

# Data Science

Tushar B. Kute,  
<http://tusharkute.com>

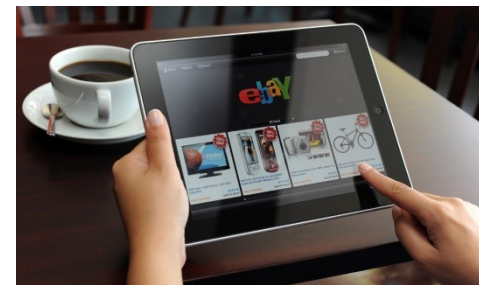


# Objectives

- Defining data science and big data
- Recognizing the different types of data
- Gaining insight into the data science process

# Data All Around

- Lots of data is being collected and war
  - Web data, e-commerce
  - Financial transactions, bank/credit transactions
  - Online trading and purchasing
  - Social Network
  - Cloud



# Data and Big Data

- “90% of the world’s data was generated in the last few years.”
- Due to the advent of new technologies, devices, and communication means like social networking sites, the amount of data produced by mankind is growing rapidly every year.
- The amount of data produced by us from the beginning of time till 2003 was 5 billion gigabytes. If you pile up the data in the form of disks it may fill an entire football field.
- The same amount was created in every two days in 2011, and in every six minutes in 2016. This rate is still growing enormously.

# Big Data Definition

- No single standard definition...

*“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it...*

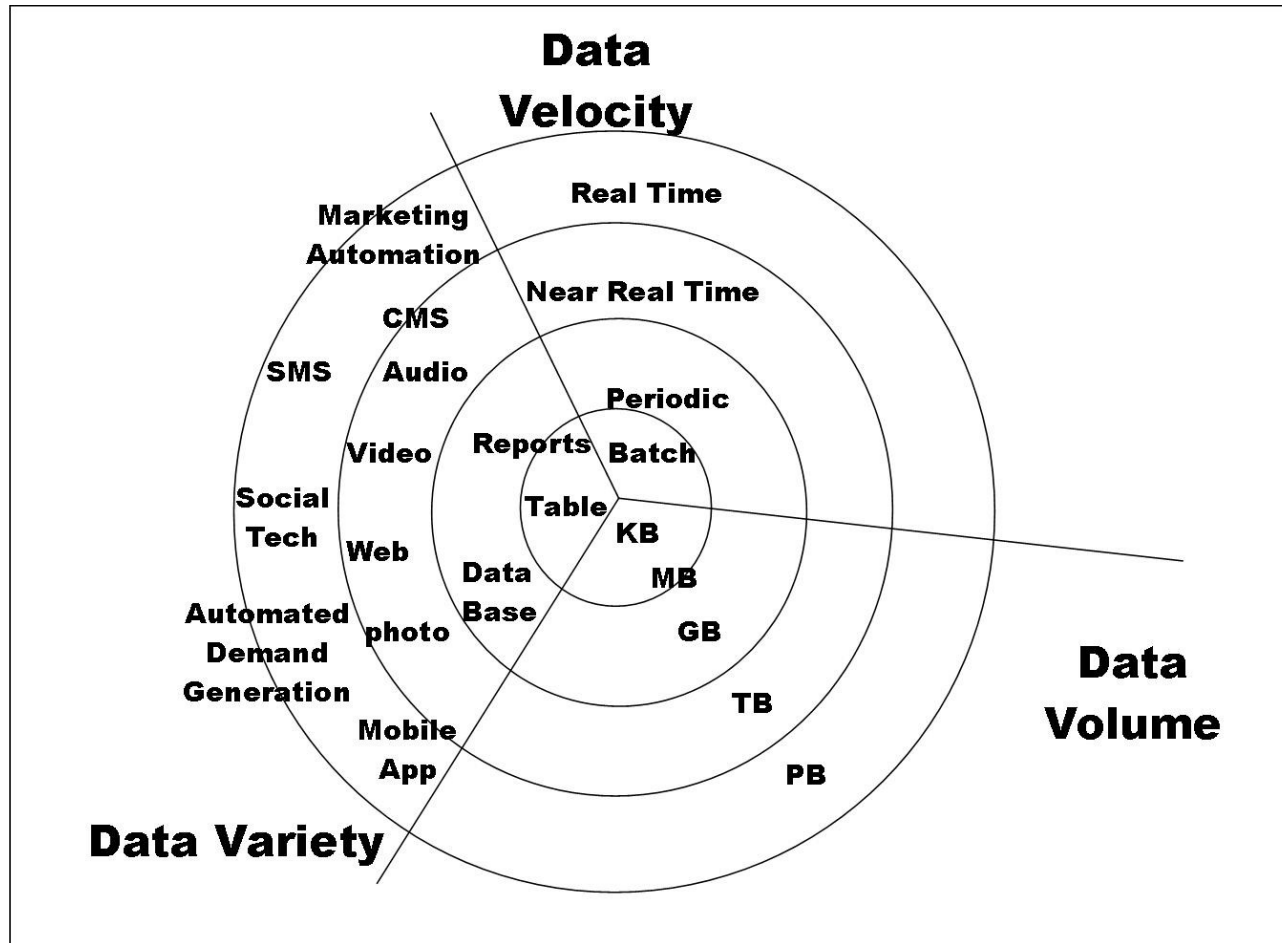
# What is Big Data

- Big Data is a collection of large datasets that cannot be processed using traditional computing techniques.
- It is not a single technique or a tool, rather it involves many areas of business and technology.

# Big Data

- Big Data is any data that is expensive to manage and hard to extract value from
  - Volume
    - The size of the data
  - Velocity
    - The latency of data processing relative to the growing demand for interactivity
  - Variety and Complexity
    - The diversity of sources, formats, quality, structures.

# Big Data

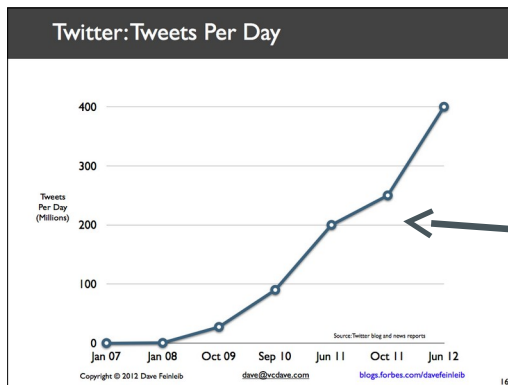
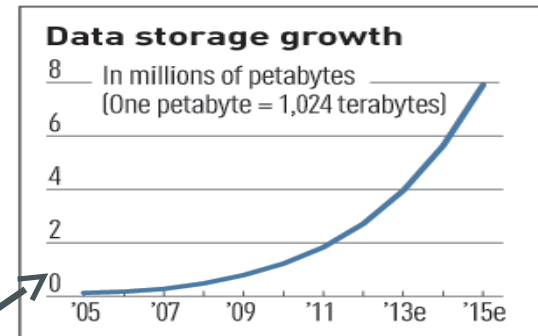
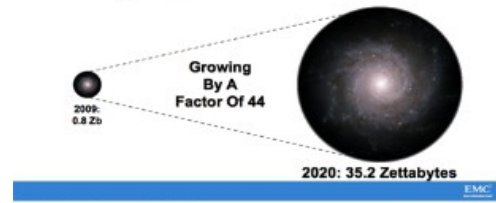




