Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

MEASURING ERRORS

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

QUIZ

How to **measure regression errors?**

Answer: Measure the distance between the data points and the fitted regression line.

How to **measure classification errors?**

Answer: Count the number of objects that are misclassified.

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

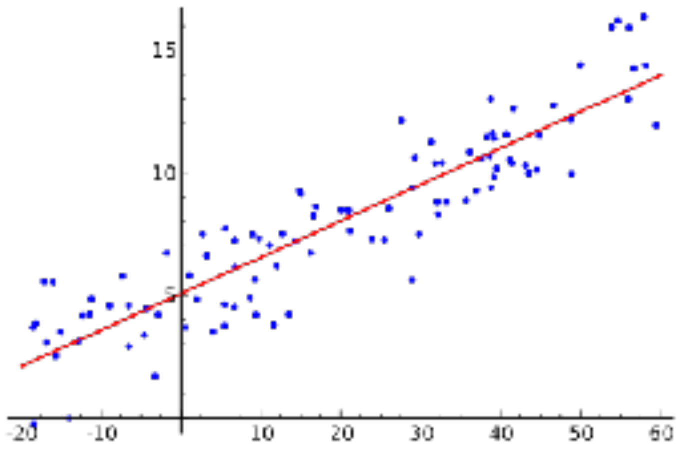
REGRESSION, A RECAP

EMMA BEAUXIS-AUSSALET

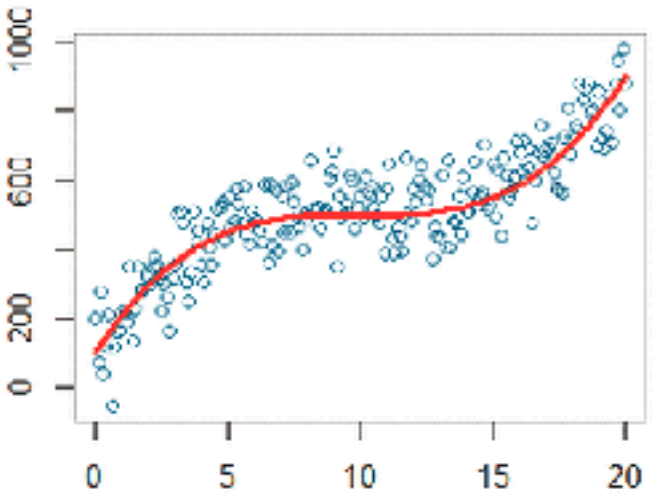
e.m.a.l.beauxis@hva.nl

REGRESSION

Regression is basically **“fitting a line”**, e.g., with **linear** functions.



Simple linear regression



Polynomial regression

Univariate

$$\mathbf{y} = a\ \mathbf{x} + b$$

$$\mathbf{y} = a_1\ \mathbf{x} + a_2\ \mathbf{x^2} + a_3\ \mathbf{x^3} + \ldots + a_i\ \mathbf{x^i} + b$$

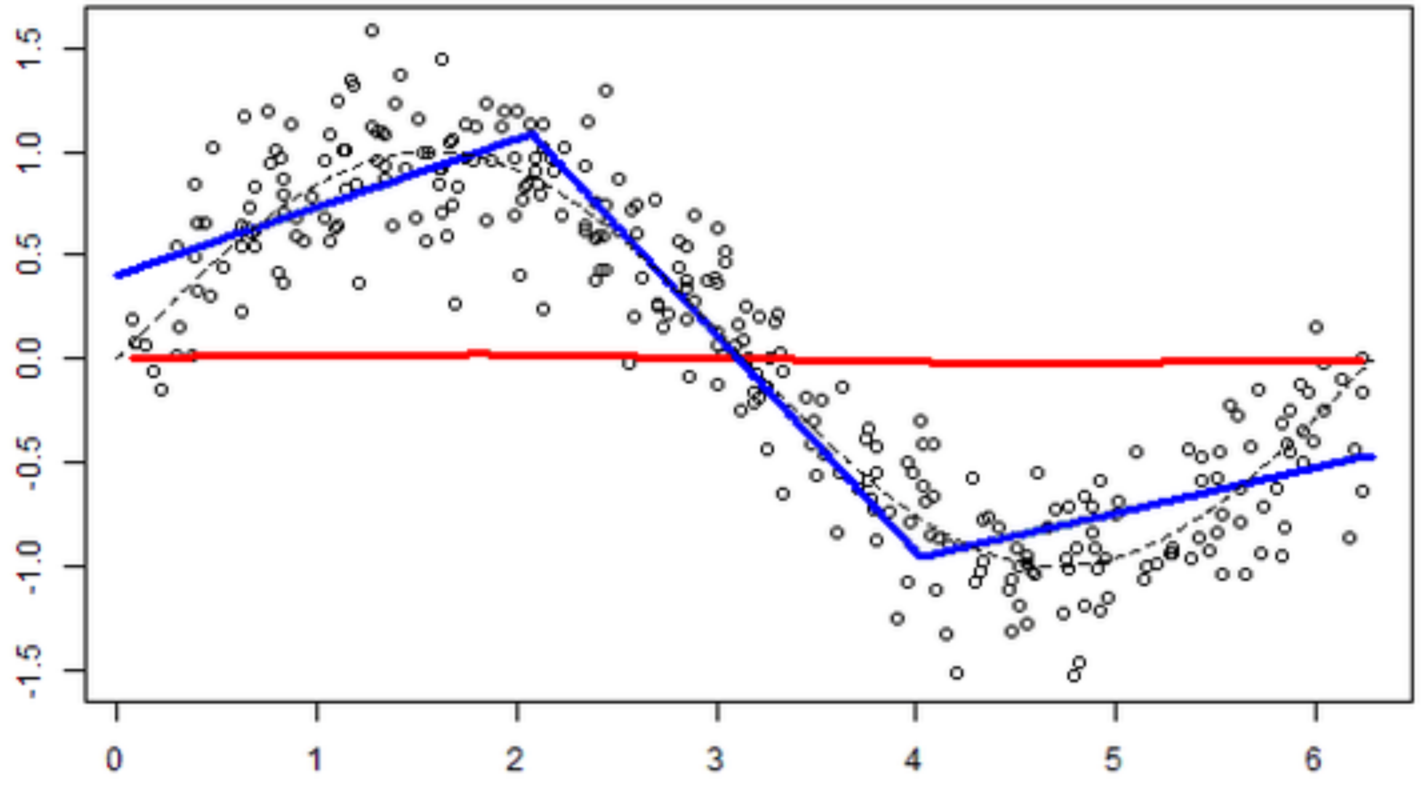
Multivariate

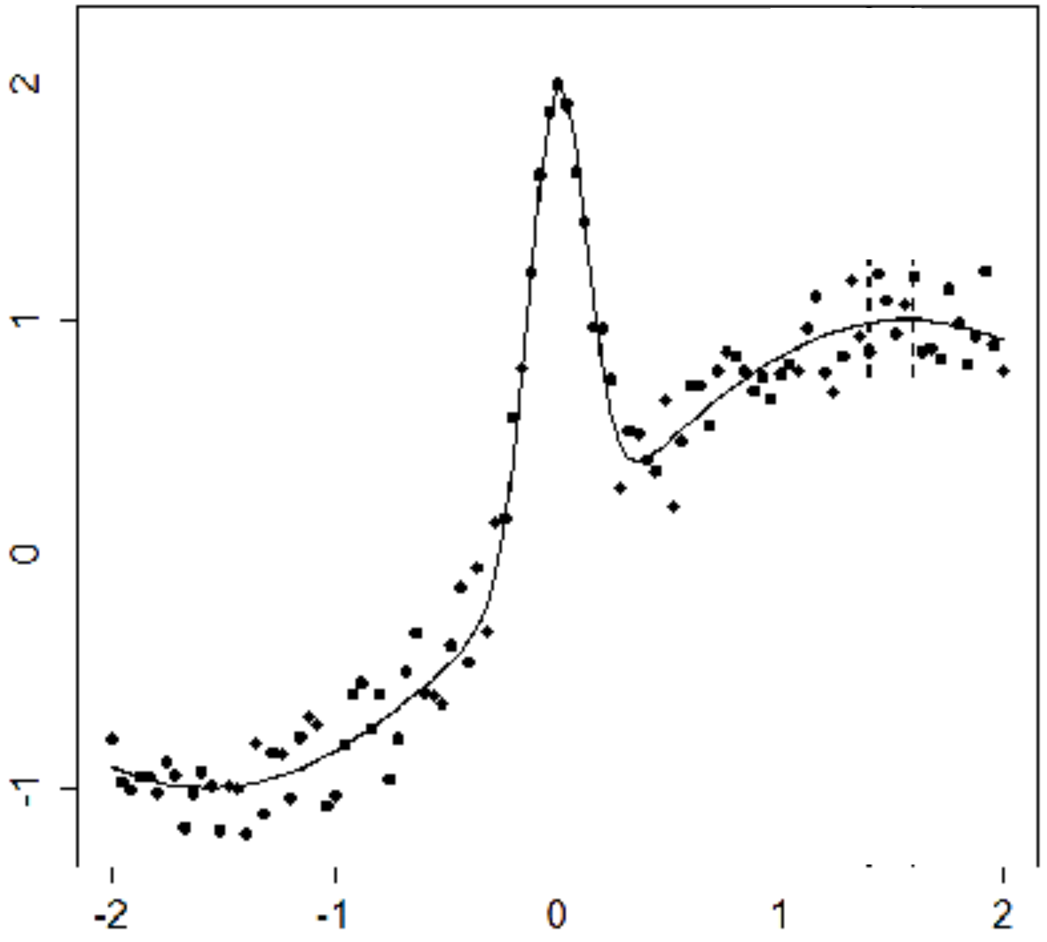
$$\mathbf{y} = a_1\ \mathbf{x_1} + a_2\ \mathbf{x_2} + \ldots + a_i\ \mathbf{x_i} + b$$

$$\mathbf{y} = a_{10}\ \mathbf{x_1} + a_{01}\ \mathbf{x_2} + a_{11}\ \mathbf{x_1 x_2} + a_{20}\ \mathbf{x_1^2} + a_{02}\ \mathbf{x_2^2} + a_{22}\ \mathbf{x_1^2 x_2^2} + \ldots + b$$

REGRESSION

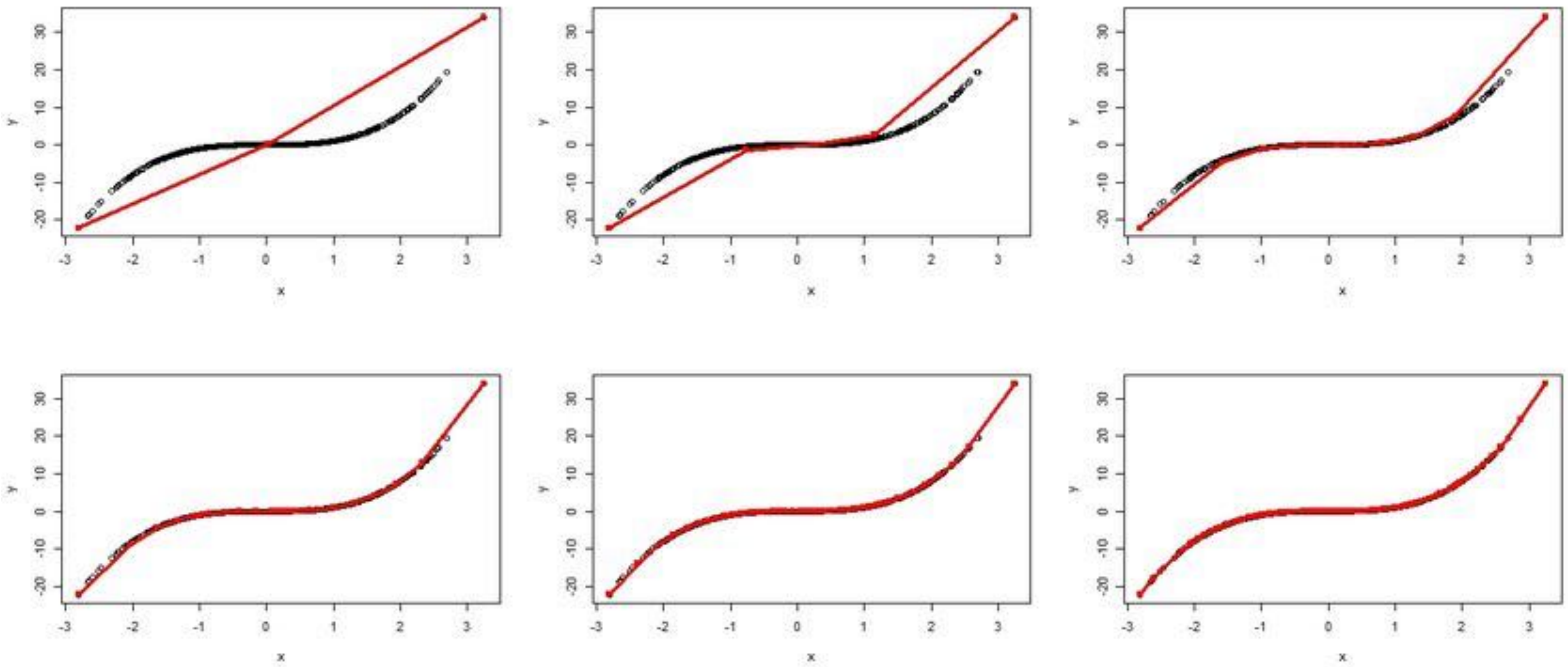
Some cases require **non-linear**, sometimes **non-parametric** methods.





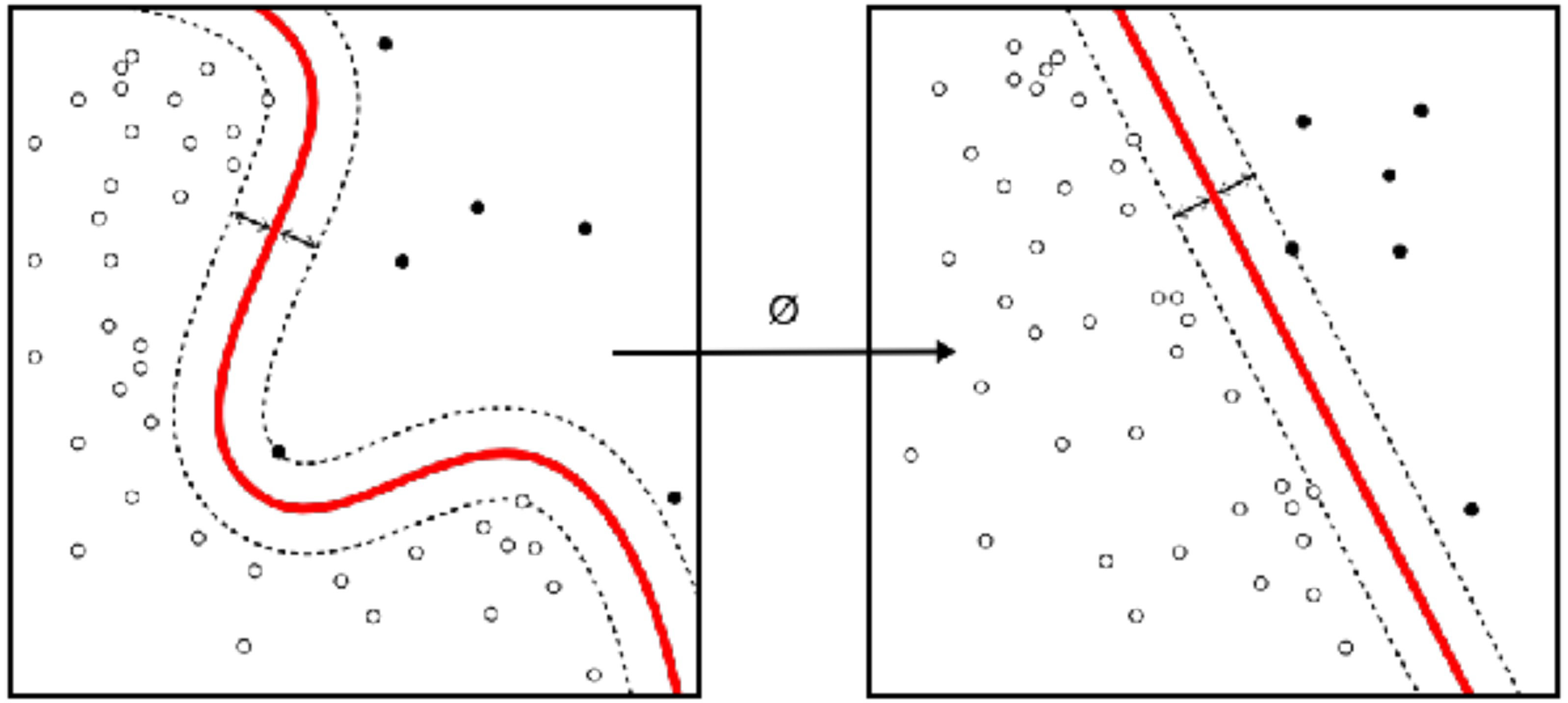
REGRESSION

Some cases require **non-linear**, sometimes **non-parametric** methods.



NON LINEARITY

Non-linear problems can be transformed into linear ones (sometimes). For instance, by transforming the data, by mapping data points on different coordinates.



(e.g., SVM uses the kernel trick)

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

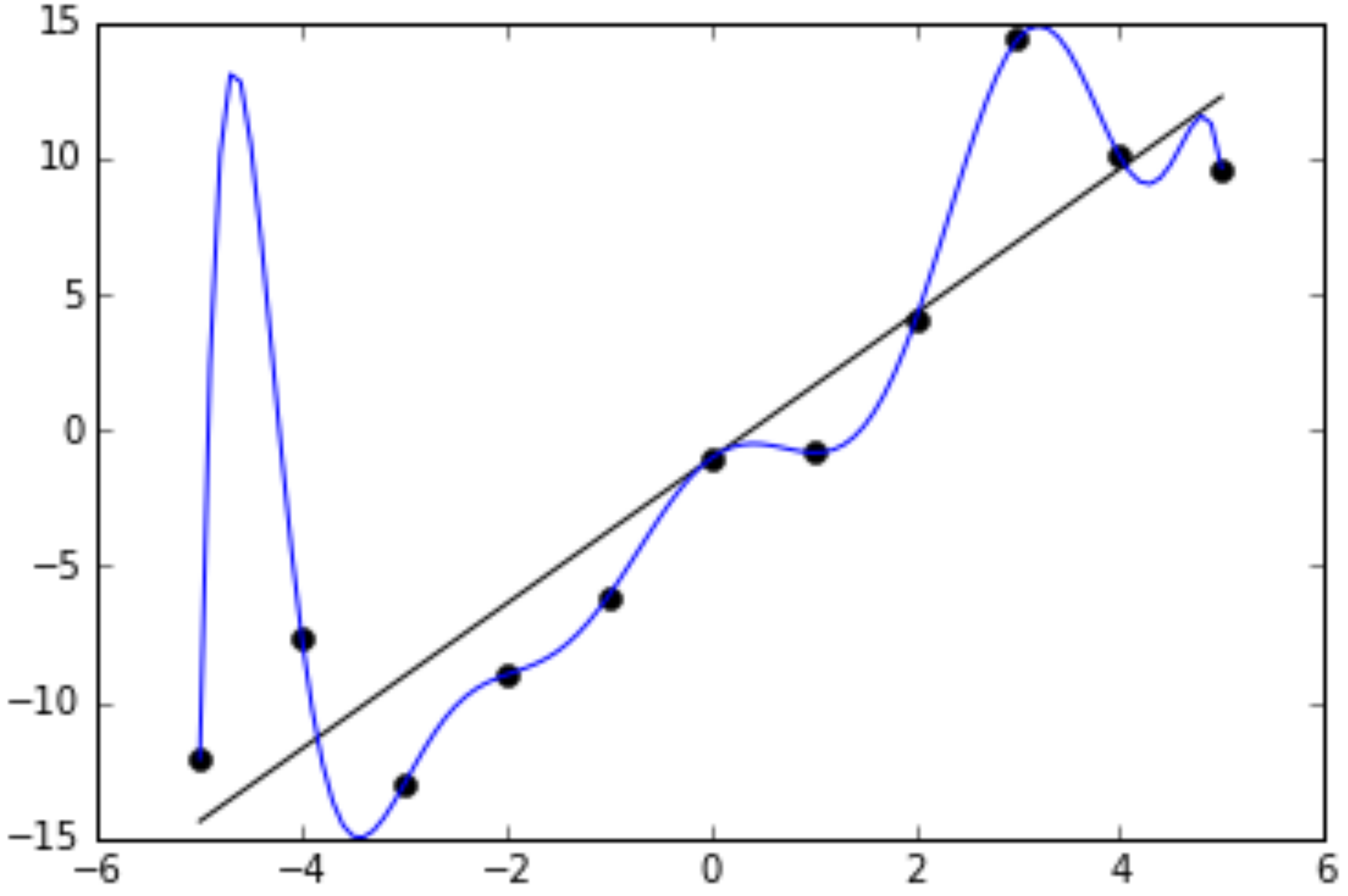
REGRESSION ERRORS

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

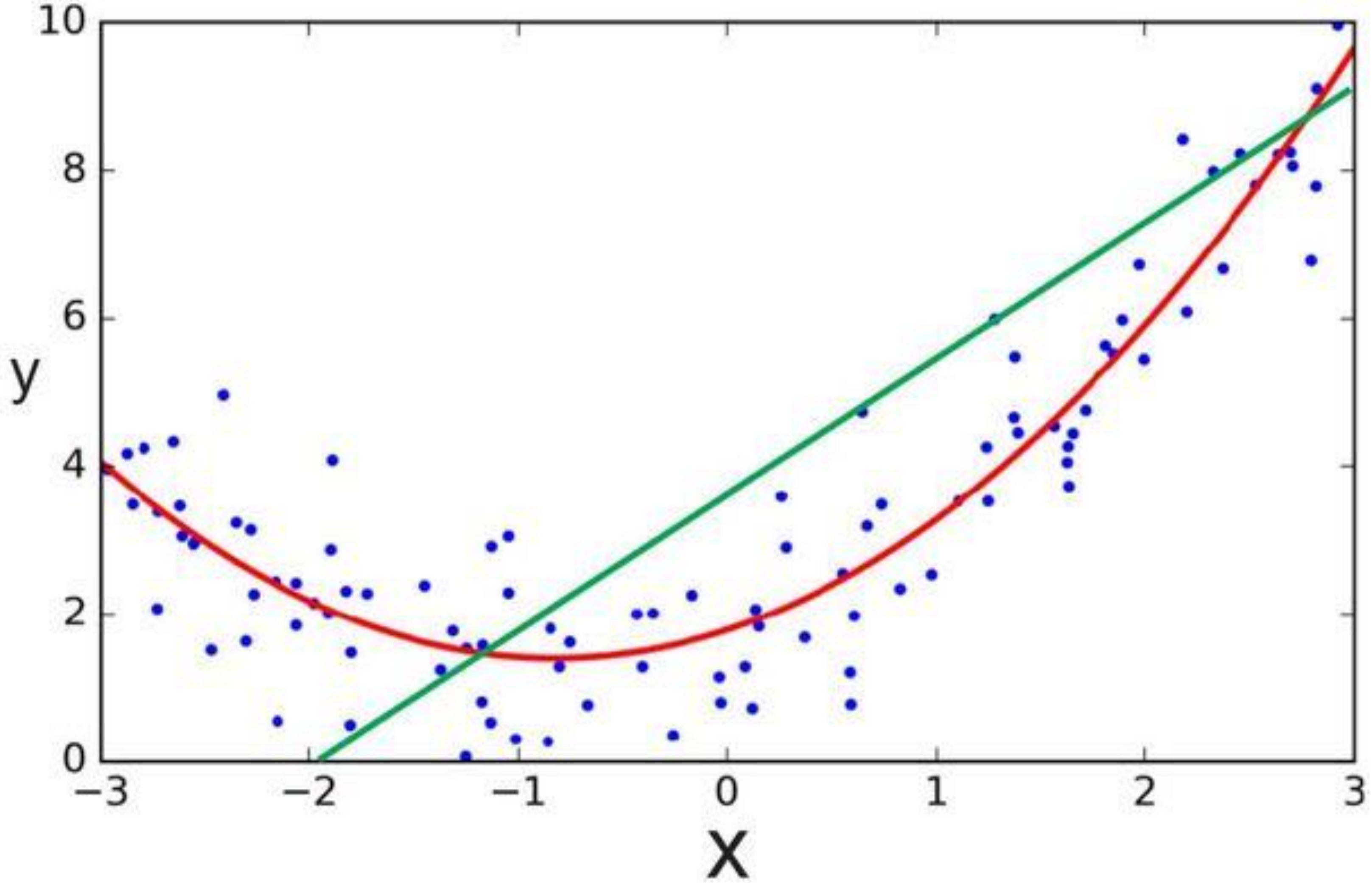
OVER-FITTING

Perfect results are **suspicious**.
Errors may be minimal for one dataset, but not for other datasets.



