Data Science

Dr. Muhammad Nouman Durrani

Introduction to the Course

- Data Science is a dynamic and fast-growing field at the interface of Statistics and Computer Science.
- It is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms (Wikipedia).
- This course will equip students with some of its basic principles and tools including data collection and integration, data cleaning, data analysis using machine learning, visualization and effective communication.
- The main focus of these topics will be on understanding and integration of concepts and their application for solving problems.

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**

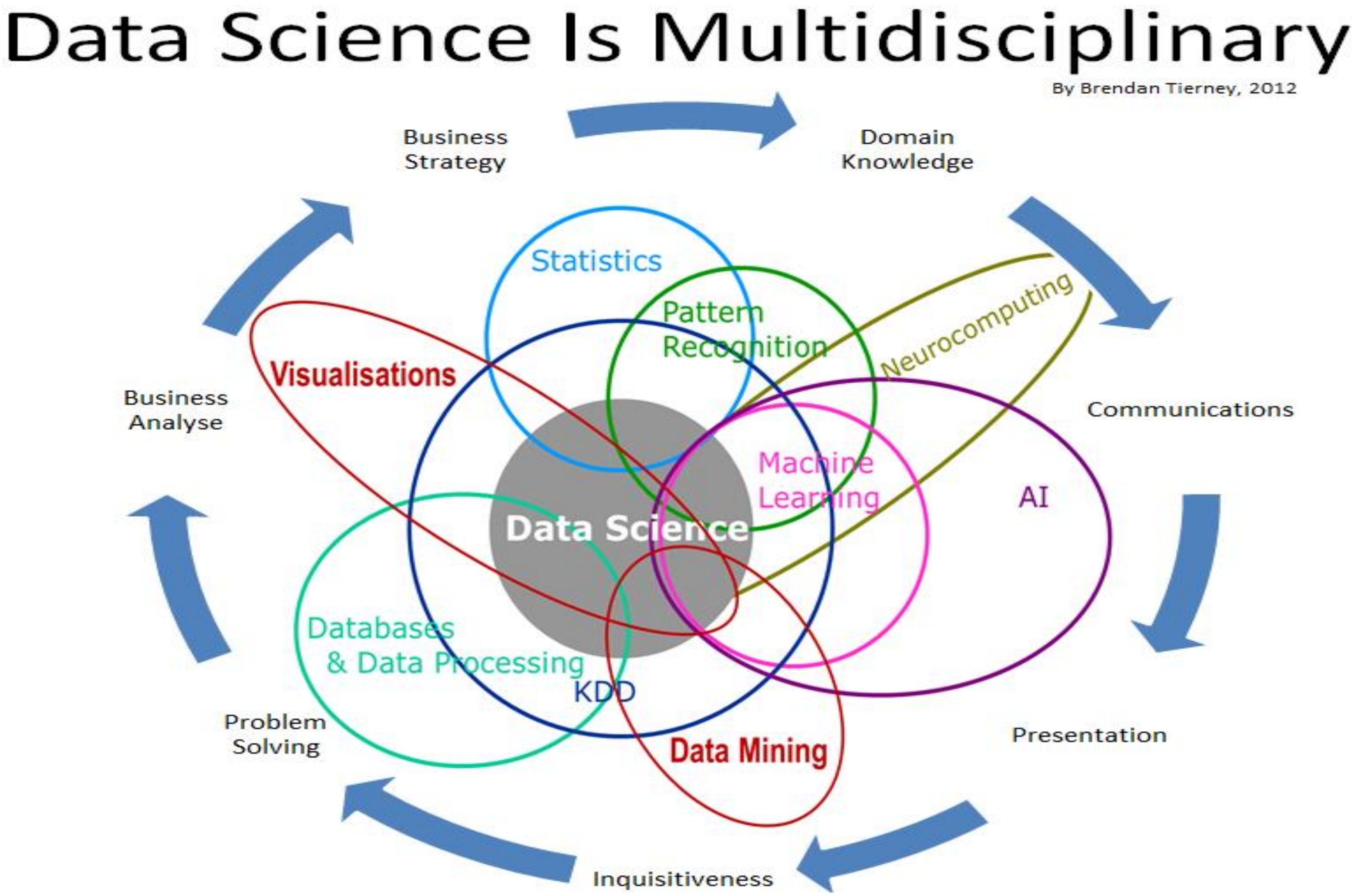
"Hot New Gig in Tech"

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

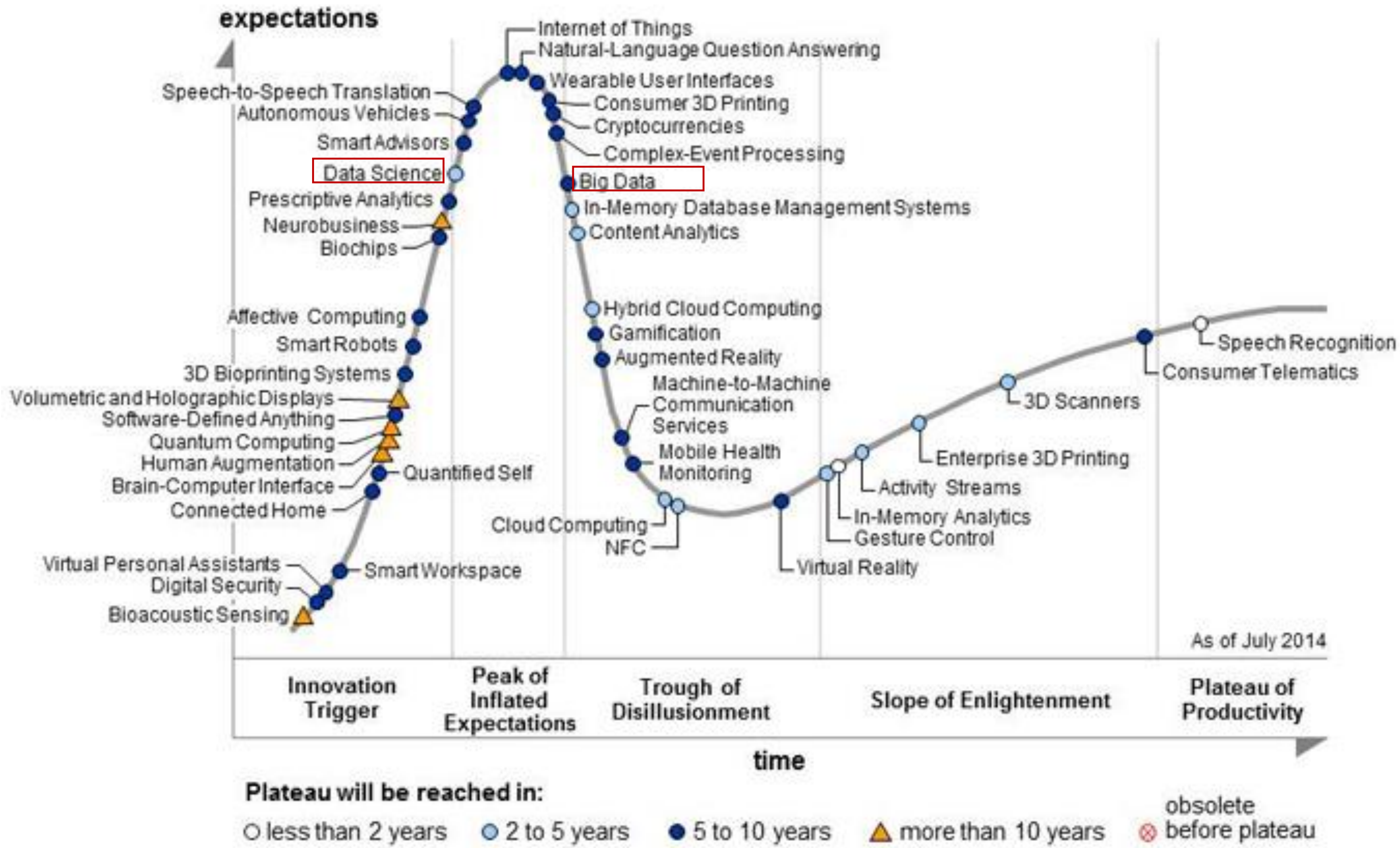
"The next sexy job"

