Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Use Blackboard´s forum if a question may be relevant to other students, too. Email always both joeran.beel@scss.tcd and doug.leith@scss.tcd.ie. Give a meaningful subject, starting with "*[ML1819]*". No file attachments.

Week 05 & 06: Machine Learning Training & Evaluation

CS7CS4/CS4404 Machine Learning
v4 2018-10-16

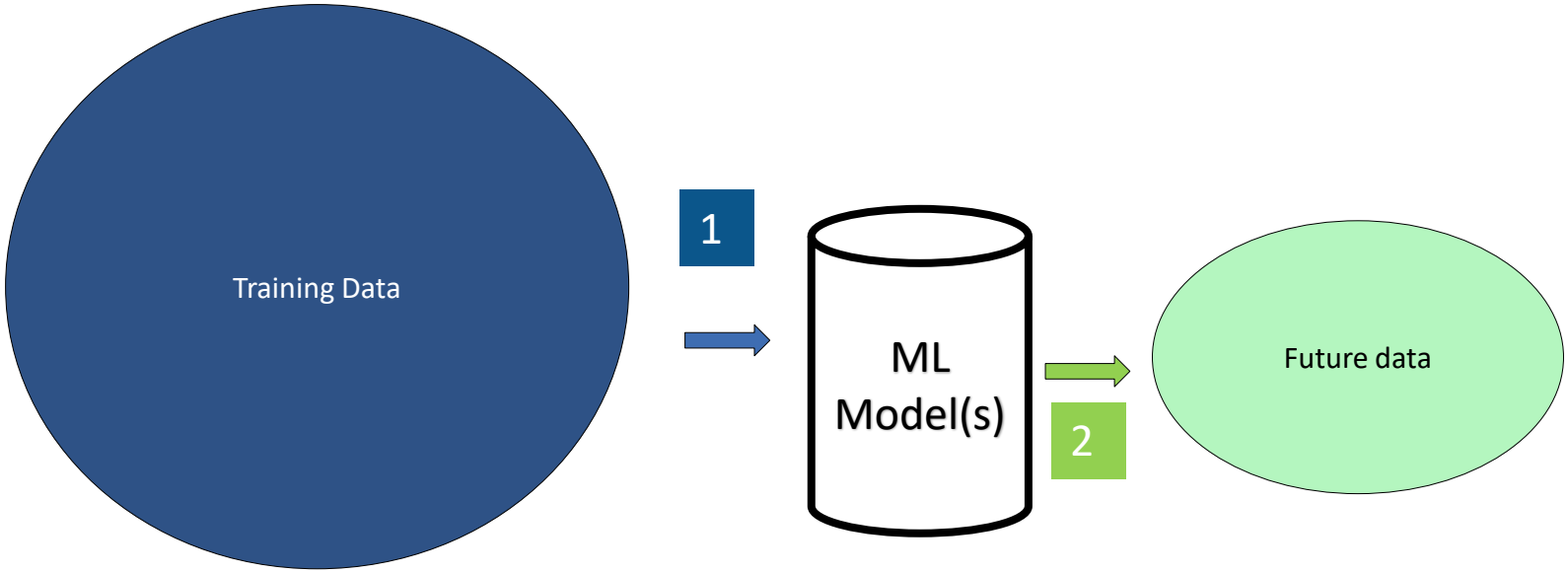
Dr Joeran Beel
Assistant Professor in Intelligent Systems
Department of Computer Science and Statistics
Trinity College Dublin, Ireland

Dr Douglas Leith
Professor in Computer Systems
Department of Computer Science and Statistics
Trinity College Dublin, Ireland

Basic idea of (Supervised) Machine Learning

1. **Build a model based on historic data (that is as effective as possible for future data)**
2. **Apply the model to future data**
3. **Major questions**

1. How to maximize performance of your model?
2. How good is the model?



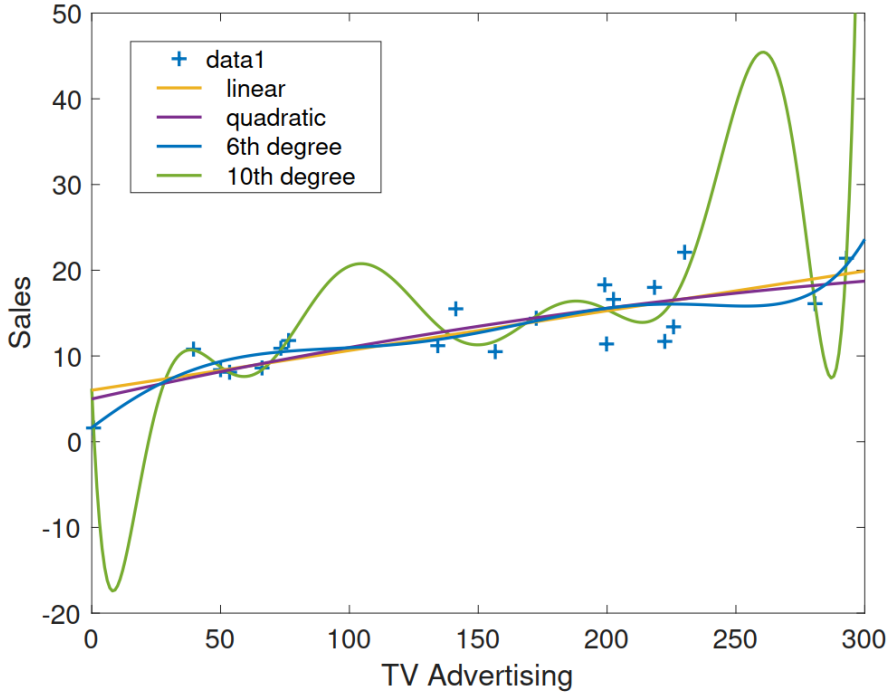
Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Training & Testing

How to maximize the performance of your model?

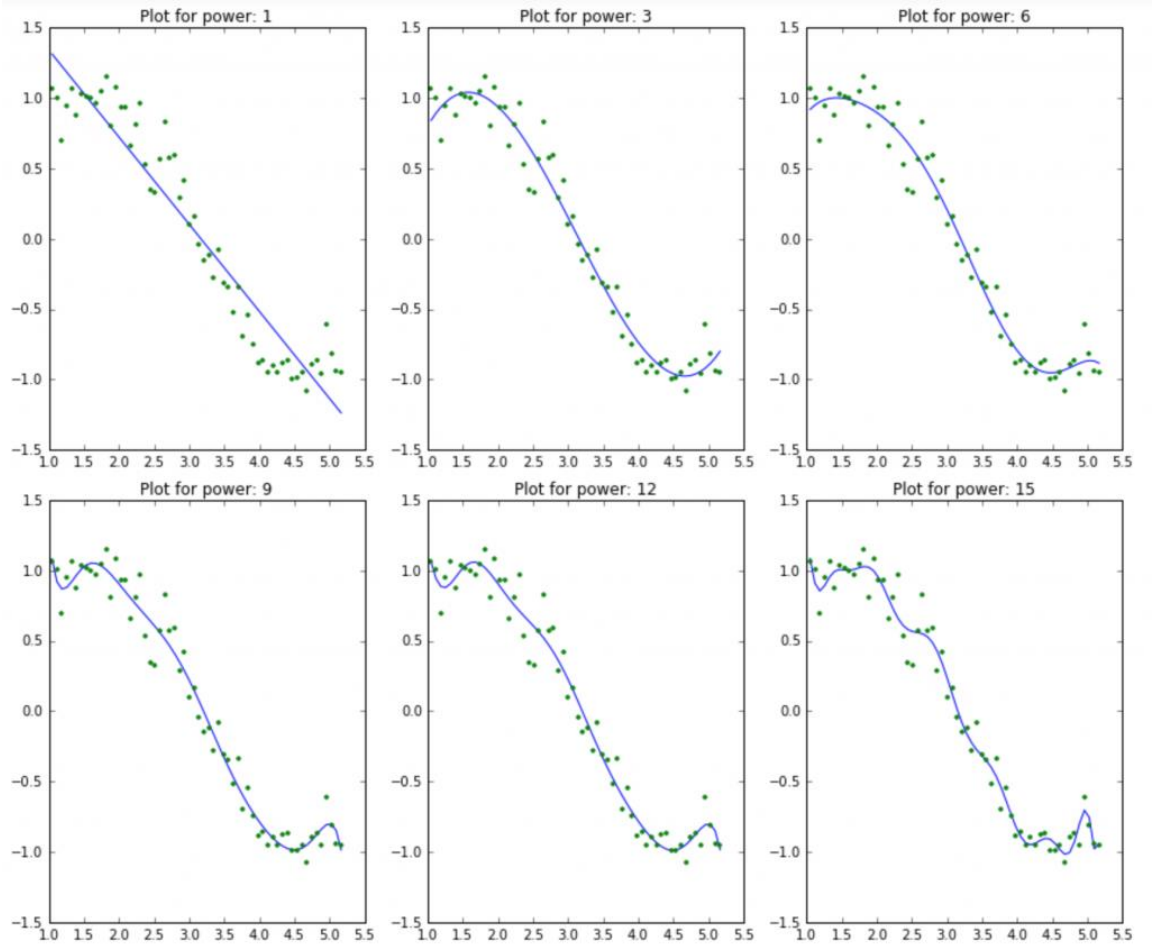
Hyperparameters

- **Degree of freedom (number of variables e.g. in regression)**
- **Number of clusters in k-nearest neighbour**
- **Batch Size**
- **Number of hidden layers in a deep neural network**
- **Features to use**
- **…**



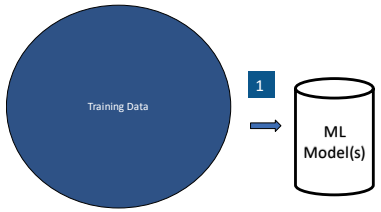
Impact of Degree of Freedom (Polynomial Regression)

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + \epsilon.$$

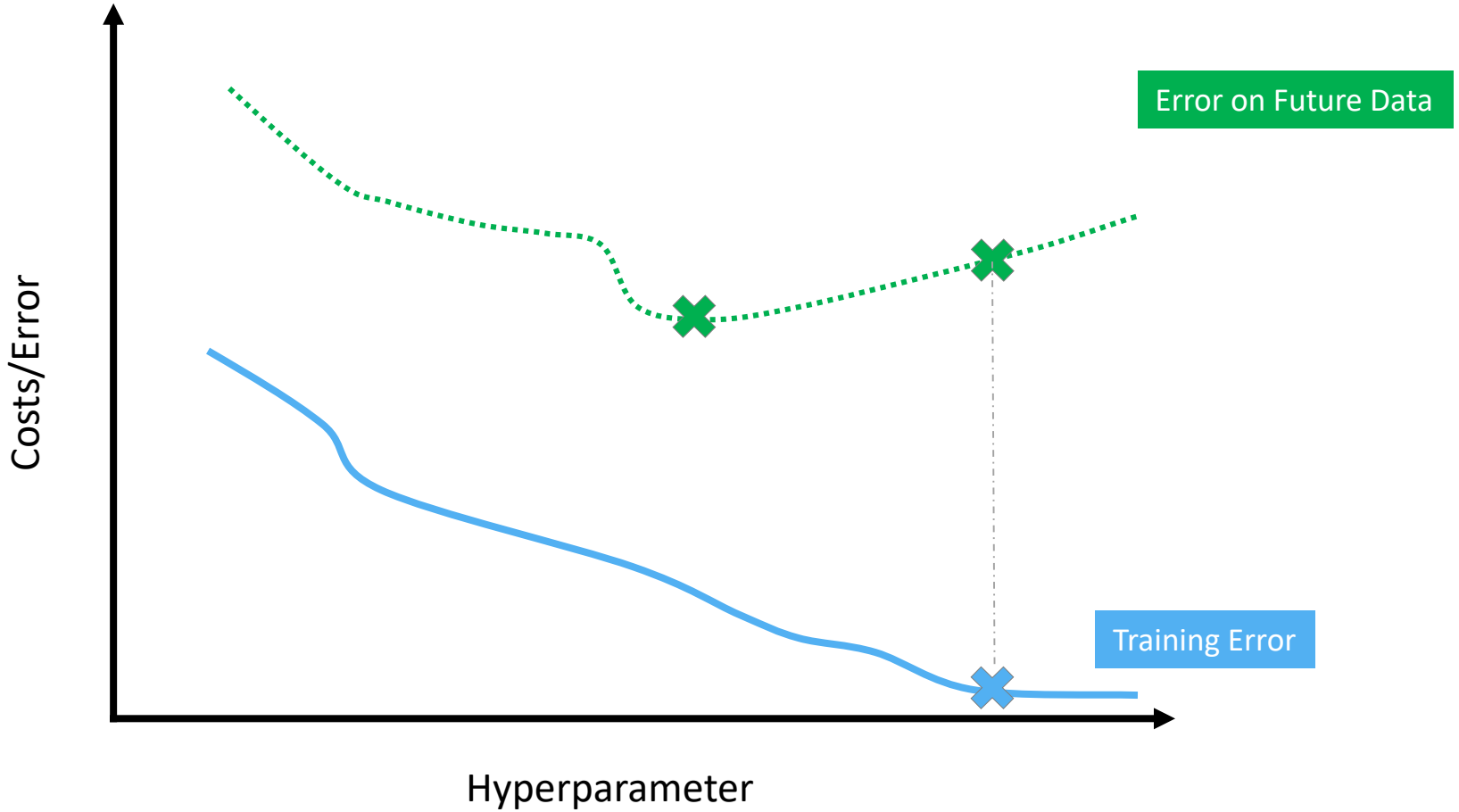


https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/
The R Software: Fundamentals of Programming and Statistical Analysis By Pierre Lafaye de Micheaux, Rémy Drouilhet,

Basic Concept

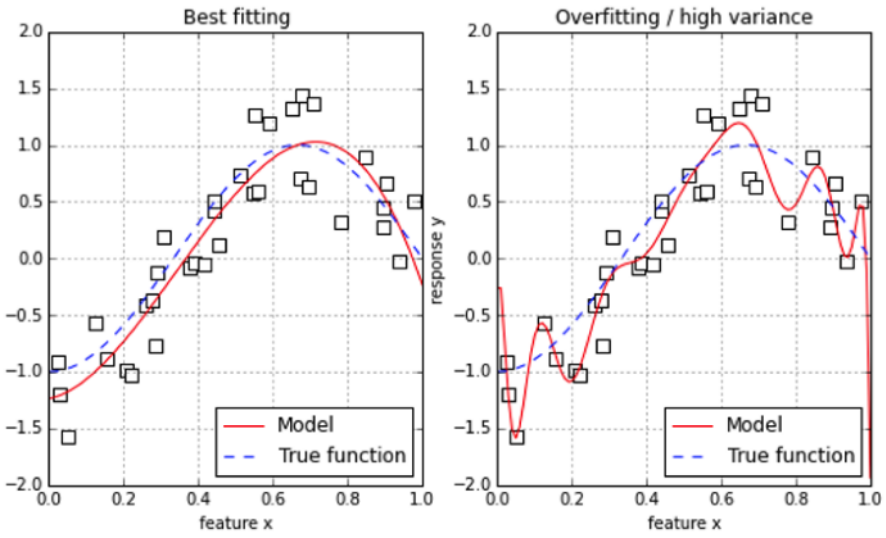


- **Experiment with hyperparameters until error is minimized**
- **Ideally, you would want to train on a test set, and evaluate on real/future data (and stop once the minima on the real data is achieved)**



Overfitting

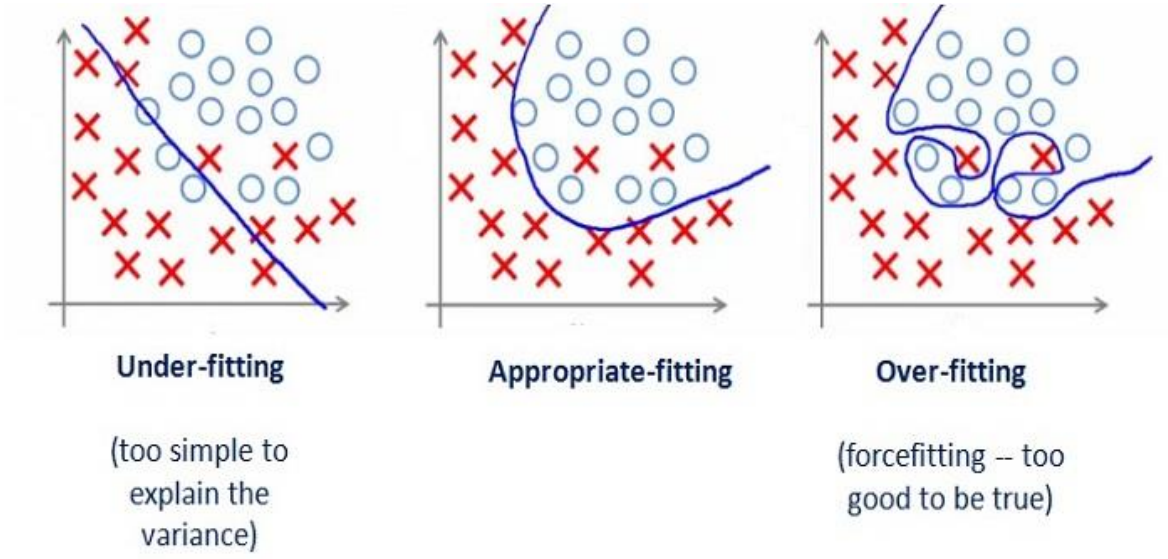
- **Complex models tend to detect patterns in the noise**
- **Those patterns do not generalize well to new instances**
- **Solutions**
  - Simpler model
  - More training data
  - Remove noise from training data



John Paul Mueller and Luca Massaron, *Machine Learning for Dummies* (John Wiley & Sons, 2016).

