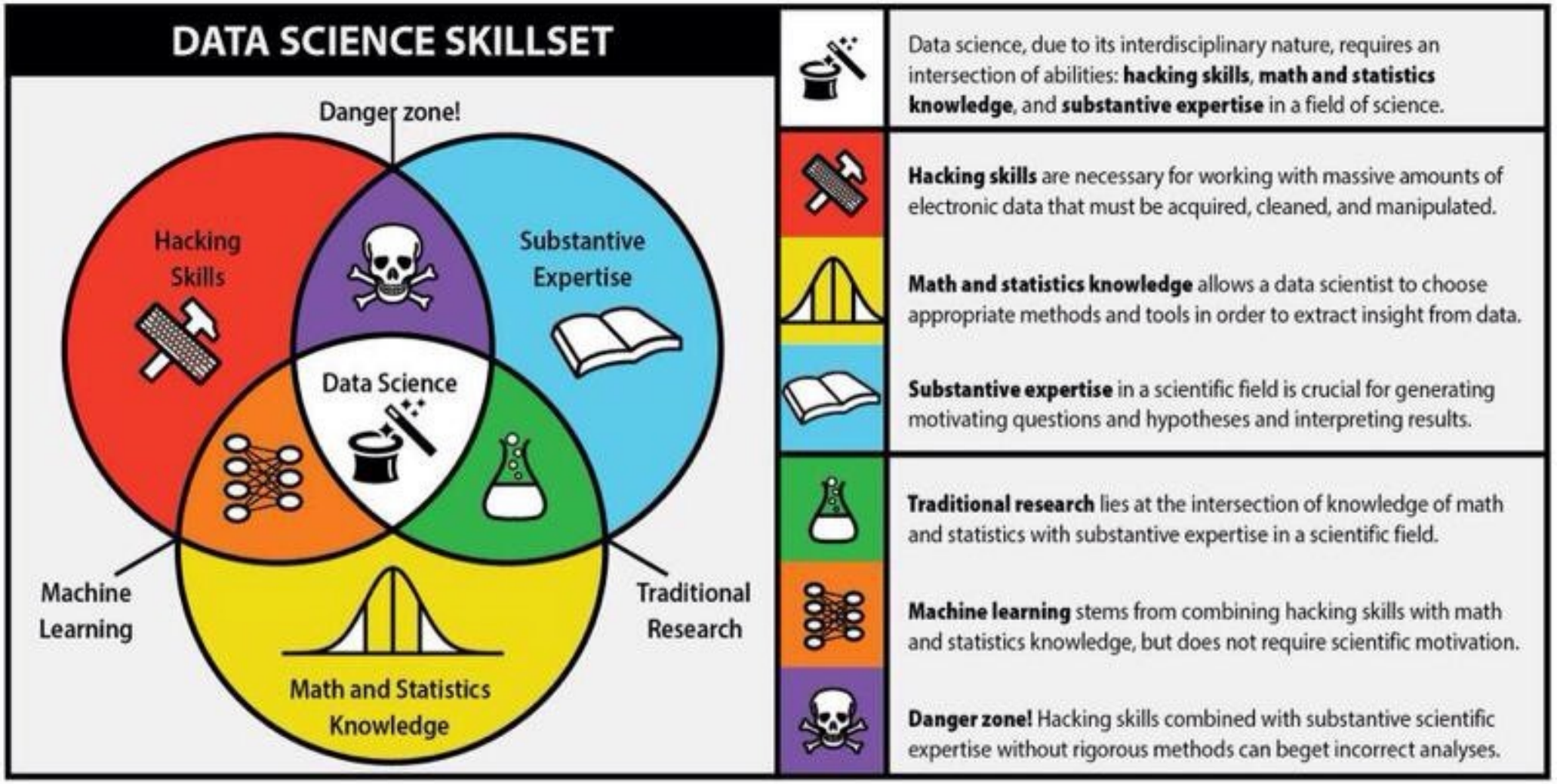
Data analysis, data science

- Data analysis: *”the process of extracting information from data”*.
- Data science: *”the process of extracting knowledge from data”*.
- Information vs knowledge (Oxford dictionary):
  - Information: *facts provided or learned about something or someone*.
  - Knowledge: *facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject*.
- Data analysis can be viewed as one part of data science.
- Both are very broad/general (and somewhat vague) concepts.

