine learning principles* with the *goal* to *discover hidden patterns* from the raw data
- It also involves *solving a problem* in various ways to arrive at the *solution*
- It involves to *design* and *construct* new processes for *data modeling* and *production* using various *prototypes*, *algorithms*, *predictive models*, and *custom analysis*

What are Big Data & Data Analytics?

- ***Big Data***
  - *Large amounts of data* which is pouring in from *various data sources* and has *different formats*
  - To analyze the *insights* which can lead to *better decisions* and *strategic* business moves
- ***Data Analytics***
  - The *science* of *examining raw data* with the purpose of *drawing conclusions* about that information
  - *Discovering useful information* from the data to *support decision-making*, involving *inspecting*, *cleansing*, *transforming* & *modeling data*

Role of Data Scientist

Predicts future based on *past patterns* making use of *AI* and *Machine Learning* algorithms

Finding co-relations, *hidden patterns*, *market trends* from data

Examines data from *multiple sources* to discover insights

Data Scientist

Skill-set Required

Skill Requirements

- Statistical & Analytical Skills
- Machine Learning Principles
  - *Data Mining Activities*
- In-depth Knowledge of Programming
  - *Python Programming*
  - *SQL Database/Coding*
  - *SAS or R Coding*
- Co-relation
  - *Data Visualization*

Role of Big Data Professional

Architect distributed systems:
Data structure & Process flow

Build large *scale data* processing system

Process data using *various big data tools* & ensure *network connectivity*

Big Data Professional

Skill-set Required

Skill Requirements

- Statistical & Analytical Skills
  - *Working with Unstructured Data*
- Distributed Technologies
  - *Hadoop, Spark, Hive, etc.*
- General Purpose Programming
  - *SQL Database/Coding*
  - *C, Java, Python, MATLAB*
- Business Skills
  - *Creativity*
  - *Data Visualization*

