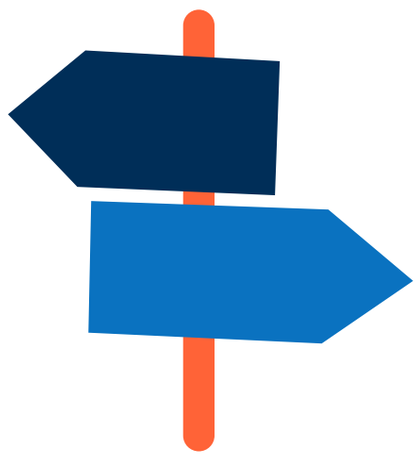
MTAT.03.319

Business Data Analytics

Lecture 4
Customer Life Cycle: Regression Problems

Rajesh Sharma
https://css.cs.ut.ee/

Recall from Lecture 3: Customer segmentation



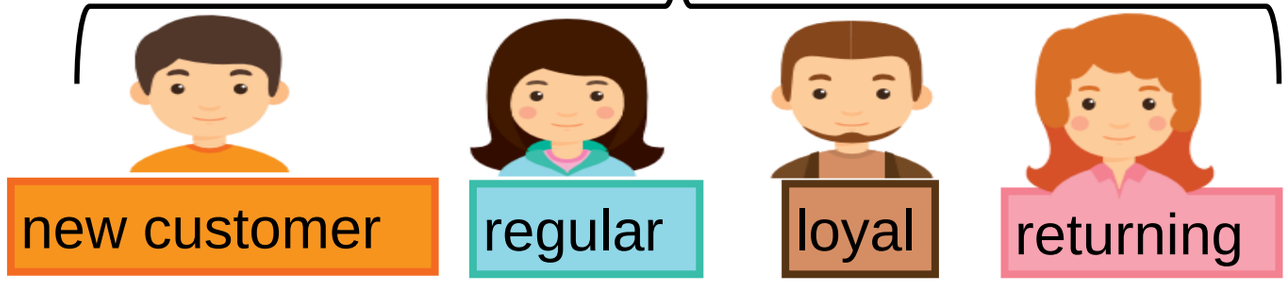
Intuition-based

Historical/behavioral-based

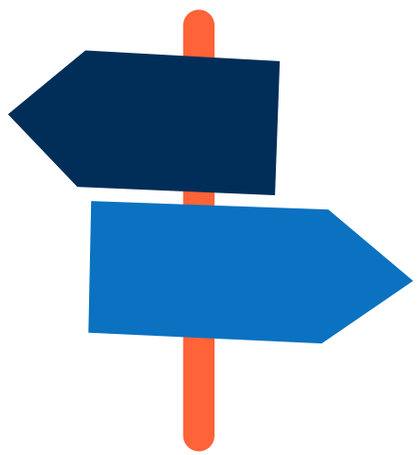
- RFM
- Value tier
- Lifecycle stage

Data-driven

- K-means
- Hierarchical clust
- DB Scan



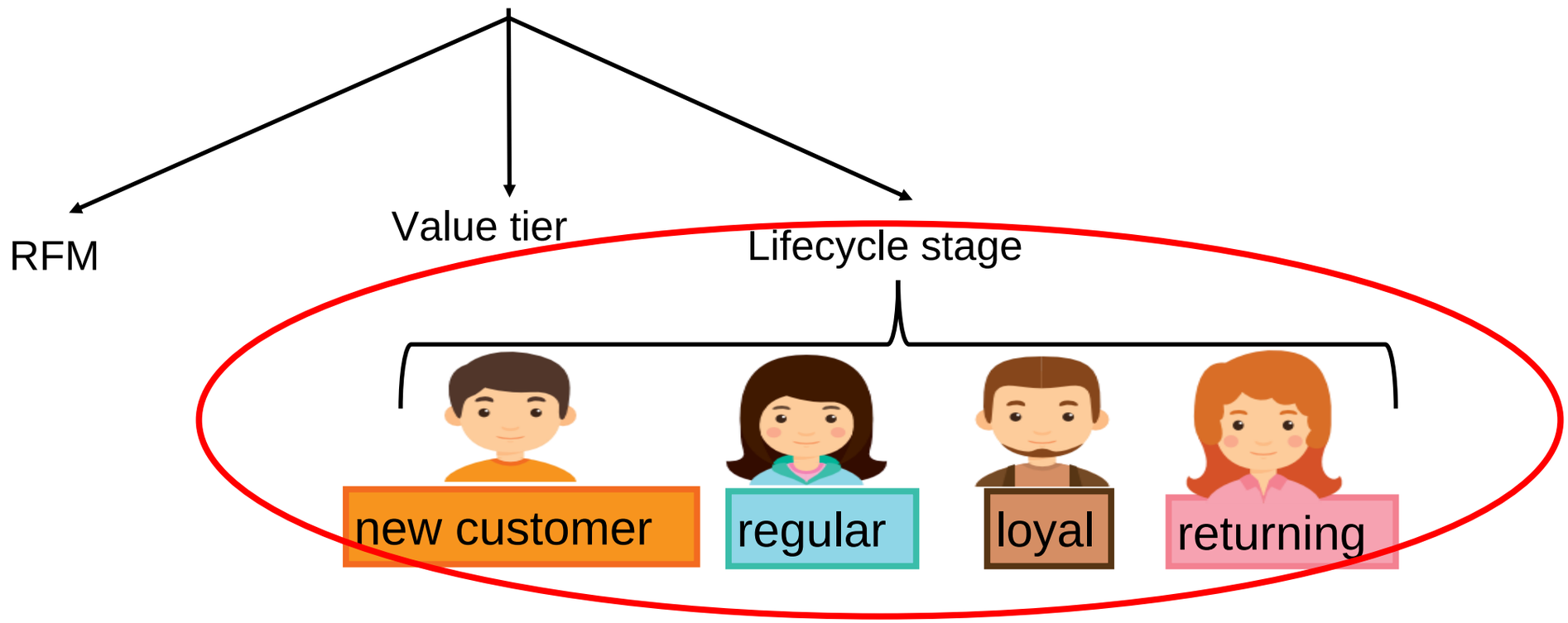
Recall from Lecture 3: Customer segmentation



Intuition-based

Historical/behavioral-based

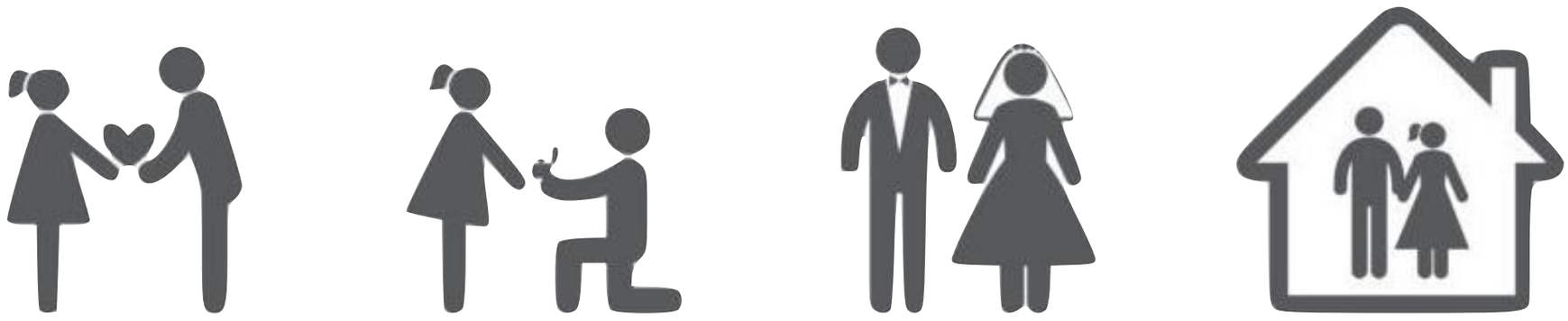
Data-driven

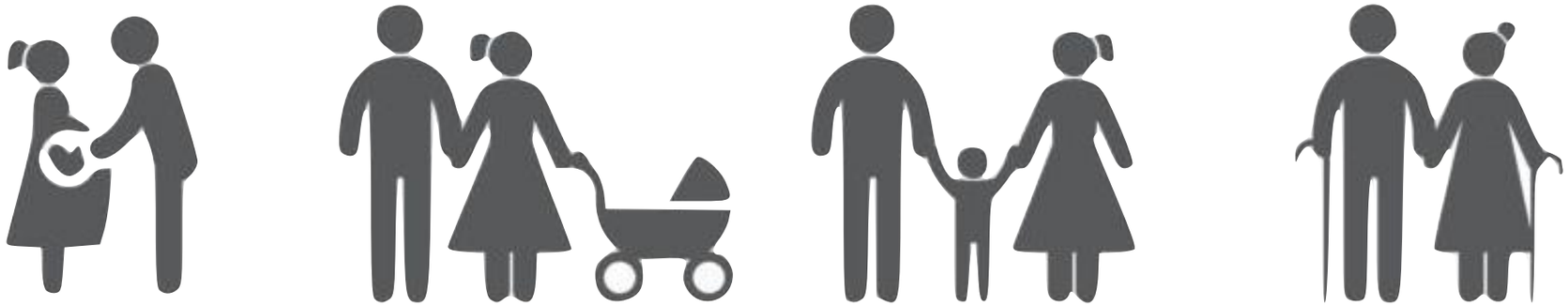


Marketing and Sales

Customer Lifecycle Management: Regression Problems

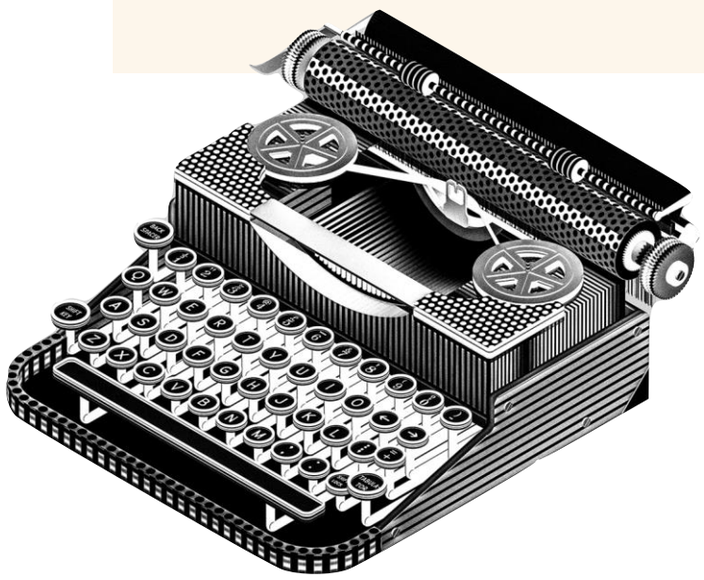
Customer lifecycle





Customer lifecycle

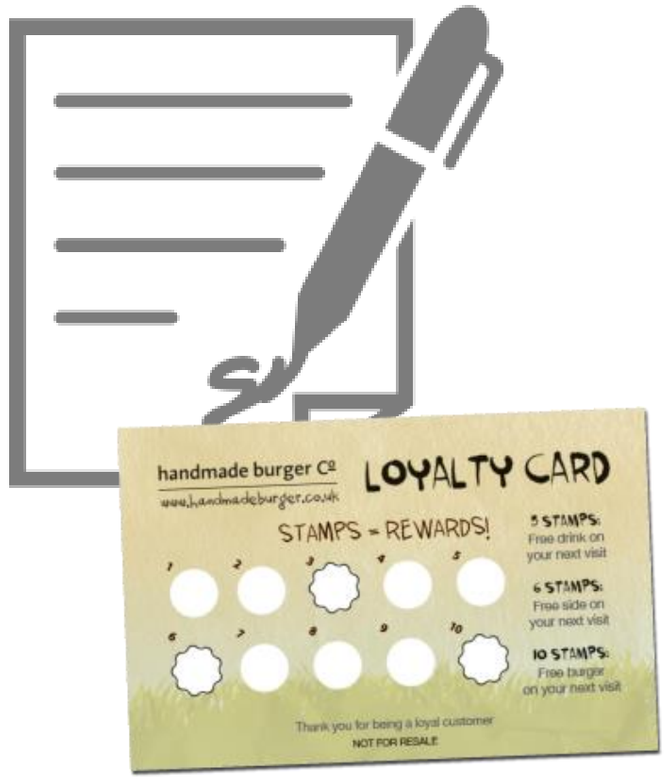
Moving companies grow not because they force people to move more often, but by attracting new customers



Relationships based on commitment

event-based

subscription-based



Relationships based on commitment

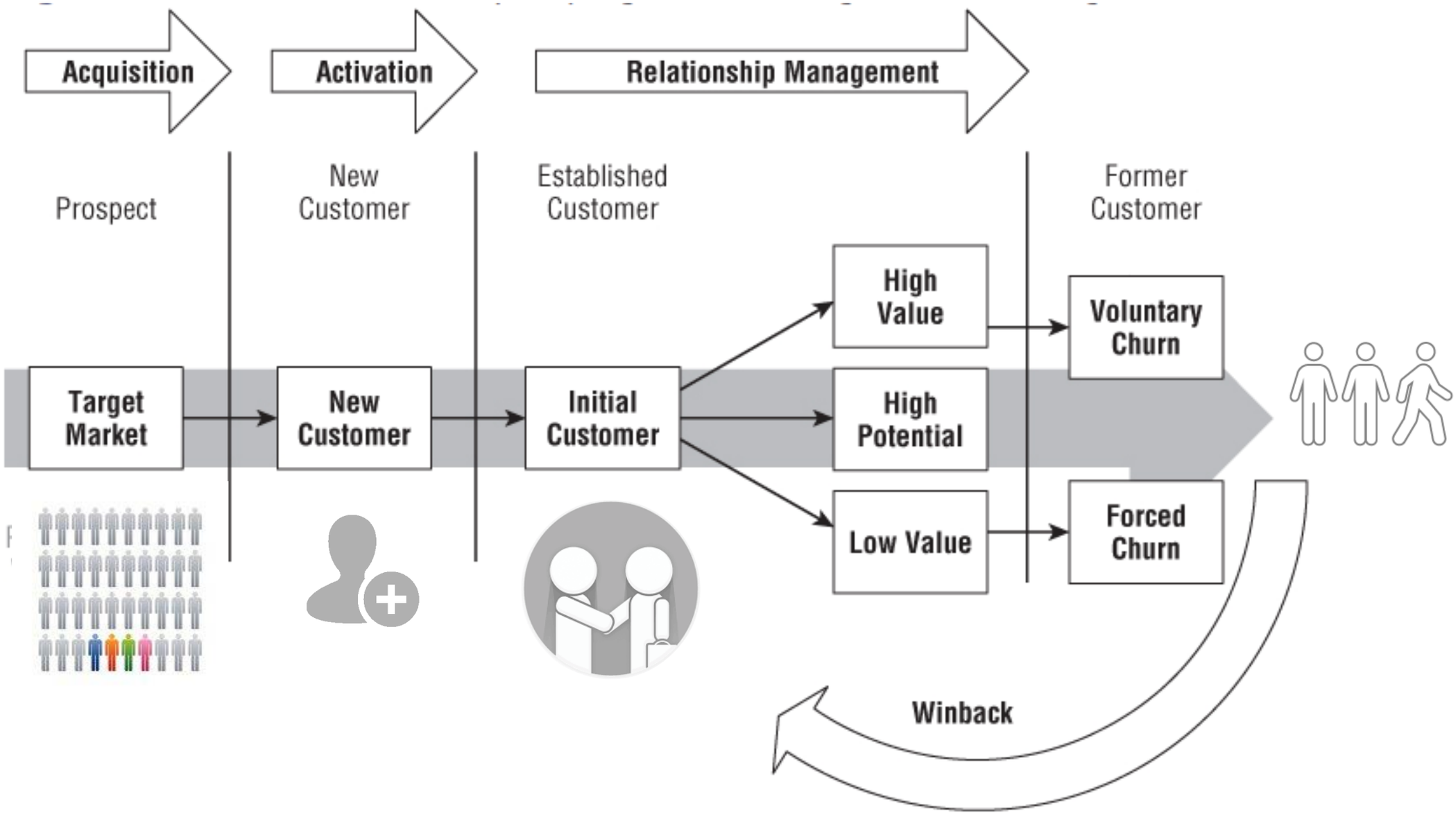
Event-based

- Packers and Movers
- Wedding Planners

Subscription-based

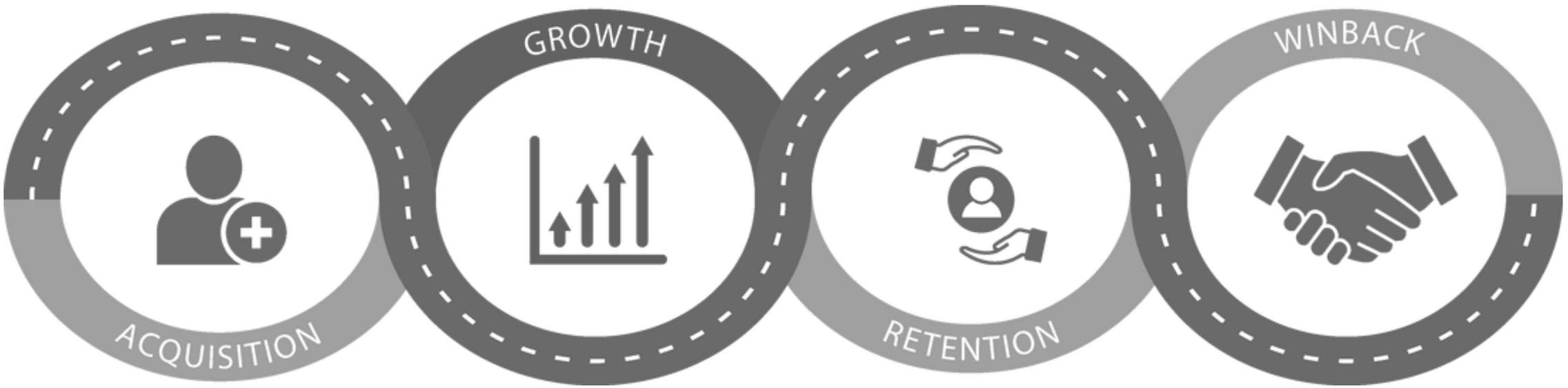
- Telco
- Banks
- Retail (Walmart, Konsume, etc)
- Hairdressers

Customer lifecycle



[Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, Third Edition]

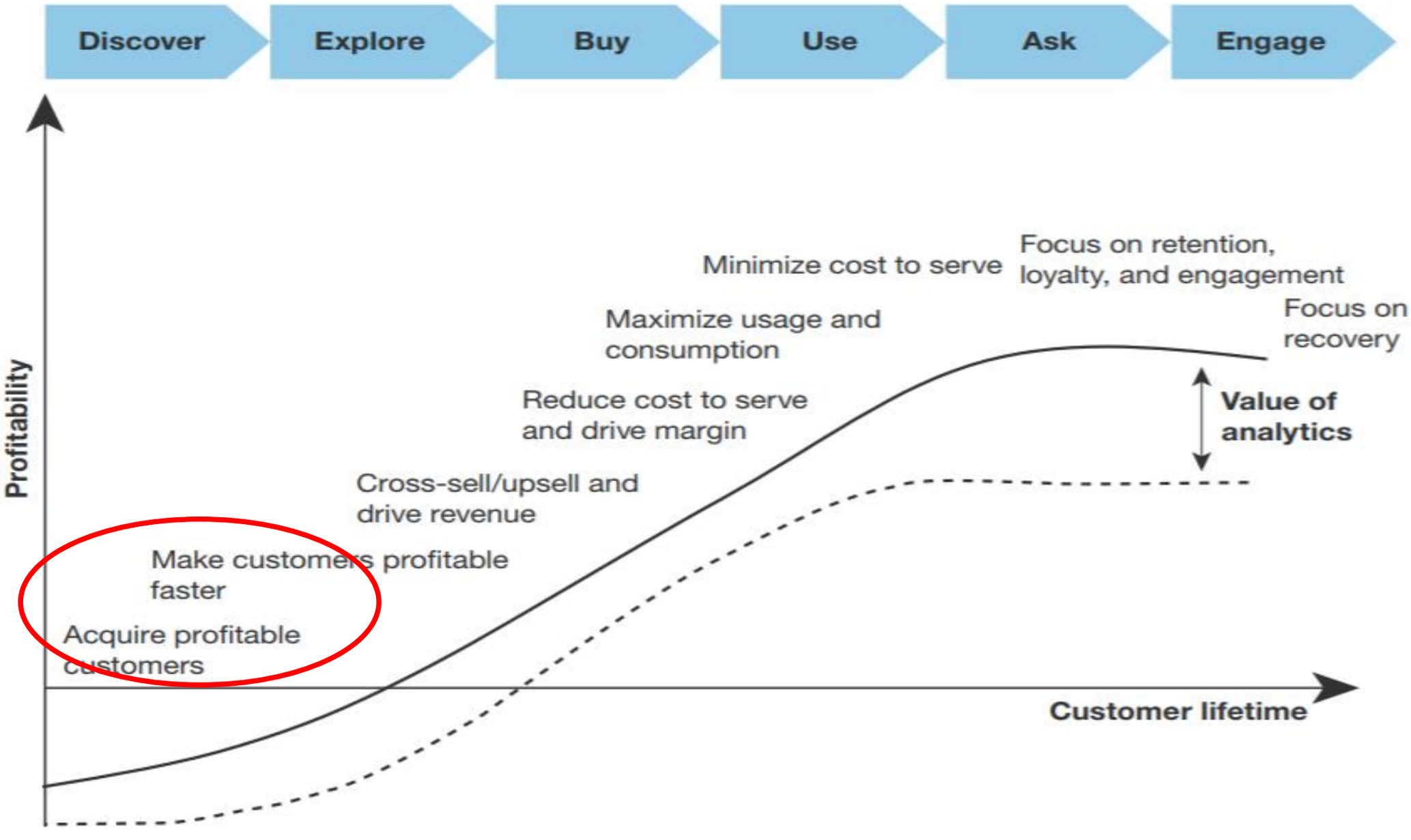
Customer lifecycle



Customer Lifecycle (Techniques/Approaches)

Problems	**Solutions**
• Start with understanding of your existing customers (Segment the customers)	• Clustering (Lecture 3)
• Acquire profitable customers	• Regression techniques (Present, Lecture 4)
• Understanding future behavior with Propensity Model	• Classification (Lecture 5)
• Convince/Influencing your customers to Spend more	• Cross-selling/Up-selling (Lecture 6)
• Causality or Just by chance?	• AB Testing (Lecture, 7)

Customer lifecycle (Returns)



https://www.sas.com/content/dam/SAS/en_us/doc/analystreport/forrester-analytics-drives-customer-life-cycle-management-108033.pdf

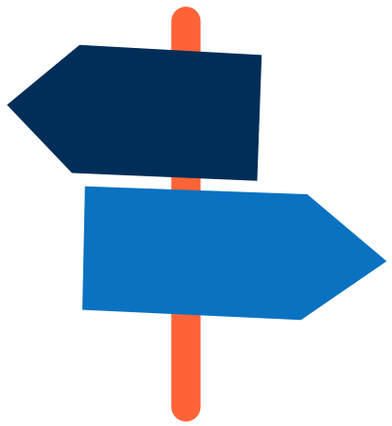
Customer lifetime value (CLV)

- or **CLTV**, lifetime customer value (**LCV**), or life-time value (**LTV**) .
- Describes the amount of profit a customer generates over his or her entire lifetime*
- We attempt to minimize “cost per acquisition” (CAC) and keeping any given customer.

*: Very much depends on the domain. For example, 1 to 20 years.

