CMPS 4450

Data Mining and Visualization

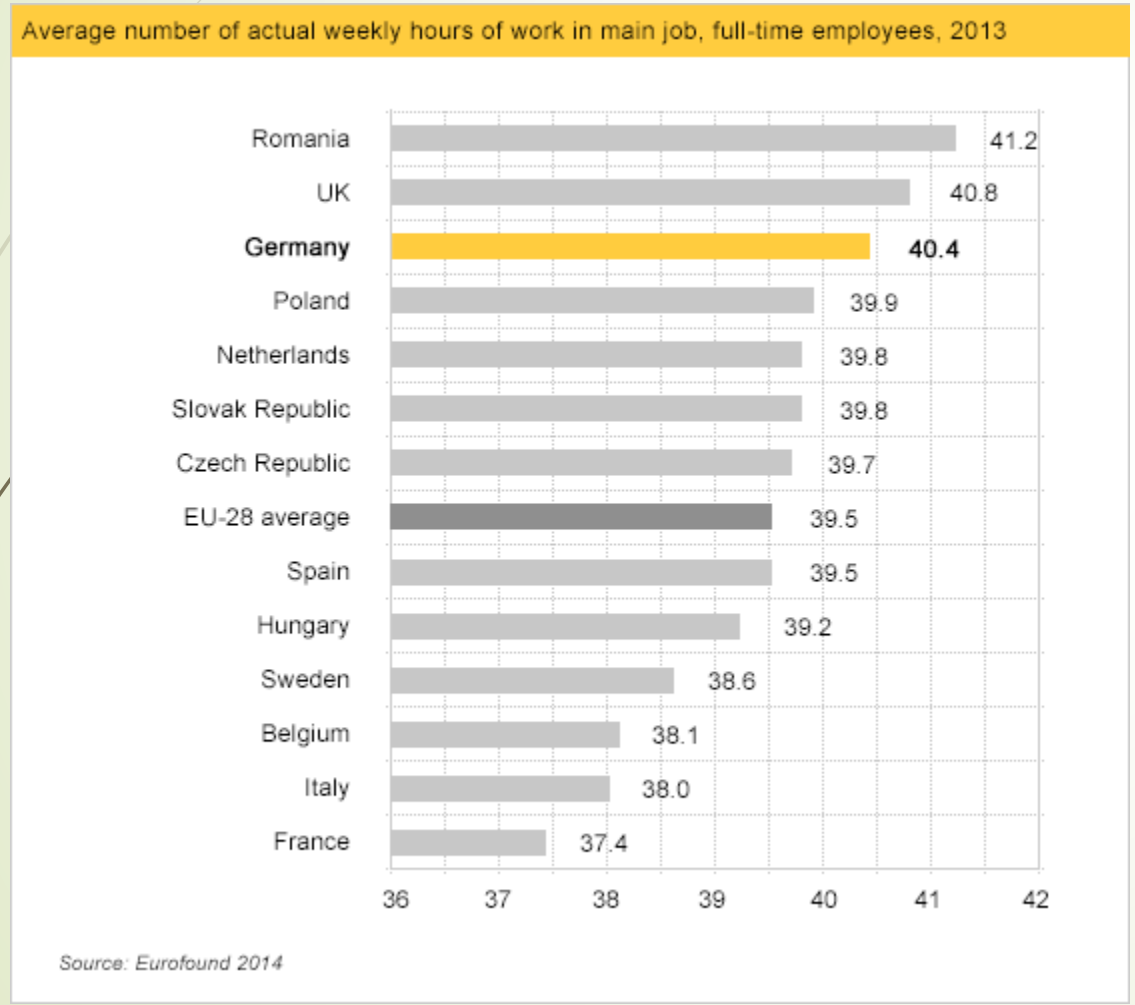
Dr. Chengwei Lei

CEECS

California State University, Bakersfield

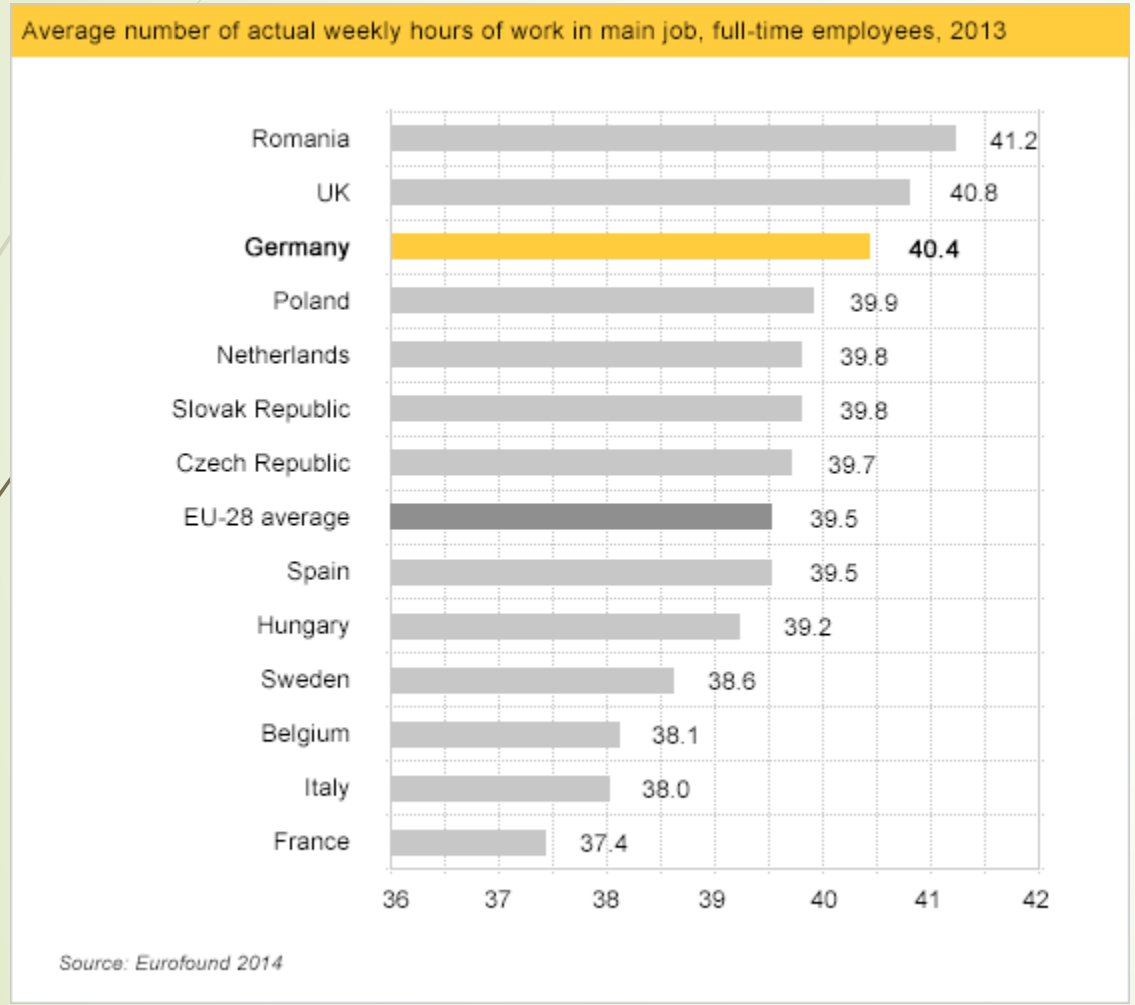
Before We start

What can you tell from the chart



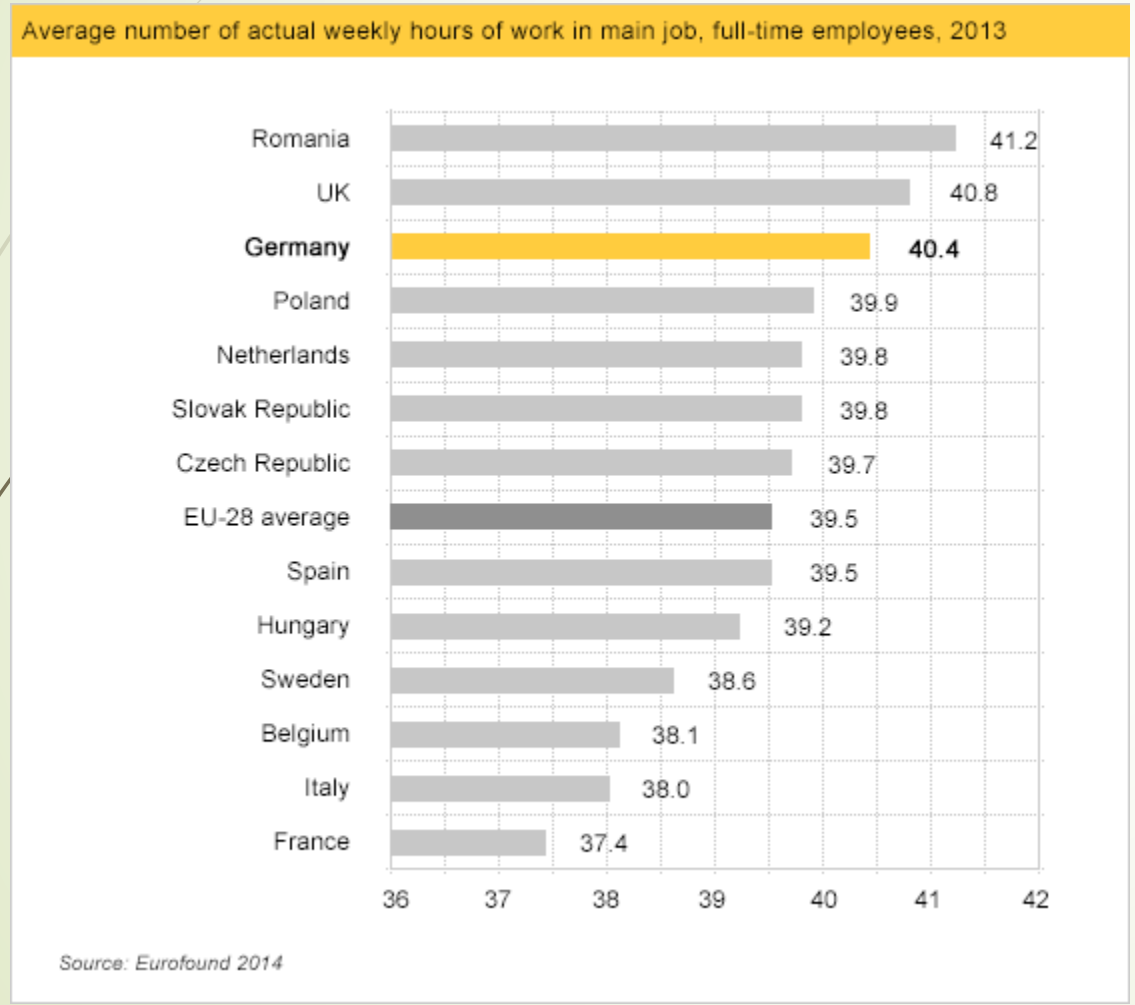
- The bar chart was created by the German economic development agency GTAI, and comes from a webpage about the German labor market. In the accompanying