Data analysis, data science

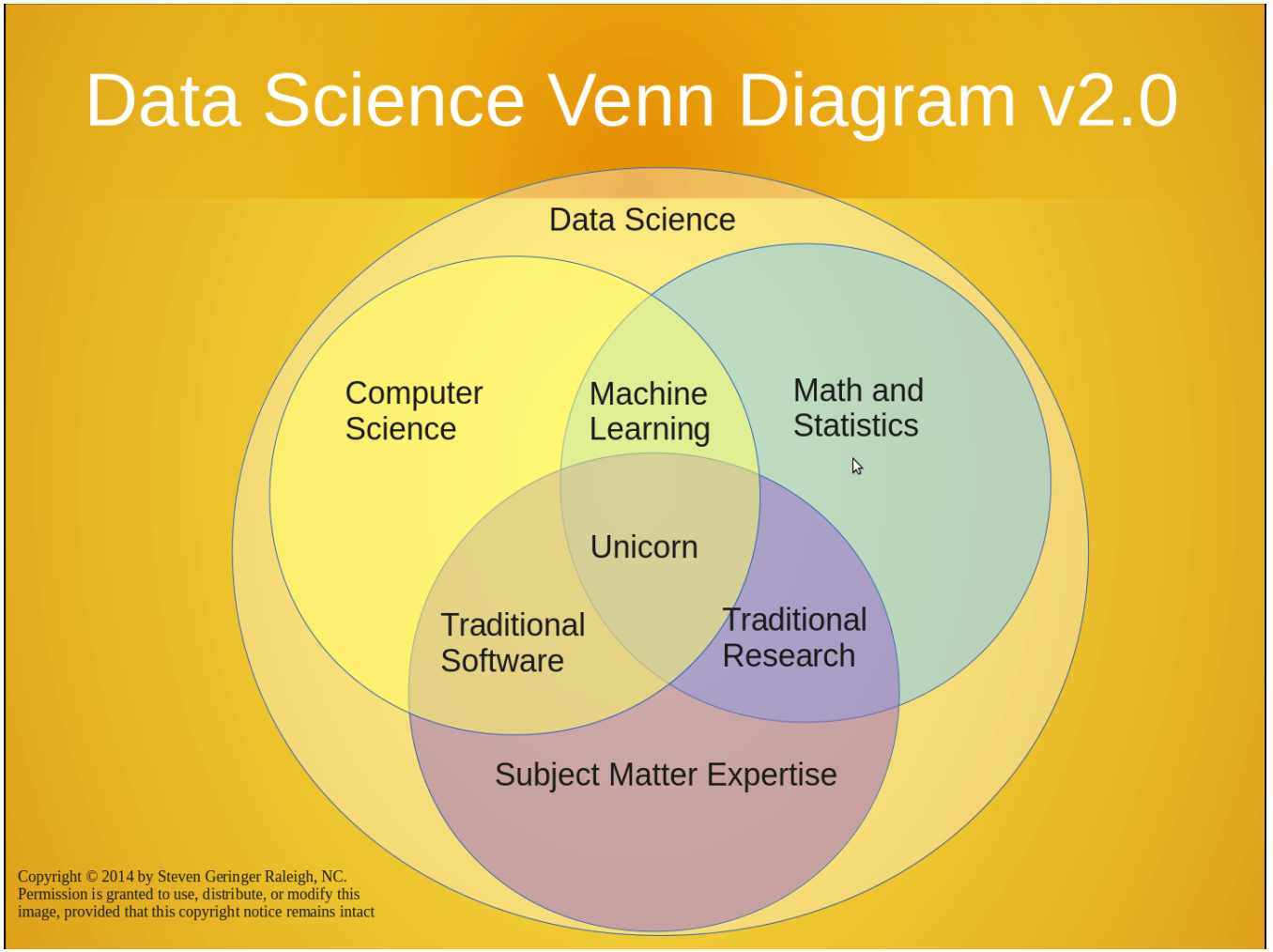
- Data science: can be seen as an umbrella term covering a wide variety of methodologies and fields related to modern data analysis.



- Figure by to Natalia Bilenko.

