

Supervised Learning From Examples

Dr. Lotfi ben Othmane
University of North Texas

Project Selection

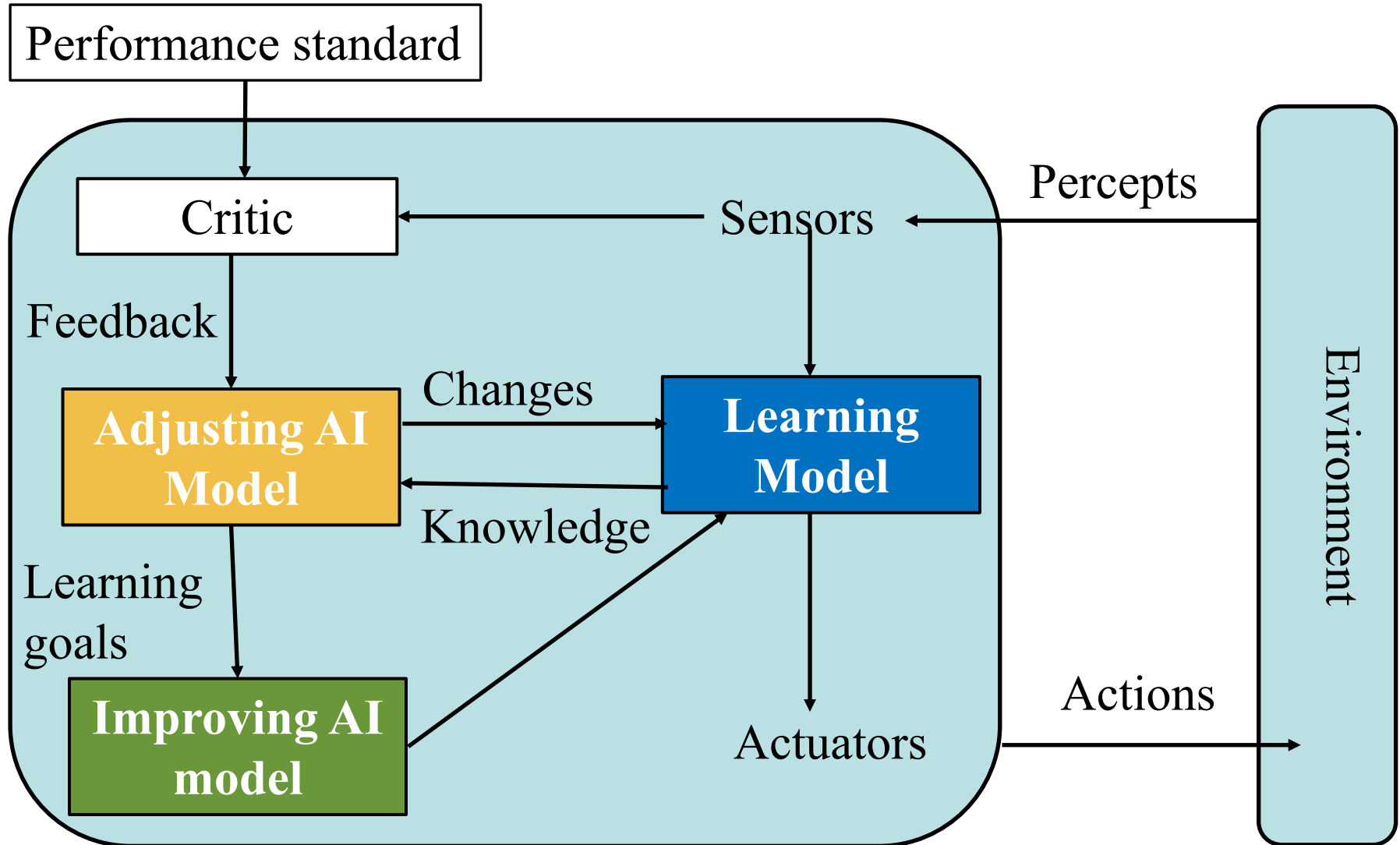
- Get a partner
- Choose a topic
 - Hardware-based
 - Cyber-security
 - Language processing
 - Etc.
- Look for open sources on Github with your criteria

