Introduction to data science

LECTURE – 1 (WEEK 1)

Overview

Course and Instructor’s introduction

Understanding data

Data science intro

- The process

Examples

Types of Data

Welcome to Intro to Data Science – fall 2022

Instructor (s):

Dr. Asma Ahmad [Sec: A, B, C]

- Email: asma.ahmad@nu.edu.pk/asma.ahmad@lhr.nu.edu.pk
- Office: E-157

Office Hours: will be displayed on office door shortly

Course material on Google classroom: xv2e6ny

TA:

- TBD

Reference Books and Resources

Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk From The Frontline. O'Reilly. 2014. ISBN 978-1-449-35865-5.

Jiawei Han, Micheline Kamber and Jian Pei. Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann Publishers. 2012. ISBN 978-0-12-381479-1.

Tom Mitchell: Machine Learning

Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. Mining of Massive Datasets. v2.1, Cambridge University Press. 2014.

Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. MIT Press. 2013. ISBN 0262018020.

Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. O'Reilly 2013. ISBN 978-1-449-36132-7.

Tools and Software Packages

WEKA

KNIME

ORANGE

CBA

MATLAB

PYTHON

R

SPSS

SAS

ArcView GIS

Maptitude for the Web

etc.,

Language: Python
Tools: NumPy, Pandas, matplotlib, scikit-learn
Other packages if needed: SciPy, and SymPy

Tentative Grading Policy

Class	
Assignments	15%
Quizzes	10%
Midterm Exam	30%
Final Exam	45%

Grading Policy

There is simply no chance of extension in any of the deadline, what so ever

You can request for re-checking of any of your evaluation as per following rules;

Exams: Same day

Assignments: 2 days after handing-over

Quizzes: 2 days after handing over

Warning! After due time, request will not be considered even if it’s genuine.

Warning! Regularly check flex and don’t come in the end with bulk of queries in hand.

General Guidelines

Visit Google classroom regularly for updates

No email submissions when Google classroom is there. *Always remember that, you are putting your task in trash by yourself when you are emailing it.*

- Cheating cases are intolerable. You will be given negative marks for cheated stuff irrespective of the fact that, you were provider or the other one. Your cheating in exam will make it easy for you to step down from the Course with an ‘F’ grade. . . ☹
- There will be no re-take of any evaluation if you haven’t informed earlier through a proper channel.

Quiz is inevitable so always, expect a One ☺ [at least, one in every week]

General Guidelines

You have to depend on yourself to have a good grade in the course

- *You need to appear in demo to have it graded otherwise you will be given 5% of the total marks even if it was best assignment/project of the class*
- *Grading will be individual even for group tasks.*
- *There will always be a quiz of written assignments whose performance will be considered as performance in that assignment.*

Tentative Outline

Introduction & Applications

Data (Acquisition, Storage, & processing)

Data Preprocessing and Mining

- Classification
- Clustering
- Association Rule Mining
- Attribute selection (Feature Selection)

Machine Learning

- Classification and Regression
- Gradient Descent
- Regularization
- Support Vector Machines
- Dimensionality Reduction
- Outlier Detection

Statistical Inference

- Statistical Modeling
- Probability distribution
- Fitting a model

Descriptive and Exploratory Analysis

Market Basket Analysis: Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns

Exploratory Data Analysis [EDA]

Data Visualization

Learning Outcomes

- Understand the basics of Data Science,
- Prepare and wrangle the data for analysis
- Perform Exploratory Data Analysis (EDA)
- Understand and apply machine learning algorithms to gain insight from the data

Why do we need Data Science?

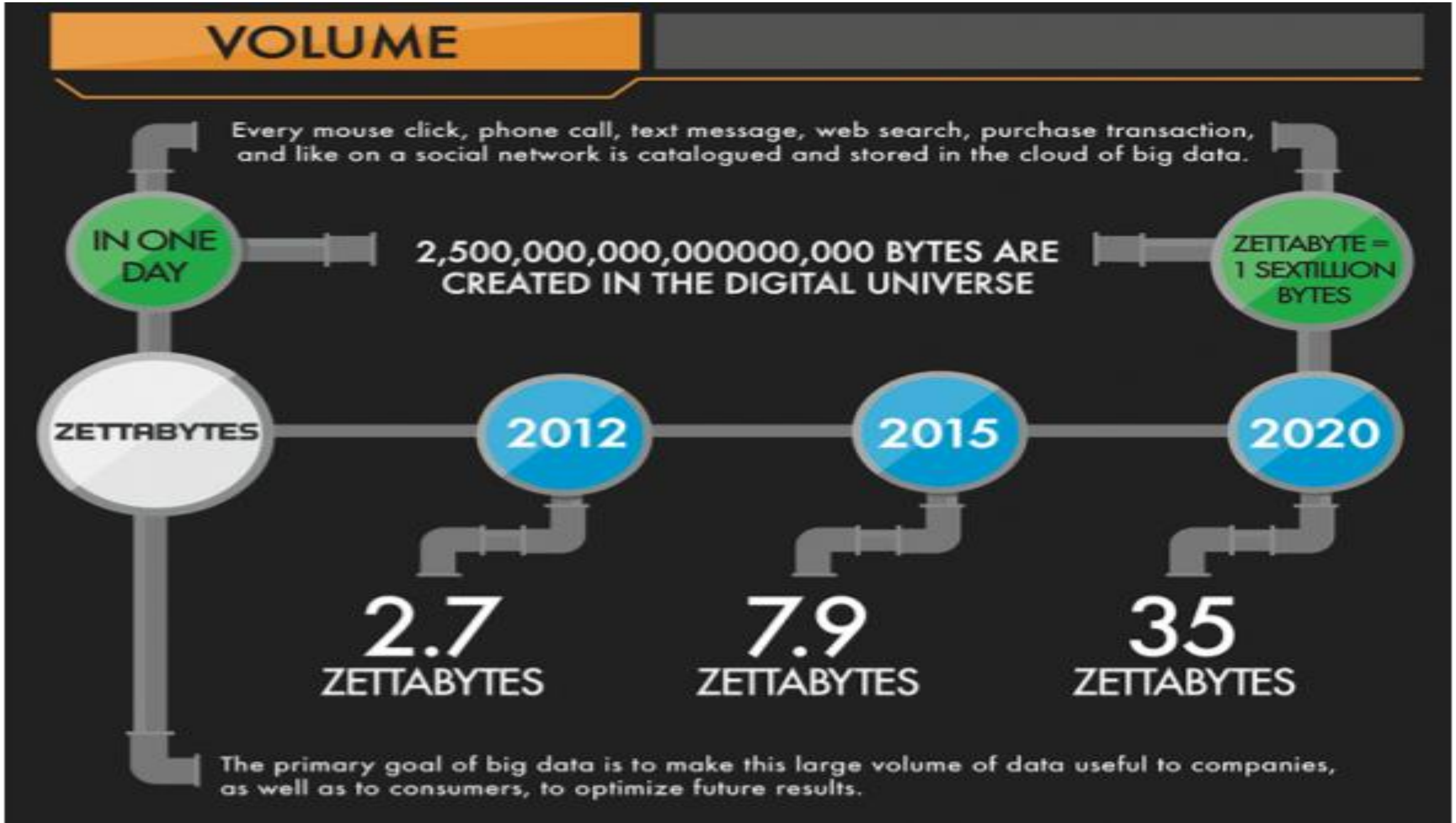
Good decisions require good information derived from raw facts

Big Data

“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.”