Role of Data Analyst

Acquire, *analyze* and *process* the *data*

Finding *insights* for the *collected data*

Create data *reports* using various *reporting tools*

Data Analyst

Skill-set Required

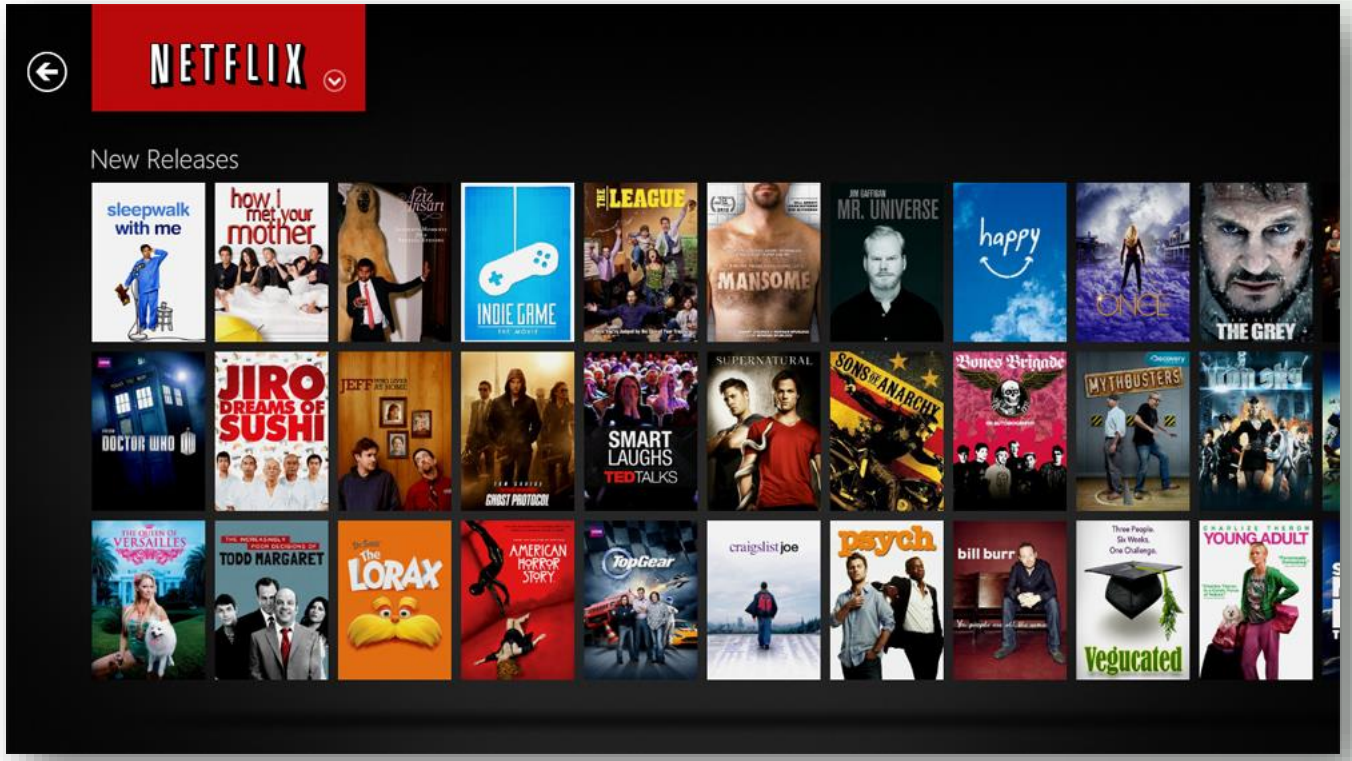
Skill Requirements

- Data Warehousing
  - *Hadoop Based Analytics*
- Adobe & Google Analytics
- Programming Skills
  - *SQL Database/Coding*
  - *Scripting & Statistical Skills*
- Data Interpretation
  - *Data Visualization*
  - *Spread-Sheet Knowledge*

Netflix: Use of DS *vs* BD *vs* DA

An Illustration Scenario

Netflix is a internet TV company providing *online streaming content* as well as *DVD/Blu Ray* rentals direct to home



2019@https://www.netflix.com/mo-en/

Netflix: Role of Big Data Professional

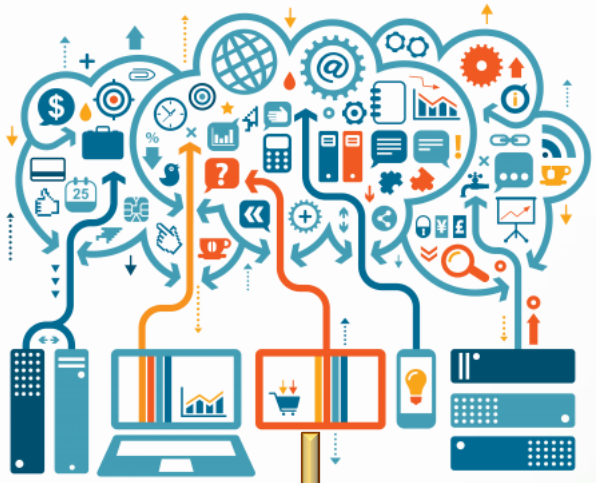
An Illustration Scenario

Netflix generates a *huge amount of data*:

- It is *unstructured*, in forms of:
  - *Text* files
  - *Audio* files
  - *Image & Video* files, and
  - User *preferences*, etc.
- It is *difficult to process* this unstructured data using the *traditional approach*
- Very *complicated task*

Netflix: Role of Big Data Professional

To Process Netflix Data



Designs and *creates* an *environment* using *Big Data tools* to store and process the Netflix Data

Big Data Professional

Netflix: Role of Data Scientist

Optimizes Streaming Experience

Understanding the impact of QoE on user behavior

Improving the streaming experience

Optimize content caching

Improving content quality

Data Scientist

Netflix: Role of Data Scientist

Optimizes Streaming Experience

Understanding the impact of QoE on user behavior

Quality of experience (**QoE**):

- How user *Interacts* with the Netflix
  - Understand and predict behavior
- *Number of hours* that members watch?
- How often *playback* is temporarily interrupted (*rebuffer rate*)?
- What is the *quality* of the picture (*bitrate*)?

Data Scientist

Netflix: Role of Data Scientist

Optimizes Streaming Experience

Improving the streaming experience

How do provide the *best user experience*?

- Look at the *algorithms* for *playback*
- Determine the *bitrate* to be *served*
- Determine which *server* to download the *content*

Data Scientist

Determine all *the factors* to improve the *streaming experience*

Netflix: Role of Data Scientist

Optimizes Streaming Experience

Optimize caching of content

Are there any problems on *content delivery side*?

- To locate the content *closer* to Netflix members, i.e. *network hops*
- Monitor *behavior* of the members being served and the *experience*

Data Scientist

Then, one can *optimize* the *decisions* around *content caching*

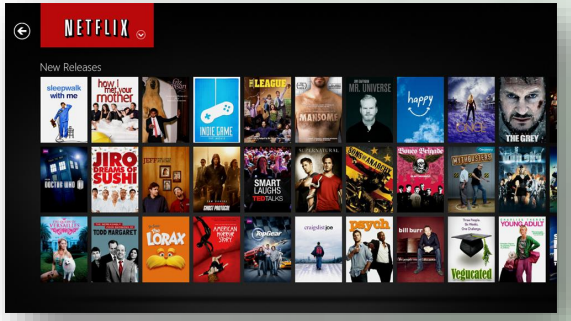
Netflix: Role of Data Scientist

Optimizes Streaming Experience

Improving content quality

User experience involving quality of content

- Look at the *quality* of *video*, *audio*, *subtitles*, *closed captions*, etc.
- *Take feedbacks* from users/members, *record* the *reported issues*



Then, combining member feedback with *intrinsic factors*, *build model to predict quality issue* using *machine learning* with *natural language processing* (NLP), *text mining techniques*, etc.

