Lecture 10 – NoSQL

CS2209
Information Storage and Management II

A TRADITION OF INDEPENDENT THINKING

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

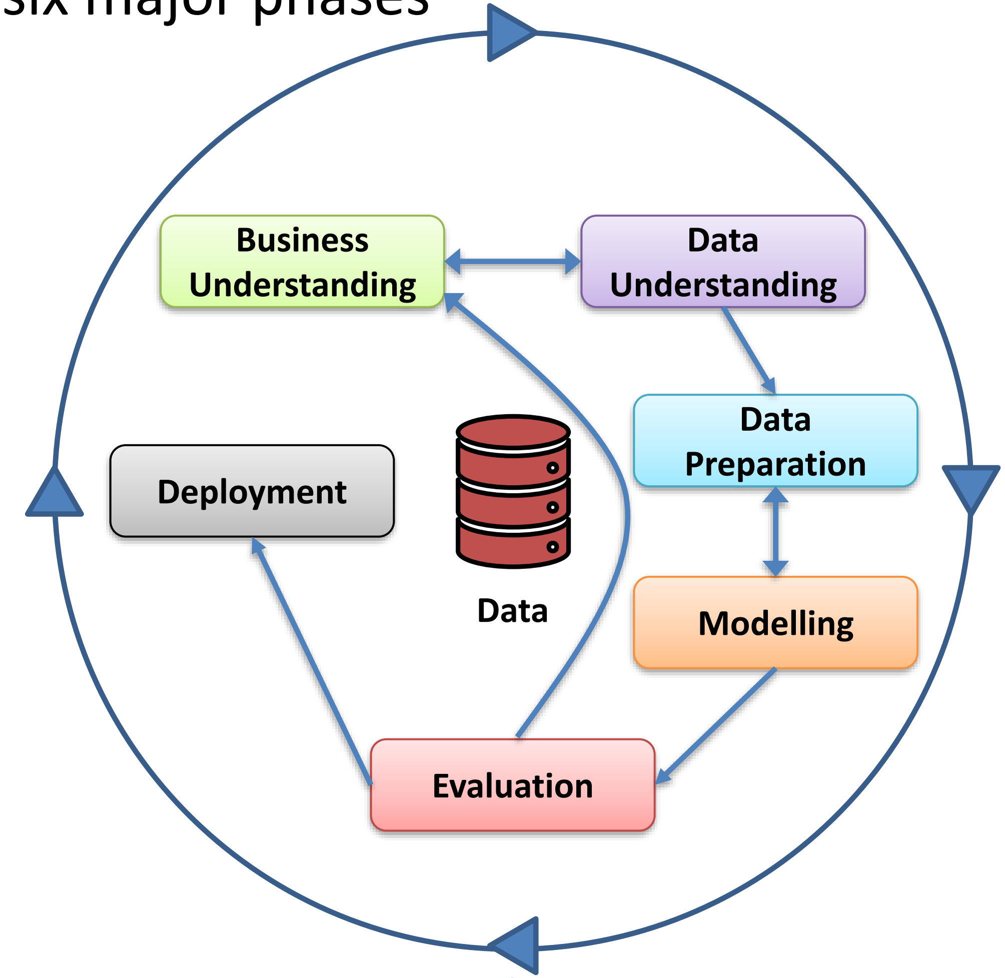
Dr Harry Nguyen
<hn@cs.ucc.ie>

What We Saw Last Lecture

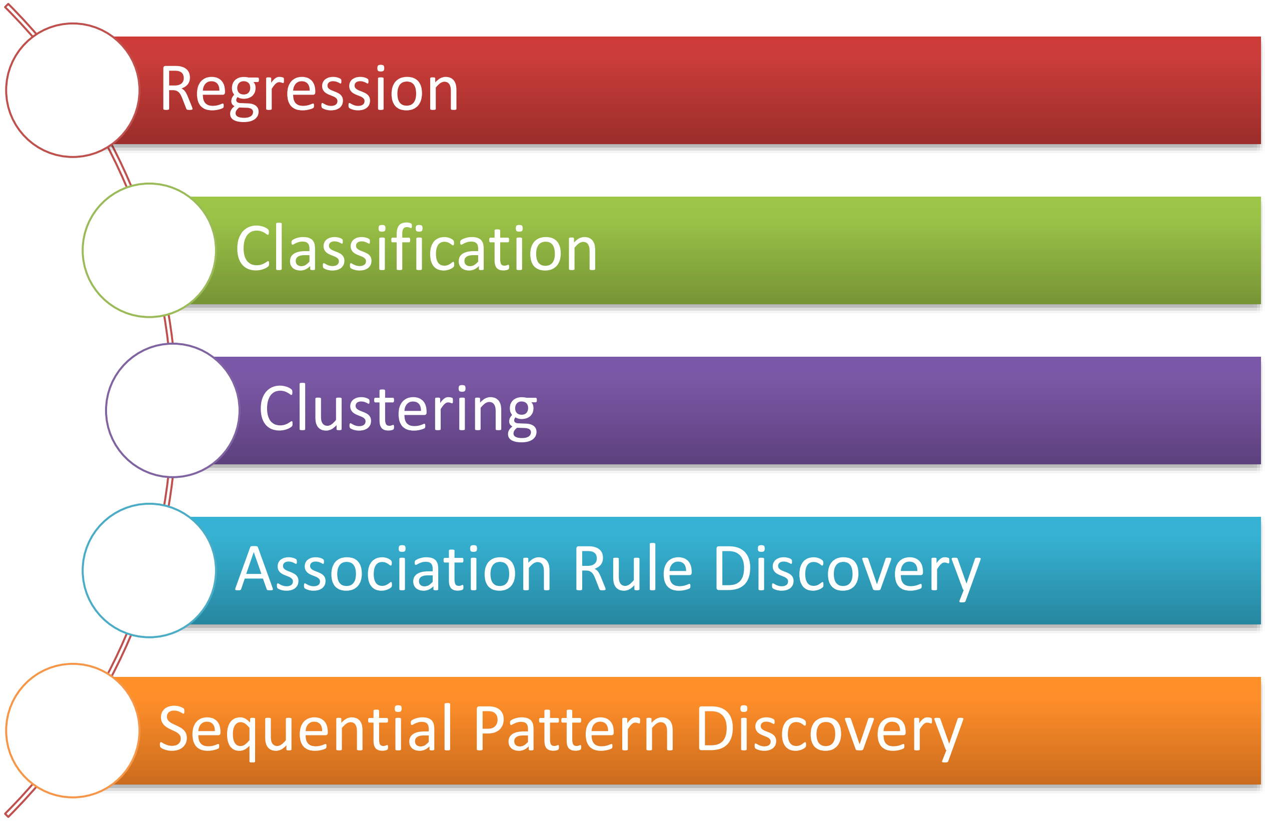
- Data Mining
- CRISP-DM
- Data Mining Techniques