- Eric Schmidt, Google

Why do we need Data Science?



VOLUME

Every mouse click, phone call, text message, web search, purchase transaction, and like on a social network is catalogued and stored in the cloud of big data.

IN ONE DAY

2,500,000,000,000000,000 BYTES ARE CREATED IN THE DIGITAL UNIVERSE

ZETTABYTE = 1 SEXTILLION BYTES

ZETTABYTES

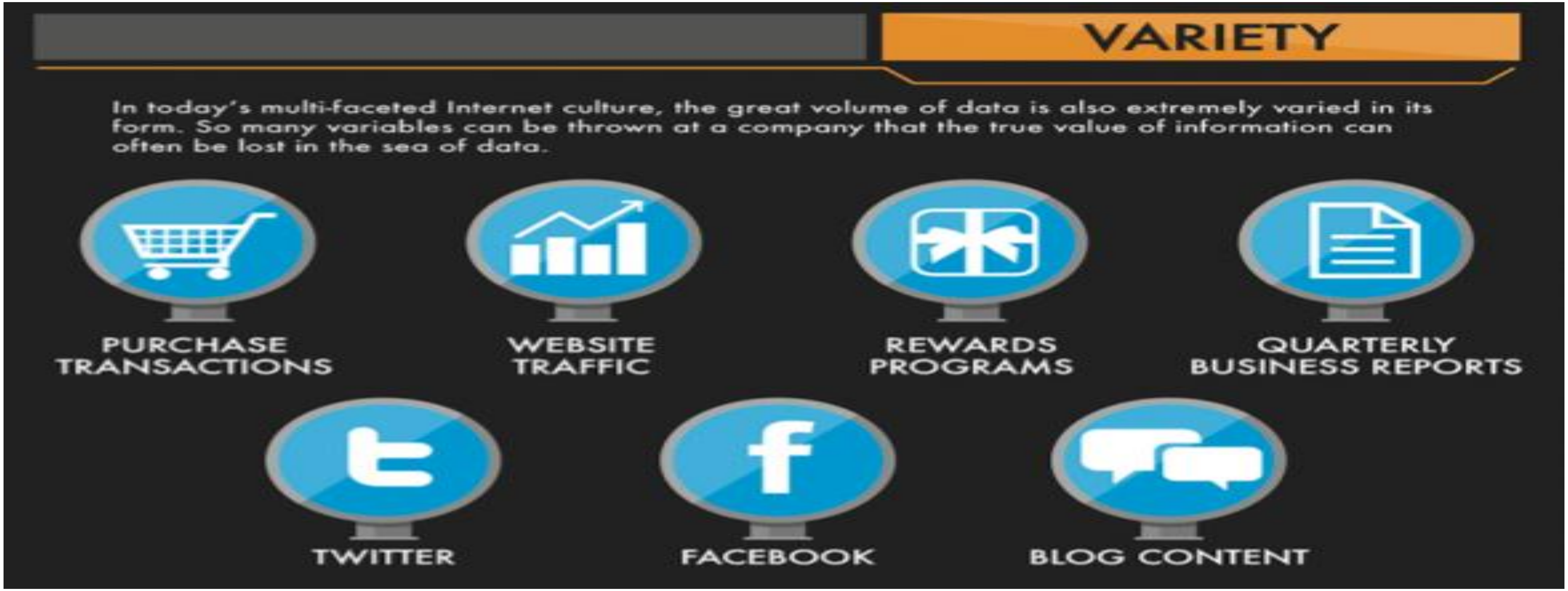
2012 – 2.7 ZETTABYTES

2015 – 7.9 ZETTABYTES

2020 – 35 ZETTABYTES

The primary goal of big data is to make this large volume of data useful to companies, as well as to consumers, to optimize future results.

Why do we need Data Science



VARIETY

In today's multi-faceted Internet culture, the great volume of data is also extremely varied in its form. So many variables can be thrown at a company that the true value of information can often be lost in the sea of data.

Nate Silver

American Data Scientist who analyzes elections and baseball.

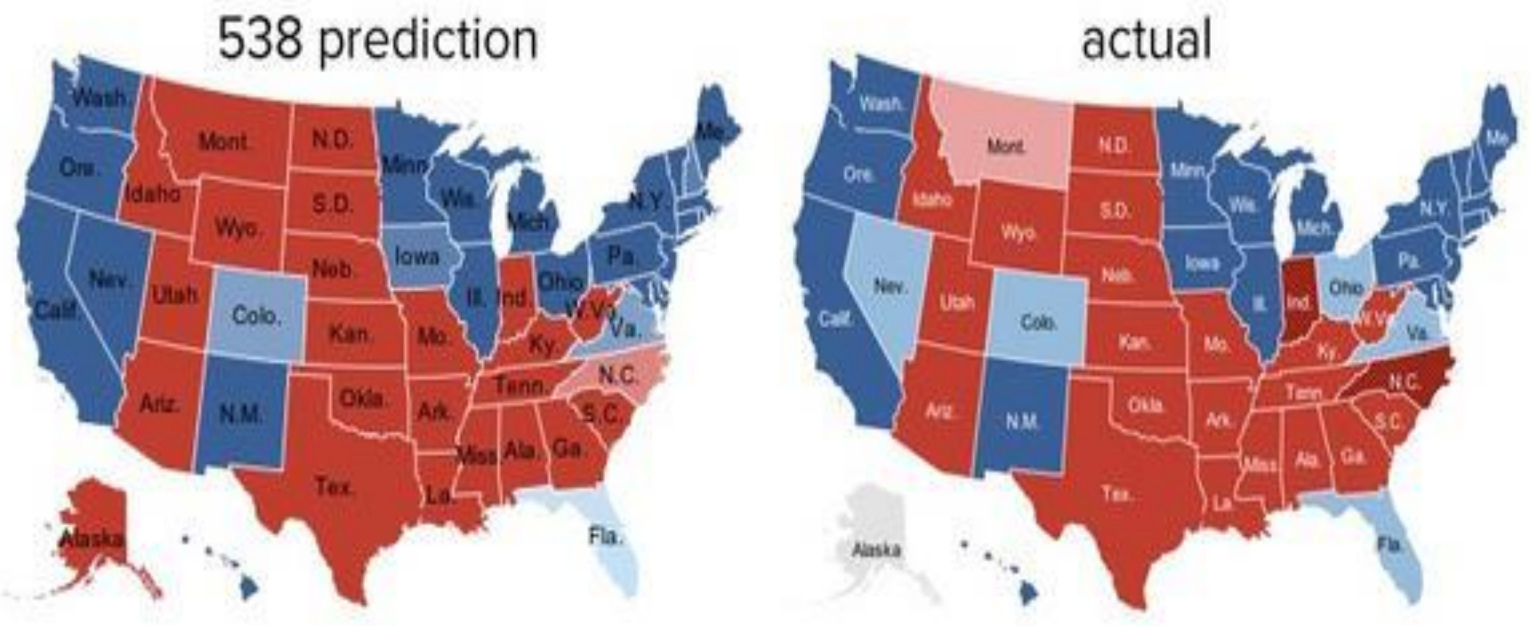
-PECOTA: a system for forecasting the performance and career development of Major -League Baseball players.

