

# Data Science Tools and Techniques

Dr. Muhammad Nouman Durrani

Disclaimer: Data from various Internet sources have been used.  
The Instructor fully acknowledge the copyrights.

# What is Data Science

- As the world entered the era of big data, the need for its storage also grew.
- It was the main challenge and concern for the enterprise industries until 2010.
- The main focus was on building framework and solutions to store data.
- Now when Hadoop and other frameworks have successfully solved the problem of storage, the focus has shifted to the processing of this data.

# 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.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.

# WHAT IS DATA SCIENCE?

- **Fortune**

“Hot New Gig in Tech”

- **Hal Varian, Google’s Chief Economist, NYT, 2009:**

“The next sexy job”

“The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill.”

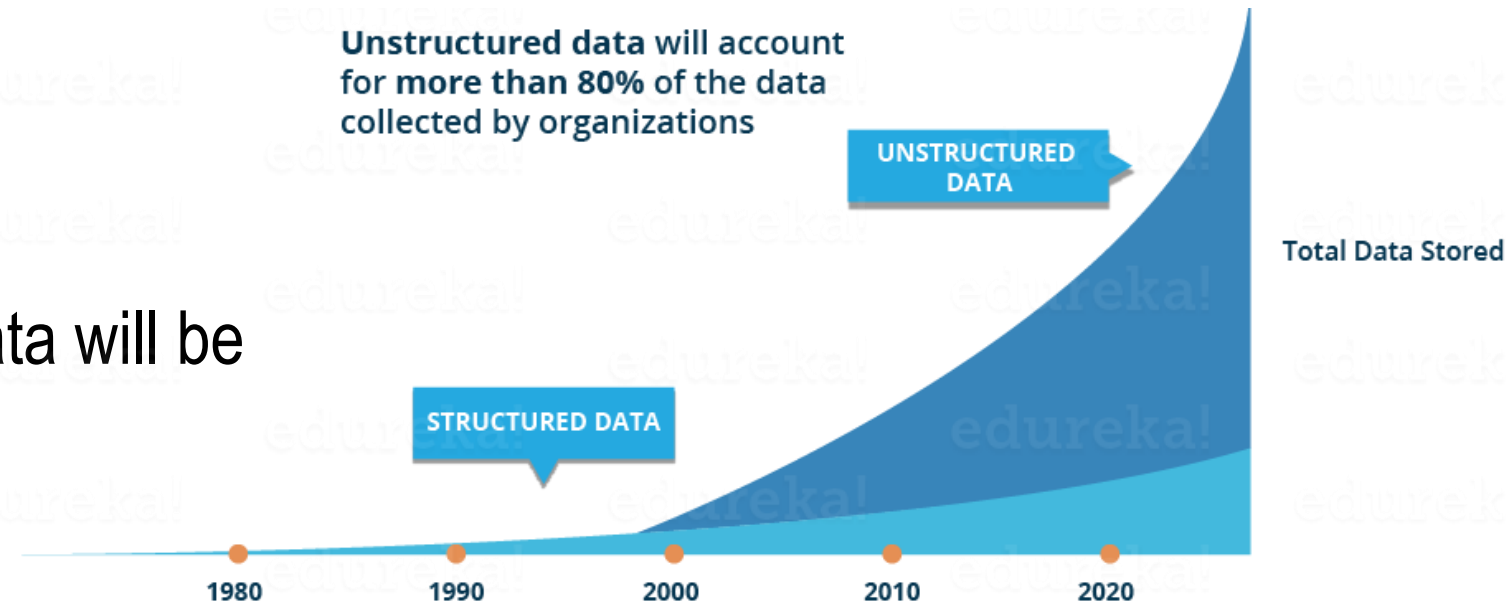
- **Mike Driscoll, CEO of metamarkets:**

“Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.”

“Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools & materials, coupled with a theoretical understanding of what's possible.”

# Why We Need Data Science

- Traditionally, the data was mostly structured and small in size, which could be analyzed using simple **traditional tools**.
- Today most of the data is unstructured or semi-structured.
- By 2020, more than 80 % of the data will be unstructured.



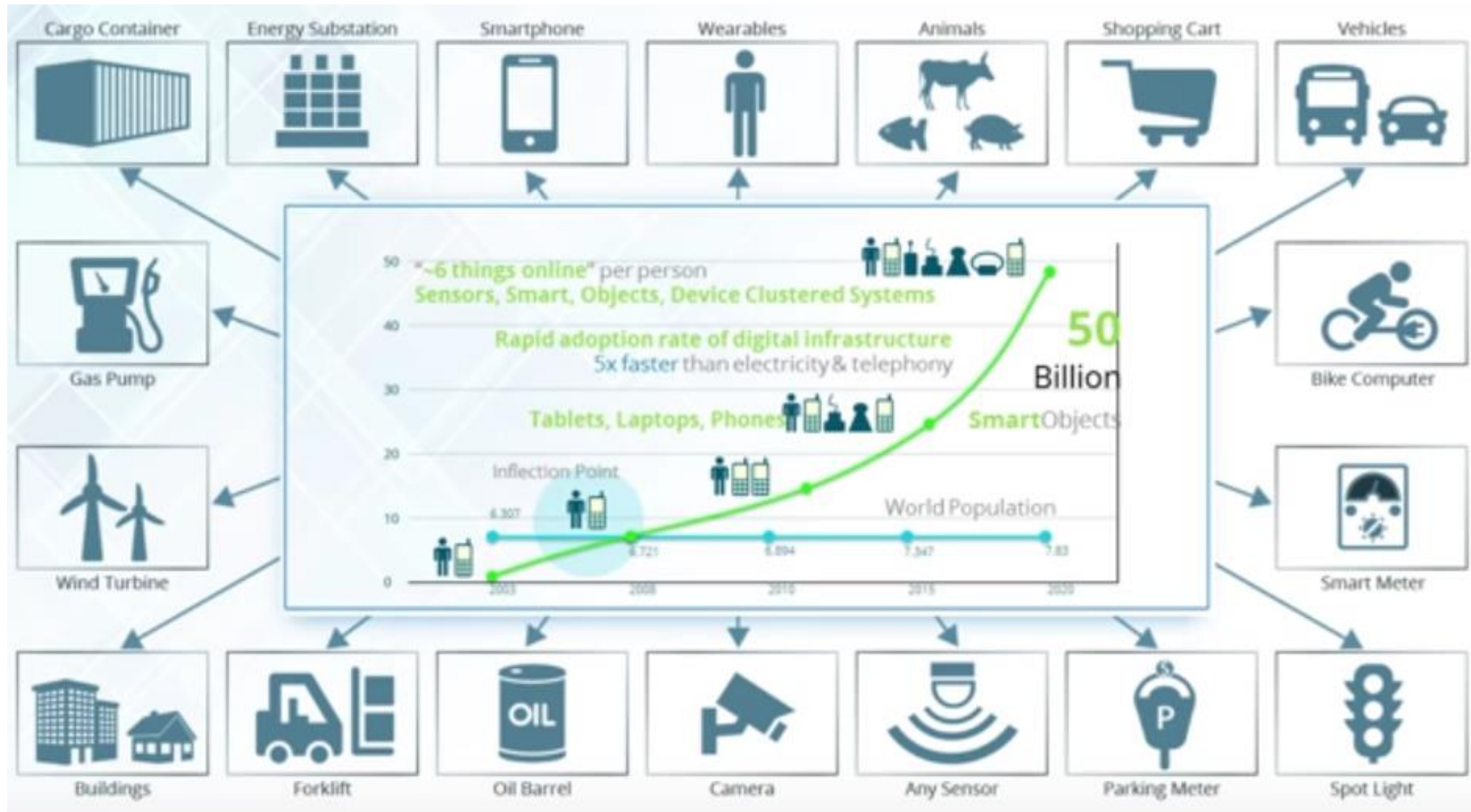
# From where the data comes from?

