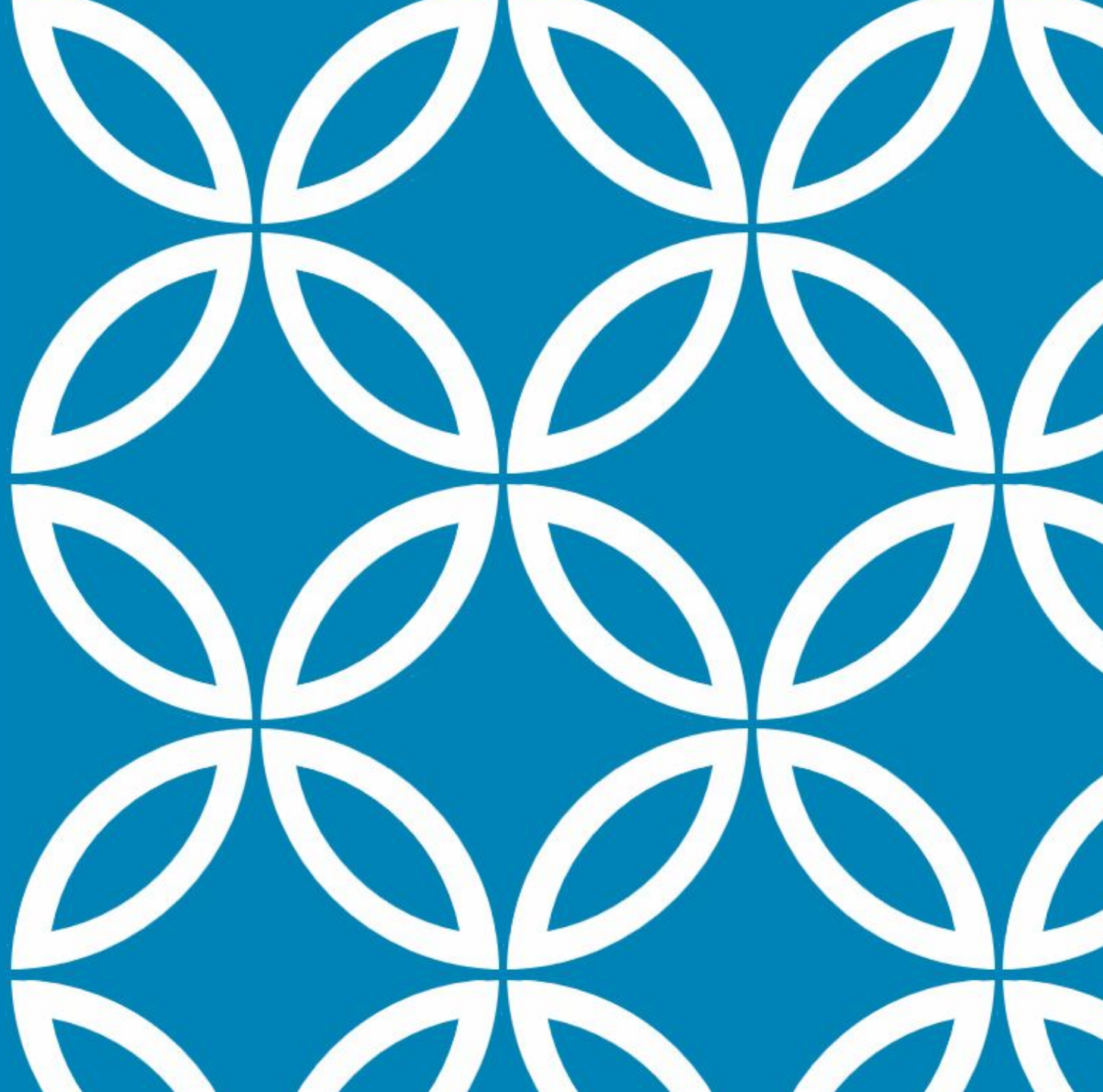


INTRODUCTION TO DATA SCIENCE

Dr. Bhakti Palkar



CONTENT

1. What is Data Science? Importance and applications of Data Science, Data Science workflow, The 5 V's of Data Science
2. Structured, semi-structured, and unstructured data. Challenges and considerations for handling different data types. Ethical Considerations in Data Science.

DATA ALL AROUND

Lots of data is being collected
and warehoused

Web data, e-commerce

Financial transactions, bank/credit transactions

Online trading and purchasing

Social Network

Google processes 20 PB a day (2008)

Facebook has 60 TB of daily logs



HOW MUCH DATA IS BIG DATA?

Google processes 20 Petabytes(PB) per day (2008)

Facebook has 2.5 PB of user data + 15 TB per day (2009)

eBay has 6.5 PB of user data + 50 TB per day (2009)

CERN's **Large Hadron Collider(LHC)** generates 15 PB a year



BIG DATA

Big Data is any data that is expensive to manage and hard to extract value from

Volume-The size of the data, Big data tools use Distributed systems

Velocity-The latency of data processing relative to the growing demand for interactivity, Eg. Social media messages going viral in seconds

Variety-the diversity of sources, formats, quality, structures. 80% of world data is unstructured.

So extremely large data sets may be analyzed computationally to reveal patterns, trends, and associations that are not transparent or easy to identify.

VOLUME

30 *BILLION* RFID TAGS TODAY
(1.3B IN 2005)

4.6
billion
camera
phones
world wide

12+ *TBs*
of tweet data
every day

? *TBs* of
data every
day



25+ *TBs* of
log data every
day



76 *million* smart meters in
2009... 200M by 2014

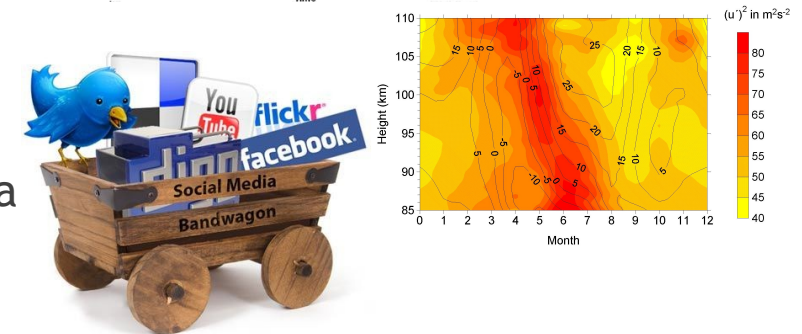
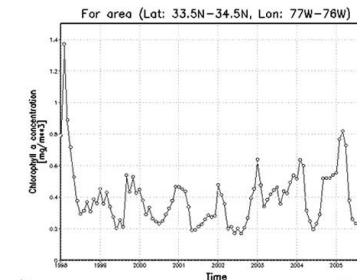
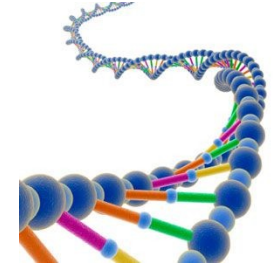
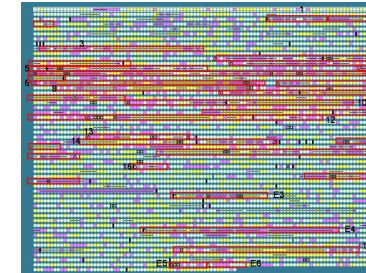