# Characteristics of Big Data: Volume

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially

The Digital Universe 2009-2020



*Exponential increase in collected/generated data*

# Computer Memory Units

## ***UNITS OF COMPUTER MEMORY***

<b>1 Bit</b>	<b>Binary Digit</b>
<b>1 Nibble</b>	<b>4 Bits</b>
<b>8 Bits</b>	<b>1 Byte</b>
<b>1024 Bytes</b>	<b>1 KB (Kilo Byte)</b>
<b>1024 Kilo Bytes</b>	<b>1 MB (Mega Byte)</b>
<b>1024 Mega Bytes</b>	<b>1 GB (Giga Byte)</b>
<b>1024 Giga Bytes</b>	<b>1 TB (Tera Byte)</b>
<b>1024 Tera Bytes</b>	<b>1 PB (Peta Byte)</b>
<b>1024 Peta Bytes</b>	<b>1 EB (Exa Byte)</b>
<b>1024 Exa Bytes</b>	<b>1 ZB (Zetta Byte)</b>
<b>1024 Zetta Bytes</b>	<b>1 YB (Yotta Byte)</b>
<b>1024 Yotta Bytes</b>	<b>1 BB (Bronto Byte)</b>
<b>1024 Bronto Bytes</b>	<b>1 GB* (Geop Byte)</b>
<b>1024 Geop Bytes</b>	<b>1 SB (Sagan Byte)</b>
<b>1024 Sagan Bytes</b>	<b>1 PB (Piya Byte)</b>
<b>1024 Piya Bytes</b>	<b>1 AB (Alpha Byte)</b>
<b>1024 Alpha Bytes</b>	<b>1 KB* (Kryat Byte)</b>
<b>1024 Kryat Bytes</b>	<b>1 AB* (Amos Byte)</b>
<b>1024 Amos Bytes</b>	<b>1 PB* (Pectrol Byte)</b>
<b>1024 Pectrol Bytes</b>	<b>1 BB* (Bolger Byte)</b>
<b>1024 Bolger Bytes</b>	<b>1 SB* (Sambo Byte)</b>
<b>1024 Sambo Bytes</b>	<b>1 QB (Quesa Byte)</b>
<b>1024 Quesa Bytes</b>	<b>1 KB** (Kinsa Byte)</b>
<b>1024 Kinsa Bytes</b>	<b>1 RB (Ruther Byte)</b>
<b>1024 Ruther Bytes</b>	<b>1 BB** (Bubni Byte)</b>
<b>1024 Bubni Bytes</b>	<b>1 SB** (Seaborg Byte)</b>
<b>1024 Seaborg Bytes</b>	<b>1 BB*** (Bohr Byte)</b>
<b>1024 Bohr Bytes</b>	<b>1 HB (Hassiu Byte)</b>
<b>1024 Hassiu Bytes</b>	<b>1 MB* (Meitner Byte)</b>
<b>1024 Meitner Bytes</b>	<b>1 DB (Darmstad Byte)</b>
<b>1024 Darmstad Bytes</b>	<b>1 RB* (Roent Byte)</b>
<b>1024 Roent Bytes</b>	<b>1 CB (Coper Byte)</b>

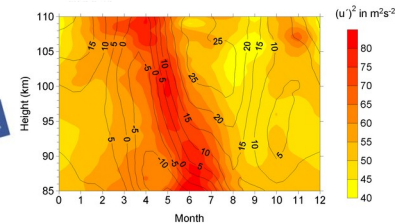
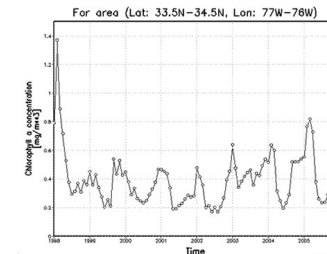
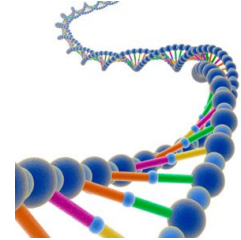
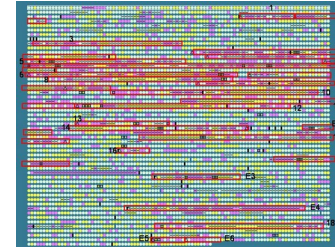
*www.raj-gigaworld.blogspot.com*

R.D

*www.facebook.com/raj-dev/36generalknowledge*

# Characteristics of Big Data: Variety

- Various formats, types, and structures
- Text, numerical, images, audio, video, sequences, time series, social media data, multi-dim arrays, etc...
- Static data vs. streaming data
- A single application can be generating/collecting many types of data



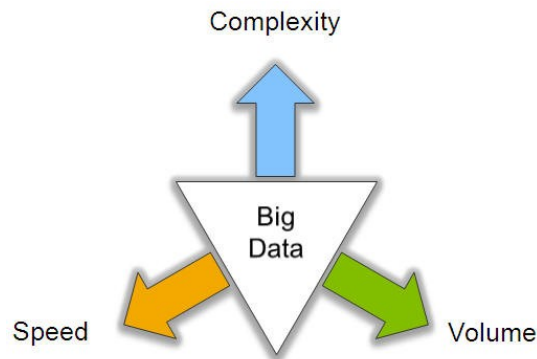
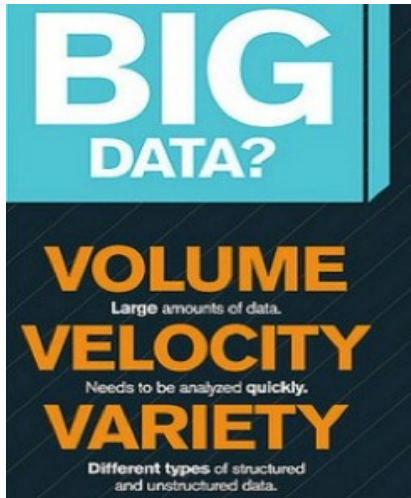
To extract knowledge all these types of data need to be linked together