- This data is generated from different sources like:
  - financial logs, text files, multimedia forms, sensors, and instruments.
- **Simple tools** are not capable of processing this huge volume and variety of data.
- This is why **we need more complex and advanced analytical tools and algorithms** for processing, analyzing and drawing meaningful insights out of it.

# Data Generated Every Minute!



# IoTs: 50 Billion Connected Devices by 2020





# THE WORLD OF DATA

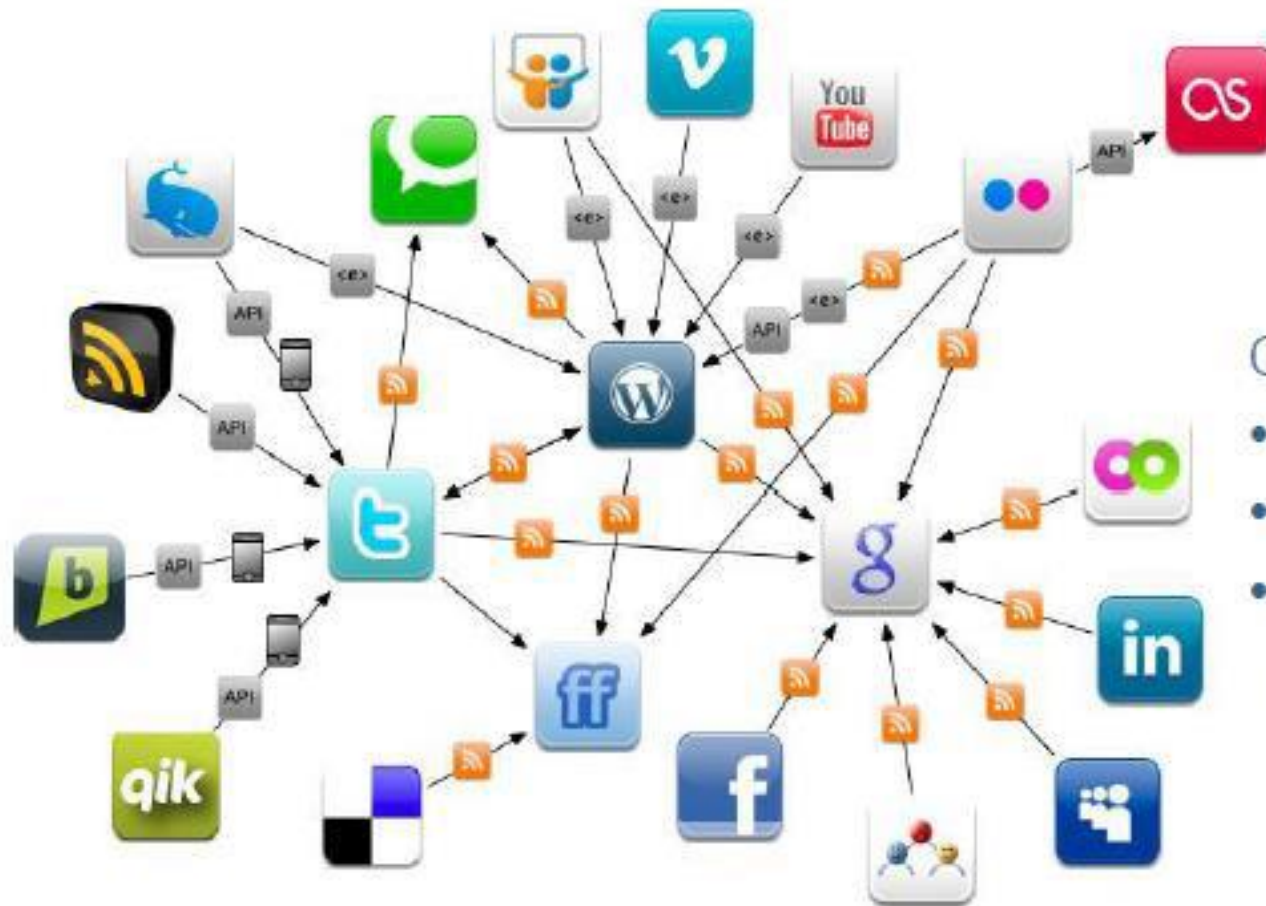
- Peta-bytes are in norm
  - Google processes 24 PB a day (2009)
  - AT&T transfers about 30 PB a day through its networks
  - Microsoft migrated 150 PB of user data from Hotmail to Outlook (2013)
  - Facebook stores about 357 PB of user uploaded images (2013)
  - eBay has 6.5 PB of user data + 50 TB/day (2009)
- How big is Internet? 672 Exabytes of accessible data (2013)

# Contributors: Surveillance guys



1 VGA resolution color camera  
produces 800 GB/hour

# Contributors: Social Networks



On a typical day:

- 500 million tweets
- 55 million FB status updates
- 1 billion pieces of content shared on FB



Anne Helmond, May 2009

# Contributors: Scientific Instruments



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



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



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

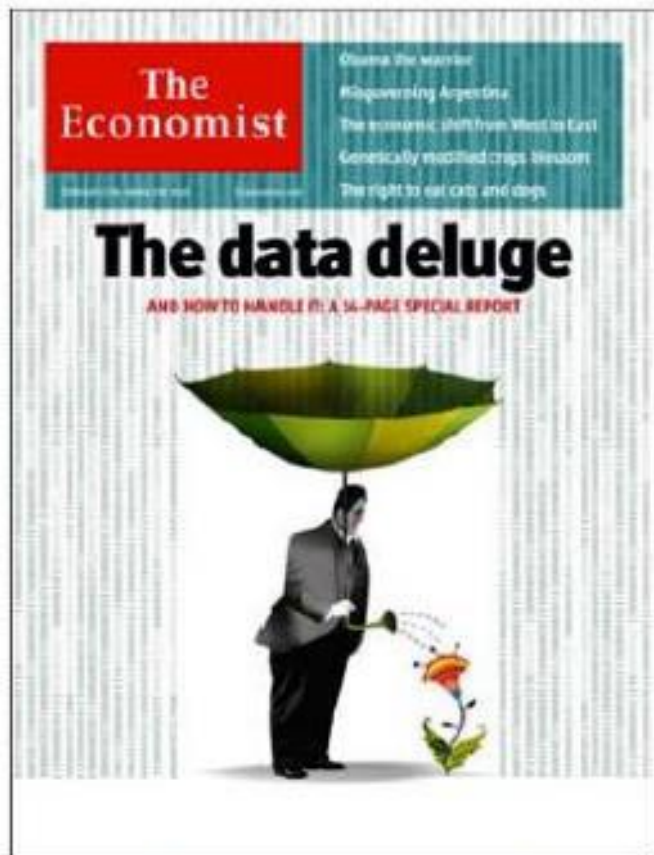


**Sensor technology and networks**  
(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion



# Customer Challenges: The Data Deluge



The Economist, Feb 25, 2010

IN 2010 THE DIGITAL UNIVERSE WAS  
**1.2 ZETTABYTES**

IN A DECADE THE DIGITAL UNIVERSE WILL BE  
**35 ZETTABYTES**

**90%** OF THE DIGITAL UNIVERSE IS  
**UNSTRUCTURED**

IN 2011 THE DIGITAL UNIVERSE IS  
**300 QUADRILLION** FILES

The **data deluge** refers to the situation where the sheer volume of new **data** being generated is overwhelming the capacity of institutions to manage it and researchers to make use of it.

**WIRED**

The New York Times

Bloomberg  
Businessweek

Forbes

WALL STREET JOURNAL

# Big Data Definition

**Big Data:** Massive sets of unstructured/semi-structured data from Web traffic, social media, sensors, etc.

- 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...