text, the agency boasts that German workers are more motivated and work more hours than do workers in other EU nations.

What can you tell from the chart



- It looks like Germany has a big edge over other nations such as Sweden, let alone France, right?

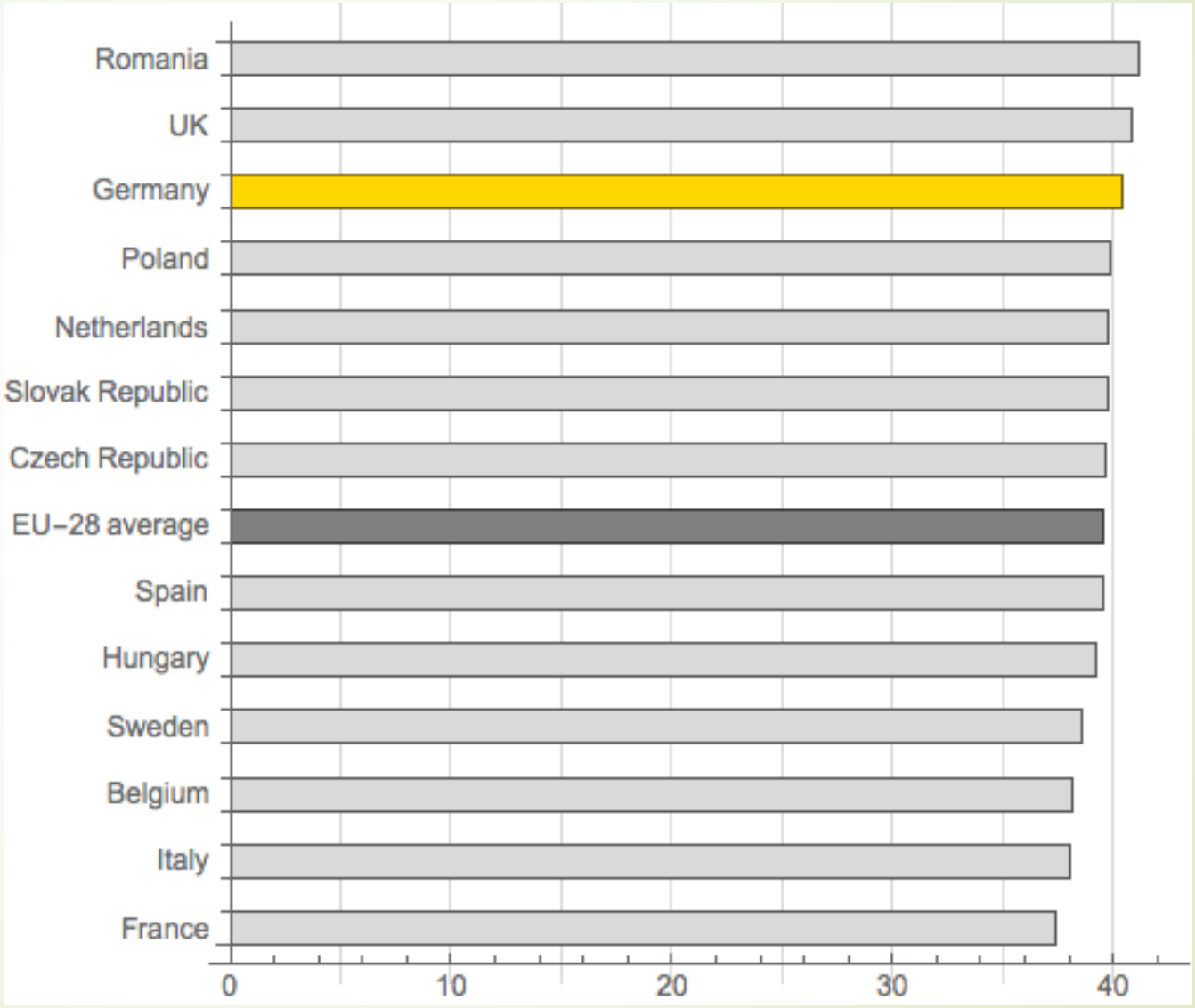
What can you tell from the chart



- It looks like Germany has a big edge over other nations such as Sweden, let alone France, right?