Definition of AI

- AI is concerned with rational actions given
 - Objectives
 - Intractability



Learning Agents



Association of Percepts to Outputs



70-80 F



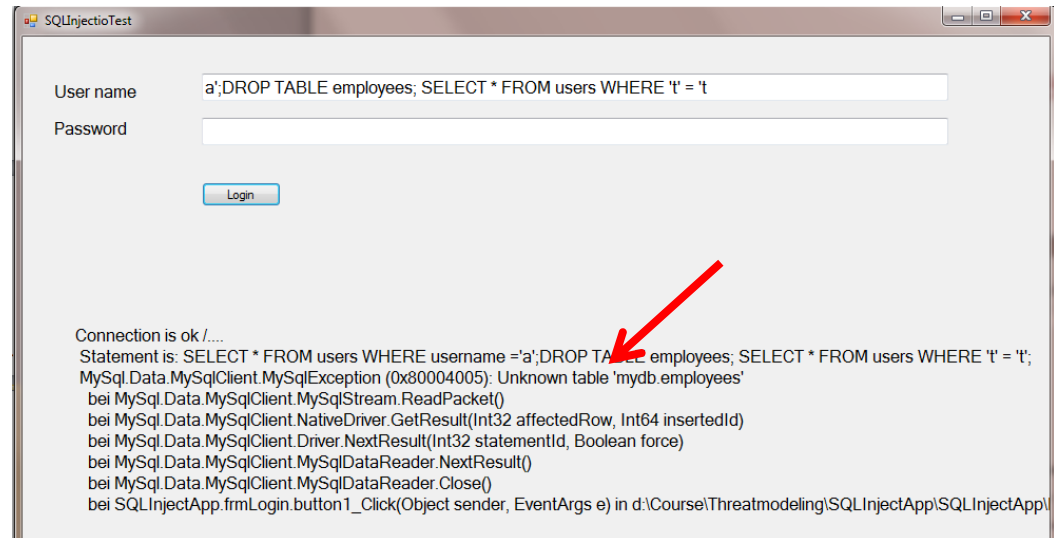
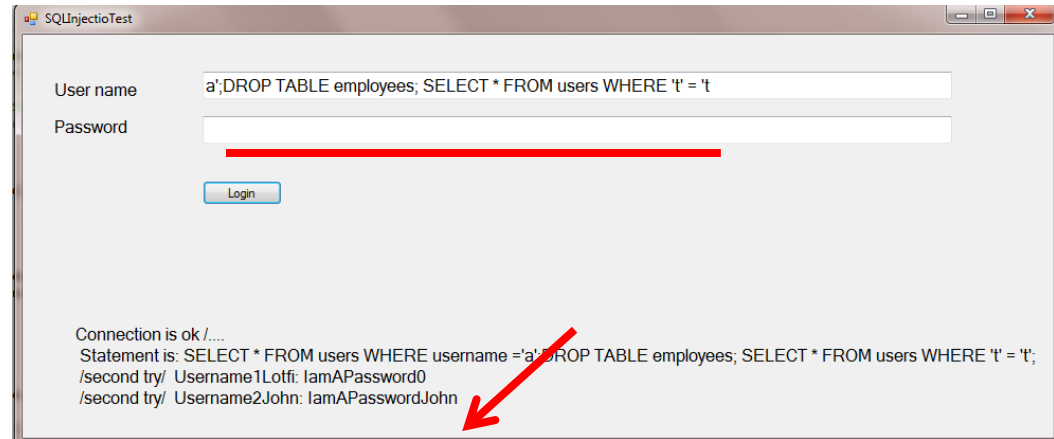
25-35 F

Now- Cost of Fixing SQL Injection Attack

Original query: *SELECT * FROM users*

WHERE username = ' + variable + ','

Malicious data: **a';DROP TABLE employees; SELECT * FROM users WHERE 't' = 't**



Now- Cost of Fixing Vulnerabilities

Assign the vulnerabilities to three experts



10 minutes



15 minutes



12 minutes

Time average = $(10+15+12)/3 = 12.3$ minutes

Now- Cost of Fixing Vulnerabilities



That is not true



Associate Percepts to Output

Vulnerabilities	Fixing time (min)
Dead code (unused methods)	2.6
Lack of authorization check	6.9
Unsafe threading	8.5
XSS (stored)	9.6
SQL injection	12.3

Now- Cost of Fixing Vulnerabilities



Now- Cost of Fixing Vulnerabilities



I am planning a new project. How much should I plan for you to fix vulnerabilities?



Should I ask for \$10k, \$200k or \$1 million?

Now- Cost of Fixing Vulnerabilities

Help me: How much should I ask?



