Lecture-4

Course: Applied Data Science

Introduction and Motivation to Data Science

By

Dr. Sibarama Panigrahi

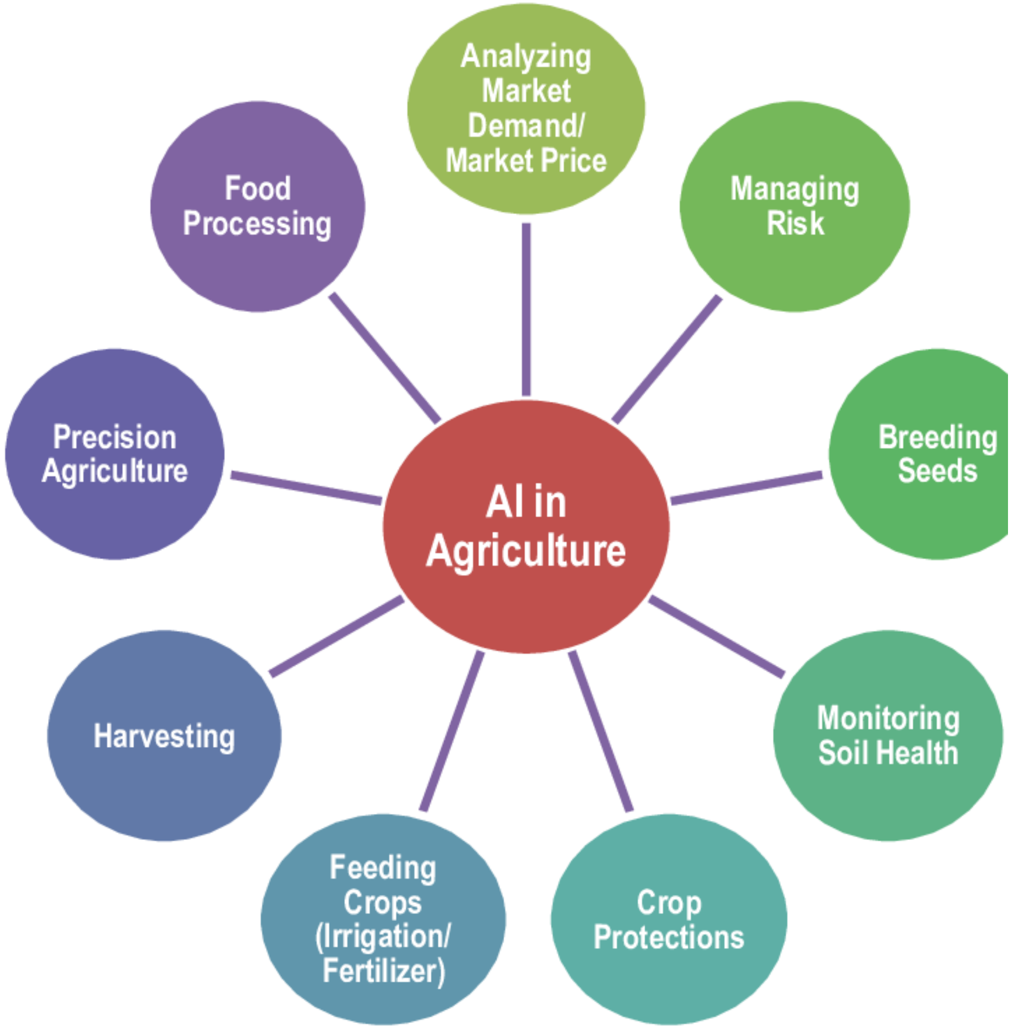
Assistant Professor, Department of Computer Sc. & Engineering
National Institute of Technology, Rourkela, Odisha, 769008, India
Mobile No.: +91-7377302566
Email: panigrahis[at]nitrkl[dot]ac[dot]in
panigrahi[dot]sibarama[at]gmail[dot]com

Outlines…

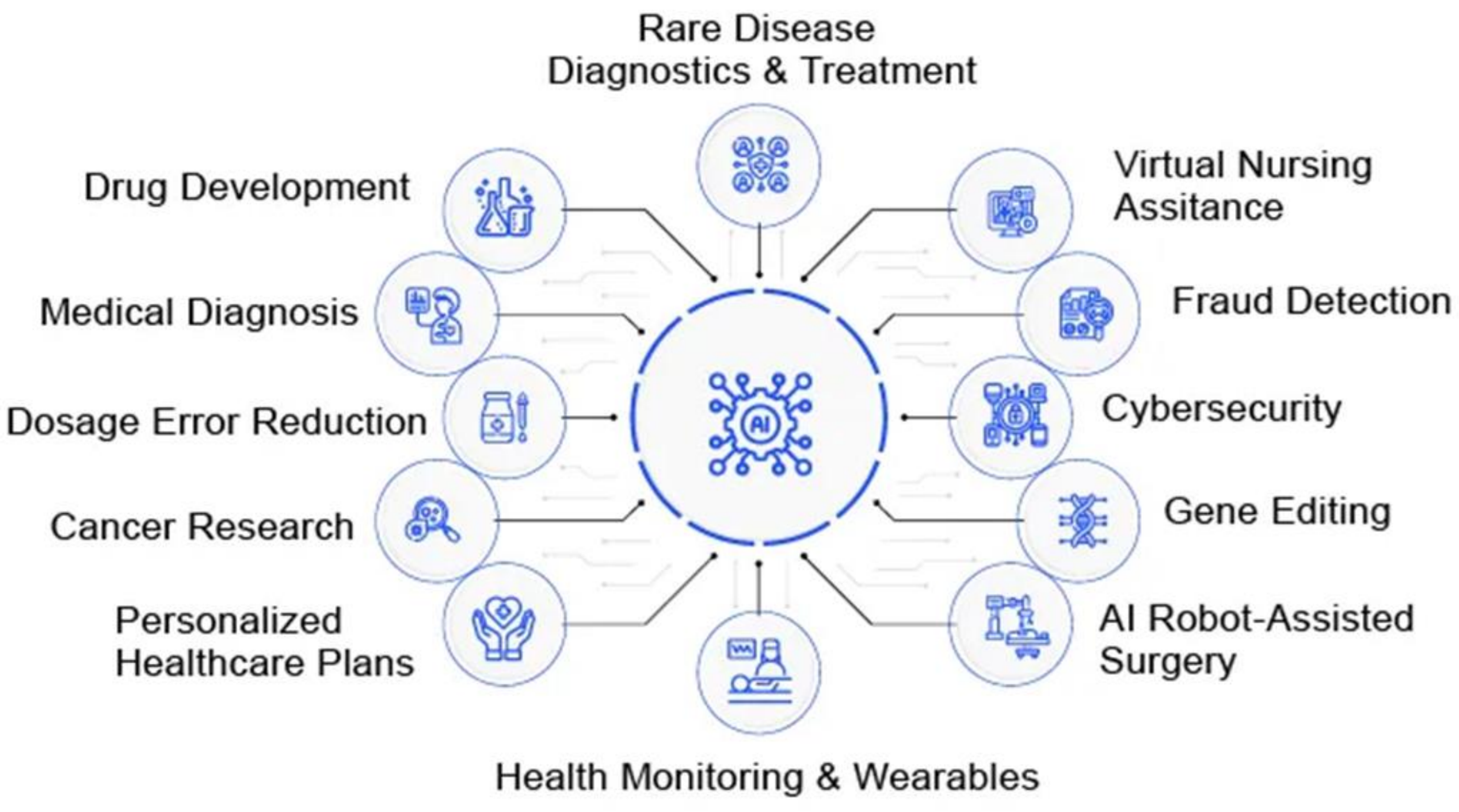
- Motivation to Study Data Science
- Data Science
- Data & Big Data
- Facets of Data