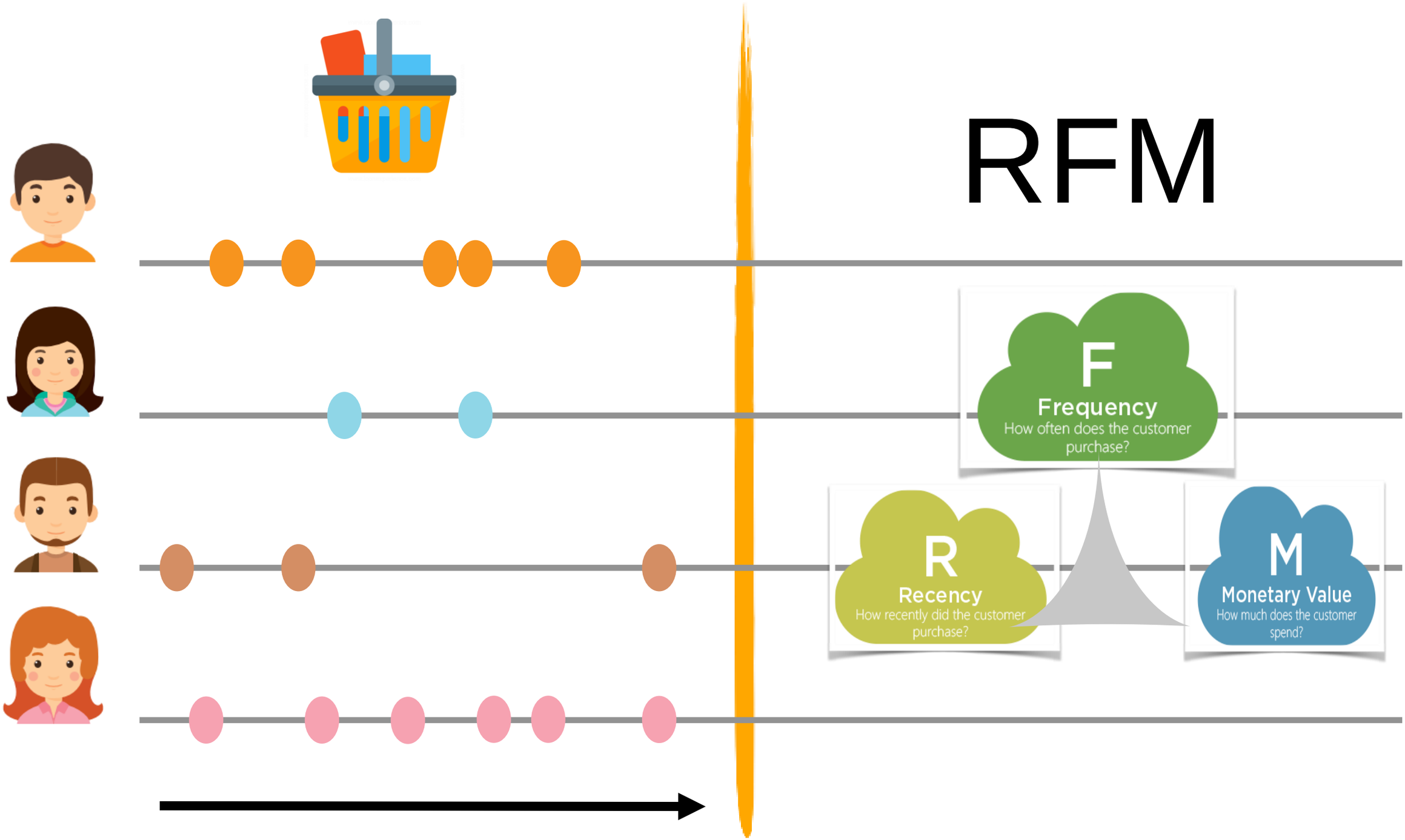
CLV is often referred to two forms of lifetime value analysis:

Historical lifetime value: simply sums up revenue or profit per customer.



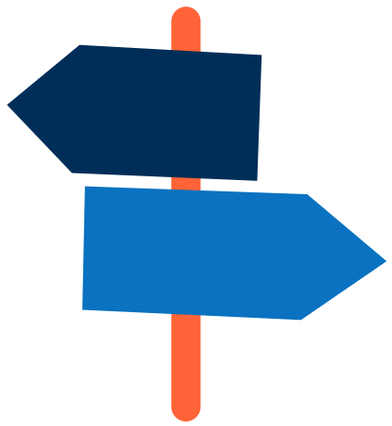
Predictive lifetime value: projects what new customers will spend over their entire lifetime.

Historical Life time value



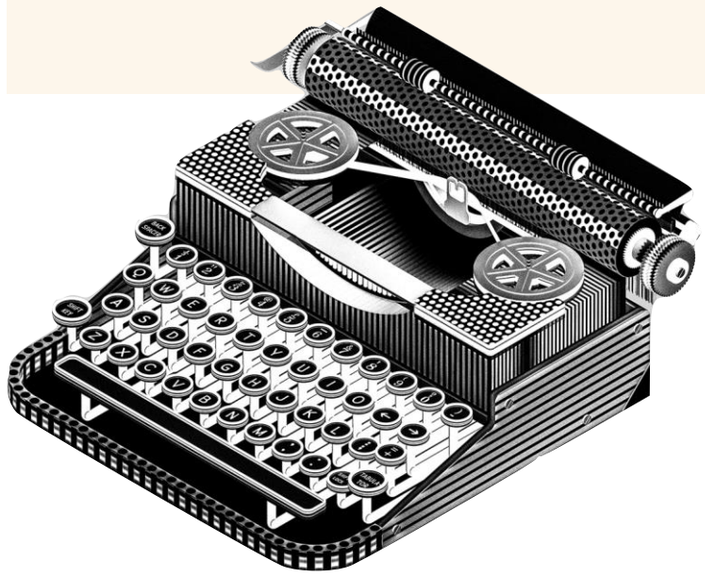
CLV is often referred to two forms of lifetime value analysis:

Historical lifetime value: simply sums up revenue or profit per customer.

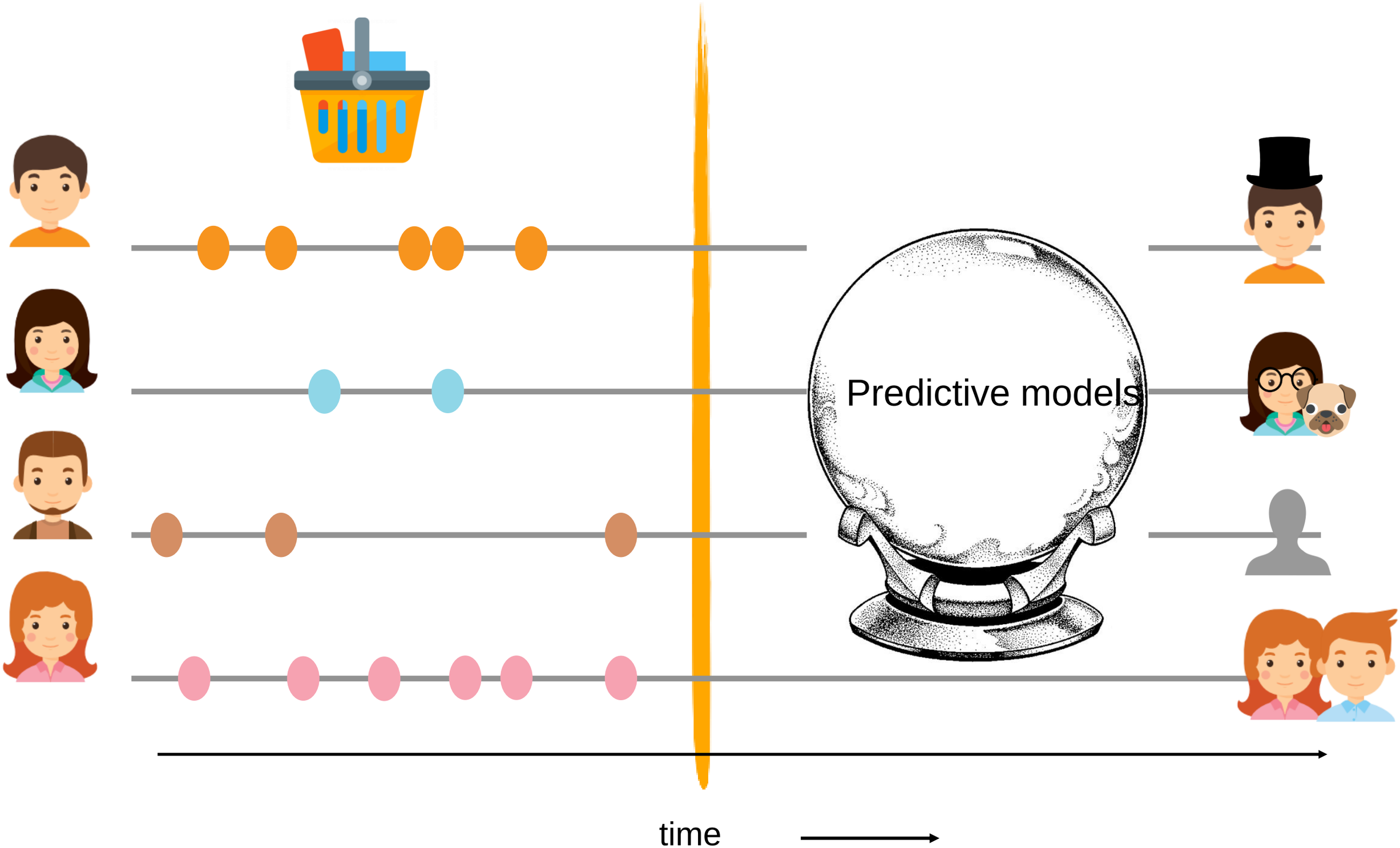


Predictive lifetime value: projects what new customers will spend over their entire lifetime.

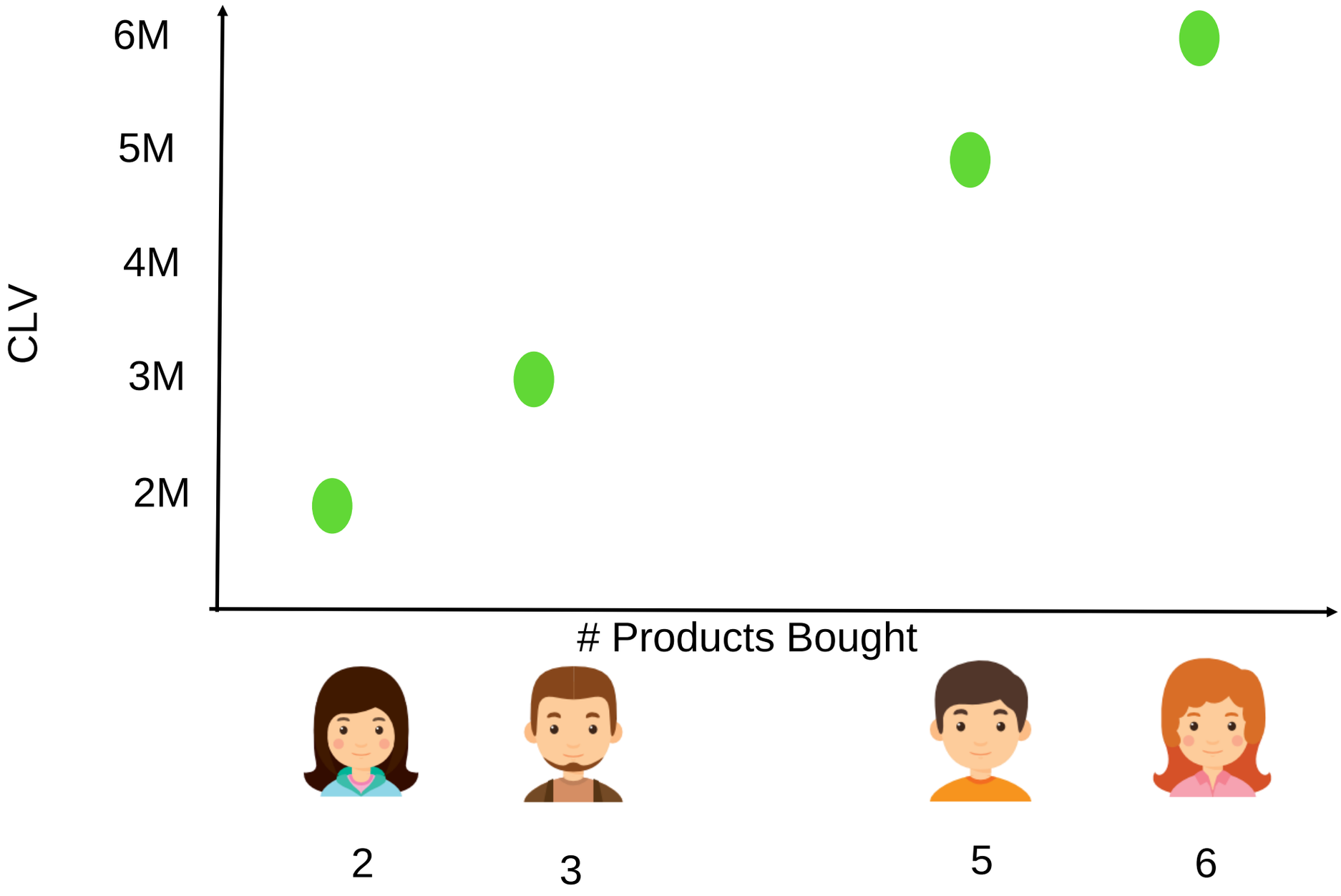
“Algorithms predict purchase frequency, average order value, and propensity to churn to create an estimate of the value of the customer to the business.
Predictive CLV is extremely useful for evaluating acquisition channel performance, using modeling to target high value customers, and identifying and cultivating VIP customers early in their brand journey.”