# Is it only the volume that makes it Big?

- “Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population.” *Teradata magazine article, 2011*
- “Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.” *The McKinsey Global Institute, 2011*

# Is it only the volume that makes it Big?

► “Big data is a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.” *Wikipedia*

“Big Data is any data that is expensive to manage and hard to extract value from.” *Michael Franklin, Univ; of California, Berkeley*



# Case 1: Recommend products to your customer

- How about if you could understand the precise requirements of your customers from the existing data like the customer's
  - past browsing history,
  - purchase history,
  - age and
  - income.
- No doubt we had all this data earlier too, but now with the vast amount and variety of data, we can train models more effectively and **recommend the product to the customers with more precision.**
- It will bring **more business to your organization.**

## Case 2: The role of Data Science in decision making

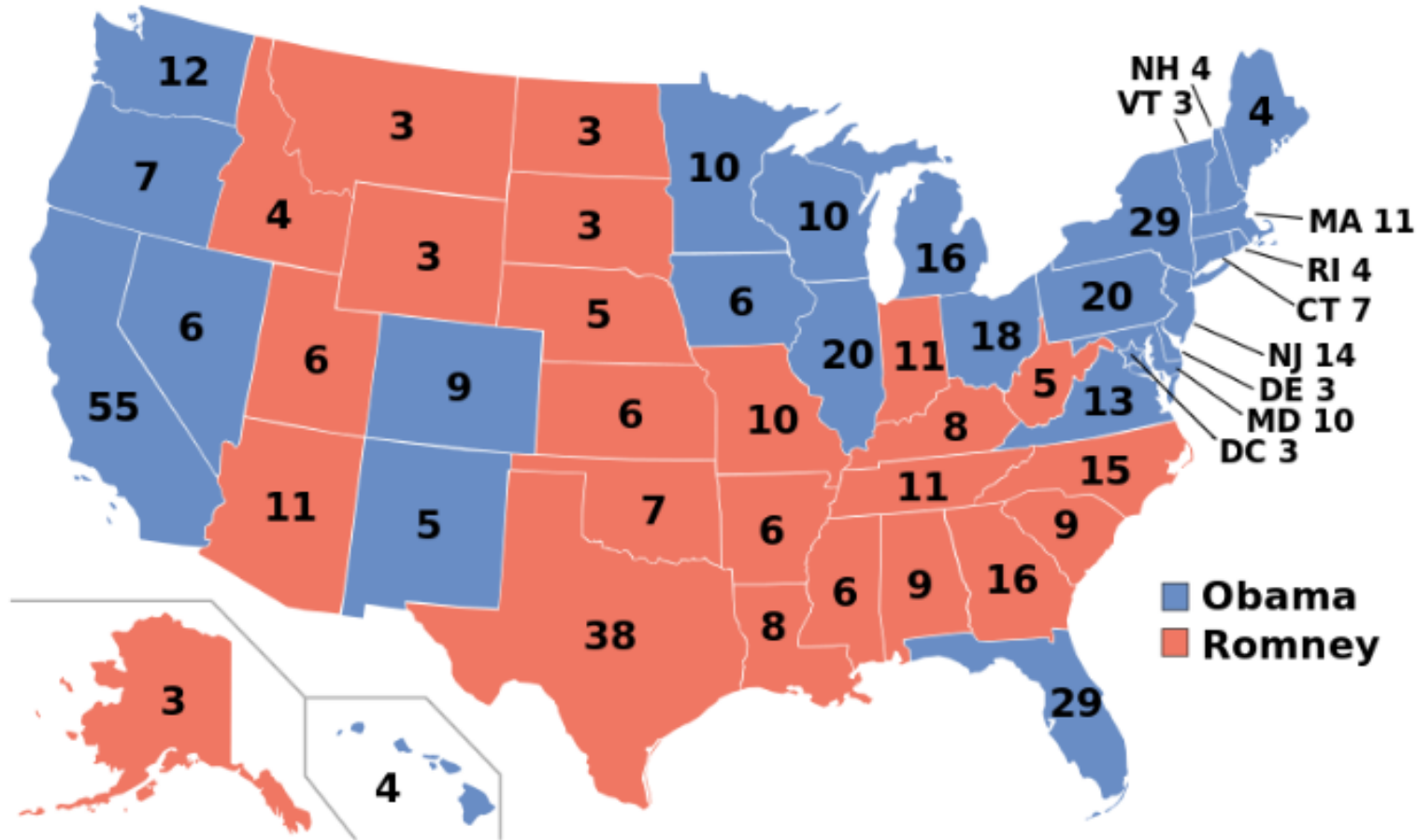
- How about if your car had the intelligence to drive you home?
- The self-driving cars collect live data from sensors, including radars, cameras and lasers to create a map of its surroundings.
- Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where to take a turn
  - making use of advanced machine learning algorithms.

## Case 3: Weather forecasting Data Science in predictive analytics

- Data from ships, aircrafts, radars, satellites can be collected and analyzed to build models.
- These models will not only forecast the weather but also help in predicting the occurrence of any natural calamities.
- It will help you to take appropriate measures beforehand and save many precious lives.

# NATE SILVER

“Silver, who made his name by using cold hard math to call **49 out of 50** states in the 2008 general election and **all 50 in 2012**”



<http://commons.wikimedia.org/wiki/File:ElectoralCollege2012.svg>  
(public domain)

## RELATED: OBAMA CAMPAIGN'S DATA-DRIVEN GROUND GAME

"In the 21st century, **the candidate with [the] best data**, merged with the best messages dictated by that data, **wins**."

Andrew Rasiej, Personal Democracy Forum

"...the **biggest win came from good old SQL** on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed."

Dan Woods

Jan 13 2013, CITO Research

"The decision was made to have **Hadoop** do the aggregate generations and anything not real-time, but then have Vertica to answer sort of 'speed-of-thought' queries about all the data."

Josh Hendler, CTO of H & K Strategies

# ELECTION 2016

## “Donald Trump Is The Nickelback Of GOP Candidates”

“[d]isliked by most, super popular with a few”

### Trump Is The 13th Most Popular GOP Candidate

Average of national, Iowa and New Hampshire polls since July 18

	CANDIDATE	FAVORABLE	UNFAVORABLE	NET FAVORABLE	FIRST CHOICE
1	Walker	56%	13%	+43%	14%
2	Rubio	56%	16%	+39%	6%
3	Carson	50%	15%	+35%	7%
4	Jindal	45%	18%	+27%	2%
5	Fiorina	44%	17%	+27%	2%
6	Cruz	49%	23%	+27%	5%
7	Huckabee	52%	27%	+26%	5%
8	Perry	50%	25%	+25%	2%
9	Santorum	44%	28%	+16%	1%
10	Bush	50%	34%	+16%	12%
11	Paul	44%	30%	+14%	5%
12	Kasich	31%	17%	+14%	4%
13	Trump	47%	43%	+4%	20%
14	Christie	35%	47%	-12%	3%

<http://fivethirtyeight.com/datalab/donald-trump-is-the-nickelback-of-gop-candidates/>



# How Nate Silver Missed Donald Trump

The election guru said Trump had no shot. Where did he go wrong?

By *Leon Neyfakh*



2.2k



318



980



Polls whiz kid Nate Silver and presidential candidate Donald Trump.

"If Silver's system depends largely on interpreting poll numbers, how reliable can that system be if the pre-Iowa and New Hampshire polls are basically worthless? **Garbage in, garbage out.**"

[http://www.slate.com/articles/news\\_and\\_politics/politics/2016/01/nate\\_silver\\_said\\_donald\\_trump\\_had\\_no\\_shot\\_where\\_did\\_he\\_go\\_wrong.2.html](http://www.slate.com/articles/news_and_politics/politics/2016/01/nate_silver_said_donald_trump_had_no_shot_where_did_he_go_wrong.2.html)

# EXPRESSION OF EMOTIONS OVER THE 20<sup>TH</sup> CENTURY

- 1) Convert all the digitized books in the 20<sup>th</sup> century into n-grams  
(Thanks, Google!)