100s of
millions of
GPS
enabled
devices
sold
annually

2+
billion people on the Web
by end 2011

VARIETY

- ❖ Relational Data (Tables/Transaction/Legacy Data)
- ❖ Text Data (Web)
- ❖ Semi-structured Data (XML)
- ❖ Graph Data
 - ❖ Social Network, Semantic Web (RDF), ...
- ❖ Streaming Data
 - ❖ You can only scan the data once
- ❖ A single application can be generating/collecting many types of data
- ❖ Big Public Data (online, weather, finance, etc)



VELOCITY

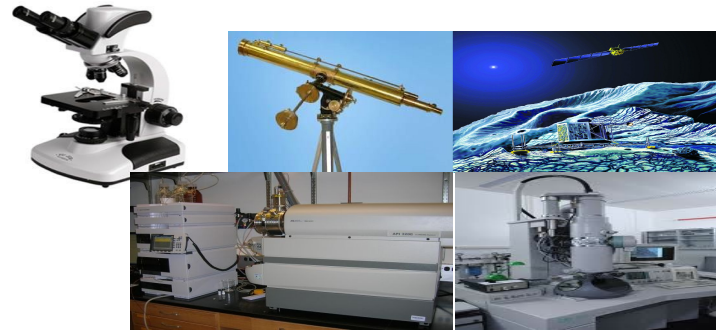
- ▶ Data is being generated fast and need to be processed fast
- ▶ Late decisions - missing opportunities
- ▶ **Examples**
 - ▶ **E-Promotions:** Based on your current location, your purchase what you like , send promotions right now for store next to you
 - ▶ **Healthcare monitoring:** sensors monitoring your activities and body, any abnormal measurements require immediate reaction



REAL-TIME/FAST DATA



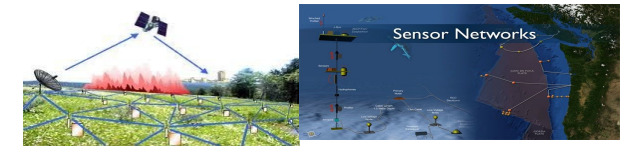
Social media and networks
(all of us are generating data)



Scientific instruments
(collecting all sorts of data)



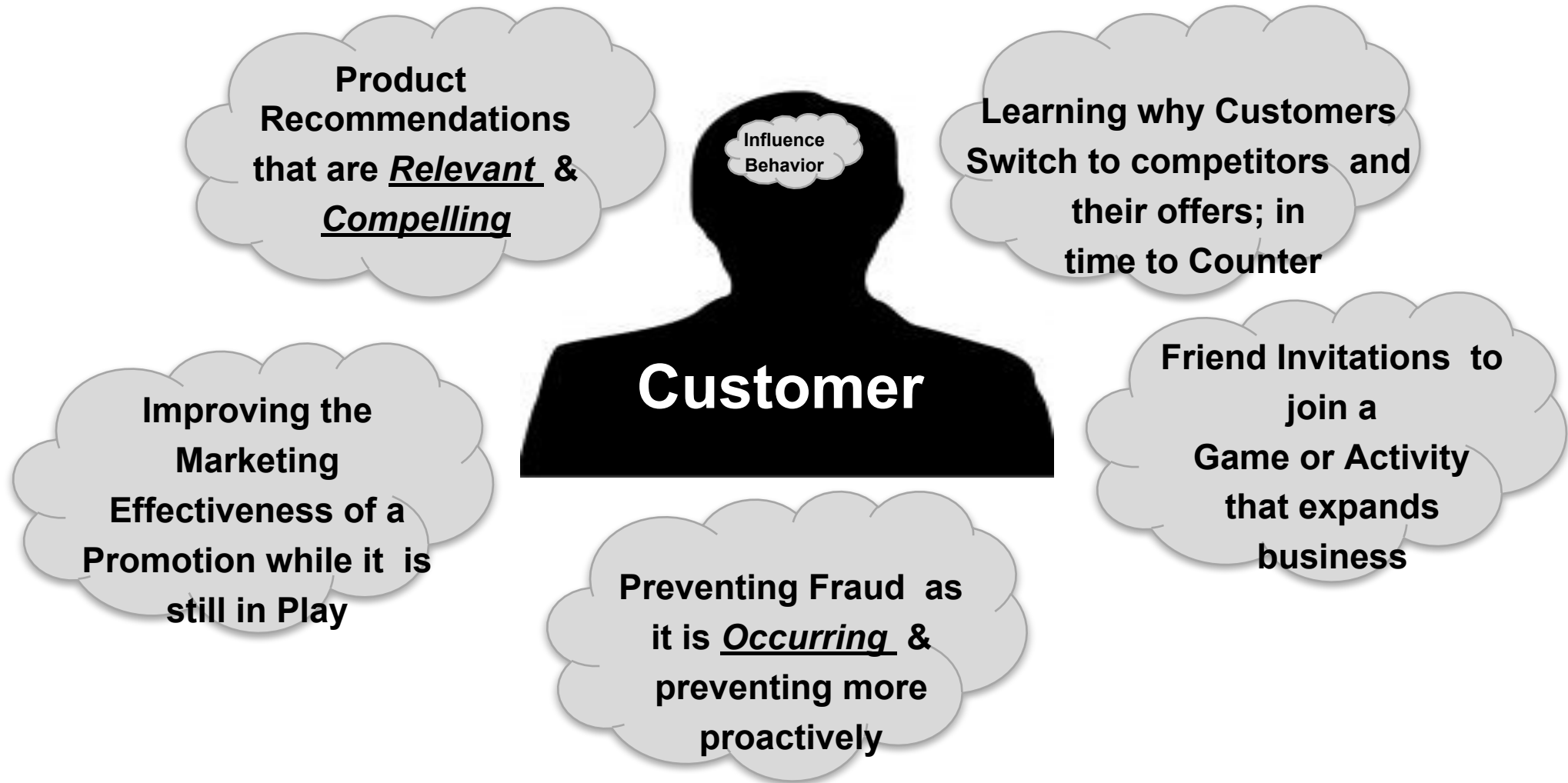
Mobile devices
(tracking all objects all the time)