- 2012 U.S. Presidential election

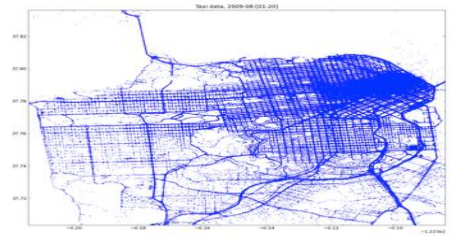
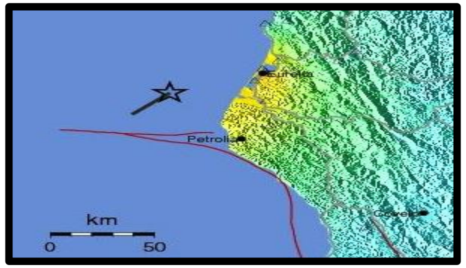
2012 U.S. Presidential election

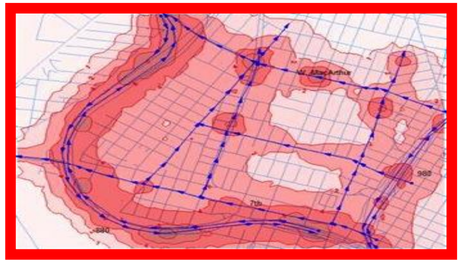
- Correctly predicted the winners of all the states.

“Nate Silver won the election”
– Harvard Business Review



What can you do with the data?



Crowdsourcing + physical modeling + sensing + data assimilation

to produce:

From Alex Bayen, UCB

Data vs. Information

Data

Raw facts

- Have not yet been processed to reveal their meaning to the end user

Building blocks of information

Data management

- Generation, storage, and retrieval of data

Information

Produced by processing raw data to reveal its meaning

Requires context

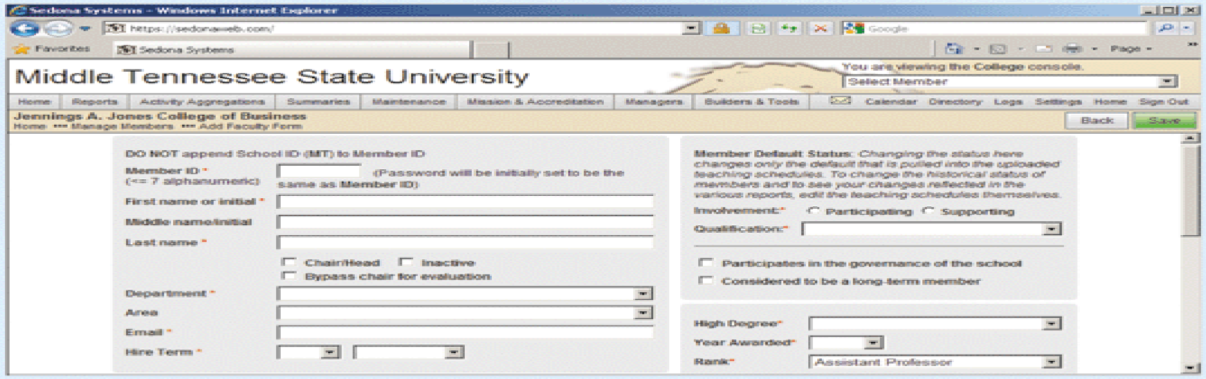
Bedrock of **knowledge**

Should be accurate, relevant, and **timely** to enable good decision making

Data vs. Information (cont'd.)

FIGURE 1.1 **Transforming raw data into information**

a) Data entry screen



Middle Tennessee State University

You are viewing the College console.

Select Member

Jennings A. Jones College of Business

Home ••• Manage Members ••• Add Faculty Form

DO NOT append School ID (MT) to Member ID

Member ID * (<= 7 alphanumeric) (Password will be initially set to be the same as Member ID)

First name or initial *

Middle name/initial

Last name *

□ Chair/Head □ Inactive

□ Bypass chair for evaluation

Department *

Area

Email *

Hire Term *

Member Default Status: *Changing the status here changes only the default that is pulled into the uploaded teaching schedules. To change the historical status of members and to see your changes reflected in the various reports, edit the teaching schedules themselves.*