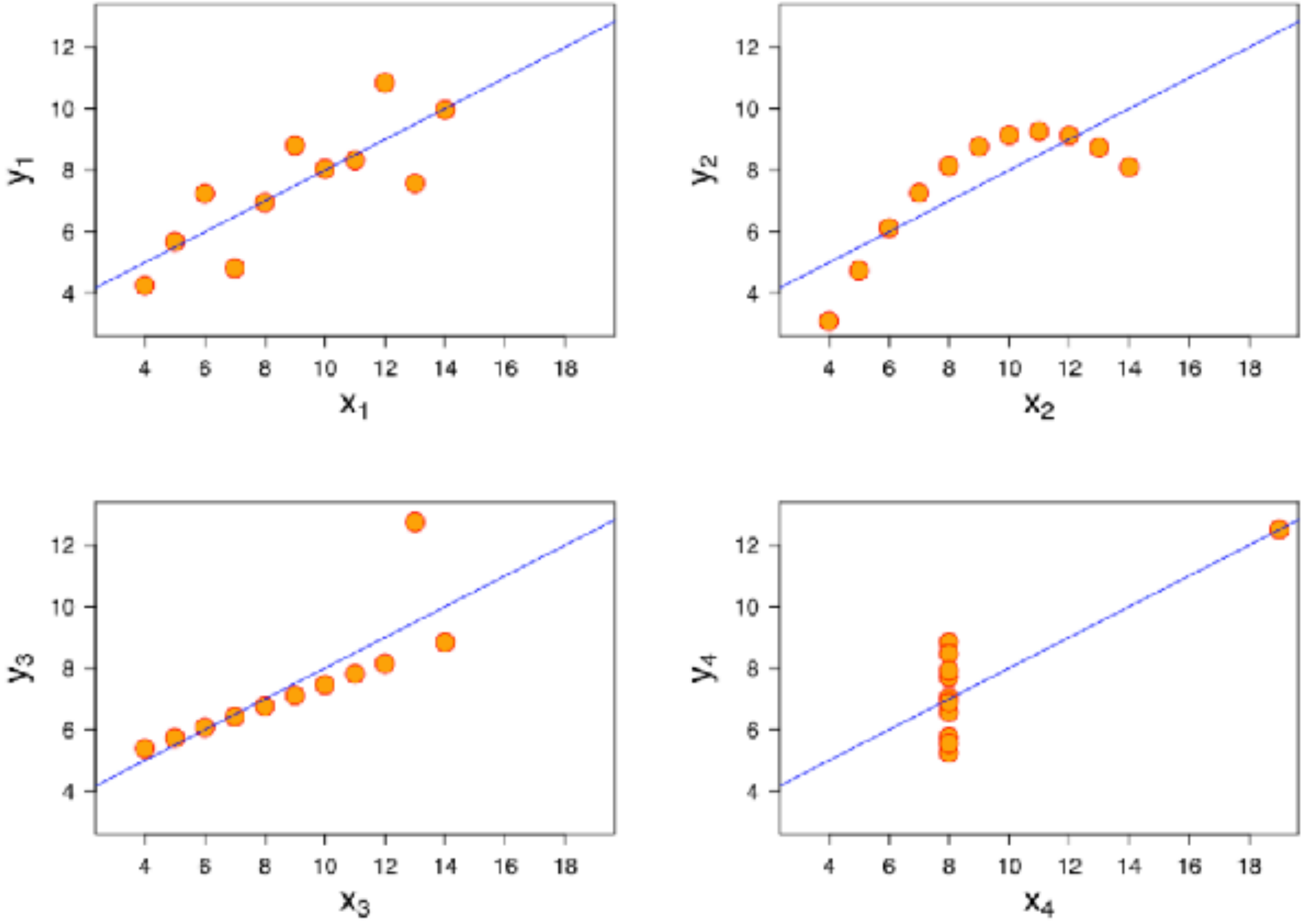
UNDER-FITTING

Low errors may **conceal underlying issues** and inaccurate assumptions.



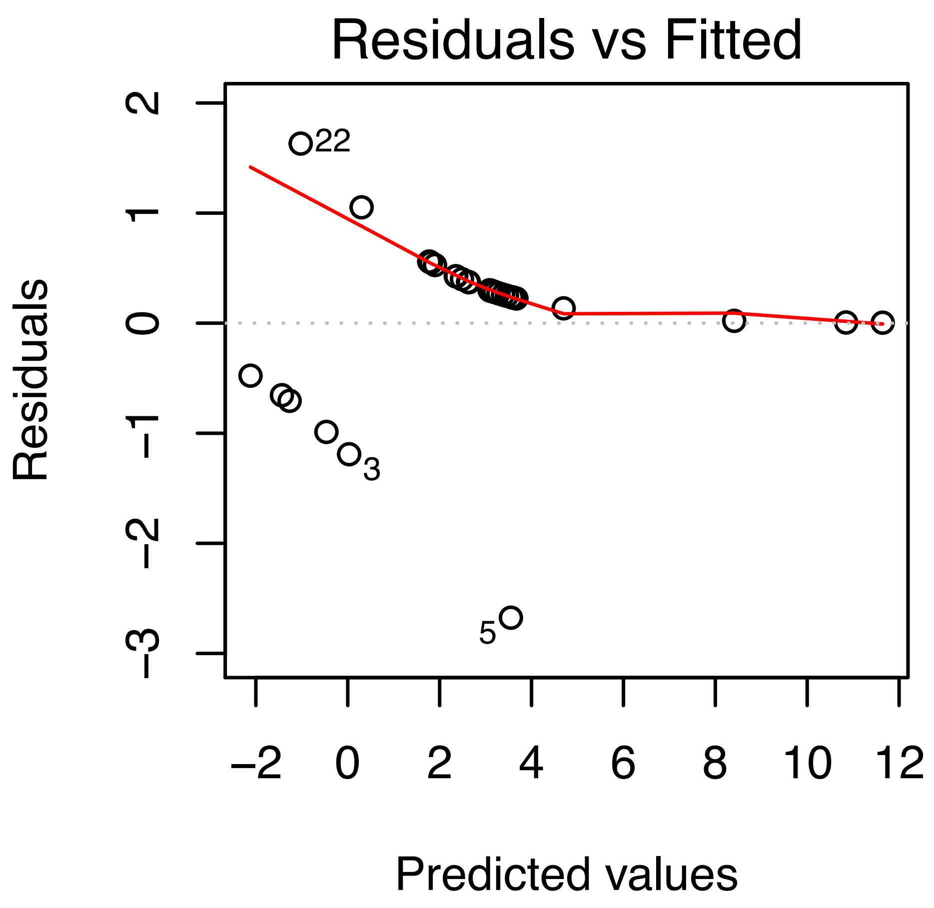
UNDER-FITTING

Low errors may **conceal underlying issues** and inaccurate assumptions.



VISUALIZING RESIDUALS

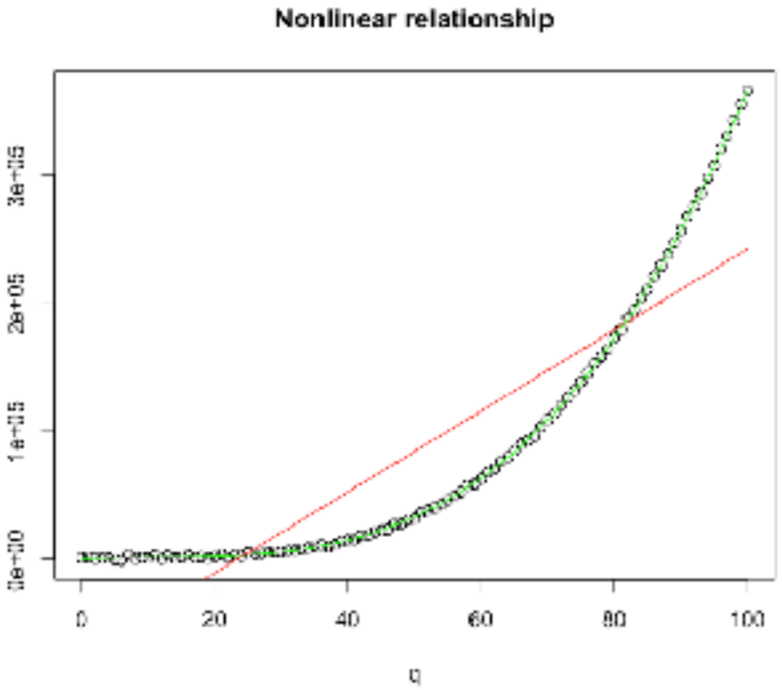
Residuals Vs Fitted and **QQ Plots** are typical graphs (e.g., in basic R output).



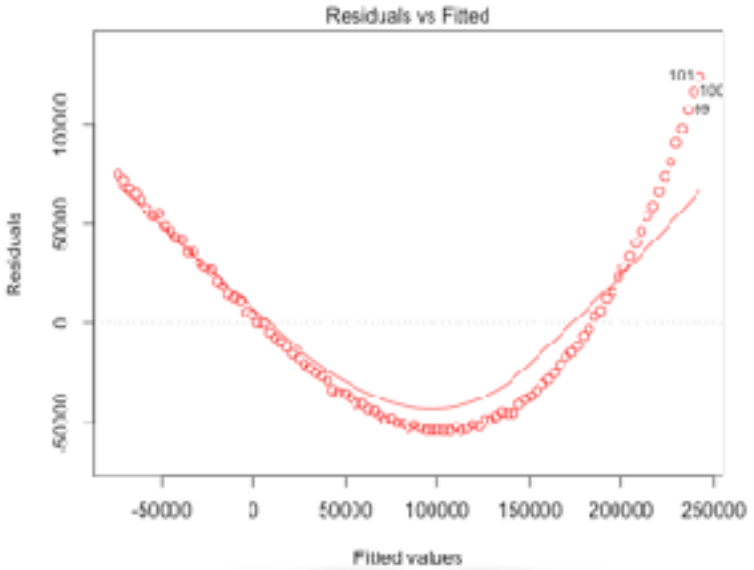
RESIDUALS VERSUS FITTED

$$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x}^2 + a_3\mathbf{x}^3 + b$$

$$\hat{y} = a\mathbf{x} + b$$

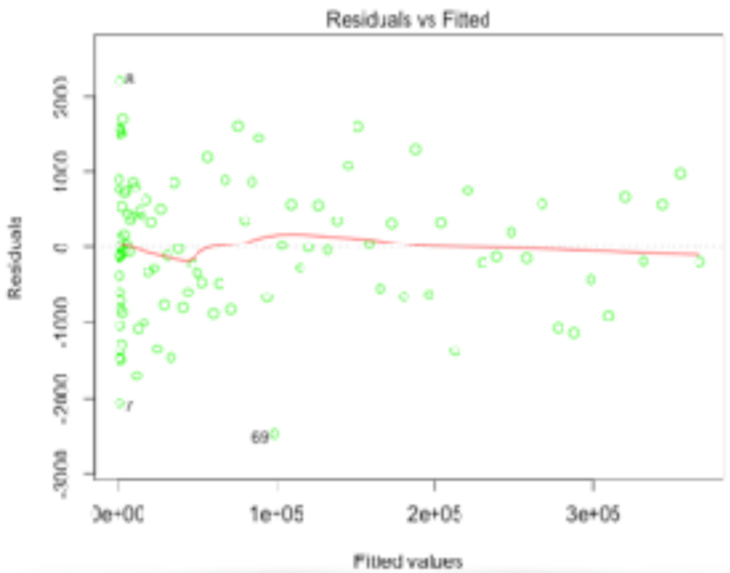


$$y - \hat{y}$$



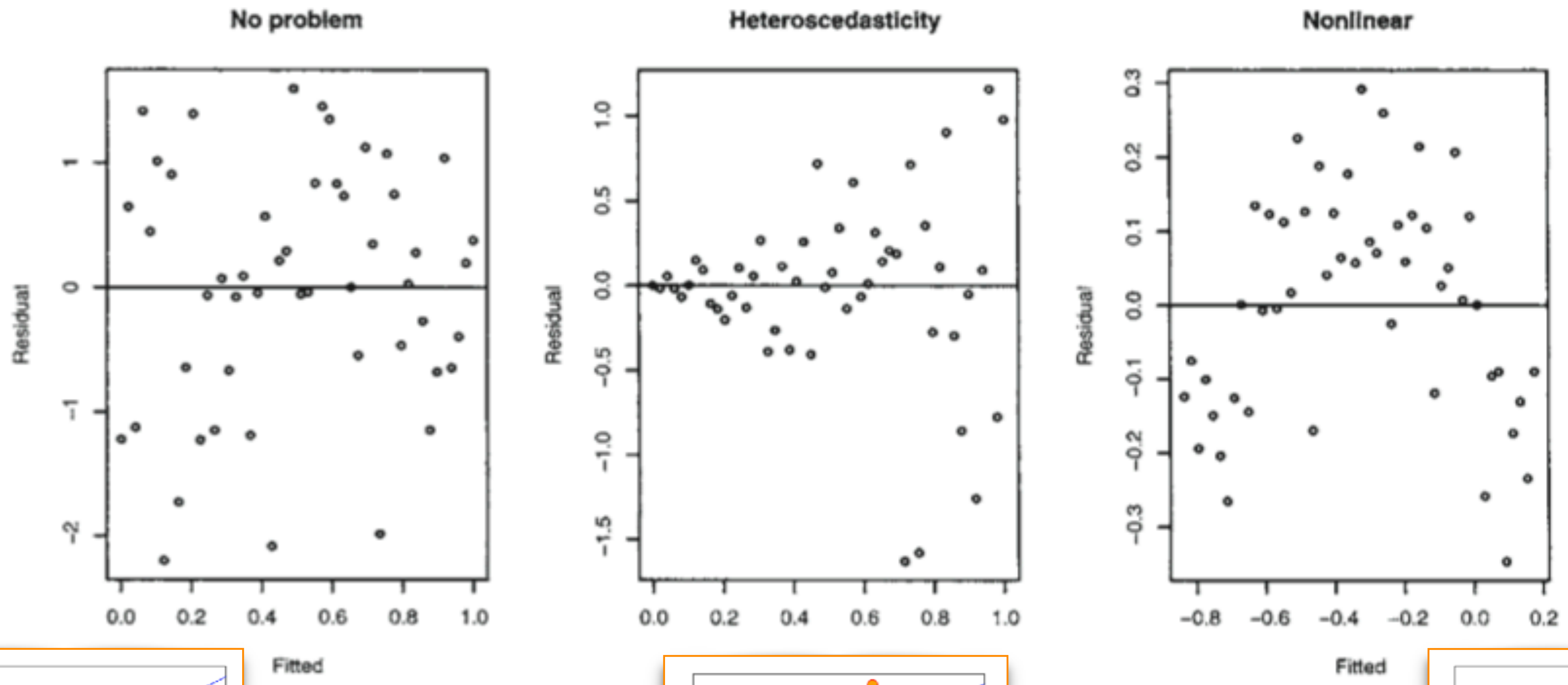
$$\hat{y} = a\mathbf{x} + b$$

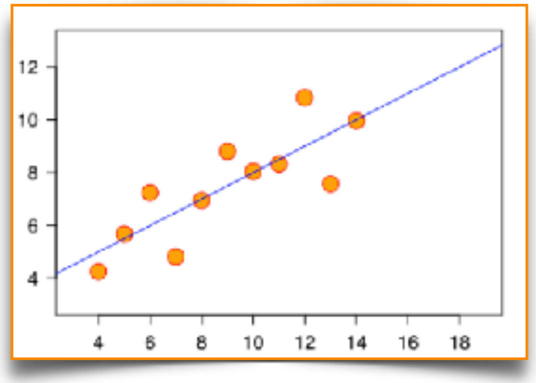
$$y - \hat{y}$$

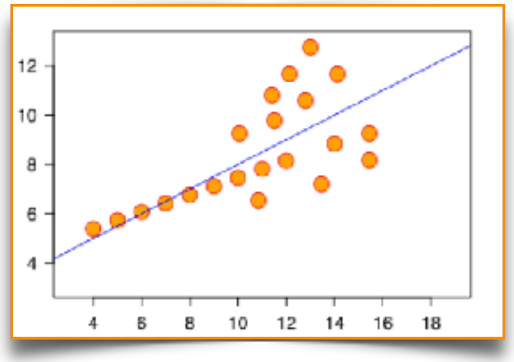


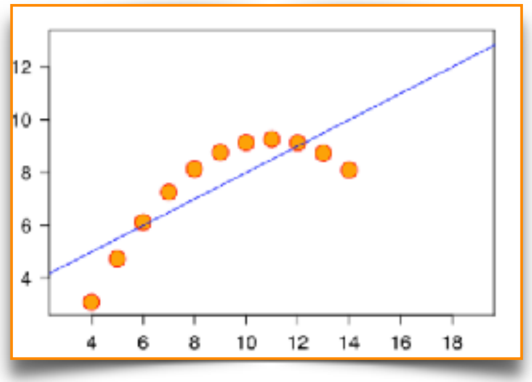
$$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x}^2 + a_3\mathbf{x}^3 + b$$

RESIDUALS VERSUS FITTED





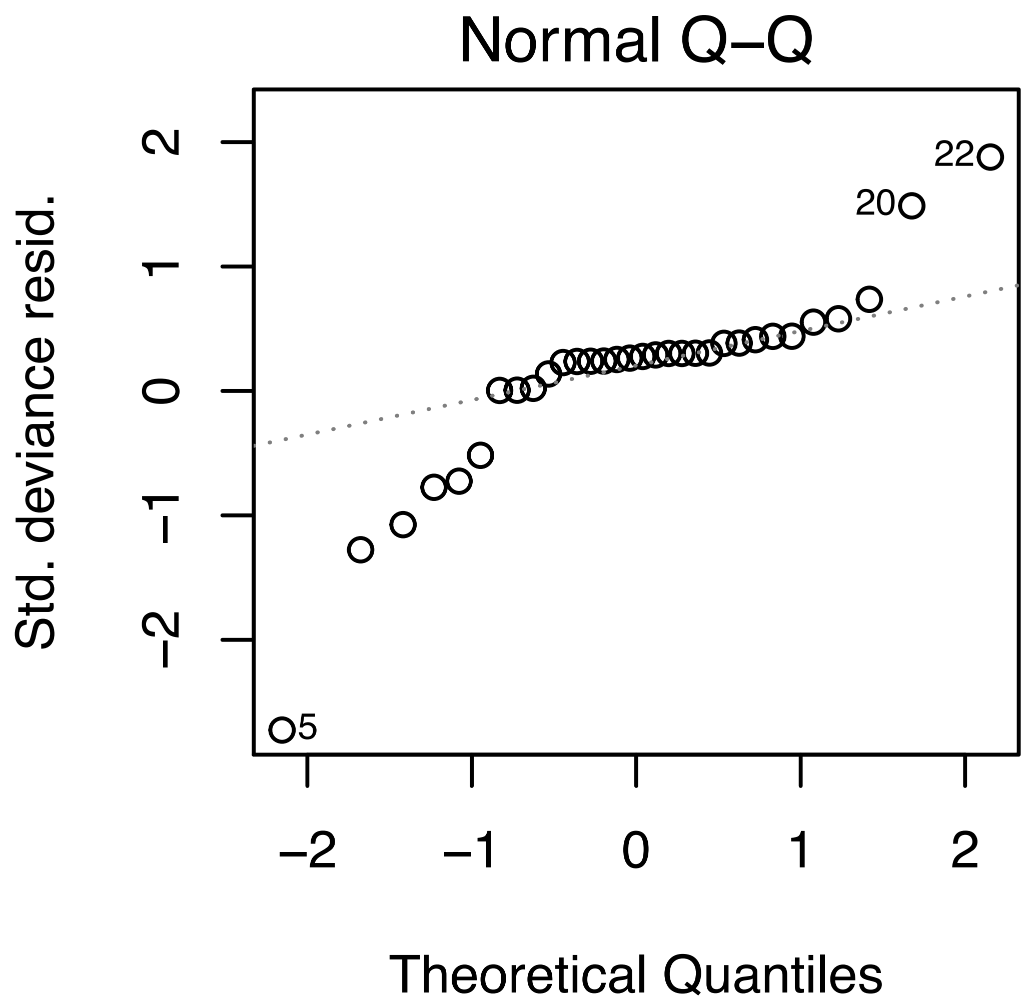




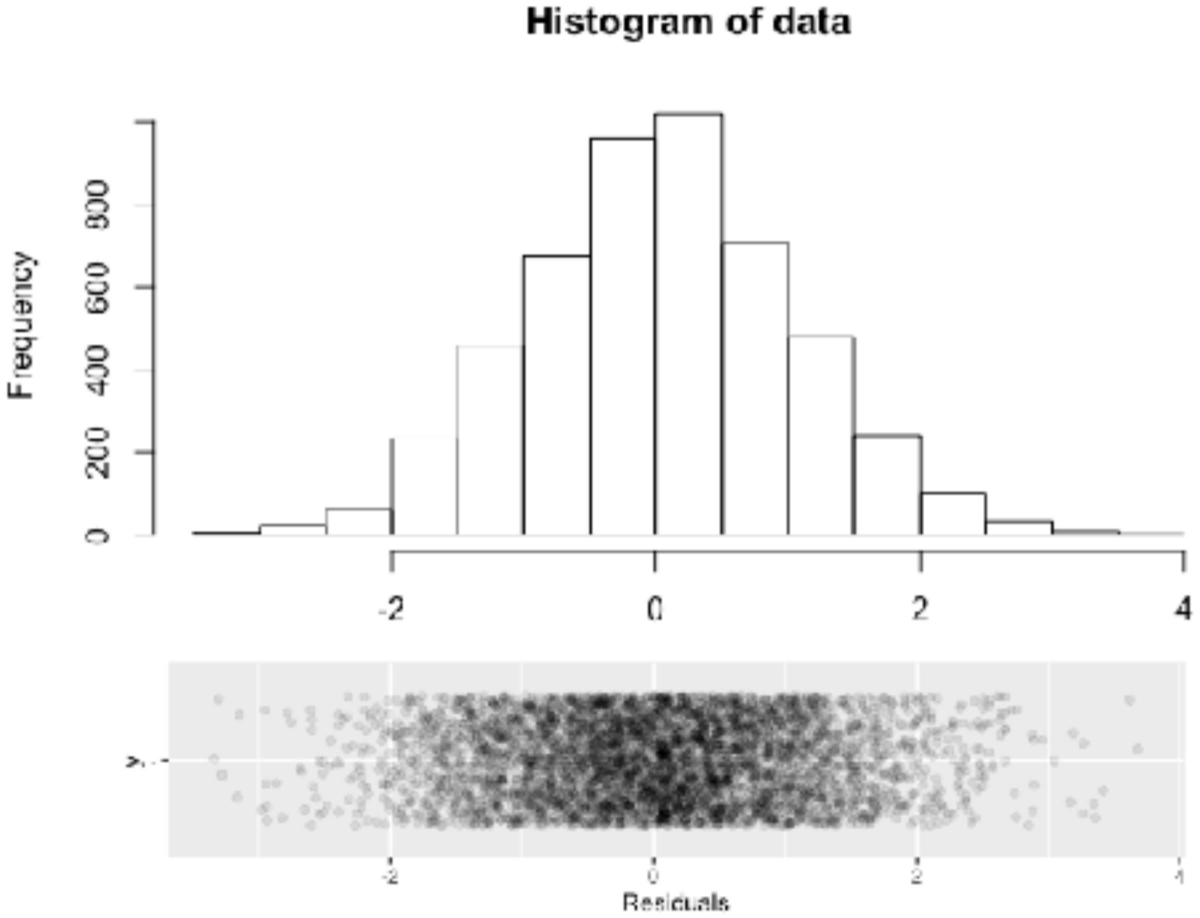
[1] Faraway, Linear Models with R (2005, p. 59)

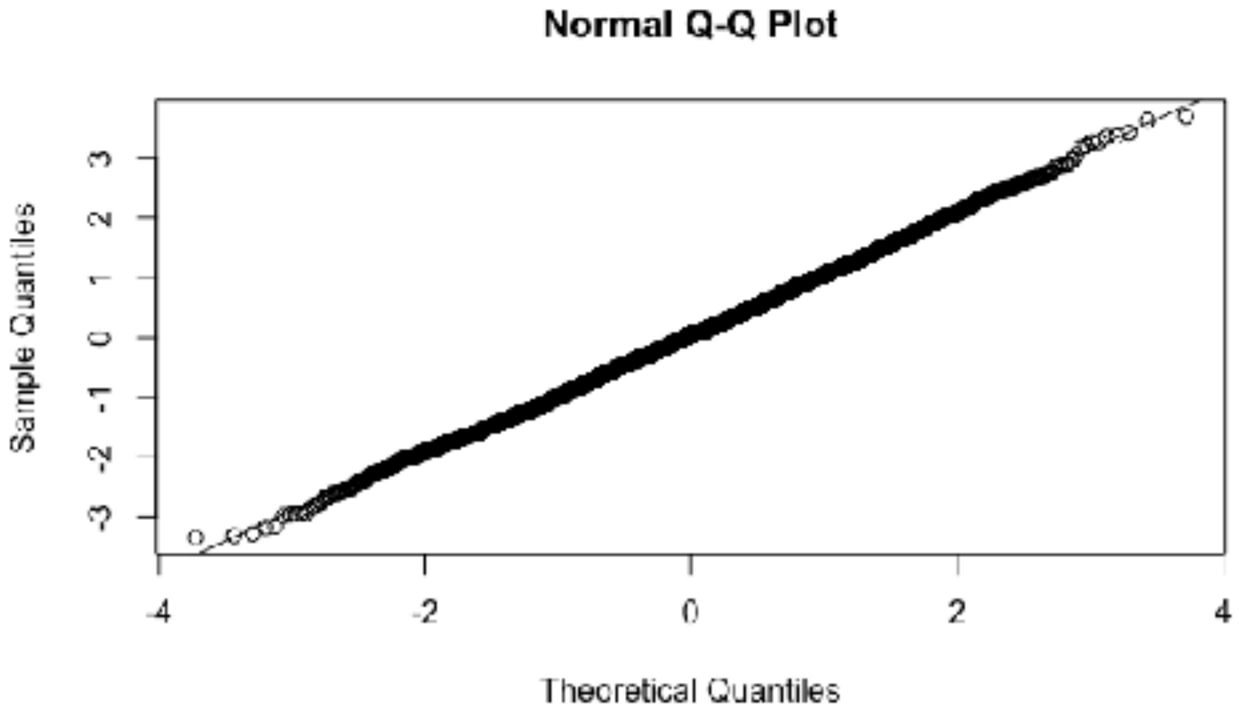
VISUALIZING RESIDUALS

Two typical plots (e.g., basic R output).