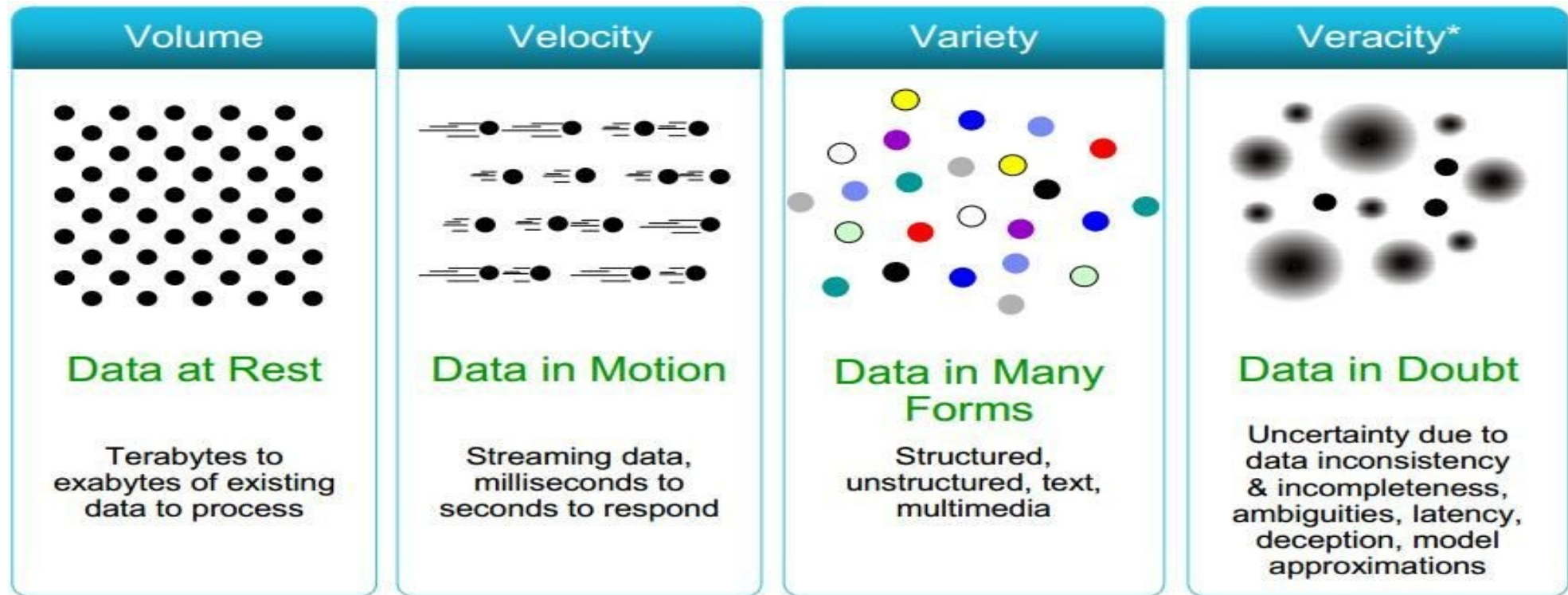
Sensor technology and networks
(measuring all kinds of data)

- ▶ The progress and innovation is no longer hindered by the ability to collect data
- ▶ But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

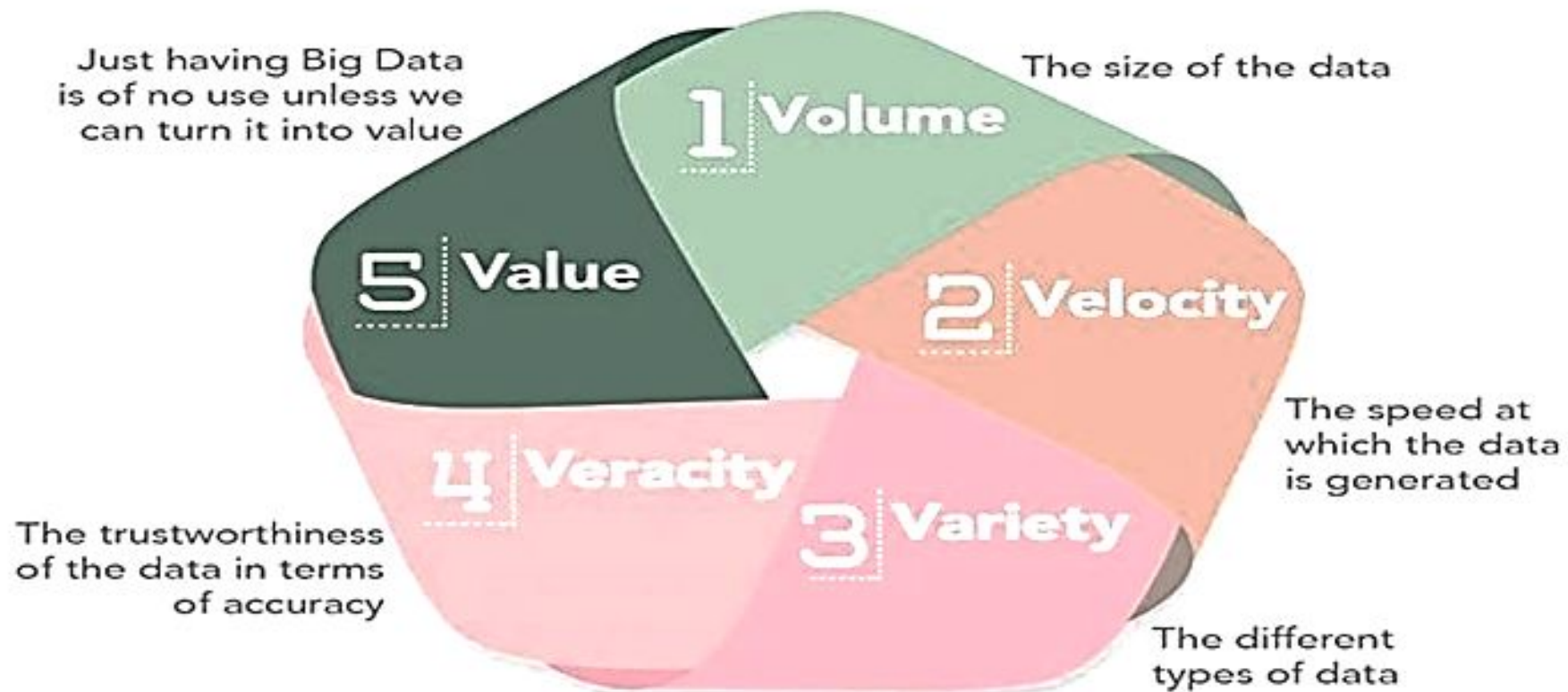
REAL-TIME ANALYTICS/DECISION REQUIREMENT



SOME MAKE IT 4V'S



5V'S OF DATA SCIENCE



THE MODEL HAS CHANGED...