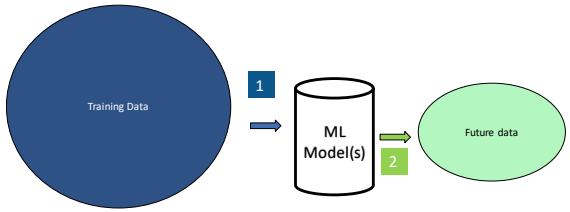
Underfitting the Training Data

- **Opposite of overfitting**
- **Model too simple**
- **Solutions**
  - Select more powerful model (more variables; or different algorithm)
  - Have better features
  - Reduce constraints / more degrees of freedom

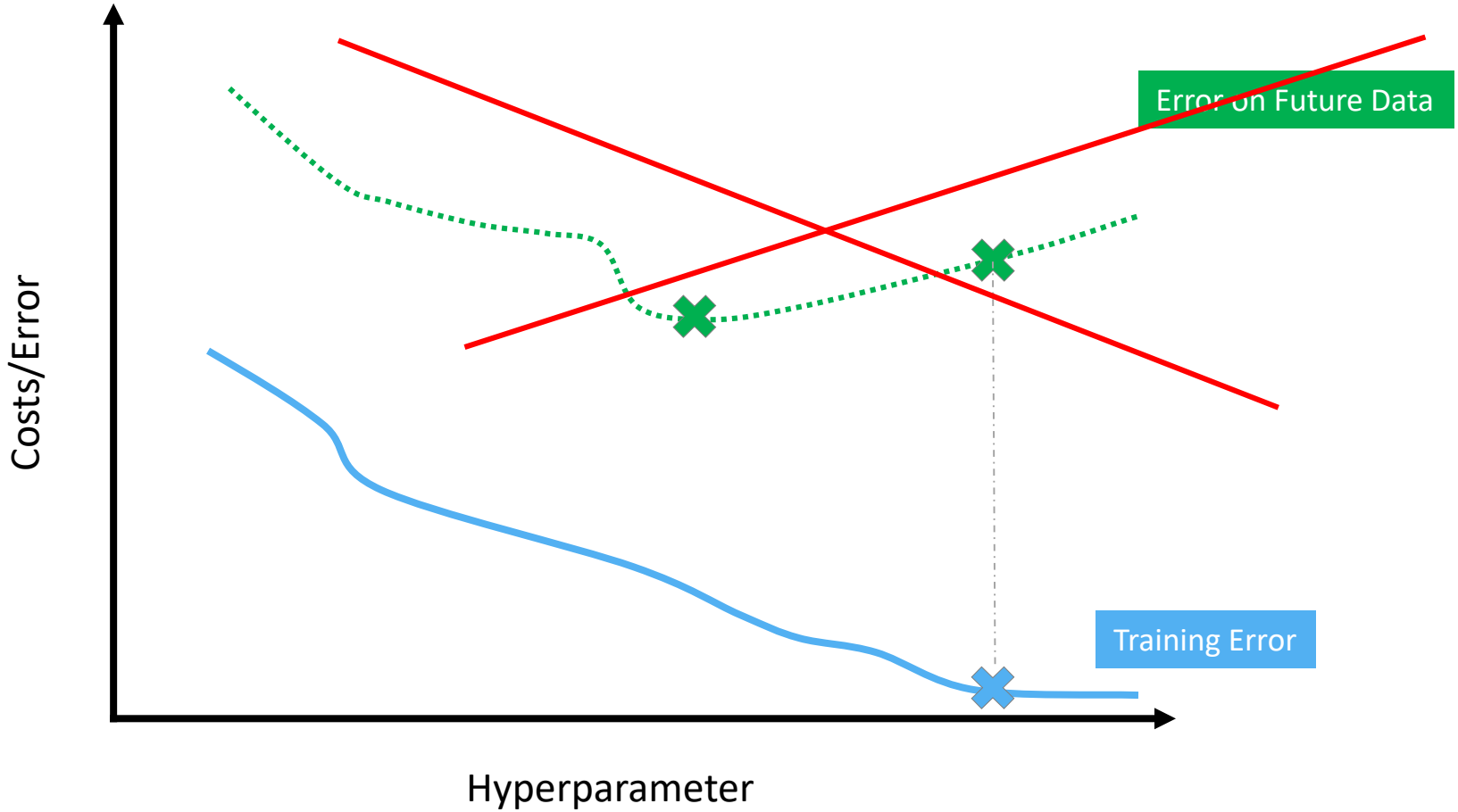


https://vitalflux.com/wp-content/uploads/2015/02/fittings.jpg

Problem



- **Future data is not available**

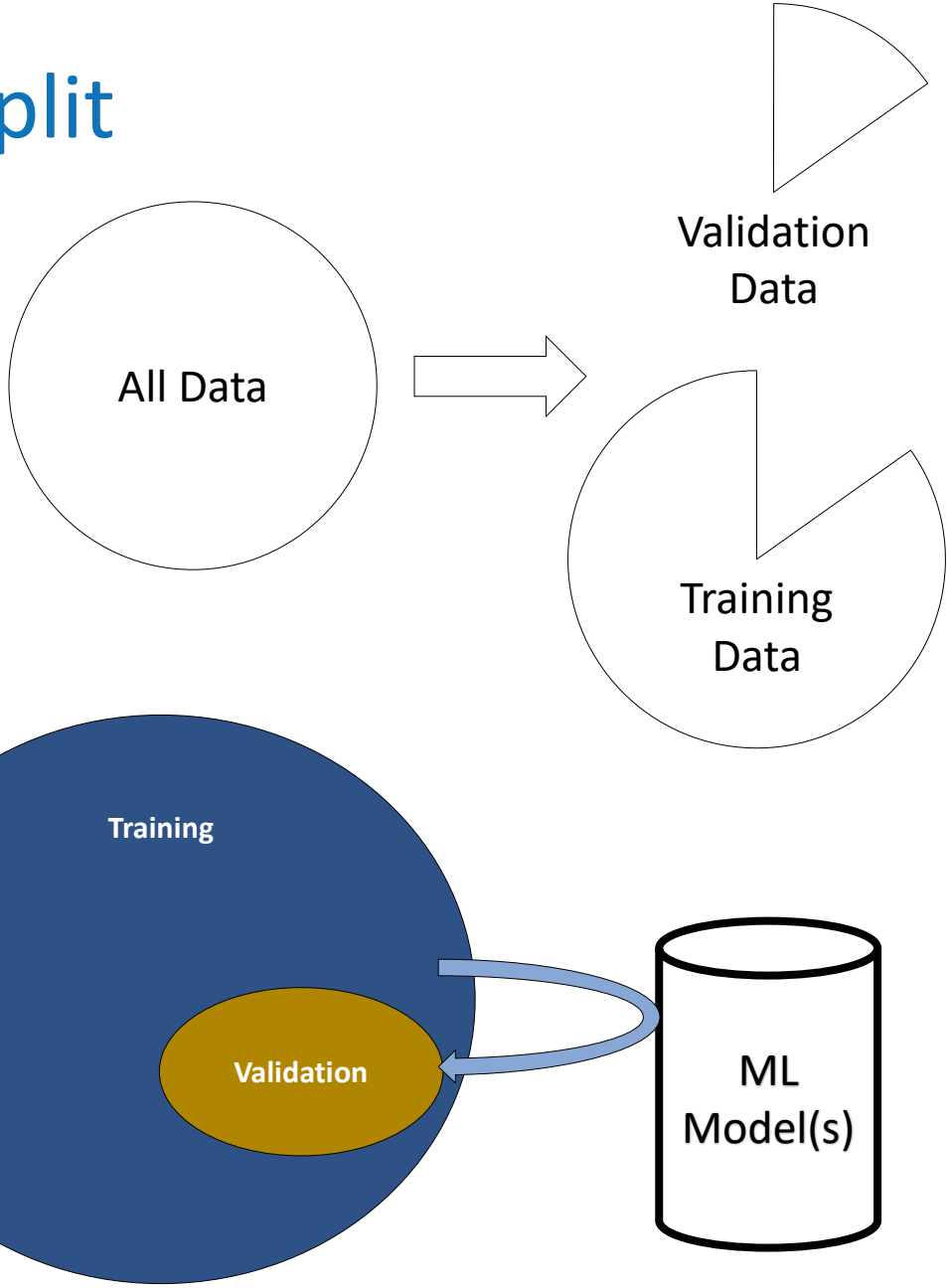


Hold-Out Evaluation

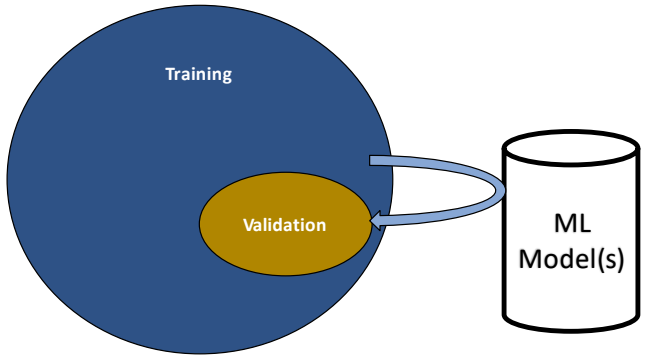
Hold-Out aka X/Y Split

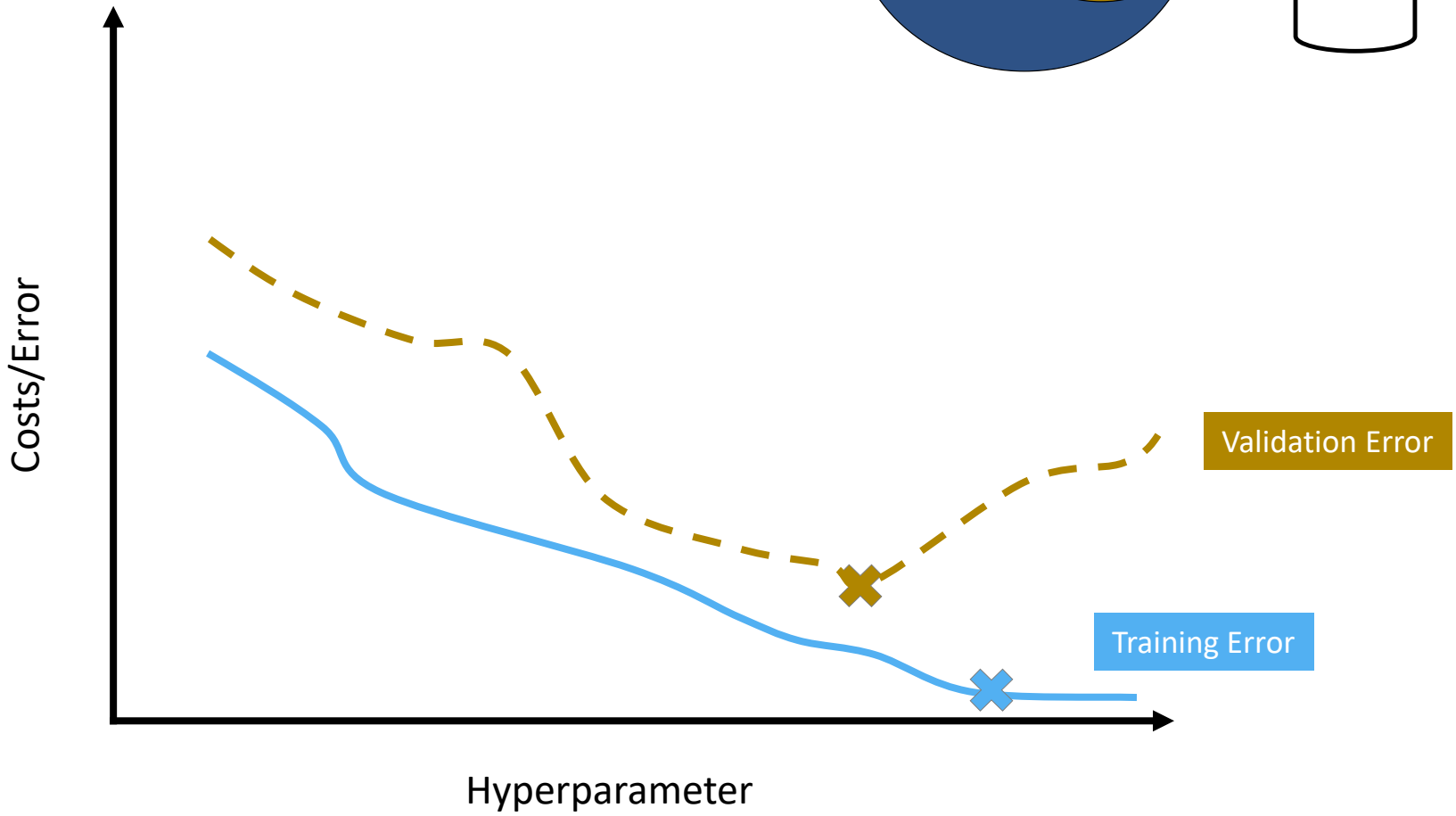
Supervised Learning

- **Split dataset into two subsets training d_t and validation d_v**
- **Typically $d_t > d_v$, e.g.**
  - 60/40
  - 70/30
  - 80/20
  - 90/10



Hold-Out Illustration

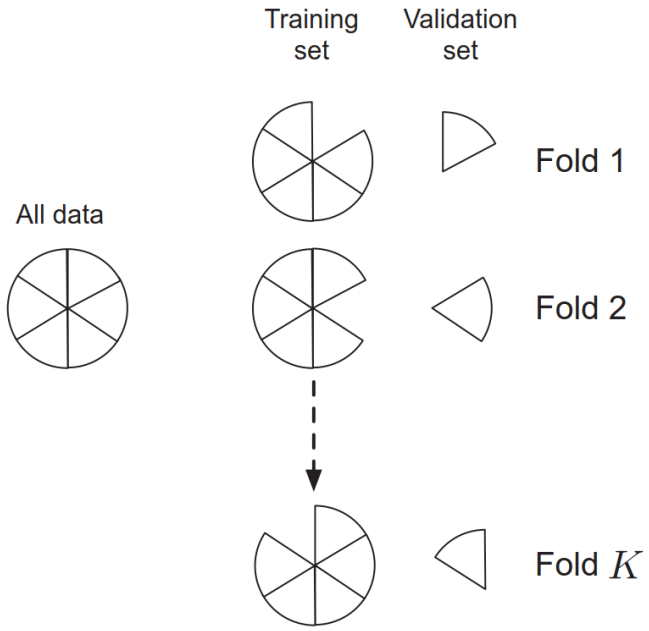




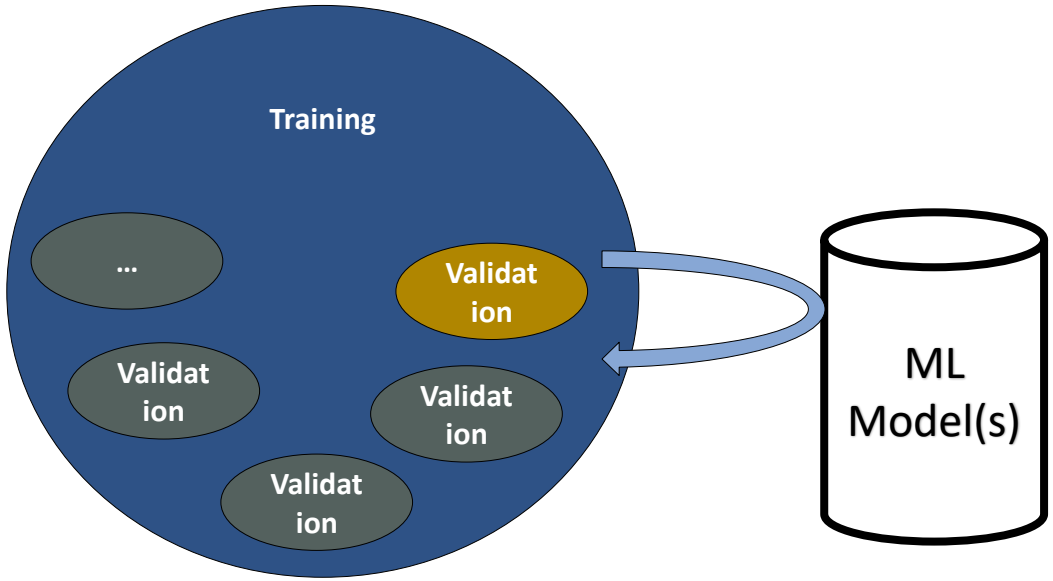
K-fold Cross Validation

Simon Rogers and Mark Girolami, A first course in machine learning (CRC Press, 2016).