# Characteristics of Big Data: Velocity

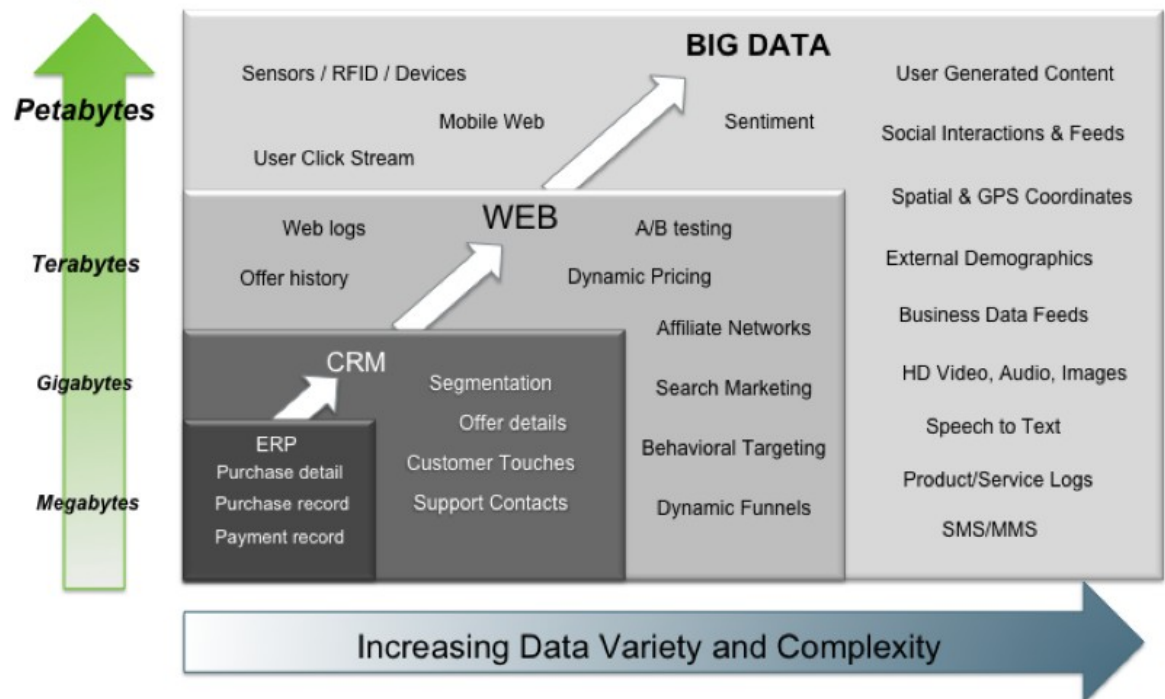
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions, missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like send promotions right now for store next to you.
  - **Healthcare monitoring:** sensors monitoring your activities and body any abnormal measurements require immediate reaction.



# Big Data: 3 Vs

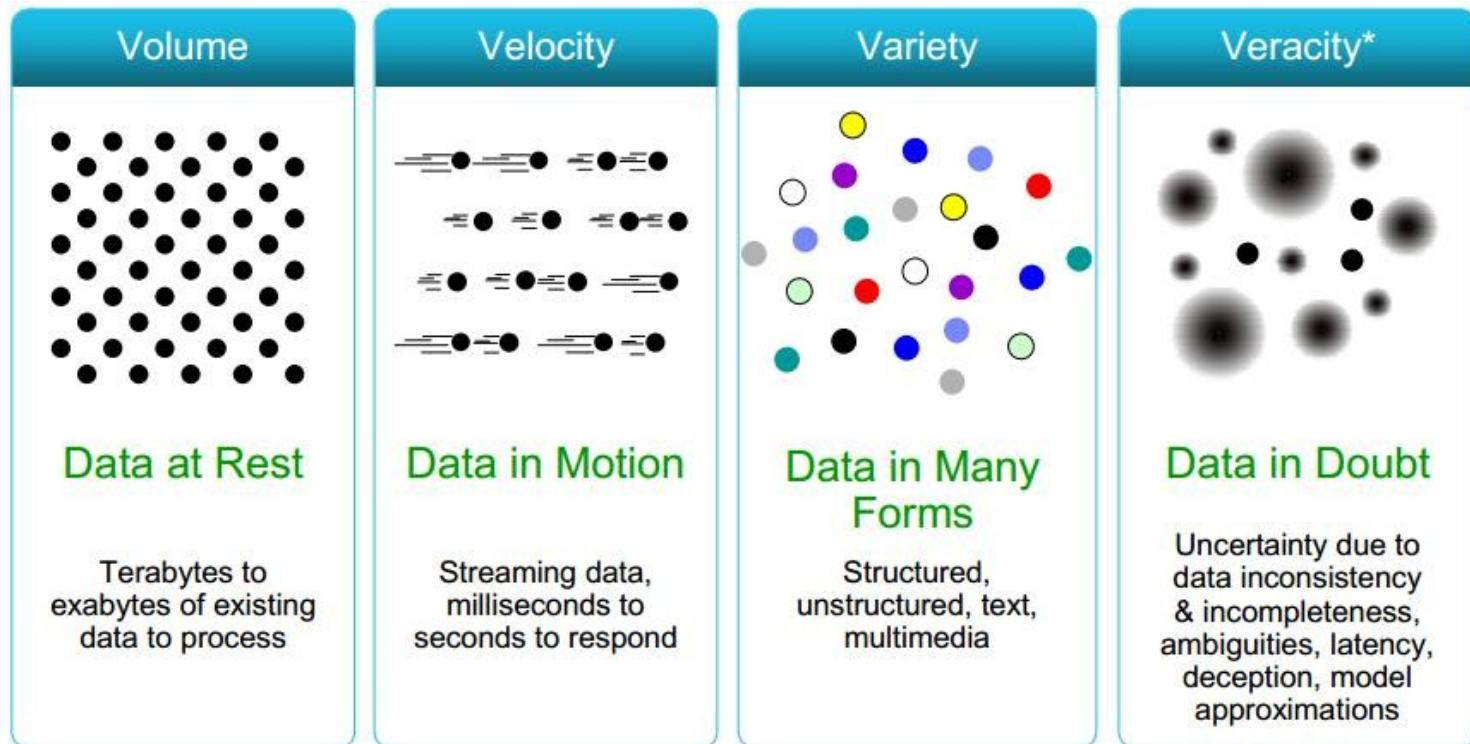


Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

# Big Data: The 4<sup>th</sup> V





# What Comes Under Big Data?

- **Black Box Data:** It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data:** Social media such as Facebook and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data:** The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Power Grid Data:** The power grid data holds information consumed by a particular node with respect to a base station.

# What Comes Under Big Data?

- Transport Data: Transport data includes model, capacity, distance and availability of a vehicle.
- Search Engine Data: Search engines retrieve lots of data from different databases.
- Structured data: Relational data.
- Semi Structured data: XML data.
- Unstructured data: Word, PDF, Text, Media Logs.



# Benefits of Big Data

- Using the information kept in the social network like Facebook, the marketing agencies are learning about the response for their campaigns, promotions, and other advertising mediums.
- Using the information in the social media like preferences and product perception of their consumers, product companies and retail organizations are planning their production.
- Using the data regarding the previous medical history of patients, hospitals are providing better and quick service.

# Big Data Technologies

- Operational Big data
- Analytical Big data

# Operational Big Data

- These include systems like MongoDB that provide operational capabilities for real-time, interactive workloads where data is primarily captured and stored.
- NoSQL Big Data systems are designed to take advantage of new cloud computing architectures that have emerged over the past decade to allow massive computations to be run inexpensively and efficiently.

