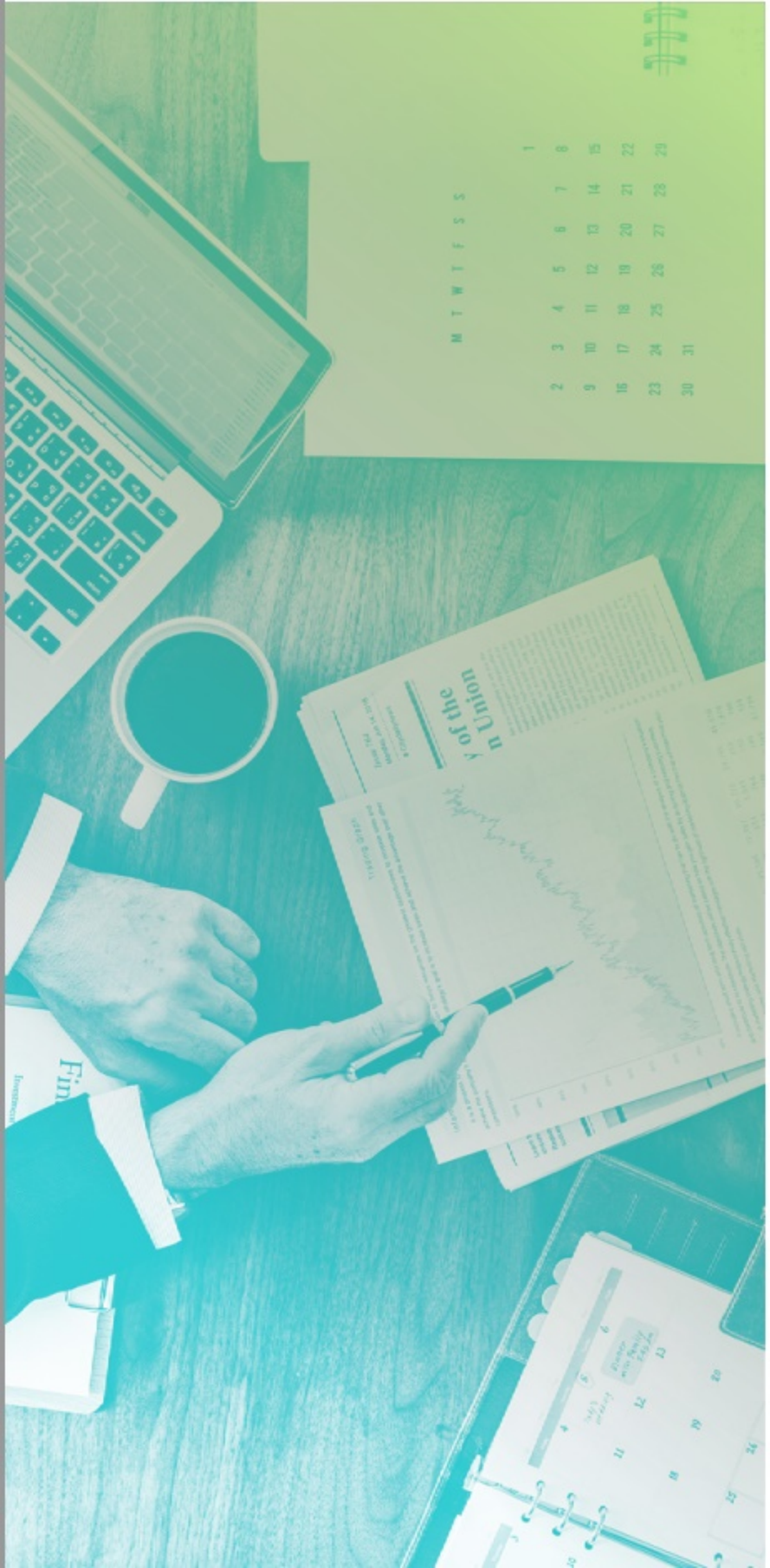
Unit I:

Introduction: What is data visualization? History, The data visualization process, Why is data visualization so important in reports and statements? Explaining, Exploring, Analyzing.

Data Science Definition – Big Data and Data Science Hype – Why data science – The Current Landscape – Skill sets required for Data Scientist



Introduction
Data visualization

The ways we structure and visualize information are changing rapidly and getting more complex with each passing day. **Thanks to the rise of social media, the ubiquity of mobile devices, and service digitalization, data is available on any human activity that utilizes technology**. The generated information is hugely valuable and makes it possible to analyze trends and patterns, and to use big data to draw connections between events. Thus, data visualization can be an effective mechanism for presenting the end user with understandable information in real time.

Every company has data, be it to communicate with clients and senior managers or to help manage the organization itself. It is only through research and interpretation that this data can acquire meaning and be transformed into knowledge.

This ebook seeks to guide readers through a series of basic references in order to help them understand data visualization and its component parts, and to equip them with the tools and platforms they need to create interactive visuals and analyze data. In effect, it seeks to provide readers with a basic vocabulary and a crash course in the principles of design that govern data visualization so that they can create and analyze interactive market research reports.

What is data visualization?

Data visualization is the process of acquiring, interpreting and comparing data in order to clearly communicate complex ideas, thereby facilitating the identification and analysis of meaningful patterns.

> Data visualization can be essential to strategic communication: it helps us interpret available data; detect patterns, trends, and anomalies; make decisions; and analyze inherent processes. All told, it can have a powerful impact on the business world.