Motivation to Study Data Science [Software Giants]

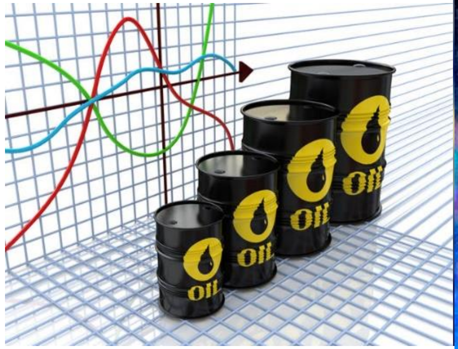
Motivation to Study Data Science [In Agriculture]



Motivation to Study Data Science [In Health Sector]


Rare Disease Diagnostics & Treatment
Drug Development
Virtual Nursing Assitance
Medical Diagnosis
Fraud Detection
Dosage Error Reduction
AI
Cybersecurity
Cancer Research
Gene Editing
Personalized Healthcare Plans
AI Robot-Assisted Surgery
Health Monitoring & Wearables

Motivation to Study Data Science [In Forecasting]



Crude Oil

Stock Price

Retail Industry

Internet Traffic

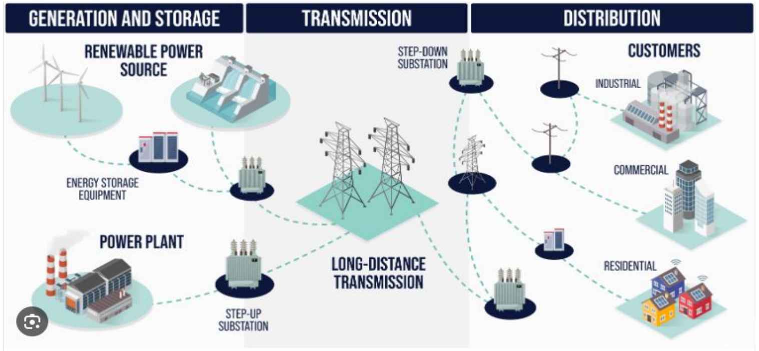
Electricity Price

Call Volume

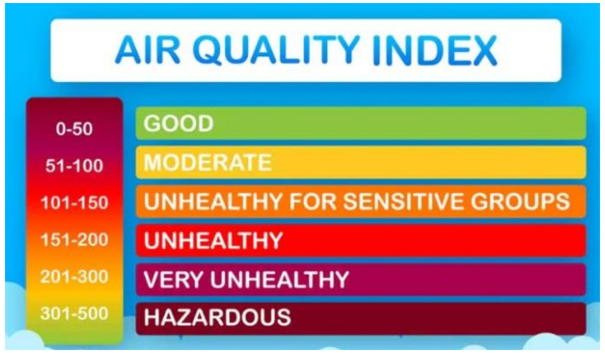
Flood

Earthquake

Motivation to Study Data Science [In Forecasting]



Electricity Load Forecasting



AIR QUALITY INDEX

0-50	GOOD
51-100	MODERATE
101-150	UNHEALTHY FOR SENSITIVE GROUPS
151-200	UNHEALTHY
201-300	VERY UNHEALTHY
301-500	HAZARDOUS

Air Quality Index Forecasting

Rainfall Forecasting

Streamflow Forecasting

Agricultural Product Price Forecasting

Seed Demand Forecasting

Wind Speed Forecasting

Temperature Forecasting

:

:

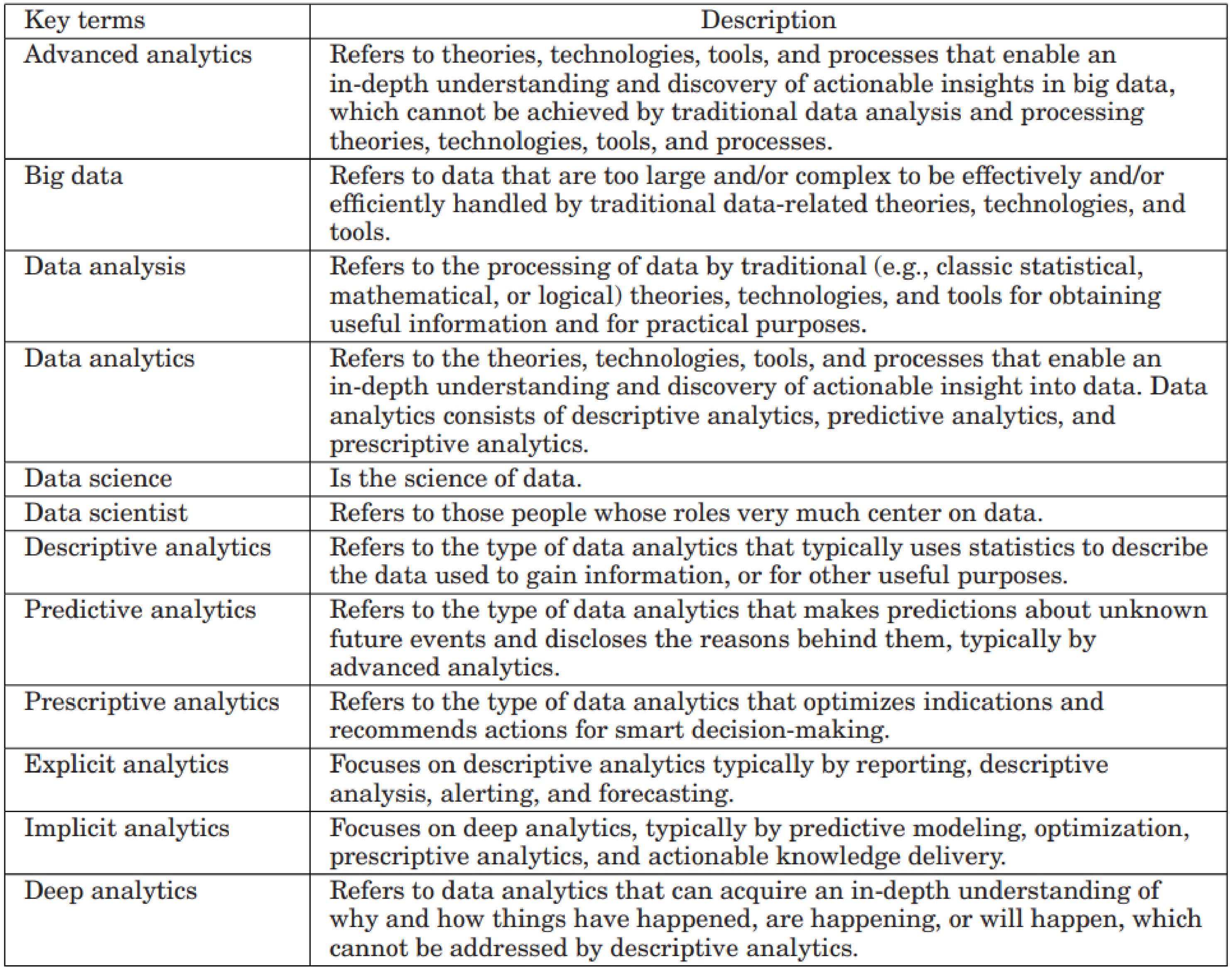
:

What is Data Science?

- **Definition:**
  - "*Data science is the science of data*" or "Data science is the study of data." [1]
  - **Data science** encompasses *a set of principles, problem definitions, algorithms, and processes* for extracting nonobvious and useful patterns from *large data sets* [2].
  - **Data science** is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology [3].

1. ACEMS. 2014. The Australian Research Council (ARC) Centre of Excellence for Mathematical and Statistical Frontiers. Retrieved from acems.org.au/.
2. John D. Kelleher, Brendan Tierney. 2018. WHAT IS DATA SCIENCE?. MIT Press, 1-38.
3. Ritu Agarwal and Vasant Dhar. 2014. Editorial-big data, data science, and analytics: The opportunity and challenge for IS research. Information Systems Research 25, 3 (2014), 443–448.

Some Key Terms in Data Science



Key terms	Description
Advanced analytics	Refers to theories, technologies, tools, and processes that enable an in-depth understanding and discovery of actionable insights in big data, which cannot be achieved by traditional data analysis and processing theories, technologies, tools, and processes.
Big data	Refers to data that are too large and/or complex to be effectively and/or efficiently handled by traditional data-related theories, technologies, and tools.
Data analysis	Refers to the processing of data by traditional (e.g., classic statistical, mathematical, or logical) theories, technologies, and tools for obtaining useful information and for practical purposes.
Data analytics	Refers to the theories, technologies, tools, and processes that enable an in-depth understanding and discovery of actionable insight into data. Data analytics consists of descriptive analytics, predictive analytics, and prescriptive analytics.
Data science	Is the science of data.
Data scientist	Refers to those people whose roles very much center on data.
Descriptive analytics	Refers to the type of data analytics that typically uses statistics to describe the data used to gain information, or for other useful purposes.
Predictive analytics	Refers to the type of data analytics that makes predictions about unknown future events and discloses the reasons behind them, typically by advanced analytics.
Prescriptive analytics	Refers to the type of data analytics that optimizes indications and recommends actions for smart decision-making.
Explicit analytics	Focuses on descriptive analytics typically by reporting, descriptive analysis, alerting, and forecasting.
Implicit analytics	Focuses on deep analytics, typically by predictive modeling, optimization, prescriptive analytics, and actionable knowledge delivery.
Deep analytics	Refers to data analytics that can acquire an in-depth understanding of why and how things have happened, are happening, or will happen, which cannot be addressed by descriptive analytics.

Source:
Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1-42.