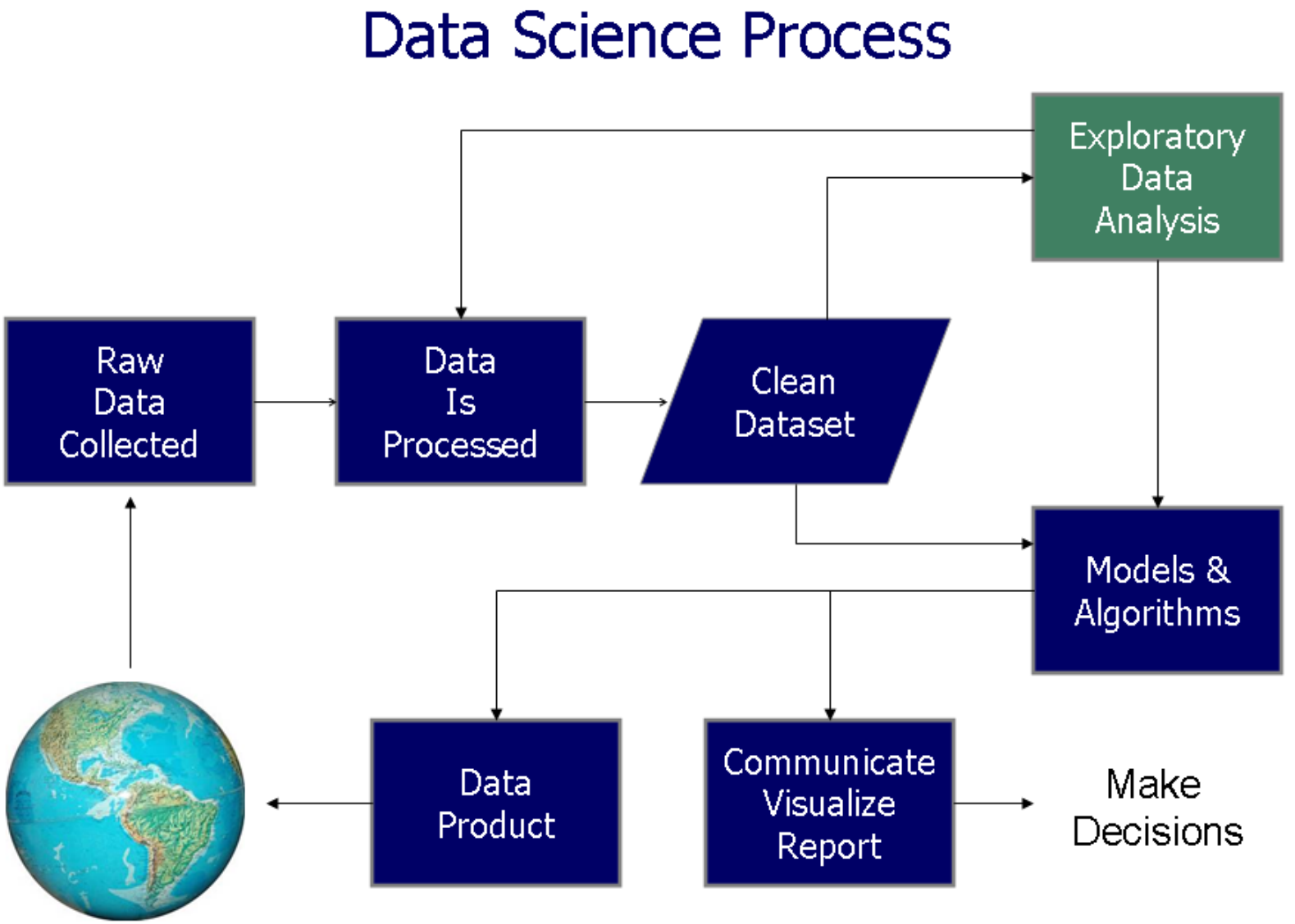
Data analysis, data science...

- Note: not realistic for a single person to be a know-it-all data scientist?



