Lecture 0

Praphul Chandra

Insofe

Why are you here?

Why are you here – putting it in context.

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Why are you here – putting it in context (again)

Data Analytics is an iterative process

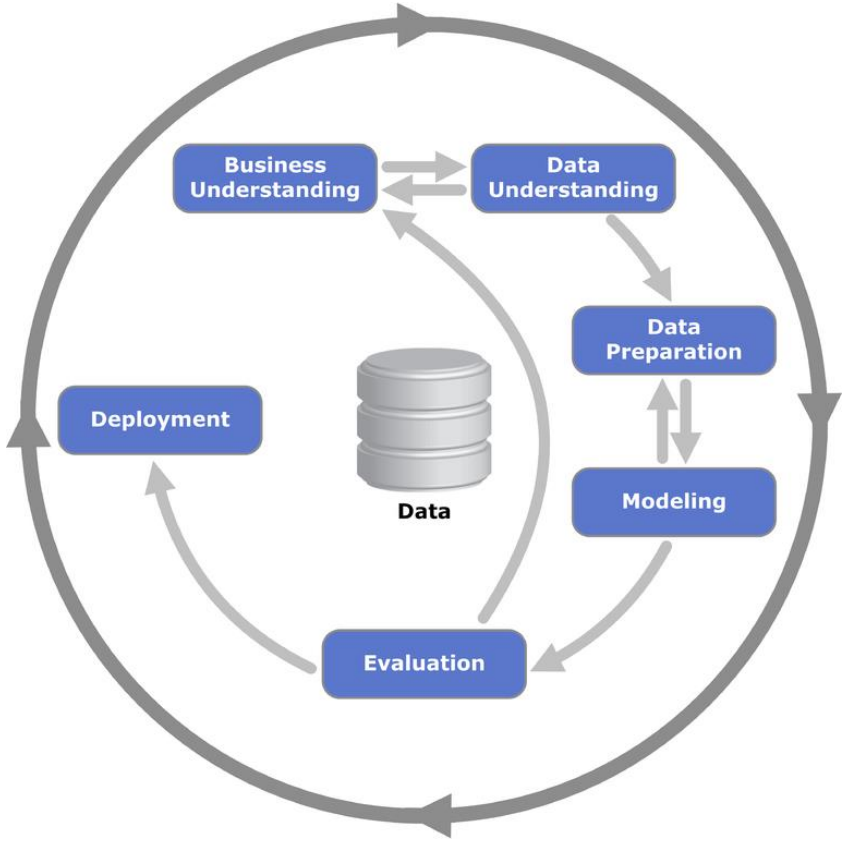
- If you “get it right” at the first go on the data, you have not explored enough – Iterate!

