MSc - Data Mining

Topic 02 : Data Handling

Part 01 : Data-Storage

Dr Bernard Butler and Dr Kieran Murphy

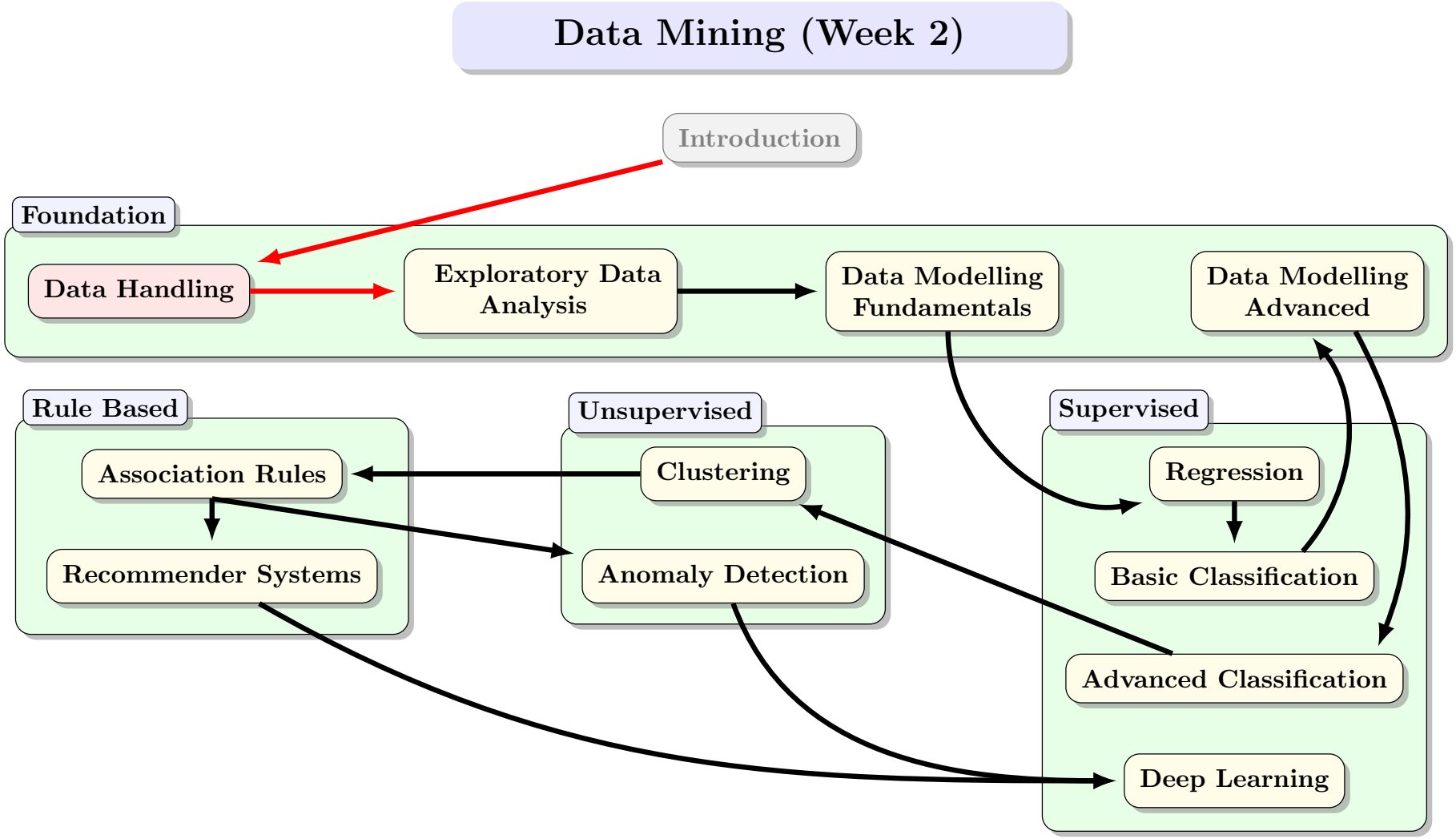
Department of Computing and Mathematics, WIT.
(bbutler@tssg.org and kmurphy@wit.ie)

