PES UNIVERSITY

Introduction to Data Science

Dr.Mamatha.H.R
Professor
Department of Computer Science and Engineering
PES University
Bangalore

Course material created using various Internet resources

Data?

- Data vs. Information
  - Data
    - Raw facts
    - Distinct pieces of information, usually formatted in a special way
  - Information
    - A collection of facts organized in such a way that they have additional value beyond the value of the facts themselves

Examples

- Data – thermometer readings of temperature taken every hour:

  16.0, 17.0, 16.0, 18.5, 17.0,15.5….

- Information – today's high: 18.5

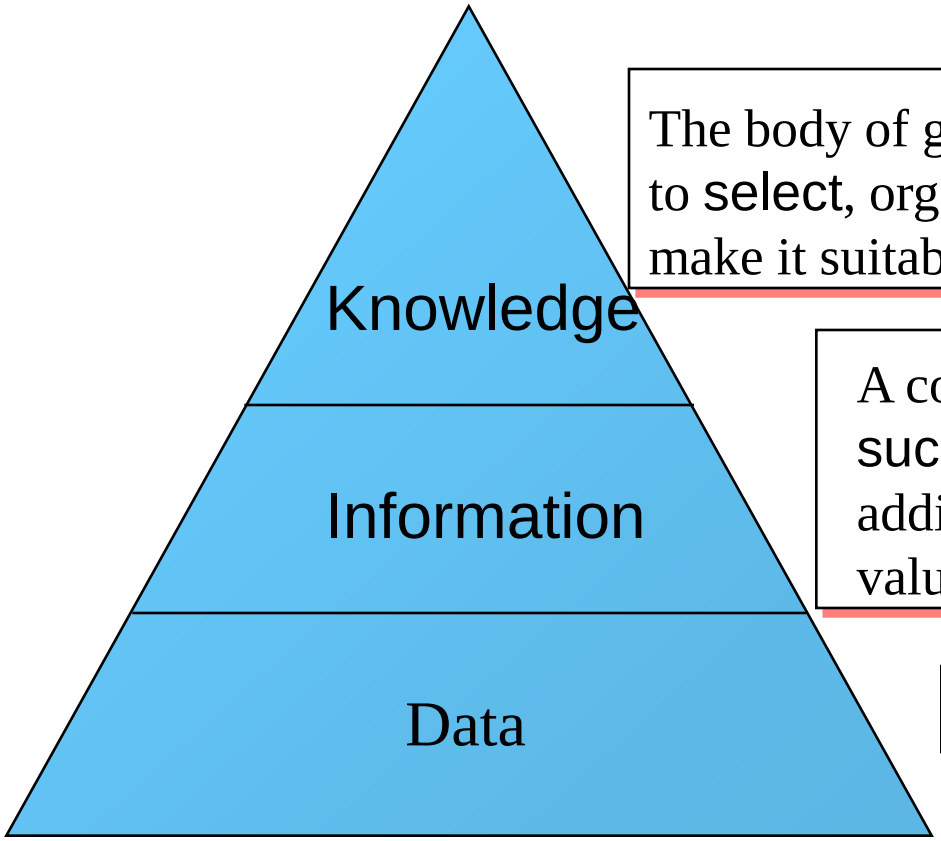
  today's low: 15.5

Types of Data

Data	Represented by
Alphanumeric data	Numbers, letters, and other characters
Image data	Graphic images or pictures
Audio data	Sound, noise, tones
Video data	Moving images or pictures

Data → Information

Information Concepts



The body of guidelines and procedures used to select, organize, and manipulate data to make it suitable for a specific task.

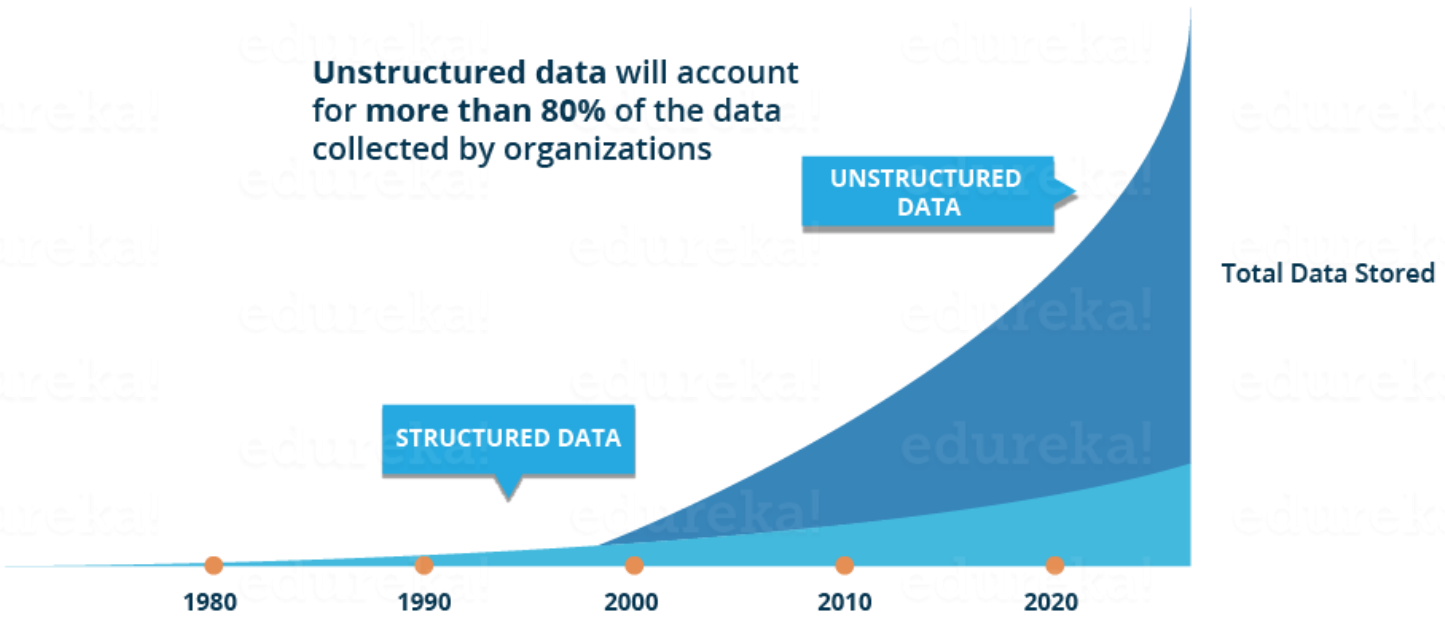
A collection of facts organized in such a way that they have additional value beyond the value of facts themselves.

Raw facts

Science

- Science-latin word Scientia
- Meaning Knowledge
- Science is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions about the universe

Why We Need Data Science



Structured , Unstructured and Semi-structured Data ?

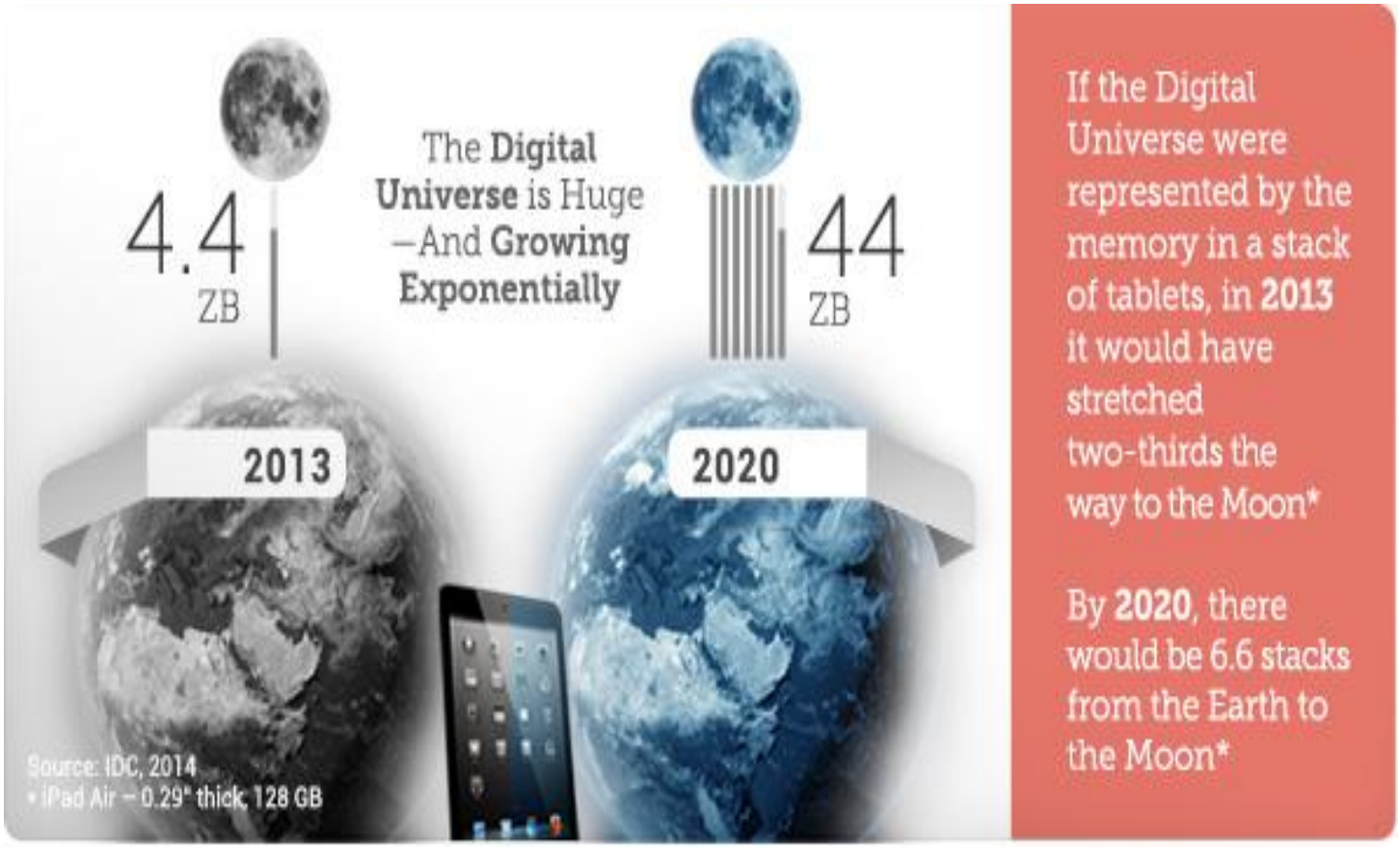
Examples

- relational databases and spreadsheets.
- text and multimedia content. photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.
- XML documents and NoSQL databases.

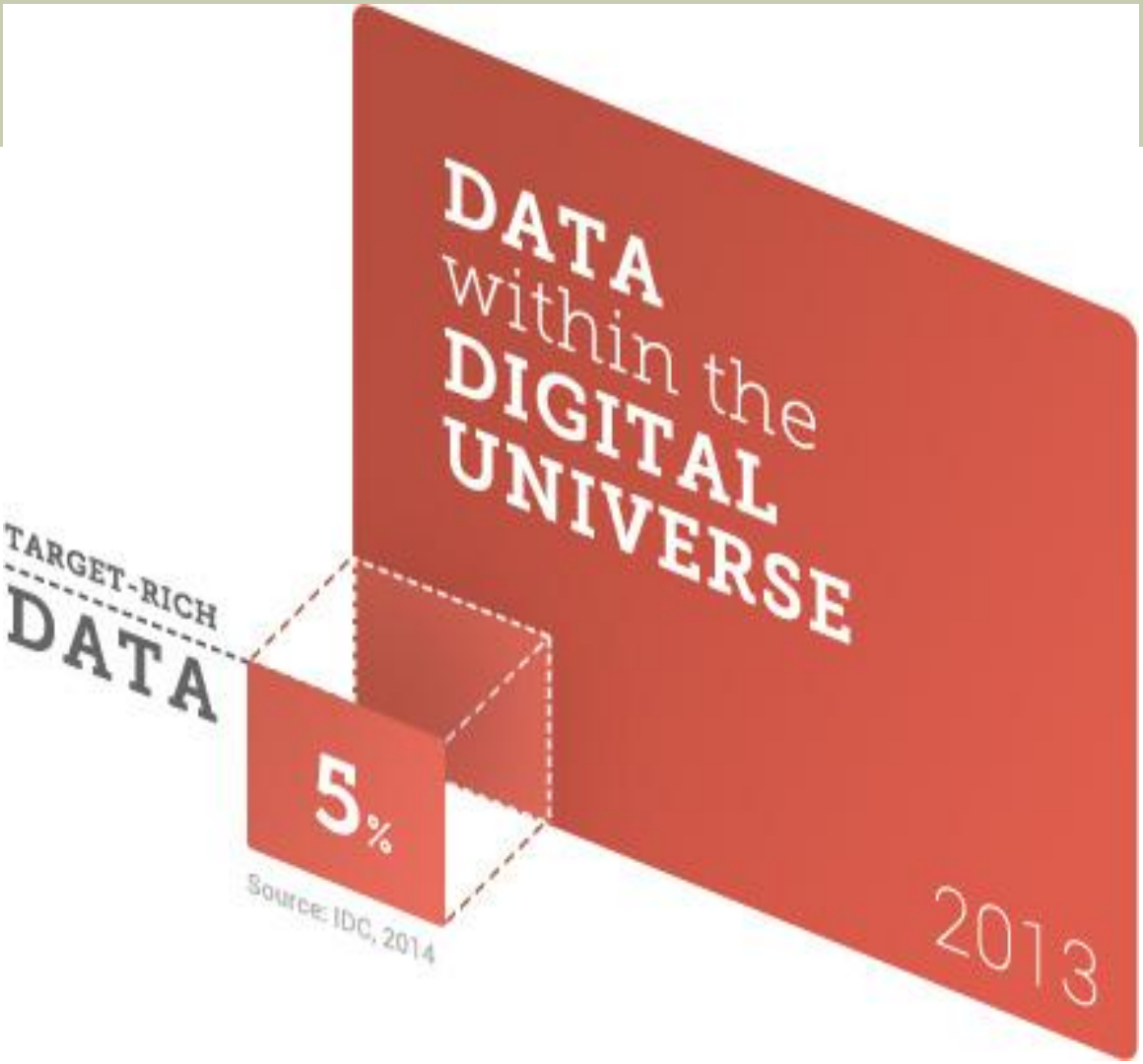
- For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text.
- Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments.
- Photos or other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics.
- XML and other markup languages are often used to manage semi-structured data.