► The Model of Generating/Consuming Data has Changed

Old Model: Few companies are generating data, all others are consuming data



New Model: all of us are generating data, and all of us are consuming data



BIG DATA

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

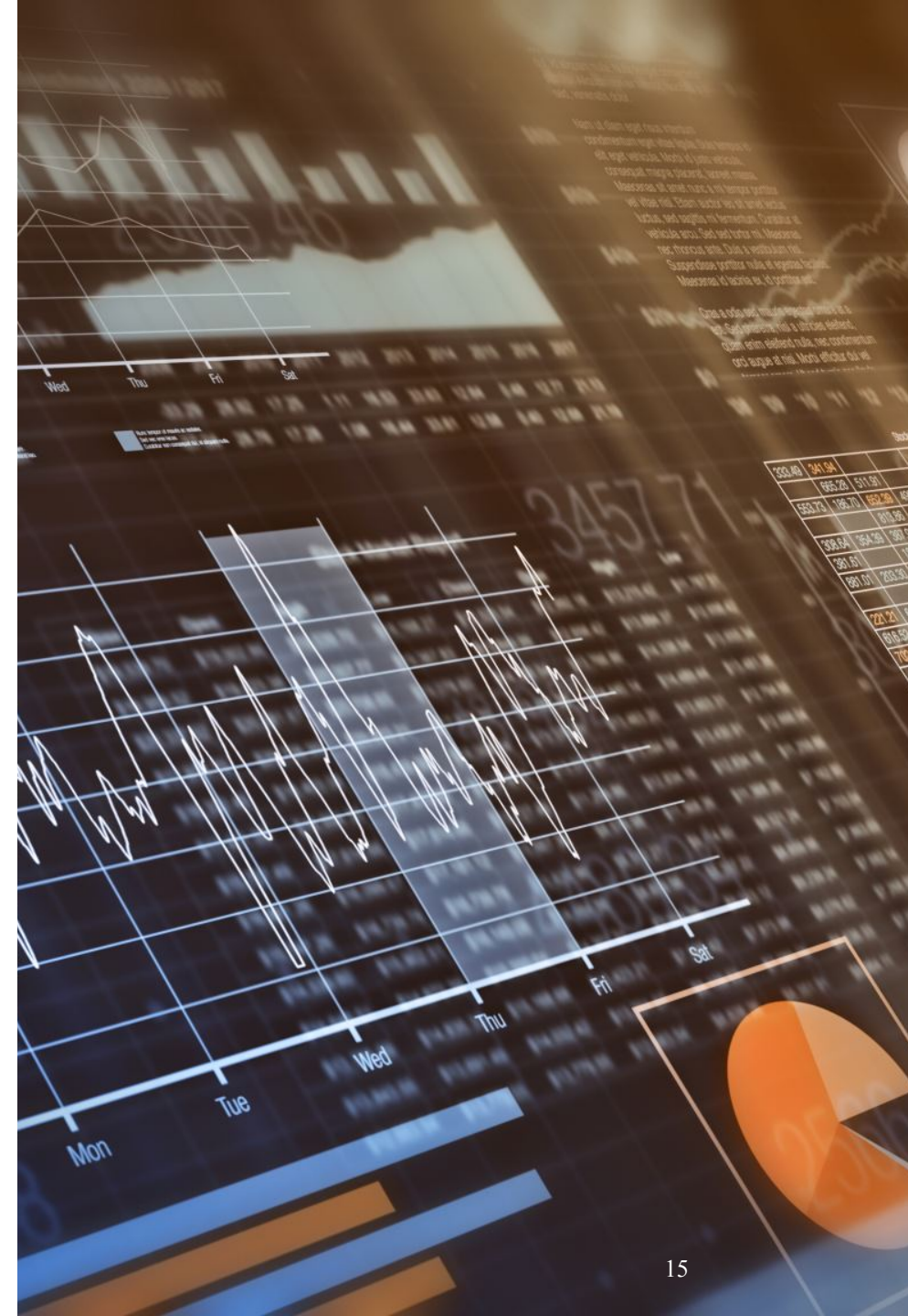


WHAT IS DATA SCIENCE?

Data Science is an interdisciplinary field that focuses on extracting knowledge from data sets which are typically huge in amount.

The field encompasses analysis, preparing data for analysis, and presenting findings to inform high-level decisions in an organization.

As such, it incorporates skills from computer science, mathematics, statistics, information visualization, graphic, and business.



3 TYPES OF DATA

Structured data –

Structured data is data whose elements are addressable for effective analysis.

It has been organized into a formatted repository that is typically a database.

It concerns all data which can be stored in database SQL in a table with rows and columns.

They have relational keys and can easily be mapped into pre-designed fields.

Today, those data are most processed in the development and simplest way to manage information. *Example:* Relational data.

3 TYPES OF DATA

Semi-Structured data –

Semi-structured data is information that does not reside in a relational database but that has some organizational properties that make it easier to analyze.

With some processes, you can store them in the relation database (it could be very hard for some kind of semi-structured data), but Semi-structured exist to ease space. *Example: XML data.*

3 TYPES OF DATA

Unstructured data –

Unstructured data is a data which is not organized in a predefined manner or does not have a predefined data model, thus it is not a good fit for a mainstream relational database.

So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications. *Example:* Word, PDF, Text, Media logs.

CHALLENGES AND CONSIDERATIONS FOR HANDLING DIFFERENT DATA TYPES.

Handling different data types in **data science** comes with unique challenges due to the diverse nature of datasets and the need for analytical rigor.

1. Data Cleaning and Preprocessing

Challenge: Raw data often contains missing, inconsistent, or irrelevant data types.

Consideration:

- Use techniques like **imputation** for missing values (mean, median, mode, or ML-based imputation).
- Standardize inconsistent formats (e.g., dates or currency units).
- Remove or encode outliers appropriately.

2. Mixed Data Types in Datasets

Challenge: Data science workflows frequently involve datasets with mixed data types (e.g., numerical, categorical, textual).

Consideration:

- Use appropriate encoding for categorical variables (e.g., one-hot, label encoding).
- Convert textual data to numerical representations (e.g., TF-IDF or word embeddings like Word2Vec).
- Maintain consistency in numerical data precision (e.g., float64 vs. float32).

3. Handling High Cardinality in Categorical Data

Challenge: High-cardinality features (e.g., unique identifiers or names) can inflate model complexity.