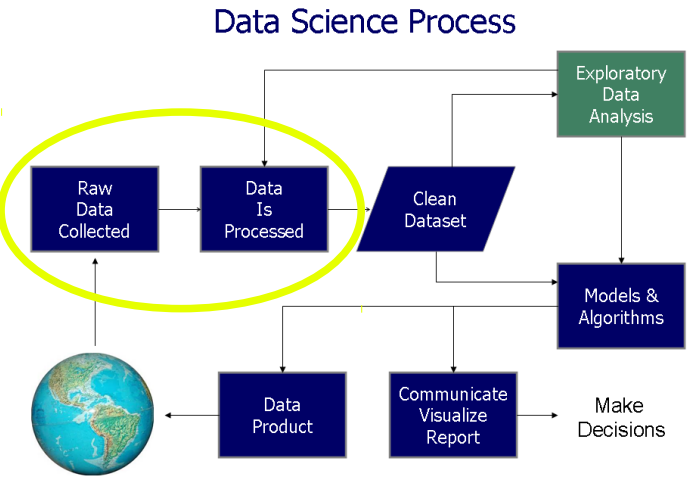
Data science process

- Figure source: Wikipedia.



Data science process: collecting and (pre)processing

- **Collecting raw data.**
  - The data is gathered from the data source in its original form (”as it is”).



- **Preprocessing the data.**
  - The data may need to be transformed into a different format in order to better facilitate storage or later processing.

- High-level data categories:
  - **Structured data**.
    - The data has a predefined fixed structure.
      - E.g. consists of an ordered set of attribute values (like an SQL row).
    - Example: data expressed as comma-separated values (CSV):

```
’Kimi Räikkönen’;’Finland’;’Ferrari’;17.10.1979
```

Structured, unstructured, semi-structured data

- Data format categories:...
  - **Structured data...**
    - Predefined structure enables direct extraction of (some) information.
  - **Unstructured data**.
    - The data lacks a predefined data model or organization.
      - Different pieces of data may have wildly different forms.
      - Extracting information requires special processing (e.g. full search).
    - Examples: natural language text, video, audio, …
  - **Semi-structured data**.
    - The data contains some description about its structure, but not all pieces of data have exactly the same form.
    - Examples: JSON, XML.

Structured, unstructured, semi-structured data...

- JSON: **J**ava**S**cript **O**bject **N**otation.
  - Attributes belonging together (an ”object”) are placed inside **{** and **}**.
  - Several objects or values may be grouped in a list of objects or values.
    - Placed inside **[** and **]**, separated by commas.
  - The value for an attribute **attr** expressed in the form **"attr": "value"**.
    - The value may also be a ”subobject” or a list.
  - Successive attributes are separated by commas.
- Example:

```
{
 [{"name": "Kimi Räikkönen",
    "team": "Ferrari",
    "birthdate": "17.10.1979",
    "country": "Finland"}
    , ... could contain also other drivers' information ... ]
}
```

Structured, unstructured, semi-structured data...

- XML: e**X**tensible **M**arkup **L**anguage.
  - The value of a given data attribute **attr**: delimited by a start tag of form <**attr**> and an end tag of form </**attr**>.
  - Attributes may also be defined inside the start tags in the form **attr=value**.

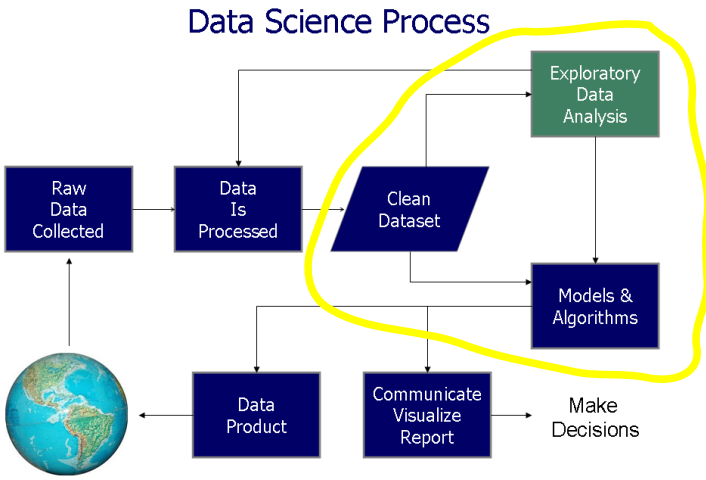
- Example:

```
<F1Drivers>
    <driver  team="Ferrari">
        <name>Kimi Räikkönen</name>
        <birthdate>17.10.1979</birthdate>
        <country>Finland</country>
    </driver>
    ... could contain also other drivers' information ...
</F1Drivers>
```

- Many tools exist (e.g. in Python) for handling CSV, XML or JSON data.

