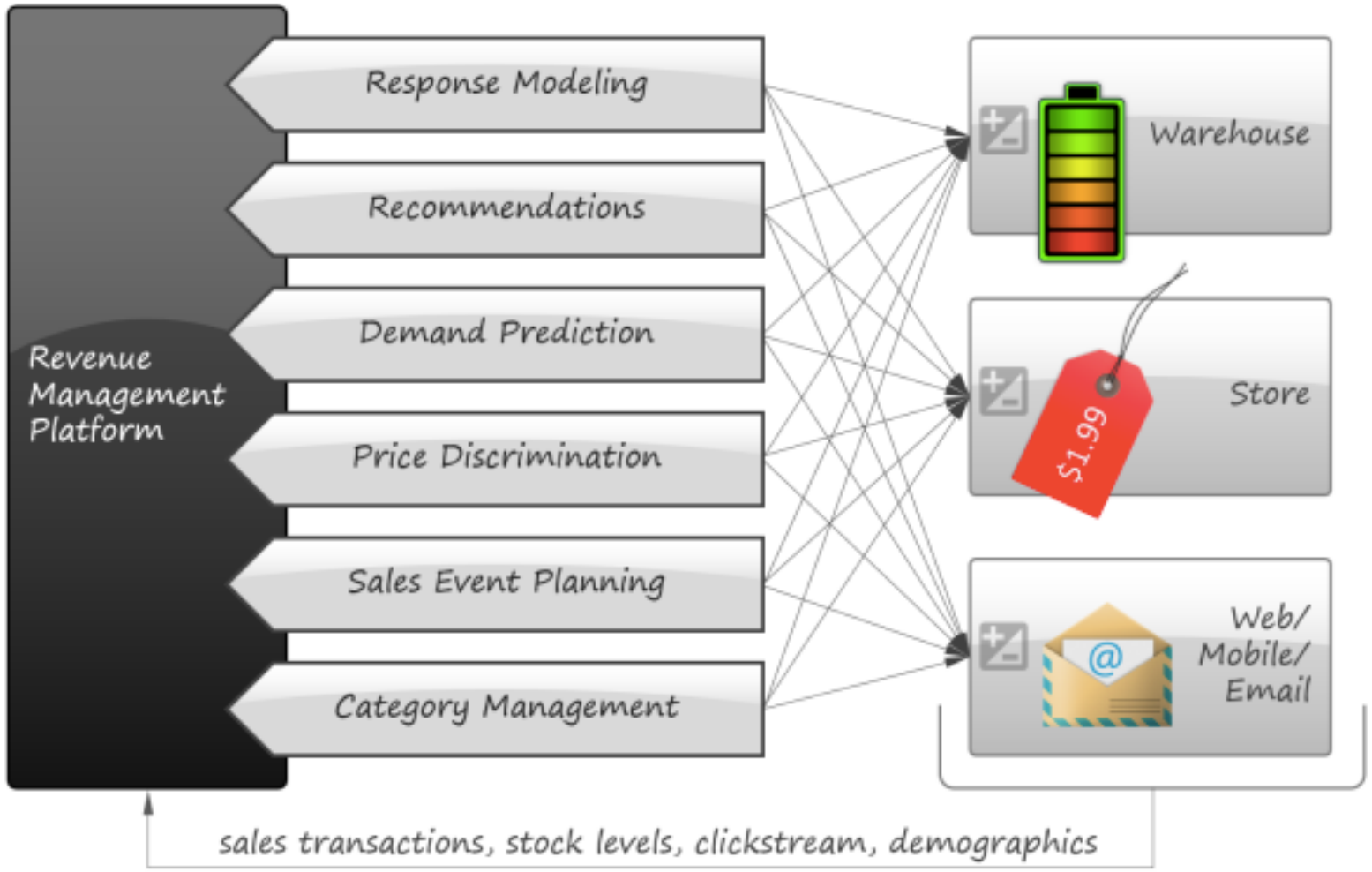
Data Mining with IPython Notebook & D3.js.

By Intellij System Solution Sdn. Bhd.

Applications



What is the course about

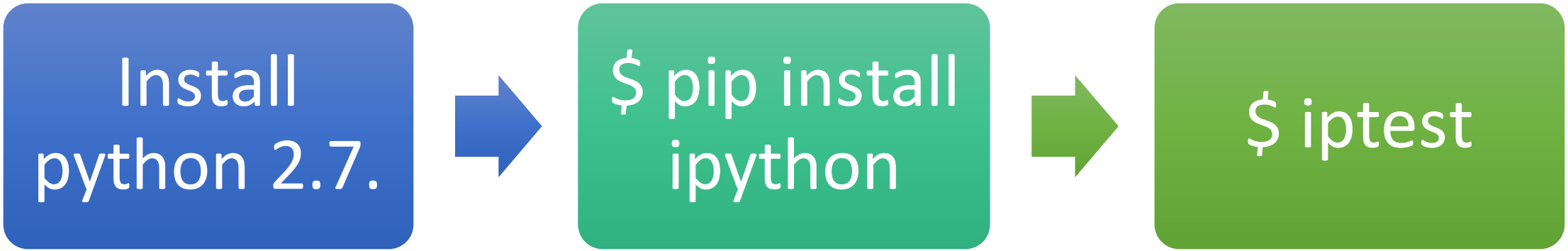
Data mining	• Using data to make predictions
Ipython notebook	• IPython Notebook is a web-based interactive computational environment for creating IPython notebooks. An IPython notebook is a JSON document containing an ordered list of input/output cells which can contain code, text, mathematics, plots and rich media.
D3.js	• JavaScript library for producing dynamic, interactive data visualizations in web browsers. It makes use of the widely implemented SVG, HTML5, and CSS standards. It is the successor to the earlier Protovis framework.