Now- Cost of Fixing Vulnerabilities

$$\text{Time} = f(x_i, x_l, x_k)$$

Identify the
research goal

Collect data

Prepare the
data

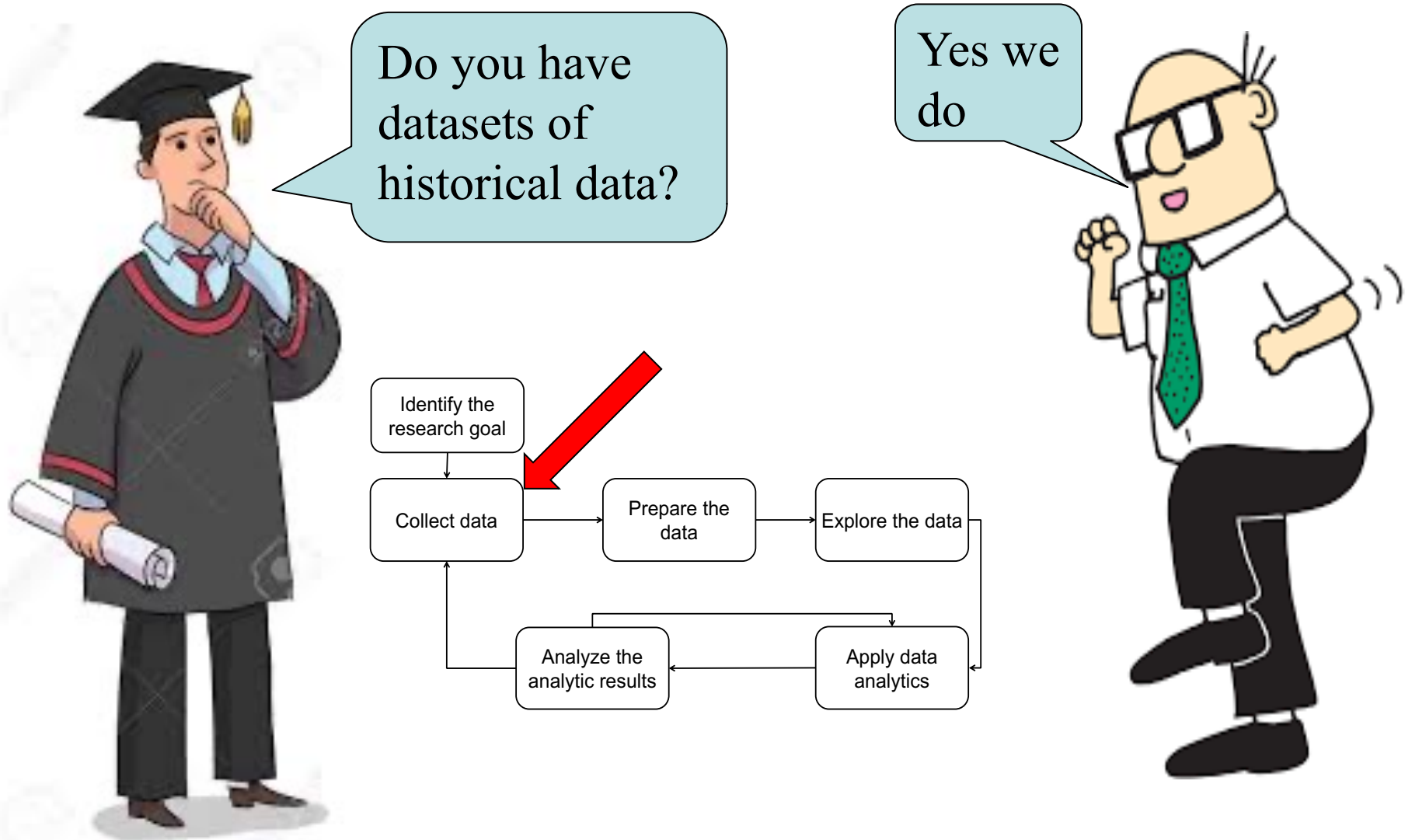
Explore the data

Analyze the
analytic results

Apply data
analytics



Now- Cost of Fixing Vulnerabilities



Datasets for Cost for Fixing Vulnerabilities

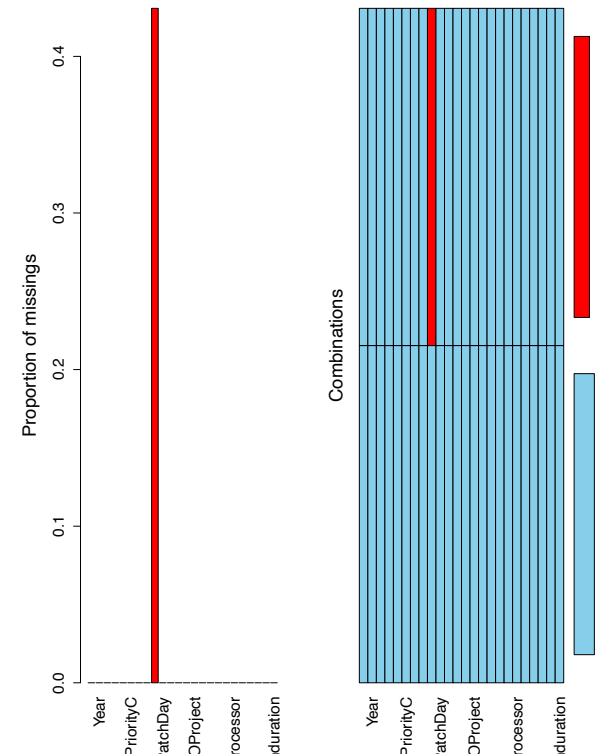
Table 4 List of the attributes of Java and C++ issue fixing (Data source 2)

Attribute	Description
Date_found	Date on which the issue was found
Date_solved	Date on which the issue was closed
Vulnerability_name	Vulnerability types such as memory corruption and buffer overflow
Scan_source	Tool that performed the scan, i.e., Coverity (for C++ code) or Fortify (for Java code)
Project_name	Project identifier
Folder_name	Indicates the required behavior of the developer toward the issue, e.g., must fix, fix one of the sets, optional, etc
Scan_status	Status of the issues, i.e., new, updated, removed, and reintroduced (i.e., removed but reopened). It allows to identify whether the issue is addressed or not, and is a false positive or not
Vulnerability_count	Number of issues of the same vulnerability found at once
Priority	The priority of fixing the vulnerability. Range: 1 to 4, with 1 highest and 4 lowest priority

Prepare the Data

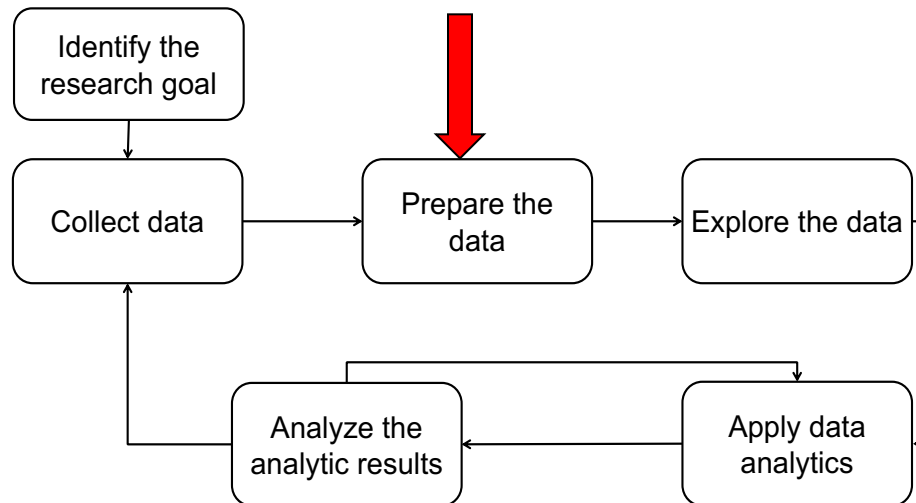
Data cleaning

1. Identify not related rows
2. Ensure that only one data collection technique is used for each attribute
3. Check missing data
4. Identification of outliers
5. Exclude records that have invalid data, e.g., invalid vulnerability type such as “not assigned”
6. Exclude records of open issues



Data Preparation – cont.

