Data Science

Data Science

1. Introduction
2. Data Pre-processing
3. Exploratory Data Analysis
4. Unstructured Data Mining
5. Social Networks Analysis
6. Model Evaluation Techniques

Text and Reference Books

Text Books

1. Daniel T. Larose and Chantal D. Larose, “**Data Mining and Predictive Analytics**”, Wiley Publication, ISBN 978-1-118-11619-7
2. Ronen Feldman, James Sanger, “The Text Mining Handbook Advanced Approaches in Analyzing Unstructured Data”, Cambridge University Press
3. Jure Leskovec, Anand Rajaraman, and Jeffery David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2 edition (13 November 2014), ISBN-10: 1107077230, ISBN-13: 978-1107077232

Reference Books

1. Cathy O'Neil and Rachel Schutt, "Doing Data Science", O'Reilly Media, October 2013, Print ISBN:978-1-449-35865-5
2. Field Cady, “The Data Science Handbook”, Wiley Publication
3. Wes McKinney, “Python for Data Analysis”, O'Reilly Media, ISBN ISBN: 978-1-449-31979-3

Course Outcomes

At the end of the course students will be able to-

1. Explain fundamentals of data science.
2. Explain and apply data processing techniques and exploratory data analysis.
3. Explain fundamentals of unstructured data mining.
4. Explain Social Network Analysis techniques.
5. Explain model evaluation techniques.

What is Data Science ?

- **Data science** is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to **data mining**.

- **Data science** is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "**understand and analyze actual phenomena**" with data.

- It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

- **Data science** is the study of the extraction of knowledge from data.

- Data science and machine learning provides the basis for business growth, cost and risk reduction and even new business model creation.

Relationship to statistics

Many critical academics and journalists see

- no distinction between data science and statistics.
- data science is a buzzword without a clear definition and has simply replaced "business analytics".
- Data science, like any other interdisciplinary field, employs methodologies and practices from across the academia and industry.
- Data science is different from the existing practice of data analysis.

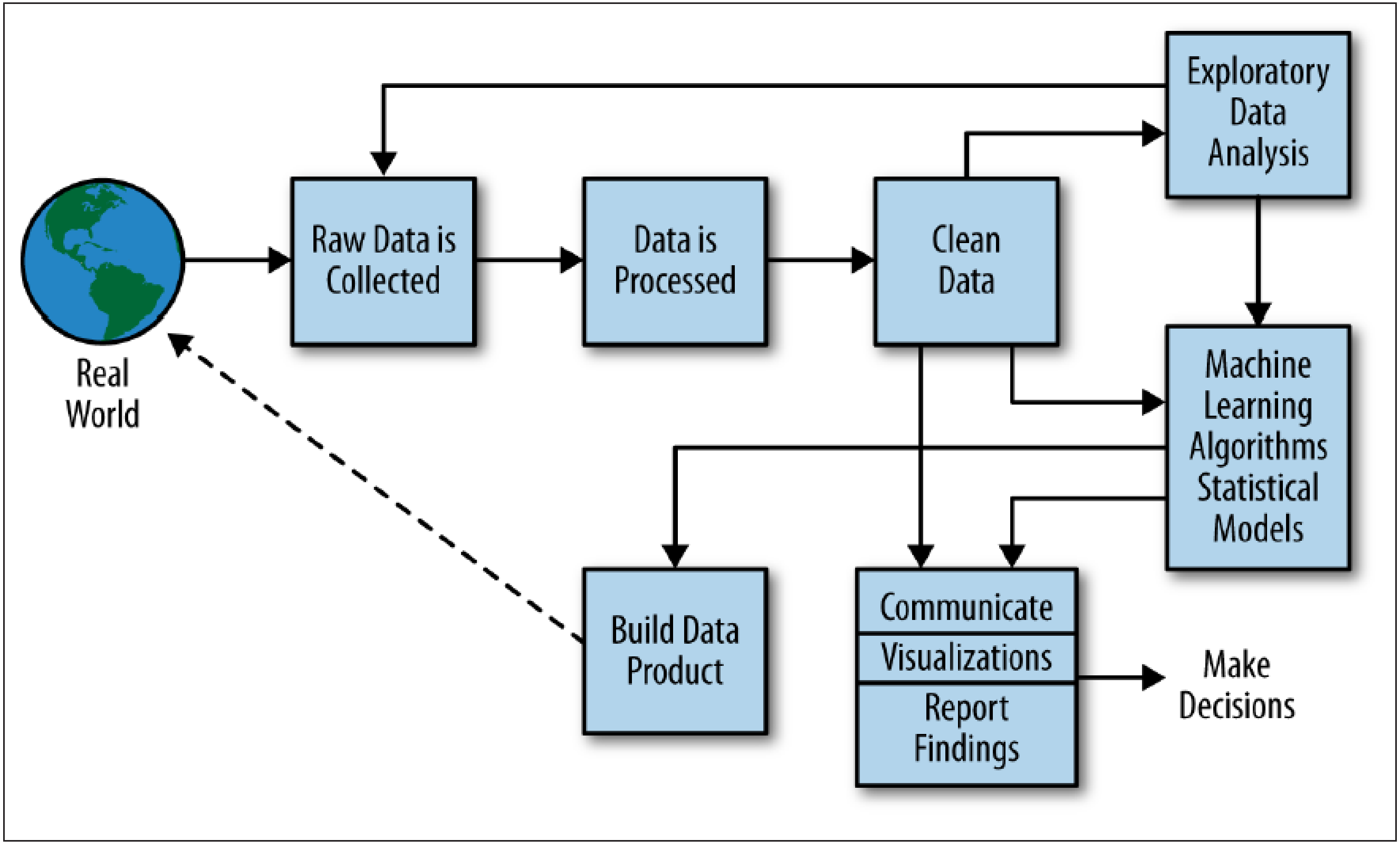
Data science – discovery of data insight

- This aspect of data science is all about **uncovering findings from data**.
- Diving in at a granular level to mine and understand **complex behaviors, trends, and inferences.**
- It's about **surfacing hidden insight** that can help enable companies to make smarter business decisions.
- For example:
  - Netflix data mines movie viewing patterns to understand what drives user interest, and uses that to make decisions on which Netflix original series to produce.
  - Proctor & Gamble utilizes time series models to more clearly understand future demand, which help plan for production levels more optimally.

Data science – development of data product

- A "data product" is a technical asset that:
  (1) utilizes data as input, and
  (2) processes that data to return algorithmically-generated results.
- The classic example of a data product is a recommendation engine.
- Here are some examples of data products:
  - **Amazon's** recommendation engines suggest items for you to buy. **Netflix** recommends movies to you. **Spotify** recommends music to you.
  - **Gmail's** spam filter is data product – an algorithm behind the scenes processes incoming mail and determines if a message is junk or not.
  - **Computer vision** used for self-driving cars is also data product – machine learning algorithms are able to recognize traffic lights, and other cars on the road.

The Data Science Process



Data Science Tools

Data Wrangling : SQL, Hive, Pig, Impala, Spark.

Machine Learning : R, Python (Sci-Kit learn), Spark (MlLib), Knime, WEKA, Rapidminer.

Visualization : GGplot and Rshiny for R, Matplotlib for Python, D3.js and Tableau.

Who is Data Scientist?

- **Data Scientists** are a new breed of analytical data expert who have the technical skills to solve complex problems – and the curiosity to explore what problems need to be solved.

- A **data scientist** is someone who is better at **statistics** than any **software engineer** and better at software engineering than any statistician**.**

- Many data scientists began their careers as statisticians or data analysts.

Data Scientist

A Data Scientist can also be divided into different roles based on their skill sets.

- Data Researcher
- Data Developers

Typical job duties for data scientists

- Collecting large amounts of noisy data and transforming it into a more usable format.
- Solving business-related problems using data-driven techniques.
- Working with a variety of programming languages, including Statistical Analysis System (SAS), R and Python.
- Having a solid grasp of statistics, including statistical tests and distributions.
- Staying on top of analytical techniques such as machine learning, deep learning and text analytics.
- Communicating and collaborating with both IT and business.
- Looking for order and patterns in data, as well as spotting trends that can help a business’s bottom line.

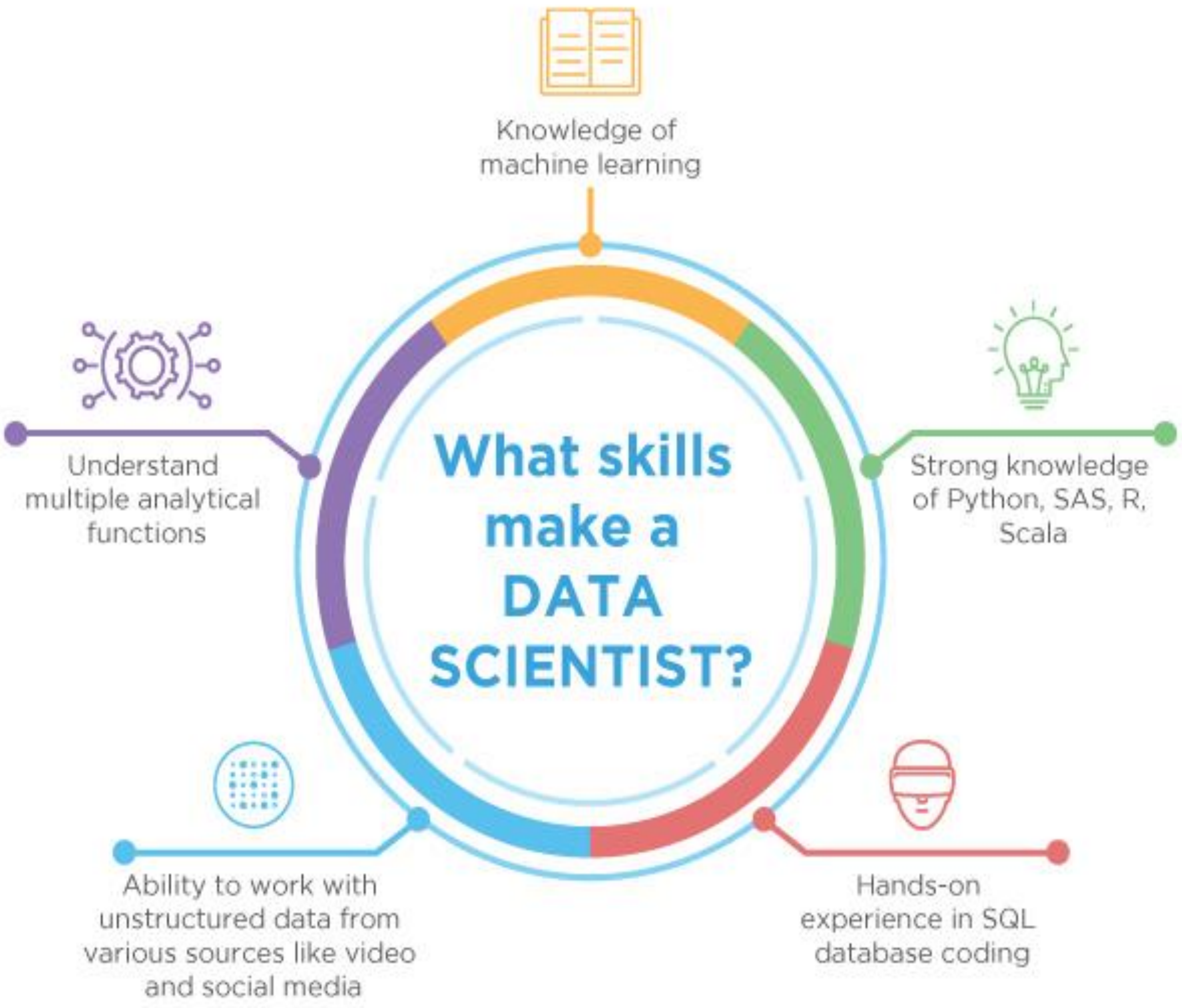
What’s in a data scientist’s toolbox?

- **Data visualization:** the presentation of data in a pictorial or graphical format so it can be easily analyzed.
- **Machine learning:** a branch of artificial intelligence based on mathematical algorithms and automation.
- **Deep learning**: an area of machine learning research that uses data to model complex abstractions.
- **Pattern recognition**: technology that recognizes patterns in data (often used interchangeably with machine learning).
- **Data preparation**: the process of converting raw data into another format so it can be more easily consumed.
- **Text analytics**: the process of examining unstructured data to glean key business insights.

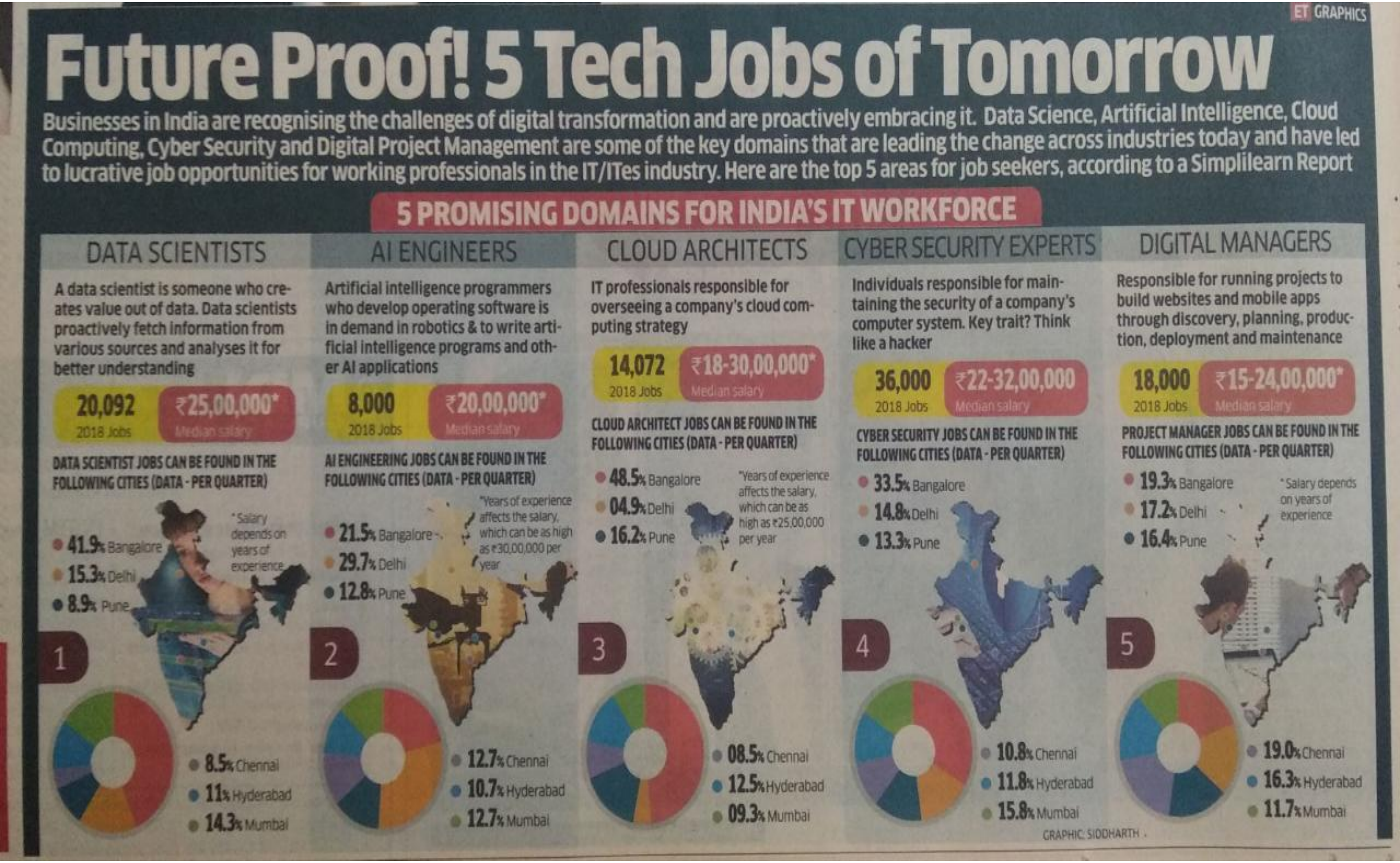
Technical Skills for Data Scientists

- Math (e.g. linear algebra, calculus and probability)
- Statistics (e.g. hypothesis testing and summary statistics)
- Machine learning tools and techniques (e.g. k-nearest neighbors, random forests, ensemble methods, etc.)
- Software engineering skills (e.g. distributed computing, algorithms and data structures)
- Data mining
- Data cleaning and munging
- Data visualization (e.g. ggplot and d3.js) and reporting techniques
- Unstructured data techniques
- R and/or SAS (Statistical Analysis System) languages
- SQL databases and database querying languages
- Python (most common), C/C++ Java, Perl
- Big data platforms like Hadoop, Hive & Pig
- Cloud tools like Amazon S3 (Simple Storage Service)

