Modern Data Science: Trends and Challenges in Colombia

Cesar O. Diaz B. PhD.

06/10/2016

About me

- Electrical engineer. 2001
- Electronical Master. 2009
- PhD in Computer Science. 2014
- Postdoc. 2015
- Professor UJTL. Today.
- Research interest: Cloud computing, Cloud Analytics, Big Data and IoT.

Outline

- Big Data Definition
- How Much Data is Created Every Minute?
- The Growth of Devices Connected to the Internet
- Data inflation
- Big Data Opportunities
- Data Science
- Data Analytics

Big Data Definition

- Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.

"much IT investment is going towards managing and maintaining big data”

- About 40,200,000 results (0.60 seconds)

Big Data Definition

McKinsey Global institute define:

"Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

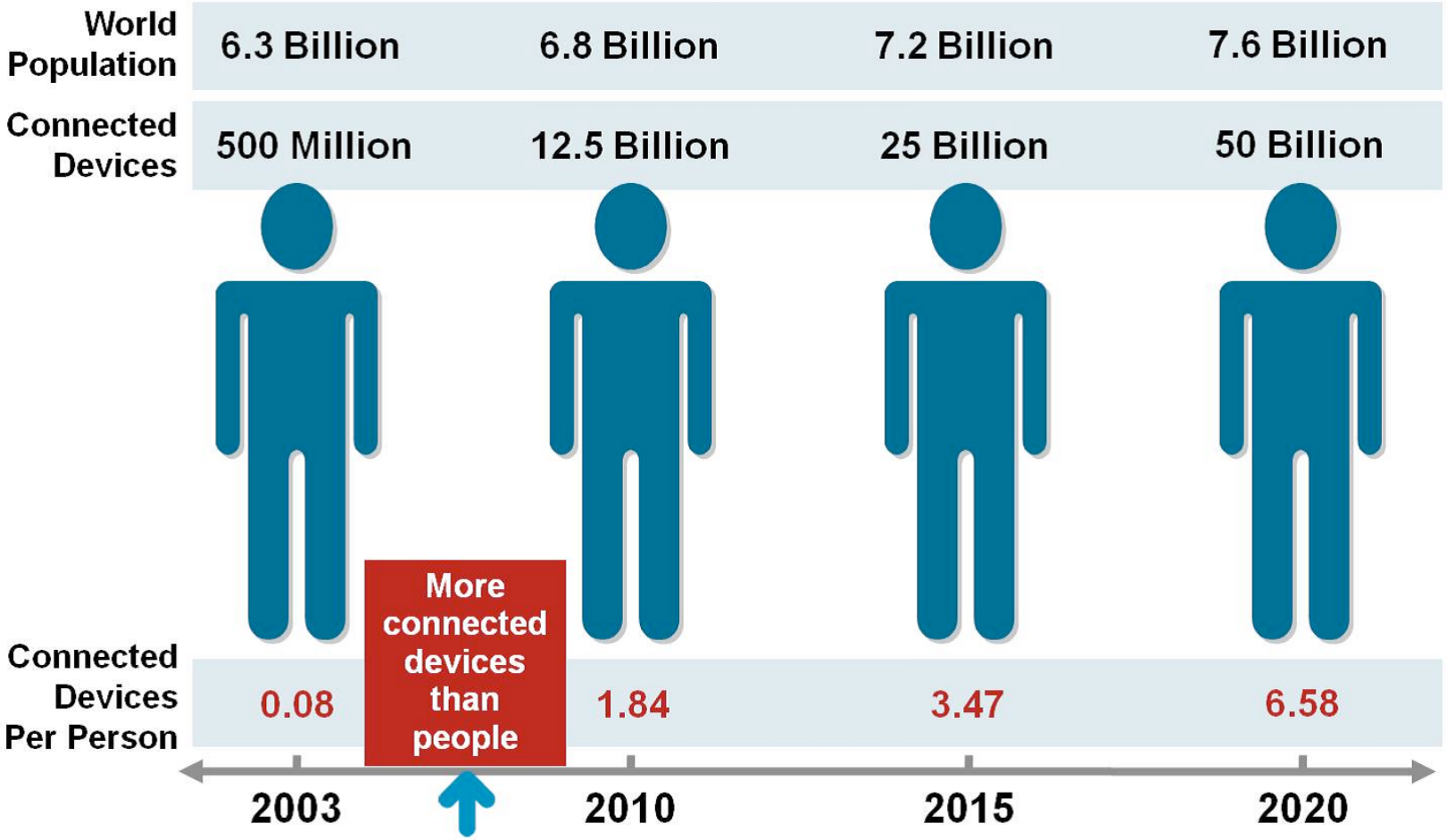
It isn't defined big data in terms of being larger than a certain number of terabytes (thousands of gigabytes)

J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. Big data: The next frontier for innovation, competition, and productivity. 2011.