Spring Semester, 2021

Outline

- Types of storage model
- Storage operations

Where we are in the module



Review of Week 1

- The Pre-Start lab provided an opportunity to look at the tips dataset, and to use your existing knowledge and skills to investigate how tips are issued to servers
- The module introduction provided an overview of the module as a whole, outlining the history of data mining, process models, the machine learning topics we cover and how the module is delivered and assessed
- The statistical review provided a refresher on the statistical reasoning (e.g., where is a data set) that is essential for this module, and the probability that enables us to make inferences
- The optimisation overview features gradient descent, because this optimisation technique is heavily used across machine learning techniques
- The pandas overview features very helpful advice from an expert pandas user (Kieran!) which will prove invaluable to you in programming assessments
- Today we continue this introduction as we cover foundational aspects, this time regarding *Data Engineering*: how data is stored and processed at scale
- In later weeks, our attention switches to *machine learning*, which forms the bulk of this module

Overview of today's lecture

Data Infrastructures

- Data infrastructures form the foundation of data mining
- Data does not exist in a vacuum
  - it is stored somewhere, if only temporarily
  - it moves somewhere else (for a variety of reasons)
- Data can be represented in many formats:
  - basic, with separate metadata: CSV, TSV, COBOL-based, ...
  - complex, with metadata included: XML, JSON, YAML, Avro, ...
- Data can be "text" or binary-valued
- Data operations can be represented as CRUD or HTTP verbs

Why is data stored?

There are two basic use cases for data storage

Persistence

Within an application, we often need to save data temporarily when handing it over to a separate entity for further processing. Often, the data is saved from/loaded into memory to make recovery easier. Generally, data persisted like this has little visibility or use for data mining purposes.

Archive

Between applications, data can be saved to more permanent storage as a way of keeping a record of key results, e.g., bank transactions, for audit, BI or similar purposes. Generally, data stored in this way *is useful* for data mining purposes.

Note that data mining workflows are themselves applications! However, the need, or lack of such, for persistence is a defining characteristic of various DM platforms and/or algorithms.

Databases and Database Management Systems: A review

Definition 1

A database is a collection of data that is organised to facilitate data management operations, especially *Reading*, *Writing*, *Updating* and *Deleting*.

Definition 2

A DBMS is *software* that manages a database, and enables users to apply data management operations, to control access to the data and to configure how it is stored.

DBMS have a long and rich history, with many interesting aspects that are out of scope here. For data mining purposes, the main interest is in the data model, format and access method.

The Data Model

Definition 3

A data model is an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world entities they represent.

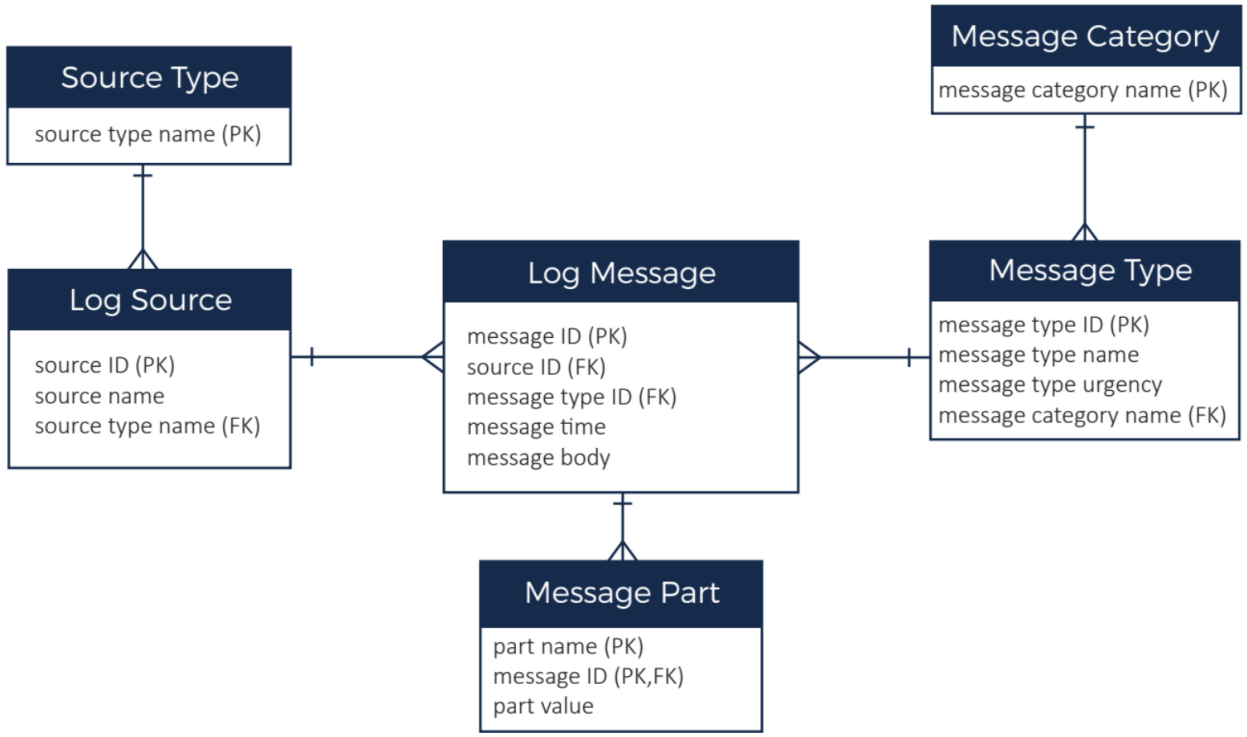
Data models come in a multitude of forms, but a common classification is into *relational* and *non-relational*.
They represent a trade-off between competing objectives, namely the model's alignment with how the data is