# Analytical Big Data

- These includes systems like Massively Parallel Processing (MPP) database systems and MapReduce that provide analytical capabilities for retrospective and complex analysis that may touch most or all of the data.
- MapReduce provides a new method of analyzing data that is complementary to the capabilities provided by SQL, and a system based on MapReduce that can be scaled up from single servers to thousands of high and low end machines.

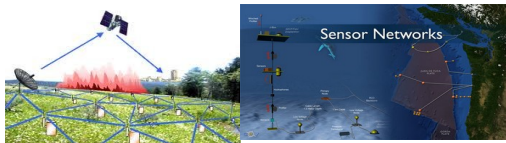
# Who generates Big Data?



**Social media and networks**  
(all of us are generating data)



**Scientific instruments**  
(collecting all sorts of data)



**Sensor technology and networks**  
(measuring all kinds of data)



**Mobile devices**  
(tracking all objects all the time)

# Big Data generation models

- The Model of Generating/Consuming Data has Changed**

**Old Model:** Few companies are generating data, all others are consuming data



**New Model:** all of us are generating data, and all of us are consuming data



# Challenges in Big Data

- The major challenges associated with big data are as follows:
  - Capturing data
  - Curation
  - Storage
  - Searching
  - Sharing
  - Transfer
  - Analysis
  - Presentation

# Types of Data

- Relational Data  
(Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), ...
- Streaming Data



# What to do with this 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

# 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.

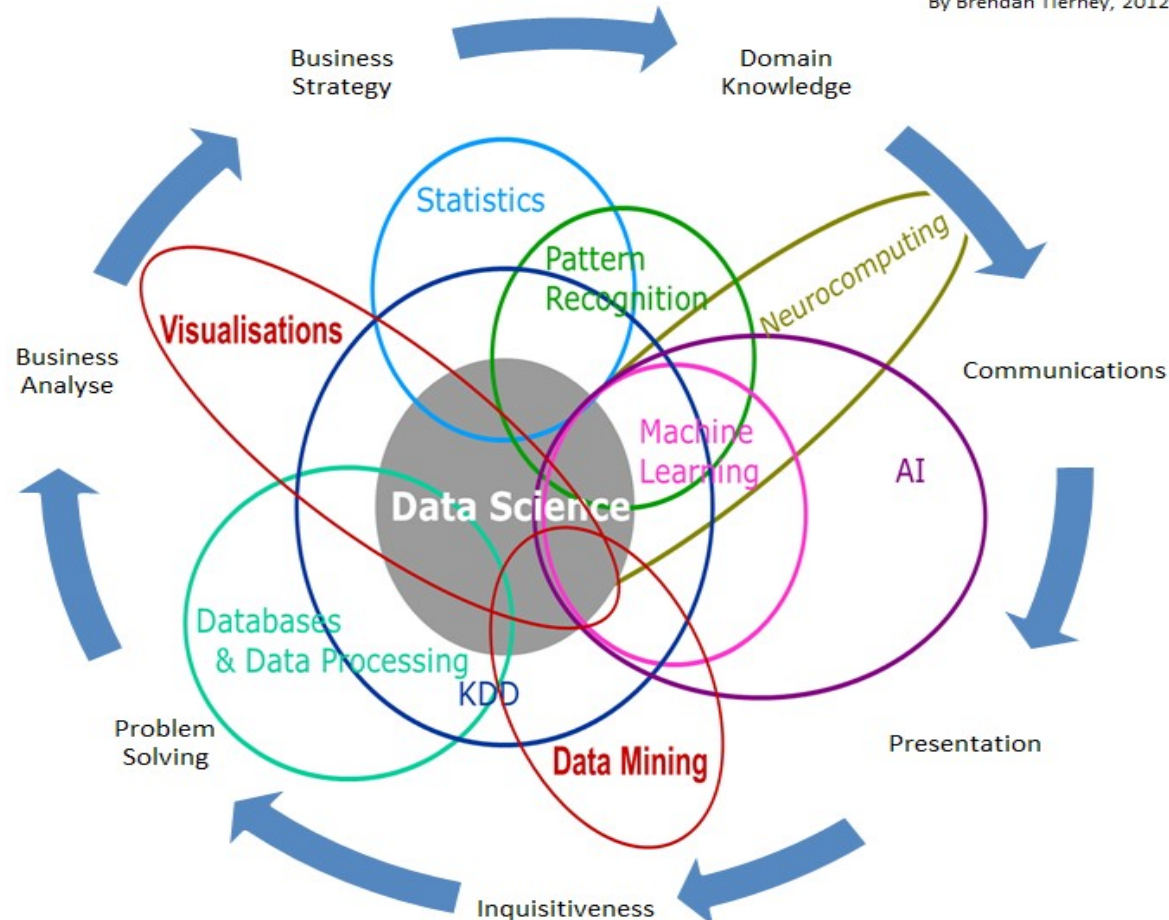
# What is Data Science ?