Involvement:* ○ Participating ○ Supporting

Qualification:*

□ Participates in the governance of the school

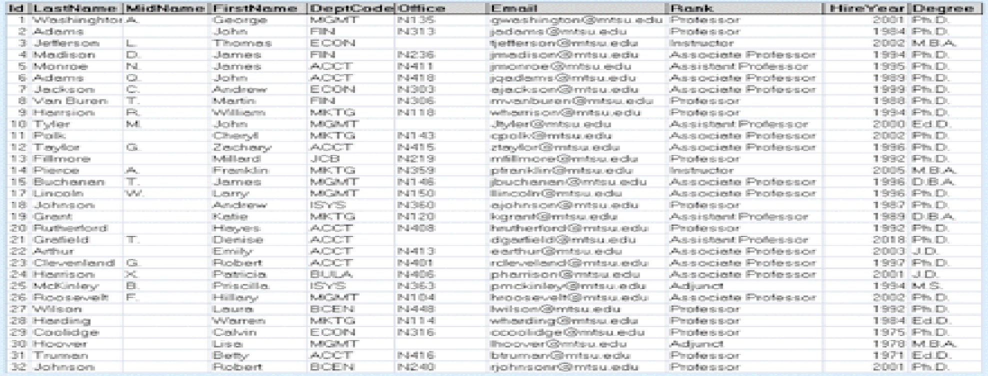
□ Considered to be a long-term member

High Degree*

Year Awarded*

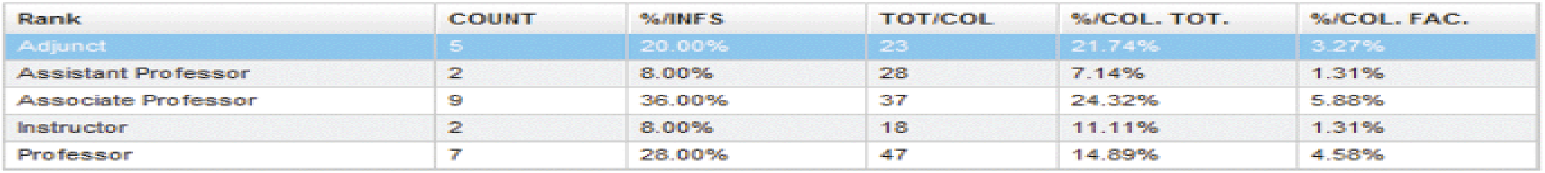
Rank* Assistant Professor

b) Raw data



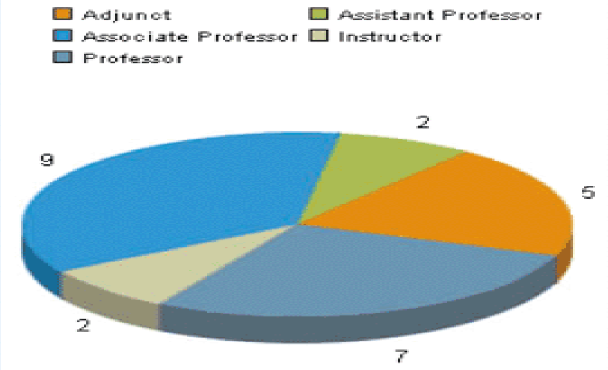
Id	LastName	MidName	FirstName	DeptCode	Office	Email	Rank	HireYear	Degree
1	Washingtor	A.	George	MGMT	N135	gwashington@mtsu.edu	Professor	2001	Ph.D.
2	Adams		John	FIN	N313	jadams@mtsu.edu	Professor	1984	Ph.D.
3	Jefferson	L.	Thomas	ECON		tjefferson@mtsu.edu	Instructor	2002	M.B.A.
4	Madison	D.	James	FIN	N236	jmadison@mtsu.edu	Associate Professor	1994	Ph.D.
5	Monroe	N.	James	ACCT	N411	jmonroe@mtsu.edu	Assistant Professor	1995	Ph.D.
6	Adams	Q.	John	ACCT	N418	jqadams@mtsu.edu	Associate Professor	1989	Ph.D.
7	Jackson	C.	Andrew	ECON	N303	ajackson@mtsu.edu	Associate Professor	1999	Ph.D.
8	Van Buren	T.	Martin	FIN	N306	mvanburen@mtsu.edu	Professor	1988	Ph.D.
9	Harrison	R.	William	MKTG	N118	wharrison@mtsu.edu	Professor	1994	Ph.D.
10	Tyler	M.	John	MGMT		Jtyler@mtsu.edu	Assistant Professor	2000	Ed.D.
11	Polk		Cheryl	MKTG	N143	cpolk@mtsu.edu	Associate Professor	2002	Ph.D.
12	Taylor	G.	Zachary	ACCT	N415	ztaylor@mtsu.edu	Associate Professor	1996	Ph.D.
13	Fillmore		Millard	JCB	N219	mfillmore@mtsu.edu	Professor	1992	Ph.D.
14	Pierce	A.	Franklin	MKTG	N359	pfranklin@mtsu.edu	Instructor	2005	M.B.A.
15	Buchanan	T.	James	MGMT	N146	jbuchanan@mtsu.edu	Associate Professor	1996	D.B.A.
17	Lincoln	W.	Larry	MGMT	N150	llincoln@mtsu.edu	Associate Professor	1996	Ph.D.
18	Johnson		Andrew	ISYS	N360	ajohnson@mtsu.edu	Professor	1987	Ph.D.
19	Grant		Katie	MKTG	N120	kgrant@mtsu.edu	Assistant Professor	1989	D.B.A.
20	Rutherford		Hayes	ACCT	N408	hrutherford@mtsu.edu	Professor	1992	Ph.D.
21	Grafield	T.	Denise	ACCT		dgarfield@mtsu.edu	Assistant Professor	2018	Ph.D.
22	Arthur		Emily	ACCT	N413	earthur@mtsu.edu	Associate Professor	2003	J.D.
23	Cleveland	G.	Robert	ACCT	N401	rcleveland@mtsu.edu	Associate Professor	1997	Ph.D.
24	Harrison	X.	Patricia	BULA	N406	pharrison@mtsu.edu	Associate Professor	2001	J.D.
25	McKinley	B.	Priscilla	ISYS	N363	pmckinley@mtsu.edu	Adjunct	1994	M.S.
26	Roosevelt	F.	Hillary	MGMT	N104	hroosevelt@mtsu.edu	Associate Professor	2002	Ph.D.
27	Wilson		Laura	BCEN	N448	lwilson@mtsu.edu	Professor	1992	Ph.D.
28	Harding		Warren	MKTG	N114	wharding@mtsu.edu	Professor	1984	Ed.D.
29	Coolidge		Calvin	ECON	N316	ccoolidge@mtsu.edu	Professor	1975	Ph.D.
30	Hoover		Lisa	MGMT		lhoover@mtsu.edu	Adjunct	1978	M.B.A.
31	Truman		Betty	ACCT	N416	btruman@mtsu.edu	Professor	1971	Ed.D.
32	Johnson		Robert	BCEN	N240	rjohnson@mtsu.edu	Professor	2001	Ph.D.