CRISP-DM

- Cross-industry standard process for data mining (CRISP-DM) consists of six major phases



Data Mining Tasks



- Regression
- Classification
- Clustering
- Association Rule Discovery
- Sequential Pattern Discovery

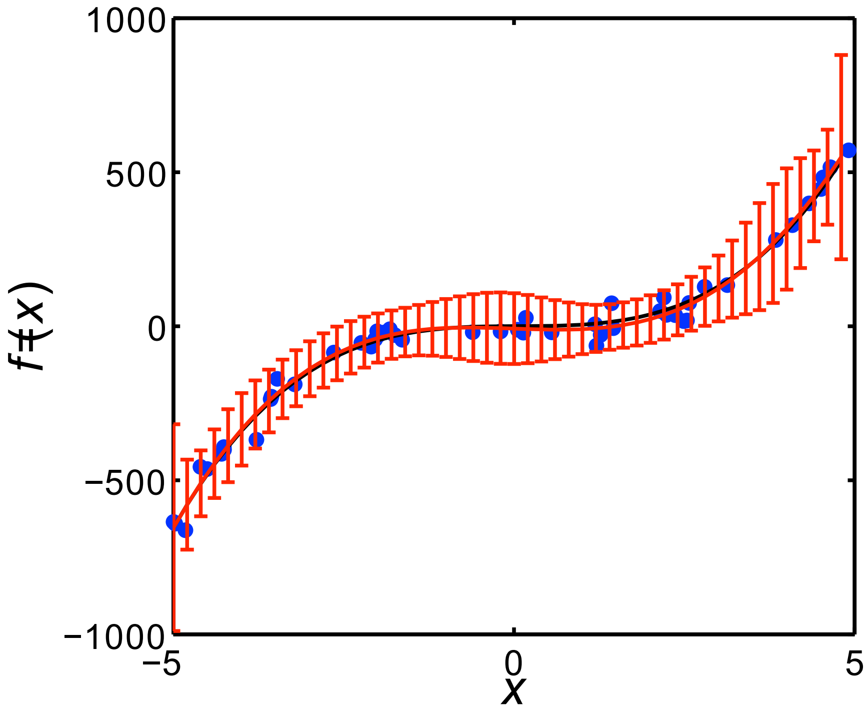
Regression

Regression

- Learning a continuous function from a set of examples

Example (s)

- Predicting stock prices (x might be time or some other variable of interest)