"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."

Data Science

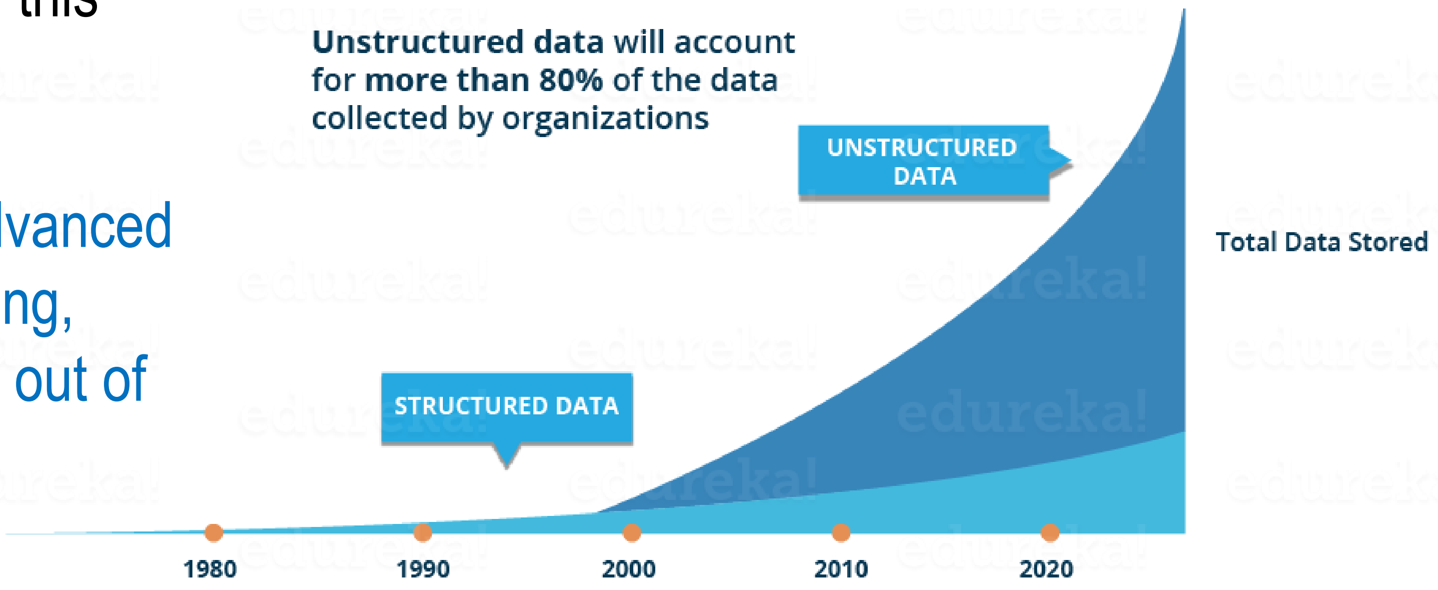


Gartner's 2014 Hype Cycle



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.
- 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



From where the data comes from?

- This data is generated from different sources like:
  - text files, multimedia forms, sensors, and instruments

