DATA ANALYTICS

UE22CS342AA2

UNIT-1

Lecture 1 : Introduction to DA , Data Sources and Representations

Gowri Srinivasa

Department of Computer Science and Engineering

Data Analytics

Unit 1

Lecture 1 : Introduction to Data Analytics , Data Sources and Representations

Slides excerpted from: U. Dinesh Kumar, “Business Analytics”, Wiley, 2nd Edition 2022

Gowri Srinivasa
Department of Computer Science and Engineering

Slides collated by:
Nishanth M S, CSE 2023, PES University
nishanthmsathish.23@gmail.com
Harshitha Srikanth, CSE 2024, PES University
harshithasrikanth13@gmail.com
Karthik Namboori, VII Sem, PESU, Department of CSE
namkarthik2003@gmail.com

With grateful thanks for contribution of slides to:
Dr. Mamatha H R, Professor at the Department of CSE, PESU

DATA ANALYTICS

Teaching Team 2024 - Course Instructors

Dr. Gowri Srinivasa

Dr. Bharathi R.

Dr. Deepu R

Dr. Jyothi R.

Dr. Prajwala T.R.

Dr. Sudeepa Roy Dey

Dr. Sujatha R. Upadhyaya

Dr. K. S. Nagegowda

Prof. Suresh Jamadagni

DATA ANALYTICS

Teaching Team 2024 - Teaching Assistants

Mr. Vibhav Vasudevan
vibhav.vasudevan@gmail.com

Mr. Anirudh Lakhotia
anirudhlakhotia5@gmail.com

Ms. Amritha GK
amrithagk12@gmail.com

Mr. Karthik Namboori
namkarthik2003@gmail.com

Mr. Anshul Ranjan
anshulpranjan@gmail.com

Ms. Sanjana Suresh
sanjanasuresh2709@gmail.com

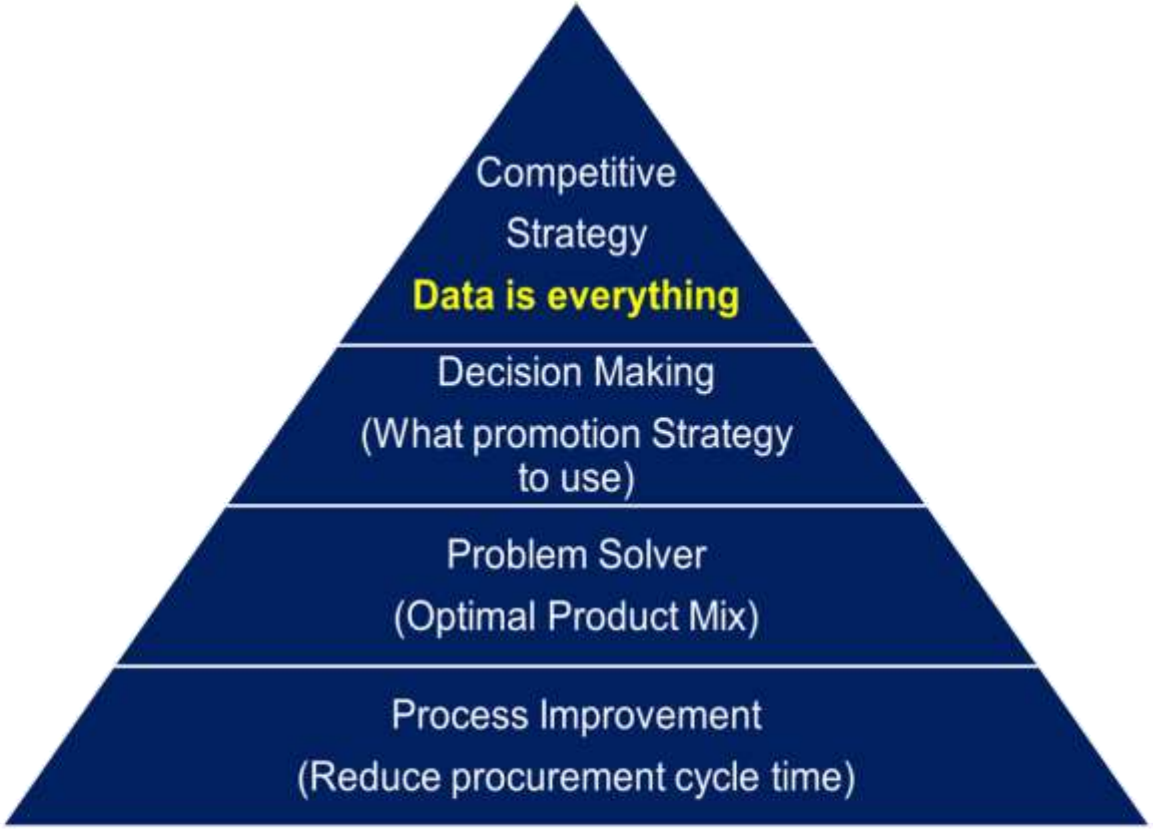
What is data analytics?

- The science of examining raw data to **elicit patterns**, **develop insights** , and **draw conclusions** to help take a **business decision.**

- **The need** : Business decisions are very complex. There exist several alternate solutions, complex interdependent factors and lack of available time to take a decision.

- Analysis vs Analytics
  - Analysis – Examining and understanding past data.
  - Analytics – Analysis + forecasting (or predictive modeling).

Pyramid of analytics applications and Data driven decision making



Competitive Strategy

Data is everything

Decision Making

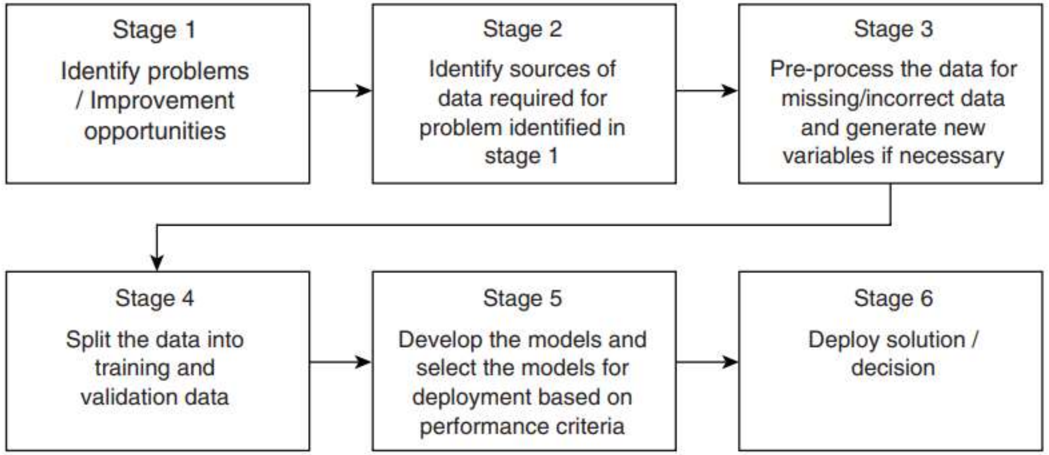
(What promotion Strategy to use)

Problem Solver

(Optimal Product Mix)

Process Improvement

(Reduce procurement cycle time)



Stage 1: Identify problems / Improvement opportunities

Stage 2: Identify sources of data required for problem identified in stage 1

Stage 3: Pre-process the data for missing/incorrect data and generate new variables if necessary

Stage 4: Split the data into training and validation data

Stage 5: Develop the models and select the models for deployment based on performance criteria