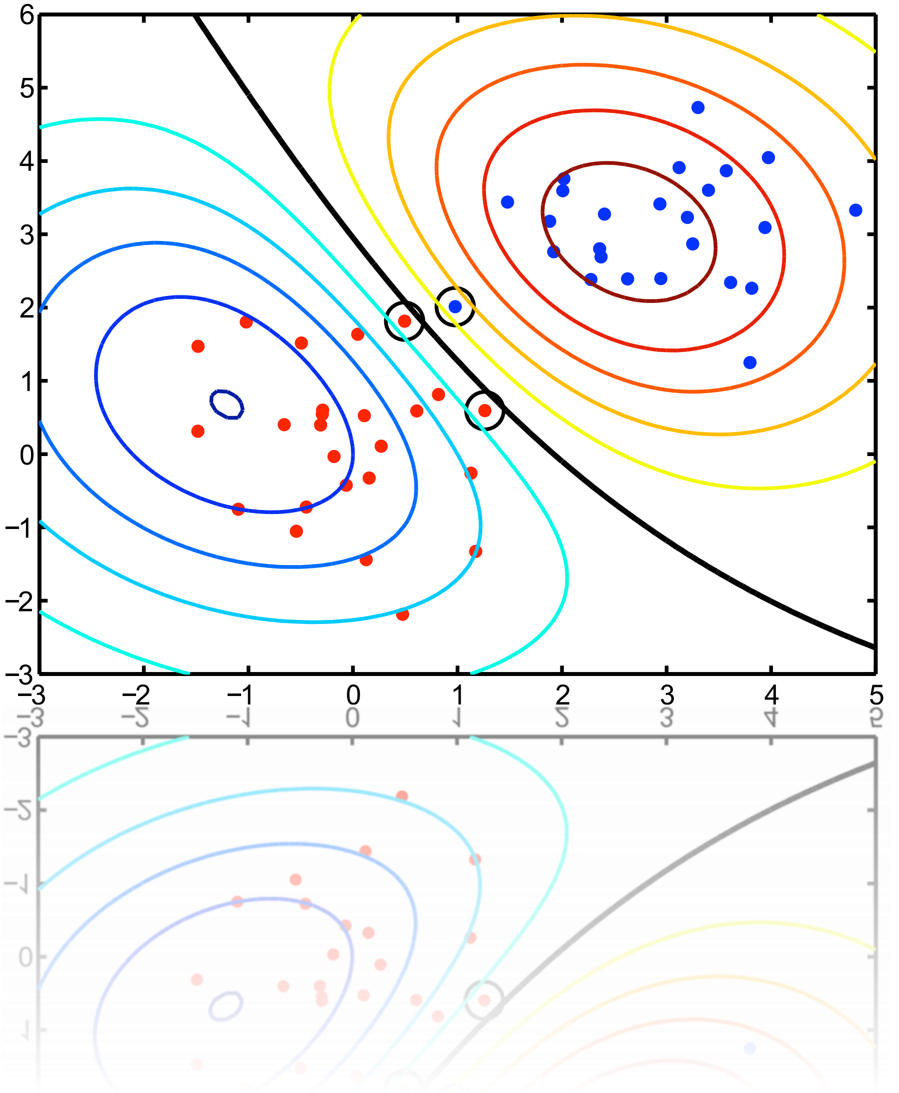
CLASSIFICATION

Classification

- Learning rules that can separate objects of different types from one another

Example(s)