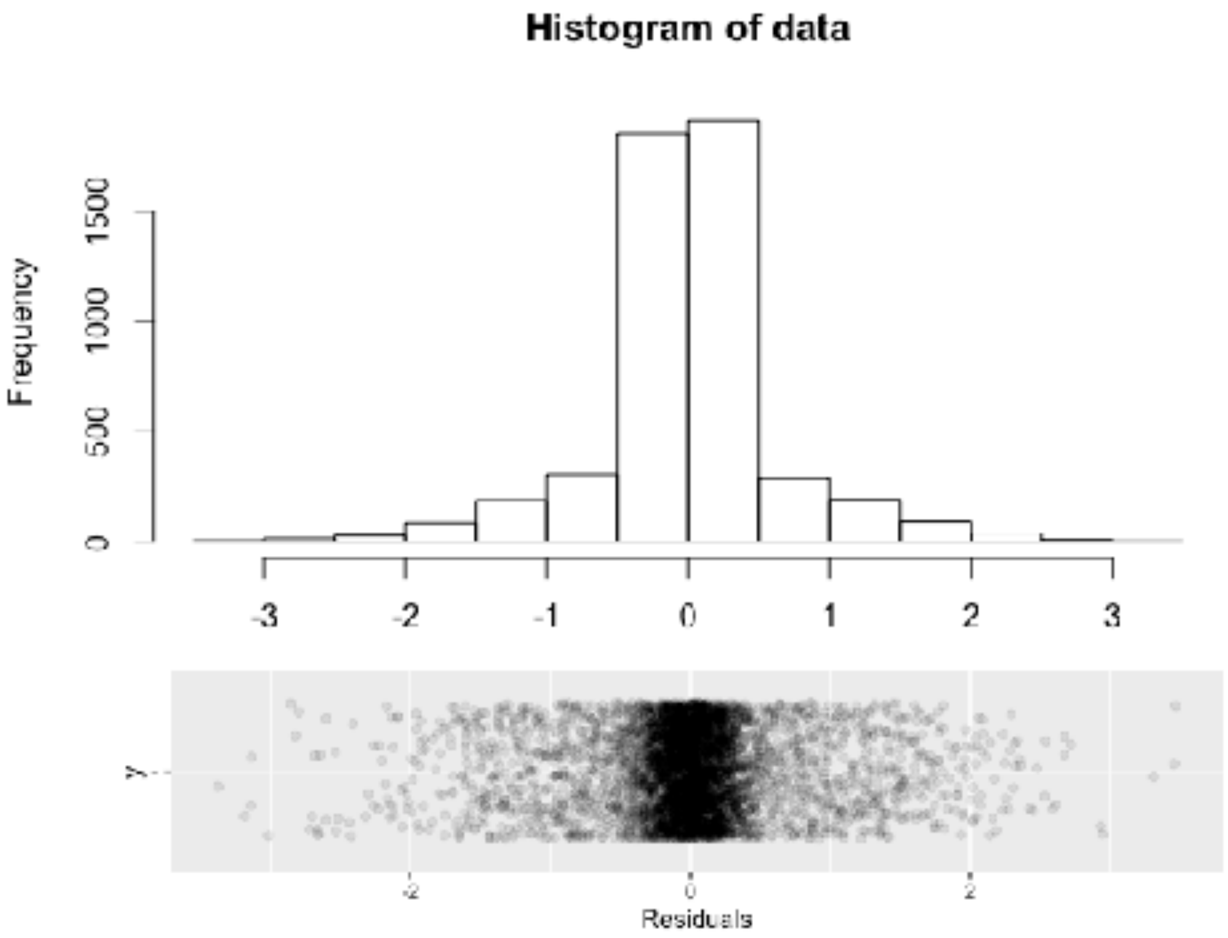


QQ PLOT

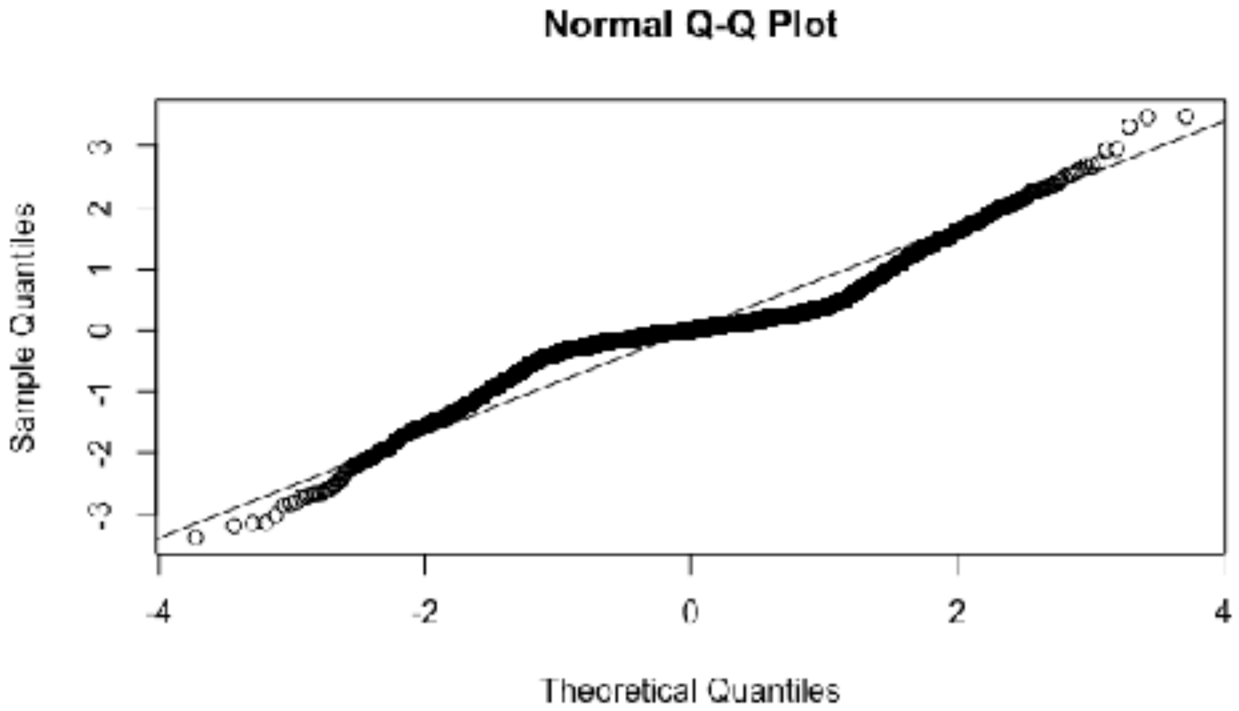

Histogram of data
Frequency
0
200
400
600
800
-2
0
2
4
Residuals


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4
-3
-2
-1
0
1
2
3

QQ PLOT

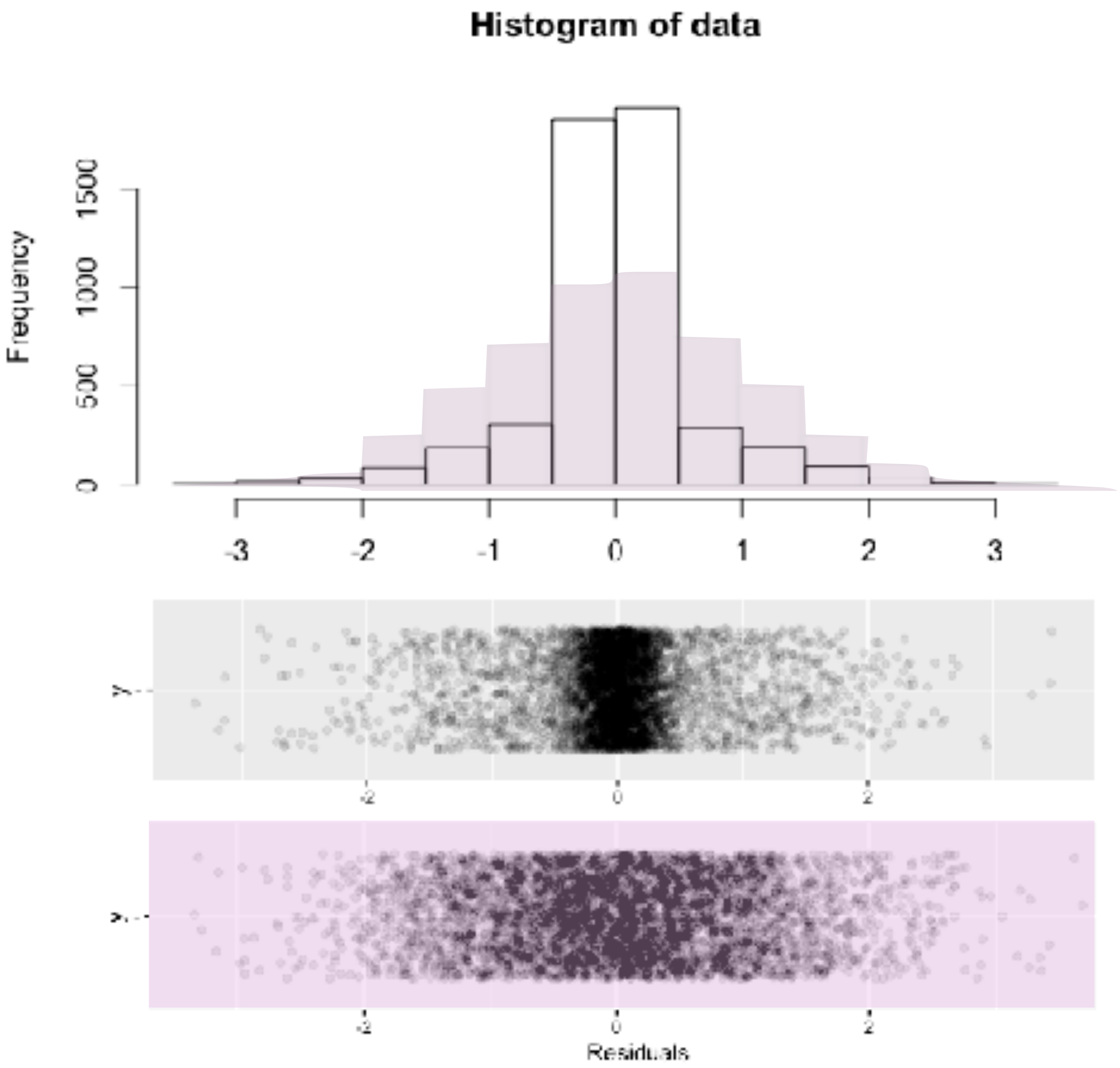


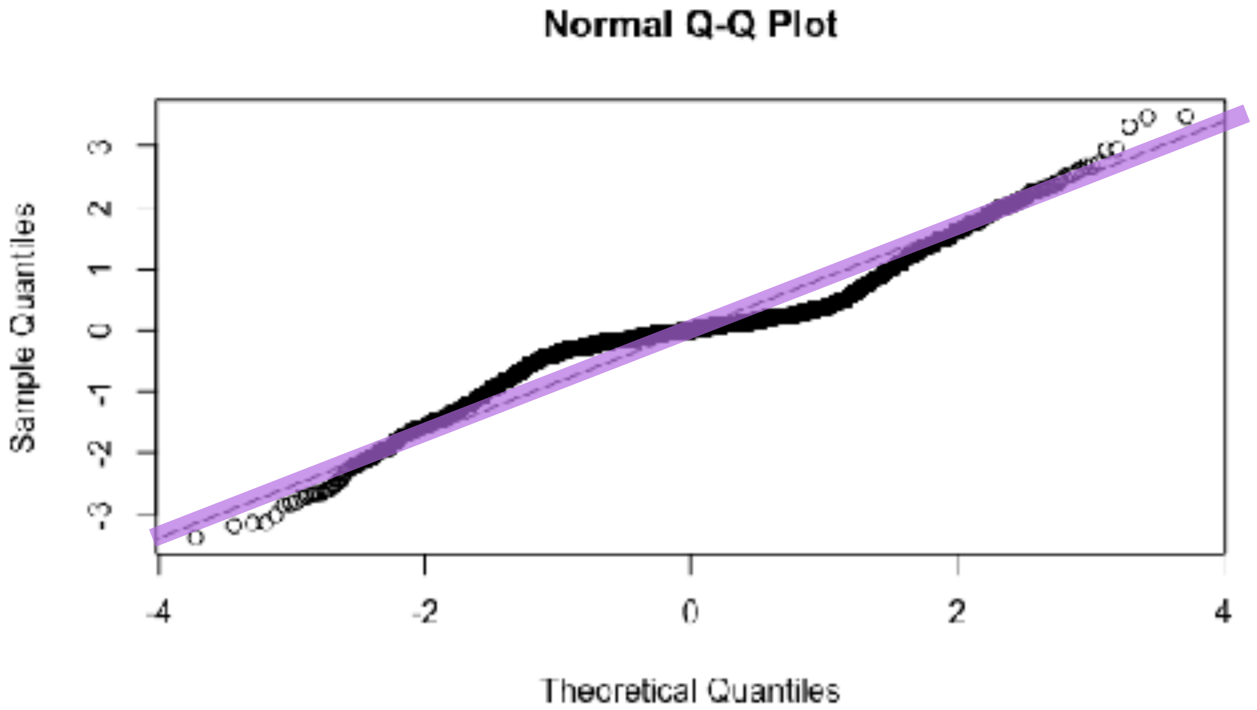
Histogram of data
Frequency
0
500
1000
1500
-3
-2
-1
0
1
2
3
y
-2
0
2
Residuals



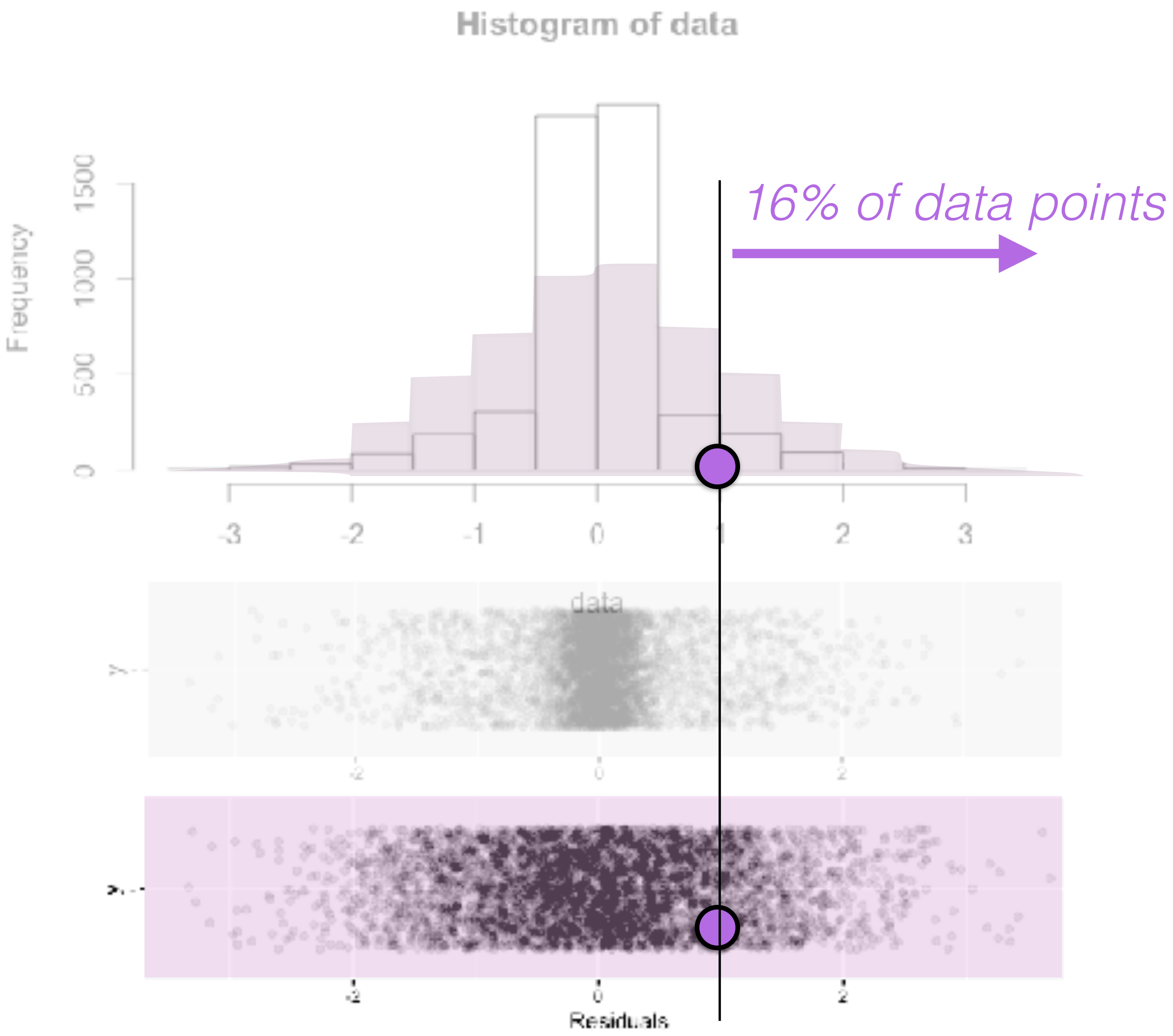
Normal Q-Q Plot
Sample Quantiles
-3
-2
-1
0
1
2
3
-4
-2
0
2
4
Theoretical Quantiles

QQ PLOT

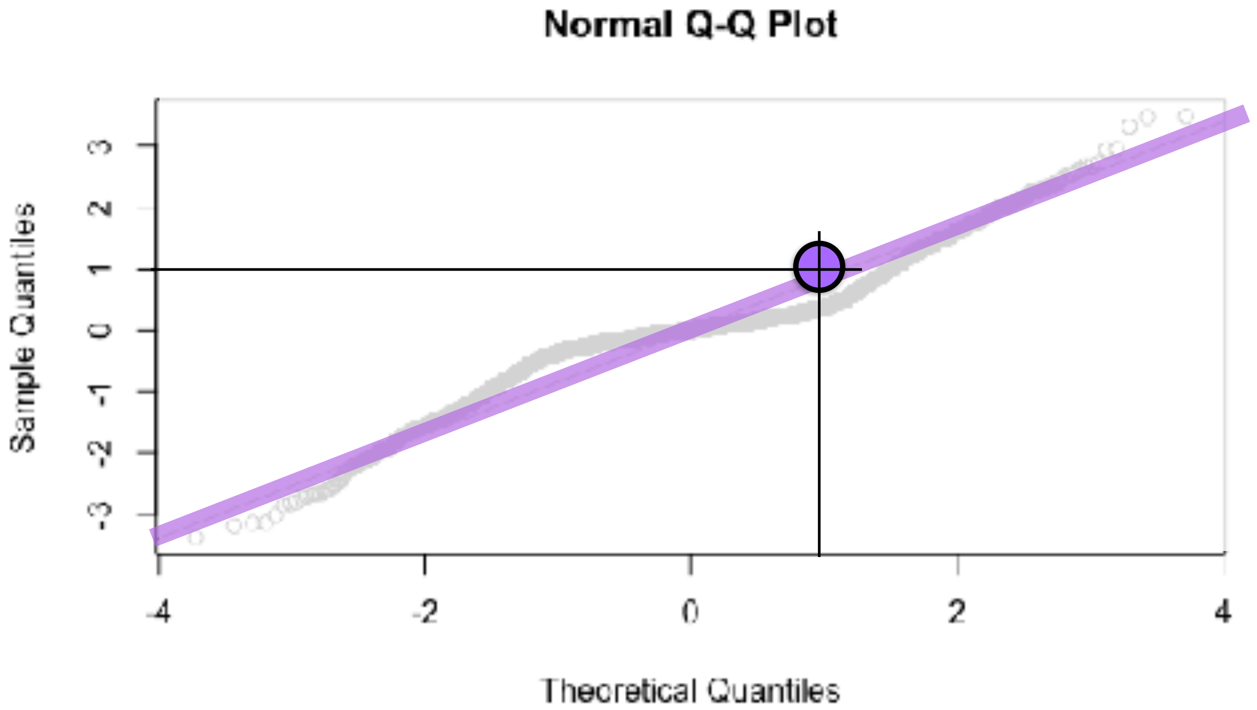

Histogram of data
Frequency
0
500
1000
1500
-3
-2
-1
0
1
2
3
-2
0
2
-2
0
2
Residuals


Normal Q-Q Plot
Sample Quantiles
-3
-2
-1
0
1
2
3
-4
-2
0
2
4
Theoretical Quantiles

QQ PLOT

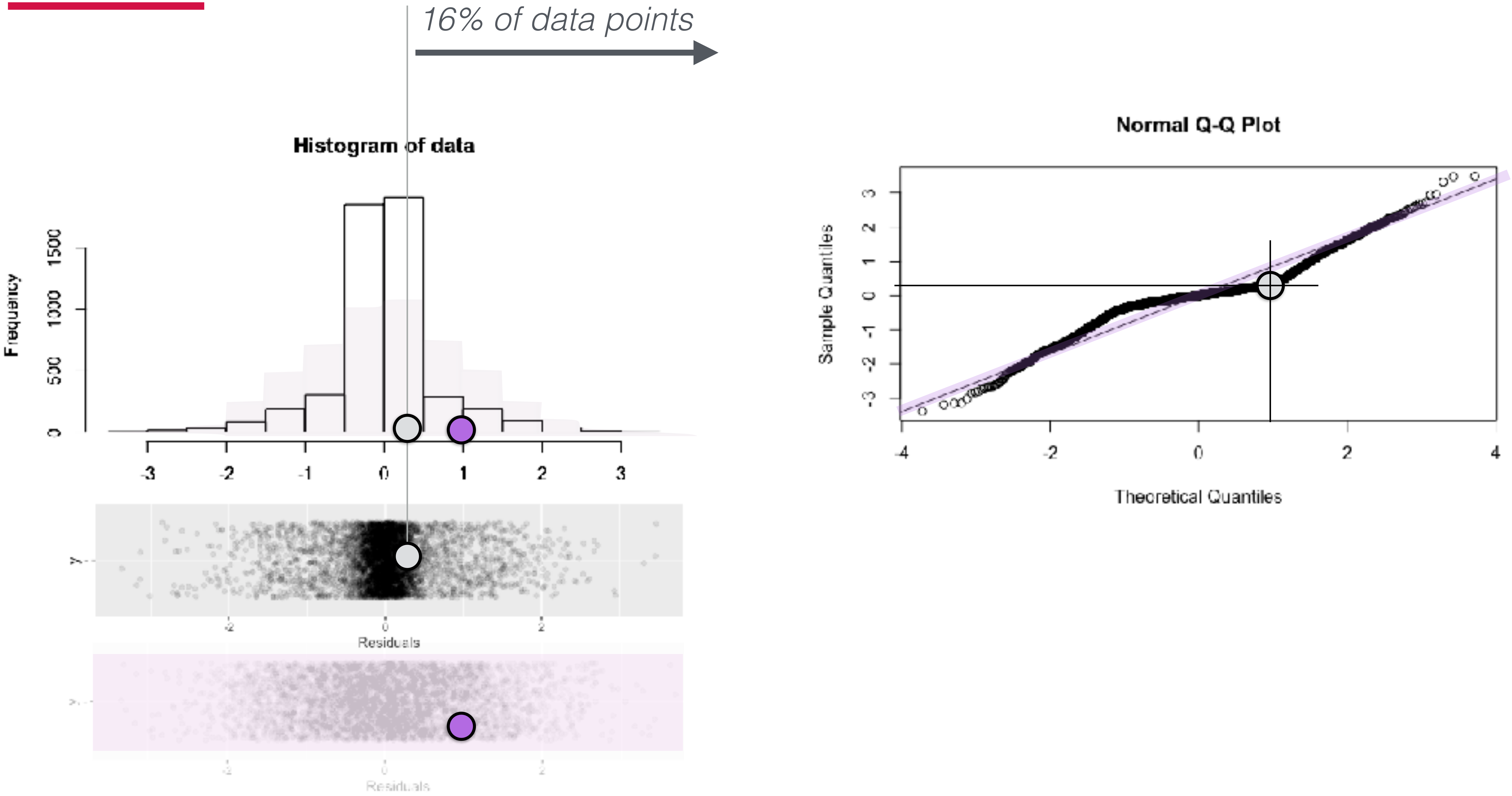


Histogram of data
Frequency
0
500
1000
1500
-3
-2
-1
0
1
2
3
data
16% of data points
Residuals