- structured for data input (Writes and/or Updates)
- structured for data output (Reads)
- viewed by its owners/users (stakeholders have their own perspectives)
- to be placed in the physical storage infrastructure, which is often distributed/replicated for scalability and/or resilience

The Relational Model

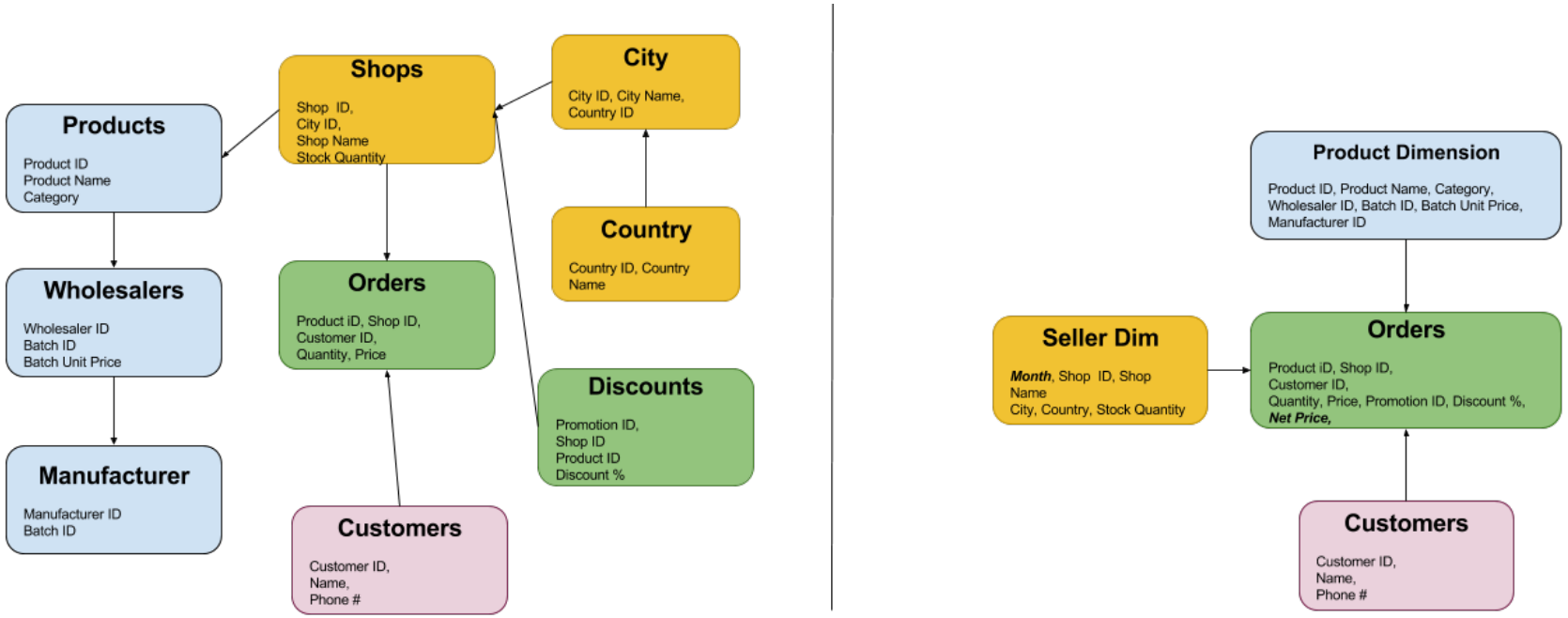
The diagram says that entities like `Message Type` and `Log Message` are linked by a Primary Key-to-Foreign Key relationship in the sense that each `Log Message` has a single `Message Type`, but the same `Message Type` can be assigned to many `Log Message`s.

The relational model is well-established, commonly used and (in normalised form) is a good fit to many OLTP workloads. For OLAP and/or mainly Read operations, a normalised model is more convenient.