Data Science Journey

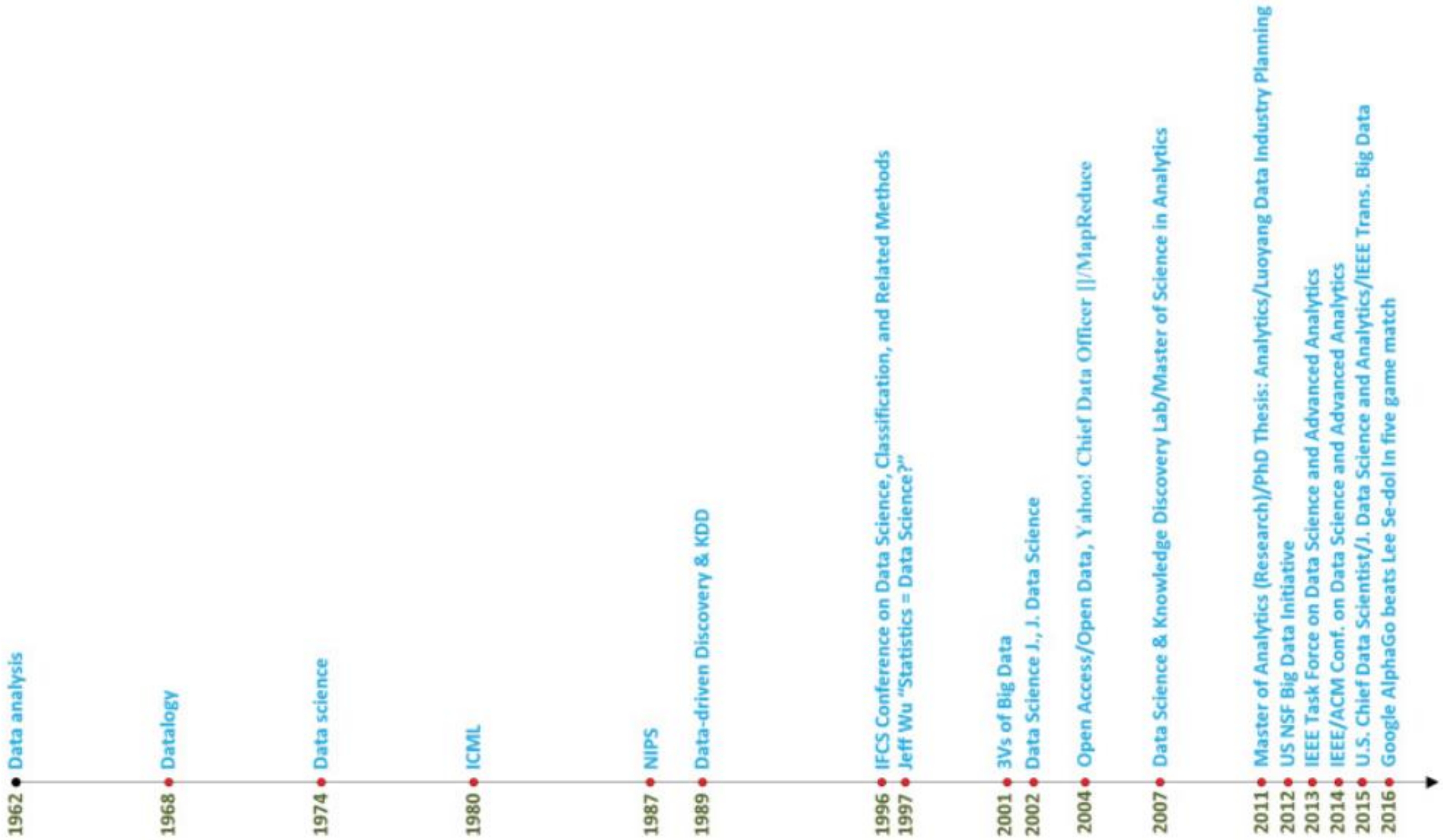


Fig. 1. Data science journey (with respect to typical events).

Source:
Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1-42.

Data Science Journey

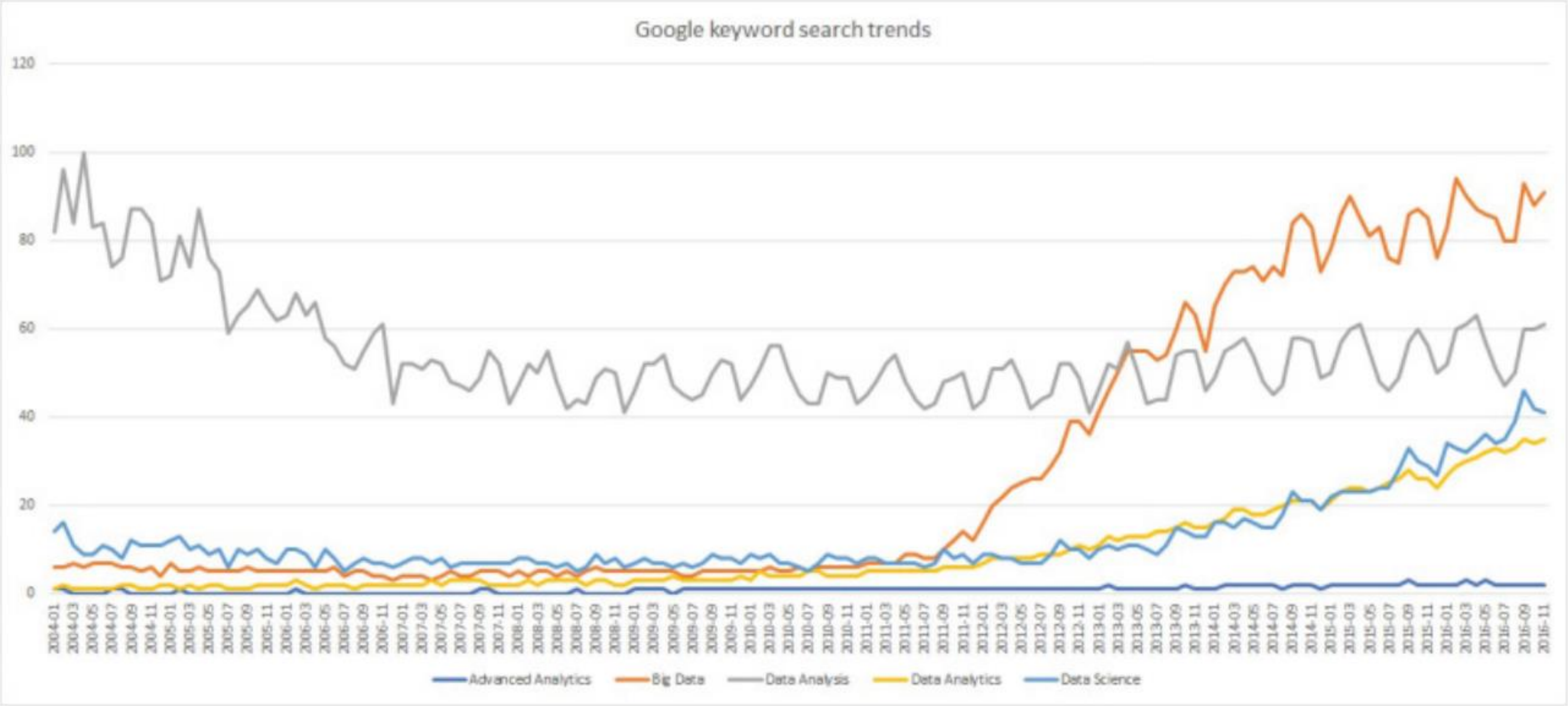


Fig. 2. Online search interest trends on data science-related keywords by Google.
Note: The data was collected on 15 November 2016.