The data visualization process

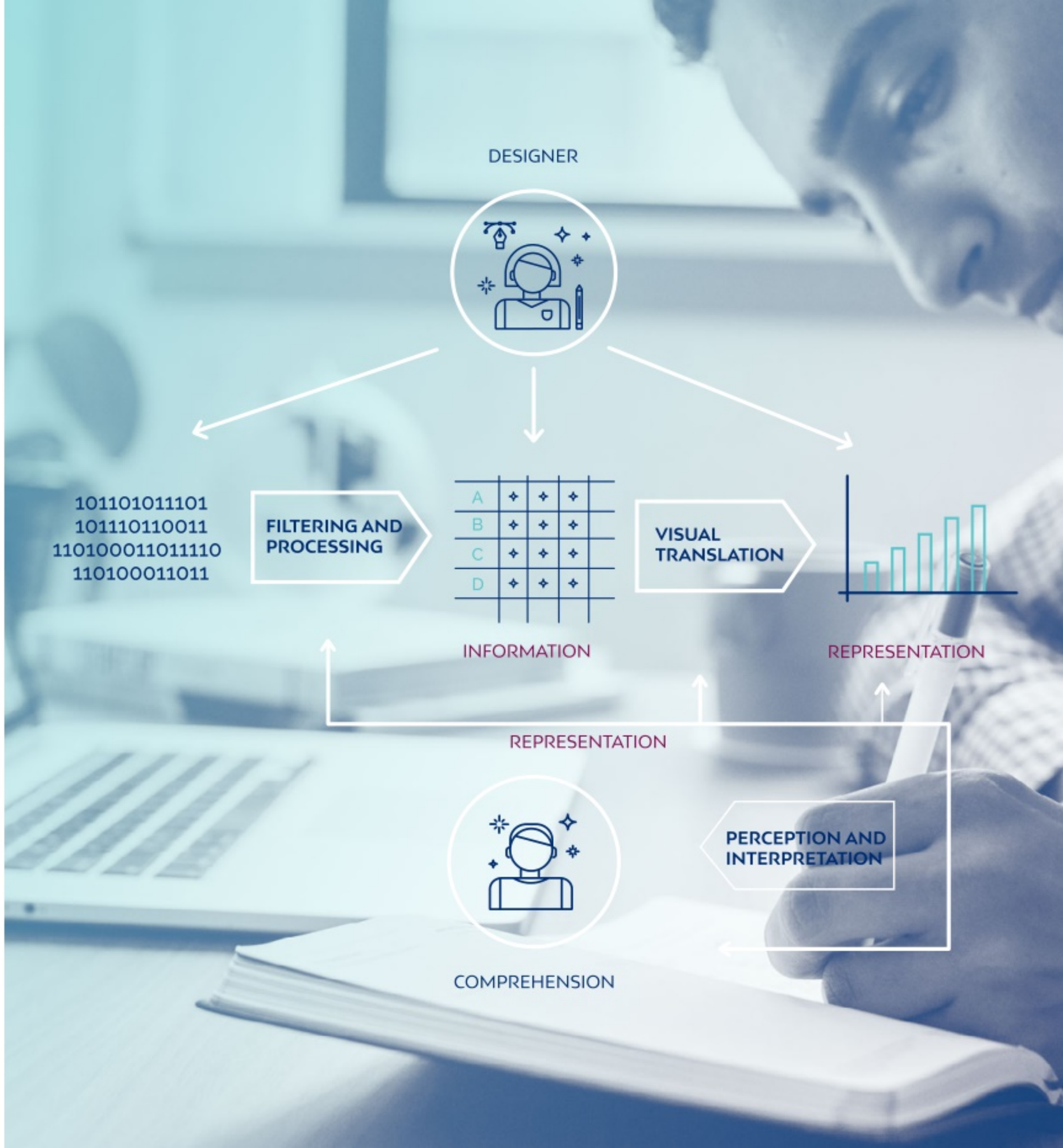
Several different fields are involved in the data visualization process, with the aim of simplifying or revealing existing relationships, or discovering something new within a data set.

Visualization process[1]

Filtering & processing. Refining and cleaning data to convert it into information through analysis, interpretation, contextualization, comparison, and research.

Translation & visual representation. Shaping the visual representation by defining graphic resources, language, context, and the tone of the representation, all of which are adapted for the recipient.

Perception & interpretation. Finally, the visualization becomes effective when it has a perceptive impact on the construction of knowledge.



1 Pérez, J. and Vialcanet, G. (2013). Guía de visualización de datos aplicada al marketing digital: Cómo transformar datos en conocimiento (p.5-6).

Why is data visualization so important in reports and statements?

We live in the era of visual information, and visual content plays an important role in every moment of our lives. A study by SHIFT Disruptive Learning demonstrated that **we typically process images 60,000 times faster than a table or a text,** and that our brains typically do a better job remembering them in the long term. That same research detected that after three days, analyzed subjects retained between 10% and 20% of written or spoken information, compared with 65% of visual information.

The rationale behind the power of visuals:

- The human mind can see an image for just **13 milliseconds** and store the information, provided that it is associated with a concept. Our eyes can take in **36,000 visual messages per hour**.
- **40%** of nerve fibers are connected to the retina.

All of this indicates that human beings are better at processing visual information, which is lodged in our long-term memory.

Consequently, for reports and statements, a visual representation that uses images is a much more effective way to communicate information than text or a table; it also takes up much less space.

This means that **data visuals are more attractive, simpler to take in, and easier to remember.**

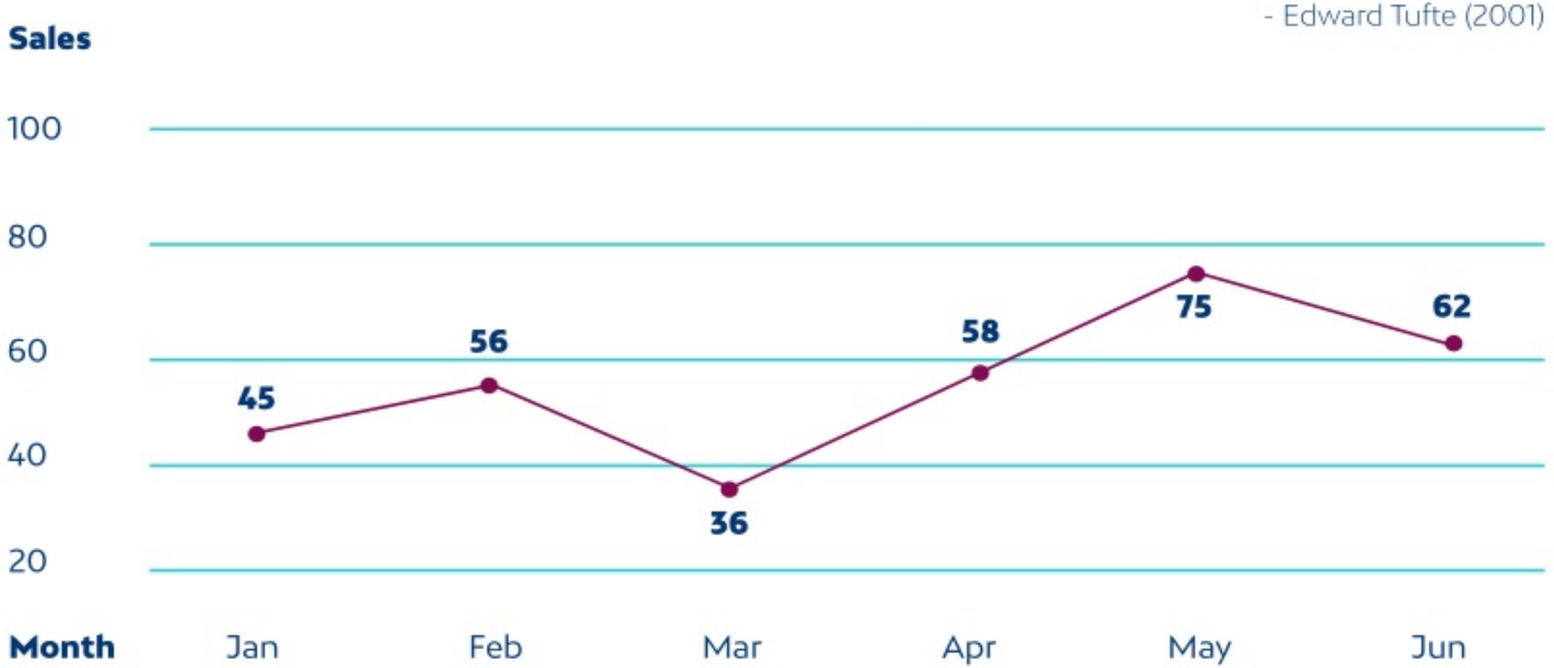
Try it for yourself. Take a look at this table:

Month	Jan	Feb	Mar	Apr	May	Jun
Sales	45	56	36	58	75	62

Identifying the evolution of sales over the course of the year isn't easy. However, when we present the same information in a visual, the results are much clearer (see the graph below).

The graph takes what the numbers cannot communicate on their own and conveys it in a visible, memorable way. This is the real strength of data visualization.

> **"Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space."**
>
> - Edward Tufte (2001)

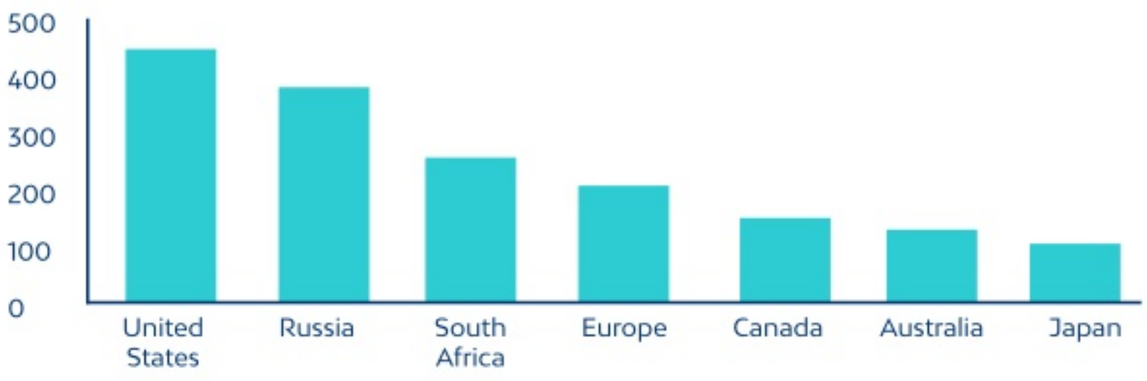


Data visualization chiefly helps in 3 key aspects of reports and statements:

1) Explaining

Visuals aim to lead the viewer down a path in order to describe situations, answer questions, support decisions, communicate information, or solve specific problems. When you attempt to explain something through data visualization, you start with a question, which interacts with the data set in such a way that enables viewers to make a decision and, subsequently, answer the question.

For example: This graphic below could clearly explain the country with the greatest demand for a certain product compared globally, in a concrete month.



2) Exploring

Some visuals are designed to lend a data set spatial dimensions, or to offer numerous subsets of data in order to raise questions, find answers, and discover opportunities. When the goal of a visual is to explore, the viewers start by familiarizing themselves with the dataset, then identifying an area of interest, asking questions, exploring, and finding several solutions or answers.

For example: an interactive graphic from *The Guardian*[2] invites us to explore how the linguistic standard of U.S. presidential addresses has declined over time. The visual is interactive and explanatory, in addition to indicating the readability score of various presidents' speeches.

3) Analyzing

Other visuals prompt viewers to inspect, distill, and transform the most significant information in a data set so that they can discover something new or predict upcoming situations.

For example: this interactive graphic about learning machine[3] invites us to explore and discover information within the visual by scrolling through it. Using the machine learning method, the visual explains the patterns detected in the data in order to categorize characteristics.

We'll close this introduction with a 2012 reflection by Alberto Cairo, a specialist in information visualization and a leader in the world of data visualization. For the author, a good visual must provide clarity, highlight trends, uncover patterns, and reveal unseen realities:

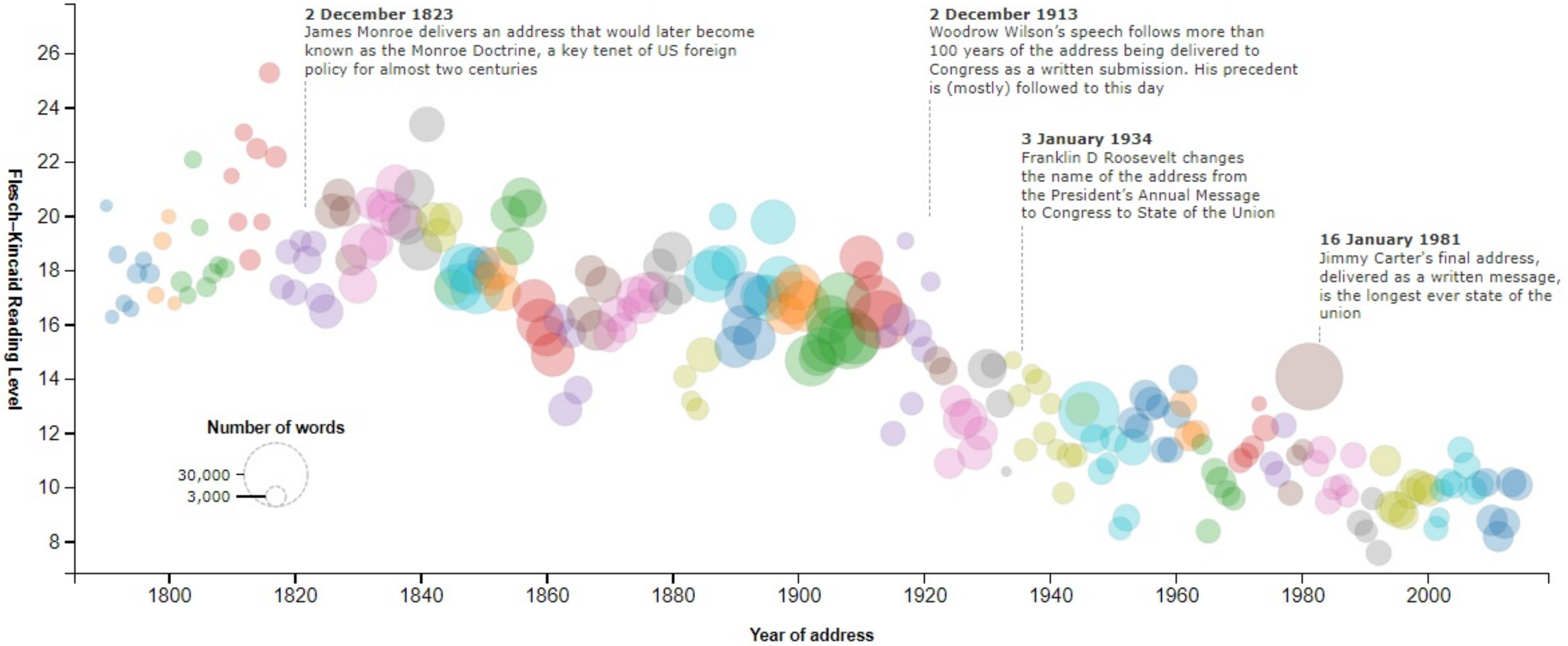
> **We create visuals so that users can analyze data and, from it, discover realities that not even the designer, in some instances, had considered."**