Source: https://www.instaclustr.com/cassandra-nosql-data-model-design-2/

Data normalisation

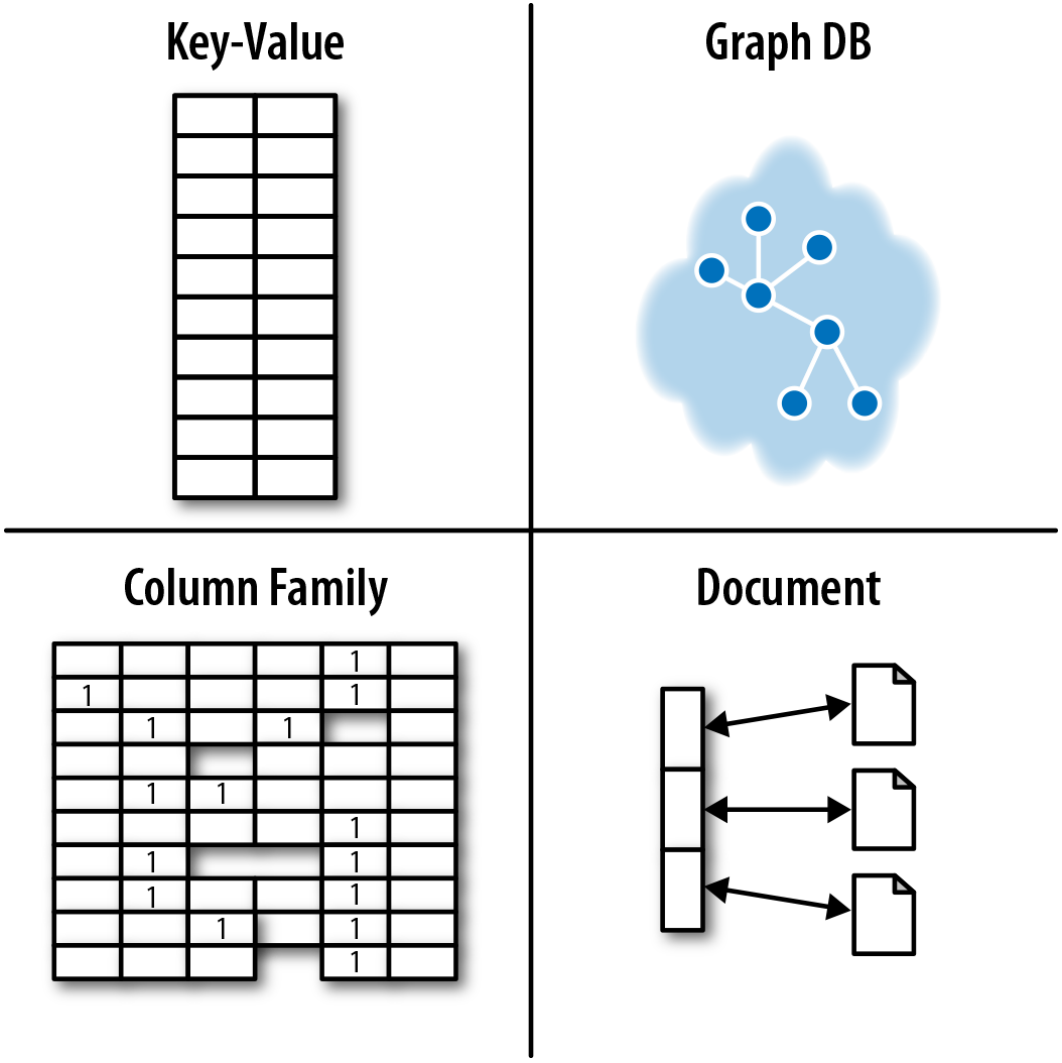


Source: https://www.techinasia.com/4-ways-fix-slow-dashboards

Normalised data is broken into smaller pieces, so Create, Update and Delete operations tend to be more specific (smaller) and the data storage requirement is generally smaller too. So (non-concurrent) Write performance can be good, but concurrent writes are slowed by the need for *locks*. By contrast, denormalised data often has redundancy but that can be good (better support for error-checking) and users of the data often prefer to get the data in one piece, rather than having to join it together (more effort and a *leaky abstraction*) at Query (Read) time.

NoSQL database options

SQL is the most common language for accessing relational databases. Non-relational DBMS often use (proprietary) languages based on SQL. NoSQL databases are often designed with high availability and scalability as critical requirements. NoSQL databases are indicated for massive, append-only use cases such as system or user-content interaction logs.