- 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 Disciplines

## Data Science Is Multidisciplinary

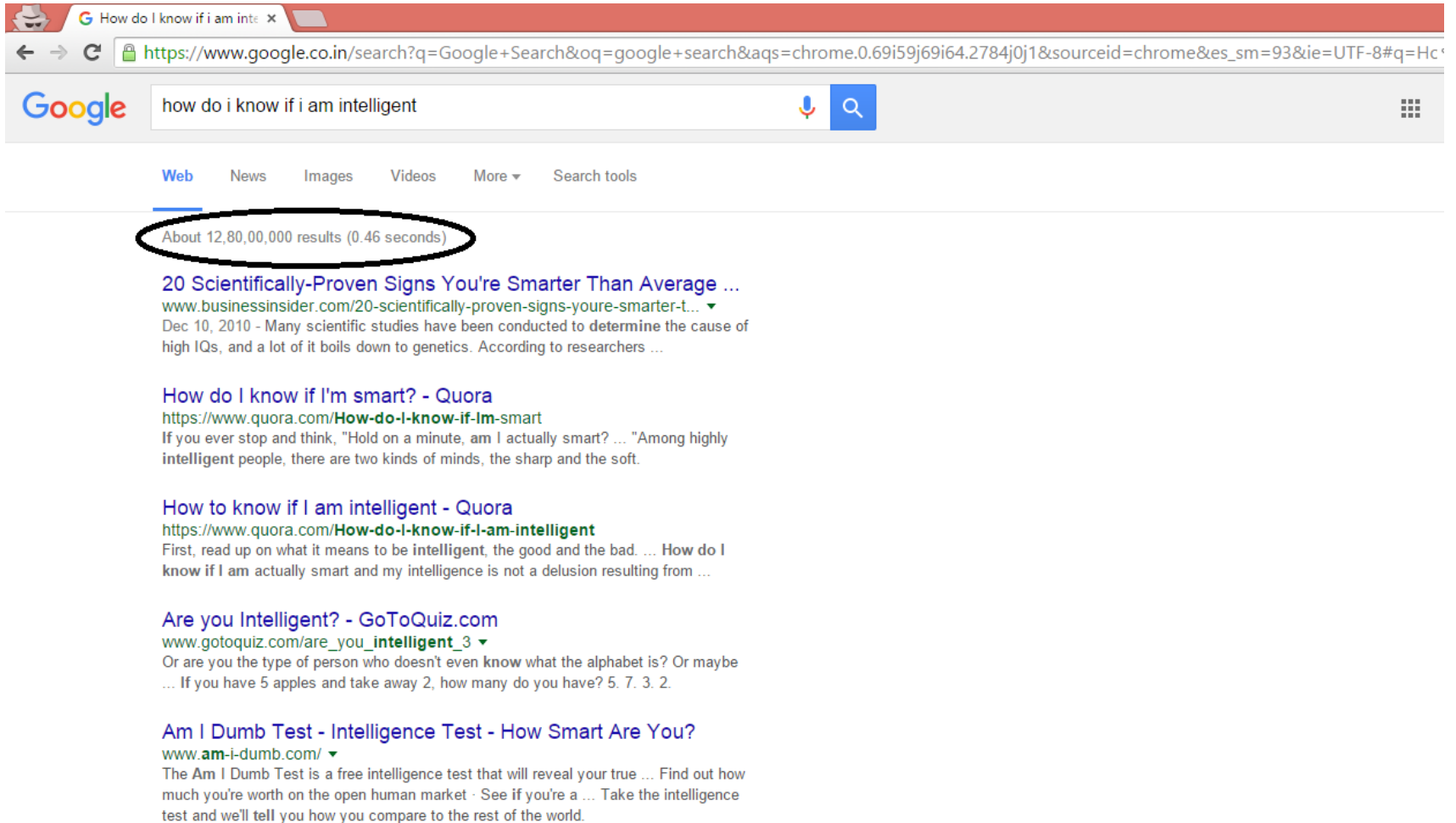
By Brendan Tierney, 2012



# Real Life Examples

- Internet Search
- Digital Advertisements (Targeted Advertising and re-targeting)
- Recommender Systems
- Image Recognition
- Speech Recognition
- Gaming
- Price Comparison Websites
- Airline Route Planning
- Fraud and Risk Detection
- Delivery logistics

# Internet Search



The screenshot shows a Google search results page. The search bar contains the text "how do i know if i am intelligent". Below the search bar, the results are displayed. The first result is titled "20 Scientifically-Proven Signs You're Smarter Than Average ..." and is from the website "www.businessinsider.com". The second result is titled "How do I know if I'm smart? - Quora" and is from the website "https://www.quora.com/How-do-I-know-if-I-m-smart". The third result is titled "How to know if I am intelligent - Quora" and is from the website "https://www.quora.com/How-do-I-know-if-I-am-intelligent". The fourth result is titled "Are you Intelligent? - GoToQuiz.com" and is from the website "www.gotoquiz.com/are\_you\_intelligent\_3". The fifth result is titled "Am I Dumb Test - Intelligence Test - How Smart Are You?" and is from the website "www.am-i-dumb.com/".

How do I know if i am inte x

https://www.google.co.in/search?q=Google+Search&oq=google+search&aqs=chrome.0.69i59j69i64.2784j0j1&sourceid=chrome&es\_sm=93&ie=UTF-8#q=Hc

Google how do i know if i am intelligent

Web News Images Videos More Search tools

About 12,80,00,000 results (0.46 seconds)

**20 Scientifically-Proven Signs You're Smarter Than Average ...**  
www.businessinsider.com/20-scientifically-proven-signs-youre-smarter-t...  
Dec 10, 2010 - Many scientific studies have been conducted to determine the cause of high IQs, and a lot of it boils down to genetics. According to researchers ...


**How do I know if I'm smart? - Quora**  
https://www.quora.com/How-do-I-know-if-I-m-smart  
If you ever stop and think, "Hold on a minute, am I actually smart? ... "Among highly intelligent people, there are two kinds of minds, the sharp and the soft.


**How to know if I am intelligent - Quora**  
https://www.quora.com/How-do-I-know-if-I-am-intelligent  
First, read up on what it means to be intelligent, the good and the bad. ... How do I know if I am actually smart and my intelligence is not a delusion resulting from ...

**Are you Intelligent? - GoToQuiz.com**  
www.gotoquiz.com/are\_you\_intelligent\_3  
Or are you the type of person who doesn't even know what the alphabet is? Or maybe ... If you have 5 apples and take away 2, how many do you have? 5. 7. 3. 2.

**Am I Dumb Test - Intelligence Test - How Smart Are You?**  
www.am-i-dumb.com/  
The Am I Dumb Test is a free intelligence test that will reveal your true ... Find out how much you're worth on the open human market - See if you're a ... Take the intelligence test and we'll tell you how you compare to the rest of the world.

# Targeting Advertisement


[Home](#)
[Resize Image](#)
[Compress Image](#)
[Password Generator](#)
[MP3 Cutter](#)
[Free Proxy](#)
[PNG to ICO](#)


**Hootsuite - 60 Days Free**

Enjoy 60 days free in celebration of our enhanced Instagram features.

**Introduction**

GIFMaker.me allows you to create animated gifs, slideshows, and video animations with music online freely and easily, no registration required.

Select multiple images:


[Upload Images](#)

4 images uploaded successfully

You can make your gif now

Drag the images to change the order:

**Control Panel:**



Canvas size:


100 % 1280 x 851 px

Animation speed:

500 milliseconds

Repeat times ( 0 = infinite loop ):