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

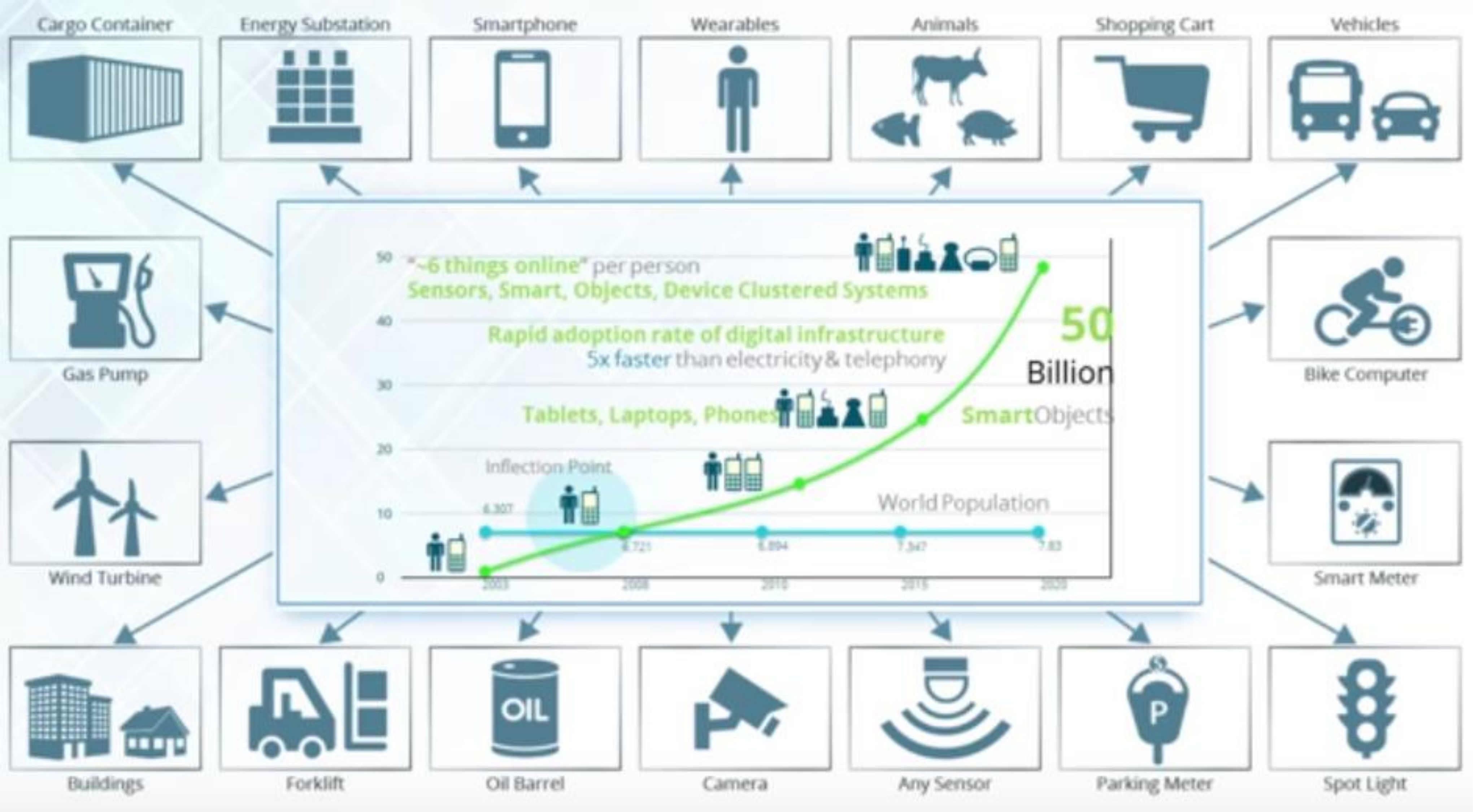
Contributors: Social Networks

Data Generated Every Minute!

Contributors:

- 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: IoTs - 50 Billion Connected Devices by 2020



Contributors: Surveillance guys

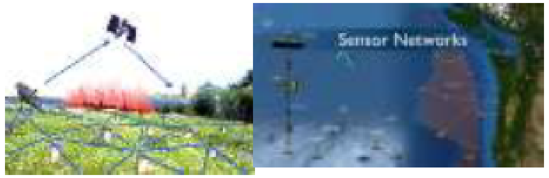
1 VGA resolution color camera produces 800 GB/hour

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

Types of Data We Have

```xml
<note>
 <date>2017-11-08</date>
 <hour>08:30</hour>
 <to>Raj</to>
 <from>Ravi</from>
 <body>Meeting at 8am.</body>
</note>
```

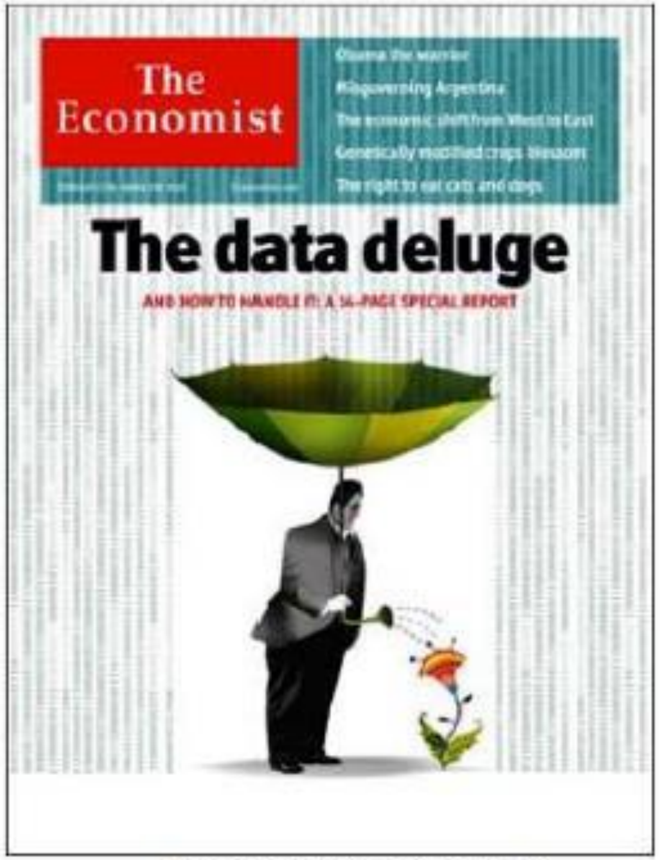
- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
- Social Network, Semantic Web (RDF), …
- Streaming Data