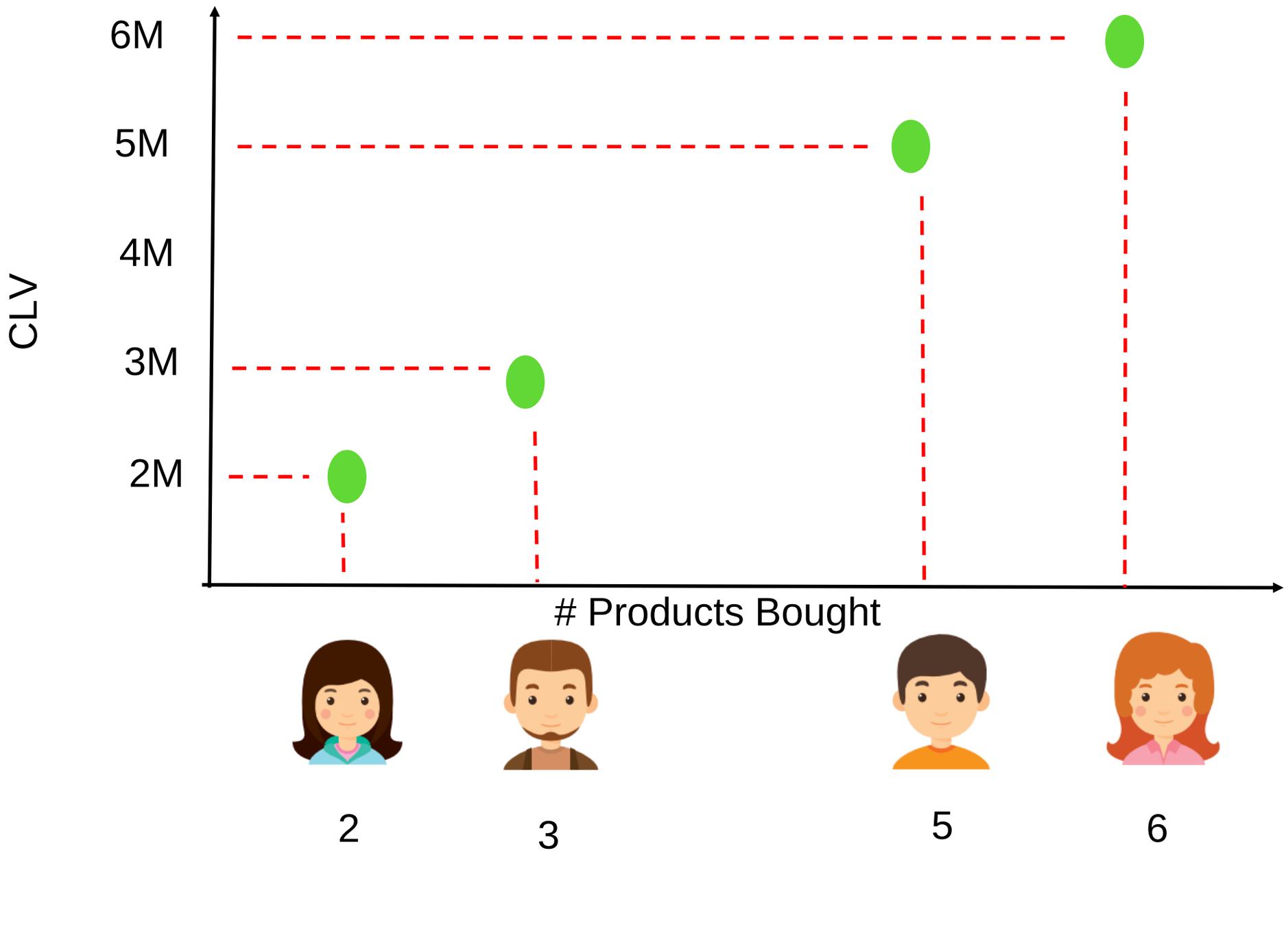
Predicting Life Time Value



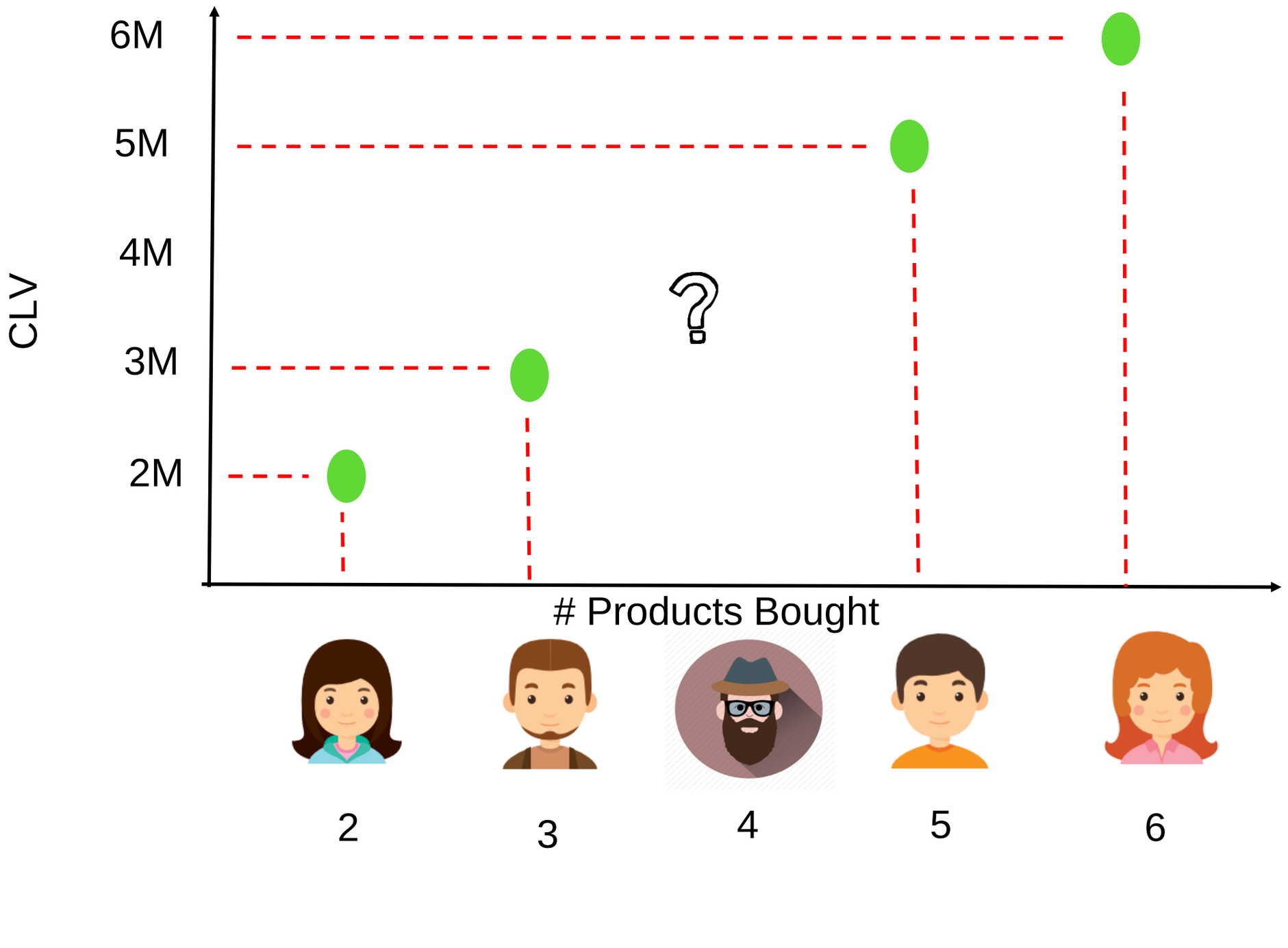
Mapping CLV



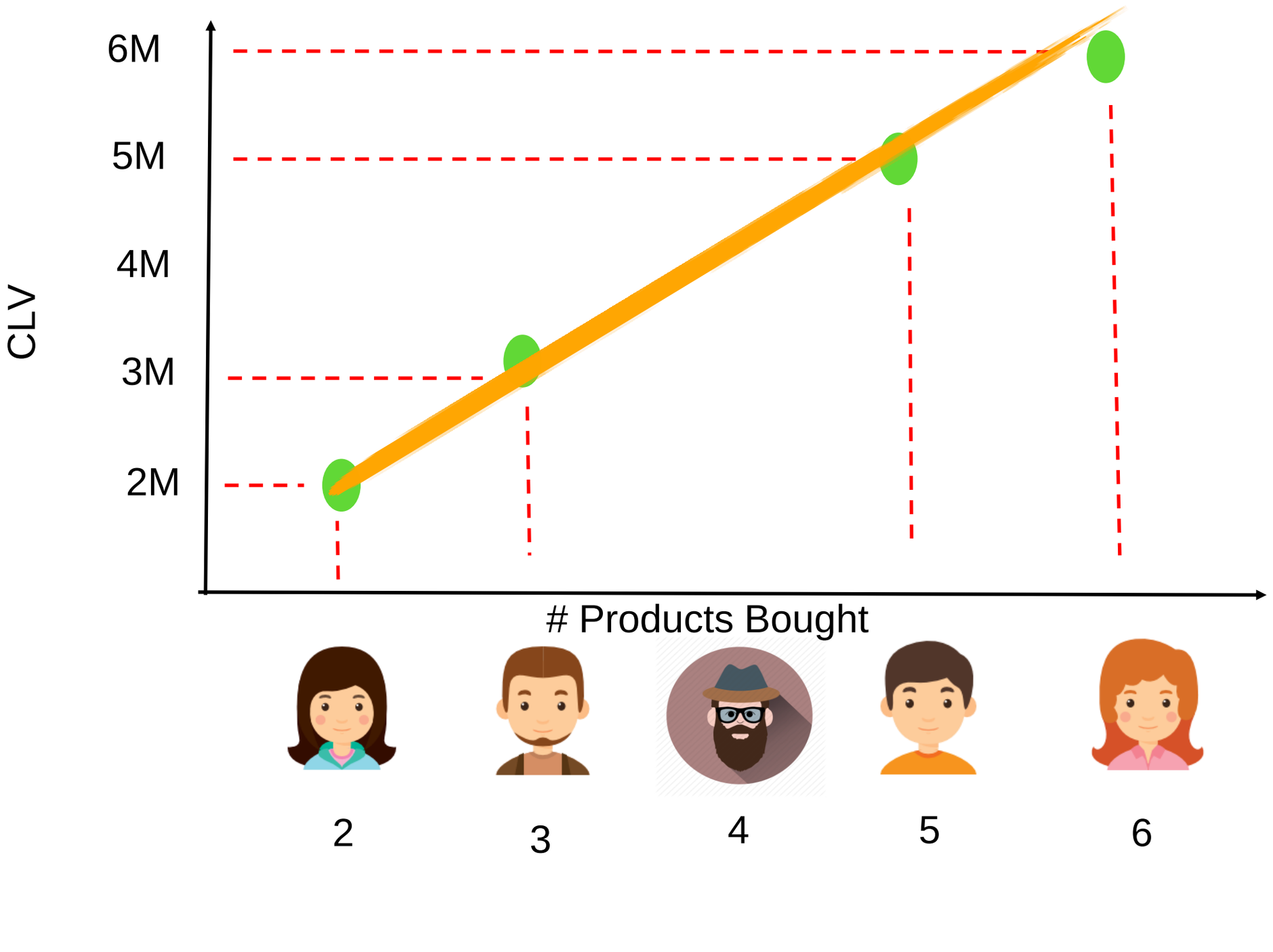
Mapping CLV



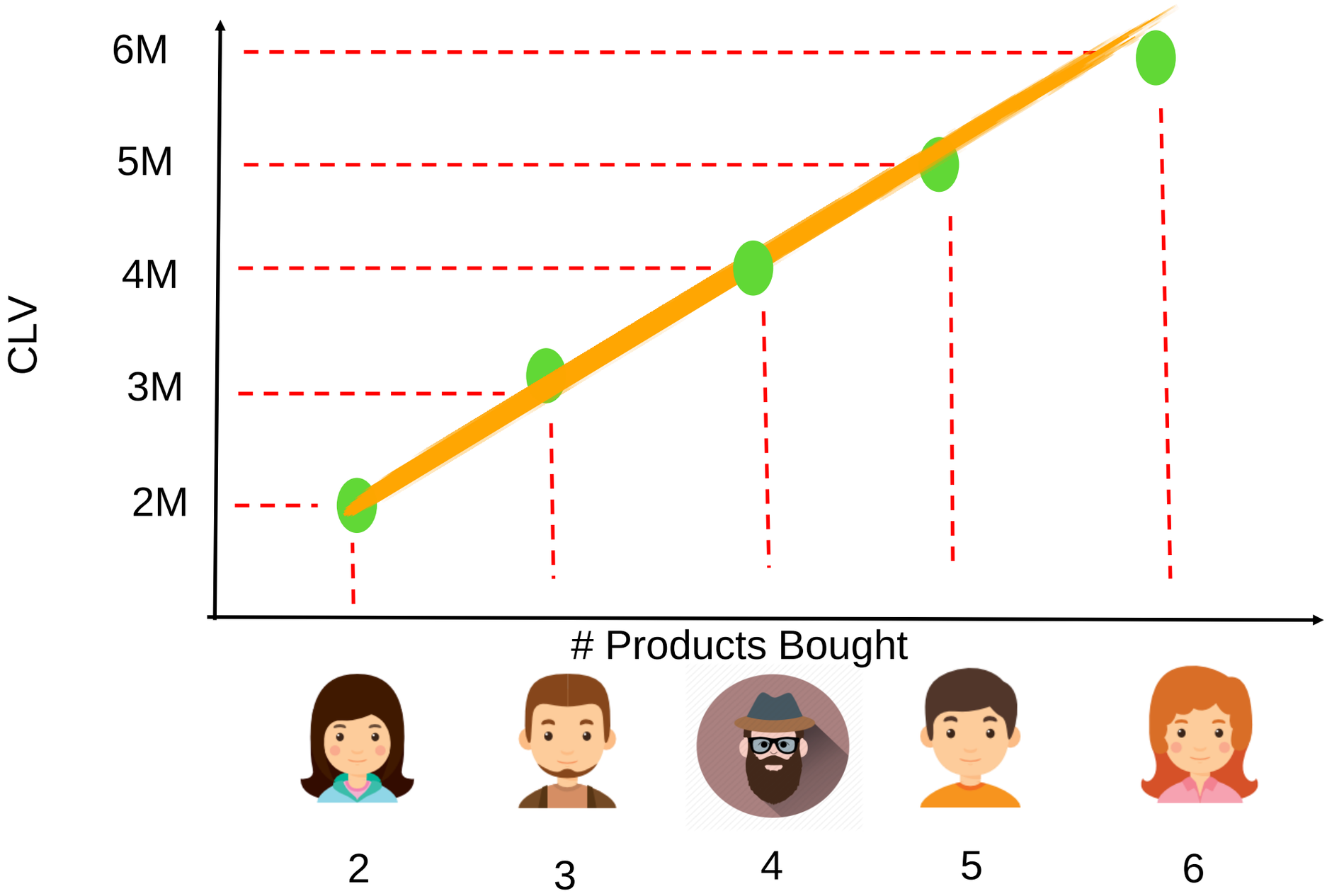
Predicting CLV



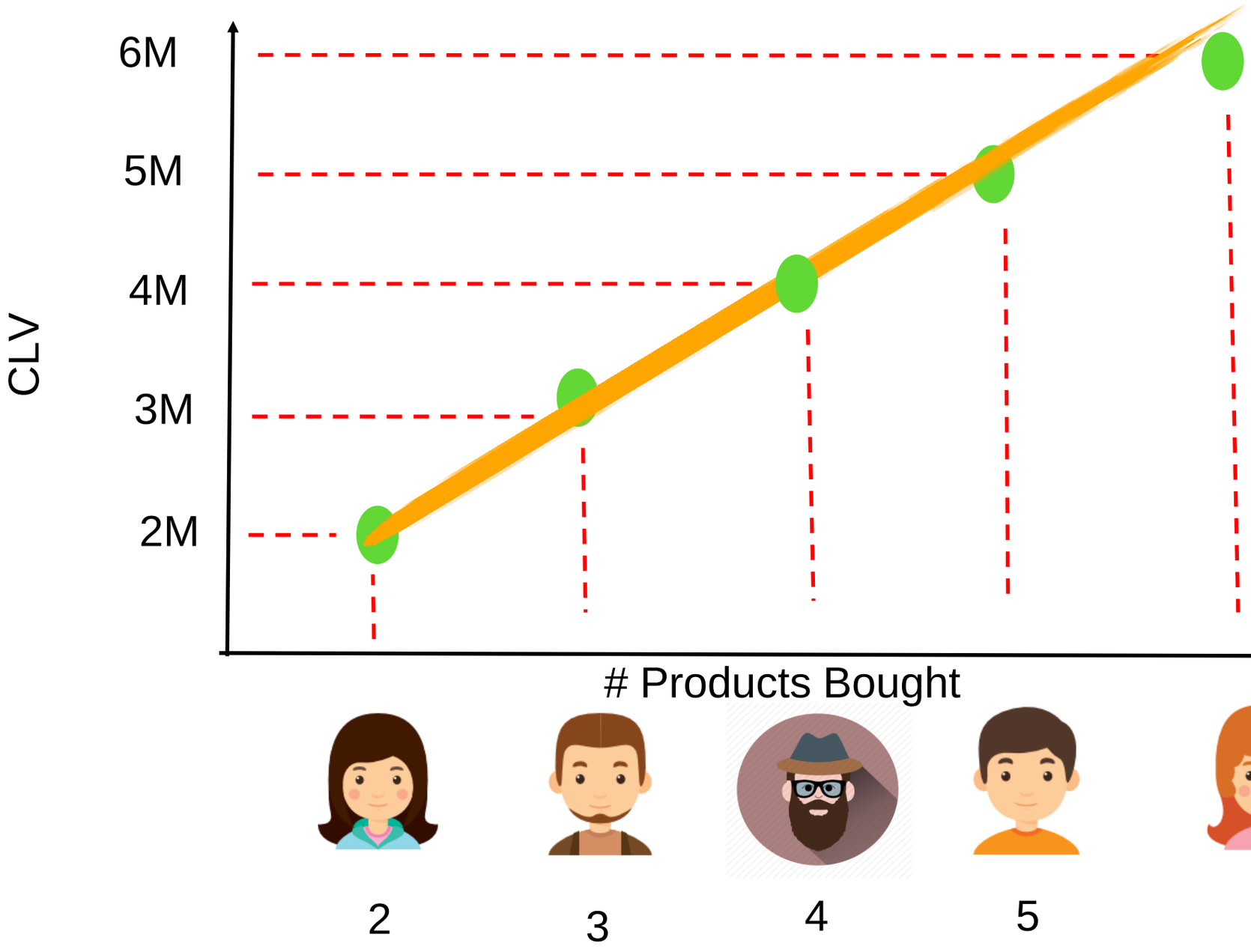
Predicting CLV



Predicting CLV



Predicting CLV

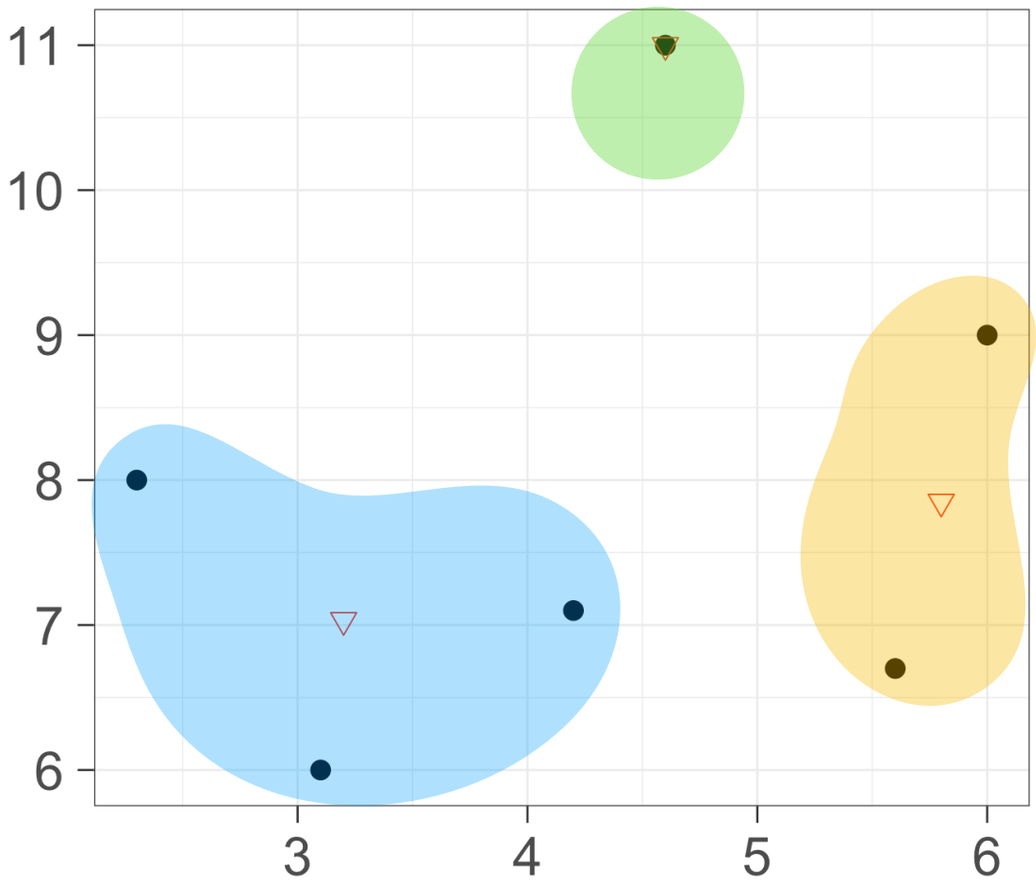


We just did a linear regression

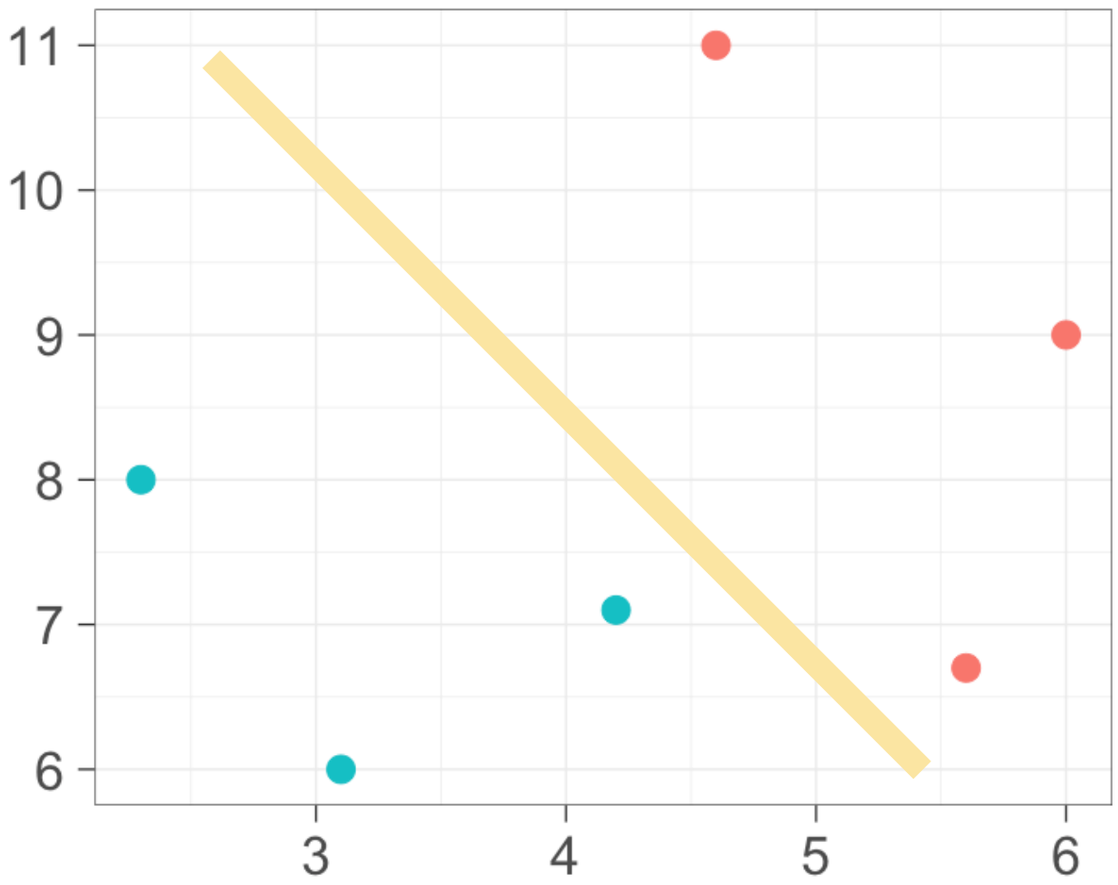
NOTE: But the data is not always spread like this (that is in this nice way)

Technical Terms

Unsupervised learning



Supervised learning



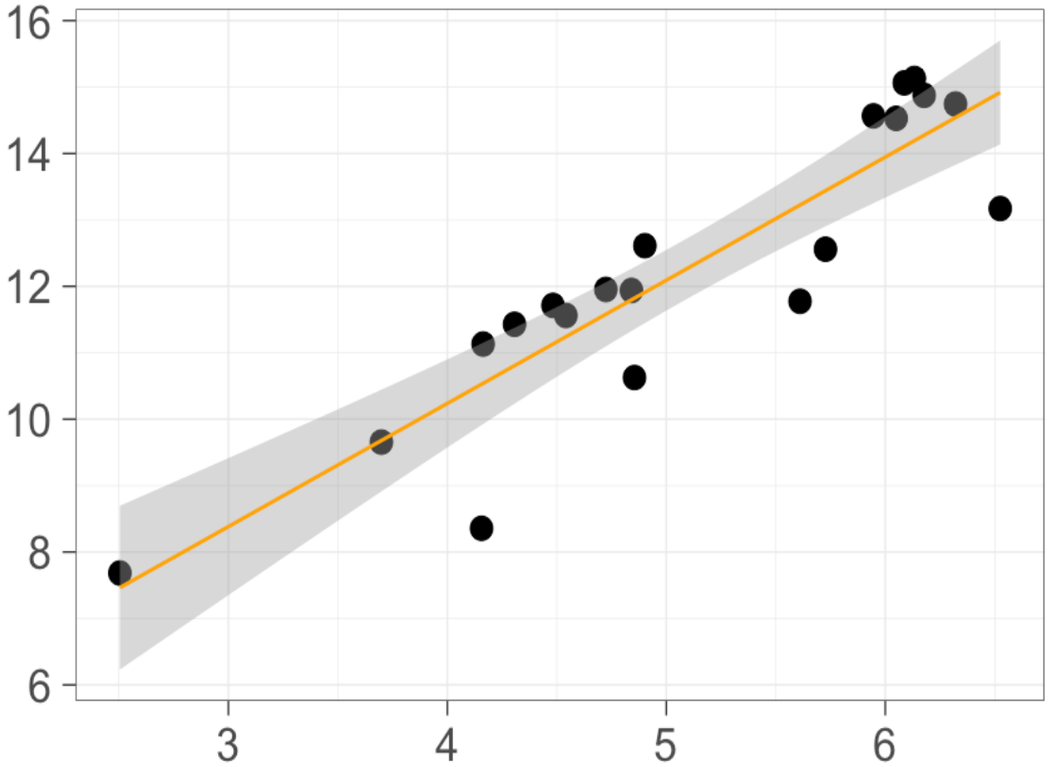
Supervised vs. Unsupervised Learning

The goal of the supervised approach is to learn function that maps input x to output y, given a labeled set of pairs $D = \{(\mathbf{x_i}, y_i)\}_{i=1}^{N}$

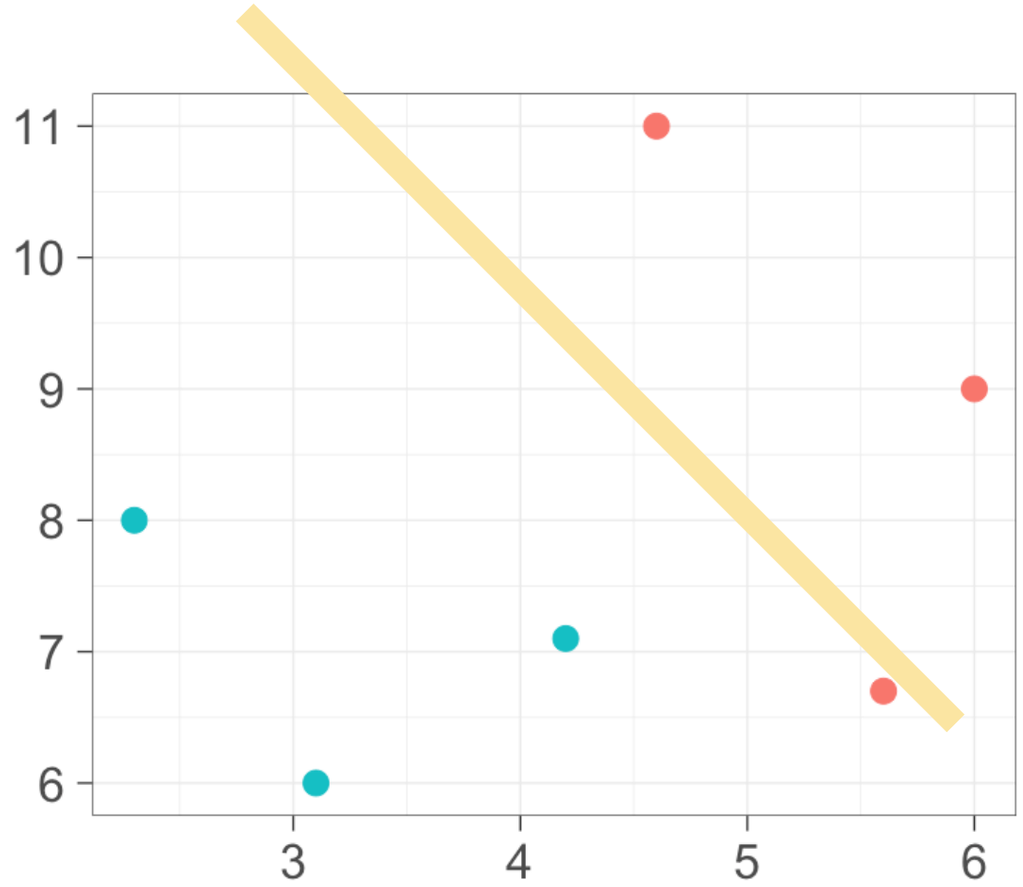
The goal of the unsupervised approach is to learn “interesting patterns” given only an input $D = \{\mathbf{x_i}\}_{i=1}^{N}$

Categorize different types of customers ?

Supervised learning
Regression vs. Classification

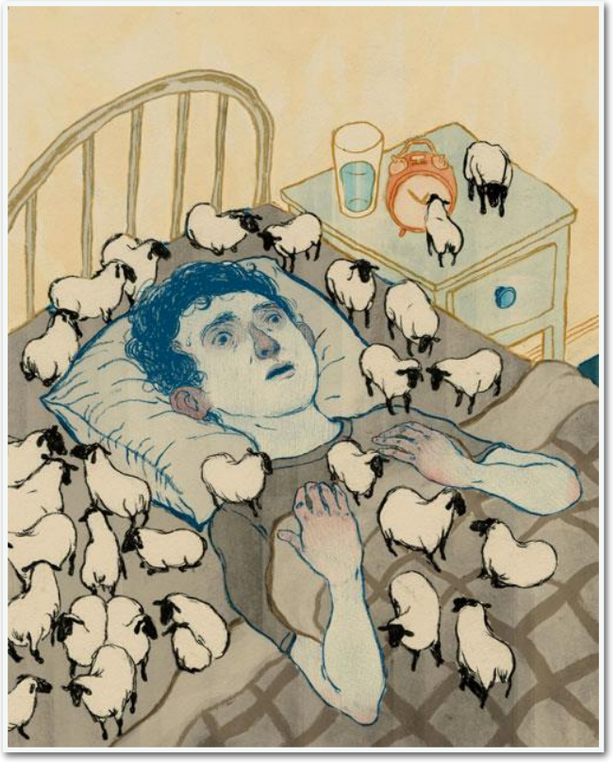


How many will leave ?



If a particular customer will leave or not ?

