MAS DSE 260: Capstone Project

İlkay ALTINTAŞ, Ph.D.

STEP 2:
Designing the Data Acquisition and Preparation Pipelines

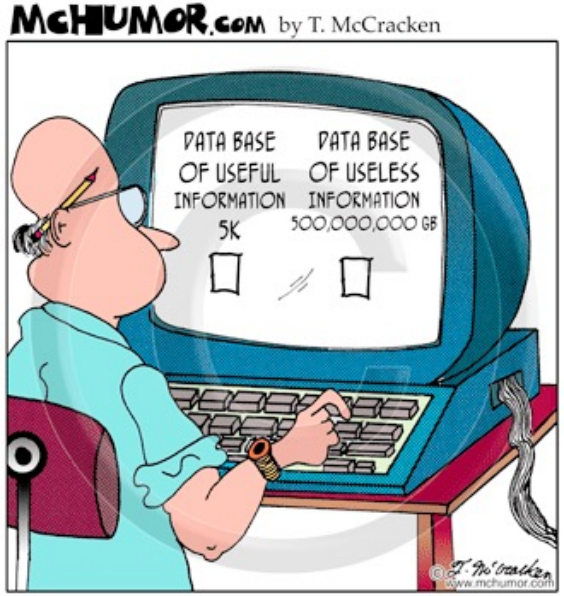
Why do we need this step?

- Data ingest and engineering is the basis for effective modeling
- Many quality issues with data that needs to be cleaned up
  - Not just missing values, but also deeper problems on uncertainty and bias related to data collection
- Defined data pipelines (potentially automated) help with data refresh and updates

If your data pipeline is strong, your modeling will be strong!

Objectives

- Understanding what data is available and useful to you
- Plan for data transport and locate your datasets in the right database, storage or analytics environment
- Model the data
- Identify data integration needs if applicable
- Set up a data pipeline that streamline (and potentially automates) data ingest and refresh
- Design programmatic access to data



©T. McCracken mchumor.com

Planning for a Good Data Pipeline

- Focus on
  - where your datasets come from (data sources)
  - adding context and model data as you collect it
  - privacy and policy issues with the data
  - data quality assurance
  - where each dataset gets stored
  - how can pieces of data be combined
  - how each datasets get used
  - programmatic access to data

Using the PPODS Approach

- Each step in your data pipelines is a separate pod
- Define success metrics for calling a step done

Metrics for accountability should be built into the process.

Cost **Timeline**

Planning of deliverables

Purpose **Expectations**

Pod ➔ sub-process

Defined by:

- Purpose and goal
- Stakeholders
- Expectations
  - Key questions to be answered, production/consumption relationships, needs, dependencies, limits, …
- Contracts
  - Performance, economic, accuracy, policy, privacy, reproducibility, political, …
- Knowns
- Known unknowns

Step II Report Guidelines

- Title, team members and advisor(s)
- Sections:
  - **Raw Data Sources**
    - Summaries of each dataset description from report 1
    - Table for Dataset name, source location, destination in your data pipeline, data movement and processing scripts and notebooks, and data size
  - **Data Exploration, Cleaning, Wrangling and Engineering**
    - Data Exploration Summary
    - Data Preprocessing Approach
    - Approach for storing processed and/or integrated data
      - Processed dataset description for each processed dataset including why you want to process it that way
      - Table for processed data sets including processed data set name, input datasets, link to the processing scripts and notebooks, and provisional data size
    - Approach for Feature Engineering and Data Modeling
      - Summary of feature sets
      - Table for feature set including links to input datasets, feature engineering scripts and notebooks, and provisional data size
    - Approach for Data Access
      - Initial design for data querying interfaces
      - Justification for manual vs. programmatic access
  - **Data Pipeline**
    - Description of the needs, approach, and data access and refresh frequency
    - Logical diagram showing major data pipeline components for data sources and sinks
  - **Set up for your data environment**
    - Cloud vs. local, database vs. flat files, etc.
  - **Bullets for each team member's individual contributions in Step 2**
  - **Any major updates to Step 1 as a result of data pipeline step**
- Keep it to 5-8 pages, 12 pt font
- Due date: 2/4/22 9am Pacific

Next Presentation: Project Pitch (2/5/22)

- **Audience:** Board Members
- **Main points** to be made
  - Problem definition
  - The opportunity for improvements: define success
  - Proposed solution/approach
  - Data sources, assessment and strategy
  - Feasibility and resources needed
  - What are the risks and how will you mitigate from those?

- **Don't forget** to introduce your team in the beginning of the presentation. Think story lines in the captions!

You will have more time to iterate. We are doing the first draft. Next, are some reminders on what to start focusing on.

Focus 1: Acquiring Data

After you are done with this part, you should be able to..

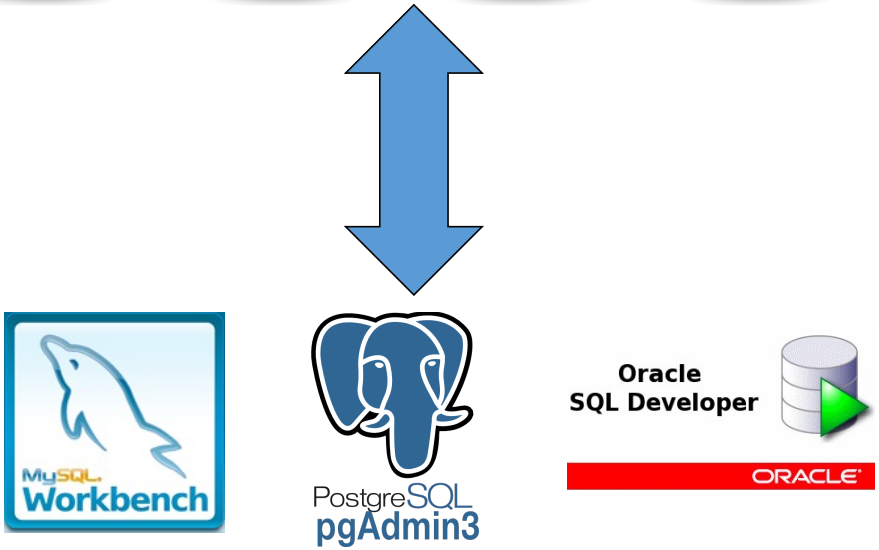
- List techniques and technologies to access and retrieve the data you need
- Describe an example scenario related to your problem accessing your datasets

Data can come from a variety of sources using different technologies

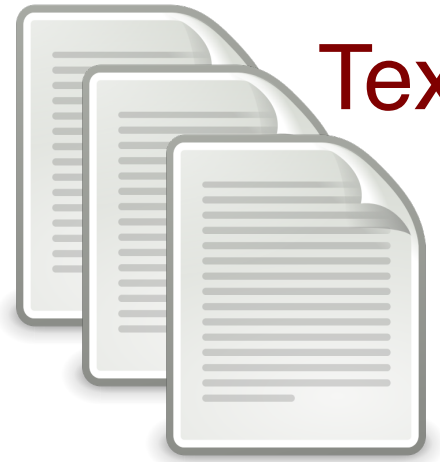
Data comes from many places…

…with many ways to access it.

Traditional databases



SQL and query browsers



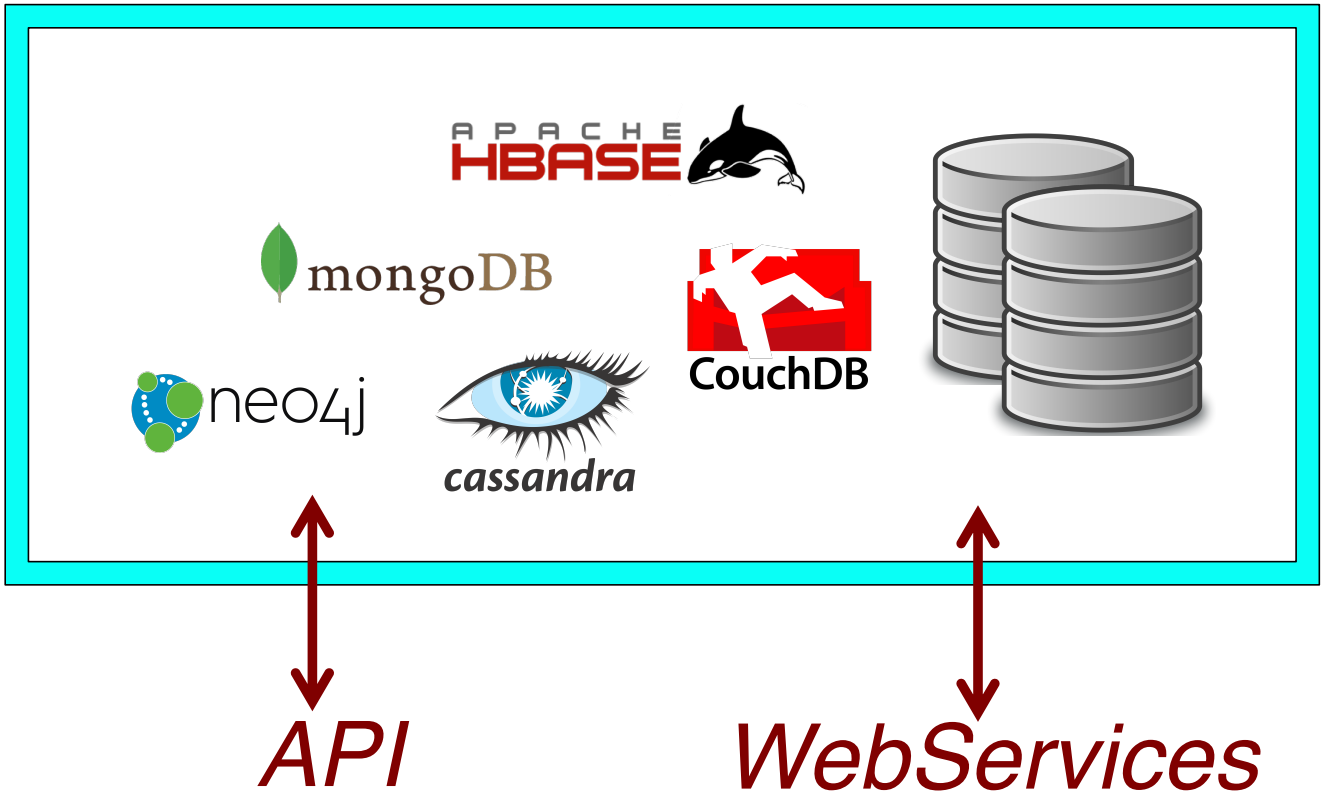
Text files

JavaScript python™ Perl R Ruby php

Scripting languages

Remote data
SOAP
HTML
JSON
XML
REST
RSS
W3C
WebSocket
Web Services

NoSQL storage



Acquiring Data From WIFIRE

Historical weather

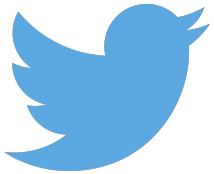
SQL

Current weather

WebSocket

Real-time tweets near fires

REST



Traditional databases

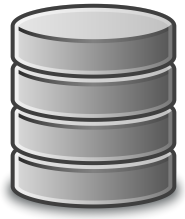
SQL and query browsers

Remote data

Web Services

Text files

Scripting languages



NoSQL storage

Web Services

Programming Interfaces

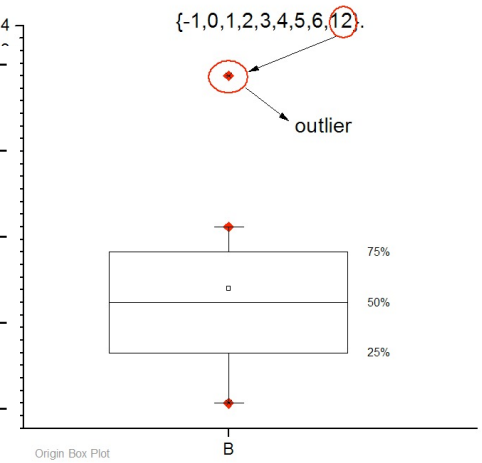
Focus 2: Exploring Data

Why Explore?