(<http://books.google.com/ngrams/>)

*A 1-gram: "yesterday"*

*A 5-gram: "analysis is often described as"*

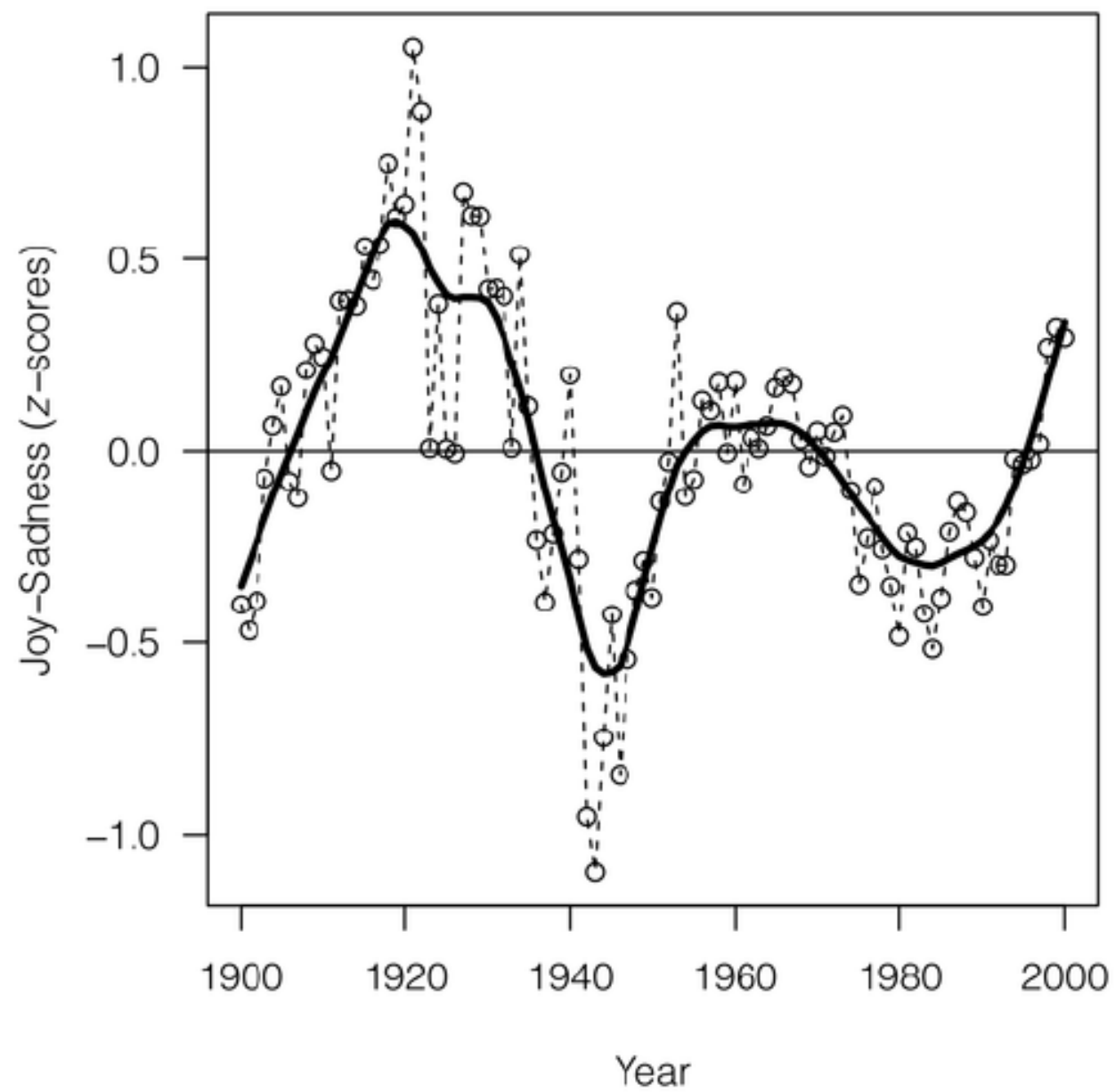
- 2) Label each 1-gram (word) with a mood score.  
(Thanks, WordNet Affect)

- 3) Count the occurrences of each mood word

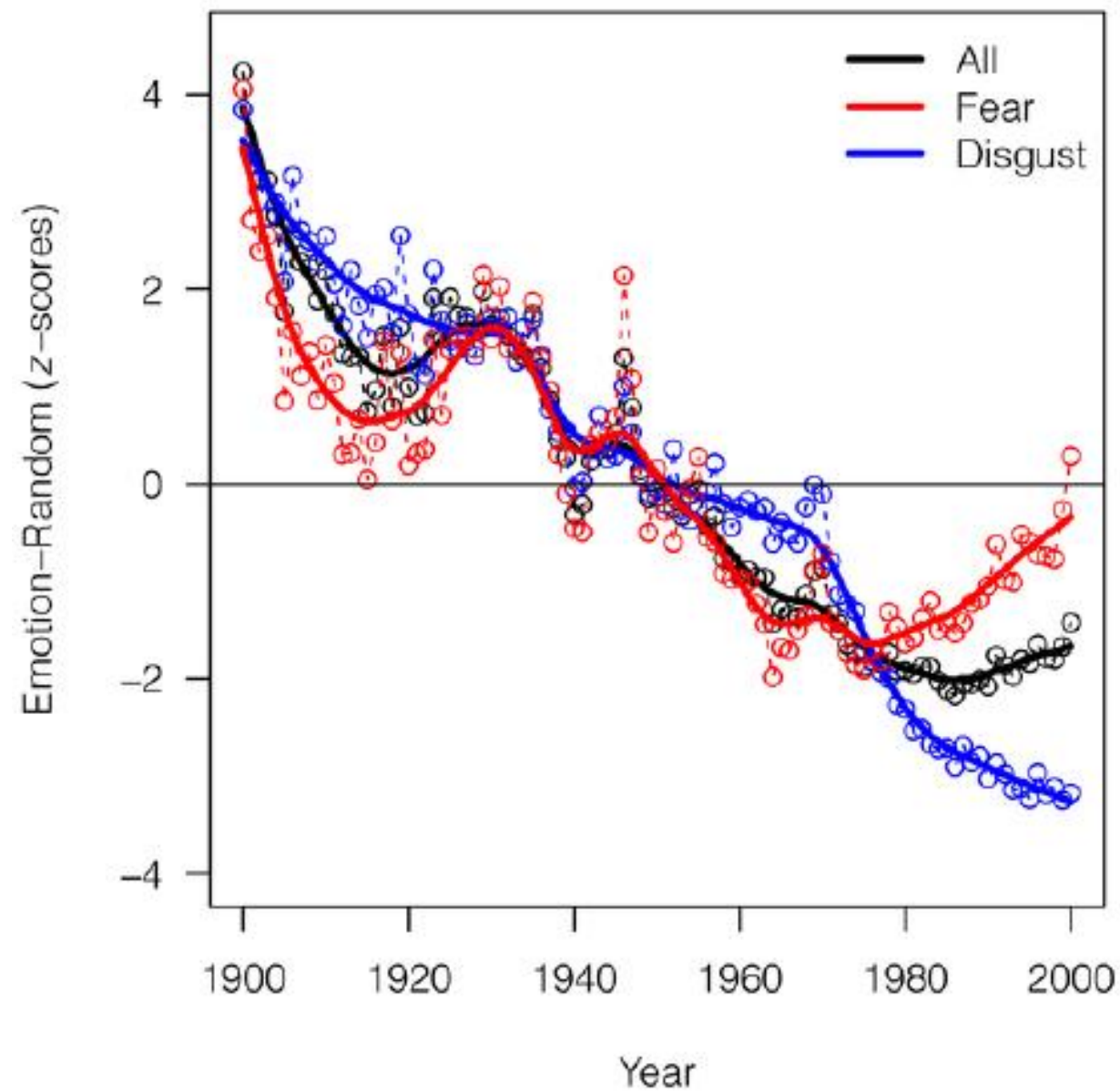
$$\mathcal{M}_Y = \frac{1}{n} \sum_{i=1}^n \frac{c_i}{C_{\text{the}}},$$

$$\mathcal{M}z_Y = \frac{\mathcal{M}_Y - \mu_{\mathcal{M}}}{\sigma_{\mathcal{M}}},$$





