Data Mining

John Samuel
CPE Lyon

Year: 2019-2020
Email: john(dot)samuel(at)cpe(dot)fr

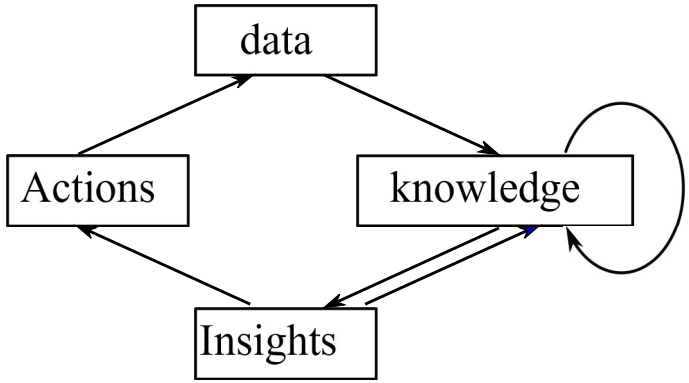
Goals

- Lifecycle of data
- Data acquistion and storage
- Data extraction and integration
- Pre-treatment of data
- Data transformation
- ETL
- Data analysis
- Data visualisation

1. Lifecycle of data

Lifecycle of Data

1. Data
2. Knowledge
3. Insights
4. Actions

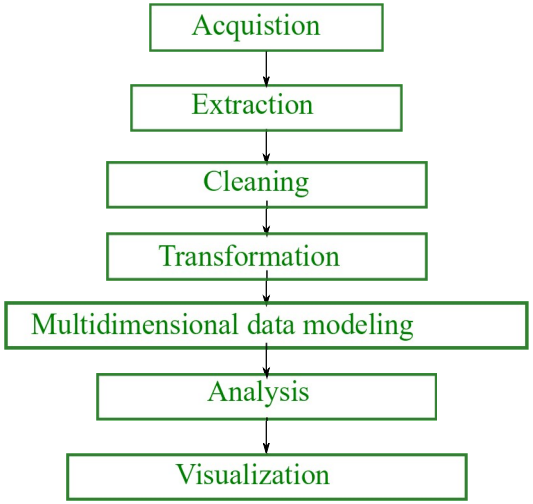


Data Lifecycle

1. Lifecycle of data

1.1. From Data to Knowledge

1. Data acquisition
2. Data Extraction
3. Data Cleaning
4. Data Transformation
5. Data analysis modeling
6. Data Storage
7. Analysis
8. Visualisation

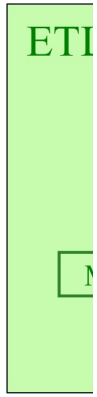


Major steps of data analysis

1.1.1. Data Acquistion

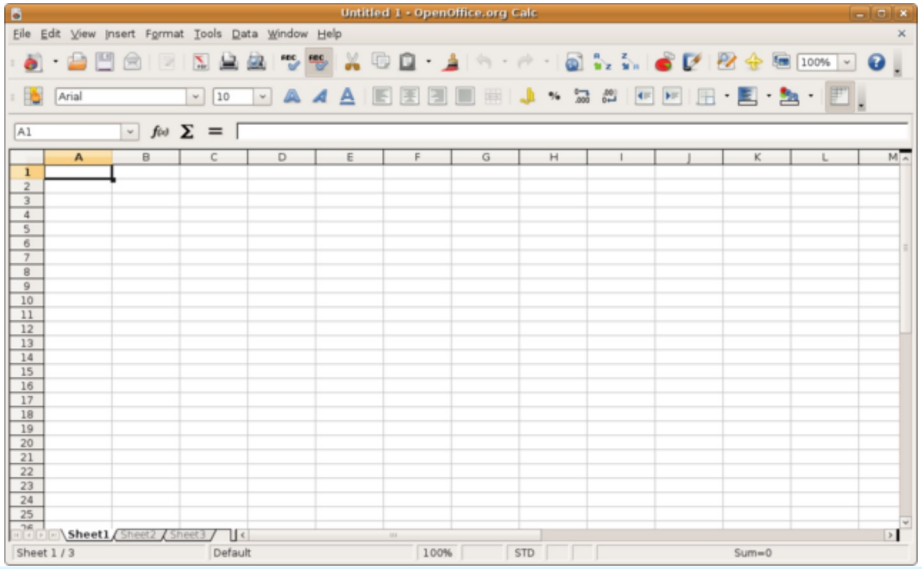
1.1.2. ETL (Extraction Transformation and Loading)