- How much data is generated every year?
- How about the storage?
- Are we using all the data that we are generating?

- Like the physical universe, the digital universe is large – by 2020 containing nearly as many digital bits as there are stars in the universe.


The Digital Universe is Huge –And Growing Exponentially
4.4 ZB
44 ZB
2013
2020
If the Digital Universe were represented by the memory in a stack of tablets, in 2013 it would have stretched two-thirds the way to the Moon*
By 2020, there would be 6.6 stacks from the Earth to the Moon*
Source: IDC, 2014
* iPad Air – 0.29" thick, 128 GB

- In terms of sheer volume, 1.8 zettabytes of data is equivalent to:
- Every person in the United States tweeting 3 tweets per minute for 26,976 years nonstop
- Every person in the world having over 215 million high-resolution MRI scans per day
- Over 200 billion HD movies (each 2 hours in length)—would take 1 person 47 million years to watch every movie 24x7
- The amount of information needed to fill 57.5 billion 32GB Apple iPads. With that many iPads we could:
  - Create a wall of iPads, 4,005-miles long and 61-feet high extending from Anchorage, Alaska to Miami, Florida.
  - Build the Great iPad Wall of China—at twice the average height of the original
  - Build a 20-foot high wall around South America
  - Cover 86% of Mexico City
  - Build a mountain 25-times higher than Mt. Fuji


DATA
within the
DIGITAL
UNIVERSE
TARGET-RICH
DATA
5%
Source: IDC, 2014
2013

- In 2014, the digital universe will equal 1.7 megabytes a minute for every person on Earth.

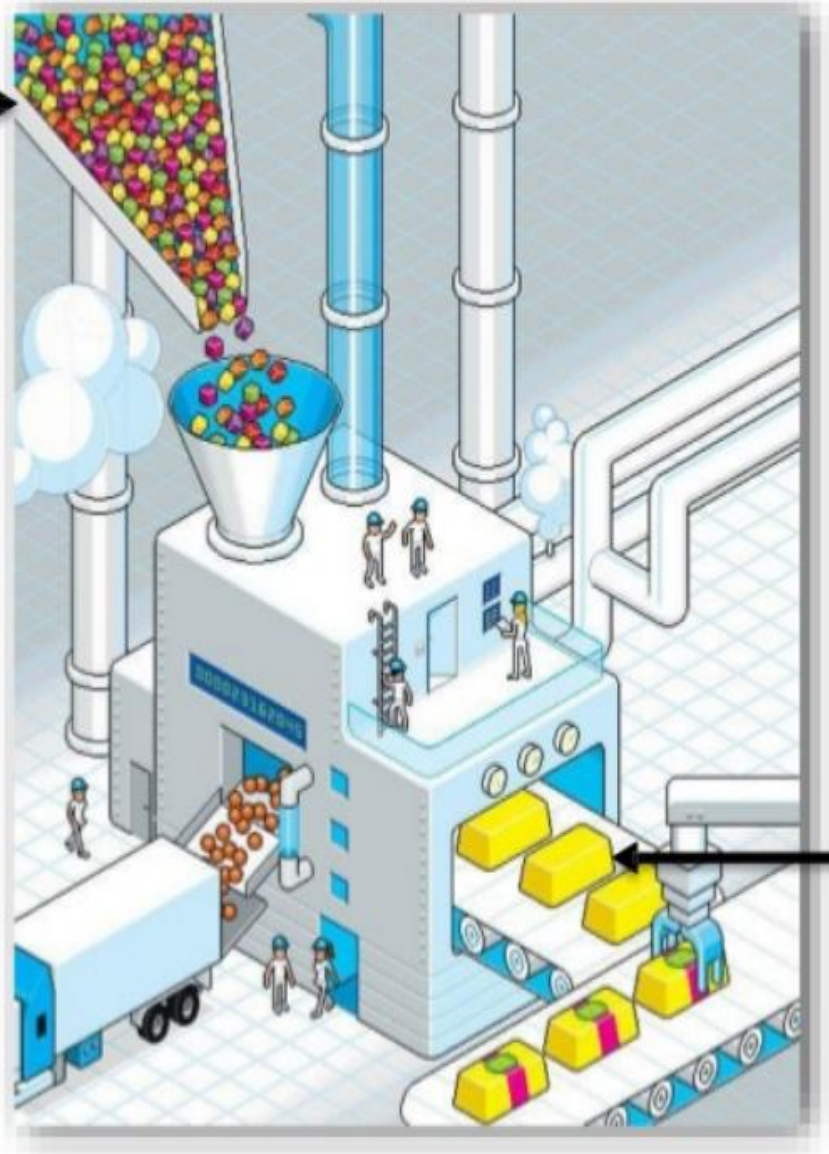
- Sophisticated quantitative analysis is being applied to many aspects of life, not just missile trajectories or financial hedging strategies, as in the past.
- For example, Farecast, a part of Microsoft's search engine Bing, can advise customers whether to buy an airline ticket now or wait for the price to come down by examining 225 billion flight and price records. The same idea is being extended to hotel rooms, cars and similar items.

- WHEN the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in its first few weeks than had been amassed in the entire history of astronomy. Now, a decade later, its archive contains a whopping 140 terabytes of information. A successor, the Large Synoptic Survey Telescope, due to come on stream in Chile in 2016, will acquire that quantity of data every five days.
- Such astronomical amounts of information can be found closer to Earth too. Wal-Mart, a retail giant, handles more than 1m customer transactions every hour, feeding databases estimated at more than 2.5 petabytes—the equivalent of 167 times the books in America's Library of Congress.
- Facebook, a social-networking website, is home to 40 billion photos. And decoding the human genome involves analysing 3 billion base pairs—which took ten years the first time it was done, in 2003, but can now be achieved in one week.

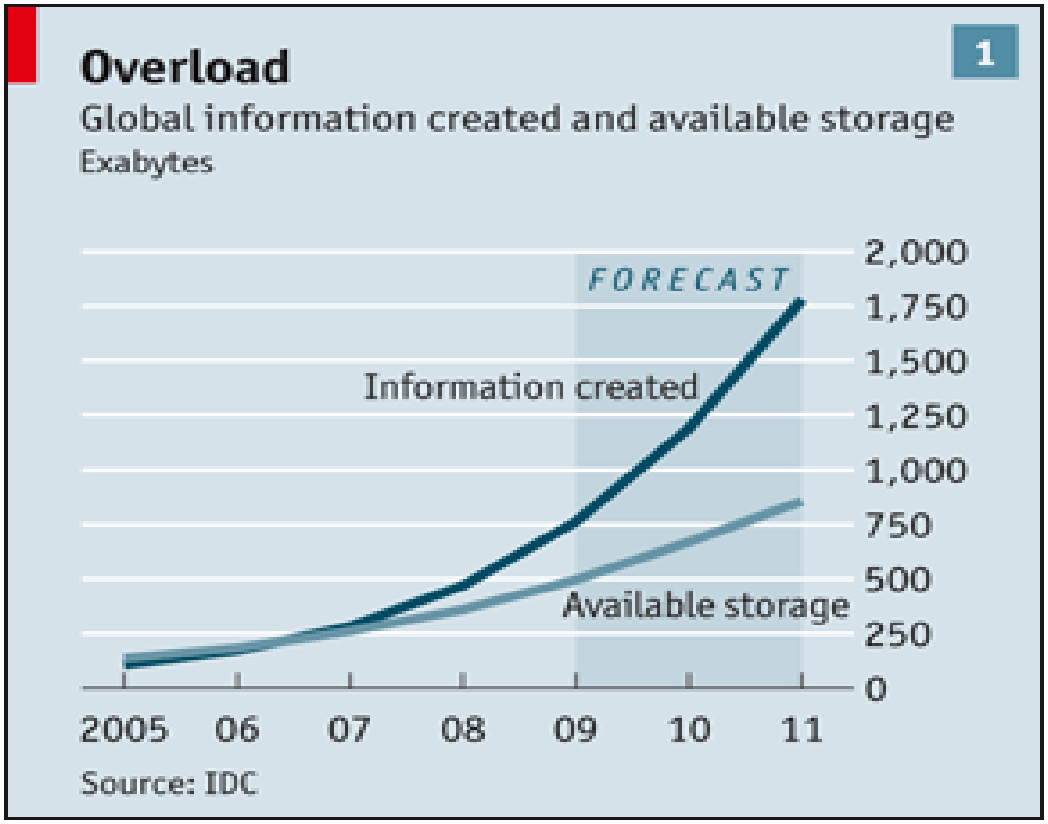

Ellen DeGeneres
@TheEllenShow
Follow
If only Bradley's arm was longer. Best photo ever. #oscars pic.twitter.com/C9U5NOtGap
Reply
Retweet
Favorite
More
RETWEETS
3,411,626
FAVORITES
1,987,141
7:06 PM - 2 Mar 2014
Flag media
Copyright Twitter.com

- Like the physical universe, it is diverse – created by everyone using a digital camera, by the more than 2 billion people and millions of enterprises living their lives and doing their work online, and by the millions of sensors and communicating devices sending and receiving data over the Internet
- . It includes Oscars-host Ellen DeGeneres' "celeb selfie" tweet that was viewed 26 million times across the Web during a 12-hour period; the more than one billion hours of TV shows and movies streamed from Netflix per month; the data collected by sensors connected to a giant gas turbine and its analysis, making electricity cheaper and cleaner; and the data streaming at 2.8 Gigabytes per second from the Australian Square Kilometer Array Pathfinder (ASKAP) radio telescope.

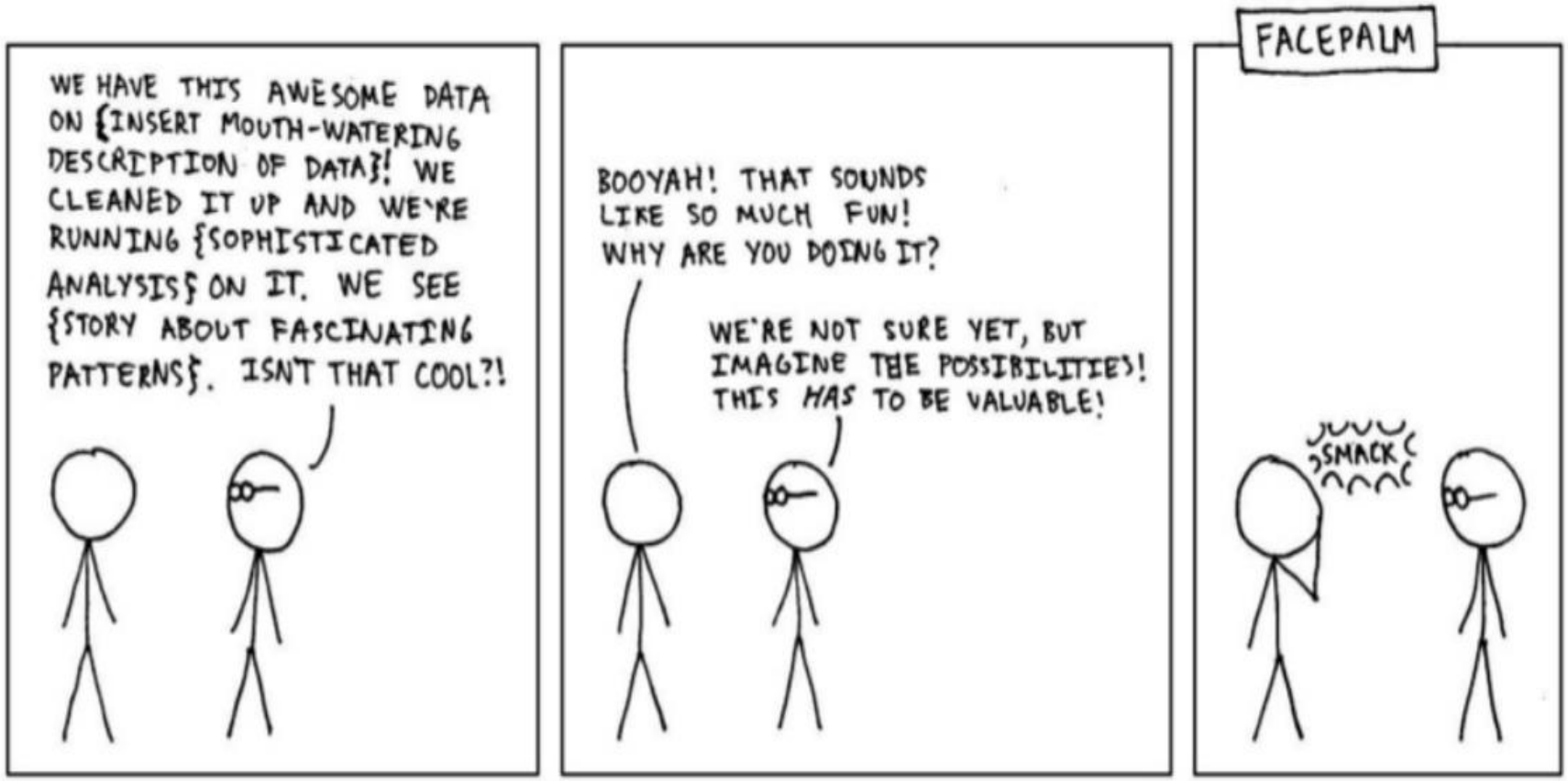
Lots of Data (Terabytes or Petabytes)