0 times

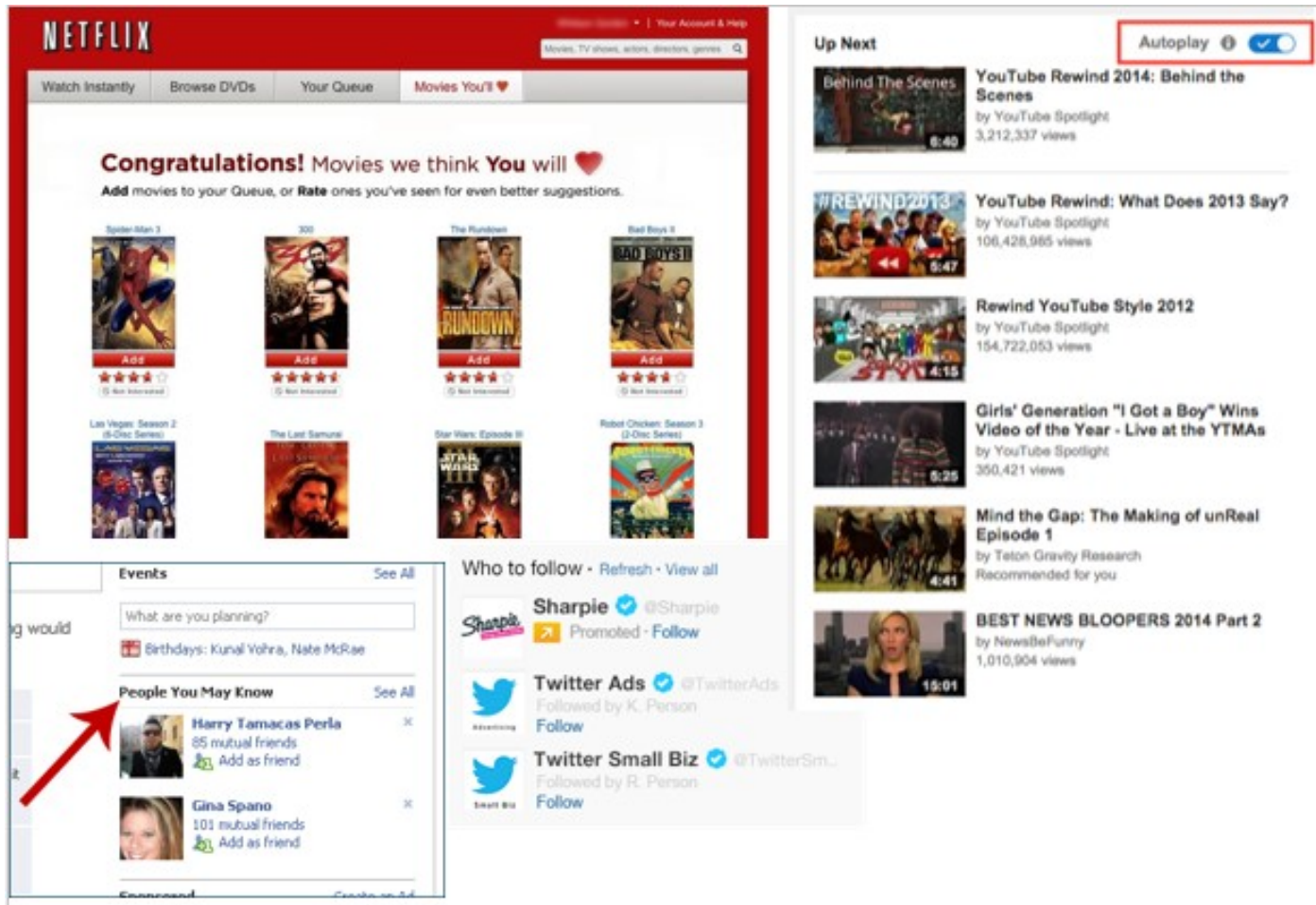


**Telecommuting & FlexJobs**

30% Off Today



# Recommender System

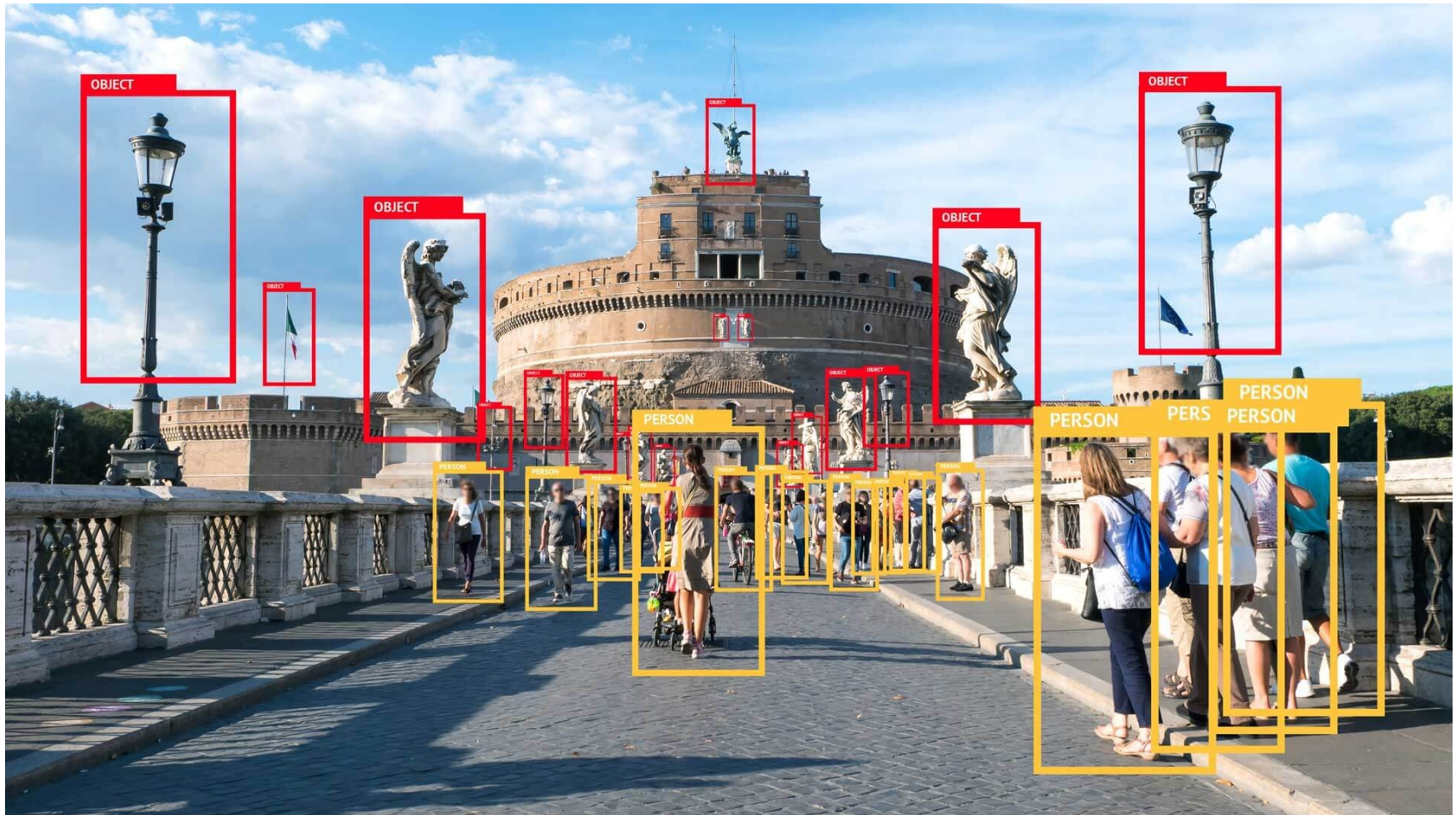


The collage illustrates three different recommendation systems:

- Netflix:** The top section shows a "Congratulations! Movies we think You will" recommendation. Below this, there are two rows of movie thumbnails with "Add" buttons and star ratings. The bottom section shows a "People You May Know" recommendation, with a red arrow pointing to the "Add as friend" button for Harry Tamacas Perla.
- YouTube:** The "Up Next" section shows a list of recommended videos, including "Behind The Scenes", "YouTube Rewind 2014: Behind the Scenes", "YouTube Rewind: What Does 2013 Say?", "Rewind YouTube Style 2012", "Girls' Generation 'I Got a Boy' Wins Video of the Year - Live at the YTMAs", "Mind the Gap: The Making of unReal Episode 1", and "BEST NEWS BLOOPERS 2014 Part 2". An "Autoplay" toggle is visible in the top right corner.
- Facebook:** The "Who to follow" section shows a list of recommended accounts, including "Sharpie", "Twitter Ads", and "Twitter Small Biz".