Key-Value databases

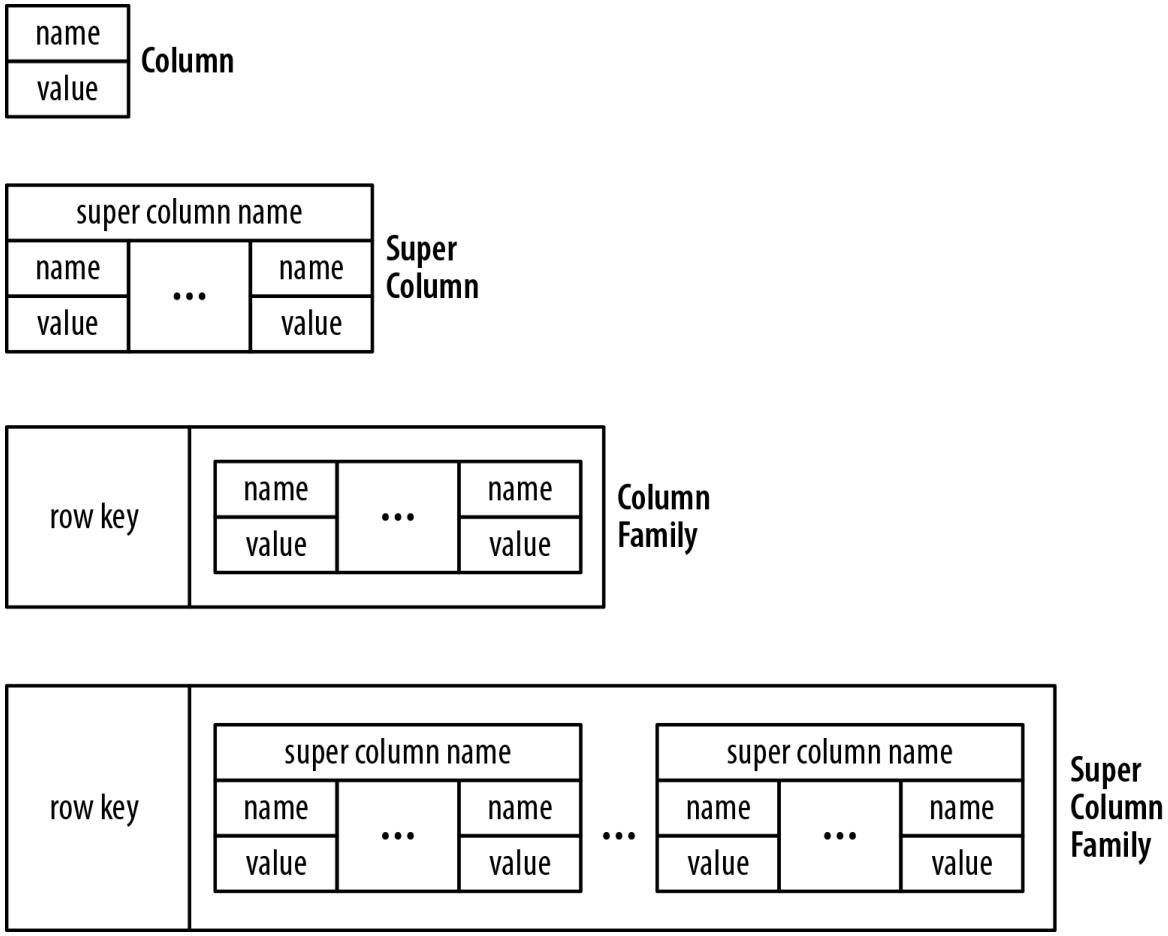
- Key-value data model is very simple.
- Logically it is a distributed hashmap of *aggregate*s (single values or data structures comprising multiple values).
- The type of each aggregate does not need to be the same
- The keys are hashed and mapped onto *buckets* that are distributed across *machines*.
- Examples include Dynamo, Redis, MemcacheDB, etc.

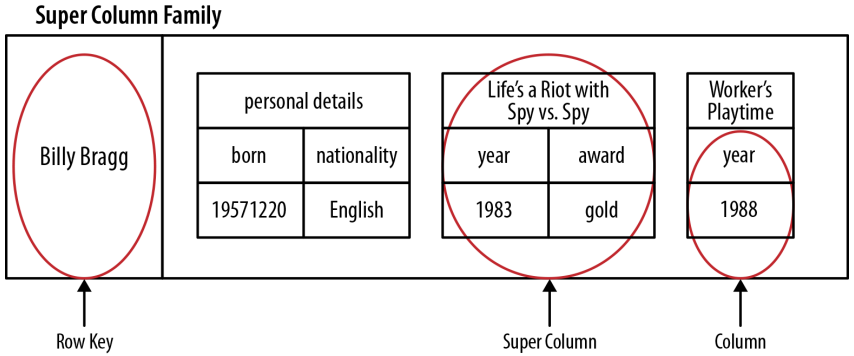
Document databases

Relational	Document-oriented
database	database
table	collection
row	document
table join	embedded/linked documents
rows have same structure	row structure can vary

Document-oriented database is a subclass of key-value stores. The value is no longer *opaque*, but it has structure (often a JSON fragment). Often explicit *sharding* is used to distribute documents. Examples include MongoDB, CouchDB.