Data science process: cleaning and analyzing

- **Cleaning the data.**
  - Remove erroneous, incomplete, duplicate etc. data items that could distort the results.

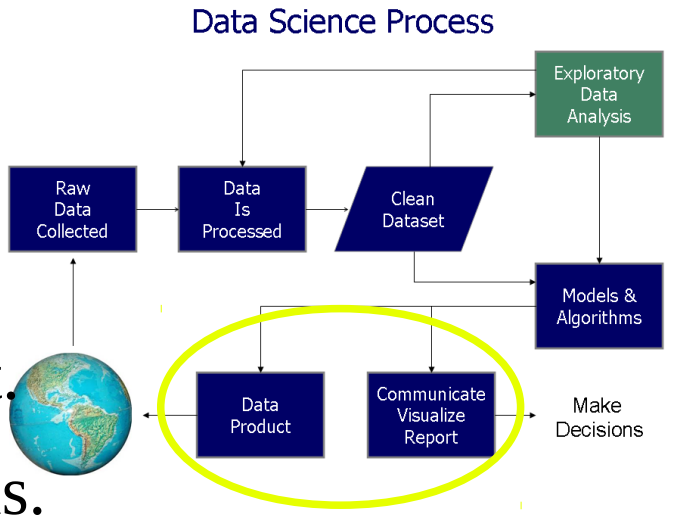


- **Exploratory data analysis.**
  - Initial analysis: may e.g. explore basic general properties of the data.
    - E.g. what kind of statistical distribution does the data seem to have?

- **Models & algorithms**:
  - The main analysis step.
  - Select, design and/or implement methods that compute the desired information/knowledge from the data.
  - Could involve e.g. statistical modeling, machine learning, data mining, information retrieval, etc.

Data science process: data product, communicating

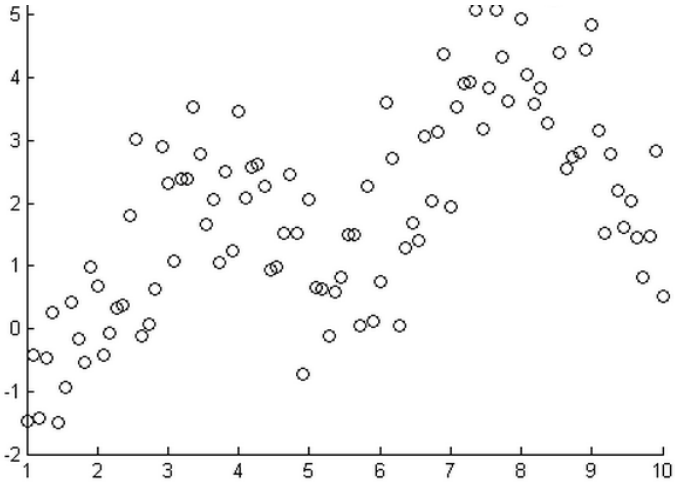
- **Data product.**
  - The (possibly reusable) program/tool that was implemented for performing the analysis.
  - Could be taken to continuous use/development.
    - Repeatedly (or constantly) performed analysis.