2 Available at: https://www.fusioncharts.com/whitepapers/downloads/Principles-of-Data-Visualization.pdf
3 Available at: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

The state of our union is … dumber:

How the linguistic standard of the presidential address has declined

Using the Flesch-Kincaid readability test the Guardian has tracked the reading level of every State of the Union



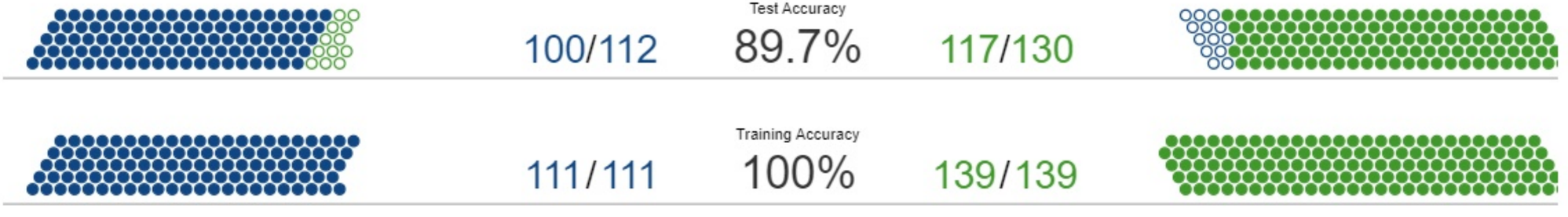
Presidents in order of reading level

Recap

1. Machine learning identifies patterns using **statistical learning** and computers by unearthing **boundaries** in data sets. You can use it to make predictions.

2. One method for making predictions is called a decision trees, which uses a series of if-then statements to identify boundaries and define patterns in the data.

3. **Overfitting** happens when some boundaries are based on on *distinctions that don't make a difference*. You can see if a model overfits by having test data flow through the model.



What is Data Science?

Data science is a deep study of the massive amount of data, which involves extracting meaningful insights from raw, structured, and unstructured data that is processed using the scientific method, different technologies, and algorithm.

It is a multidisciplinary field that uses tools and techniques to manipulate the data so that you can find something new and meaningful.

Data science uses the most powerful hardware, programming systems, and most efficient algorithms to solve the data related problems. It is the future of artificial intelligence.

In short, we can say that data science is all about:

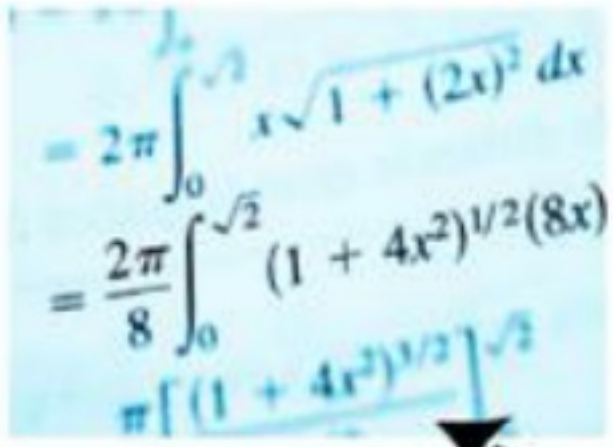
- Asking the correct questions and analyzing the raw data.
- Modeling the data using various complex and efficient algorithms.
- Visualizing the data to get a better perspective.
- Understanding the data to make better decisions and finding the final result.

Example:

Let suppose we want to travel from station A to station B by car. Now, we need to take some decisions such as which route will be the best route to reach faster at the location, in which route there will be no traffic jam, and which will be cost-effective. All these decision factors will act as input data, and we will get an appropriate answer from these decisions, so this analysis of data is called the data analysis, which is a part of data science.