c) Information in summary format



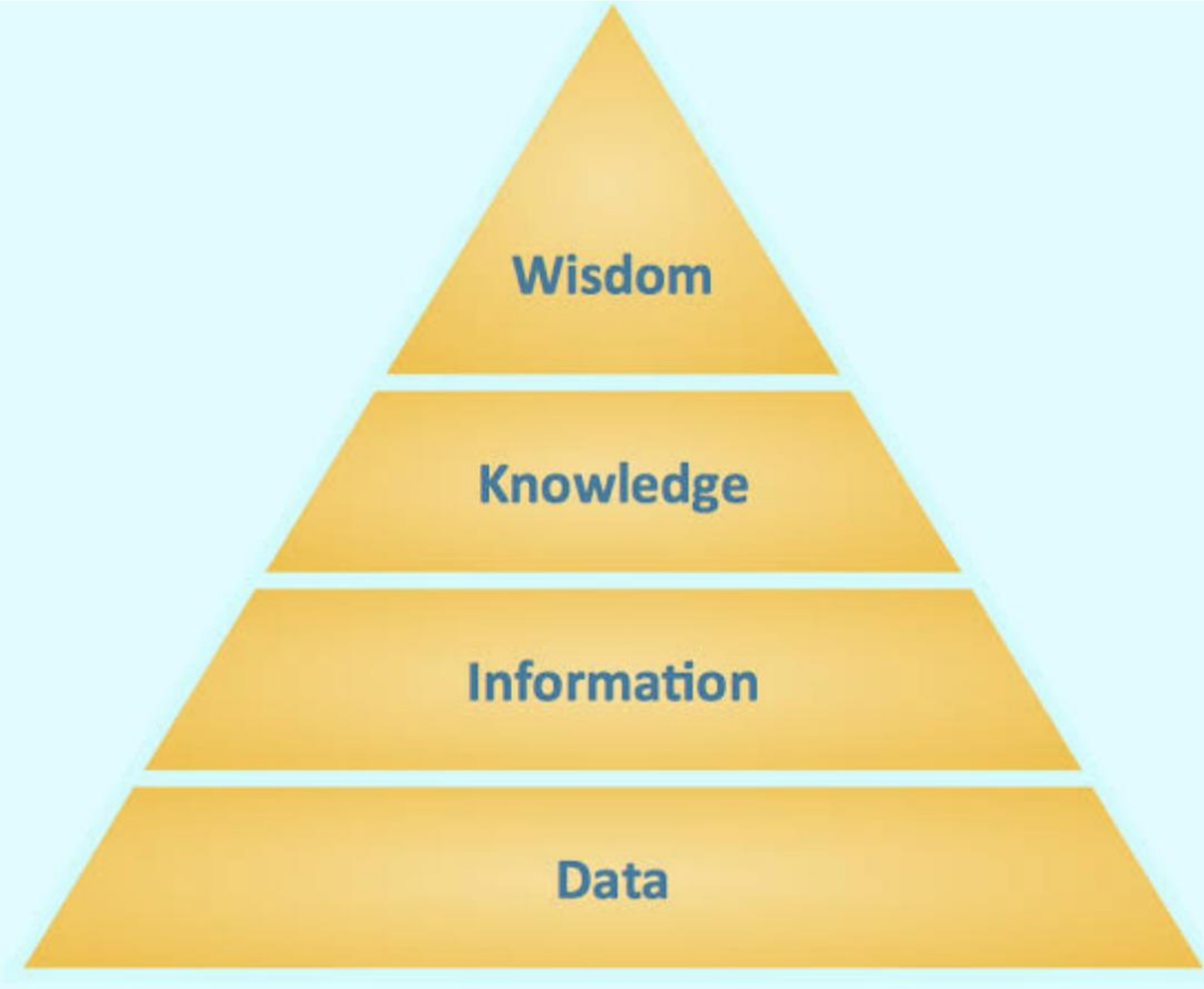
Rank	COUNT	%/INFS	TOT/COL	%/COL. TOT.	%/COL. FAC.
Adjunct	5	20.00%	23	21.74%	3.27%
Assistant Professor	2	8.00%	28	7.14%	1.31%
Associate Professor	9	36.00%	37	24.32%	5.88%
Instructor	2	8.00%	18	11.11%	1.31%
Professor	7	28.00%	47	14.89%	4.58%

d) Information in graphical format



SOURCE: Course Technology/Cengage Learning
Data entry screen courtesy of Sedona Systems, 2011.
Information screens courtesy of JCBDashboard, 2011.

Data, Information, and Beyond



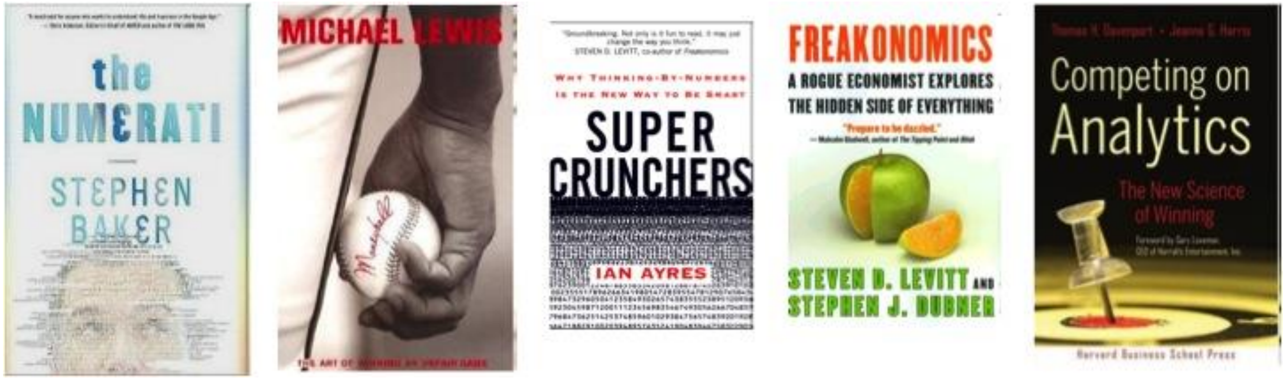
What is Data Science?