1. Data Extraction
2. Data Cleaning
3. Data Transformation
4. Loading data to information stores



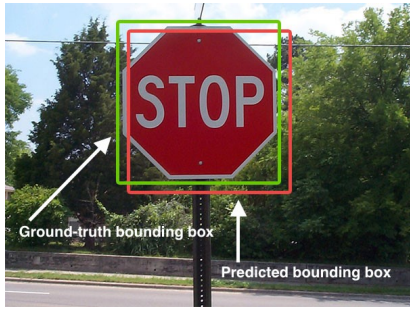
ETL (Extra

1.1.3. Data Analysis

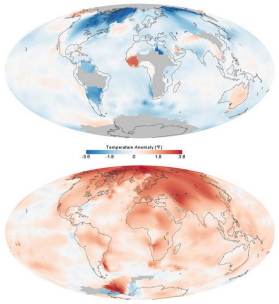


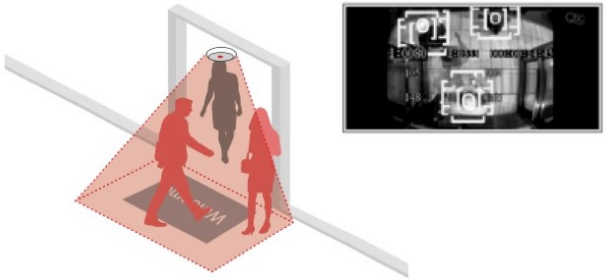
1.1.3. Data analysis

1.1.4. Data Visualization



1.1.4. Data Visualization





2.1. Data acquisition

1. Surveys
   - Manual surveys
   - Online surveys
3. Sensors[1]
   - Temperature, pressure, humidity, rainfall
   - Acoustic, navigation
   - Proximity, presence sensors
4. Social networks
5. Video surveillance cameras
6. Web

1. https://en.wikipedia.org/wiki/List_of_sensors

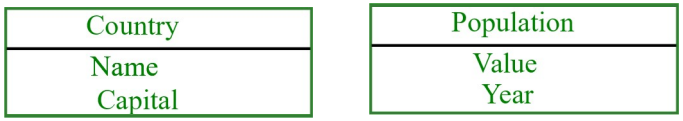
2.2. Data storage formats

- Binary and Textual Files
- CSV/TSV
- XML
- JSON
- Media (Images/Audio/Video)

2.2 Types of data stores

1. Structured data stores
   - Relational databases
   - Object-oriented databases
2. Unstructured data stores
   - Filesystems
   - Content-management systems
   - Document collections
3. Semi-structured data stores
   - Filesystems
   - NoSQL data stores

Paris is the capital of France. In 2015,
its population was recorded as 2,206,488




```
<xml>
 <country>
   <name>France</name>
   <capital>
     <name>Paris</name>
     <population>
       <value>2,206,488</value>
       <year>2015</year>
     </population>
   </capital>
 </country>
</xml>
```

Unstructured vs. Structured vs. Semi-structured

2.3.1. ACID Transactions[1]

- **Atomicity**: Each transaction must be "all or nothing".
- **Consistency**: Any transaction must bring database from one valid state to another.
- **Isolation**: Both concurrent execution and sequential execution of transactions must bring the database to same state.
- **Durability**: Irrespective of power losses, crashes, a transaction once committed to the database must remain in that state.

1. https://en.wikipedia.org/wiki/ACID

2.3.1. ACID Transactions

- Ensure validity of databases even in case of errors, power failures
- Important in banking sector

2.3.2. Types of data stores

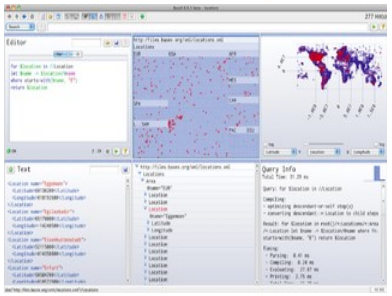
- Relational databases
- Object-oriented databases
- NoSQL (Not only SQL) data stores
- NewSQL