Skills for Data Scientists


Knowledge of machine learning
Understand multiple analytical functions
What skills make a DATA SCIENTIST?
Strong knowledge of Python, SAS, R, Scala
Ability to work with unstructured data from various sources like video and social media
Hands-on experience in SQL database coding

Career as Data Scientist



ET GRAPHICS

Future Proof! 5 Tech Jobs of Tomorrow

Businesses in India are recognising the challenges of digital transformation and are proactively embracing it. Data Science, Artificial Intelligence, Cloud Computing, Cyber Security and Digital Project Management are some of the key domains that are leading the change across industries today and have led to lucrative job opportunities for working professionals in the IT/ITes industry. Here are the top 5 areas for job seekers, according to a Simplilearn Report

5 PROMISING DOMAINS FOR INDIA'S IT WORKFORCE

DATA SCIENTISTS

A data scientist is someone who creates value out of data. Data scientists proactively fetch information from various sources and analyses it for better understanding

20,092 2018 Jobs

₹25,00,000* Median salary

DATA SCIENTIST JOBS CAN BE FOUND IN THE FOLLOWING CITIES (DATA - PER QUARTER)

- 41.9% Bangalore
- 15.3% Delhi
- 8.9% Pune

*Salary depends on years of experience

1

- 8.5% Chennai
- 11% Hyderabad
- 14.3% Mumbai

AI ENGINEERS

Artificial intelligence programmers who develop operating software is in demand in robotics & to write artificial intelligence programs and other AI applications

8,000 2018 Jobs

₹20,00,000* Median salary

AI ENGINEERING JOBS CAN BE FOUND IN THE FOLLOWING CITIES (DATA - PER QUARTER)

- 21.5% Bangalore
- 29.7% Delhi
- 12.8% Pune

*Years of experience affects the salary, which can be as high as ₹30,00,000 per year

2

- 12.7% Chennai
- 10.7% Hyderabad
- 12.7% Mumbai

CLOUD ARCHITECTS

IT professionals responsible for overseeing a company's cloud computing strategy

14,072 2018 Jobs

₹18-30,00,000* Median salary

CLOUD ARCHITECT JOBS CAN BE FOUND IN THE FOLLOWING CITIES (DATA - PER QUARTER)

- 48.5% Bangalore
- 04.9% Delhi
- 16.2% Pune

*Years of experience affects the salary, which can be as high as ₹25,00,000 per year

3

- 08.5% Chennai
- 12.5% Hyderabad
- 09.3% Mumbai

CYBER SECURITY EXPERTS

Individuals responsible for maintaining the security of a company's computer system. Key trait? Think like a hacker

36,000 2018 Jobs

₹22-32,00,000 Median salary

CYBER SECURITY JOBS CAN BE FOUND IN THE FOLLOWING CITIES (DATA - PER QUARTER)

- 33.5% Bangalore
- 14.8% Delhi
- 13.3% Pune

4

- 10.8% Chennai
- 11.8% Hyderabad
- 15.8% Mumbai

DIGITAL MANAGERS

Responsible for running projects to build websites and mobile apps through discovery, planning, production, deployment and maintenance

18,000 2018 Jobs

₹15-24,00,000* Median salary

PROJECT MANAGER JOBS CAN BE FOUND IN THE FOLLOWING CITIES (DATA - PER QUARTER)

- 19.3% Bangalore
- 17.2% Delhi
- 16.4% Pune

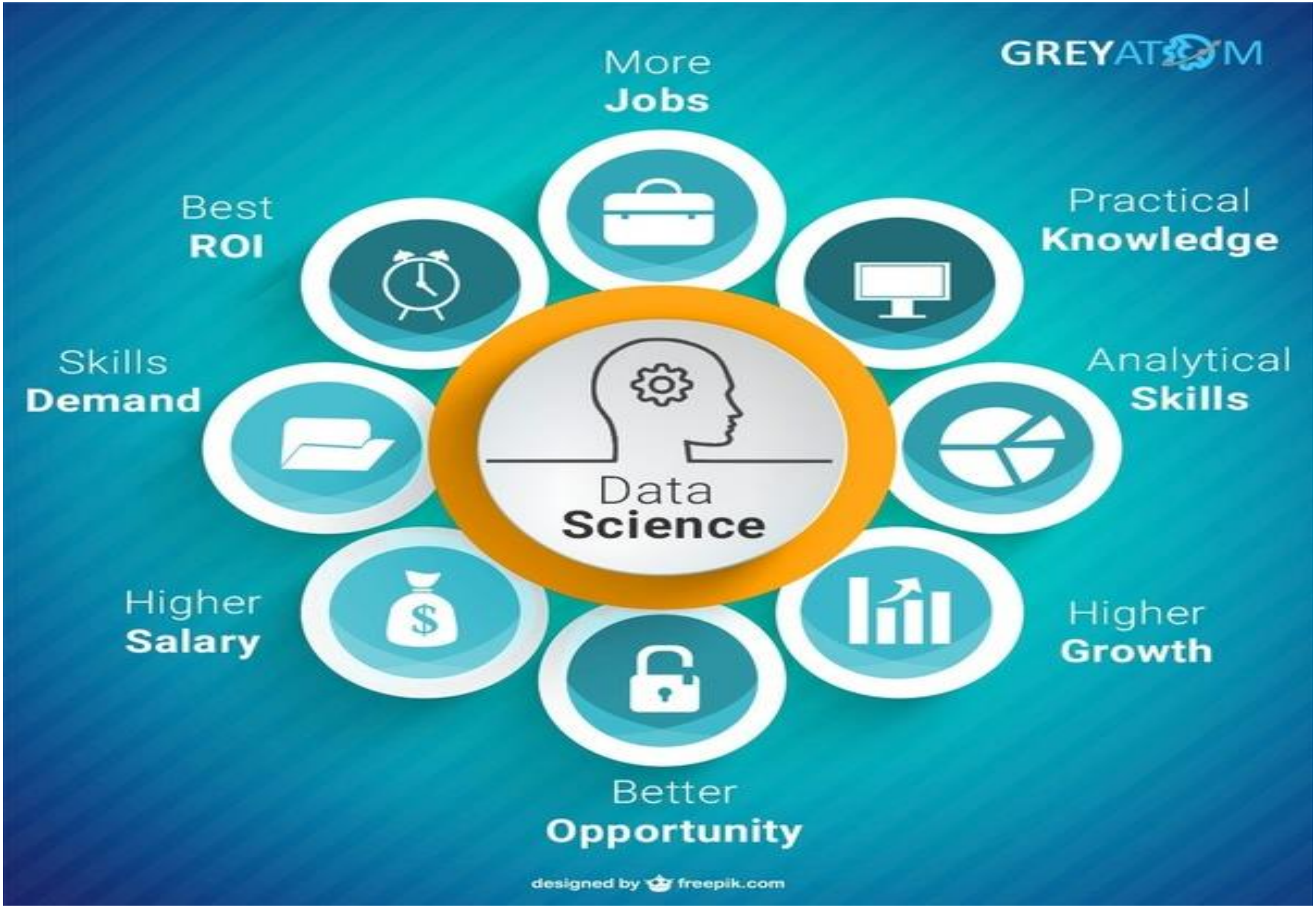
*Salary depends on years of experience

5

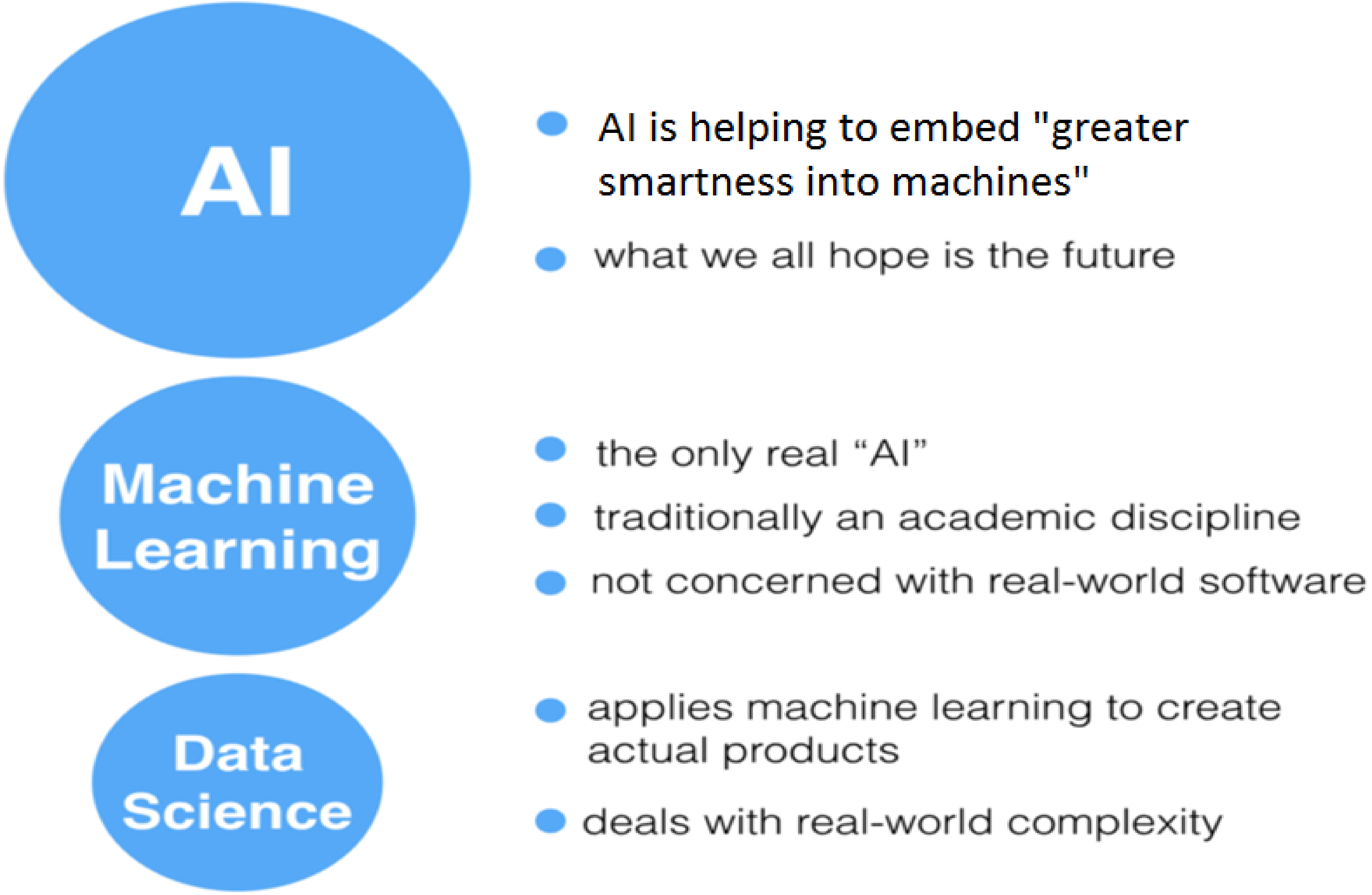
- 19.0% Chennai
- 16.3% Hyderabad
- 11.7% Mumbai

GRAPHIC: SIDDHARTH

Career as Data Scientist Cont….



Data Science, Machine Learning, & AI



AI

- AI is helping to embed "greater smartness into machines"
- what we all hope is the future

Machine Learning

- the only real “AI”
- traditionally an academic discipline
- not concerned with real-world software

Data Science

- applies machine learning to create actual products
- deals with real-world complexity

Difference between AI, Machine Learning and Data Science

Artificial intelligence:

- Wide term with applications ranging from robotics to text analysis.
- It is still a technology under evolution
- Arguments of whether we should be aiming for high-level AI or not.

Machine learning:

- Subset of AI that focuses on a narrow range of activities.
- It is, in fact, the only real artificial intelligence with some applications in real-world problems.

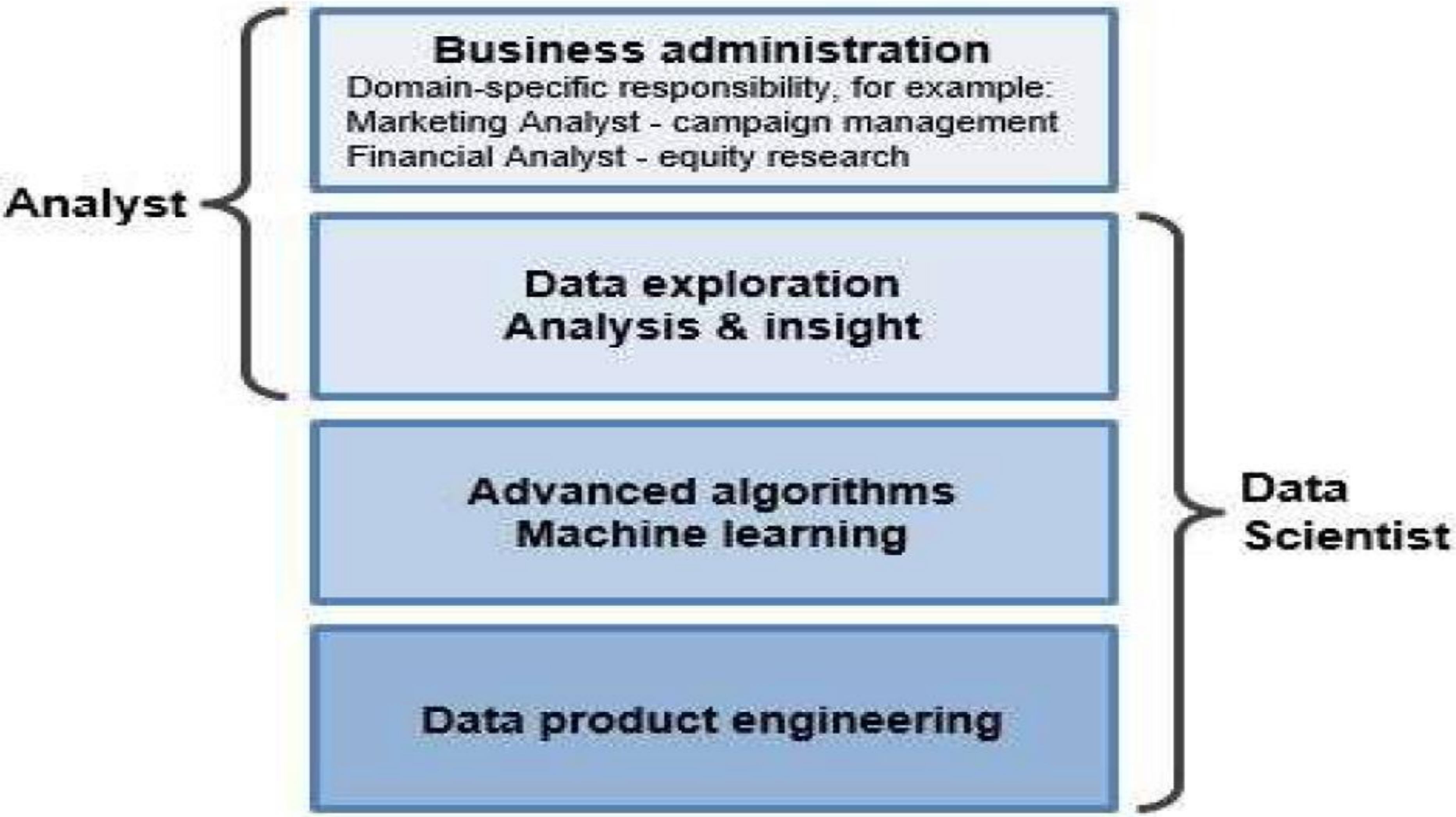
Data science:

- Isn't exactly a subset of machine learning but it uses ML to analyze data and make predictions about the future.
- It combines machine learning with other disciplines like big data analytics and cloud computing.
- Data science is a practical application of machine learning with a complete focus on solving real-world problems.