- Data transformation
 1. Compute factors from other data,
 - E.g., compute duration from end and start date
 2. Derive categorical information from text

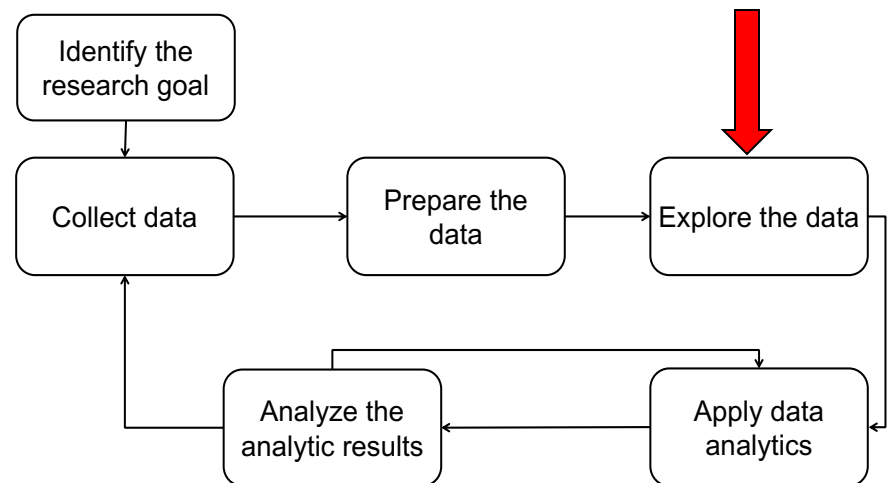


Data Exploration Techniques

- Deriving knowledge from data requires understanding of patterns and hidden facts from data.

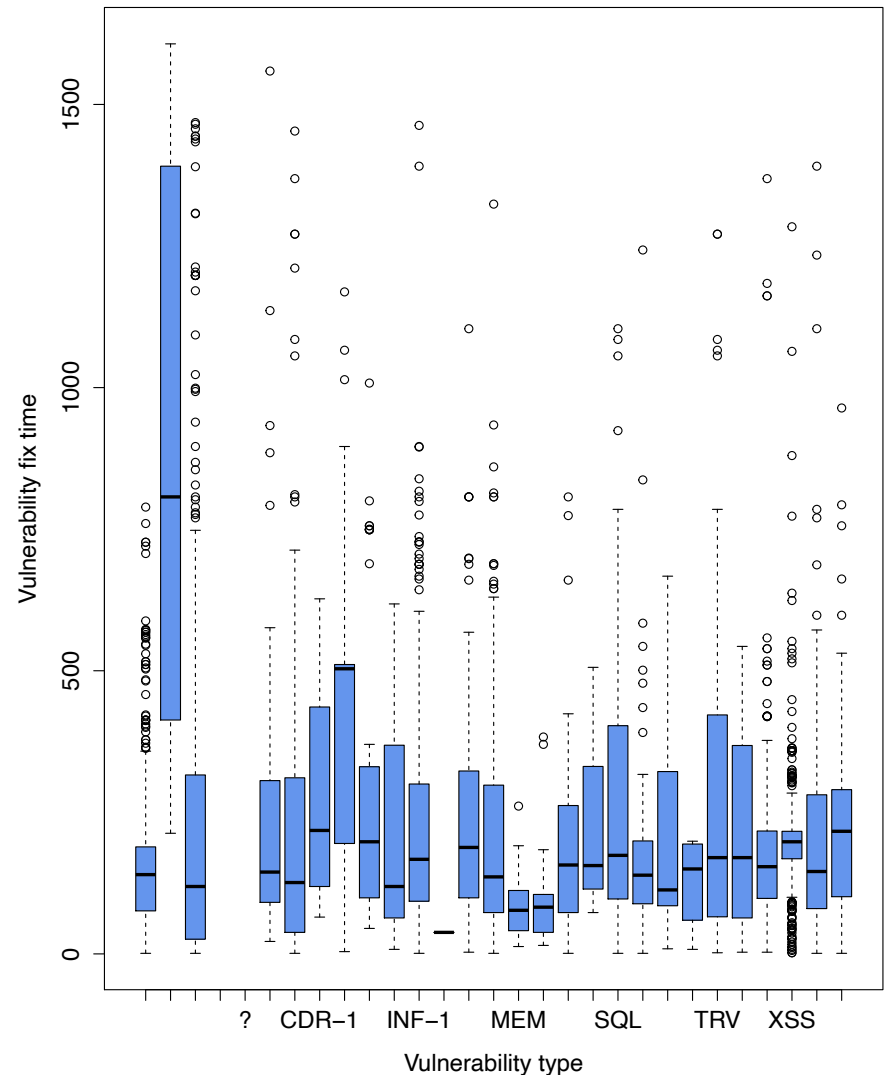
- Techniques:

- Data visualization
- Correlation analysis
- Hypothesis testing



Data Exploration Techniques– Data Visualization

- Concerns use of plots to visualize the characteristics of the distribution of the data such as frequency and variability
- Possible plots
 - Box plots
 - Line charts
 - Etc.



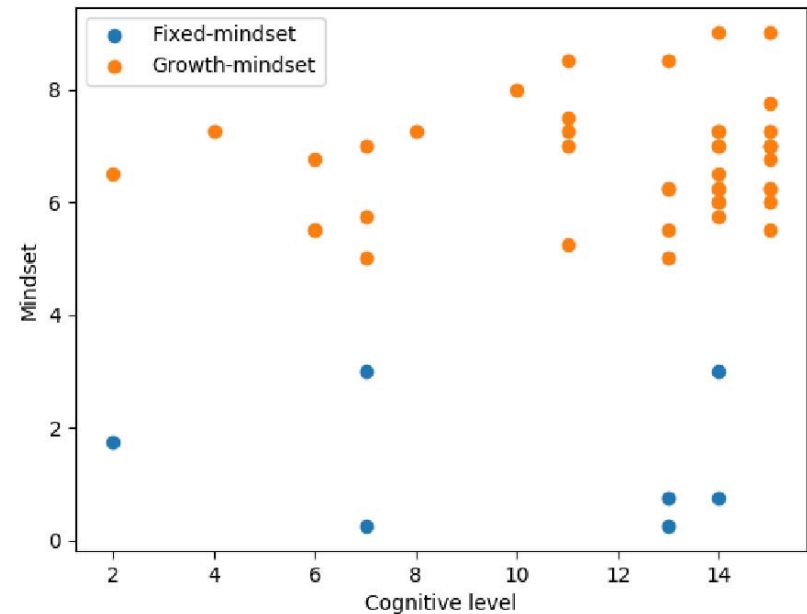
Example of Python Code to Visualize the Data

```
import matplotlib.pyplot as plt

objects = ('Group assignment', 'Ind. assignment', 'Case studies',)
y_pos = np.arange(len(objects))
plt.bar(y_pos, count_learnMethod, align='center')
plt.xticks(y_pos, objects, rotation='vertical')
plt.subplots_adjust(bottom=0.3) #space for the ticks
plt.ylabel('Frequency')
plt.xlabel('Learning methods')
plt.title('Learning methods frequencies')
plt.show()
```