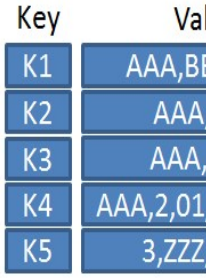
2.3.3. NoSQL

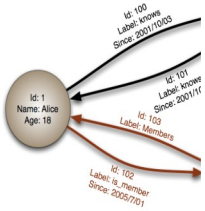
- Comprises consistency
- Focus on availability and speed

2.3.3. Types of NoSQL stores

- Column-oriented database
- Document-oriented database
- Key-value database
- Graph-oriented database







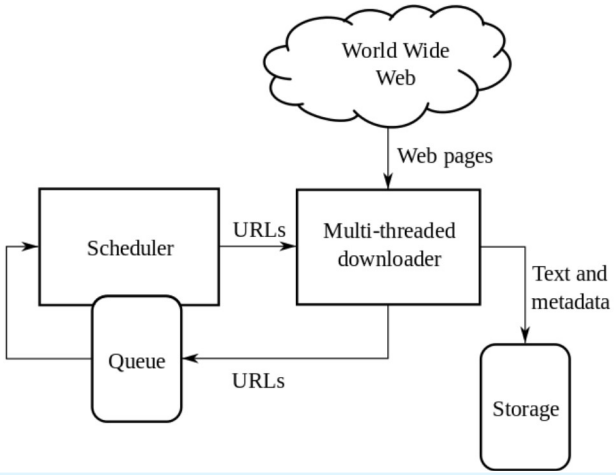
3.1. Data extraction techniques

- Data dumps
  - Downloading complete data dumps
  - Downloading selective data dumps
- Periodical polling of data feeds (e.g., blogs, news feeds)
- Data streams
  - Subscrbing to data streams (push notifications)

3.2. Query interfaces

- Query endpoints supporting declarative languages
  - SQL
  - SPARQL
- Automated Manual search (and filter) options

3.3. Crawlers for web pages



Web crawlers: navigating the entire using hyperlinks

3.4. Application Programming Interface (API)

- Web operations (CRUD) to manipulate externally managed resources
- Requires programmers to develop wrappers for web service integration

API (Inte

4.1 Data Cleaning: Types of Errors

- Syntactical errors
- Semantical errors
- Data coverage errors

4.1.1. Syntactical errors

- Lexical errors (e.g., user entered a string instead of a number)
- Data format errors (e.g, order of last name, first name)
- Irregular data errors (e.g., usage of different metrics)

4.1.2. Semantic errors

- Violation of integrity constraints
- Contradiction
- Duplication
- Invalid data (unable to detect despite presence of triggers and integrity constraints)

4.1.3. Coverage errors

- Missing values
- Missing data

4.2.1. Handling Syntactical errors

- Validation using schema (e.g., XSD, JSONP)
- Data transformation

4.2.2. Handling Semantic errors

- Duplicate elimination using techniques like specifying integrity constraints like functional dependencies