Source: Cao, L. (2017). Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 1-42.

Data & Big Data

- **Data:** Known facts that can be recorded and that have implicit meaning.

- **Big Data:** Big Data refers to extremely large and complex data sets that are difficult to process, store, and analyze using traditional data processing techniques.
  - The concept of Big Data is characterized by the "Three Vs":
    - **Volume:** The sheer amount of data generated and collected, often measured in terabytes, petabytes, or even exabytes. This includes data from various sources such as social media, sensors, devices, transactions, and more.
    - **Velocity:** The speed at which data is generated, processed, and analyzed. Big Data often involves real-time or near-real-time data streams, requiring quick processing to gain insights and make decisions.
    - **Variety:** The diversity of data types and sources. Big Data encompasses structured data (like databases), semi-structured data (like XML or JSON), and unstructured data (like text, images, videos, and social media posts).
  - Sometimes, two additional "Vs" are also considered:
    - **Veracity:** The quality and accuracy of the data. Big Data often involves large amounts of noisy, inconsistent, or incomplete data, making it challenging to ensure its reliability.
    - **Value:** The potential insights and business value that can be derived from analyzing Big Data. The goal of Big Data analytics is to extract meaningful patterns, trends, and correlations that can inform decision-making.

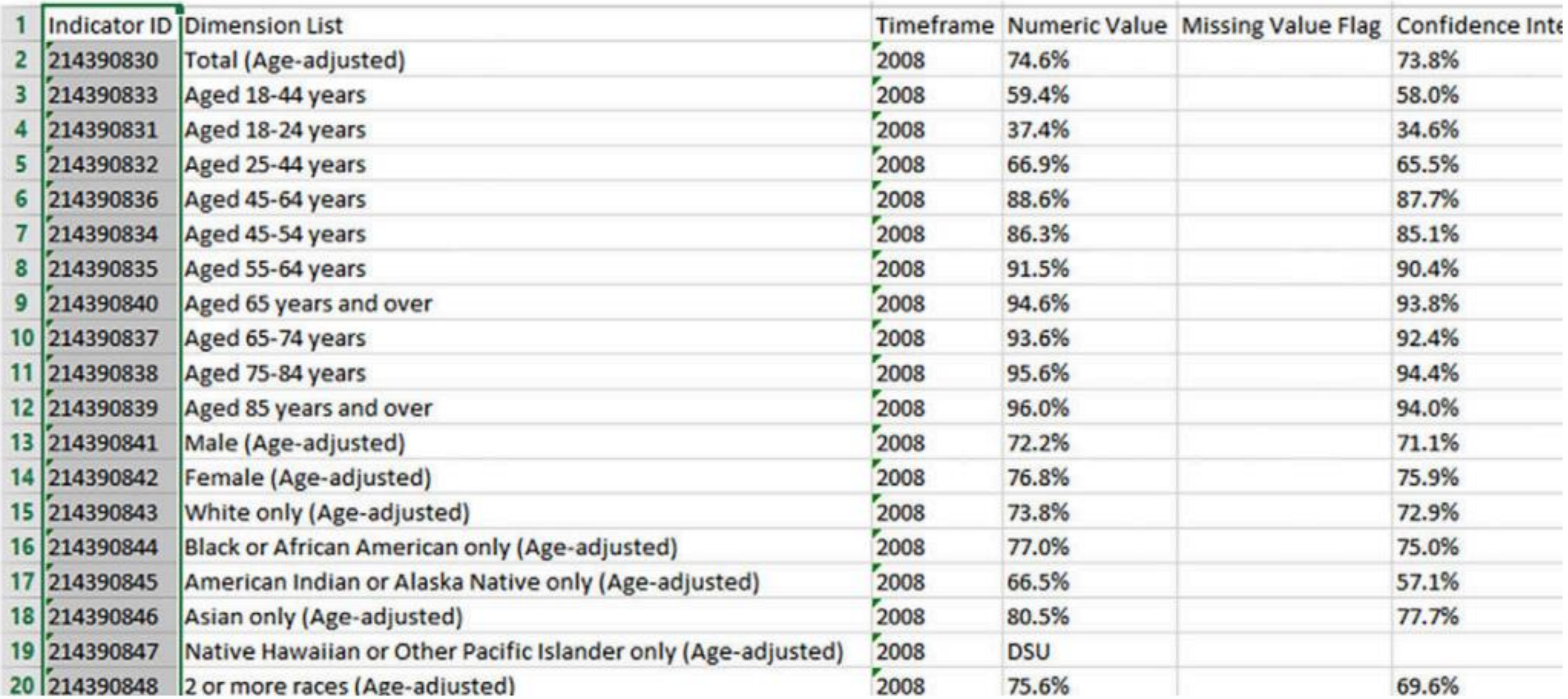
Facets of Data

- Structured
- Unstructured
- Semi-Structured
- Natural language
- Machine-generated
- Graph-based
- Audio, video, and images
- Streaming
- Ordered

Source:
Davy Cielin, Arno Meysman, Mohamed Ali, Introducing Data Science, Manning
Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, Pearson.

Facets of Data

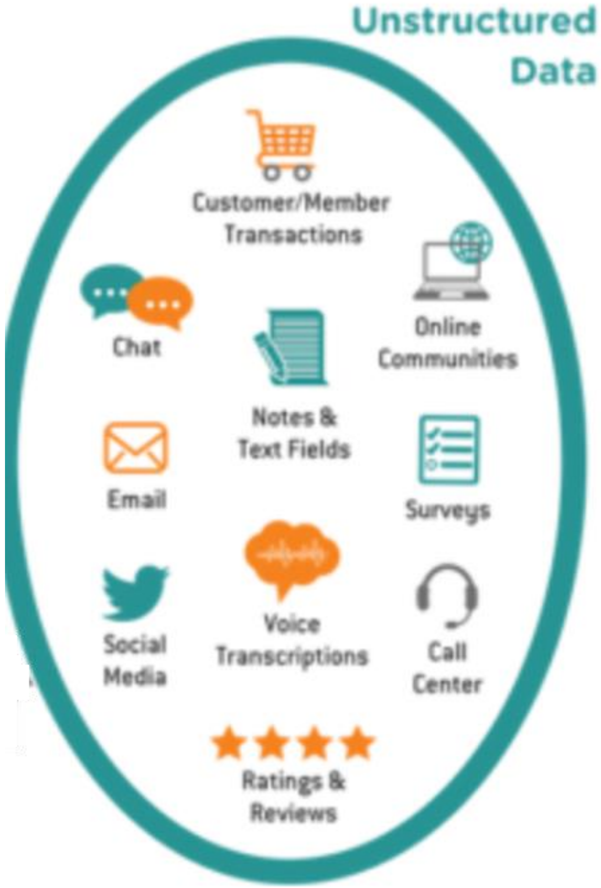
- **Structured Data:** Organized data that is easily searchable in databases, such as tables with rows and columns (e.g., SQL databases).