Data Scientist

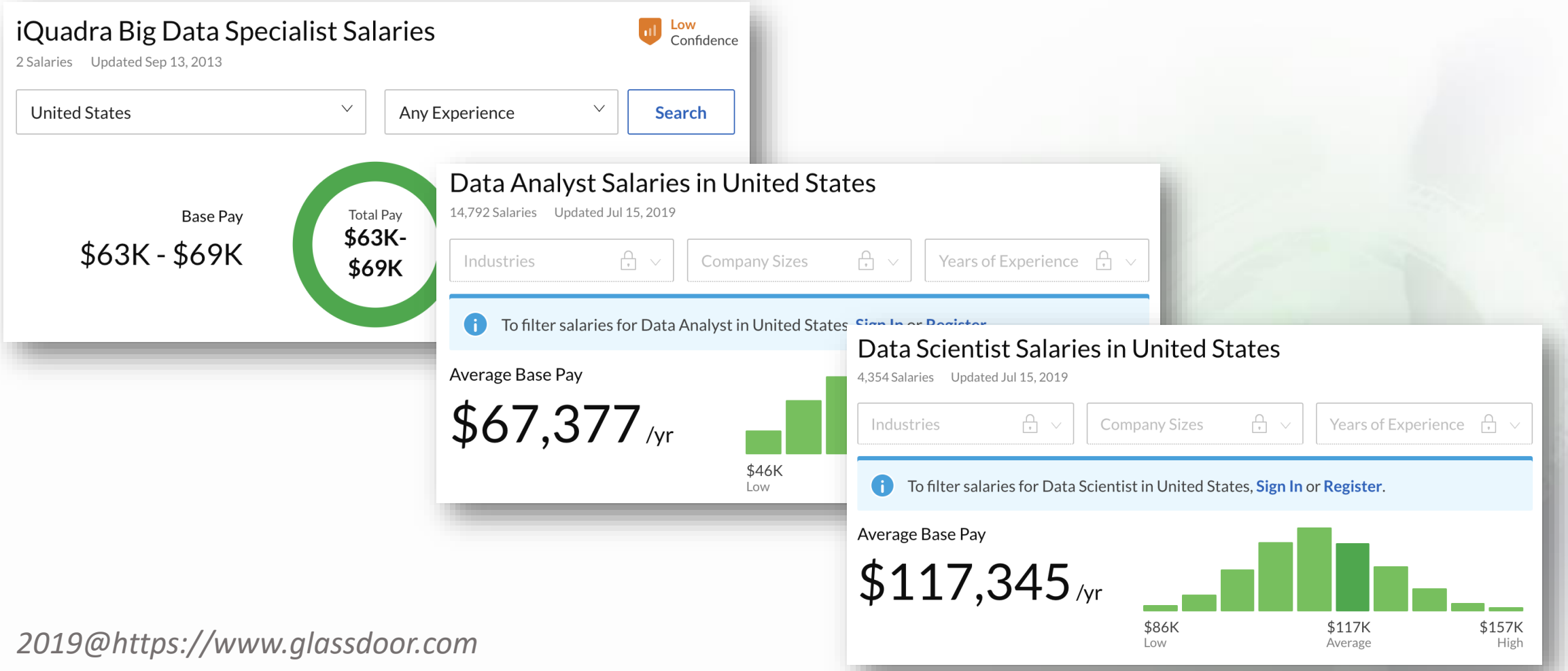
Netflix: Role of Data Analytics

Drive Netflix Success

Data Analyst

- *Capture* user *activities*, analyze *preferences of user*
- *Create personalized member profile* based on preferences
- *Predict* and *recommend videos* to members