4.2.3. Handling Coverage errors

- Interpolation techniques
- External data sources

4.2.4. Administrators and handling errors

- User feedback
- Alerts and triggers

5.1 Languages

- Template languages
- XSLT
- AWK
- Sed
- Programming languages like PERL

6.1. ETL (Extraction Transformation and Loading)

1. Data Extraction
2. Data Cleaning
3. Data Transformation
4. Loading data to information stores

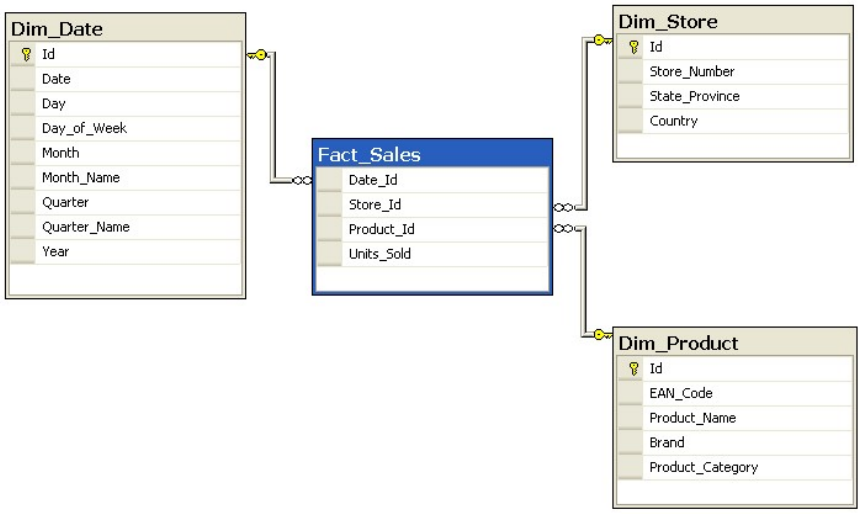
6.2.1. Models for data analysis

- Multidimensional data analysis
  - Dimensions
    - Attributes
    - Levels
    - Hierarchies
  - Facts
    - Measures

6.2.1. Models for data analysis

- Multidimensional data analysis: Examples
  - Dimensions (e.g.Spatio-temporal dimensions, Product)
    - Attributes (e.g. Name, Manufactures etc.)
    - Levels (e.g., Day, Month, Quarter, Store, City, Country etc.)
    - Hierarchies (e.g., Day-Month-Quarter-Year, Store-City-Country etc.)
  - Facts
    - Measures (e.g., Number of products sold/unsold)

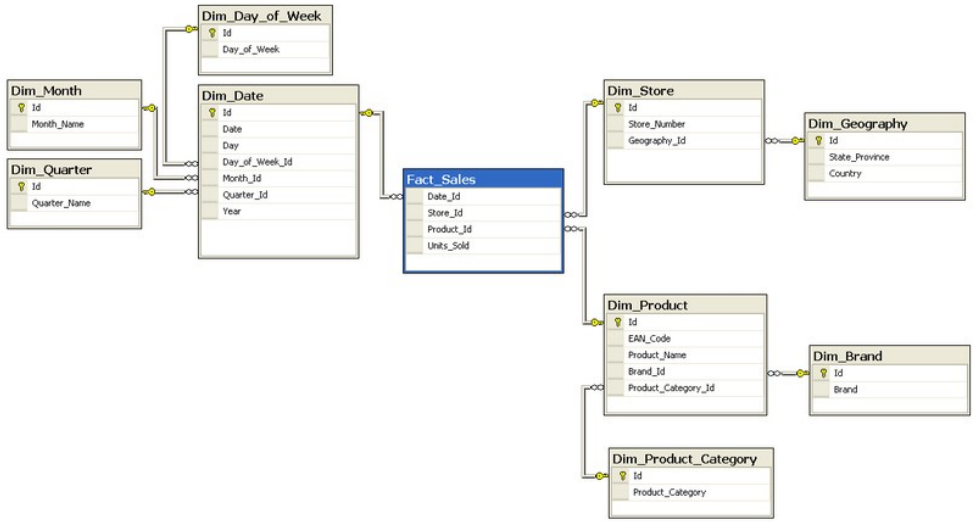
6.2.3. Star Schema



6.2.3. Data Cubes

- Data cubes for online analytical processing (OLAP)
- OLAP Cube operations
  - Slice
  - Dice
  - Drill up/down
  - Pivot

6.2.4. Snow Schema



6.2. ETL: From one data store to another

- From: Data sources
  - Internal or external databases
  - Web Services
- To: Data warehouses
  - Enterprise warehouses
  - Web warehouses

Activities of data analysis

1. Retrieving values
2. Filter
3. Compute derived values
4. Find extremum
5. Sort
6. Determine range
7. Characterize distribution
8. Find analysis
9. Cluster
10. Correlate
11. Contextualization

1. https://en.wikipedia.org/wiki/Data_analysis

8.1. Data Visualization

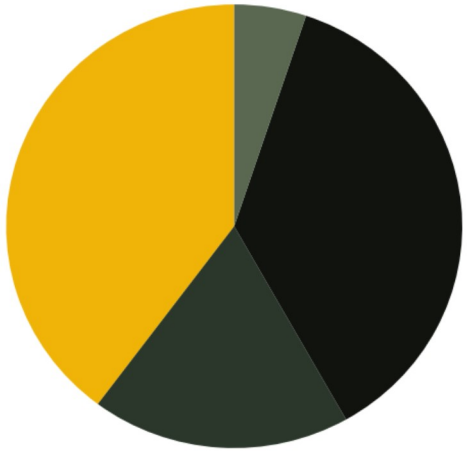
1. Time-series
2. Ranking
3. Part-to-whole
4. Deviation
5. Sort
6. Frequency distribution
7. Correlation
8. Nominal comparison
9. Geographic or geospatial

