

# Adversarial examples, face verification

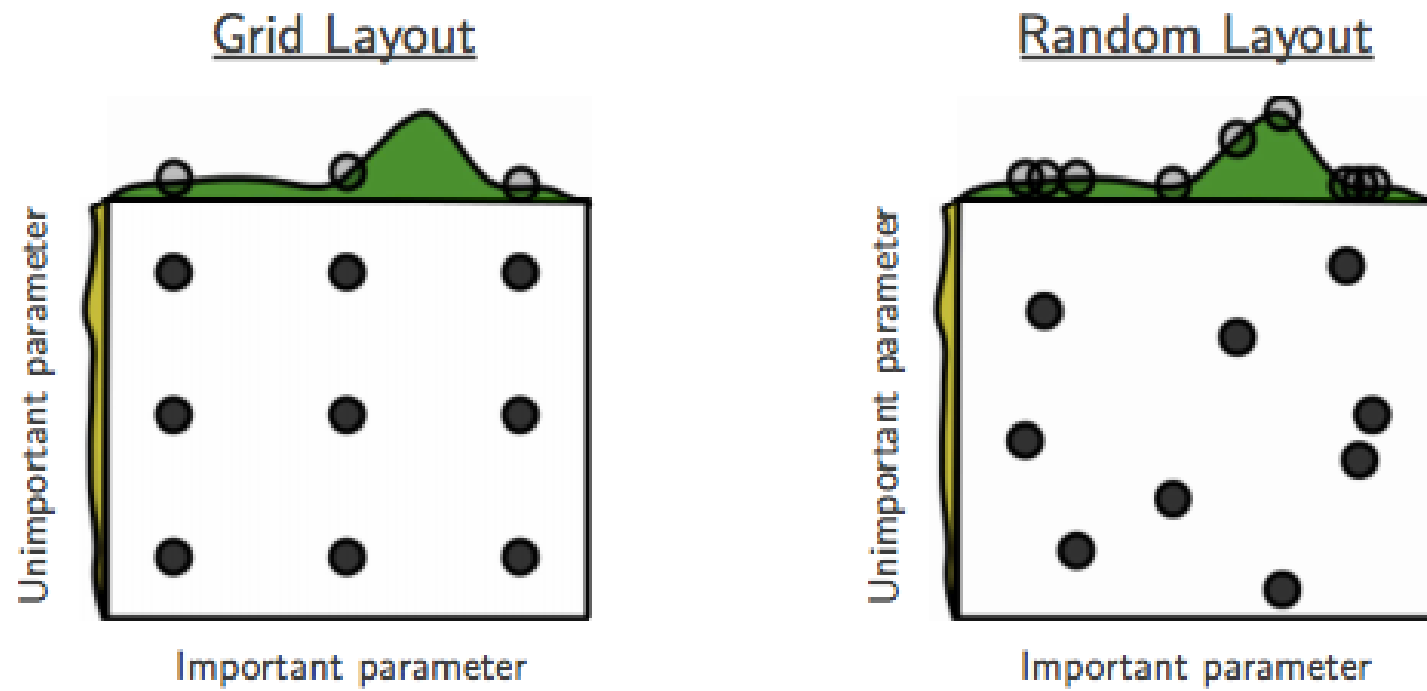
Bálint Ármin Pataki

# Hyperparameter search

**Hyperparameter:** the one are tuned by hand

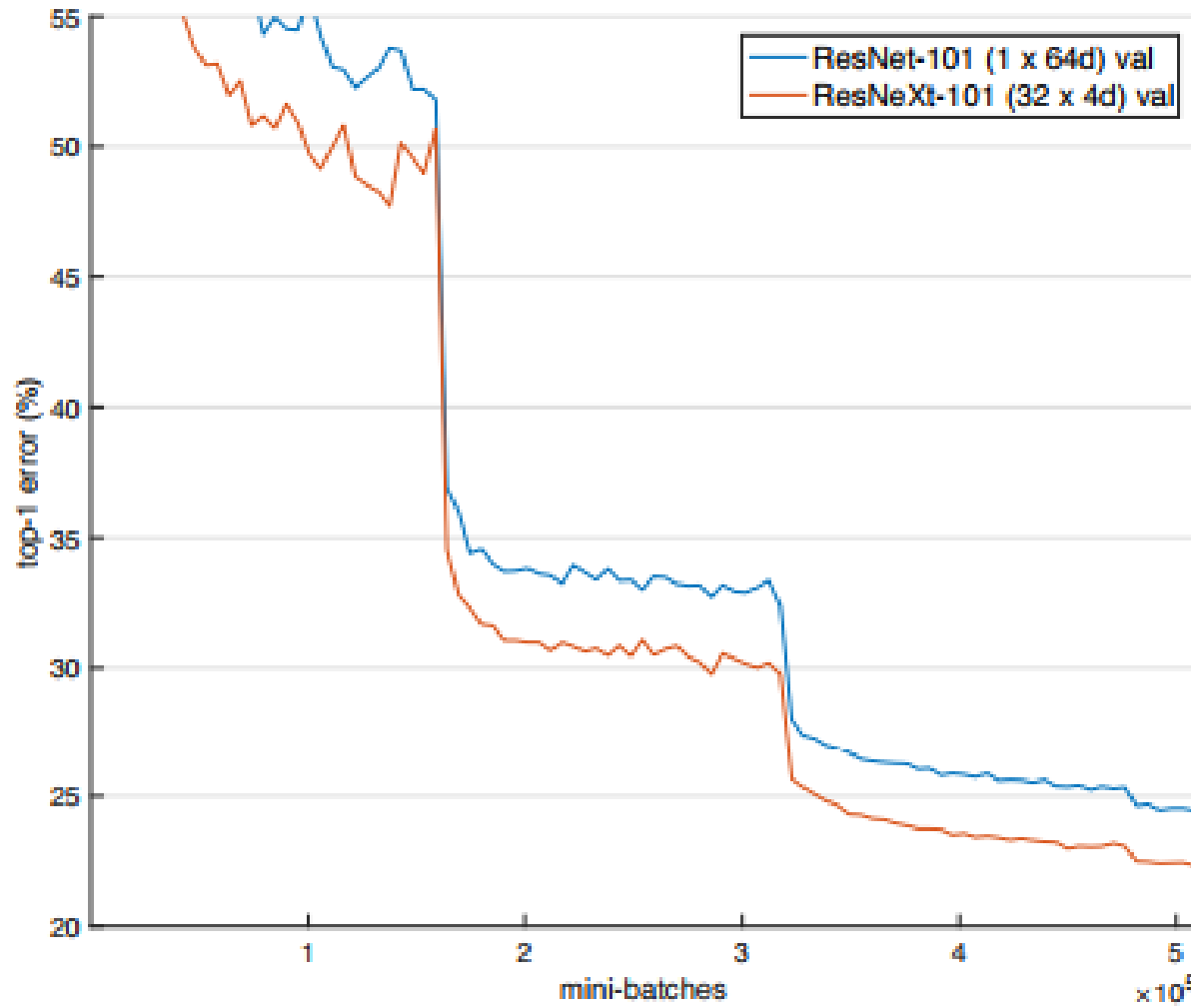
**Parameter:** the ones are tuned by the optimization