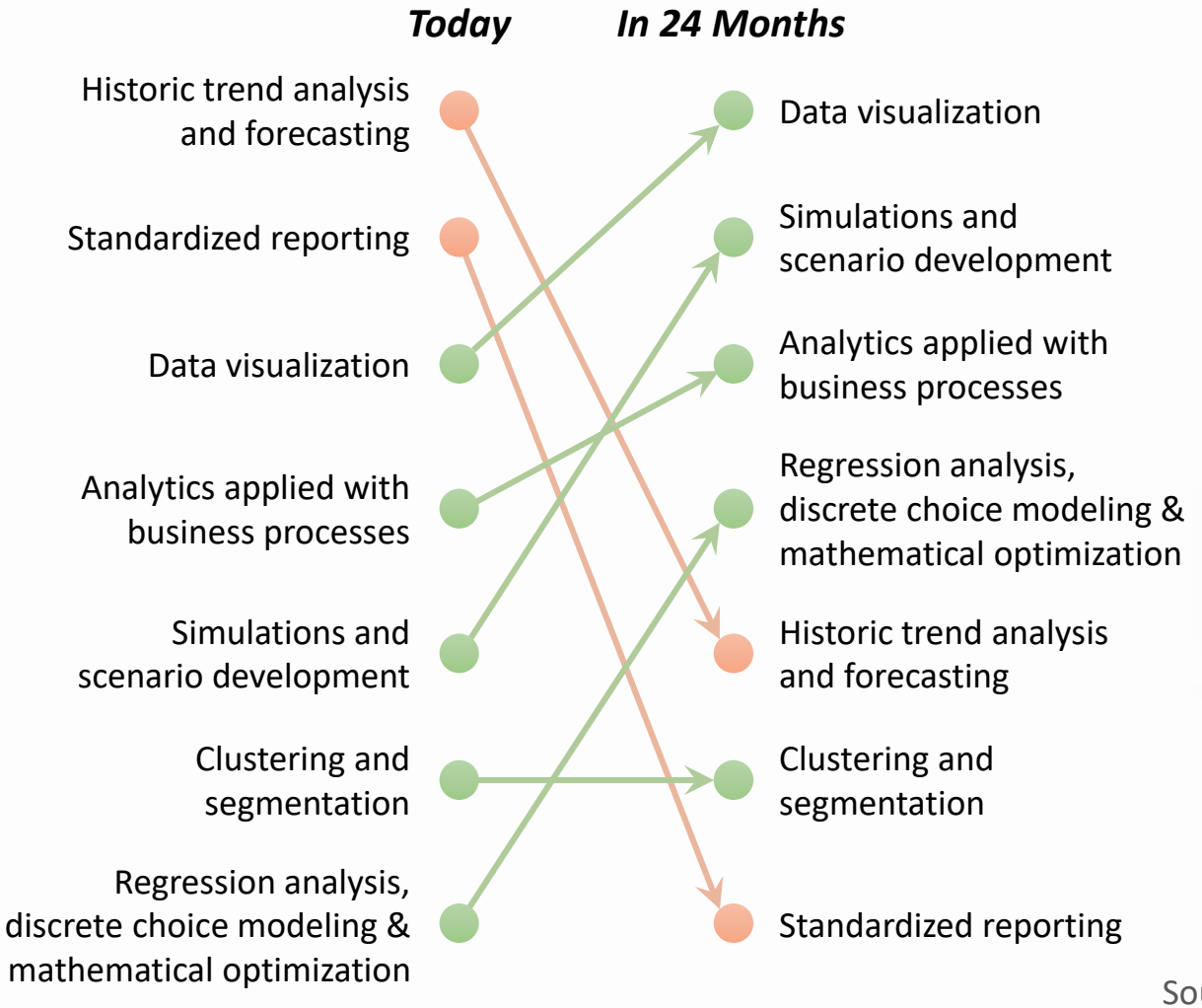
What Salaries Do They Get?