Consideration:

- Use dimensionality reduction techniques (e.g., embedding layers in neural networks).
- Group categories with low occurrences into an "Other" category.

4. Numerical Data Precision and Scaling

Challenge: Large ranges or differences in scales (e.g., age vs. income) can affect model performance.

Consideration:

- Apply normalization (e.g., Min-Max scaling) or standardization (z-scores).
- Be cautious of **loss of precision** in floating-point operations.

5. Text Data

Challenge: Text data is unstructured and requires significant preprocessing.

Consideration:

- Preprocess using techniques like stemming, lemmatization, and stopword removal.
- Handle multilingual data with language detection and tokenization libraries.
- Use advanced NLP methods (e.g., transformers) for contextual understanding.

6. Time Series Data

Challenge: Temporal dependencies and irregular intervals make time series unique.

Consideration:

- Convert timestamps to consistent formats and extract features like day, month, year, or lags.
- Resample irregular intervals to a uniform frequency.
- Use models designed for sequential data (e.g., ARIMA, LSTMs).

7. Image and Multimedia Data

Challenge: Handling large file sizes and extracting meaningful features from unstructured image or video data.

Consideration:

- Use pre-trained models (e.g., ResNet, EfficientNet) for feature extraction.
- Optimize storage using compressed formats without losing crucial information.

8. Data Type Mismatches

Challenge: Errors arise when data is incorrectly labeled or stored (e.g., numerical data stored as strings).

Consideration:

- Use automated tools (e.g., Pandas `infer_objects`) to identify mismatched types.
- Validate and convert data types explicitly (e.g., `astype()` in Python).

9. Missing Data

Challenge: Missing values in different data types (numerical vs. categorical) require tailored handling.

Consideration:

- For numerical data, use mean/median imputation or interpolation.
- For categorical data, use mode imputation or "Unknown" labels.

10. Large-Scale Data

Challenge: Datasets with millions of rows or high-dimensional data strain memory and processing power.

Consideration:

- Use distributed frameworks like Apache Spark or Dask for big data.
- Reduce dimensionality using PCA or feature selection techniques.

11. Bias in Data Types

Challenge: Bias in categorical data (e.g., imbalanced classes) can affect model performance.

Consideration:

- Apply oversampling (SMOTE) or undersampling to address class imbalance.
- Use balanced accuracy or F1-score instead of simple accuracy for evaluation.

12. Data Visualization

Challenge: Visualizing mixed data types can be non-trivial.

Consideration:

- Use specialized plots (e.g., scatter plots for numerical, bar plots for categorical).
- Combine data types into summary metrics for easier visualization.

ETHICAL CONSIDERATIONS IN DATA SCIENCE.

Ethical considerations in data science are critical to ensure responsible and fair use of data, algorithms, and outcomes.

1. Data Privacy

Concern: Collection and use of personal data can infringe on individuals' privacy.

Examples:

- Tracking users' online behavior without consent.
- Exposing sensitive information through data breaches.

Best Practices:

- Implement data anonymization techniques (e.g., pseudonymization, differential privacy).
- Use encryption for sensitive data storage and transfer.

2. Bias and Fairness

Concern: Models trained on biased data can perpetuate or exacerbate discrimination.

Examples:

- A hiring algorithm discriminates against women due to historical bias in the training data.
- Facial recognition systems misclassify certain ethnic groups at higher rates.

Best Practices:

- Conduct fairness audits to identify and mitigate bias.
- Use diverse and representative datasets for training.
- Regularly evaluate models for fairness using metrics like demographic parity or equalized odds.

3. Informed Consent

Concern: Individuals may not be fully aware of how their data is being collected and used.

Examples:

- Hidden clauses in terms of service agreements.
- Using location data from mobile apps without user consent.

Best Practices:

- Clearly communicate data collection purposes.
- Obtain explicit consent for data usage, especially for sensitive data.
- Provide opt-out options and transparency.

4. Accountability

Concern: Lack of accountability can lead to misuse or harm from data science applications.

Examples:

- No clear ownership of errors in a predictive healthcare model.
- Deploying automated decision systems without oversight.

Best Practices:

- Assign accountability to specific roles or teams.
- Maintain audit logs to track decisions made by models.
- Ensure human oversight for critical decisions.

5. Transparency and Explainability

Concern: Complex models like deep learning are often "black boxes," making decisions hard to interpret.

Examples:

- A credit scoring algorithm denies a loan without explaining why.
- Autonomous vehicles make critical decisions that cannot be understood or challenged.

Best Practices:

- Use explainable AI techniques (e.g., SHAP, LIME).
- Clearly document data sources, model assumptions, and limitations.
- Provide end-users with interpretable explanations of outcomes.

6. Data Security

Concern: Inadequate security measures can expose sensitive data to cyberattacks.

Examples:

- Ransomware attacks targeting healthcare datasets.
- Leaks of personally identifiable information (PII).

Best Practices:

- Employ robust cybersecurity measures like multi-factor authentication and firewalls.
- Conduct regular security audits and penetration testing.
- Limit access to sensitive data on a need-to-know basis.

7. Misuse of Data

Concern: Data science tools and techniques can be exploited for unethical purposes.

Examples:

- Deepfake generation for spreading misinformation.
- Targeted advertising that manipulates vulnerable populations.

Best Practices:

- Define ethical use policies for data and models.
- Monitor and restrict usage of tools that can cause harm.
- Collaborate with ethics boards to review applications.

8. Environmental Impact

Concern: Large-scale data processing and training of AI models have significant energy costs.

Examples:

- Training a large NLP model emitting CO₂ equivalent to multiple flights.

Best Practices:

- Optimize algorithms for computational efficiency.
- Use cloud services powered by renewable energy.
- Explore greener machine learning techniques.

9. Intellectual Property

Concern: Unauthorized use of proprietary datasets or algorithms can lead to legal and ethical issues.

Examples:

- Using copyrighted content for training AI without permission.

Best Practices:

- Respect copyright and licensing agreements.
- Properly cite data sources and algorithms.
- Develop proprietary datasets ethically and transparently.

10. Impact on Jobs and Society

Concern: Automation and data-driven systems may displace workers or reinforce societal inequalities.

Examples:

- AI replacing customer service roles without a transition plan for affected employees.
- Predictive policing algorithms targeting marginalized communities disproportionately.

Best Practices:

- Engage with stakeholders to assess societal impacts.
- Create transition plans for workers affected by automation.
- Use data science to identify and mitigate inequalities.

11. Dual-Use Dilemma

Concern: Technologies developed for positive applications can be repurposed for harm.

Examples:

- AI used for medical diagnostics being weaponized in military applications.