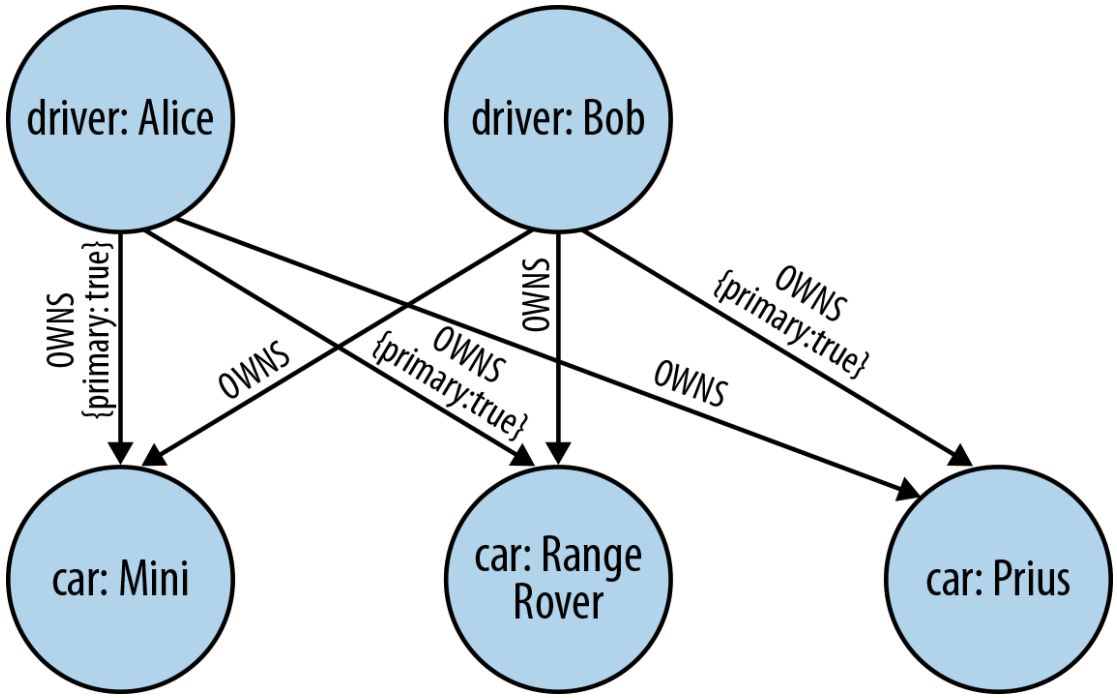
Column Family databases





Logically it is a hashmap of hashmaps (nested key-value pairs, grouped). The inner hashmaps are like rows/documents. There is lots of flexibility, e.g., they handle sparsity well. The model was introduced by Google BigTable, and can be found in Hadoop HBase (realised as partitions containing buckets) and (in modified form) in Cassandra.