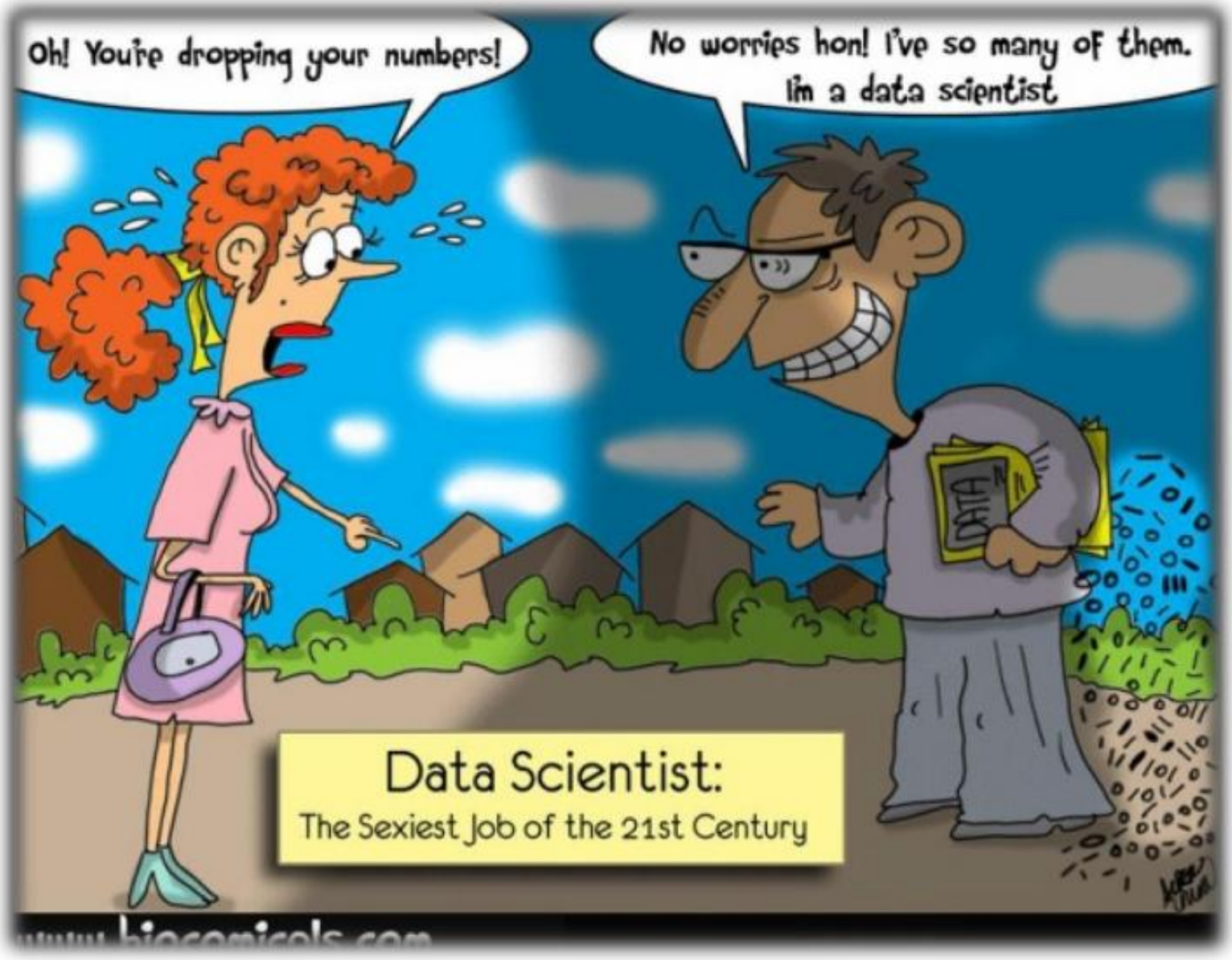
Systems/Enterprises generate huge amount of data from Terabytes to and even Petabytes of information

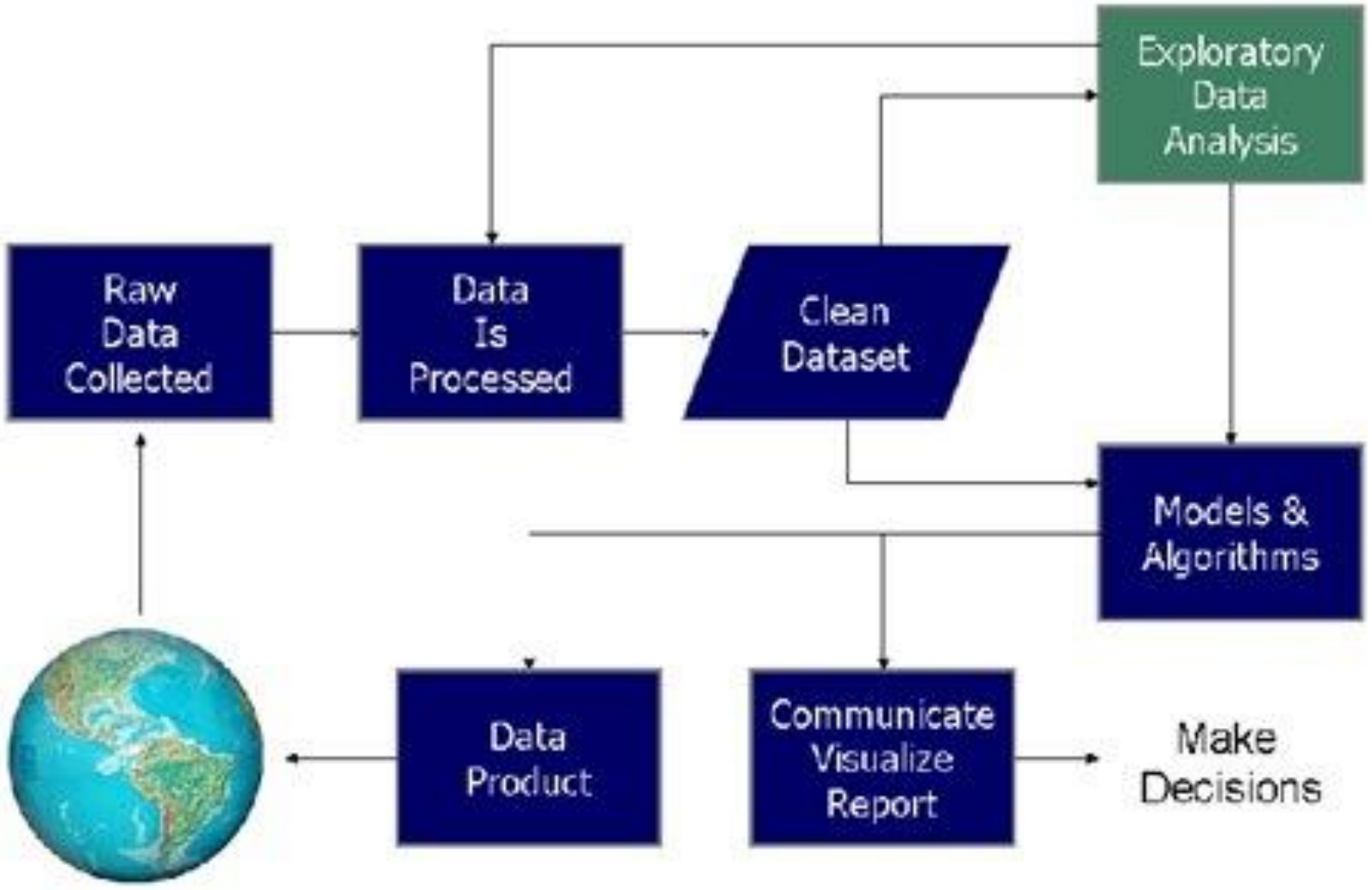

Overload
1
Global information created and available storage
Exabytes
FORECAST
Information created
Available storage
2,000
1,750
1,500
1,250
1,000
750
500
250
0
2005
06
07
08
09
10
11
Source: IDC