- No. The size of this gap is an illusion. The graph is misleading because the horizontal axis representing working hours does not go to zero, but rather cuts off at 36.

- Below, we've redrawn the graph with an axis going all the way to zero. Now the differences between countries seem negligible.

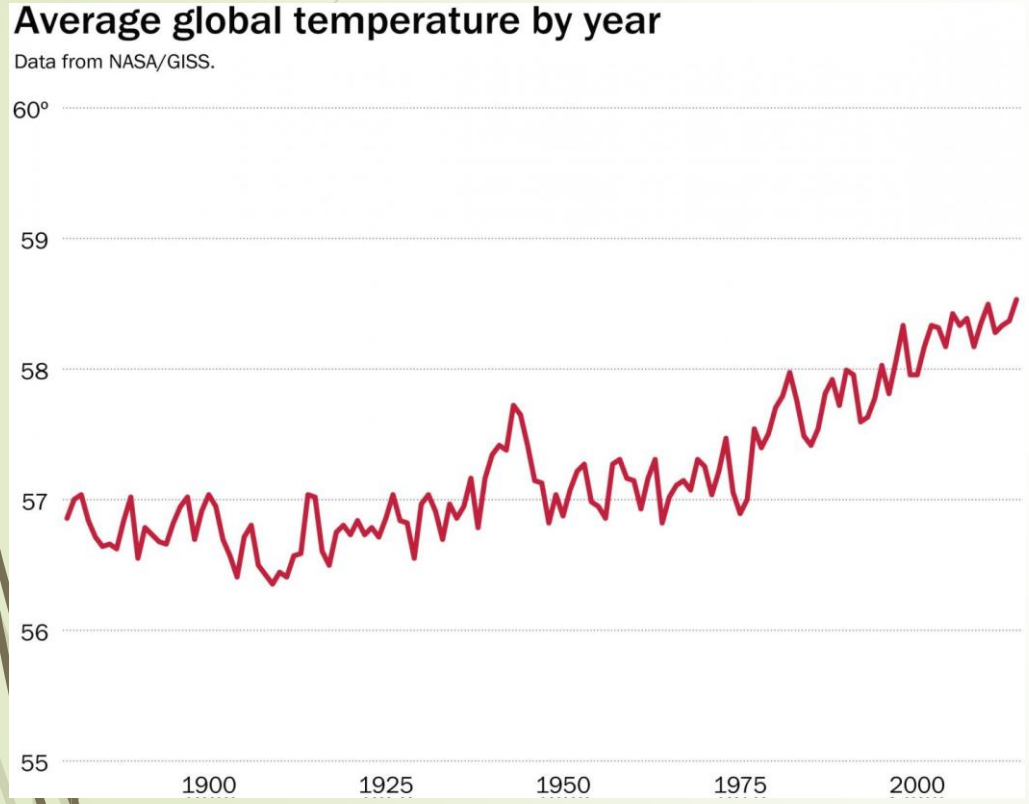


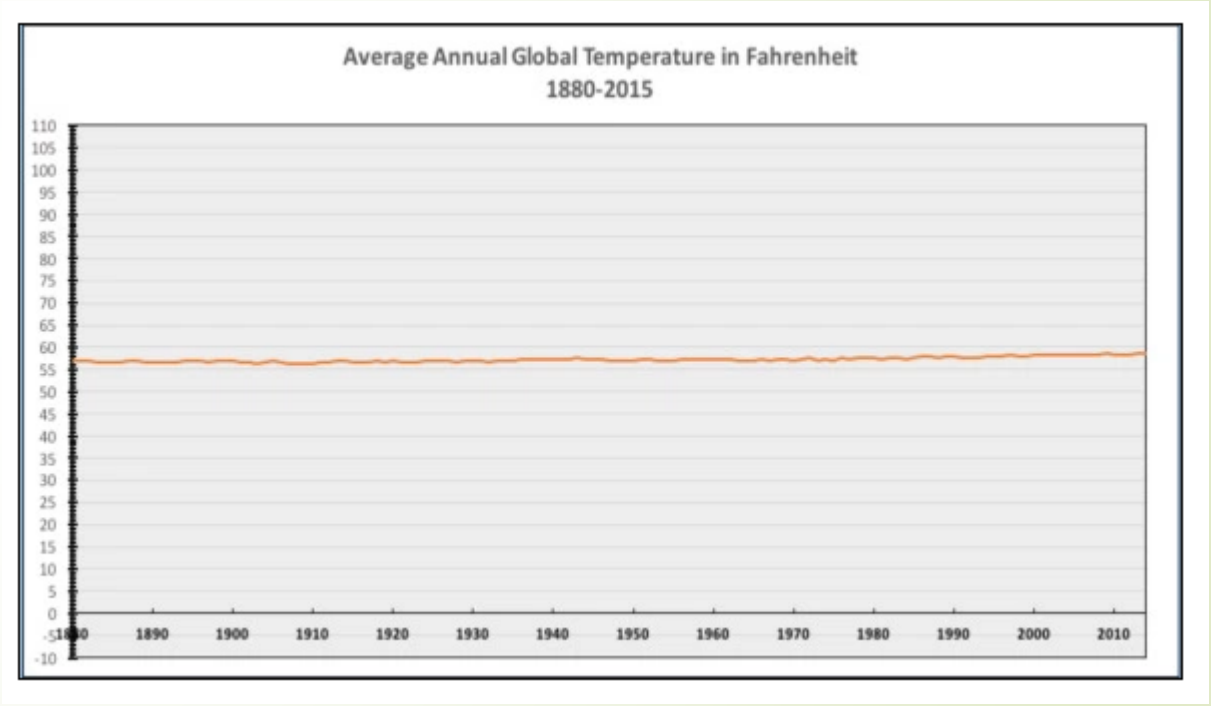
- This is a a well-known issue: drawing bar charts with a measurement (dependent variable) axis that does not go to zero.

- You might notice that in the redrawn graph we've removed the horizontal gridlines separating the countries. These were not particularly misleading, but they add visual clutter without serving any purpose whatsoever.

- Check here for more **Misleading axes on graphs**
  - http://callingbullshit.org/tools/tools_misleading_axes.html

Example 2


Average global temperature by year
Data from NASA/GISS.
60°
59
58
57
56
55
1900
1925
1950
1975
2000


Average Annual Global Temperature in Fahrenheit
1880-2015
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
-10
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010

That's why we need data mining tool

- "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."