How Much Data is Created Every Minute?

- Data never sleeps. Every minute massive amounts of it are being generated from every phone, website and application across the Internet. Just how much data is being created and where does it come from?

The Growth of Devices Connected to the Internet



Source: Cisco IBSG: April 2011

Data Inflation

TABLE 18.1 **Data inflation**

Unit	Size	Description
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data.
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing.
Kilobyte (KB)	1,024 bytes; 2^{10} bytes	From "thousand" in Greek. One typed page of typed text is 2KB.
Megabyte (MB)	1,024KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB.
Gigabyte (GB)	1,024MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1–2GB.
Terabyte (TB)	1,024GB; 2^{40} bytes	From "monster" in Greek. All the printed information in America's Library of Congress totals 15TB.
Petabyte (PB)	1,024TB; 2^{50} bytes	All letters delivered by America's postal service in 2010 amounted to around 5PB. Google processed around 1PB every hour in 2009.
Exabyte (EB)	1,024PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*.
Zettabyte (ZB)	1,024EB; 2^{70} bytes	The total amount of information in existence in 2011 was estimated to be around 1.8ZB.
Yottabyte (YB)	1,024ZB; 2^{80} bytes	Currently too large to imagine.

Note: The prefixes are determined by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991, but terms for larger amounts have yet to be established.
Source: *The Economist*

Cost per Megabyte

The cost of hard drives, used in computers for storing data in large quantities, has been falling rapidly for many years

- 1956 – U$10.000
- 1981 – U$700
- 1983 – U$316
- 1984 – U$280
- 1987 – U$90
- 1988 – U$40
- 1989 – U$12
- 1990 – U$9
- 1991 – U$7
- 1992 – U$4
- 1993 – U$2
- 1994 – 95¢
- 1995 – 85¢
- 1996 – 29¢
- 1997 – 14¢
- 1998 – 5¢
- 1999 – 2¢
- 2000 – 1¢

I. Smith. Cost of hard drive storage space. http://ns1758.ca/winch/winchest.html, 2016

Cost per Megabyte

The cost of hard drives, used in computers for storing data in large quantities, has been falling rapidly for many years

- 1956 – U$10.000
- 1981 – U$700
- 1983 – U$316
- 1984 – U$280
- 1987 – U$90
- 1988 – U$40
- 1989 – U$12
- 1990 – U$9
- 1991 – U$7
- 1992 – U$4
- 1993 – U$2
- 1994 – 95¢
- 1995 – 85¢
- 1996 – 29¢
- 1997 – 14¢
- 1998 – 5¢
- 1999 –
- 200

per gigabyte (below)
instead of per megabyte (above)

I. Smith. Cost of hard drive storage space. http://ns1758.ca/winch/winchest.html, 2016

How to Benefit from Big Data[1]

1 Multiple Data Sources

- Creatively source internal and external data
- Upgrade IT architecture and infrastructure for easy merging of data.

2 Prediction and Optimization Models

- Focus on the biggest drivers of performance
- Build models that balance complexity with ease of use