This module will

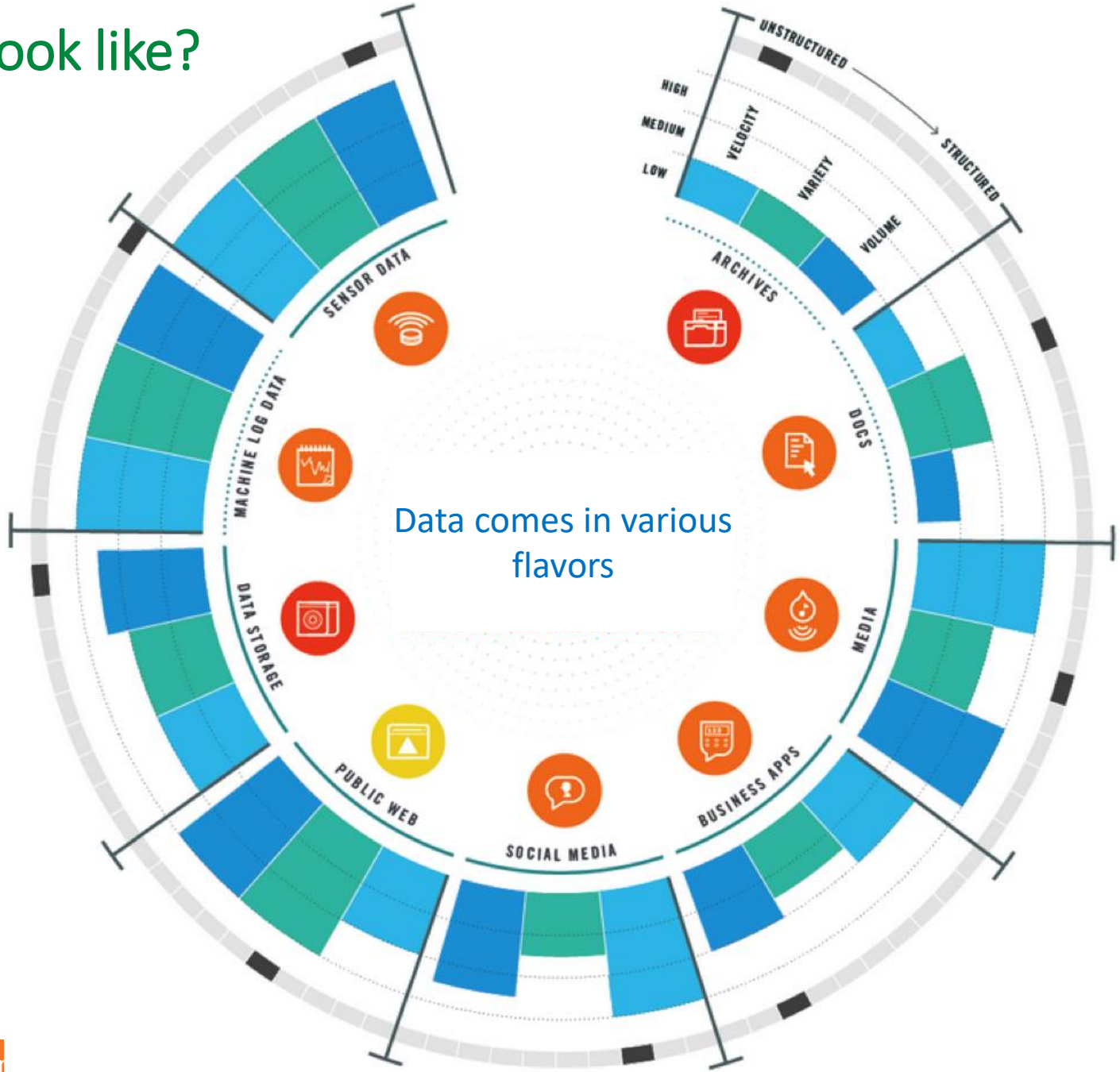
- Introduce key concepts

Once you know the science, it really is an art

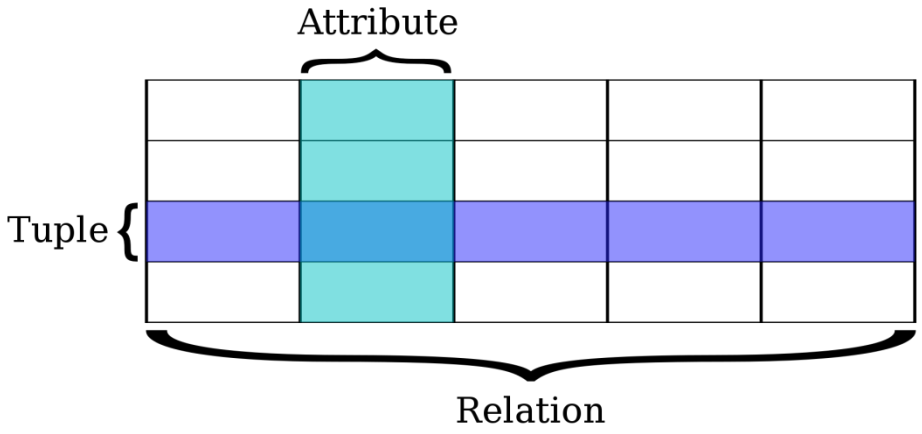
- Learn by doing.
- And doing again…

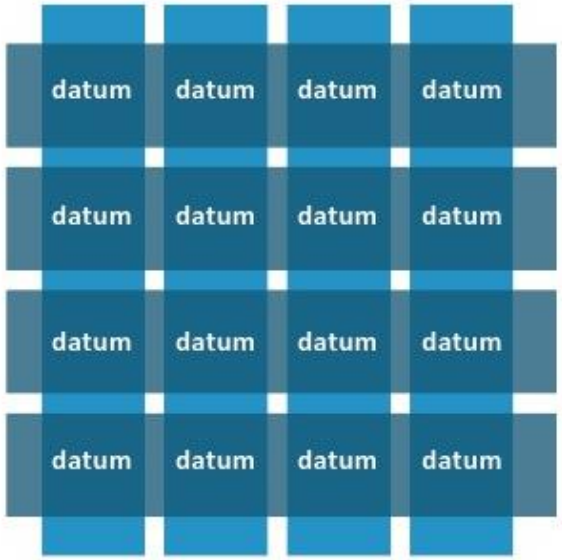