2019@https://www.glassdoor.com

Analytics: A New Path to Value

MIT 2010

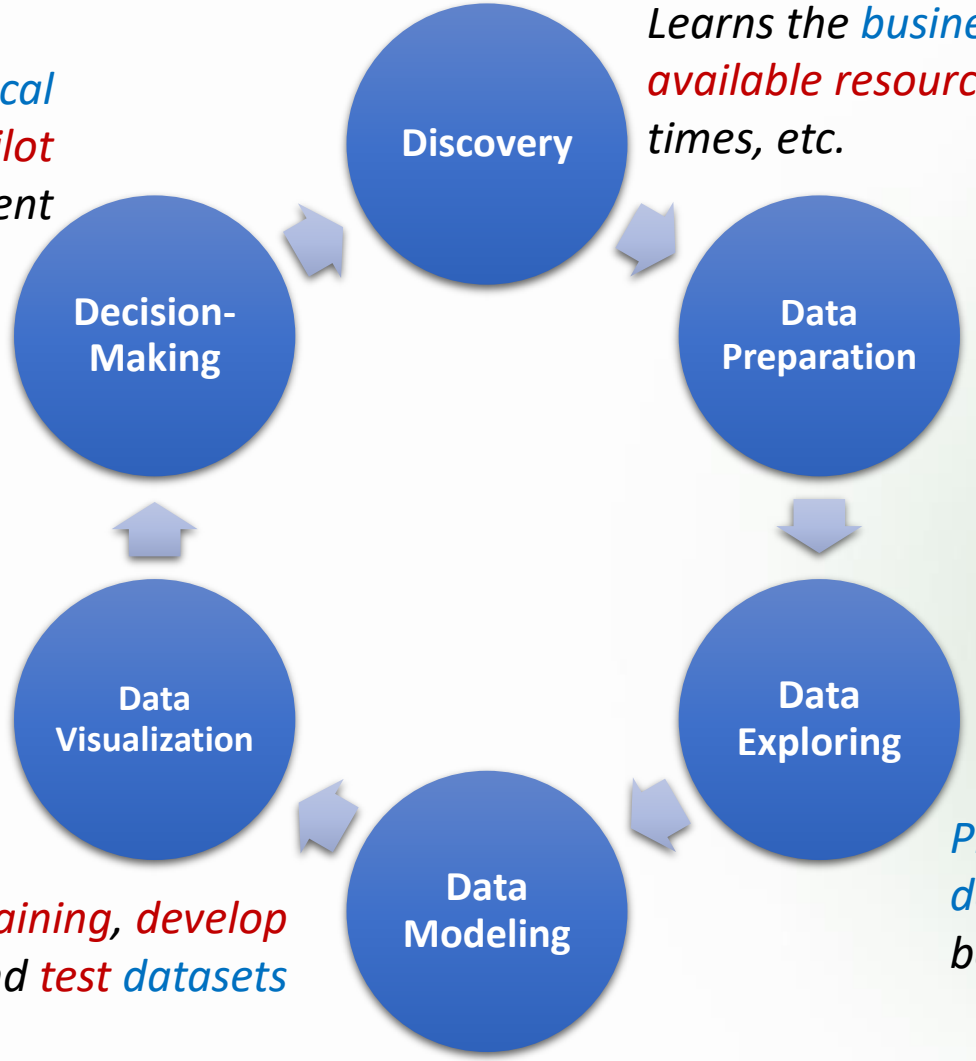


Respondents were asked to identify the *top three analytic techniques* creating *value* for the organization, and predict which three would be creating the most value in *24 months*

Source: Big Data, Analytics and the Path From Insights to Value
MIT Sloan Management Review, December 2010

Life Cycle of Data Science Project

Data Analytics Life Cycle



Learns the business domain, accesses available resources, e.g. data, technologies, times, etc.

Presence of an analytic sandbox, get data into the sandbox, and get familiar with the data

Plan a model used to explore the data, learn about the relationships between variables

Build a model based on training, develop and test datasets

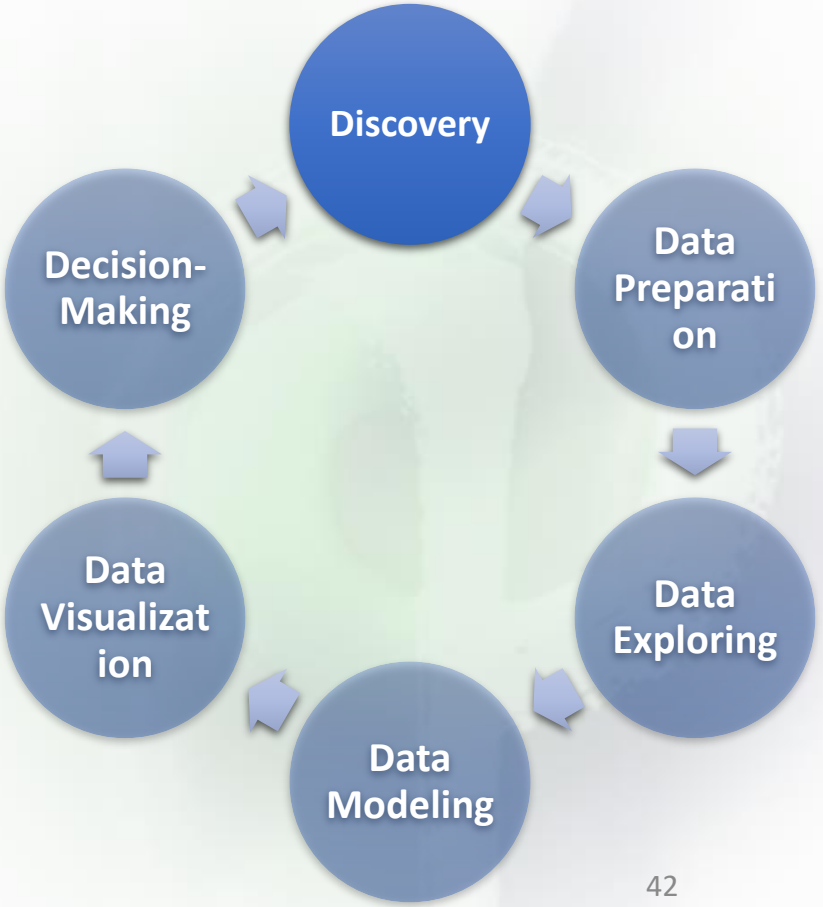
Analyze and summarize the findings, quantify the business value and convey findings to stakeholders

Deliver final reports, technical documents, and implement pilot project in production environment

Data Analytics Life Cycle

Discovery/Acquisition

- Learning the *business domain*
  - *To understand the problem, determine business knowledge needed*
- *Resources*
  - *To assess the resources available to support a project, e.g. tools, technologies, data*
- Framing the *problem*
  - *To state the analytics problem to be solved, and objective*
- Identifying *key stakeholders*
  - *To identify the key stakeholders and their interests in the project*
- Developing *Initial hypotheses*
  - *To define ideas that the team can test with data*
- Identifying *potential* data sources
  - *Identify data sources, capture aggregate data sources, review the raw data, evaluate the data structures, scope of data infrastructure*