Prerequisites for this course

Github accounts.

Willingness and perseverance.

Setup Environment for IPython Notebook



UI Components IPython Notebook

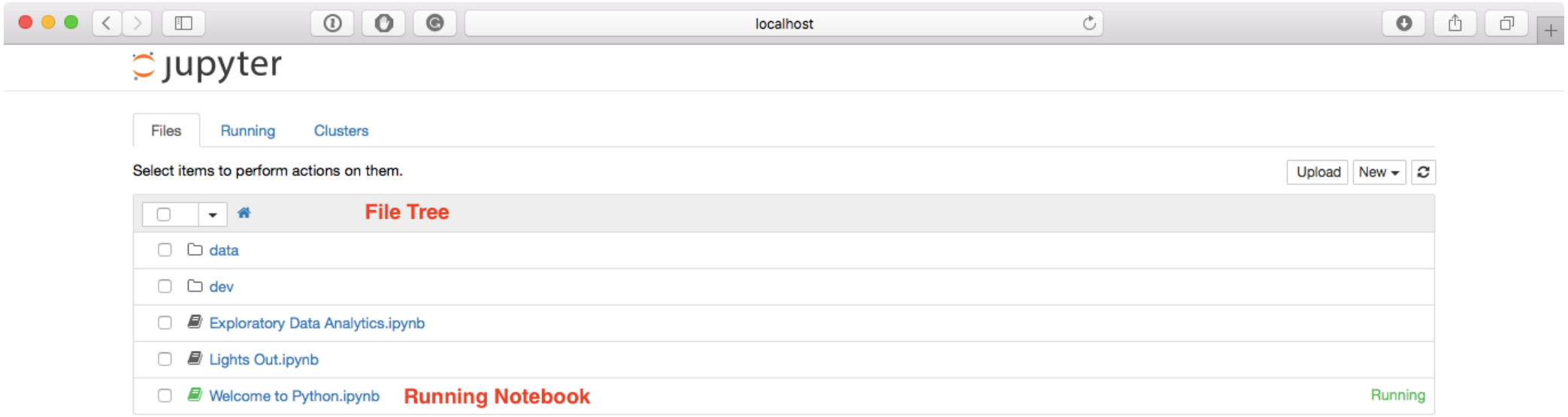
Notebook Dashboard

Notebook Editor

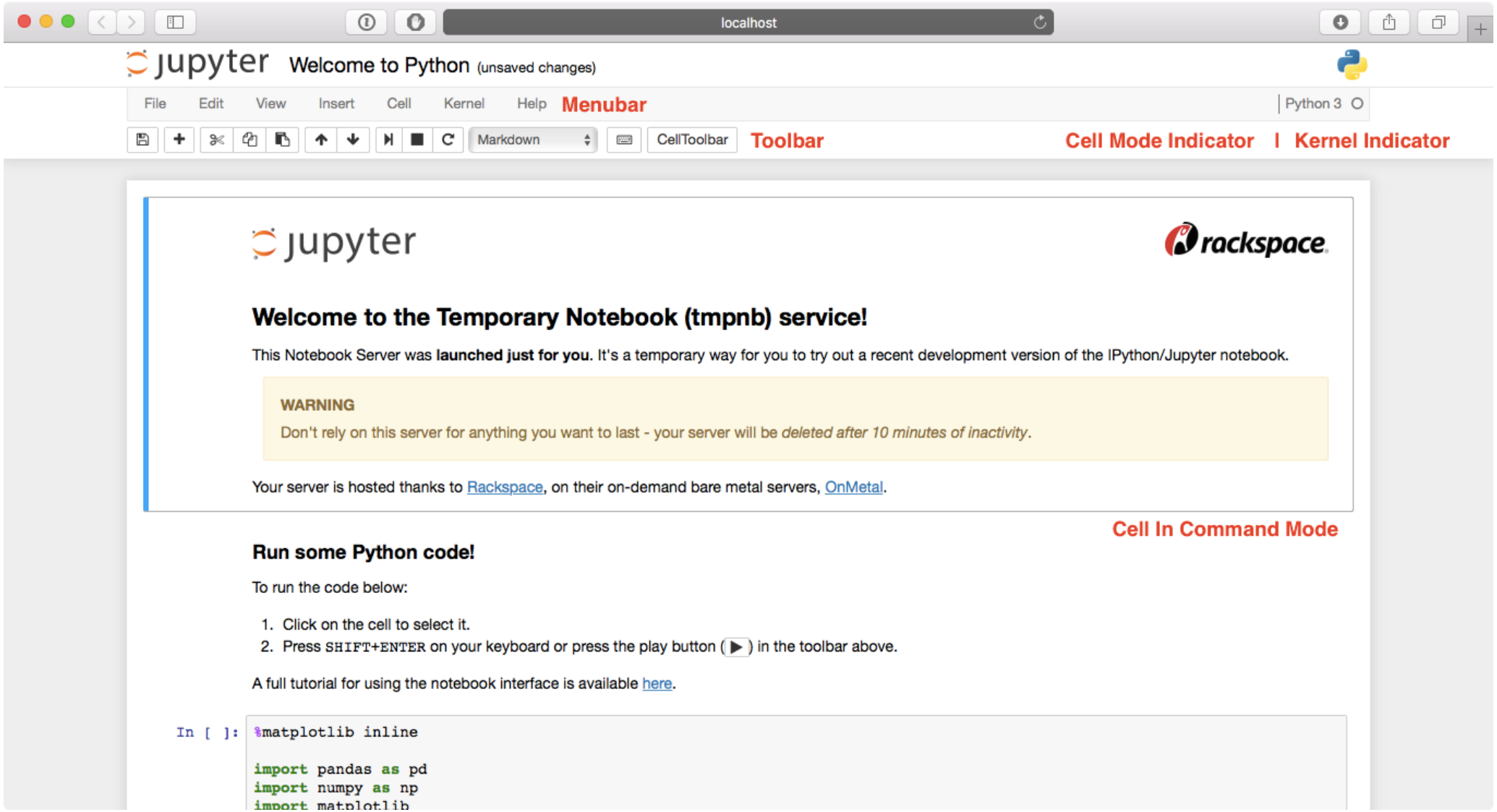
Edit Mode and Notebook Editor

File Editor

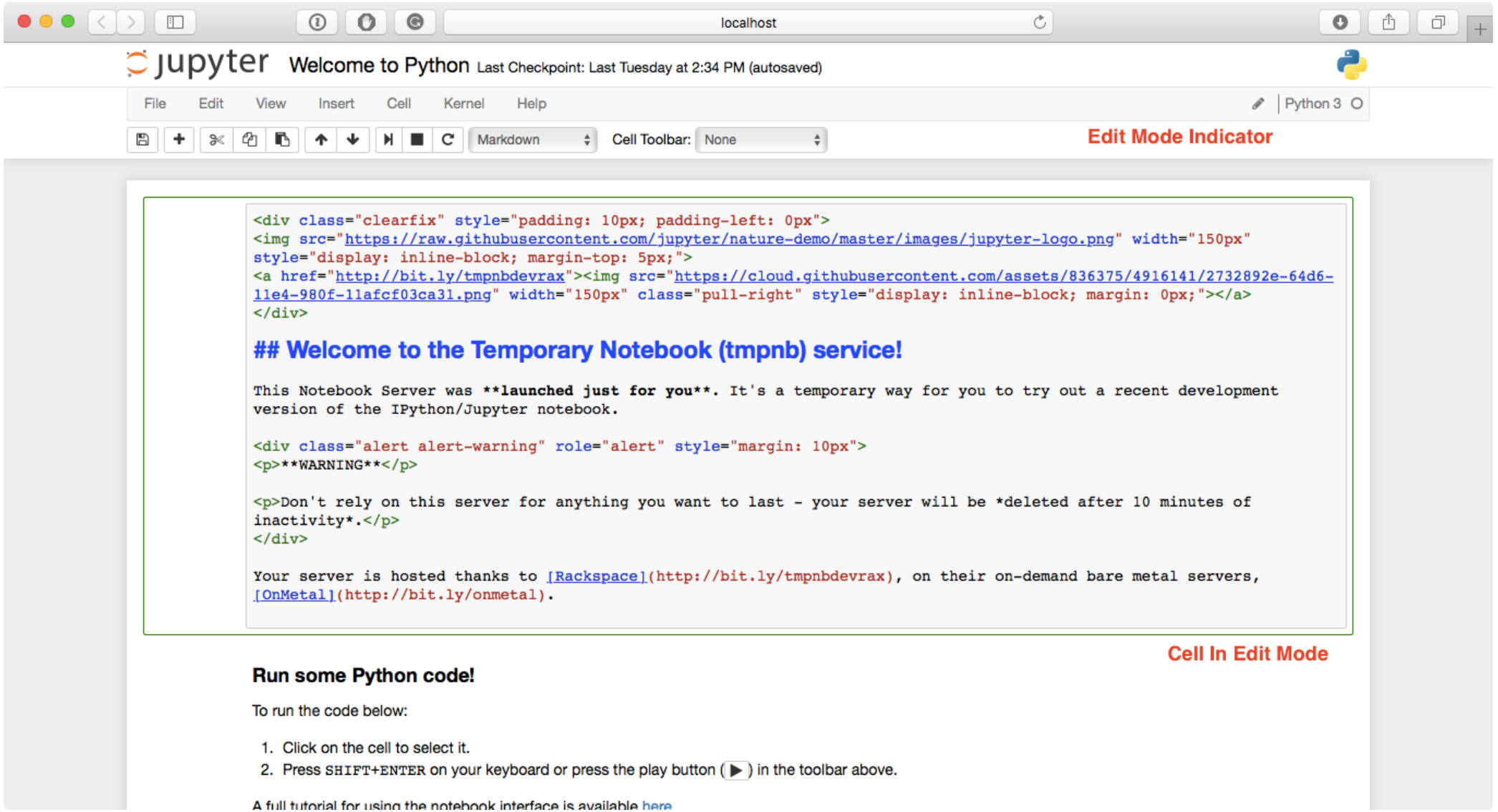
Notebook Dashboard



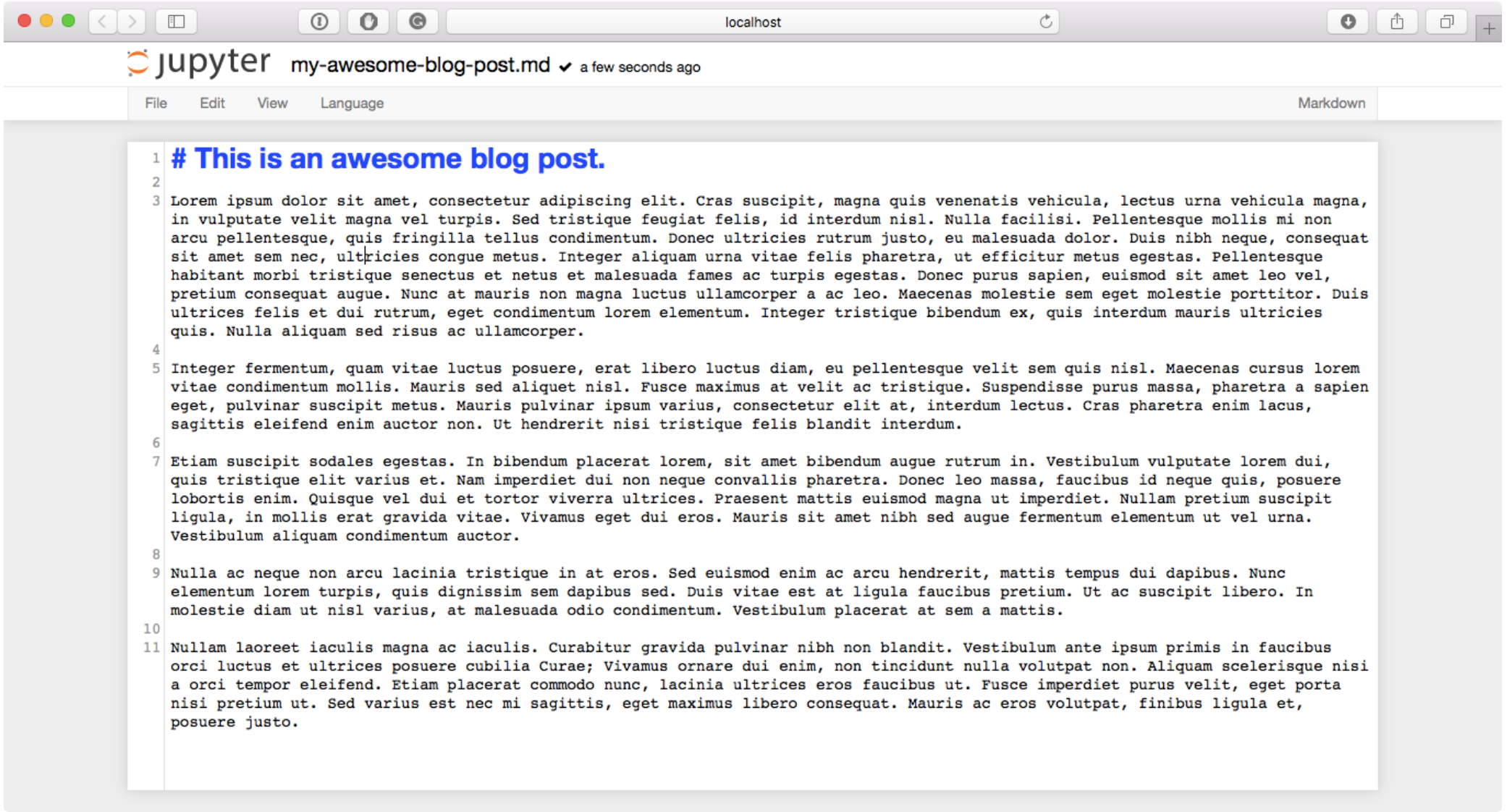
Notebook Editor



Edit Mode and Notebook Editor



File Editor


localhost
jupyter
my-awesome-blog-post.md
a few seconds ago
File
Edit
View
Language
Markdown
This is an awesome blog post.
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras suscipit, magna quis venenatis vehicula, lectus urna vehicula magna, in vulputate velit magna vel turpis. Sed tristique feugiat felis, id interdum nisl. Nulla facilisi. Pellentesque mollis mi non arcu pellentesque, quis fringilla tellus condimentum. Donec ultricies rutrum justo, eu malesuada dolor. Duis nibh neque, consequat sit amet sem nec, ultricies congue metus. Integer aliquam urna vitae felis pharetra, ut efficitur metus egestas. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec purus sapien, euismod sit amet leo vel, pretium consequat augue. Nunc at mauris non magna luctus ullamcorper a ac leo. Maecenas molestie sem eget molestie porttitor. Duis ultrices felis et dui rutrum, eget condimentum lorem elementum. Integer tristique bibendum ex, quis interdum mauris ultricies quis. Nulla aliquam sed risus ac ullamcorper.
Integer fermentum, quam vitae luctus posuere, erat libero luctus diam, eu pellentesque velit sem quis nisl. Maecenas cursus lorem vitae condimentum mollis. Mauris sed aliquet nisl. Fusce maximus at velit ac tristique. Suspendisse purus massa, pharetra a sapien eget, pulvinar suscipit metus. Mauris pulvinar ipsum varius, consectetur elit at, interdum lectus. Cras pharetra enim lacus, sagittis eleifend enim auctor non. Ut hendrerit nisi tristique felis blandit interdum.
Etiam suscipit sodales egestas. In bibendum placerat lorem, sit amet bibendum augue rutrum in. Vestibulum vulputate lorem dui, quis tristique elit varius et. Nam imperdiet dui non neque convallis pharetra. Donec leo massa, faucibus id neque quis, posuere lobortis enim. Quisque vel dui et tortor viverra ultrices. Praesent mattis euismod magna ut imperdiet. Nullam pretium suscipit ligula, in mollis erat gravida vitae. Vivamus eget dui eros. Mauris sit amet nibh sed augue fermentum elementum ut vel urna. Vestibulum aliquam condimentum auctor.
Nulla ac neque non arcu lacinia tristique in at eros. Sed euismod enim ac arcu hendrerit, mattis tempus dui dapibus. Nunc elementum lorem turpis, quis dignissim sem dapibus sed. Duis vitae est at ligula faucibus pretium. Ut ac suscipit libero. In molestie diam ut nisl varius, at malesuada odio condimentum. Vestibulum placerat at sem a mattis.
Nullam laoreet iaculis magna ac iaculis. Curabitur gravida pulvinar nibh non blandit. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Vivamus ornare dui enim, non tincidunt nulla volutpat non. Aliquam scelerisque nisi a orci tempor eleifend. Etiam placerat commodo nunc, lacinia ultrices eros faucibus ut. Fusce imperdiet purus velit, eget porta nisi pretium ut. Sed varius est nec mi sagittis, eget maximus libero consequat. Mauris ac eros volutpat, finibus ligula et, posuere justo.