# Image Recognition



# Speech Recognition





# Computer Games



# Price Comparison Website



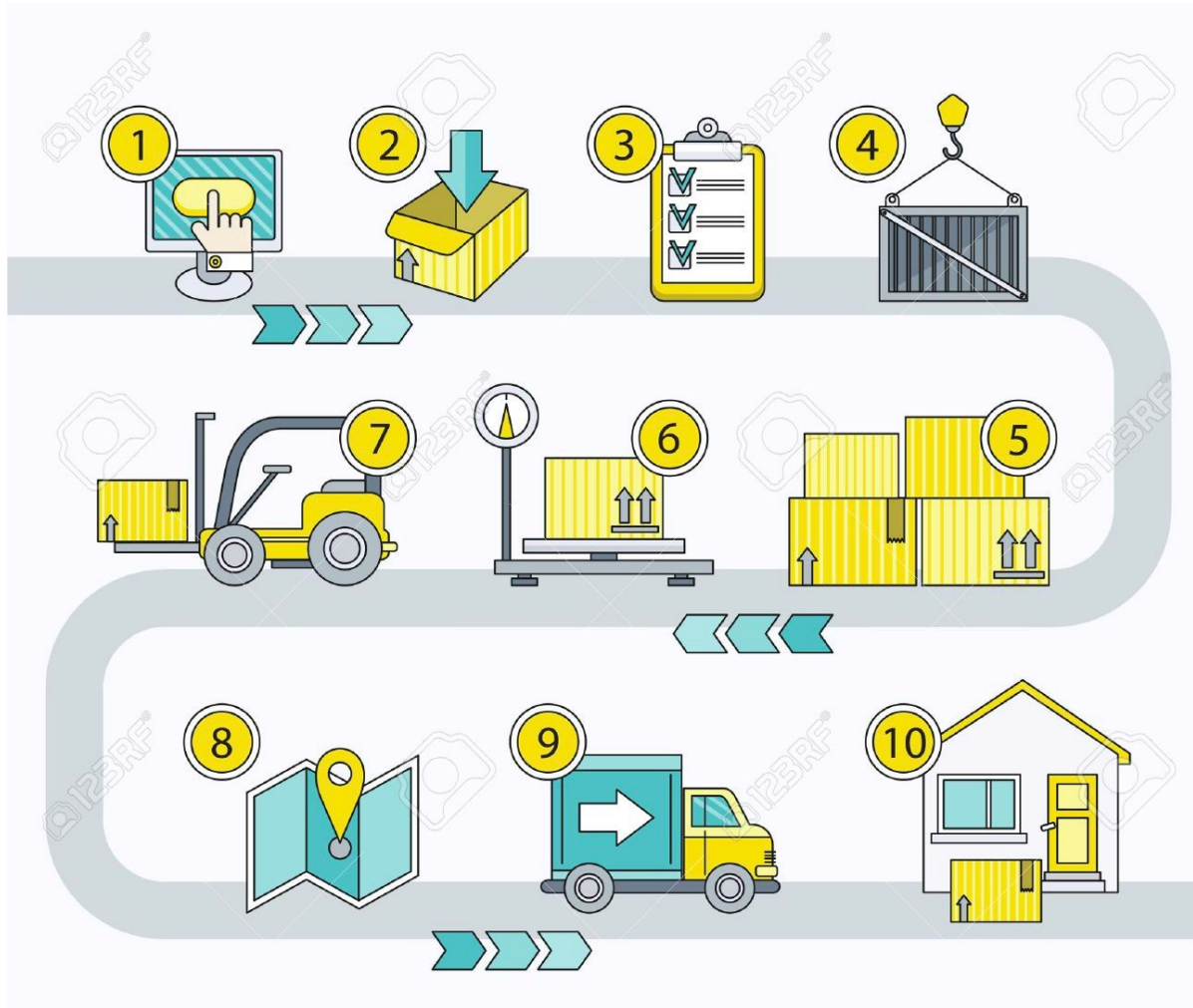




# Fraud Detection



# Delivery Logistics



# Facets of Data

- In data science and big data you'll come across many different types of data, and each of them tends to require different tools and techniques. The main categories of data are these:
  - Structured
  - Unstructured
  - Natural language
  - Machine-generated
  - Graph-based
  - Audio, video, and images
  - Streaming



# Strutured Data

- Structured data is data that depends on a data model and resides in a fixed field within a record.
- As such, it's often easy to store structured data in tables within databases or Excel files, SQL , or Structured Query Language, is the preferred way to manage and query data that resides in databases.
- You may also come across structured data that might give you a hard time storing it in a traditional relational database.

# Strutured Data

1	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Int
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

# Unstructured Data

- Unstructured data is data that isn't easy to fit into a data model because the content is context-specific or varying. One example of unstructured data is your regular email
- Although email contains structured elements such as the sender, title, and body text, it's a challenge to find the number of people who have written an email complaint about a specific employee because so many ways exist to refer to a person, for example.
- The thousands of different languages and dialects out there further complicate this.
- A human-written email, as shown in next figure, is also a perfect example of natural language data.

# Unstructured Data



# Natural Language

- Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques and linguistics.
- The natural language processing community has had success in entity recognition, topic recognition, summarization, text completion, and sentiment analysis, but models trained in one domain don't generalize well to other domains.
- Even state-of-the-art techniques aren't able to decipher the meaning of every piece of text. This shouldn't be a surprise though: humans struggle with natural language as well. It's ambiguous by nature.

# Machine Generated Data

- Machine-generated data is information that's automatically created by a computer, process, application, or other machine without human intervention.
- Machine-generated data is becoming a major data resource and will continue to do so. Wikibon has forecast that the market value of the industrial Internet (a term coined by Frost & Sullivan to refer to the integration of complex physical machinery with networked sensors and software) will be approximately \$540 billion in 2020.
- IDC (International Data Corporation) has estimated there will be 26 times more connected things than people in 2020. This network is commonly referred to as the internet of things.