Data Scientist and Data Analyst

- ***Data scientist*** *- predict the future based on past patterns whereas a data analyst - curates meaningful insights from data.*
- ***Data scientist*** *- estimates the unknown whereas data analyst looks at the known from new perspectives.*
- ***Data scientist*** *is expected to generate their own questions while a data analyst finds answers to a given set of questions from data.*
- ***Data analyst*** *- addresses business problems and scientist not just addresses business problems but picks up those problems that will have the most business value once solved.*

Data Scientist and Data Analyst



Analyst

Business administration

Domain-specific responsibility, for example:
Marketing Analyst - campaign management
Financial Analyst - equity research

Data exploration
Analysis & insight

Advanced algorithms
Machine learning

Data product engineering

Data Scientist

Applications / Examples of Data Science

- **Internet Search**
- **Digital Advertisements (Targeted Advertising and re-targeting)**
- **Recommender Systems**
- **Image Recognition**
- **Speech Recognition**
- **Gaming**
- **Price Comparison Websites**
- **Airline Route Planning**
- **Fraud and Risk Detection**
- **Delivery logistics**
- **Miscellaneous (**Marketing, Finance, Human Resources, Health Care, Government Policies)

Data Sources

- According to Wall Street Journal, the digital universe will reach 180 zetabytes by 2025.
- The new economy is more about analyzing rapid real-time flows of data, often unstructured.
  - The streams of photos and videos generated by users of social networks
  - The ream of information produced by commuters on their way to work
  - The flood of data from hundreds of sensors in a jet engine
  - Data from subway trains and wind turbines
  - Uber, known for cheap taxi rides, owns the biggest pool of data about supply and demand for personal transportation.
  - Tesla, maker of fancy electric cars collect mountains of data, which allow the firm to optimize its self-driving algorithms and then update the software accordingly.

Data Sources

- More and more data generated through:

1. Facebook: data generated through photo sharing, text-photo messaging
2. Alphabet – mapping and navigation information
3. IBM – meterology data(weather data), health care data
4. INTEL –self-driving cars(Mobileye 2017- *Mobileye*, an *Intel* company, is a leader in automated technology and the world's largest supplier of cameras for advanced driver assistance systems (ADAS))
5. Microsoft – keyboard/ AI generated data , business networking data(LinkedIn)
6. Oracle – Cloud data platform, marketing data

Data Sources

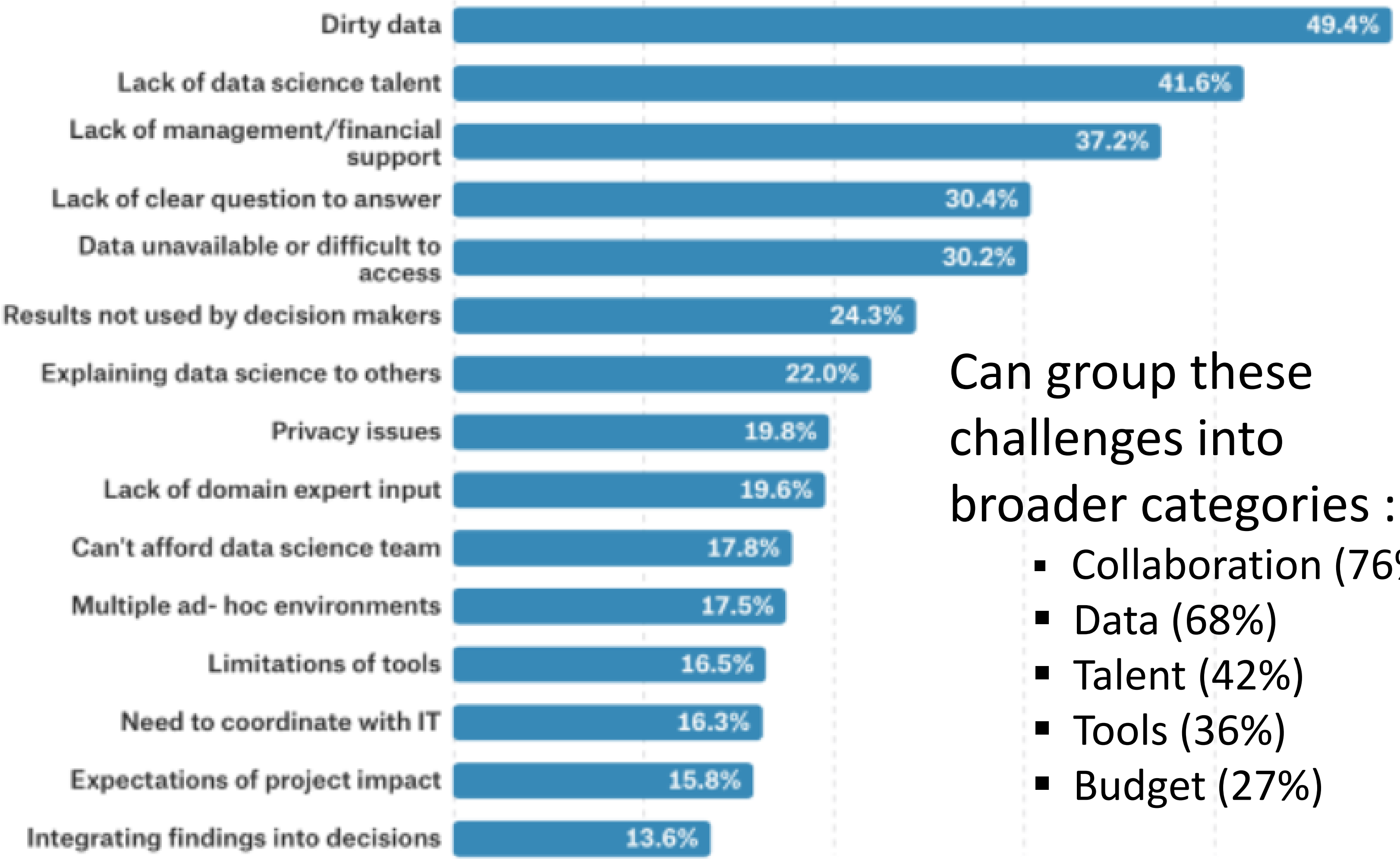
- **The application and usage of publicly available datasets is only limited by your creativity and application.**
- **Simple & Generic datasets to get Started**
  - **data.gov**
  - **data.gov.in**
  - **data.worldbank.org**
  - **rbi.org.in/Scripts/Statistics.aspx**
  - **github.com/fivethirtyeight/data**
- **Huge Datasets**
  - **Amazon Web Services (AWS) datasets**
  - **Google datasets**
  - **Youtube labeled Video Dataset**
- **Datasets for predictive modeling & machine learning**
  - **UCI Machine Learning Repository**
  - **Kaggle**
  - **Analytics Vidhya**
  - **Quandl**
  - **Past KDD Cups**
  - **Driven Data**

Data Sources Cont…

- **Image classification datasets**
  - The MNIST (Modified National Institute of Standards and Technology) Database
  - Chars74K
  - Frontal Face Images
  - ImageNet
- **Text Classification datasets**
  - Spam – Non Spam
  - Twitter Sentiment Analysis
  - Movie Review Data
- **Datasets for Recommendation Engine**
  - MovieLens
  - Jester
- **Websites which Curate list of datasets from various sources**
  - KDNuggets
  - Awesome Public Datasets
  - Reddit Datasets Subreddit

Challenges

Kaggle conducted survey from August 7 to August 25, 2017 - 16,716 people responded



Can group these challenges into broader categories :

- Collaboration (76%)
- Data (68%)
- Talent (42%)
- Tools (36%)
- Budget (27%)

Barriers and Challenges at Work

1. Dirty data (36% reported)
2. Lack of data science talent (30%)
3. Company politics (27%)
4. Lack of clear question (22%)
5. Data inaccessible (22%)
6. Results not used by decision makers (18%)
7. Explaining data science to others (16%)
8. Privacy issues (14%)
9. Lack of domain expertise (14%)
10. Organization small and cannot afford data science team (13%)

Comparative Study of data science with databases

- A database simply is the place where you store the data.
- A **data engineer** (former database admin) is responsible for setting up and maintaining the infrastructure for the database.
- In contrast, a **data scientist** is not concerned about storing data. His/Her job is to actually derive meaningful insights.
- **Databases** are logical structures that are based on set theory that have relationships based on unique primary and foreign keys. These links allow processing of data between tables while holding the integrity of the data.
- **Data Science** is the study and analysis of data using methods, processes, and insights that corresponds to structured or unstructured data.
- The difference is that databases are actual objects and data science are methods, processes that corresponds to structured or unstructured data.

What is the role of SQL in data science?

- Data Scientist works with data and all the structured data is stored in databases. So, if one needs to play with data, he must need SQL
- For querying and manipulating the data we use a language that is similar to SQL known as HiveQL.
- For creating a table and test environment, Data Scientists use SQL.
- For doing analytics tasks over the data that is stored in Oracle DB or SQL Server, SQL proves better.
- When working with Big Data processing tools, you will use SQL for data preparation and wrangling

Scientific Computing/computational science

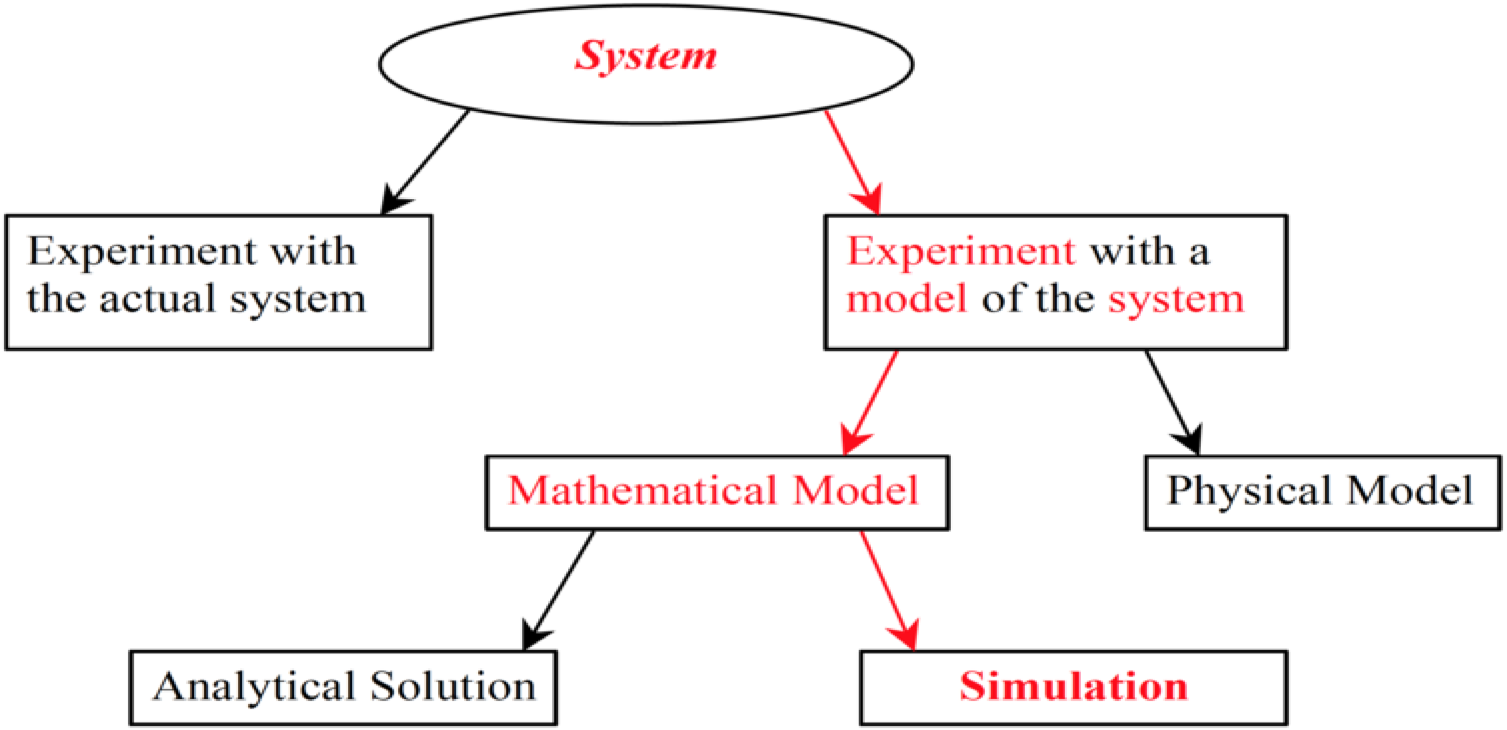
- **Scientific computing** is the science of solving problems with computers.
- **Computational science** (also **scientific computing** or **scientific computation** (**SC**)) is a rapidly growing multidisciplinary field that uses advanced computing capabilities to understand and solve complex problems.
- Considered a third mode of science

Scientific Computing/computational science

- The term **computational scientist** is used to describe someone skilled in scientific computing. This person is usually a scientist, an engineer or an applied mathematician who applies high-performance computing in different ways to advance the state-of-the-art in their respective applied disciplines in physics, chemistry or engineering.