- Disease diagnosis
- Spam email detection



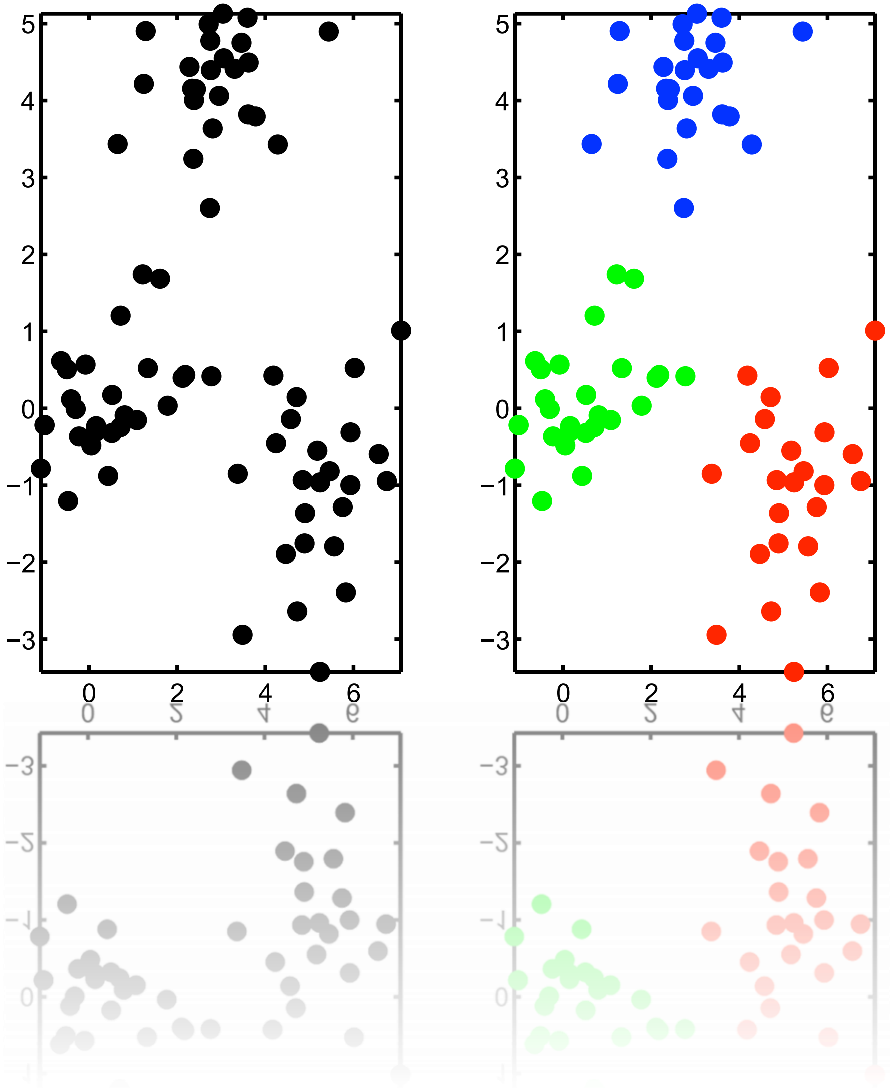
Clustering

Clustering

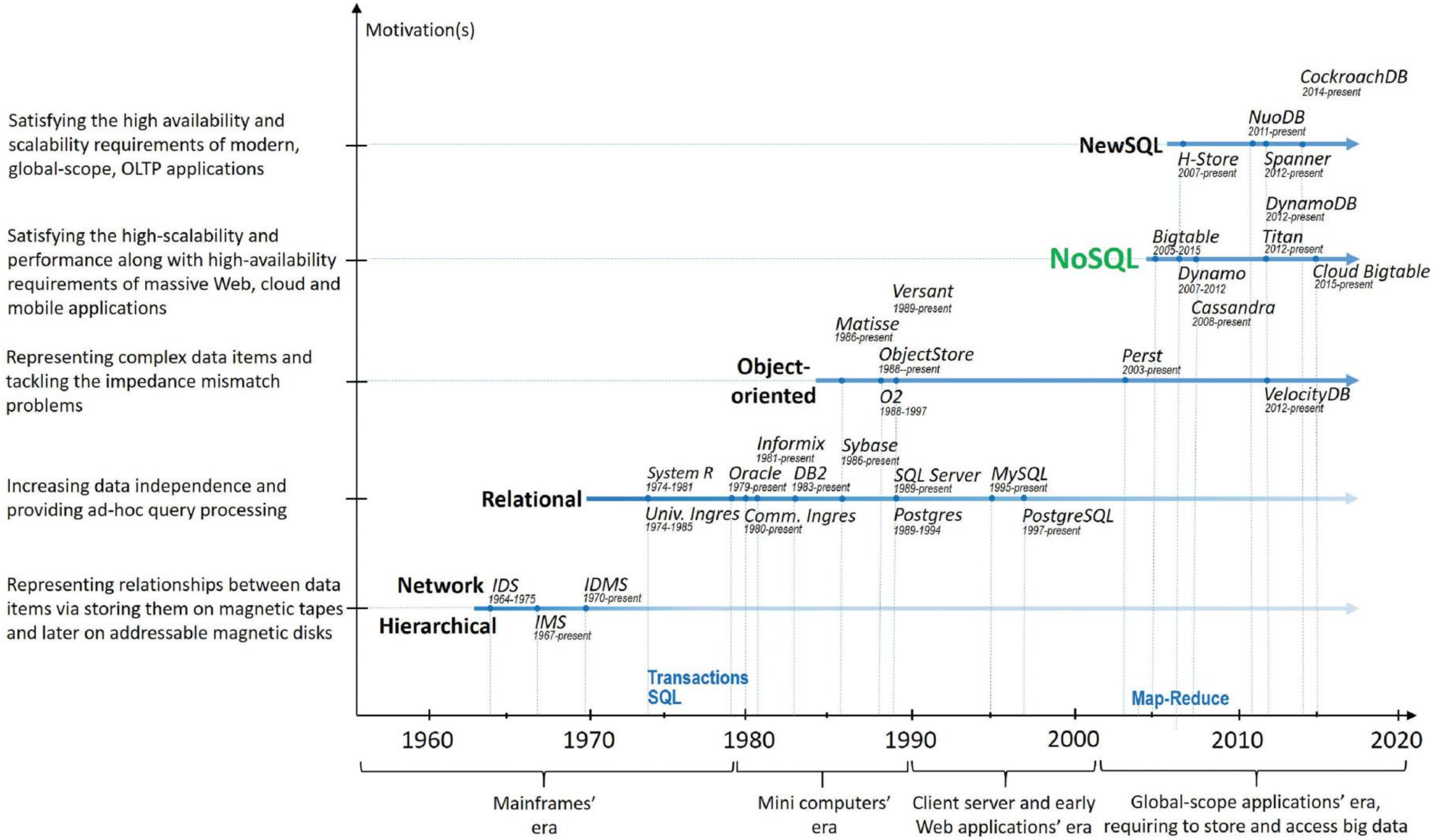
- Finding groups of similar objects

Example(s)

- People with similar “preferences”
- Genes with similar functions



THE DB LANDSCAPE



Davoudian, Ali, Liu Chen, and Mengchi Liu. "A survey on NoSQL stores." *ACM Computing Surveys (CSUR)* 51.2 (2018): 1-43.

Issues with Relational Databases

Impedance mismatch for application developers

Persisted Representation is different to that used by a client application

The data must be retrieved, organized and new objects created from this. Often this data is organized in memory as hierarchies of objects (data structures - dictionaries of lists of dictionaries….)

Movement towards services where data integration occurs in code and individual data sources' structures are relatively simple.