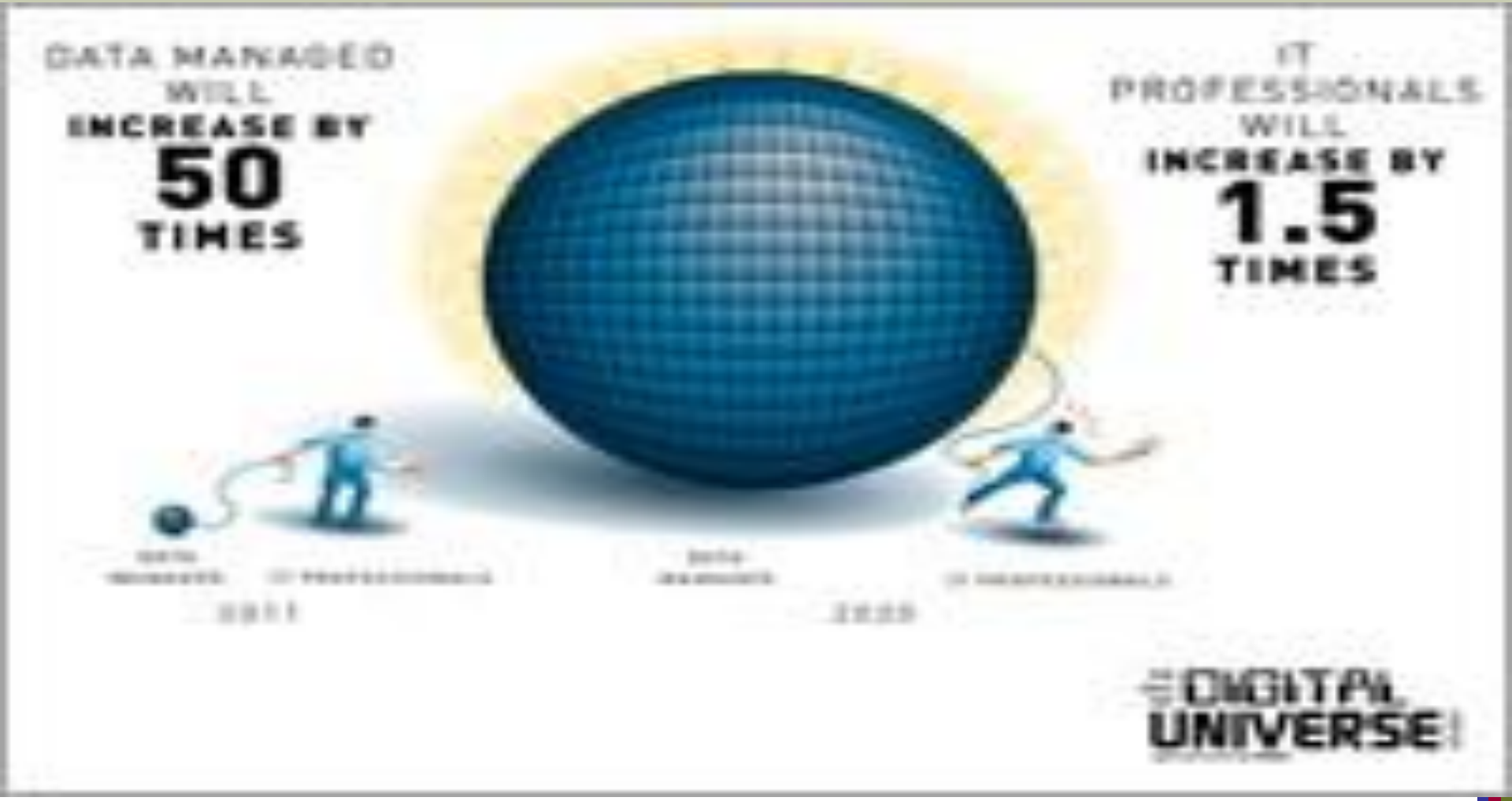
No one knows how to use it



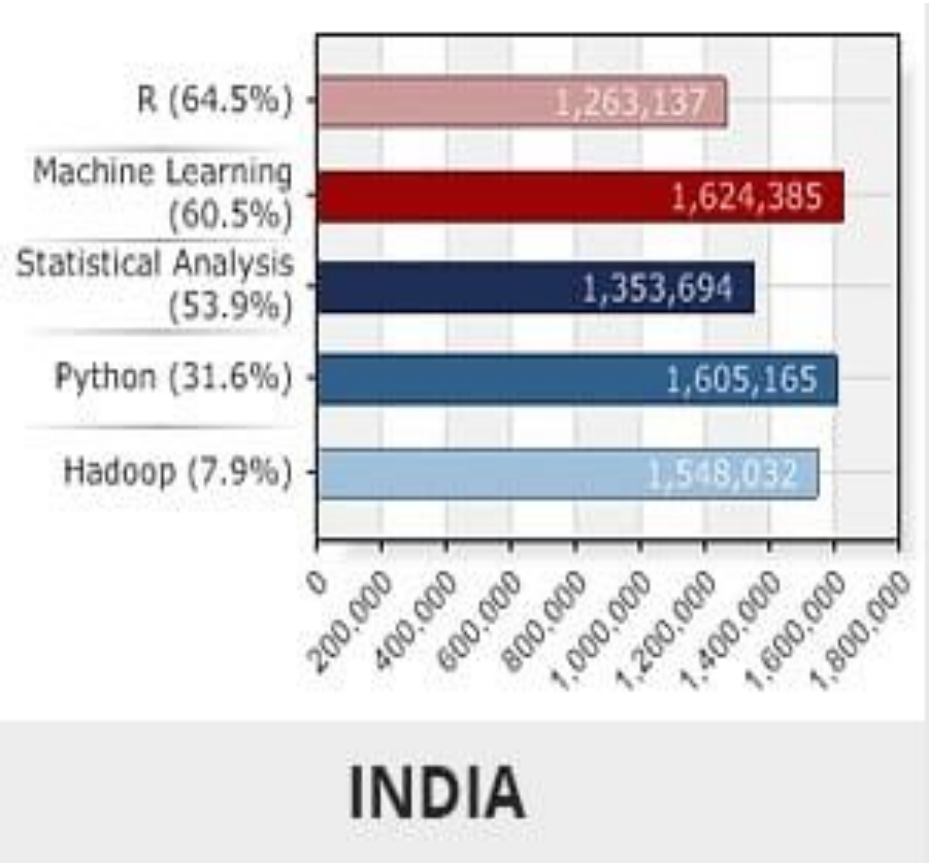

22%
Data that is Useful if Tagged & Analyzed
Source IDC, 2014
37%
2013
2020


Oh! You're dropping your numbers!
No worries hon! I've so many of them. I'm a data scientist
DATA
Data Scientist:
The Sexiest Job of the 21st Century

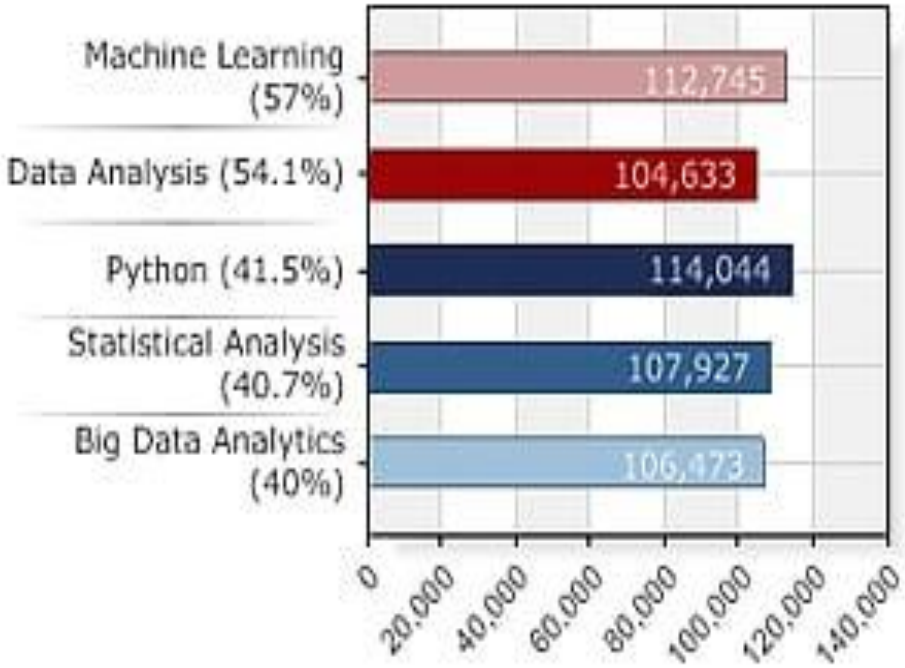

Exploratory Data Analysis
Raw Data Collected
Data Is Processed
Clean Dataset
Models & Algorithms
Data Product
Communicate Visualize Report
Make Decisions


DATA MANAGED
WILL
INCREASE BY
50
TIMES
IT
PROFESSIONALS
WILL
INCREASE BY
1.5
TIMES
THE DIGITAL UNIVERSE

"Data Scientist" Job Trends

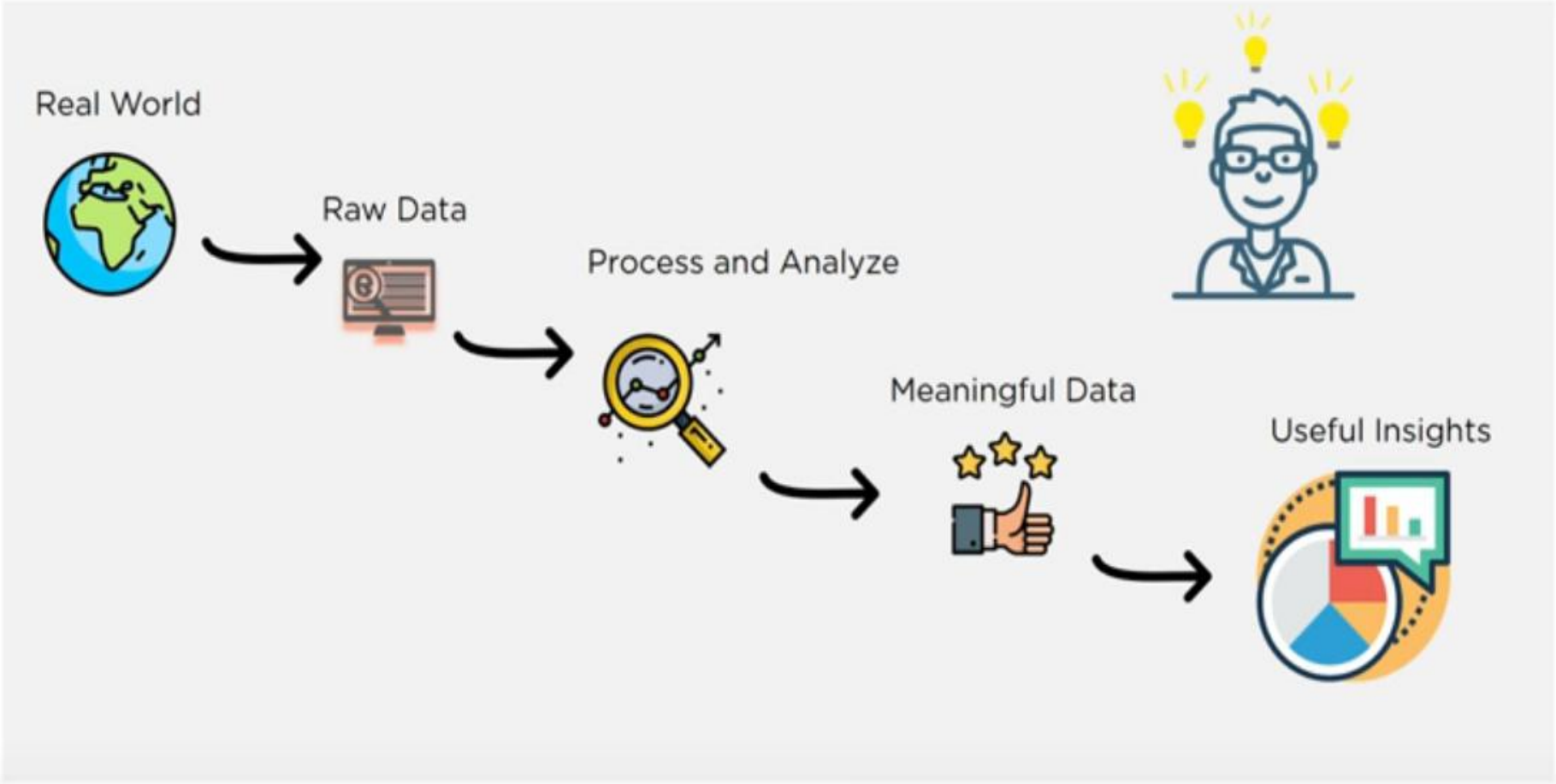

R (64.5%)
1,263,137
Machine Learning (60.5%)
1,624,385
Statistical Analysis (53.9%)
1,353,694
Python (31.6%)
1,605,165
Hadoop (7.9%)
1,548,032
0
200,000
400,000
600,000
800,000
1,000,000
1,200,000
1,400,000
1,600,000
1,800,000

INDIA


Machine Learning (57%)
112,745
Data Analysis (54.1%)
104,633
Python (41.5%)
114,044
Statistical Analysis (40.7%)
107,927
Big Data Analytics (40%)
106,473
0
20,000
40,000
60,000
80,000
100,000
120,000
140,000

US

What does a Data Scientist do?



Dear Flyer, We regret to inform you that your flight has been cancelled due to delay from Airbus on account of engine delivery

1. Due to lack of data available, flights are often delayed or cancelled at the last minute
2. Due to improper route planning, customers don't get the flight for desired time and duration
3. Incorrect decisions in selection of right equipment leads to unplanned delays and cancellations

simplilearn

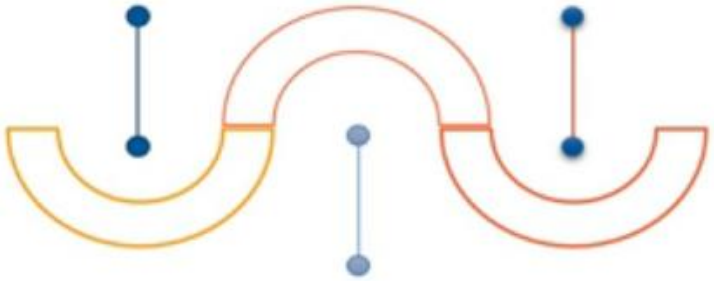
Using Data Science, we can achieve the following:

- Route Planning: Whether to schedule direct or connecting flights
- Predictive analytics model can be built to foresee flight delays
- Promotional offers depending on customer booking patterns
- Deciding which class of planes to purchase for better performance

Logistics companies like FedEx are using Data Science models for operational efficiency