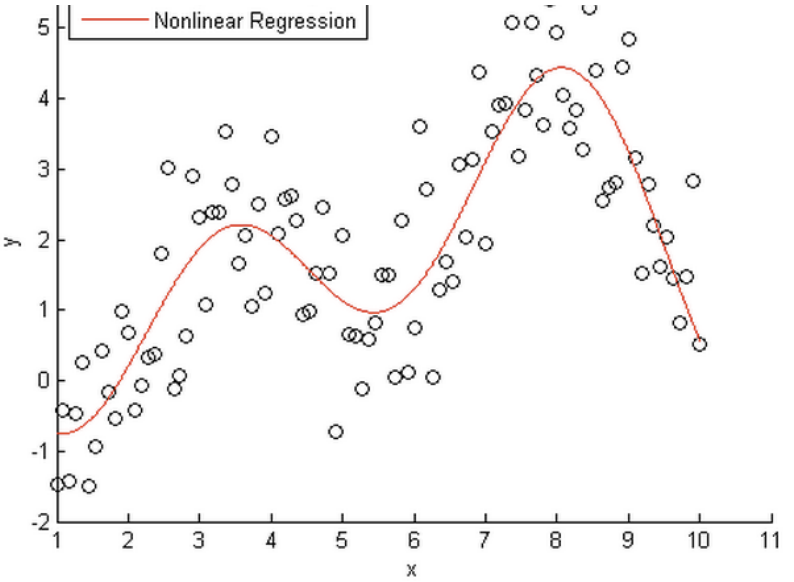


- **Communicating.**
  - Reporting the results of the analysis.
    - Could e.g. involve visualization.

Machine learning, data mining...

- Data mining.
  - Computational data analysis to find interesting properties from data.
  - Emphasis is on analyzing current data.
- Examples of common general data mining tasks:
  - Regression analysis.
    - Form a statistical model that describes the data (shape, trend etc.).
      - ”Function/curve fitting”.

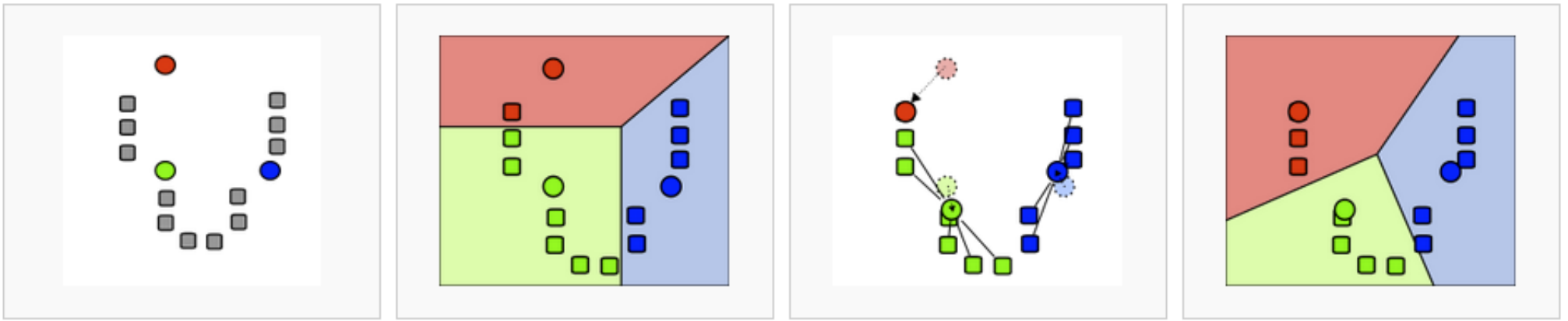




- Source of images: MathWorks.

Machine learning, data mining...

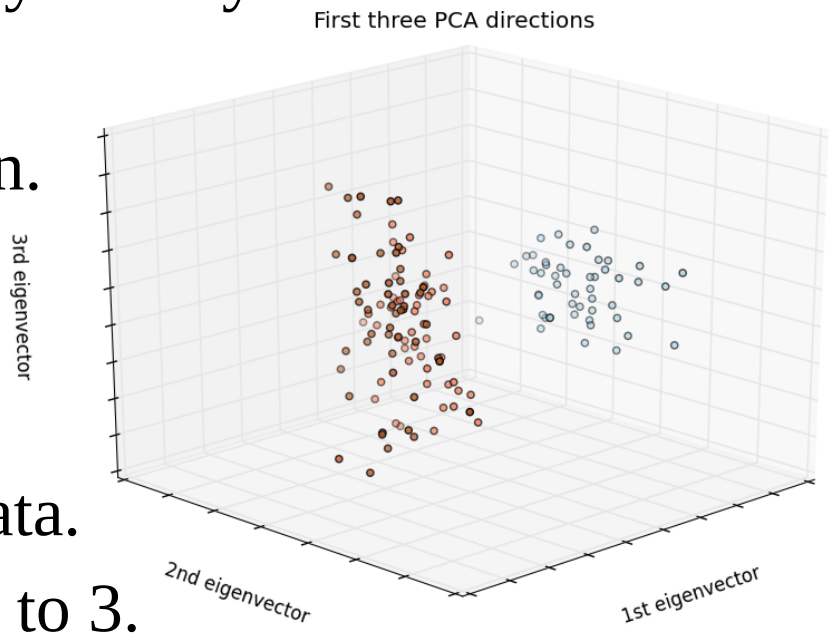
- Examples of common general data mining tasks:...
  - Clustering.
    - Group similar/related data items together (into ”clusters”).
    - Example: *k*-means clustering.
      - Divide a given data point set into *k* clusters.
      - Each point is assigned into the cluster with the nearest centroid.



