DATA SCIENCE AND VISUALIZATION			
Course Code	21CS644	CIE Marks	50
Teaching Hours/Week (L:T:P: S)	3:0:0:0	SEE Marks	50
Total Hours of Pedagogy	40	Total Marks	100
Credits	03	Exam Hours	03

Course Learning Objectives

- CLO 1. To introduce data collection and pre-processing techniques for data science
- CLO 2. Explore analytical methods for solving real life problems through data exploration techniques
- CLO 3. Illustrate different types of data and its visualization
- CLO 4. Find different data visualization techniques and tools
- CLO 5. Design and map element of visualization well to perceive information

Syllabus

- **Module1: Introduction to Data Science**
- **Module2:Exploratory Data Analysis and the Data Science Process**
- **Module3:Feature Generation and Feature Selection Extracting Meaning from Data**
- **Module4: Data Visualization and Data Exploration**
- **Module5: A Deep Dive into Matplotlib Introduction**

Syllabus

- **Module1**: **Introduction to Data Science Introduction**: What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? – Datafication, Current landscape of perspectives, Skill sets. Needed Statistical Inference: Populations and samples, Statistical modelling, probability distributions, fitting a model.

- **Module2:Exploratory Data Analysis and the Data Science Process**: Basic tools (plots, graphs and summary statistics) of EDA, Philosophy of EDA, The Data Science Process, Case Study: Real Direct (online real estate firm). Three Basic Machine Learning Algorithms: Linear Regression, k-Nearest Neighbours (k- NN), k-means.

- **Module3:Feature Generation and Feature Selection Extracting Meaning from Data**: Motivating application: user (customer) retention. Feature Generation (brainstorming, role of domain expertise, and place for imagination), Feature Selection algorithms. Filters; Wrappers; Decision Trees; Random Forests. Recommendation Systems: Building a User-Facing Data Product, Algorithmic ingredients of a Recommendation Engine, Dimensionality Reduction, Singular Value Decomposition, Principal Component Analysis, Exercise: build your own recommendation system.

Syllabus

- **Module4: Data Visualization and Data Exploration Introduction**: Data Visualization, Importance of Data Visualization, Data Wrangling, Tools and Libraries for Visualization Comparison Plots: Line Chart, Bar Chart and Radar Chart; Relation Plots: Scatter Plot, Bubble Plot , Correlogram and Heatmap; Composition Plots: Pie Chart, Stacked Bar Chart, Stacked Area Chart, Venn Diagram; Distribution Plots: Histogram, Density Plot, Box Plot, Violin Plot; Geo Plots: Dot Map, Choropleth Map, Connection Map; What Makes a Good Visualization?
- **Module5: A Deep Dive into Matplotlib Introduction**, Overview of Plots in Matplotlib, Pyplot Basics: Creating Figures, Closing Figures, Format Strings, Plotting, Plotting Using pandas DataFrames, Displaying Figures, Saving Figures; Basic Text and Legend Functions: Labels, Titles, Text, Annotations, Legends; Basic Plots:Bar Chart, Pie Chart, Stacked Bar Chart, Stacked Area Chart, Histogram, Box Plot, Scatter Plot, Bubble Plot; Layouts: Subplots, Tight Layout, Radar Charts, GridSpec; Images: Basic Image Operations, Writing Mathematical Expressions

Course Outcomes

At the end of the course the student will be able to:

- **CO 1**. Understand the data in different forms
- **CO 2**. Apply different techniques to Explore Data Analysis and the Data Science Process
- **CO 3**. Analyze feature selection algorithms & design a recommender system.
- **CO 4**. Evaluate data visualization tools and libraries and plot graphs.
- **CO 5**. Develop different charts and include mathematical expressions.

Assessment Details (both CIE and SEE)

- The weightage of Continuous Internal Evaluation (CIE) is 50% and for Semester End Exam (SEE) is 50%.
- The minimum passing mark for the CIE is 40% of the maximum marks (20 marks).
- The minimum passing mark is 35% (18 Marks out of 50) in the semester-end examination (SEE), and a
- Minimum of 40% (40 marks out of 100) in the sum total of the CIE (Continuous Internal Evaluation) and SEE (Semester End Examination) taken together .

Continuous Internal Evaluation:

- **Three Unit Tests each of 20 Marks (duration 01 hour)**
  - 1. First test at the end of 5th week of the semester
  - 2. Second test at the end of the 10th week of the semester
  - 3. Third test at the end of the 15th week of the semester
- **Two assignments each of 10 Marks**
  - 4. First assignment at the end of 4th week of the semester
  - 5. Second assignment at the end of 9th week of the semester
- 6. At the end of the 13th week of the semester **Group discussion/Seminar/quiz** any one of three suitably planned to attain the COs and POs for 20 Marks (duration 01 hours)
- The sum of three tests, two assignments, and quiz/seminar/group discussion will be out of 100 marks and will be scaled down to 50 marks

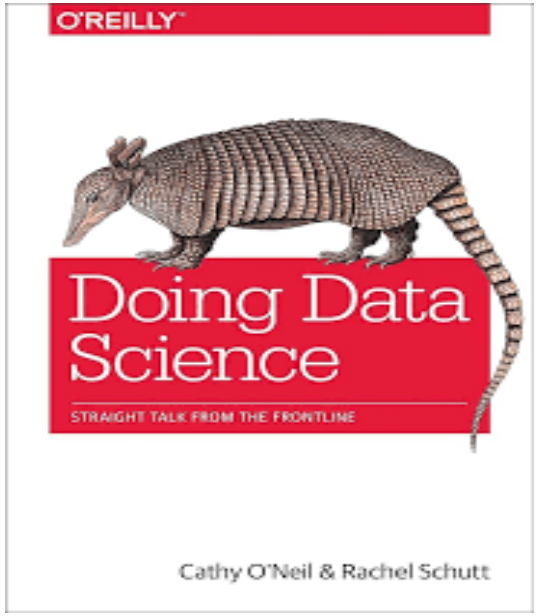
Semester End Examination:

- Theory SEE will be conducted by University as per the scheduled timetable, with common question papers for the subject (duration 03 hours) Max Marks: 100
  - 1. The question paper will have ten questions. Each question is set for 20 marks. Marks scored shall be proportionally reduced to 50 marks
  - 2. There will be 2 questions from each module. Each of the two questions under a module (with a maximum of 3 sub-questions), should have a mix of topics under that module.

The students have to answer 5 full questions, selecting one full question from each module.

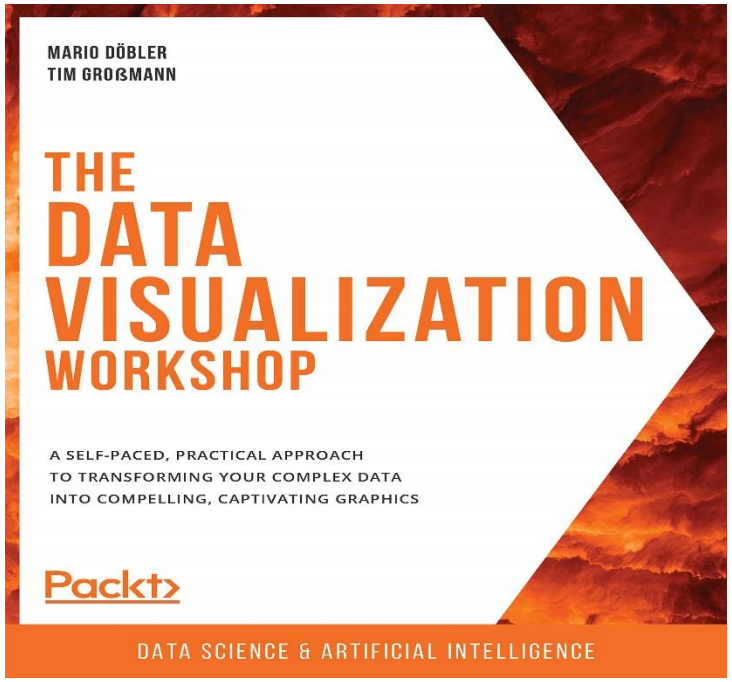
Textbooks

- 1. Doing Data Science, Cathy O’Neil and Rachel Schutt, O'Reilly Media, Inc O'Reilly Media, Inc, 2013



Textbooks

- 2. **Data Visualization workshop**, Tim Grobmann and Mario Dobler, Packt Publishing, ISBN 9781800568112



Reference:

- 1. Mining of Massive Datasets, Anand Rajaraman and Jeffrey D. Ullman, Cambridge University Press, 2010
- 2. Data Science from Scratch, Joel Grus, Shroff Publisher /O'Reilly Publisher Media
- 3. A handbook for data driven design by Andy krik

Weblinks and Video Lectures (e-Resources):

- 1. https://nptel.ac.in/courses/106/105/106105077/
- 2. https://www.oreilly.com/library/view/doing-data-science/9781449363871/toc01.html
- 3. http://book.visualisingdata.com/
- 4. https://matplotlib.org/
- 5. https://docs.python.org/3/tutorial/
- 6. https://www.tableau.com/

Module1 : Introduction to Data Science and Statistical Inference Needed:

1. **Introduction** : What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? Datafication, Current landscape of perspectives, Skill sets.
2. **Needed Statistical Inference**: Populations and samples, Statistical modelling, probability distributions, fitting a model.

Background of Book

- Cathy and Rachel were unsure and puzzled about all the excitement around **data science and Big Data in the media**.
- They discussed their confusion over breakfast, wondering if this new trend might be important and match their skills.
- Instead of ignoring it, they decided to learn more.
- Eventually, **Rachel started** teaching a data science course at Columbia University, **Cathy wrote** about it on her blog, and now you're reading a book based on their experiences.