Discover the best routes to ship

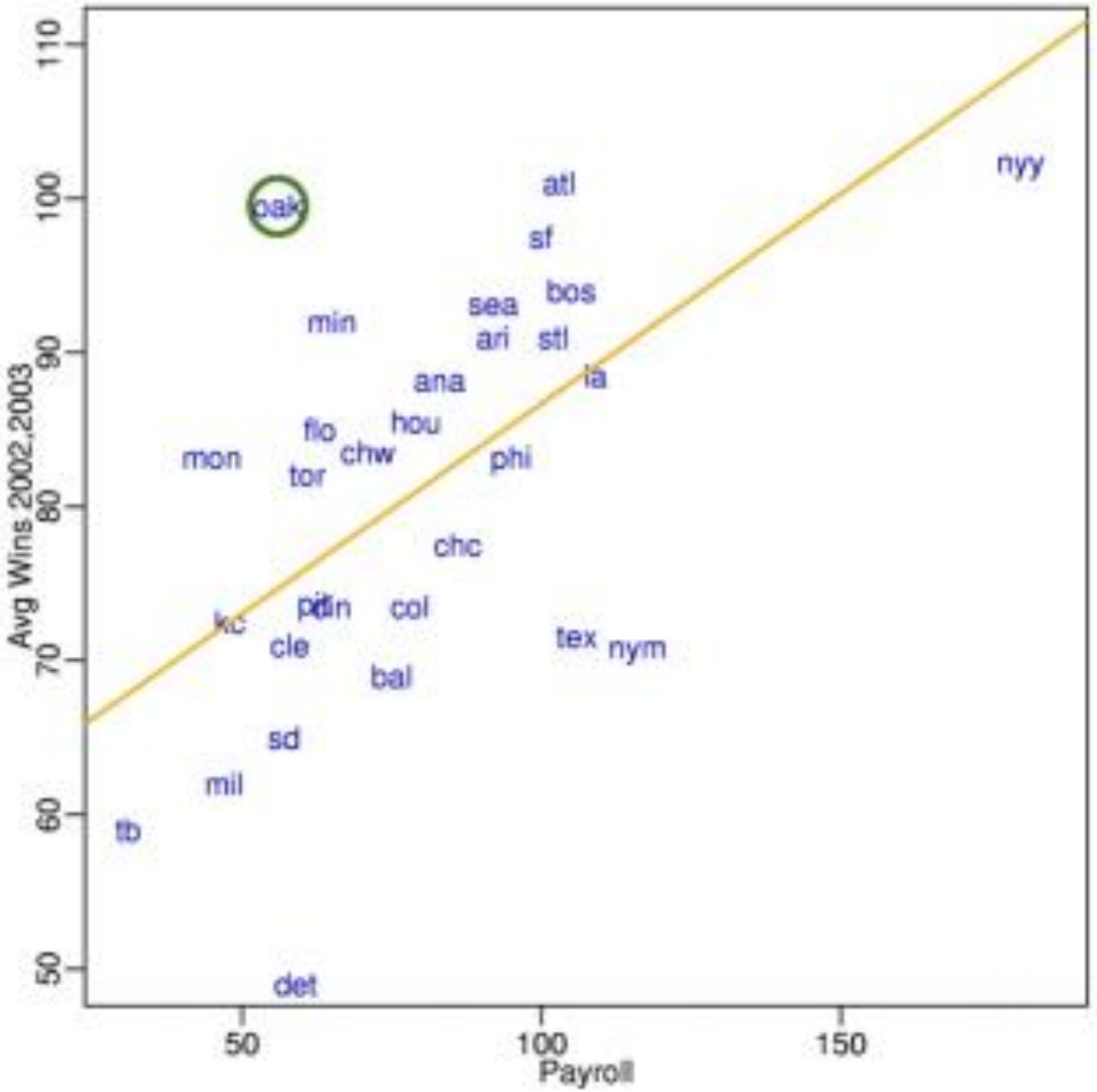
The best suited time to deliver



The best mode of transport

Data can do a lot: sports

Data can do a lot: sports



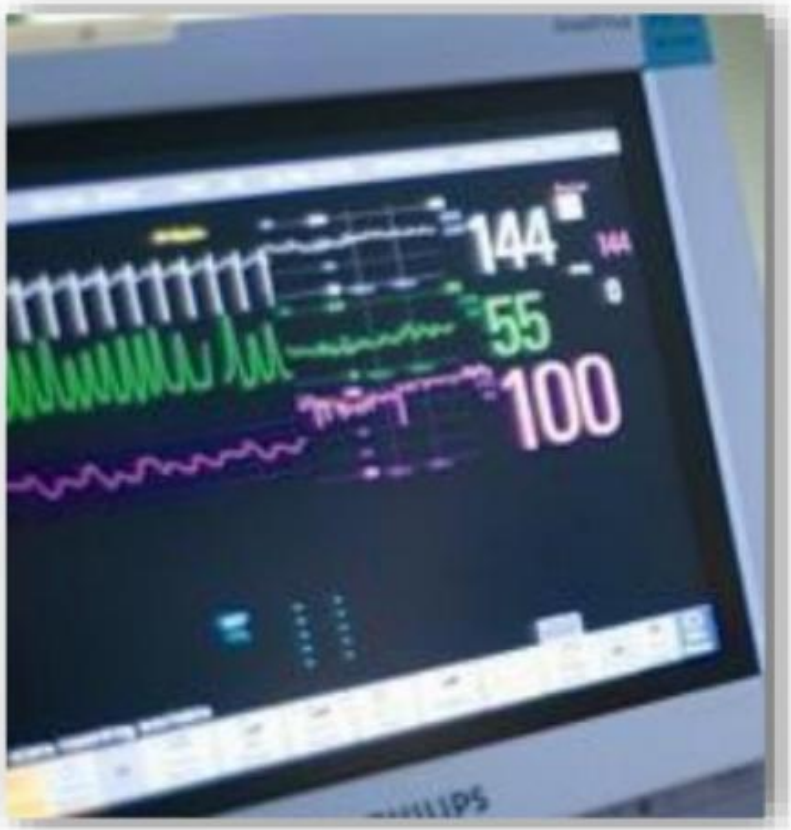
Sports teams are using data for tracking ticket sales and even for tracking team strategies.

Advertising and marketing agencies are tracking social media to understand responsiveness to campaigns, promotions, and other advertising mediums

Data can do a lot: medicine

The National Institutes of Health has announced a new opportunity for organizations interested in helping engage volunteers in the *All of Us* Research Program, part of the Precision Medicine Initiative. This funding opportunity, open to national and regional organizations, as well as local community groups, will support activities to promote enrollment and retention in the *All of Us* Research Program across diverse communities.

All of Us is an ambitious effort to gather data over time from 1 million or more people living in the United States, with the ultimate goal of accelerating research and improving health. Unlike research studies that are focused on a specific disease or population, *All of Us* will serve as a national research resource to inform thousands of studies, covering a wide variety of health conditions. Researchers will use data from the program to learn more about how individual differences in lifestyle, environment and biological make-up can influence health and disease. By taking part, people will be able to learn more about their own health and contribute to an effort that will advance the health of generations to come. NIH plans to launch the program later this year.



Hospitals are analyzing medical data and patient records to predict those patients that are likely to seek readmission within a few months of discharge. The hospital can then intervene in hopes of preventing another costly hospital stay.

Medical diagnostics company analyzes millions of lines of data to develop first non-intrusive test for predicting coronary artery disease. To do so, researchers at the company analyzed over 100 million gene samples to ultimately identify the 23 primary predictive genes for coronary artery disease

Amazon has an unrivalled bank of data on online consumer purchasing behaviour that it can mine from its 152 million customer accounts.

Amazon also uses Big Data to monitor, track and secure its 1.5 billion items in its retail store that are laying around it 200 fulfilment centres around the world. Amazon stores the product catalogue data in S3.

S3 can write, read and delete objects up to 5 TB of data each. The catalogue stored in S3 receives more than 50 million updates a week and every 30 minutes all data received is crunched and reported back to the different warehouses and the website.

Netflix uses 1 petabyte to store the videos for streaming.

BitTorrent Sync has transferred over 30 petabytes of data since its pre-alpha release in January 2013.

The 2009 movie **Avatar** is reported to have taken over 1 petabyte of local storage at **Weta Digital** for the rendering of the 3D CGI effects.

One petabyte of average **MP3**-encoded songs (for mobile, roughly one megabyte per minute), would require 2000 years to play.

"More data usually beats better algorithms,"

Such as: Recommending movies or music based on past preferences.

Your FB posts, tweets are writing H'wood's next hit

Producers Look At Social Media To Gauge Trend

Getty Images

WHAT'S TRENDING?

Your Facebook posts and tweets may contain hidden creativity. In fact, they could be helping to write the next Hollywood blockbuster.

As once-monolithic TV audiences splinter and migrate to the internet, viewers have unwittingly turned the creative process upside down. Their social media posts, blogs and file downloads are telling streaming companies and producers what actors, writers and themes to weave together on-screen for the best chance of bottom-line success.

The prize is clear for entertainment companies ranging from Netflix Inc., which has more than 81 million subscribers in 190 countries, to Stan Entertainment Pty, a startup battling the $42 billion Nasdaq-listed rival in Australia. Giving customers what they want — before they've even asked for it — increases the odds of a show's success while forging loyalty for the content-provider, industry executives maintain.

"We're trying to get a sense of what people are talking about, what's trending, and what's relevant out there," said Chris Oliver-Taylor, managing director of Matchbox Pictures Pty, the Australian producer of Glitch. "If the entire Twittersphere is talking about this particular thing, or if there's a trend in this area and we create a piece of content that talks directly to it, then logic suggests that those people will engage with it," Oliver-Taylor said.

The evolution of home entertainment to content streamed over the internet to multiple devices has facilitated greater insight into what, when and for how long customers are watching. It's that knowledge that informed the writing and casting of "Wolf Creek".

"When we sit down with the creators and producers of original productions, we effectively give them a high-level brief of what we're looking for," said Mike Sneesby, Stan's CEO. "Part of what's informed that brief is the data that comes from our platform."

Knowing what elements could make for a great show isn't enough to guarantee success, Stan's Sneesby said. "It's not always possible to match what the data says with the availability of projects," said Sneesby

Amazon.com Inc. says it releases pilots at Amazon Studios periodically for customers to watch and review. Their feedback is taken into account when executives decide which pilots will be co-

Netflix, which distributes shows such as "House of Cards" and "Orange Is the New Black," pioneered the use of mathematical equations to promote titles that a subscriber might enjoy. That's based on variables such as previously downloaded content, the subscriber's location and the show's broader popularity. BLOOMBERG

Data can do a lot: discover bias

Stanford | News

Search Stanford news...

Home Find Stories For Journalists Contact

JUNE 15, 2016

Stanford big data study finds racial disparities in Oakland, Calif., police behavior, offers solutions

Stanford researchers analyzing thousands of data points found racial disparities in how Oakland Police Department officers treated African Americans on routine traffic and pedestrian stops. The researchers suggest 50 measures to improve police-community relations, such as better data collection, bias training and changes in cultures and systems.

BY CLIFTON B. PARKER

New Stanford research on thousands of police interactions found significant racial differences in Oakland, California, police conduct toward African Americans in traffic and pedestrian stops, while offering a big data approach to improving police-community relationships there and elsewhere.

Data can do a lot: politics

FAQ | Today's Polls | Pollster Ratings | Contact | Electoral History

FiveThirtyEight Politics Done Right

2010 SENATE RANKINGS

1	Missouri	Open
2	Nevada ▲	Reid
3	Ohio	Open
4	Connecticut ▼	Dodd
5	Colorado ▲	Bennet
6	New Hampshire ▼	Open
7	Kentucky	Open
8	Arkansas ▲	Lincoln
9	Illinois	Burris
10	North Carolina	Burr
11	Delaware ▼	Open
12	Pennsylvania ▼	Specter
13	Texas	Open?
14	Louisiana	Vitter
15	Iowa ▲	Grassley

Advertise @ 538!