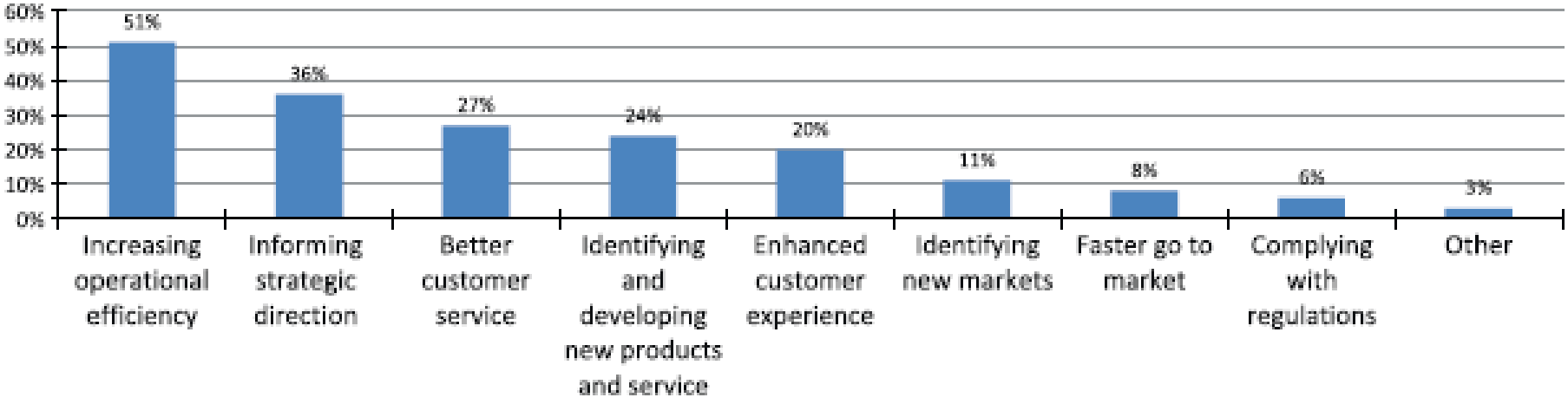
3 Organizational Transformation

- Create simple, understandable tools for people on the front lines.
- Update processes and develop capabilities to enable tool use.

[1]Barton D, Court D. "*Making advanced analytics work for you*". Harvard Business Review, octubre 2012; volumen 90, número 10: 78–83

Big Data Opportunities

Big Data Opportunities: above 50% of 560 enterprises think Big Data will help them in increasing operational eficciency, etc.



UTADEO
UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO

4 AÑOS ACIS
ASOCIACIÓN COLOMBIANA DE INGENIEROS DE SISTEMAS

Data Science

It's 2016 and there is still no unique definition of Data Science

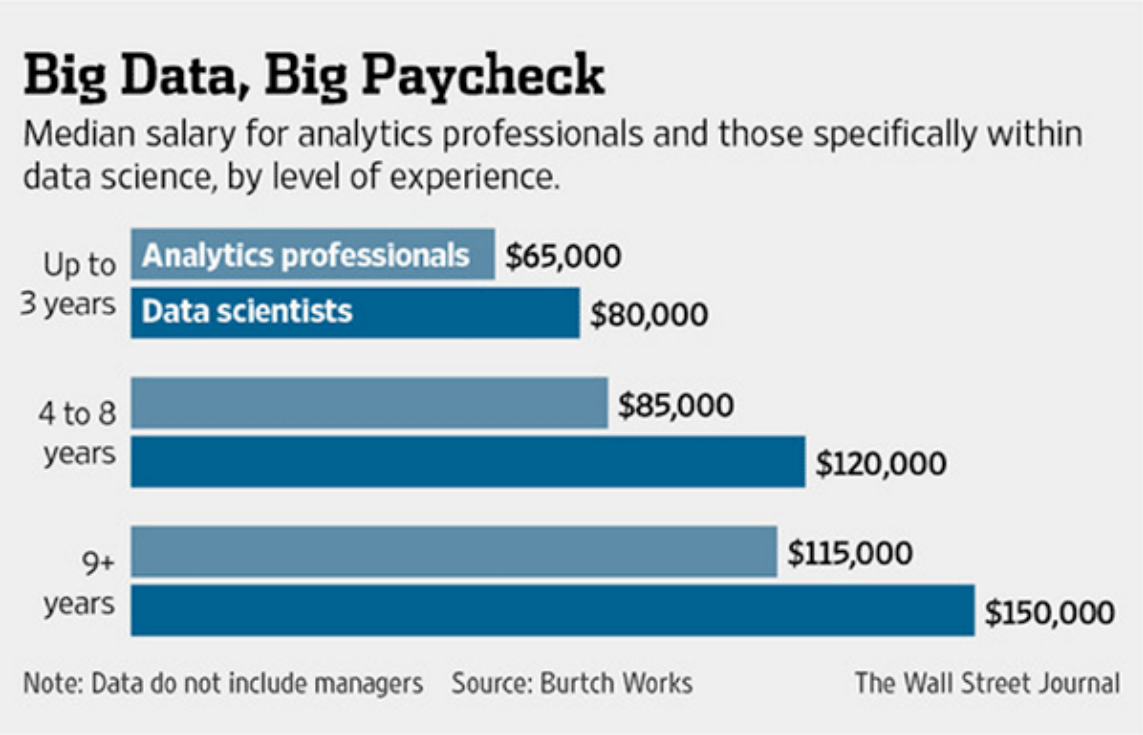
Data Science

Data Science is like teenage sex:

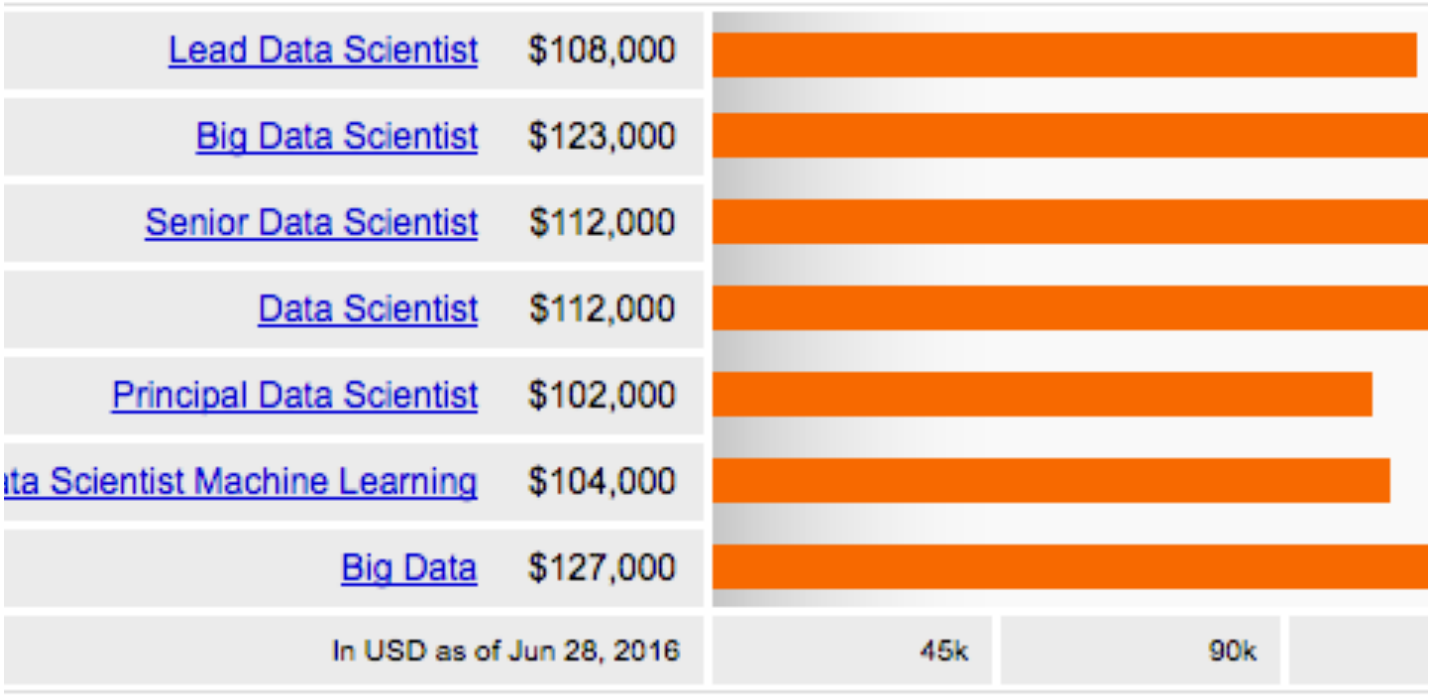
- everyone talks about it,
- nobody really knows how to do it,
- everyone thinks everyone else is doing it,
- so everyone claims they are doing it...

Even worse, people use several words interchangeable

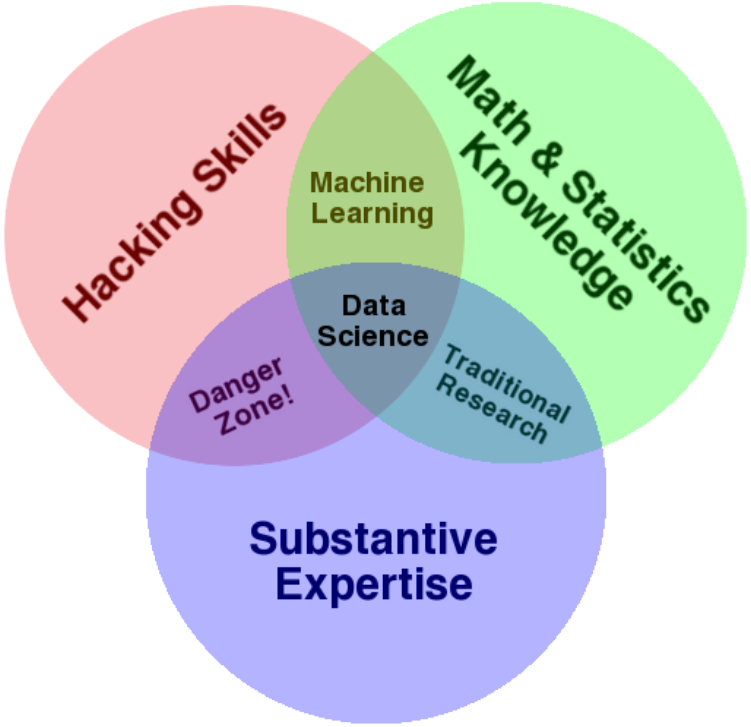
Data Science



Average Salary of Jobs with Related Titles



Data Science



Data Science is the intersection of Hacking Skills, Math & Statistics Knowledge and Substantive Expertise

Those are the pillars of data science: computing, statistics, mathematics and quantitative disciplines combined to analyze data for better decision making

Hacking Skills

Ability to build things and find clever solutions to problems.