Data Exploration Techniques– Correlation Coefficients

- Correlation tests the relationships of two variables
- Pearson correlation is used for continuous variables
- Select appropriate correlation techniques based on the type of variables: continuous, categorical, nominal.



$$\rho = 0.149$$

P-value = 0.34

Growth mindset may not indicate the cognitive level

Example of Python Code for Correlations

```
import numpy as np  
from scipy.stats import stats
```

```
cm, p = stats.pearsonr(Effectiveness, ConfidenceTotal)  
print ("correlation of effectiveness and total confidence: " +  
str(cm) + " p " + str(p))
```

Data Exploration Techniques–Dependency Testing

Perception	Confident	Fair confidence	Moderate confidence	No confidence
Architecture knowledge	7	9	12	1
Design of architecture	15	11	16	3
Development	3	4	4	1
No expectation	8	5	8	3
Other	6	5	7	1

The adjusted Chi-square test confirms the independence between the students' self-confidence levels and their expectations with χ^2 of 3.4832, a p-value of 0.995, and Cramer V 0.00.

Example of Python Code for Dependency Testing

```
chi, pval, cdf = stats.chi2_contingency(observation) [0:3]
```

```
p = 1 - significance
```

```
critical_value = chi2.ppf(p, cdf)
```

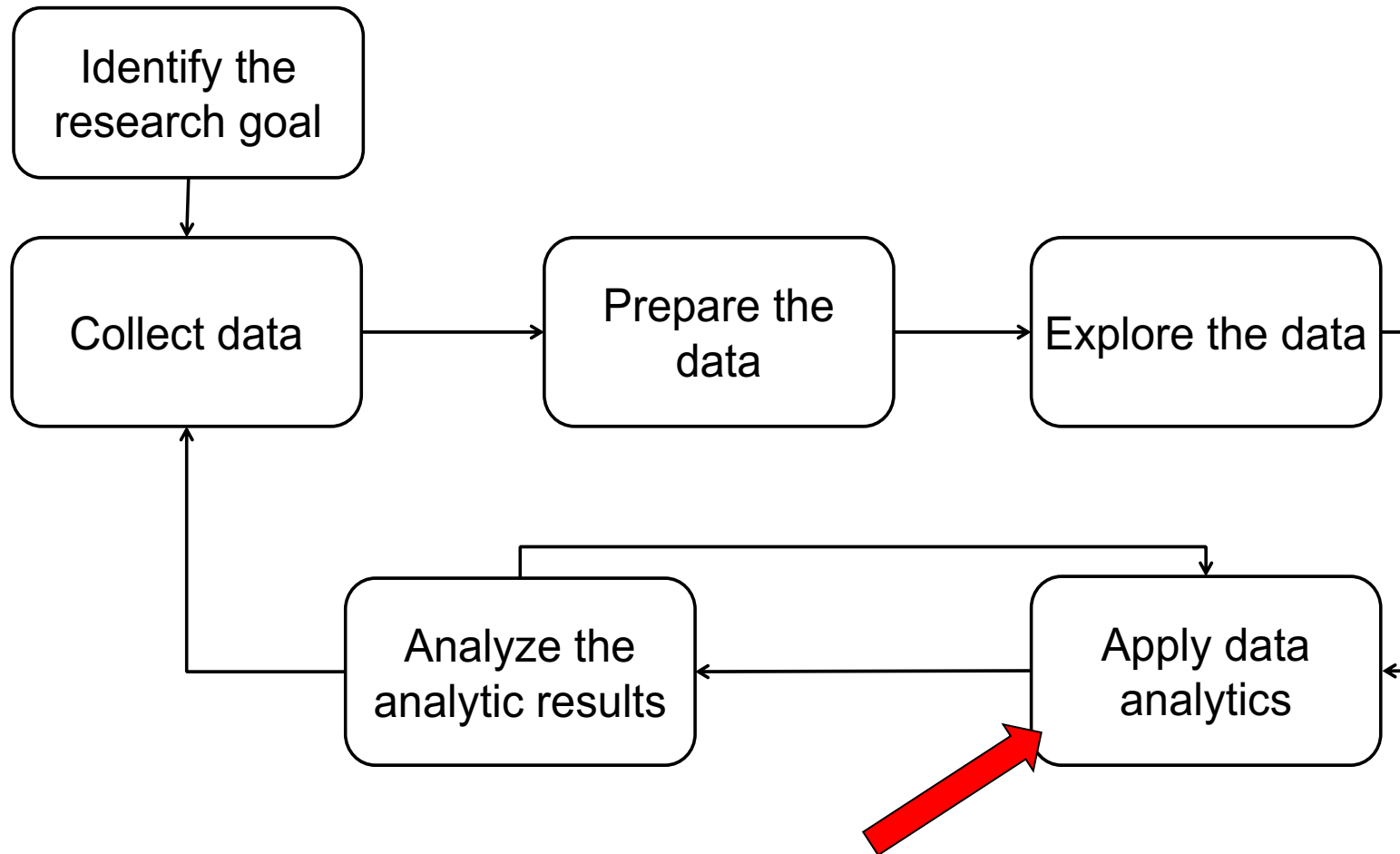
```
if critical_value < chi:
```

```
    print ("reject null hypothesis")
```

```
else:
```

```
    print ("accept null hypothesis")
```