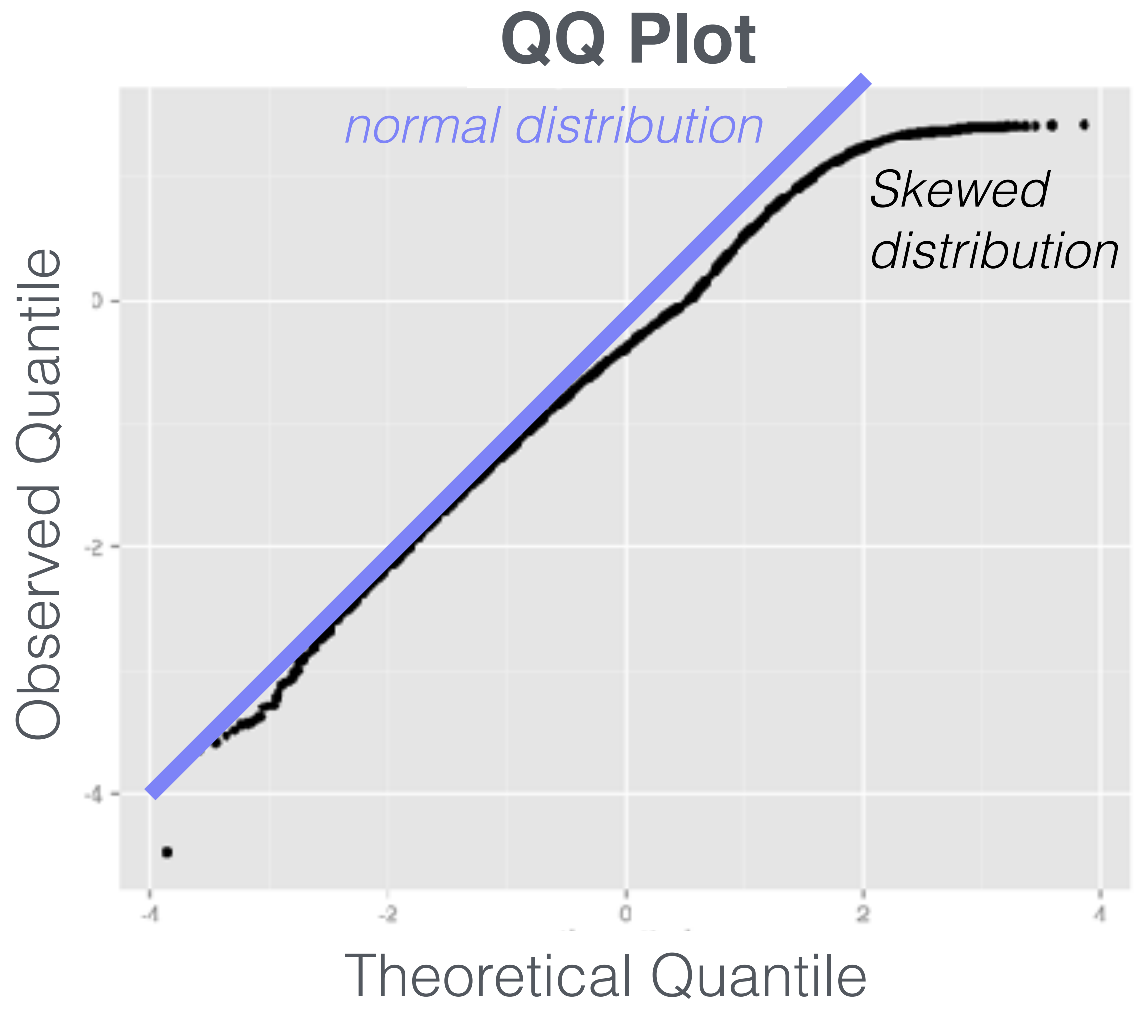


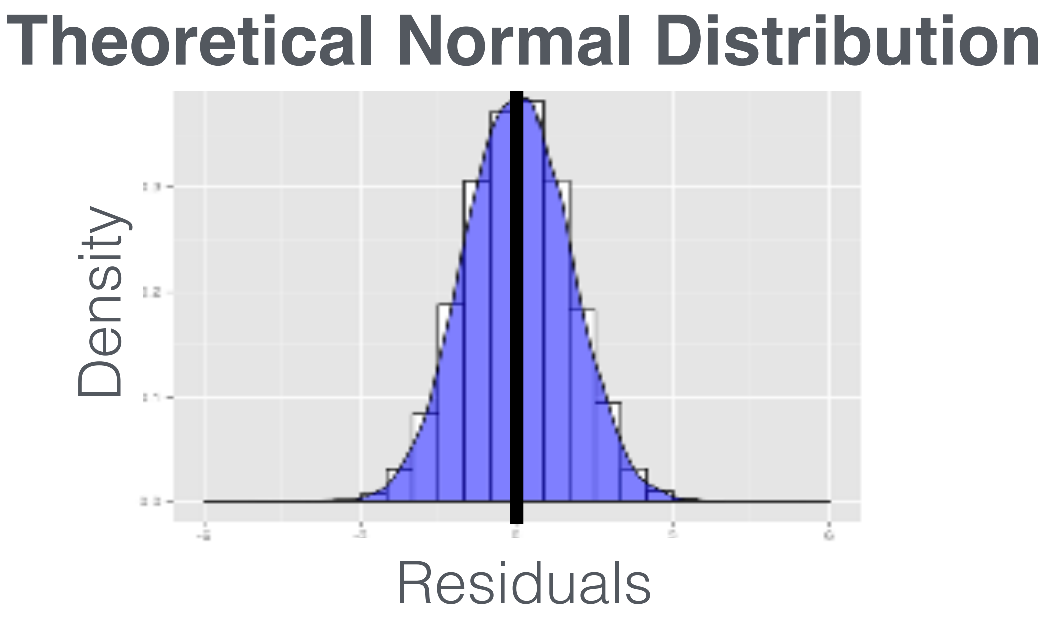
Normal Q-Q Plot
Sample Quantiles
-3
-2
-1
0
1
2
3
-4
-2
0
2
4
Theoretical Quantiles

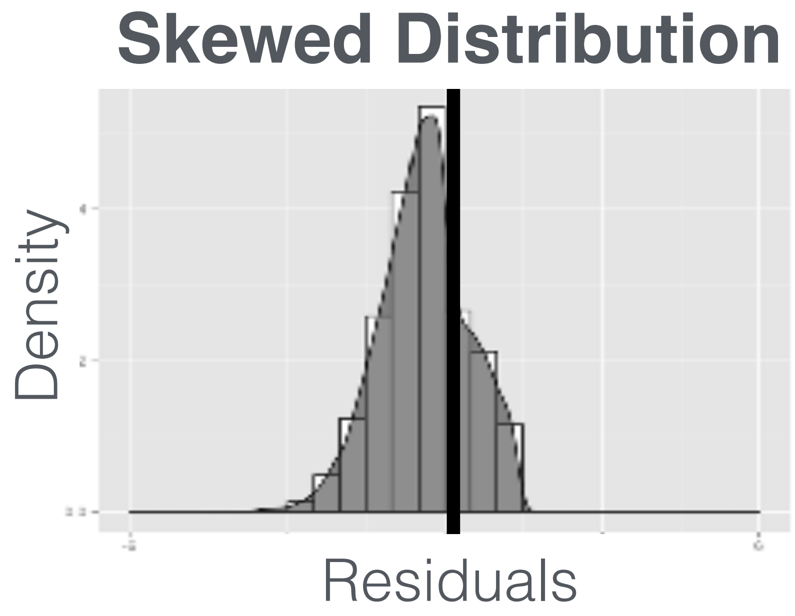
QQ PLOT


16% of data points
Histogram of data
Frequency
0
500
1000
1500
-3
-2
-1
0
1
2
3
Residuals
Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-4
-2
0
2
4