Acerbi A, Lamos V, Garnett P, Bentley RA (2013) **The Expression of Emotions in 20th Century Books**. PLoS ONE 8(3): e59030. doi:10.1371/journal.pone.0059030





# Flavor network and the principles of food pairing

Yong-Yeol Ahn, Sebastian E. Ahnert, James P. Bagrow & Albert-László Barabási

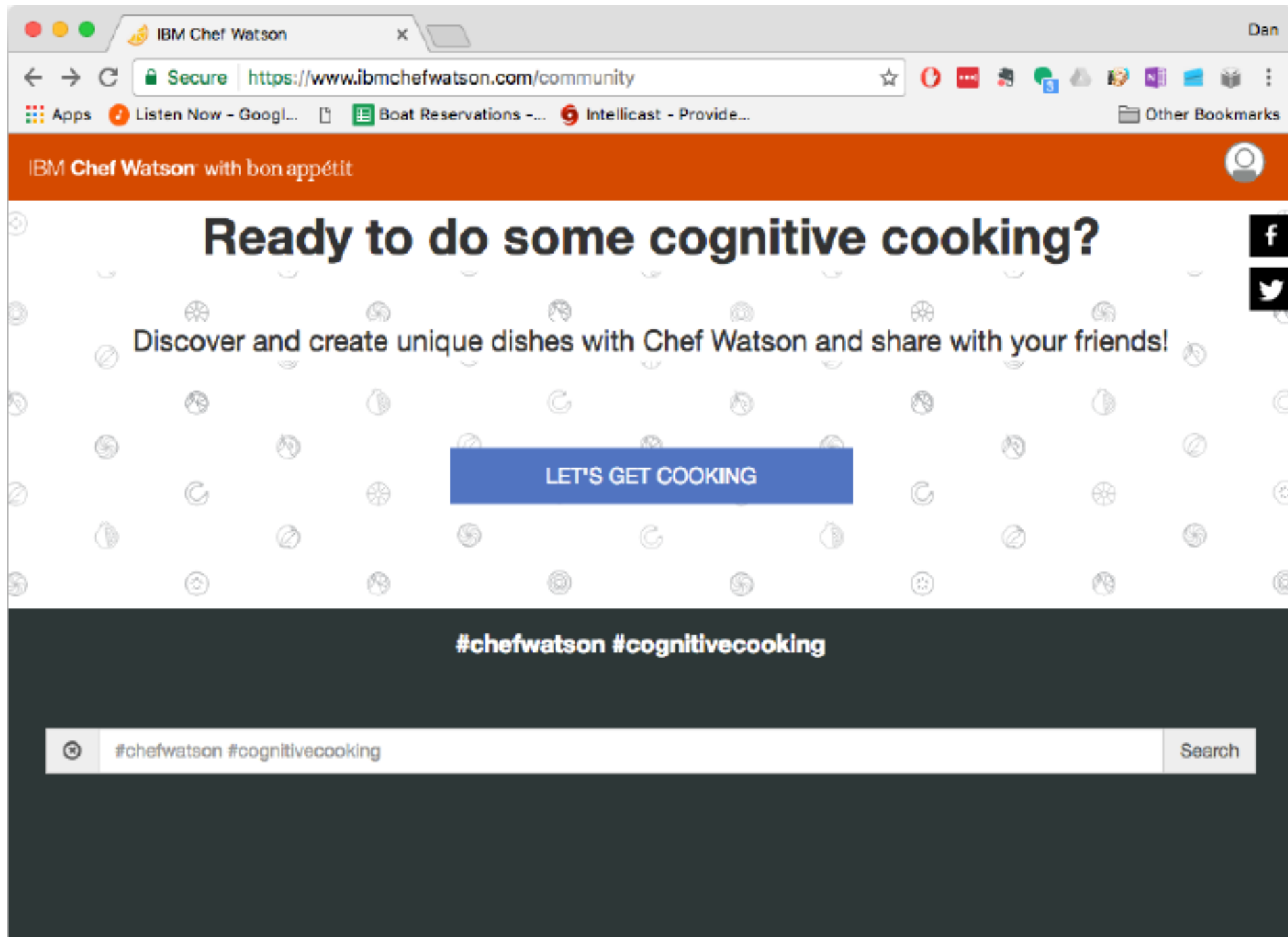
[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

*Scientific Reports* 1, Article number: 196 | doi:10.1038/srep00196

Received 18 October 2011 | Accepted 24 November 2011 | Published 15 December 2011

*Idea: Analyze the co-occurrence graph of ingredients in recipes to analyze the underlying principles of food pairing.*



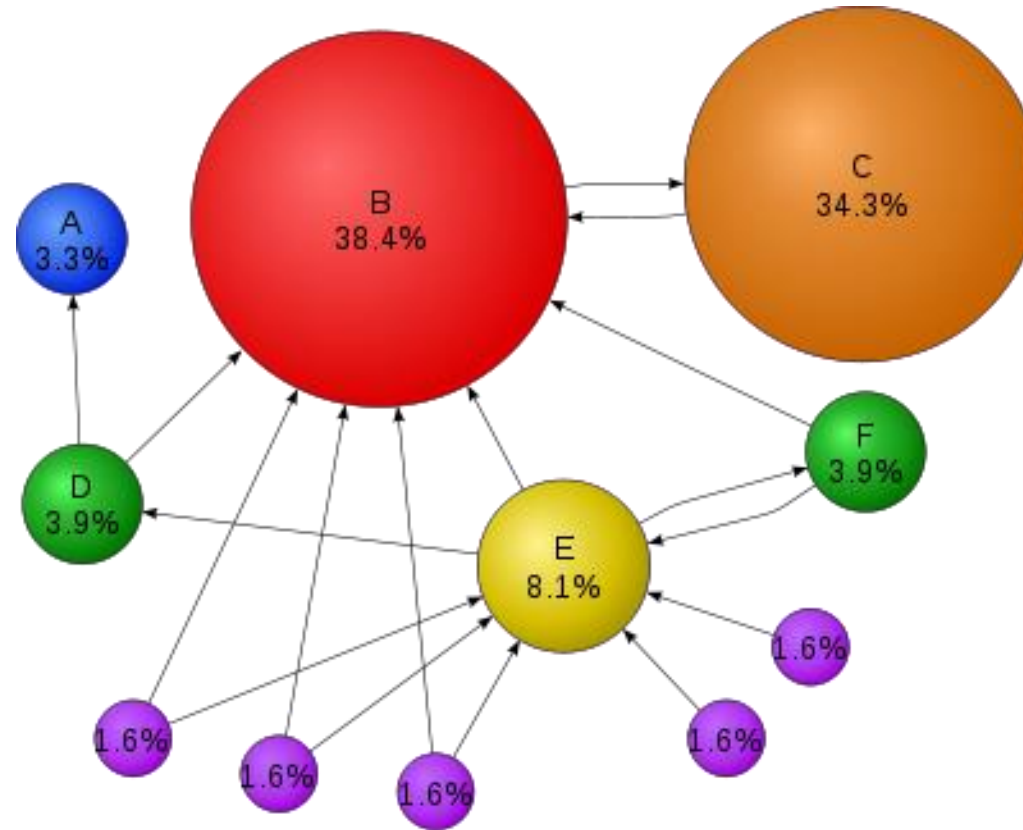


TRENDS + NEWS

## We Spent a Year Cooking With the World's Smartest Computer — and Now You Can, Too



## PageRank: The web as a behavioral dataset



# ONLINE EXAMPLES

- **Cooking with Watson**

- <https://www.ibmchefwatson.com/community>

- **Google Flu Data:**

- [https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac\\_](https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_)