1	Indicator ID	Dimension List	Timeframe	Numeric Value	Missing Value Flag	Confidence Inte
2	214390830	Total (Age-adjusted)	2008	74.6%		73.8%
3	214390833	Aged 18-44 years	2008	59.4%		58.0%
4	214390831	Aged 18-24 years	2008	37.4%		34.6%
5	214390832	Aged 25-44 years	2008	66.9%		65.5%
6	214390836	Aged 45-64 years	2008	88.6%		87.7%
7	214390834	Aged 45-54 years	2008	86.3%		85.1%
8	214390835	Aged 55-64 years	2008	91.5%		90.4%
9	214390840	Aged 65 years and over	2008	94.6%		93.8%
10	214390837	Aged 65-74 years	2008	93.6%		92.4%
11	214390838	Aged 75-84 years	2008	95.6%		94.4%
12	214390839	Aged 85 years and over	2008	96.0%		94.0%
13	214390841	Male (Age-adjusted)	2008	72.2%		71.1%
14	214390842	Female (Age-adjusted)	2008	76.8%		75.9%
15	214390843	White only (Age-adjusted)	2008	73.8%		72.9%
16	214390844	Black or African American only (Age-adjusted)	2008	77.0%		75.0%
17	214390845	American Indian or Alaska Native only (Age-adjusted)	2008	66.5%		57.1%
18	214390846	Asian only (Age-adjusted)	2008	80.5%		77.7%
19	214390847	Native Hawaiian or Other Pacific Islander only (Age-adjusted)	2008	DSU		
20	214390848	2 or more races (Age-adjusted)	2008	75.6%		69.6%

Facets of Data

- **Unstructured Data:** Data that does not have a predefined structure, such as text, images, videos, and social media posts.

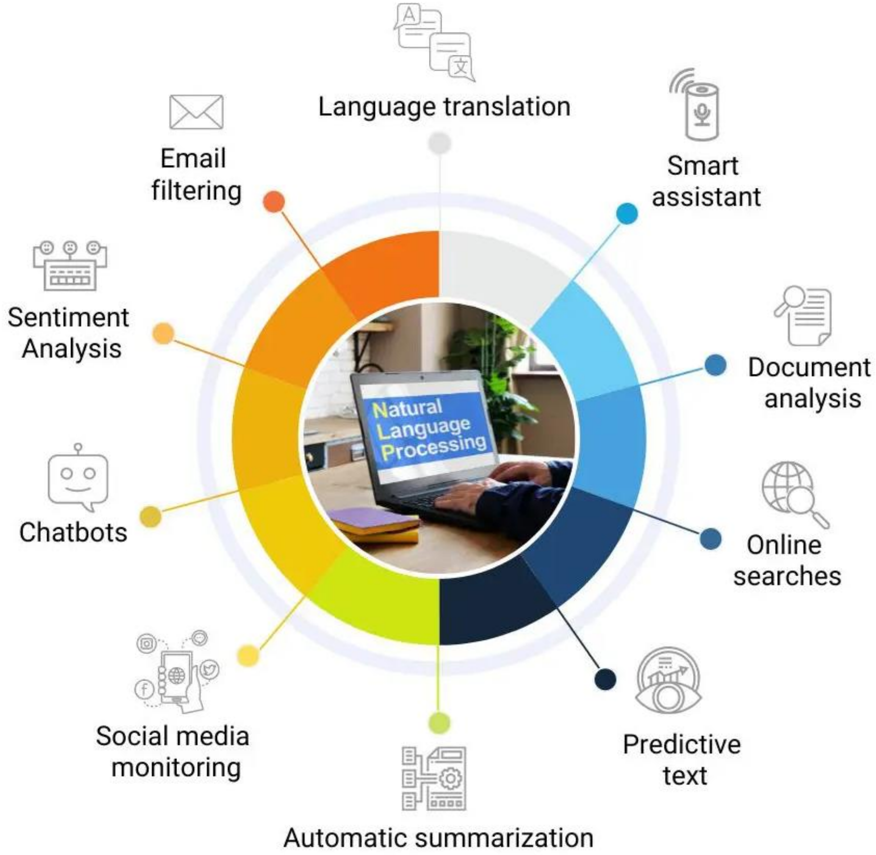


Facets of Data

- **Semi-Structured:** Data that has some organizational properties but doesn't fit neatly into structured formats, like XML or JSON files.

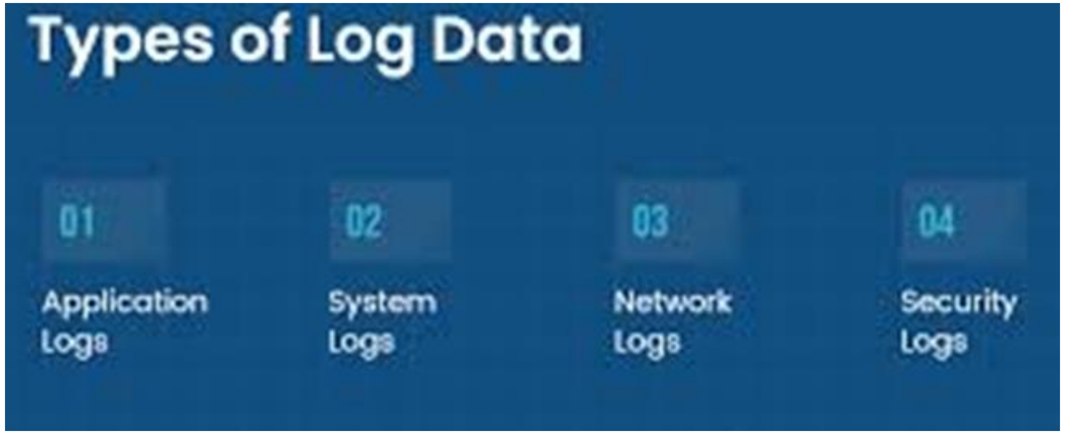
Facets of Data

- **Natural language Data:** Natural language is a special type of unstructured data; it's challenging to process because it requires knowledge of specific data science techniques and linguistics.