algorithm

Diagrammatic representation of Grid Search by Bergstra & Bengio



# Learning rate is special

## Pre-defined schedule vs baby-sitting

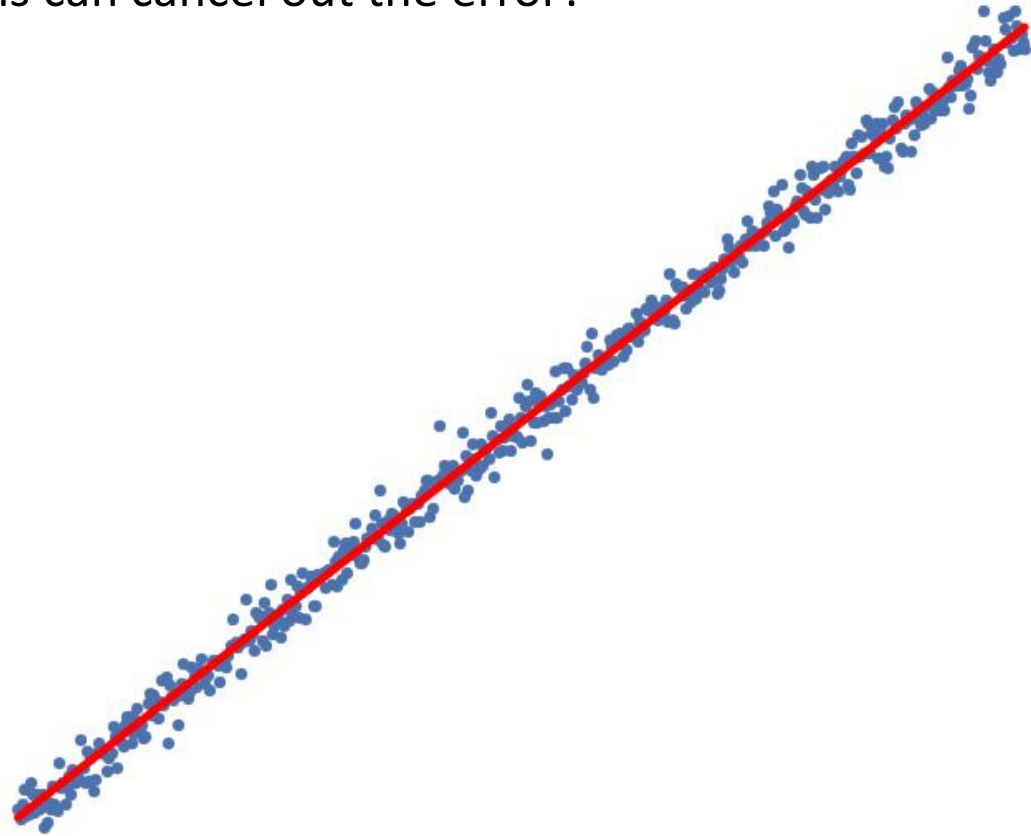


[Xie: Aggregated Residual Transformations for Deep Neural Networks arXiv:1611.05431v2]