Big Data and Data Science Hype:

Here are five reasons why **Big Data** and **Data Science** have caused a lot of excitement:

1. **Confusing Terminology**
2. **Lack of Recognition for Previous Work**
3. **Exaggerated Hype**
4. **Confusion with Statistics**
5. **Debates about Science vs. Craft**

Getting Past the Hype

- **Rachel** studied statistics in school and then started working at Google. Initially didn't believe about data science hype, Rachel found her job at Google revealed its validity.
- School laid a foundation, but c**oding, data visualization, and domain knowledge were crucial** for her role, highlighting a significant gap between academia and industry.
- This made her curious about a new field called **data science**. Data science combines **statistics and computer science** to solve real-world problems using data.
- Rachel investigated this field by talking to people at Google, start-ups, and universities. She even taught a course on data science to understand it better.

Getting Past the Hype (Continued--)

- From those meetings she started to form a clearer picture of the new thing that's emerging. She ultimately decided to continue the investigation by giving a course at Columbia called "**Introduction to Data Science**," which Cathy covered on her blog.
- The author believes data science is a real, new field that brings together different subjects and has a new way of working with data
- Now, ***Rachel and Cathy*** want to share their knowledge about data science with many more people through this book.

Why now the Data Science is Emerging?

The two key factors that explains the significance of the current time period for the emergence of data science are:

1. **Availability of massive amounts of data:**
   - Our online activities like shopping, communicating, reading news, listening to music, searching for information, and expressing opinions are being tracked and generating a lot of data.
   - Additionally, there is increasing "datafication" of our offline behaviors as well, mirroring the online data collection.
   - Data is being collected across various sectors like finance, healthcare, pharmaceuticals, bioinformatics, social welfare, government, education, retail, and many others.
2. **Abundance of inexpensive computing power:**
   - Along with the availability of massive amounts of data, there is also an abundance of inexpensive computing power to process and analyze this data.

Datafication and its Implications

- **Datafication** is defined as the process of taking all aspects of life and turning them into data. It involves quantifying and recording various activities and behaviors, both online and offline, for later examination and analysis.

Examples of datafication include:

- "Liking" something on social media platforms
- Google's augmented reality glasses recording what people look at
- Twitter capturing people's thoughts
- LinkedIn capturing professional networks

Datafication and its Implications (Continued)

Spectrum of intentionality in datafication:

- People **intentionally** participate in some forms of datafication, like social media interactions.
- People **unintentionally or passively** get datafied through activities like browsing the web (via cookies) or walking in public spaces (via sensors and cameras).

"**Once we datafy things**, we can transform their purpose and turn the information into new forms of value."

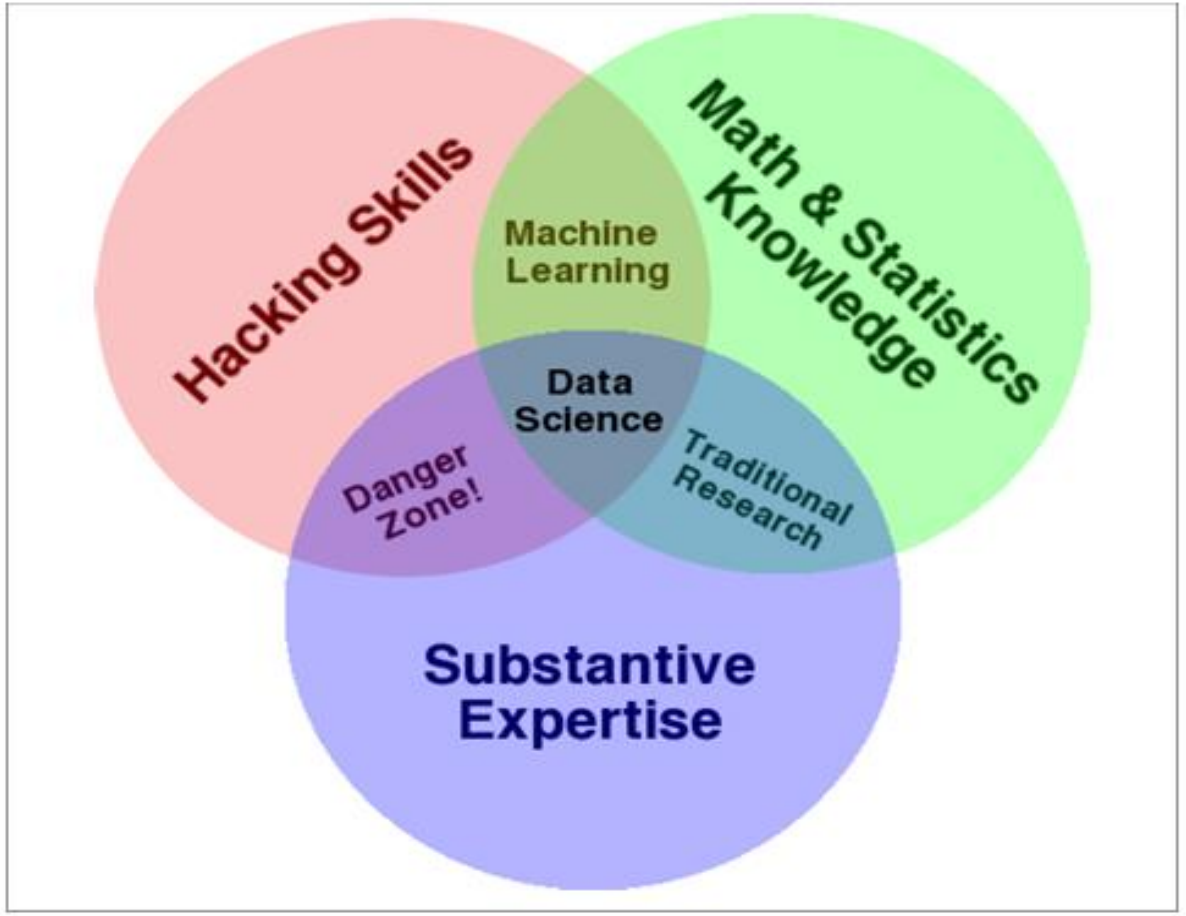
- Here "we" likely refers to **modelers** and **entrepreneurs** who monetize the data for purposes like **targeted advertising** and increased efficiency through automation.
- It also raises the concern that if "we" is meant to refer to people in general, this perspective of datafication **creating value primarily for commercial interests** goes against the **broader interests of society**.

What is Data Science? The Current Landscape (with a Little History)/Different Perspectives

- What data science is and whether it is distinct from statistics/machine learning.

What is Data Science? The Current Landscape (with a Little History)/Different Perspectives

- The perspective from the Quora discussion and blog posts that view data science as encompassing skills like *statistics, data munging (parsing/formatting data), coding, visualization, etc.* Suggesting it is a broad set of skills and techniques.
- Driscoll then refers to Drew Conway's Venn diagram of data science from 2010, shown in Figure below:



What is Data Science? The Current Landscape (with a Little History)/Different Perspectives

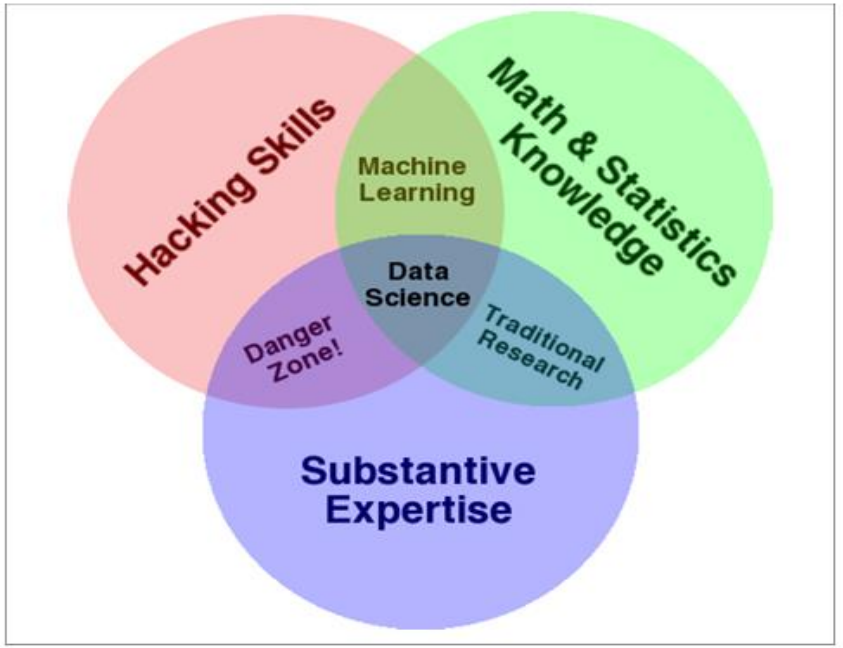
- One view, argued by statistician **Cosma Shalizi**, is that data science is just **statistics rebranded**, and any good **statistics department** already does what data science claims to do.

What is Data Science? The Current Landscape (with a Little History)/Different Perspectives

- The perspective that "data scientist" emerged as a **new job title** in 2008 at companies like **LinkedIn and Facebook** to describe a hybrid skillset combining *statistics, computer science, curiosity and persistence* for working on data problems
- William Cleveland's 2001 position paper viewing **"data science**" as an **action plan to expand the field of statistics, suggesting data science** existed conceptually before the job title.i.e in the high-tech industry.

The Role of the Social Scientist in Data Science

- Providing expertise in understanding and **analyzing human/user behavior data**, which was particularly relevant for early data science roles at social network companies **like LinkedIn and Facebook**.
- Contributing the **"substantive expertise"** component required for data science problems that involve social phenomena, as represented in Drew Conway's data science Venn diagram.[Fig1]



The Role of the Social Scientist in Data Science

- Asking ***good investigative questions and having strong inquiry skills***, which are valuable qualities for a data scientist.
- Bringing a combination of ***quantitative, programming, and social science skills***, which can make them well-suited for data science roles focused on social science-related problems.
- Potentially being a part of the emerging field of ***"computational social sciences,"*** which is described as a subset of data science focused on social phenomena.