Demo

http://bit.ly/2rNIZLY

Exploring datasets

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attributes

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color* = {*auburn, black, blond, brown, grey, red, white*}
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size* = {*small, medium, large*}, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Comparison Supervised and Unsupervised Learning

Supervised	Unsupervised
Classification	Clustering
• known number of classes • based on a training set • used to classify future observations	• unknown number of classes • no prior knowledge • used to understand (explore) data

Classifier

- Naïve Bayes Classifier Algorithm.
- K Means Clustering Algorithm.
- Support Vector Machine Algorithm.
- Apriori Algorithm.
- Linear Regression.
- Logistic Regression.
- Artificial Neural Networks.
- Random Forests.

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

  Note: *n* is sample size and *N* is population size.

  - Weighted arithmetic mean:

  $$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

  - Trimmed mean: chopping extreme values

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right) width$$

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)
  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

Detailed Accuracy By Class

TP Rate

- http://bit.ly/2qj5BHP

FP Rate

- http://bit.ly/2qn1vdN

Precision & Recall

- http://bit.ly/2piTCZv

F-Measure

- http://bit.ly/2qmYdqU

ROC Area

- http://bit.ly/1In2v72

Confusion Matrix

- http://bit.ly/25JMSDF

Building a classifier

Lab 1

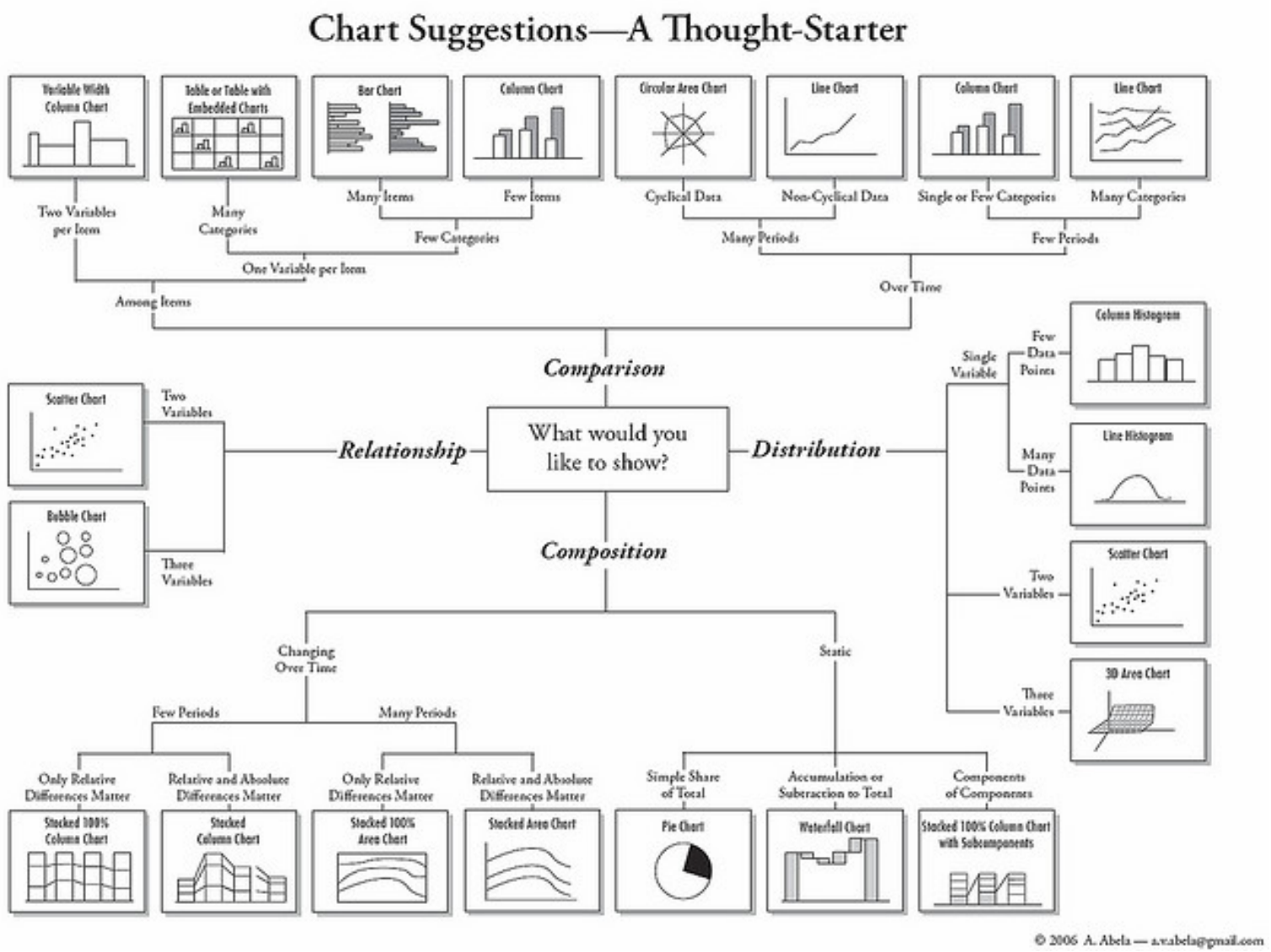
Data treatment

- Variable Identification
- Univariate Analysis
- Bivariate Analysis
- Missing Values Imputation
- Outliers Treatment
- Variable transformation
- Variable / Feature creation