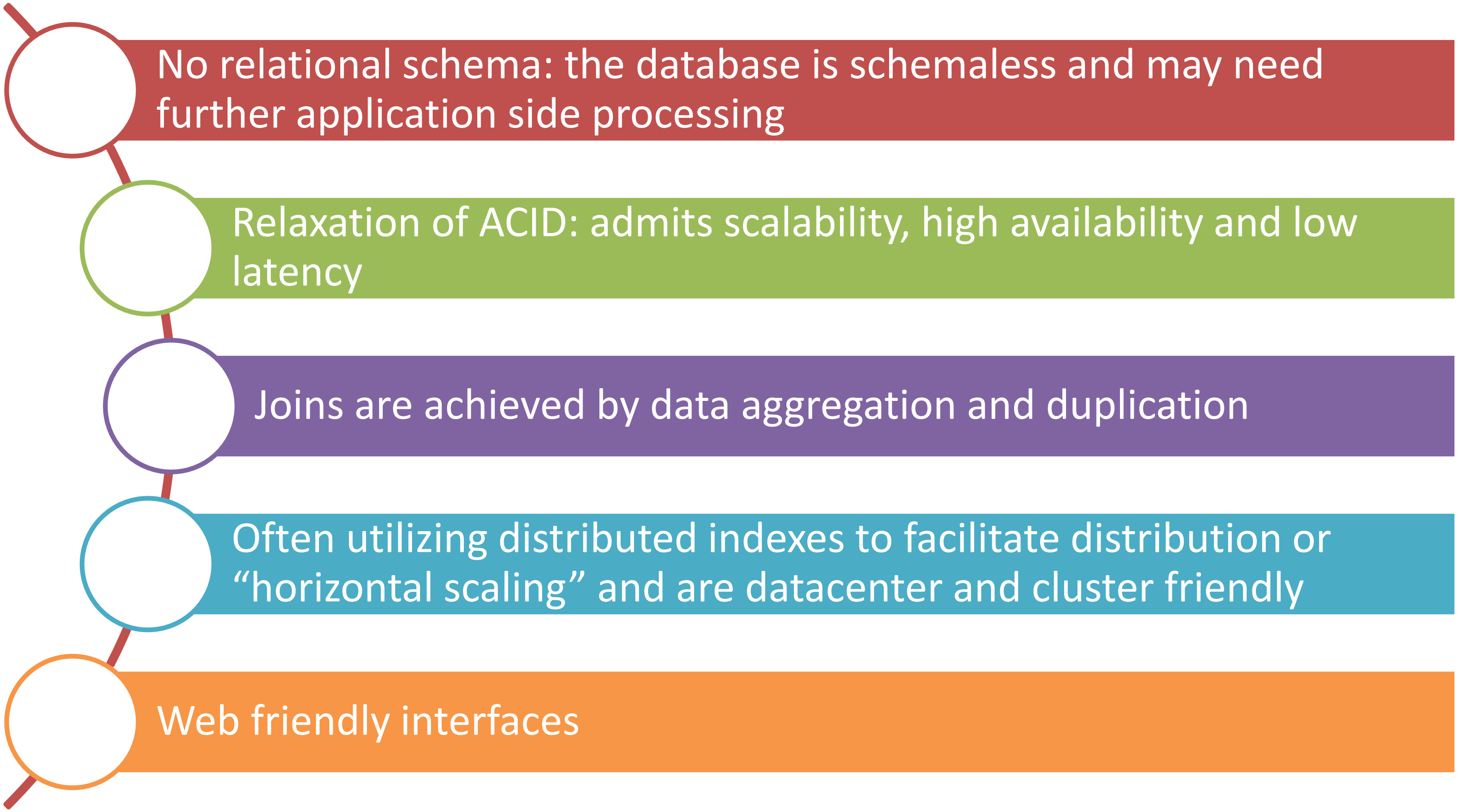
Also, a need to run on Clusters (expand horizontally) and support processing of massive scales of data.

NoSQL – What's in a name?

- **NoSQL** – its not even certain what it means – **No SQL**, **Non SQL**, **N**ot **O**nly **SQL**

NoSQL

- There is no single definition – but a series of characteristics



No relational schema: the database is schemaless and may need further application side processing

Relaxation of ACID: admits scalability, high availability and low latency

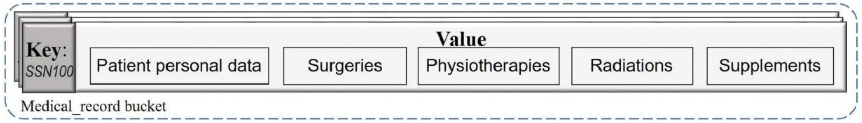
Joins are achieved by data aggregation and duplication

Often utilizing distributed indexes to facilitate distribution or “horizontal scaling” and are datacenter and cluster friendly