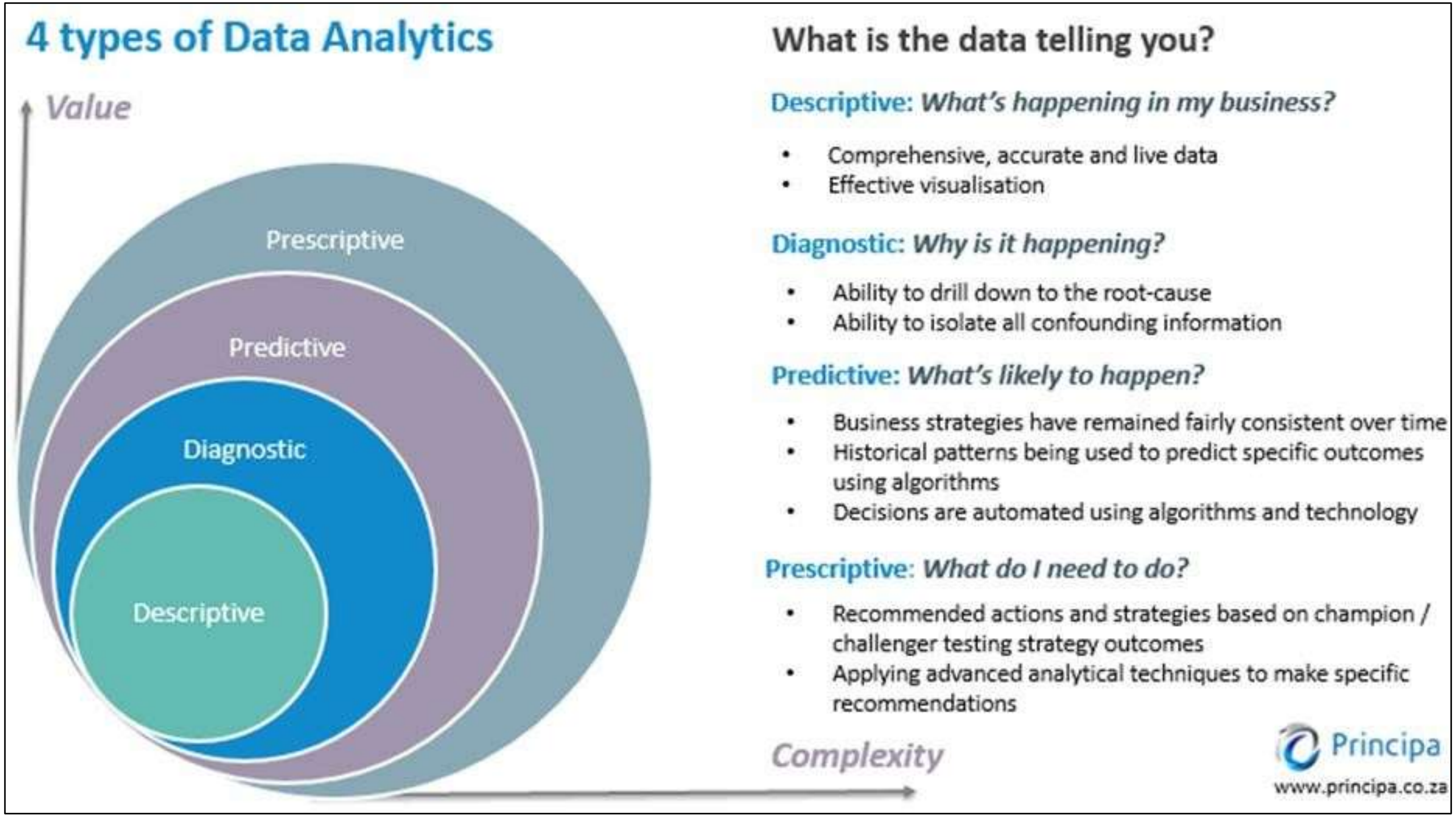
Stage 6: Deploy solution / decision

FIGURE 1.1 Business analytics – Data-driven decision-making flow diagram.

Slide courtesy of **Dr. U. Dinesh Kumar**, Professor, IIM-B (author, ‘Business Analytics’, 2nd Ed., Wiley, 2022)

DATA ANALYTICS

What does it involve?



4 types of Data Analytics

What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

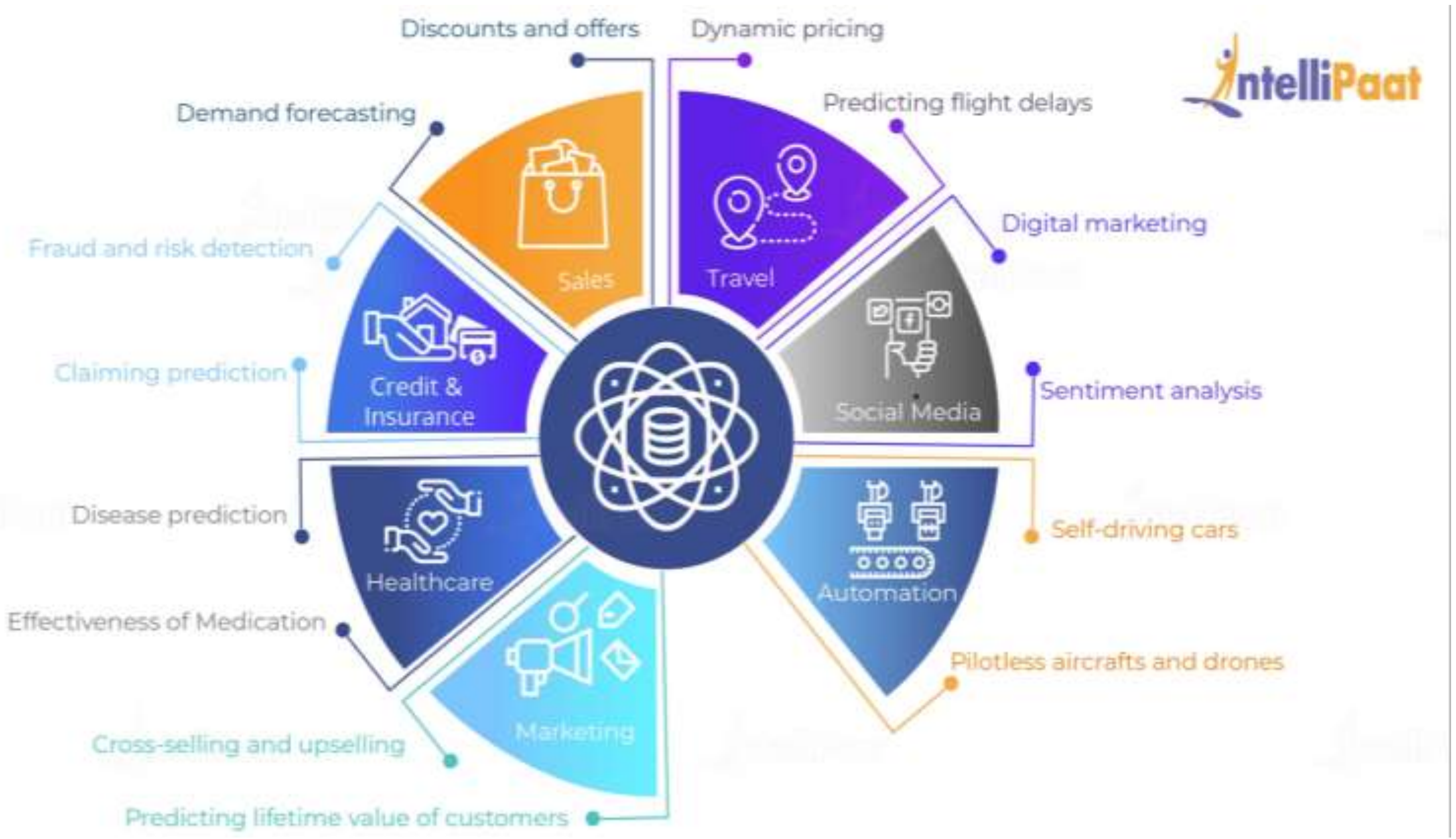
- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

Slide courtesy of **Dr. Mamatha H. R.**, Professor, Dept. of CSE, PES University

Why is it important?

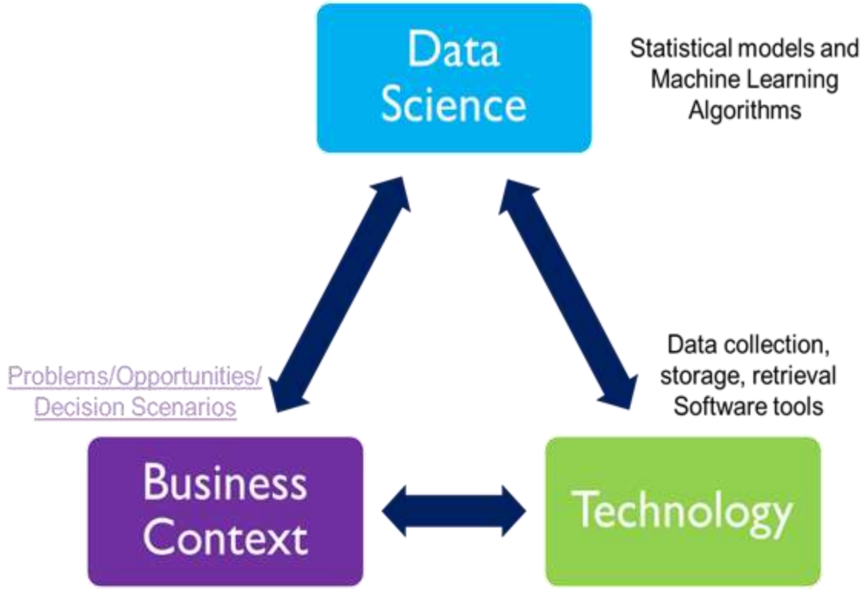
Data Analytics is used in all these application areas…



… and more!