11.04.2008

Today's Polls and Final Election Projection: Obama 349, McCain 189

by Nate Silver @ 1:16 PM

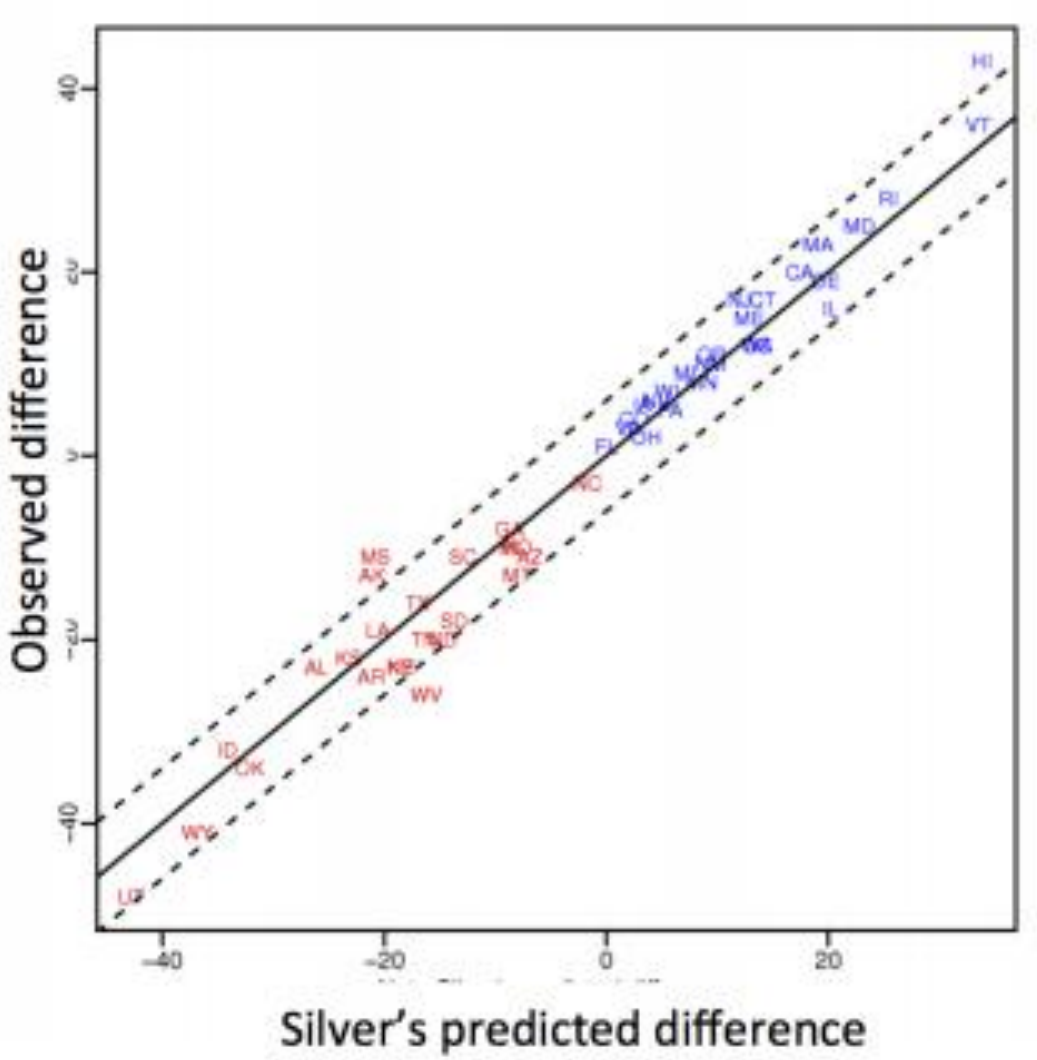
Share This Content

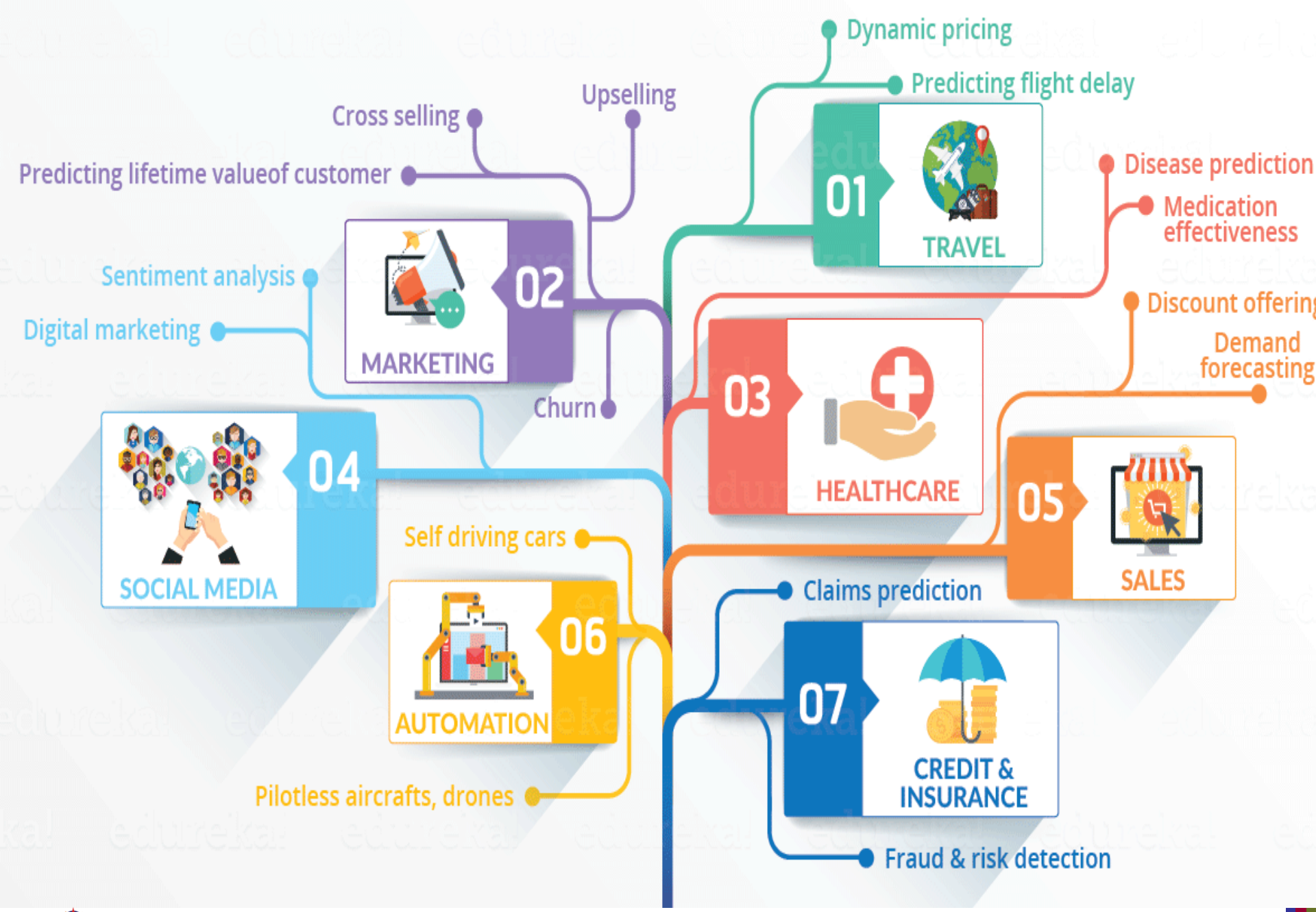
It's Tuesday, November 4th, 2008, Election Day in America. The last polls have straggled in, and show little sign of mercy for John McCain. Barack Obama appears poised for a decisive electoral victory.

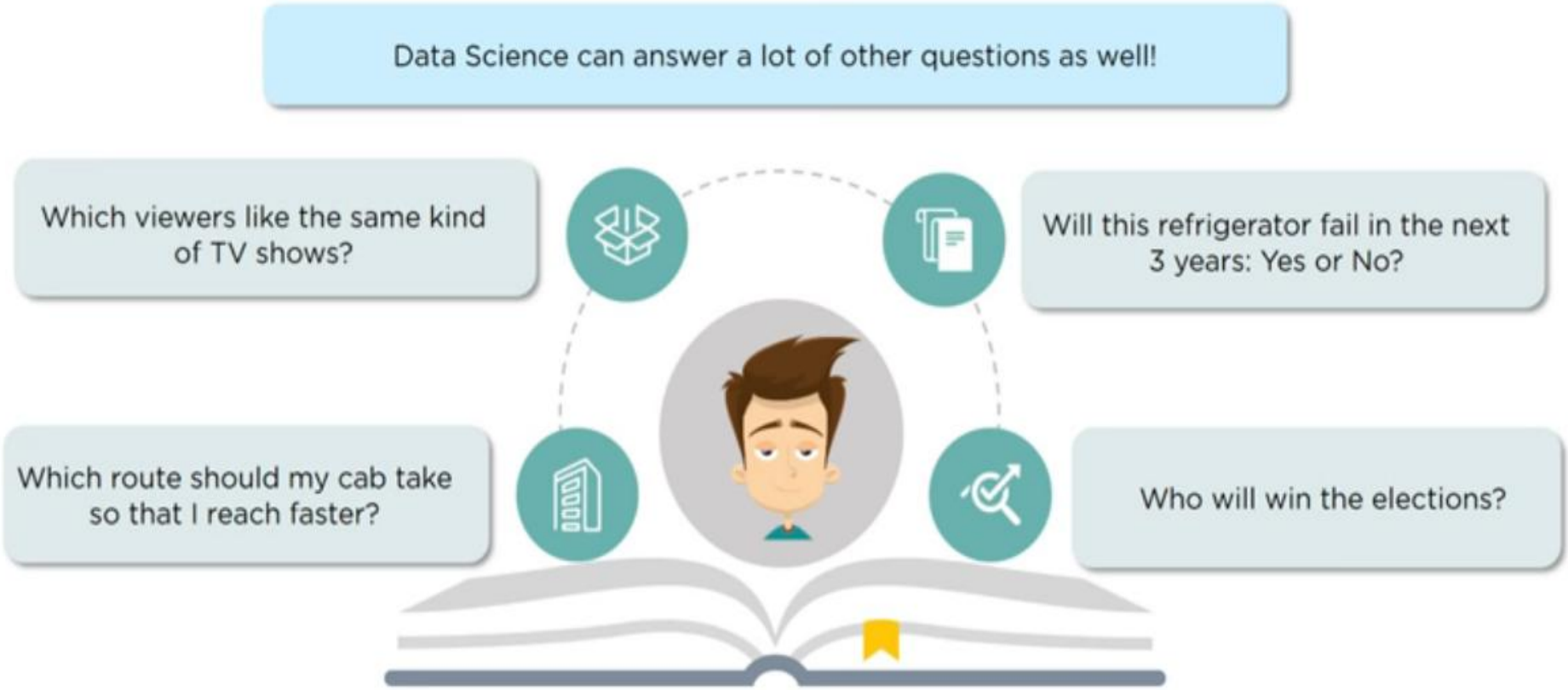
Our model projects that Obama will win all states won by John Kerry in 2004, in addition to Iowa, New Mexico, Colorado, Ohio, Virginia, Nevada, Florida and North Carolina, while narrowly losing Missouri and Indiana. These states total 353 electoral votes. Our official projection, which looks at these outcomes probabilistically -- for instance, assigns North Carolina's 15 electoral votes to Obama 59 percent of the time -- comes up with an incrementally more conservative projection of 348.6 electoral votes.

We also project Obama to win the popular vote by 6.1 points; his lead is slightly larger than that in the polls now, but our model accounts for the fact that candidates with large leads in the polls typically underperform their numbers by a small margin on Election Day.

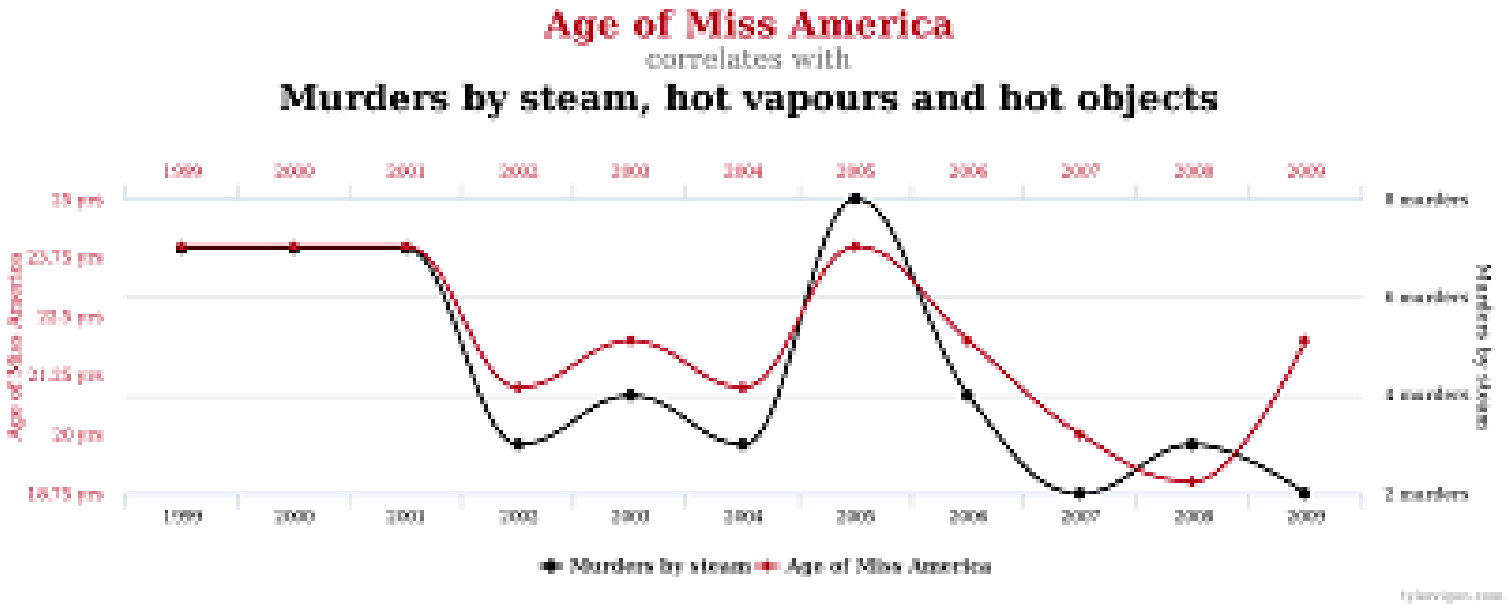
Data can do a lot: politics




Dynamic pricing
Predicting flight delay
01
TRAVEL
Disease prediction
Medication effectiveness
Discount offering
Demand forecasting
05
SALES
Upselling
Cross selling
Predicting lifetime valueof customer
02
MARKETING
Sentiment analysis
Digital marketing
Churn
03
HEALTHCARE
04
SOCIAL MEDIA
Self driving cars
06
AUTOMATION
Pilotless aircrafts, drones
Claims prediction
07
CREDIT & INSURANCE
Fraud & risk detection


Data Science can answer a lot of other questions as well!
Which viewers like the same kind of TV shows?
Will this refrigerator fail in the next 3 years: Yes or No?
Which route should my cab take so that I reach faster?
Who will win the elections?

Is data all we need?



- It is easy to find "interesting" patterns where none exist!
- How should we judge whether a "pattern" is interesting?
- When should we worry about falsely labelling patterns "interesting"? (E.g. Google mistranslates a sentence vs. incorrect cancer diagnosis. . .)

Is data all we need?

Cornell University Library

arXiv.org > cs > arXiv:1607.06520

Computer Science > Computation and Language

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai

Submitted on 21 Jul 2016

The blind application of machine learning runs the risk of amplifying biases present in data. Such a danger is facing us with word embedding, a popular framework to represent text data as vectors which has been used in many machine learning and natural language processing tasks. We show that even word embeddings trained on Google News articles exhibit female/male gender stereotypes to a disturbing extent. This raises concerns because their widespread use, as we describe, often tends to amplify these biases. Geometrically, gender bias is first shown to be captured by a direction in the word embedding. Second, gender neutral words are shown to be linearly separable from gender definition words in the word embedding. Using these properties, we provide a methodology for modifying an embedding to remove gender stereotypes, such as the association between between the words receptionist and female, while maintaining desired associations such as between the words queen and female. We define metrics to quantify both direct and indirect gender biases in embeddings, and develop algorithms to "debias" the embedding. Using crowd-worker evaluation as well as standard benchmarks, we empirically demonstrate that our algorithms significantly reduce gender bias in embeddings while preserving the its useful properties such as the ability to cluster related concepts and to solve analogy tasks. The resulting embeddings can be used in applications without amplifying gender bias.