# Ensemble - motivation

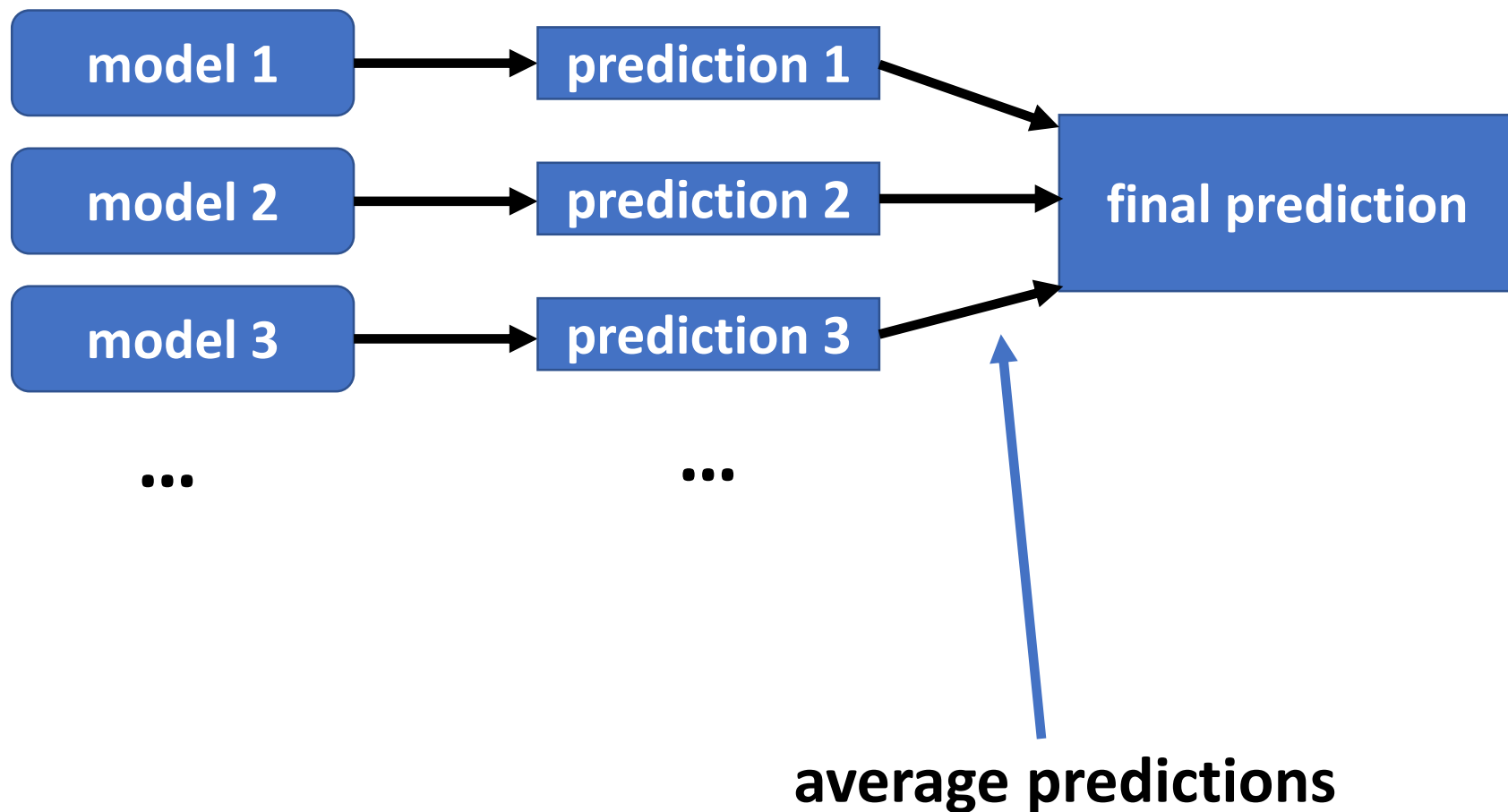
What if the error is random, but model dependent?

→ training different models can cancel out the error?