Domain Expertise

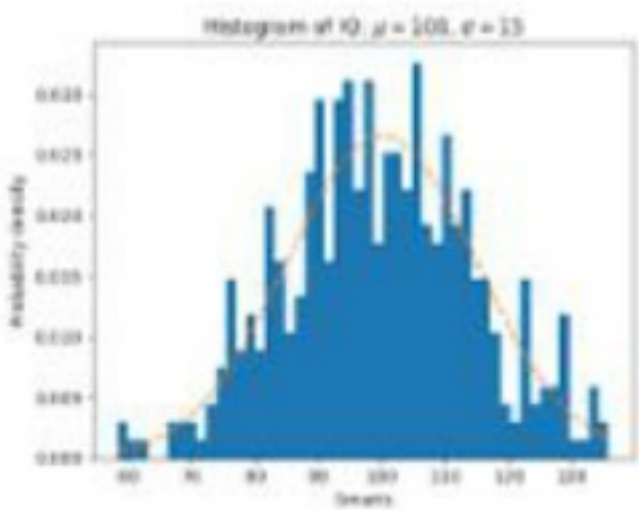
Statistics



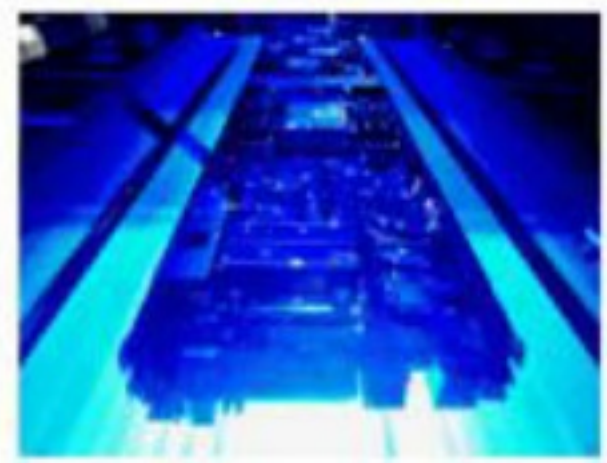
Data Engineering

Data Science

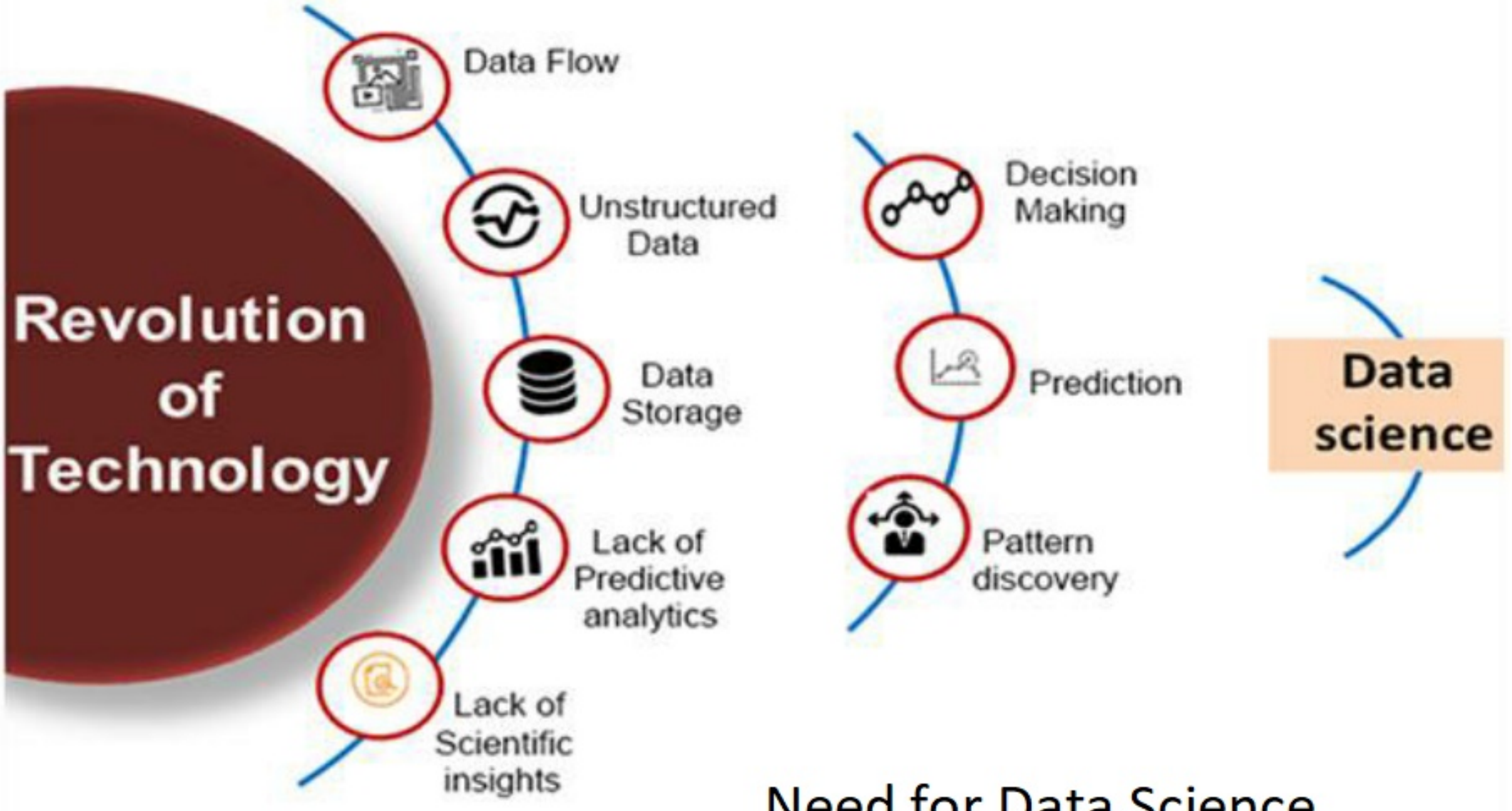
Visualization



Advanced Computing



Applications of Data Science



Need for Data Science

Hype

Hype means "a situation in which something is advertised and discussed in newspapers, on television, etc. a lot in order to attract everyone's interest"

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

Harvard Business Review, https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

16,566 views | Jun 26, 2014, 11:00am

The Hottest Jobs In IT: Training Tomorrow's Data Scientists

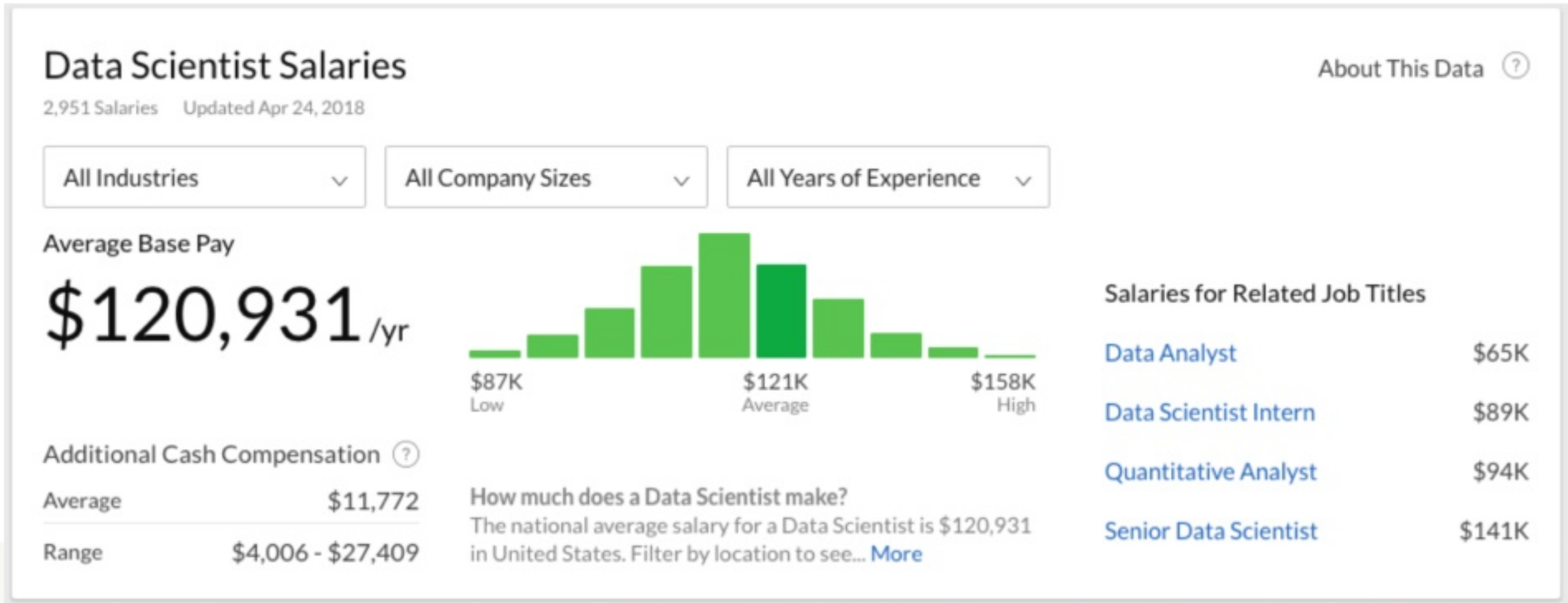
EMC Contributor Brand Contributor
EMC BRANDVOICE