Data Science is the application of computational and statistical techniques to address or gain insight into some problem in the real world

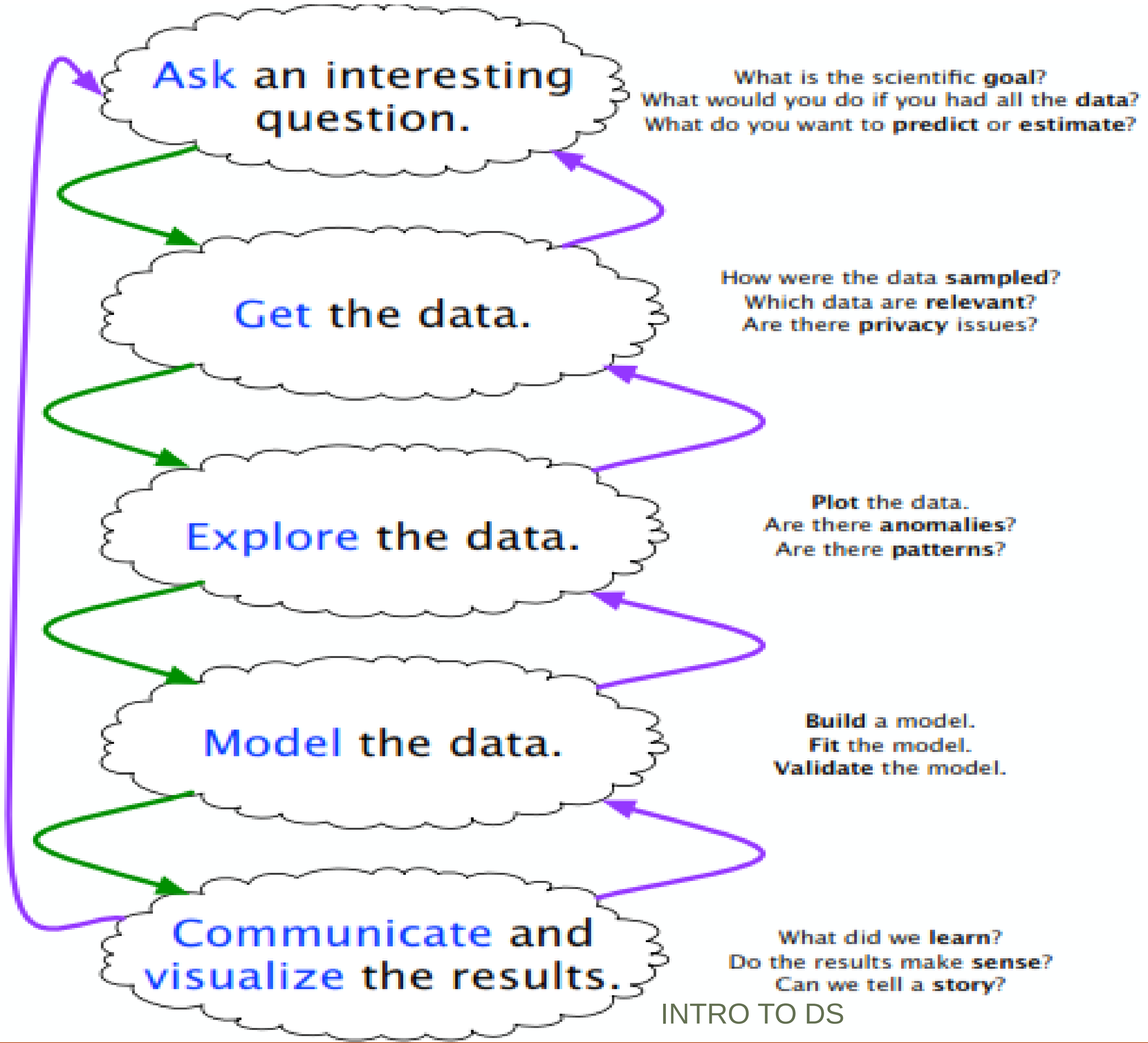
Data Science = Statistics + data processing + machine learning + scientific inquiry + visualization + business analytics + big data + …

Contrast: Databases

Datawarehouse	Data Science
Querying the past	Querying the future


the NUMERATI
STEPHEN BAKER
MICHAEL LEWIS
SUPER CRUNCHERS
IAN AYRES
FREAKONOMICS
A ROGUE ECONOMIST EXPLORES THE HIDDEN SIDE OF EVERYTHING
STEVEN D. LEVITT AND STEPHEN J. DUBNER
Competing on Analytics
The New Science of Winning
Harvard Business School Press

Business intelligence (**BI**) is the transformation of raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities - Wikipedia



Ask an interesting question.

What is the scientific **goal**?
What would you do if you had all the **data**?
What do you want to **predict** or **estimate**?

Get the data.

How were the data **sampled**?
Which data are **relevant**?
Are there **privacy** issues?

Explore the data.

Plot the data.
Are there **anomalies**?
Are there **patterns**?

Model the data.

Build a model.
Fit the model.
Validate the model.

Communicate and visualize the results.

What did we **learn**?
Do the results make **sense**?
Can we tell a **story**?

Big Data – Five V of Data

Volume:

- How much data is really relevant to the problem solution? Cost of processing?
- *So, can you really afford to store and process all that data?*

Velocity:

- Much data coming in at high speed
- Need for streaming versus block approach to data analysis
- *So, how to analyze data in-flight and combine with data at-rest*

Variety:

- A small fraction is structured formats, Relational, XML, etc.
- A fair amount is semi-structured, as web logs, etc.
- The rest of the data is unstructured text, photographs, etc.
- *So, no single data model can currently handle the diversity*

Big Data – Five V of Data (Cont.)

Veracity: cover term for …

- Accuracy, Precision, Reliability, Integrity
- *So, what is it that you don't know you don't know about the data?*

Value:

- How much value is created for each unit of data (whatever it is)?
- *So, what is the contribution of subsets of the data to the problem solution?*

Types of Data Analytics

Descriptive: A set of techniques for reviewing and examining the data set(s) to understand the data and analyze business performance.

Diagnostic: A set of techniques to determine what has happened and why

Predictive: A set of techniques that analyse current and historical data to determine what is most likely to (not) happen

Types of Data Analytics (Cont.)

Prescriptive: A set of techniques for computationally developing and analyzing alternatives that can become courses of action – either tactical or strategic – that may discover the unexpected

Decisive: A set of techniques for visualizing information and recommending courses of action to facilitate human decision-making when presented with a set of alternatives.

Data Science Life Cycle

Problem Understanding: It all starts with understanding the problem at hand, the questions, and the answers we are trying to find from the dataset at hand.

Data Acquisition: Data Acquisition, as the name suggests, is about retrieving the data with the help of Data Engineers where required. It also consolidates all of the data required to answer the question or to solve the problem at hand.

Data Wrangling: Data wrangling is about using knowledge to preprocess data. It involves looking for missing values and asking business questions like why they are missing. Furthermore, it uses knowledge to give shape to the dataset appropriate for visualizations and to support the coming steps in the life cycle.