QQ PLOT

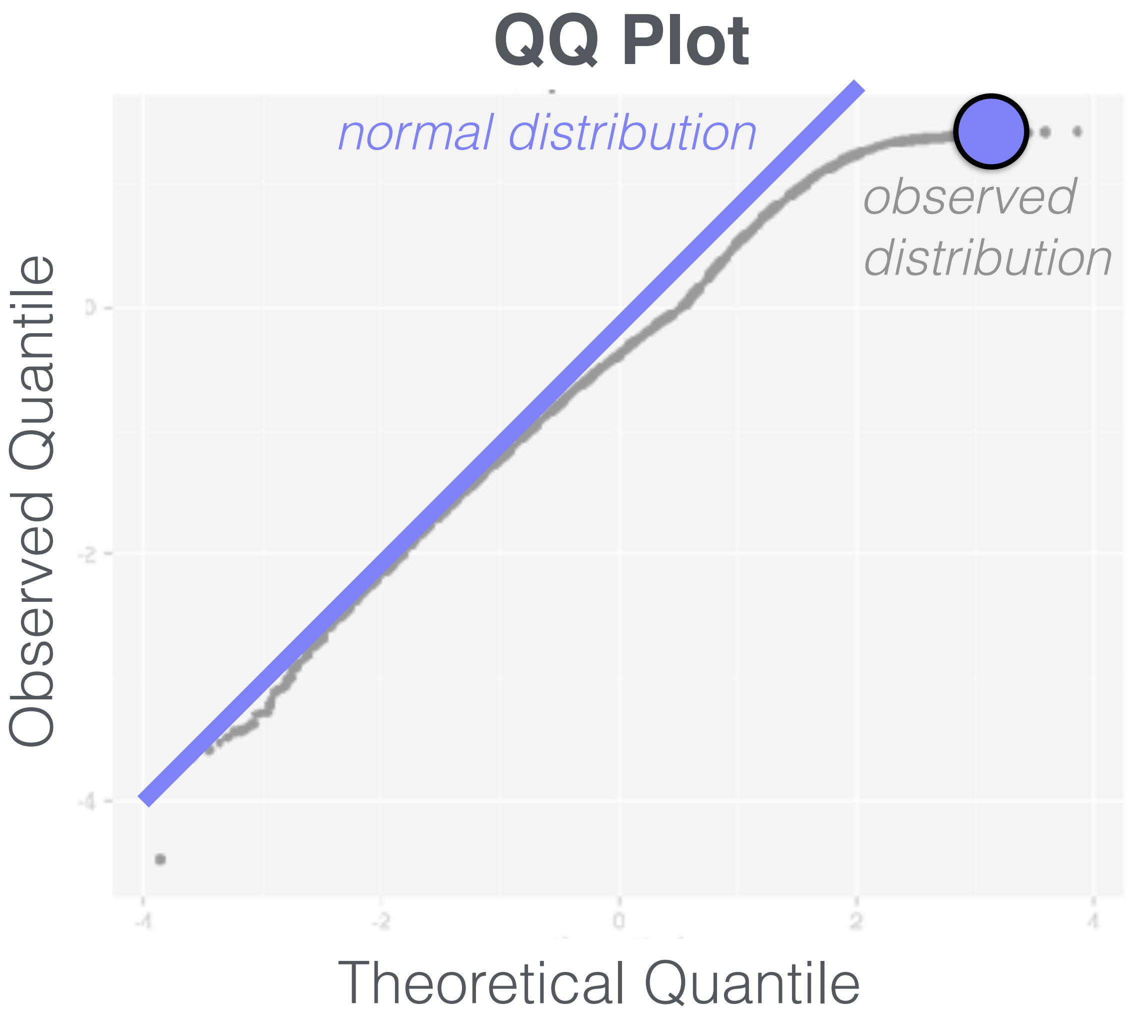




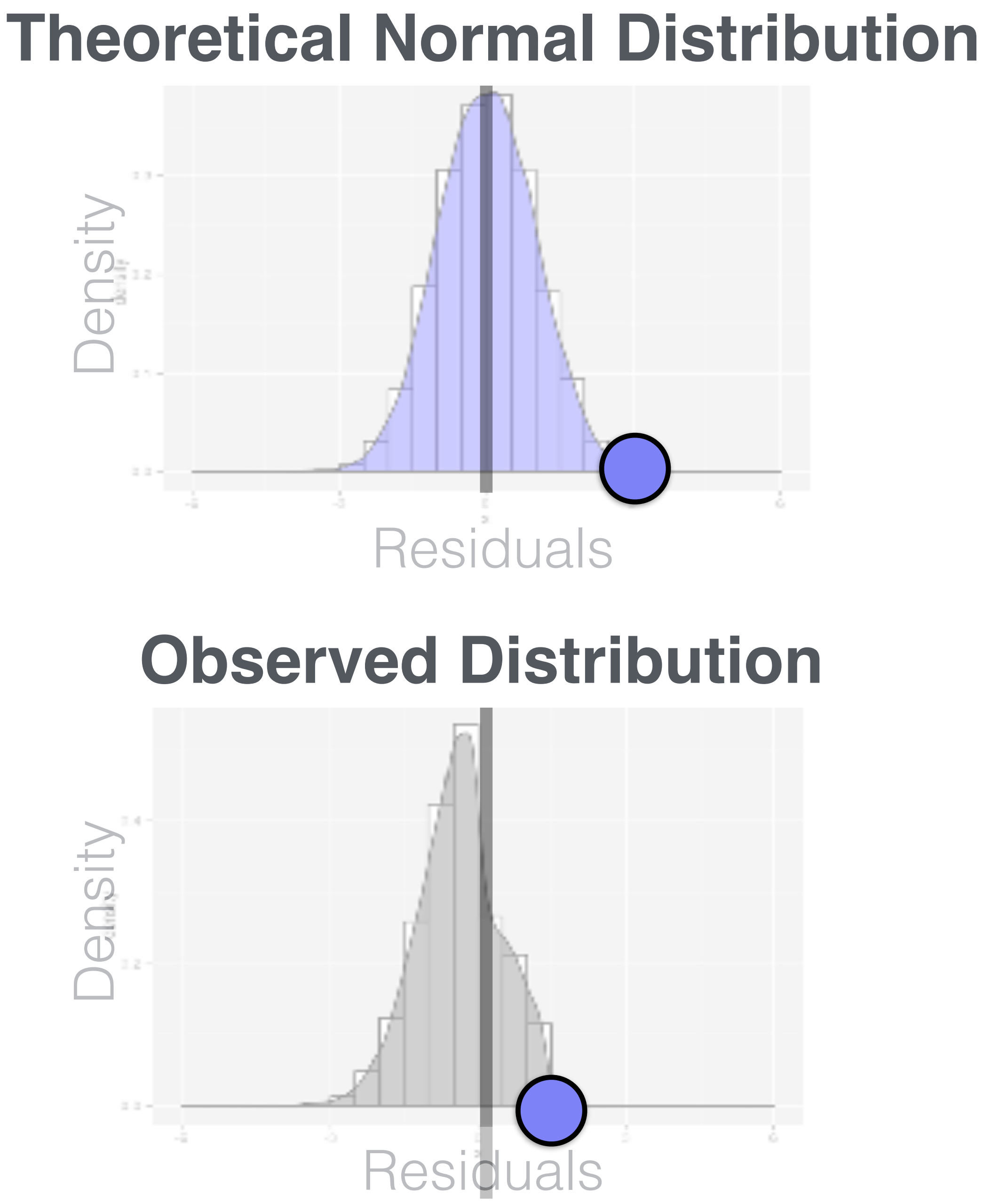


https://xiongge.shinyapps.io/QQplots/

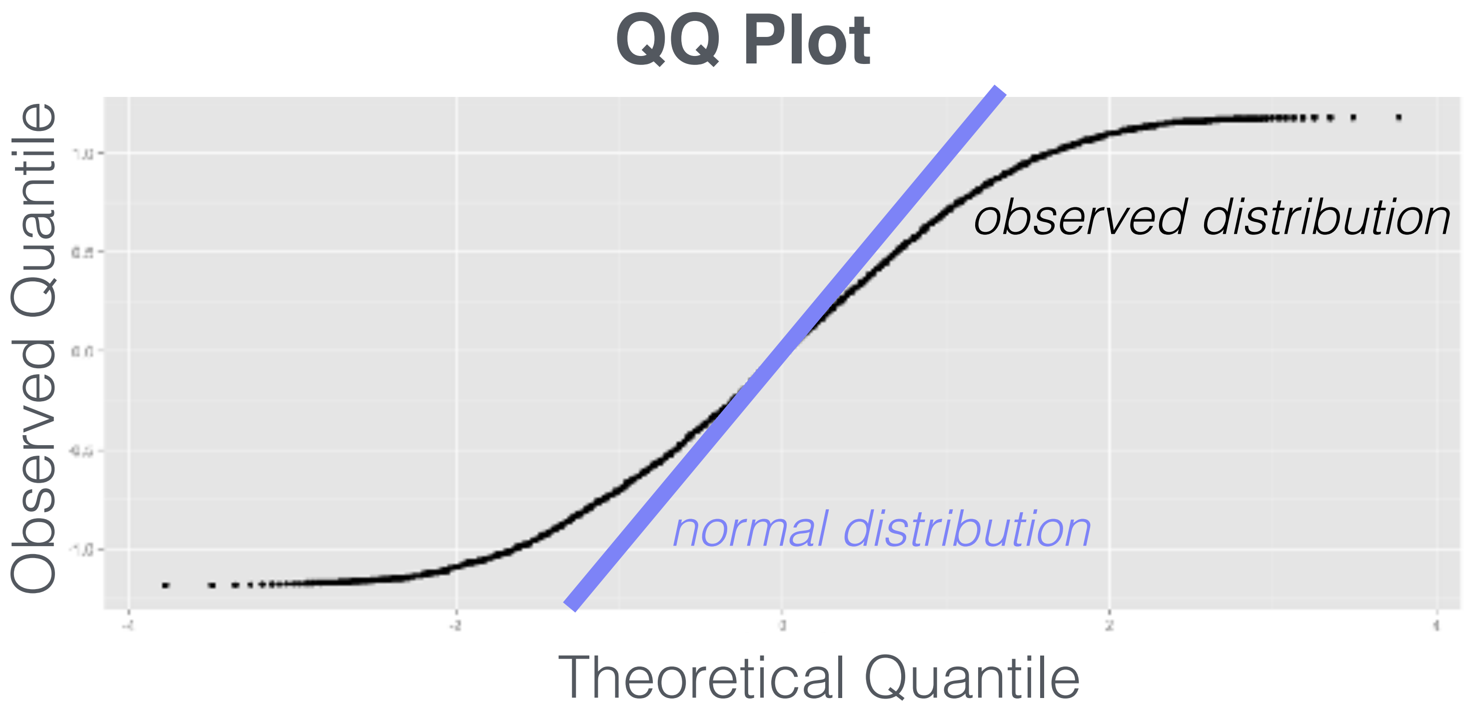
QQ PLOT



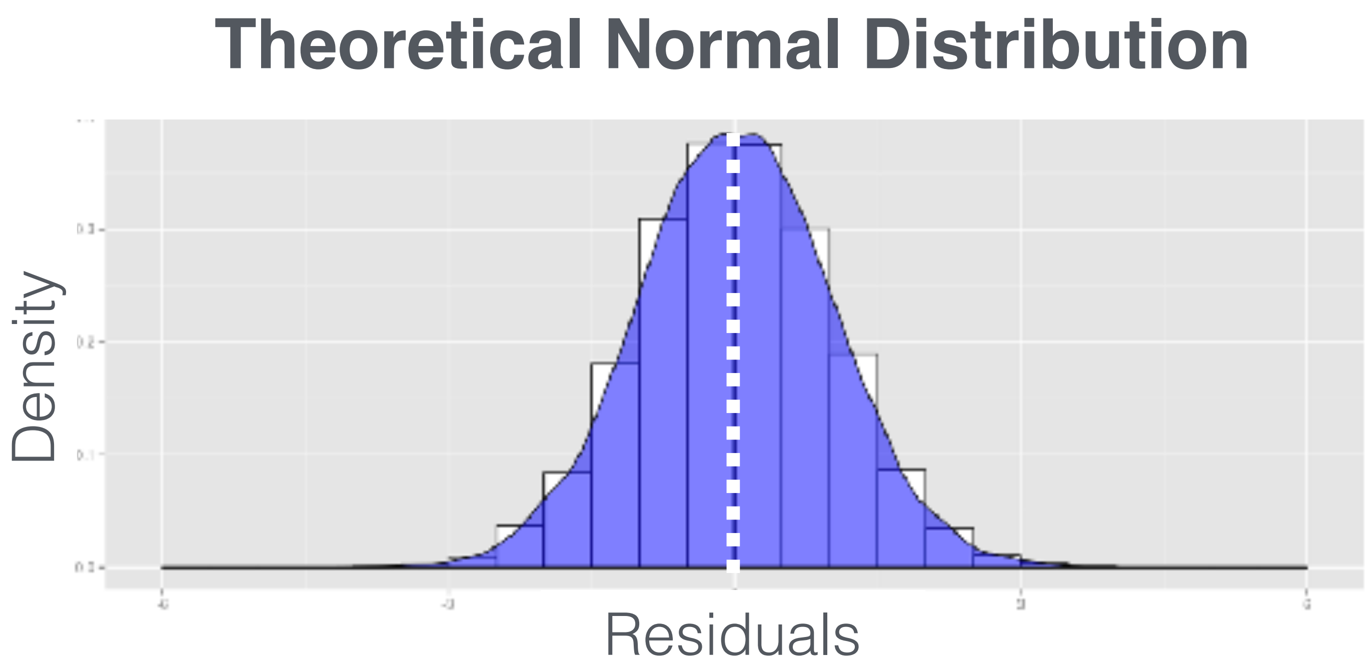
https://xiongge.shinyapps.io/QQplots/

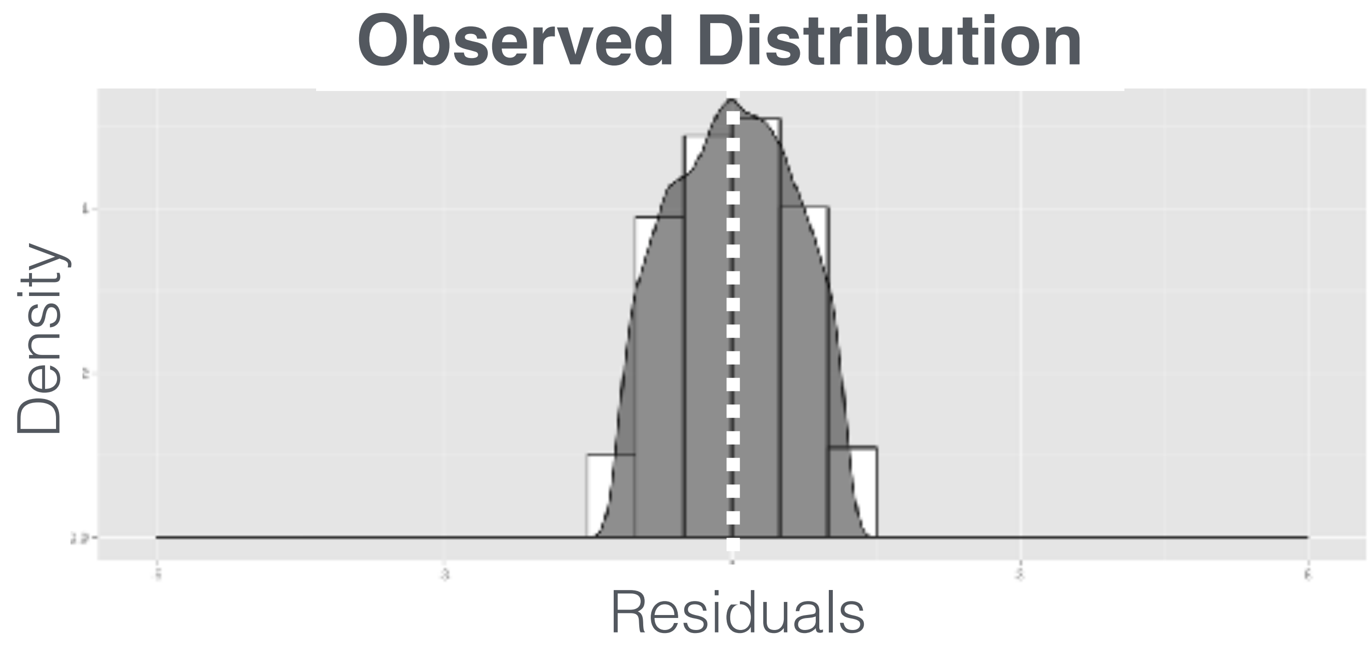


QQ PLOT

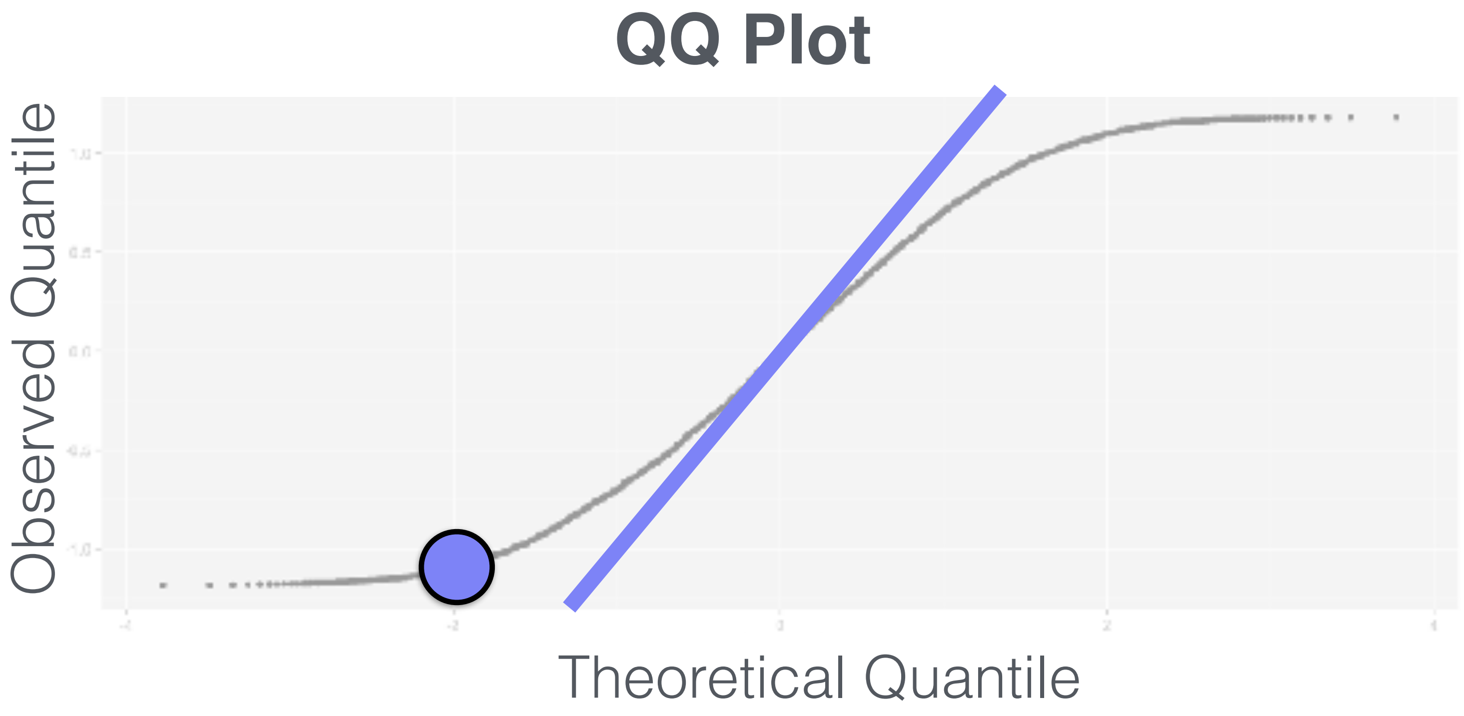


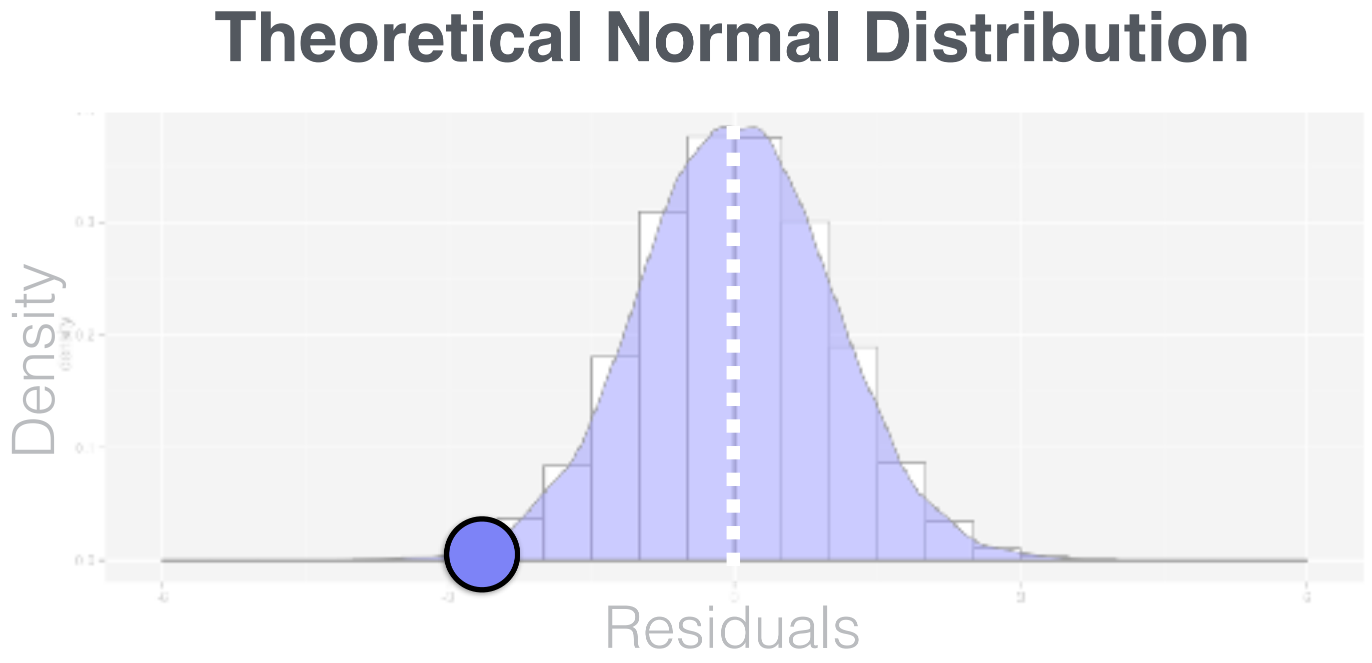
https://xiongge.shinyapps.io/QQplots/

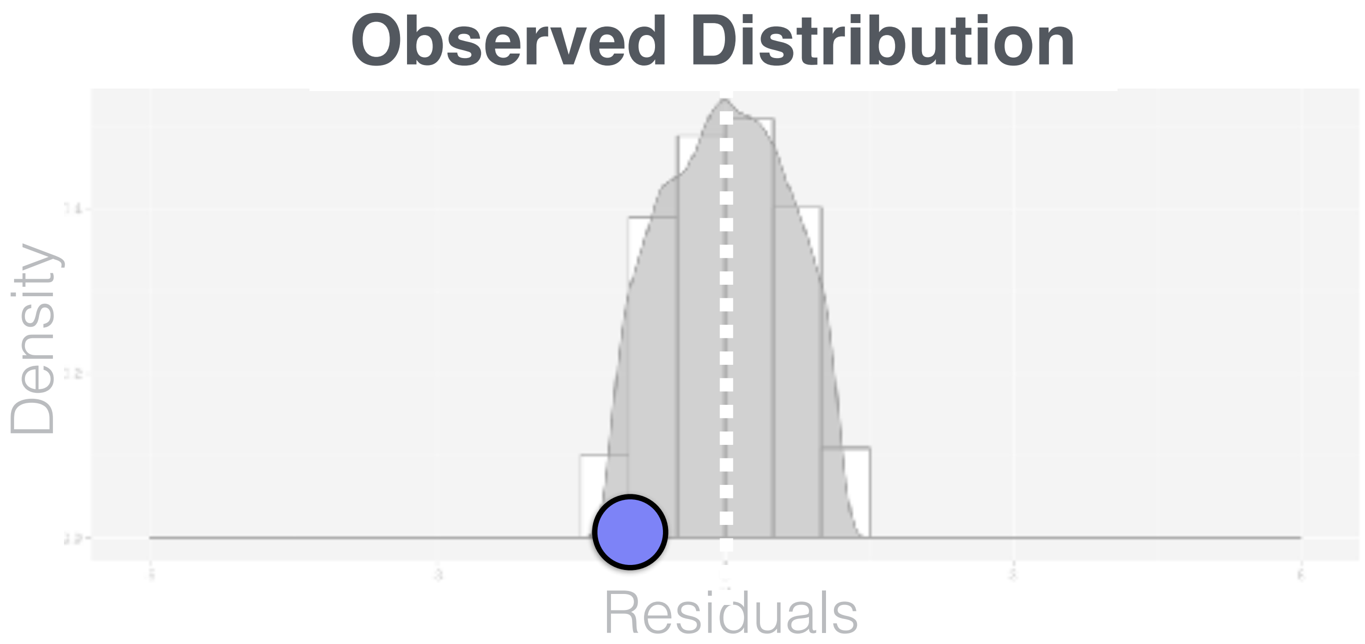




QQ PLOT

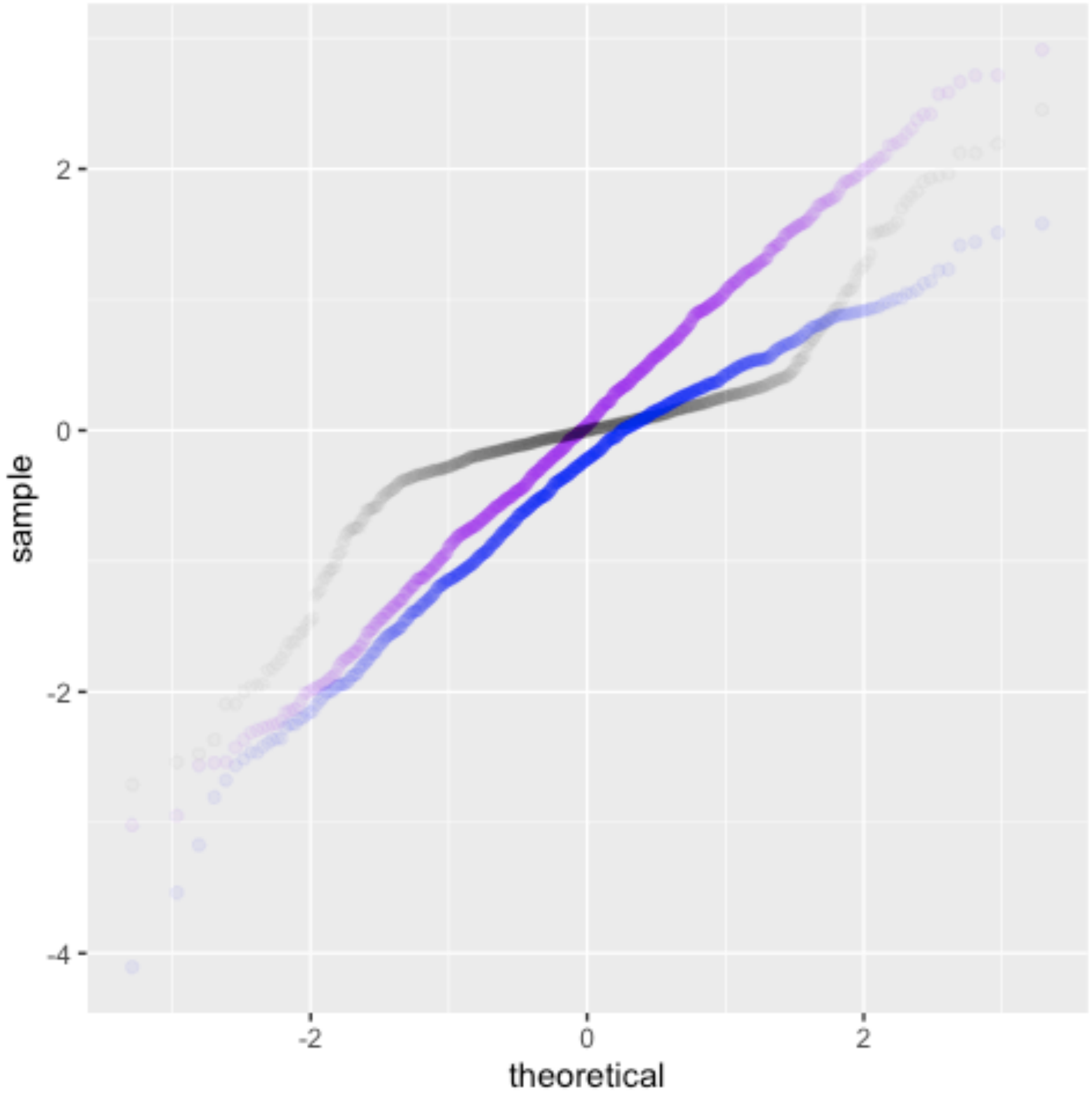




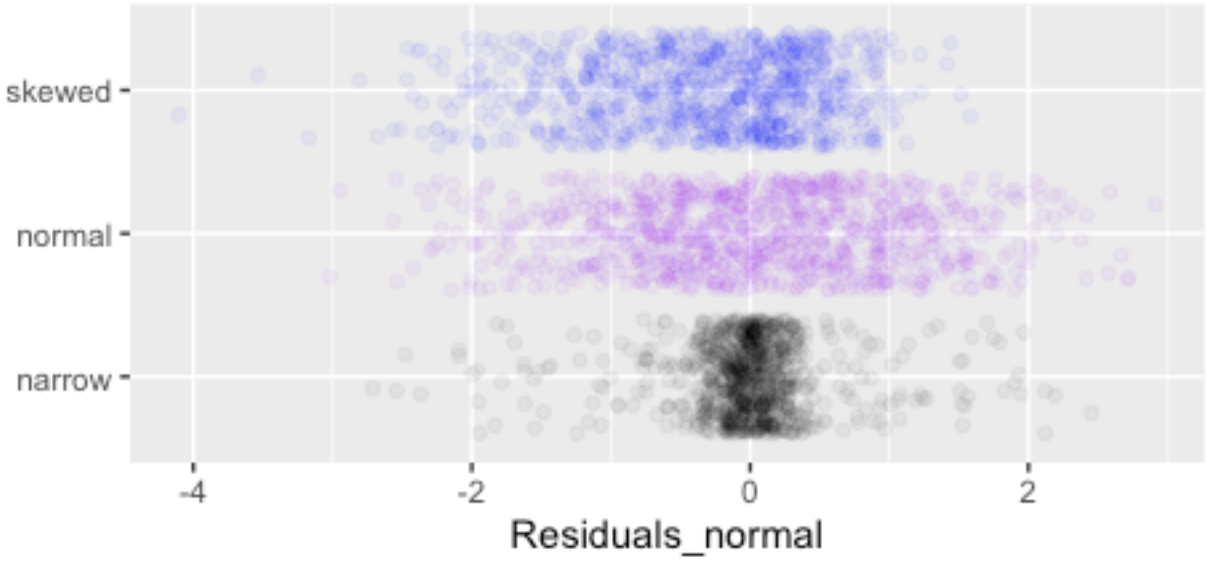


https://xiongge.shinyapps.io/QQplots/

QQ PLOT



2
0
-2
-4
sample
-2
0
2
theoretical

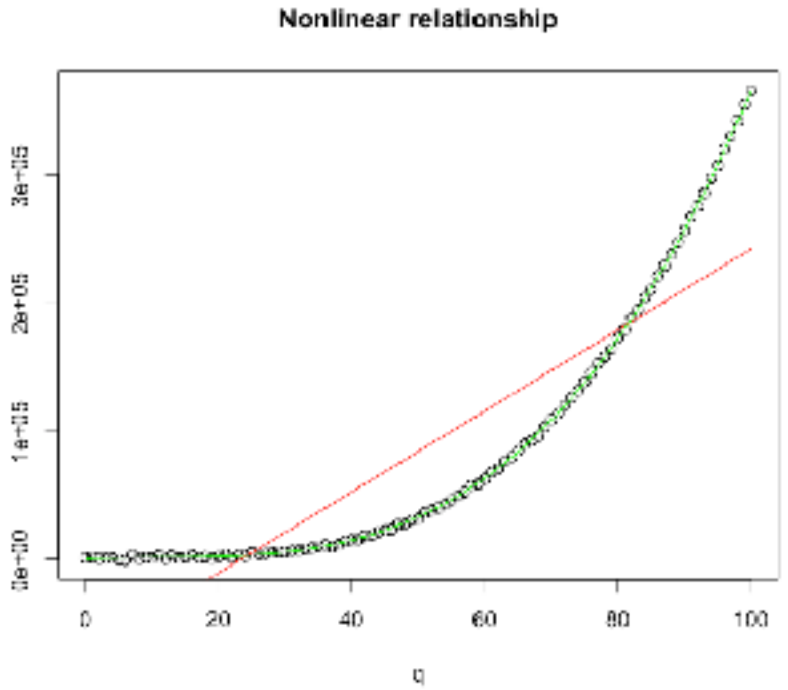


skewed
normal
narrow
-4
-2
0
2
Residuals_normal

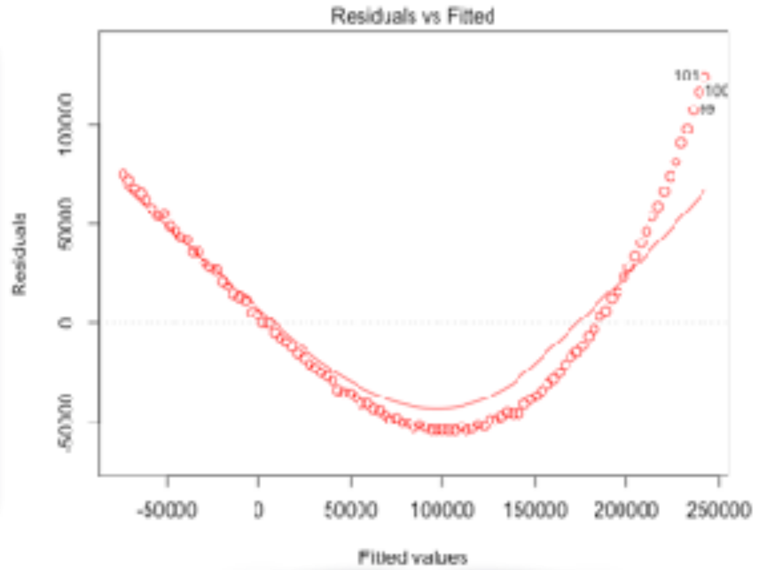
EXAMPLE

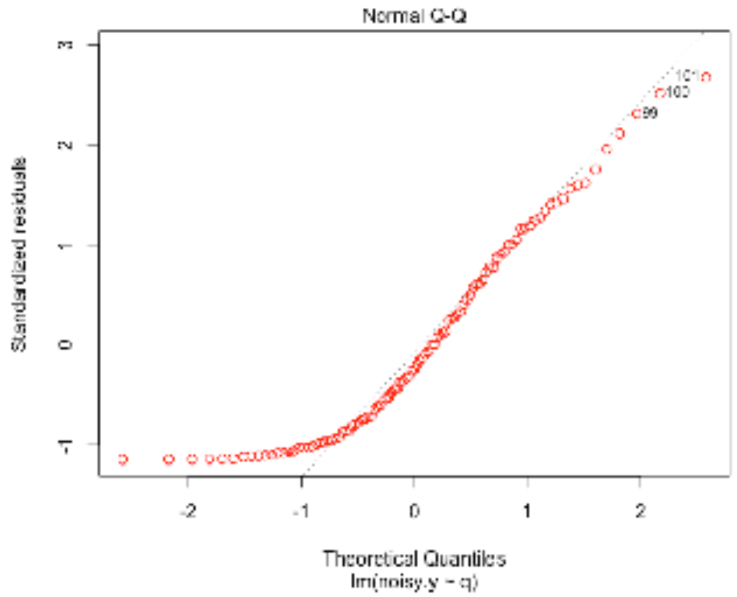
$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x}^2 + a_3\mathbf{x}^3 + b$

$\hat{y} = a\mathbf{x} + b$



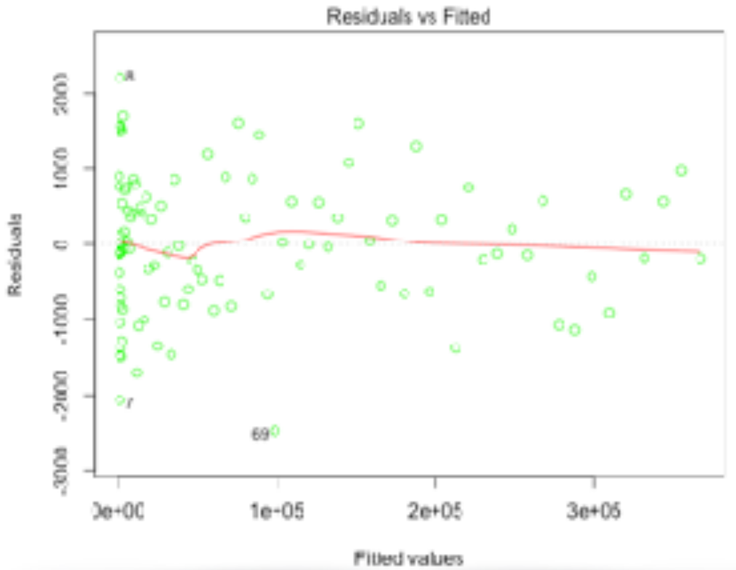
$y - \hat{y}$

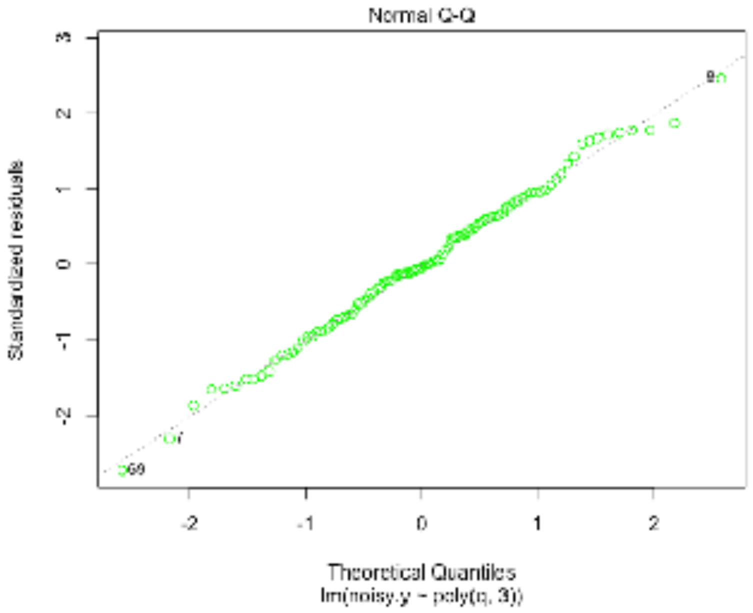




$\hat{y} = a\mathbf{x} + b$

$y - \hat{y}$





$\hat{y} = a_1\mathbf{x} + a_2\mathbf{x}^2 + a_3\mathbf{x}^3 + b$

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

TEST SET
TRAINING SET
TARGET SET

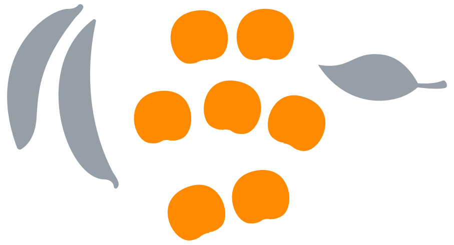
EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

BASIC ERRORS

- **Classification errors** are like grain quality.

"How many stones and straws are in this bag of grain?"

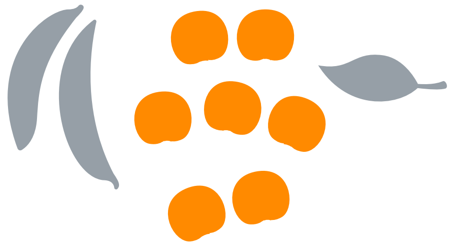


Elements are in the right category, or not.

BASIC ERRORS

- **Classification errors** are like grain quality.

"How many stones and straws are in this bag of grain?"



Elements are in the right category, or not.

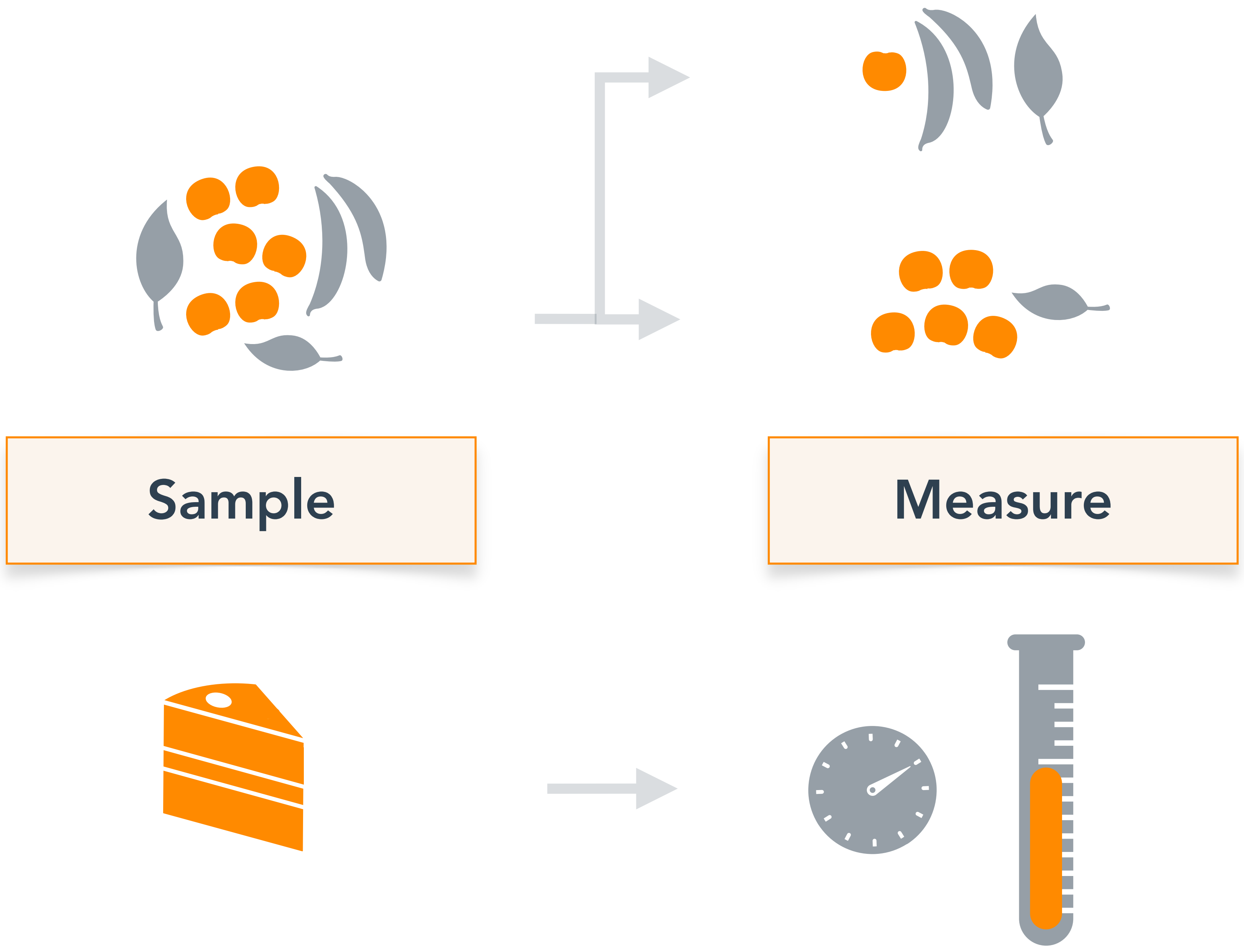
- **Regression errors** are like nutritional content.