Forbes, https://www.forbes.com/sites/emc/2014/06/26/the-hottest-jobs-in-it-training-tomorrows-data-scientists/

Hype

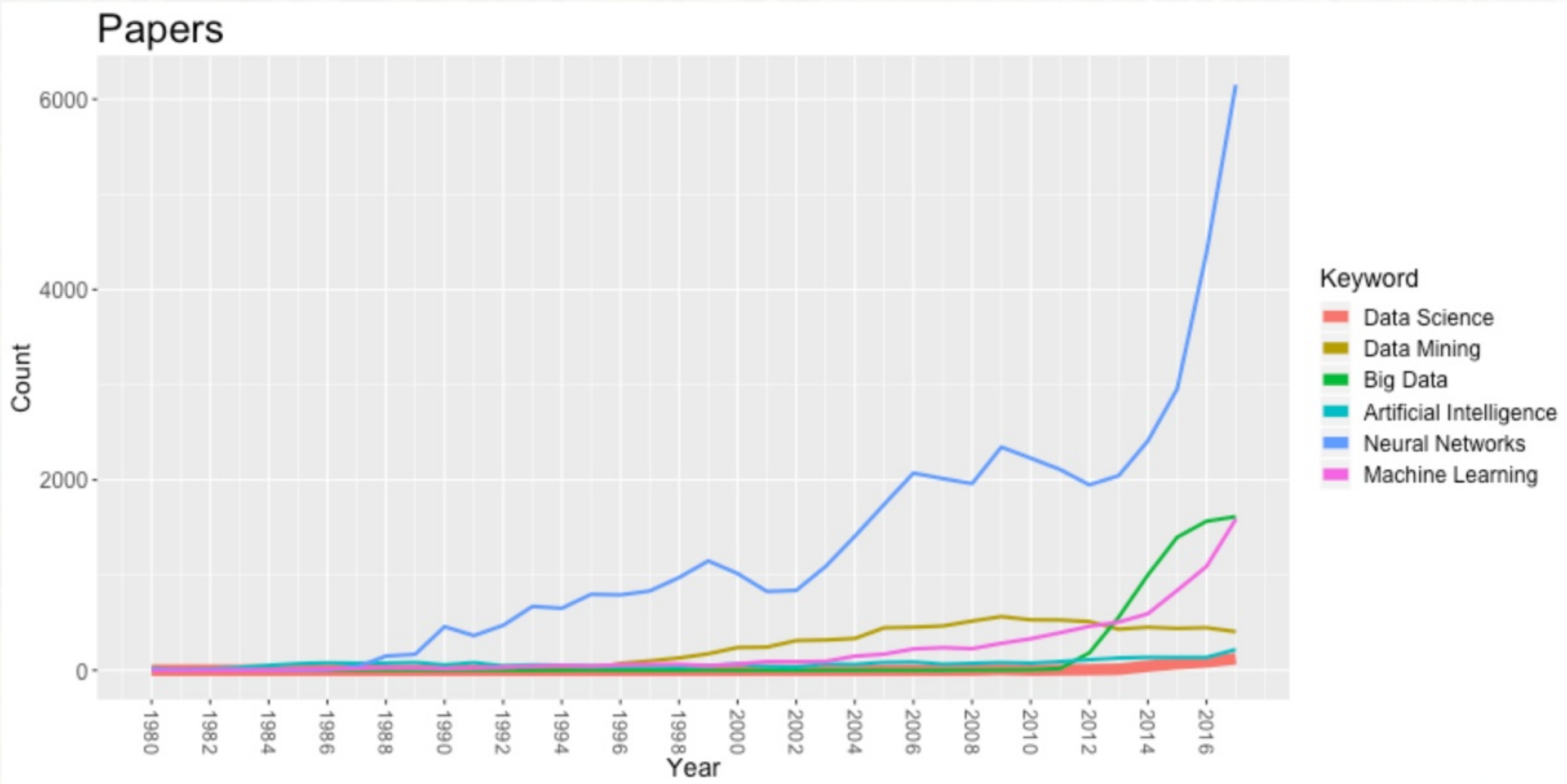
By 2018, the United States will experience a shortage of 190,000 skilled data scientists, and 1.5 million managers and analysts capable of reaping actionable insights from the big data deluge.

Susan Lund et al., "Game Changers: Five Opportunities for US Growth and Renewal," McKinsey Global Institute Report, July 2013. http://www.mckinsey.com/insights/americas/us_game_changers