DATA SCIENCE AND VISUALIZATION			
Course Code	21CS644	CIE Marks	50
Teaching Hours/Week (L:T:P: S)	3:0:0:0	SEE Marks	50
Total Hours of Pedagogy	40	Total Marks	100
Credits	03	Exam Hours	03

Course Learning Objectives

- CLO 1. To introduce data collection and pre-processing techniques for data science
- CLO 2. Explore analytical methods for solving real life problems through data exploration techniques
- CLO 3. Illustrate different types of data and its visualization
- CLO 4. Find different data visualization techniques and tools
- CLO 5. Design and map element of visualization well to perceive information

Syllabus

- **Module1**:
  - **Introduction to Data Science and**
  - **Statistical Inference**
- **Module2:**
  - **Exploratory Data Analysis and**
  - **The Data Science Process**
- **Module3:**
  - **Feature Generation and Feature Selection**
  - **Extracting Meaning from Data**
- **Module4:**
  - **Introduction to Data Visualization and Data Exploration**
  - **Different types of Plots**
- **Module5: A Deep Dive into Matplotlib Introduction**

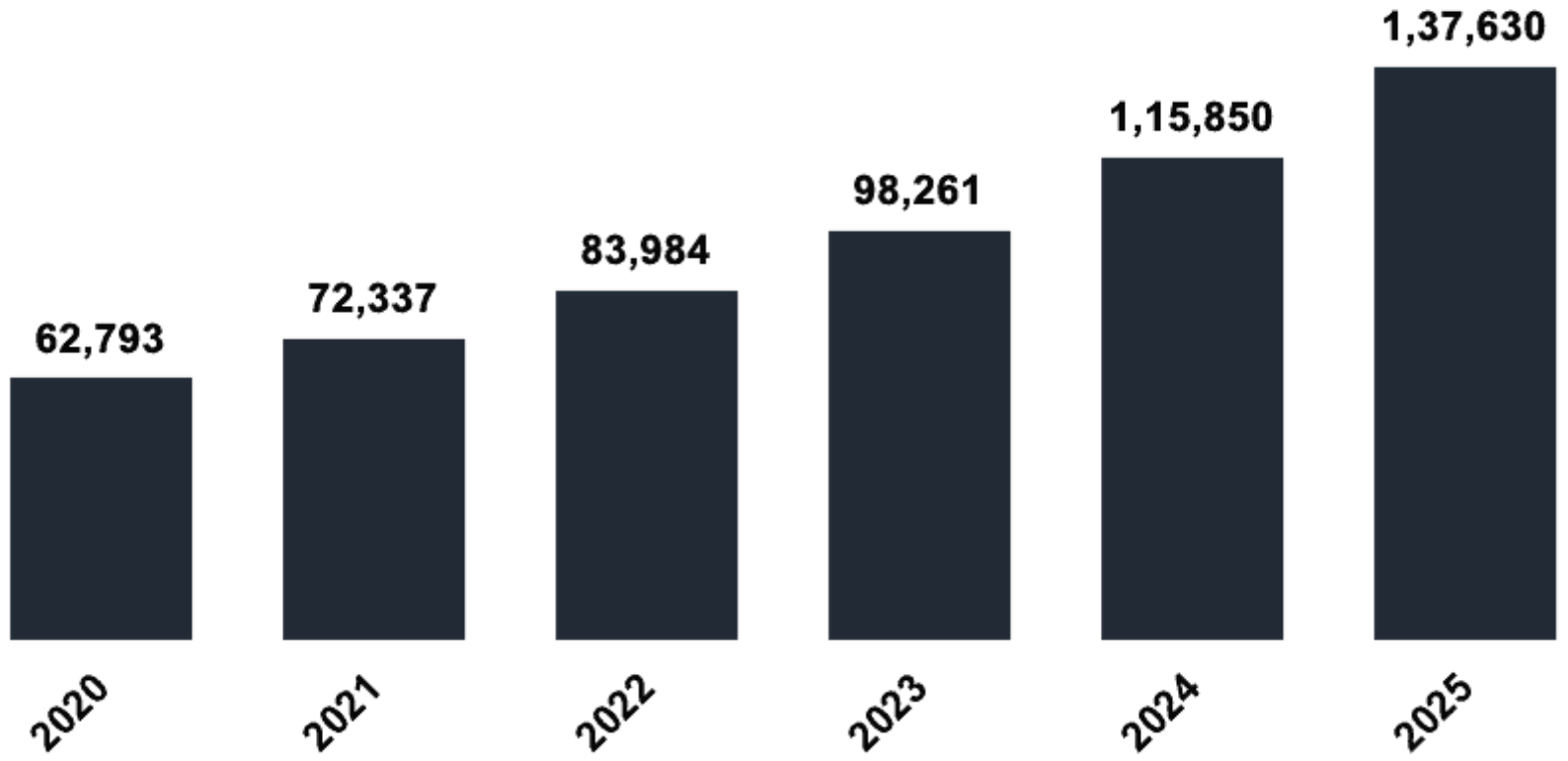
Module1 : Introduction to Data Science and Statistical Inference Needed:

1. **Introduction** : What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? Datafication, Current landscape of perspectives, Skill sets.
2. **Needed Statistical Inference**: Populations and samples, Statistical modelling, probability distributions, fitting a model.

Data Scientist Jobs [During 2010 – 2014]

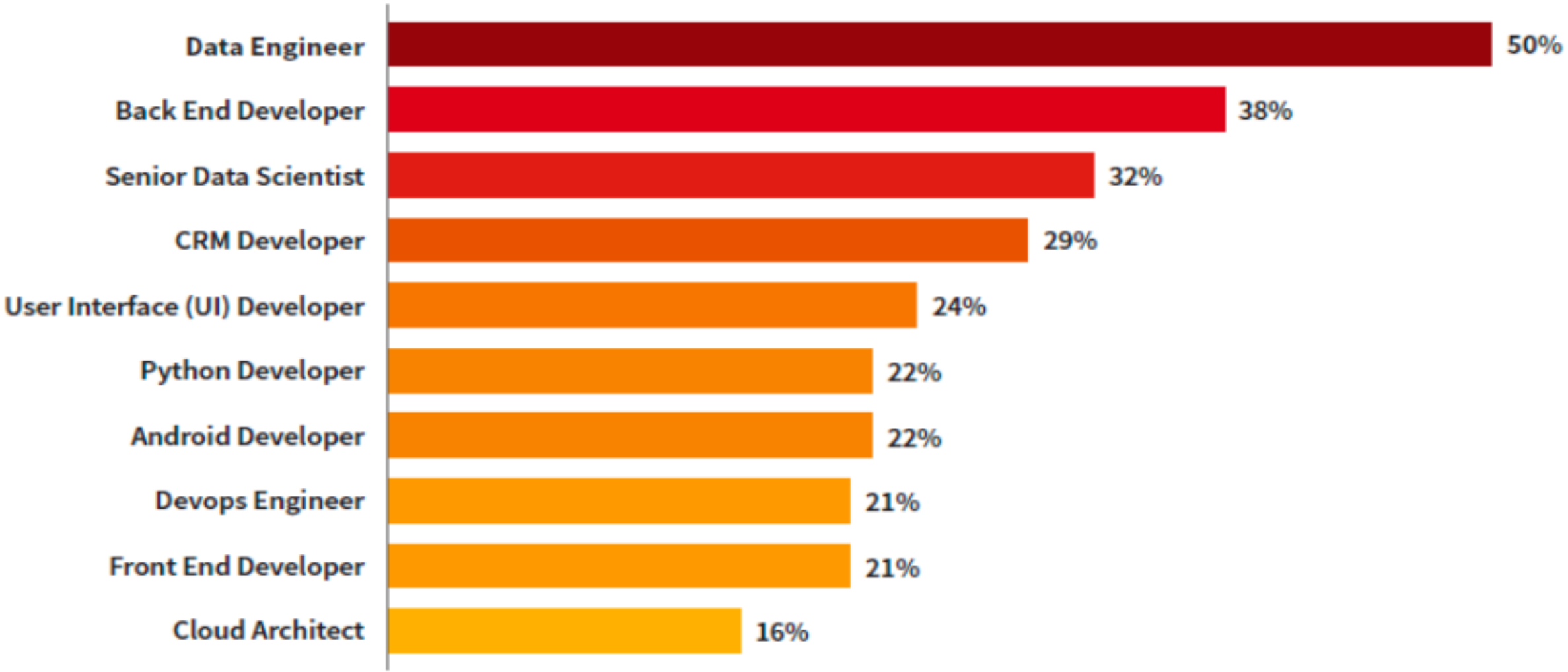
- The specific mention of **465** job openings for data scientists in New York City alone at the time the content was written. This number is described as "**a lot**", highlighting the significant demand for data science roles in just one major city.
- The observation that most job descriptions for data scientists require expertise in multiple areas such as computer science, statistics, communication, data visualization, and extensive domain expertise. This implies a high demand for professionals with a diverse and multidisciplinary skillset