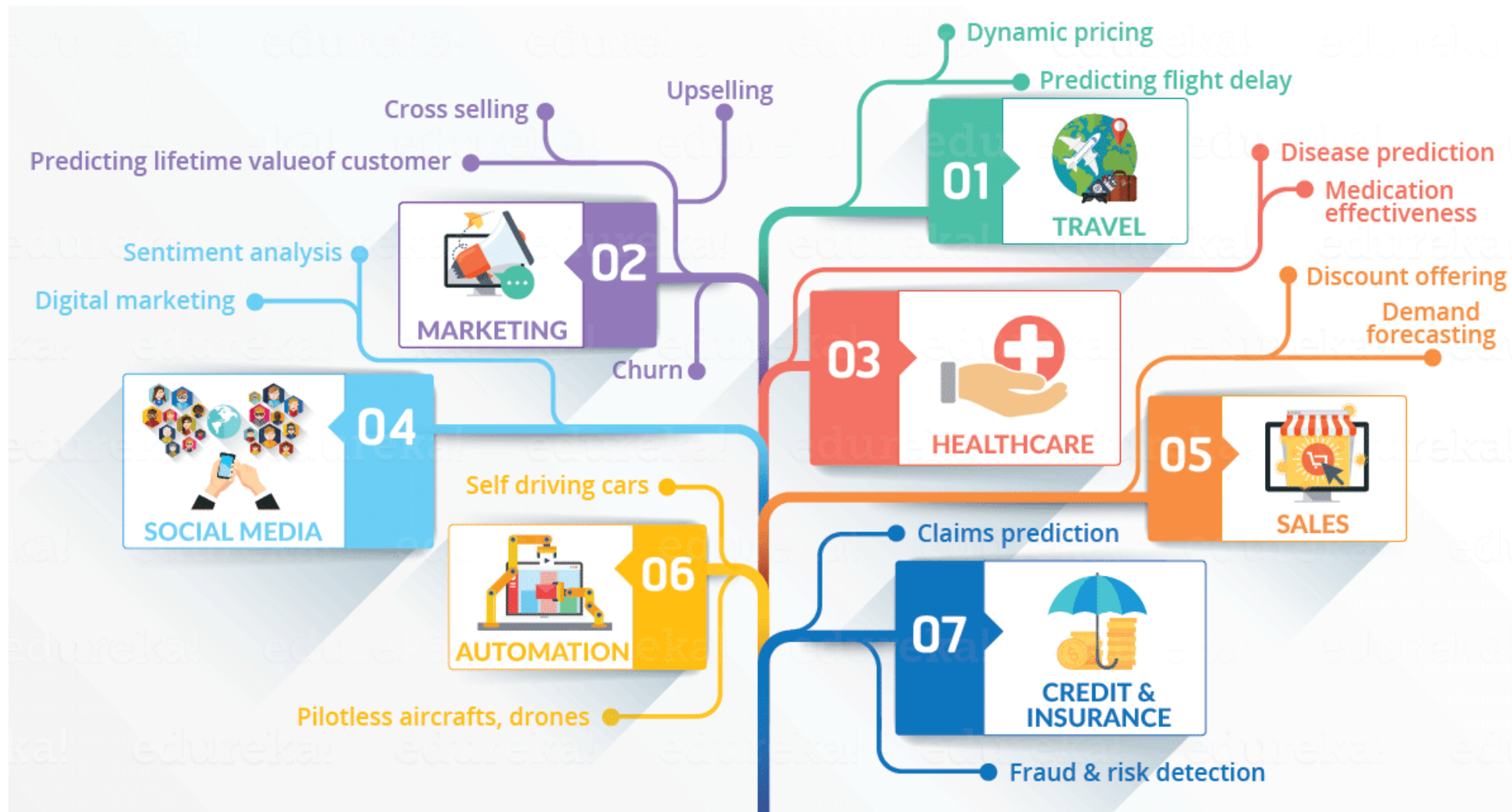
- **3<sup>rd</sup> Party Google Flu Data in D3**

- <http://stat4701-edav-d3.github.io/viz/cities/cities.html>

- **Global Burden of Disease in D3**

- <http://www.healthdata.org/gbd/data-visualizations>

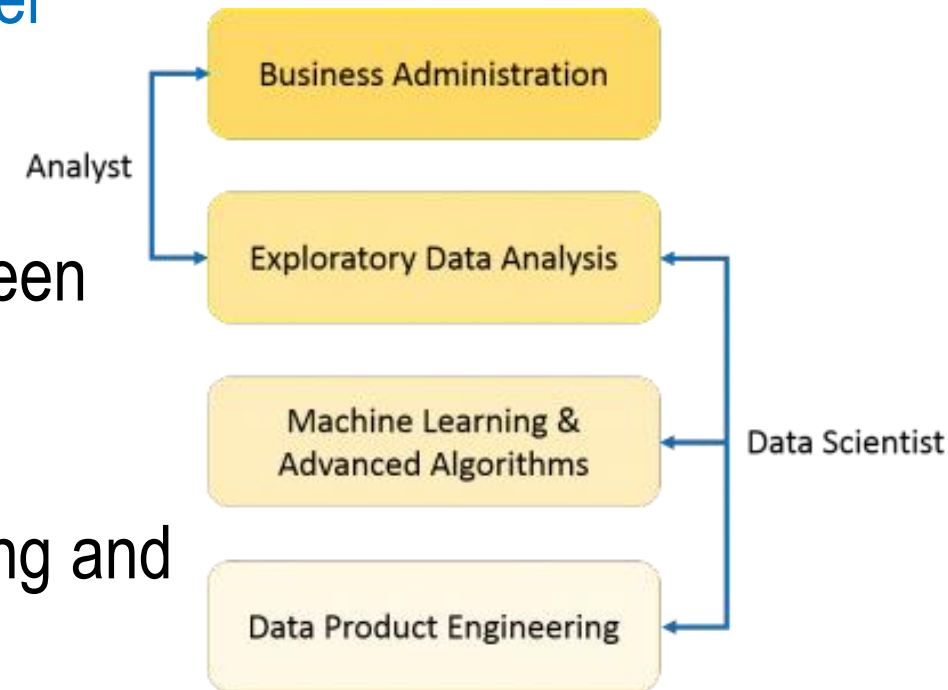
An infographic to see domains where Data Science is creating its impression.





# 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.
- How is this different from what statisticians have been doing for years?
- The answer lies in the difference between explaining and predicting.



# Data Analyst and a Data Scientist

- A Data Analyst usually explains what is going on by processing history of the data.
- Data Scientist not only does the **exploratory analysis** to discover insights from it, but also **uses** various advanced **machine learning algorithms** to identify the occurrence of a particular event in the future.
- Data Science is primarily used **to make decisions and predictions** making use of:
  - Predictive causal analytics,
  - Prescriptive analytics (predictive plus decision science) and
  - Machine Learning.

# Predictive causal analytics –

- If you want a model which can predict the possibilities of a particular event in the future, you need to apply **predictive causal analytics**.
- **For Example:** If you are providing money on credit, then the probability of customers making future credit payments on time is a matter of concern for you.
- Here, you can build a model which can perform predictive analytics on the payment history of the customer to predict if the future payments will be on time or not.