Sleeping habits

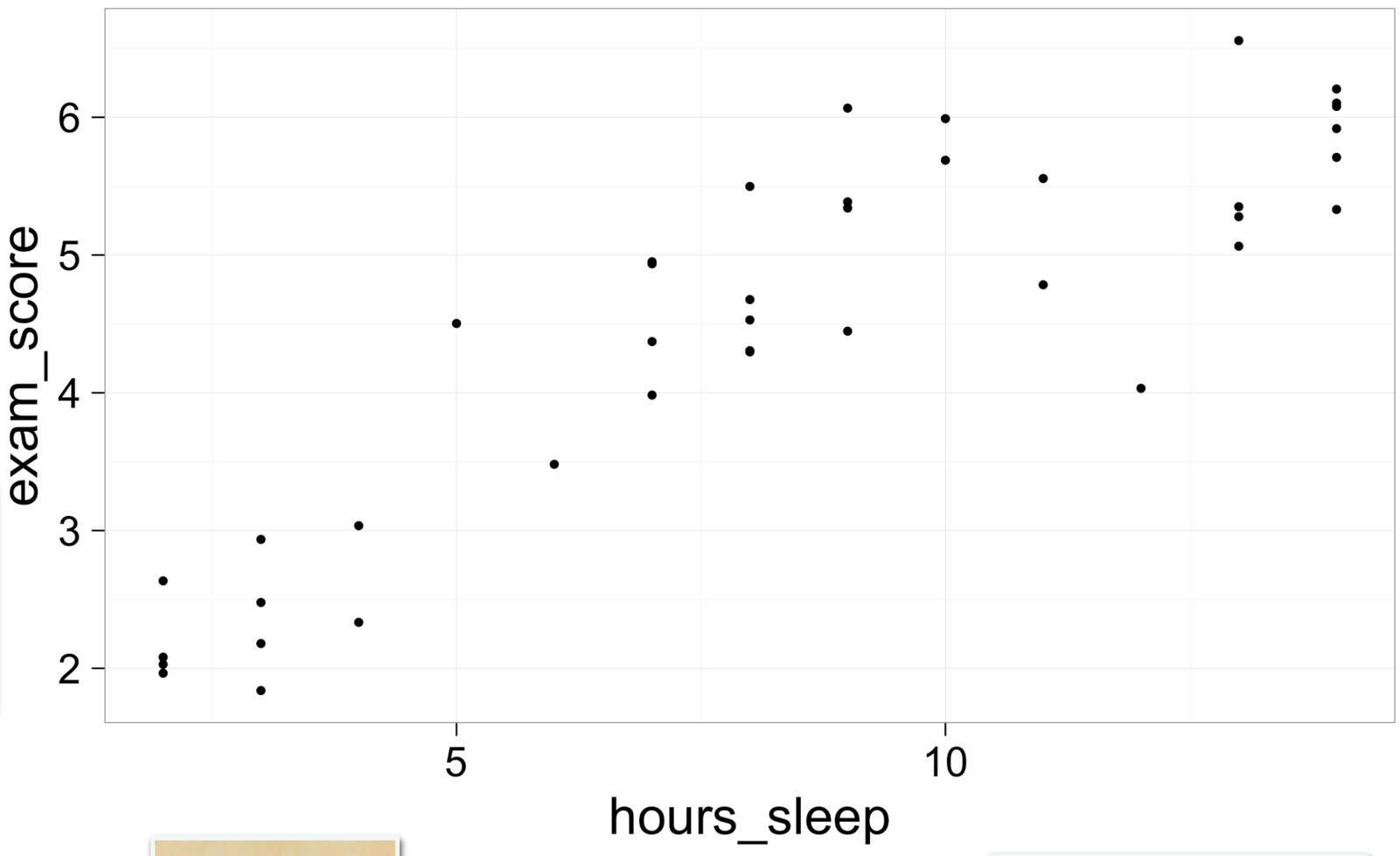


4 hours of sleep

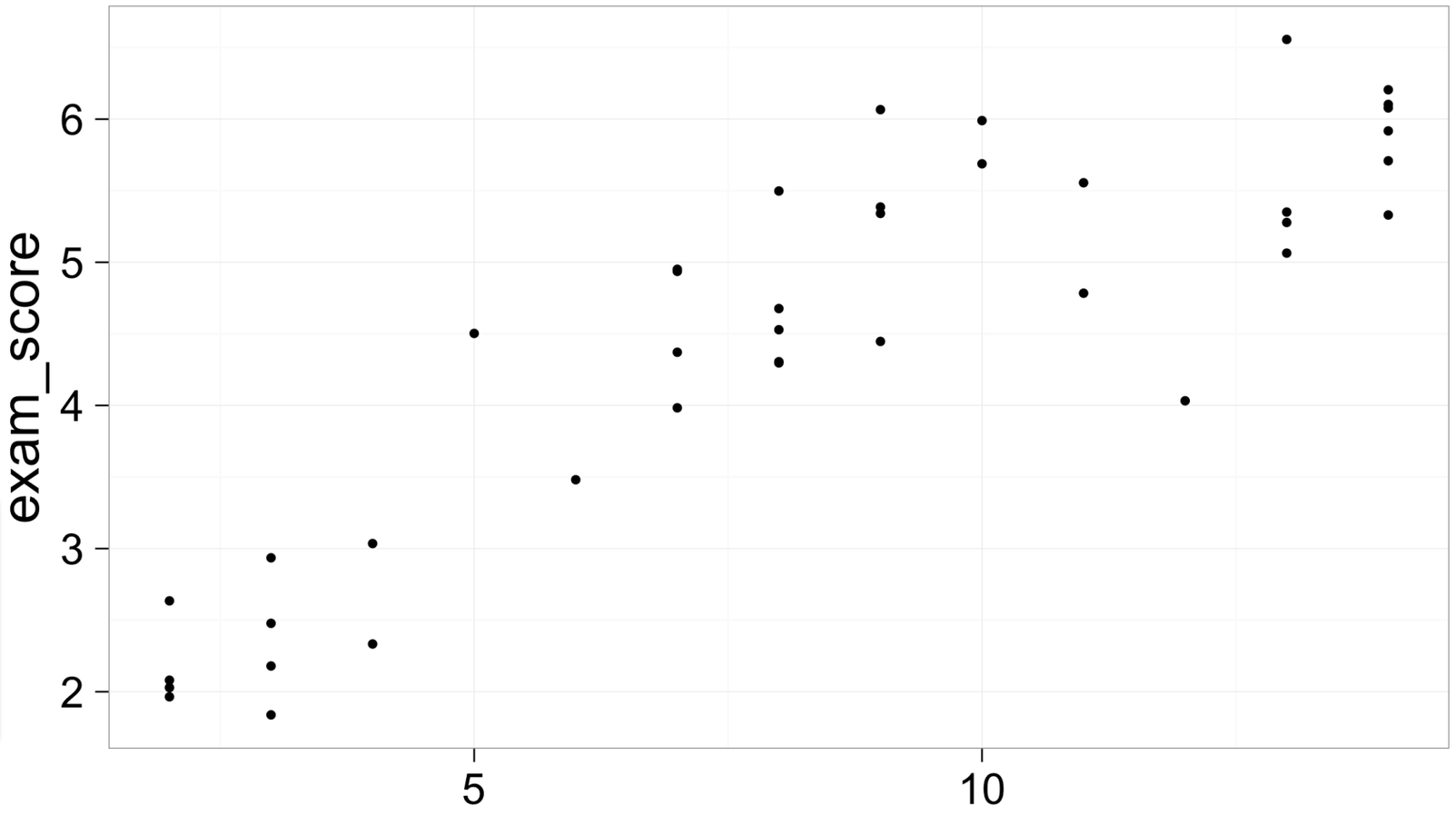
8 hours of sleep

exam performance

Simple Linear regression

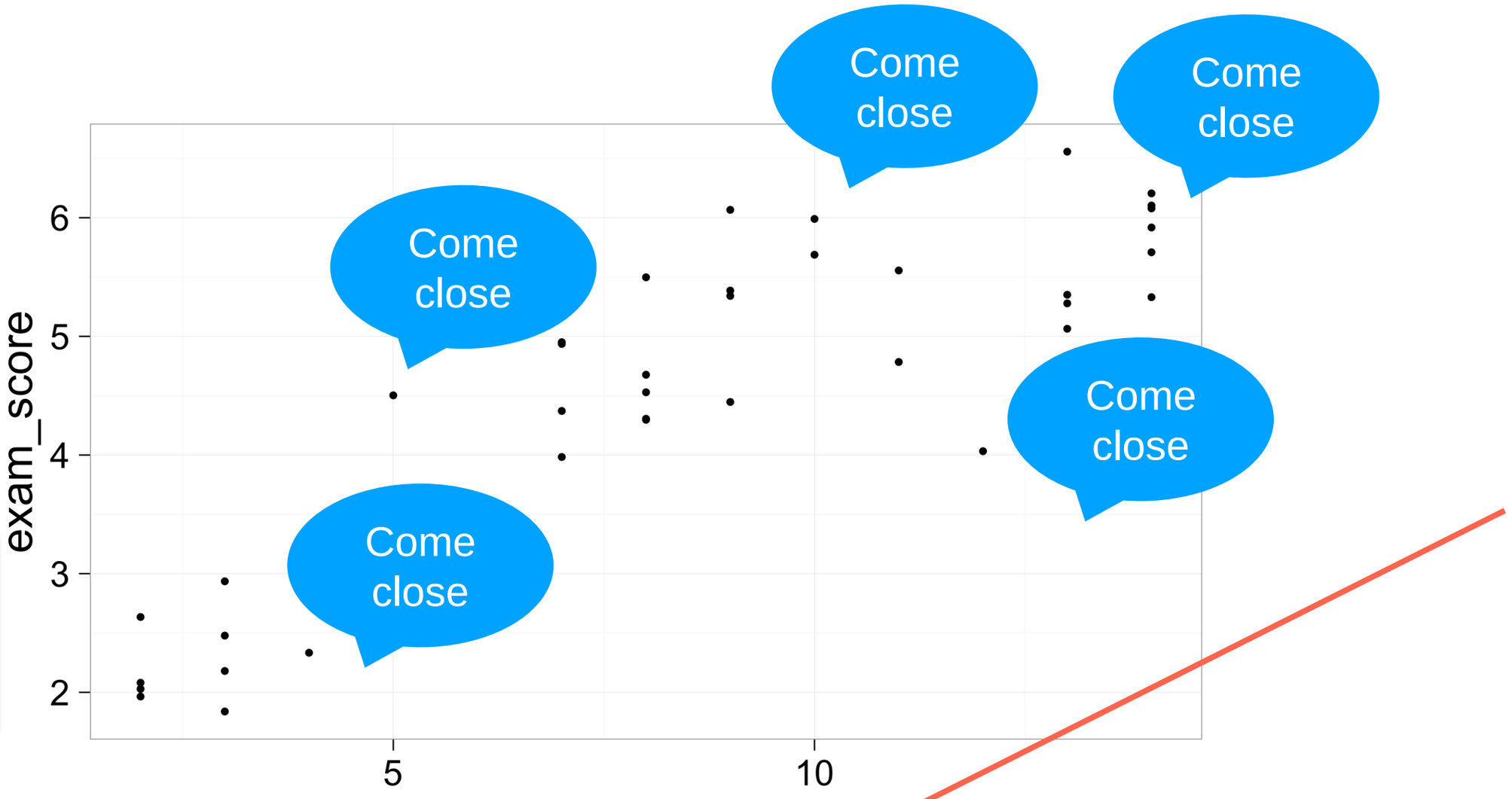


Simple Linear regression

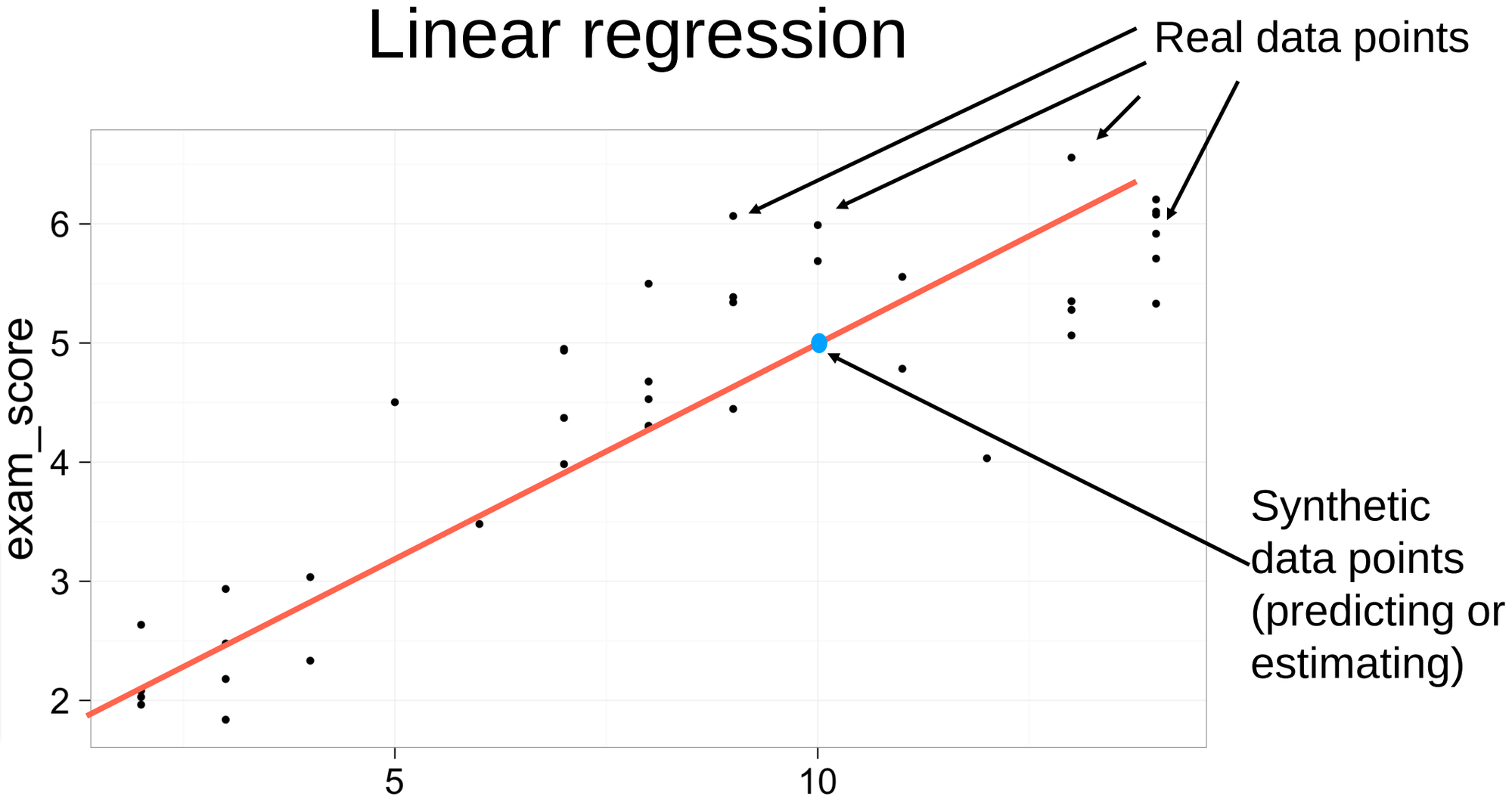


How to fit the line ?

Simple Linear regression



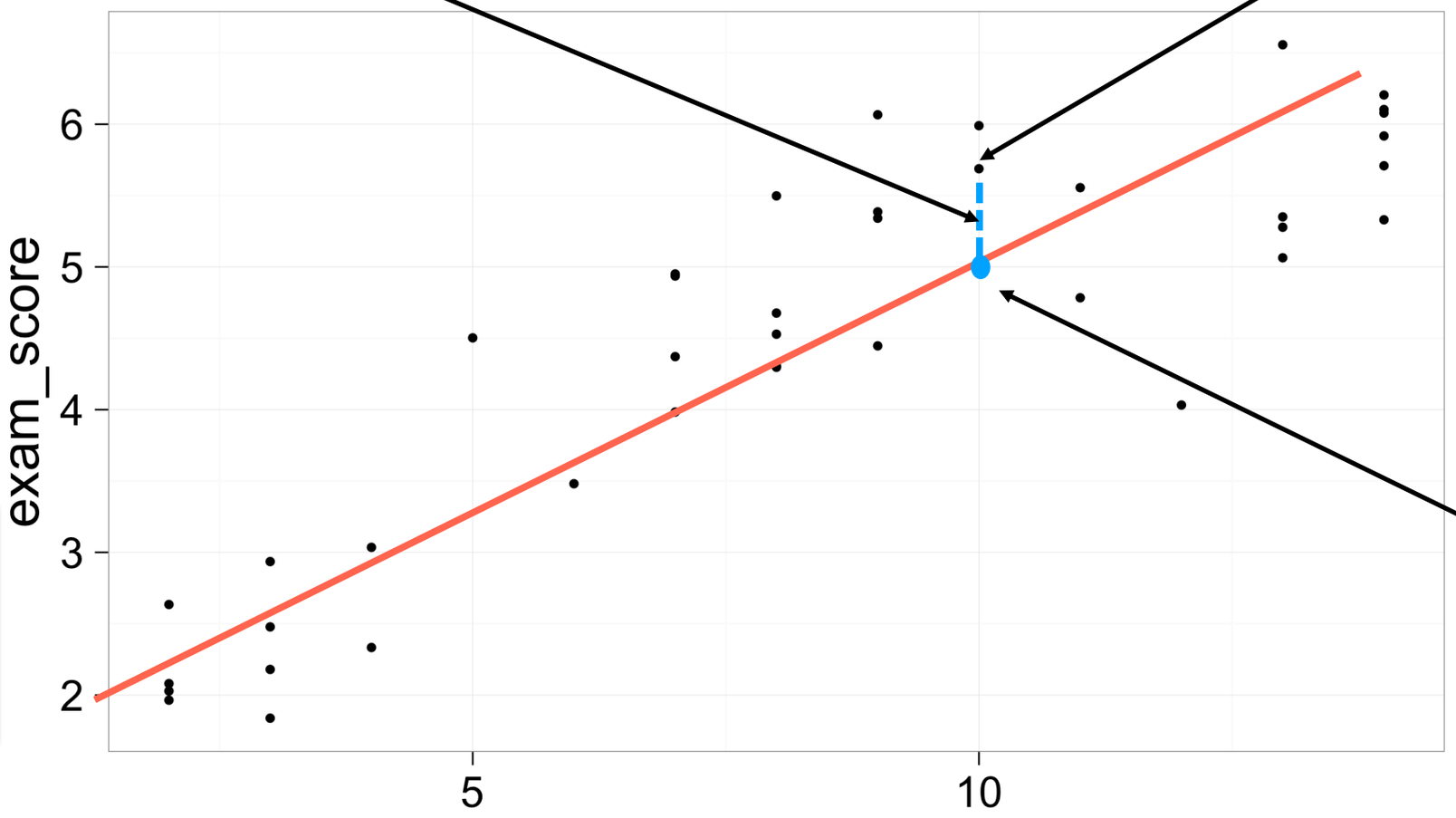
Linear regression



Error = (obs. - Pred.)

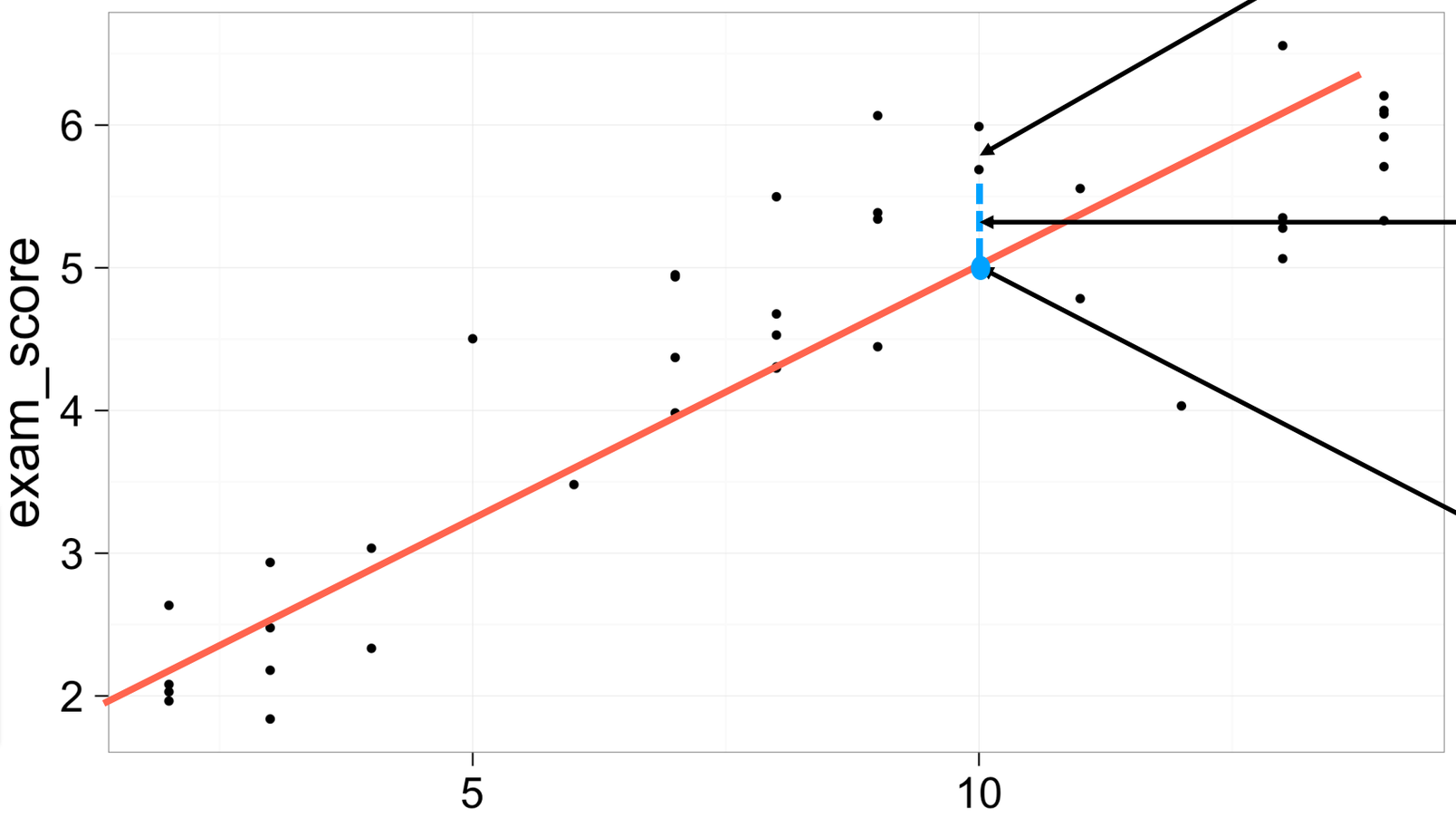
Linear regression

Real data point
(10,5.25)



Synthetic
data point
(estimating)
(10,5)

Linear regression

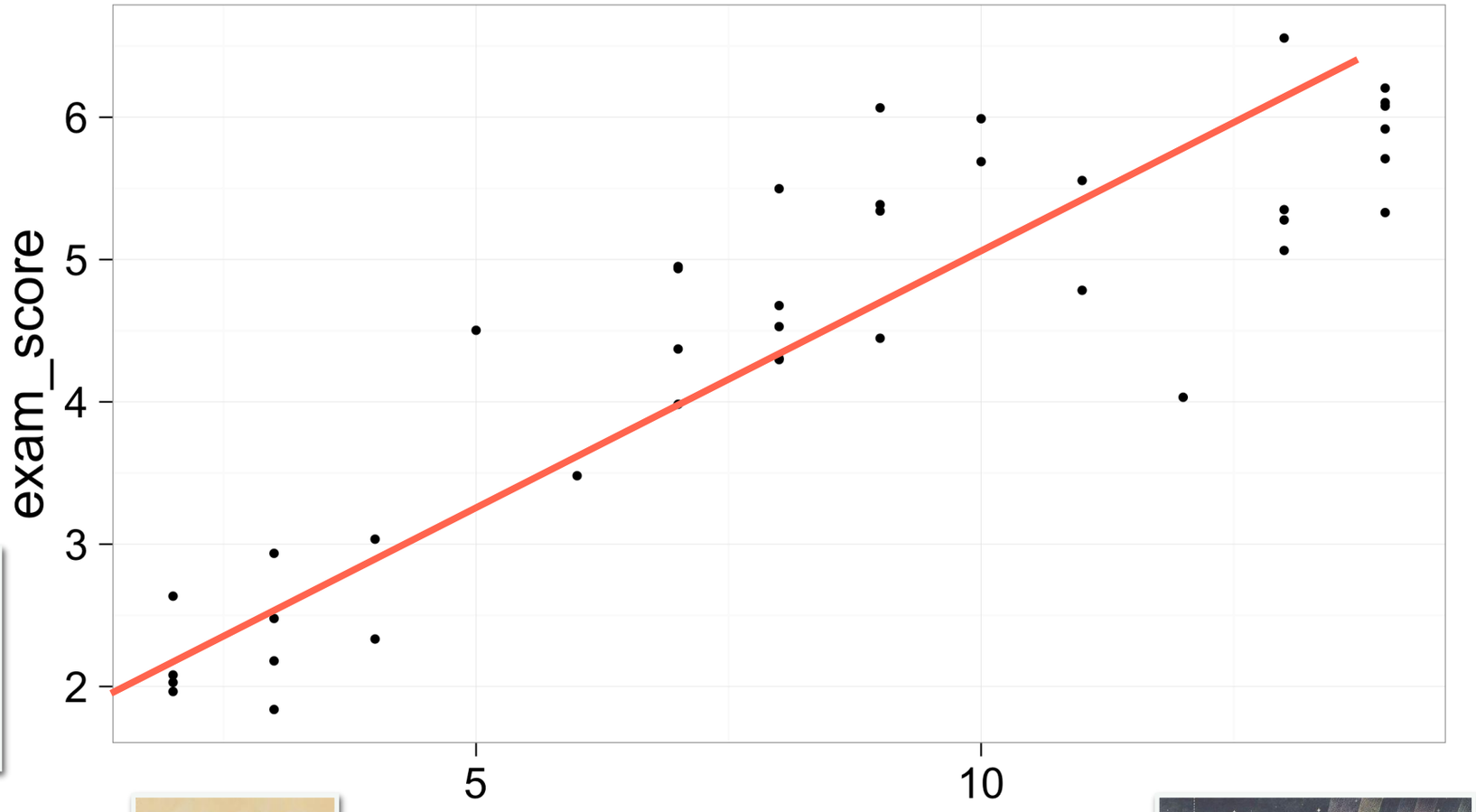


Real data point
(10,5.25)

Error

Synthetic
data point
(predicting or
estimating)
(10,5)