Use Data Analytics



Types of Machine Learning Techniques

$y = f(x_1, x_2, \dots, x_n)$ —
y: output and x_i are the attribute values

Estimate function f using $\hat{f} / \hat{y} = \hat{f}(x_1, x_2, \dots, x_n)$
 $\Rightarrow \hat{f}$ is the model

Goal: minimize $\{|\hat{y} - y|\}$ for current and future data records

Types of Data

There are four types of data:

1. Continuous and categorical – represented using numerical values
2. Text -- represented using numerical values or matrices
3. Images -- represented using matrices
4. Graphs -- represented using matrices

Category of ML Problems

There are two categories

1. Supervised machine learning: derive \hat{f} from a sample of data **given X and Y**
2. Unsupervised machine learning: Derive \hat{f} from a sample of data **given X but not Y**

Machine Learning Techniques

$$\hat{y} = \hat{f}(x_1, x_2, \dots, x_n)$$

- **Regression**: The outcome variable is continuous
 - **Classification**: The outcome variable is categorical
 - **Clustering**: The outcome variable is categorical
 - **Forecasting**: Any
-
- Regression and classification techniques are supervised methods
 - Clustering techniques are unsupervised methods

Examples of Statistical ML Methods

Response variable	ML types	Algorithms
Categorical	Classification	Logistic regression Naïve Bays Support Vector machine (SVM)
Continuous	Regression	Linear regression Tree-based regression Neural-network regression
Continuous	Forecasting	Exponential smoothing Auto reg. integrated moving avg.

Exercise - AI Problems vs Types of Data

What machine learning approach would you use to:

1. Summarize a text
2. Detect object in an image
3. Predict the time to fix vulnerabilities
4. Identify the leader in a group
5. Identify the components of a software
6. Identify attacks in cars