Scientific Computing/computational science

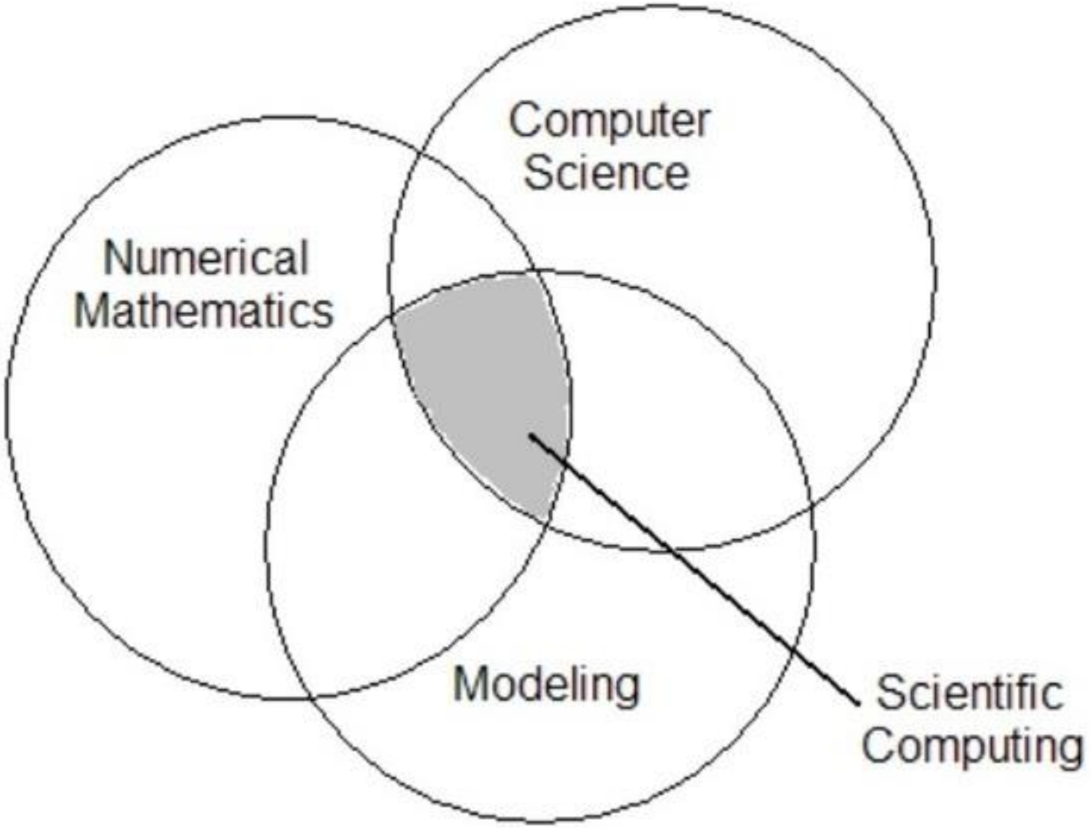
Ways to study a system


System
Experiment with the actual system
Experiment with a model of the system
Mathematical Model
Physical Model
Analytical Solution
Simulation

Scientific Computing/computational science

- A computational scientist should be capable of
  - recognizing **complex problems**
  - adequately **conceptualise** the system containing these problems
  - design a framework of algorithms suitable for studying this system: the **simulation**
  - choose a suitable **computing infrastructure** (parallel computing/grid computing/supercomputers)
  - maximising the computational power of the simulation
  - the model is **validated**
  - adjust the conceptualisation of the system accordingly
  - repeat cycle until a suitable level of validation is obtained

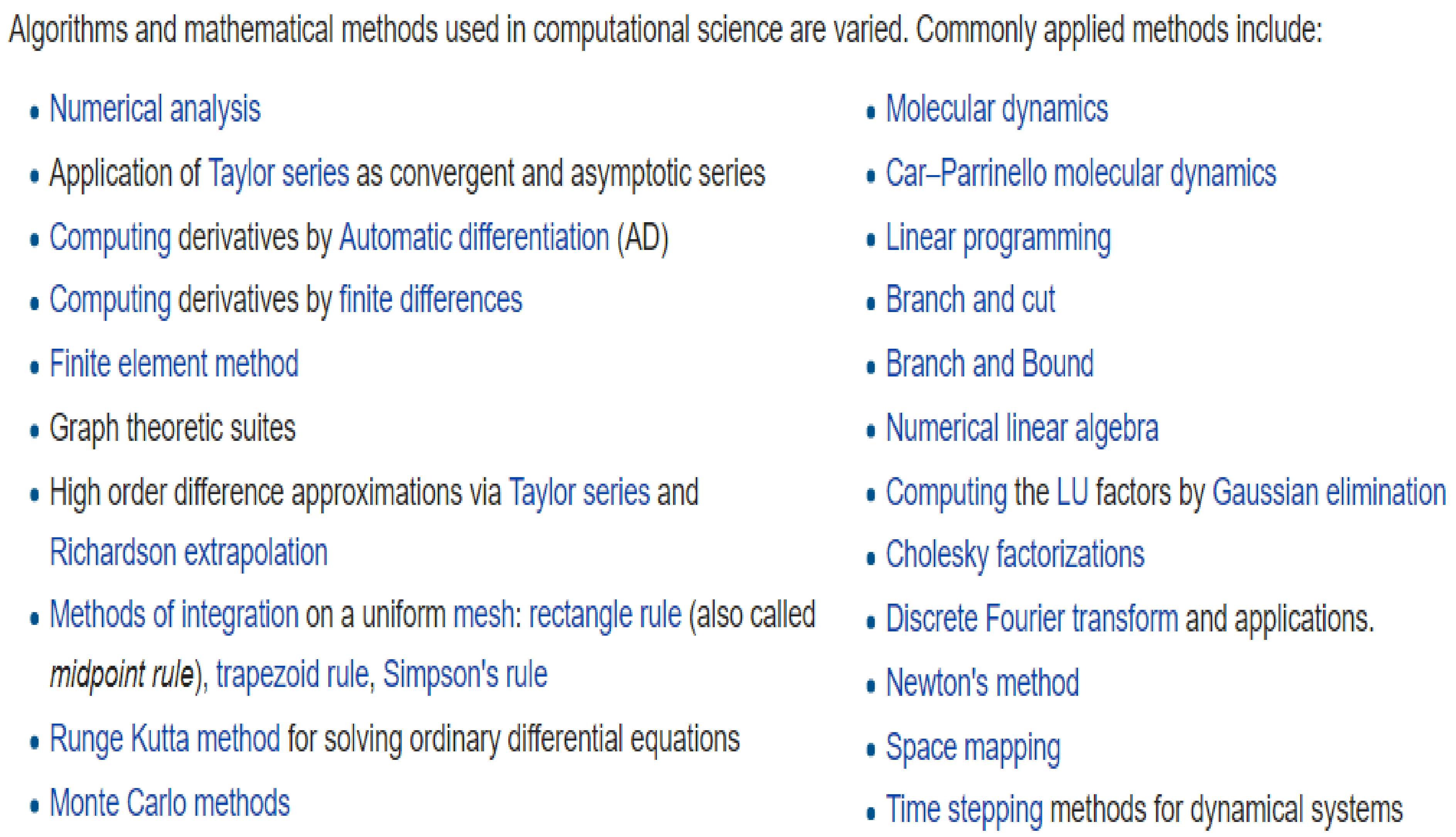
Scientific Computing/computational science



Applications of computational science

- **Urban complex systems**
- **Computational finance**
- **Computational biology**
- **Computational science in engineering**

Methods and Algorithms



Algorithms and mathematical methods used in computational science are varied. Commonly applied methods include:

- Numerical analysis
- Application of Taylor series as convergent and asymptotic series
- Computing derivatives by Automatic differentiation (AD)
- Computing derivatives by finite differences
- Finite element method
- Graph theoretic suites
- High order difference approximations via Taylor series and Richardson extrapolation
- Methods of integration on a uniform mesh: rectangle rule (also called *midpoint rule*), trapezoid rule, Simpson's rule
- Runge Kutta method for solving ordinary differential equations
- Monte Carlo methods
- Molecular dynamics
- Car–Parrinello molecular dynamics
- Linear programming
- Branch and cut
- Branch and Bound
- Numerical linear algebra
- Computing the LU factors by Gaussian elimination
- Cholesky factorizations
- Discrete Fourier transform and applications.
- Newton's method
- Space mapping
- Time stepping methods for dynamical systems

Data Science and Scientific Computing

They are not interchangeable.

- Computational science tends to refer more to HPC, simulation techniques (differential equations, molecular dynamics, etc.), and is usually referred to as scientific computing.
- Data science tends to refer to computationally-intensive data analysis, like "big data", bioinformatics, machine learning, Bayesian analyses.

Machine Learning

- **Machine learning** gives computers the ability to learn without being explicitly programmed.
- Machine learning is a subset of AI.
- **Data Science** is a broad term comprising of statistics, programming, data visualization, big data, machine learning and much more.

Data Modelling

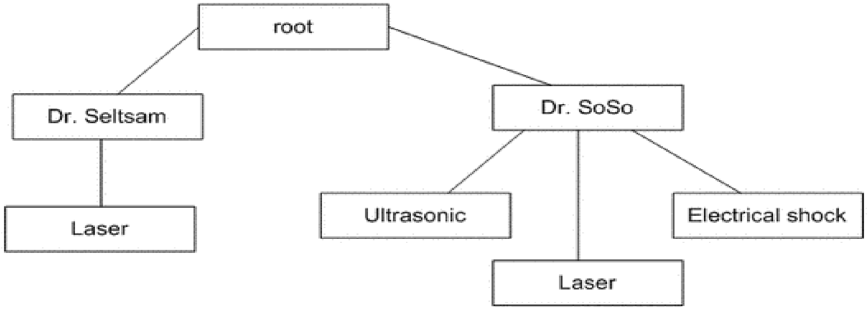
- Data modeling is the analysis of data objects and their relationships to other data objects.
- Data modeling is often the first step in database design and object-oriented programming
- Data modeling involves a progression from conceptual model to logical model to physical schema.
- Data modeling is an **important skill for data scientists** or others involved with data analysis.
- Data models have been built during the analysis and design phases of a project.

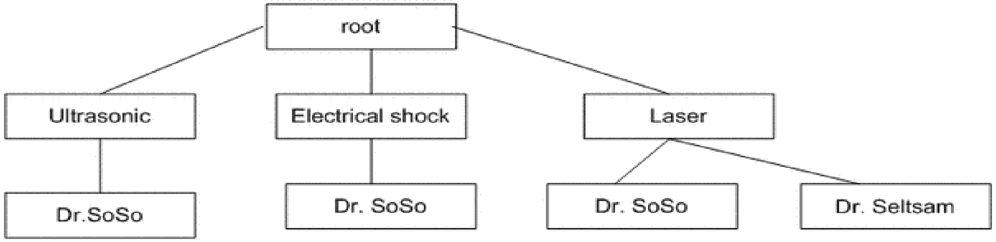
Data Modeling Approaches

- Hierarchical data modeling
- Relational data modeling
- The entity relationship model
- Graph data models

Data Modeling Approaches

- **Hierarchical data modeling**
  - Array data in treelike, one-to-many arrangements
  - IBM's Information Management System (IMS) is a primary example
  - Method is common still in **XML,** geographic information systems (**GISes)**

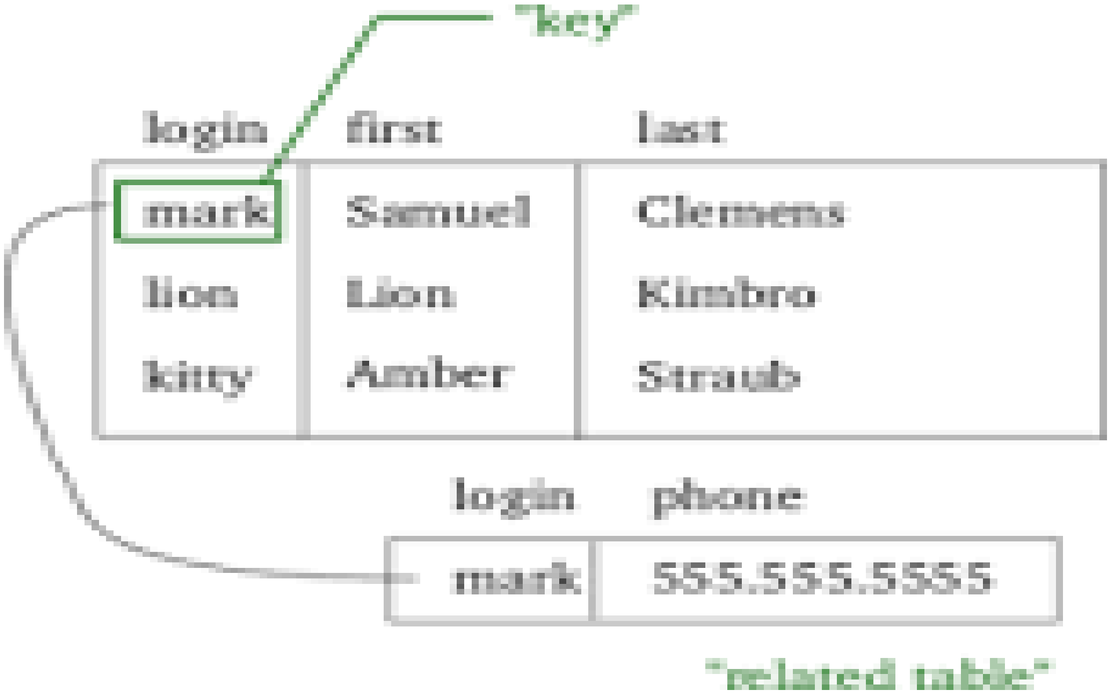




Data Modeling Approaches

❑ Relational data modeling

- ➢ Relational data modeling was first described in a 1970
- ➢ Data segments are explicitly joined by use of tables
- ➢ Relational data model was coupled with SQL



"key"

login	first	last
mark	Samuel	Clemens
lion	Lion	Kimbro
kitty	Amber	Straub

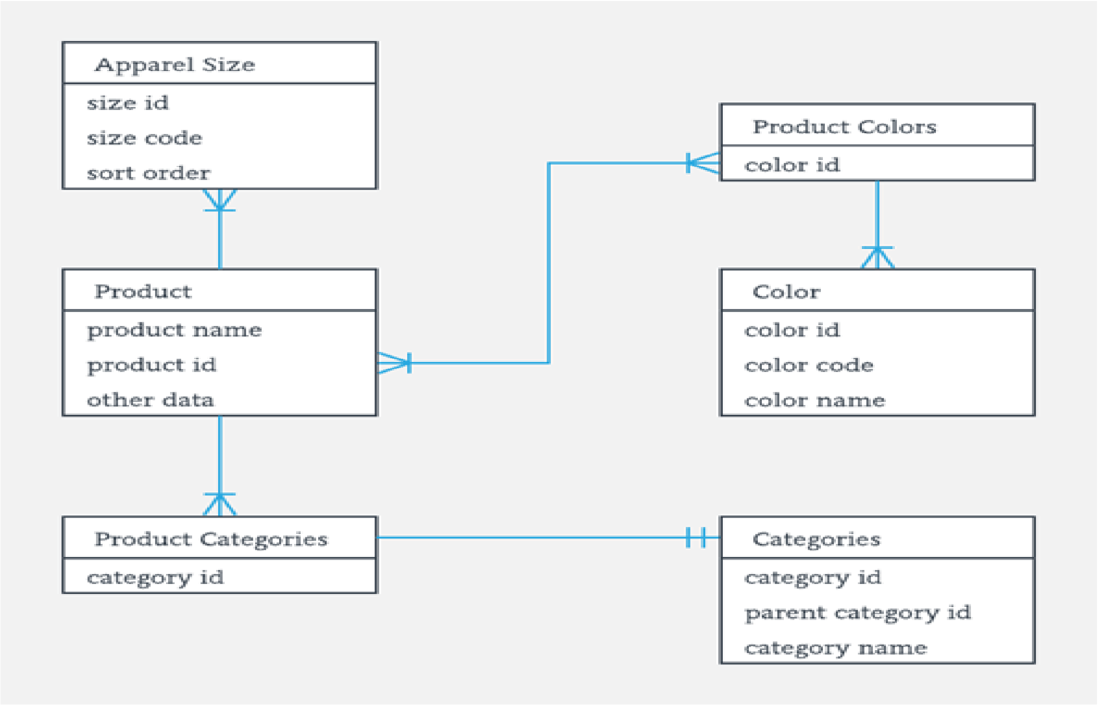
login	phone
mark	555.555.5555

"related table"

Data Modeling Approaches

❑ Entity Relationship Model

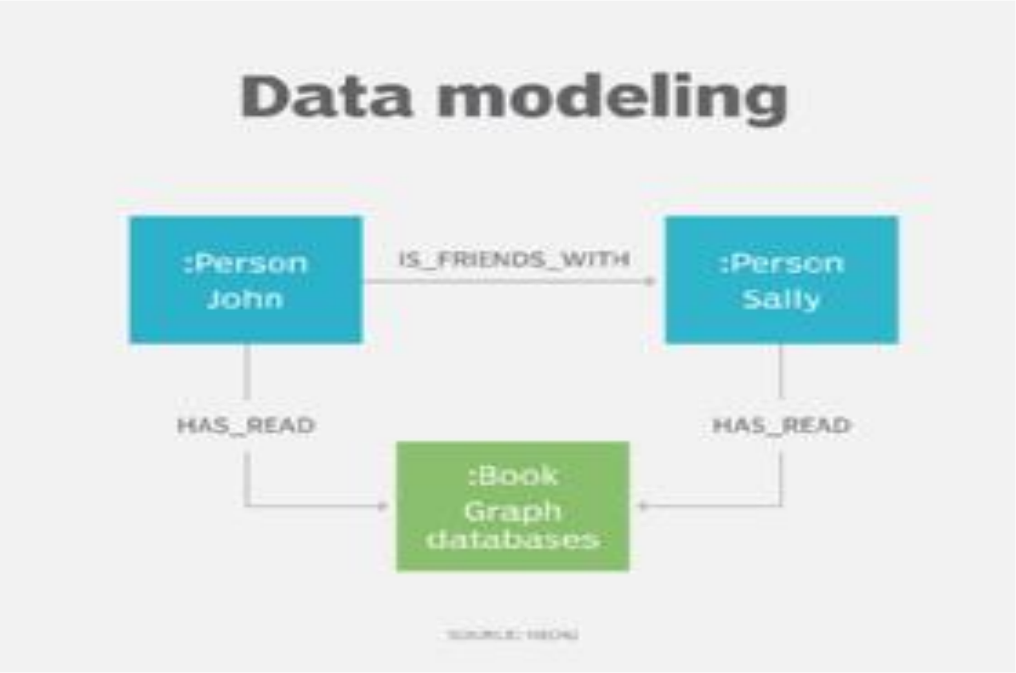
- ➢ Closely integrated with relational data models
- ➢ ER models use diagrams to graphically depict the elements in a database
- ➢ Relationships are visually mapped, providing a ready means to communicate data design



Data Modeling Approaches

❑ Graph data models

- ➢ Used with graph databases
- ➢ Used for describing complex relationships within data sets particularly in Social Media



SOURCE: NEO4J

Statistical Data Modeling

- ❑ A statistical model is a special class of mathematical model.
- ❑ Statistical model is non-deterministic
- ❑ **Statistical modeling** is a simplified, mathematically-formalized way to approximate reality and optionally to make predictions from this approximation.
- ❑ The statistical model is the mathematical equation that is used.