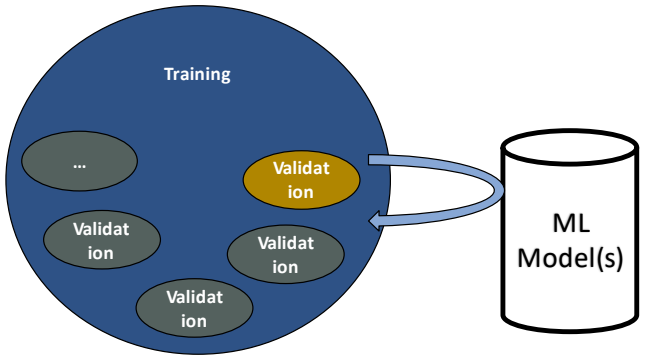
K-fold Cross Validation

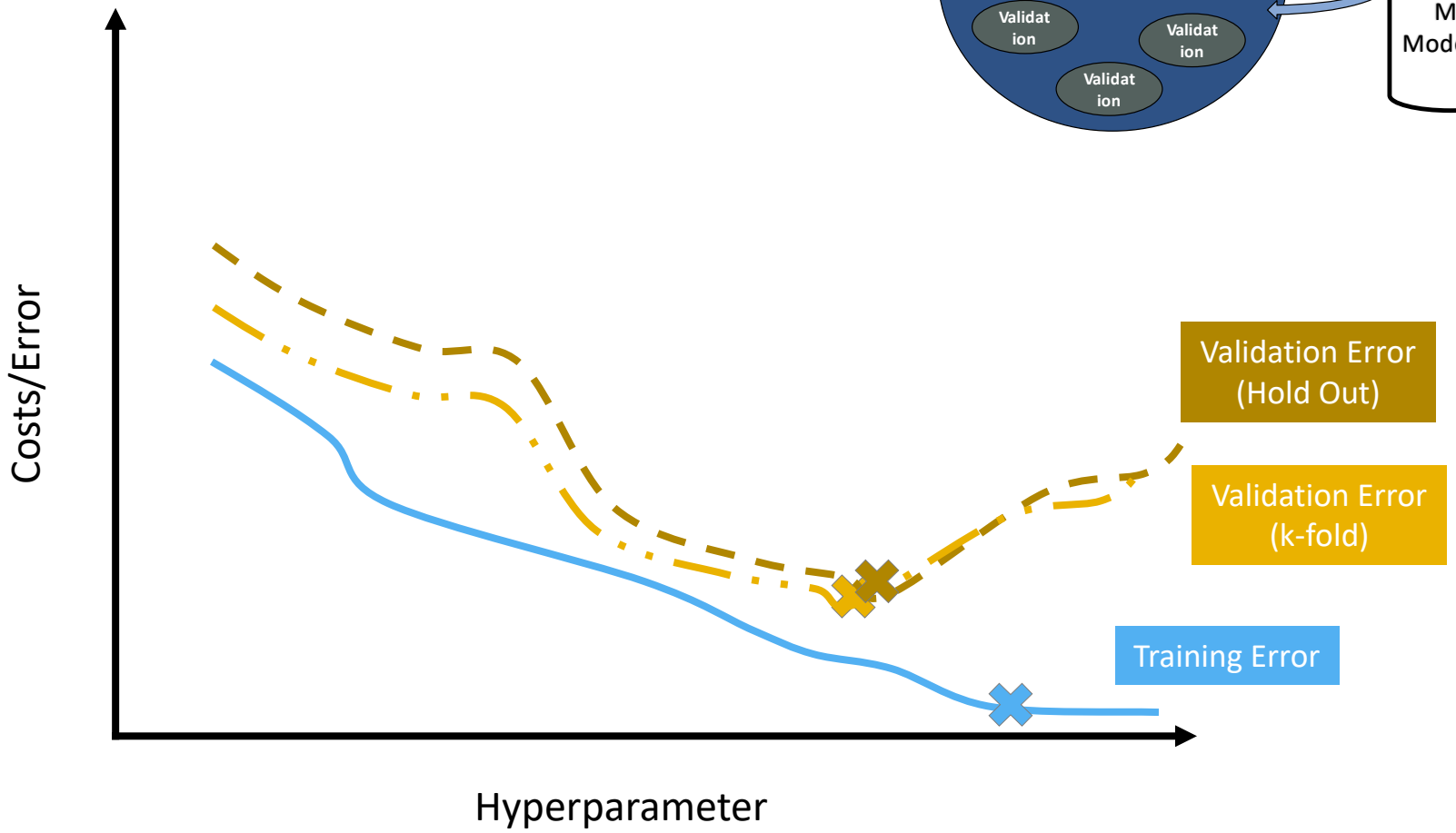


- **Split data in k blocks of equal size**
- **k is often 10**
- **Conduct k evaluations**
  - Train on fold 1…k
  - Evaluate with remaining block
- **Particularly sensible for small datasets**
- **Increases runtime by around factor k**



Hold-Out Illustration





Averaging the folds

A

	Number of Instances	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
Fold 1	2	1	-	1	-	0.50	0.50	1.00	0.67
Fold 2	2	1	1	-	-	1.00	1.00	1.00	1.00
Fold 3	2	1	-	-	1	0.50	1.00	0.50	0.67
Fold 4	2	1	1	-	-	1.00	1.00	1.00	1.00
Fold 5	2	1	1	-	-	1.00	1.00	1.00	1.00
Fold 6	2	1	1	-	-	1.00	1.00	1.00	1.00
Fold 7	2	1	1	-	-	1.00	1.00	1.00	1.00
Fold 8	2	1	1	-	-	1.00	1.00	1.00	1.00
Fold 9	2	1	1	-	-	1.00	1.00	1.00	1.00
Fold 10	2	1	1	-	-	1.00	1.00	1.00	1.00
Average (1)						0.900	0.950	0.950	0.933
Average (2)	20	10	8	1	1	0.900	0.909	0.909	0.909

B

Accuracy of Hold-out vs k-fold

TABLE I
PAGE BLOCKS CLASSIFICATION

Algorithm	*50% Hold-out*	*2-fold CV*	*20% Hold-out*	*5-fold CV*
Complex Tree	96.3	**96.3**	96.3	**96.4**
Medium Tree	96.7	**96.7**	96.2	**96.6**
Simple Tree	**95.9**	95.7	94.3	**95.7**
Linear SVM	**96.4**	96.1	96.1	**96.2**
Quadratic SVM	**96.9**	96.7	96.9	**96.9**
Cubic SVM	96.3	**96.6**	96.5	**96.9**
Fine Gaussian	94.4	**94.9**	94.7	**95.5**
Medium Gaussian	95.2	**96.1**	95.9	**96.5**
Coarse Gaussian	93.3	**94.2**	94.2	**94.8**
Fine KNN	95.8	**96.1**	95.8	**96.3**
Medium KNN	95.2	**95.9**	95.3	**96.1**
Coarse KNN	92.1	**92.5**	92.9	**93.4**
Cosine KNN	95.1	**95.7**	95.2	**95.8**
Cubic KNN	95.2	**95.8**	95.3	**96.0**
Weighted KNN	96.5	**96.6**	96.0	**96.9**
Boosted Ensemble	92.7	**93.2**	92.9	**93.1**
Bagged Ensemble	**97.6**	97.3	97.0	**97.5**
SD	92.4	**92.9**	92.5	**92.7**
Subspace KNN	95.4	**95.6**	94.9	**95.9**
RUSBoost	7.4	**37.5**	7.2	**26.7**

TABLE II
OPTICAL RECOGNITION OF HANDWRITTEN DIGITS

Algorithm	*50% Hold-out*	*2-fold CV*	*20% Hold-out*	*5-fold CV*
Complex Tree	87.1	**87.3**	87.7	**88.5**
Medium Tree	**70.3**	69.6	70.8	**70.9**
Simple Tree	38.4	**39.1**	38.4	**40.0**
Linear SVM	96.9	**97.4**	97.1	**97.5**
Quadratic SVM	98.1	**98.4**	98.0	**98.6**
Cubic SVM	98.5	**98.6**	**98.2**	98.0
Fine Gaussian	67.2	**67.6**	69.2	**70.0**
Medium Gaussian	97.3	**97.6**	97.2	**97.9**
Coarse Gaussian	94.6	**95.3**	95.1	**95.8**
Fine KNN	97.5	**97.7**	97.8	**97.8**
Medium KNN	96.9	**96.9**	96.5	**97.2**
Coarse KNN	92.0	**92.5**	**94.5**	93.8
Cosine KNN	95.8	**96.0**	96.6	**96.7**
Cubic KNN	**96.0**	95.9	**97.0**	96.5
Weighted KNN	97.4	**97.4**	97.0	**97.6**
Boosted Ensemble	**40.8**	39.7	**42.1**	38.1
Bagged Ensemble	97.3	**97.8**	98.0	98.0
Subspace Discriminant	94.5	**94.9**	95.0	**95.3**
Subspace KNN	98.4	**98.7**	98.7	**98.8**
RUSBoost	19.0	**19.4**	19.1	**21.4**

TABLE III
MAGIC GAMMA TELESCOPE

Algorithm	*50% Hold-out*	*2-fold CV*	*20% Hold-out*	*5-fold CV*
Complex Tree	84.2	**84.2**	85.2	**85.6**
Medium Tree	82.6	**83.0**	82.3	**84.0**
Simple Tree	79.1	**79.4**	**79.3**	79.2
Linear SVM	**79.7**	79.2	79.1	**79.2**
Quadratic SVM	86.0	**86.1**	86.4	**87.1**
Cubic SVM	86.2	**86.9**	86.5	**86.9**
Fine Gaussian	85.5	**85.6**	**86.2**	86.1
Medium Gaussian	**87.2**	86.6	**86.4**	86.0
Coarse Gaussian	82.2	**82.6**	83.5	**83.7**
Fine KNN	80.7	**80.9**	81.6	**82.3**
Medium KNN	**83.0**	82.9	83.3	**83.5**
Coarse KNN	81.0	**81.0**	82.2	**82.2**
Cosine KNN	83.1	**83.7**	84.0	**84.2**
Cubic KNN	82.4	**82.9**	83.0	**83.5**
Weighted KNN	83.1	**84.0**	84.3	**85.1**
Boosted Ensemble	**82.0**	81.4	**81.9**	81.3
Bagged Ensemble	87.6	**87.8**	88.1	**88.5**
Subspace Discriminant	77.0	**77.6**	76.6	**77.8**
Subspace KNN	**82.6**	82.2	**83.0**	82.6
RUSBoost	71.7	**72.5**	73.7	**74.0**

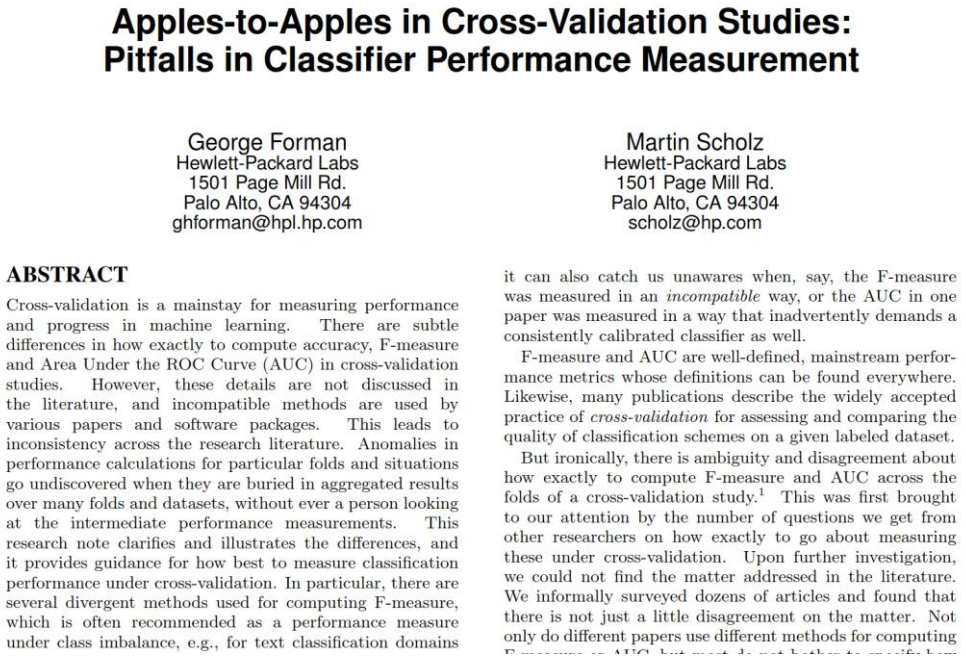
TABLE IV
LETTER RECOGNITION

Algorithm	*50% Hold-out*	*2-fold CV*	*20% Hold-out*	*5-fold CV*
Complex Tree	**62.6**	62.4	61.2	**62.0**
Medium Tree	**37.0**	36.4	35.7	**36.3**
Simple Tree	15.5	**16.0**	15.9	**15.9**
Linear SVM	83.8	**83.9**	84.4	**84.7**
Quadratic SVM	94.1	**94.4**	95.4	**95.7**
Cubic SVM	95.0	**95.4**	96.5	**96.6**
Fine Gaussian	**89.4**	88.0	91.1	**91.5**
Medium Gaussian	93.0	**93.1**	94.5	**94.8**
Coarse Gaussian	**79.4**	79.1	80.6	**81.5**
Fine KNN	93.7	**93.8**	94.8	**95.2**
Medium KNN	91.5	**91.8**	93.5	**93.7**
Coarse KNN	77.7	**77.9**	**82.8**	82.6
Cosine KNN	**90.5**	90.2	**92.6**	92.4
Cubic KNN	90.1	**90.4**	92.4	**92.6**
Weighted KNN	93.5	**93.7**	95.0	**95.0**
Bagged Ensemble	95.0	**95.4**	96.6	**96.6**
Boosted Ensemble	**13.1**	10.3	10.3	**10.9**
RUSBoost	**7.0**	6.9	6.8	**6.9**
Subspace Discriminant	66.7	**67.4**	66.4	**66.9**
Subspace KNN	95.2	**95.3**	96.2	**96.3**

Sanjay Yadav and Sanyam Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," in Advanced Computing (IACC), 2016 IEEE 6th International Conference on (IEEE, 2016), 78–83.

Want to know more?

- **E.g. how to draw the folds**
- **9-page Paper http://www.kdd.org/exploration_files/v12-1-p49-forman-sigkdd.pdf**
- **Search for…**
  - Leave-one-out cross validation (LOOCV)
  - Bootstrapping
  - Monte Carlo Cross Validation



Apples-to-Apples in Cross-Validation Studies: Pitfalls in Classifier Performance Measurement

George Forman
Hewlett-Packard Labs
1501 Page Mill Rd.
Palo Alto, CA 94304
ghforman@hpl.hp.com

Martin Scholz
Hewlett-Packard Labs
1501 Page Mill Rd.
Palo Alto, CA 94304
scholz@hp.com

ABSTRACT

Cross-validation is a mainstay for measuring performance and progress in machine learning. There are subtle differences in how exactly to compute accuracy, F-measure and Area Under the ROC Curve (AUC) in cross-validation studies. However, these details are not discussed in the literature, and incompatible methods are used by various papers and software packages. This leads to inconsistency across the research literature. Anomalies in performance calculations for particular folds and situations go undiscovered when they are buried in aggregated results over many folds and datasets, without ever a person looking at the intermediate performance measurements. This research note clarifies and illustrates the differences, and it provides guidance for how best to measure classification performance under cross-validation. In particular, there are several divergent methods used for computing F-measure, which is often recommended as a performance measure under class imbalance, e.g., for text classification domains it can also catch us unawares when, say, the F-measure was measured in an *incompatible* way, or the AUC in one paper was measured in a way that inadvertently demands a consistently calibrated classifier as well.

F-measure and AUC are well-defined, mainstream performance metrics whose definitions can be found everywhere. Likewise, many publications describe the widely accepted practice of *cross-validation* for assessing and comparing the quality of classification schemes on a given labeled dataset.