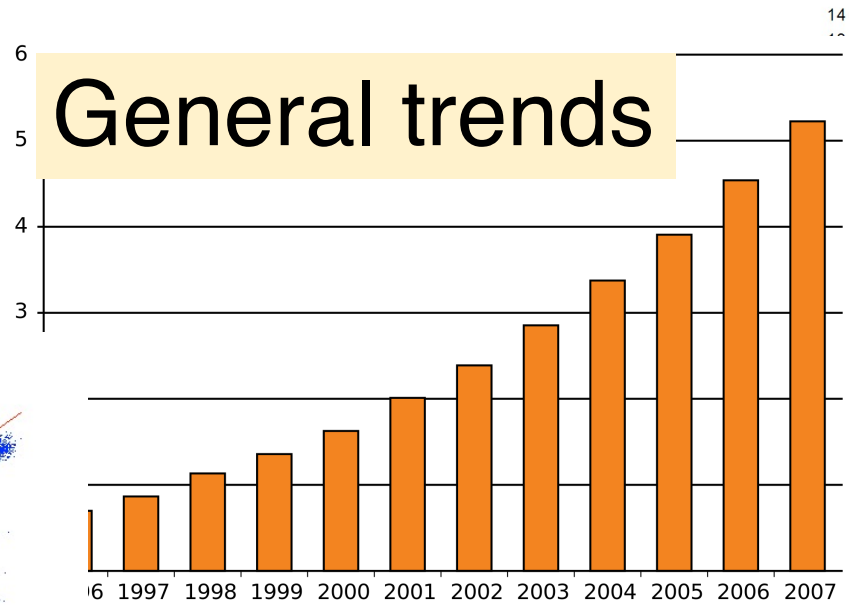
Goal: Understand your data

Why Explore?