You can afford to scan the data once

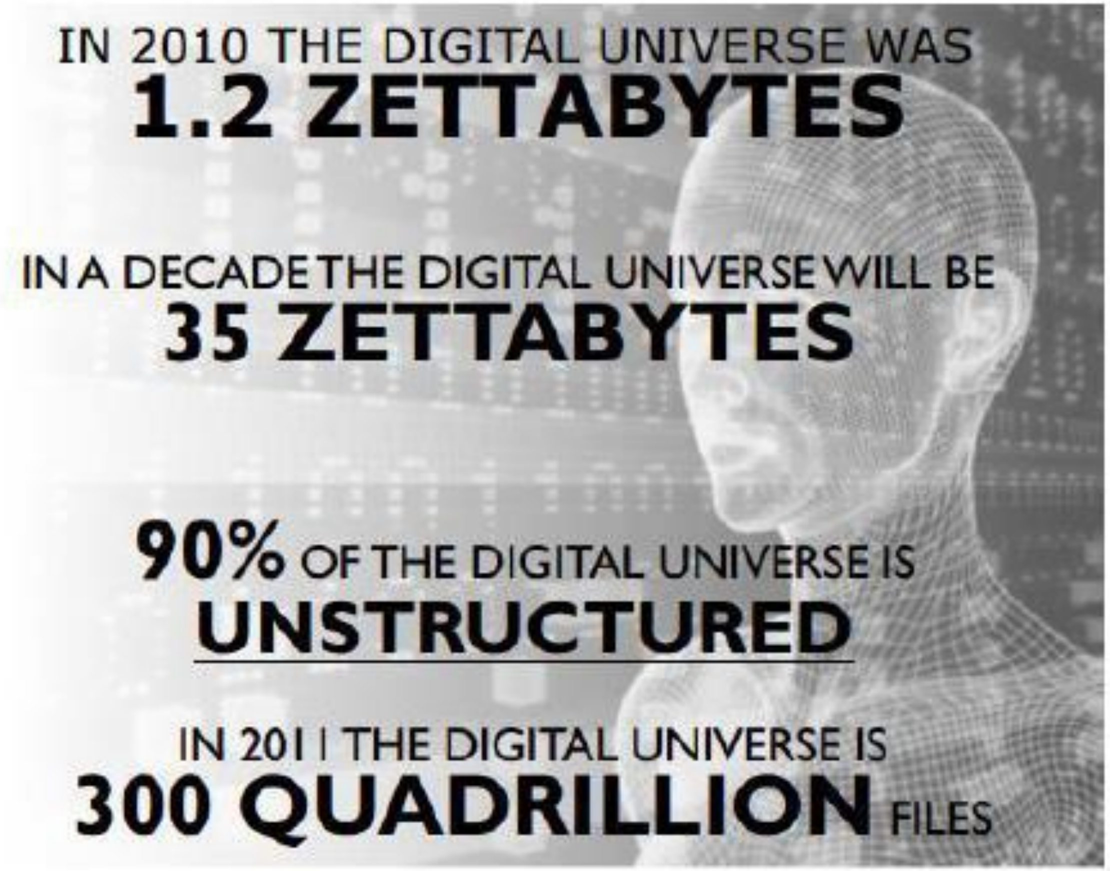
What To Do With These Data?

- Aggregation and Statistics
  - Data warehousing and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling

Customer Challenges: The Data Deluge



The Economist, Feb 25, 2010



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

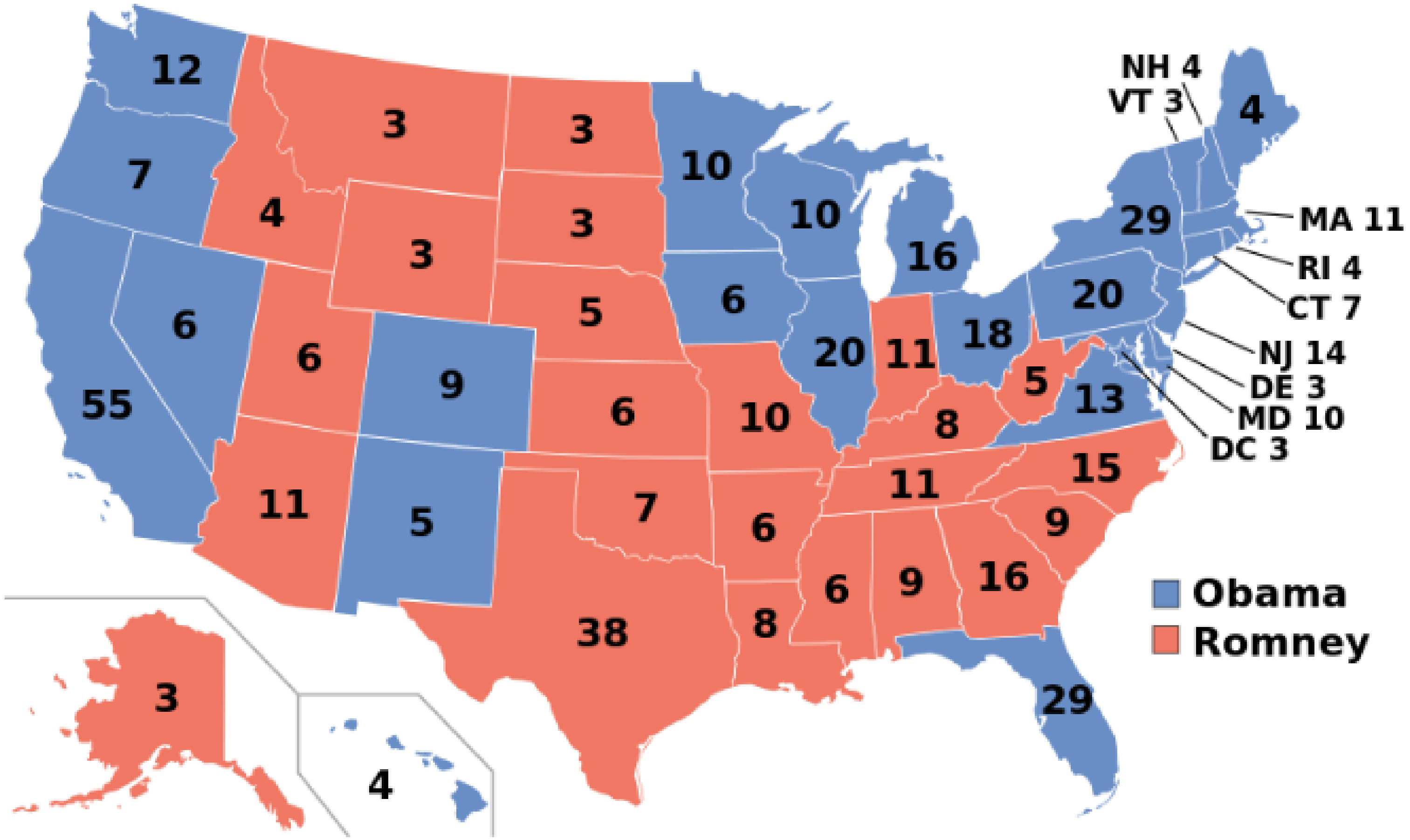
- How about if your car had the intelligence to drive you home?
- The self-driving cars collect live data from sensors, including radars, cameras and lasers to create a map of its surroundings.
- Based on this data, it takes decisions like when to speed up, when to speed down, when to overtake, where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

Polls whiz kid Nate Silver and presidential candidate Donald Trump.

Photo illustration by *Slate*. Images by Slaven Vlasic/Getty Images and Ethan Miller/Getty Images.