Statistical Data Modeling techniques

- ❑ Linear Regression
- ❑ Non-linear regression
- ❑ Logistic Regression
- ❑ Multivariate analysis

Statistical limits on data- Bonferroni's principle

- Suppose you have a certain amount of data, and you look for events of a certain type within that data.
- Calculate the expected number of occurrences of the events you are looking for, on the assumption that data is random.
- If this number is significantly larger than the number of real instances you hope to find, then you must expect almost anything you find to be bogus.
- i.e., a statistical artifact rather than evidence of what you are looking for.
- This observation is the informal statement of Bonferroni’s principle.

Statistical limits on data- Bonferroni's principle

- **Bonferroni's Principle** is an informal presentation of a statistical theorem that states if your method of finding significant items returns significantly more items that you would expect in the actual population, you can assume most of the items you find with it **are bogus**.

- This essentially means that an algorithm or method we think is useful for finding a particular set of data actually returns **more false positives** as it returns larger portion of the data than should be within that category.

Statistical limits on data- Bonferroni's principle

❑ Applying Bonferroni's Principle to an algorithm or system for identifying or classifying data gives an **upper bound on the accuracy** of your methods.

❑ Not to say that the **algorithm is correct** in the case that it matches a number relatively close to what you would expect.

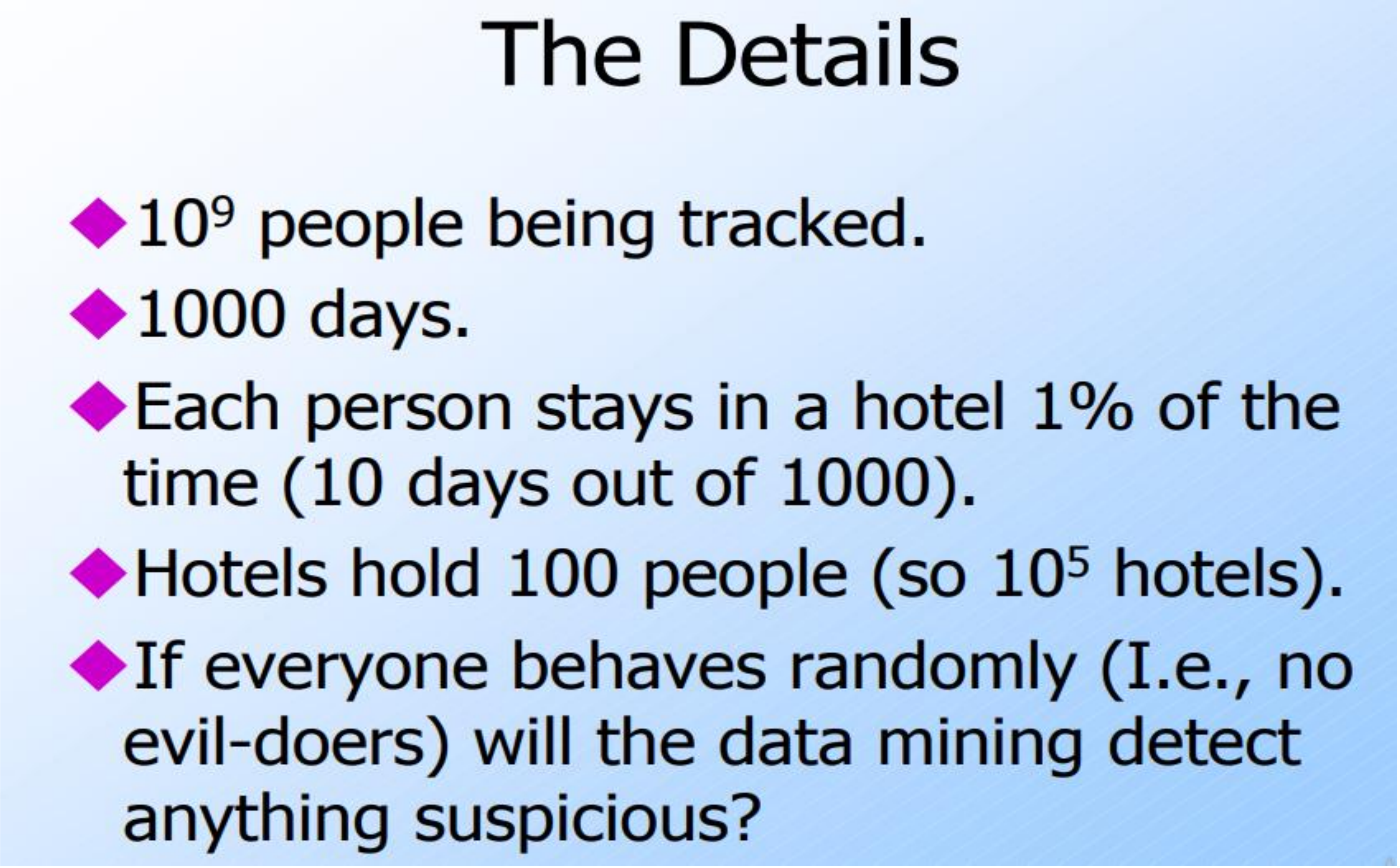
❑Bonferroni Principle is a statistical method for accounting for random events.

Statistical limits on data- Bonferroni's Principle

- ❑ **Problem Definition:** Suppose that "evil-doers" periodically gather at a hotel to plot their evil. We want to detect them based on the following assumptions:

- ❑ To find "evil-doers", we shall look for people who, on two different days, were both at the same hotel.

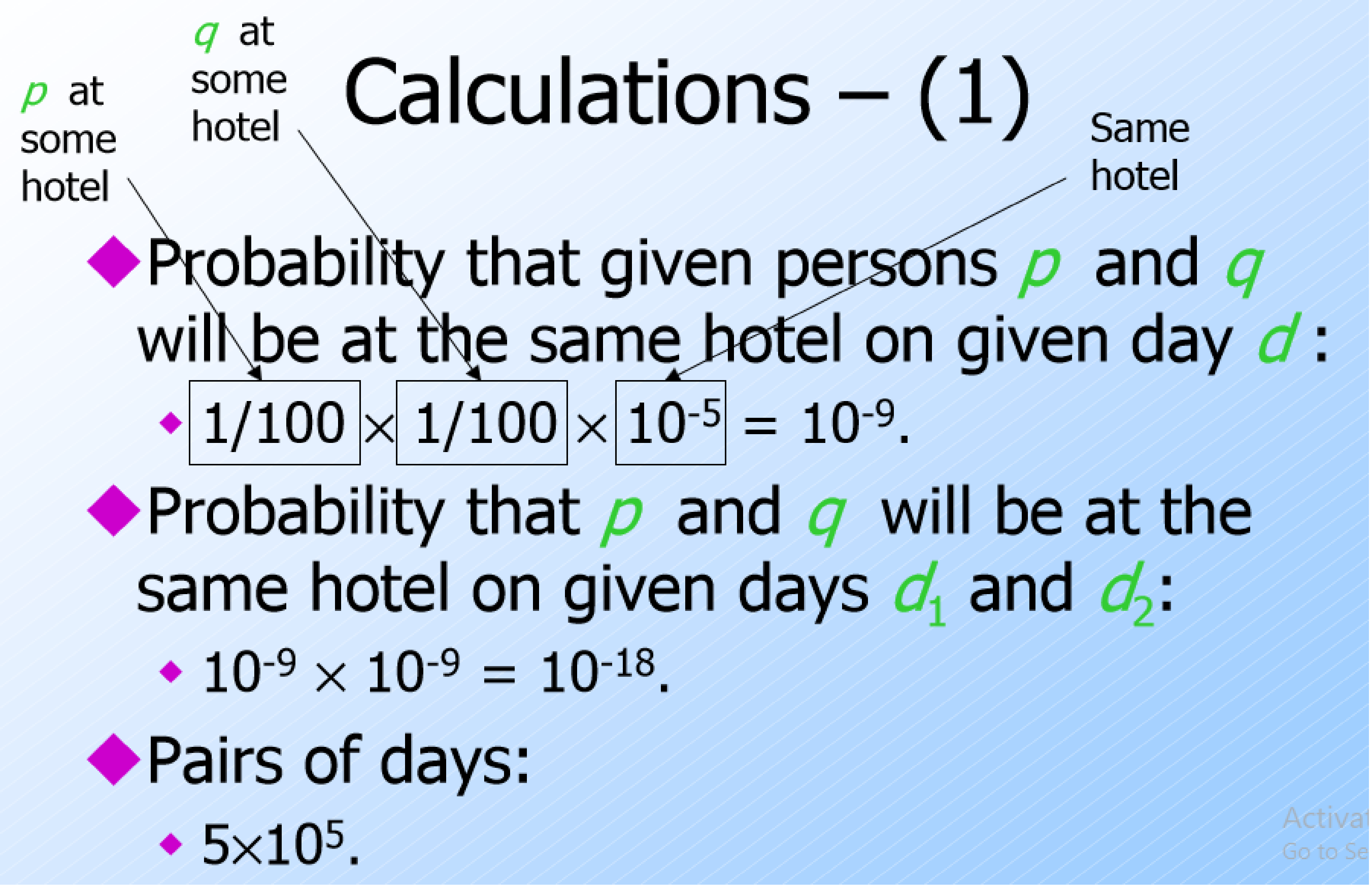
Bonferroni's Principle



The Details

- 10^9 people being tracked.
- 1000 days.
- Each person stays in a hotel 1% of the time (10 days out of 1000).
- Hotels hold 100 people (so 10^5 hotels).
- If everyone behaves randomly (I.e., no evil-doers) will the data mining detect anything suspicious?

Bonferroni's Principle



p at some hotel

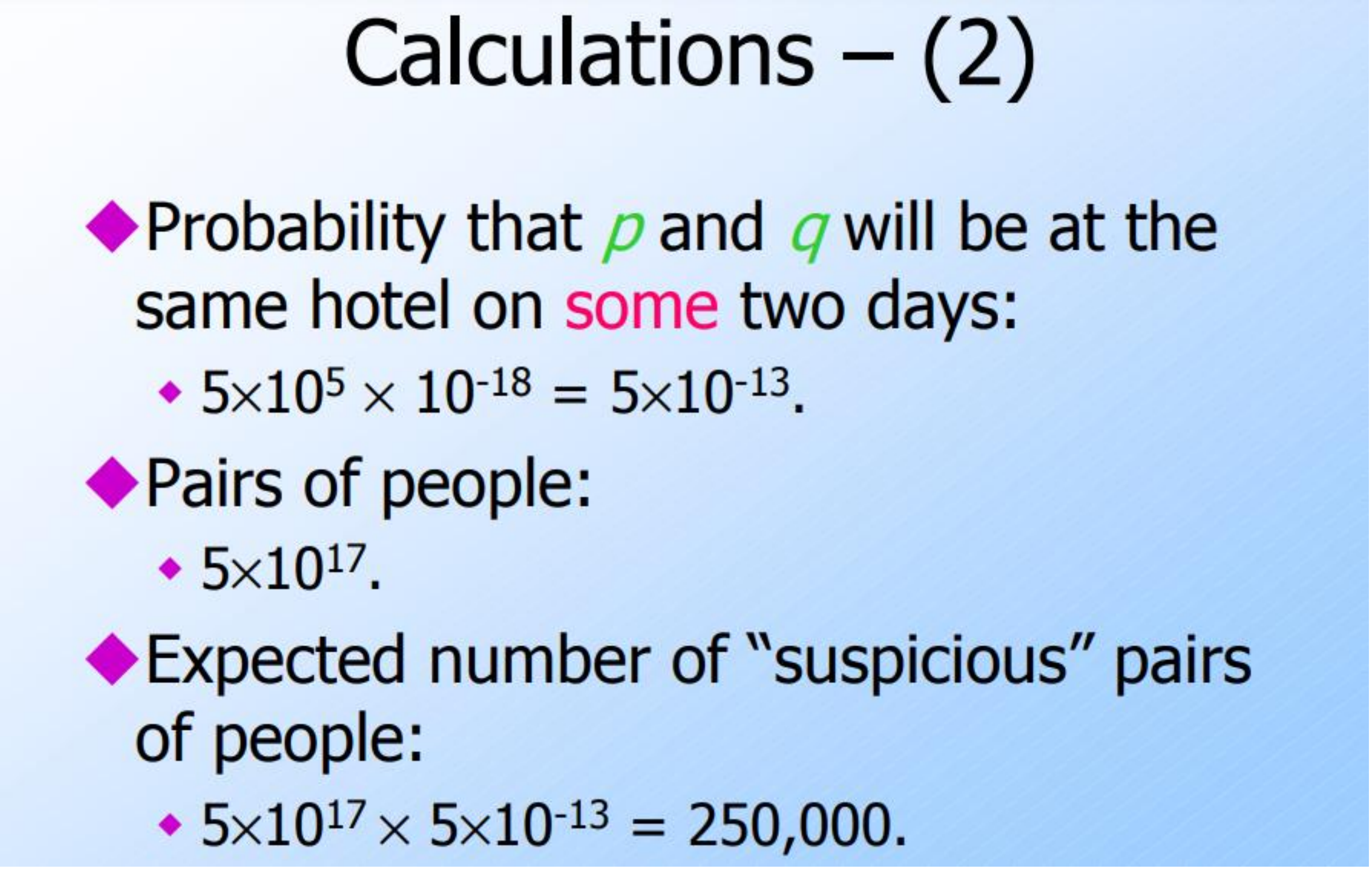
q at some hotel

Calculations – (1)

Same hotel

- Probability that given persons *p* and *q* will be at the same hotel on given day *d* :
  - $\boxed{1/100} \times \boxed{1/100} \times \boxed{10^{-5}} = 10^{-9}$.
- Probability that *p* and *q* will be at the same hotel on given days d_1 and d_2:
  - $10^{-9} \times 10^{-9} = 10^{-18}$.
- Pairs of days:
  - 5×10^5.