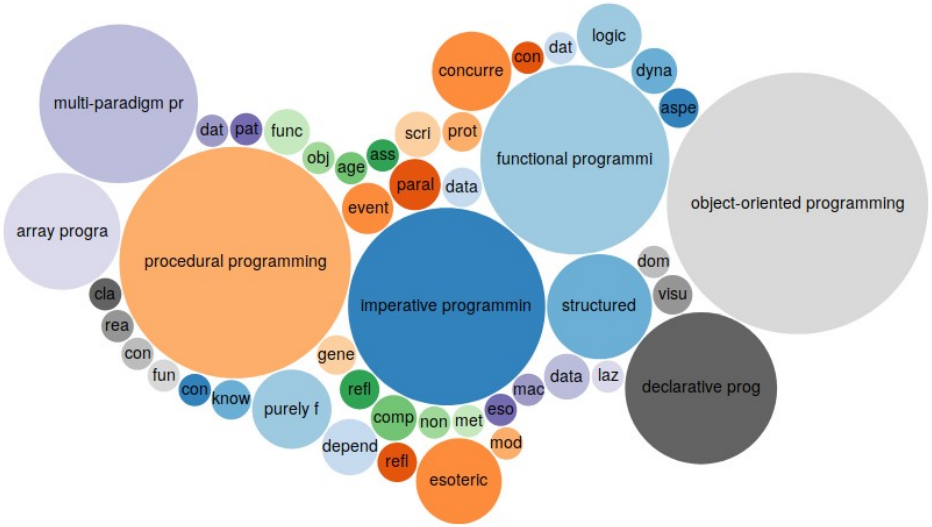
1. https://en.wikipedia.org/wiki/Data_visualization

8.2. Data Visualization: Examples

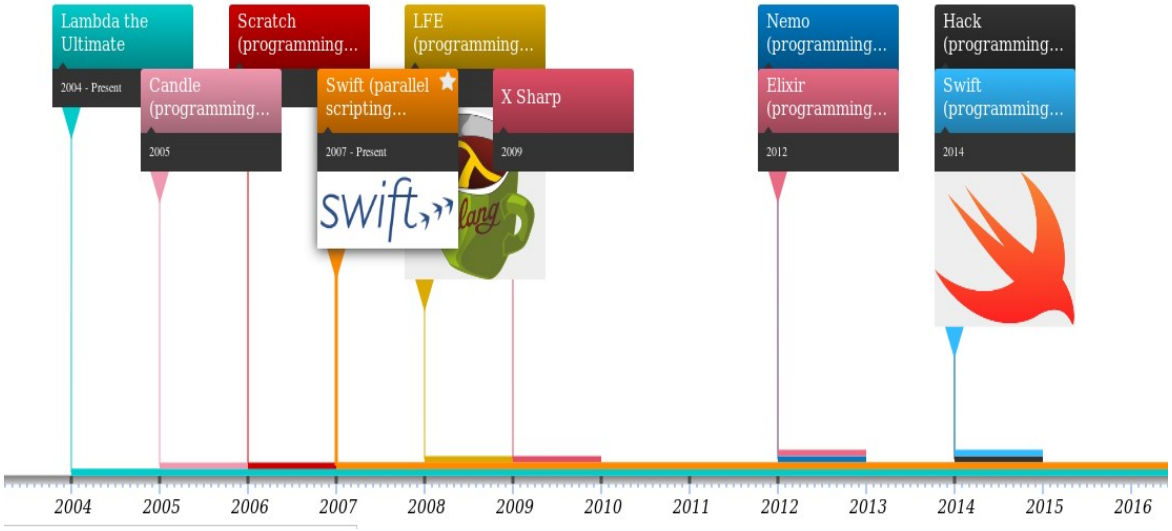
1. Bar-chart (Nominal comparison)
2. Pie-chart (part-to-whole)
3. Histograms (frequency-distribution)
4. Scatter-plot (correlation)
5. Network
6. Line-chart (time-series)
7. Treemap
8. Gantt chart
9. Heatmap