Subjects: **Computation and Language** (cs.CL); Artificial Intelligence (cs.AI); Learning (cs.LG); Machine Learning (stat.ML)

Cite as: arXiv:1607.06520 [cs.CL]

(or arXiv:1607.06520v1 [cs.CL] for this version)

Blindly used, machine learning algorithms can reinforce biases hidden in data.

Learn how to use data

- **Explore**: identify patterns
- **Predict**: make informed guesses
- **Infer**: quantify what you know

So Data Science is mainly needed for:

Better Decision Making

Whether A or B?

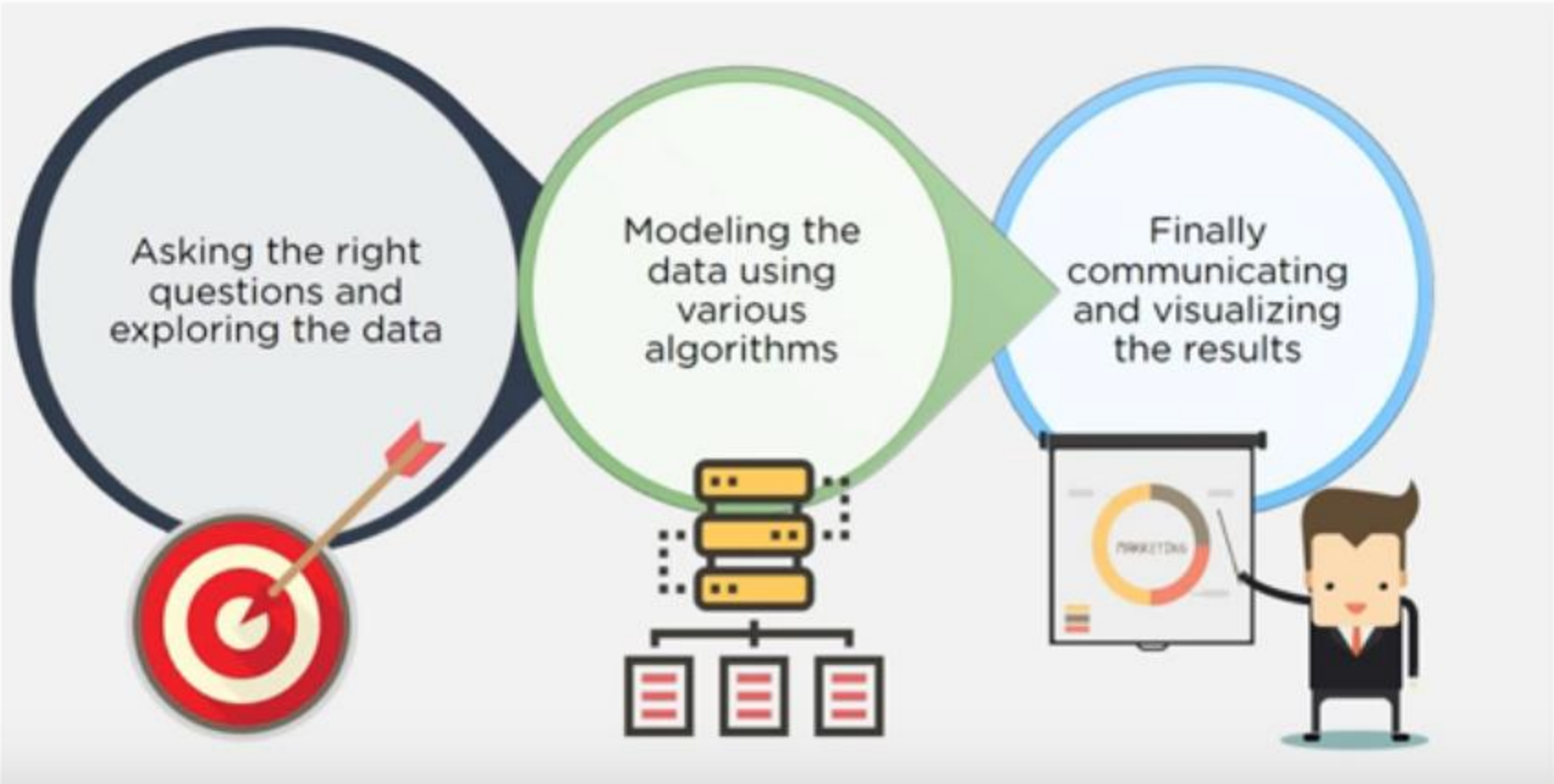
Predictive Analysis

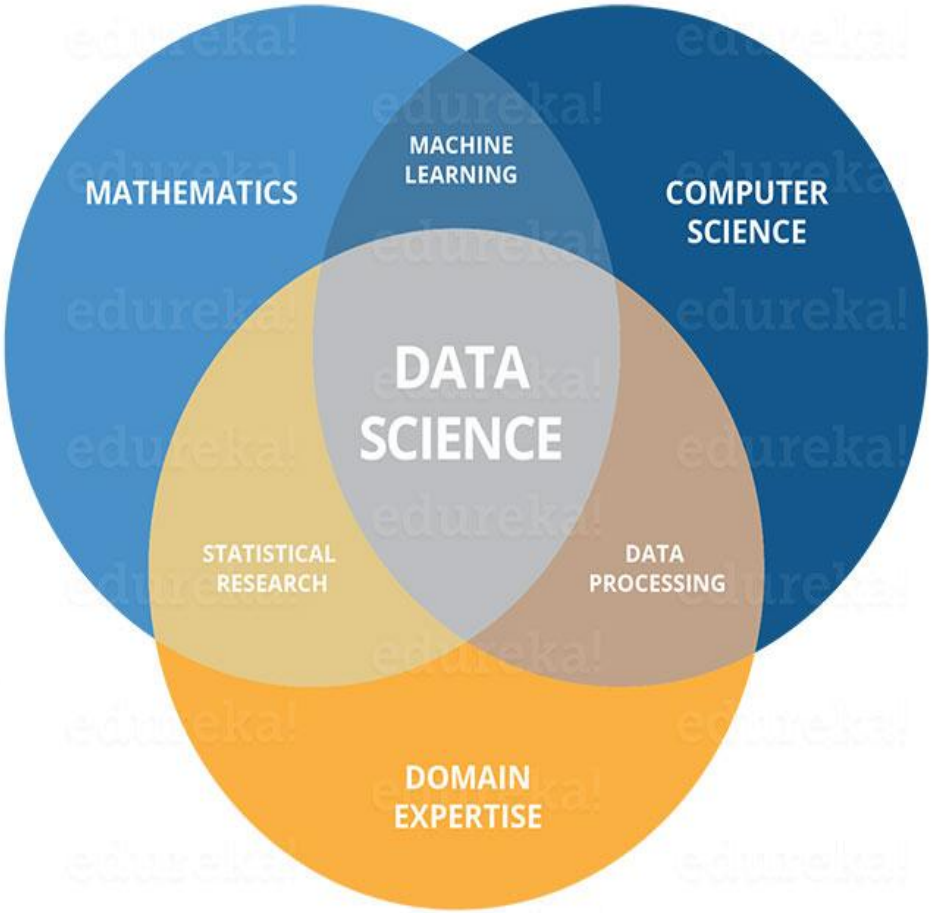
What will happen next?

Pattern Discovery

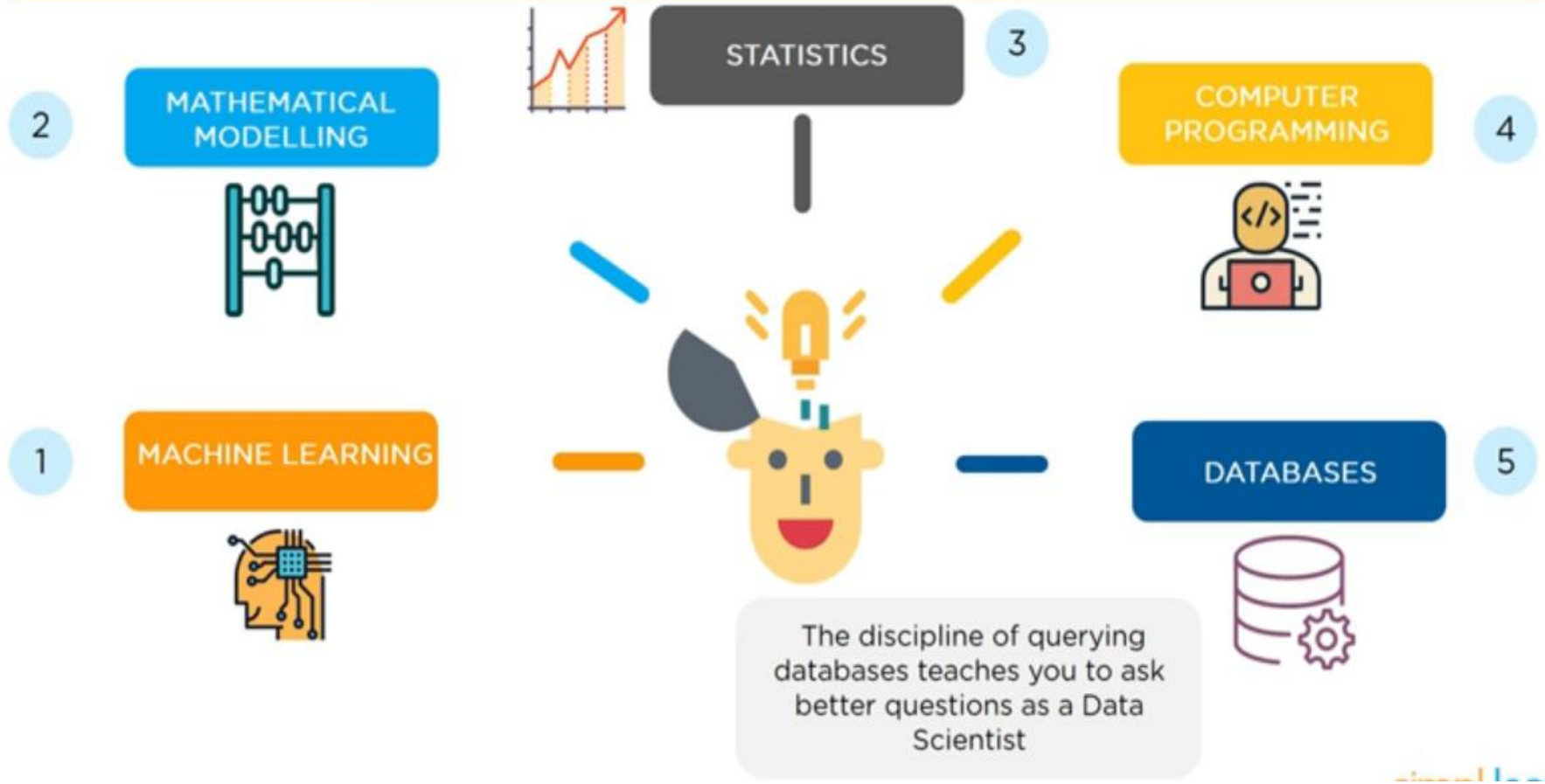
Is there any hidden information in the data?