Data Exploration: Data Exploration is about visualization and other statistics’ measures to see whether the questions we asked, in the beginning, are being answered or not? The data analyst’s job ends here.

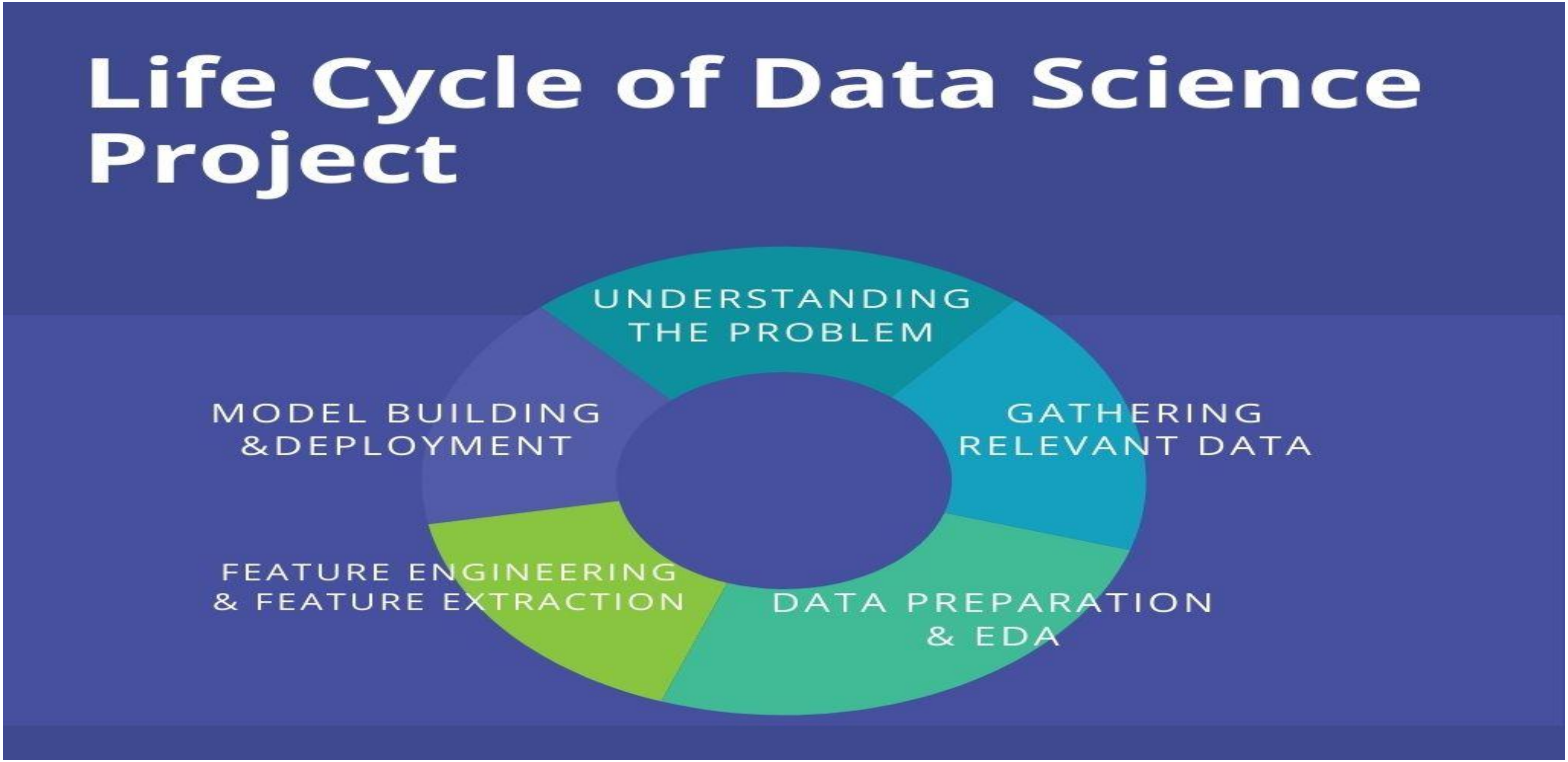
Data Science Life Cycle

Feature Engineering and Selection: It is a preprocessing step before modeling in both Machine Learning and Deep Learning. We will look into these fields in the coming sections. It has similar steps to Data Wrangling apart from some algorithms for Feature Selection and transformation.

Modeling: Modeling is the process that uncovers the meaning of the data. It is about capturing underlying trends and the data’s behavior to make the model, which can be used for predictive analytics as described in the previous section.

Deployment: After we build the model we’ll deploy it in the most efficient and optimized manner so that real-world people can use it. It can be deployed on mobile applications and web applications.

Monitoring: After we have deployed the model, we will want to monitor it. Monitoring is about familiarizing the model with the new dataset and tracking the number of requests that the model receives. It also involves making changes to the analysis and starting over if required.



Life Cycle of Data Science Project

UNDERSTANDING THE PROBLEM

GATHERING RELEVANT DATA

DATA PREPARATION & EDA

FEATURE ENGINEERING & FEATURE EXTRACTION

MODEL BUILDING &DEPLOYMENT

Asking Good Questions

Software developers are not encouraged to ask questions, but data scientists are:

- What exciting things might you be able to learn from a given data set?
- What things do you/your people really want to know?
- What data sets might get you there?