- Programming/Coding: Python and R (and others)
- Databases: MySQL, PostgreSQL, Cassandra, MongoDB and CouchDB.
- Visualization: D3, Tableau, Qlikview and Markdown.
- Big Data: Hadoop, MapReduce and Spark.

Hacking Skills

Ability to build things and fi
problems.

Being able to manipulate text files at the command-line, understanding vectorized operations, thinking algorithmically; these are the hacking skills that make for a successful data hacker.

and Spark.

Math and Statistics

Being able understand the right solution to each problem

- Linear algebra: Matrix manipulation
- Machine Learning: Random Forests, SVM,
- Boosting Descriptive statistics: Describe, Cluster
- Statistical inference: Generate new knowledge .

Math and Statistics

Being able understand the right so
problem

- Linea

This is not to say that a PhD in statistics is required to be a competent data scientist, but it does require knowing what an ordinary least squares regression is and how to interpret it.

Substantive Expertise

Ability to ask good questions requires domain understanding, that's why a data scientist can't create data based solutions without a good industry knowledge

- Is this A or B or C? (classification)
- Is this weird? (anomaly detection).
- How much/how many? (regression).
- How is it organized? (clustering).
- What should I do next? (reinforcement learning)

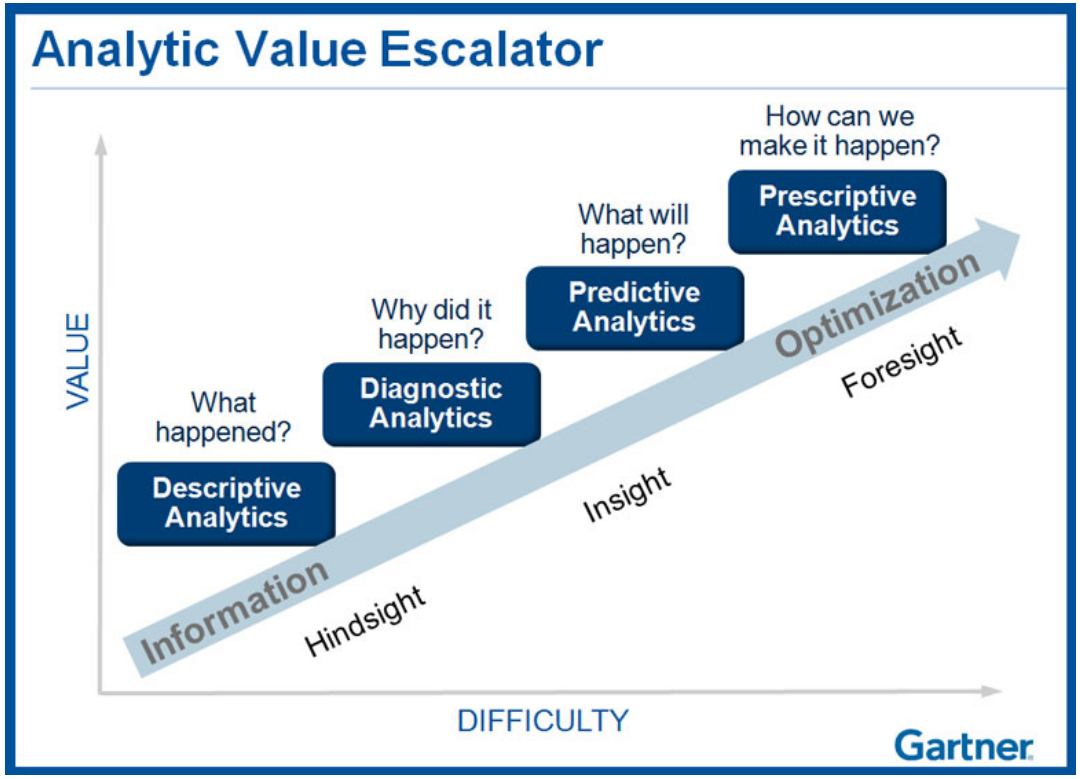
Substantive Expertise

Ability to ask good questions requi
understanding, that's wh
create data ba
kn

Science is about discovery and building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods.