Graph databases



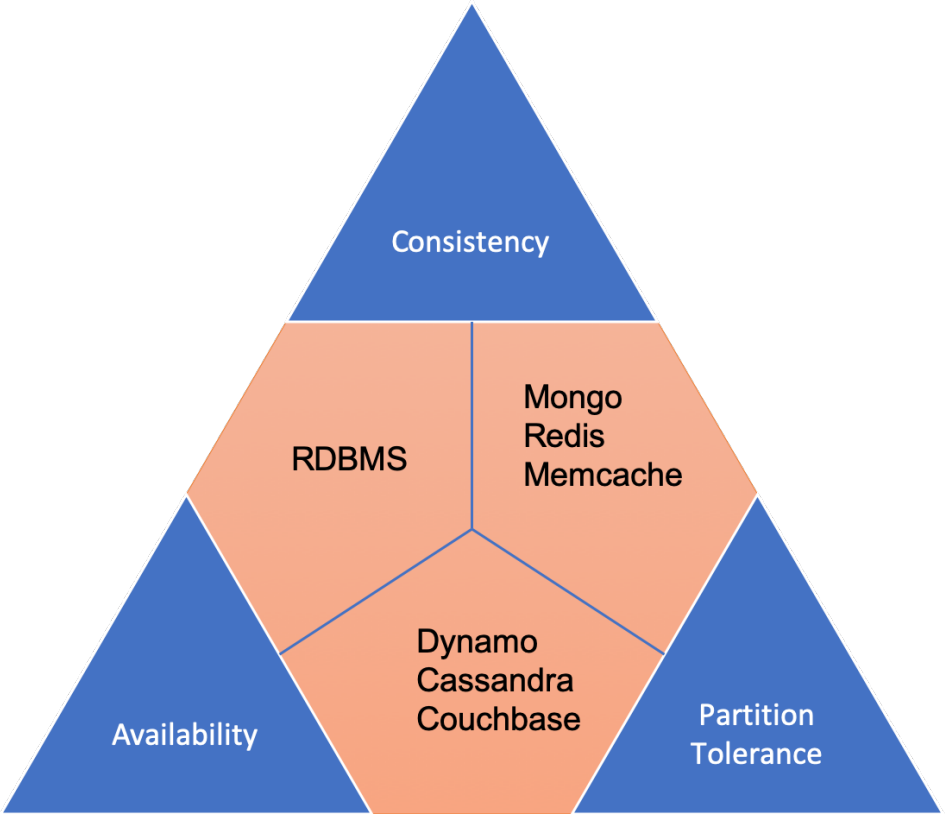
Labeled property graph model shown, e.g., Neo4j, can be implemented with index-free adjacency. Alternative graph model is RDF triple store, for semantic web linked data. Triple stores (e.g., Jena) use extra nodes instead of properties. Languages include cypher, gremlin and SPARQL. Excellent for (social) network applications.

NewSQL databases

Feature	OldSQL	NoSQL	NewSQL
Relational	✔	✘	✔
SQL	✔	✘	✔
ACID transactions	✔	✘	✔
Horizontal scalability (deployment)	✘	✔	✔
Scalability (performance)	✘	✔	✔
Schema-less	✘	✔	✘

NewSQL databases share many of the advantages of "OldSQL" and NoSQL. Three categories exist: optimised storage engines, automatic sharding, and new architectures where distributed computing is central, rather than an add-on. Examples include Google Spanner, MS Azure Cosmos DB, VoltDB, NuoDB.

CAP Theorem



Source: https://www.linkedin.com/pulse/cap-theorem-architectural-tool-vinay-avasthi/

- Conjecture (Brewer, 2000): databases can offer no more than 2 of C,A,P at a time when a failure occurs.
- Thus databases tend to pick 2 of the 3 to guarantee.
- When *partitions* occur, a relational DBMS cannot guarantee C and A simultaneously - it has to address one first.
- By contrast, A CP DBMS like Mongo would have difficulty with Availability when a failure occurs.
- Clever database administration can minimise problem, but the trade-off remains a key factor when choosing DBMS type.

DBMS "Chemistry": ACID vs BASE

ACID	BASE
Strong consistency	Weak consistency
Isolation	Availability first; last write wins
Pessimistic (uses locks))	Optimistic
Nested transactions	Best effort: approximate answers
Available/Consistent	Available/Partition-tolerant
Robust Database/Simple Code	Simple Database/Complex Code

- Atomicity, Consistency, Isolation, Durable: strong guarantees on Consistency - transactions must complete or roll back to a consistent state; transactions do not interfere with each other.
- Basic Availability, Soft state, Eventual consistency: strong guarantees on Availability - transaction handling is approximate, so Consistency might suffer in the short term and might not ever be fully achieved.