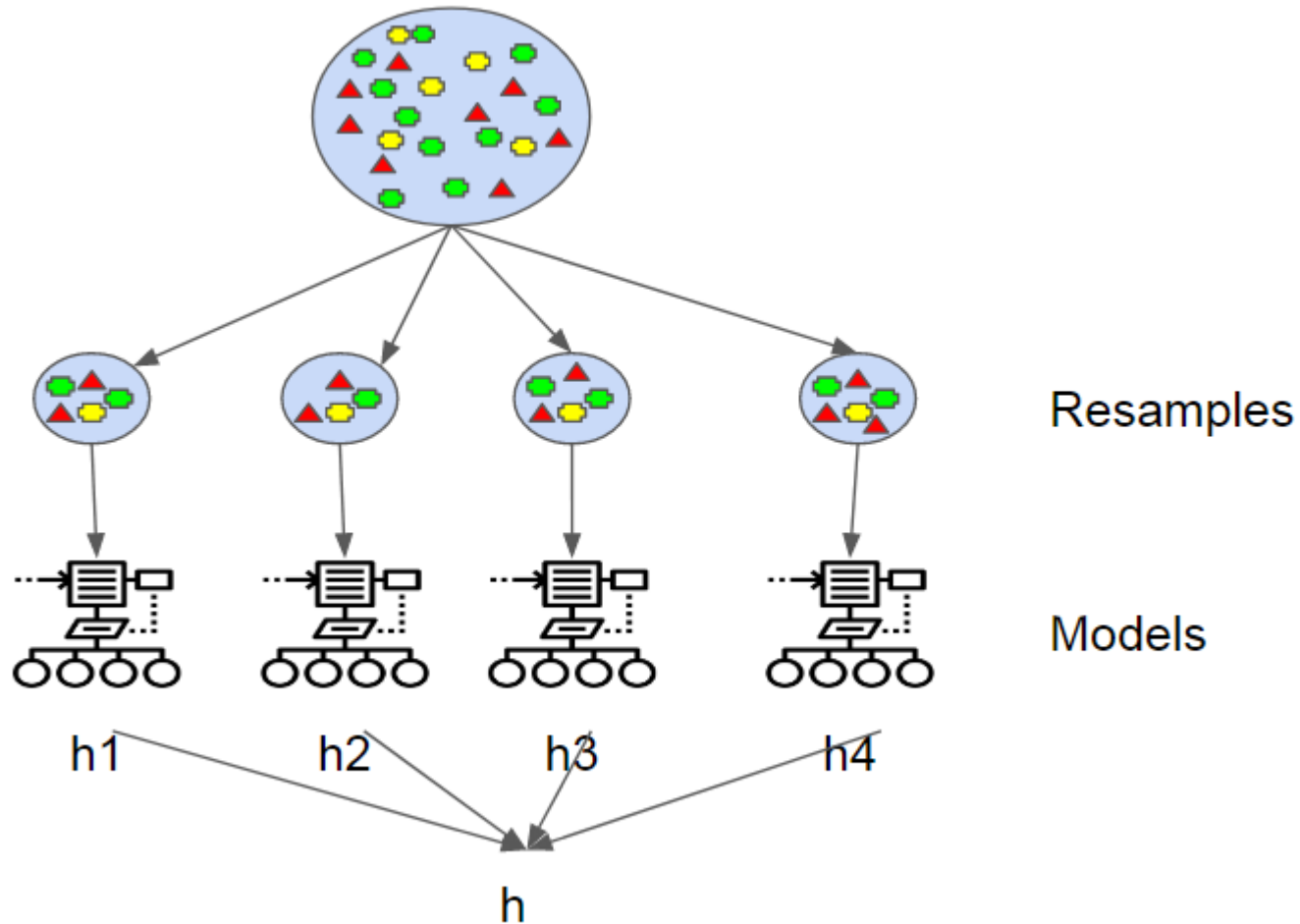
At Kaggle always an ensemble wins.

Often it has no practical relevance, but it can increase the score with epsilon.



Special case: Bagging (Bootstrap AGGREGatING)  
- same models trained on subset of train data

# Bagging



<https://medium.com/@SeattleDataGuy/how-to-develop-a-robust-algorithm-c38e08f32201>