  - --------Gartner Group

Drivers

- Market: From focus on product/service to focus on customer
- IT: From focus on up-to-date balances to focus on patterns in transactions - Data Warehouses - OLAP
- Dramatic drop in storage costs : Huge databases – e.g Walmart: 20 million transactions/day, 10 terabyte database, Blockbuster: 36 million households
- Automatic Data Capture of Transactions – e.g. Bar Codes , POS devices, Mouse clicks, Location data (GPS, cell phones)
- Internet: Personalized interactions, longitudinal data

Target marketing

- Business problem: Use list of prospects for direct mailing campaign
- Solution: ??????

Target marketing

- Business problem: Use list of prospects for direct mailing campaign
- Solution: Use Data Mining to identify most promising respondents combining demographic and geographic data with data on past purchase behavior
- Benefit: Better response rate, savings in campaign cost

Example: Fleet Financial Group

- Redesign of customer service infrastructure, including $38 million investment in data warehouse and marketing automation
- Used logistic regression to predict response probabilities to home-equity product for sample of 20,000 customer profiles from 15 million customer base
- Used CART to predict profitable customers and customers who would be unprofitable even if they respond

Churn Analysis: Telcos

- Business Problem: Prevent loss of customers, avoid adding churn-prone customers
- Solution: ????

Churn Analysis: Telcos

- Business Problem: Prevent loss of customers, avoid adding churn-prone customers
- Solution: Use neural nets, time series analysis to identify typical patterns of telephone usage of likely-to-defect and likely-to-churn customers
- Benefit: Retention of customers, more effective promotions

Example: France Telecom

- CHURN/Customer Profiling System implemented as part of major custom data warehouse solution
- Preventive CPS based on customer characteristics and known cases of churning and non-churning customers identify significant characteristics for churn
- Early detection CPS based on usage pattern matching with known cases of churn customers.

Fraud Detection

- Business problem: Fraud increases costs or reduces revenue
- Solution: ?????

Fraud Detection

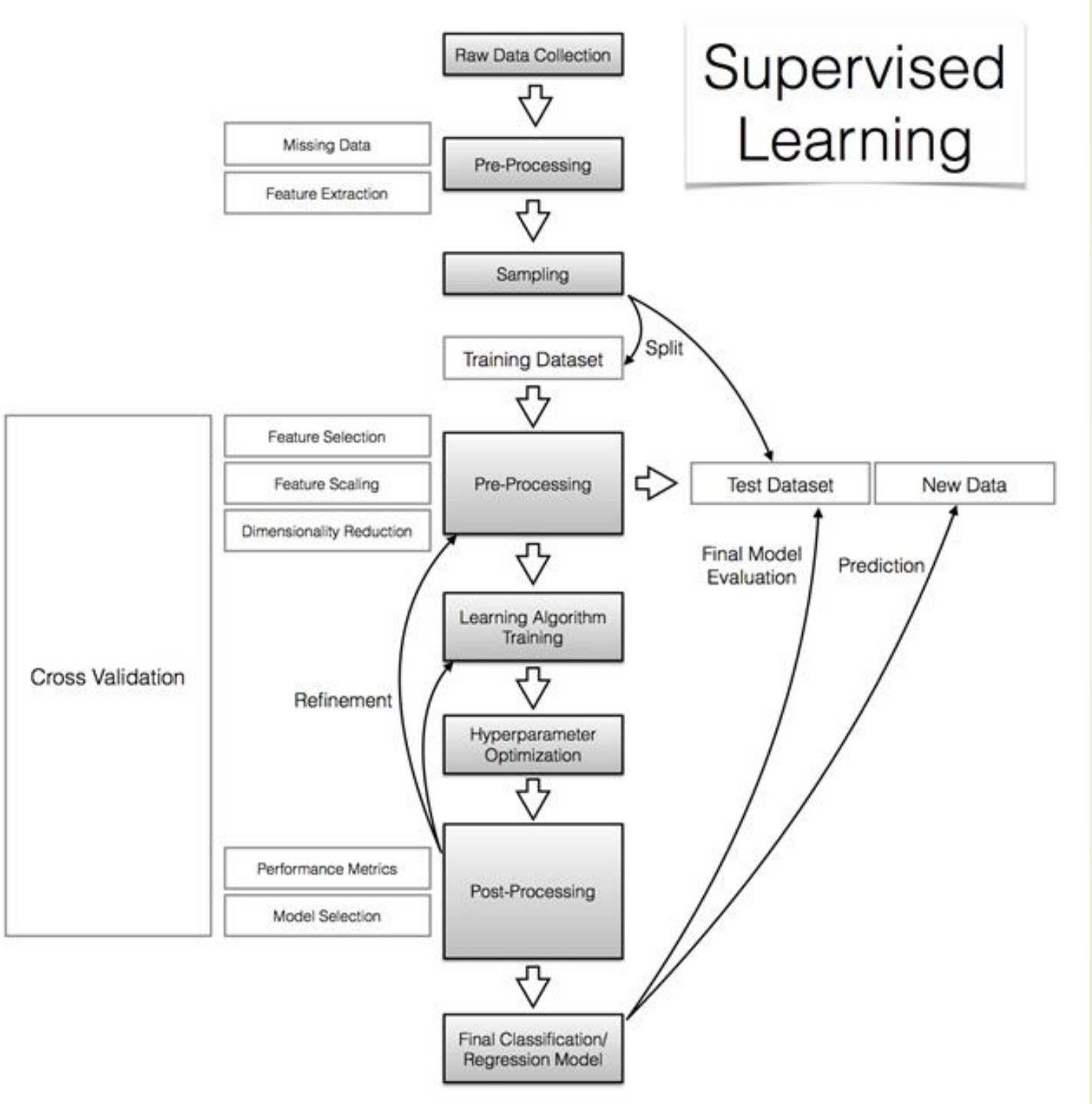
- Business problem: Fraud increases costs or reduces revenue
- Solution: Use logistic regression, neural nets to identify characteristics of fraudulent cases to prevent in future or prosecute more vigorously
- Benefit: Increased profits by reducing undesirable customers

Example: Automobile Insurance Bureau of Massachusetts

- Past reports on claims adjustors scrutinized by experts to identify cases of fraud
- Several characteristics (over 60) of claimant, type of accident, type of injury/treatment coded into database
- Dimension Reduction methods used to obtain weighted variables. Multiple Regression Step-wise Subset selection methods used to identify characteristics strong correlated with fraud

Supervised learning

- **Supervised learning** is the machine **learning** task of inferring a function from labeled training data.
- The training data consist of a set of training examples.