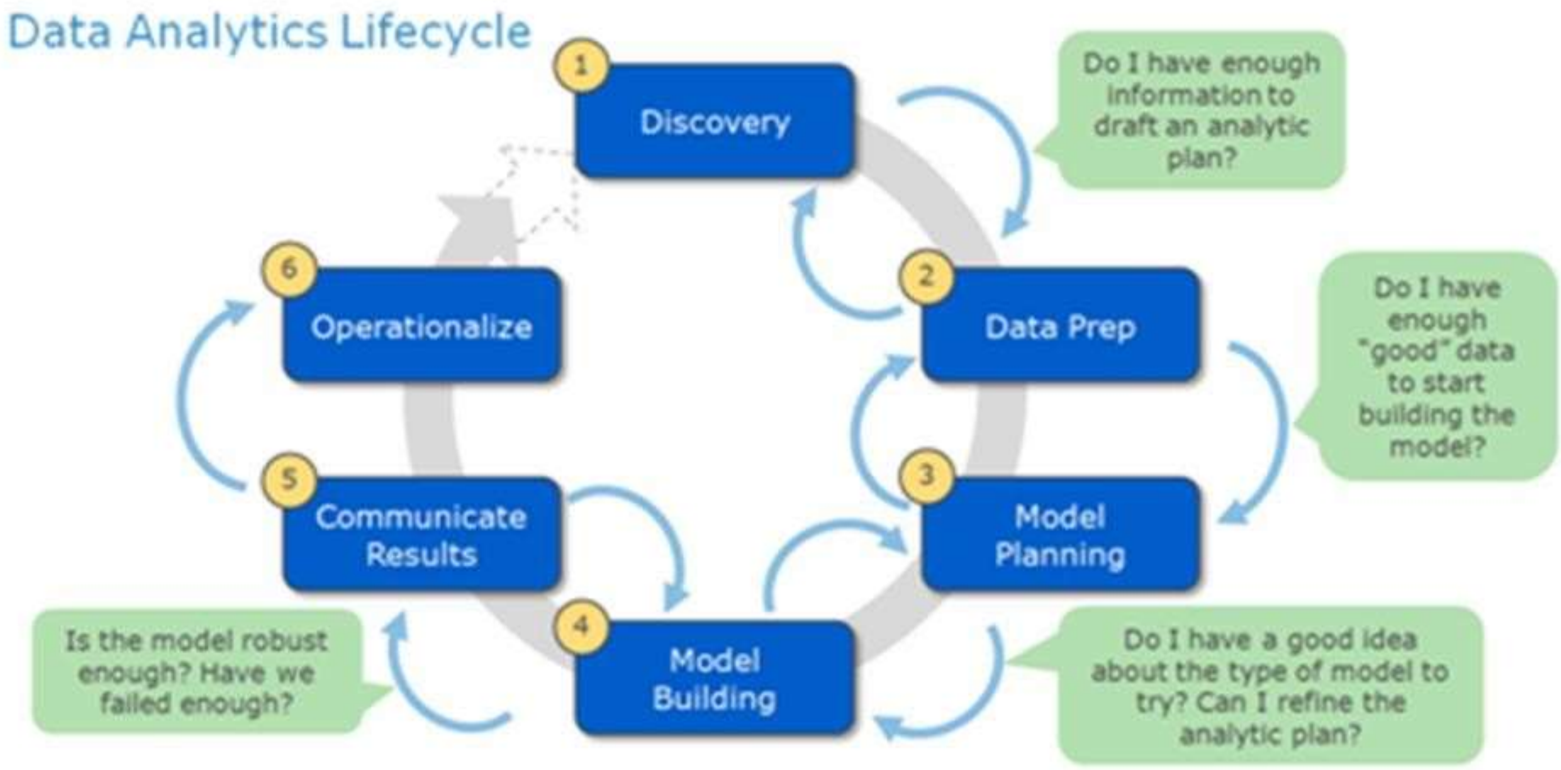
Few examples

- **Banking** : To reduce cheque clearance time, in determining loan approval and interest rate.
- **E-Commerce** : To analyze buyer behavior to plan inventory and recommend products.
- **Retail stores** : Shelf space allocation to drive the profits up.
- **OTT Platforms** : Recommend content a user would like.


Data Science
Statistical models and Machine Learning Algorithms
Problems/Opportunities/ Decision Scenarios
Data collection, storage, retrieval Software tools
Business Context
Technology

DATA ANALYTICS

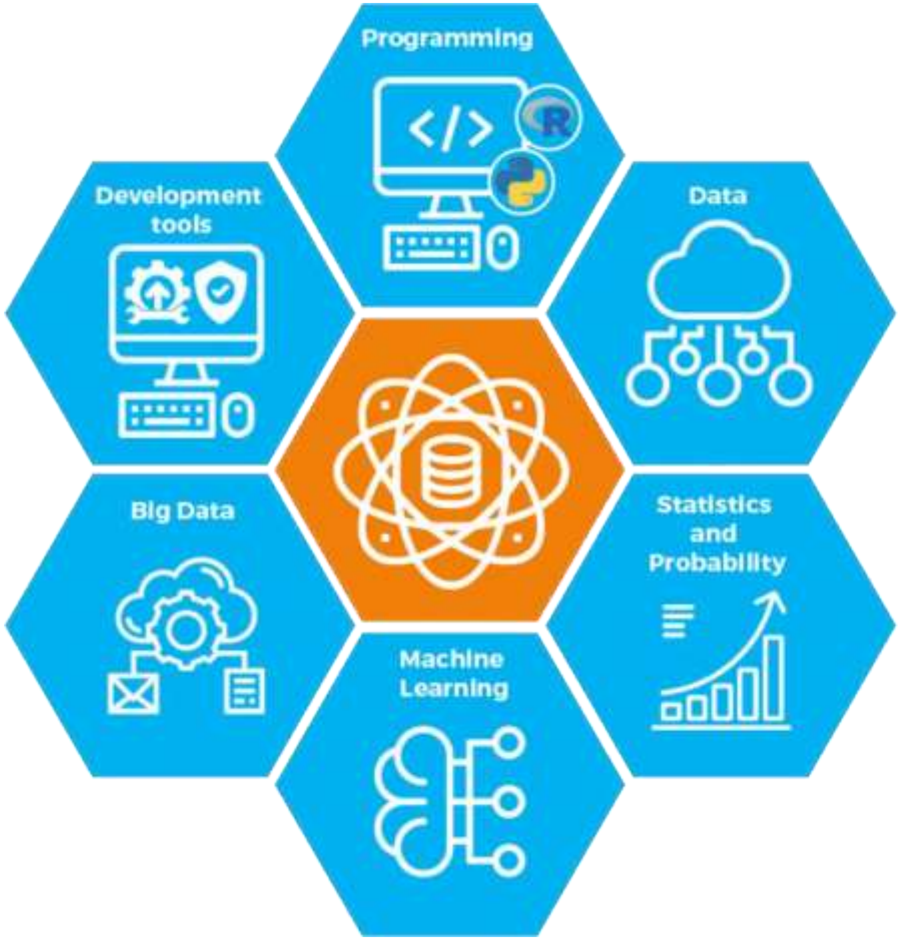
Lifecycle


Data Analytics Lifecycle
1
Discovery
2
Data Prep
3
Model Planning
4
Model Building
5
Communicate Results
6
Operationalize
Do I have enough information to draft an analytic plan?
Do I have enough "good" data to start building the model?
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
Is the model robust enough? Have we failed enough?

Slide courtesy of **Dr. Mamatha H. R.**, Professor, Dept. of CSE, PES University

DATA ANALYTICS

Skills Required



And…
a secret ingredient

Intuition or
deductive reasoning
and domain knowledge

Case Study

Indian online grocery store bigbasket.com

Problem context driving analytics : “Did you forget?” feature

- The ability to predict the items that a customer may have forgotten to order can have a significant impact on the profits of online grocers such as bigbasket.com

- The ability to ask right questions is an important success criteria for analytics projects.

Case Study

Indian online grocery store bigbasket.com

Technology:

- To find out whether a customer has forgotten to place an order for an item

- Information technology is used for data capture, data storage, data preparation, data analysis, data share and to deploy solution

- An important output of analytics is automation of actionable items derived from analytical models which is usually achieved using IT

Indian online grocery store bigbasket.com

Data science is the most important component of analytics, it consists of statistical and operations research techniques, machine learning and deep learning algorithms.

The objective of the data science component of analytics is to identify the most appropriate statistical model/machine learning algorithm that is best based on a measure of accuracy.