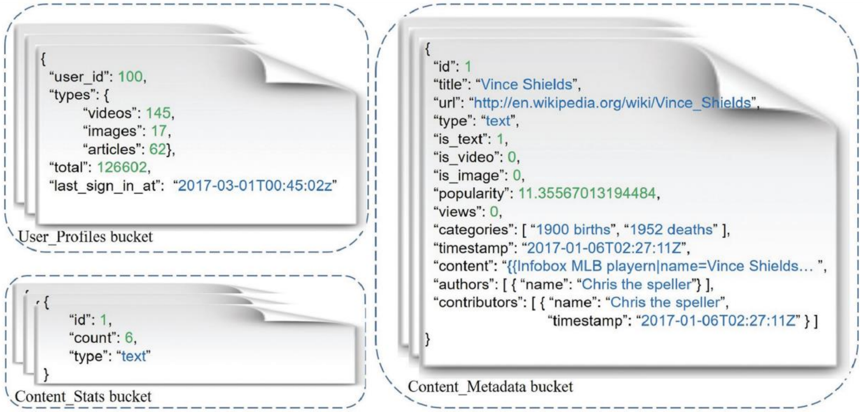
Web friendly interfaces

High Level NoSQL Data Models

Key Value and Wide Column Stores



Document



```
{
  "user_id": 100,
  "types": {
        "videos": 145,
        "images": 17,
        "articles": 62},
  "total": 126602,
  "last_sign_in_at":  "2017-03-01T00:45:02z"
```

User_Profiles bucket

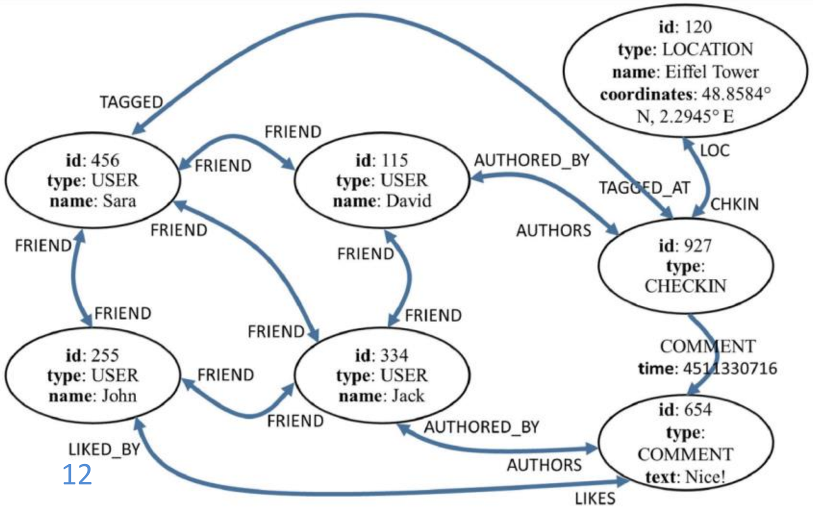
```
{
    "id": 1,
    "count": 6,
    "type": "text"
}
```

Content_Stats bucket

```
{
  "id": 1
  "title": "Vince Shields",
  "url": "http://en.wikipedia.org/wiki/Vince_Shields",
  "type": "text",
  "is_text": 1,
  "is_video": 0,
  "is_image": 0,
  "popularity": 11.35567013194484,
  "views": 0,
  "categories": [ "1900 births", "1952 deaths" ],
  "timestamp": "2017-01-06T02:27:11Z",
  "content": "{{Infobox MLB playern|name=Vince Shields… ",
  "authors": [ { "name": "Chris the speller"} ],
  "contributors": [ { "name": "Chris the speller",
                       "timestamp": "2017-01-06T02:27:11Z" } ]
}
```

Content_Metadata bucket

Graph



DATA MODELS

Key Values

- Simplest and most popular NoSQL stores
- Data are managed and represented as (key, value) pairs
- Efficient, highly scalable, key-based lookup structures such as Distributed Hash Tables (DHTs)

Document

- Extended key-value stores
- Represented as a document encoded in standard semi-structured formats such as XML, JSON, or BSON (Binary JSON)
- Has a flexible schema through adding or removing its attributes at runtime, when an attribute has a name along with one or more values

MongoDB

Interact with the data using functions

Interactions with Mongo are driven by the clients

- Schema less – schema doesn’t exist until we add documents
- Easily expand the ‘schema’ – just add documents with the data you want to persist

Database

- Collections
  - Documents
    - Key Value Pairs

Collections can be made up of documents with different keys or keys can have different types of values

- Have to have some structure to provide uniformity to processing

JavaScript Object Notation (JSON)

Plain text file format

Although generally derived from Javascript scripting language, JSON is a language independent format