ext? (reinforcement learning)

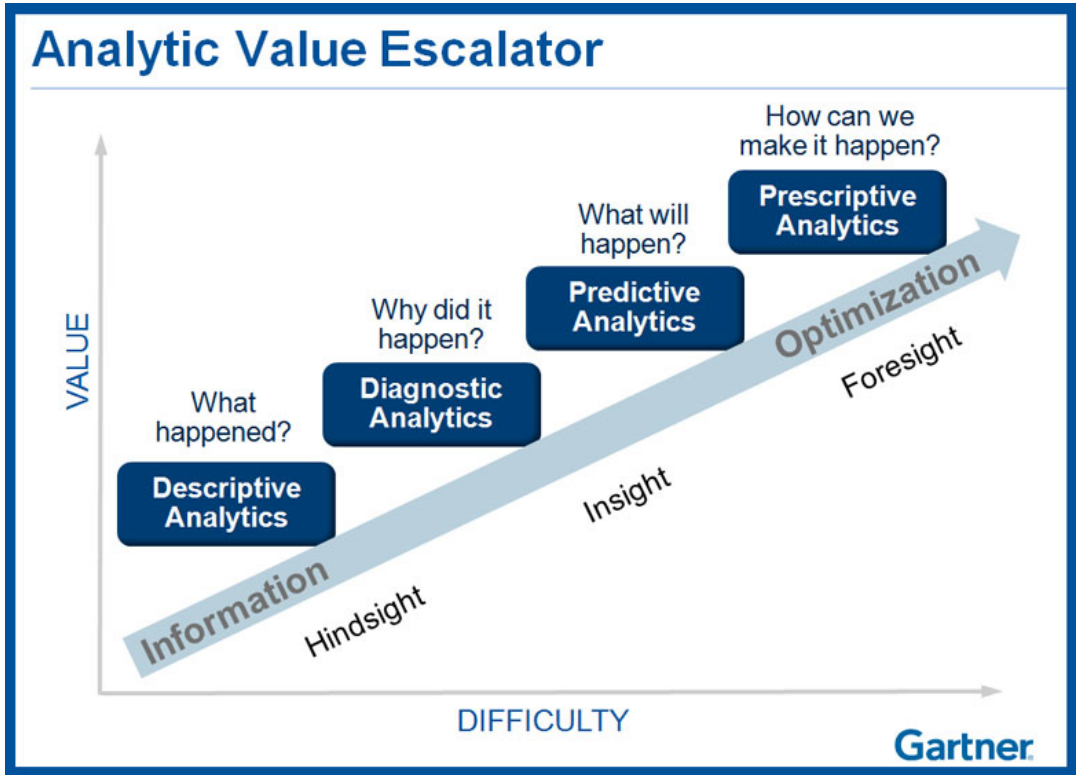
Data Analytics



Descriptive Analytics:

- Most marketers today utilize descriptive analytics tools to show how a specific campaign or marketing channel has performed.
- Google Analytics shows us how many visitors came to each page, whether they clicked on any links and how long they spent on the site, among other things. Social media tools, such as Radian 6, that provide an overview of what is happening on social media sites are another form.