"This much cholesterol is in my cake, really?"

Quantities are over- or under-estimated, or not.

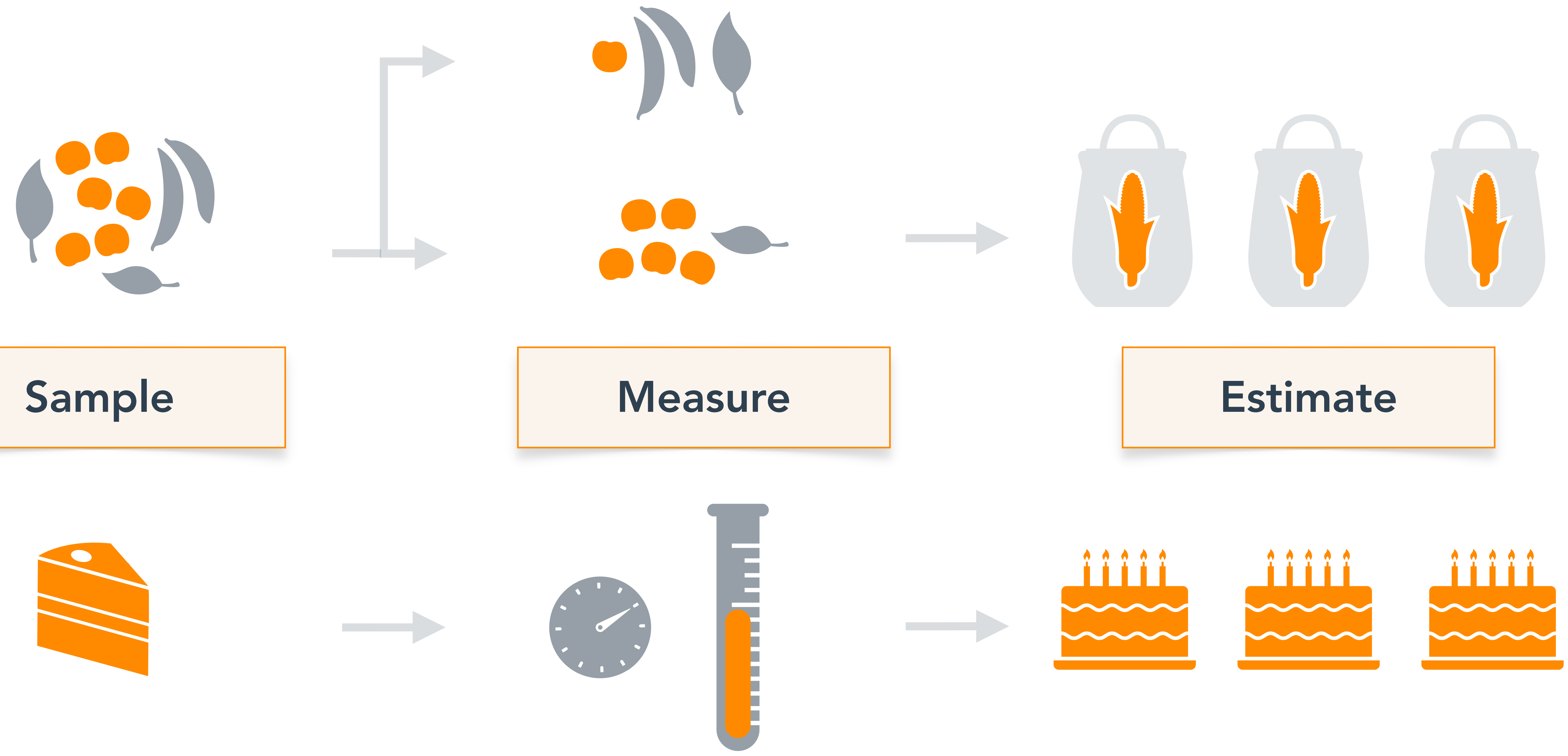
TEST SETS

- Only a sample is tested



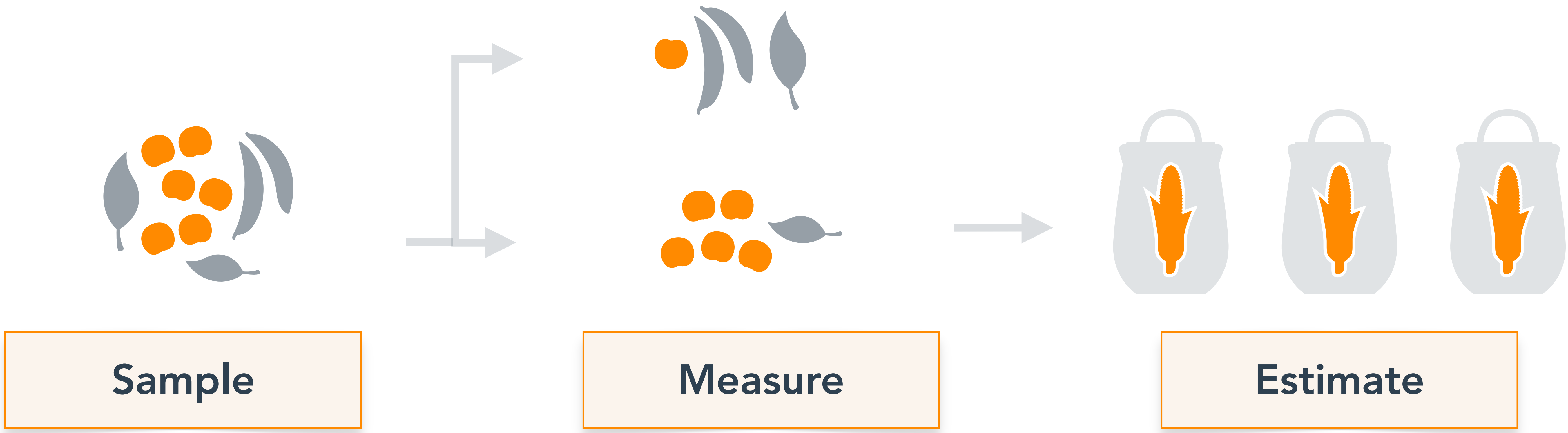
TEST SETS

- **Only a sample is tested** to estimate the errors in entire batches.



TEST SETS

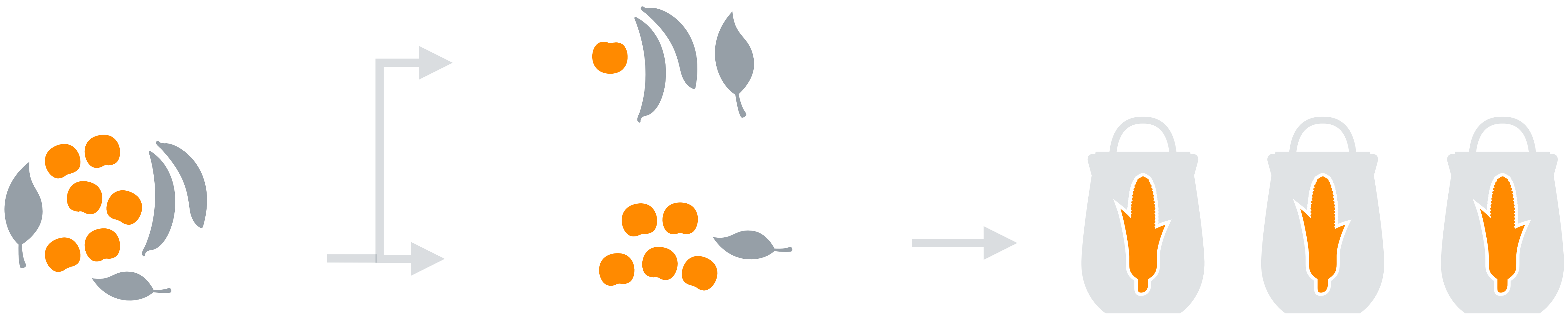
- **Only a sample is tested** to estimate the errors in entire batches.



"How many errors for this test set?"

TEST SETS

- **Only a sample is tested** to estimate the errors in entire batches.



Sample

“How many errors for this test set?”

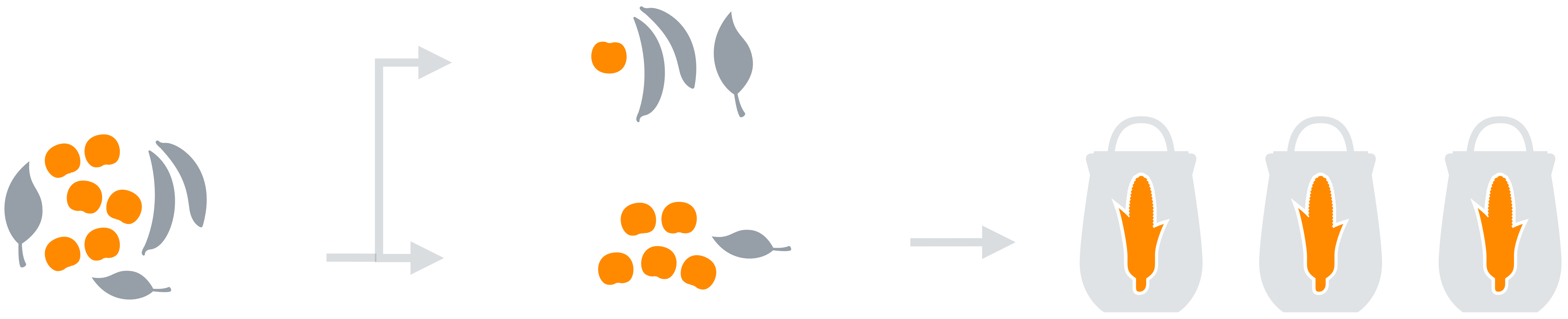
Measure

“Let's run the AI and count them.”

Estimate

TEST SETS

- **Only a sample is tested** to estimate the errors in entire batches.



Sample

"How many errors for this test set?"

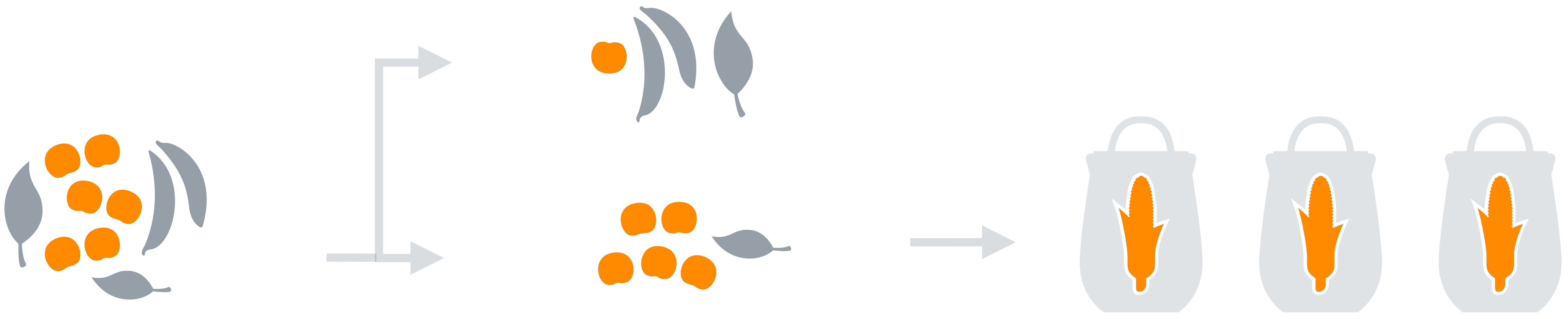
Measure

"Let's run the AI and count them."

Estimate

"So how many errors in this other set?"

TEST SET vs. TRAINING SET vs. TARGET SET



Sample

“How many errors for this test set?”

Measure

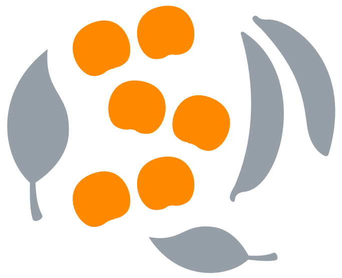
“Let's run the AI and count them.”

Estimate

“So how many errors in this other set?”

TEST SET vs. TRAINING SET vs. TARGET SET

- **Try the AI** with test sets.



Test set

"How many errors for this test set?"

Should be a random sample.

Measure

"Let's run the AI and count them."

Estimate

"So how many errors in this other set?"

TEST SET vs. TRAINING SET vs. TARGET SET

- **Try the AI** with test sets. **Make the AI model** with training sets.

Test set

"How many errors for this test set?"

Should be a random sample.

Training set

"Let's run the AI and count them."

May be a non-random sample.

Estimate

"So how many errors in this other set?"

TEST SET vs. TRAINING SET vs. TARGET SET

- **Try the AI** with test sets. **Make the AI model** with training sets. **Apply the AI** on target sets.

Test set

"How many errors for this test set?"

Should be a random sample.

Training set

"Let's run the AI and count them."

May be a non-random sample.

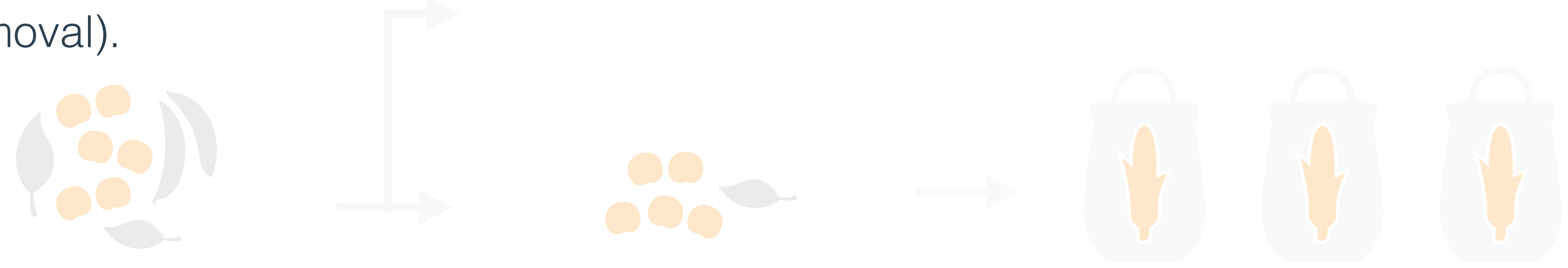
Target set

"So how many errors in this other set?"

May be a non-random sample.

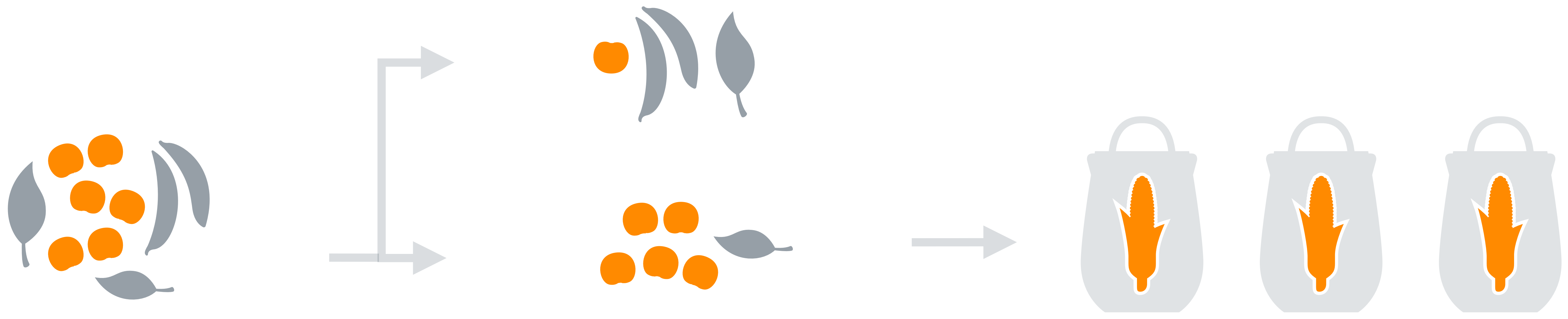
CHOOSING TEST & TRAINING SETS

- **Test sets are randomly sampled** to represent the target set. **Training sets may not**. AI models may work best if training sets are adjusted (e.g., downsampling or upsampling, outlier removal).



Test set	Training set	Target set
"How many errors for this test set?"	*"Let's run the AI and count them."*	*"So how many errors in this other set?"*
Should be a random sample.	May be a non-random sample.	May be a non-random sample.

VARIANCE IN PRACTICE



Test set

"How many errors for this test set?"

Should be a random sample.

Training set

"Let's run the AI and count them."

May be a non-random sample.

Target set

"So how many errors in this other set?"

May be a non-random sample.

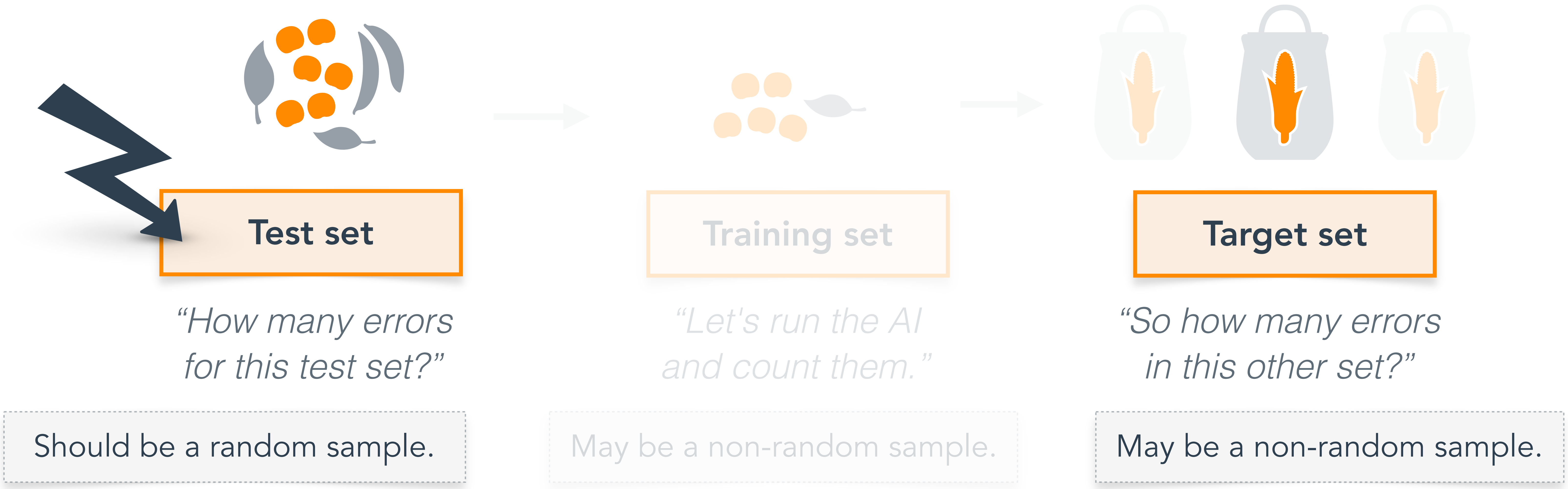
VARIANCE IN PRACTICE

- The **training set** is fixed.

Test set	Training set	Target set
"How many errors for this test set?"	*"Let's run the AI and count them."*	*"So how many errors in this other set?"*
Should be a random sample.	May be a non-random sample.	May be a non-random sample.

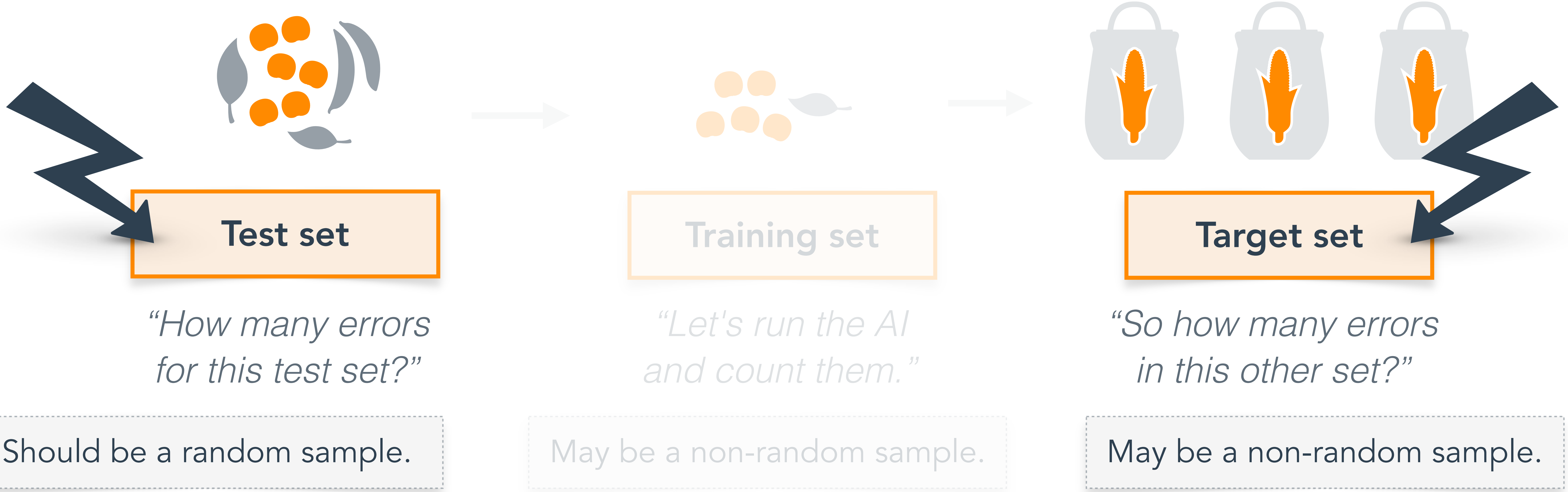
VARIANCE IN PRACTICE

- The **test set** may **differ** from the **target set** to assess.

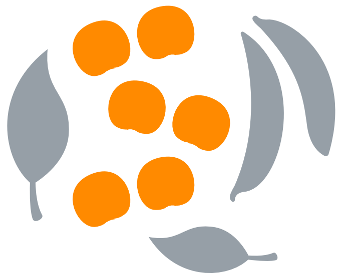


VARIANCE IN PRACTICE

- The **test set** may **differ** from the **target set** to assess.
- The **target sets** may also **differ among each other**.



RANDOM VARIANCE

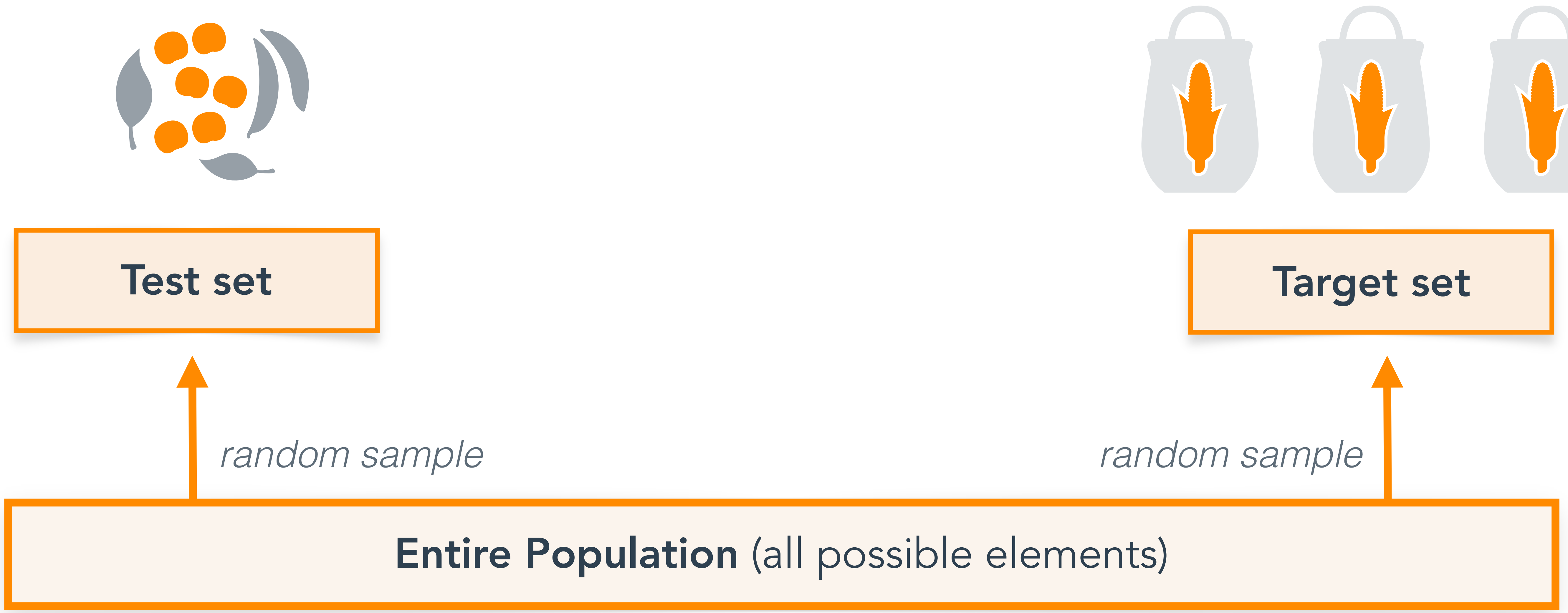


Test set

Target set

RANDOM VARIANCE

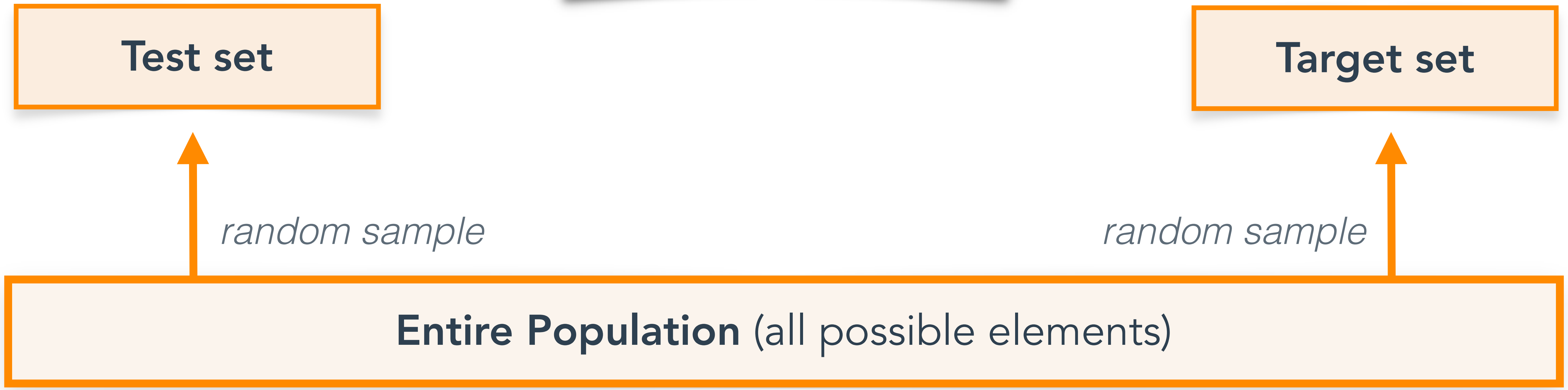
- Test and target sets are **random samples** from the same **population**.