- Source of figures: Wikipedia.

Machine learning, data mining...

- Examples of common general data mining tasks:...
  - Reduce the dimensionality of the data by principal component analysis.
    - Typical motivations:
      - Simplifying visualisation of data that originally has > 3 dimensions.
      - May reduce computational resource requirements in later analysis.
    - Determines maximally informative axes for the reduced dimensions.
      - But note that reducing dimensionality usually loses information!
    - An example using 4-dimensional "Iris"-data provided e.g. in scikit-learn.
      - An example of a data item: [5.4, 3.9, 1.7, 0.4]
      - The figure shows a 3D-plot of the data.
        - The original 4 dimensions reduced to 3.



Machine learning, data mining...

- Machine learning.
  - A class of computational analysis methods that build (and update) a general data model based on known data.
  - Emphasis is on analyzing future data.
    - Use experience from existing data in order to improve the analysis of future (yet unseen) data.
  - *”A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”* -T. Mitchell.
- Machine learning tasks often concern some type of classification.
  - Given a classification for current data, classify also new data items.
    - What is the most probable class for a previously unseen item?

Machine learning, data mining...

- Machine learning...
  - Typically involves two aspects:
    - Training (or fitting).
      - Build a model by using existing/selected training data with known classification. (Neural networks, support vector machine, etc.)
    - Prediction.
      - Use the model to classify previously unknown data.
  - About the terminology used in the figure:
    - ”input” means data.
    - ”label” corresponds to our term ”class” (in terms of classifying data).