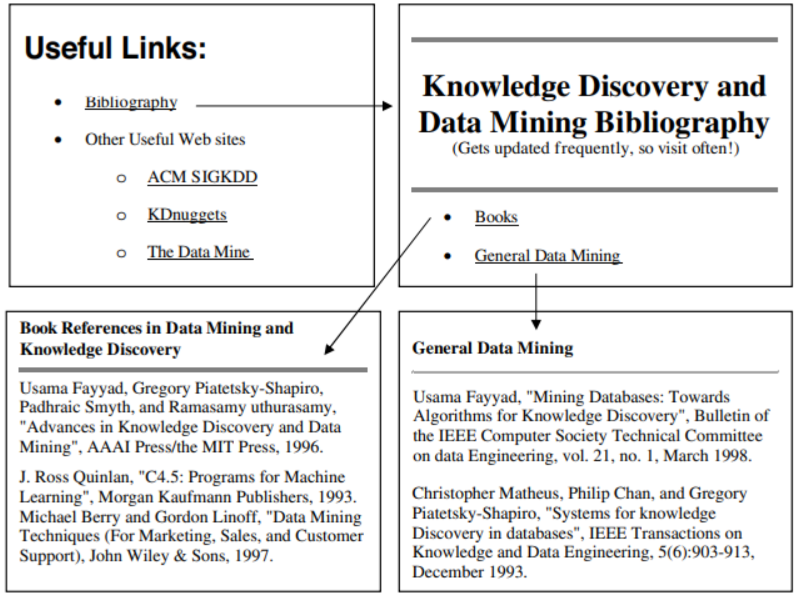
Facets of Data

- **Machine-generated Data:** Machine-generated data is information that’s automatically created by a computer, process, application, or other machine without human intervention.



Facets of Data

- **Graph-based Data:** Graph or network data is, in short, data that focuses on the relationship or adjacency of objects.



Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

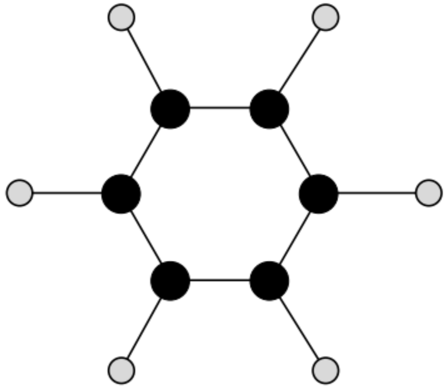
J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

General Data Mining

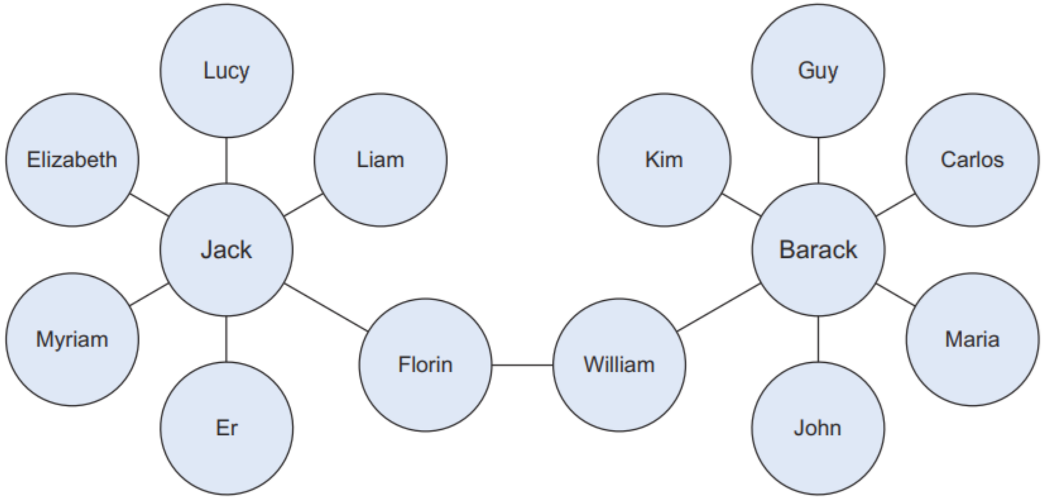
Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

(a) Linked Web pages.



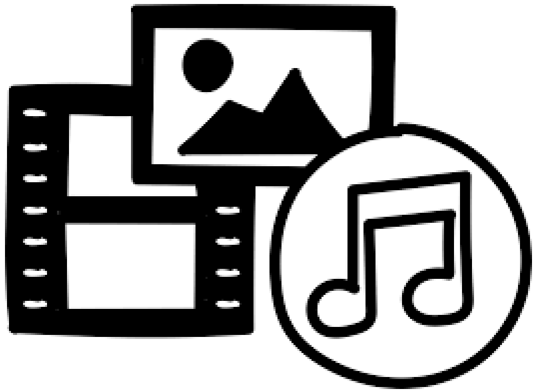
(b) Benzene molecule.



Friends in a social network are an example of graph-based data.