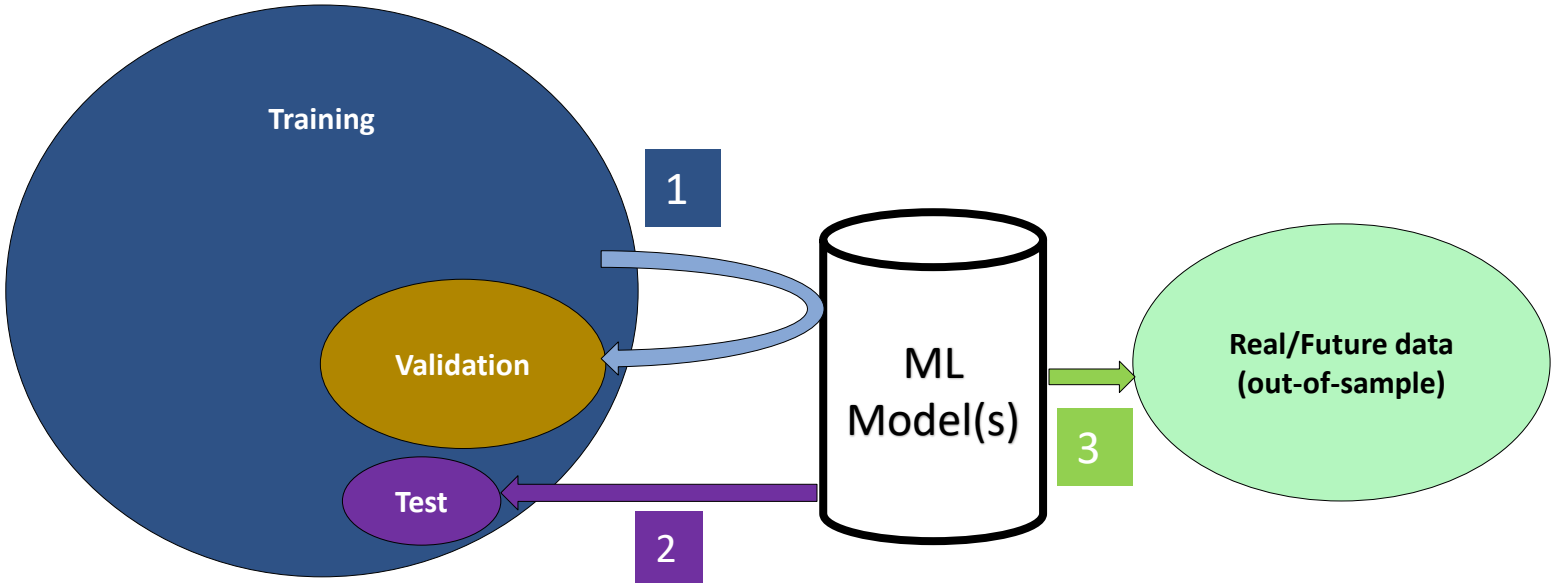
But ironically, there is ambiguity and disagreement about how exactly to compute F-measure and AUC across the folds of a cross-validation study.[1] This was first brought to our attention by the number of questions we get from other researchers on how exactly to go about measuring these under cross-validation. Upon further investigation, we could not find the matter addressed in the literature. We informally surveyed dozens of articles and found that there is not just a little disagreement on the matter. Not only do different papers use different methods for computing

Test Set

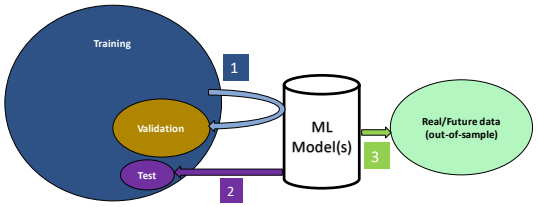
Entire Process

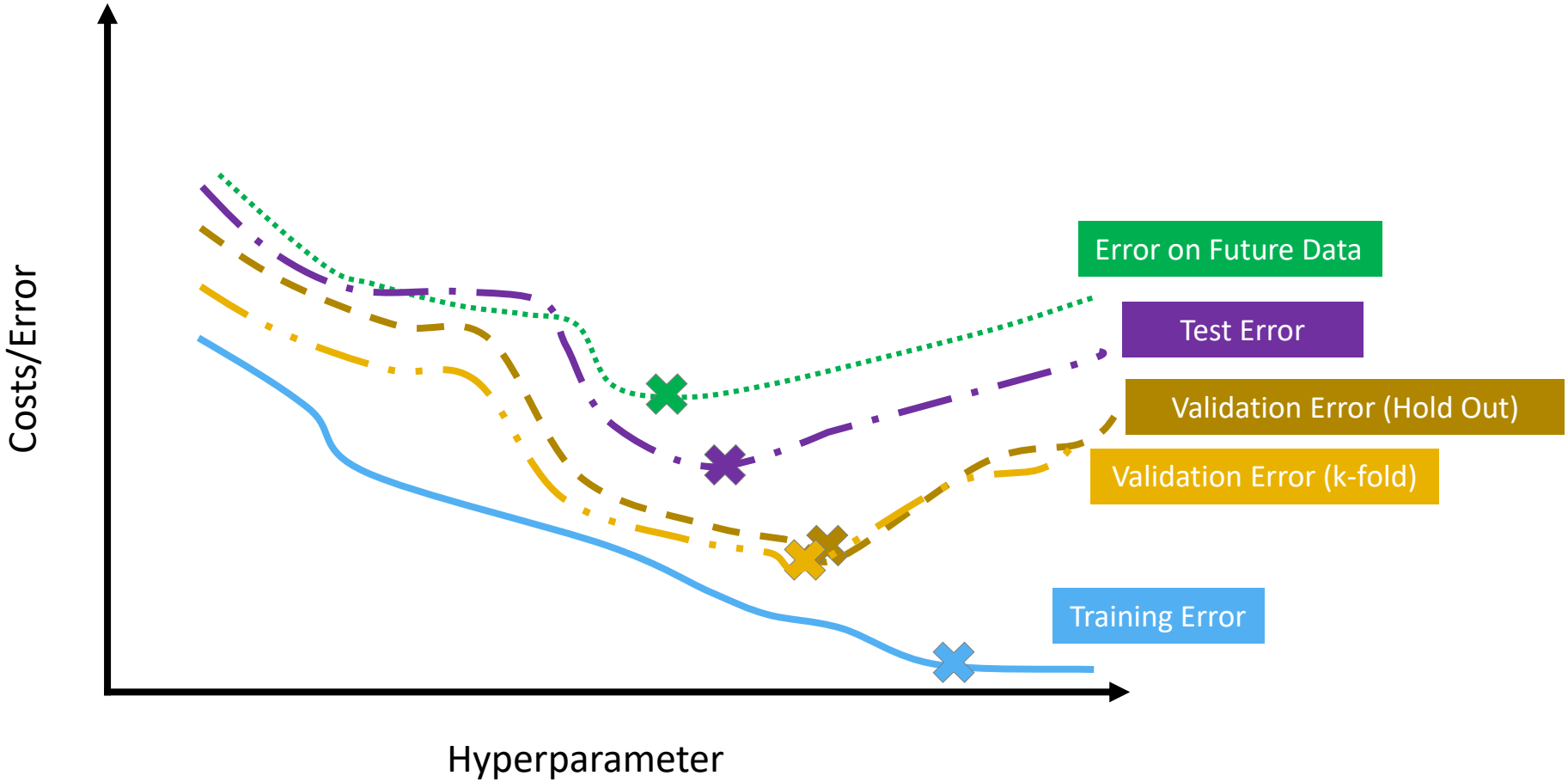
Supervised Learning

- **Optimizing for the validation dataset creates (potentially) some bias towards the validation data**
- **Solution: Hold out a test set (e.g. 10%)**
- **Use it only at the very end; never look at it during training and validation**

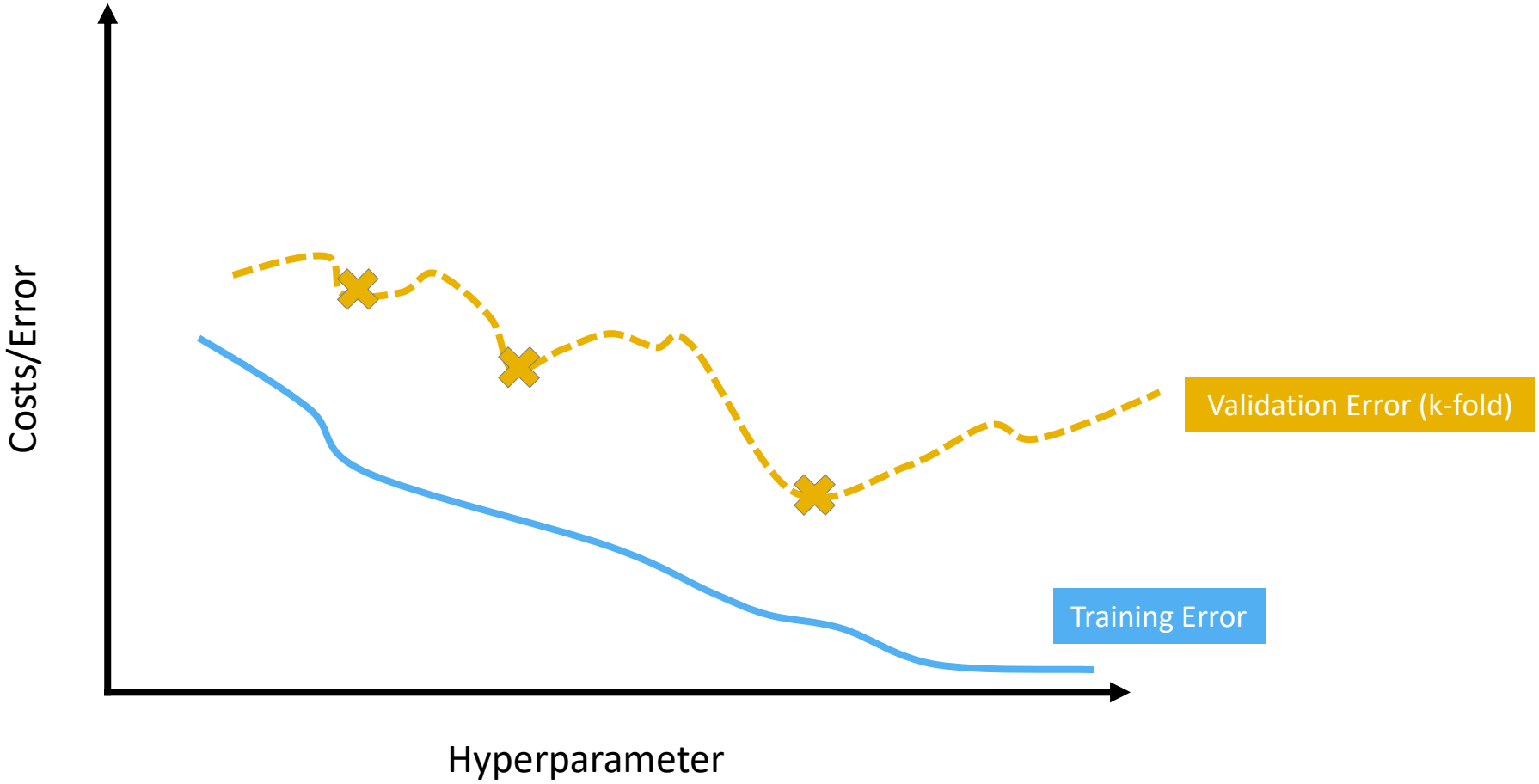


Illustration



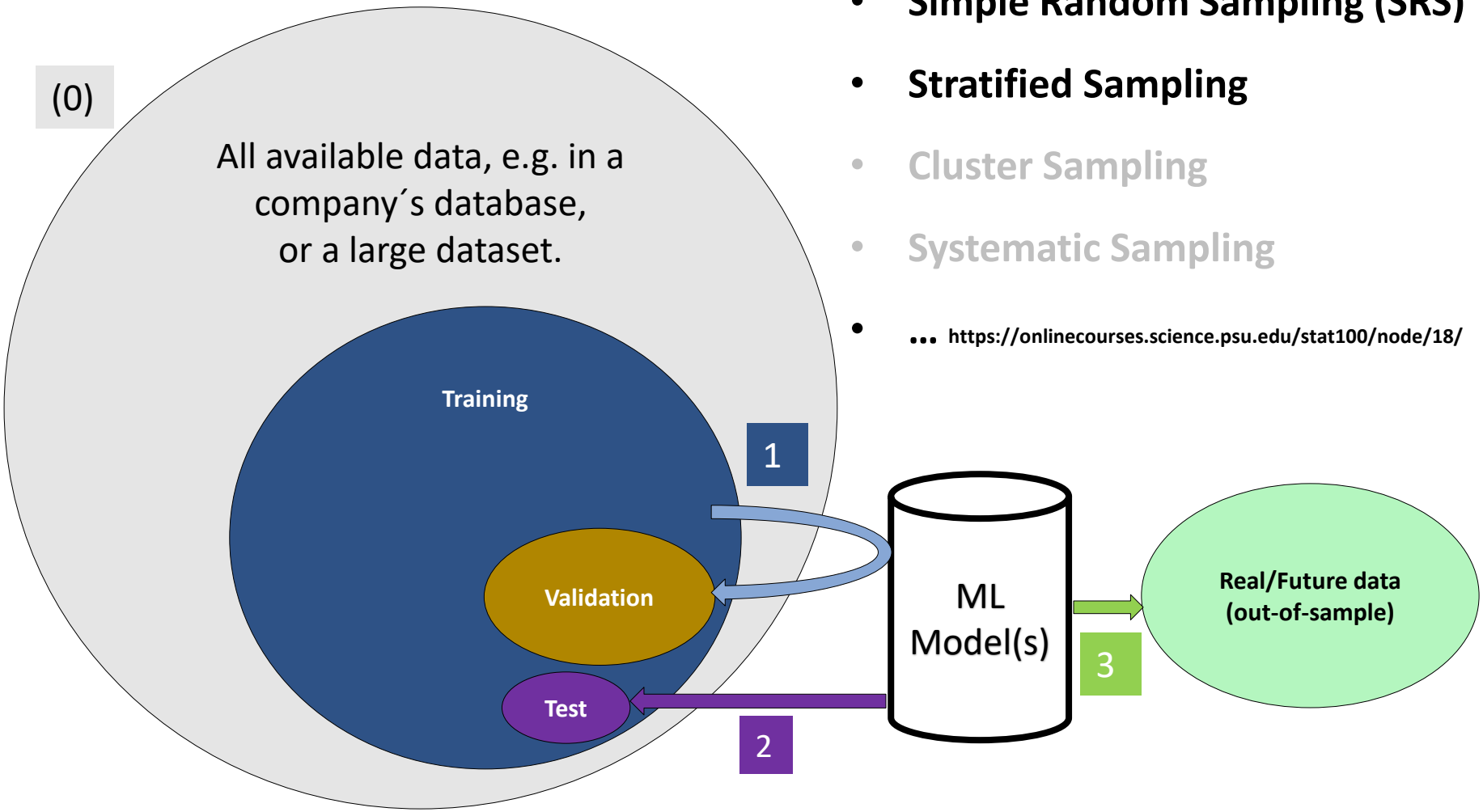


When to stop?