Example: "did you forget?" prediction is a classification problem in which customers are classified into:

1. Forget
2. Not forget

Data Sources

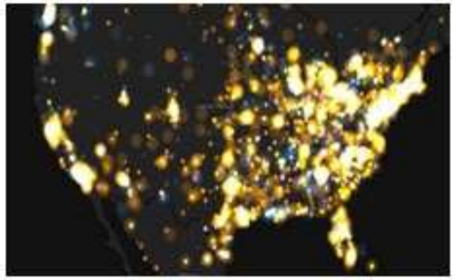
- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies.
- New mantra
  - Gather whatever data you can whenever and wherever possible
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned

Cyber Security

[NEWSFACTOR NETWORK]

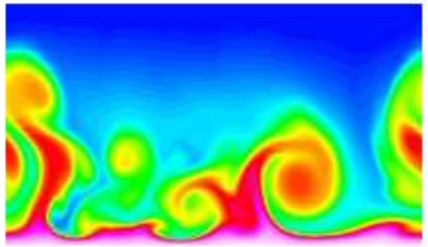
E-Commerce

Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

Data Sources

- Lots of data is collected and warehoused every day
- Yahoo has peta bytes of web data
- Facebook has billions of active users
- Purchases at department/ grocery stores, e-commerce
    - Amazon handles millions of visits/day
- Bank/Credit Card transactions

How large is *big* (data)?

- 1 bit
- 1 byte = 8 bits
- 1 KB = 1024 bytes
- 1 MB = 1024 KB (kilobytes)
- 1 GB = 1024 MB (megabytes)
- 1 TB = 1024 GB (gigabytes) ≈ 10^{12} bytes
- 1 PB = 1024 TB (terabytes) ≈ 10^{15} bytes

20 PB = amt of data processed by Google per day!

- 1 EB = 1024 PB (petabytes)
- 1 ZB = 1024 EB (exabytes)
- 1 YB = 1024 ZB (zettabytes)

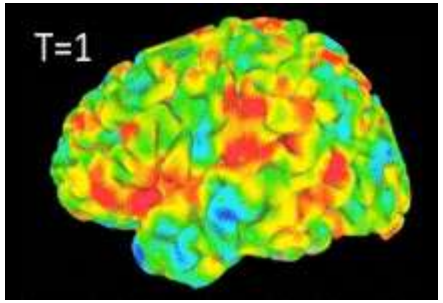
- What is a Domegemegrottebyte?

DATA ANALYTICS

Data Sources

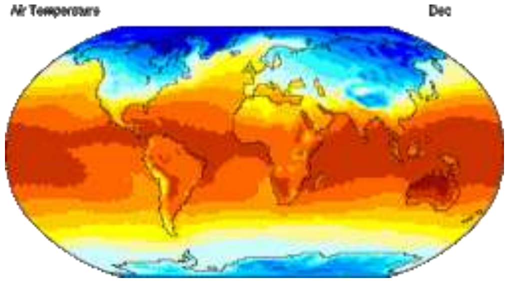
Data collected and stored at enormous speeds

- Remote sensors on a satellite - NASA EOSDIS archives over petabytes of earth science data / year
- Telescopes scanning the skies - Sky survey data
- High-throughput biological data
- Scientific simulations - terabytes of data generated in a few hours

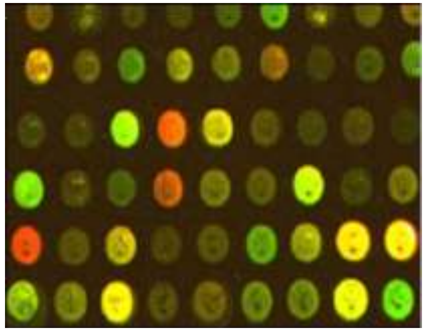


fMRI Data from Brain

Sky Survey Data



Surface Temperature of Earth



Gene Expression Data

What is Data?

- Collection of ***data objects*** and their ***attributes.***
- An ***attribute*** is a property or characteristic of an object.
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature.
- A collection of attributes describe an ***object.***
  - An object is also known as a record, point, case, sample, entity, or instance.

Attributes

Objects

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Attribute Values

- ***Attribute values*** are numbers or symbols assigned to an attribute for a particular object

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters

  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different

Types of Attributes

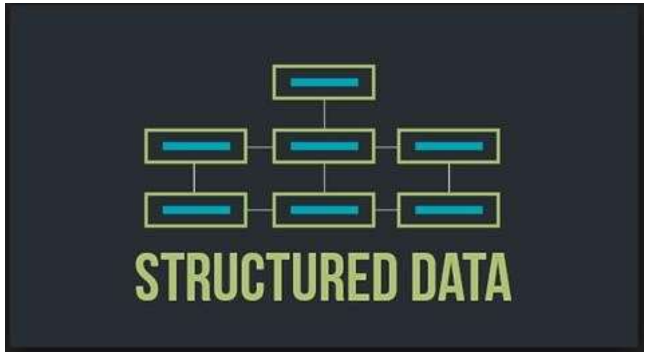
- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: Calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: Temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, orthe set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating point variables.

Data Representations

- Structured Data:Structured data means that the data is described in a matrix form with labelled rows and columns.
- Unstructured Data:Any data that is not originally in the matrix form with rows and columns is an unstructured data.
- Semi structured:Semi-structured data (also known as partially structured data) is a type of data that doesn't follow the tabular structure associated with relational databases or other forms of data tables but does contain tags and metadata to separate semantic elements and establish hierarchies of records and fields.



Data Representations

- Relational databases and spreadsheets. – Structured Data

- Text and multimedia content. Photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents. - Unstructured Data

- XML documents and NoSQL databases. – Semi structured Data

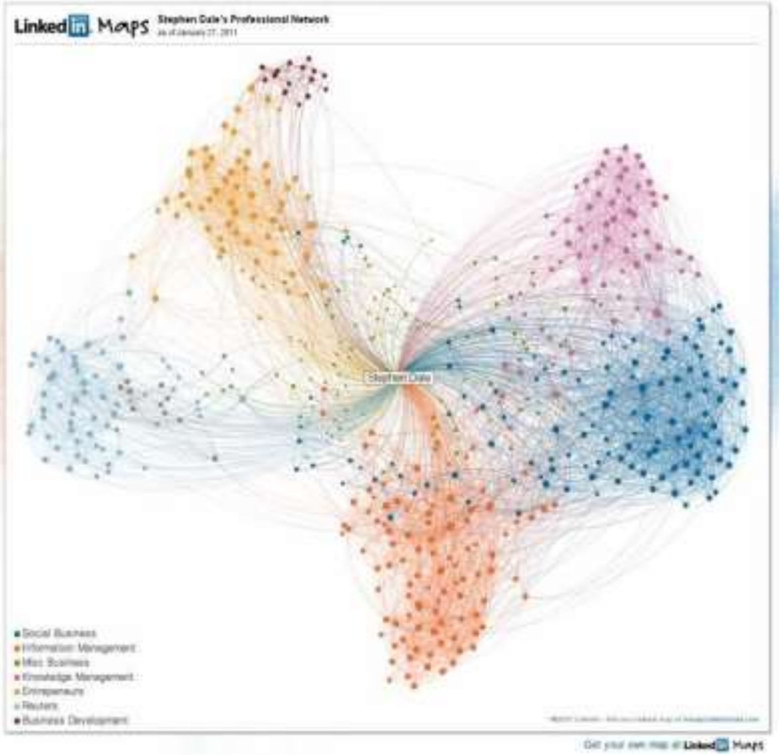
- For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text.

DATA ANALYTICS

Data Representations

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data

- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures



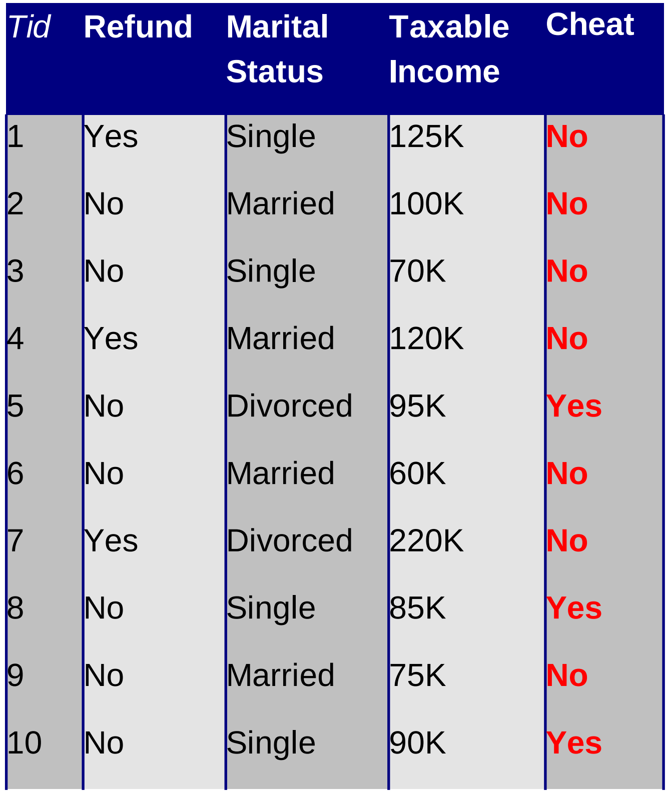
Data Representations

- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data

- Spatial, image and multimedia
  - Spatial data: maps
  - Image data
  - Video data