# Prescriptive analytics:

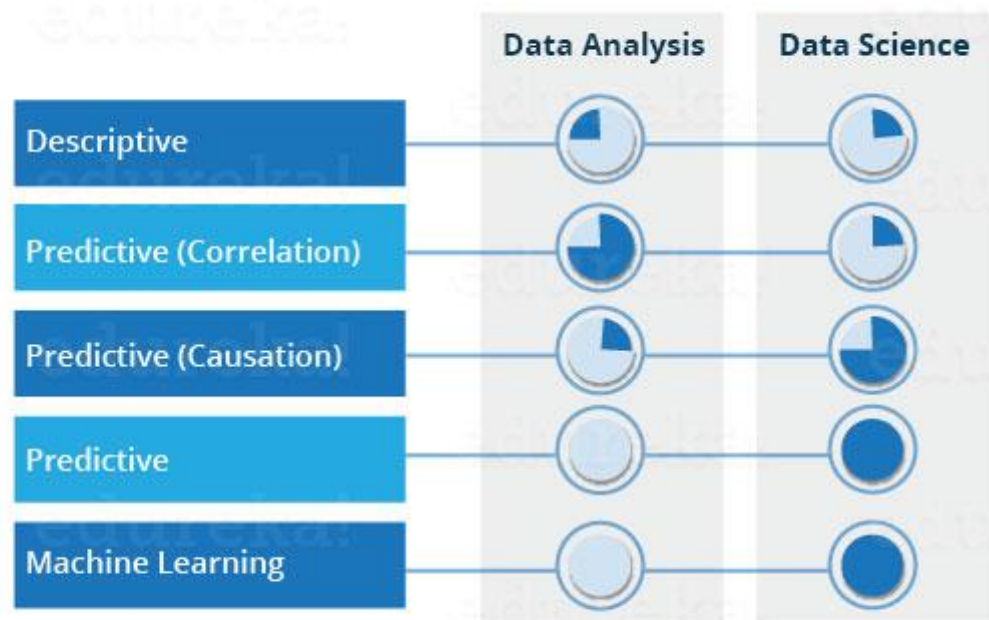
- If you want a model which has the intelligence of taking its own decisions and the ability to modify it with dynamic parameters, you need prescriptive analytics for it.
- This relatively new field is all about providing advice.
- In other terms, it not only predicts but suggests a range of prescribed actions and associated outcomes.
- The best example for this is Google's self-driving car.
  - The data gathered by vehicles can be used to train self-driving cars.
  - We run algorithms on this data to bring intelligence to it.
  - This will enable your car to take decisions like when to turn, which path to take, when to slow down or speed up.

# Machine learning for making predictions

- If you have transactional data of a finance company and need to build a model to determine the future trend, then machine learning algorithms are the best bet.
- This falls under the paradigm of supervised learning because you already have the data based on which you can train your machines.
- **For example**, a fraud detection model can be trained using a historical record of fraudulent purchases.

# Machine learning for pattern discovery

- If you don't have the parameters based on which you can make predictions, then you need to find out the hidden patterns within the dataset to be able to make meaningful predictions.
- This falls under the paradigm of unsupervised model as you don't have any predefined labels for grouping.
- The most common algorithm used for pattern discovery is Clustering.
  - Let's say you are working in a telephone company and you need to establish a network by putting towers in a region.
  - Then, you can use the clustering technique to find those tower locations which will ensure that all the users receive optimum signal strength.



# Business Intelligence (BI) vs. Data Science

- BI basically analyzes the previous data to find hindsight and insight to describe the business trends.
- BI enables you to take data from external and internal sources, prepare it, run queries on it and create dashboards to answer the questions like quarterly revenue analysis or business problems.
- BI also evaluate the impact of certain events in the near future.
- Data Science is a more forward-looking approach, an exploratory way with the focus on analyzing the past or current data and predicting the future outcomes with the aim of making informed decisions.



# Business Intelligence (BI) vs. Data Science

Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Both Structured and Unstructured ( logs, cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP)
Focus	Past and Present	Present and Future
Tools	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

# Contrast: Databases

	Databases	Data Science
Data Value	“Precious”	“Cheap”
Data Volume	Modest	Massive
Examples	Bank records, Personnel records, Census, Medical records	Online clicks, GPS logs, Tweets, Building sensor readings
Priorities	Consistency, Error recovery, Auditability	Speed, Availability, Query richness
Structured	Strongly (Schema)	Weakly or none (Text)
Properties	Transactions, ACID*	CAP* theorem (2/3), eventual consistency
Realizations	SQL	NoSQL: MongoDB, CouchDB, Hbase, Cassandra, Riak, Memcached, Apache River, ...

ACID = Atomicity, Consistency, Isolation and Durability

CAP = Consistency, Availability, Partition Tolerance

# Contrast: Machine Learning

## Machine Learning

Develop new (individual) models

Prove mathematical properties of models

Improve/validate on a few, relatively clean, small datasets

Publish a paper

## Data Science

Explore many models, build and tune hybrids













Understand empirical properties of models

Develop/use tools that can handle massive datasets

Take action!

Some recent ML Competitions at <https://www.kaggle.com/>

NIST Pre-Pilot Data Science Evaluation – likely to be incorporated to be part of Labs/Final project

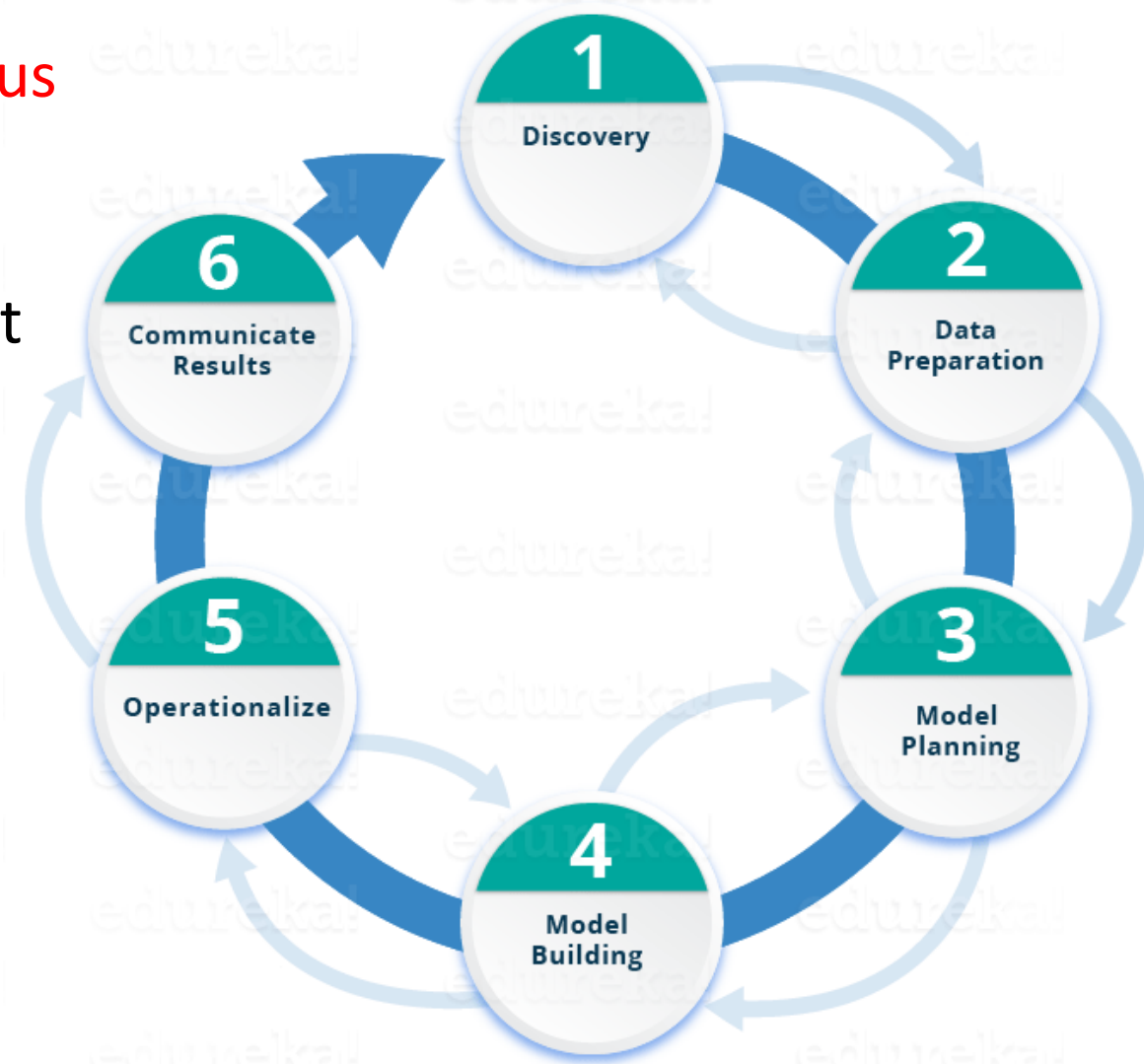
Active Competitions				kaggle	
		<b>Flight Quest 2: Flight Optimization</b> Final Phase of Flight Quest 2	33 days Coming soon \$220,000		
		<b>Packing Santa's Sleigh</b> He's making a list, checking it twice; to fill up his sleigh, he needs your advice	5.8 days 338 teams \$10,000		
		<b>Flu Forecasting</b>  Predict when, where and how strong the flu will be	41 days 37 teams		
		<b>Galaxy Zoo - The Galaxy Challenge</b> Classify the morphologies of distant galaxies in our Universe	2 months 160 teams \$16,000		
		<b>Loan Default Prediction - Imperial College Lon...</b> Constructing an optimal portfolio of loans	52 days 82 teams \$10,000		
		<b>Dogs vs. Cats</b> Create an algorithm to distinguish dogs from cats	11 days 166 teams Swag		