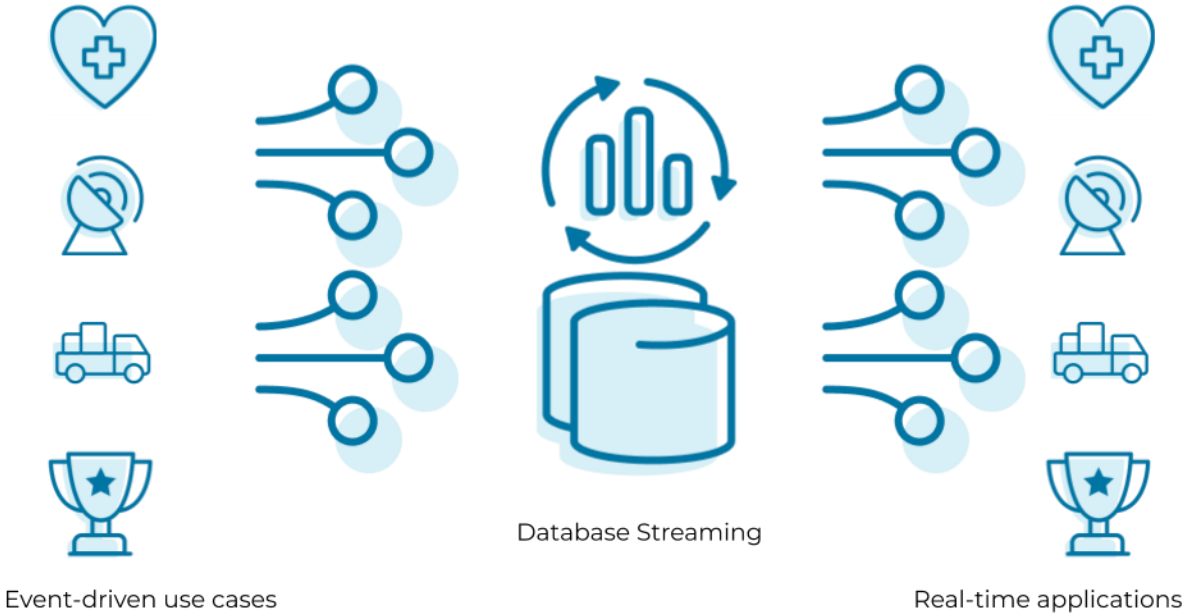
Facets of Data

- **Audio, video, and images Data:** Audio, image, and video are data types are unstructured data that pose specific challenges to a data scientist. Tasks that are trivial for humans, such as recognizing objects in pictures, turn out to be challenging for computers.



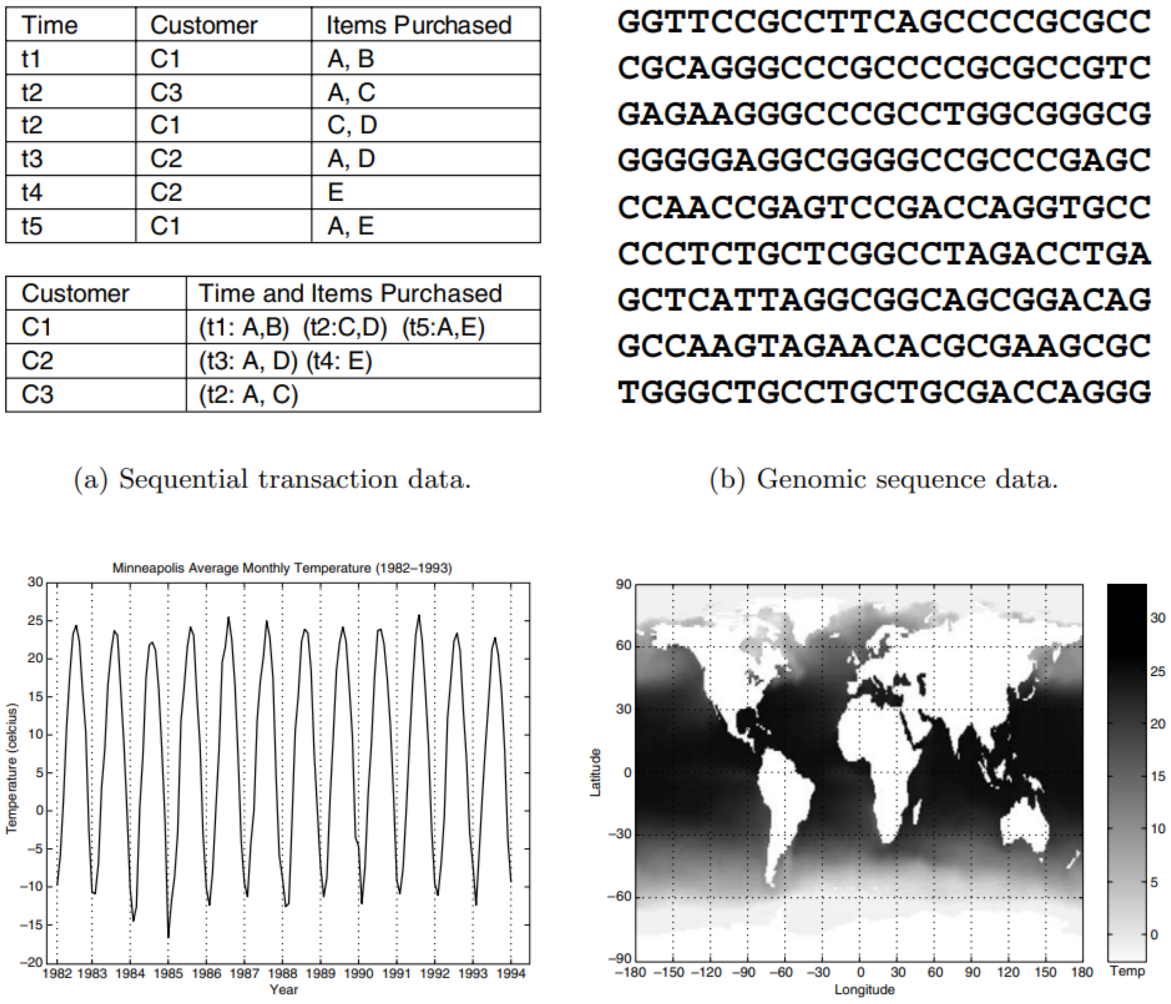
Facets of Data

- **Streaming Data:** While streaming data can take almost any of the previous forms, it has an extra property. The data flows into the system when an event happens instead of being loaded into a data store in a batch. Although this isn’t really a different type of data, we treat it here as such because you need to adapt your process to deal with this type of information.



Facets of Data

- **Ordered Data:**
  - Sequential Data
  - Genomic Sequence Data
  - Time series Data
  - Spatial Data



Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

(a) Sequential transaction data.

(b) Genomic sequence data.

(c) Temperature time series.

(d) Spatial temperature data.

For Your Valuable Time.