Data Science Jobs in India



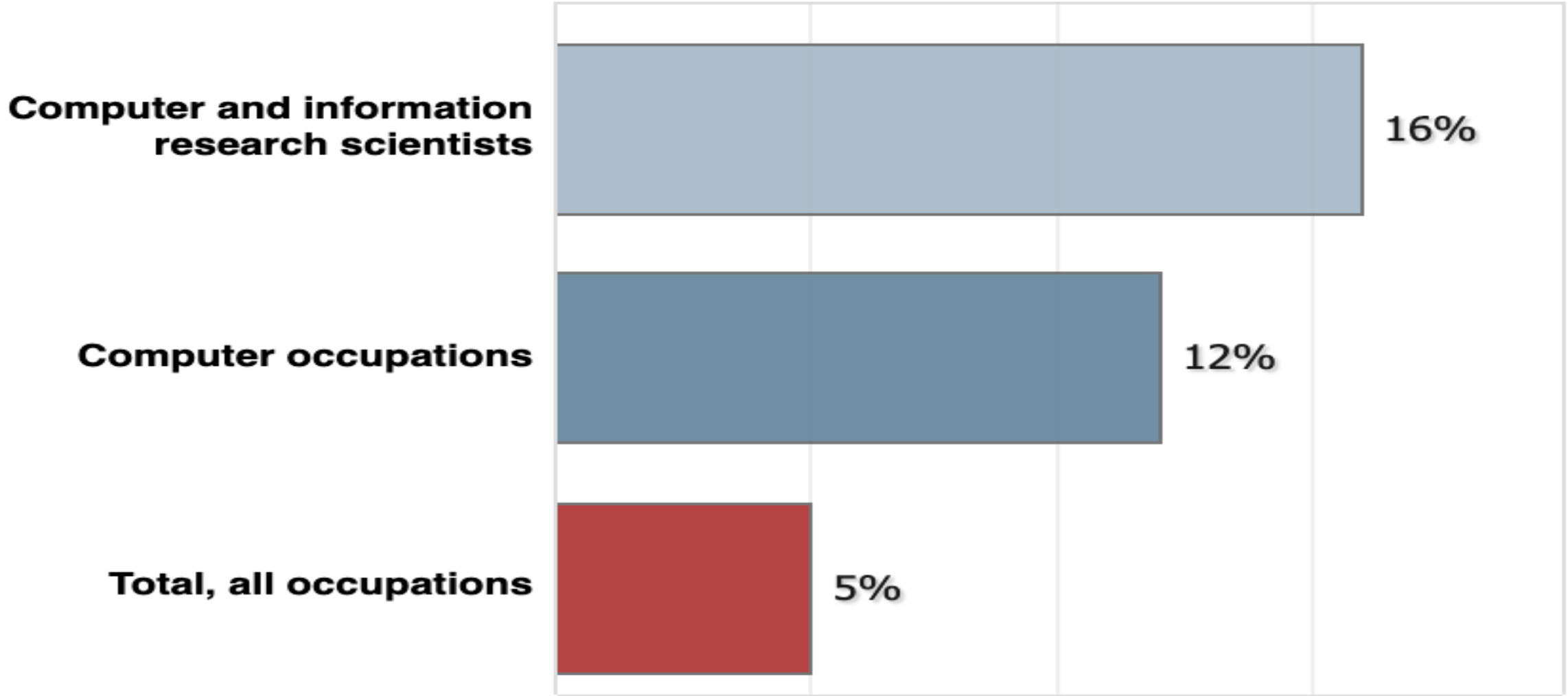
FASTEST GROWING TECH OCCUPATIONS

YEAR-OVER-YEAR GROWTH



Computer and Information Research Scientists

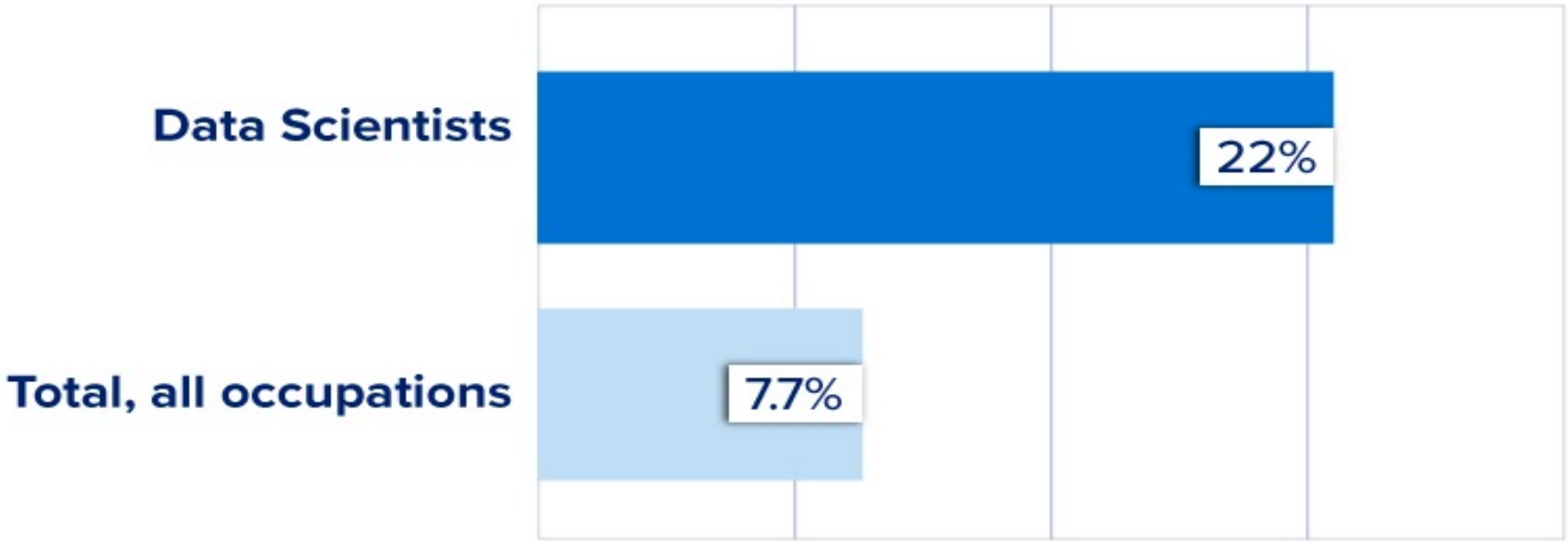
Percent change in employment, projected 2018-28



Note: All Occupations includes all occupations in the U.S. Economy.

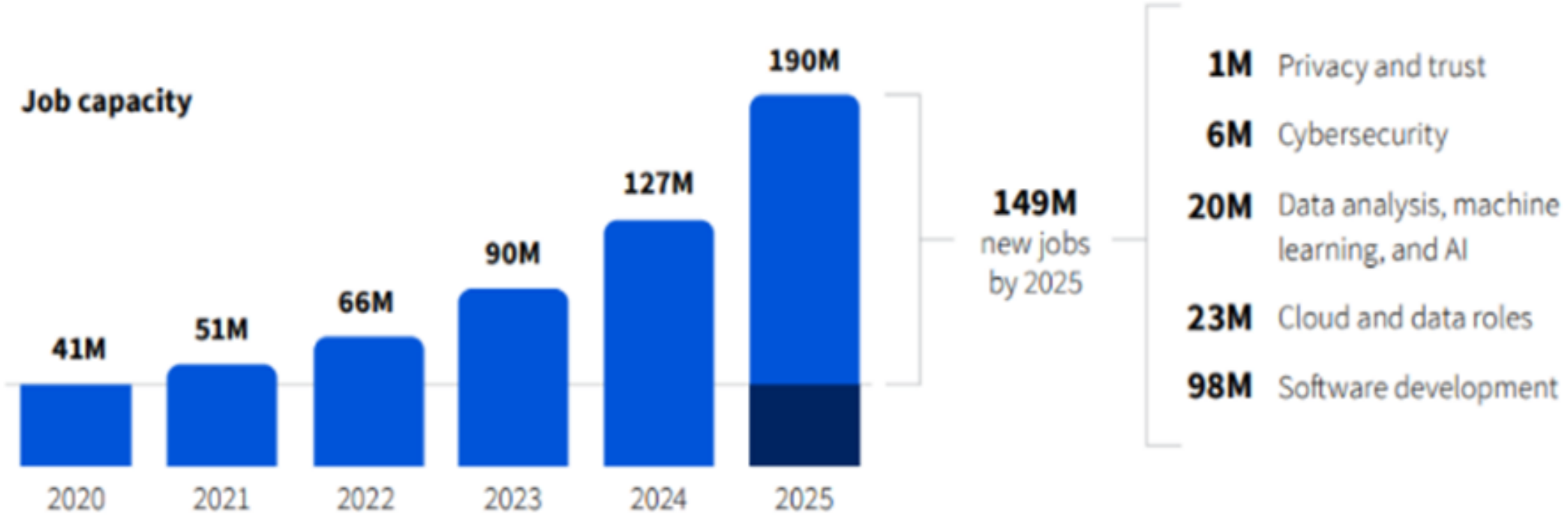
Source: U.S. Bureau of Labor Statistics, Employment Projections program

Computer and Information Research Scientists Job Outlook 2020-30



U.S. Bureau of Labor Statistics, "Occupational Outlook Handbook, Computer and Information Research Scientists." (2020)
https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm

Digital job growth from 2020 to 2025



Data Source: Microsoft Data Science utilizing LinkedIn data. Methodology and assumptions can be found in the white paper "Methodology: Digitization Capacity of the World Economy."