Supervised Learning
Raw Data Collection
Missing Data
Pre-Processing
Feature Extraction
Sampling
Training Dataset
Split
Feature Selection
Feature Scaling
Dimensionality Reduction
Pre-Processing
Test Dataset
New Data
Final Model Evaluation
Prediction
Learning Algorithm Training
Cross Validation
Refinement
Hyperparameter Optimization
Performance Metrics
Model Selection
Post-Processing
Final Classification/ Regression Model

Supervised learning

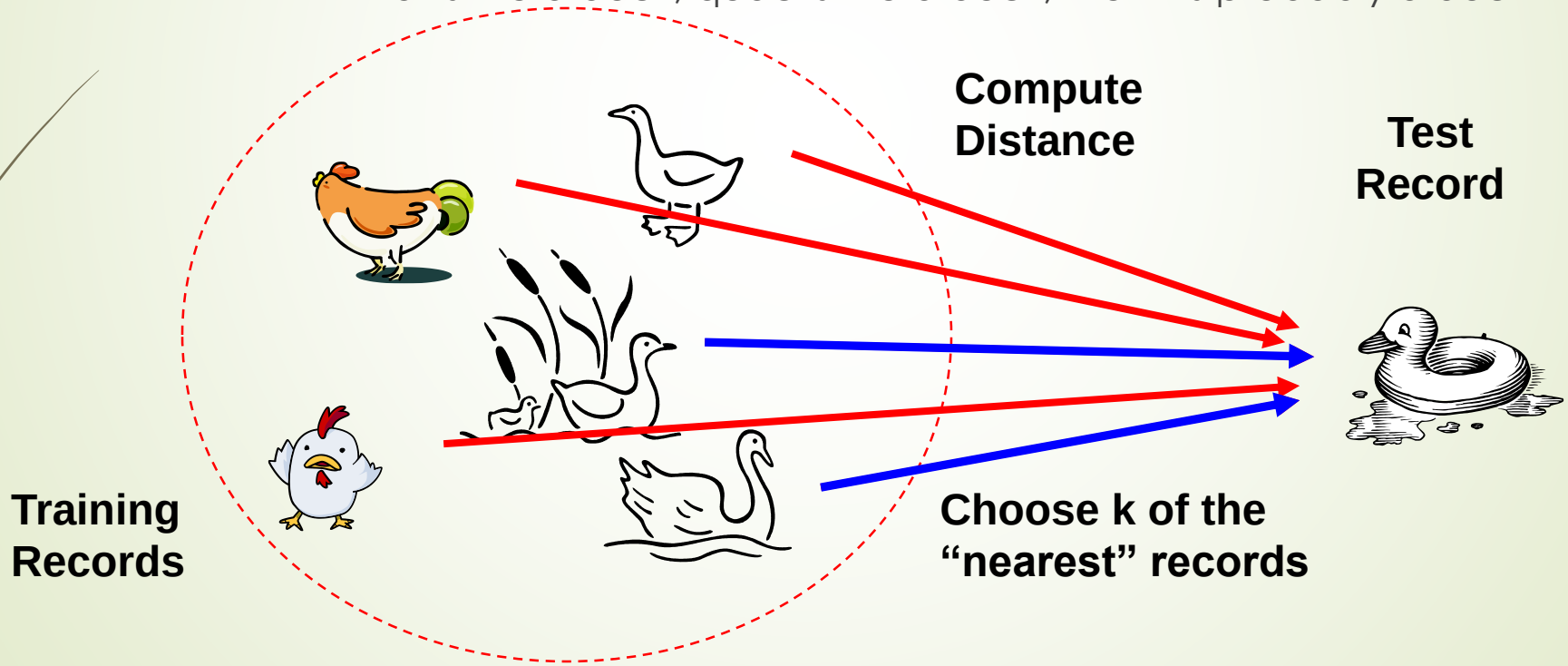
- Decision trees.
- k-NN.
- Linear regression.
- Naive Bayes.
- Neural networks.
- Logistic regression.
- Perceptron.
- Ensembles (Bagging, Boosting, Random forest)
- …..

Lazy algorithm

- k-NN classifiers are lazy learners
  - It does not build models explicitly
  - Unlike eager learners such as decision tree induction and rule-based systems
  - Classifying unknown records are relatively expensive

Nearest Neighbor Classifiers

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



KNN - Applications

- Classification and Interpretation
  - legal, medical, news, banking
- Problem-solving
  - planning, pronunciation
- Function learning
  - dynamic control
- Teaching and aiding
  - help desk, user training

KNN

- KNN is conceptually simple, yet able to solve complex problems
- Can work with relatively little information
- Learning is simple (no learning at all!)
- Memory and CPU cost
- Feature selection problem
- Sensitive to representation

KNN

- k-Nearest Neighbor Classification

KNN

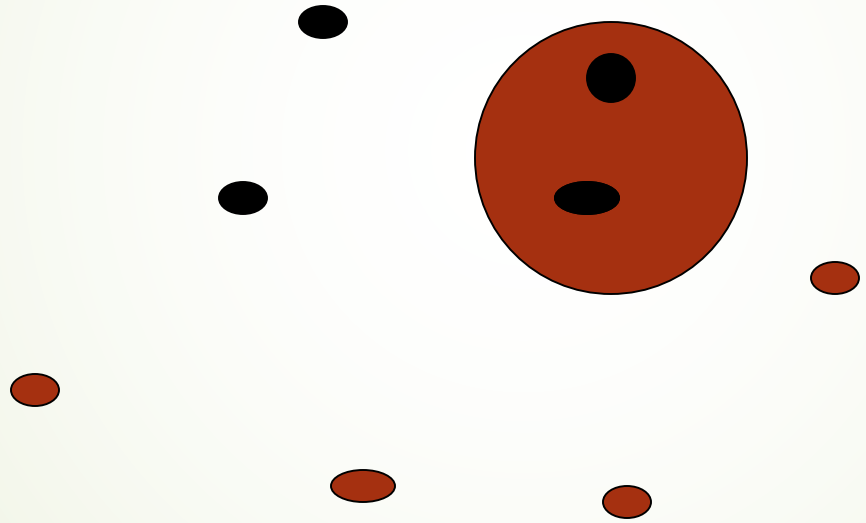
KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure

- The idea behind the k-Nearest Neighbor algorithm is to build a classification method using no assumptions about the form of the function, $y = f(x^1, x^2, \ldots x^p)$ that relates the dependent (or response) variable, y, to the independent (or predictor) variables $x^1, x^2, \ldots x^p$. The only assumption we make is that it is a "smooth" function. This is a non-parametric method because it does not involve estimation of parameters in an assumed function form such as the linear form that we encountered in linear regression.