Data treatment

Lab 2

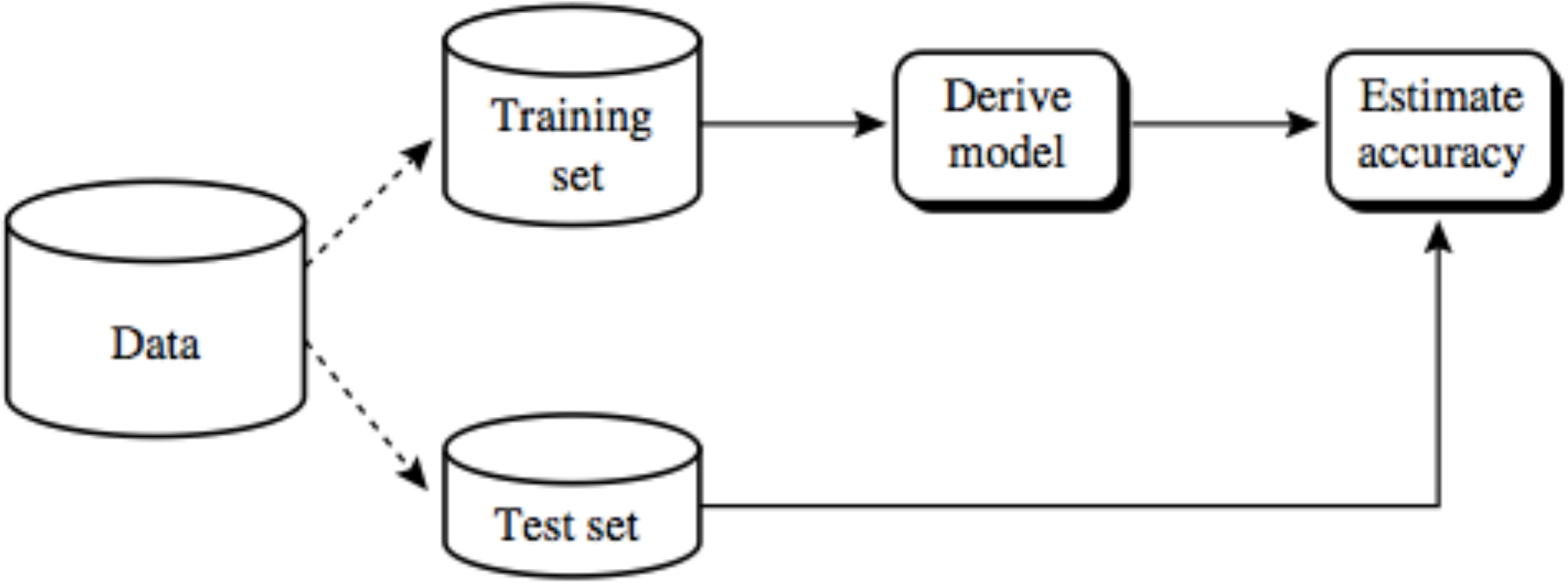
Visualizing your data



Visualizing your data

Lab 3

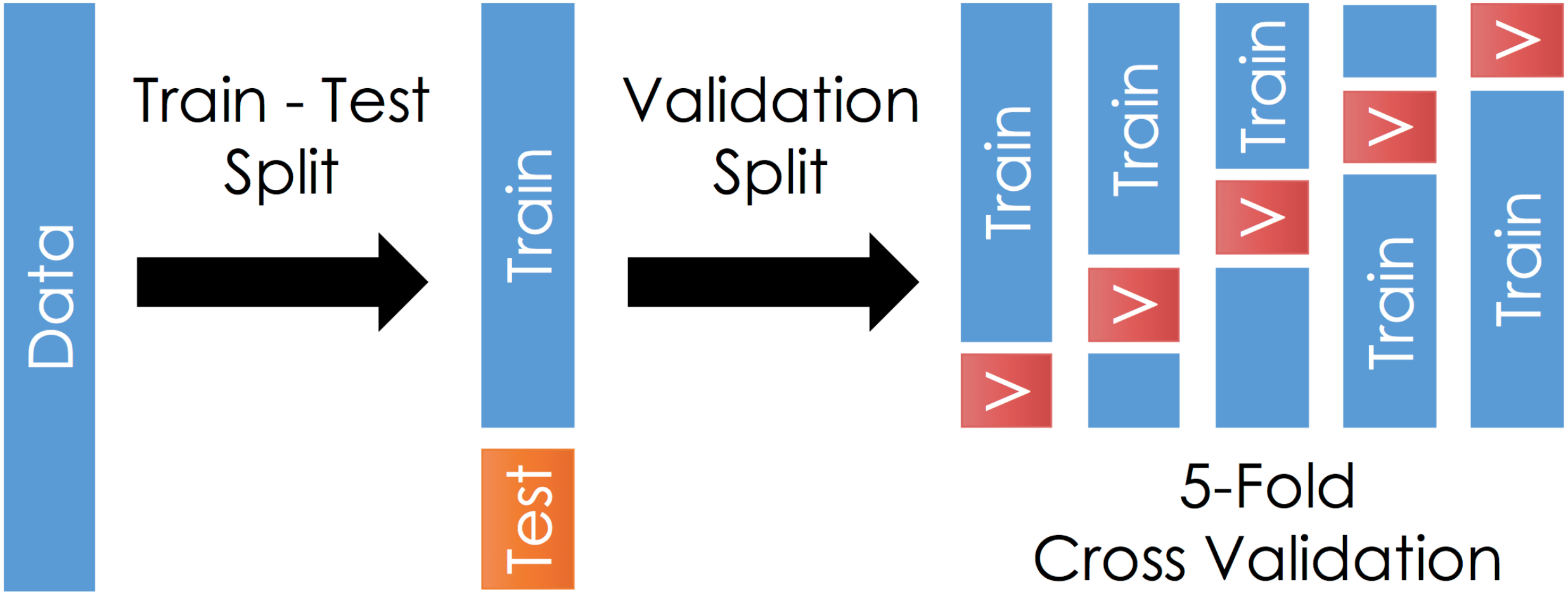
Training & Testing



Training & Testing

Lab 4

Cross Validation



Cross Validation

Lab 5

Random Forest

- What is Random Forest?
- How does it work?
- Advantages of Random Forest.
- Disadvantages of Random Forest.

Naïve Bayes Classifier Algorithm

- What is Naïve Bayes Classifier Algorithm?
- How does it work?
- Advantages of Naïve Bayes Classifier Algorithm.
- Disadvantages of Naïve Bayes Classifier Algorithm.

K-Nearest Neighbors algorithm

- What is K-Nearest Neighbors algorithm?
- How does it work?
- Advantages of K-Nearest Neighbors Algorithm.
- Disadvantages of K-Nearest Neighbors Algorithm.

Linear Regression

- What is Linear Regression?
- How does it work?
- Advantages of Linear Regression.
- Disadvantages of Linear Regression.