FIGURE 1

Data Scientist Profile

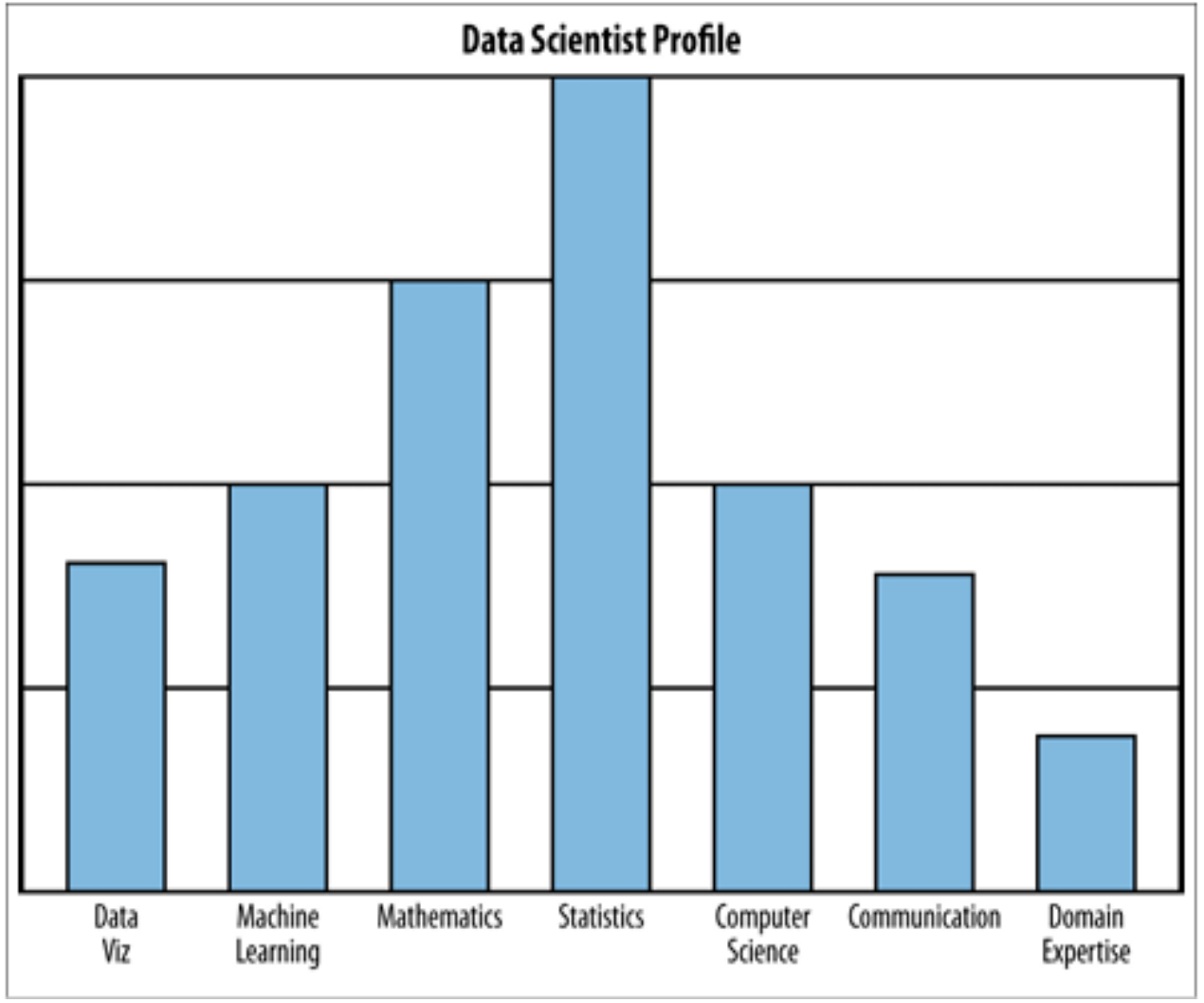
The key elements that define the Data Scientist Profile are:

1. Computer science skills
2. Mathematical skills
3. Statistical skills
4. Machine learning skills
5. Domain expertise (subject matter knowledge)
6. Communication and presentation skills
7. Data visualization skills

Data Scientist Profile

Fig:

- Rachel's data science profile, which she created to illustrate trying to visualize oneself as a data scientist;
- she wanted students and guest lecturers to "riff" on this—to add buckets or remove skills, use a different scale or visualization method, and think about the drawbacks of self-reporting



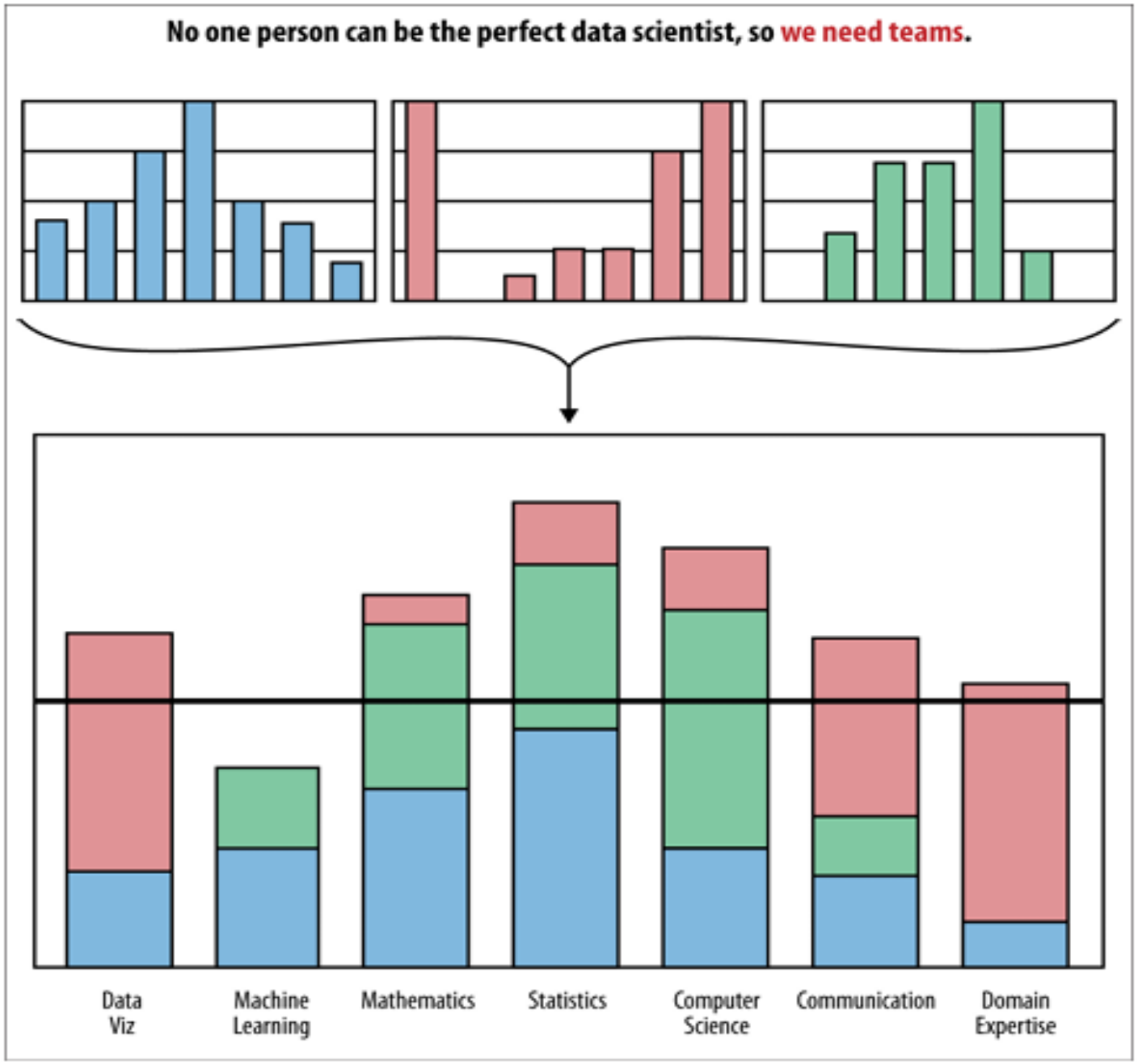
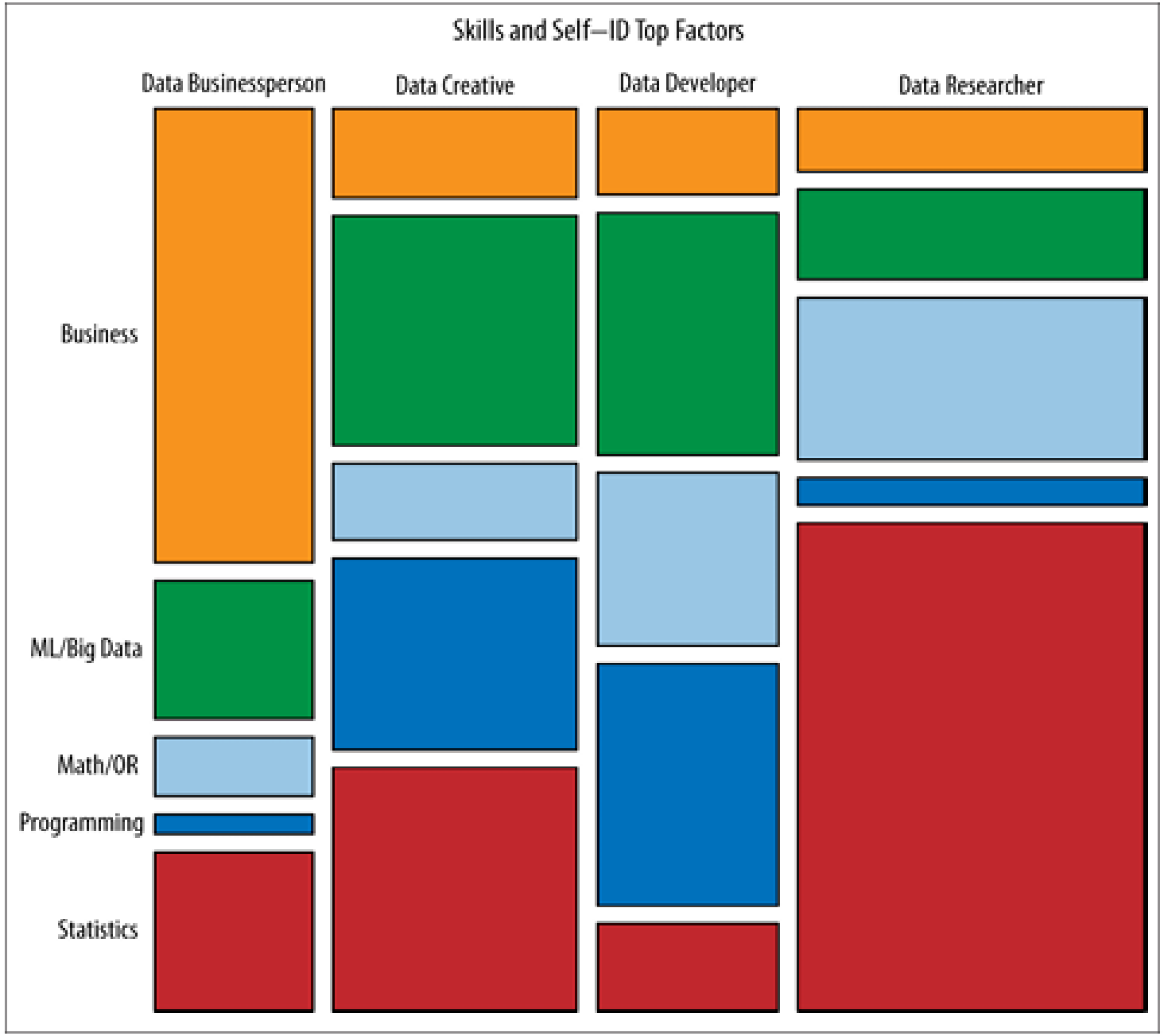