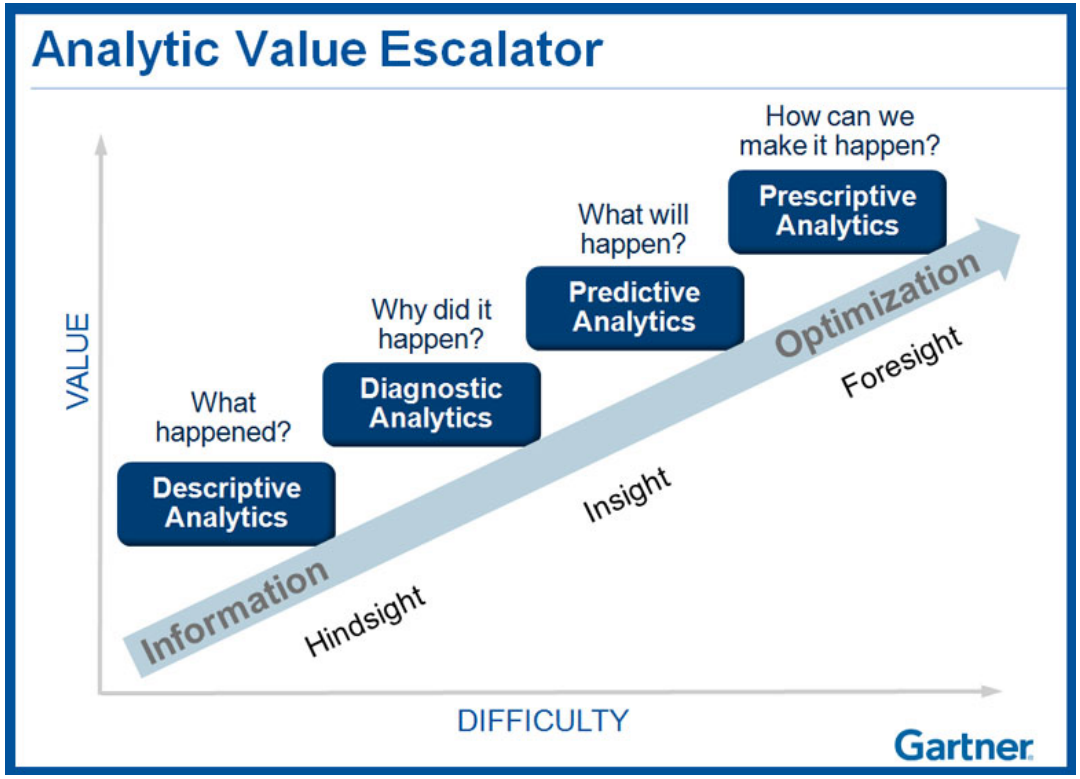
Data Analytics



Diagnostic Analytics:

- This type of analytics includes the use of tools like decision trees. One use of a decision tree is to map out the step-by-step journey that a customer takes in the purchasing process.
- BestBuy and Walmart use diagnostic analytics to help them answer the question, "Why did it happen?" when a customer visits a location and purchases a DVD player made by one company when there are five alternatives next to it on the shelf and even more online

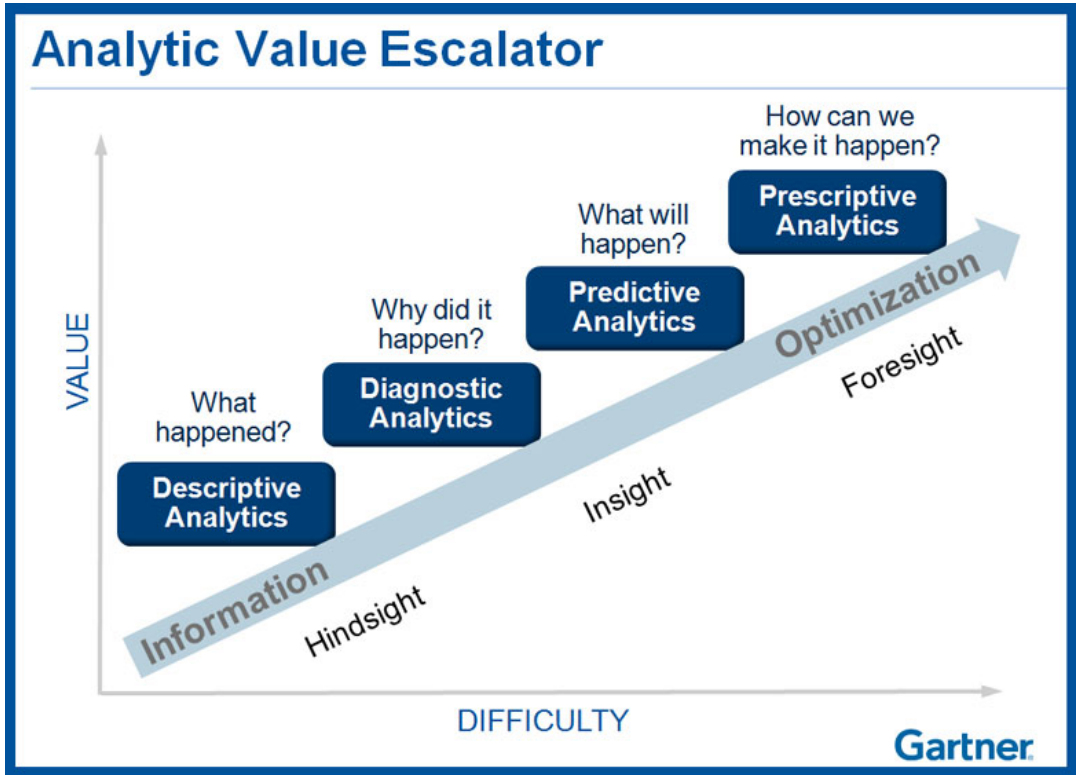
Data Analytics



Predictive Analytics:

- This form of analytics assists those interested in forecasting the outcome of an event before it happens.
- One example of this is lead scoring, which is the practice of ranking potential leads based on set pre-established factors, assessing them, and marketing to those who have the greatest odds of conversion.