Best Practices:

- Evaluate potential risks during the development phase.
- Restrict access to dual-use technologies.

12. Long-Term Consequences

Concern: Decisions made today might have unforeseen negative consequences in the future.

Examples:

- Data retention policies leading to misuse decades later.

Best Practices:

- Regularly reassess models and data policies.
- Adopt a precautionary principle when deploying high-impact applications.

DATA SCIENCE IS THE FUTURE

One of the reasons for the acceleration of data science in recent years is the enormous volume of data (**e.g Big Data**) currently available and being generated.

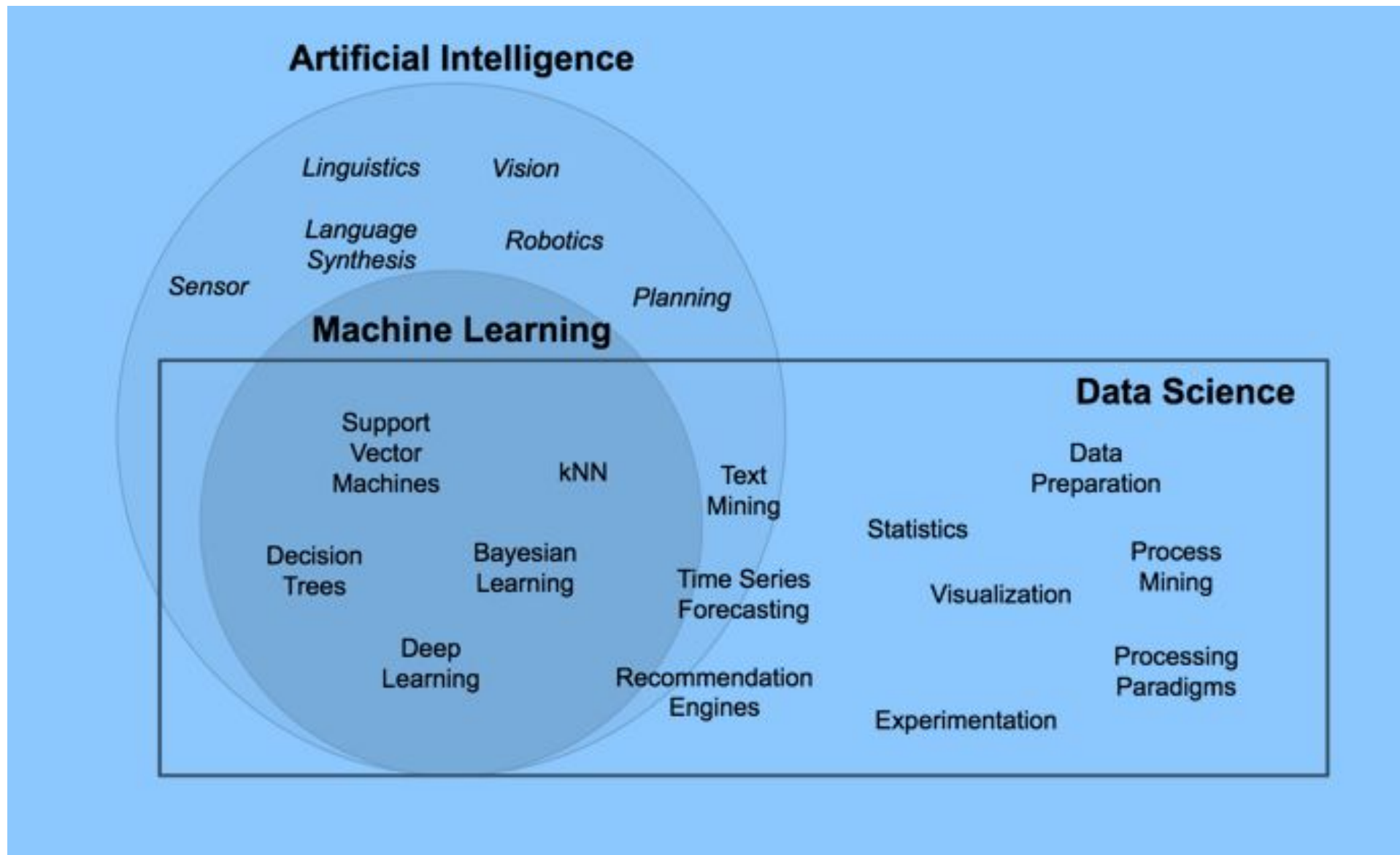
Not only are huge amounts of data being collected about many aspects of the world and our lives, but we concurrently have the rise of inexpensive computing. This has formed the perfect storm in which we have rich data and the tools to analyze it.

Advancing computer memory capacities, more enhanced software, more competent processors, and now, more numerous data scientists with the skills to put this to use and solve questions using the data!

And that's the big reason why do we need data science in the future.

DATA SCIENCE VS BIG DATA

Data Science	Big data
It is about the collection, processing, analyzing, and utilizing of data in various operations. It is more conceptual.	Big Data is a technique to collect, maintain and process huge information.
It is a field of study just like Computer Science, Applied Statistics, or Applied Mathematics.	It is a technique for tracking and discovering trends in complex data sets.
The goal is to build data-dominant products for a venture.	The goal is to make data more vital and usable i.e. by extracting only important information from the huge data within existing traditional aspects.
Tools mainly used in Data Science include SAS, R, Python, etc	Tools mostly used in Big Data include Hadoop, Spark, Flink, etc.



DATAFICATION

Datafication is a process of “taking all aspects of life and turning them into data.”

Examples: Twitter datafies stray thoughts. LinkedIn datafies professional networks.”

Example: “Wal-Mart is able to take data from your past buying patterns, their internal stock information, your mobile phone location data, social media as well as external weather information and analyse all of this in seconds so it can send you a voucher for a BBQ cleaner to your phone— but only if you own a barbeque, the weather is nice and you currently are within a 3 miles radius of a Wal-Mart store that has the BBQ cleaner in stock.”

We are being datafied, or rather our actions are, and when we “like” someone or something online, we are intending to be datafied, or at least we should expect to be. But when we merely browse the Web, we are unintentionally, or at least passively, being datafied through cookies that we might or might not be aware of. And when we walk around in a store, or even on the street, we are being datafied in a completely unintentional way, via sensors, cameras, or Google glasses.