"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

EXPRESSION OF EMOTIONS OVER THE 20TH CENTURY

1) Convert all the digitized books in the 20th century into n-grams
   (Thanks, Google!)
   (http://books.google.com/ngrams/)

A 1-gram: "yesterday"
A 5-gram: "analysis is often described as"

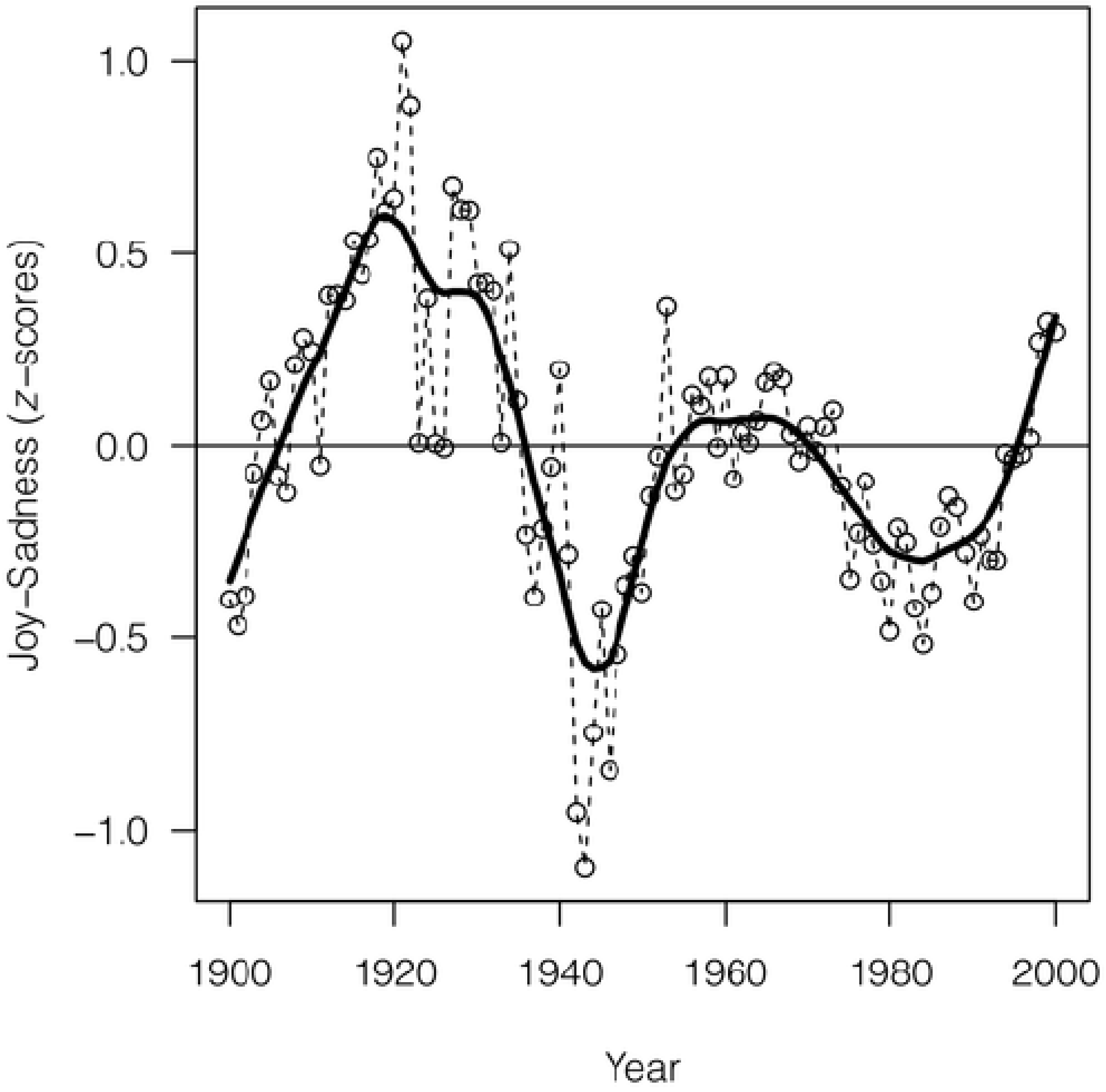
2) Label each 1-gram (word) with a mood score.
   (Thanks, WordNet Affect)

3) Count the occurences of each mood word

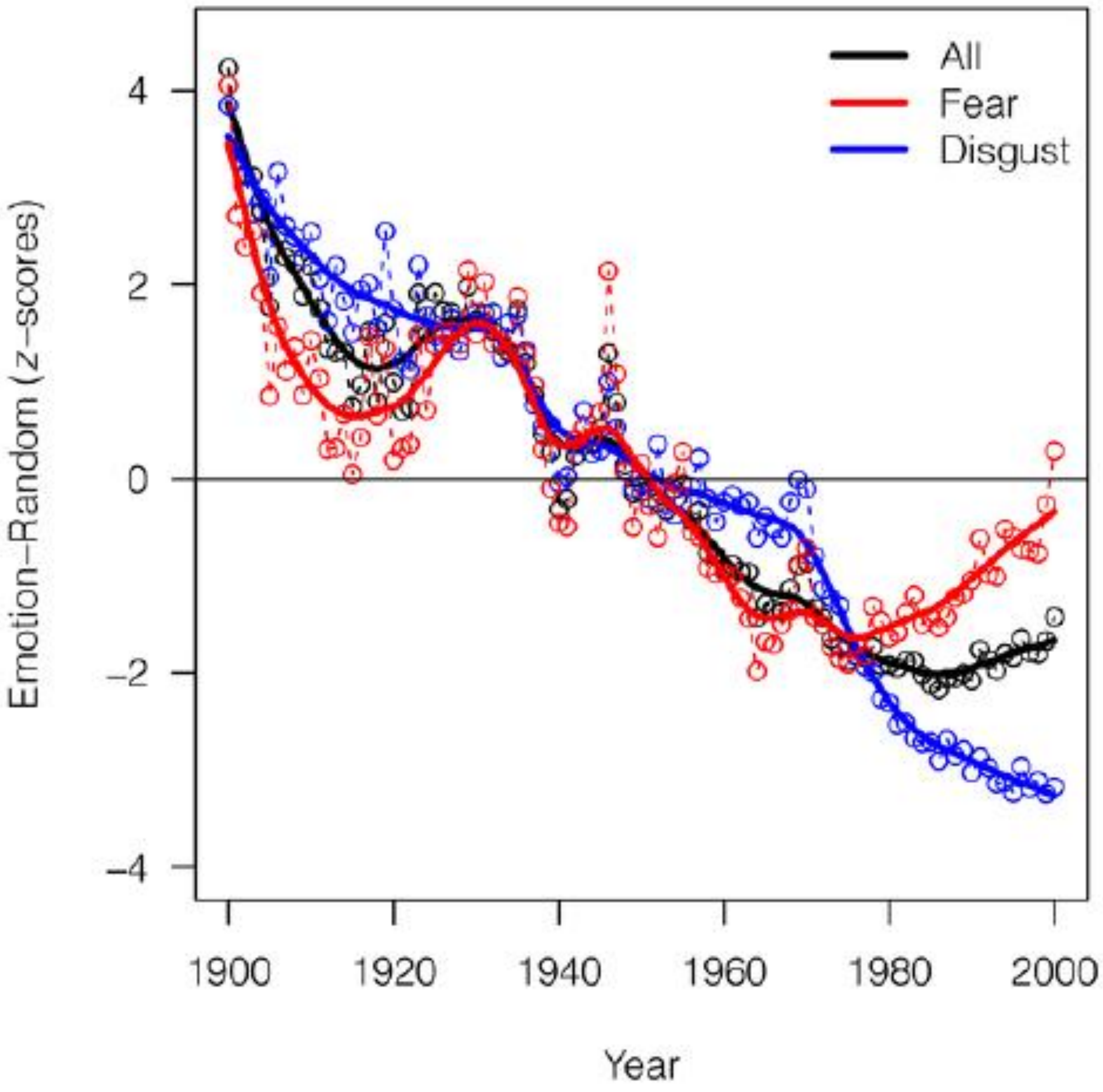
$$\mathcal{M}_Y = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{C_{\text{the}}},$$