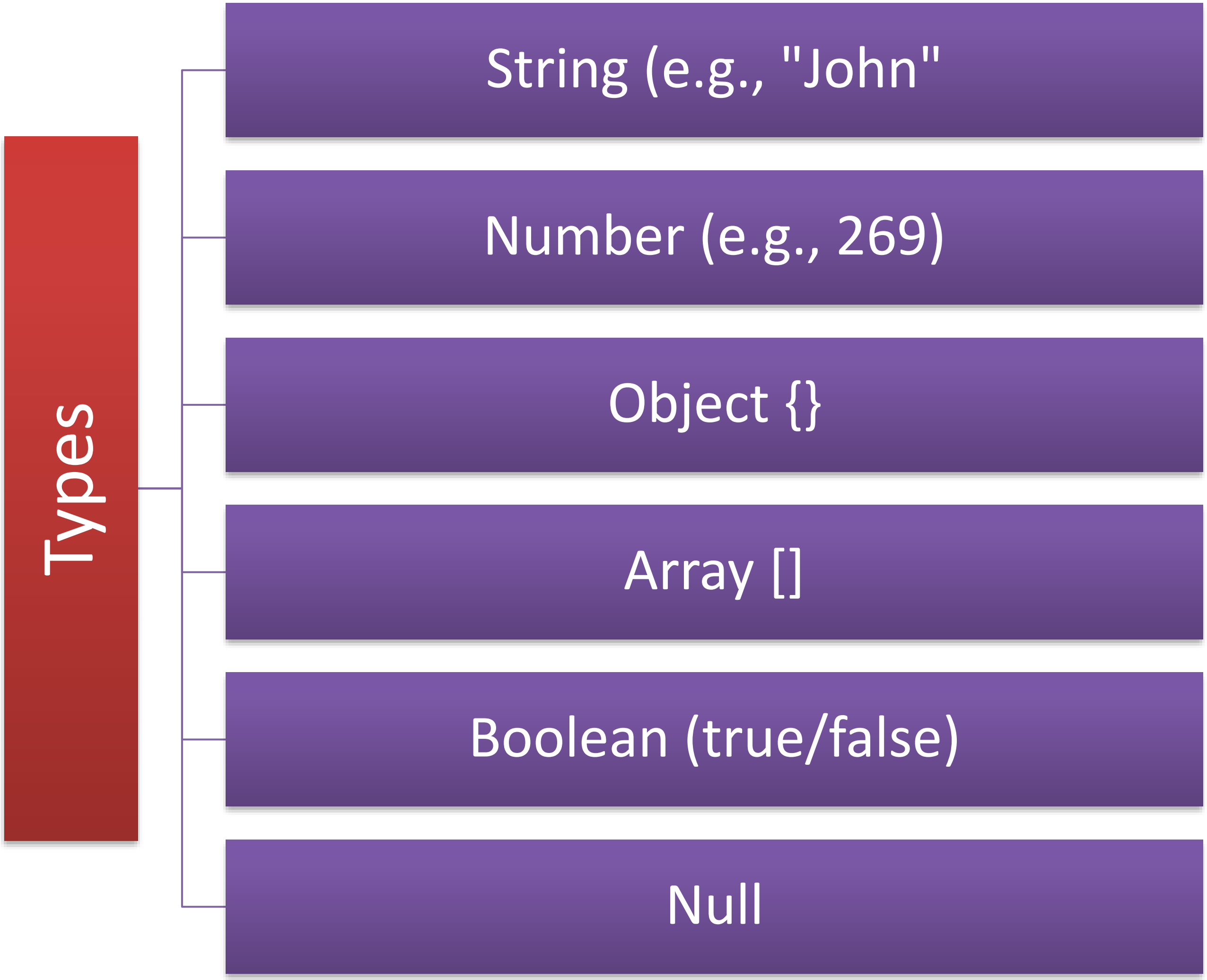
Also closely related to representations of dictionaries in Python

Python JSON to Dictionary is included in the language

JSON - Example

```
{
    "name":"josh",
    "pets": [{
        "name": "zoe",
        "age":7,
        "birthday":null,
        "species":"dog",
        "commands":["fetch", "shake", "come"]
    }],
    "employed":true
}
```

JSON - Types



Types

String (e.g., "John")

Number (e.g., 269)

Object {}

Array []

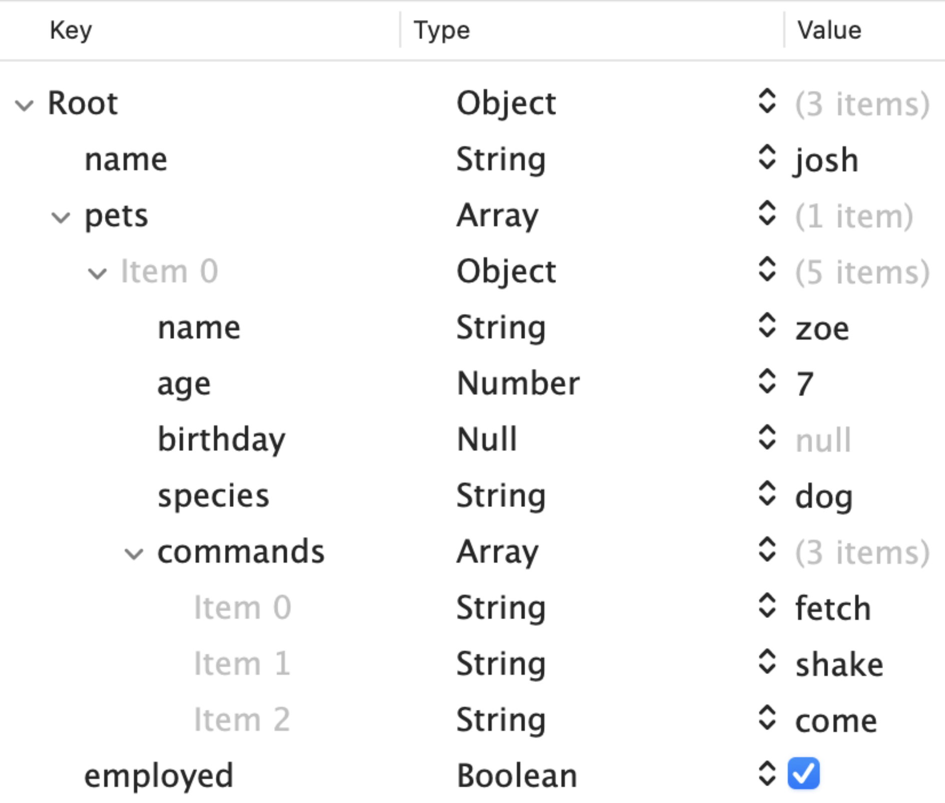
Boolean (true/false)