Data Representations-Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Data Representations-Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Data Representations-Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	Play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

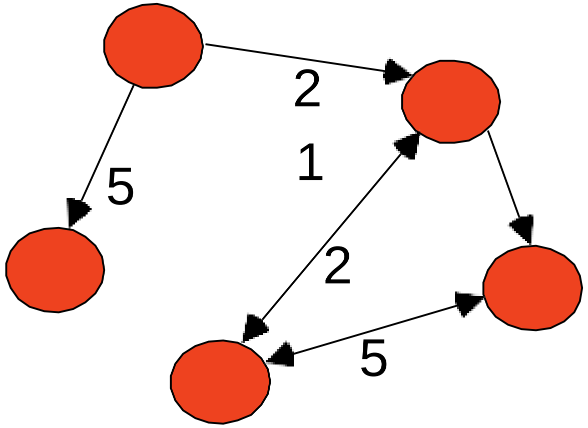
Data Representations-Transaction data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

TID	*Items*
1	Bread,Coke,Milk
2	Beer,Bread
3	Beer,Coke,Diaper,Milk
4	Beer,Bread,Diaper,Milk
5	Coke,Diaper,Milk

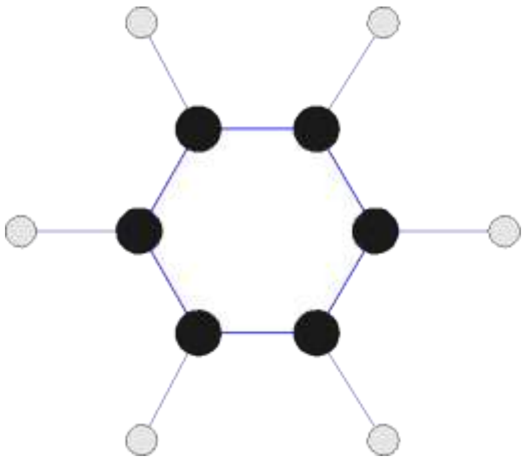
Data Representations

- **Graph Data**
- Examples: Generic graph and HTML Links



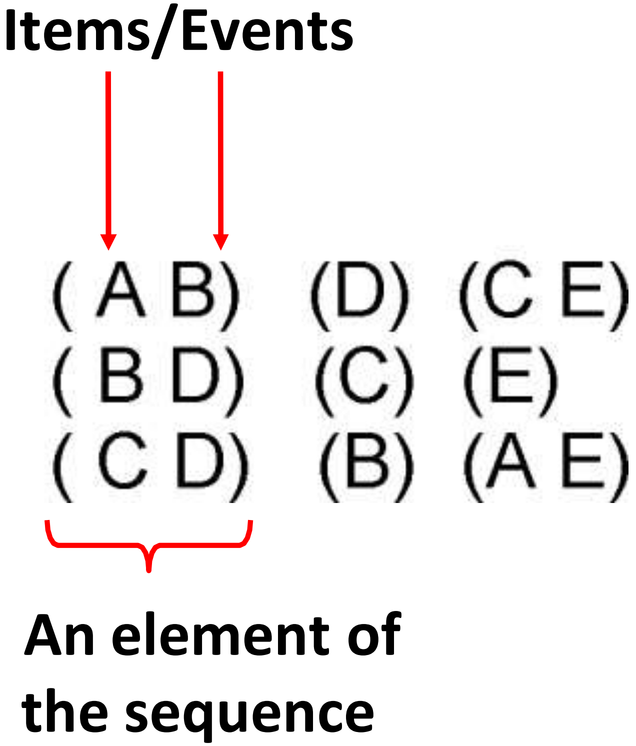
```
<a>Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

- **Chemical Data**
- Benzene Molecule: C_6H_6



Data Representations-Ordered Data

- Sequences of transactions

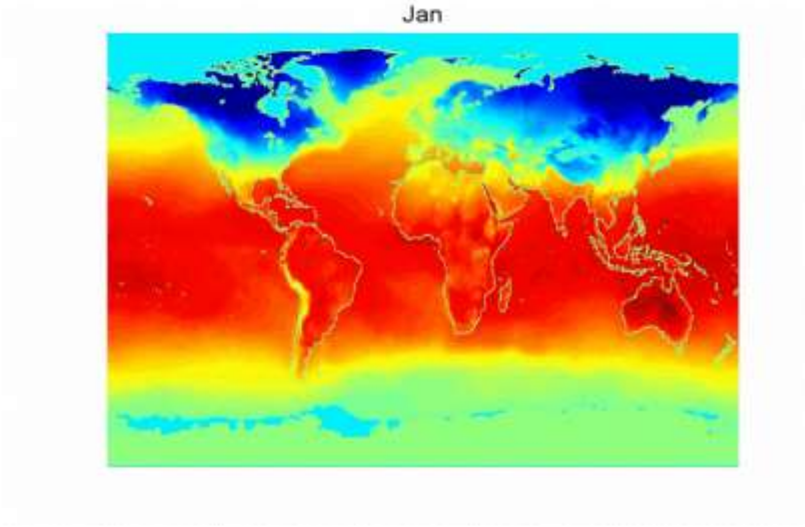


Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

- Spatio-Temporal Data



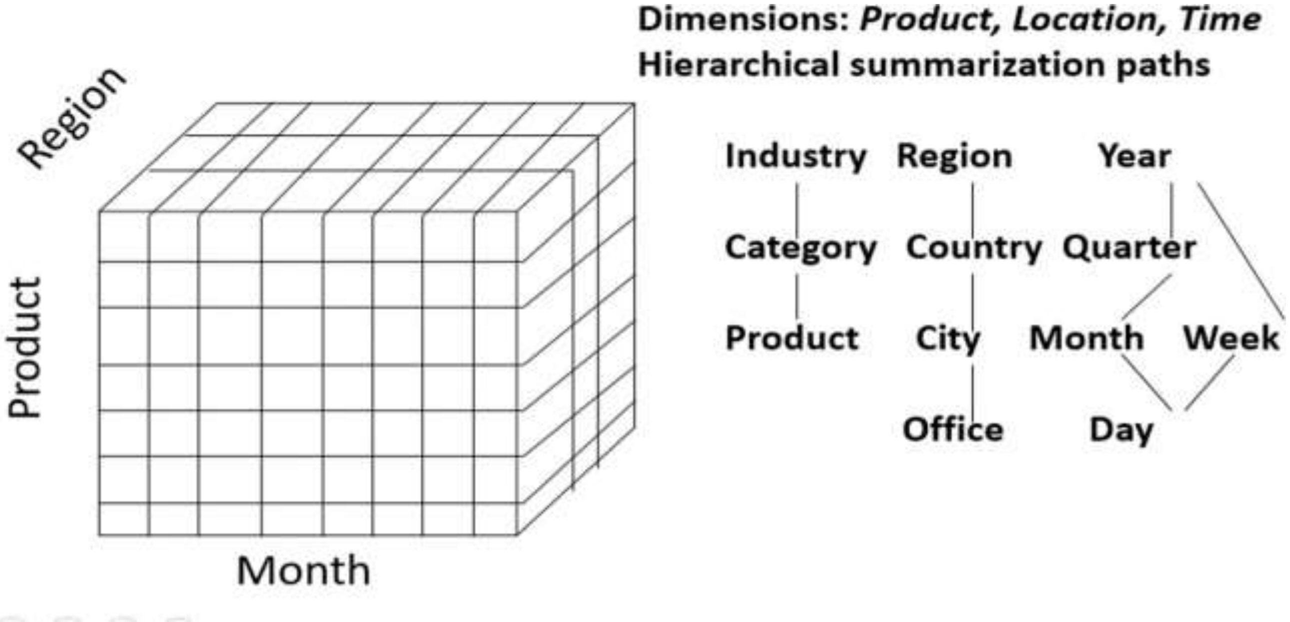
- Genomic sequence data

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

Data Representations- Data Warehouse

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

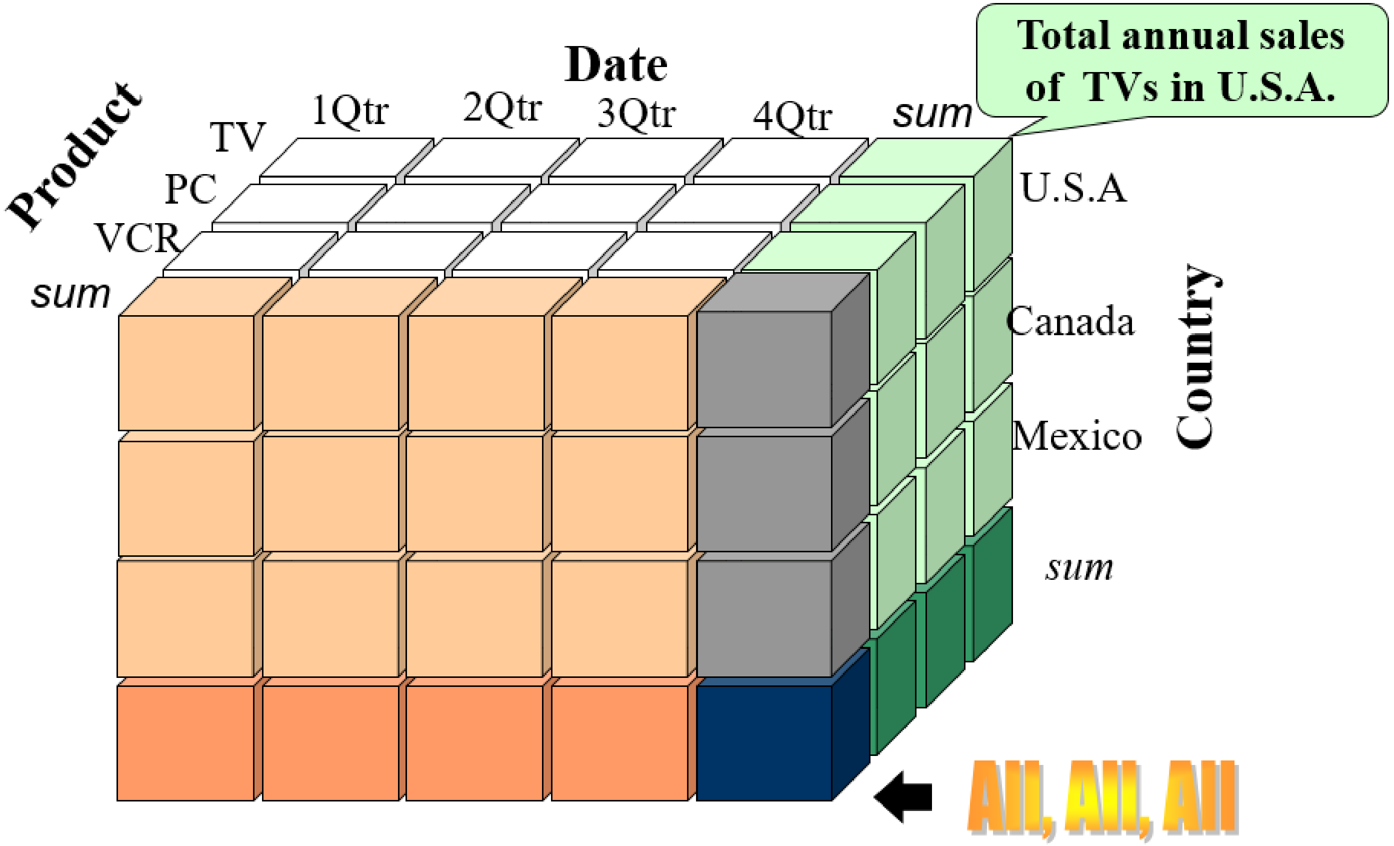
Sales volume as a function of product, month, and region