Pie Chart



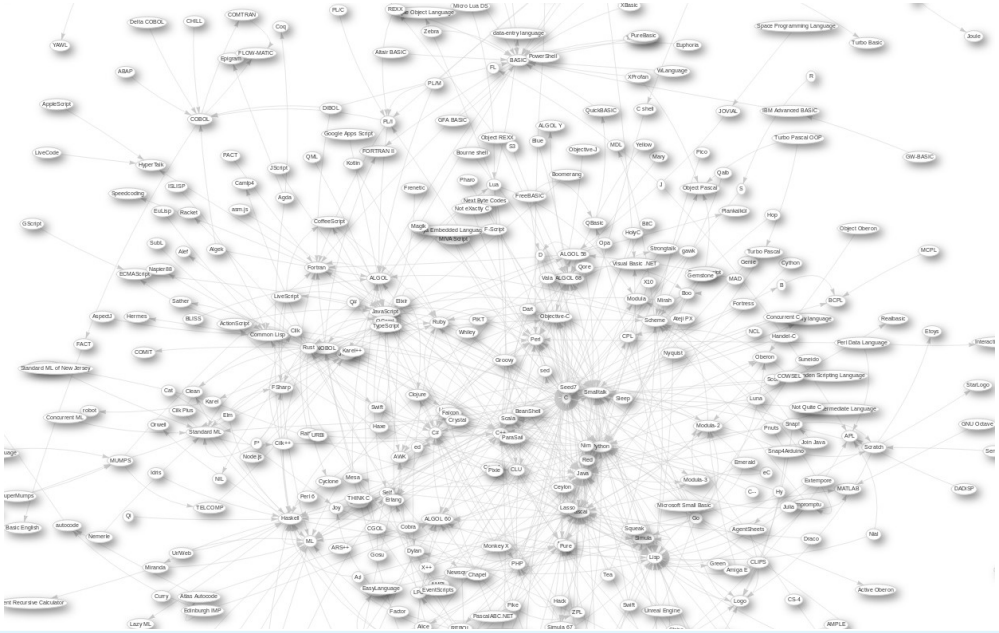


Programming Language Paradigms (Bubble Chart)



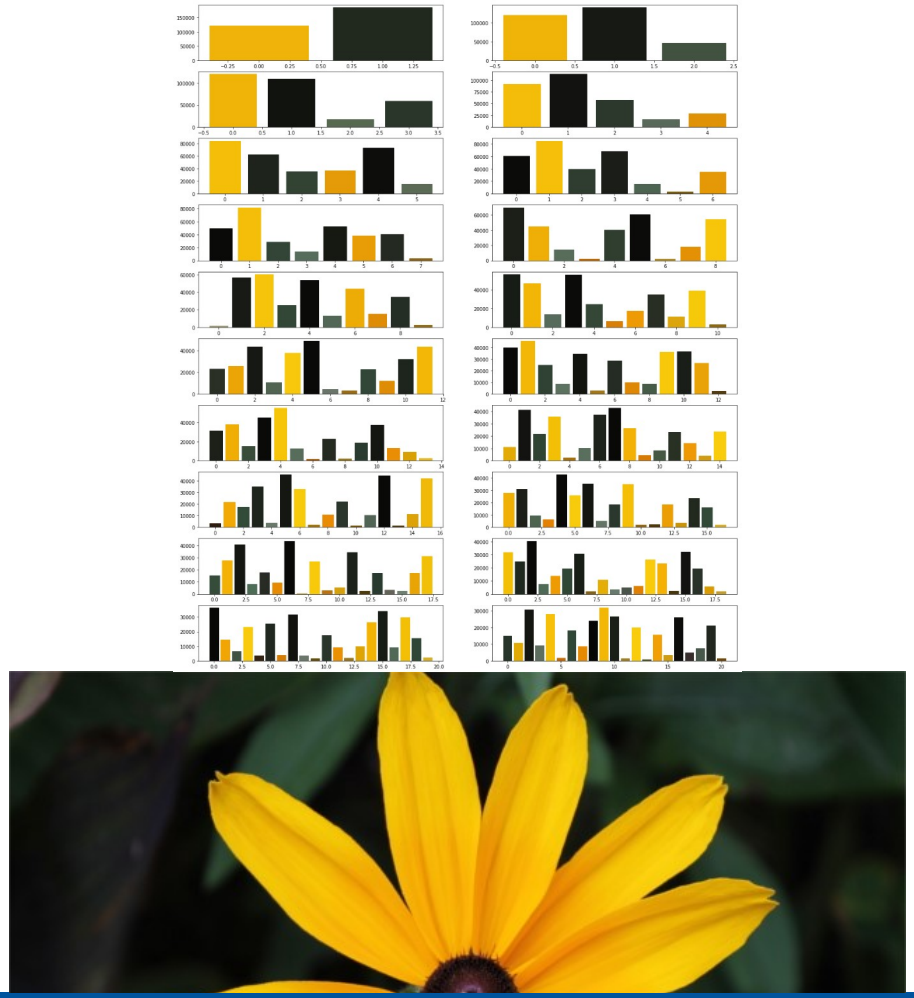
Timeline of Programming Languages (using Histropedia)