What does data look like?



Data Model: Relational





Salient Features

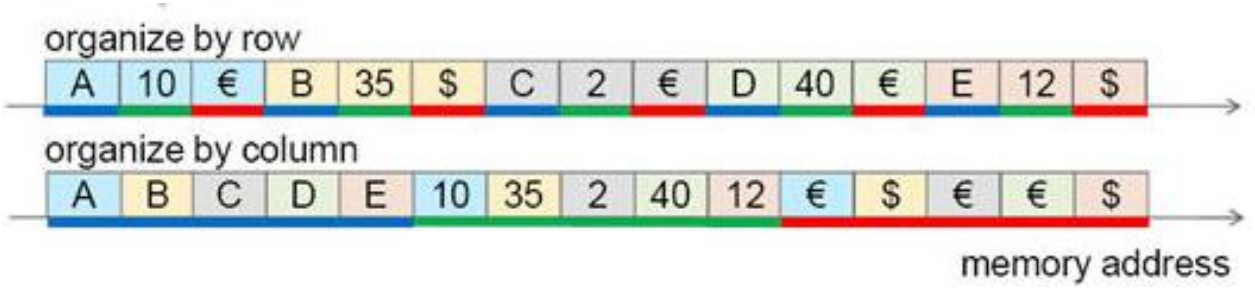
Structure before Store

Rigid “Tabular” Structure

Relational Algebra *(Attributes are “somehow” related)*

Data Model : Relational (Columnar Store)