Dimensions: *Product, Location, Time*
Hierarchical summarization paths

Industry Region Year
Category Country Quarter
Product City Month Week
Office Day

A Sample Data Cube

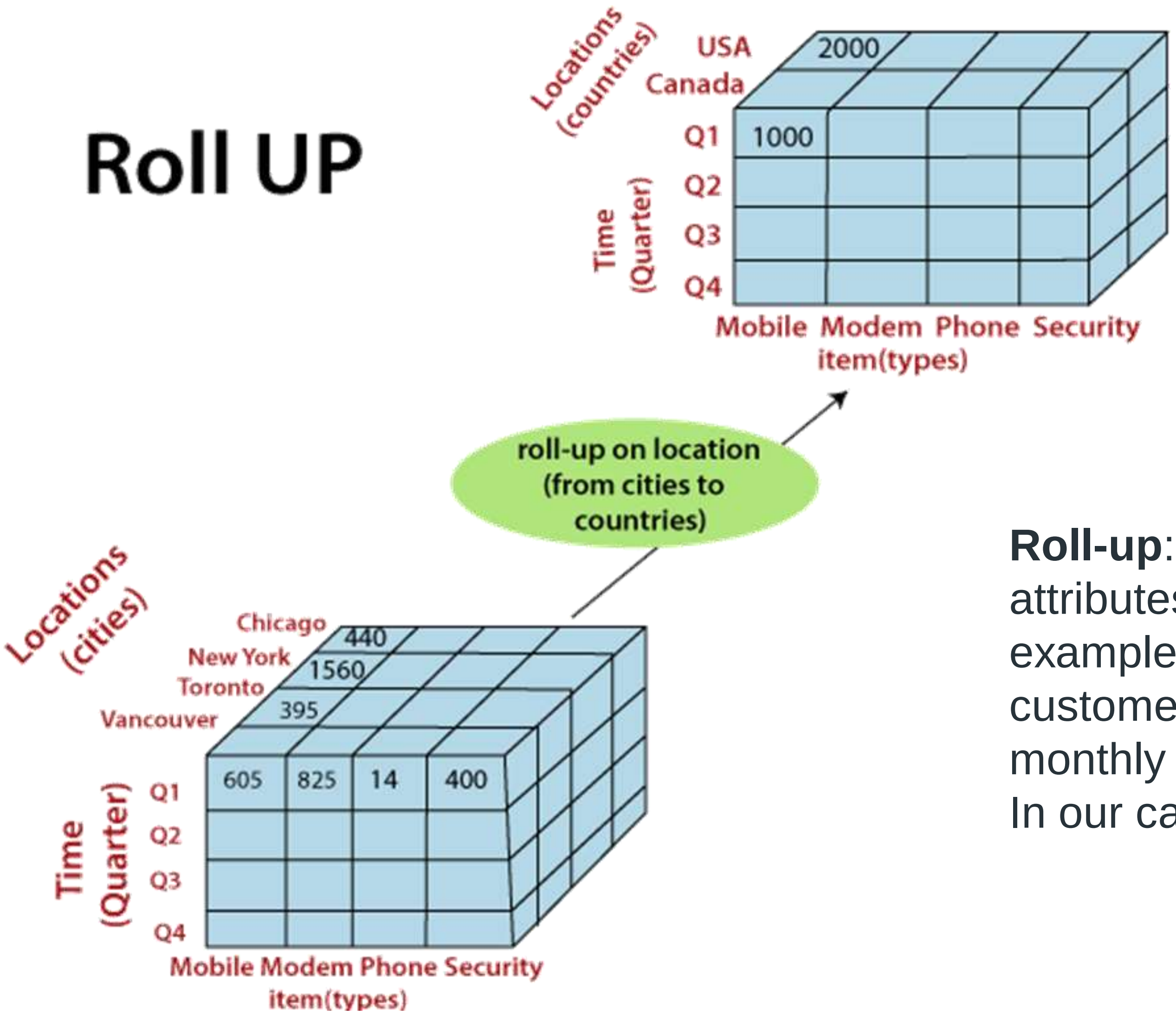


Typical OLAP Operations

- Roll up (drill-up): summarize data
  - by climbing up hierarchy or by dimension reduction
- Drill down (roll down): reverse of roll-up
  - from higher level summary to lower level summary or detailed data, or introducing new dimensions
- Slice and dice: project and select
- Pivot (rotate):
  - reorient the cube, visualization, 3D to series of 2D planes
- Other operations
  - drill across: involving (across) more than one fact table
  - drill through: through the bottom level of the cube to its back-end relational tables (using SQL)

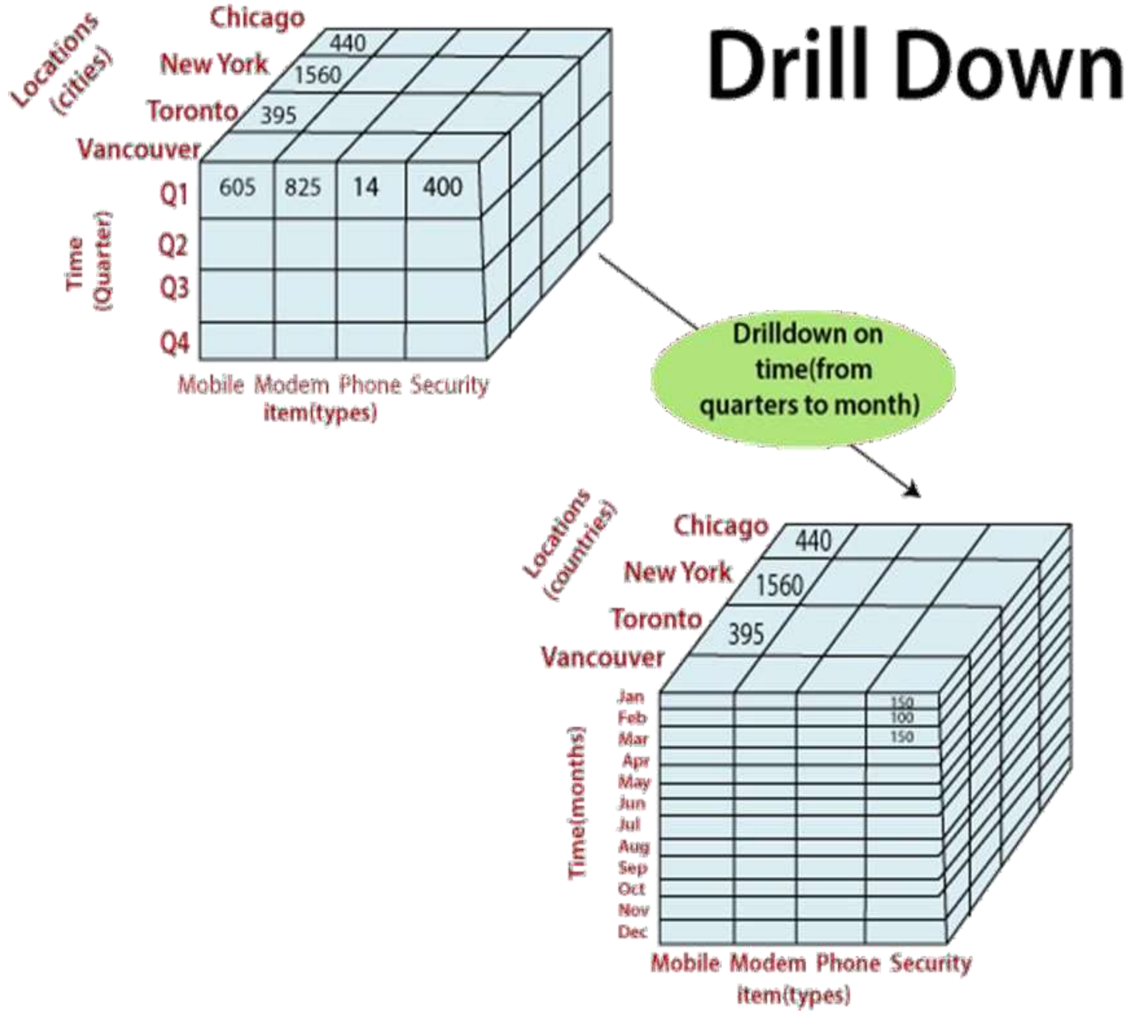
Typical OLAP Operations

Roll UP



Roll-up: operation and aggregate certain similar data attributes having the same dimension together. For example, if the data cube displays the daily income of a customer, we can use a roll-up operation to find the monthly income of his salary.
In our case we roll-up from cities to countries as shown

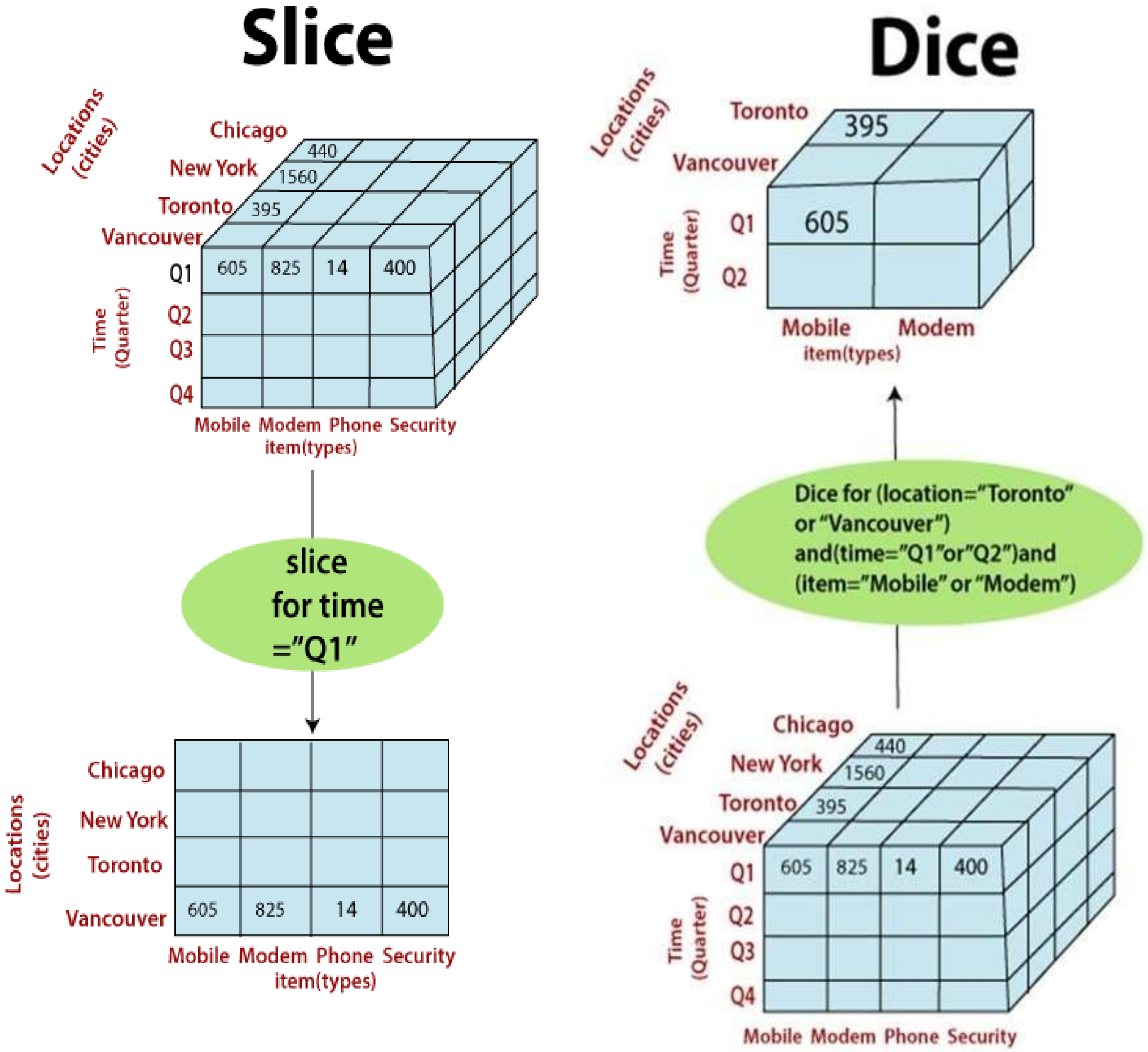
Typical OLAP Operations