Costs/Error
Validation Error (k-fold)
Training Error
Hyperparameter

Sampling

ML Data != All Available data



- **Simple Random Sampling (SRS)**
- **Stratified Sampling**
- **Cluster Sampling**
- **Systematic Sampling**
- **… https://onlinecourses.science.psu.edu/stat100/node/18/**

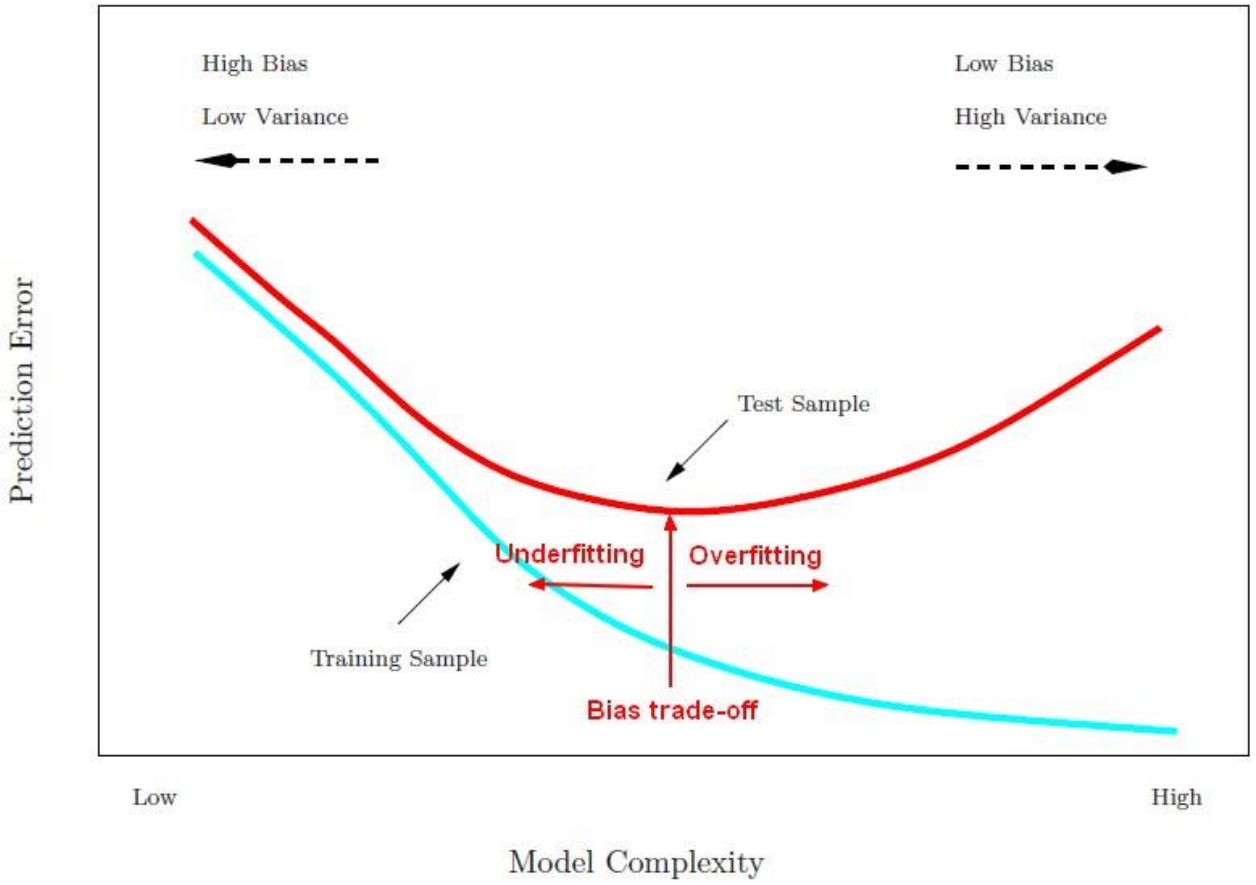
Stratified Sampling

- **https://onlinecourses.science.psu.edu/stat100/node/18/**

	Example 1	Example 2	Example 3
Population	All people in U.S.	All PSU intercollegiate athletes	All elementary students in the local school district
Groups (Strata)	4 Time Zones in the U.S. (Eastern,Central, Mountain,Pacific)	26 PSU intercollegiate teams	11 different elementary schools in the local school district
Obtain a Simple Random Sample	500 people from each of the 4 time zones	5 athletes from each of the 26 PSU teams	20 students from each of the 11 elementary schools
Sample	$4 \times 500 = 2000$ selected people	$26 \times 5 = 130$ selected athletes	$11 \times 20 = 220$ selected students

Bias-Variance Trade Off

- **High Bias occurs when an algorithm has not enough flexibility**
- **Variances increases the more sensitive an algorithm reacts to the data**



https://gerardnico.com/wiki/_detail/data_mining/model_complexity_error_training_test.jpg?id=data_mining%3Abias_trade-off

Bias-Variance Tradeoff (2)

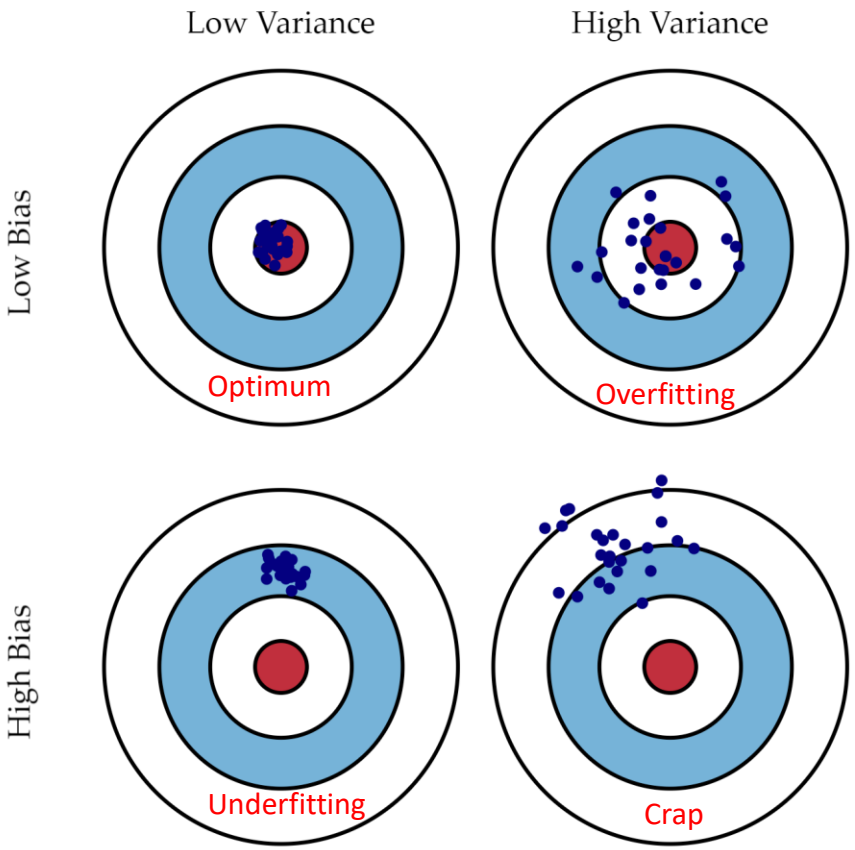


Fig. 1 Graphical illustration of bias and variance.

http://scott.fortmann-roe.com/docs/BiasVariance.html

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Evaluation

How good is your model (compared to others)?

The overall question

What is a good machine learning model?

→ A model that achieves its objective.

1. What is the objective?
2. How can the objective be framed as a machine learning task?
3. How can it be measured?

https://image.freepik.com/free-vector/the-winner_23-2147506357.jpg

Defining the specific objective and metric is not easy

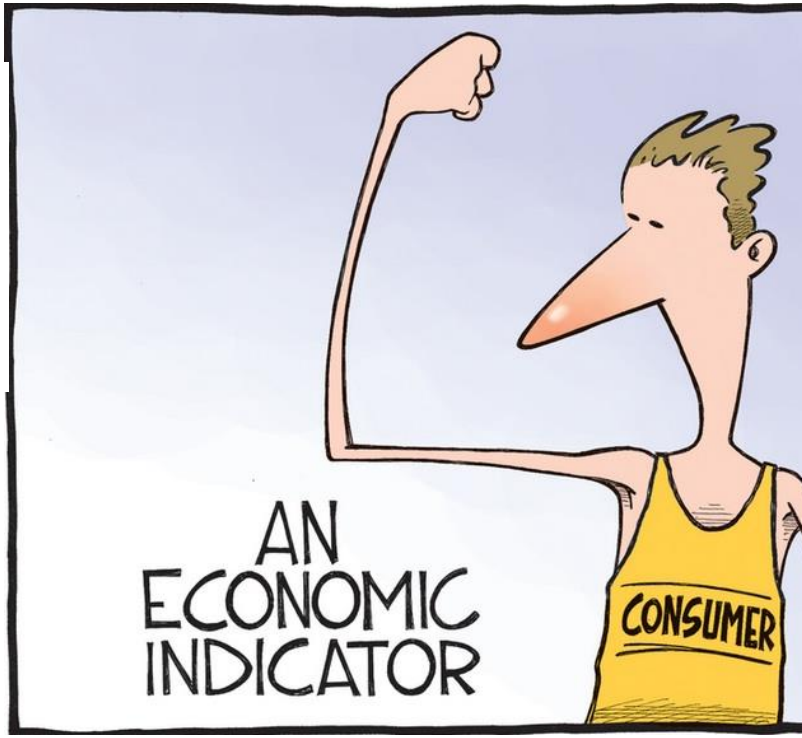
Objective: Maximize economic strength of Ireland

What is economic strength?

?

How to measure it?
(e.g. unemployment rate)

?



http://www.w-t-w.org/en/wp-content/uploads/2015/02/Economic-Strength.jpg

Pitfalls

“When a measure becomes a target, it ceases to be a good measure.”
Goodhart’s law

Example: Emergency Treatment in hospitals

- **Metric: Waiting time**
- **Consequences**
  - Keep patients in ambulance
  - Favor younger patients / easy cases over older patients / more complicated cases
  - Treat them quickly and tell them to come back
  - …

http://cyberlibris.typepad.com/blog/files/Goodharts_Law.pdf

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

(Offline) Evaluation Metrics

Classification Metrics

Confusion Matrix

In (binary) classification, there are four potential outcomes

- Correctly classified as „true“
- Correctly classified as „false“
- Incorrectly classified as „true“
- Incorrectly classified as „false“

		Predicted Class	
		POSITIVE	**NEGATIVE**
Actual Class	**POSITIVE**	True Positives (TP)	False Negative (FN)
	NEGATIVE	False Positive (FP)	True Negatives (TN)

Error Rate (ERR)

- **How many of the classifications were incorrectly predicted?**

$$ERR = \frac{\#\ Incorrect\ Predictions}{\#\ All\ Predictions}$$

$$ERR = \frac{FP + FN}{TP + TN + FP + FN}$$

		Predicted Class	
		POSITIVE	**NEGATIVE**
Actual Class	**POSITIVE**	True Positives (TP)	False Negative (FN)
	NEGATIVE	False Positive (FP)	True Negatives (TN)

Accuracy

- **How many of the classifications were correctly predicted?**

$$Accuracy = \frac{\#\ Correct\ Predictions}{\#\ All\ Predictions}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = 1 - \mathrm{ERR}$$

		Predicted Class	
		POSITIVE	**NEGATIVE**
Actual Class	**POSITIVE**	True Positives (TP)	False Negative (FN)
	NEGATIVE	False Positive (FP)	True Negatives (TN)

Meaningfulness of Accuracy

- **Not always ideal, especially with skewed distributions of classes.**
- **Example**
  - Dataset with cancer screenings.
  - 10 Cancer Case; 99,999 non-cancer cases
  - Simple heuristic „Always predict not cancer“
  - Accuracy: 99.99%
- **When TP < FP, accuracy will increase when we simply always predict „negative“ (and vice versa).**

https://static.seekingalpha.com/uploads/2010/1/18/saupload_big_cap_vs_small_cap.jpg

Easily Improving Accuracy: Real-World Example

Audio Captchas

- **Goal: Recognize audio captchas (digits)**
- **Split audio captcha into sequences**
- **Send to multiple speech recognition APIs**
- **Combine results with ensemble**
- **Solve 80.31% of captchas**
- **Analyse problems (per class accuracy)**
- **Add simple heuristic: Classify everything that couldn‘t be classified as digit as „six“**
- **Improve accuracy by ~6%**

N	Best Ensemble of N Services	Per-digit Accuracy	Captcha Success
1	Google Cloud	55.81%	2.19%
2	Bing + IBM	82.87%	47.26%
3	+ Google Cloud	88.25%	66.96%
4	+ Wit	91.36%	78.12%
5	+ Sphinx	91.93%	80.09%
6	+ Google	91.99%	80.31%
	+ “X” $\mapsto$ “6”	**93.41%**	**85.15%**

+ ~6%

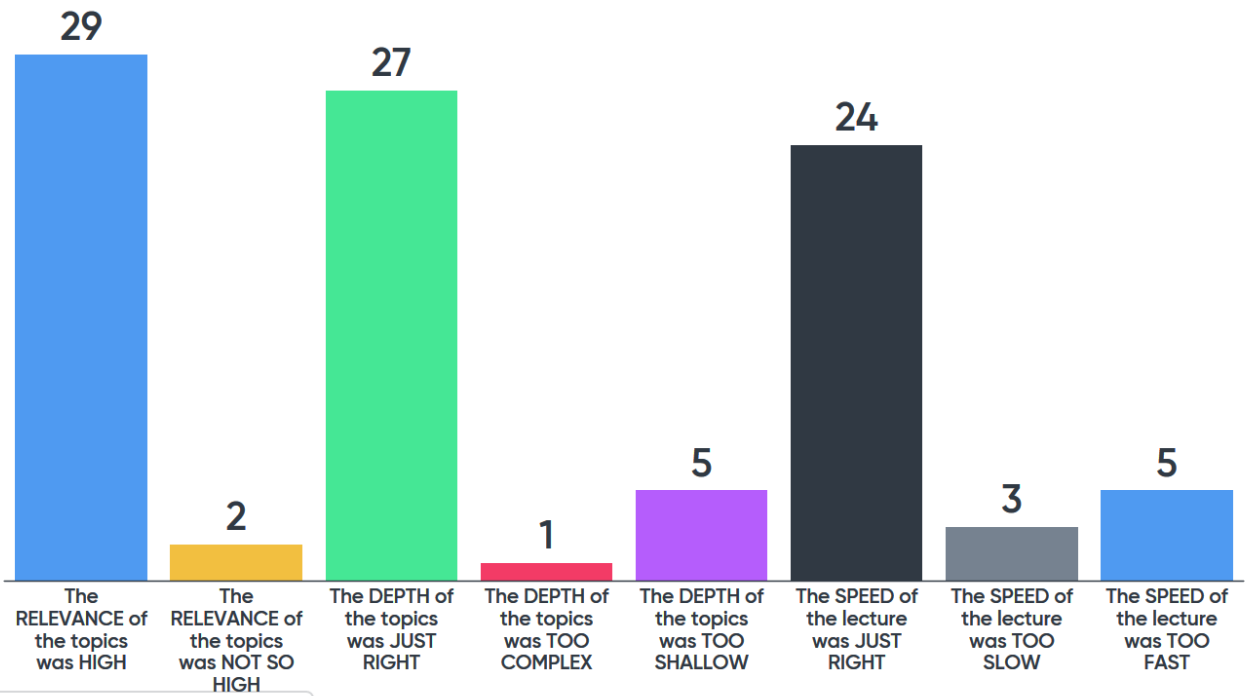
unCaptcha: A Low-Resource Defeat of reCaptcha’s Audio Challenge
http://uncaptcha.cs.umd.edu/papers/uncaptcha_woot17.pdf

If you attended the lecture yesterday, please vote

Go to **www.menti.com** and use the code **51 13 86**

Lecture Evaluation

Mentimeter



Slide is not active Activate

32

Sensitivity / Recall / True Positive Rate (TPR)

- **How many of the positive classes were correctly predicted?**
- **How many of the correct results were retrieved?**

$$Recall = \frac{TP}{TP + FN} = 1 - FNR$$

- **Relevant e.g. for retrieval and medicine**

		Predicted Class	
		POSITIVE	**NEGATIVE**
Actual Class	**POSITIVE**	True Positives (TP)	False Negative (FN)
	NEGATIVE	False Positive (FP)	True Negatives (TN)

And more

- **False Positive Rate (FPR): How many of the positive predictions were incorrect (probability of a false alarm)?**
- **True Negative Rate (TNR) / Specificity**
- **False Negative Rate (FNR)**

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

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

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

		Predicted Class	
		POSITIVE	**NEGATIVE**
Actual Class	**POSITIVE**	True Positives (TP)	False Negative (FN)
	NEGATIVE	False Positive (FP)	True Negatives (TN)

Example

		Predicted Class	
		Cancer predicted (Σ 0)	**No cancer predicted (Σ 0)**
Actual Class	**Patients with cancer (Σ 100)**	60	40
	Patients without cancer (Σ 85000)	5000	80,000

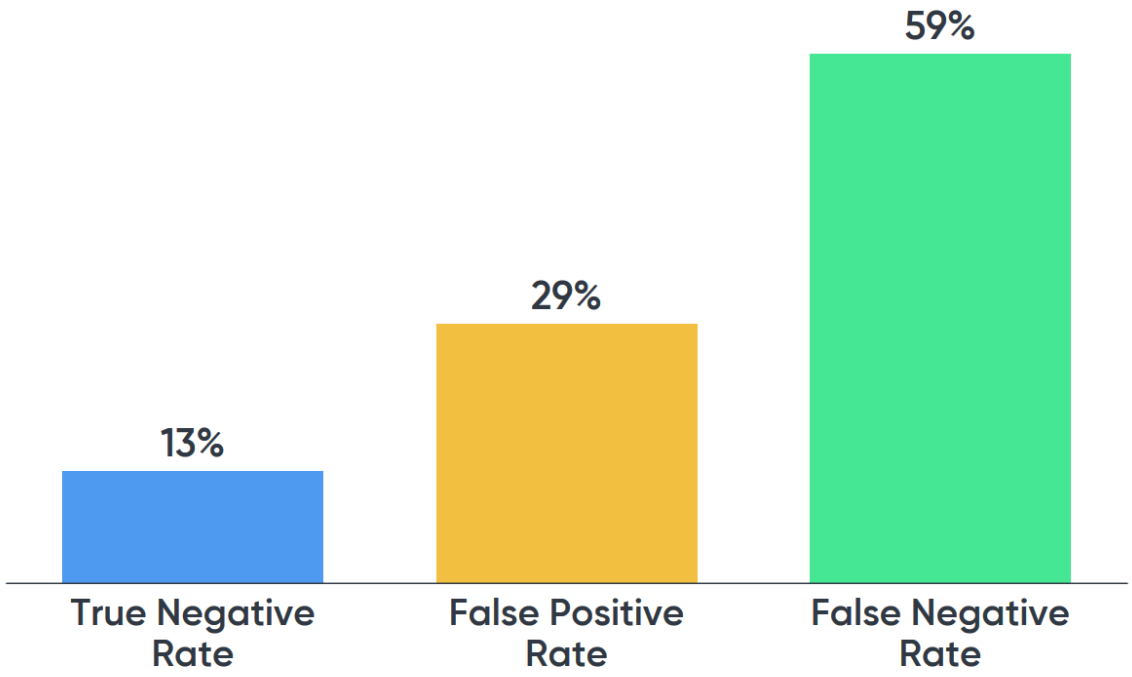
Error Rate	0.059
Accuracy	0.941
Recall (TPR)	0.600
FPR	0.059
TNR	0.941
FNR	0.400

How to get a Recall of 1?