1. Do I have the outcome variable?
2. Is the outcome variable continuous?



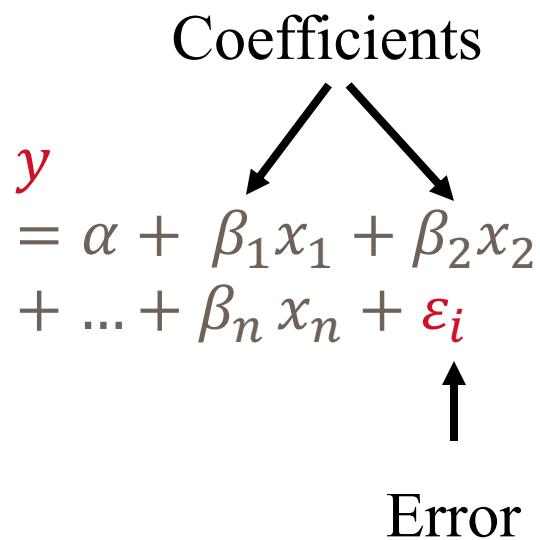
ML Methods – Linear Regression

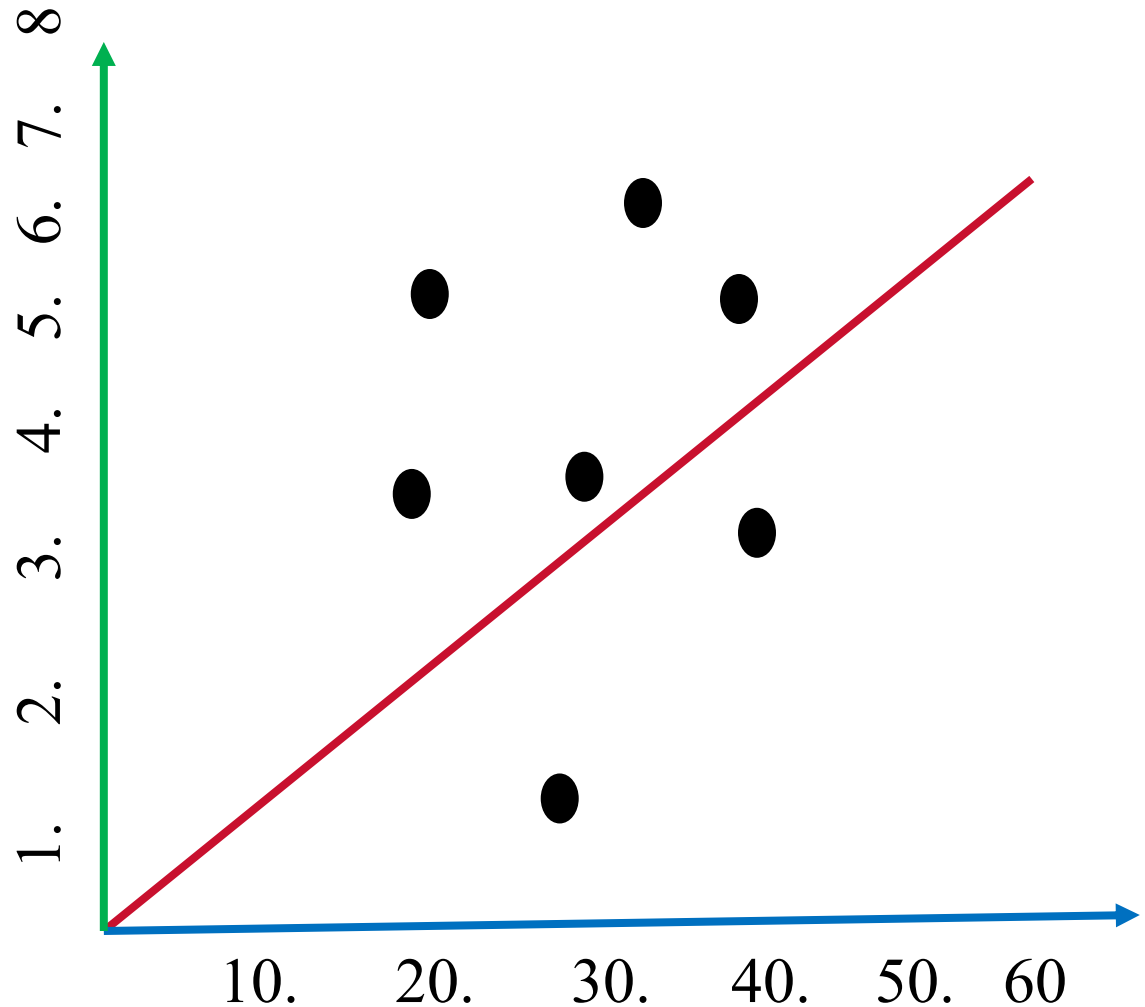
y

Coefficients

$$= \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon_i$$

Error





ML Methods – Linear Regression

Goal: minimize $\{\varepsilon_i\}$ / $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon_i$

That is: “summarize” how well the regression model fits the data.

The linear regression uses the **coefficient of determination**

$$R^2 = 1 - \frac{\sum_{k=0}^n (y_i - \hat{y}_i)^2}{\sum_{k=0}^n (y_i - \bar{y})^2}$$

R^2 is the **proportion** of **variation** of **estimated output** from **real output** to the **variation** computed using the **null model**, i.e., variation of real values from the **mean** of the values.

Regression Methods – Linear Regression

Goal: minimize $\{\varepsilon_i\}$ / $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon_i$

- There are mathematical methods to estimate the coefficients such as **ordinary least squares (OLS)**

https://en.wikipedia.org/wiki/Ordinary_least_squares

Example of Python Code for Linear Regression

```
from sklearn import datasets, linear_model
from sklearn.metrics import mean_squared_error, r2_score

# Create linear regression object
regr = linear_model.LinearRegression()

# Train the model using the training sets
regr.fit(diabetes_X_train, diabetes_y_train)

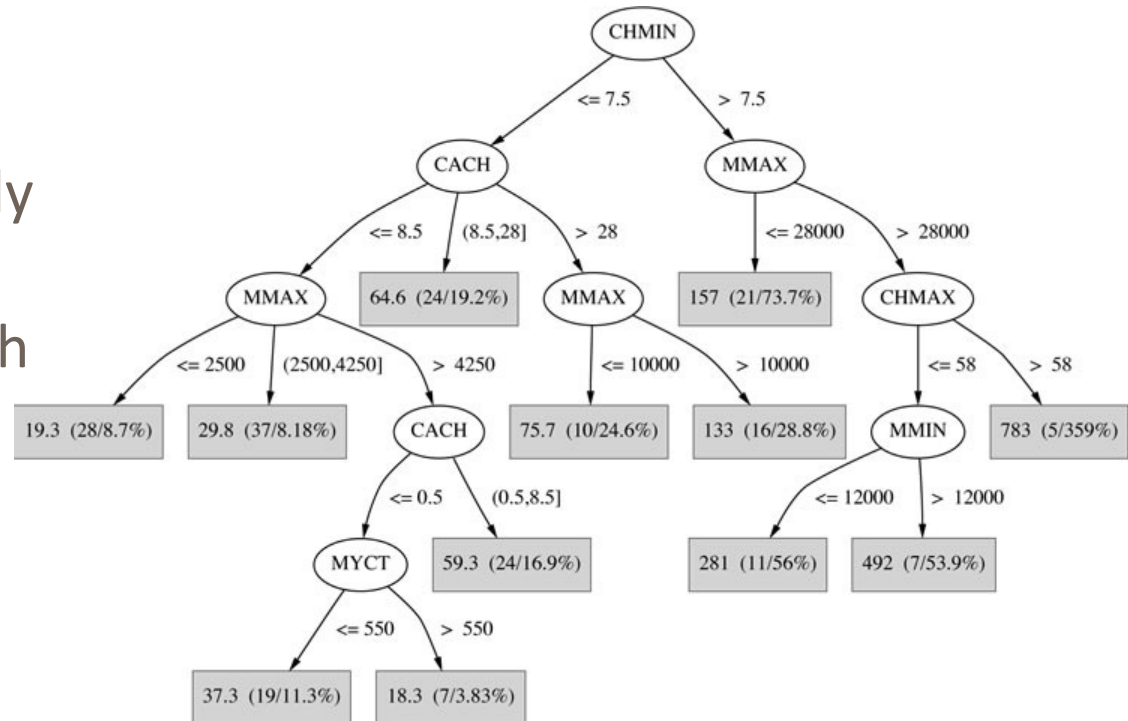
# Make predictions using the testing set
diabetes_y_pred = regr.predict(diabetes_X_test)

# The coefficients
print("Coefficients: \n", regr.coef_)
```

Source: https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html

Regression Method - Tree Regression

Tree-based regression recursively partitions the observations for each of the predicted factors such that it reduces the metric that measures the error.