Simple Linear regression

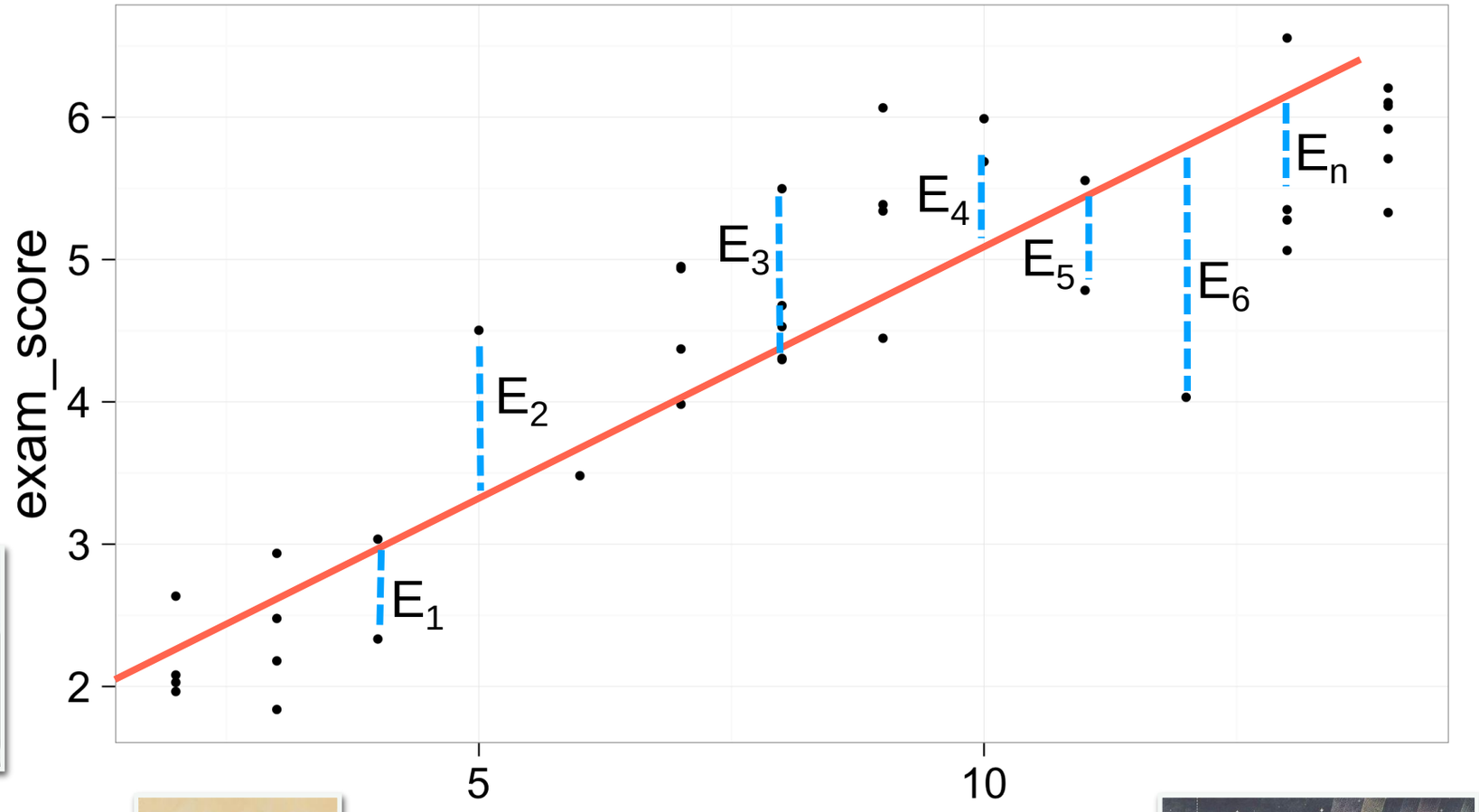


Task: given a list of observations $D = \{(x_i, y_i)\}_{i=1}^{N}$ find a line

$$\hat{y} = ax + b$$

that **approximates** the correspondence in the data

Simple Linear regression



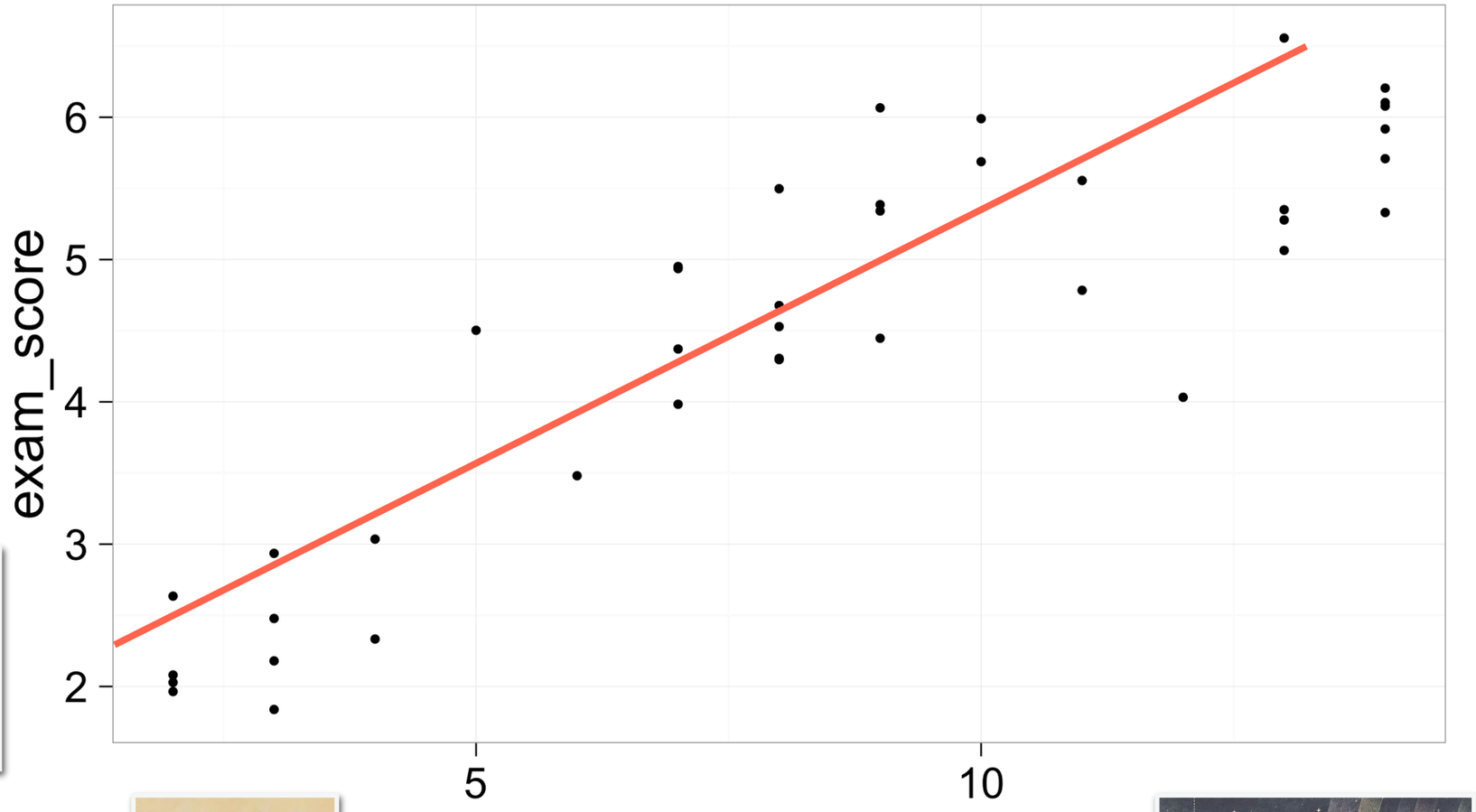
$$SS_A = E_1^2 + E_2^2 + E_3^2 + \ldots + E_n^2$$

Sum of Squares: to find the optimal line

*Sum of Squares: is a type of loss function.

*Sum of Squares: Least Squared error

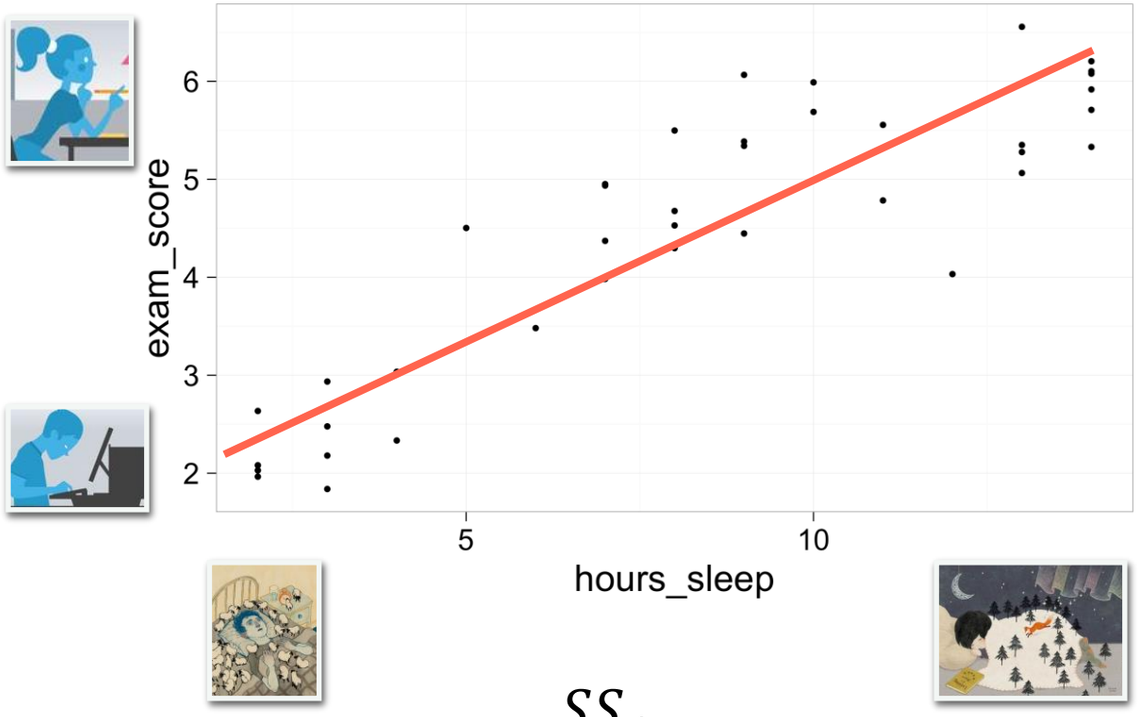
Simple Linear regression



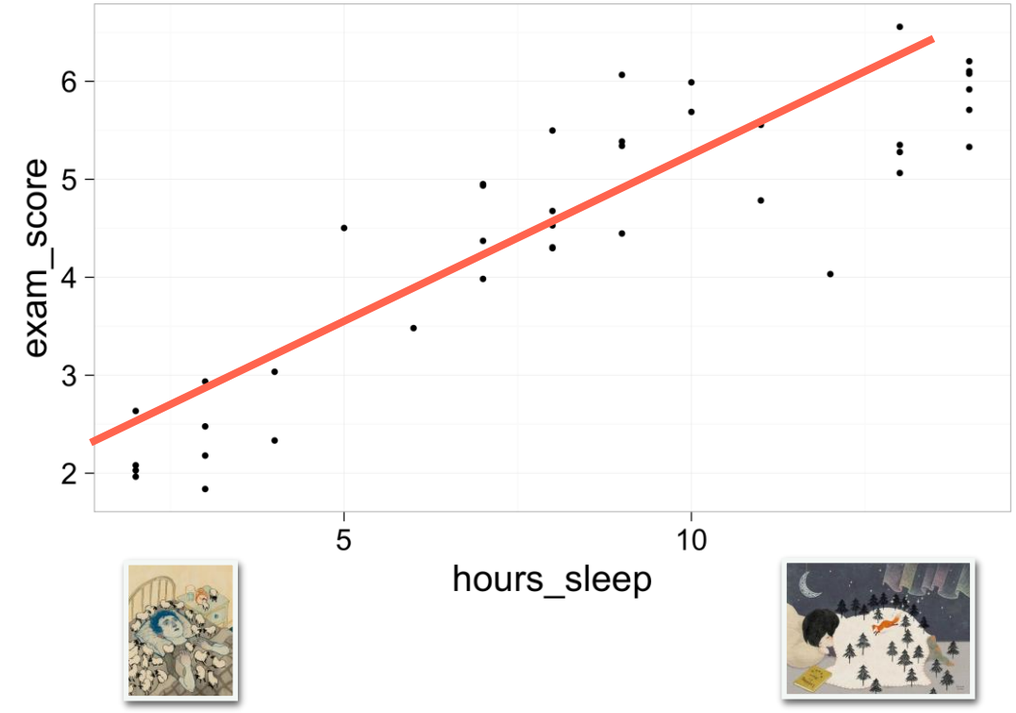
Case B: We move the line bit upward

Sum of Error for case B: SS_B

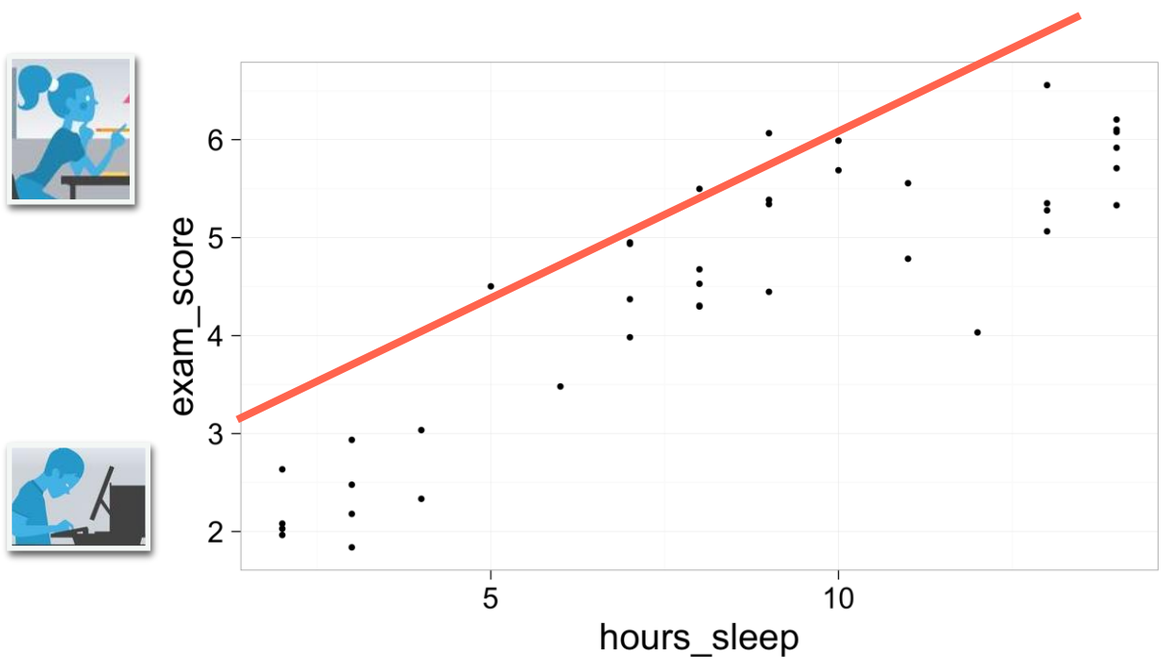
Finding optimal function



SS_A



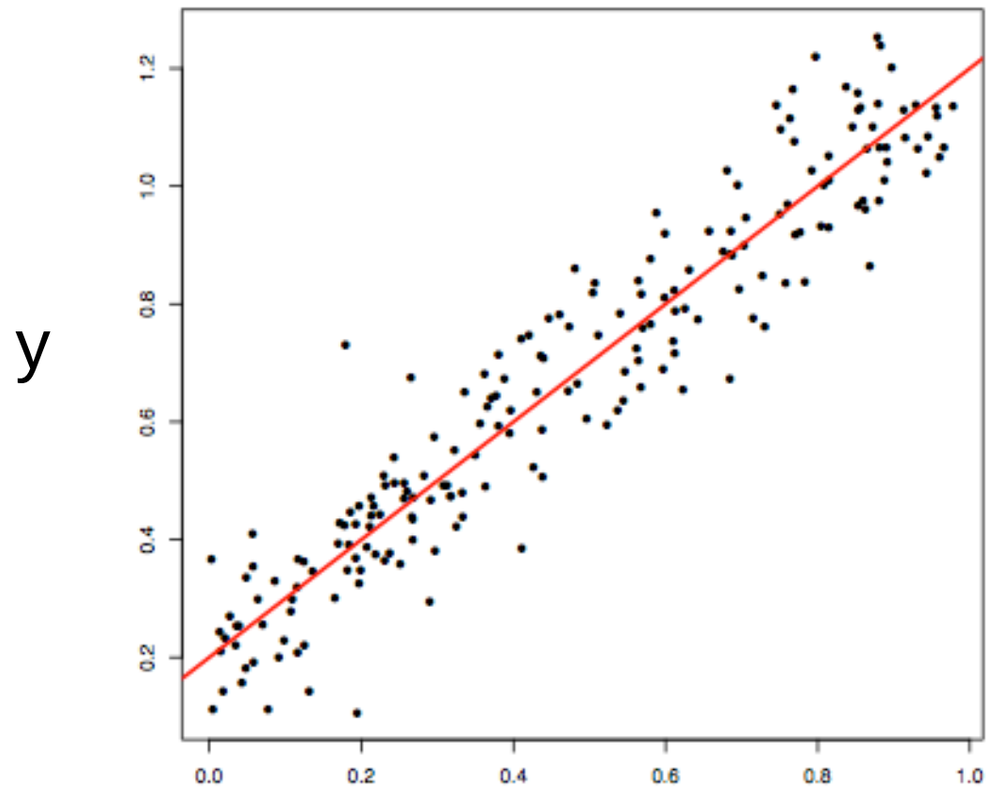
SS_B



SS_C

Q: Which line fits the well ?
Ans: Which has the minimum SS Error

Simple linear regression



Task: given a list of observations $D = \{(x_i, y_i)\}_{i=1}^{N}$ find a line

$$\hat{y} = ax + b$$

that approximates the correspondence in the data

Simple linear regression

$$y = \alpha + \beta x + \epsilon$$

output

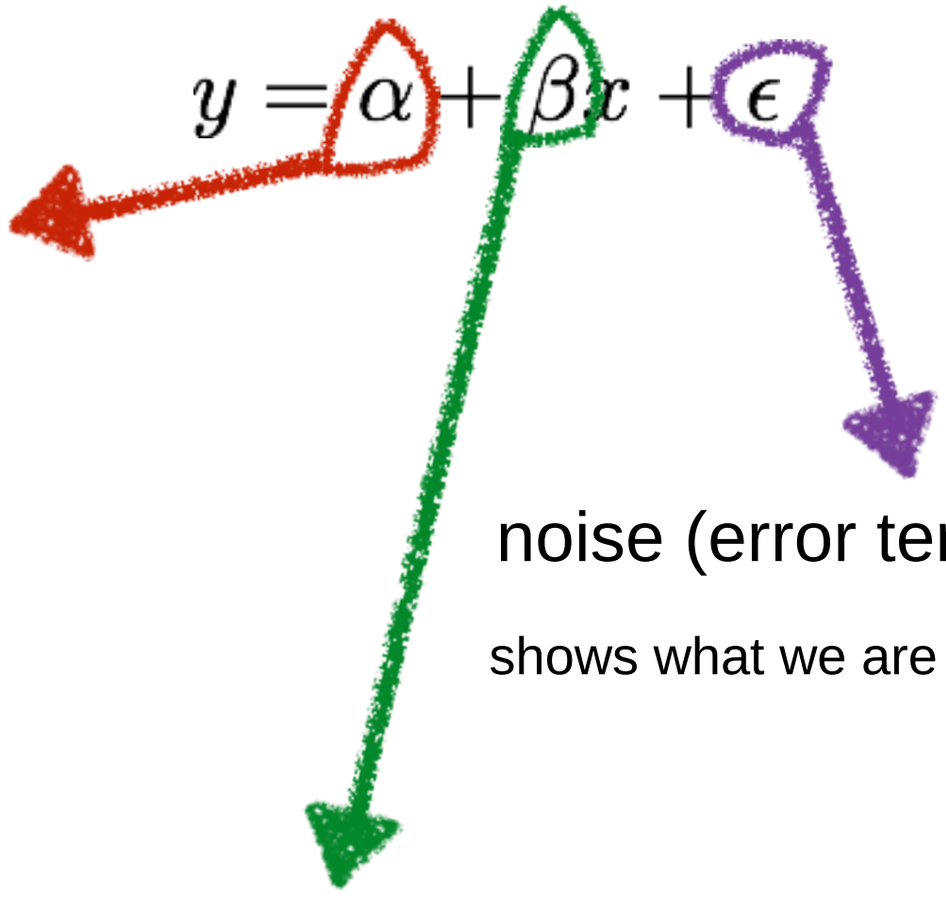
(dependent variable, response)

input

(independent variable, feature, explanatory variable, Predictor etc)

Simple linear regression

intercept (bias)



noise (error term, residual)

shows what we are not able to predict with x

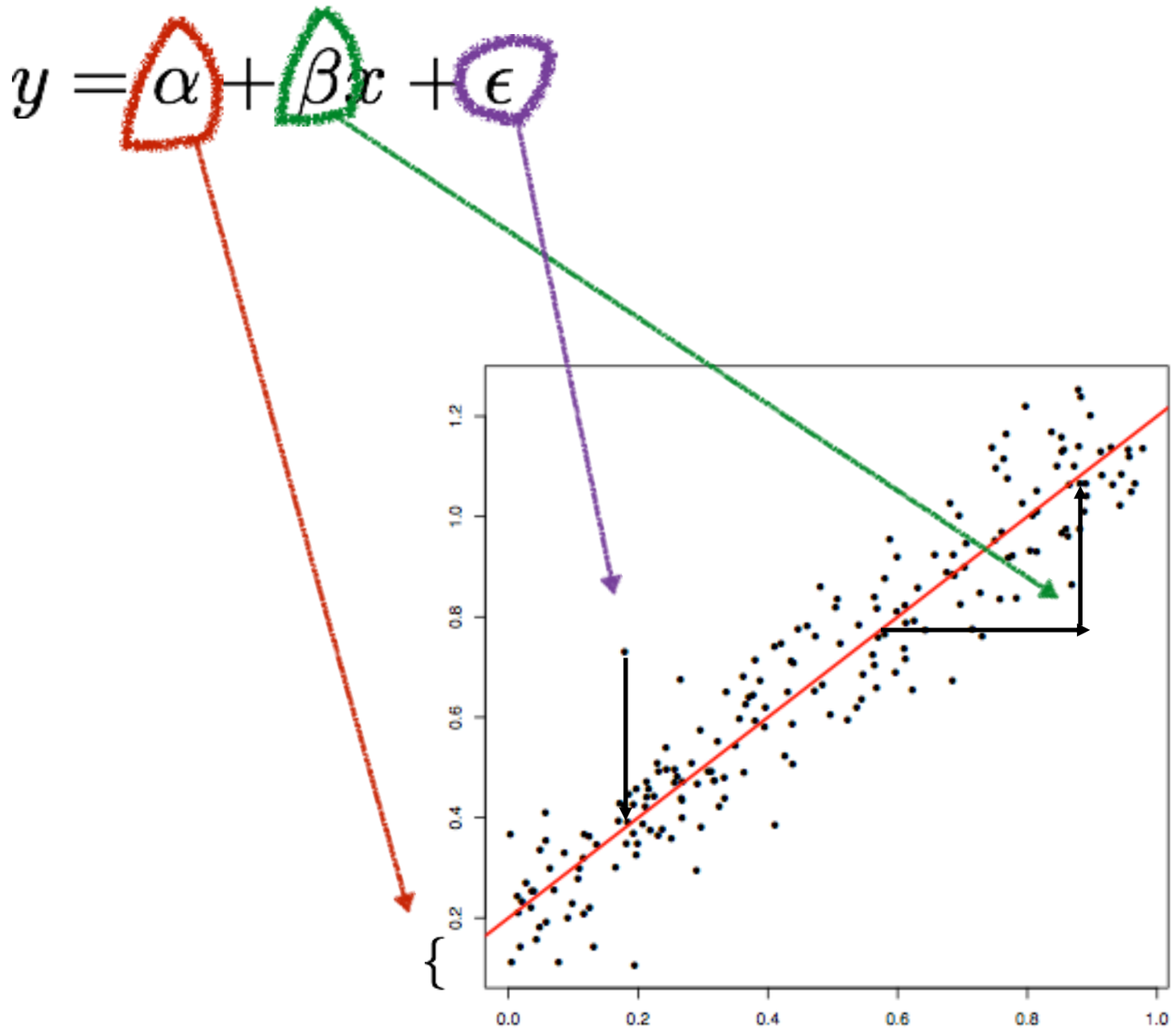
coefficient (slope, or weight w)

shows how increases output

if input increases by one unit

Simple linear regression

$$y = \alpha + \beta x + \epsilon$$



Simple linear regression

We search for a function $\hat{y} = f(x)$

such that minimizes mean squared error (MSE) :