8. Data Visualization

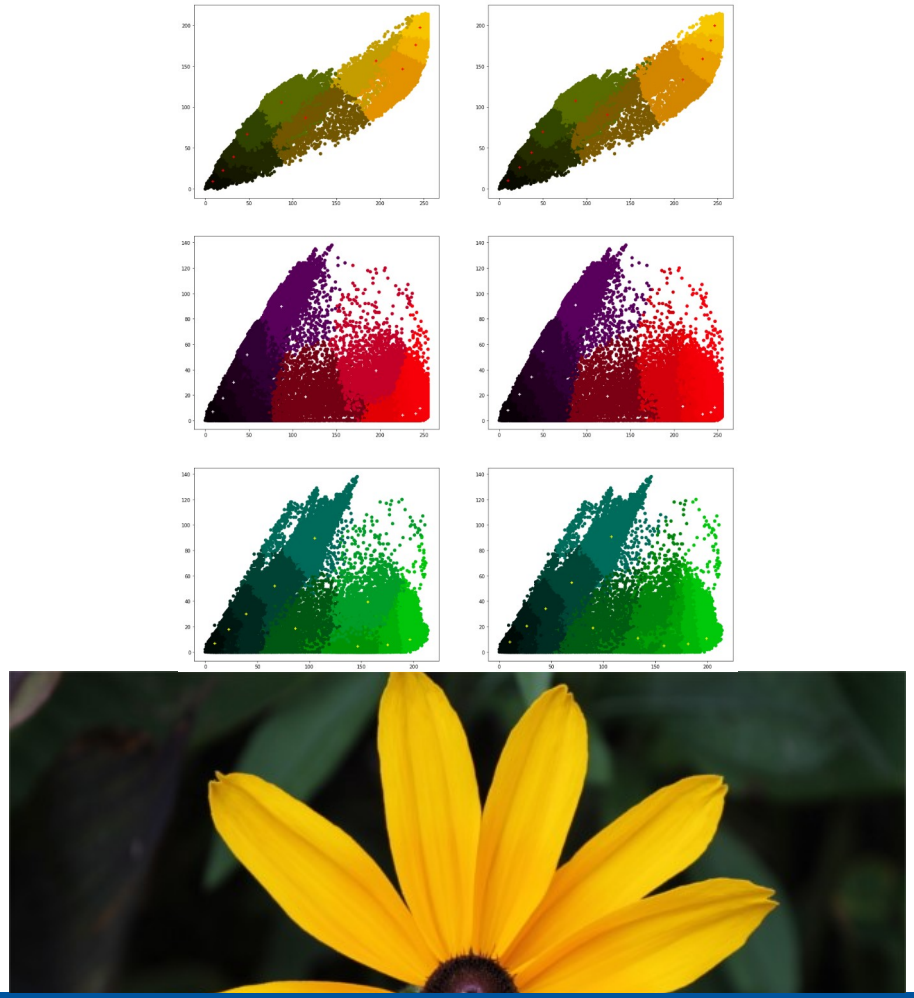


Influence Graph of Programming Languages

k Predominant colours



RGB Scatter plots (Comparison)



References

Colors

- Color Tool - Material Design

Images

- Wikimedia Commons