# Machine Generated Data

```
CSIPERF:TXCOMMIT;313236
2014-11-28 11:36:13, Info
69), objectname [6]"(null)"
2014-11-28 11:36:13, Info
result 0x00000000, handle @0x4e54
2014-11-28 11:36:13, Info
Beginning NT transaction commit...
2014-11-28 11:36:13, Info
trace:
CSIPERF:TXCOMMIT;273983
2014-11-28 11:36:13, Info
70), objectname [6]"(null)"
2014-11-28 11:36:13, Info
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:13, Info
Beginning NT transaction commit...
2014-11-28 11:36:14, Info
trace:
CSIPERF:TXCOMMIT;386259
2014-11-28 11:36:14, Info
71), objectname [6]"(null)"
2014-11-28 11:36:14, Info
result 0x00000000, handle @0x4e5c
2014-11-28 11:36:14, Info
Beginning NT transaction commit...
2014-11-28 11:36:14, Info
trace:
CSIPERF:TXCOMMIT;375581
```

```
CSI 00000153 Creating NT transaction (seq
CSI 00000154 Created NT transaction (seq 69)
CSI 00000155@2014/11/28:10:36:13.471
CSI 00000156@2014/11/28:10:36:13.705 CSI perf
CSI 00000157 Creating NT transaction (seq
CSI 00000158 Created NT transaction (seq 70)
CSI 00000159@2014/11/28:10:36:13.764
CSI 0000015a@2014/11/28:10:36:14.094 CSI perf
CSI 0000015b Creating NT transaction (seq
CSI 0000015c Created NT transaction (seq 71)
CSI 0000015d@2014/11/28:10:36:14.106
CSI 0000015e@2014/11/28:10:36:14.428 CSI perf
```

# Graph or Network Data

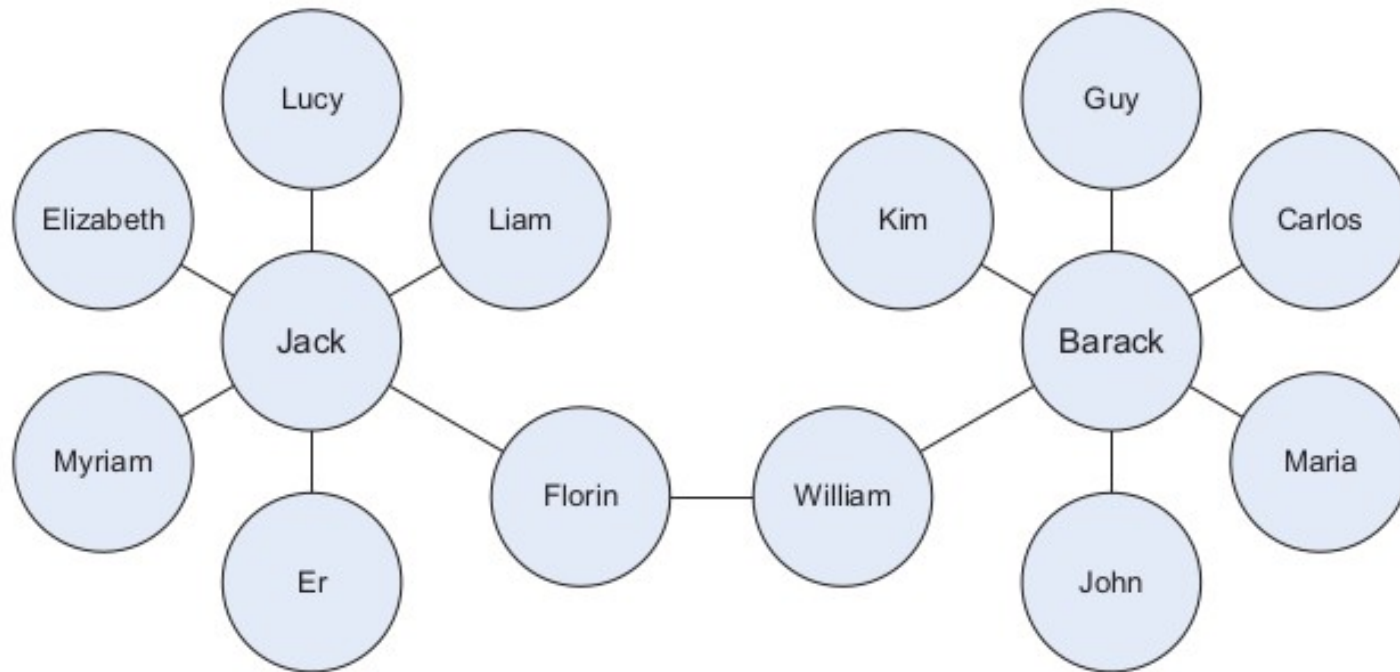
- “Graph data” can be a confusing term because any data can be shown in a graph.
- “Graph” in this case points to mathematical graph theory. In graph theory, a graph is a mathematical structure to model pair-wise relationships between objects. Graph or network data is, in short, data that focuses on the relationship or adjacency of objects.
- The graph structures use nodes, edges, and properties to represent and store graphical data. Graph-based data is a natural way to represent social networks, and its structure allows you to calculate specific metrics such as the influence of a person and the shortest path between two people.



# Graph or Network Data

- Examples of graph-based data can be found on many social media websites (For instance, on LinkedIn you can see who you know at which company.
- Your follower list on Twitter is another example of graph-based data. The power and sophistication comes from multiple, overlapping graphs of the same nodes. For example, imagine the connecting edges here to show “friends” on Facebook.
- Imagine another graph with the same people which connects business colleagues via LinkedIn.
- Imagine a third graph based on movie interests on Netflix. Overlapping the three different-looking graphs makes more interesting questions possible.

# Graph or Network Data



# Audio, Video and Image

- Audio, image, and video are data types that pose specific challenges to a data scientist.
- Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers. MLBAM (Major League Baseball Advanced Media) announced in 2014 that they'll increase video capture to approximately 7 TB per game for the purpose of live, in-game analytics.
- High-speed cameras at stadiums will capture ball and athlete movements to calculate in real time, for example, the path taken by a defender relative to two baselines.

# Audio, Video and Image

- Recently a company called DeepMind succeeded at creating an algorithm that's capable of learning how to play video games.
- This algorithm takes the video screen as input and learns to interpret everything via a complex process of deep learning. It's a remarkable feat that prompted Google to buy the company for their own Artificial Intelligence ( AI ) development plans.
- The learning algorithm takes in data as it's produced by the computer game; it's streaming data.

# Streaming Data

- While streaming data can take almost any of the previous forms, it has an extra property.
- The data flows into the system when an event happens instead of being loaded into a data store in a batch.
- Although this isn't really a different type of data, we treat it here as such because you need to adapt your process to deal with this type of information.
- Examples are the “What’s trending” on Twitter, live sporting or music events, and the stock market.

# Thank you

*This presentation is created using LibreOffice Impress 5.1.6.2, can be used freely as per GNU General Public License*



@mitu\_skillologies



/MITuSkillologies



@mitu\_group



/company/mitu-  
skillologies



MITUSkillologies

## Web Resources

<https://mitu.co.in>

<http://tusharkute.com>

[contact@mitu.co.in](mailto:contact@mitu.co.in)

[tushar@tusharkute.com](mailto:tushar@tusharkute.com)