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad = \quad \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta x_i - \alpha)^2$$

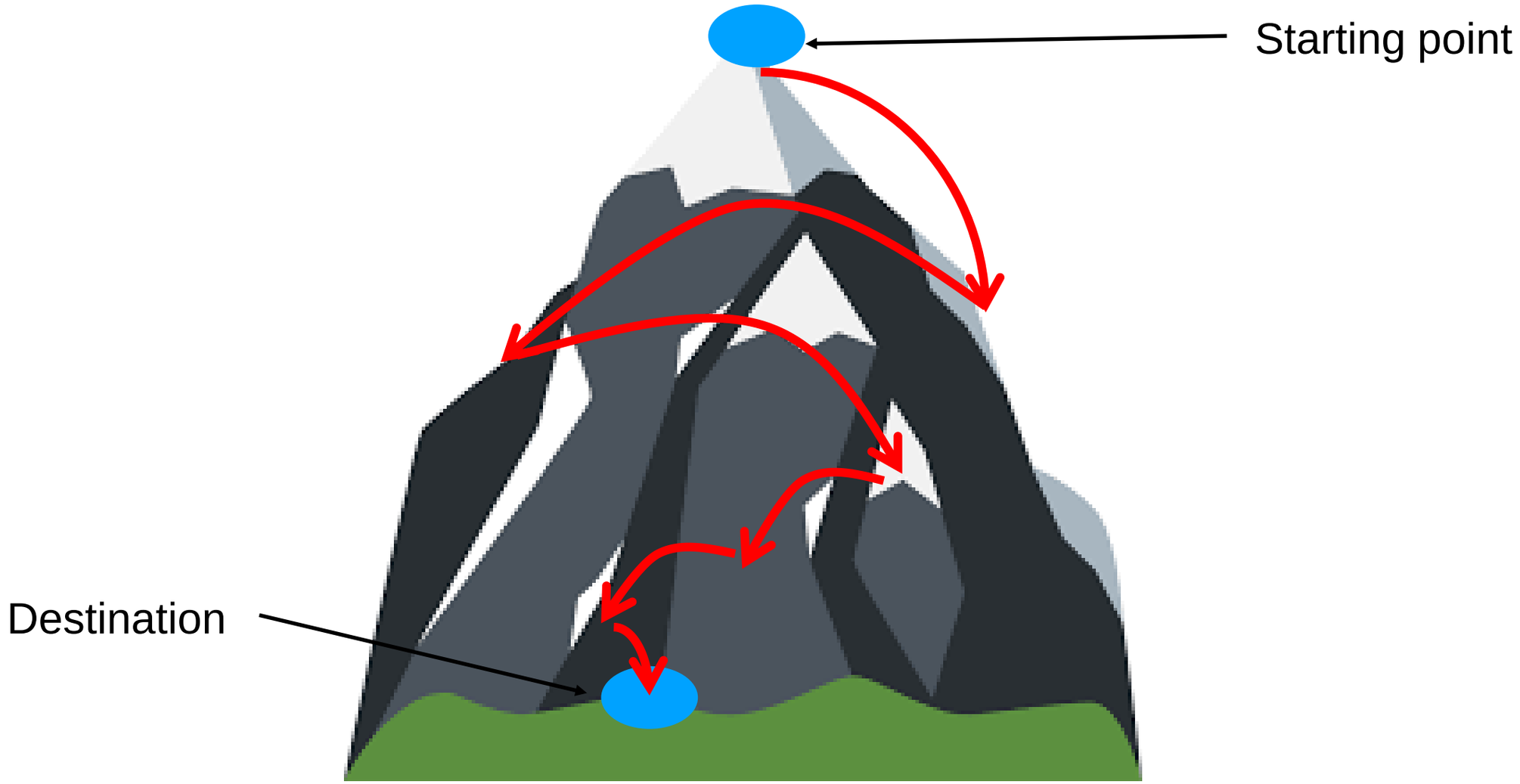
which means to find derivatives wrt α and β

and solve the system of equations:

$$\begin{cases} \frac{\partial MSE}{\partial \alpha} = 0 \\ \frac{\partial MSE}{\partial \beta} = 0 \end{cases}$$

Q: How to optimally fit the line?

Ans: Gradient Descent



Suggested Reading

https://medium.com/code-heroku/gradient-descent-for-machine-learning-3d871fa48b4c (concise but misses some stuff)
https://towardsdatascience.com/understanding-the-mathematics-behind-gradient-descent-dde5dc9be06e (elaborate)
https://www.youtube.com/watch?v=sDv4f4s2SB8&t=202s (video by stat quest)
https://www.kdnuggets.com/2017/04/simple-understand-gradient-descent-algorithm.html (Housing price)

Simple linear regression: example in python

A collection of observations of the Old Faithful geyser

in the USA Yellowstone National Park


```
df=  pd.read_csv("faithful)
```

```
df.head(6)
eruptions waiting
0     3.600      79
1     1.800      54
2     3.333      74
3     2.283      62
4     4.533      85
5     2.883      55
```

```
df.shape
 272   2
```

X: length of the waiting period until the next one (in mins)

Y: the duration of the geyser eruptions (in mins)

```
model = LinearRegression().fit(x, y)
y_pred = model.predict(x)
```

What do we want to model here?

i.e. What is input and output?

Simple linear regression

The fitted model is: eruptions = -1.87 + 0.08 x waiting

What is the eruption time if waiting was 70?

Simple linear regression

The fitted model is: eruptions = -1.87 + 0.08 x waiting

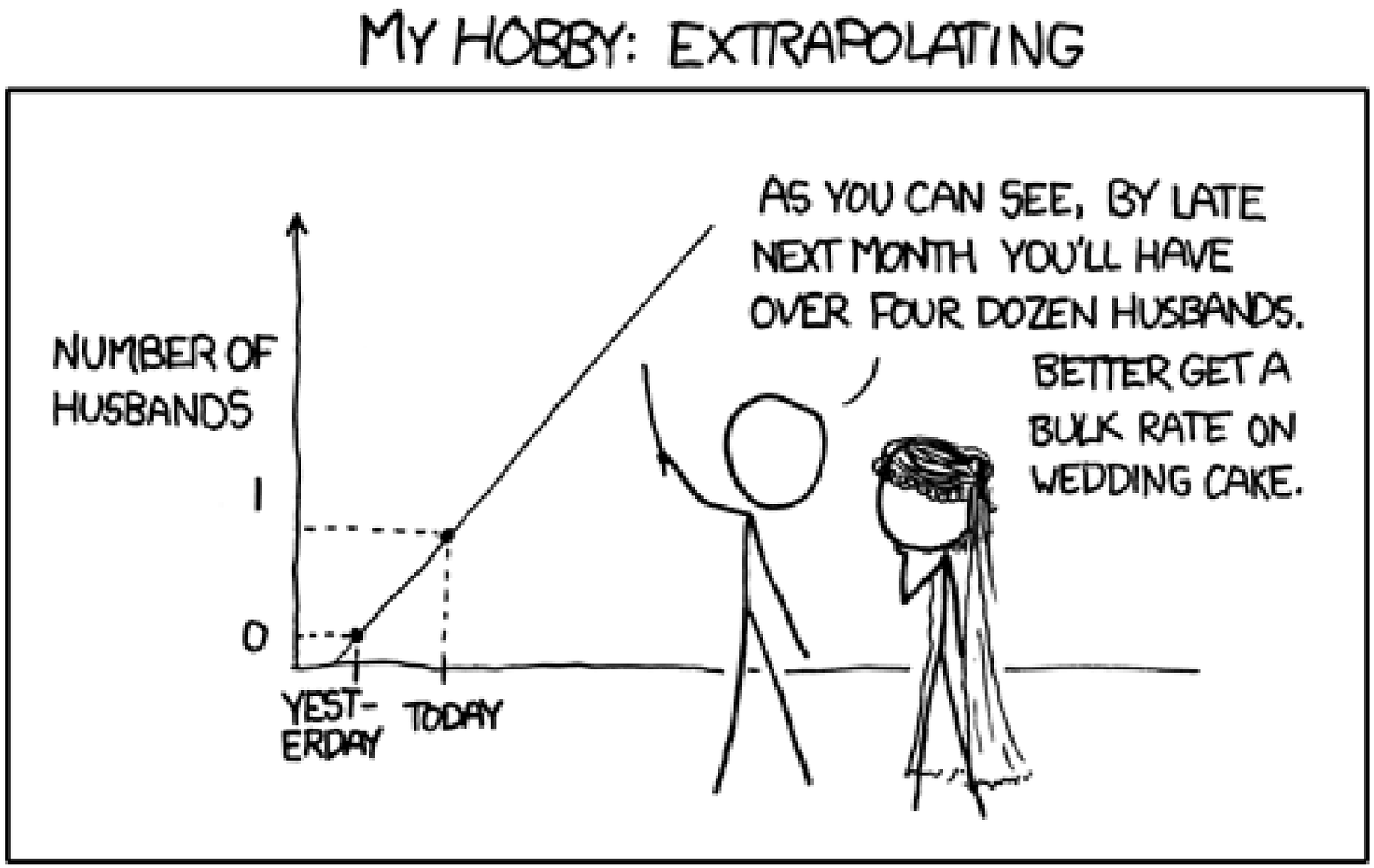
What is the eruption time if waiting was 70?

```
> -1.874016 + 70*0.075628
[1] 3.419944
> coef(model)[[1]] + coef(model)[[2]]*70
```

Prediction

using linear regression

- Doesn’t not apply in every case
  example: not in event based scenarios



Machine Learning

commandment

	F1	F2	F3	Label
Obs 1				
Obs 2				
Obs 3				
Obs 4				
Obs 5				

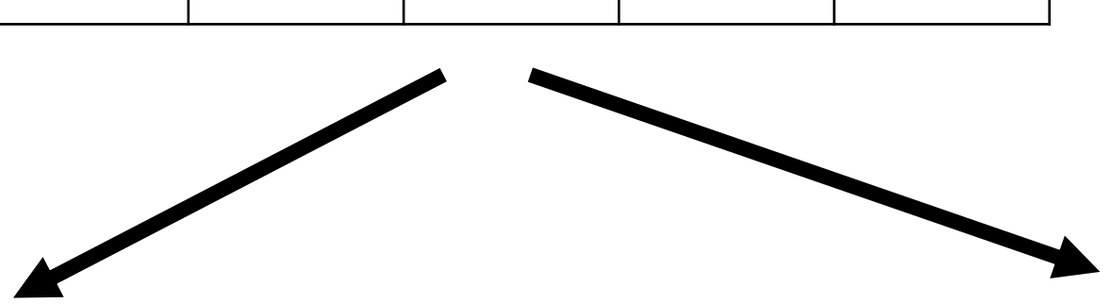
Original Data

Machine Learning

commandment

Original Data

	F1	F2	F3	Label
Obs 1				
Obs 2				
Obs 3				
Obs 4				
Obs 5				



	F1	F2	F3	Label
Obs 1				
Obs 2				
Obs 3				
Obs 4				

Training Data

	F1	F2	F3	Label
Obs 5				

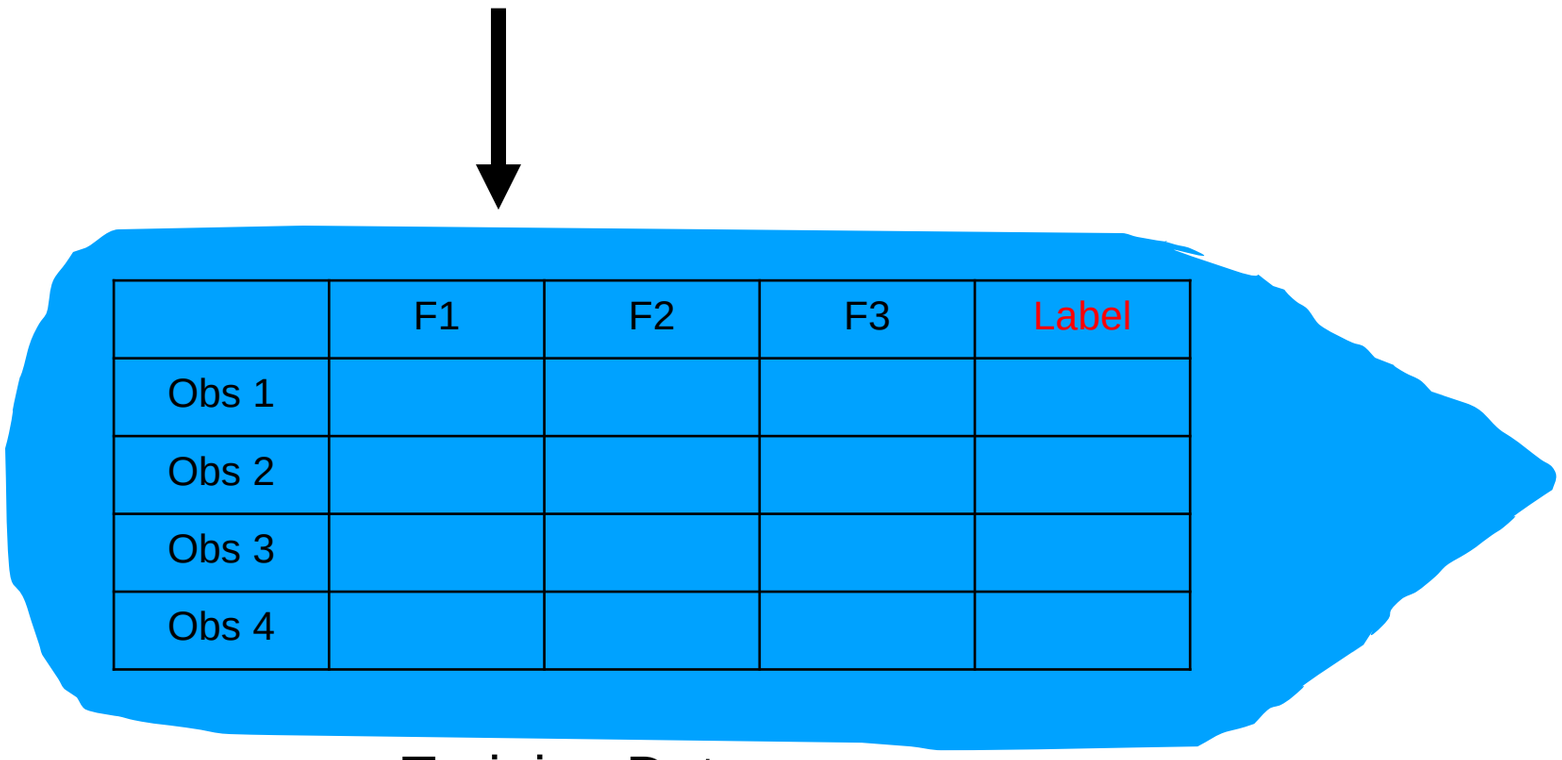
Test Data

Machine Learning

commandment

Original Data

	F1	F2	F3	Label
Obs 1				
Obs 2				
Obs 3				
Obs 4				
Obs 5				



Training Data

Algorithm
(Eg.Linear Regression)

Model

Machine Learning

commandment

Original Data

	F1	F2	F3	Label
Obs 1				
Obs 2				
Obs 3				
Obs 4				
Obs 5				

Training Phase

Testing Phase

Training Data

	F1	F2	F3	Label
Obs 1				
Obs 2				
Obs 3				
Obs 4				

Algorithm (Eg.Linear Regression)

Model

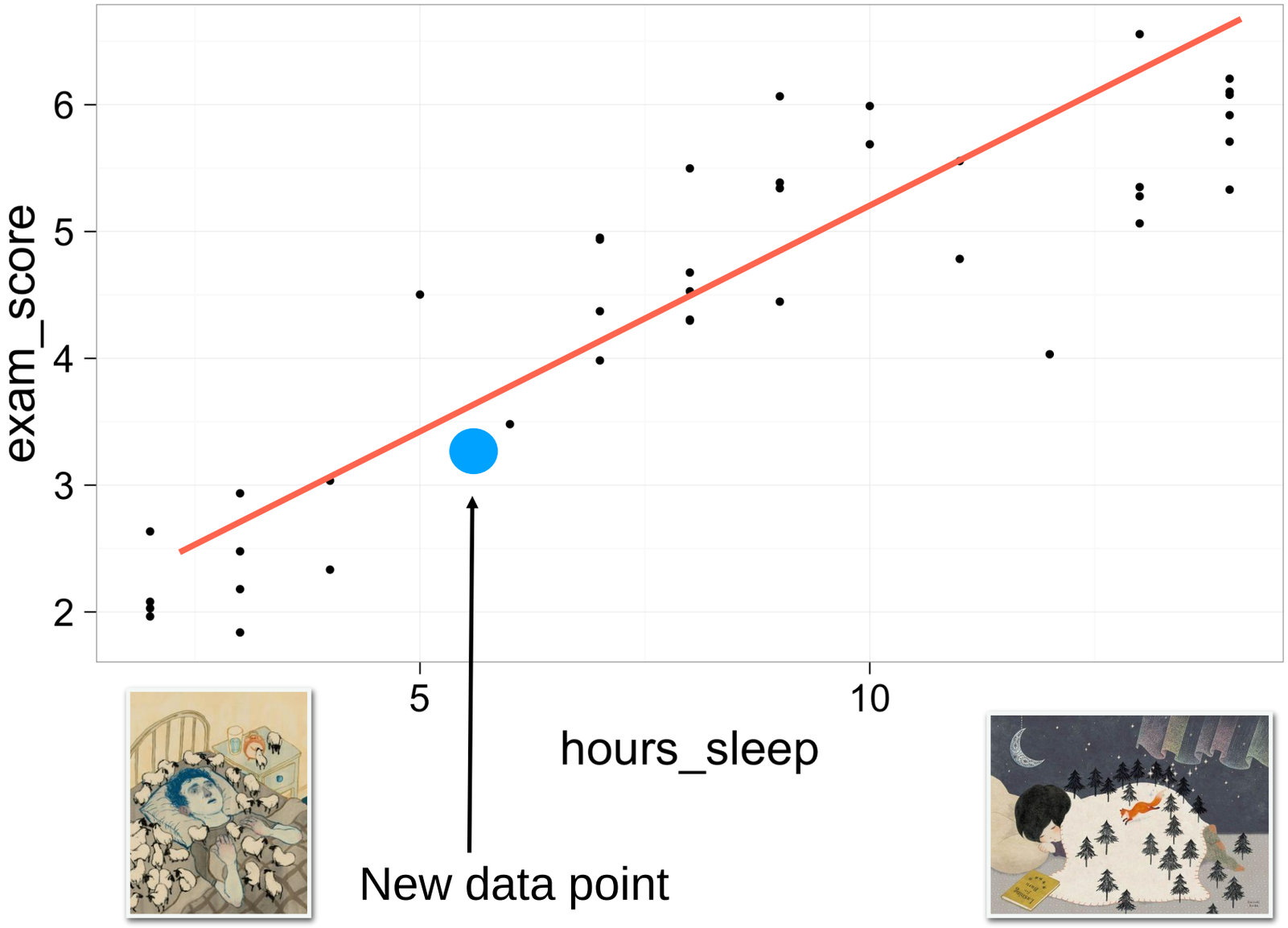
	F1	F2	F3
Obs 5			

Model

	Label
Obs 5	**Pred. Value**

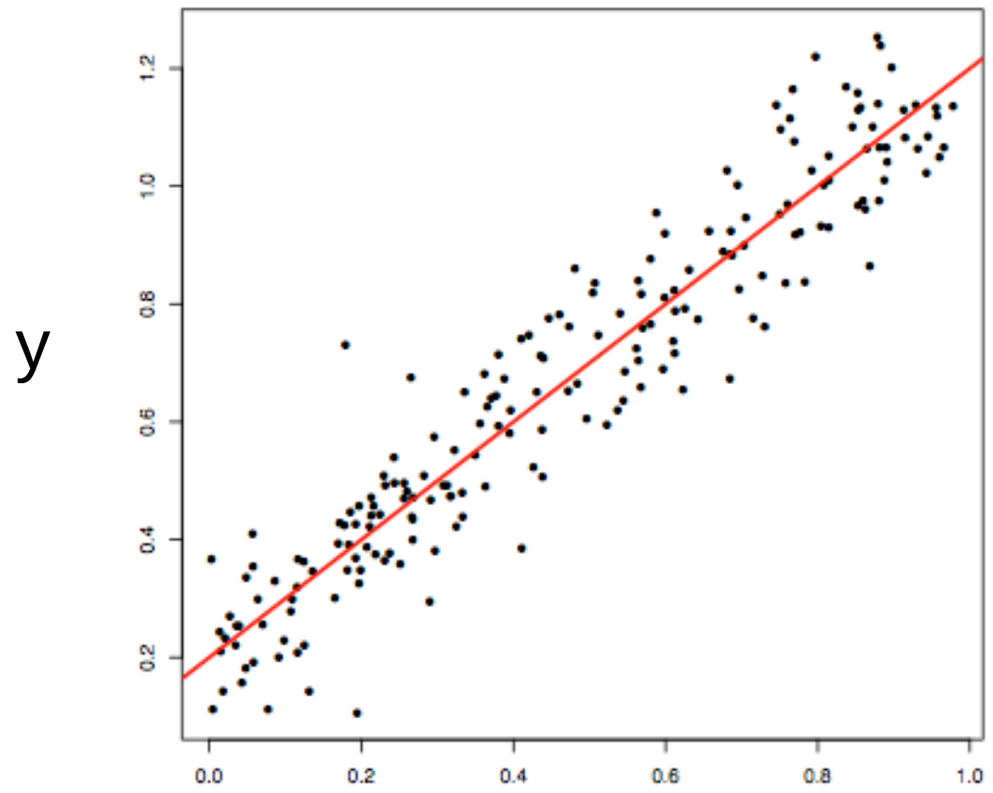
	Label
Obs 5	**Ground Truth**

Simple Linear regression
(How to use it for prediction?)



Given a hours of sleep of a student predict the exam score ?

Simple linear regression



Predict this Label

Input

Task: given a list of observations $D = \{(x_i, y_i)\}_{i=1}^{N}$ find a line

$$\hat{y} = ax + b$$

that approximates the correspondence in the data

2. Multiple regression

all the same, but instead of one feature, x is a k-dimensional vector

$$\mathbf{x}_i = (x_{i1}, x_{i2}, .., x_{ik})$$

the model is the linear combination of all features:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

intercept (bias)

Coefficients

NOTE: error term is assumed to be zero

Multiple regression
Interpreting coefficients

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Each coefficient is interpreted as the estimated change in $\hat{y}$ corresponding to a one unit change in a variable, when all other variables are held constant.