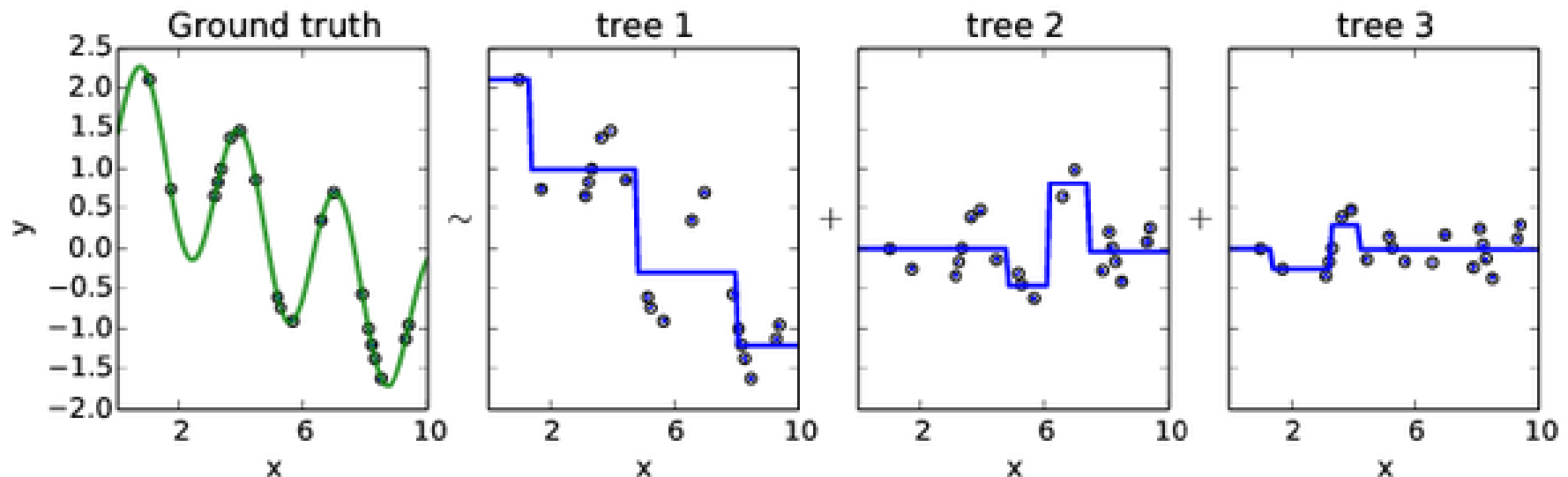
# Gradient boosting

Model 1 fits the original data.

Model 2 fits original data – model 1 prediction = the residual

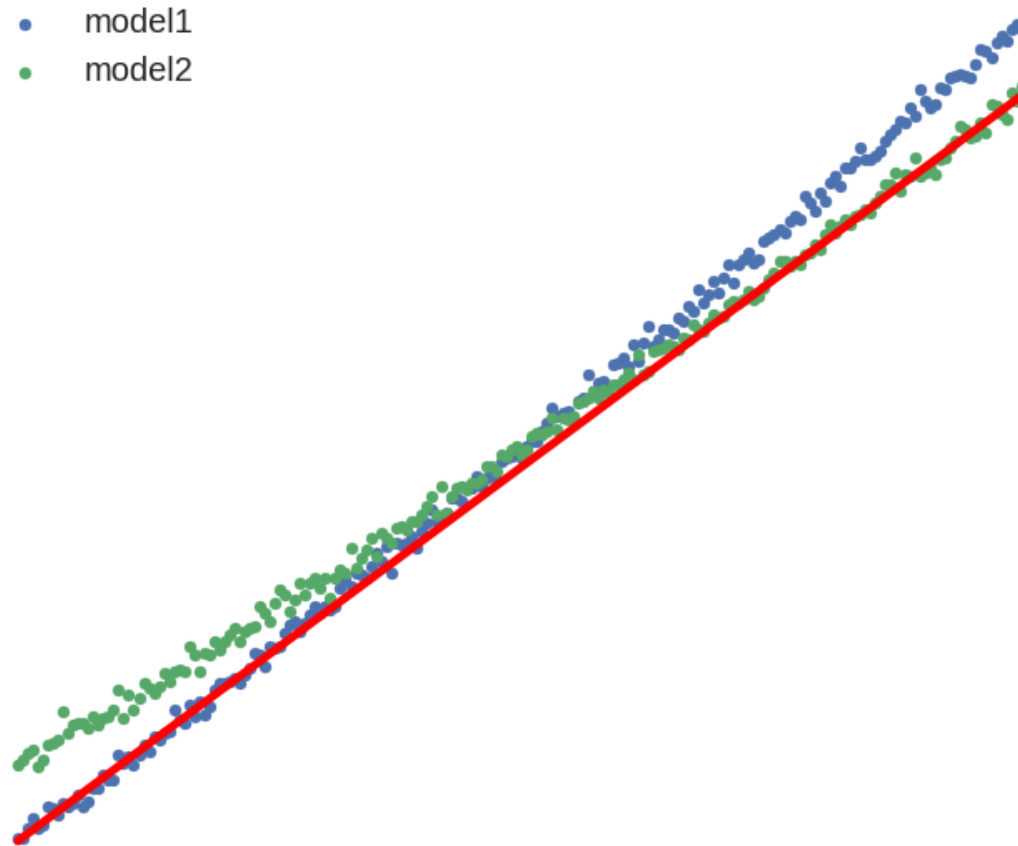
Model 3 fits original data – model 1 prediction – model 2 prediction

...



<https://www.quora.com/How-would-you-explain-gradient-boosting-machine-learning-technique-in-no-more-than-300-words-to-non-science-major-college-students>