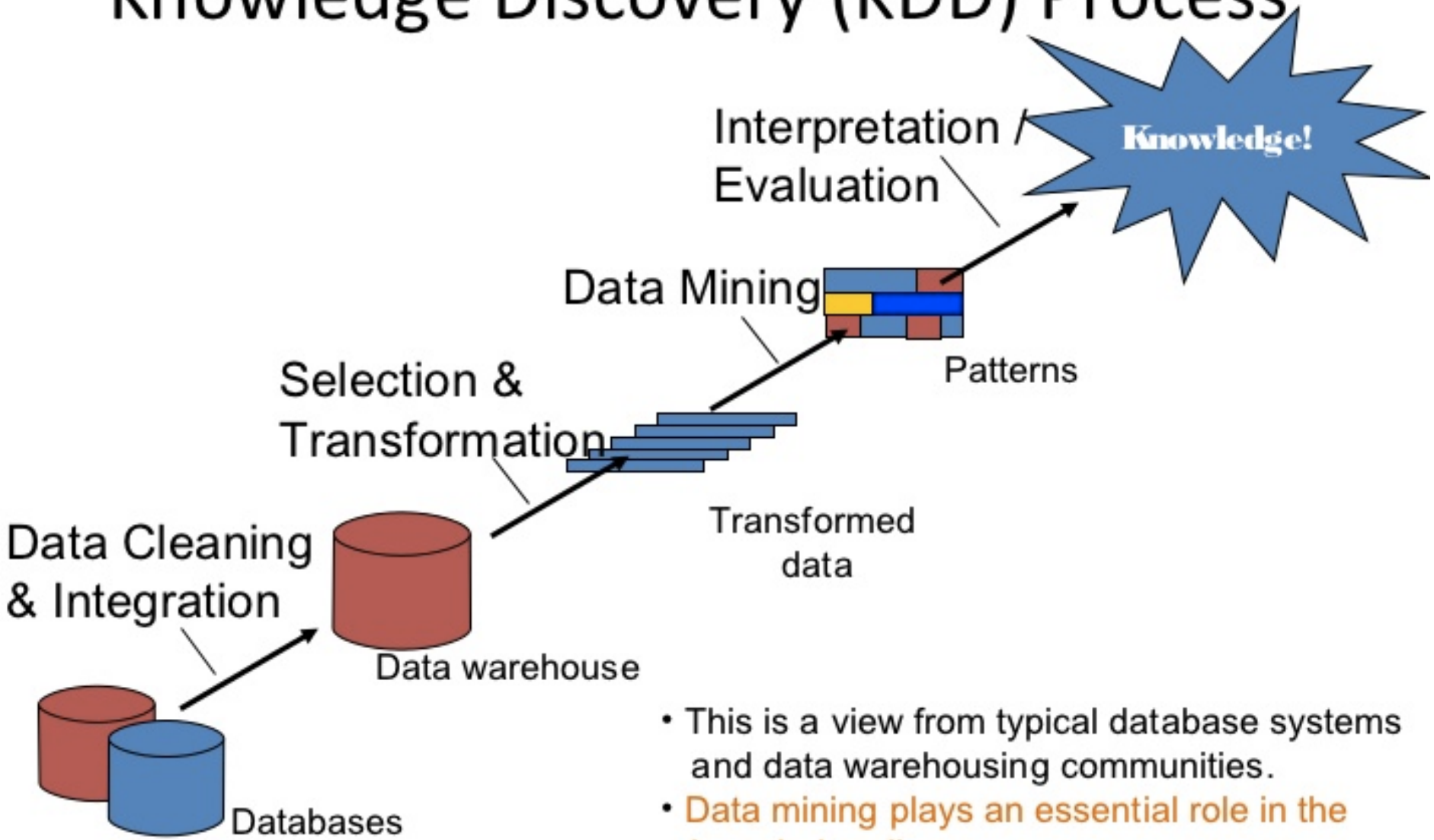
Logistic Regression

- What is Logistic Regression?
- How does it work?
- Advantages of Logistic Regression.
- Disadvantages of Logistic Regression.

Support Vector Machine

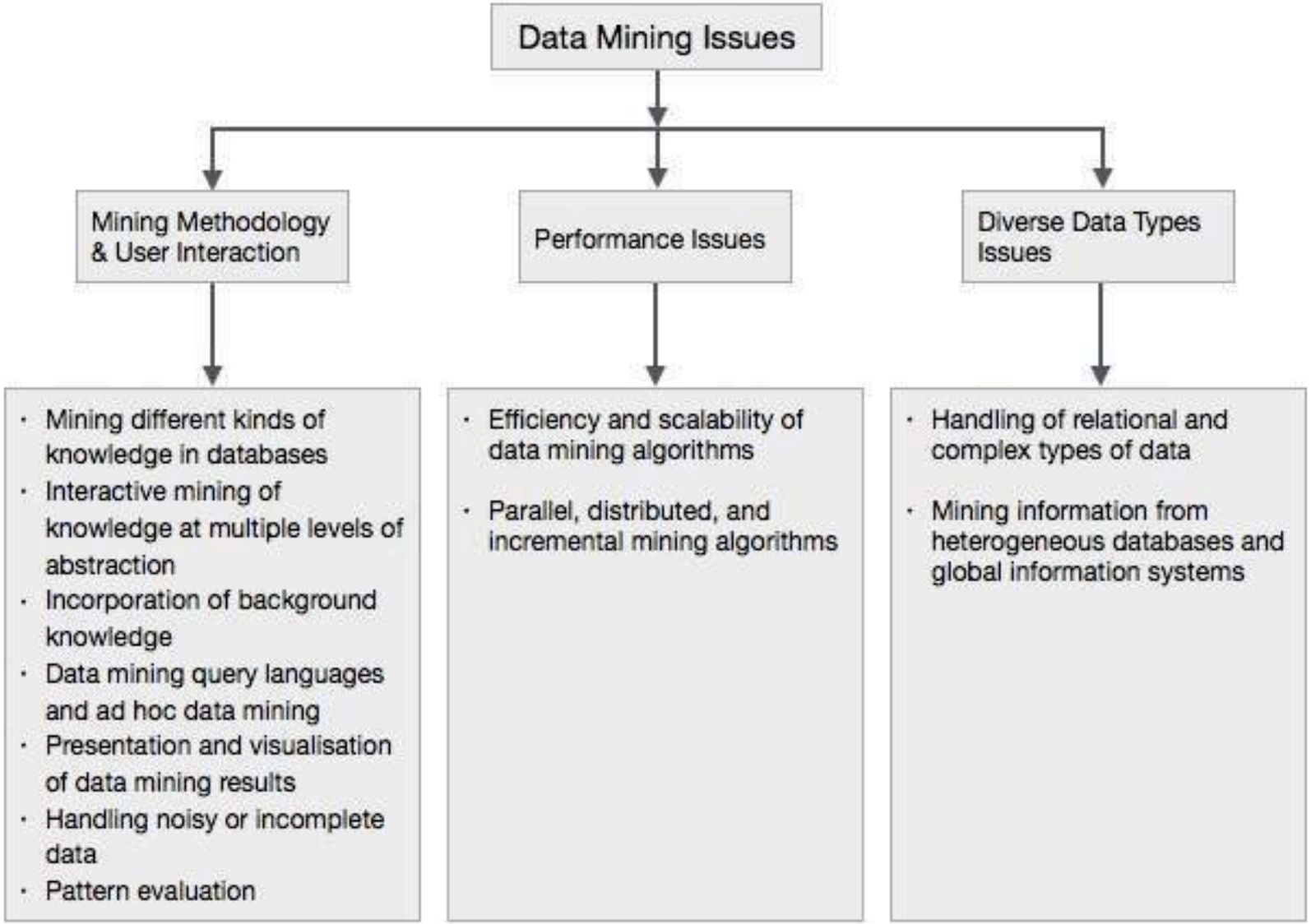
- What is Support Vector Machine?
- How does it work?
- Advantages of Support Vector Machine.
- Disadvantages of Support Vector Machine.

Knowledge Discovery (KDD) Process



- This is a view from typical database systems and data warehousing communities.
- Data mining plays an essential role in the knowledge discovery process.

Top Challenges



Data Mining Ethics

Privacy

Trust Between Customers and Business

Topics we not covered in this course

Natural Language Processing

Image Processing

Questions?

Contact Us

Email: nurdin@intellij.my

Phone: +6 011 2625 2058

Address:
E-28-1, Jalan Multimedia 7/AG
City Park, i-City
40000 Shah Alam, Selangor
Malaysia.