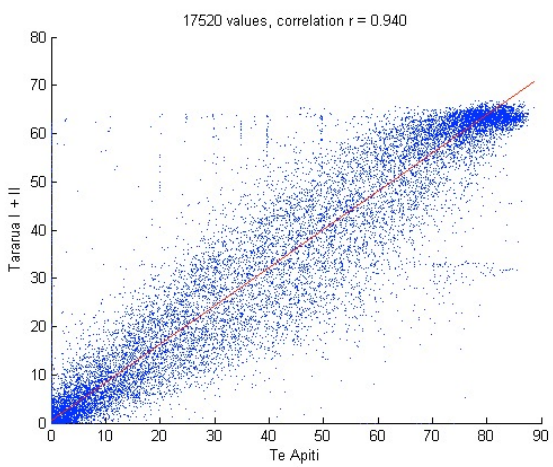
Outliers



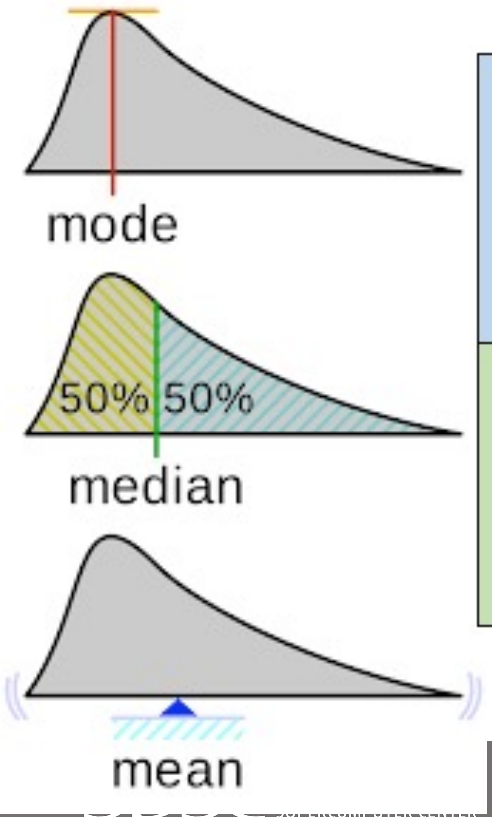
General trends

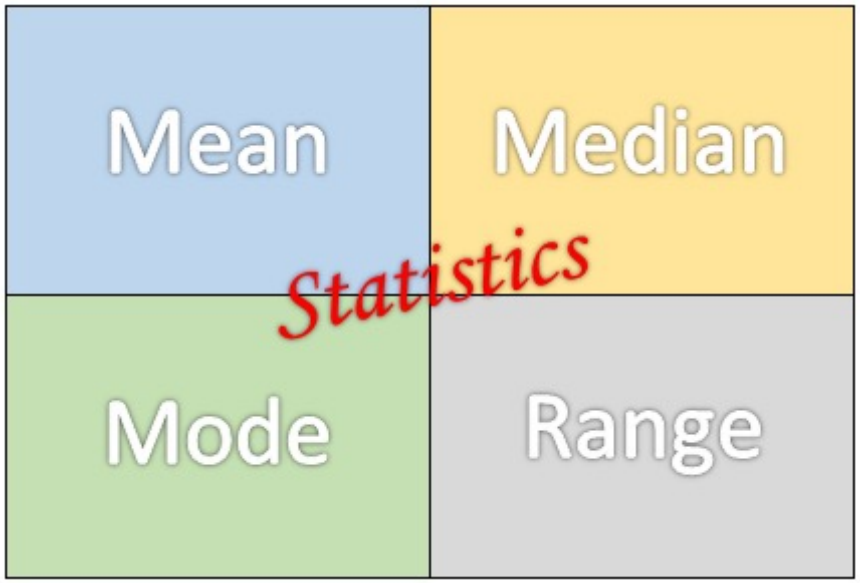


Correlations



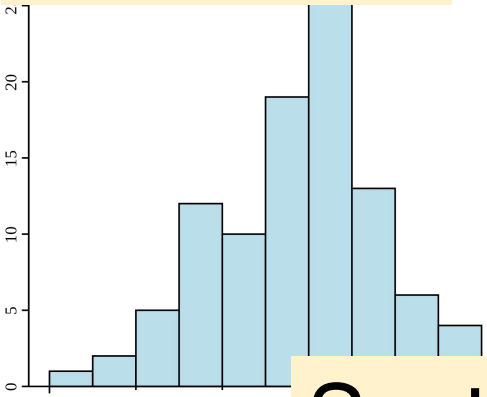
Describe Your Data



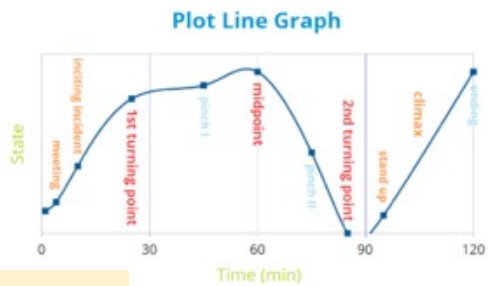


Visualize Your Data

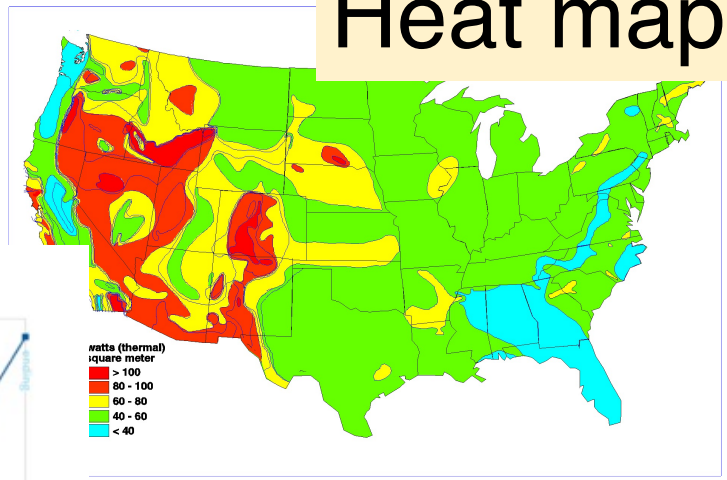
Histogram



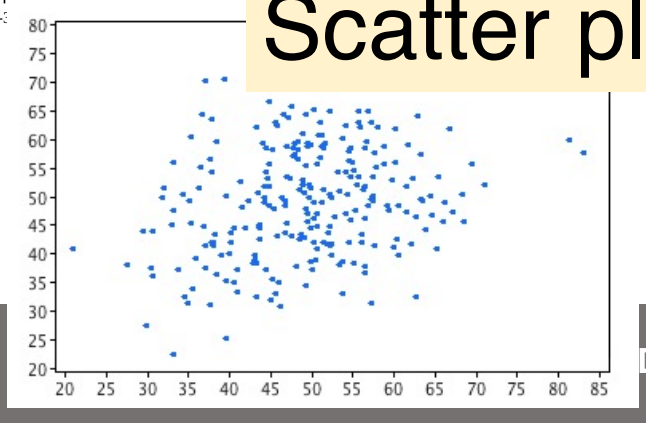
Line graphs



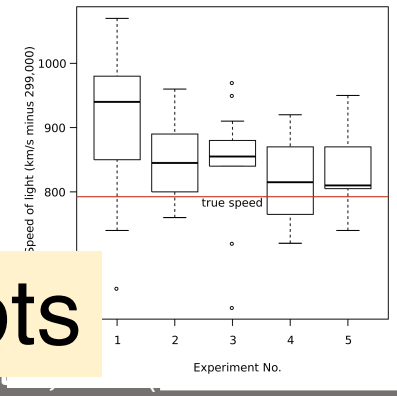
Heat maps

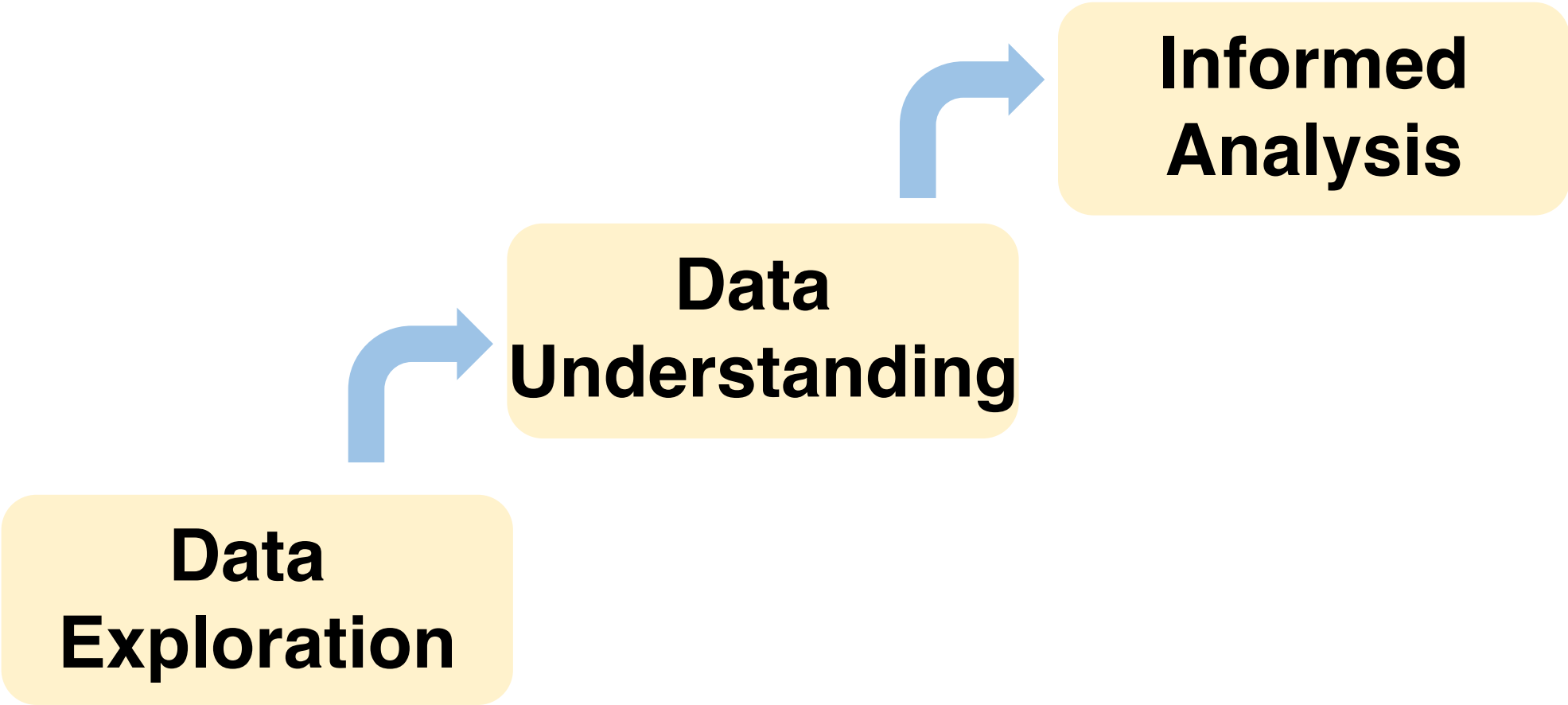


Scatter plots



Boxplots




Informed Analysis
Data Understanding
Data Exploration

Focus 3: Pre-processing Data

Describe what is needed to transform raw data to data that can be used for analysis.

Clean + Transform

Real-world data is messy!

- Inconsistent values
- Duplicate records
- Missing values
- Invalid data
- Outliers