$\hat{y} = 20 + 9x_1 + 10x_2$

$\hat{y}$ = Estimated time

x_1 = Distance

x_2 = # deliveries

9 times is an estimate of the expected increase in estimated time

In delivery time corresponding to a unit increase in distance

when # deliveries are held constant

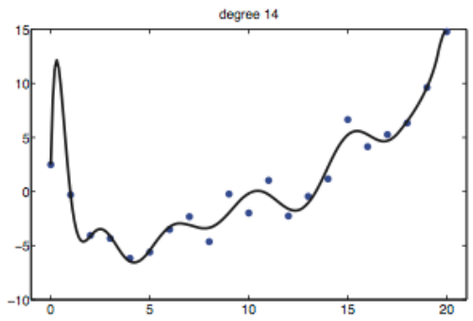
Multiple regression

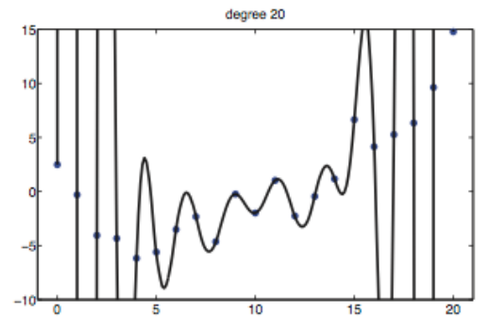
```
> head(train)
    CustomerID TransPerCustomer_1 AmountPerCust_1 AmountPerTr_1 AmountPerCust_2 gender age discount_proposed
534        821                 2           18.02      9.010000            12.92      0  41                 1
285        442                 1            0.62      0.620000             2.78      1  19                 0
572        883                 2            9.23      4.615000            14.10      1  41                 1
652        993                 1            3.29      3.290000             0.24      0  25                 0
230        359                 1            5.03      5.030000             6.06      0  41                 0
471        727                 3           28.03      9.343333            14.73      0  41                 1
    clicks_in_eshop
534               3
285               5
572               3
652               5
230               5
471               3
```

model_1 <- lm(data=train[,-1], AmountPerCust_2 ~ AmountPerCust_1 + TransPerCustomer_1 + AmountPerTr_1 + gender + age + discount_proposed + clicks_in_eshop)

Multiple regression

Linear model requires parameters to be linear, not features!





This is linear model

$$y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2$$

$$x' = \phi(x)$$

This is (polynomial) linear model

$$y = \beta_0 + \beta_1 x_1^7 + \beta_2 x_1^3 + \beta_3 x_1 + \beta_4 x_2^2$$

$$x^z, \sqrt{(x)}, log(x)...$$

This is not linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2^2 x_2$$

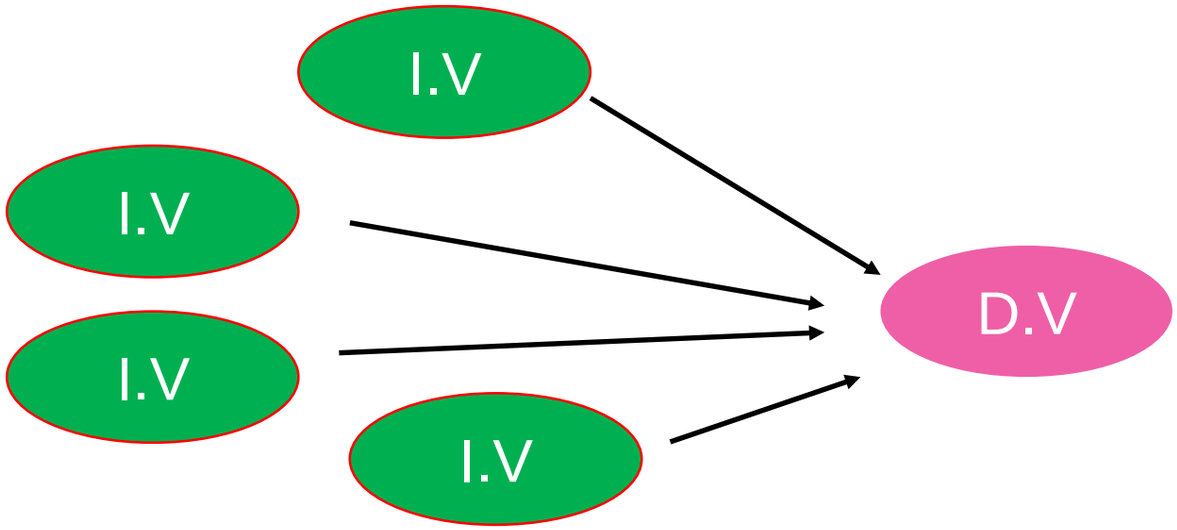
Linear Vs. Multiple Regression

Linear Regression

1 to 1

Multiple Regression

Many to 1



D.V: Dependent Variable or Predictive variable

I.V: Independent Variable or Input variable/features

Multiple regression

Use case: Regional Delivery service

Problem: Estimate the delivery time based on

1) Total distance of the trip
2) # of deliveries that have to be made during the trip

Total Distance	# Deliveries	Delivery time
2	4	5
4	3	6
3	4	4

Source: Brandon Foltz learning material.

Multiple regression

Use case: Regional Delivery service

Problem: Estimate the delivery time based on

1) Total distance of the trip
2) # of deliveries that have to be made during the trip

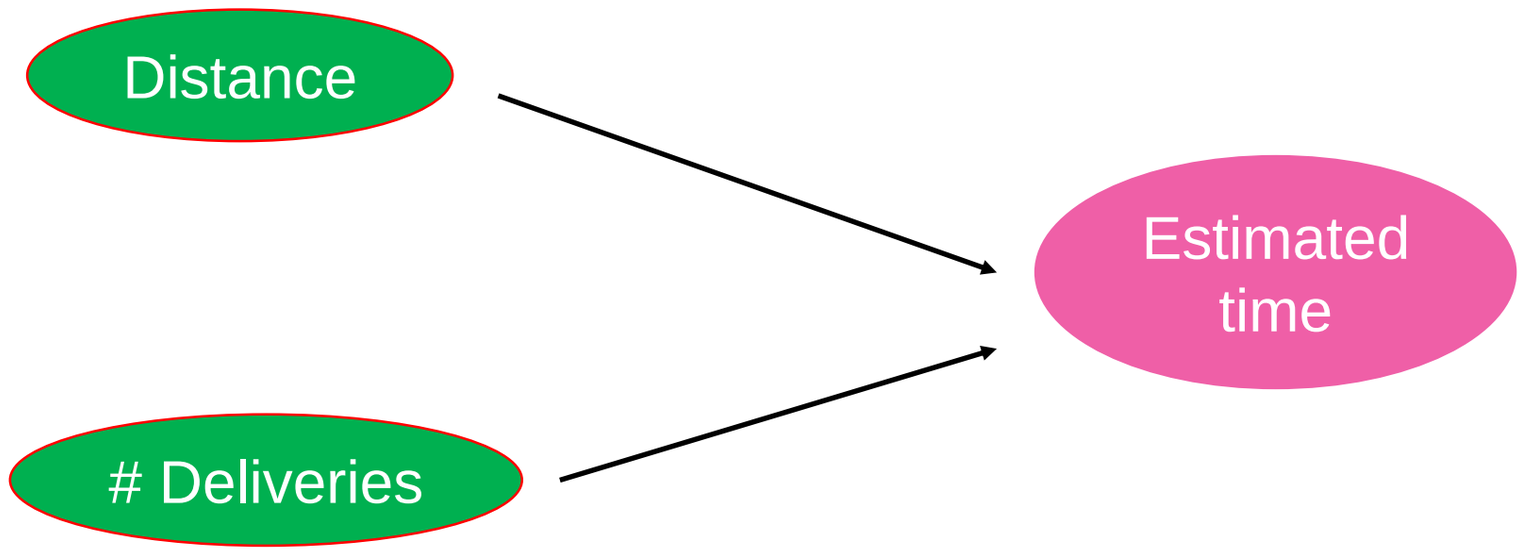
Input or Independent variable (I.V)

Predict or Dependent variable (D.V)

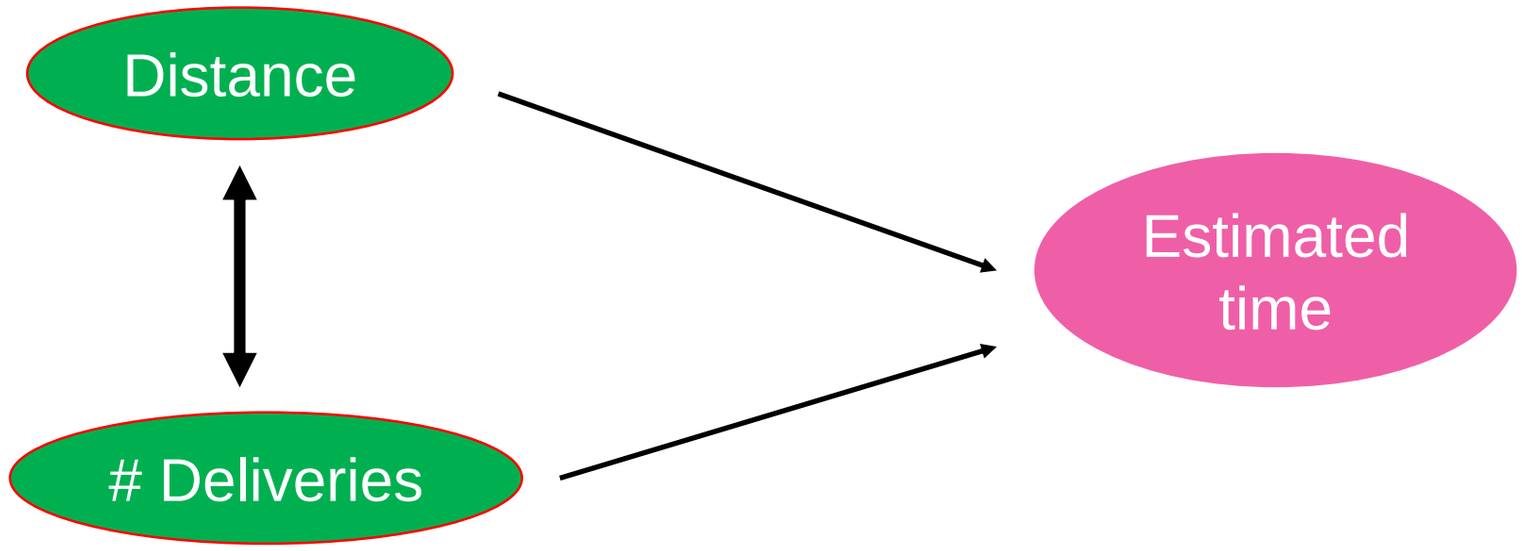
Total Distance	# Deliveries	Delivery time
2	4	5
4	3	6
3	4	4

Source: Brandon Foltz learning material.

Some Considerations



Some Considerations



Some Considerations

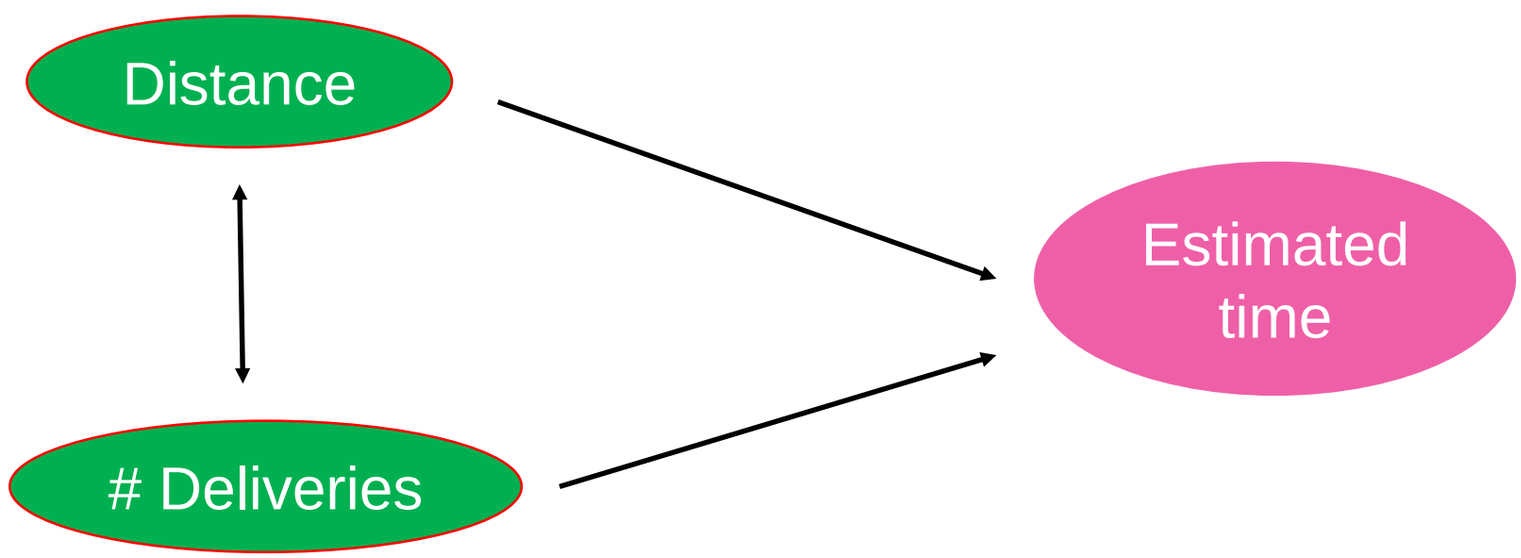
collinearity problem

Independent variables could not only be related (in some proportion) with dependent variable but they could be related with each other (called as multicollinearity)

Ideally, all the independent variables to be correlated with the dependent variable but not with each other.

Some ways to avoid multicollinearity :

- Correlations
- Scatter plots
- Simple regressions.



Some Considerations

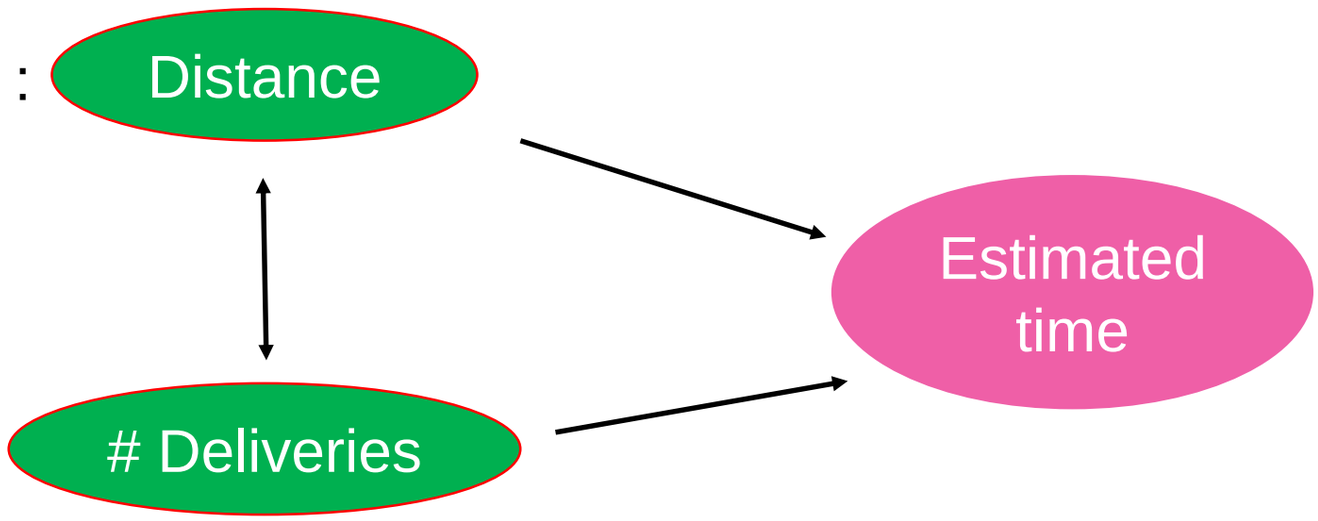
Adding more variables does not mean will make things better: It can lead to problem of overfitting

Independent variables could not only be related (in some proportion) with dependent variable but they could be related with each other (called as multicollinearity)

Ideally, all the independent variables to be correlated with the dependent variable but not with each other.

Some ways to avoid multicollinearity :

- Correlations
- Scatter plots
- Simple regressions.

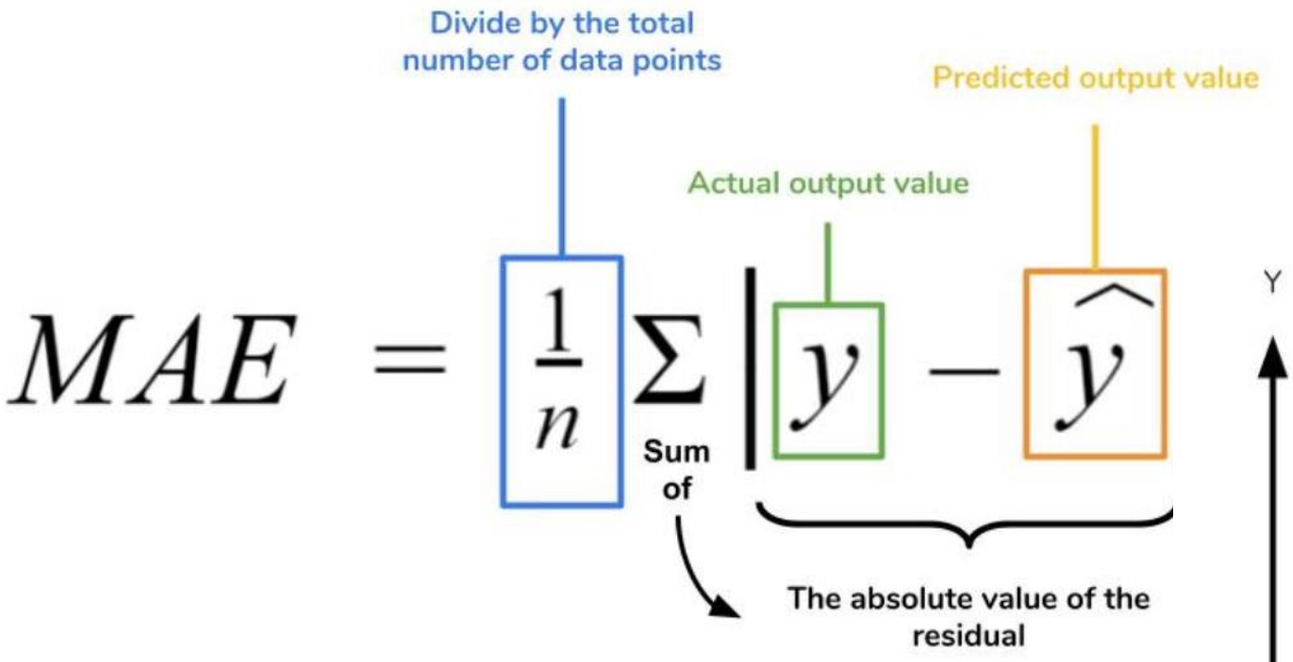


Quality Assessment

- MAE: Mean Absolute Error
- MSE: Mean Square Error
- RMSE: Root Mean Square Error
- MAPE: Mean Absolute Percentage Error
- SMAPE: Symmetric Mean Absolute Percentage Error
- R^2
- Adjusted R^2

https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70

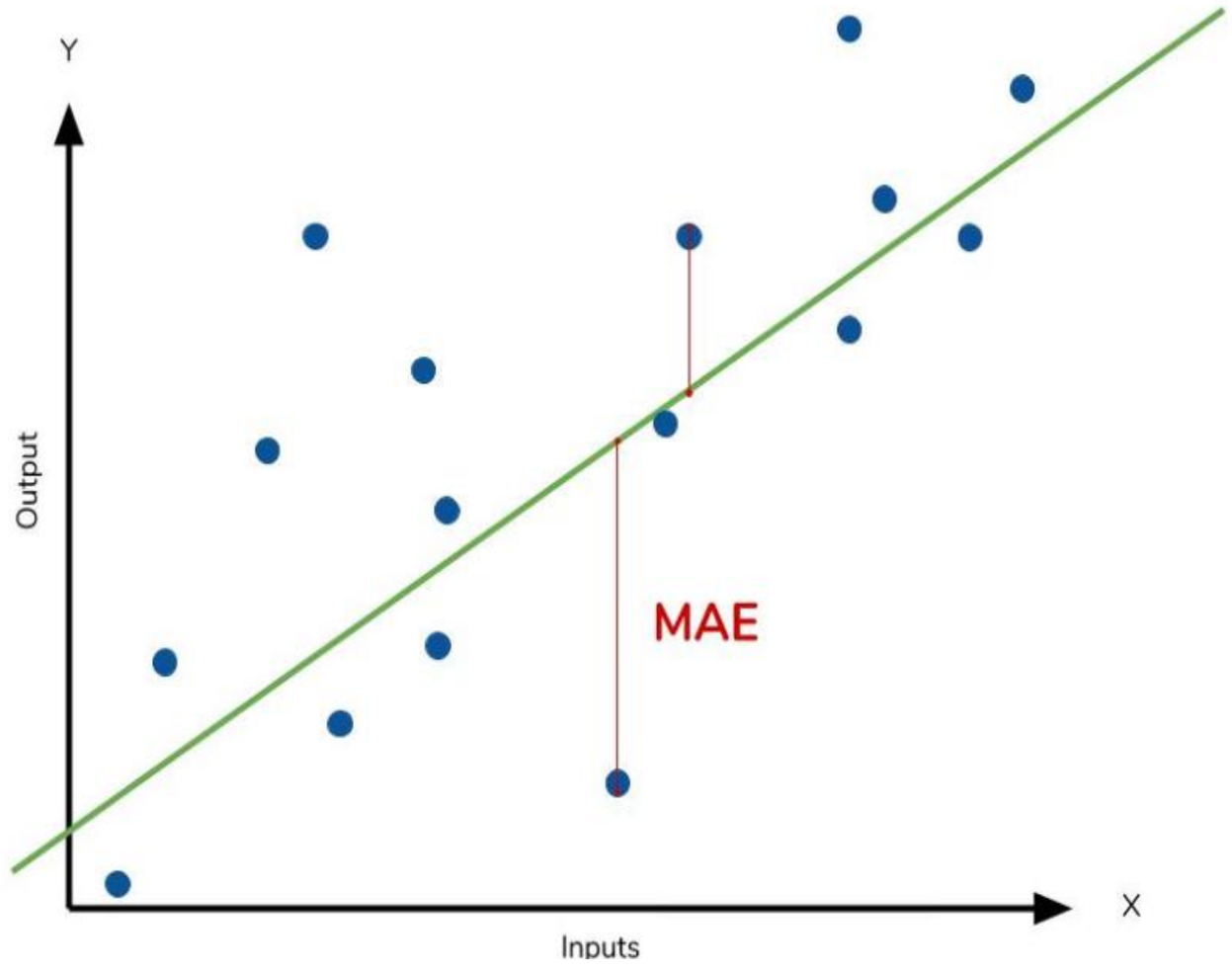
MAE



absolute difference between the data and the model's predictions.

small MAE suggests the model is great at prediction, while a large MAE suggests that your model may have trouble in certain areas

Does not does not indiacate underperformance or **overperformance** of the model (whether or not the model under or overshoots actual data).



MSE/RMSE

What about outliers ?