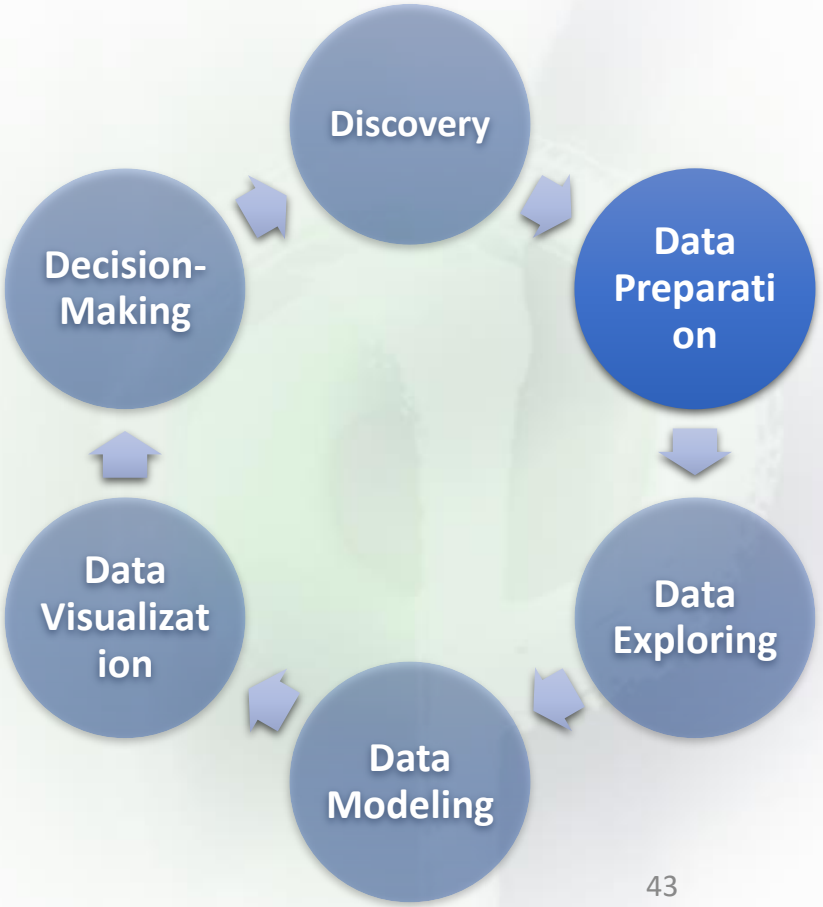


Data Analytics Life Cycle

Data Preparation

The *steps* to *explore*, *preprocess*, and *condition* data *prior* to modeling and analysis

- Preparing the *analytic sandbox*
  - *The workspace to explore the data without interfering with live production Database*
- Performing *ETLT*
  - *Perform extract, transform, load processes to extract data from a datastore, perform data transformations and load back into datastore*
- *Learning* about the data
  - *Classify the data, highlight gaps, identify useful data*
- Data *conditioning*
  - *Process of cleaning data, normalizing datasets, and performing transformations on the data*

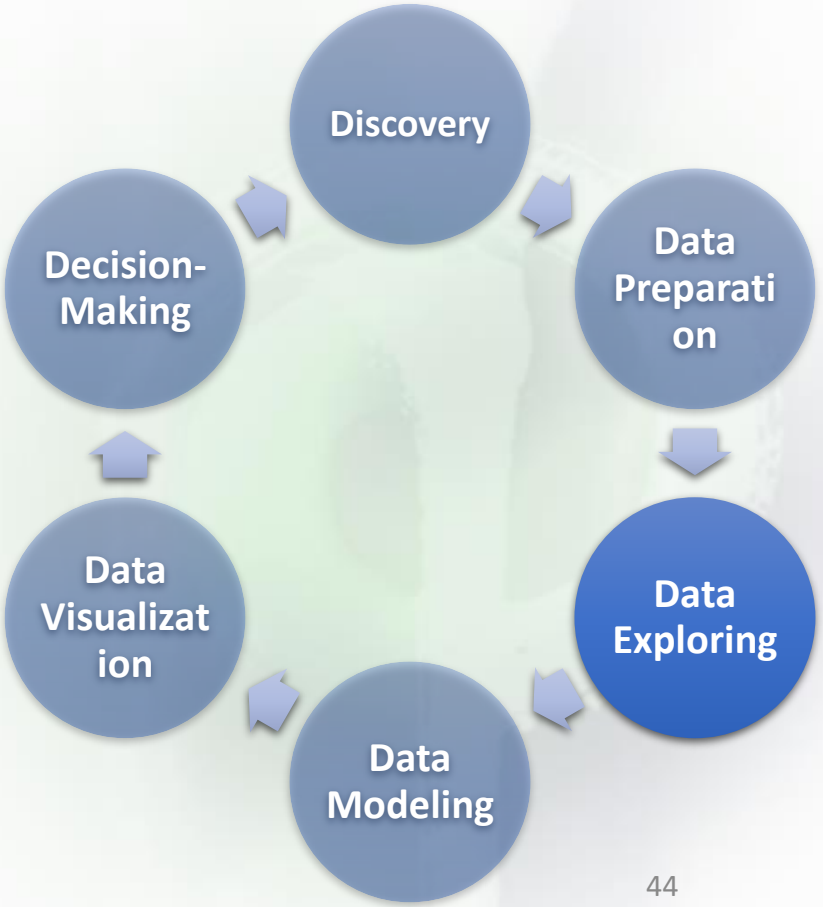


Data Analytics Life Cycle

Data Exploring

This phase aims to identify the *structure of data*, ensure *analytical tools* are available to *achieve its objectives*