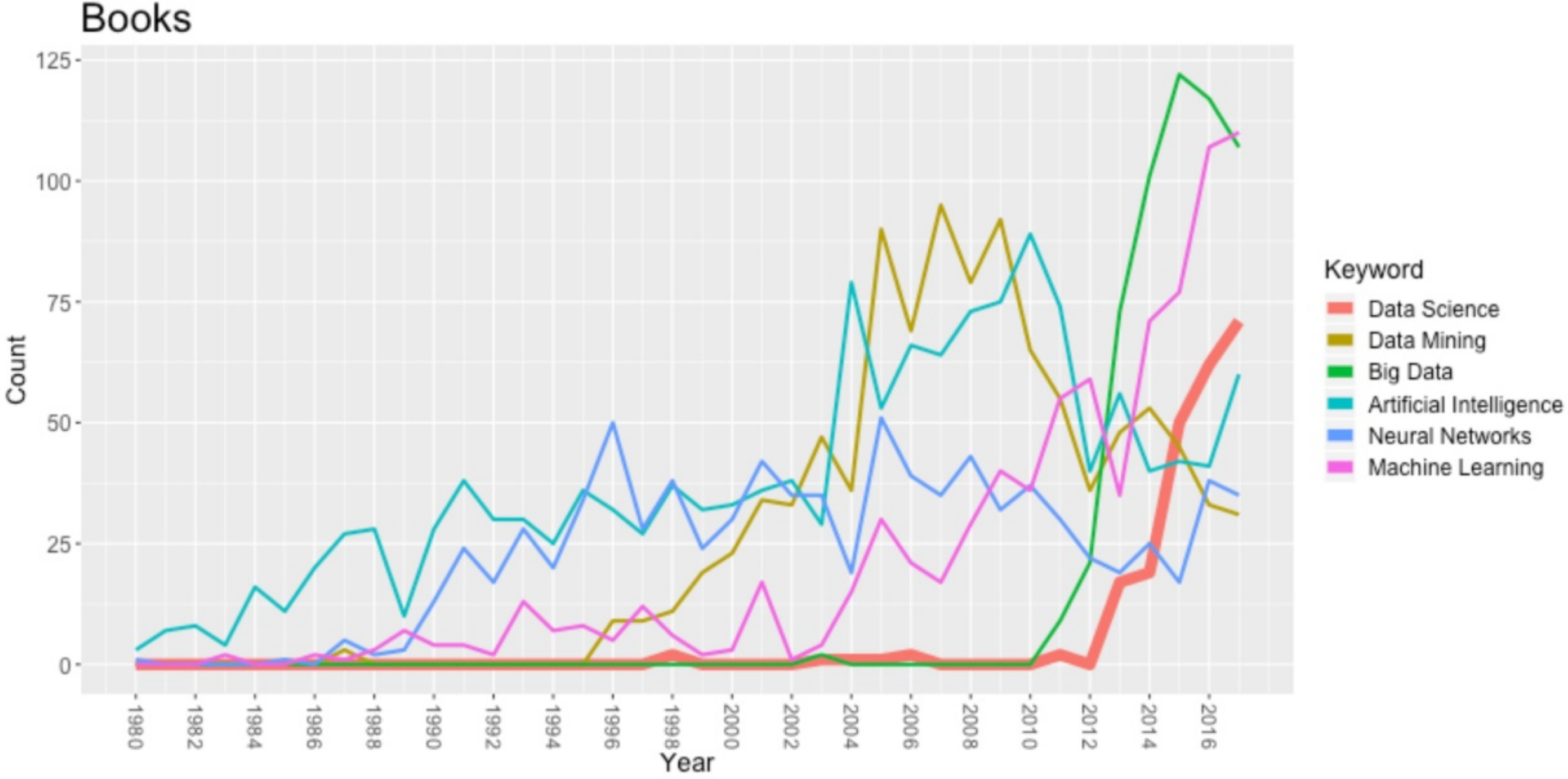


glassdoor.com

Hype


Papers
6000
4000
2000
0
Count
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
Keyword
Data Science
Data Mining
Big Data
Artificial Intelligence
Neural Networks
Machine Learning

Hype


Books
125
100
75
50
25
0
Count
1980
1982
1984
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
Keyword
Data Science
Data Mining
Big Data
Artificial Intelligence
Neural Networks
Machine Learning

Hype

Data science and machine learning are nothing new, but several high-level trends continue to push technologies into the spotlight and generate attention and enthusiasm:

- Growing interest (and hype) around artificial intelligence (AI), fueled by vendor marketing combined with the understandable but erroneous conflation of AI with data science and machine learning.

- The data science and machine-learning talent shortage, and efforts to combat it with education, upskilling and smarter tools using more automation.

- Increases in computing power and availability of advanced system architectures... These advances have also fueled the hype and interest around deep learning.

- The explosion in popularity of open-source tools and libraries for data science and machine learning. The data science and machine-learning market is one of the most vibrant and collaborative technology market that strongly embraces open-source technologies.

Big Data and Data Science Hype

- Big Data, how big?
- Data Science, who is doing it?
- Academia have been doing this for years
- Statisticians have been doing this work.

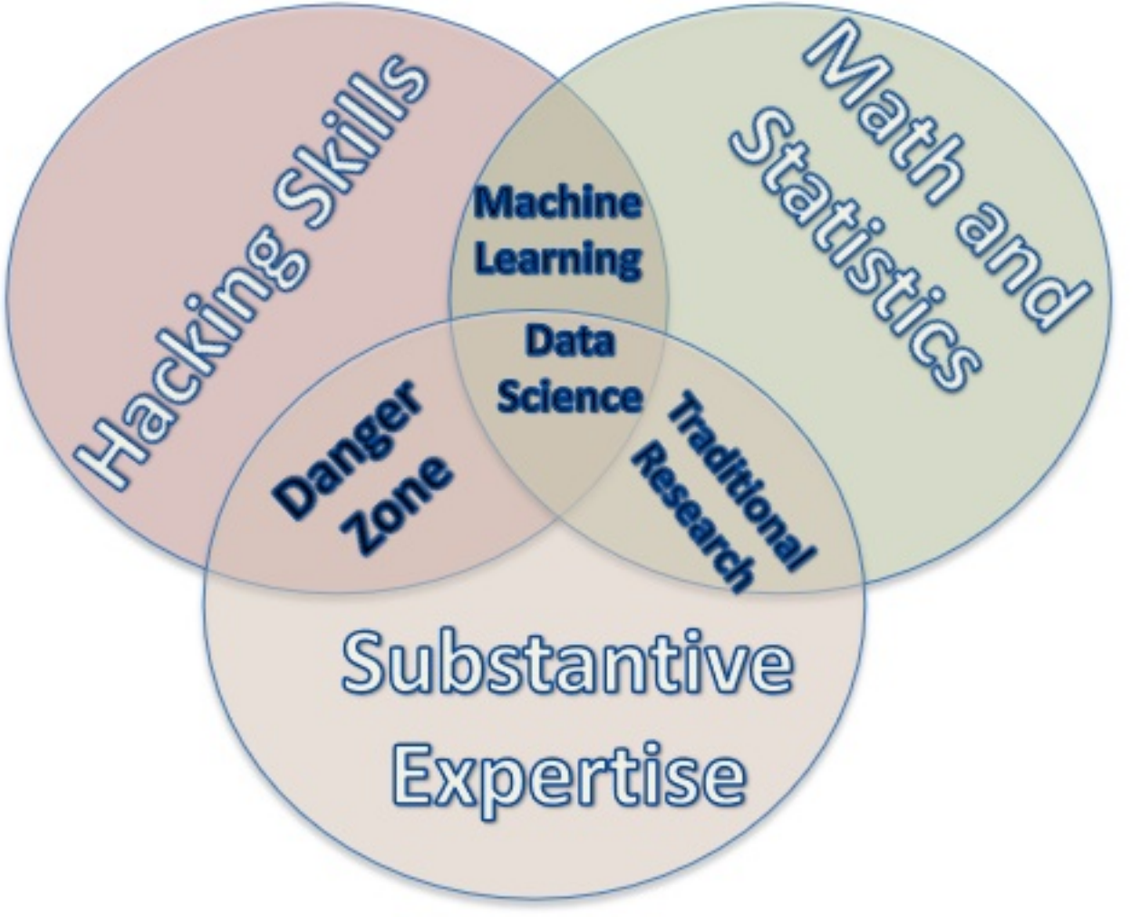
Conclusion: The terms have lost their basic meaning and now are too ambiguous, thus, today they are now meaningless.