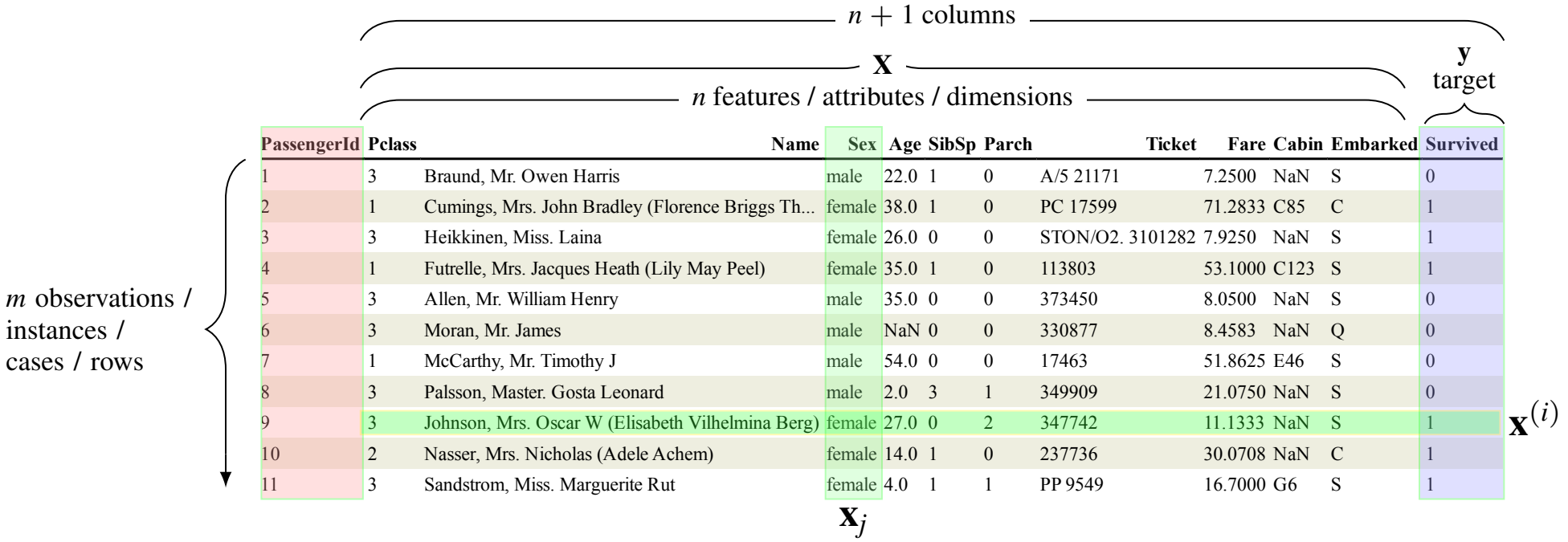
Summary of Data Storage

- Data flows through data mining workflows, so its management and provenance is critical
- Operational databases are a common source of such data
- Data structure and relationships need to be considered in any workflow
- Many data mining tools prefer data structured as *dataframes* (generalised tables)
  - Denormalised data is preferred by ML tools
  - Operational data from relational and graph databases tend to be normalised
  - Other data stores tend to be denormalised, but need code to "unpick" the data elements
- BASE might not be good enough for anomaly detection systems: higher probability of false positives!

References

Learning Spark: Lightning-Fast Big Data Analysis, Holden Karau, Andy Konwinski, Patrick Wendell and Matei Zaharia. 2015 O'Reilly Media.
High Performance Spark: Best practices for scaling & optimizing Apache Spark Holden Karau and Rachel Warren. 2017 O'Reilly Media.
Graph Databases: New opportunities for connected data, Ian Robinson, Jim Webber and Emil Eifrem. 2015 O'Reilly Media.

Terminology / Notation (sneak peek of Week 3!)



PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S	0
2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C	1
3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S	1
4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S	1
5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S	0
6	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q	0
7	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S	0
8	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S	0
9	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S	1
10	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C	1
11	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7000	G6	S	1

- A labeled dataset consists of m rows $\times$ $(n + 1)$ columns.
- Use bold to represent vectors and matrices.
- Use subscripts to indicate particular feature / attribute / column . . . . . $\mathbf{x}_j$
- Use superscript in parenthesis to indicate particular observation / instance/ case / row . . . . . $\mathbf{x}^{(i)}$
- So $x_j^{(i)}$ (or $x_{i,j}$) is the i-th observation in the j-th feature . . . . . $x_j^{(i)}$

Entering data in a dataframe

Tip: load data into dataframe using pandas: `pandas.read_*`

- Refer to `https://pandas.pydata.org/pandas-docs/stable/reference/io.html` for documentation on pandas functions to read from text files, SQL databases, HDFS, etc.
- When reading from a normalised SQL database, it is probably better to use `pandas.read_sql_query(..)` rather than collecting the tables individually and denormaliseing into a dataframe in the notebook
- A pandas dataframe is a good fit to tabular data and key-value stores, but more effort is needed with recursive data structures (column families, document stores, graphs etc.)
- Notebooks are generally applied to bounded data (e.g., from a batch process) rather than unbounded data from a stream pipeline
- Using `pyarrow` and Apache Arrow, it is possible to create a pandas dataframe from a Spark dataframe.
- Using `pyspark` it is possible to create a Spark dataframe from a pandas dataframe.
- However, it is more idiomatic to use python just as a way of orchestrating Spark pipelines, calling the Spark ML libraries and related libraries as needed: keep the processing in Spark.