A	10	€
B	35	$
C	2	€
D	40	€
E	12	$



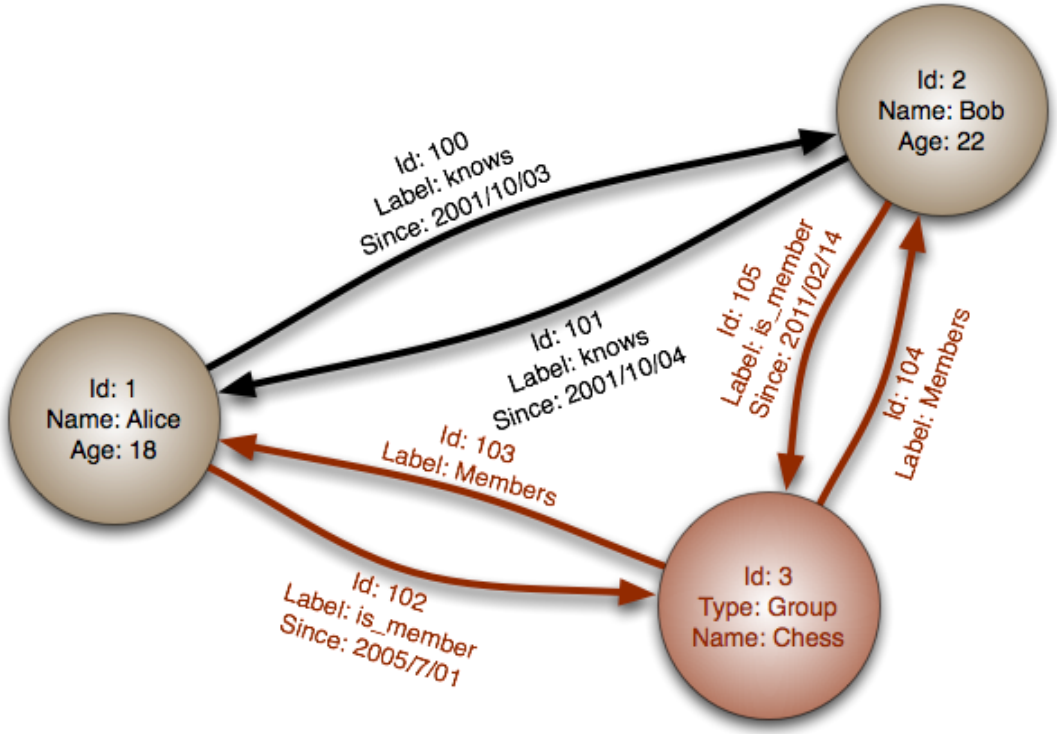
Salient Features

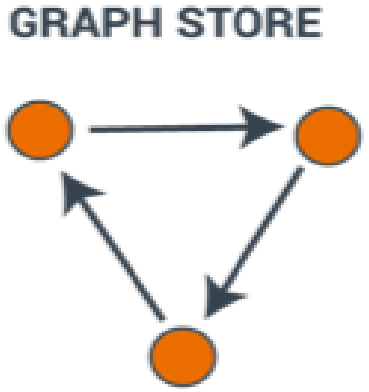
Relational Data Model

Change in how we store data in physical memory

Columnar store ➔ faster to access a “feature” / “attribute” for all rows

Data Model : Graph (Network)





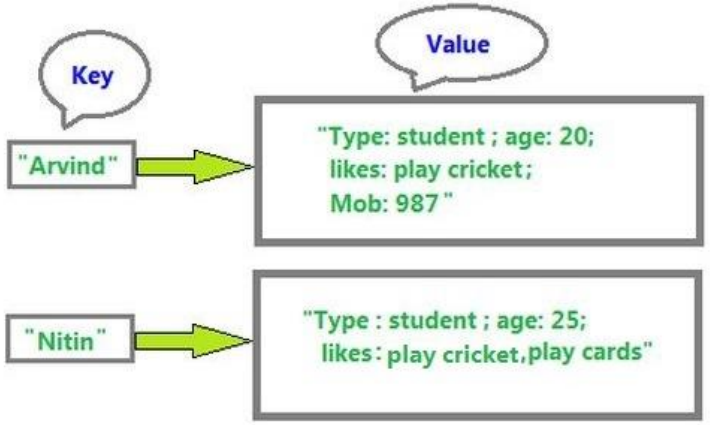
Salient Features

Structure before Store

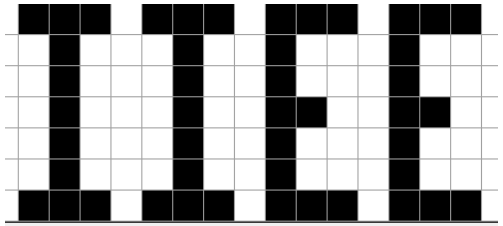
Nodes & Edges Structure

Graph Theory *(Entities & Relationship)*

Data Model: "Unstructured"