THE DATA SCIENCE LIFE CYCLE

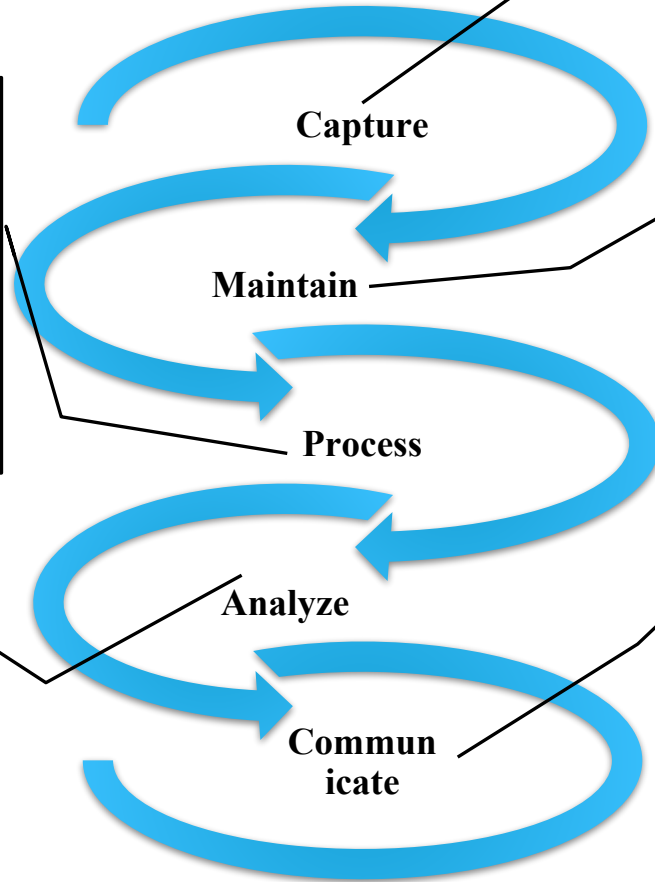
Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data.

Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.

Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis.

Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.

Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports.



A DATA SCIENCE PROFILE

On a daily basis, a data scientist may do the following tasks:

1. Discover patterns and trends in datasets to get insights.
2. Create forecasting algorithms and data models.
3. Improve the quality of data or product offerings by utilising machine learning techniques.
4. Distribute suggestions to other teams and top management.
5. In data analysis, use data tools such as R, SAS, Python, or SQL.
6. Top the field of data science innovations.

DATA SCIENTIST MUST-HAVE SKILLS

MATH & STATISTICS

- Machine Learning
- Statistical Modeling
- Exploratory Analysis
- Clustering
- Regression Analysis

DOMAIN KNOWLEDGE & SOFT SKILLS

- Inclination towards business operations
- Keen on working with data
- Problem solver
- Strategic, proactive, and cooperative
- Interested in hacking



PROGRAMMING & DATABASE

- Computer Science Fundamentals
- Database Management System
- Data Visualization
- Python
- Big Data

COMMUNICATION & VISUALIZATION

- Storytelling skills
- Convert data-based insights into decisions
- Collaborative with Sr. Management
- Knowledge of tools like Tableau
- Visual art design

ARE DATA SCIENCE JOBS IN DEMAND IN INDIA?

The Indian analytics industry is predicted to escalate to USD 98 billion in 2025 and nearly USD 119 billion in 2026. Currently, **the demand for data scientists is at an all-time high in India**. Analysts have predicted around 11 million job openings in data science by 2026 in India alone.

Average Data Scientist, IT Salary in India

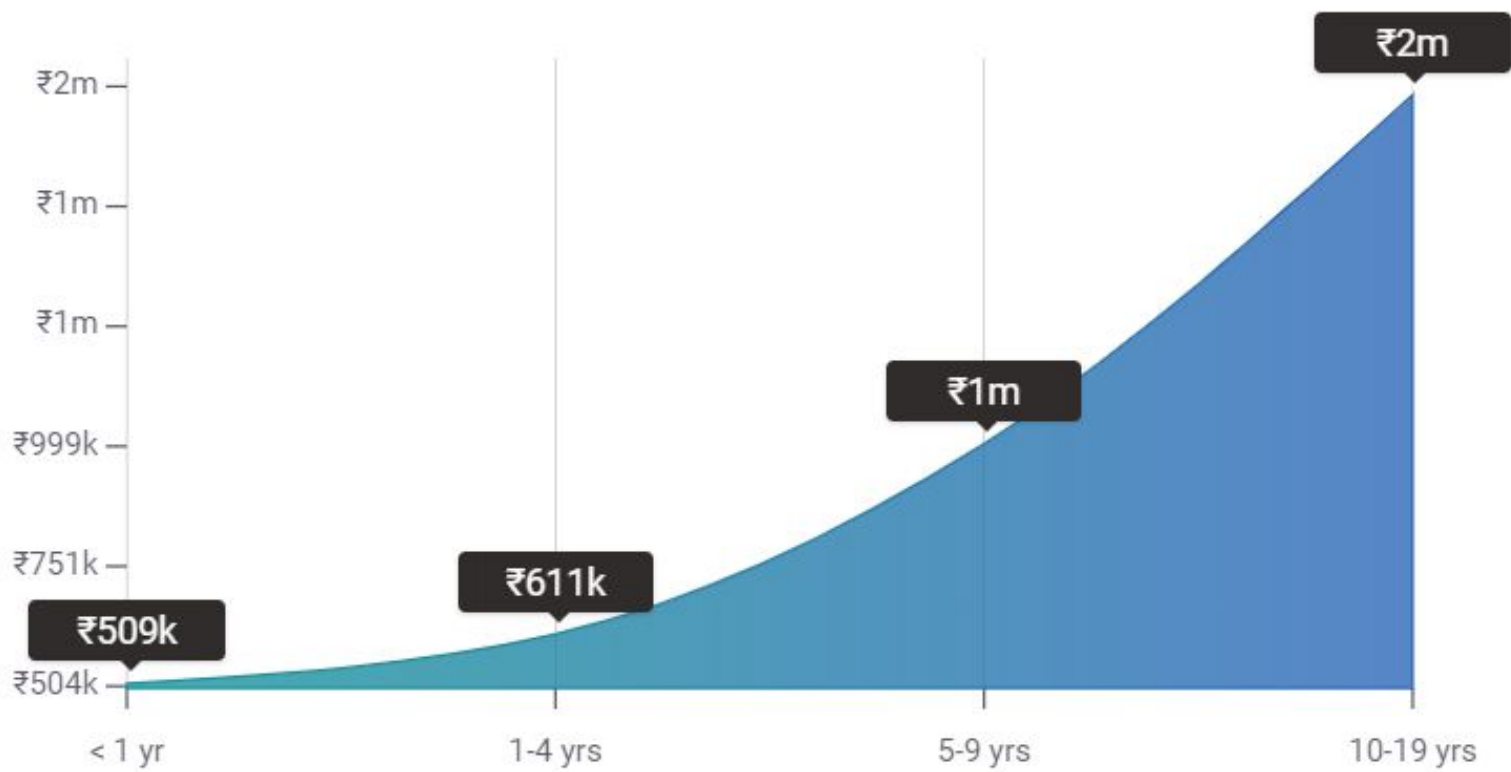
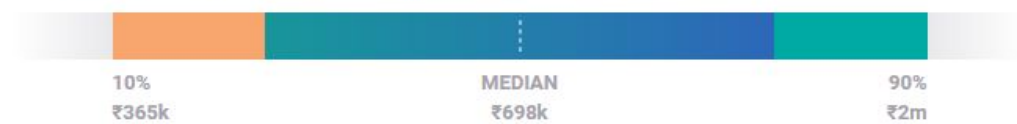
₹698,413

Avg. Salary [Show Hourly Rate](#)

₹60,678
BONUS

₹35,000
PROFIT SHARING

The average salary for a Data Scientist, IT in India is ₹698,413.