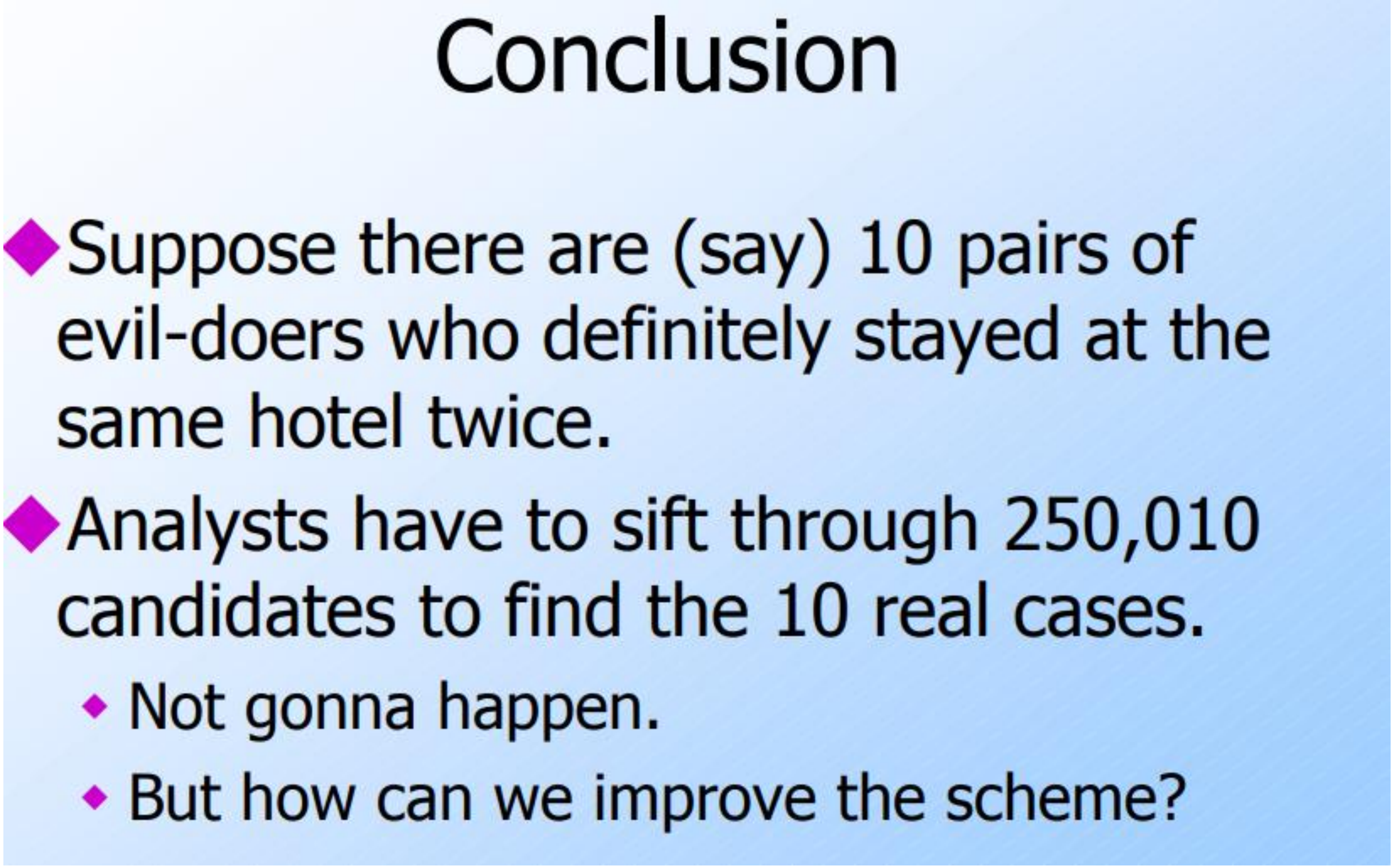
Bonferroni's principle



Calculations – (2)

- Probability that *p* and *q* will be at the same hotel on some two days:
  - $5 \times 10^{5} \times 10^{-18} = 5 \times 10^{-13}$.
- Pairs of people:
  - 5×10^{17}.
- Expected number of "suspicious" pairs of people:
  - $5 \times 10^{17} \times 5 \times 10^{-13} = 250{,}000$.

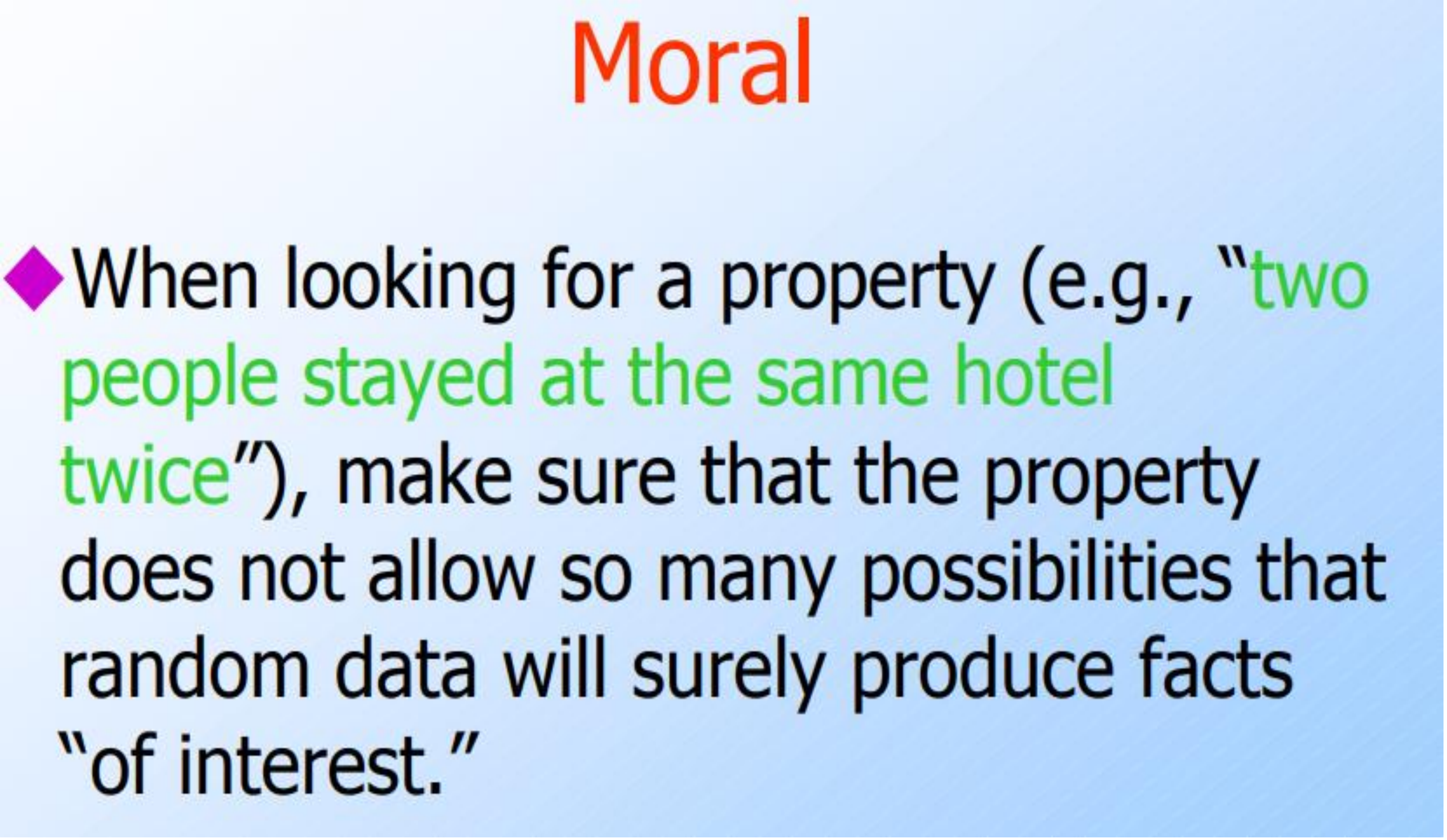
Bonferroni's Principle



Conclusion

- Suppose there are (say) 10 pairs of evil-doers who definitely stayed at the same hotel twice.
- Analysts have to sift through 250,010 candidates to find the 10 real cases.
  - Not gonna happen.
  - But how can we improve the scheme?

Bonferroni's Principle



Moral

- When looking for a property (e.g., "two people stayed at the same hotel twice"), make sure that the property does not allow so many possibilities that random data will surely produce facts "of interest."

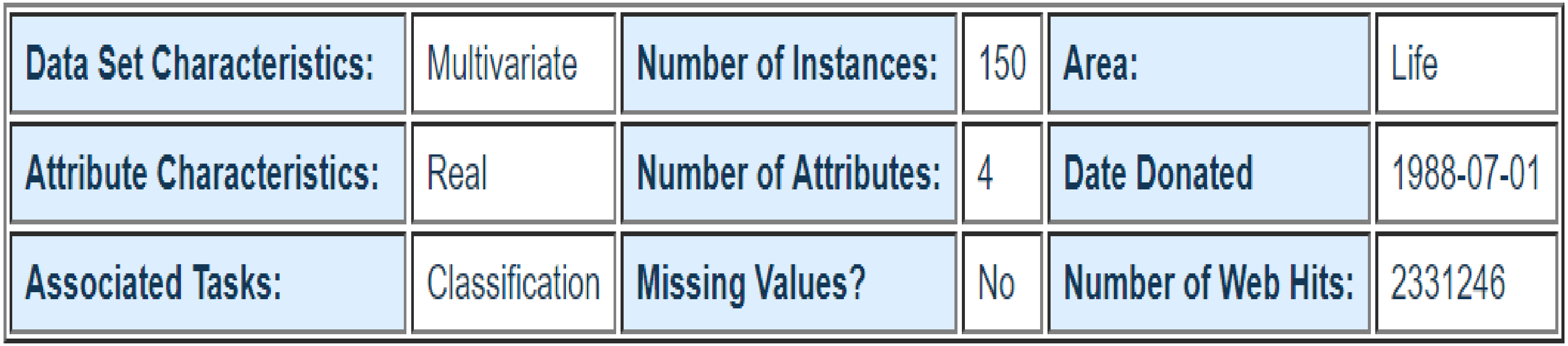
Data Visualization Techniques

- **Data visualization** is viewed by many disciplines as a modern equivalent of visual communication.
- It involves the creation and study of the visual representation of data.
- To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools.
- It makes complex data more accessible, understandable and usable.
- Data visualization is both an art and a science.

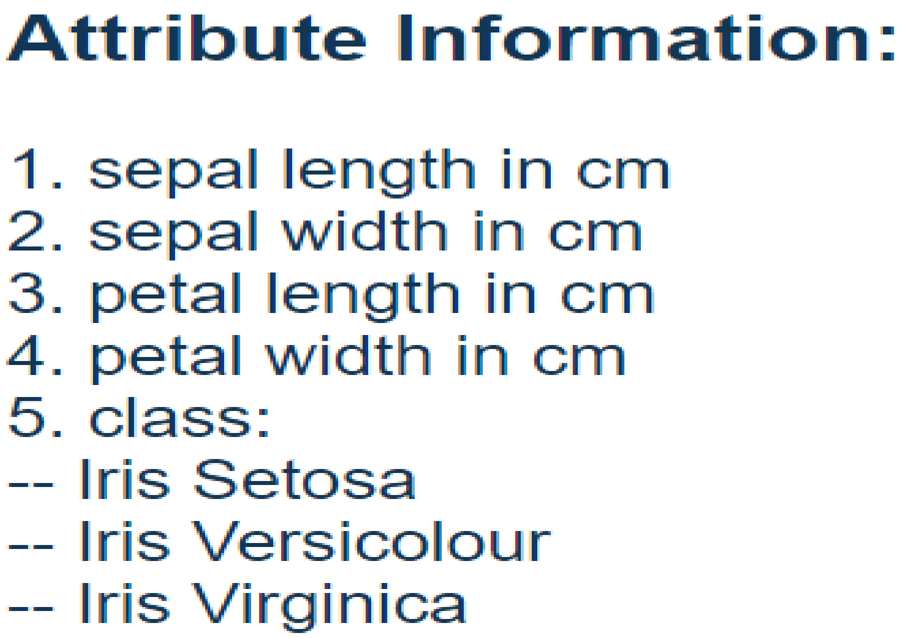
Data Visualization Techniques

- ❑ Pie Charts
- ❑ Bar Charts
- ❑ Histograms
- ❑ Boxplots
- ❑ Scatterplots
- ❑ Scatterplots with logarithmic Axes
- ❑ Scatter Matrices

Iris Data Set



Data Set Characteristics:	Multivariate	**Number of Instances:**	150	**Area:**	Life
Attribute Characteristics:	Real	**Number of Attributes:**	4	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	No	**Number of Web Hits:**	2331246



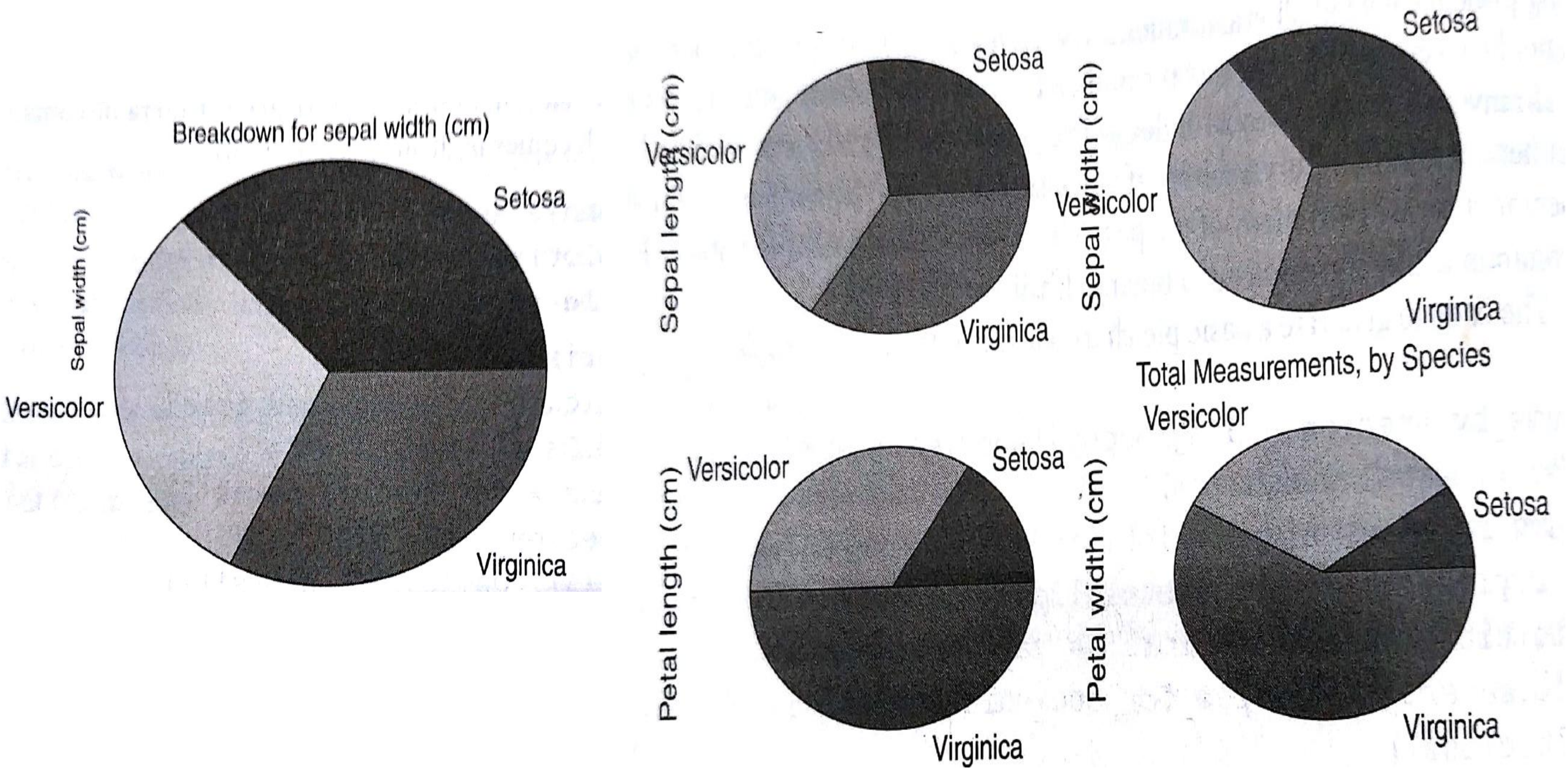
Attribute Information:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:

-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

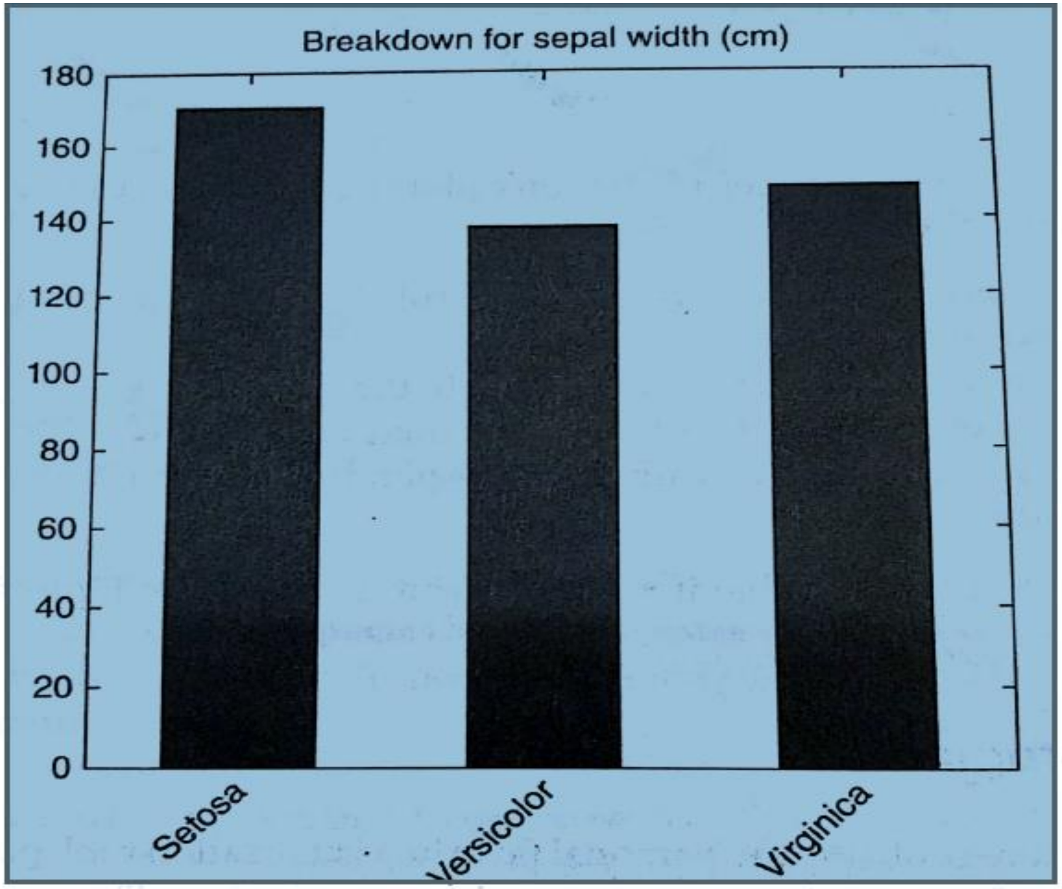
Pie Charts

- ❑ One of clearest way to present data
- ❑ You could equally get by looking at numbers

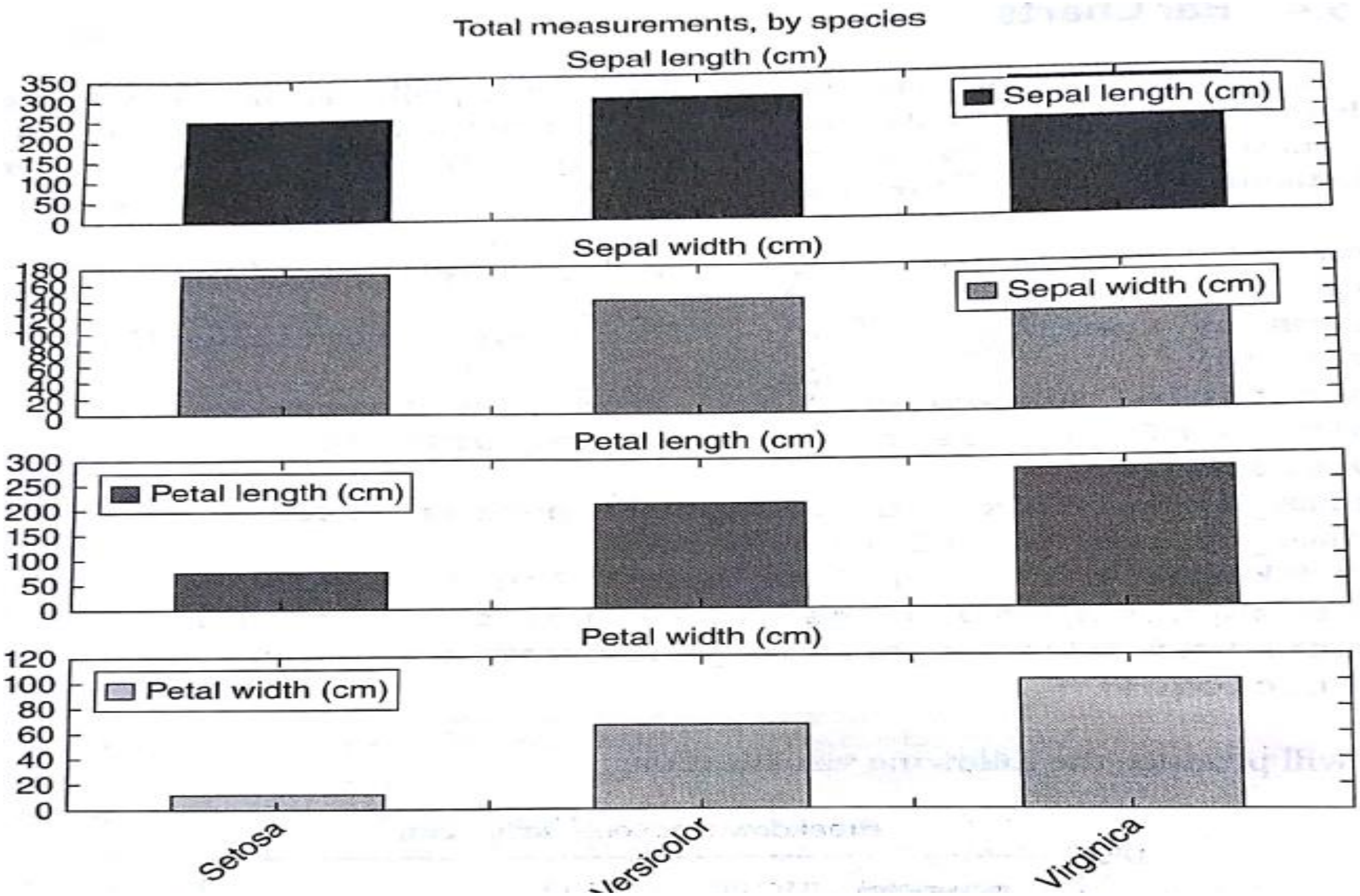


Bar Charts

- ❑ Same information as Pie Chart can be conveyed
- ❑ Interested in relative sizes of different flowers rather than how big a slice of the flower spices



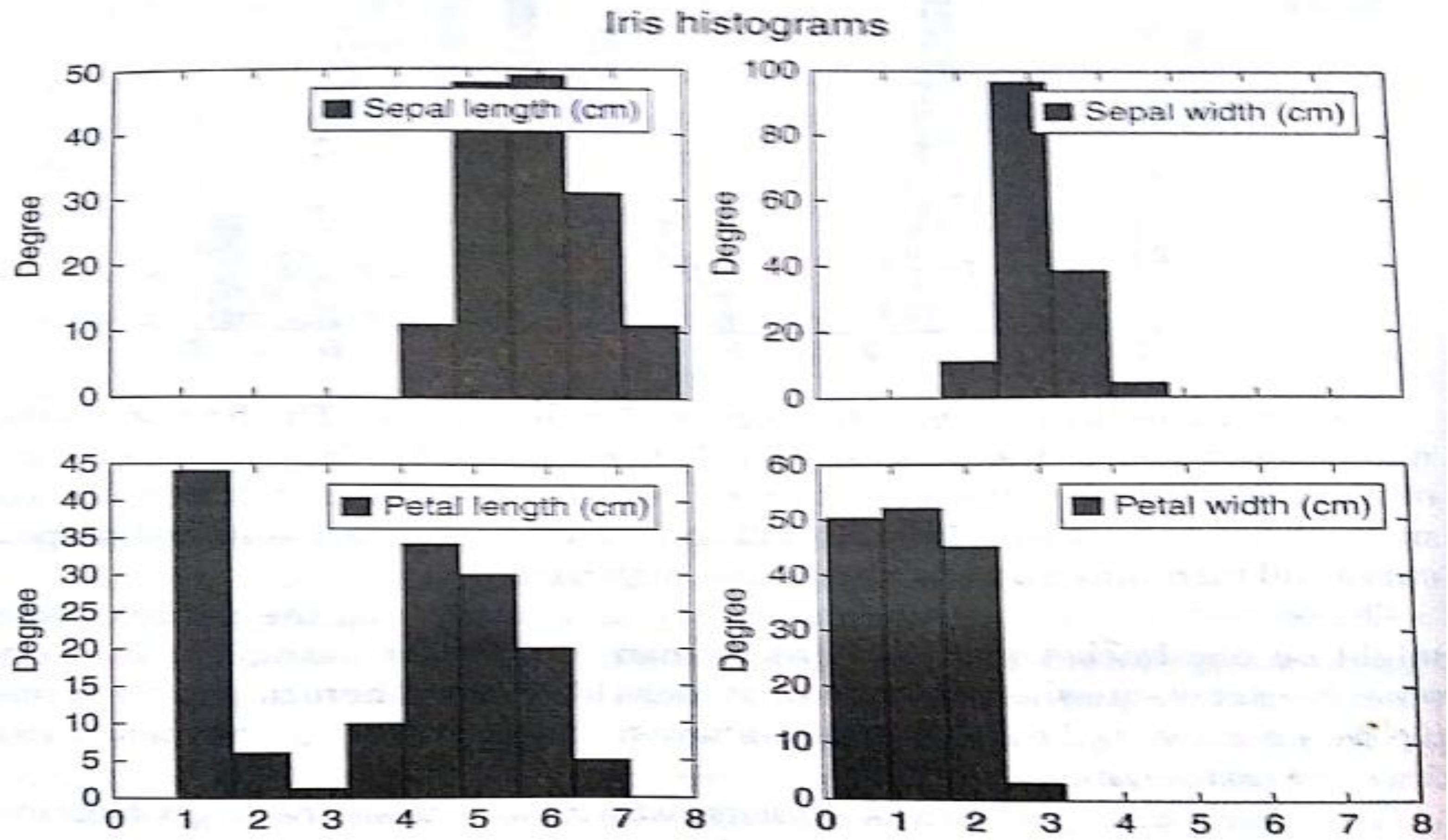
Bar Charts



Total measurements, by species
Sepal length (cm)
350
300
250
200
150
100
50
0
Sepal length (cm)
Sepal width (cm)
180
160
140
120
100
80
60
40
20
0
Sepal width (cm)
Petal length (cm)
300
250
200
150
100
50
0
Petal length (cm)
Petal width (cm)
120
100
80
60
40
20
0
Petal width (cm)
Setosa
Versicolor
Virginica

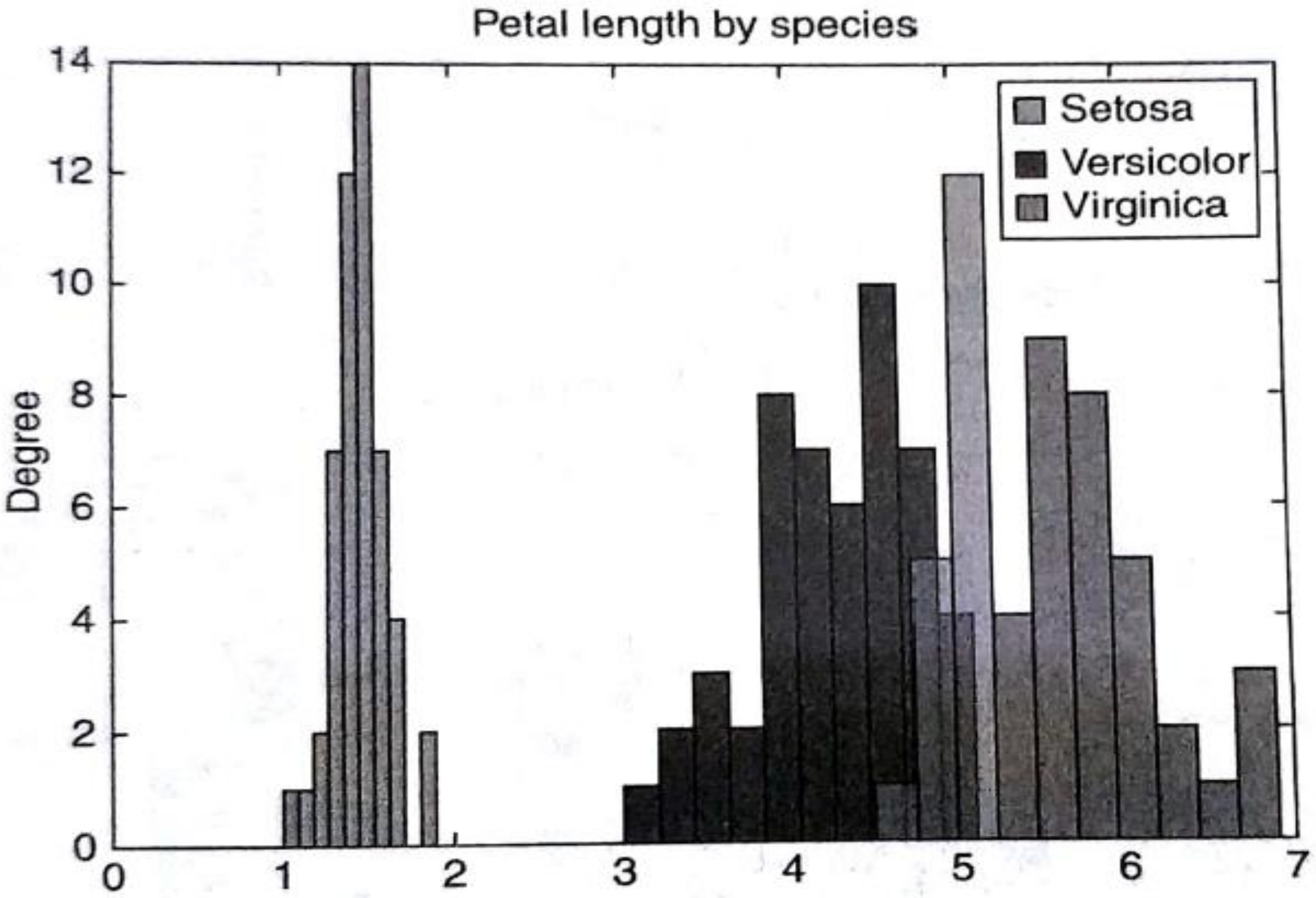
Histograms

- ❑ Favorite visualization tool, as it contains interesting information
- ❑ Bumps in the histogram corresponds to classes of real-world entities
- ❑ Can get sense of few outliers, variation in the population and so on