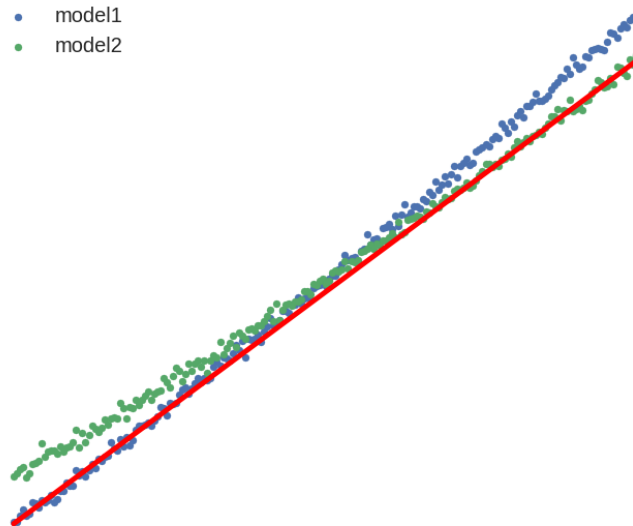
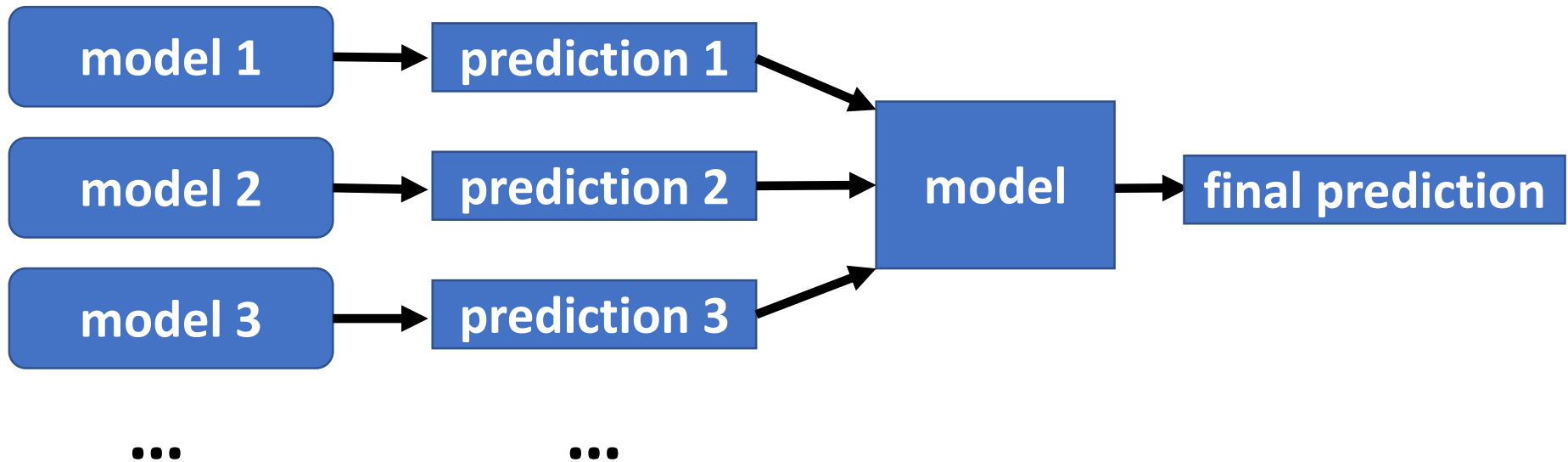
# Stacking - motivation



Model 1 is strong for smaller  $X$ , model 2 is for larger  $X$ .  
Would be great to combine them!