Go to **www.menti.com** and use the code **51 13 86**

For cancer screening, which metric is most important besides True Positive Rate?

Mentimeter

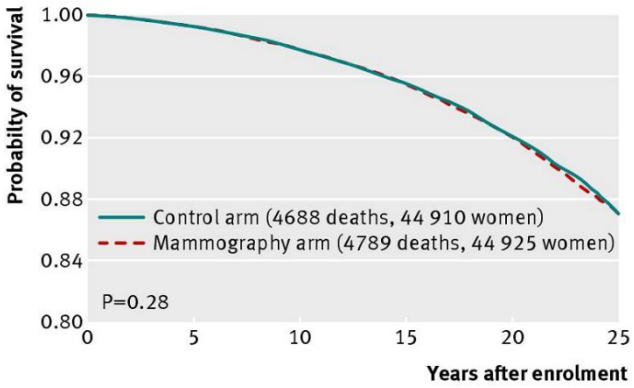


Slide is not active **Activate**

80

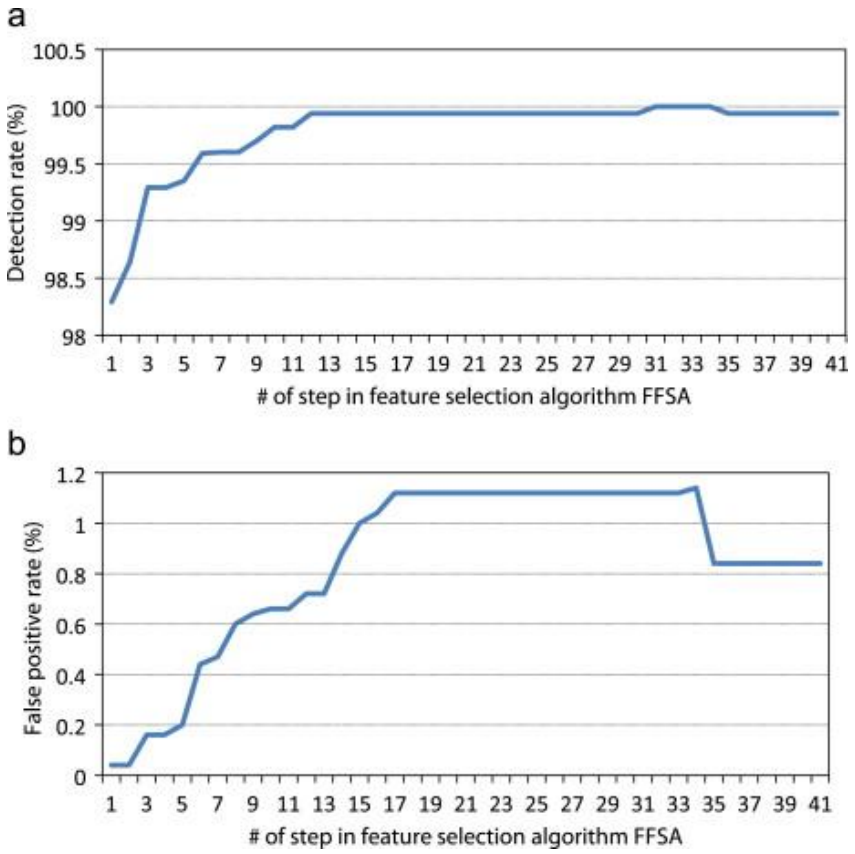
More Details on Metrics in Medicine

- **https://andrewgelman.com/2018/09/10/38592/**
- **https://andrewgelman.com/2017/09/02/cause-breast-cancer-specific-mortality-assignment-mammography-control/**
- **https://www.bmj.com/content/362/bmj.k3702**
- **https://www.nytimes.com/2009/11/17/health/17cancer.html**
- **JAMA https://www.health.harvard.edu/blog/new-mammography-guidelines-call-for-starting-later-and-screening-less-often-201510218466**
- **https://jamanetwork.com/journals/jama/fullarticle/2680553**
- **https://jamanetwork.com/journals/jama/fullarticle/2679928**
- **https://jamanetwork.com/journals/jama/fullarticle/2679928**



Meaningfulness of Recall / TP Rate

- **True positive rate alone has little meaning**
- **TP Rate can always be 100% if you compromise on False Positives**
- **Keep in mind that costs of mis-classification may differ significantly**
- **E.g. cancer screening**
  - Costs for „false positive“ is typically rather low
  - Costs for „false negative“ might be deadly

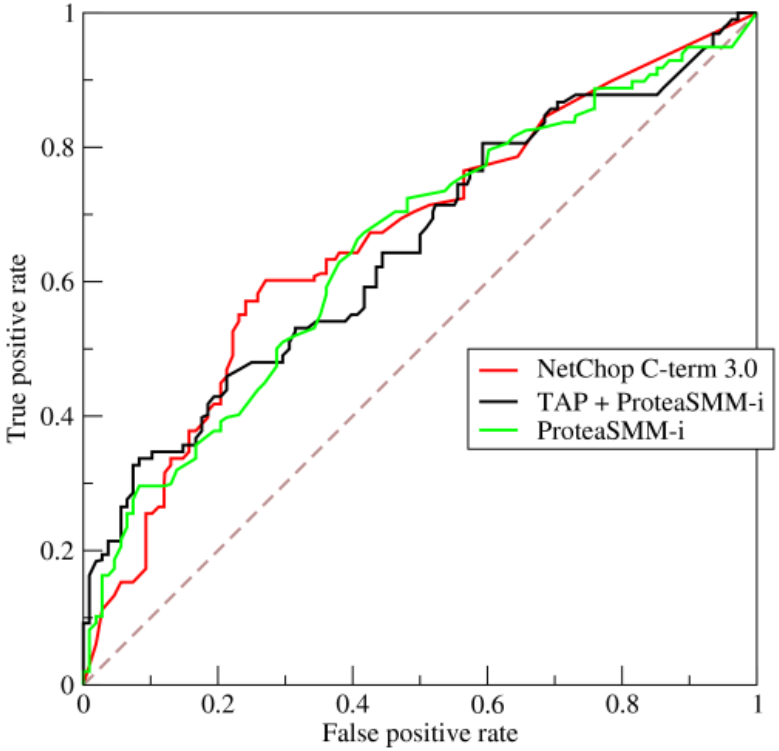


http://www.sciencedirect.com/science/article/pii/S1084804511000038#f0005

Can you think of an opposite example? High cost for False Positive and rahter low costs for False Negative?

Receiver Operating Characteristic Curve (ROC Curve)

- **Plots True Positive Rate vs. False Positives**
- **Shows how many TP can be gained, by accepting more FP (typically, TP can be increased by increasing „sensitivity“ of method, though FP also increase)**
- **In ideal world TP Rate = 100% -- nothing to gain**
- **Often, difficult to see, which algorithm is better**
- **Area Under The Curve (AUC)**
  - In theory values between 0 and 1
  - In practice between 0.5 and 1
  - AUC = 0.5 is random classifier
- **More details: http://www.dataschool.io/roc-curves-and-auc-explained/**

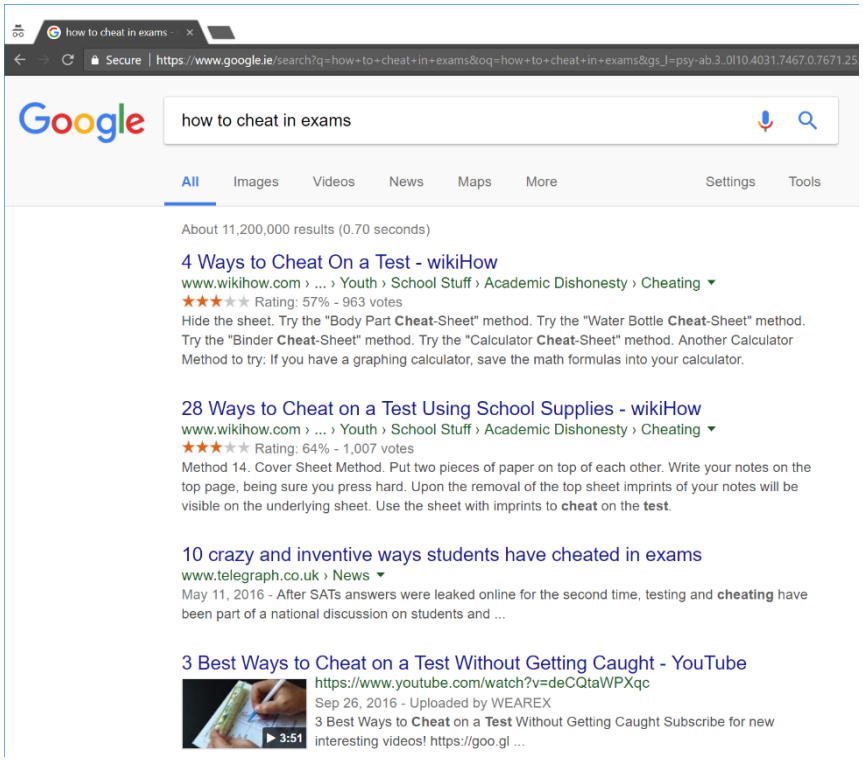


https://upload.wikimedia.org/wikipedia/commons/6/6b/Roccurves.png

Precision / Positive predictive value

- **In search: How many of the retrieved results are correct/relevant?**
- **In machine learning: of all the instances classified as „positive“, how many were classified correctly?**
- **Goal: You don‘t want to have irrelevant items shown in search results**
- ***p@n* (or p@k) describes precision among the top *n* results**

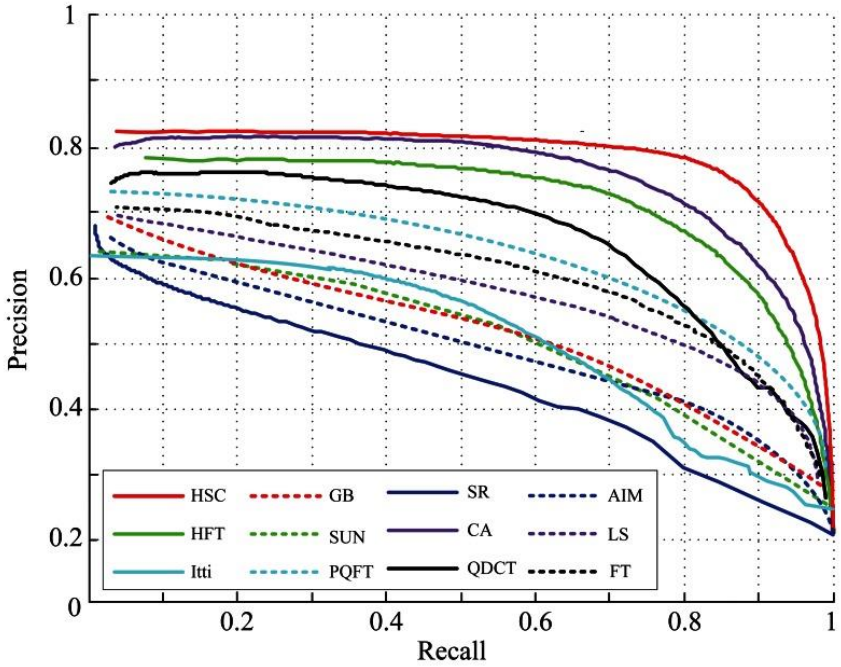
$$P = \frac{TP}{TP + FP}$$



How to get a precision of 1?

Precision-Recall Curve / F-measure / F_1 Score

- **Precision-Recall Curve plots precision vs. recall**
- **F-measure combines Precision and Recall into one metric (harmonic mean)**
- **Precision-Recall Curve & F-Measure is similar to ROC Curve and AUC**



$$F_\beta = (\beta^2 + 1) \times \frac{P \times R}{(\beta^2 \times P) + R}$$

$\beta < 1$ emphasizes precision
$\beta > 1$ emphasizes recall

$$F_{\beta=1} = F_1 = 2 \times \frac{P \times R}{P + R}$$

https://www.researchgate.net/profile/Jianru_Xue/publication/236038455/figure/fig6/AS:213888827170830@1428006478416/The-Precision-recall-curve-for-naive-thresholding-of-saliency-maps-using-1000-publicly.png

Log-Loss

- „Soft“ measure for accuracy
- For probabilistic classifiers (e.g. probability of 0.73 that it‘s class A)
- Considers how close the predicted class was to true class
- E.g. if classifier incorrectly classifies instance as class A with probability of 0.51 that‘s still better than a probability of 0.93.
- log-loss = cross entropy between the distribution of the true labels and the predictions
- “Closely related to what’s known as the relative entropy, or Kullback–Leibler divergence
- Entropy measures the unpredictability of something.
- Cross entropy incorporates the entropy of the true distribution, plus the extra unpredictability when one assumes a different distribution than the true distribution.
- log-loss is an information-theoretic measure to gauge the “extra noise” that comes from using a predictor as opposed to the true labels.
- By minimizing the cross entropy, we maximize the accuracy of the classifier”
- Equal weight for false positives and false negatives

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}[\,(y_i * \log(p_i)) + ((1-y_i) * \log(1-p_i))]$$

$p_i = calculated\ probability\ (or\ confidence)\ that\ the\ ith\ data\ point\ belongs\ to\ class\ 1$
$y_i = true\ label\ of\ the\ ith\ instance\ [0|1]$
$N = number\ of\ instances$

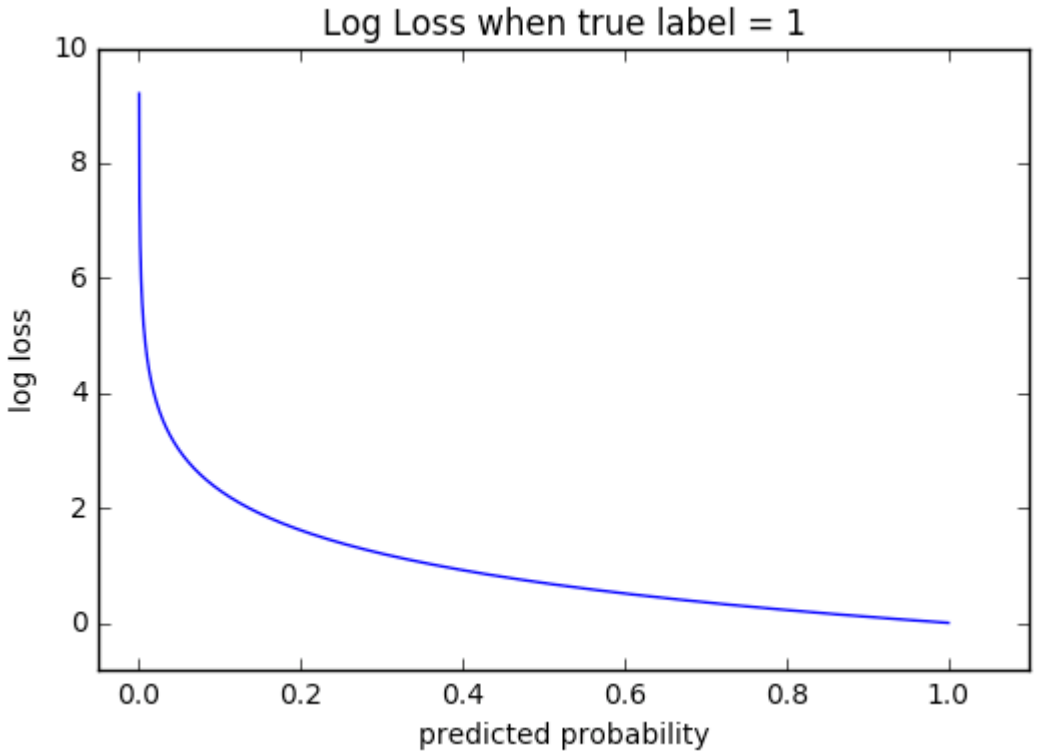
$$LogLoss_{MultiClass} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

$N = number\ of\ instances$
$M = number\ of\ classes, i.e.\ labels$
$y_{i,j} = binary\ indicator\ if\ label\ j\ is\ correctly\ predicted\ for\ instance\ i\ [0/1]$
$p_{i,j} = probability\ that\ j\ is\ the\ correct\ label\ for\ i$

Alice Zheng, “Evaluating Machine Learning Models” (O’Reilly Media, Inc, 2015).

http://wiki.fast.ai/images/4/43/Log_loss_graph.png

Example



True Label	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p_i	0.00000001	0.000001	0.0001	0.001	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.999
LogLoss	8	6	4	3	2	1	0.7	0.5	0.4	0.3	0.2	0.15	0.10	0.05	0.0004