Data Quality Issues

Addressing Data Quality Issues

Remove data with missing values

Generate best estimate for invalid values

Merge duplicate records

Remove outliers

Addressing Data Quality Issues

Remove data with missing values

Generate best estimate for invalid values

Merge duplicate records

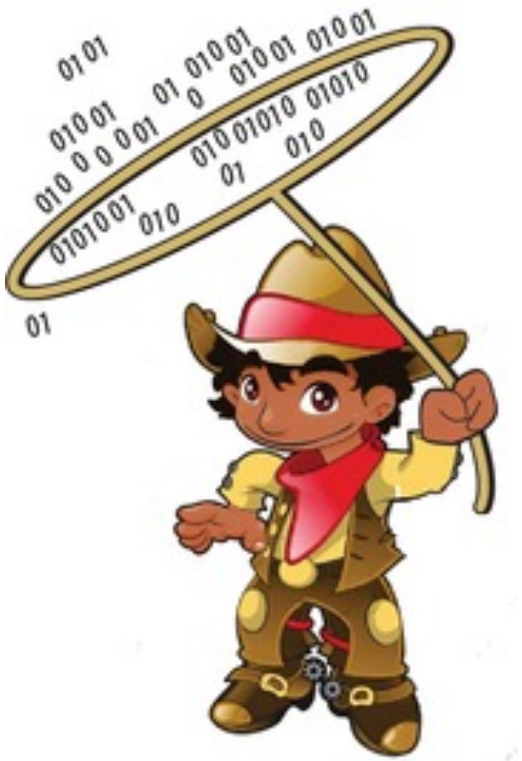
Remove outliers

Domain Knowledge

Getting Data in Shape

Data Munging

Data Preprocessing



Data Wrangling

Data Munging

Data Manipulation

Dimensionality Reduction

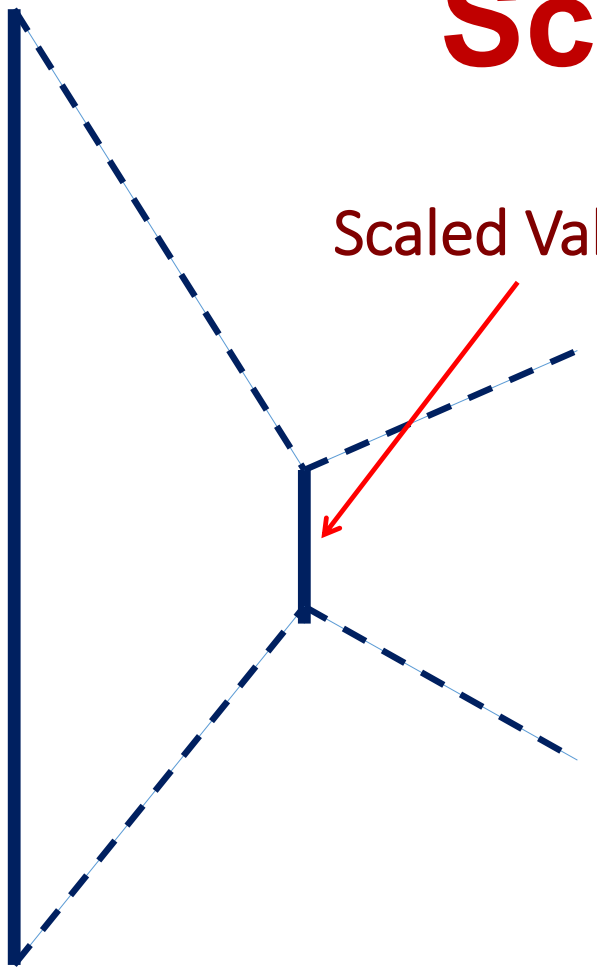
Transformation

Feature Selection

Scaling

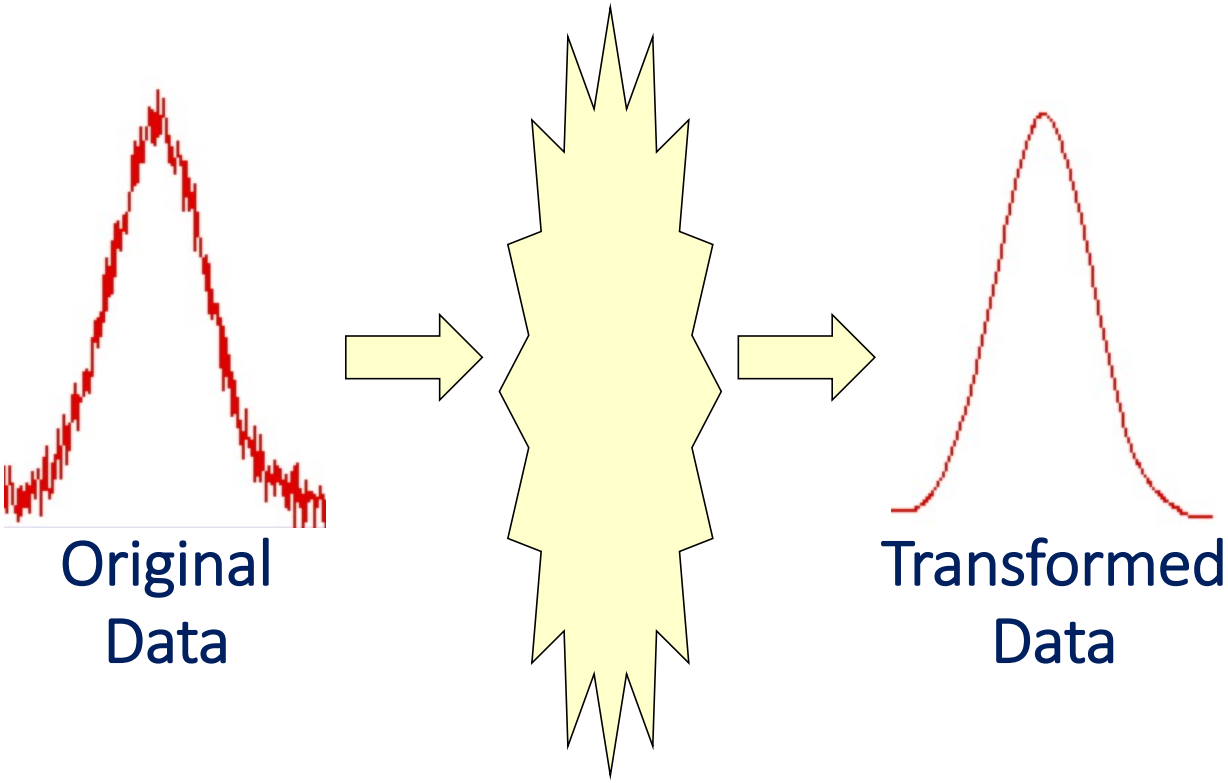
Scaling

Weight

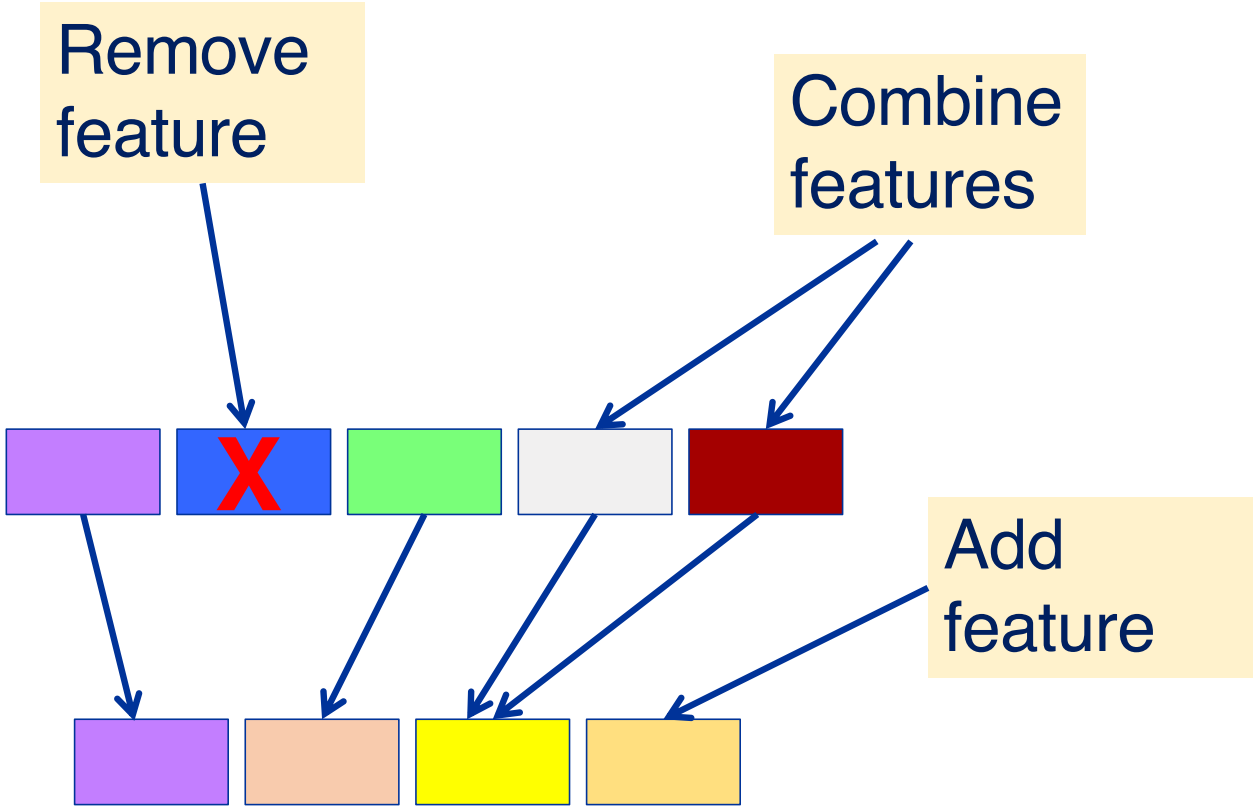


Height

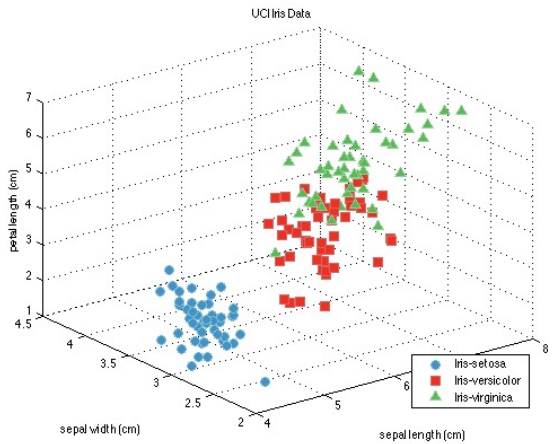
Transformation



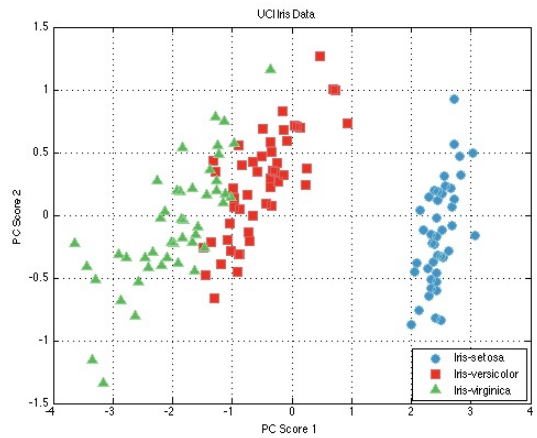
Feature Selection



Dimensionality Reduction

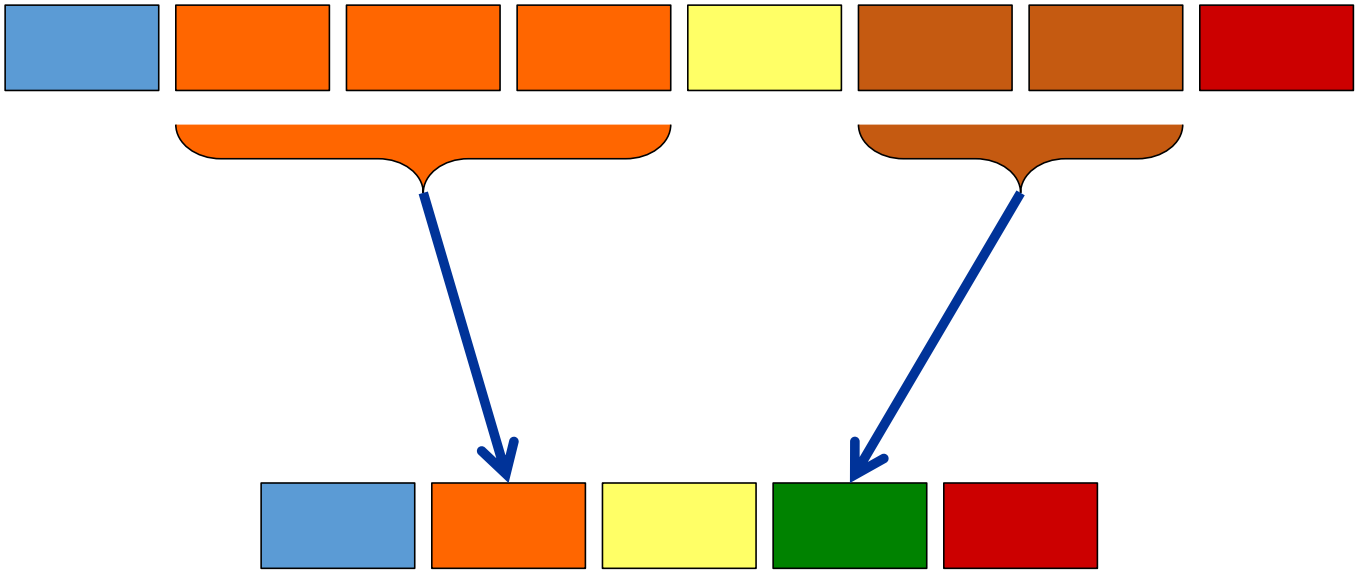


3D



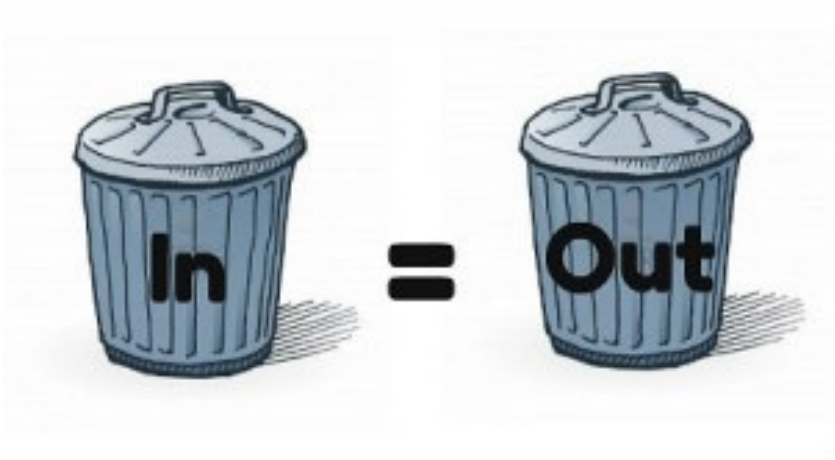
2D

Data Manipulation



Always Remember!

Garbage in = Garbage out



Data preparation is very important for meaningful analysis!

Focus 4: Designing Your Data Pipelines

Draft solution architecture to turn data operations into an ongoing and accountable process.

Solution Architecture Diagram

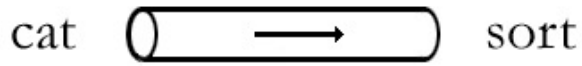
Step by step flow diagram

Environment

Consider…

- Batch vs. streaming or both
- Local vs. cloud or both
- Database vs. object storage or both
- Manual vs. automated
- Refresh frequency
- Data provenance