# A common mistake in Data Science Projects

- A common mistake made in Data Science projects is rushing into data collection and analysis, without understanding the requirements or even framing the business problem properly.
- Important for you to follow all the phases throughout the lifecycle of Data Science to ensure the smooth functioning of the project.

# Lifecycle of Data Science

- **Phase 1—Discovery:** Before you begin the project, it is important to **understand various specifications, requirements, priorities and required budget**.
- You must possess the ability to ask the right questions.
- Here, you **assess** if you have the **required resources present** in terms of people, technology, time and data to support the project.
- In this phase, you also need to frame the business problem and formulate initial hypotheses (IH) to test.

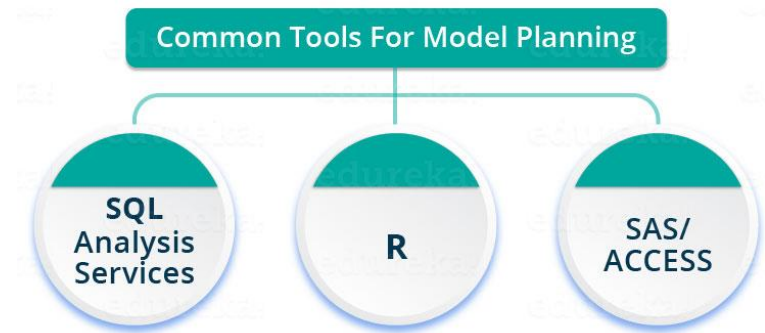


## Phase 2—Data preparation:

- In this phase, you require analytical sandbox in which you can perform analytics for the entire duration of the project.
- You need to explore, and preprocess data prior to modeling.
- Further, you will perform ETLT (extract, transform, load and transform) to get data into the sandbox.
- You can use R for data cleaning, transformation, and visualization.
- This will help you to spot the outliers and establish a relationship between the variables.



## Phase 3—Model planning:



Determine the methods and techniques to draw the relationships between variables.

These relationships will set the base for the algorithms which you will implement in the next phase.

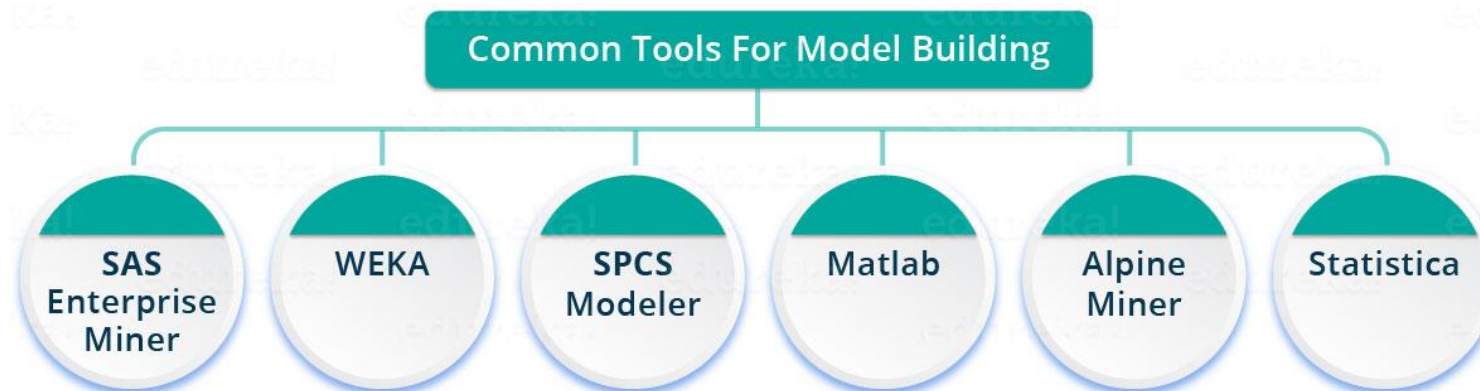
You will apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.

Various model planning tools:

- R has a complete set of modeling capabilities and provides a good environment for building interpretive models.
- SQL Analysis services can perform in-database analytics using common data mining functions and basic predictive models.
- SAS/ACCESS can be used to access data from Hadoop and is used for creating repeatable and reusable model flow diagrams

## Phase 4—Model building:

- In this phase, you will develop datasets for training and testing purposes.
- You will consider whether your existing tools will suffice for running the models or it will need a more robust environment (like fast and parallel processing).
- You will analyze various learning techniques like classification, association and clustering to build the model.
- You can achieve model building through the following tools.



## Phase 5—Operationalize:

- In this phase, you deliver final reports, briefings, code and technical documents.
- In addition, sometimes a pilot project is also implemented in a real-time production environment.
- This will provide you a clear picture of the performance and other related constraints on a small scale before full deployment.

## Phase 6—Communicate results:

- Now it is important to evaluate if you have been able to achieve your goal that you had planned in the first phase.
- In this phase, you identify all the key findings, communicate to the stakeholders and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1.

# WHAT DO DATA SCIENTISTS DO?

**“They need to find nuggets of truth in data and then explain it to the business leaders”**

-- Richard Snee, EMC

**Data scientists “tend to be “hard scientists”, particularly physicists, rather than computer science majors. Physicists have a strong mathematical background, computing skills, and come from a discipline in which survival depends on getting the most from the data. They have to think about the big picture, the big problem.”**

-- DJ Patil, Chief Scientist at LinkedIn

# MIKE DRISCOLL'S THREE SEXY SKILLS OF DATA GEEKS

“data wrangling”  
“data jujitsu”  
“data munging”



## **Data Wrangling**

- parsing, scraping, and formatting data

## **Statistics**

- traditional analysis

## **Visualization**

- graphs, tools, etc.



# DOING DATA SCIENCE (PETER HUBER)

1. **Inspection**
2. **Error checking**
3. **Modification**
4. **Comparison**
5. **Modeling and model fitting**
6. **Simulation**
7. **What-if analyses**
8. **Interpretation**
9. **Presentation of conclusions**

# DOING DATA SCIENCE (BEN FRY)

1. **Acquire**
2. **Parse**
3. **Filter**
4. **Mine**
5. **Represent**
6. **Refine**
7. **Interact**

# DOING DATA SCIENCE (COLIN MALLOWS)

1. **Identify data to collect and its relevance to your problem**
2. **Statistical specification of the problem**
3. **Method selection**
4. **Analysis of method**
5. **Interpret results for non-statisticians**

# A PRACTICAL DEFINITION

**Data Science is about the whole processing pipeline to extract information out of data**

**Data Scientist understand and care about the whole data pipeline**

**A data pipeline consists of 3 steps:**

**1) Preparing to run a model**

Gathering, cleaning, integrating, restructuring, transforming, loading, filtering, deleting, combining, merging, verifying, extracting, shaping

**2) Running the model**

**3) Communicating the results**