Fig: Data science **team profiles** can be constructed from data scientist profiles; there should be alignment between the data science team profile and the profile of the data problems they try to solve

Thought Experiment:

The **"thought experiment"** explores the idea of using data science itself to define what data science is - a meta-definition of sorts. Here are the key points regarding this thought experiment:

- The experiment poses the open-ended question: "**Can we use data science to define data science?"**
- One approach suggested is **text mining** - performing a Google search for "**data science**" and applying text mining models on the results
- An alternative is to look at **how practitioners of data science** describe what they do (e.g**. via word clouds**) and compare.
- This data could then be fed into **clustering algorithms** or other models to see if the model can accurately predict which field a person belongs to based on the "**stuff they do**."



Harlan Harris, did survey and clustering to define sub-fields within data science (illustrated in Figure).

- In essence, the thought experiment explores using ***data-driven techniques like text mining, natural language processing, and clustering*** to analyze descriptions of what data scientists and other professionals do, in order to derive a definition of data science in a bottom-up manner from the data itself.

4 PILLARS OF DATA SCIENCE

Domain Knowledge

> Business knowledge

> Expert systems

> User testing

Math & Statistics skills

> Linear algebra

> Calculus

> Descriptive statistics

> Inferential statistics

Computer science

> Big data technologies

> Programming

> Database

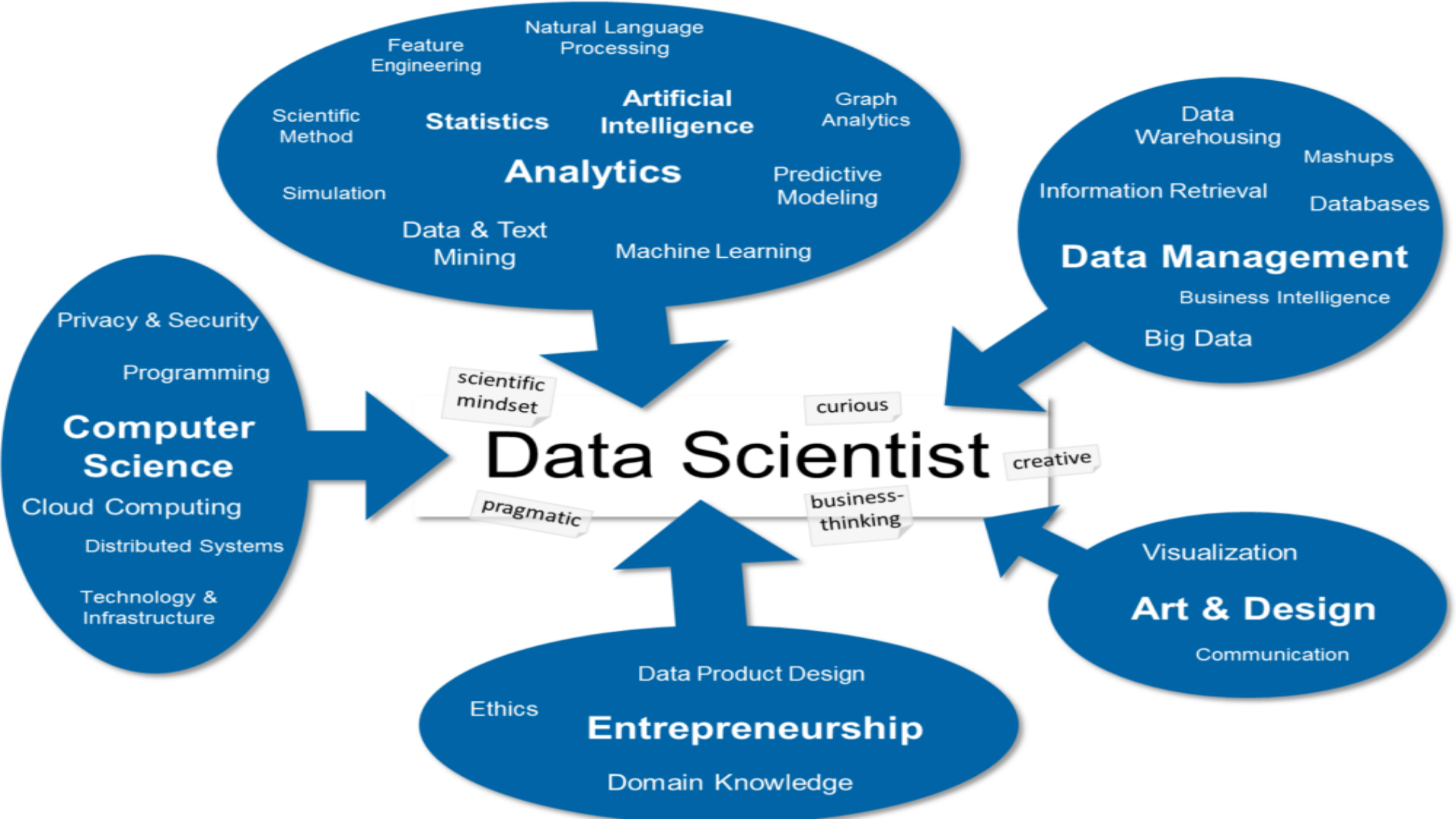
Communication & Visualization

> Storytelling skills

> Visual art design

> Able to engage with senior manager

> R packages


Natural Language Processing
Feature Engineering
Scientific Method
Statistics
Artificial Intelligence
Graph Analytics
Analytics
Simulation
Predictive Modeling
Data & Text Mining
Machine Learning
Data Warehousing
Mashups
Information Retrieval
Databases
Data Management
Business Intelligence
Big Data
Privacy & Security
Programming
Computer Science
Cloud Computing
Distributed Systems
Technology & Infrastructure
scientific mindset
curious
Data Scientist
creative
pragmatic
business-thinking
Visualization
Art & Design
Communication
Data Product Design
Ethics
Entrepreneurship
Domain Knowledge

Who is Data Scientist in the context of academia and industry?

- Data scientists in academia can be defined as interdisciplinary researchers working on computational challenges **with large, complex datasets to solve real-world problems across domains**.
- In industry, especially tech, data scientists are depicted as strategic decision-makers and hands-on practitioners who **wrangle data, build models, and communicate insights** to drive data-driven products and decisions.
- In general, a **data scientist** extracts meaning from data using statistics, machine learning, and domain knowledge. They spend significant time collecting, cleaning, and wrangling messy data, which requires **persistence, statistics, and software engineering skills.**

Academic

- In Academic students interested in becoming data scientists come from diverse backgrounds like *statistics, applied math, computer science, social sciences, journalism, biomedical informatics,* etc. They are interested in using data to solve important real-world problems.
- Academic data scientist can be defined as: ***"A scientist, trained in anything from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while simultaneously solving a real-world problem."***

Industry

- For the internet/online industry where the term originated, it **describes data scientists at different levels of seniority.**
- A **chief data scientist** sets the company's data strategy, *including data collection infrastructure, privacy concerns, user-facing data products, data-driven decision making, managing teams of engineers/scientists/analysts, communicating with leadership, patenting solutions, and setting research goals.*
- **Key responsibilities include** *exploratory data analysis, visualization, finding patterns, building models/algorithms to understand product usage, designing experiments, and communicating insights* clearly to teams and leadership through data visualizations.
- The data scientist plays a critical role in ***data-driven decision making*** and prototyping data products to be integrated into the company's offerings.

Terminology: Big Data

The different perspectives on what constitutes "Big Data":

- **"Big" is a relative term**, not an absolute threshold like 1 petabyte. Data qualifies as "Big Data" when its size outgrows the current computational capabilities (memory, storage, processing speed) available to handle it effectively.
- **"Big Data" refers to datasets that can't fit or be processed on a single machine**. Once data exceeds what a single computer can handle, new tools and methods are required to work with it, marking it as "Big Data."
- **Big Data is also described as a cultural phenomenon** - it highlights how massive amounts of data have become pervasive in our lives due to rapid technology advances.

Terminology: Big Data

- The different perspectives on what constitutes "Big Data":
- **The 4 Vs are commonly used to characterize Big Data:**
  - **Volume** (huge quantities)
  - **Variety** (diverse formats - structured, unstructured etc.)
  - **Velocity** (high speed of data generation/processing)
  - **Value** (analyzing for extracting valuable insights)
- In essence, "Big Data" is a relative concept describing data that is too large, too varied, too rapidly changing or too complex for traditional data processing systems. It requires new age computational tools and represents the data-intensive modern era we live in.