- Data exploration and *variable selection*
  - *To understand the relationships among the variables to inform selection of the variables and methods*
- *Model* Selection
  - *Select an proper analytical technique, or a list of candidate techniques, based on the goal of project*
- Common *Tools*
  - *Python/R programming with modeling capabilities*
  - *SQL Analysis services provide with in-databased analytics of data mining functions*
  - *SAS, SPSS, MatLAB, etc. provide with analytics enterprise applications*

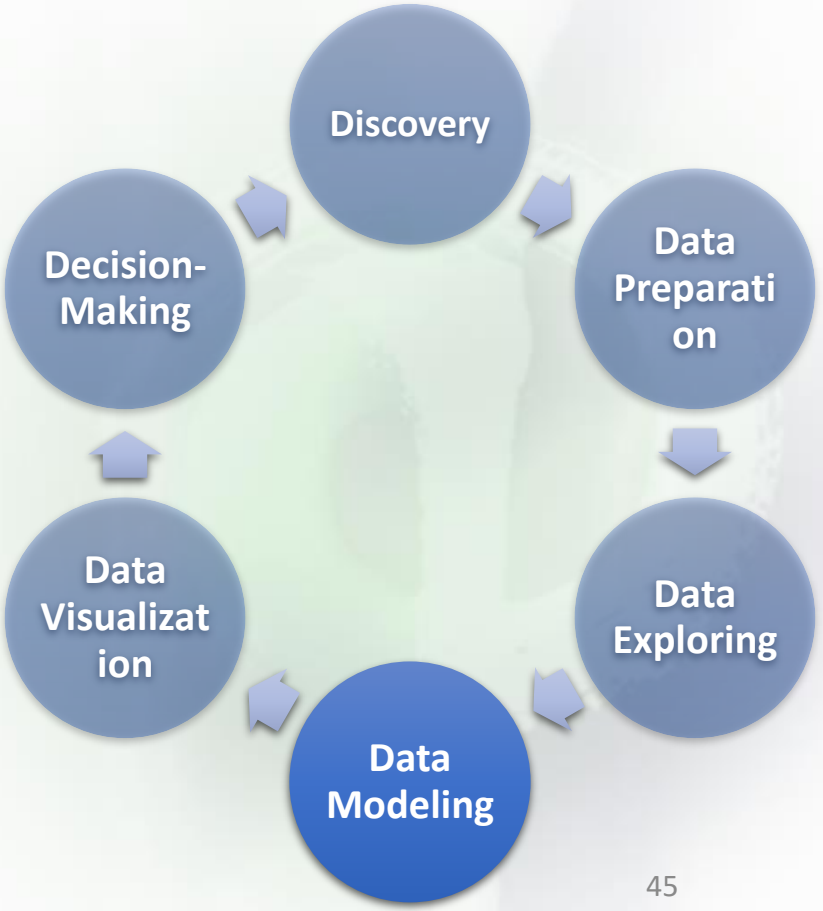


Data Analytics Life Cycle

Data Modeling

This phase aims to *design datasets* for *training*, *testing*, and *production* purposes

- Training data
  - *Enable data scientists to develop the analytical model and construct an initial model*
- Test data
  - *Disjoin from training data, it is used for validating the constructed model*
- Address outlined objectives
  - *Does the model appear valid and accurate on the test data?*
  - *Does the model output/behavior make sense to the domain experts?*
  - *Is the model sufficiently accurate to meet the goal?*
  - *Does the model avoid intolerable mistakes?*