A UNIX PIPE

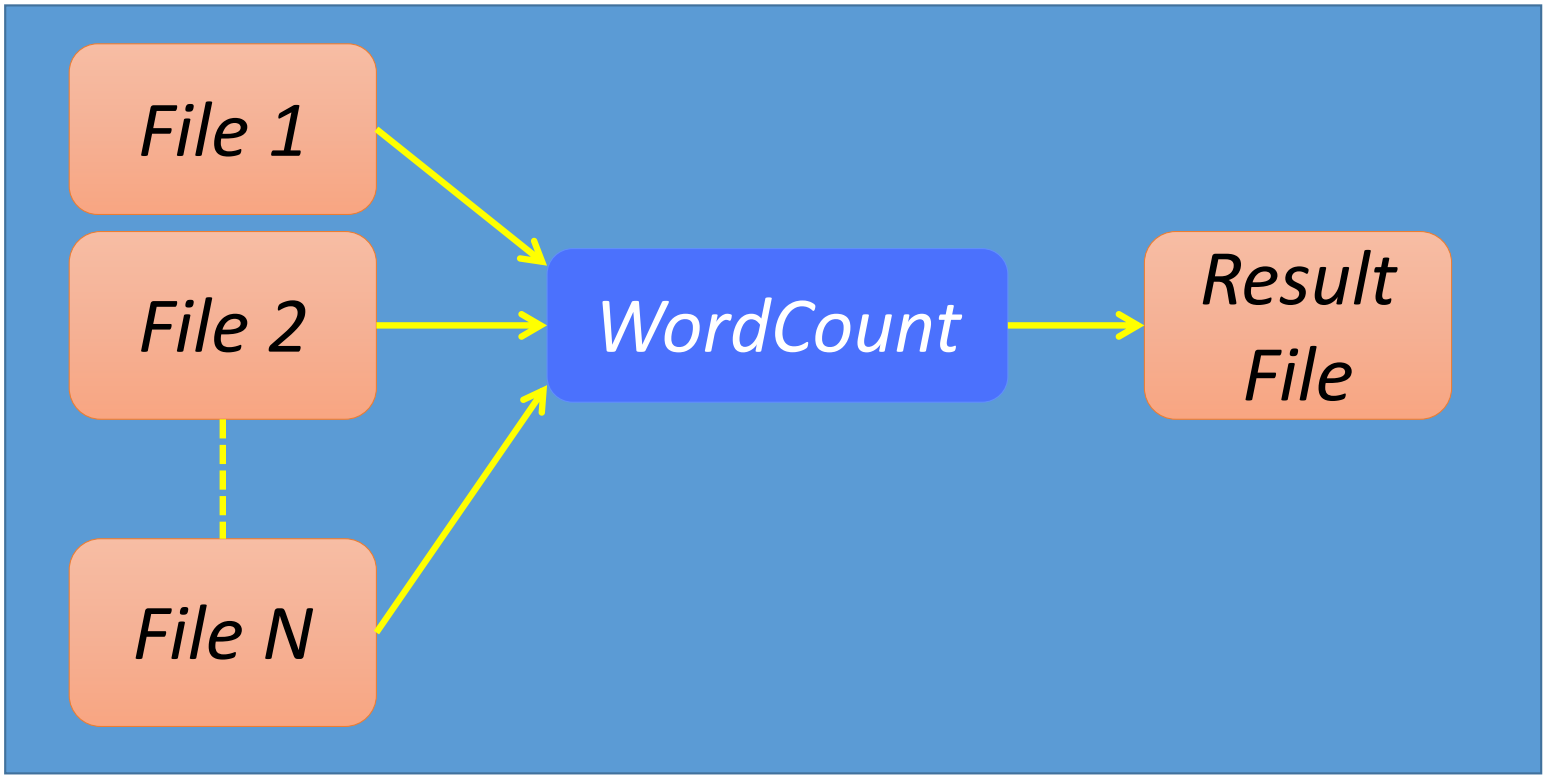


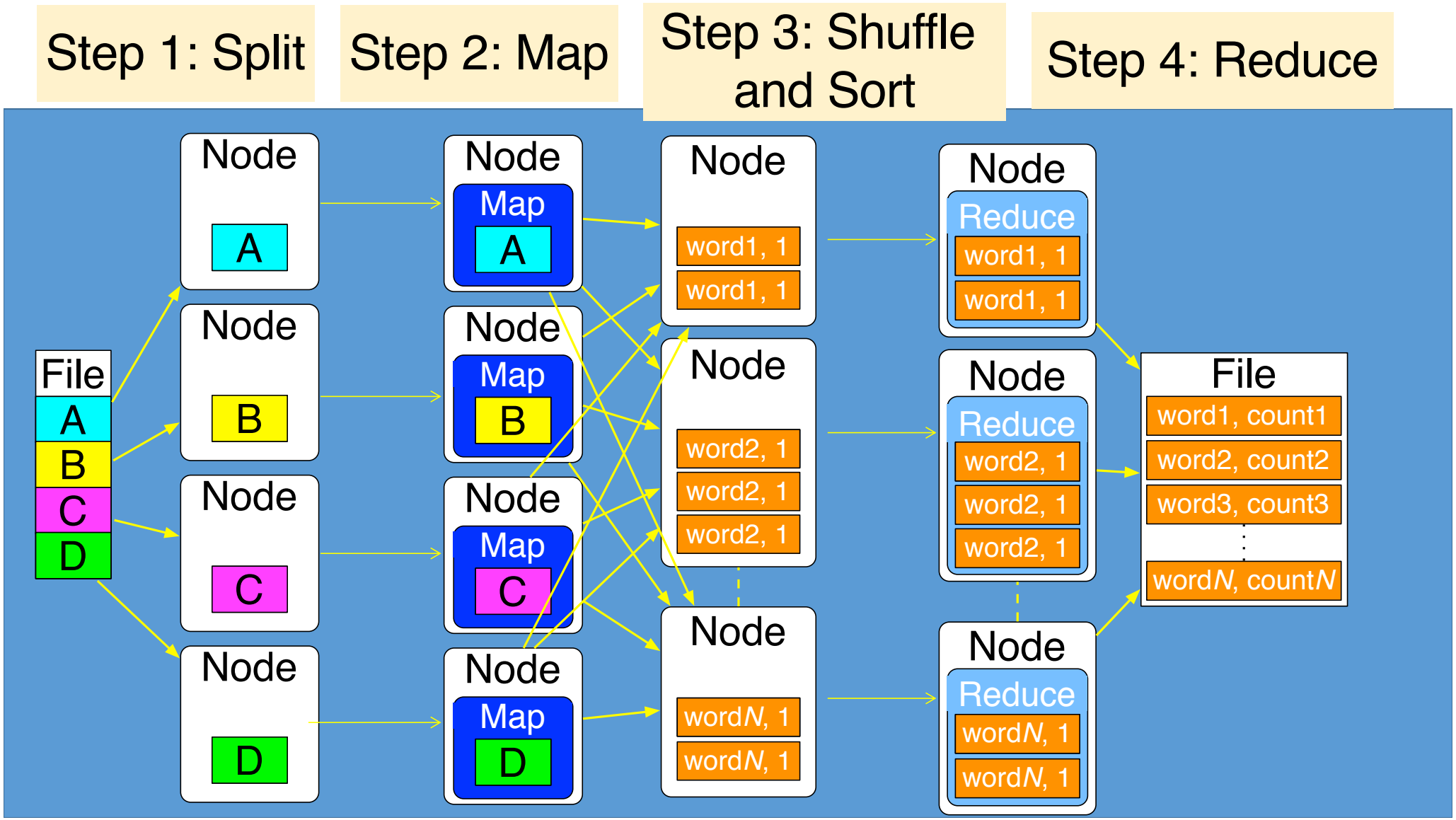
A UNIX pipe provides one-way communication between two processes on the same computer

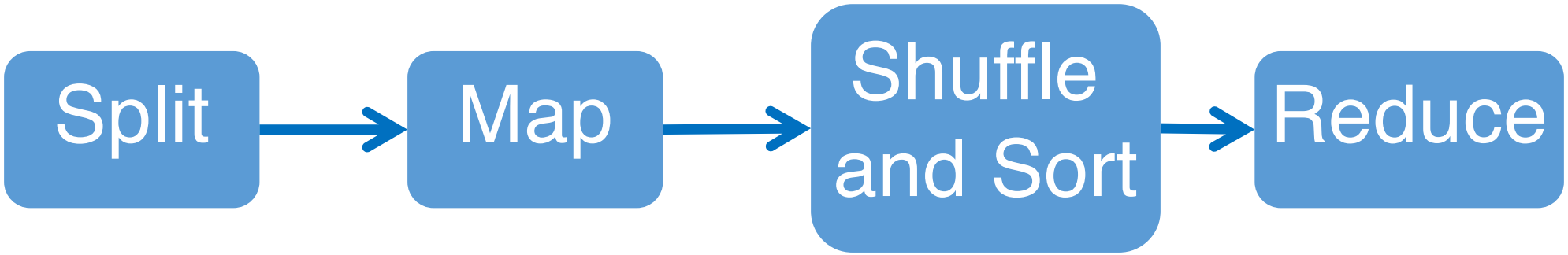
Batch Example Data Pipeline

Represents a large number of applications.

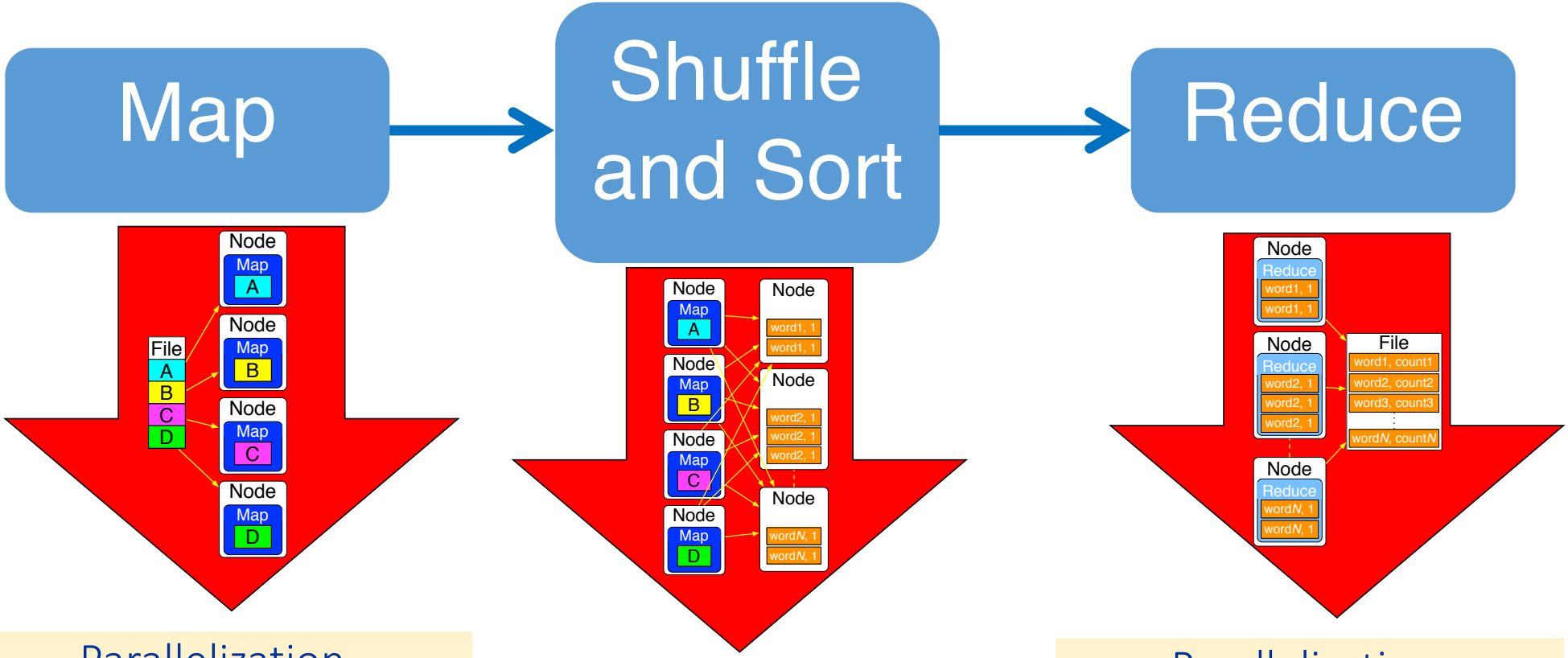
Example MapReduce Application: WordCount




Step 1: Split
Step 2: Map
Step 3: Shuffle and Sort
Step 4: Reduce
File
A
B
C
D
Node
A
Node
B
Node
C
Node
D
Node
Map
A
Node
Map
B
Node
Map
C
Node
Map
D
Node
word1, 1
word1, 1
Node
word2, 1
word2, 1
word2, 1
Node
wordN, 1
wordN, 1
Node
Reduce
word1, 1
word1, 1
Node
Reduce
word2, 1
word2, 1
word2, 1
Node
Reduce
wordN, 1
wordN, 1
File
word1, count1
word2, count2
word3, count3
wordN, countN