What is Big Data?

- **Big Data** refers to datasets whose size or complexity is beyond the ability of traditional database software tools to capture, store, manage and analyze.
- **Key characteristics of Big Data include:**
- **Volume**: Big Data involves **massive amounts** of data, ranging from terabytes to petabytes or even more. This massive volume is generated from various sources like social media, sensors, digital transactions, etc.
- **Variety:** Big Data comes in **different formats** like structured data (databases), semi-structured (XML, JSON) and unstructured (text, audio, video, etc). This heterogeneity of data types makes processing complex.
- **Velocity:** The **speed at which Big Data is generated, accessed, processed and analyzed** is **extremely high and ever-increasing**. Handling high velocity data for timely decision making is a challenge.
- **Veracity**: Big Data comes from **numerous sources** and is of widely ***varying quality and accuracy levels***. Ensuring veracity of analyzed results is crucial.
- **Value**: The core reason to deal with Big Data is to ***uncover hidden insights, patterns and correlations to extract substantial value*** from the ocean of data.

Definition of Big Data According to Steve Lohr

- **Steve Lohr defines Big Data as follows:**
- **"A bundle of technologies**" - This aligns with the technological aspect of Big Data, referring to the various tools, frameworks and platforms like Hadoop, Spark, NoSQL databases etc. that enable storage, processing and analysis of massive, complex datasets.
- **"A potential revolution in measurement"** - This points to the **transformative impact Big Data** can have on measurement and metrics across industries/domains by allowing organizations to capture, quantify and analyze more data points than ever before.
- **"A philosophy about how decisions will/should be made"** - This highlights the paradigm shift Big Data brings in terms of data-driven decision making. With insights extracted from large datasets, decisions can potentially be made based more on hard data/evidence rather than just intuition or experience alone.

Examples of Big Data

- **Social Media**:
  - **Facebook processes** over 500 terabytes of data every day from user activities like posts, comments, likes, shares, etc.
  - **Twitter generates** over 12 terabytes of data every day from tweets, retweets, and other user interactions.
- **E-commerce:**
  - **Amazon processes** millions of transactions every day and captures vast amounts of data related to customer preferences, browsing histories, and purchase patterns.
  - **Walmart handles** over 1 million customer transactions per hour, generating huge volumes of data related to inventory, sales, and customer behavior.

Examples of Big Data

- **Healthcare:**
  - **Electronic Health Records (EHRs)** store massive amounts of patient data, including medical histories, test results, imaging data, and treatment plans.
  - **Genomic sequencing projects** like the Human Genome Project generate petabytes of data from DNA sequencing and analysis.
- **Internet of Things (IoT):**
  - **Smart cities generate** massive data streams from sensors monitoring traffic, weather, air quality, and infrastructure.
  - **Industrial IoT devices** in manufacturing plants produce continuous data about machine performance, quality control, and supply chain operations.

Examples of Big Data

- **Scientific Research:**
  - The **Large Hadron Collider** at CERN generates around 30 petabytes of data annually from particle collisions and experiments.
  - Astronomical projects like the Sloan Digital Sky Survey have captured terabytes of data from telescopic observations of galaxies and celestial objects.
- **Telecommunications** : Telecom companies handle billions of call detail records, user locations, and network data every day from mobile devices and network infrastructure.
- **Finance and Banking**: Financial institutions process millions of transactions, trades, and market data every minute, generating massive volumes of data for analysis and compliance.
- **Transportation and Logistics**: Fleet management systems track real-time location, performance, and maintenance data from thousands of vehicles, generating terabytes of data.

New Kinds of Data

- **Traditional data**: Numerical, categorical, or binary data.
- **Text data:** Emails, tweets, news articles, etc.
- **Record data**: User-level data, timestamped event data, JSON formatted log files.
- **Geo-based location data**: Spatial data related to geographic locations.
- **Network data**: Data representing connections and relationships between entities.
- **Sensor data**: Data generated from various sensors and IoT devices.
- **Image data**: Visual data in the form of images.

Statistical Thinking in the Age of Big Data

When developing skills as a data scientist one should have the following foundational skills:

- **Statistics,**
- **Linear algebra, and**
- **Programming**

- Data scientists need to develop several **interdependent skill sets** in parallel, such as
  - **Data preparation,**
  - **Modeling,**
  - **Coding,**
  - **Visualization, and**
  - **Communication.**

Statistics for Data Science

- **Descriptive Statistics**: Measures of central tendency (mean, median, mode) Measures of dispersion (variance, standard deviation, range)
- Probability
- Statistical Inference
- Bayesian Statistics
- Time Series Analysis:
- Multivariate Analysis:

Statistical Inference

- The world we live in is **complex, random, and uncertain**, but it is also a "data-generating machine" through various processes and activities. Some of the **examples of potential data-generating processes are**: Counting people passing by, collecting email data, or Analyzing DNA samples, etc.
- *The statistical inference is a discipline/field that deals with understanding and making sense of the complex, random, and* ***uncertain real-world processes*** *by collecting data, developing statistical procedures, methods, models and theorems for extracting meaningful information from the data generated by* ***stochastic (random) processes.***

What is Population?

- A **population** refers to the entire group of individuals, objects, or items that a researcher is interested in studying. It is the complete set of elements that share some common characteristics, which the researcher wants to draw conclusions about.
- **Example**: If you want to study the average height of all adults in a particular city, the population would be all the adults living in that city.

Sample:

- **A sample** is a subset of the population that is selected for observation and analysis. It is a smaller, manageable group that represents the characteristics of the larger population.
- **Example:** If the population is all adults in a city, a sample could be 500 randomly selected adults from different neighborhoods within that city.

The main reasons for using a sample instead of studying the entire population are:

- **Cost and time efficiency**: Studying an entire population can be expensive and time-consuming, especially when the population is large.
- **Accessibility**: In some cases, it may not be possible or practical to study the entire population due to geographical constraints or other limitations.
- **Destructive testing**: If the study involves destructive testing, it is not feasible to test the entire population.

Goal of sampling

- The goal of sampling is to select a **representative sample** from the **population,** so that the characteristics observed in the sample can be generalized to the larger population with a known degree of accuracy and precision.
- **Example:** To estimate the average income of households in a city, a researcher might randomly select **1000 households** from different neighborhoods and use their income data as a sample to make inferences about the entire population of households in that city.
- ***In summary, the population*** *is the entire group of interest,* ***while a sample is a smaller****, manageable subset selected from the population for the purpose of studying and making inferences about the population as a whole.*

Example 1: Studying student performance in a school district

- **Population**: All students enrolled in the school district
- **Sample**: A random selection of 500 students from different schools within the district

Example 2: Estimating customer satisfaction for a retail chain

- **Population**: All customers who have shopped at the retail chain in the last year
- **Sample**: A random sample of 2,000 customers from the company's customer database

Example 3: Measuring public opinion on a political issue

- **Population**: All eligible voters in a particular country
- **Sample:** A representative sample of 1,500 voters from different regions, age groups, and demographic backgrounds

Example 4: Testing the effectiveness of a new medication

- **Population:** All patients diagnosed with a specific medical condition
- **Sample**: A random sample of 300 patients who meet the inclusion criteria for the clinical trial

Example 5: Analyzing consumer preferences for a new product

- Population: All potential customers in the target market
- Sample: A focus group of 20 individuals representative of the target demographic

Big Data Can Mean Big Assumptions

- The major assumptions and pitfalls associated with Big Data analysis are as follows:

1. **The claim that with Big Data we have "N=ALL" i.e. all the data**
2. **Data is not Objective:**
3. **Ignoring causation**
4. **User-level modelling (n=1)**

Modeling

- Models are **simplified representations or abstractions** of reality that humans create to understand the world around them. Whether *architectural blueprints, molecular visualizations, or statistical functions*, models attempt to capture the essence of the underlying phenomena or processes generating the observed data.
- The term "**mode**l" can have different meanings in different contexts, like **data models for databases** versus **statistical/mathematical models**.
- Models necessarily involve **removing or abstracting** away **extraneous details** from the **full complexity of reality**
- **Statistical models** specifically aim to capture the uncertainty and randomness inherent in data-generating processes through mathematical functions expressing the shape and structure of the data

The key aspects of modeling in data science include

1. **Data Representation**
2. **Abstraction and Simplification**
3. **Variable Selection**
4. **Assumption Making**
5. **Model Fitting**
6. **Evaluation**
7. **Interpretation**
8. **Prediction**

2.2.4 What is Model?

- A **model** is a simplified representation or abstraction of reality that humans create in an attempt to understand the nature of the world around them.
- It is an **artificial construction** that captures the essence of a phenomenon or process through a particular lens or perspective, such as architectural, biological, or mathematical.

Key points about models:

- Models are human efforts to understand and represent reality by focusing on specific aspects
- They involve **removing or abstracting** away **extraneous details** from the full complexity of reality.
- **Models in different domains serve this purpose** - architects use **blueprints**, molecular biologists use **3D visualizations**, statisticians use **mathematical functions**

Key points about models:

- Models make assumptions and simplifications **about the underlying reality**. For example, a protein model may not account for **quantum mechanics** governing **electron behaviour**.
- **Statistical models** aim to capture the uncertainty and randomness in data-generating processes using **mathematical functions** representing the data's shape and structure.