# Stacking



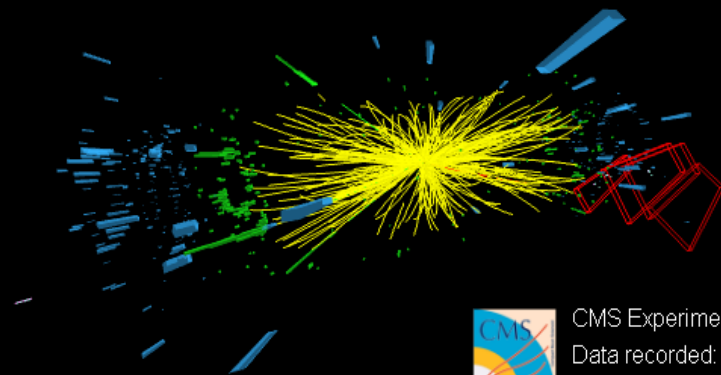
# Proper validation strategy



CMS Experiment at the LHC, CERN

Data recorded: 2012-May-07 07:46:20.384985 GMT

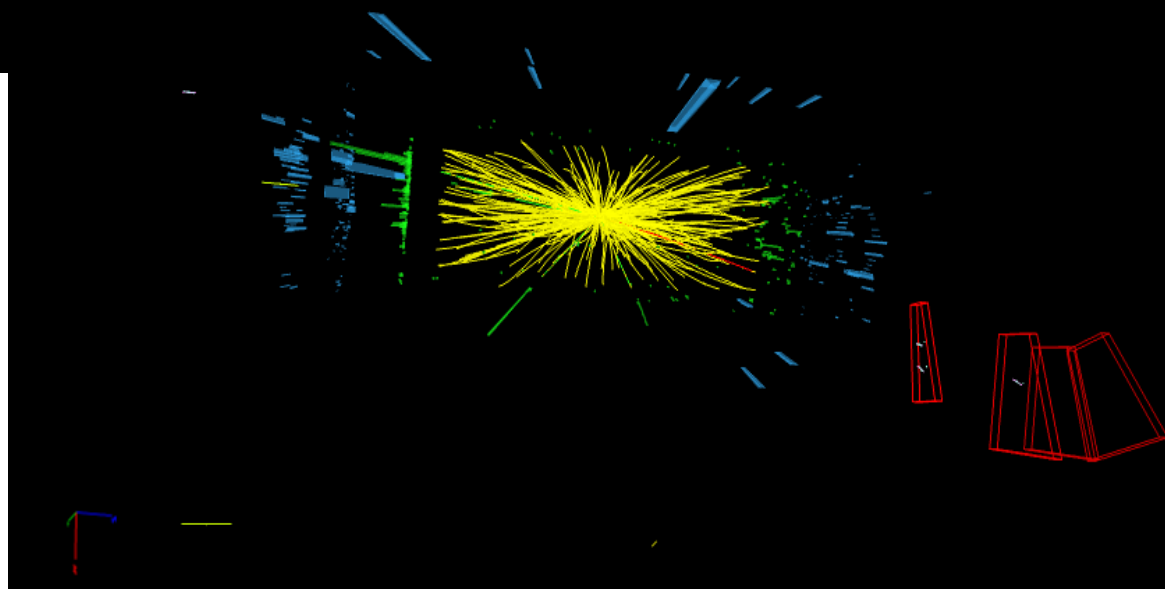
Run / Event / LS: 193575 / 400912970 / 523



CMS Experiment at the LHC, CERN

Data recorded: 2012-May-07 07:46:20.384985 GMT

Run / Event / LS: 193575 / 400912970 / 523



- **diff. view of same event**
- **diff. measurement of same person**
- **time series**
- ...

# DEMO notebook

# Other PPT

<http://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/presentation.pptx>

# Importance of the correct metric

<https://www.walesonline.co.uk/news/wales-news/facial-recognition-wrongly-identified-2000-14619145>

Facial recognition software wrongly identified more than 2,000 people as potential criminals as police patrolled the Champions League final in Cardiff.

The technology provided hundreds of “false positives” wrongly marking out innocent people as possible troublemakers when an estimated 170,000 people descended on the city for the showpiece match between Real Madrid and Juventus.

A South Wales Police spokesman admitted “no facial recognition system is 100% accurate under all conditions” but added that in the months since it was first deployed “no-one has been arrested where a ‘false positive alert’ has occurred and no members of the public have complained”.


Data published by the force showed police covering the Champions League final at the Principality Stadium on June 3 last year were alerted to 2,470 potential matches with custody pictures by the facial recognition programme.

But of these 92% – a total of 2,297 – were incorrect, with just 173 providing ‘true positive alerts’.

# Importance of the correct metric



## MNIST

- one vs all classifier (zero or not zero)
- metric: accuracy
- 90% accuracy. Is it good? 

# Instead of accuracy

Positive = predicted class

True = prediction is correct



True Positive (TP)	True Negative (TN)
False Positive (FP)	False Negative (FN)

Is this digit 0 (actually it is)?

Prediction: yes

→ True positive

Is this digit 0 (actually it is not)?

Prediction: no

→ True negative

Is this digit 0 (actually it is not)?

Prediction: yes

→ False positive

Is this digit 0 (actually it is)?

Prediction: no

→ False negative

# Instead of accuracy – AUC (ROC)

You predict a probability of being positive.

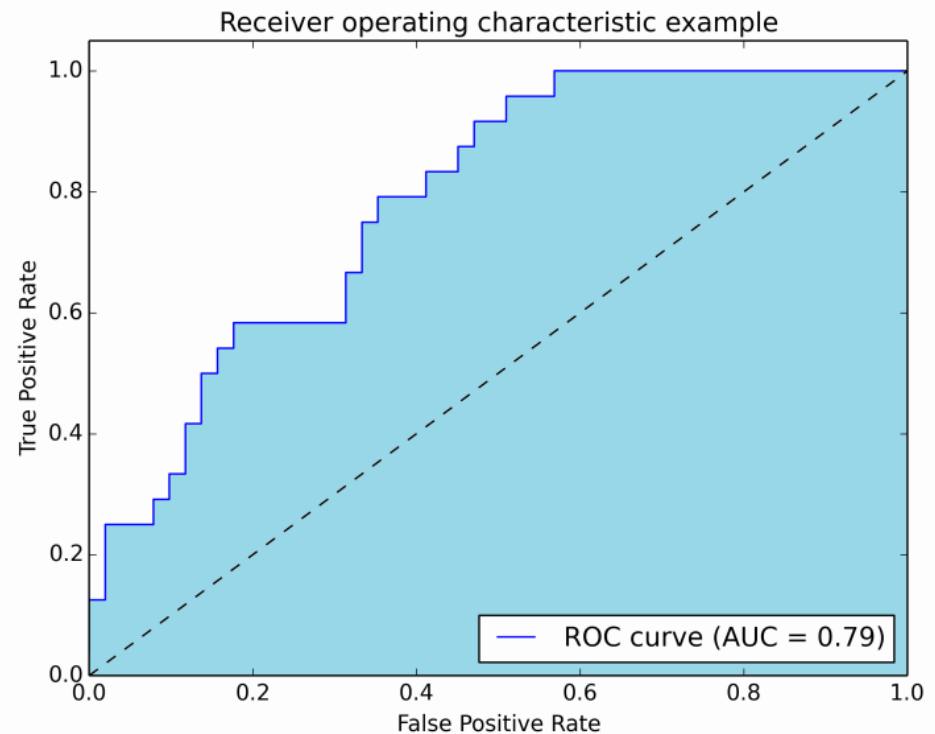
Then a threshold is applied (eq 50%) and if the probability is above, then prediction is positive