Drill-down: this operation is the reverse of the roll-up operation. It allows us to take particular information and then subdivide it further for coarser granularity analysis. It zooms into more detail. For example- if India is an attribute of a country column and we wish to see villages in India, then the drill-down operation splits India into states, districts, towns, cities, villages and then displays the required information.

In our case we drill down from quarters to months.

Typical OLAP Operations

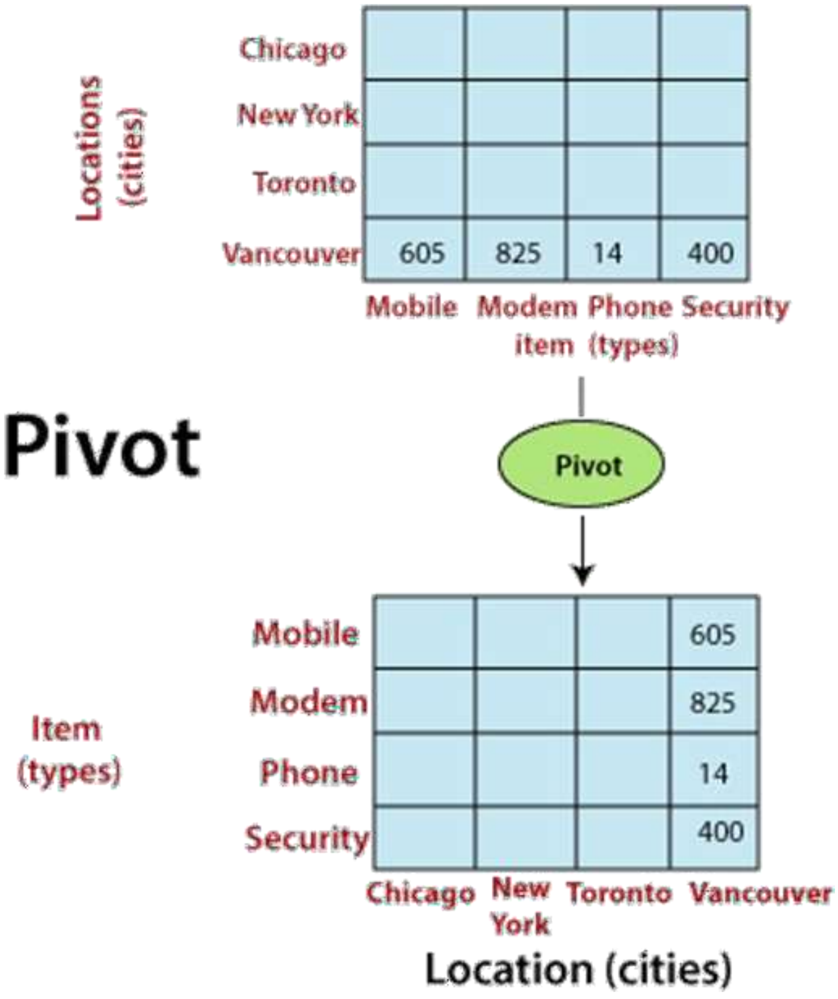


•**Slicing**: this operation filters the unnecessary portions. Suppose in a particular dimension, the user doesn't need everything for analysis, rather a particular attribute.

•**Dicing**: this operation does a multidimensional cutting, that not only cuts only one dimension but also can go to another dimension and cut a certain range of it. As a result, it looks more like a subcube out of the whole cube(as depicted in the figure).

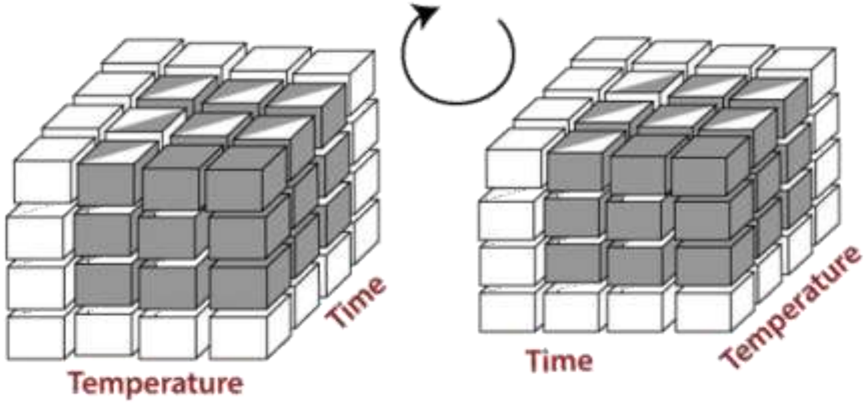
Typical OLAP Operations

Pivot



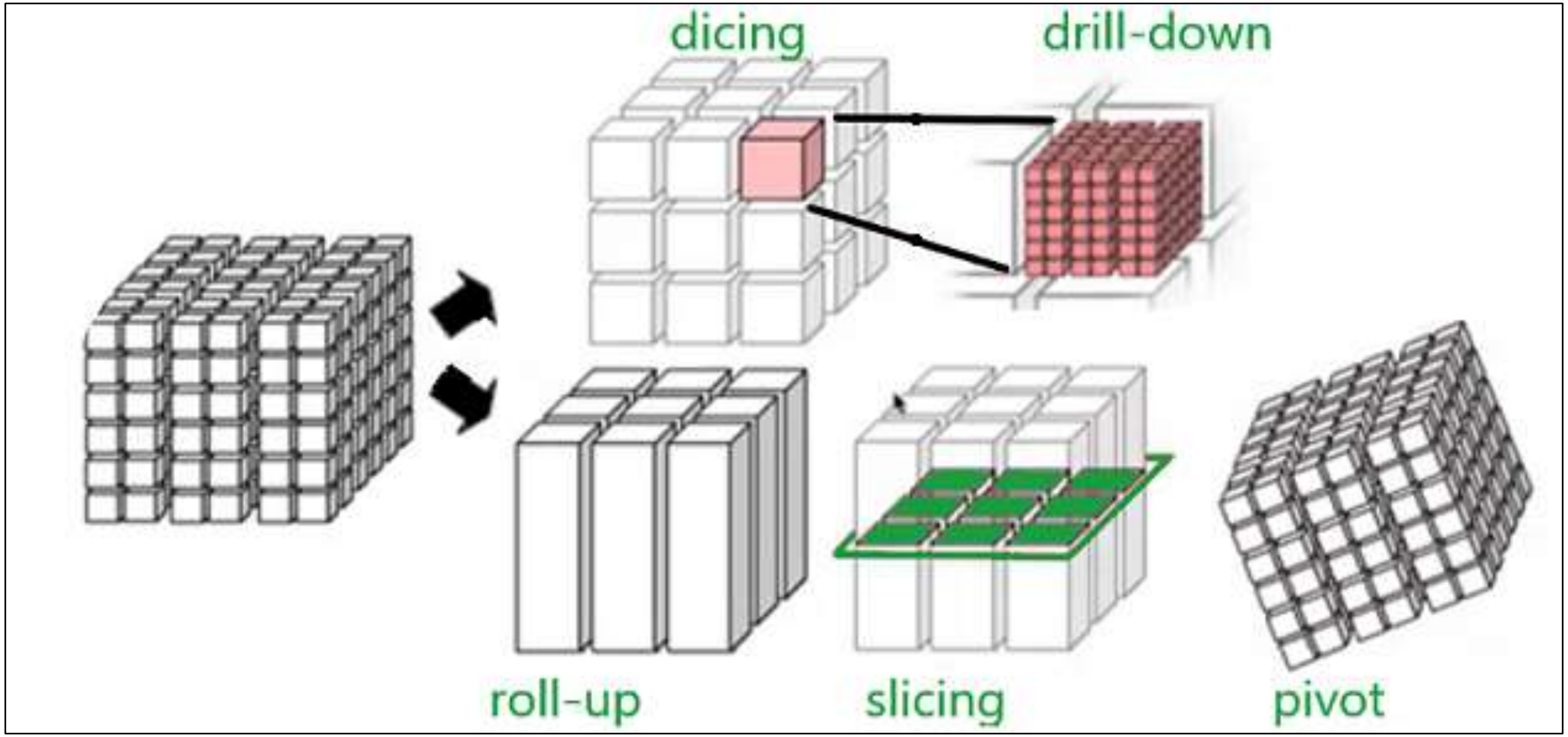
Locations (cities)	Mobile	Modem	Phone	Security
Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

Item (types)	Chicago	New York	Toronto	Vancouver
Mobile				605
Modem				825
Phone				14
Security				400