Statistical Modeling

- Statistical modeling is the process of representing the underlying data-generating process or phenomenon using mathematical or statistical expressions

Key aspects of Statistical Modeling

- **Conceptualizing the Process**
- **Mathematical Representation:** These expressions contain parameters (represented by Greek letters like **β**) whose values are initially unknown and need to be estimated from the data
- **Variable Relationships:** For example, if there is a hypothesized linear relationship between two variables **x and y**, it could be expressed as **y = β0 + β1x**, where **β0** and **β1** are the parameters to be estimated.
- **Visualizing Data Flow**
- **Parameter Estimation:** estimate the values of the unknown parameters (like **β0 and β1**) using the available data

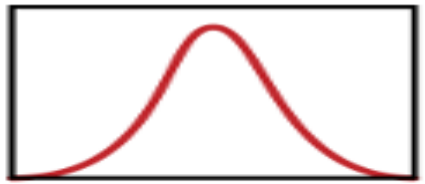
Key points: How to build a Model?

- **Model building** is not straightforward, and there are no global standards or obvious starting points. It requires making assumptions about the underlying structure of reality.
- **Exploratory data analysis (EDA),** which involves plotting and visualizing the data, can help build intuition about the dataset and guide the modeling process.
- **The recommended approach** is to start simple and then gradually increase the complexity of the model
- **Writing down assumptions** in the form of equations and code
- **There is a trade-off between simplicity and accuracy in modeling**

Probability Distributions

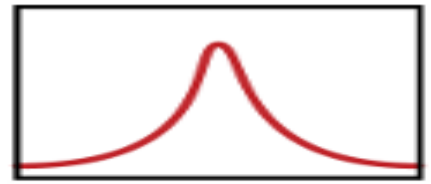
- They describe the probability of different outcomes or values of a random variable. Many common probability distributions like ***normal, Poisson, Weibull, etc***.
- For example, the **normal or Gaussian distribution** was named after **Gauss** who noticed human, heights followed a **bell-shaped curve.** For instance, if **X** represents human height, it can be modeled by a normal distribution:

$$N(x|\mu,\sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

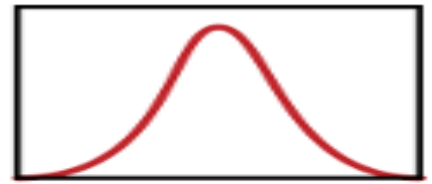


Normal Distribution

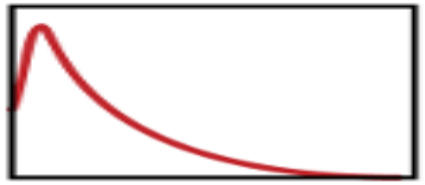
Uniform Distribution



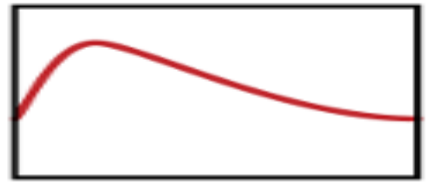
Cauchy Distribution



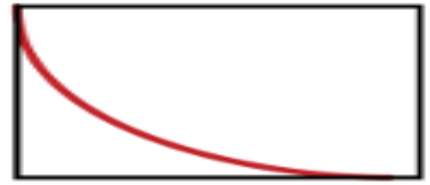
t Distribution



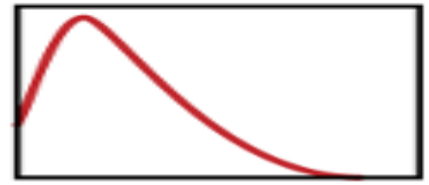
F Distribution



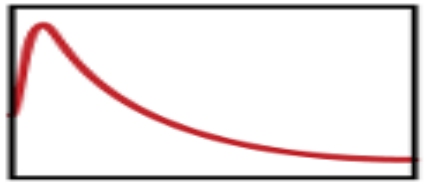
Chi-Square Distribution



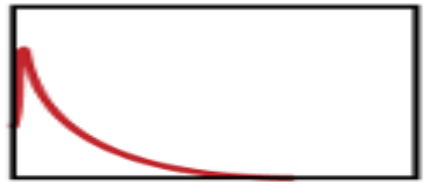
Exponential Distribution



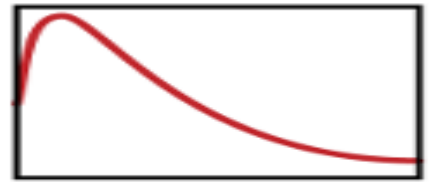
Weibull Distribution



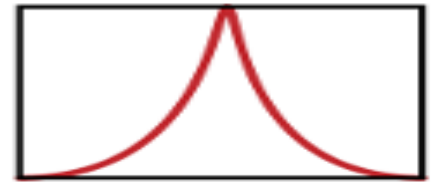
Lognormal Distribution



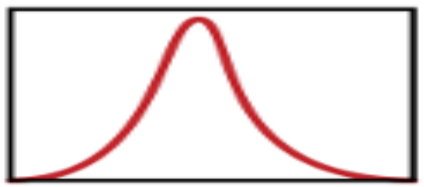
Birnbaum-Suanders
(Fatigue Life) Distribution



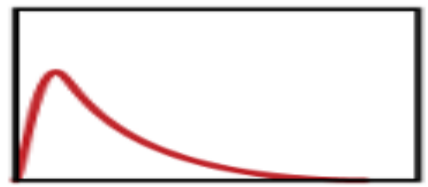
Gamma Distribution



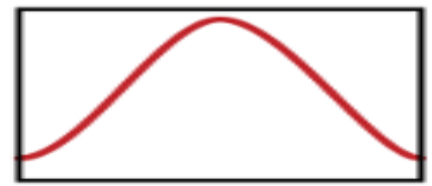
Double Exponential
Distribution



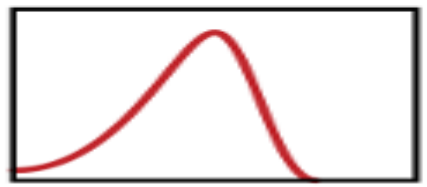
Power Normal Distribution



Power Lognormal
Distribution



Tukey-Lambda Distribution



Extreme Value Distribution

Beta Distribution

Discrete and Continuous Probability Distributions

- For discrete X, the probability distribution is described by:

$$P(X = x) = f(x)$$

where f(x) is the probability mass function.

- For continuous X, the probability distribution is described by:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

where f(x) is the probability density function.

- **Example 1 (Discrete)**: Consider a fair six-sided die roll X. The probability distribution P(X) can be described as:

$$P(X = x) = \frac{1}{6} \quad \text{for } x = 1, 2, 3, 4, 5, 6$$

Here, $f(x) = \frac{1}{6}$ for each outcome x.

- **Example 2 (Continuous)**: Suppose X follows a normal distribution $N(\mu,\sigma^2)$, where μ is the mean and σ^2 is the variance. The probability distribution P(X) is given by its probability density function:

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx$$

Here, $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

Univariate Probability distribution

- A univariate probability distribution describes the probabilities associated with a single random variable
- Let a **single variable X represent** the **heights** of adult males. X can be modeled by a normal distribution, with parameters μ (mean height) and σ (standard deviation of heights) estimated from data.

Multivariate Probability Distribution

- A multivariate probability distribution describes the probabilities associated with multiple random variables simultaneously.
- Example: Let X represent the **heights of individuals** and Y represent **their weights**. The joint distribution of X and Y can be modeled by a bivariate normal distribution, with parameters **μ_x, μ_y** (means for height and weight), **σ_x, σ_y** (standard deviations for height and weight), and ρ (correlation coefficient between height and weight).

Conditional Probability Distribution

- Conditional probability distribution refers to the probability distribution of one or more random variables given specific values or conditions of other random variables.
- **Example:** Let X represent the insurance claims of drivers. The condition is whether the driver is over 65 years old or not, represented by a binary variable A (1 for over 65, 0 for under 65).
- P(X|A==1) or P(X|A==0)

Fitting a Model:

Fitting a model involves estimating the parameters of a mathematical model using observed or given data.

- **Estimating Parameters**: When fitting a model, you estimate parameters (like coefficients in a regression model) , The aim is to find the best-fitting model that explains the relationship **between variables based on the assumption of a certain mathematical form (e.g., linear, exponential**).
- **Expressing the Model**: Once the model is fitted, it can be expressed in a mathematical form like $y = 7.2 + 4.5x$.
- **Coding the Model**: Implementing the fitted model involves coding the specified mathematical form (e.g., y=7.2+4.5x) into programming languages like **R or Python**.

Fitting a Model:

- **Optimization Methods**: Optimization methods are used to find the best values of model parameters that maximize the likelihood of observing the given data.
- **Sophistication and Expertise**: As you become more experienced or specialized in modeling, you may explore and customize optimization methods.
- In summary, fitting a model involves the iterative process of ***parameter estimation, model specification, coding, and optimization*** to develop a statistical representation that explains the relationship within the observed data.

Overfitting

- **Overfitting** in model development refers to the phenomenon where a model learns not only the underlying patterns in the training data but also captures noise, random fluctuations, or outliers that are specific to the training dataset. This results in a model that performs extremely well on the training data but fails to generalize effectively to new, unseen data.
- When a model is overfitted, **it may exhibit poor performance** when applied to new data that was not part of the training dataset.
- Overfitted models tend to **be too complex**, capturing spurious relationships and details that are not reflective of the true underlying patterns in the data.
- Overfitting can lead to **misleading conclusions** and **inaccurate predictions** when deployed in real-world scenarios.

Strategies to Mitigate Overfitting and Improve Model Generalizability:

- **Simplify the Model:**
- **Regularization:**
- **Cross-Validation:**
- **Feature Selection**
- **Early Stopping**

Module1 : Introduction to Data Science and Statistical Inference Needed:

1. **Introduction** : What is Data Science? Big Data and Data Science hype – and getting past the hype, Why now? Datafication, Current landscape of perspectives, Skill sets.
2. **Needed Statistical Inference**: Populations and samples, Statistical modelling, probability distributions, fitting a model.

End of Module1