Split
Map
Shuffle and Sort
Reduce

Map
Shuffle and Sort
Reduce



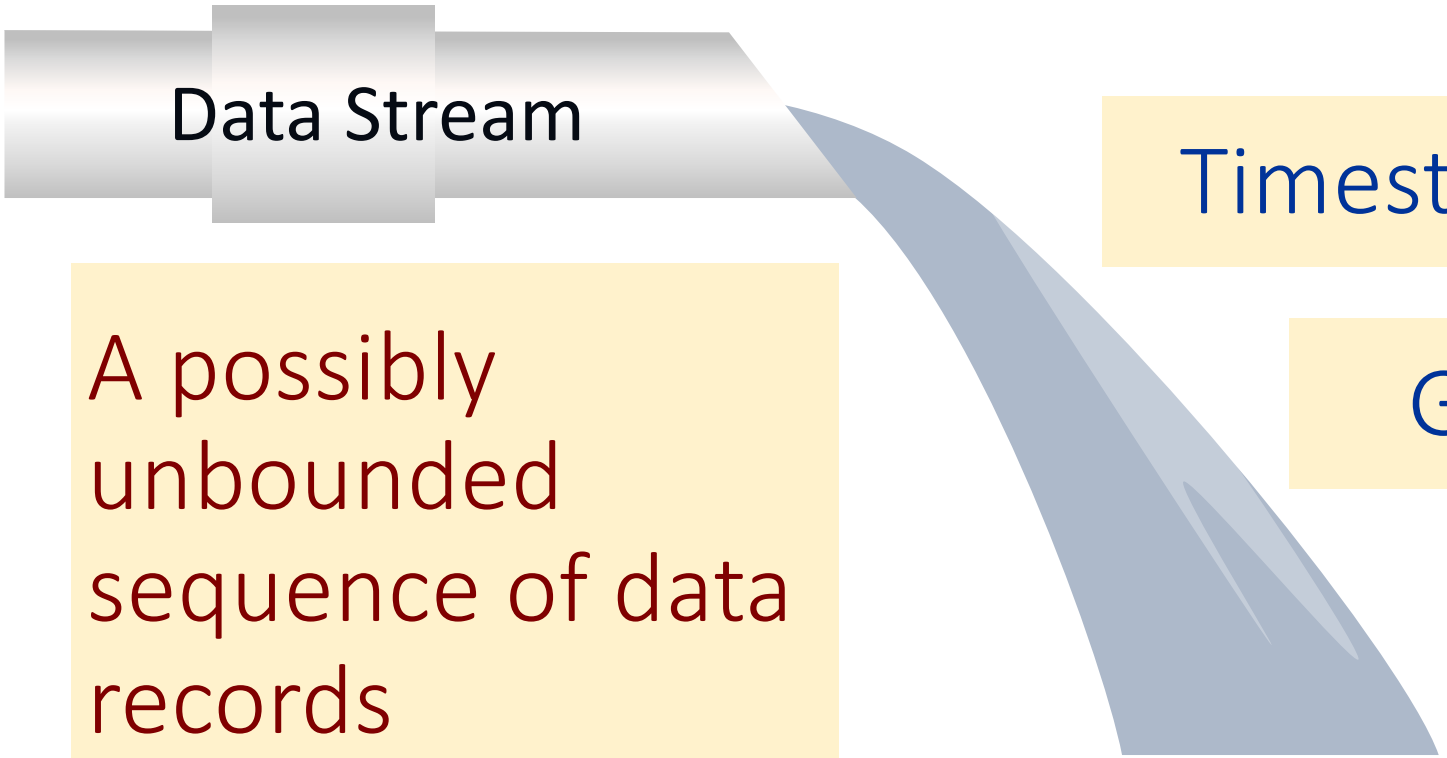
Parallelization over the input

Parallelization over intermediate data

Parallelization over data groups

Streaming Data

What is a Data Stream?