For different thresholds there is different prediction.

For different tasks you want different goals:

- identification of criminals: avoid False Negative
- unlock your phone with face recog: avoid False Positive

<http://www.navan.name/roc/>



<https://stats.stackexchange.com/questions/132777/what-does-auc-stand-for-and-what-is-it>



# Instead of accuracy – many more

Problem dependent. Objectives:

- easy to understand/interpret
- significantly better model should have significantly better score
- cover your exact need (as possible)

		True condition			
Total population		Condition positive	Condition negative	Prevalence = $\frac{\sum \text{Condition positive}}{\sum \text{Total population}}$	Accuracy (ACC) = $\frac{\sum \text{True positive} + \sum \text{True negative}}{\sum \text{Total population}}$
Predicted condition	Predicted condition positive	True positive, Power	False positive, Type I error	Positive predictive value (PPV), Precision = $\frac{\sum \text{True positive}}{\sum \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\sum \text{False positive}}{\sum \text{Predicted condition positive}}$
	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = $\frac{\sum \text{False negative}}{\sum \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\sum \text{True negative}}{\sum \text{Predicted condition negative}}$
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\sum \text{True positive}}{\sum \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\sum \text{False positive}}{\sum \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$  F <sub>1</sub> score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\sum \text{False negative}}{\sum \text{Condition positive}}$	True negative rate (TNR), Specificity (SPC) = $\frac{\sum \text{True negative}}{\sum \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)

# Instead of accuracy – many more and a few more

sensitivity, recall, hit rate, or true positive rate (TPR)

$$\text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

specificity or true negative rate (TNR)

$$\text{TNR} = \frac{\text{TN}}{N} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

precision or positive predictive value (PPV)

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

negative predictive value (NPV)

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

miss rate or false negative rate (FNR)

$$\text{FNR} = \frac{\text{FN}}{P} = \frac{\text{FN}}{\text{FN} + \text{TP}} = 1 - \text{TPR}$$

fall-out or false positive rate (FPR)

$$\text{FPR} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR}$$

false discovery rate (FDR)

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} = 1 - \text{PPV}$$

false omission rate (FOR)

$$\text{FOR} = \frac{\text{FN}}{\text{FN} + \text{TN}} = 1 - \text{NPV}$$

accuracy (ACC)

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{P + N} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

F1 score

is the harmonic mean of precision and sensitivity

$$F_1 = 2 \cdot \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Matthews correlation coefficient (MCC)

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

Informedness or Bookmaker Informedness (BM)

$$\text{BM} = \text{TPR} + \text{TNR} - 1$$

Markedness (MK)

$$\text{MK} = \text{PPV} + \text{NPV} - 1$$

Sources: Fawcett (2006), Powers (2011), and Ting (2011) <sup>[4]</sup> <sup>[1]</sup> <sup>[5]</sup>

[https://en.wikipedia.org/wiki/Precision\\_and\\_recall](https://en.wikipedia.org/wiki/Precision_and_recall)

# Importance of the correct metric

<https://www.walesonline.co.uk/news/wales-news/facial-recognition-wrongly-identified-2000-14619145>

Facial recognition software wrongly identified more than 2,000 people as potential criminals as police patrolled the **Champions League final in Cardiff**.

The technology provided hundreds of “false positives” wrongly marking out innocent people as possible troublemakers when an estimated 170,000 people descended on the city for the showpiece match between Real Madrid and Juventus.

A **South Wales Police** spokesman admitted “no facial recognition system is 100% accurate under all conditions” but added that in the months since it was first deployed “no-one has been arrested where a ‘false positive alert’ has occurred and no members of the public have complained”.

Data **published by the force** showed police covering the Champions League final at the Principality Stadium on June 3 last year were alerted to 2,470 potential matches with custody pictures by the facial recognition programme.

But of these 92% – a total of 2,297 – were incorrect, with just 173 providing ‘true positive alerts’.

<https://blog.openai.com/adversarial-example-research/>

<https://blog.openai.com/robust-adversarial-inputs/>

<http://www.navan.name/roc/>

[http://arogozhnikov.github.io/2016/07/05/gradient\\_boosting\\_playground.html](http://arogozhnikov.github.io/2016/07/05/gradient_boosting_playground.html)

<http://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/presentation.pptx>