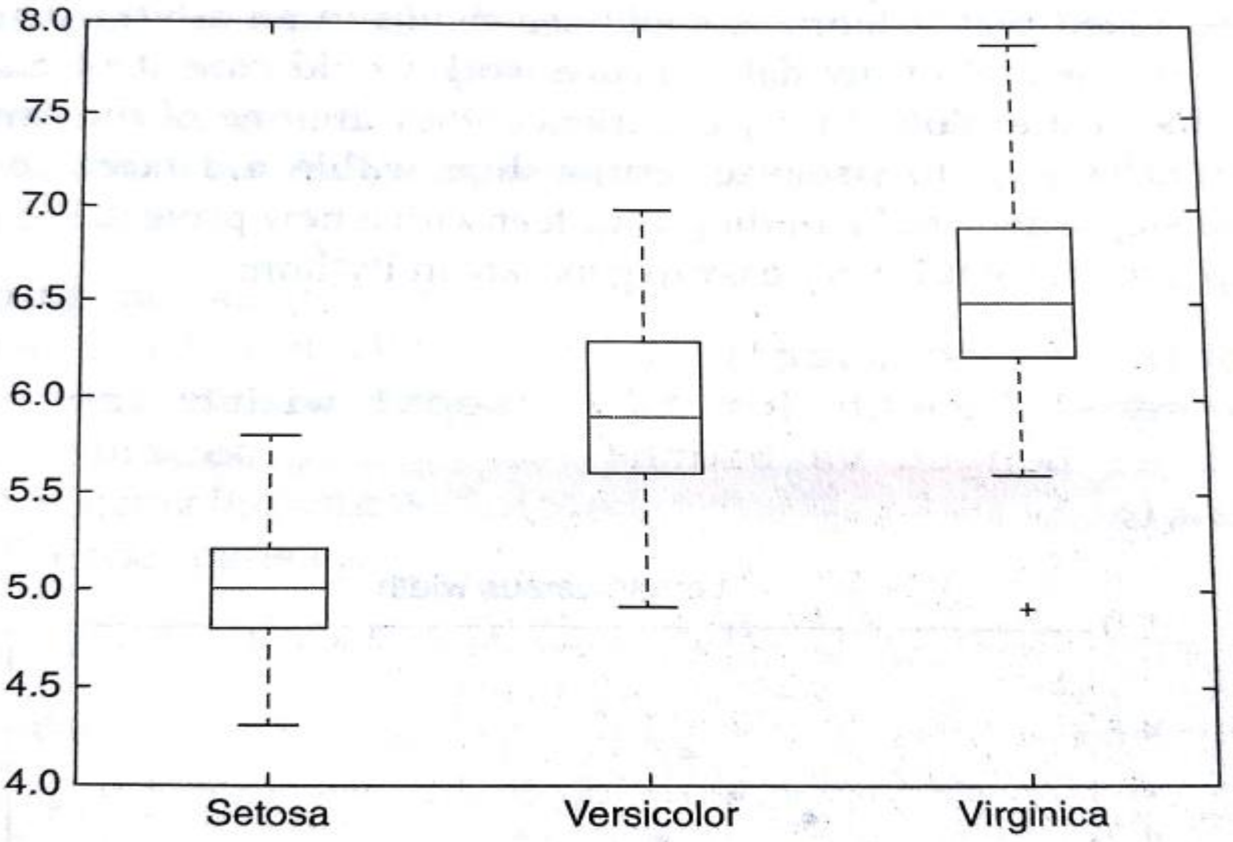
Histograms

- ❑ Can plot each spices separately



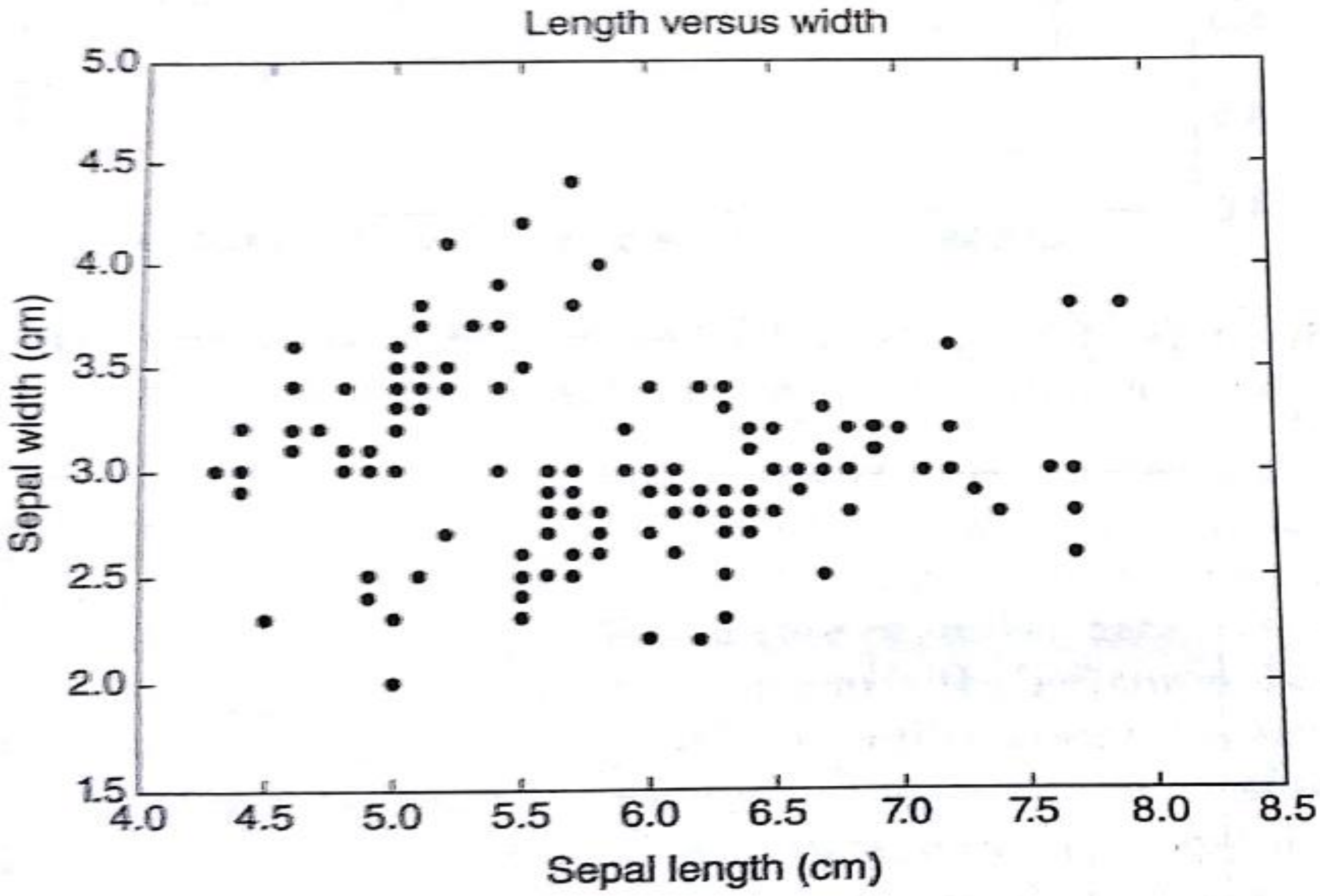
Boxplots

- ❑ Convenient way to summarize the dataset by showing median quantities, min/max values for each of the variable
- ❑ Following boxplot of sepal length for each of spices in Iris dataset



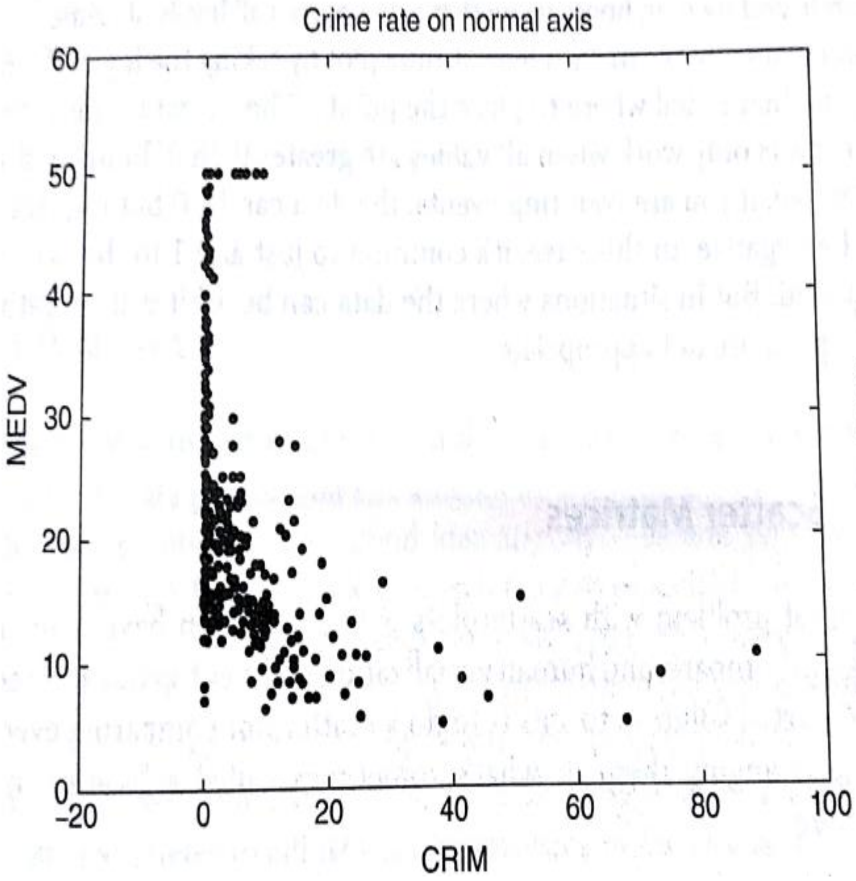
Scatterplots

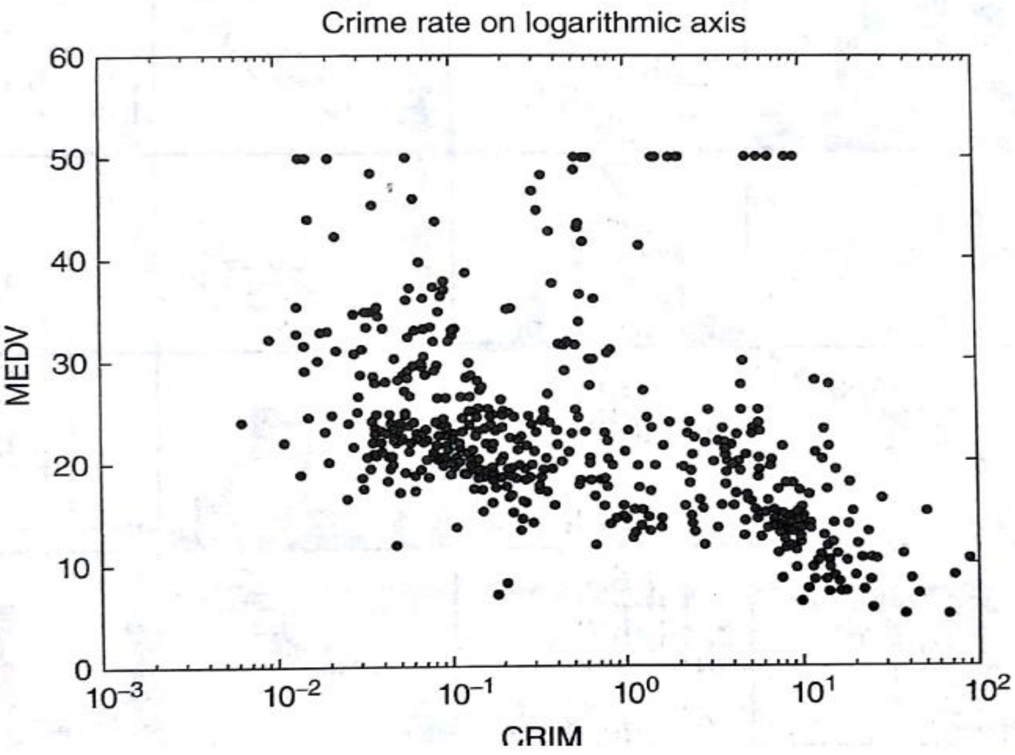
- ❑ One of simplest but powerful ways to visualize relationship within dataset
- ❑ First step when you are finding your feet with new project



Scatterplots with Logarithmic Axes

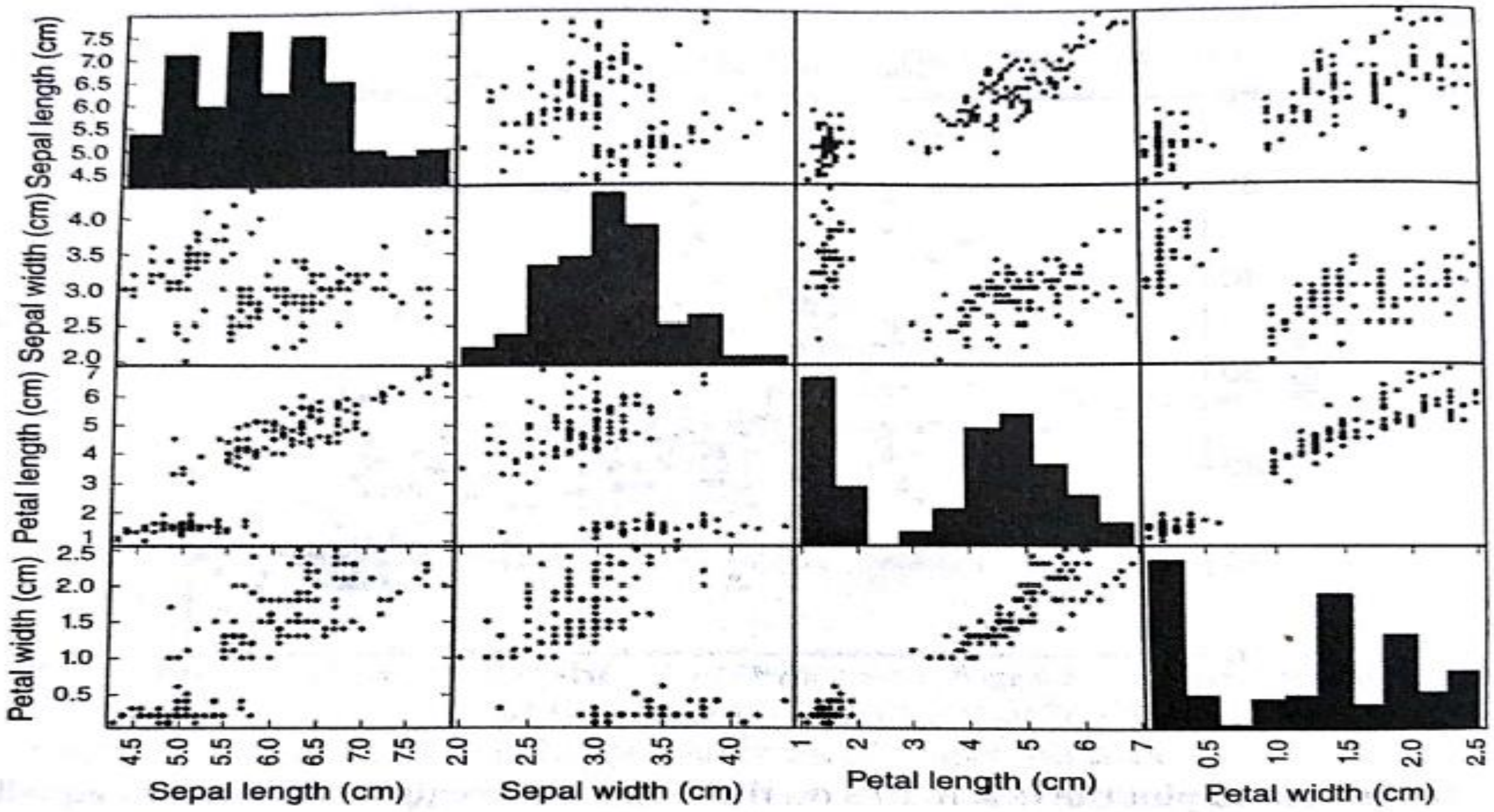
- Numbers plotted may vary by order of magnitude
- With normal scatterplot data points are squashed to left





Scatter Matrices

- ❑ Problem with scatterplots is often have to compare many variables
- ❑ Difficult with scatterplots to compare different variables
- ❑ Arrange features in scatter matrix



Thank You !!!