Getting Past the Hype / Why Now

- **The Hype:** Understanding the cultural phenomenon of data science and how others were experiencing it. Study how companies, and universities are “doing data science”.
- **Why Now:** Technology makes this possible: infrastructure for large-scale data processing, increased memory, and bandwidth, as well as a cultural acceptance of technology in the fabric of our lives. This wasn't true a decade ago.
- Consideration should be to the ethical and technical responsibilities for the people responsible for the process.

Current Landscape of Data Science

- Drew Conway's Venn diagram of data science from 20l0,



Data science Jobs:

As per various surveys, data scientist job is becoming the most demanding Job of the 21st century due to increasing demands for data science. Some people also called it "the **hottest job title of the 21st century**". Data scientists are the experts who can use various statistical tools and machine learning algorithms to understand and analyze the data.

Types of Data Science Job

If you learn data science, then you get the opportunity to find the various exciting job roles in this domain. The main job roles are given below:

1. Data Scientist
2. Data Analyst
3. Machine learning expert
4. Data engineer
5. Data Architect

Applications of Data Science:

- Image recognition and speech recognition
- Gaming world
- Internet search
- Transport
- Healthcare
- Risk detection

Skills Required For Data Scientists

The 2 types of important skills are:

1. Technical
2. Non-technical skills

Technical Skills Required For Data Scientists

Some of the most important technical data scientist skills are:

- Statistical analysis and computing
- Machine Learning
- Deep Learning
- Processing large data sets
- Data Visualization
- Data Wrangling
- Mathematics
- Programming
- Statistics
- Big Data

Some data scientists have a Ph.D. or Master's degree in statistics, computer science, or engineering. This educational background provides a strong foundation for any aspiring data scientist and also teaches the essential data scientist skills and Big Data skills needed to succeed in the field, including:

There are some schools that now offer specialized programs tailored to the educational requirements for pursuing a career in data science, giving students the option to focus on the field of study they are most interested in, and in a shorter period of time.

Some of the many options available include Massive Open Online Courses (MOOCs) or bootcamps, such as Simplilearn's Big Data & Analytics certification courses. These types of programs offer practical learning methods that you will not find in the confines of the textbook, including a hands-on approach to learning in-demand data science skills, Capstone projects, and other exercises that help prepare students to become data scientists.

Other technical data scientist skills required include:

1.Programming

You need to have knowledge of various programming languages, such as Python, Perl, C/C++, SQL, and Java, with Python being the most common coding language required in data science roles. These programming languages help data scientists organize unstructured data sets.

2.Knowledge of SAS and Other Analytical Tools

Understanding analytical tools is one of the most helpful data scientist skills for extracting valuable information from an organized data set. SAS, Hadoop, Spark, Hive, Pig, and R are the most popular data analytical tools that data scientists use. Certifications can help you establish your expertise in these analytical tools and help you gain this valuable data science skill!

3. Adept at Working with Unstructured Data

Data scientists should have experience working with unstructured data that comes from different channels and sources. For example, if a data scientist is working on a project to help the marketing team provide insightful research, the professional should be well adept at handling social media as well.Some of the other data scientist skills required are Machine Learning, Artificial intelligence, Deep learning, Probability and Statistics.Moving forward, let's discuss the non-technical skills.

4.Web Scraping

Web scraping is the automated process of extracting data from webpages.

5. ML with AI and DL with NLP:

Deep learning (DL) with natural language processing (NLP) focuses on using neural networks to process and understand human language. Machine learning (ML) and artificial intelligence (AI) are both concerned with teaching computers to learn from data.

6. Problem-Solving Skills:

Skills for Solving Issues the capacity to evaluate challenging issues and develop workable answers.

7. Probability and Statistics:

Statistics and probability is the study of randomness and uncertainty in statistics, and the application of mathematical tools to decision-making.

8. Multivariate Calculus and Linear Algebra:

Advanced mathematical ideas used in machine learning and data analysis include multivariate calculus and linear algebra.

9. Database Management:

The procedure of arranging, saving, and accessing data in a database system is known as database management.

10. Cloud Computing: Utilizing remote servers to store, control, and handle data and applications online is known as cloud computing

11. Microsoft Excel: Microsoft Excel is a spreadsheet program used for data display and analysis.

12. DevOps:

A technique of developing software that places a strong emphasis on teamwork and communication between the development and operations teams.

13. Data Extraction, Transformation, and Loading:

Data collection, cleansing, and preparation for analysis is known as data extraction, transformation, and loading.

14. Business Intelligence:

Business intelligence is the process of using tools and techniques for data analysis to acquire knowledge and guide business decisions.

15. Neural Networks:

A data scientist should possess skills in designing, training, and fine-tuning neural networks for various use cases, as well as knowledge of different neural network architectures and frameworks.

16. Model Deployment:

Data scientists need expertise in model deployment, which involves making trained machine-learning models available for use in production environments.

17. Data Structures and Algorithms:

The fundamental ideas in computer science that underpin effective data storage, retrieval, and computational problems are known as data structures and algorithms.

Non-Technical Skills Required For Data Scientists

Along with the technical data scientist skills, we will now shift our focus on non-technical skills that are required to become a data scientist. These refer to personal skills and can be difficult to assess simply by looking at educational qualifications, certifications, and so on. They include:

18. A Strong Business Acumen

The best way to productively channel technical skills is to have strong business acumen. Without it, an aspiring data scientist may not be able to discern the problems and potential challenges that need to be solved in order for an organization to grow. This is essential for helping the organization you're working for explore new business opportunities.

19. Strong Communication Skills

Next on the list of top data scientist skills is communication. Data scientists clearly understand how to extract, understand, and analyze data. However, for you to be successful in your role, and for your organization to benefit from your services, you should be able to successfully communicate your findings with team members who don't have the same professional background as you.

20. Great Data Intuition

This is perhaps one of the most significant non-technical data scientist skills. Valuable data insights are not always apparent in large data sets, and a knowledgeable data scientist has intuition and knows when to look beyond the surface for insightful information. This makes data scientists more efficient in their work, and gaining this skill comes from experience and the right training. However, this data scientist skill comes with experience and bootcamps are a great way of polishing it.

21. Analytical Mindset:

The capacity to dissect complicated issues into their component parts, analyze those parts, and derive conclusions from the data.

22. "Out-of-the-Box" Thinking:

Using creative and innovative thinking to generate novel ideas and unconventional answers.

23. Critical Thinking:

The process of evaluating and analyzing data in order to make a judgment or choice is known as critical thinking.

24. Decision Making:

Making decisions entails choosing the best course of action from a range of alternatives after carefully weighing all pertinent information.