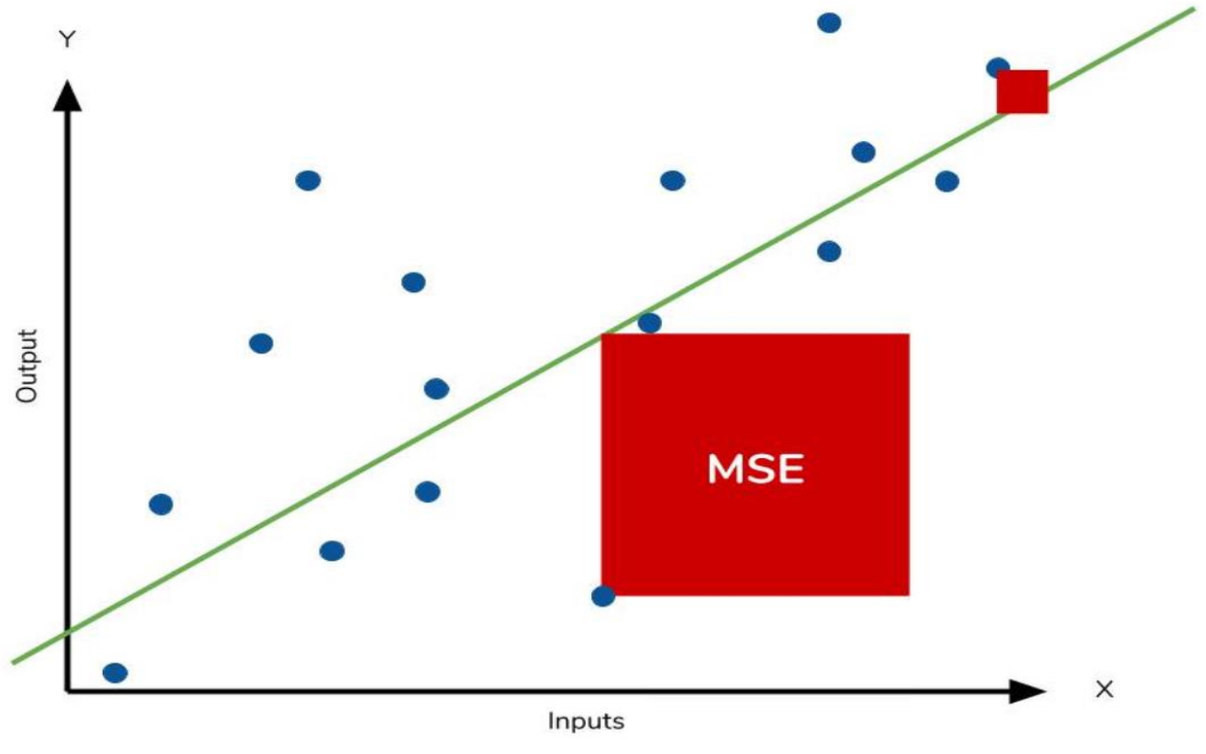
While each residual in MAE contributes **proportionally** to the total error, the error grows **quadratically** in MSE. What it means:

- outliers in our data will contribute to much higher total error in the MSE than they would in MAE.
- our model will be penalized more for making predictions that differ greatly from the corresponding actual value.

- Reference: https://www.dataquest.io/blog/understanding-regression-error-metrics/

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted



Source: https://www.dataquest.io/blog/understanding-regression-error-metrics/

MAE and (R)MSE

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Loss/cost function

MAE is more robust to outliers since it does not make use of square.

MSE is more useful if we are concerned about large errors.

With errors, MAE is steady

With increase in errors, RMSE increases as the variance associated with the frequency distribution of error magnitudes also increases.

Reference: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

RMSE Vs. MAE

CASE 1: Evenly distributed errors

ID	Error	|Error|	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9	2	2	4
10	2	2	4

MAE	RMSE
2.000	2.000

CASE 2: Small variance in errors

ID	Error	|Error|	Error^2
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	3	3	9
7	3	3	9
8	3	3	9
9	3	3	9
10	3	3	9

MAE	RMSE
2.000	2.236

CASE 3: Large error outlier

ID	Error	|Error|	Error^2
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	20	20	400

MAE	RMSE
2.000	6.325

MAE and RMSE for cases of increasing error variance

RMSE should be more useful when large errors are particularly undesirable.

https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

RMSE Vs. MAE

CASE 4: Errors = 0 or 5

ID	Error	|Error|	Error^2
1	5	5	25
2	0	0	0
3	5	5	25
4	0	0	0
5	5	5	25
6	0	0	0
7	5	5	25
8	0	0	0
9	5	5	25
10	0	0	0

var	MAE	RMSE
6.944	2.500	3.536

CASE 5: Errors = 3 or 4

ID	Error	|Error|	Error^2
1	3	3	9
2	4	4	16
3	3	3	9
4	4	4	16
5	3	3	9
6	4	4	16
7	3	3	9
8	4	4	16
9	3	3	9
10	4	4	16

var	MAE	RMSE
0.278	3.500	3.536

RMSE does not necessarily increase with the variance of the errors. RMSE increases with the variance of the frequency distribution of error magnitudes

https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

MAPE

Mean Absolute Percentage error (MAPE)

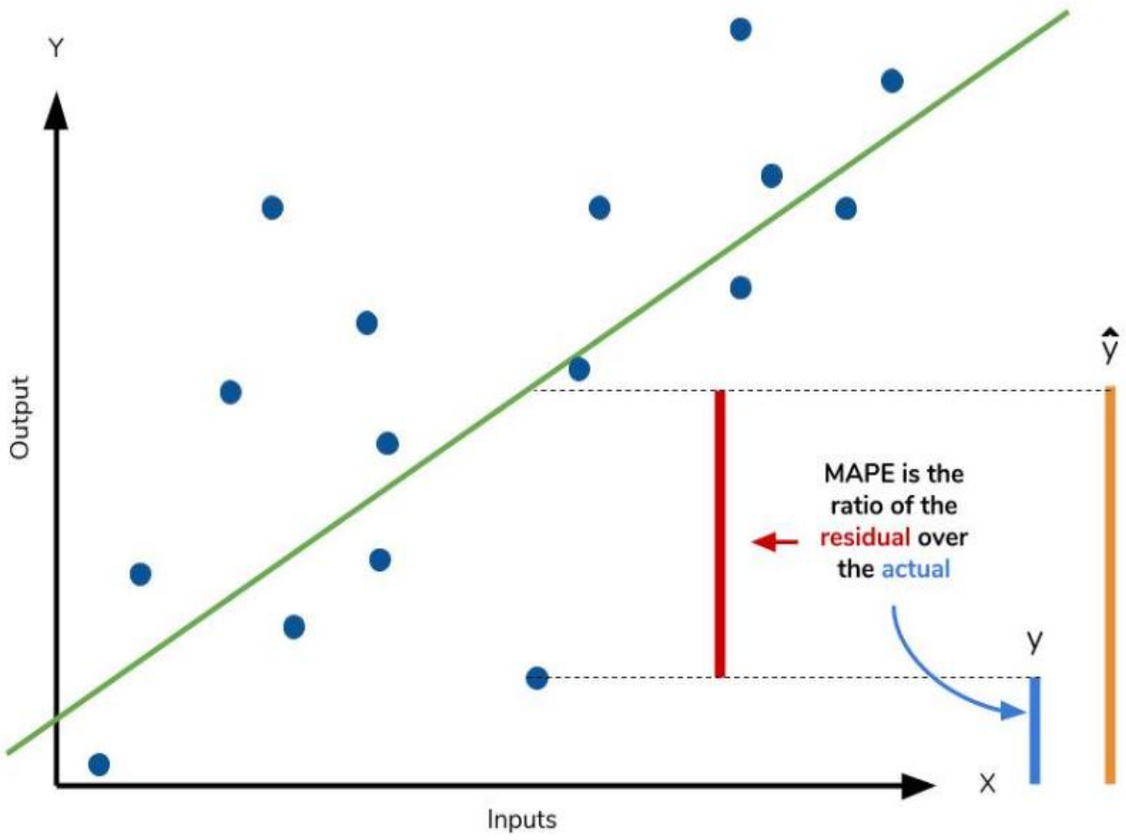
$$MAPE = \frac{100\%}{n} \sum \left| \frac{y - \hat{y}}{y} \right|$$

Multiplying by 100% converts to percentage

The residual

Each residual is scaled against the actual value

MAE is the average magnitude of error produced by the model, & the MAPE is how far the model's predictions are off from their corresponding outputs on average.



ŷ is smaller than the actual value

n = 1 ŷ = 10 y = 20

MAPE = 50%

ŷ is greater than the actual value

n = 1 ŷ = 20 y = 10

MAPE = 100%

MAPE is biased towards predictions that are systematically less than the actual values themselves

SMAPE

Symmetric mean absolute percentage error

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

A_t is the actual value
F_t is the forecast value.
n : Number of observations

1) Range between 0 and 100 (so the error is in percentage) and so models are comparable (this is not possible using MAE and RMSE).

2) Also it penalizes the error if only some of the errors are outliers but they are very high

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{|A_t| + |F_t|}$$

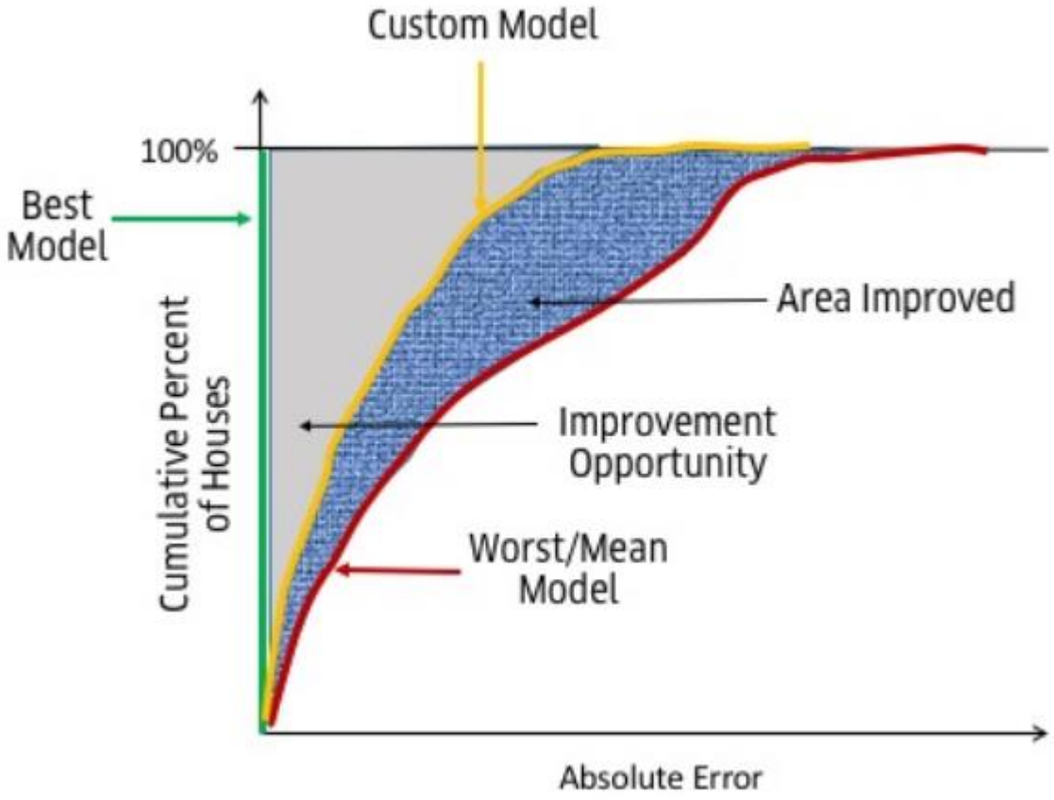
Not symmetric since over- and under-forecasts are not treated equally

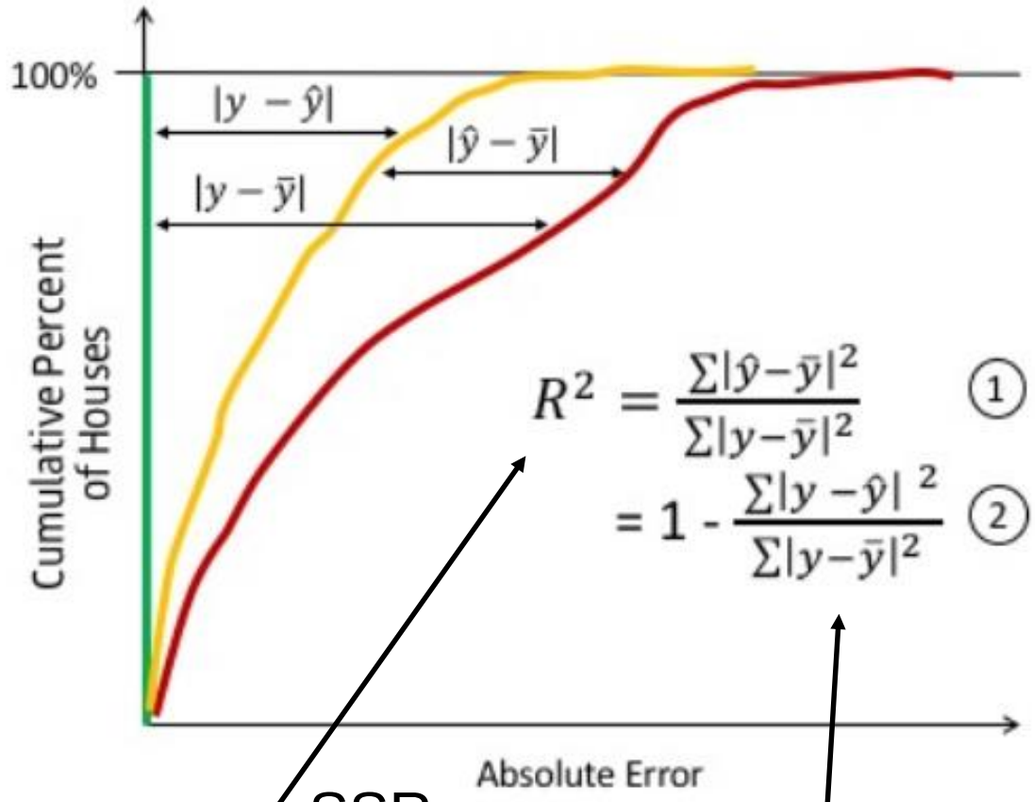
- Over-forecasting: A_t = 100 and F_t = 110 give SMAPE = 4.76%
- Under-forecasting: A_t = 100 and F_t = 90 give SMAPE = 5.26%.

Further Reading: https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error

R^2: What about improvement?

Coefficient of Determination





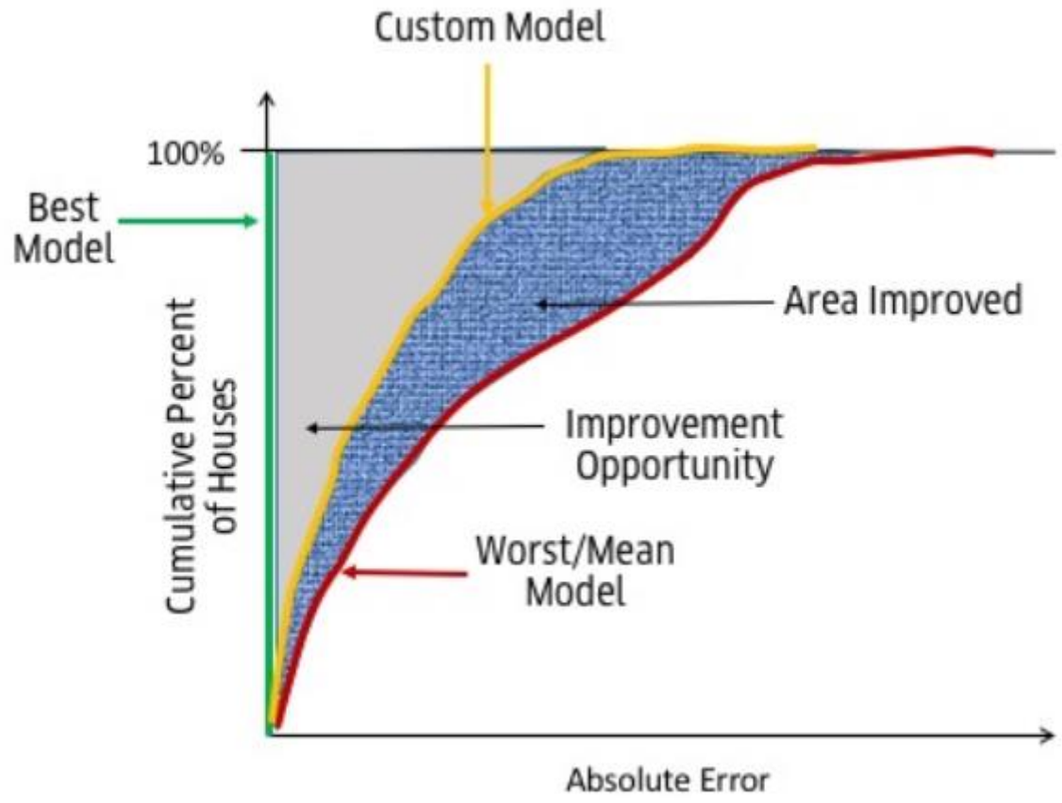
- SSE (Sum of Squares error) = $\sum(y - \hat{y})^2$
- SST (Sum of Squares total) = $\sum(y - \bar{y})^2$
- SSR (Sum of Squares Regression = $\sum(\hat{y} - \bar{y})^2$

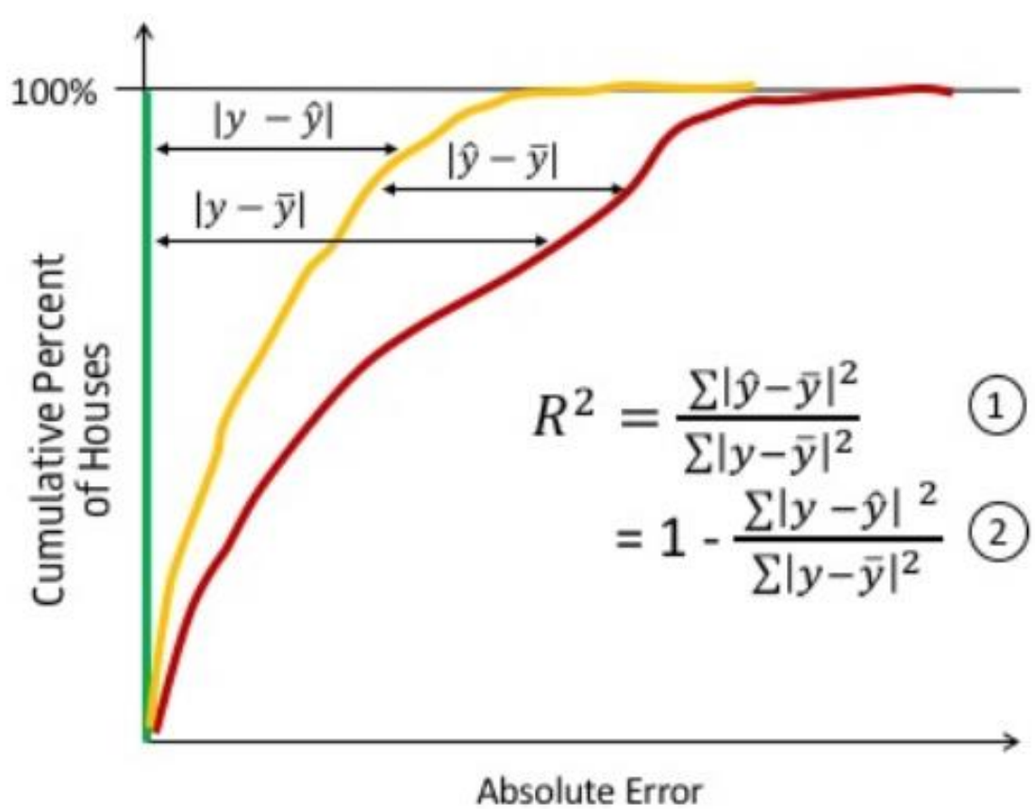
- $R^2 = \frac{SSR}{SST}$
- $R^2 = 1 - \frac{SSE}{SST}$

- R^2 ranges from 0 to 1 (as mentioned in wikipedia) or from -1 to 1 (in libraries).

Reference: https://ragrawal.wordpress.com/2017/05/06/intuition-behind-r2-and-other-regression-evaluation-metrics/

R2: What about improvement?





$$R^2 = \frac{\sum|\hat{y}-\bar{y}|^2}{\sum|y-\bar{y}|^2} \quad (1)$$

$$= 1 - \frac{\sum|y-\hat{y}|^2}{\sum|y-\bar{y}|^2} \quad (2)$$

- R2 ranges from 0 to 1 (as mentioned in wikipedia) or from -1 to 1 (in libraries).

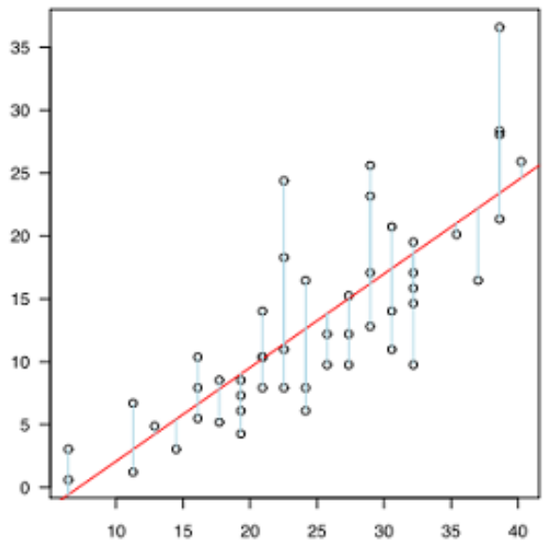
$$R2 = \frac{\text{Area between our mean – mean model}}{\text{Area between best and mean model}}$$

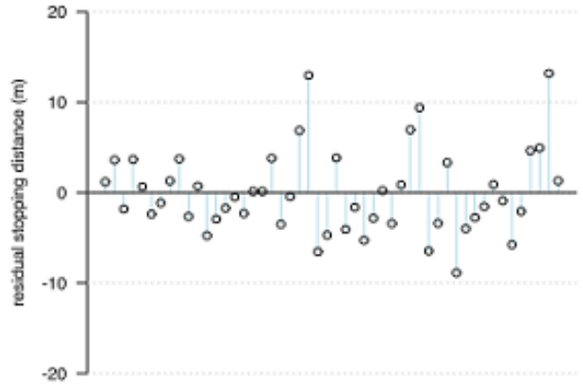
- Equation1: made an assumption that our model will be always better than mean model and hence will be in between mean model and the best model.
- Equation 2: in practices its possible that our model is worst than mean model and it falls on right side of the mean model.

Reference: https://ragrawal.wordpress.com/2017/05/06/intuition-behind-r2-and-other-regression-evaluation-metrics/

R2 and more

- R-squared *cannot* determine whether the coefficient estimates and predictions are biased, which is why you must assess the residual plots.





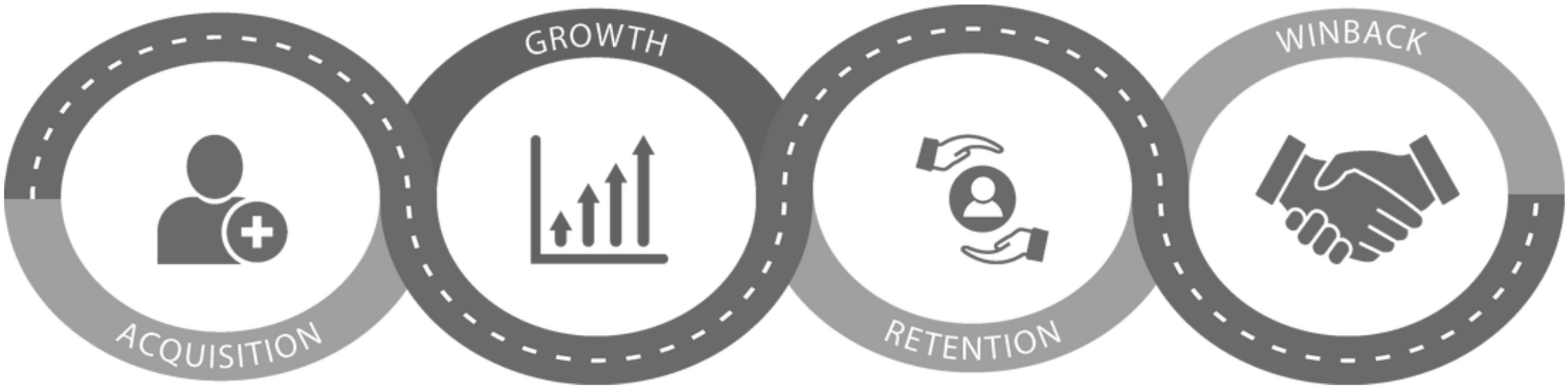
- “Adjusted R-square” penalizes you for adding variables which do not improve your existing model.

- Typically, the more non-significant variables you add into the model, the gap in R-squared and Adjusted R-squared increases.

https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

Customer lifecycle

Where does CLV falls among 4 ?



Summary

Customer LifeTime Value

Regression Problems – Continuous values.

Various Kinds of Regression Techniques:

- Linear Regression
- Multi Linear Regression
- Concept: Gradient Descent

Metrics:

- Mean Absolute Error
- (Root) Mean Squared Error
- MAPE and SMAPE
- R2 and Adjusted R2