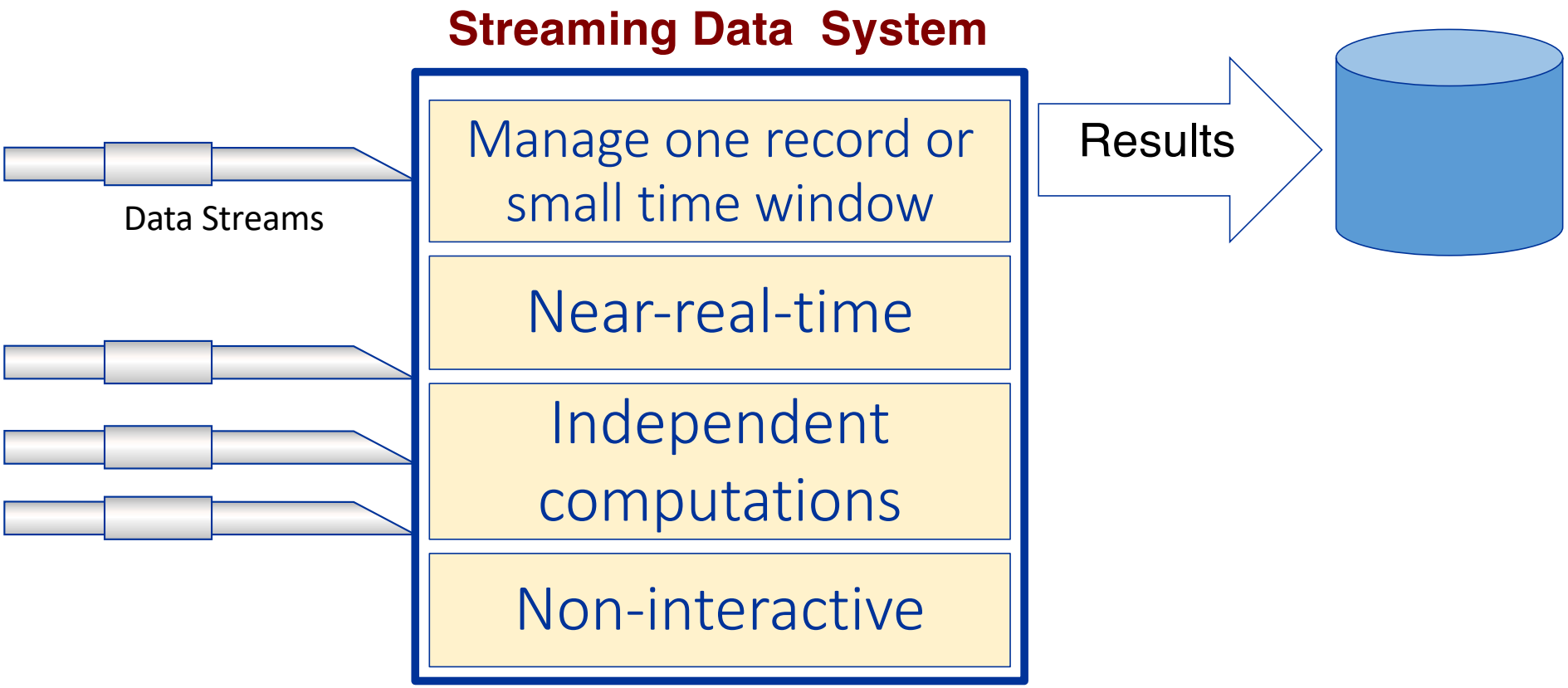
Timestamped

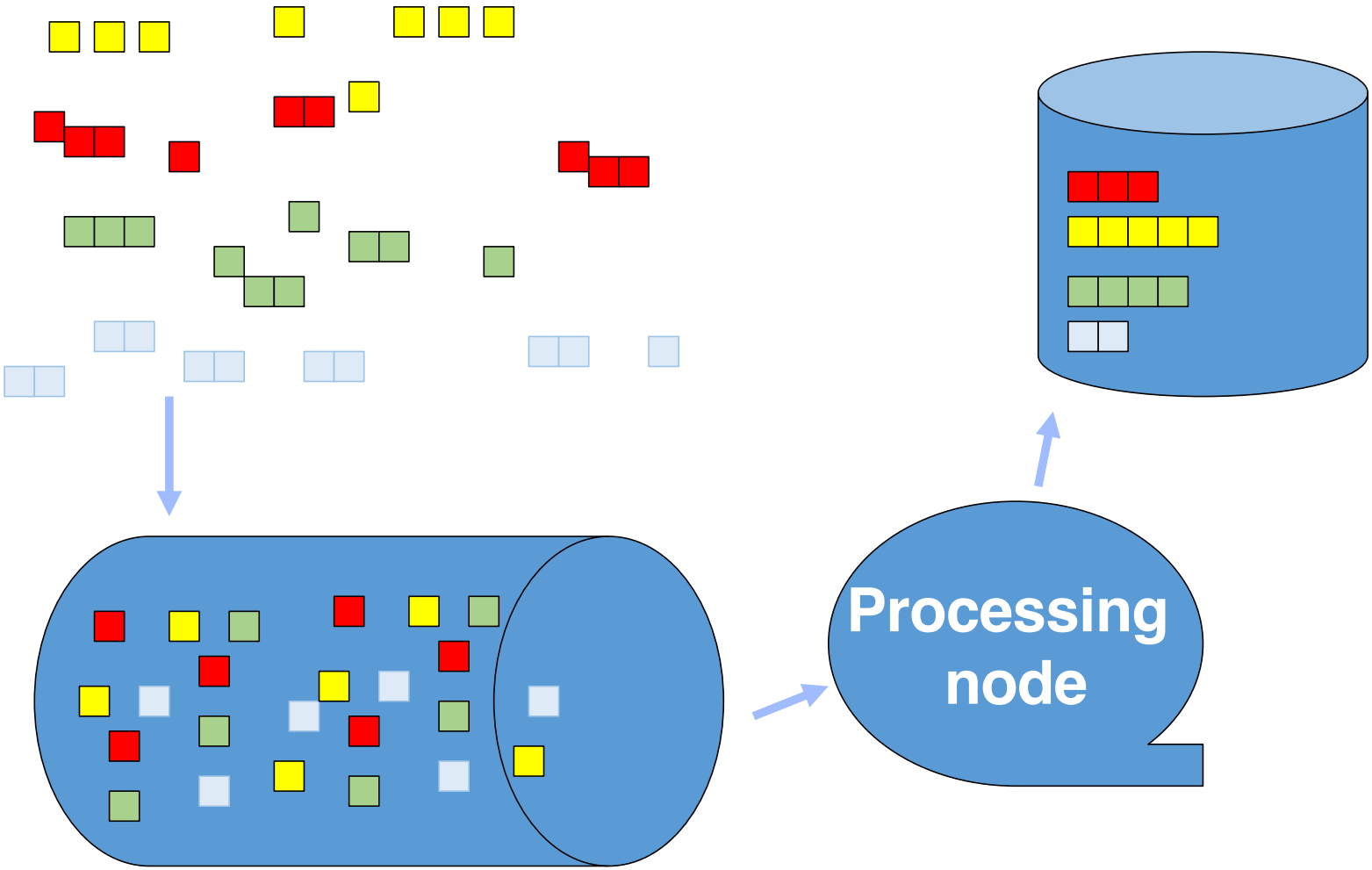
Geo-tagged

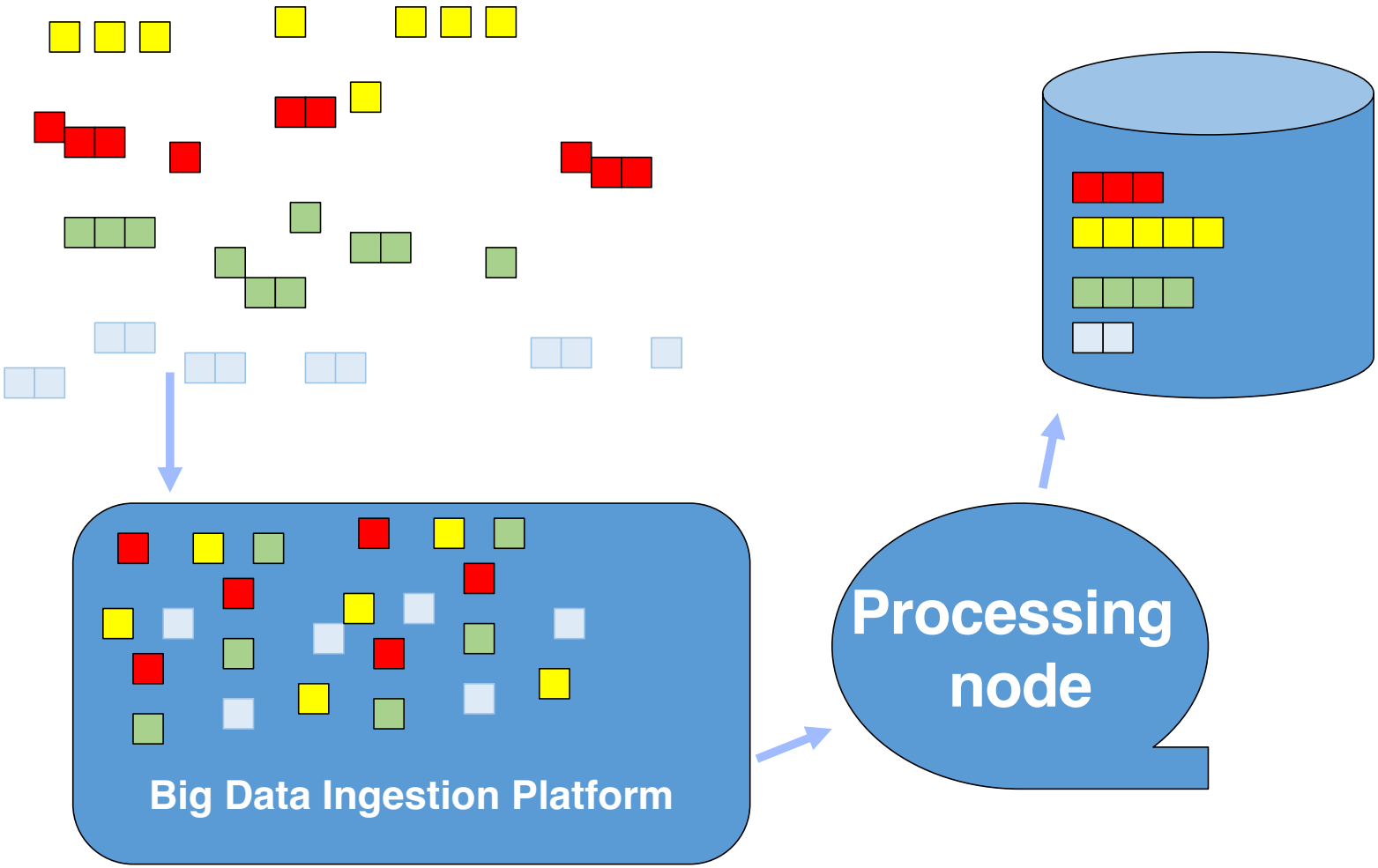
A possibly unbounded sequence of data records