$$
\begin{array}{lcccc}
 & s_1 & s_2 & s_3 & s_4 \\
\text{how} & 1 & 0 & 0 & 0 \\
\text{much} & 1 & 1 & 0 & 0 \\
\text{wood} & 2 & 2 & 0 & 2 \\
\text{would} & 1 & 1 & 0 & 1 \\
\text{a} & 2 & 2 & 0 & 1 \\
\text{woodchuck} & 2 & 3 & 1 & 2 \\
\text{chuck} & 2 & 3 & 1 & 2 \\
\text{if} & 1 & 1 & 0 & 1 \\
\text{could} & 1 & 2 & 1 & 1 \\
35 & 0 & 0 & 1 & 0 \\
\text{cubic} & 0 & 0 & 1 & 0 \\
\text{feet} & 0 & 0 & 1 & 0 \\
\text{of} & 0 & 0 & 1 & 1 \\
\text{dirt} & 0 & 0 & 1 & 0 \\
700 & 0 & 0 & 0 & 1 \\
\text{pounds} & 0 & 0 & 0 & 1
\end{array}
\rightarrow A_0 =
\begin{pmatrix}
1 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
2 & 2 & 0 & 2 \\
1 & 1 & 0 & 1 \\
2 & 2 & 0 & 1 \\
2 & 3 & 1 & 2 \\
2 & 3 & 1 & 2 \\
1 & 1 & 0 & 1 \\
1 & 2 & 1 & 1 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1
\end{pmatrix}.
$$



Salient Features

Key Idea: Store Now; Structure Later *(Schema on Read: Store as files; Impose structure later)*

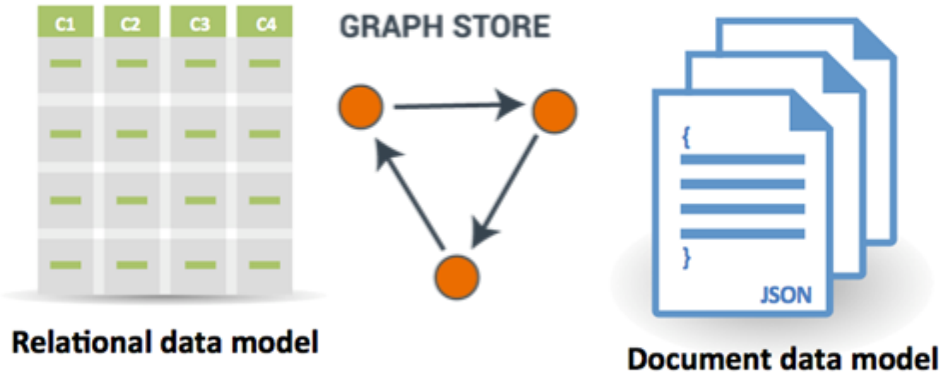
Why? : Lots of data formats can "eventually" be structured; *Structure may change with time, instance, analytics-requirements*

Very Flexible; Highly Scalable

Data Models : Summary

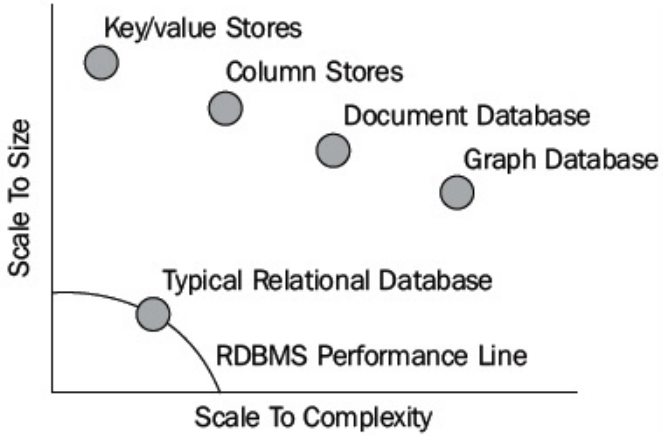
How is the data organized / structured?