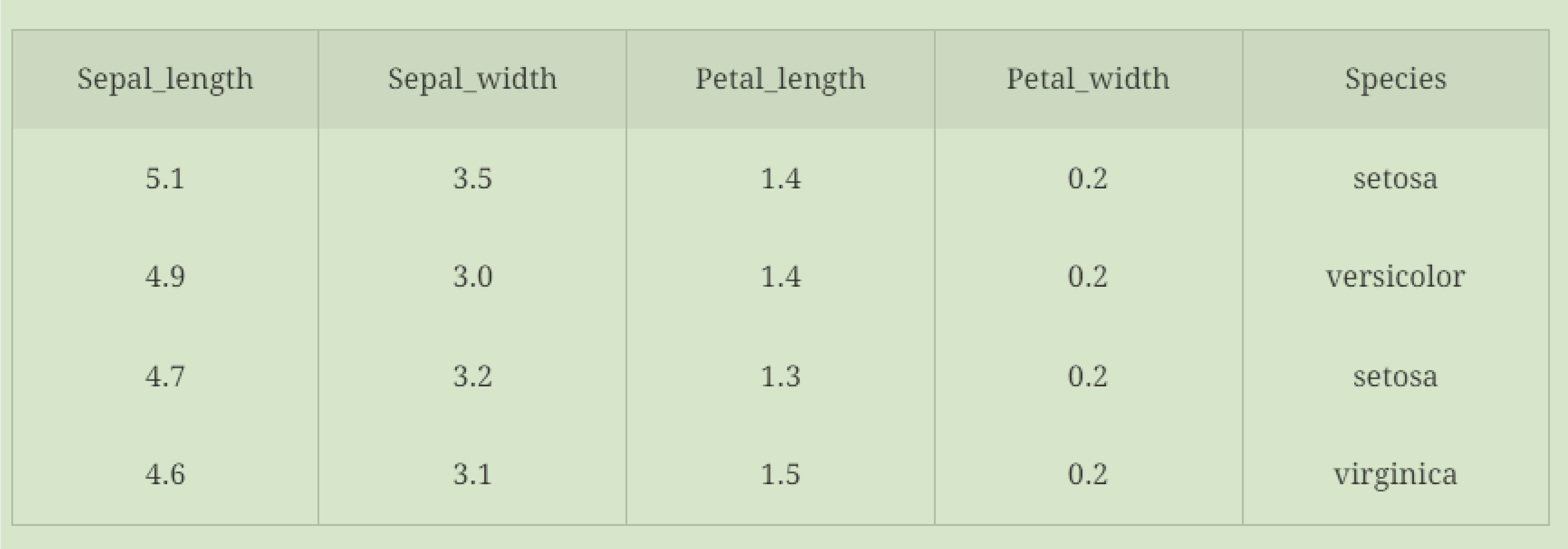
e.g., Baseball

- How to best measure individual player’s skill, value or performance?
- How fair do trades between teams work out?
- What is the trajectory of player’s performances as they mature and age?
- To what extent does batting performance correlate with the position played?

Structured vs. Semi-Structured vs. Unstructured Data

Structured Data

It comes with a predefined format and structure. Structured Data is usually stored in Relational Databases. It is easy to deal with in the Data Science domain.

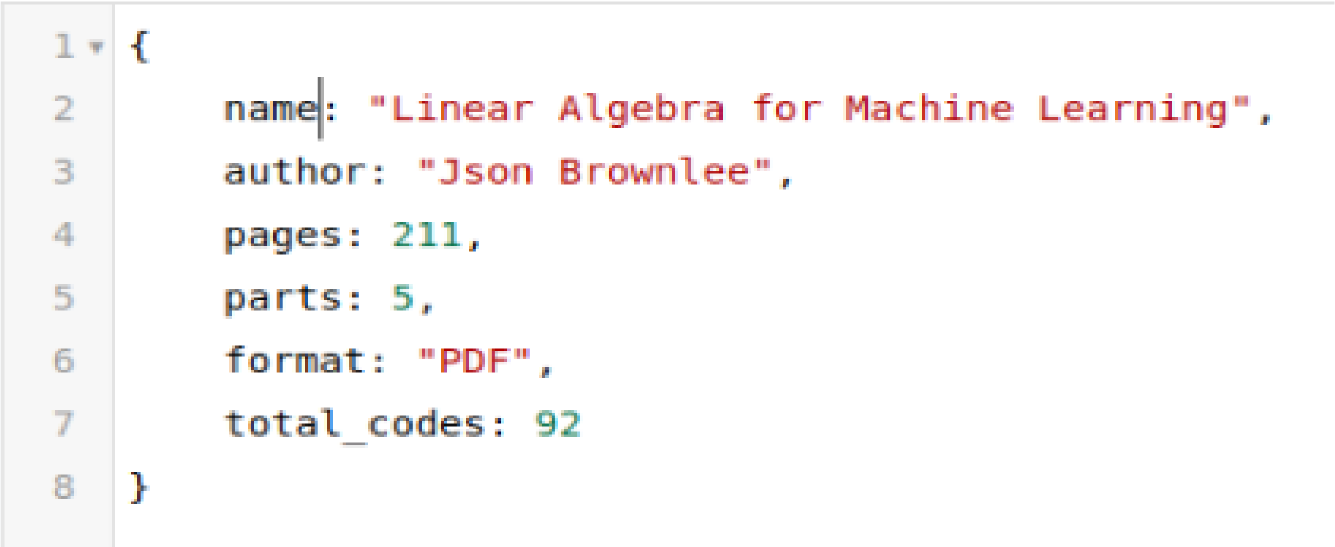


Sepal_length	Sepal_width	Petal_length	Petal_width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	versicolor
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	virginica

Semi-Structured Data

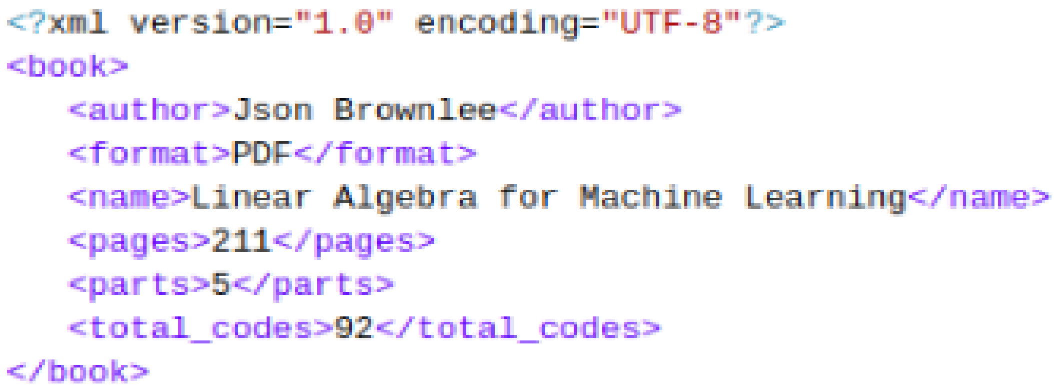
It comes with a predefined format and structure but is not stored in the Relational Database.

- JSON (Javascript Object Notation)



```
{
    name: "Linear Algebra for Machine Learning",
    author: "Json Brownlee",
    pages: 211,
    parts: 5,
    format: "PDF",
    total_codes: 92
}
```

- XML (Extensible Markup Language)



```
<?xml version="1.0" encoding="UTF-8"?>
<book>
   <author>Json Brownlee</author>
   <format>PDF</format>
   <name>Linear Algebra for Machine Learning</name>
   <pages>211</pages>
   <parts>5</parts>
   <total_codes>92</total_codes>
</book>
```

Unstructured Data

It does not have a specific format and lacks structure. It is the type of data that presents many challenges to handle in the Data Science domain

Examples:

- Images
- Videos
- Speech

THAT’S IT FOR TODAY!