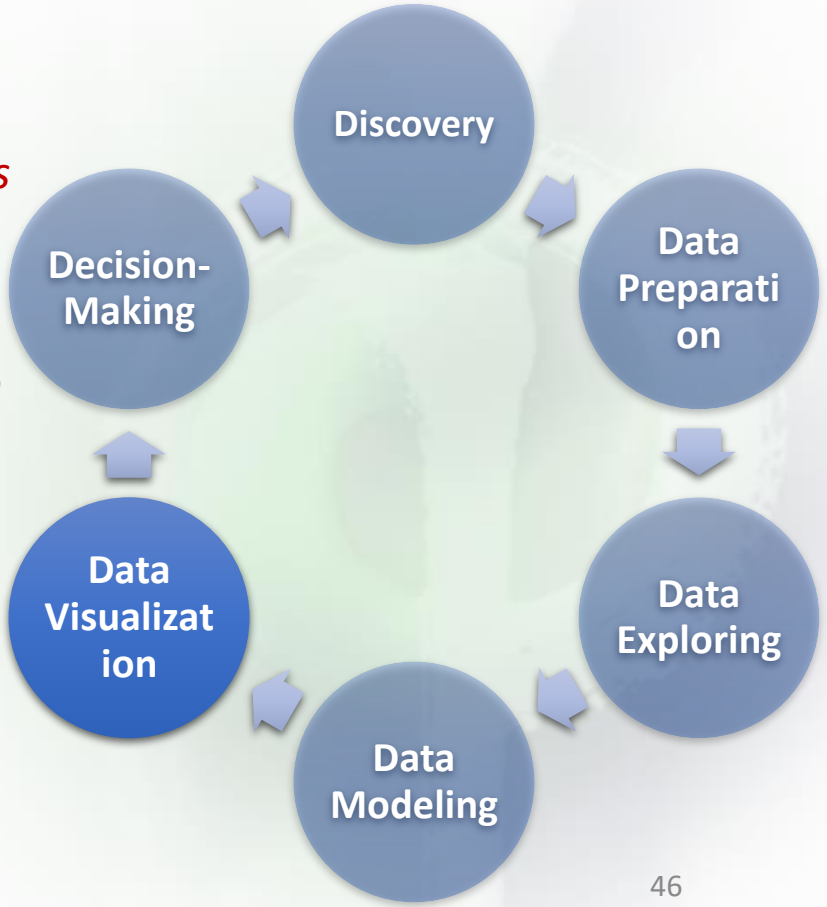


Data Analytics Life Cycle

Visualize & Communicate Results

The team needs to *articulate* the *findings* and *outcomes*, *interprets* and presents it in a *pictorial* or *graphic* format

- Validation
  - *Determine if it succeeded or failed, by performing very robust analysis and determining if the results are statistically significant and valid*
- Analysis
  - *Determine which model or models address the analytical challenge in the most appropriate way*
  - *Reflect on the implications of findings, measure the business value*
- Presentation
  - *Find the best way to prepare the presentation and demonstrate the value of the findings*
  - *Make recommendations for future work or improvements*

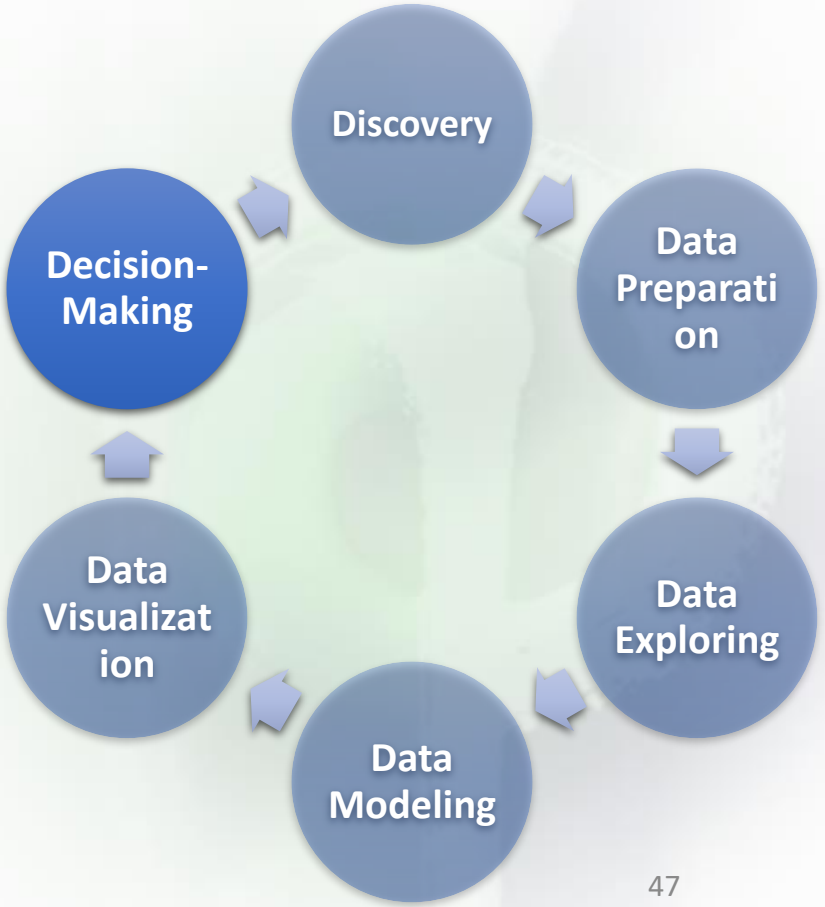


Data Analytics Life Cycle

Decision-Making & Operationalize

Once the benefits of the proposal have been identified, this phase aims to *deploy the work*

- Pilot project
  - *Setup a pilot project in a controlled way before broadening the work to a full enterprise of users*
  - *Risk can be managed more effectively on a small scope, before a wide-scale rollout*
- Model refinement
  - *Test the model in a live setting*
  - *Learn from the deployment, make any necessary adjustments before launching the model across the enterprise*
- Deliverables
  - *A presentation, technical specification documentations, well-annotated production code*



References

- Michael Sandberg (2013), *DataViz History: Charles Minard's Flow Map of Napoleon's Russian Campaign of 1812*.
- Costigan-Eaves, P., & Macdonald-Ross, M (1990), *William Playfair (1759-1823)*. Statistical Science, 318-326.

Acknowledgements

Some of the materials are adapted from:

- Neha Vaidya, 2019
- Teemu Roos, 2018