- Relational *(Table / Spreadsheet)*
- Key Value *(Unstructured)*
- Graph
- Document



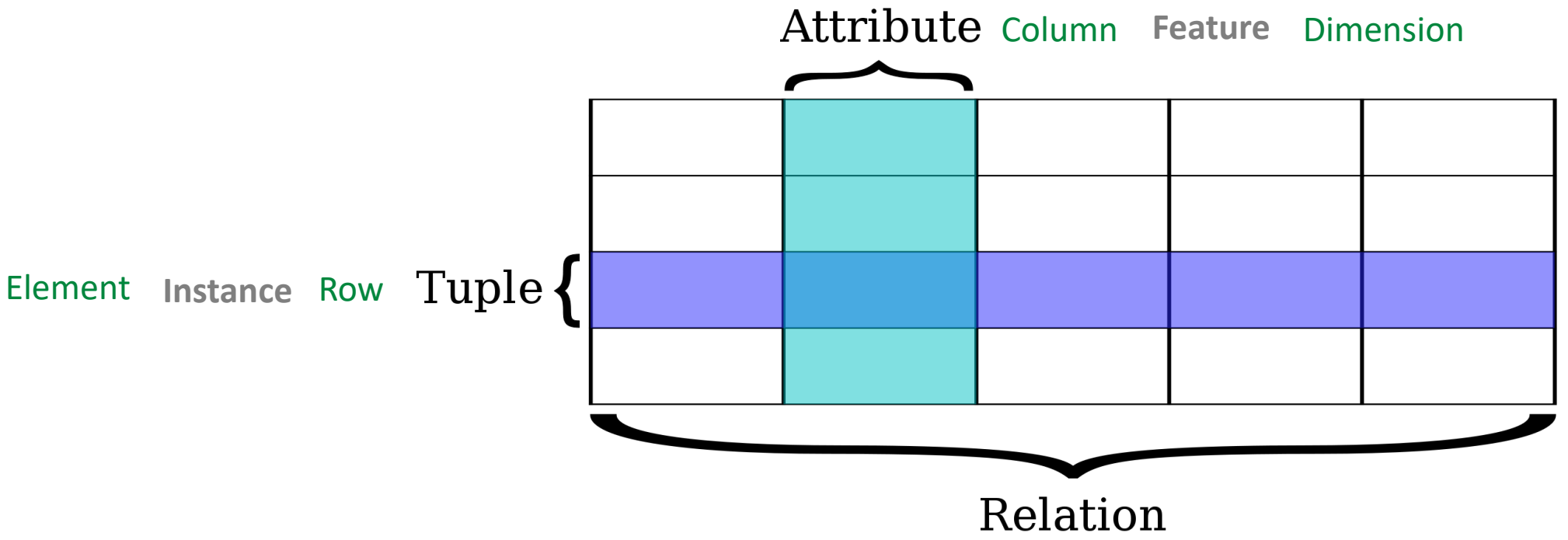
Is it BIG?

- Long *(# of rows; Large n)*
- Wide (a.k.a. High dimensional) *(# of columns; Large p)*
- Can it fit in your RAM?



Patterns in Data

Focus on Relational Data Model



$$\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$$

$$X \in \mathbb{R}^{n \times p}$$

What does data “really” look like?

If you look carefully, data has patterns.

Data Analysis is about finding patterns…

What “patterns”?

- Statistics
  - Summary Statistics (Descriptive)
  - Probability Theory
  - Inferential Statistics (Sample vs. Population)
  - Design of Experiments
- Learning from Data
  - Learning patterns to make predictions
  - Statistical Modeling
  - Function Approximation
- Supervised Learning
  - Regression
  - Classification
- Unsupervised Learning
  - Clustering
  - Association Rules
  - Dimensionality Reduction
- Deep Learning
  - Neural Networks On Steroids + Some insights
  - “Built-in” Feature Engineering
- Reinforcement Learning
  - State, Actions & Rewards
  - Learning from feedback
- Optimization
  - The underlying computational problem

Q?

Praphul Chandra

Insofe