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

Acerbi A, Lampos 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



Acerbi A, Lampos 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



Acerbi A, Lampos 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

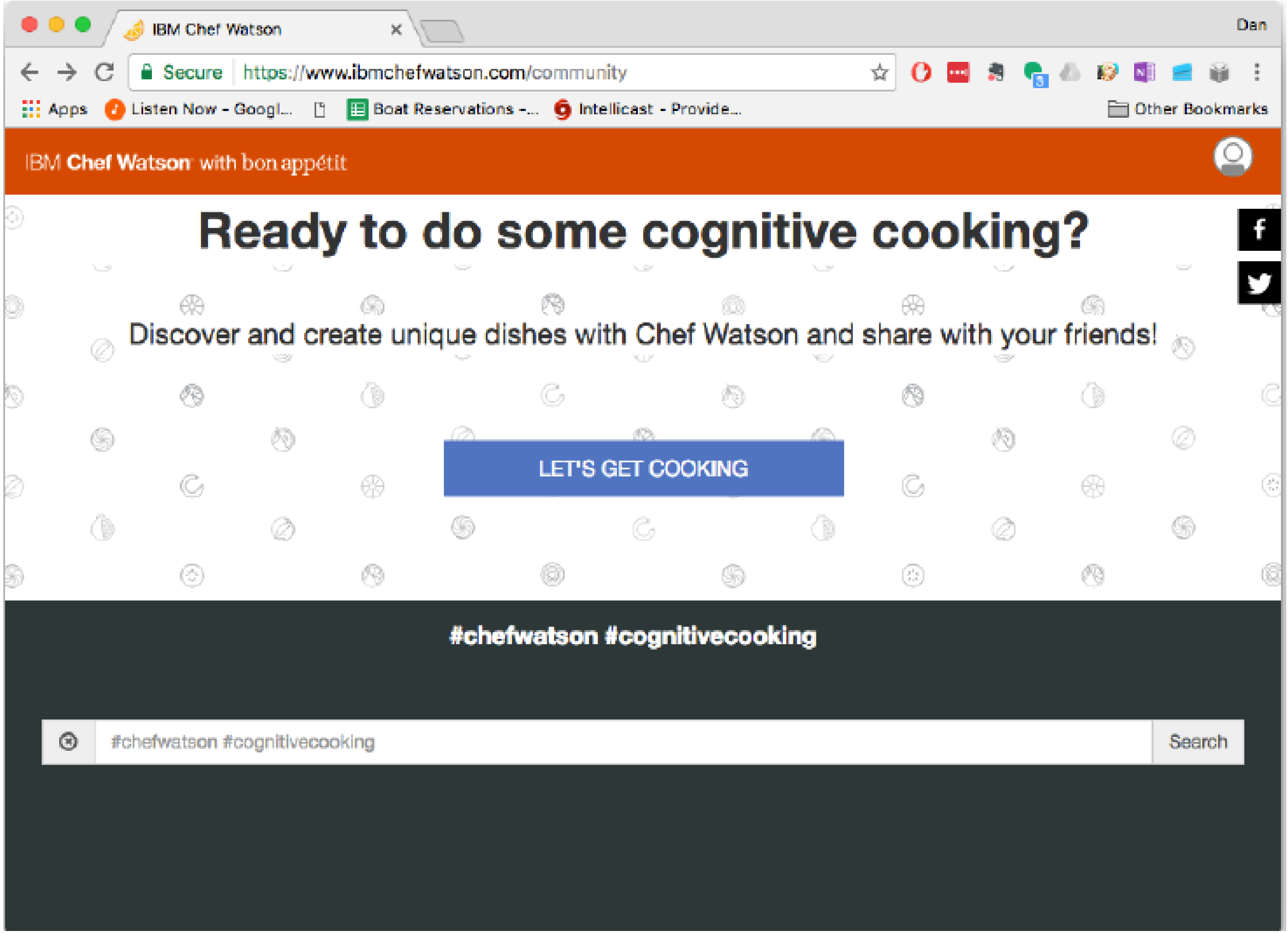
Flavor network and the principles of food pairing