So, Data Science or Data-driven Science is about:

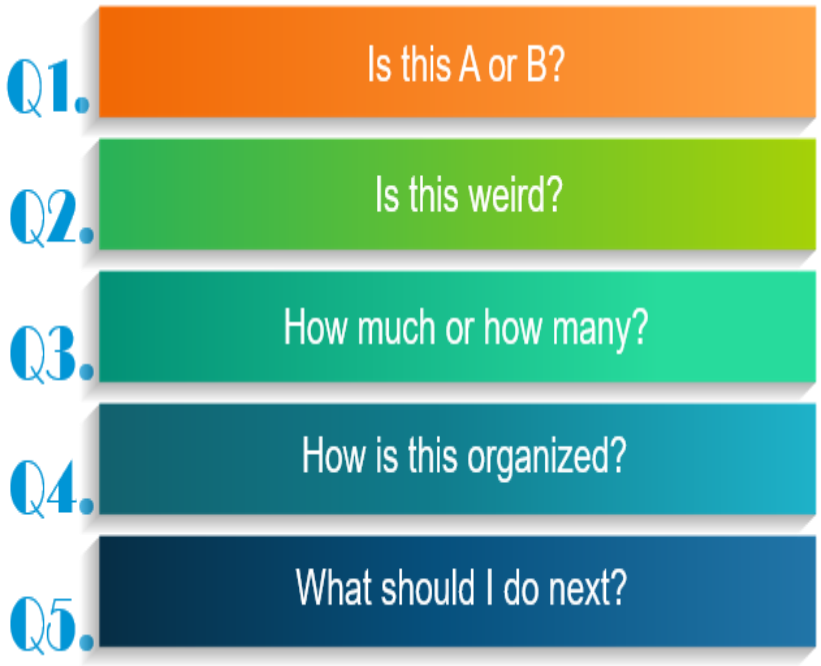



MATHEMATICS
MACHINE LEARNING
COMPUTER SCIENCE
DATA SCIENCE
STATISTICAL RESEARCH
DATA PROCESSING
DOMAIN EXPERTISE

Prerequisites for Data Science




MATHS
Statistics
Discrere Maths
Information Theory
Combinatorics
Decision Theory
Machine Learning
Econometricians
Management
Saientists
Actuaries
DATA SCIENTIST
Dta Viz Builders
Statistical
Programmers
BUSINESS
Economics
Finance
Marketing
Operations
Management
BI Developers
Data Analysis
INFORMATION SYSTEMS
Computer Science
Software Engineerinf
Systems Development



Classification Algorithm

Anomaly Detection Algorithm

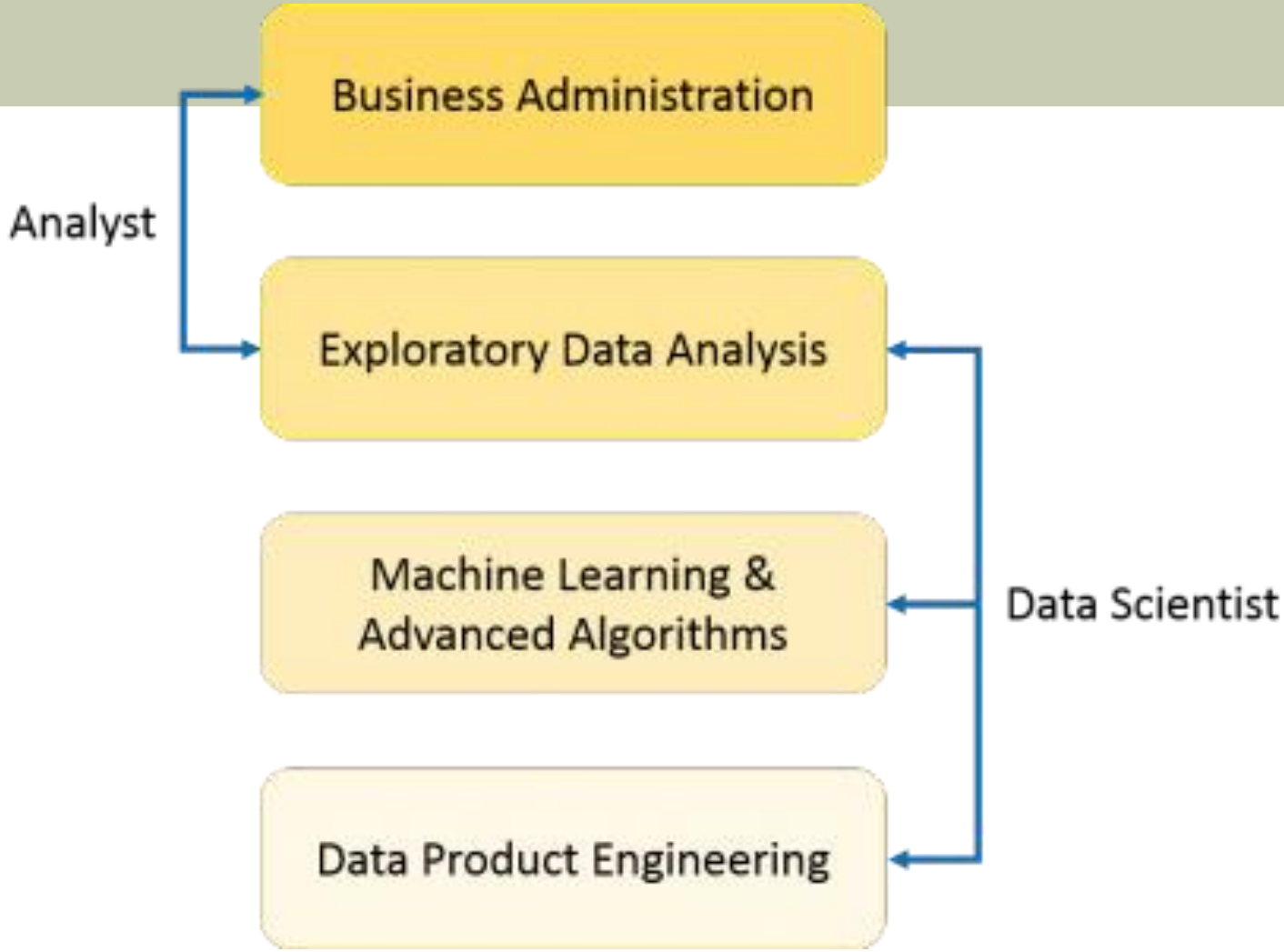
Regression Algorithms

Clustering Algorithms

Reinforcement Learning

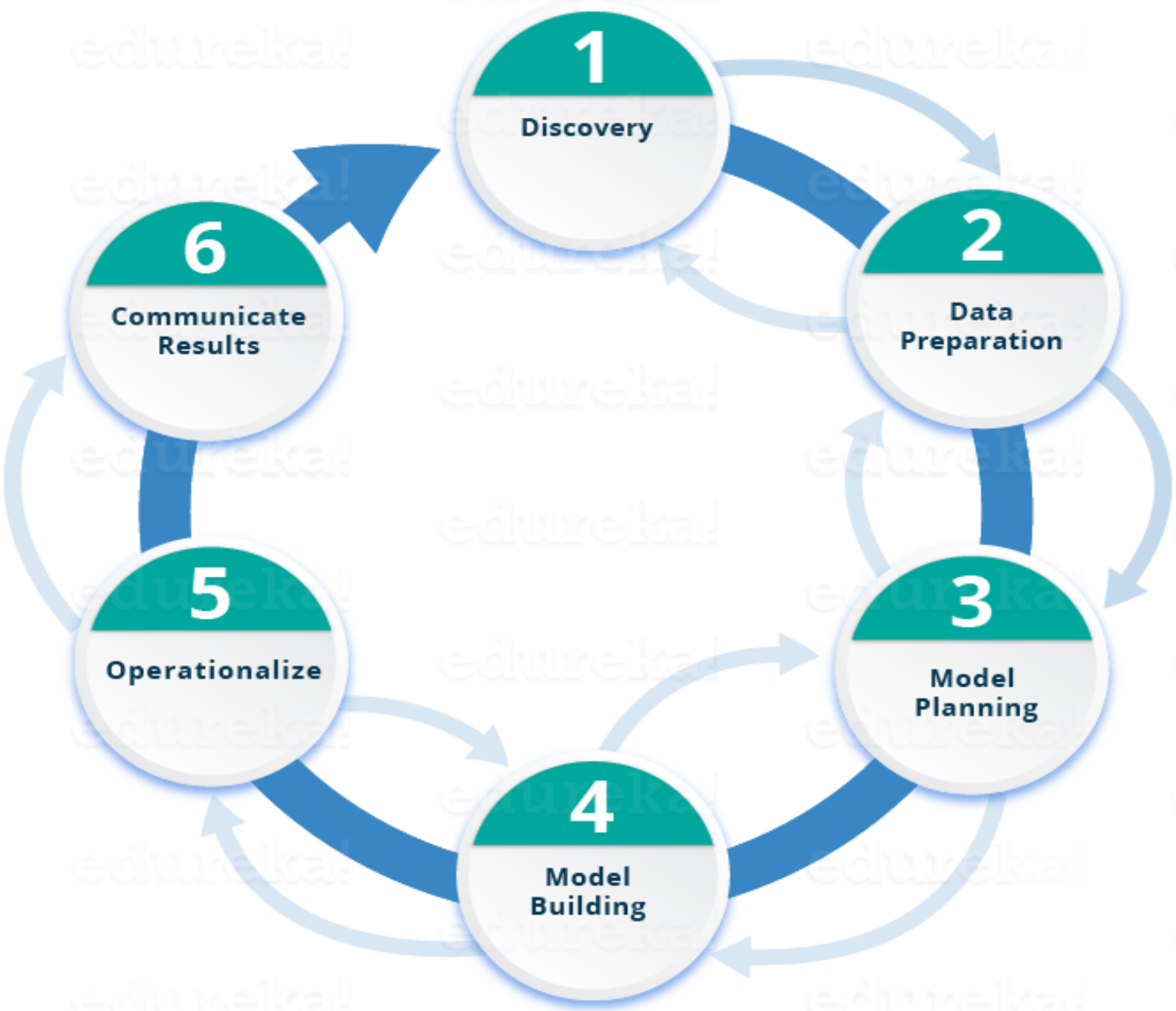
What is Data Science?

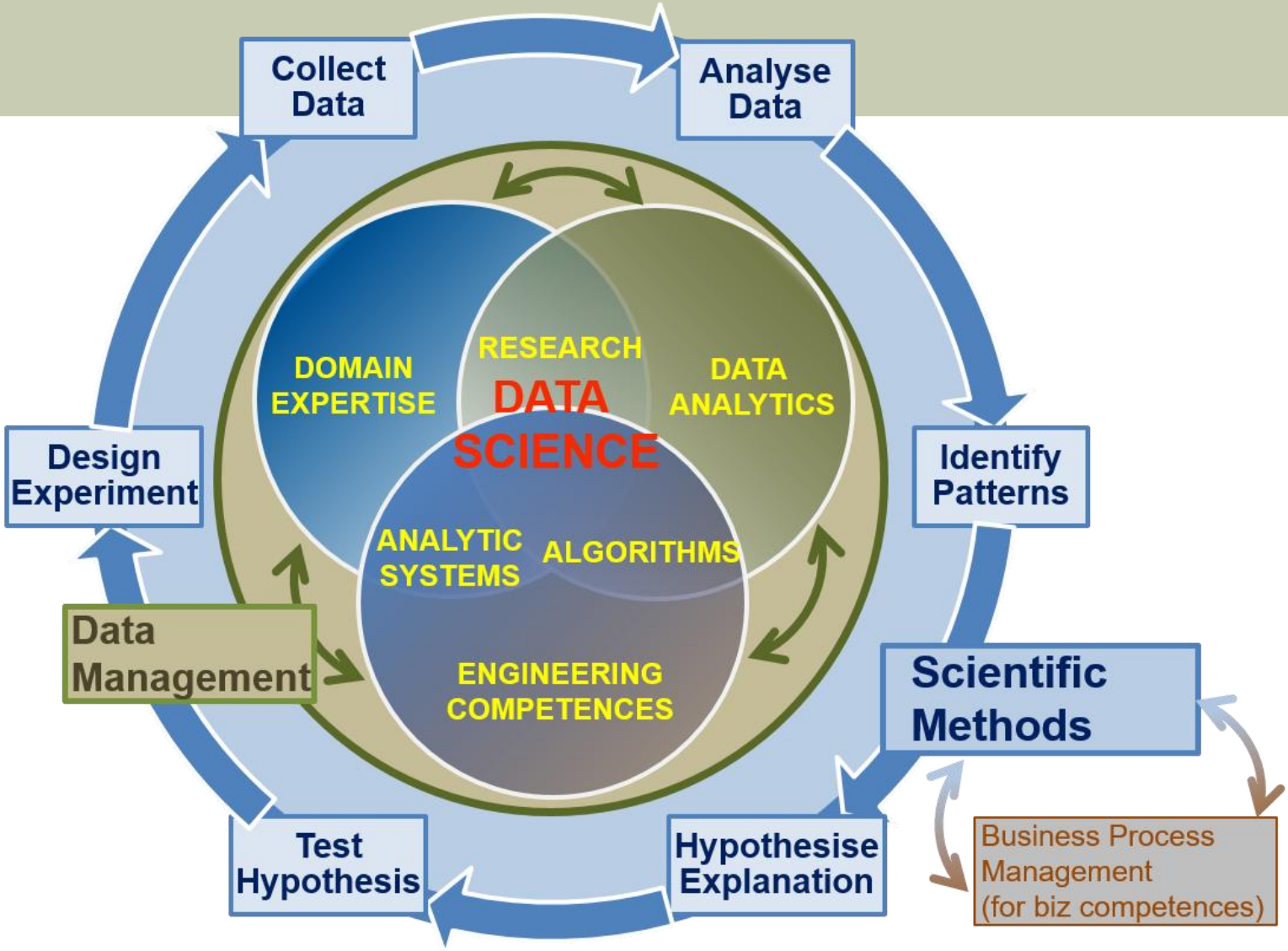
- Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.
- How is this different from what statisticians have been doing for years?


Business Administration
Analyst
Exploratory Data Analysis
Machine Learning &
Advanced Algorithms
Data Scientist
Data Product Engineering

- 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.
- Data Analysis includes descriptive analytics and prediction to a certain extent. On the other hand, Data Science is more about Predictive Causal Analytics and Machine Learning.


edureka!
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


1
Discovery
2
Data Preparation
3
Model Planning
4
Model Building
5
Operationalize
6
Communicate Results


Collect Data
Analyse Data
Identify Patterns
Scientific Methods
Hypothesise Explanation
Test Hypothesis
Design Experiment
Data Management
DOMAIN EXPERTISE
RESEARCH
DATA
SCIENCE
DATA ANALYTICS
ANALYTIC SYSTEMS
ALGORITHMS
ENGINEERING COMPETENCES
Business Process Management (for biz competences)