Source: https://www.payscale.com/research/IN/Job=Data_Scientist%2C_IT/Salary

TOP DATA SCIENTIST RECRUITERS



IMPACT OF APPLYING DATA SCIENCE IN BUSINESS SCENARIO

As a business, your goals are sure to provide better products to the consumers so that they come back.

Data science helps in developing user-centric products using analysis of customer reviews, analyzing current market trends, comparing two products, and finalizing the best products with the ability to attract customers and hold on to them for a longer time.

IMPACT OF APPLYING DATA SCIENCE IN BUSINESS SCENARIO

1. Reduces Inefficiencies

Inefficiencies often cost businesses up to 30% of their revenue. Data scientists track a range of company-wide metrics – factory production times, delivery expenditure, employee productivity, and more – and pinpoint areas for improvement.

By limiting wasted resources, it's possible to lower overall costs and boost return-on-investment.

2. Predicts Trends and Customer Behavior

On a practical level, data-based predictions have an array of applications. **It's possible, for example, to determine peak customer shopping times and adjust staff levels accordingly,** or to identify early buyer trends and implement appropriate promotional campaigns.

IMPACT OF APPLYING DATA SCIENCE IN BUSINESS SCENARIO

3. Enables Competitor Research

Data scientists are responsible for understanding and gleaning insights from data about competitors. Effective competitor research helps businesses make competitive pricing decisions, reach new markets, and stay up to date with changes in consumer behavior.

4. Allows Testing of Business Initiatives

Consistent, long-term testing enables companies to drive incremental revenue gains. **Data scientists are responsible for conducting extensive tests to guarantee successful marketing campaigns, product launches, employee satisfaction, website optimization, and more.**

IMPACT OF APPLYING DATA SCIENCE IN BUSINESS SCENARIO

5. Develops Market Understanding

Data science enables businesses to consistently reshape their products and services to fit with a shifting marketplace.

Data about customers is available from a variety of sources, and mining information from third-party platforms, like social media, search engines, and purchased datasets, presents a unique challenge.

6. Informs Hiring Decisions

One of the big problems faced by businesses when searching for new employees is the disconnect between prospects that look good on paper and perform well in practice. **Data science seeks to bridge this gap by using evidence to improve hiring practices.**

By combining and analyzing a variety of data-points about candidates, it's possible to move towards an ideal 'company-employee fit'.



How Amazon uses Data Science

U
Uses
recommendation
based system(RBS)


T
Tracking the user
to understand
the mindset

U
Understands
the technicalities
(habits)

F
Faster process
of shipping

- Predictive analytics
- Drones

U T U F



DATA SCIENCE USE CASES



<https://data-flair.training/blogs/data-science-use-cases/>

WHAT DATA SCIENTISTS AT UBER ARE DOING WITH YOUR DATA?

Going from one place to another has never been simpler, for all you have to do is open the Uber app, type your destination, and click on “Find Ride.” That’s all. But have you ever wondered what goes on behind the scenes or the complex processes behind that simple click on the app that enables you to travel hassle-free?

While you wait for your cab and forget all about it once you reach your destination, a team of data scientists at Uber works tirelessly to solve existing challenges, provide a better user experience, improve its geo-mapping, and research on driver-less cars. These curious souls use machine learning to not only improve your experience but also make their technologies learn to serve you better in the future.

CASE STUDY-UBER

Intelligent Ride Management

Every ride you book gives the AI team at Uber an enormous amount of information about you — right from your preferred pick-up point to your most frequent destinations. Uber records your behavior, preferences, interests, and even your phone's battery level. This helps them determine demand, the resources to allocate, and set fares to maximize profit. While sometimes the team would use data month to month, a larger set is required for a seasonal understanding of certain tasks that may go back years. A data expert's idea is to collect data smartly and clean it to find relevance in it.

the AI team at Uber extracts in terms of information. They collect even more data from the drivers, irrespective of whether they are carrying passengers or not. Drivers' speed and acceleration, their location, data on whether they work for a competitor (like Ola) is all possible to retrieve. The AI team then draws inferences from understanding traffic patterns, ETA, journey times, and surge pricing.

CASE STUDY-UBER

Pricing

Longer or shorter journey times give them the advantage of real-time pricing (remember the surge price that you see?). Data Scientists use predictive modelling in real time to estimate demand and execute informed pricing. Uber has even applied for a patent on this method of calculating “surge pricing”. Hotel chains and airlines also use dynamic pricing, in the same way, to meet demand and adjust prices, especially during public holidays, weekends, and popular events.

The supply and demand analysis algorithms developed by the AI team capture and monitor traffic conditions and journey duration in real time. As the demand for cabs changes, the pricing is adjusted in real-time. As the traffic conditions change, so do the journey ETAs. This is convenient for drivers as they can decide to proactively hit areas slated for higher demand and stay low when the demand drops.

CASE STUDY-UBER

Mapping

Data scientists build models using several map services like Google Maps. But a lot of information on those maps is either irrelevant or inaccurate (at least not accurate to the point they would like it to be). For example, if you visit the Ambience Mall in Gurgaon, a map would show you and the Uber driver that you have arrived at your destination while you could be almost 500 meters away from the gate you want to enter from. Data from a million trips enabled the team to offer pick/drop services to specific gates that may seem so basic to an end-user. More information that they retrieve is translating into enhanced mapping that tackles existing challenges and helps move to the next level.

CASE STUDY-UBER

Driver-Less Cars

The Uber AI Labs at Pittsburgh is running experiments with autonomous cars, and it won't be long before driverless cars become humdrum. By improving their mapping system with deep learning (meaning machines learn better with more exposure similar to how humans learn), scientists are taking us a tad bit faster into the future. The best part is that the machines are not only learning just once through algorithms but throughout the process of testing and implementation. As more trips are taken, as more models are built using AI, the machines are getting smarter about handling unprecedented scenarios.