RANDOM VARIANCE

- Error rates in random samples have **known variance** and distribution from **sampling theory** [3].

$$V(\boldsymbol{\theta}_{\boldsymbol{xy}}) = \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n_{x.}}$$



[3] Cochran, Sampling techniques. 1977.

RANDOM VARIANCE

- Error rates in random samples have **known variance** and distribution from **sampling theory** [3].
- **Smaller samples** give estimates with **higher variance**.

$$V(\theta_{xy}) = \frac{\theta^*_{xy}(1-\theta^*_{xy})}{n_{x.}}$$

test set size

Test set

Target set

random sample

random sample

Entire Population (all possible elements)

[3] Cochran, Sampling techniques (1977).

VARIANCE IN PRACTICE

- We use a test set to **estimate errors in a target set**.

VARIANCE IN PRACTICE

- We use a test set to **estimate errors in a target set**. These error estimates have added variance [4].

$$\widehat{V}\left(\widehat{\theta'_{xy}}\right) = \frac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}}$$

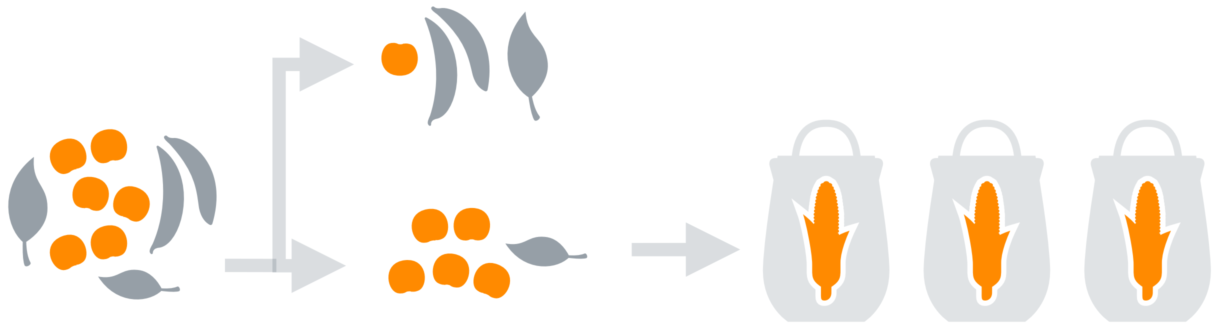
[4] Beauxis-Aussalet & Hardman, Extended Methods to Handle Classification Bias (2017).

VARIANCE IN PRACTICE

- We use a test set to **estimate errors in a target set**. These error estimates have added variance [4].

$$\widehat{V}\left(\widehat{\theta'_{xy}}\right) = \frac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}}$$

test set size



Test set

estimate errors in target set

Target set

random sample

random sample

Entire Population (all possible elements)

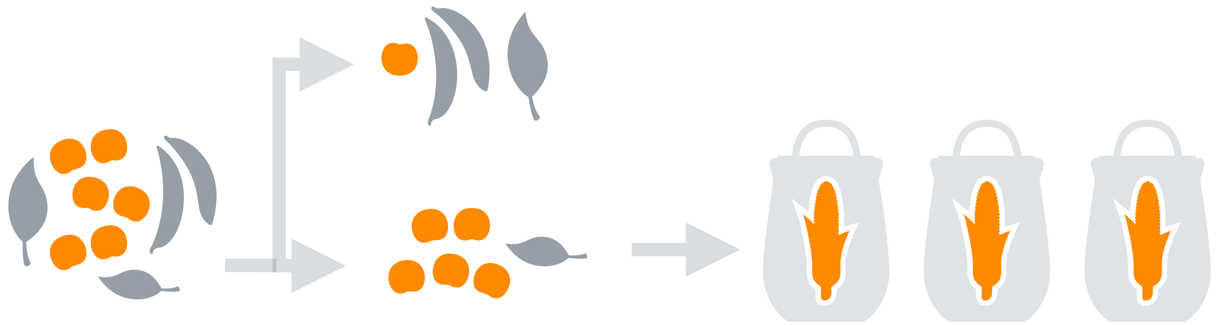
[4] Beauxis-Aussalet & Hardman, Extended Methods to Handle Classification Bias (2017).

VARIANCE IN PRACTICE

- We use a test set to **estimate errors in a target set**. These error estimates have added variance [4].

$$\widehat{V}\left(\widehat{\theta'_{xy}}\right) = \frac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}}$$

test set size *target set size*



Test set → *estimate errors in target set* → **Target set**

random sample *random sample*

Entire Population (all possible elements)

[4] Beauxis-Aussalet & Hardman, Extended Methods to Handle Classification Bias (2017).

VARIANCE IN PRACTICE

- We use a test set to **estimate errors in a target set**. These error estimates have added variance [4].
- **Smaller test or target sets** give estimates with **higher variance**.

$$\widehat{V}\left(\widehat{\theta'_{xy}}\right) = \frac{\theta_{xy}(1-\theta_{xy})}{n_{x.}} + \frac{\theta_{xy}(1-\theta_{xy})}{\widehat{n'_{x.}}}$$

test set size *target set size*

[4] Beauxis-Aussalet & Hardman, Extended Methods to Handle Classification Bias (2017).

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

CLASSIFICATION ERRORS

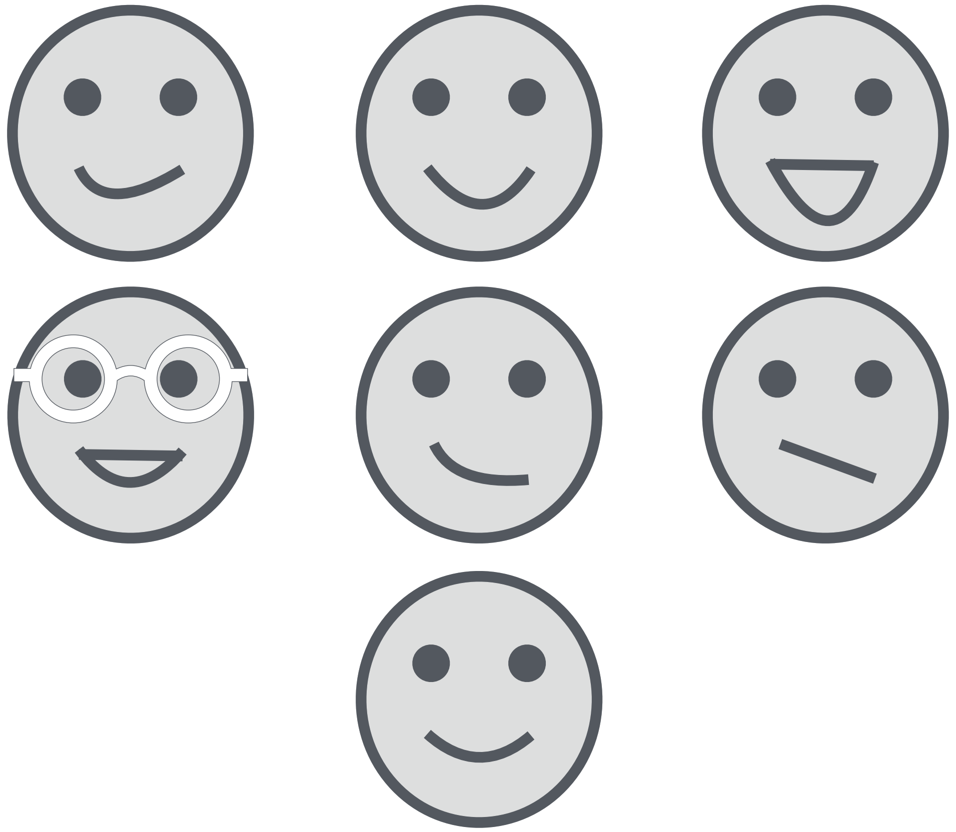
EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

CLASSIFICATION ERRORS

Test sets contain examples of correct classifications, and are used to measure the errors.

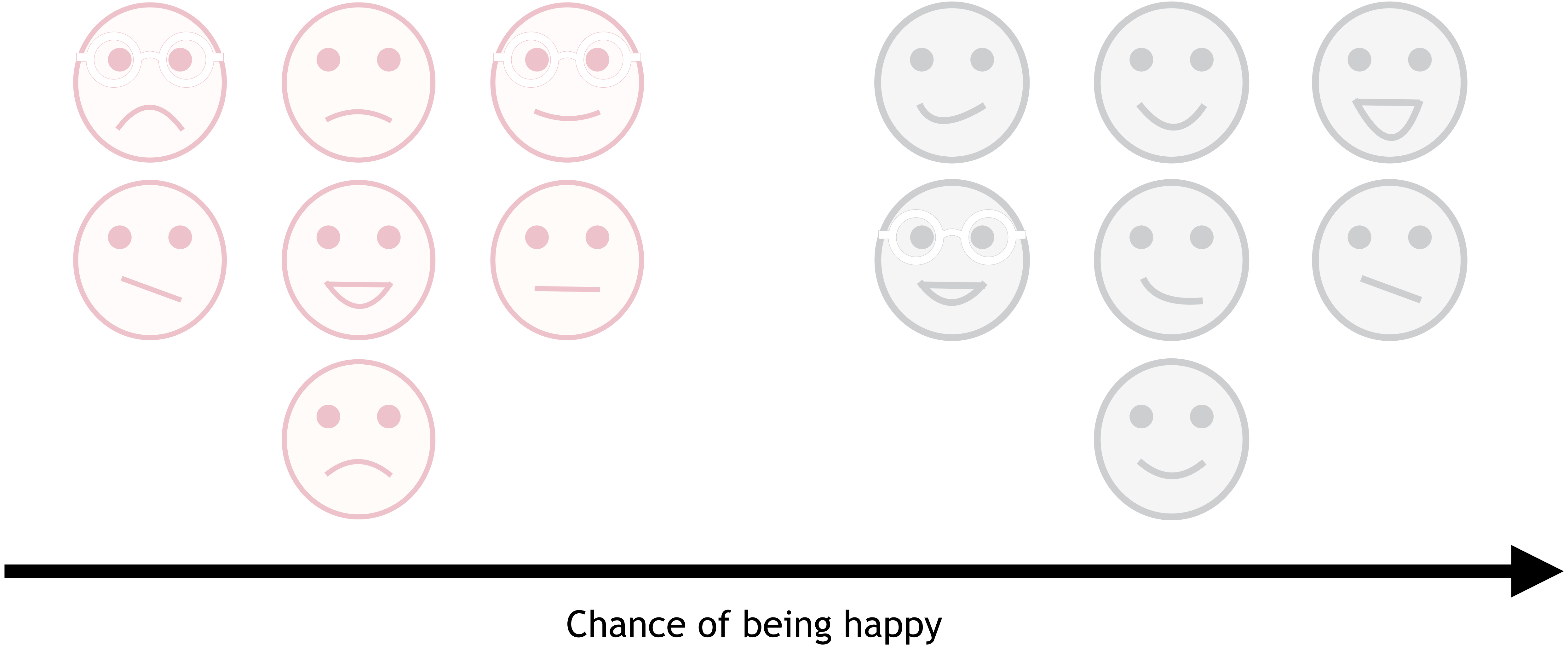
Examples of
Sad Faces



Examples of
Happy Faces

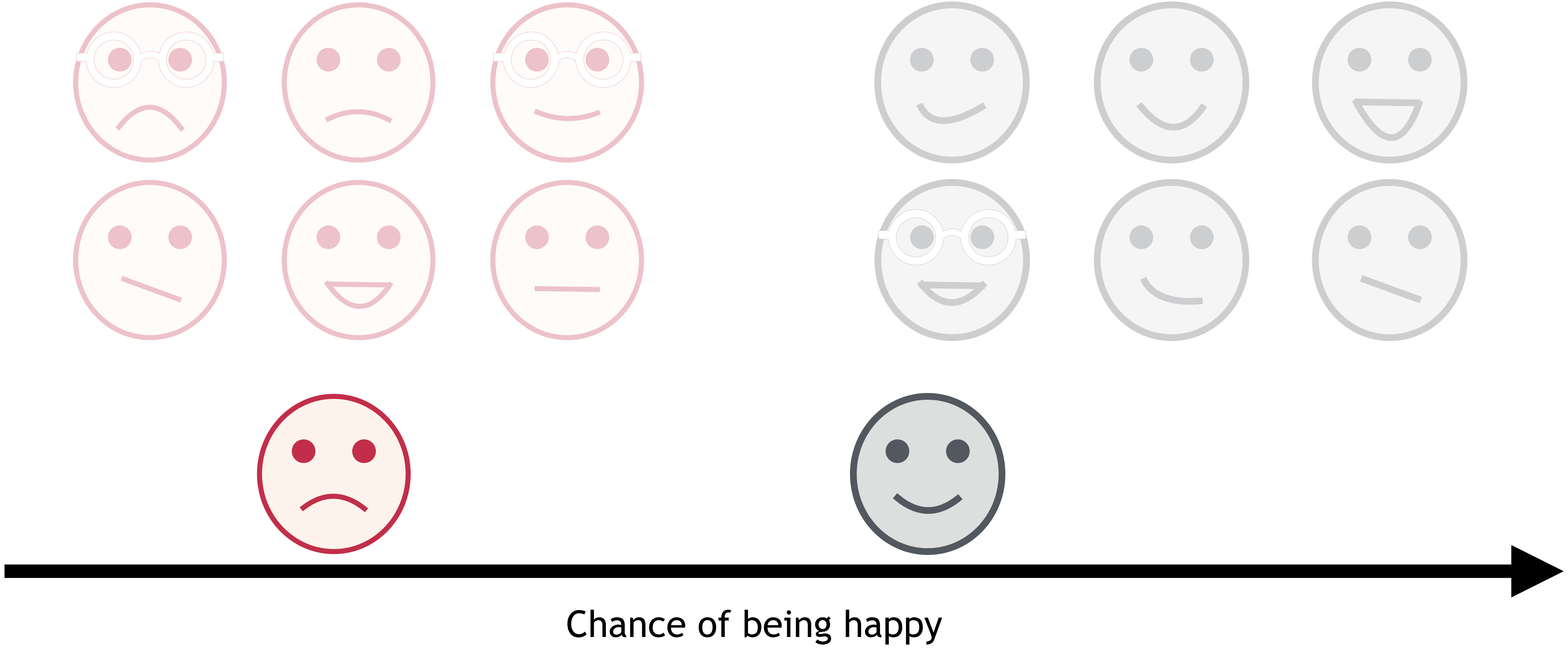
CLASSIFICATION ERRORS

Classifiers often have **tuning parameters**.



CLASSIFICATION ERRORS

Classifiers often have **tuning parameters**,
such as **thresholds** for separating the classes.

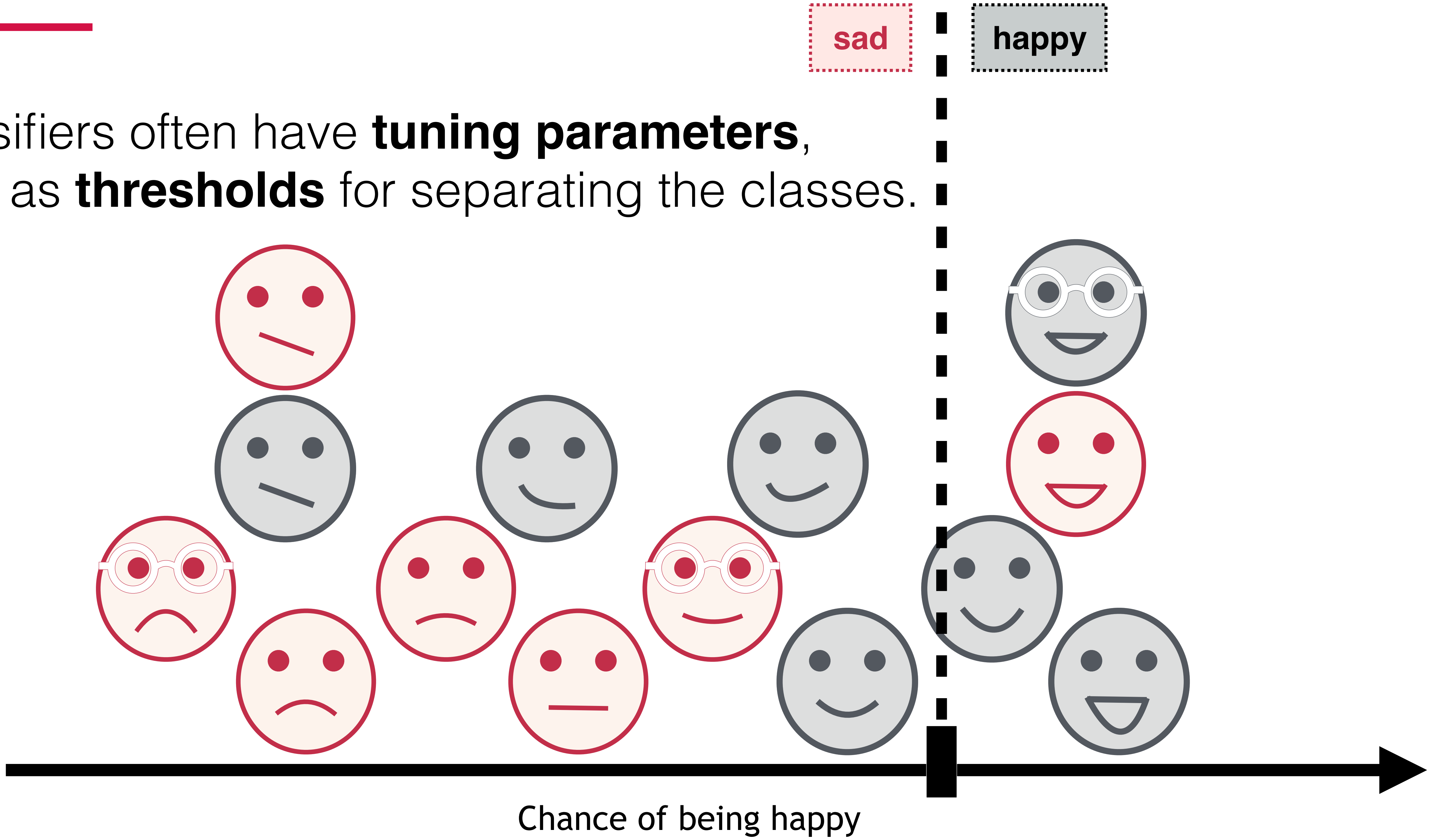


CLASSIFICATION ERRORS

Classifiers often have **tuning parameters**,
such as **thresholds** for separating the classes.

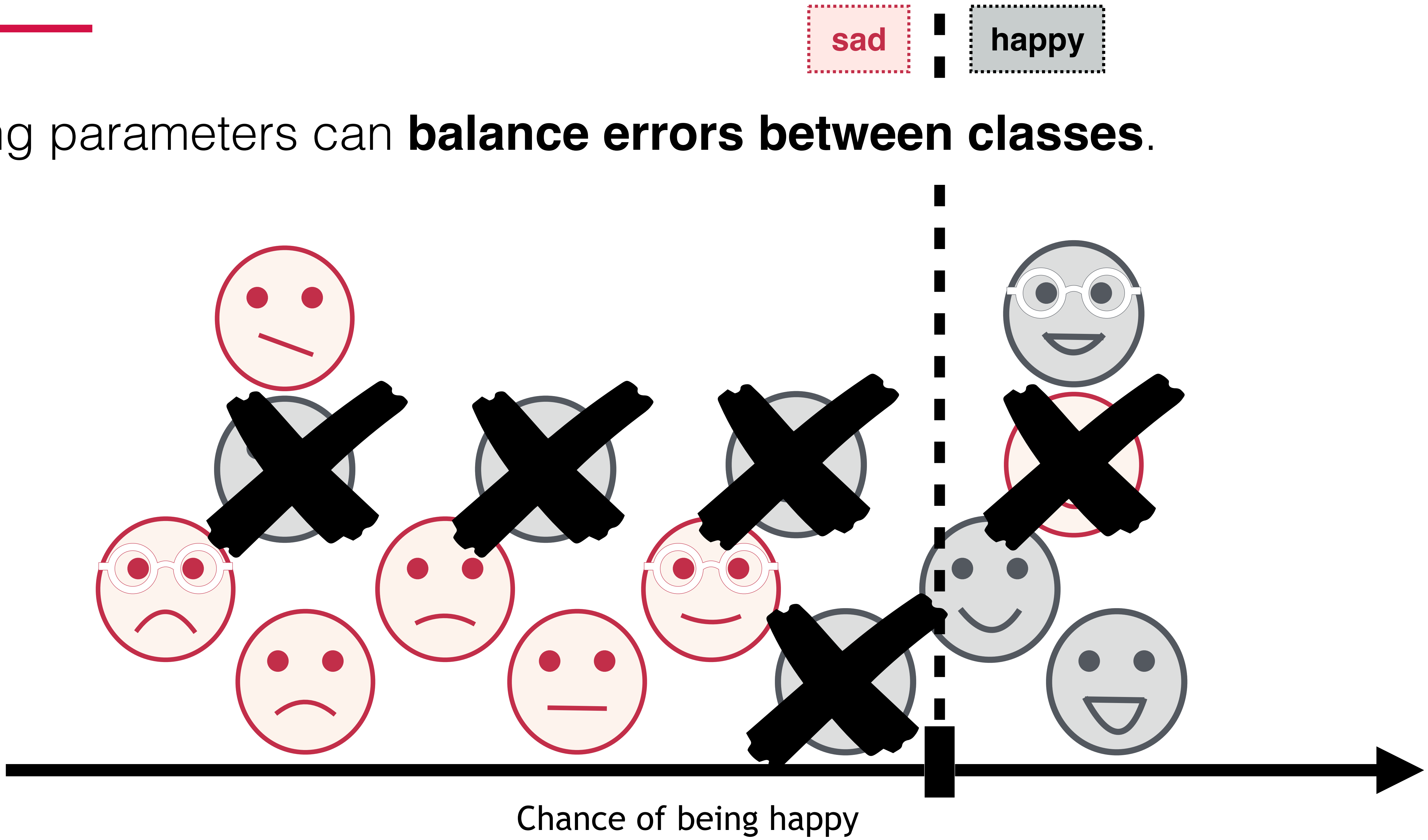
CLASSIFICATION ERRORS

Classifiers often have **tuning parameters**, such as **thresholds** for separating the classes.