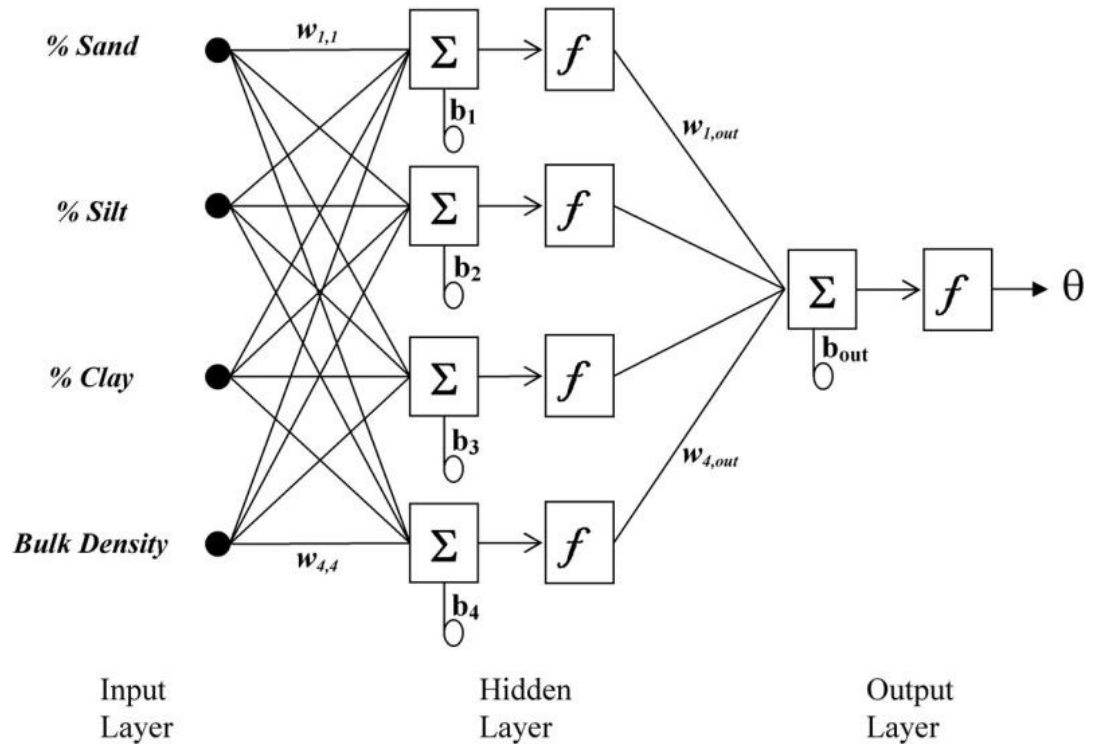


<http://jcsites.juniata.edu/faculty/rhodes/ida/dmclose.html>

Regression Method - Neural Networks

Regression

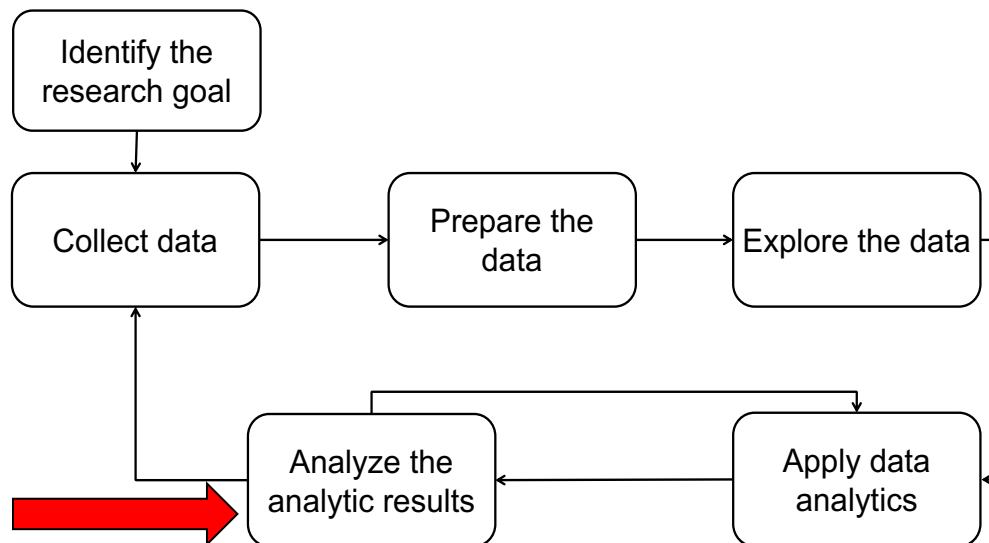
Neural-Network regression uses multi-layer network that relates the input to the output through immediate nodes. The output of the immediate nodes is the sum of the weighted inputs of the nodes of the previous layer.



Analyze the Generated Models

Tasks include

- Measure the performance of the fit using test data
- Analyze the performance of the models
- Extract the relative importance of the factors
- Identify bad row data to discard
- Identify data sets that may help to improve the models



Prediction Performance Metrics

$$R^2 = 1 - \frac{\sum_{i=0}^n (y_i - \hat{y}_i)^2}{\sum_{i=0}^n (y_i - \bar{y}_i)^2}$$

Other Performance metrics include

$$\text{PRED (h)} = \sum_{i=0}^n \begin{cases} 1 & \text{if } \frac{y_i - \hat{y}_i}{y_i} \leq h \\ 0 & \text{otherwise} \end{cases}$$

$$\text{AIC} = n \log(\hat{\sigma}^2) + 2k$$

K: number of variables

PRED: Percentage of predictions falling within a threshold **h**

The Akaike's Information Criteria computes the loss of information

Example of Results for Regression

Independent factors : Source + CVSSScore + Priority + vulnerabilitytype + Component + processor + reporter

Regression summary

Machine learning method	R ²	PRED	AIC
Linear regression	0.07	34.81	6567
Recursive partitioning and regression trees	0.21	33.92	6428
Neural networks regression	-0.98	0.71	7000

Example of Results of Regression

Factors importance based on tree model

Attribute	Overall
Component	2.75
Developer experience	2.63
Reporting source	1.26
Vulnerability type	1.11
Discovery method	0.44
CVSS	0.08
Priority	0.03

Administrative

- Assignment 1 is due on Feb 20, 2024
- Project phase 1 is due on Feb 22, 2024 – Select a project from the provided list
- Quiz 2 is Feb 22, 2024

Thank you

Any Question?