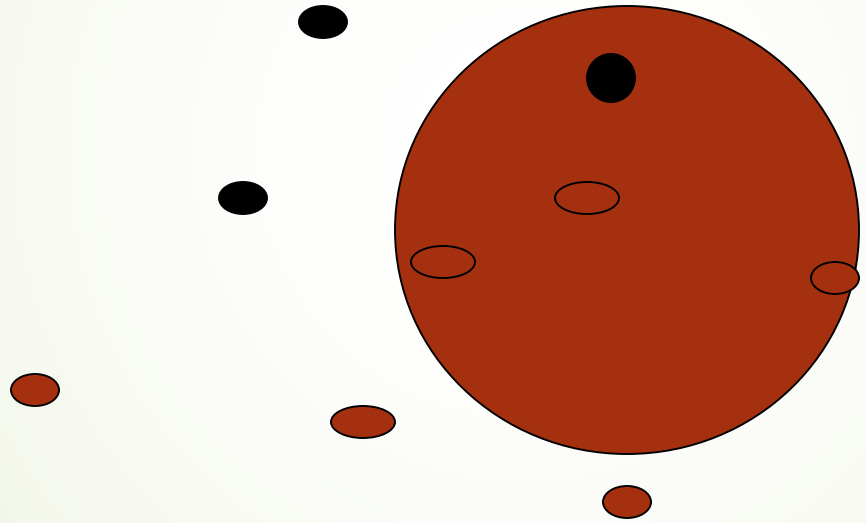
K-Nearest Neighbor

- Features
  - All instances correspond to points in an n-dimensional Euclidean space
  - Classification is delayed till a new instance arrives
  - Classification done by comparing feature vectors of the different points
  - Target function may be discrete or real-valued

1-Nearest Neighbor



3-Nearest Neighbor



K-Nearest Neighbor

- An arbitrary instance is represented by $(a_1(x), a_2(x), a_3(x),..., a_n(x))$
  - $a_i(x)$ denotes features
- Euclidean distance between two instances

  $d(x_i, x_j)=\text{sqrt}\ (\text{sum for r=1 to n}\ (a_r(x_i) - a_r(x_j))^2)$
- Continuous valued target function
  - mean value of the k nearest training examples

k-Nearest Neighbor Algorithms for Classification and Prediction

- The simplest case is k = 1 where we find the observation that is closest (the nearest neighbor)

- For k-NN we extend the idea of 1-NN as follows. Find the nearest k neighbors and then use a majority decision rule to classify a new observation.
  - The advantage is that higher values of k provide smoothing that reduces the risk of over fitting due to noise in the training data.
  - In typical applications k is in units or tens rather than in hundreds or thousands.

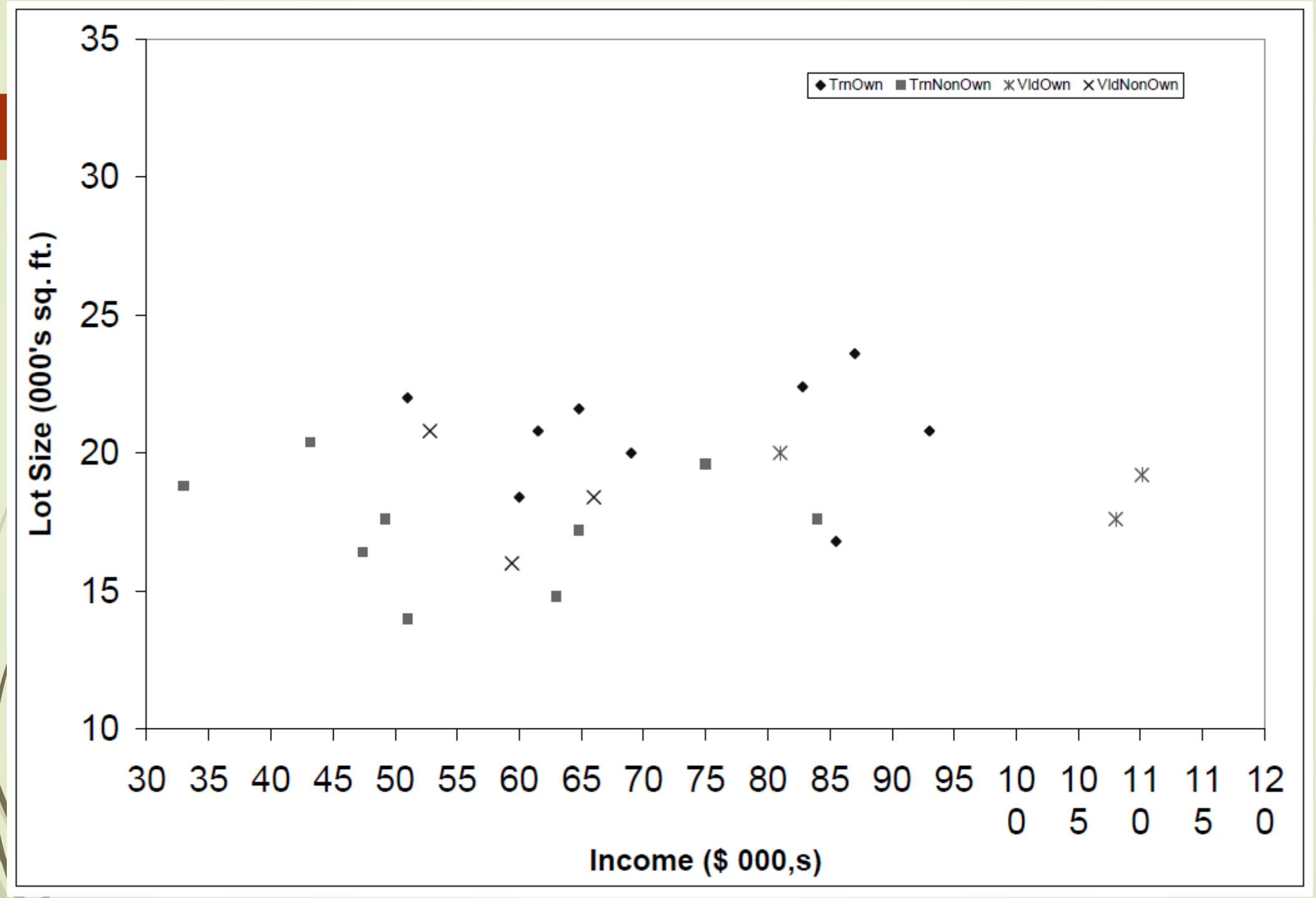
Example 1

- A riding-mower manufacturer would like to find a way of classifying families in a city into those that are likely to purchase a riding mower and those who are not likely to buy one. A pilot random sample of 12 owners and 12 non-owners in the city is undertaken. The data are shown in Table 1

Observation	Observation Income ($000's)	Lot Size (000's sq. ft.)	Owners=1, Non-owners=2
1	60	18.4	1
2	85.5	16.8	1
3	64.8	21.6	1
4	61.5	20.8	1
5	87	23.6	1
6	110.1	19.2	1
7	108	17.6	1
8	82.8	22.4	1
9	69	20	1
10	93	20.8	1
11	51	22	1
12	81	20	1
13	75	19.6	2
14	52.8	20.8	2
15	64.8	17.2	2
16	43.2	20.4	2
17	84	17.6	2
18	49.2	17.6	2
19	59.4	16	2
20	66	18.4	2
21	47.4	16.4	2
22	33	18.8	2
23	51	14	2
24	63	14.8	2

- How do we choose k? In data mining we use the training data to classify the cases in the validation data to compute error rates for various choices of k.

- For our example we have randomly divided the data into a training set with 18 cases and a validation set of 6 cases.
  - Of course, in a real data mining situation we would have sets of much larger sizes. The validation set consists of observations 6, 7, 12, 14, 19, 20 of Table 1.
  - The remaining 18 observations constitute the training data. Figure 1 displays the observations in both training and validation data sets.


TrnOwn
TrnNonOwn
VldOwn
VldNonOwn
Lot Size (000's sq. ft.)
35
30
25
20
15
10
30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120
Income ($ 000,s)

- Notice that if we choose k=1 we will classify in a way that is very sensitive to the local characteristics of our data.