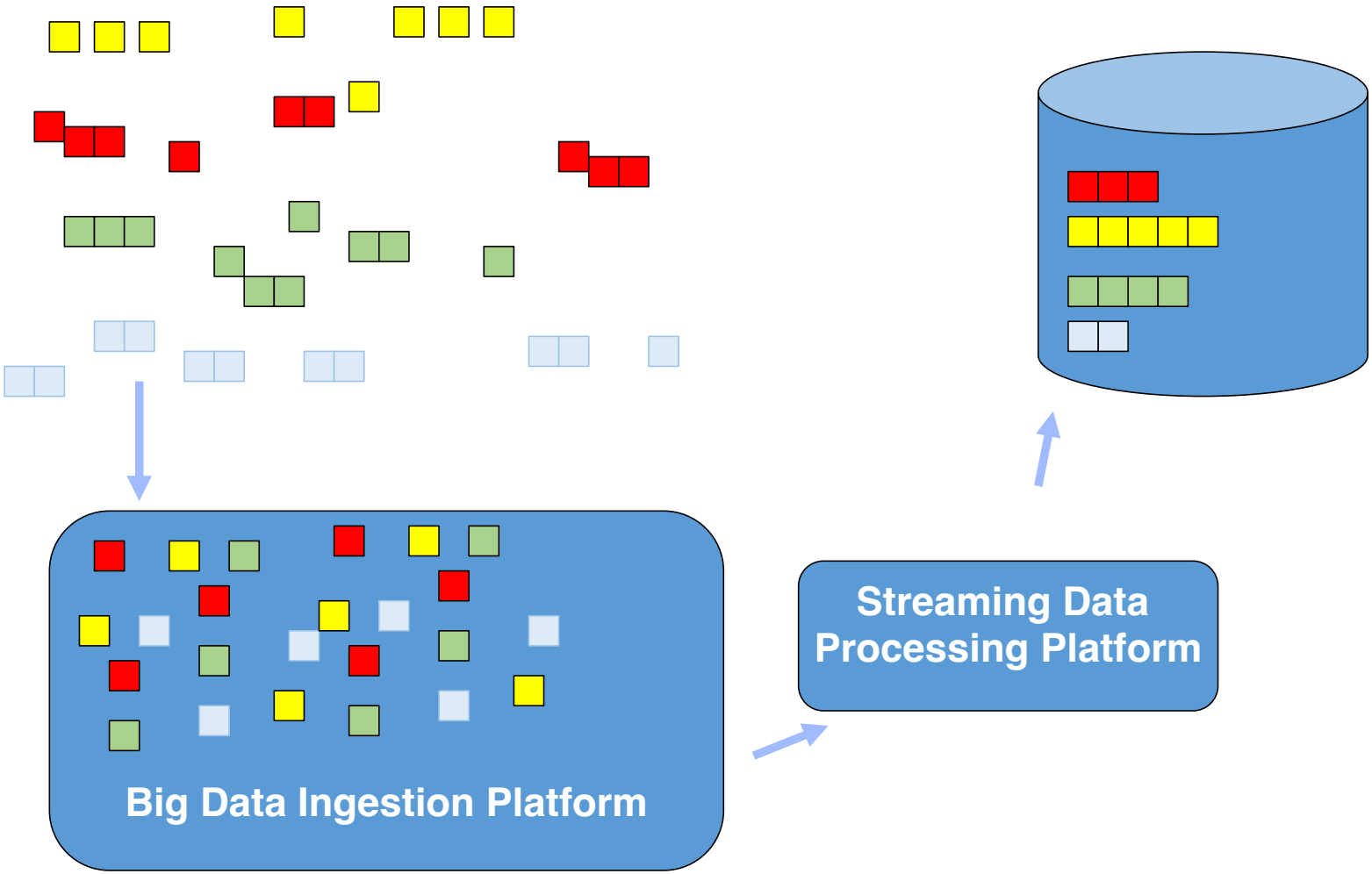
Synchronized sequence of events

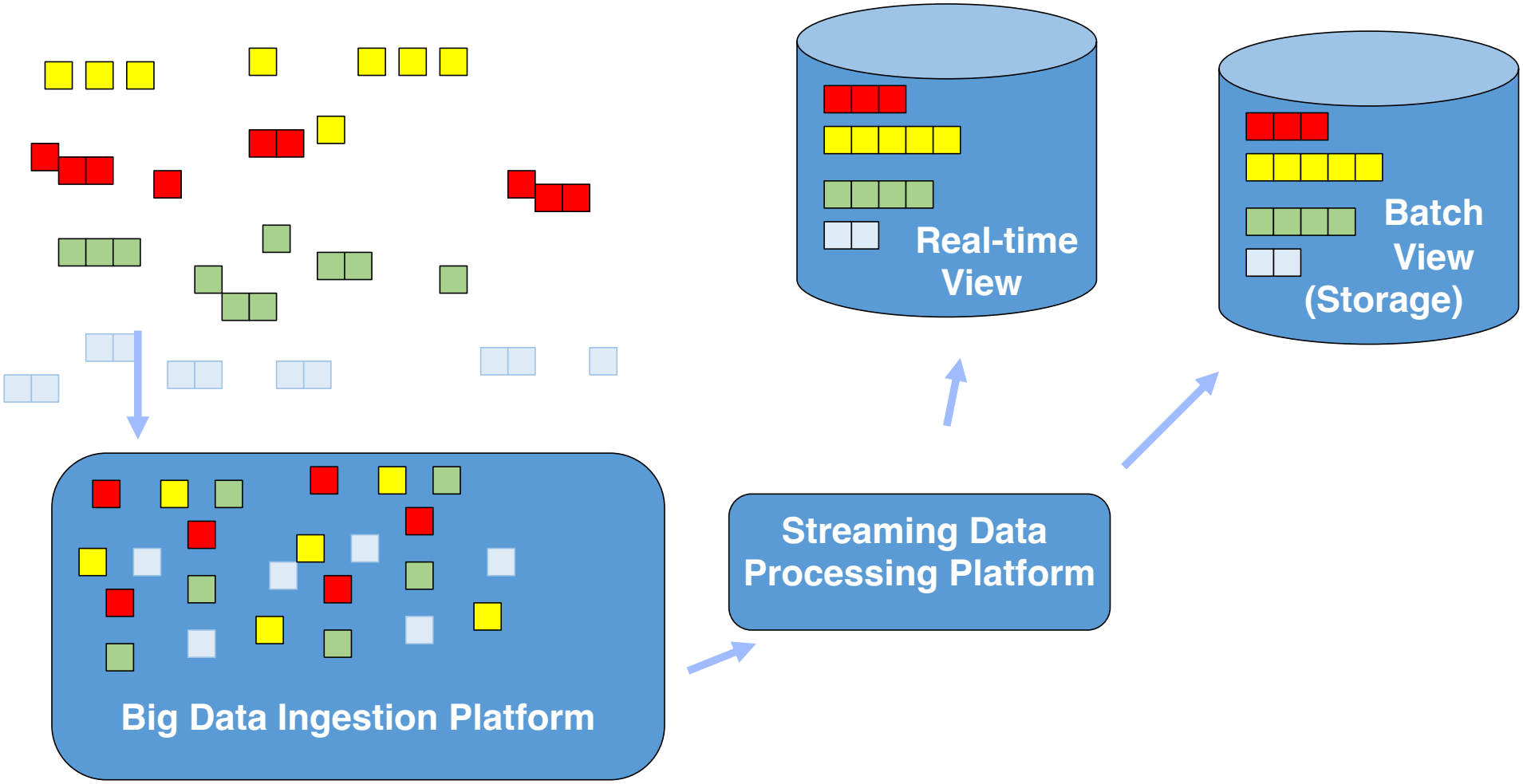
Streaming Data Systems




Processing node

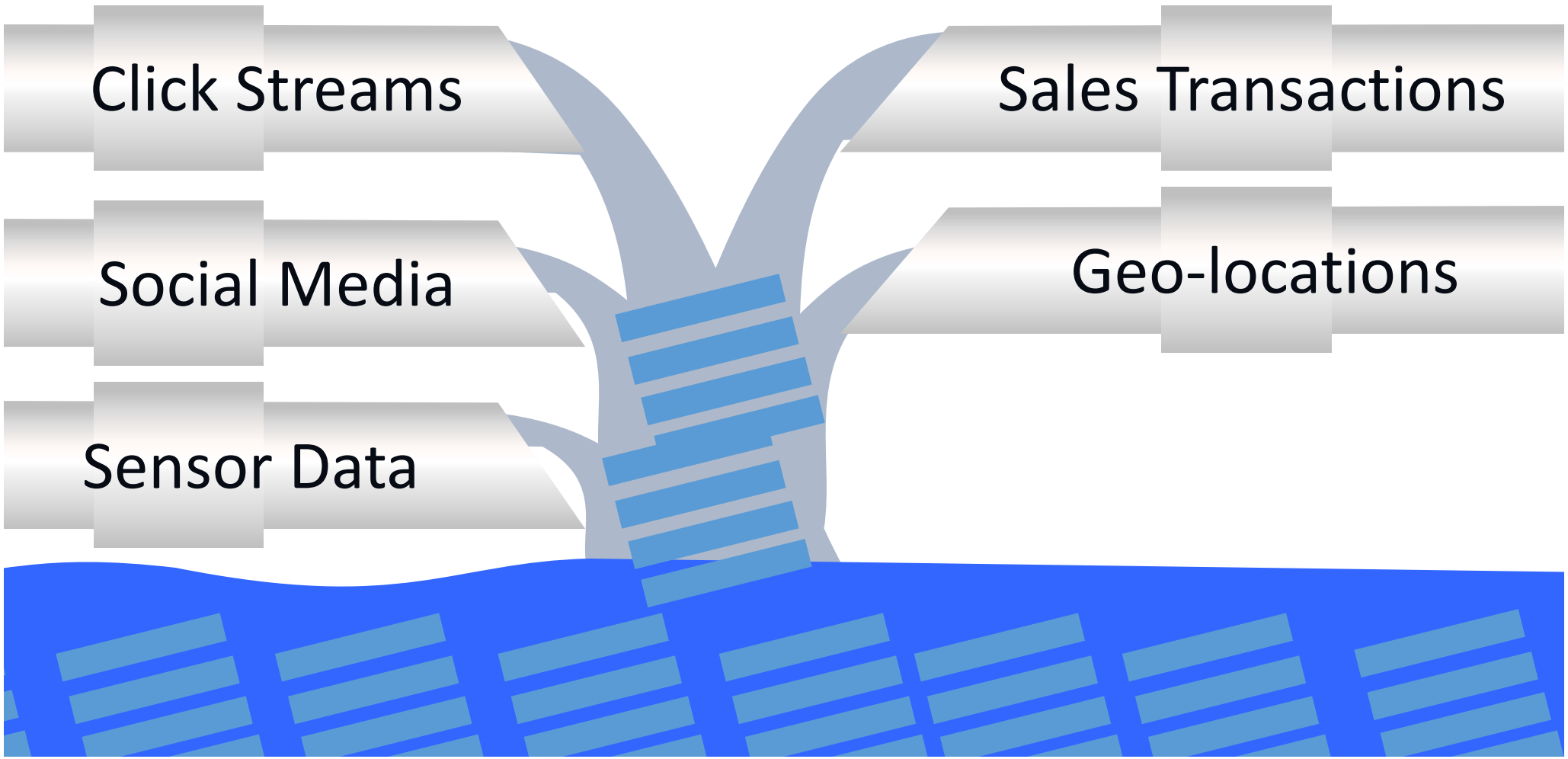

Big Data Ingestion Platform
Processing node


Big Data Ingestion Platform
Streaming Data Processing Platform


Real-time View
Batch View (Storage)
Big Data Ingestion Platform
Streaming Data Processing Platform

Databases vs. Data Lakes

Short definition of a data lake



“ If you think of a datamart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples. ”

James Dixon
CTO, Pentaho Corporation

Load data from source

Store raw data

Add data model on read

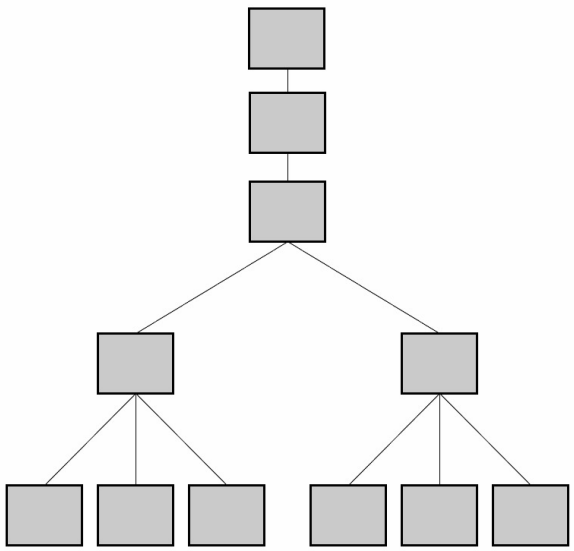
Data Lake

- Schema-on-read

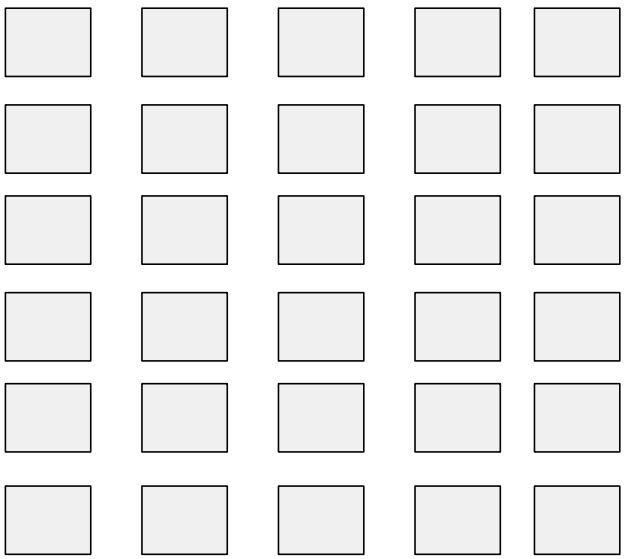
Data Warehouse

- Schema-on-write
- Transform and structure before load

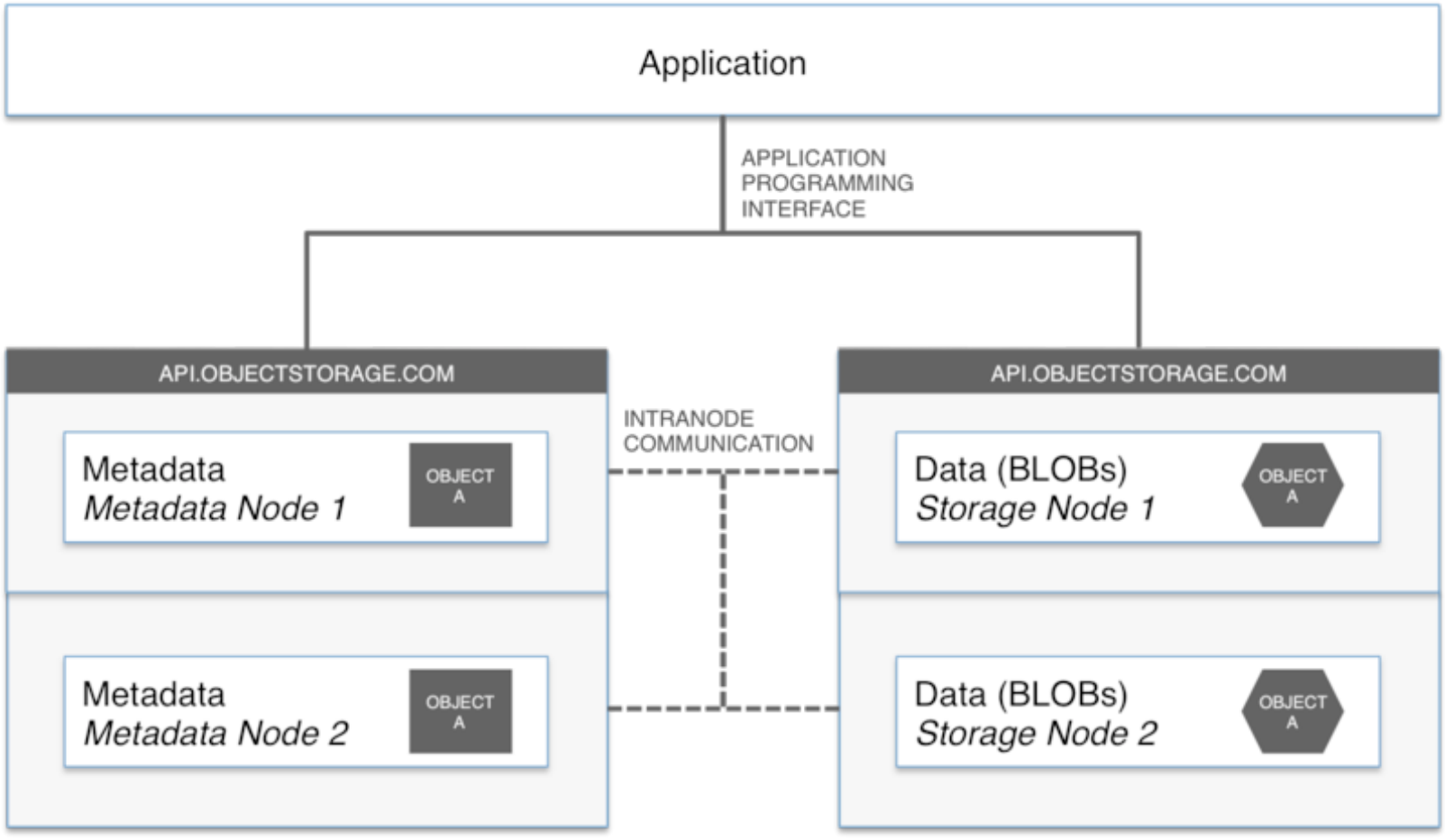
Hierarchical File System



Object storage



Object Storage



Questions?

ILkay Altintas, Ph.D.
Email: ialtintas@ucsd.edu