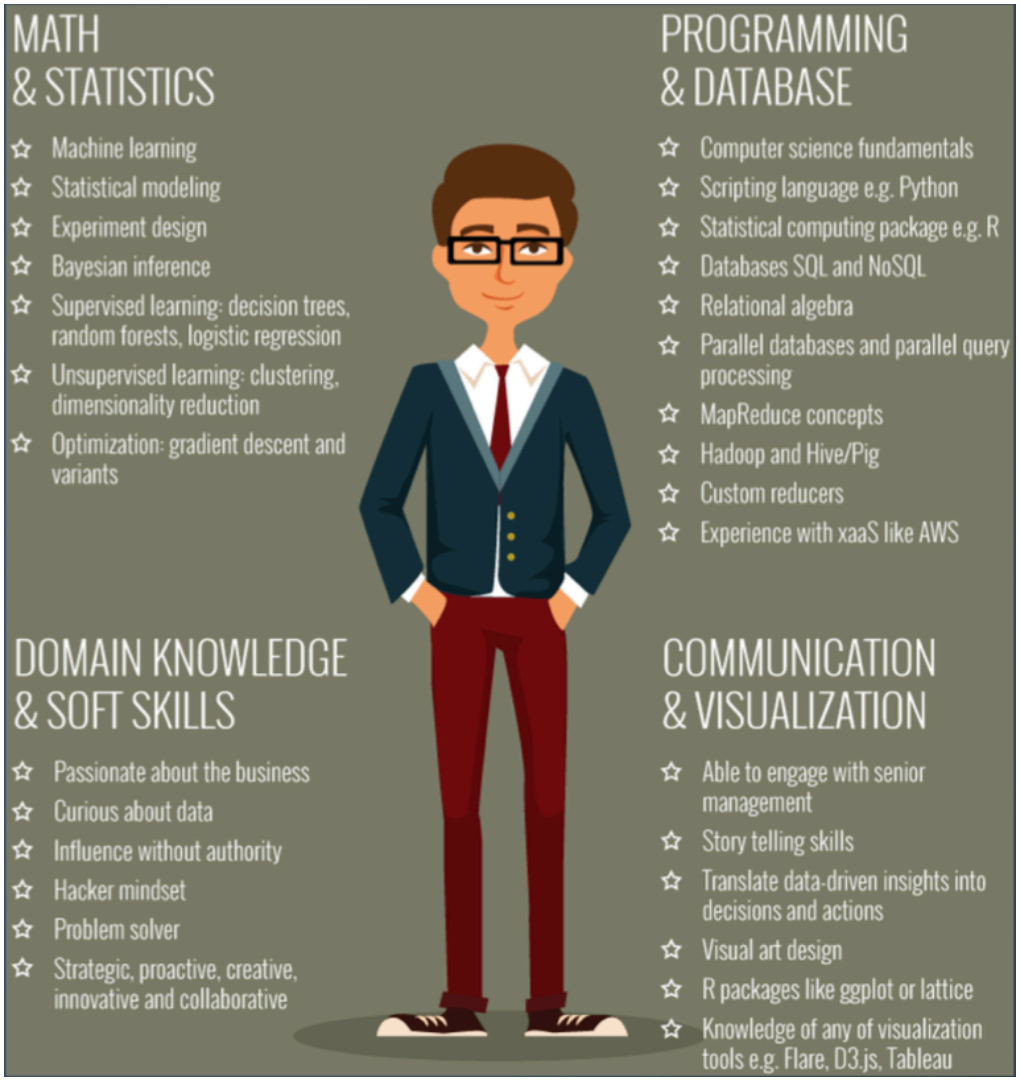
Data Analytics



Prescriptive Analytics:

- This represents the highest level of marketing optimization as this analysis allows marketers to go even further than predicting outcomes by factoring in both the results of predictions and the potential risks or rewards that could occur into one analysis.
- *Netflix and Amazon*. Whether you search through or actually make a purchase on either of those sites, from your very first visit, product recommendations are constantly being provided to you. These recommendations are based on the results of automated prescriptive analytics that occur with each interaction.

What about in Colombia?



Modern Data Scientist

The sexiest job of the 21th century

UTADEO
UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO

40 AÑOS ACIS
ASOCIACIÓN COLOMBIANA DE INGENIEROS DE SISTEMAS

GRACIAS

cesaro.diazb@utadeo.edu.co