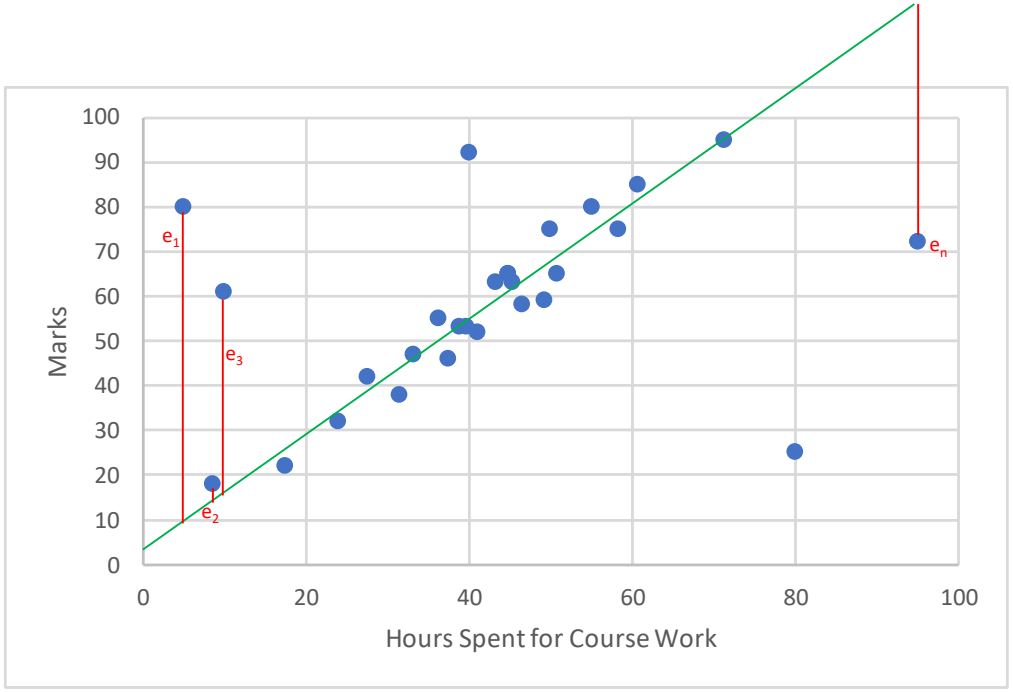
True Label	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_i	0.001	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.999	0.99999	0.999999	0.99999999
LogLoss	0.0004	0.0044	0.05	0.10	0.15	0.22	0.3	0.4	0.5	0.7	1	3	5	6	8

Regression Metrics

Mean Absolute Error (MAE)

- **Average error between prediction and observation**

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i|$$



(Root) Mean Square Error -- (R)MSE

- **Measures the standard deviation of errors that a system makes in its predictions**
- **n = number of instances in a dataset**
- **Sensitive to outliers**

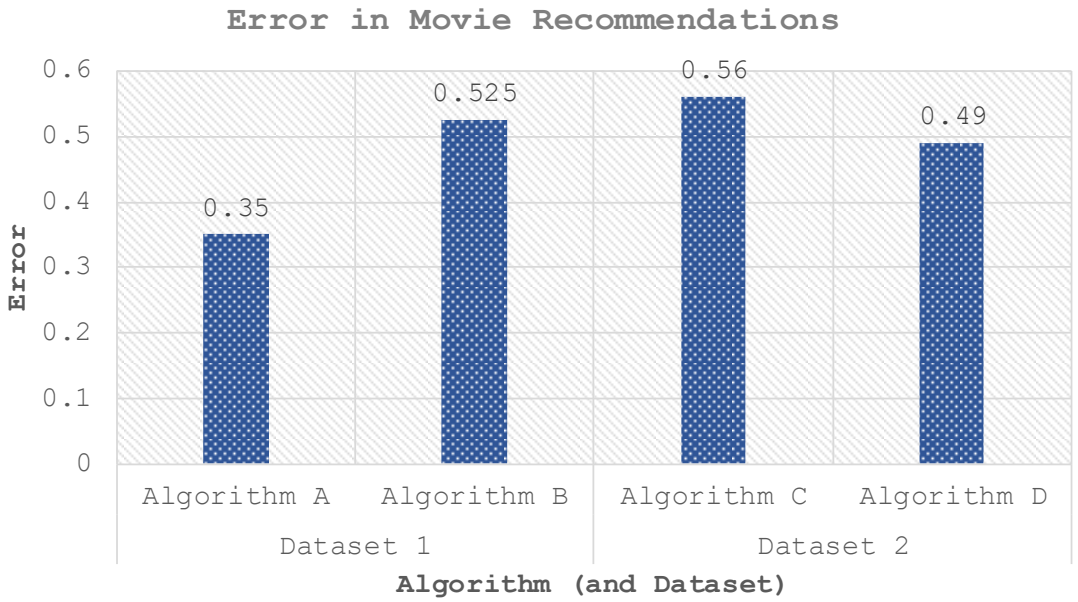
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}{n}}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{e_1^2 + e_2^2 + e_3^2 + \ldots + e_n^2}{n}$$

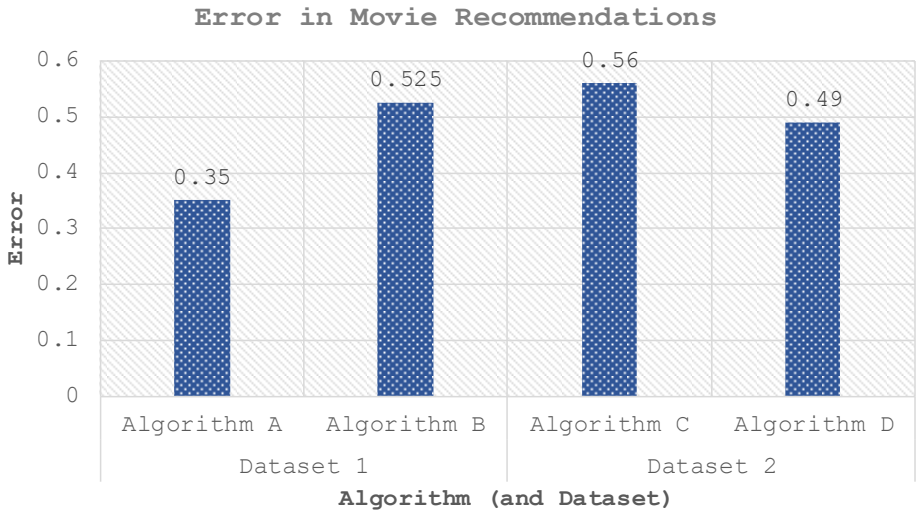
A. Géron, Hands on Machine Learning with scikit-learn and Tensorflow. O'Reilly Media, 2017.

Understanding Error Metrics

- **You need an algorithm for your movie recommender system (predict movie ratings; regression)**
- **You found two research papers. One paper compares algorithms A and B on dataset 1, and one paper compares algorithms C and D on dataset 2. Both measure error as MAE**
- **Which algorithm would you try first in your own recommender system?**



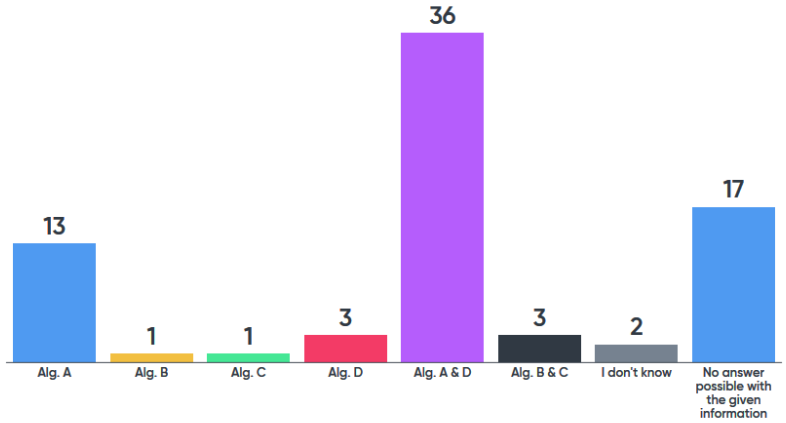
Mentimeter



Go to **www.menti.com** and use the code **51 13 86**

Which algorithm would you try first in your own recommender system?

Mentimeter



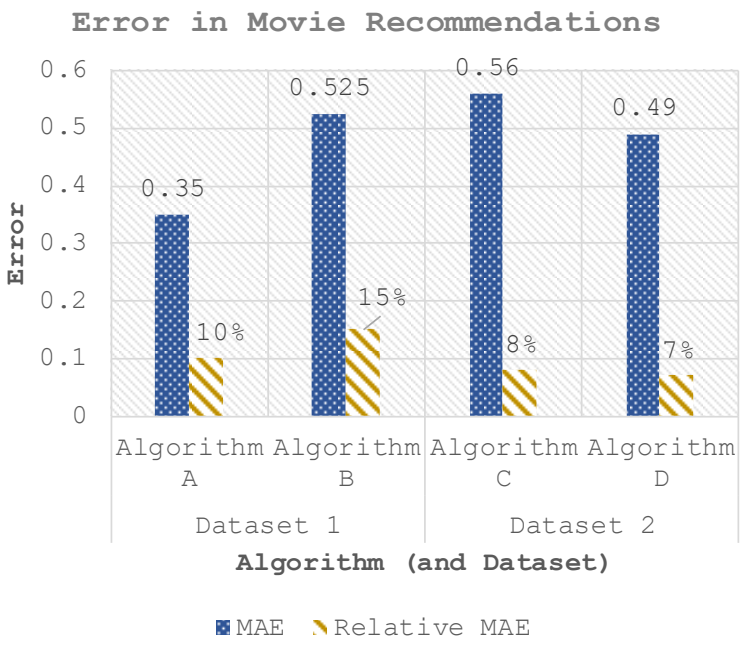
Slide is not active Activate

76

The data in detail

Dataset 1					
		Absolute Prediction Error		Relative Prediction Error	
Data Point	Actual Rating [1-5]	Algorithm A	Algorithm B	Algorithm A	Algorithm B
1	2	0.2	0.3	10%	15%
2	3	0.3	0.45	10%	15%
3	4	0.4	0.6	10%	15%
4	4	0.4	0.6	10%	15%
5	3	0.3	0.45	10%	15%
6	5	0.5	0.75	10%	15%
	(R)MAE	**0.35**	**0.525**	**10%**	**15%**

Dataset 2					
		Absolute Prediction Error		Relative Prediction Error	
Data Point	Actual Rating [1-10]	Algorithm C	Algorithm D	Algorithm C	Algorithm D
1	4	0.32	0.28	8%	7%
2	6	0.48	0.42	8%	7%
3	8	0.64	0.56	8%	7%
4	8	0.64	0.56	8%	7%
5	6	0.48	0.42	8%	7%
6	10	0.8	0.7	8%	7%
	(R)MAE	**0.56**	**0.49**	**8%**	**7%**



Classification/Ranking Metrics

- **Regression tasks usually can be evaluated as classification/ranking problem**
- **Build intervals and treat these as classes (and use classification algorithm instead of regression algorithm)**
- **Example**
  - Regression Task: Income prediction (the exact amount)
  - Classification Task: low income (€0-€10,000), medium income (€10,001-€30,000), high income (30,001-70,000), very high income (70,001+)

(Ranked) Retrieval Metrics

Mean Reciprocal Rank (MRR)

- **Measures at which rank the first relevant result is displayed.**
- **Reciprocal Rank of the first relevant result**
- **Mean Reciprocal Rank is the average of the Reciprocal Ranks**
- **MRR only cares about the first relevant result**
- **Typical Information Retrieval metric (search), not so much machine learning**

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

$rank_i = rank\ of\ first\ relevant\ result\ for\ query\ i$
$|Q| = number of\ search\ queries$

Mean Average Precision (MAP)

- **Average Precision (for one search query)**

$$\mathrm{AP}(Q_i) = \frac{1}{|R|}\sum_{j=1}^{|R|} p@k$$

Q_i = the i-*th* query
$|R| = number\ of\ relevant\ results$
$R = ranks\ of\ relevant\ results$
$p@k = precision\ at\ rank\ k\ (k \in R)$

- **Mean Average Precision (over all queries)**

$$\mathrm{MAP} = \frac{1}{|Q|}\sum_{i=1}^{|Q|} AP(Q_i)$$

$|Q| = number of\ search\ queries$

MAP Example

		Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Average Precision
Query 1	**Relevance**	Not relevant	Relevant	Not relevant	Not relevant	Relevant	(1/2 + 2/5) / 2 = 0.35
	Precision@rank_i	0/1 = 0	1/2 = 0.5	1/3 = 0.33	1/4 = 0.25	2/5 = 0.4	
Query 2	**Relevance**	Relevant	Relevant	Not relevant	Relevant	Not relevant	(1/1 + 2/2 + 3/4) / 3 = 0.92
	Precision@rank_i	1/1 = 1	2/2 = 1	2/3 = 0.67	3/4 = 0.75	3/5 = 0.6	
						MAP	(0.35 + 0.92) / 2 = 0.635

Normalized Discounted Cumulative Gain (nDGC)

- **More relevant items should be ranked higher than less relevant items → if an algorithm presents results of little relevance at the top, this should be punished more as if the results were presented at the bottom of the list.**
- **Cumulative gain = sum of top k items' relevancies**

$$CG_k = \sum_{i=1}^{k} rel_i \qquad rel_i = relevance\ at\ postion\ i$$

- **Discounted cumulative gain (DCG): penalizes relevant items that are ranked lower than they should**

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)} \qquad Alt.DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

- **Normalized Discounted Cumulative Gain (nDCG) = normalizes DCG to a value between 0 and 1**

$$nDCG_k = \frac{DCG_k}{IDCG_k} \qquad IDCG_k = Ideal\ DCG$$

Example

Item	True Relevance	IDCG =6.77
A	0.85	
B	0.80	
C	0.75	
D	0.40	
E	0.35	

Algorithm A		
Rank	**Item**	CG = 3.15
1	A	DCG = 6.77
2	B	nDCG = 1
3	C	
4	D	
5	E	

Algorithm B		
Rank	**Item**	CG = 3.15
1	E	DCG = 5.48
2	D	nDCG = 0.81
3	C	
4	B	
5	A	

More on nDCG

- **30-pages paper**
- **http://proceedings.mlr.press/v30/Wang13.pdf**

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Please submit your attendance https://www.scss.tcd.ie/doug.leith/CS4404/attendance.html or https://goo.gl/hN7ZnM

Baselines, Truth, and Gold Standards

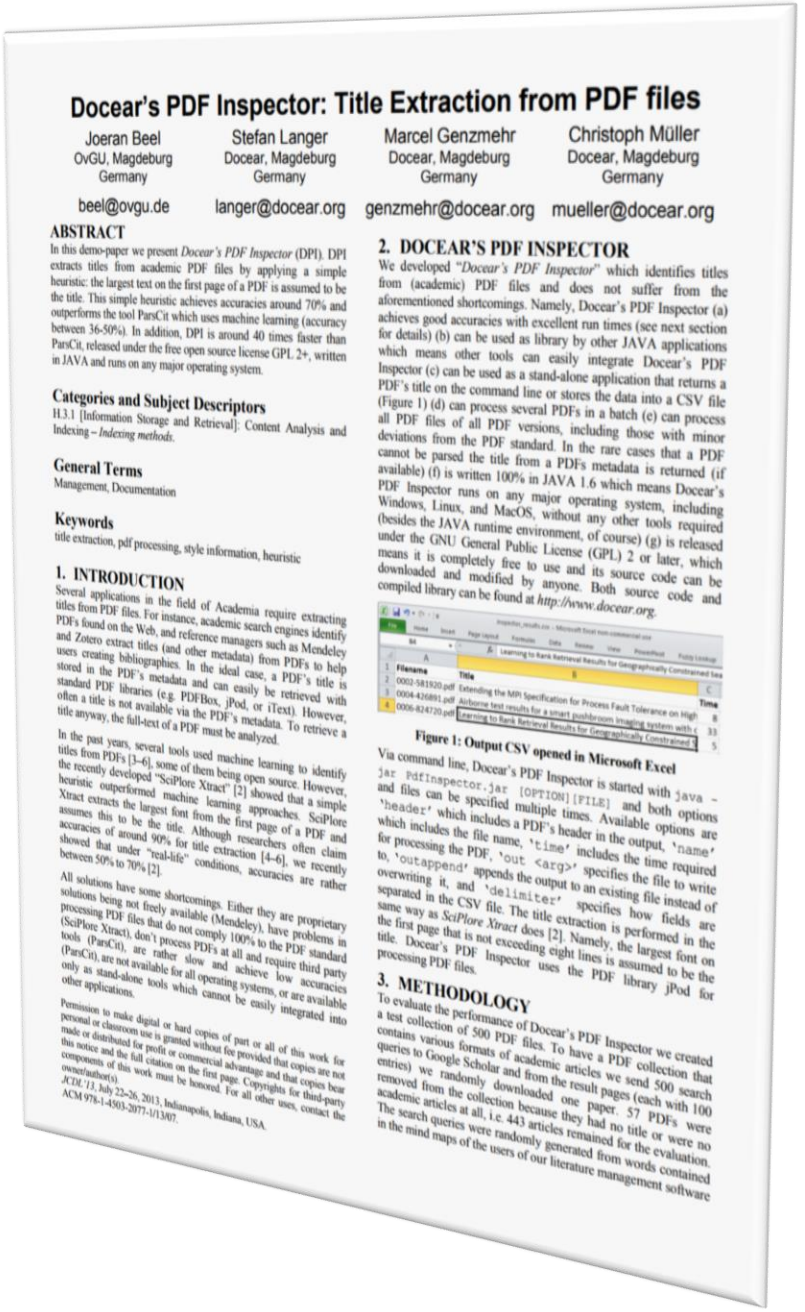
Baselines

Example

Title Detection from PDFs

„Our novel machine learning algorithm detects titles in PDF files with an accuracy of 76%“

Good or Bad?