- On the other hand if we choose a large value of k we average over a large number of data points and average out the variability due to the noise associated with individual data points.

- If we choose k=18 we would simply predict the most frequent class in the data set in all cases. This is a very stable prediction but it completely ignores the information in the independent variables.

- Table 2 shows the misclassification error rate for observations in the validation data for different choices of k.

k	1	3	5	7	9	11	13	18
Misclassification Error %	33	33	33	33	33	17	17	50

- We would choose k=11 (or possibly 13) in this case. This choice optimally trades off the variability associated with a low value of k against the over smoothing associated with a high value of k.

k-Nearest Neighbor Prediction

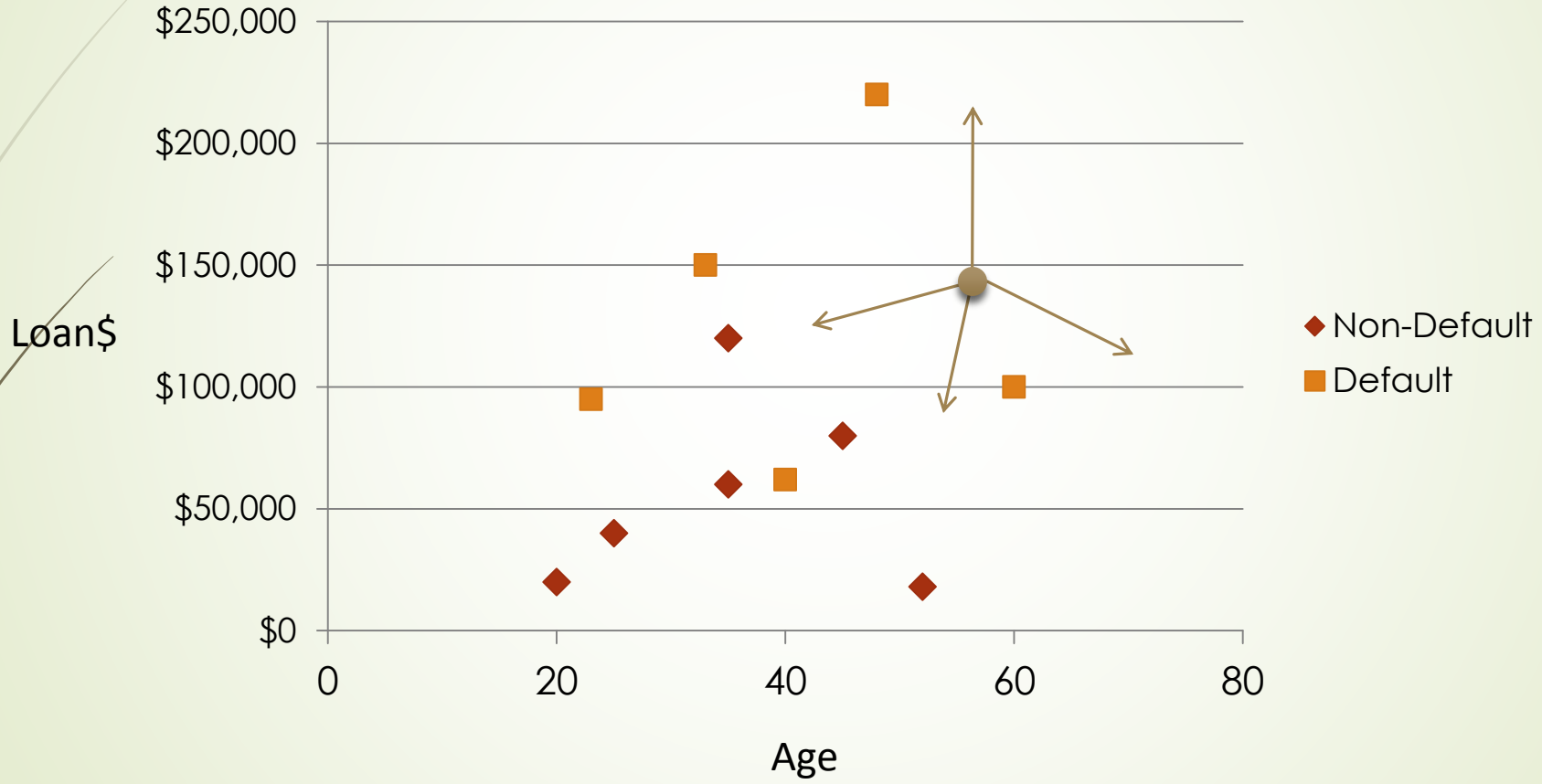
- The idea of k-NN can be readily extended to predicting a continuous value, by simply predicting the average value of the dependent variable for the k nearest neighbors.

- Often this average is a weighted average with the weight decreasing with increasing distance from the point at which the prediction is required.

Shortcomings of k-NN algorithms

- There are two difficulties with the practical exploitation of the power of the k-NN approach.
  - First, while there is no time required to estimate parameters from the training data the time to find the nearest neighbors in a large training set can be prohibitive.
  - Second, the number of observations required in the training data set to qualify as large increases exponentially with the number of dimensions p.

KNN Classification



KNN Classification – Distance

Age	Loan	Default	Distance
25	$40,000	N	102000
35	$60,000	N	82000
45	$80,000	N	62000
20	$20,000	N	122000
35	$120,000	N	22000
52	$18,000	N	124000
23	$95,000	Y	47000
40	$62,000	Y	80000
60	$100,000	Y	42000
48	$220,000	Y	78000
33	$150,000	Y	8000
		↓	
48	**$142,000**	**?**	