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

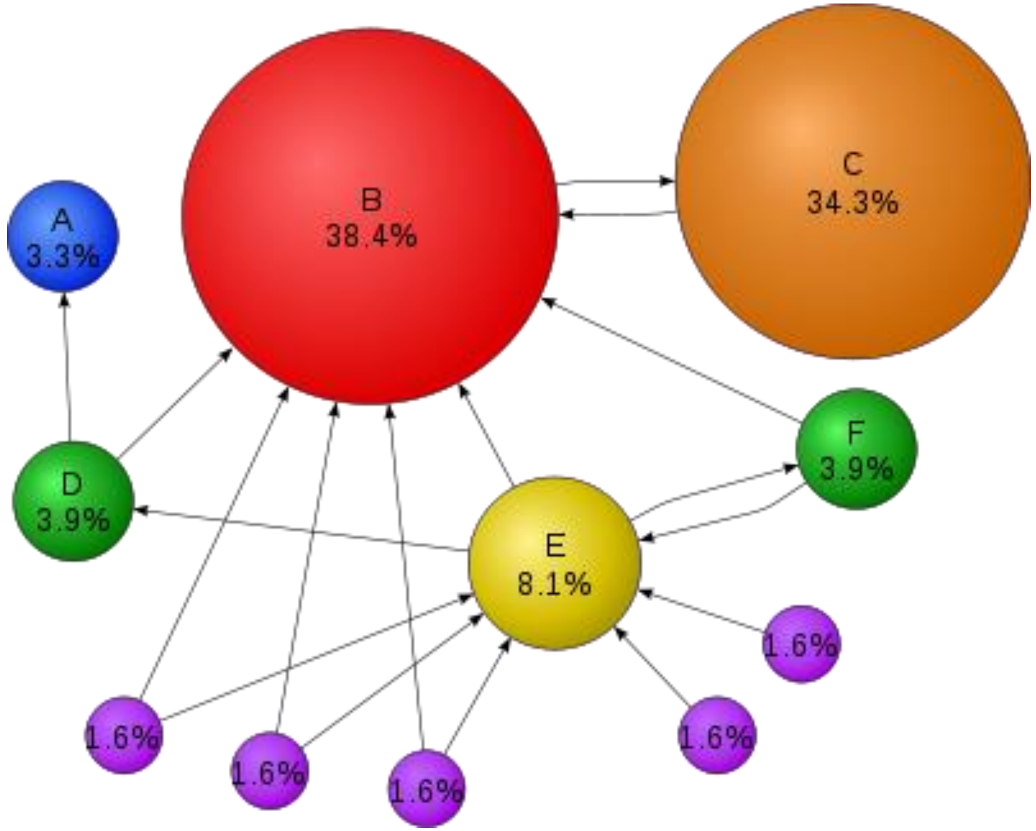
Idea: Analyze the co-occurrence graph of ingredients in recipes to analyze the underlying principles of food pairing.


IBM Chef Watson
Dan
Secure
https://www.ibmchefwatson.com/community
Apps
Listen Now - Googl...
Boat Reservations -...
Intellicast - Provide...
Other Bookmarks
IBM Chef Watson with bon appétit
Ready to do some cognitive cooking?
Discover and create unique dishes with Chef Watson and share with your friends!
LET'S GET COOKING
#chefwatson #cognitivecooking
#chefwatson #cognitivecooking
Search

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



Data Science: Why all the Excitement?

Exciting new effective applications of data analytics

e.g.,
Google Flu Trends:

Detecting outbreaks two weeks ahead of CDC data

New models are estimating which cities are most at risk for spread of the Ebola virus.

Prediction model is built on Various data sources, types and analysis.

ONLINE EXAMPLES

- Cooking with Watson
- https://www.ibmchefwatson.com/community
- Google Flu Data: https://www.google.com/publicdata/explore?ds=z3bsqef7ki44ac_
- 3rd Party Google Flu Data in D3