Null

JSON Parsing

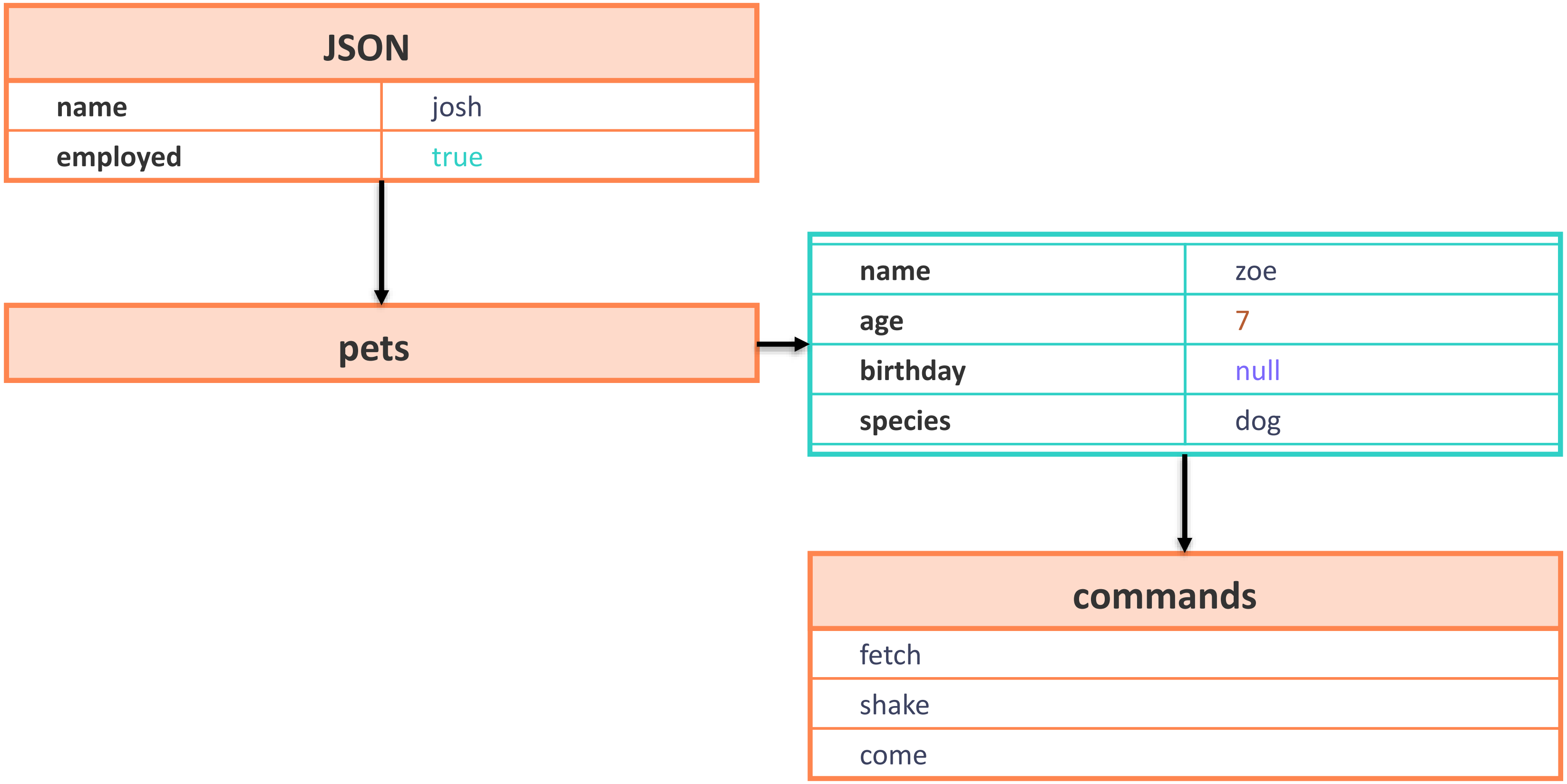
CS2209
Information
Storage and
Management II

```
{
   "name":"josh",
   "pets": [{
      "name": "zoe",
      "age":7,
      "birthday":null,
      "species":"dog",
      "commands":["fetch",
      "shake", "come"]
   }],
   "employed":true
}
```



Key	Type	Value
Root	Object	(3 items)
name	String	josh
pets	Array	(1 item)
Item 0	Object	(5 items)
name	String	zoe
age	Number	7
birthday	Null	null
species	String	dog
commands	Array	(3 items)
Item 0	String	fetch
Item 1	String	shake
Item 2	String	come
employed	Boolean	☑

JSON Data Model



JSON	
name	josh
employed	true

pets

name	zoe
age	7
birthday	null
species	dog

commands
fetch
shake
come

SUMMARY

DB Landscape

NoSQL

JSON