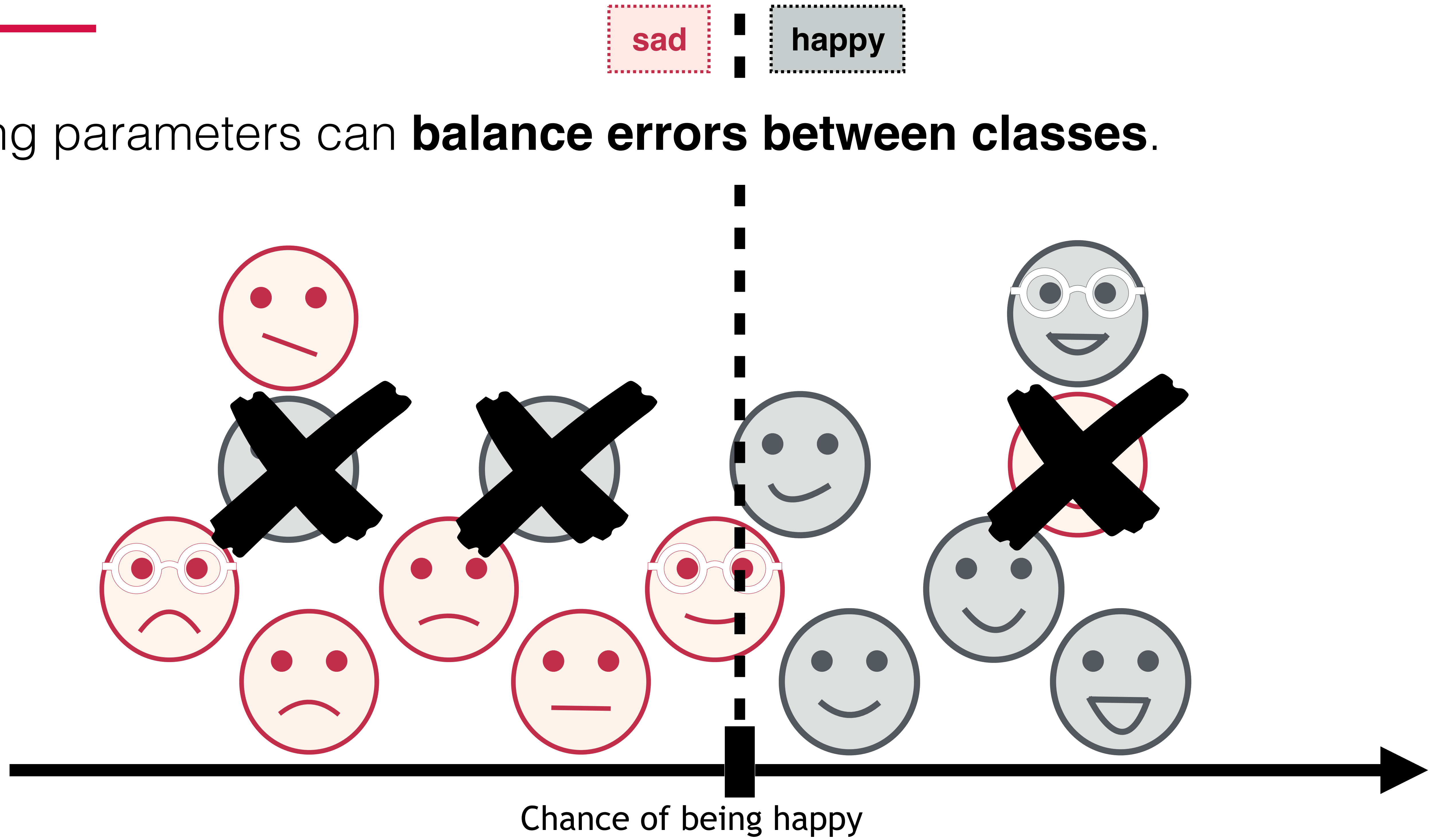
TUNING THE ERRORS

Tuning parameters can **balance errors between classes**.

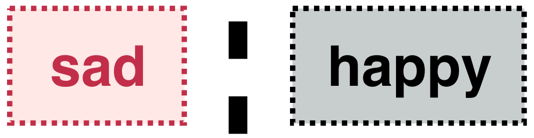


TUNING THE ERRORS

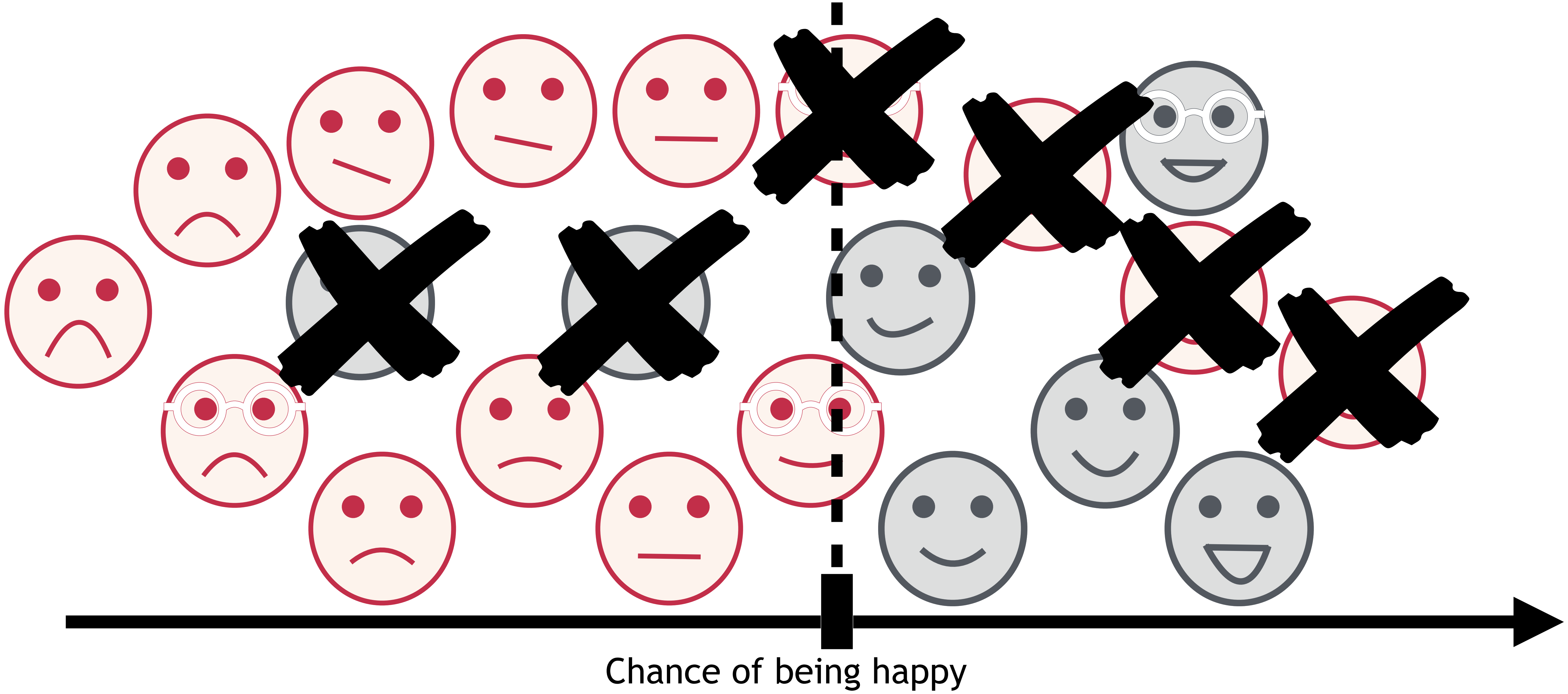
Tuning parameters can **balance errors between classes**.



TUNING THE ERRORS

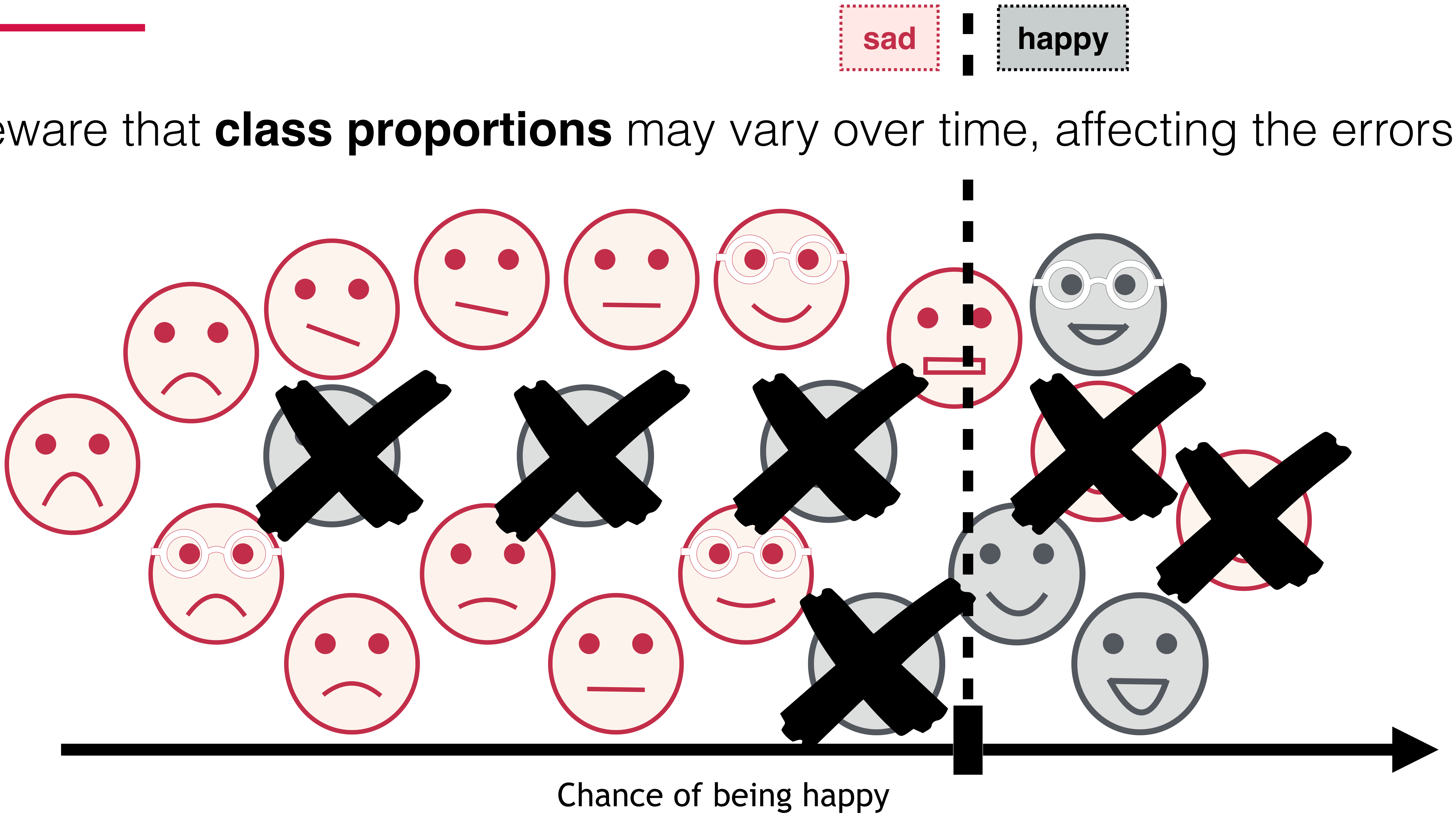


Beware that **class proportions** may vary over time, and affect the errors.



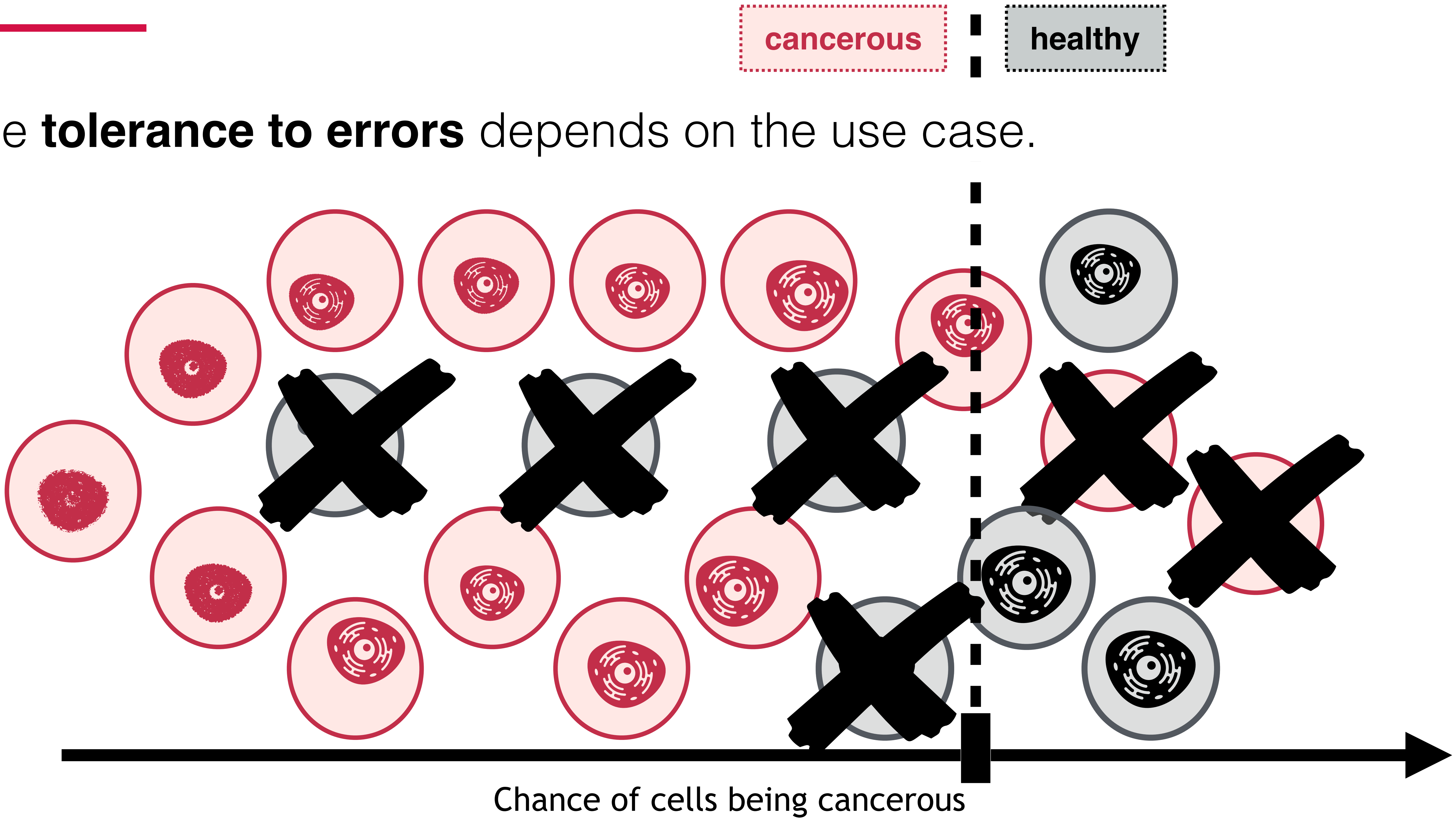
TUNING THE ERRORS

Beware that **class proportions** may vary over time, affecting the errors.



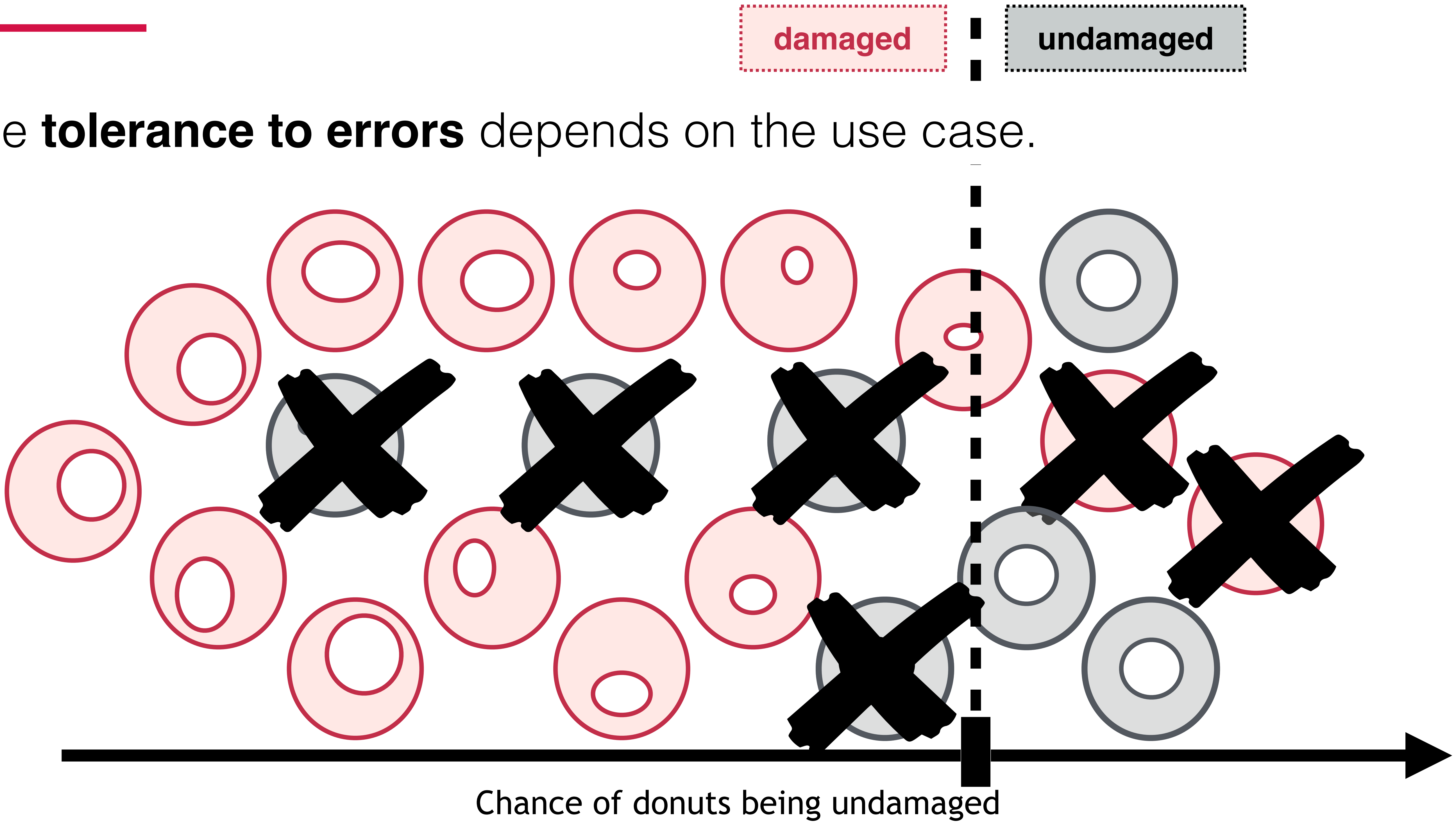
TUNING THE ERRORS

The **tolerance to errors** depends on the use case.



TUNING THE ERRORS

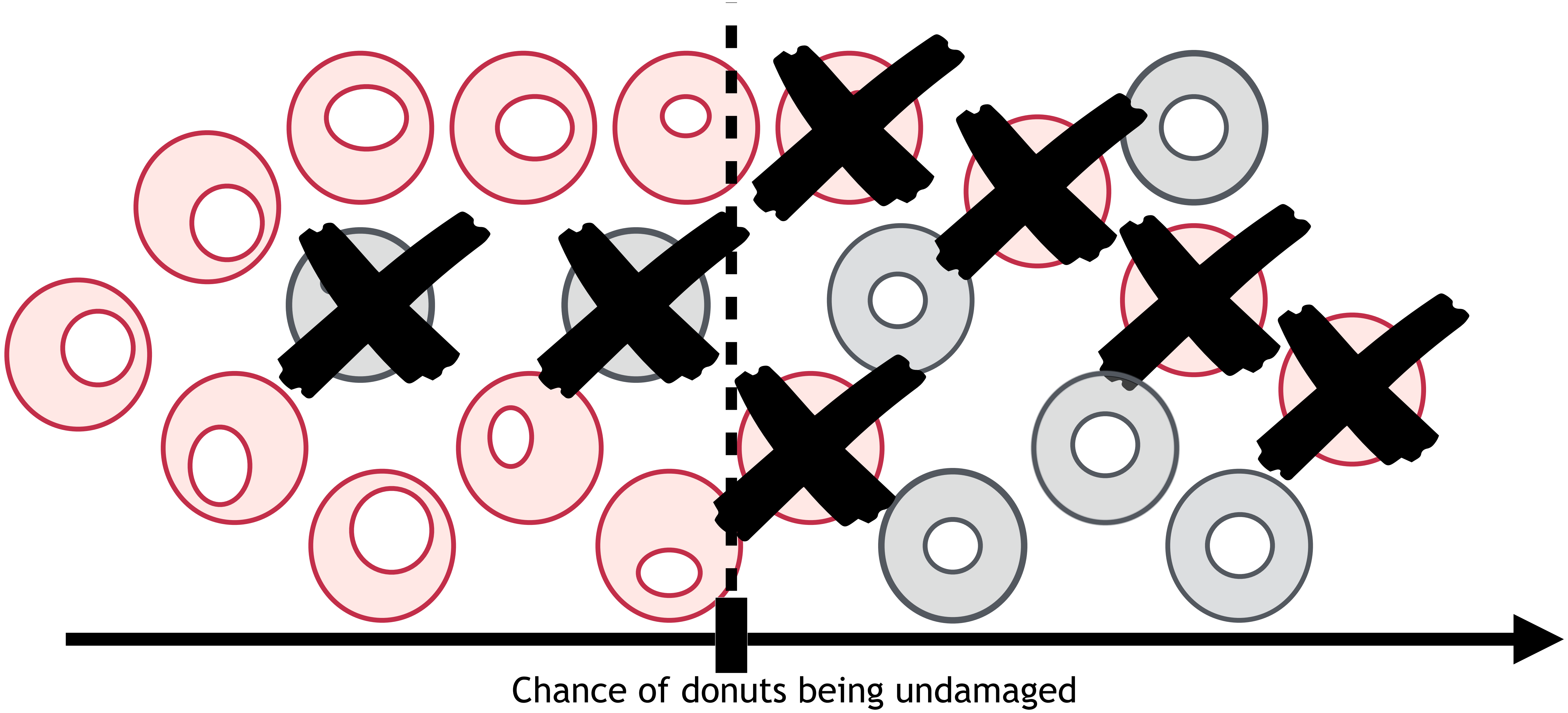
The **tolerance to errors** depends on the use case.



TUNING THE ERRORS

damaged | undamaged

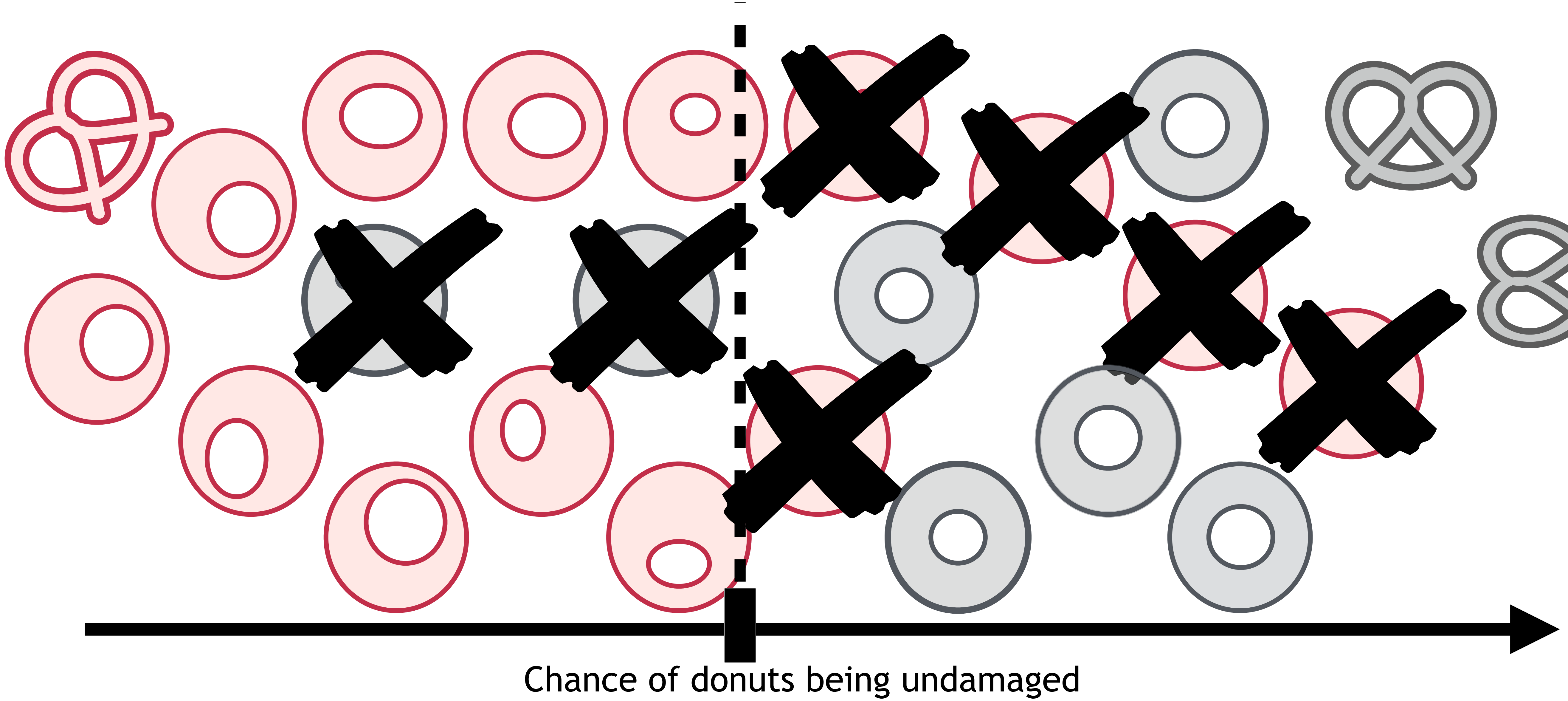
The **tolerance to errors** depends on the use case.



TUNING THE ERRORS

damaged | undamaged

Beware that **real life conditions** entail unexpected anomalies.



PRACTICAL ISSUES

Choices & tradeoffs are involved **at all steps** of the implementation.

- **Datasets** are only samples (outliers, biases, variability).
- **Tuning parameters** cannot optimise all real-life cases.
- **Error measurements** may be abstract, complex and incomplete.
- **Real-life conditions** may differ from the test conditions.

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

QUESTION / DISCUSSION