•**Pivot**: this operation is very important from a viewing point of view. It basically transforms the data cube in terms of view. It doesn’t change the data present in the data cube. For example, if the user is comparing year versus branch, using the pivot operation, the user can change the viewpoint and now compare branch versus item type.

A Quick Glance on OLAP operations



Test your understanding

- To what type of an attribute does **shoe size** belong to?

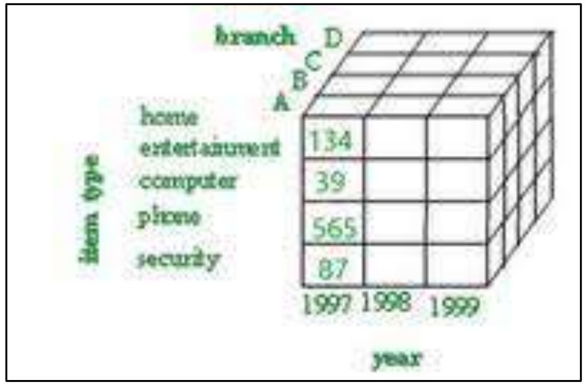
  Interval

- Which OLAP operation are you likely to perform at the **end** of the financial year?

  Roll-Up

- The example here is a 3D cube having attributes like branch(A,B,C,D),item type (home, entertainment, computer, phone,security),year(1997,1998,1999). If user wants to observe only "branch A" data then which OLAP operation must be performed?

  Slicing



References

- Business Analytics by U. Dinesh Kumar – Wiley 2nd Edition, 2022 Chapter : 1.1-1.7
- Data Mining : Concepts and Techniques by Han, Kamber and Pei , The Morgan Kaufmann Series in Data Management Systems ,3rd Edition Chapter : 4.2.5
- https://www.geeksforgeeks.org/data-cube-or-olap-approach-in-data-mining/

THANK YOU

Dr. Gowri Srinivasa
Professor, Department of Computer Science and Engineering, PES University, Bengaluru
Email: gsrinivasa@pes.edu