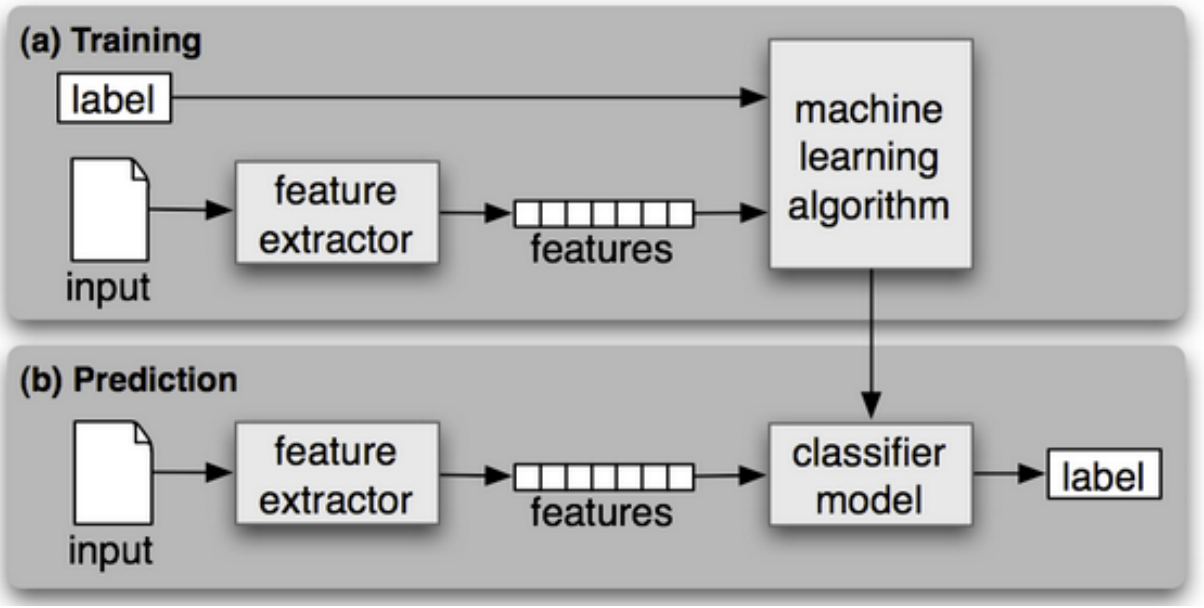
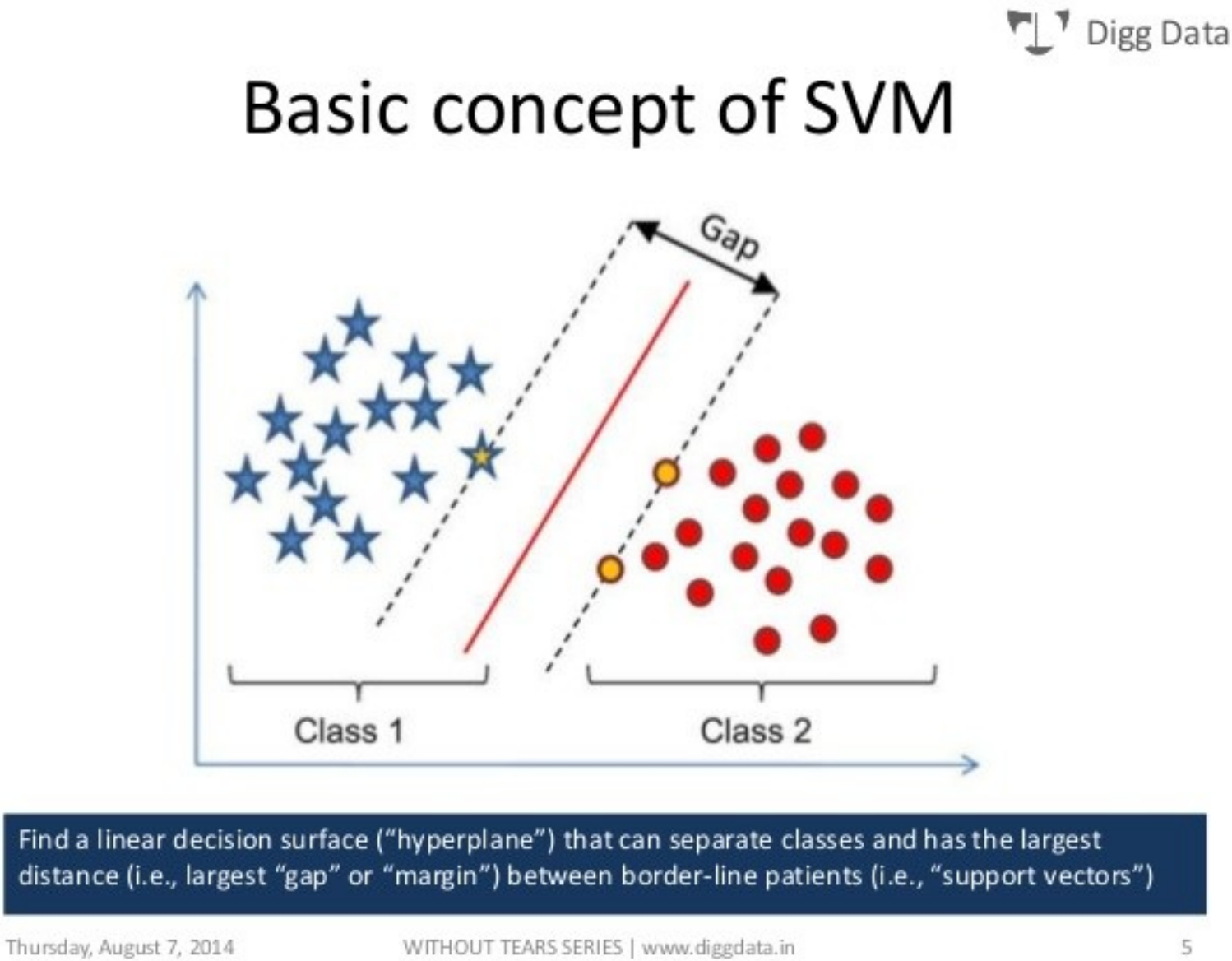


Image source: nltk.org.

Machine learning, data mining...

- An example: the working principle of a Support Vector Machine



- Image source: to www.diggdata.in