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

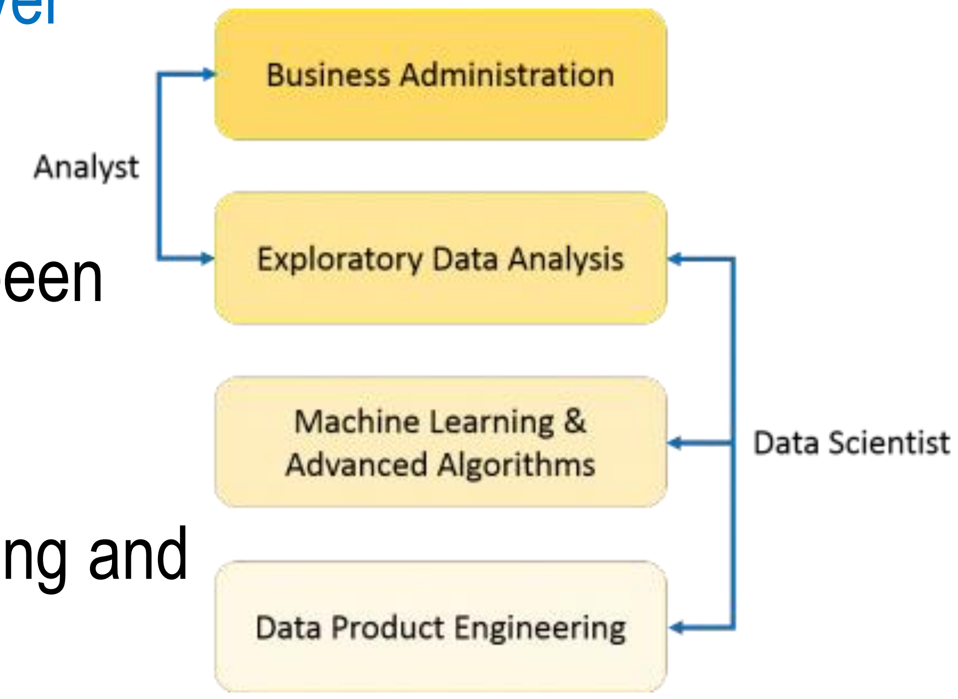
- Global Burden of Disease in D3

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

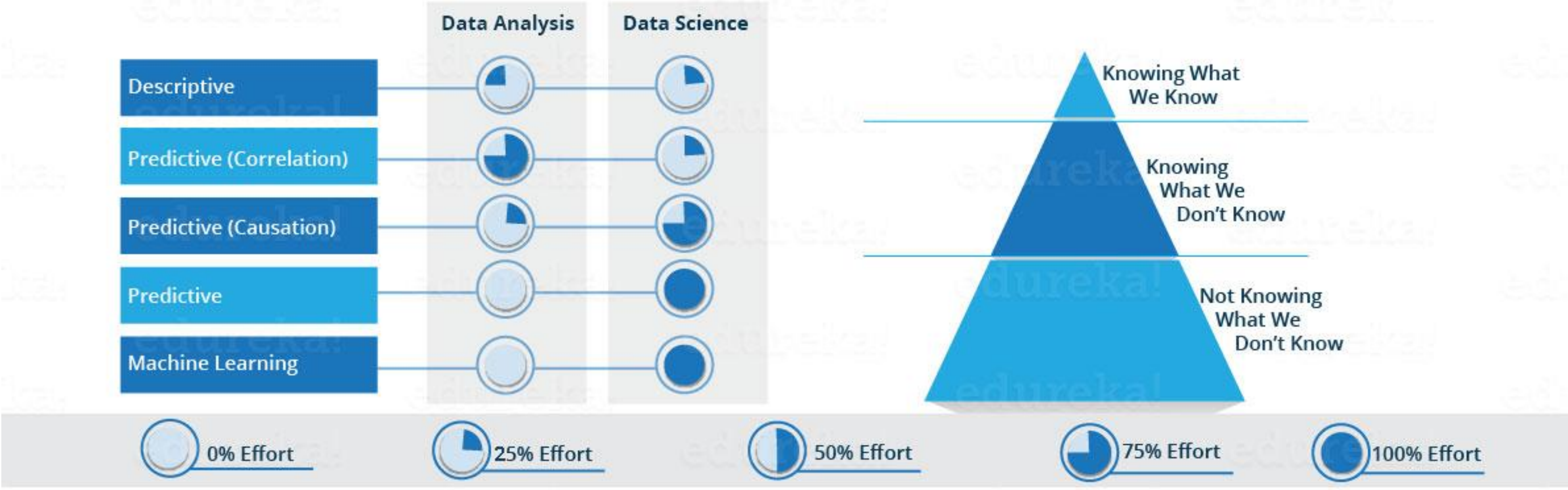
- Machine learning can be defined as the practice of using algorithms to extract data, learn from it, and then forecast future trends for that topic.
- Facebook's machine learning algorithms gather behavioral information for every user on the social platform.
    - Based on one's past behavior, the algorithm predicts interests and recommends articles and notifications on the news feed.
- Similarly, when Amazon recommends products, or when Netflix recommends movies based on past behaviors, machine learning is at work.

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.


Data Analysis
Data Science
Descriptive
Predictive (Correlation)
Predictive (Causation)
Predictive
Machine Learning
Knowing What We Know
Knowing What We Don't Know
Not Knowing What We Don't Know
0% Effort
25% Effort
50% Effort
75% Effort
100% Effort

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

Data Science vs. Machine Learning

- Data science is a field of study that aims to use a scientific approach to extract meaning and insights from data.
- Dr. Thomas Miller of Northwestern University describes data science as "a combination of information technology, modeling, and business management".

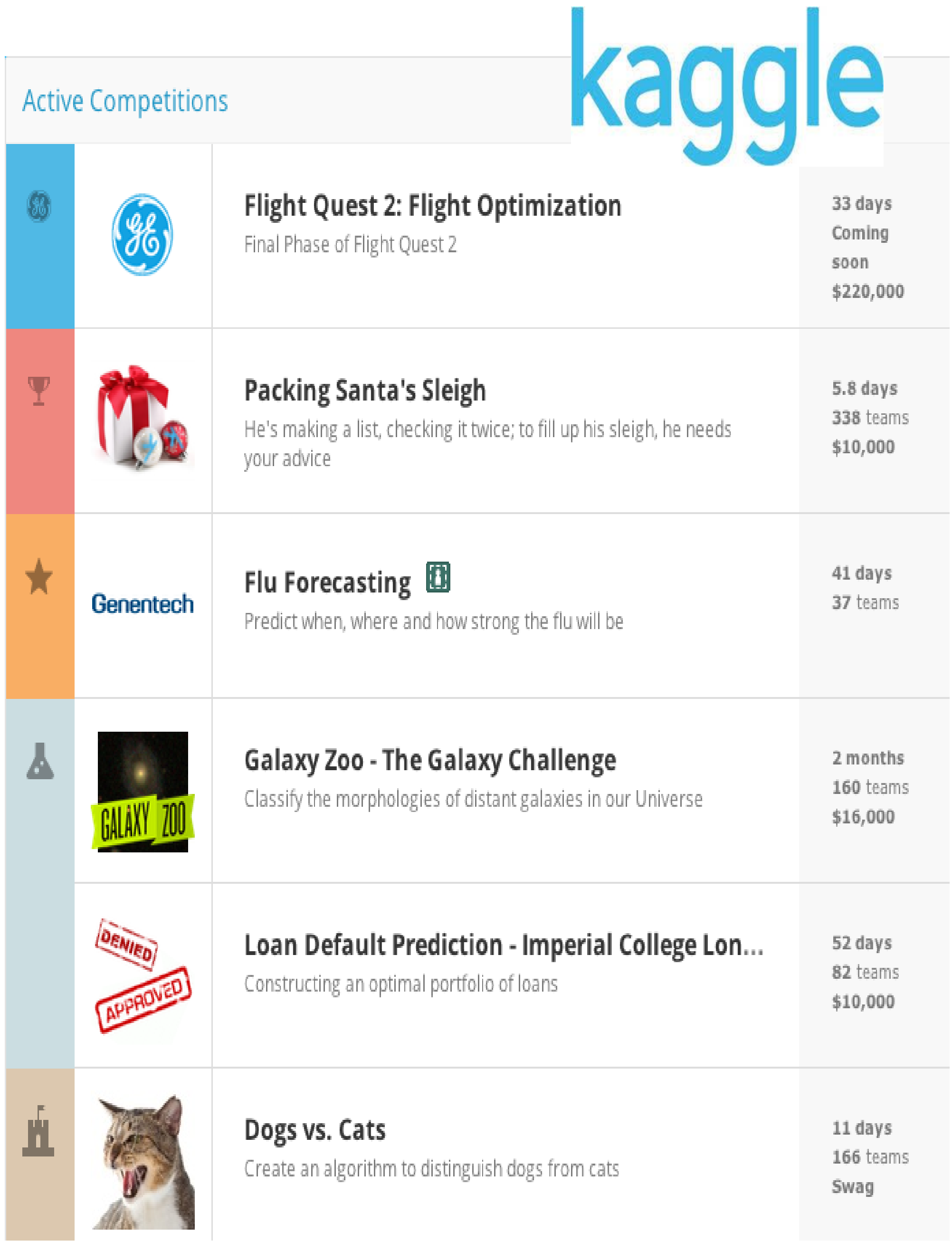
- Machine learning, on the other hand, refers to a group of techniques used by data scientists that allow computers to learn from data.

Contrast: Machine Learning

Machine Learning	Data Science
Develop new (individual) models	Explore many models, build and tune hybrids
Prove mathematical properties of models	Understand empirical properties of models
Improve/validate on a few, relatively clean, small datasets	Develop/use tools that can handle massive datasets
Publish a paper	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



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**