Euclidean Distance

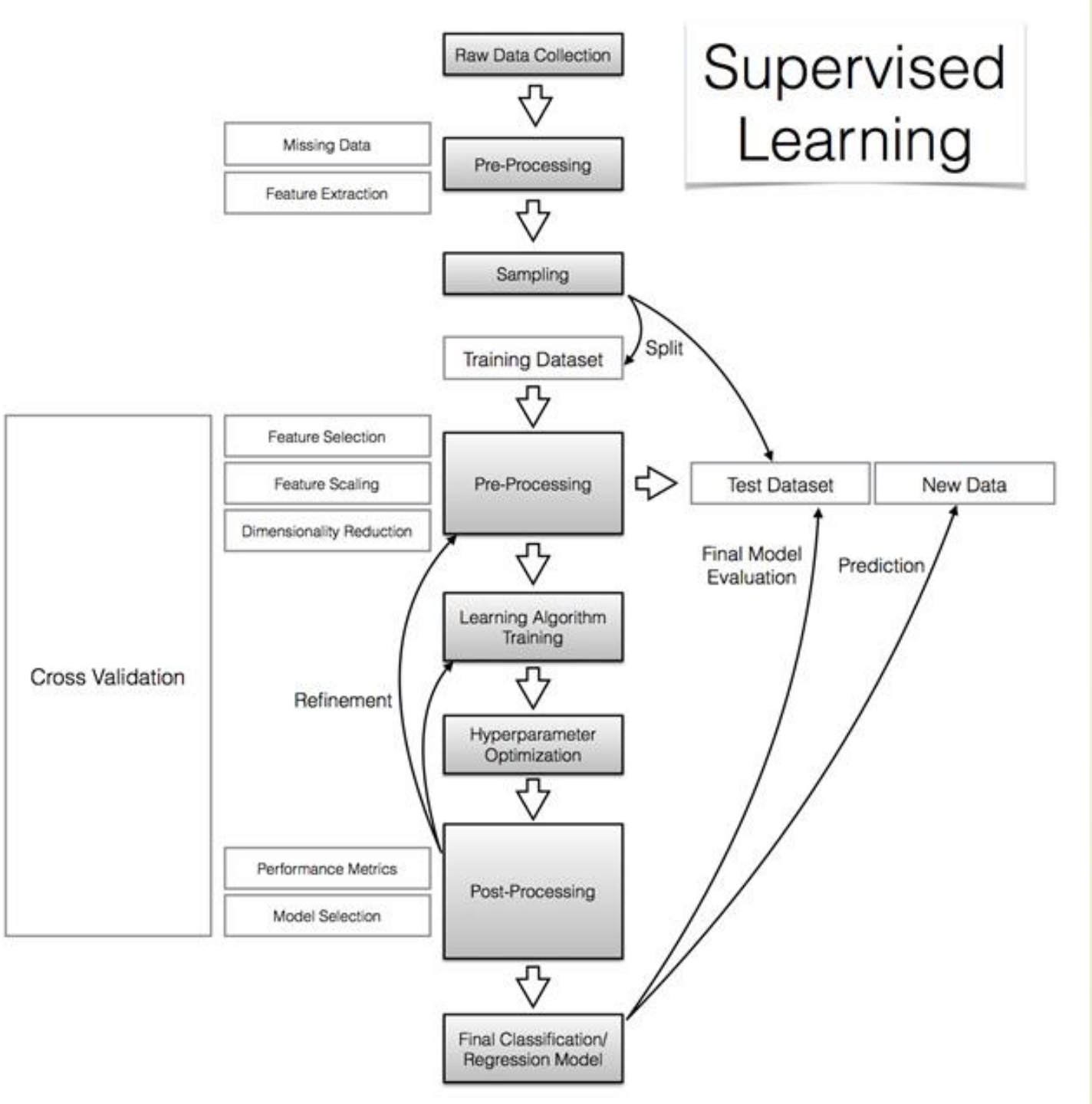
$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

KNN Classification – Standardized Distance

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

Standardized Variable

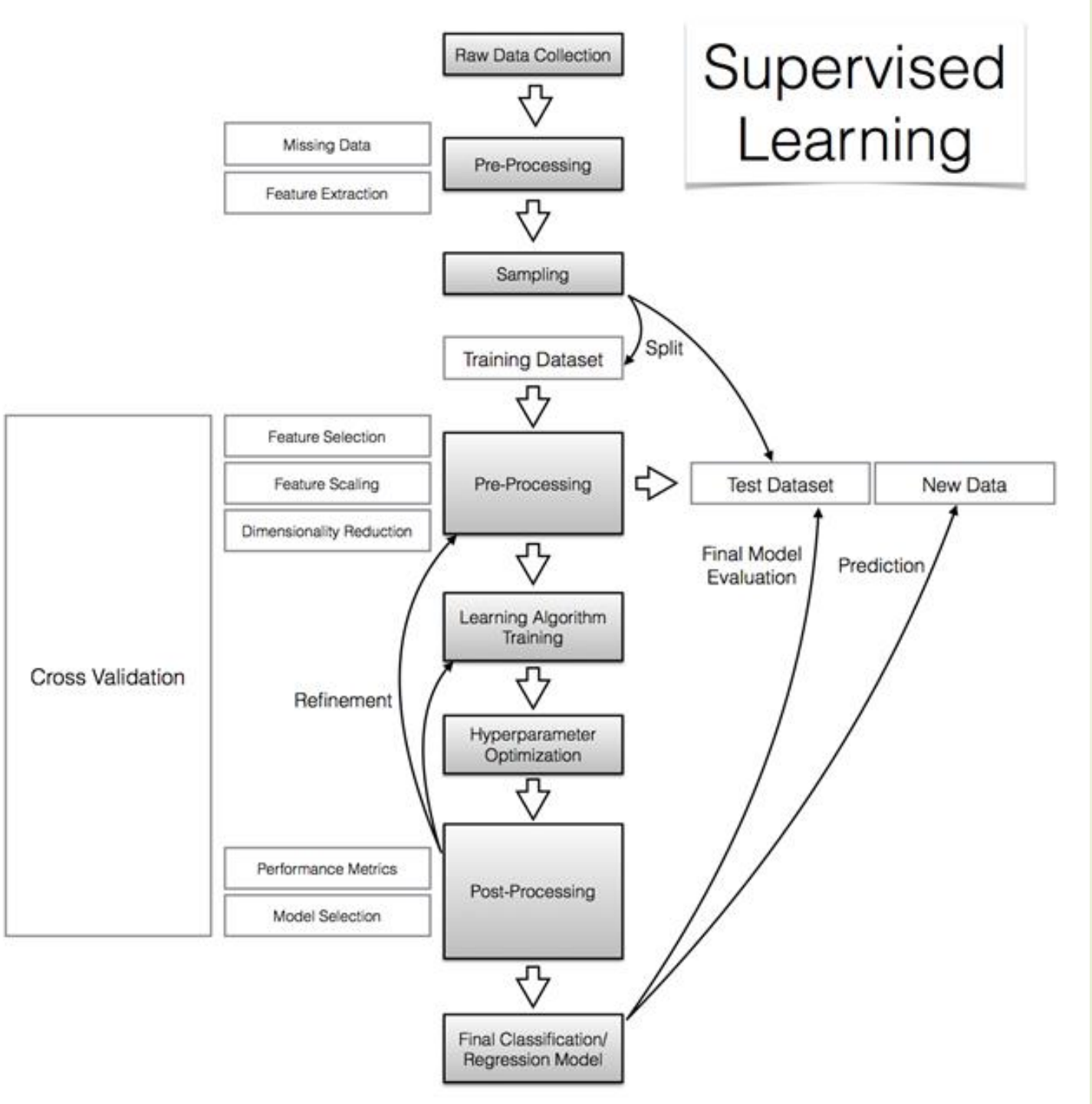
$$X_s = \frac{X - Min}{Max - Min}$$


Supervised Learning
Raw Data Collection
Missing Data
Pre-Processing
Feature Extraction
Sampling
Training Dataset
Split
Feature Selection
Feature Scaling
Dimensionality Reduction
Pre-Processing
Test Dataset
New Data
Final Model Evaluation
Prediction
Learning Algorithm Training
Cross Validation
Refinement
Hyperparameter Optimization
Performance Metrics
Model Selection
Post-Processing
Final Classification/ Regression Model

For KNN

- Define a metric for measuring the distance between the query point and cases from the examples sample.

- Selecting the value of *k*


Supervised Learning
Raw Data Collection
Missing Data
Pre-Processing
Feature Extraction
Sampling
Training Dataset
Split
Feature Selection
Feature Scaling
Dimensionality Reduction
Pre-Processing
Test Dataset
New Data
Final Model Evaluation
Prediction
Learning Algorithm Training
Cross Validation
Refinement
Hyperparameter Optimization
Performance Metrics
Model Selection
Post-Processing
Final Classification/ Regression Model