Docear's PDF Inspector: Title Extraction from PDF files

Joeran Beel, OvGU, Magdeburg, Germany, beel@ovgu.de
Stefan Langer, Docear, Magdeburg, Germany, langer@docear.org
Marcel Genzmehr, Docear, Magdeburg, Germany, genzmehr@docear.org
Christoph Müller, Docear, Magdeburg, Germany, mueller@docear.org

ABSTRACT

In this demo-paper we present *Docear's PDF Inspector* (DPI). DPI extracts titles from academic PDF files by applying a simple heuristic: the largest text on the first page of a PDF is assumed to be the title. This simple heuristic achieves accuracies around 70% and outperforms the tool ParsCit which uses machine learning (accuracy between 36-50%). In addition, DPI is around 40 times faster than ParsCit, released under the free open source license GPL 2+, written in JAVA and runs on any major operating system.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *Indexing methods*.

General Terms

Management, Documentation

Keywords

title extraction, pdf processing, style information, heuristic

1. INTRODUCTION

Several applications in the field of Academia require extracting titles from PDF files. For instance, academic search engines identify PDFs found on the Web, and reference managers such as Mendeley and Zotero extract titles (and other metadata) from PDFs to help users creating bibliographies. In the ideal case, a PDF's title is stored in the PDF's metadata and can easily be retrieved with standard PDF libraries (e.g. PDFBox, jPod, or iText). However, often a title is not available via the PDF's metadata. To retrieve a title anyway, the full-text of a PDF must be analyzed.

In the past years, several tools used machine learning to identify titles from PDFs [3–6], some of them being open source. However, the recently developed "SciPlore Xtract" [2] showed that a simple heuristic outperformed machine learning approaches. SciPlore Xtract extracts the largest font from the first page of a PDF and assumes this to be the title. Although researchers often claim accuracies of around 90% for title extraction [4–6], we recently showed that under "real-life" conditions, accuracies are rather between 50% to 70% [2].

All solutions have some shortcomings. Either they are proprietary solutions being not freely available (Mendeley), have problems in processing PDF files that do not comply 100% to the PDF standard (SciPlore Xtract), don't process PDFs at all and require third party tools (ParsCit), are rather slow and achieve low accuracies (ParsCit), are not available for all operating systems, or are available only as stand-alone tools which cannot be easily integrated into other applications.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
ACM 978-1-4503-2077-1/13/07.

2. DOCEAR'S PDF INSPECTOR

We developed *"Docear's PDF Inspector"* which identifies titles from (academic) PDF files and does not suffer from the aforementioned shortcomings. Namely, Docear's PDF Inspector (a) achieves good accuracies with excellent run times (see next section for details) (b) can be used as library by other JAVA applications which means other tools can easily integrate Docear's PDF Inspector (c) can be used as a stand-alone application that returns a PDF's title on the command line or stores the data into a CSV file (Figure 1) (d) can process several PDFs in a batch (e) can process all PDF files of all PDF versions, including those with minor deviations from the PDF standard. In the rare cases that a PDF cannot be parsed the title from a PDFs metadata is returned (if available) (f) is written 100% in JAVA 1.6 which means Docear's PDF Inspector runs on any major operating system, including Windows, Linux, and MacOS, without any other tools required (besides the JAVA runtime environment, of course) (g) is released under the GNU General Public License (GPL) 2 or later, which means it is completely free to use and its source code can be downloaded and modified by anyone. Both source code and compiled library can be found at *http://www.docear.org*.

Figure 1: Output CSV opened in Microsoft Excel

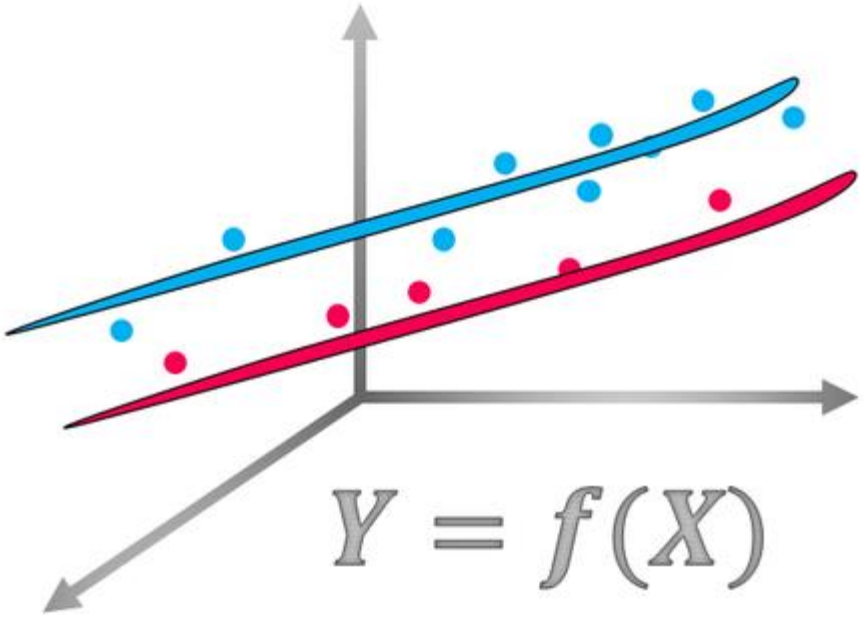
Via command line, Docear's PDF Inspector is started with `java -jar PdfInspector.jar [OPTION][FILE]` and both options and files can be specified multiple times. Available options are `'header'` which includes a PDF's header in the output, `'name'` which includes the file name, `'time'` includes the time required for processing the PDF, `'out <arg>'` specifies the file to write to, `'outappend'` appends the output to an existing file instead of overwriting it, and `'delimiter'` specifies how fields are separated in the CSV file. The title extraction is performed in the same way as *SciPlore Xtract* does [2]. Namely, the largest font on the first page that is not exceeding eight lines is assumed to be the title. Docear's PDF Inspector uses the PDF library jPod for processing PDF files.

3. METHODOLOGY

To evaluate the performance of Docear's PDF Inspector we created a test collection of 500 PDF files. To have a PDF collection that contains various formats of academic articles we send 500 search queries to Google Scholar and from the result pages (each with 100 entries) we randomly downloaded one paper. 57 PDFs were removed from the collection because they had no title or were no academic articles at all, i.e. 443 articles remained for the evaluation. The search queries were randomly generated from words contained in the mind maps of the users of our literature management software

Baselines

- **Alternative algorithm(s) to compare current model(s)**
  - State of the Art Algorithm / Currently used algorithm (e.g. in your company)
  - Simple Algorithm (e.g. linear regression)
  - Mean (or Median)
  - Most popular/frequent
  - Some simple rules
  - Random
- **Without a baseline, performance assessments of an algorithm are usually of little or no relevance**
- **Baseline gives your results meaning**



http://wiki.epidemium.cc/images/thumb/8/89/Logo2.png/515px-Logo2.png

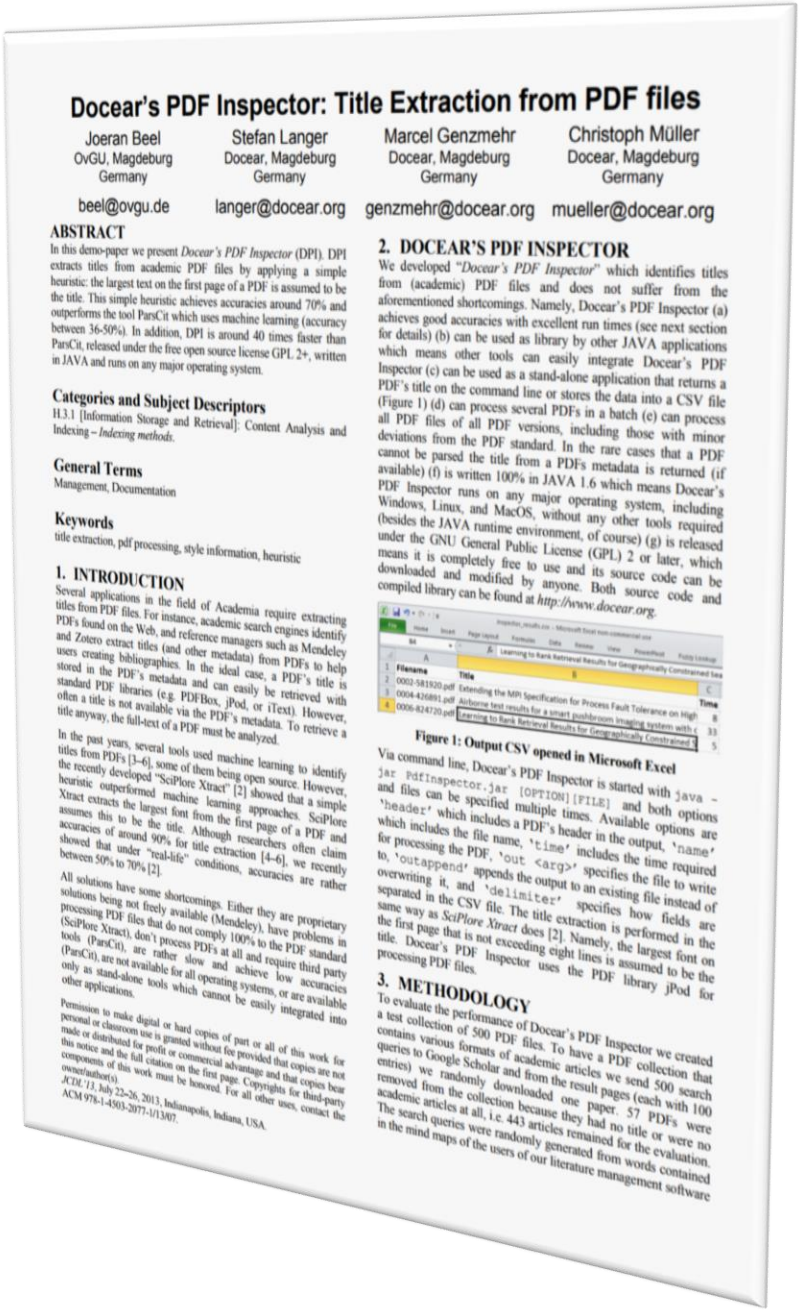
Example

Title Detection from PDFs

„Our novel machine learning algorithm detects titles in PDF files with an accuracy of 76%“

Good or Bad?

A simple rule: largest font on first page



Docear's PDF Inspector: Title Extraction from PDF files

Joeran Beel, OvGU, Magdeburg, Germany, beel@ovgu.de
Stefan Langer, Docear, Magdeburg, Germany, langer@docear.org
Marcel Genzmehr, Docear, Magdeburg, Germany, genzmehr@docear.org
Christoph Müller, Docear, Magdeburg, Germany, mueller@docear.org

ABSTRACT

In this demo-paper we present *Docear's PDF Inspector* (DPI). DPI extracts titles from academic PDF files by applying a simple heuristic: the largest text on the first page of a PDF is assumed to be the title. This simple heuristic achieves accuracies around 70% and outperforms the tool ParsCit which uses machine learning (accuracy between 36-50%). In addition, DPI is around 40 times faster than ParsCit, released under the free open source license GPL 2+, written in JAVA and runs on any major operating system.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *Indexing methods*.

General Terms

Management, Documentation

Keywords

title extraction, pdf processing, style information, heuristic

1. INTRODUCTION

Several applications in the field of Academia require extracting titles from PDF files. For instance, academic search engines identify PDFs found on the Web, and reference managers such as Mendeley and Zotero extract titles (and other metadata) from PDFs to help users creating bibliographies. In the ideal case, a PDF's title is stored in the PDF's metadata and can easily be retrieved with standard PDF libraries (e.g. PDFBox, jPod, or iText). However, often a title is not available via the PDF's metadata. To retrieve a title anyway, the full-text of a PDF must be analyzed.

In the past years, several tools used machine learning to identify titles from PDFs [3–6], some of them being open source. However, the recently developed "SciPlore Xtract" [2] showed that a simple heuristic outperformed machine learning approaches. SciPlore Xtract extracts the largest font from the first page of a PDF and assumes this to be the title. Although researchers often claim accuracies of around 90% for title extraction [4–6], we recently showed that under "real-life" conditions, accuracies are rather between 50% to 70% [2].

All solutions have some shortcomings. Either they are proprietary solutions being not freely available (Mendeley), have problems in processing PDF files that do not comply 100% to the PDF standard (SciPlore Xtract), don't process PDFs at all and require third party tools (ParsCit), are rather slow and achieve low accuracies (ParsCit), are not available for all operating systems, or are available only as stand-alone tools which cannot be easily integrated into other applications.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
ACM 978-1-4503-2077-1/13/07.

2. DOCEAR'S PDF INSPECTOR

We developed "*Docear's PDF Inspector*" which identifies titles from (academic) PDF files and does not suffer from the aforementioned shortcomings. Namely, Docear's PDF Inspector (a) achieves good accuracies with excellent run times (see next section for details) (b) can be used as library by other JAVA applications which means other tools can easily integrate Docear's PDF Inspector (c) can be used as a stand-alone application that returns a PDF's title on the command line or stores the data into a CSV file (Figure 1) (d) can process several PDFs in a batch (e) can process all PDF files of all PDF versions, including those with minor deviations from the PDF standard. In the rare cases that a PDF cannot be parsed the title from a PDFs metadata is returned (if available) (f) is written 100% in JAVA 1.6 which means Docear's PDF Inspector runs on any major operating system, including Windows, Linux, and MacOS, without any other tools required (besides the JAVA runtime environment, of course) (g) is released under the GNU General Public License (GPL) 2 or later, which means it is completely free to use and its source code can be downloaded and modified by anyone. Both source code and compiled library can be found at *http://www.docear.org*.

Figure 1: Output CSV opened in Microsoft Excel

Via command line, Docear's PDF Inspector is started with `java -jar PdfInspector.jar [OPTION][FILE]` and both options and files can be specified multiple times. Available options are `'header'` which includes a PDF's header in the output, `'name'` which includes the file name, `'time'` includes the time required for processing the PDF, `'out <arg>'` specifies the file to write to, `'outappend'` appends the output to an existing file instead of overwriting it, and `'delimiter'` specifies how fields are separated in the CSV file. The title extraction is performed in the same way as *SciPlore Xtract* does [2]. Namely, the largest font on the first page that is not exceeding eight lines is assumed to be the title. Docear's PDF Inspector uses the PDF library jPod for processing PDF files.

3. METHODOLOGY

To evaluate the performance of Docear's PDF Inspector we created a test collection of 500 PDF files. To have a PDF collection that contains various formats of academic articles we send 500 search queries to Google Scholar and from the result pages (each with 100 entries) we randomly downloaded one paper. 57 PDFs were removed from the collection because they had no title or were no academic articles at all, i.e. 443 articles remained for the evaluation. The search queries were randomly generated from words contained in the mind maps of the users of our literature management software

Continue from last lecture

Log-Loss

- „Soft“ measure for accuracy
- For probabilistic classifiers (e.g. probability of 0.73 that it‘s class A)
- Considers how close the predicted class was to true class
- E.g. if classifier incorrectly classifies instance as class A with probability of 0.51 that‘s still better than a probability of 0.93.
- log-loss = cross entropy between the distribution of the true labels and the predictions
- “Closely related to what’s known as the relative entropy, or Kullback–Leibler divergence

- Entropy measures the unpredictability of something.
- Cross entropy incorporates the entropy of the true distribution, plus the extra unpredictability when one assumes a different distribution than the true distribution.
- log-loss is an information-theoretic measure to gauge the “extra noise” that comes from using a predictor as opposed to the true labels.
- By minimizing the cross entropy, we maximize the accuracy of the classifier”
- Equal weight for false positives and false negatives

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}[\,(y_i * \log(p_i)) + ((1-y_i) * \log(1-p_i))]$$

$p_i = calculated\ probability\ (or\ confidence)\ that\ the\ ith\ data\ point\ belongs\ to\ class\ 1$
$y_i = true\ label\ of\ the\ ith\ instance\ [0|1]$
$N = number\ of\ instances$

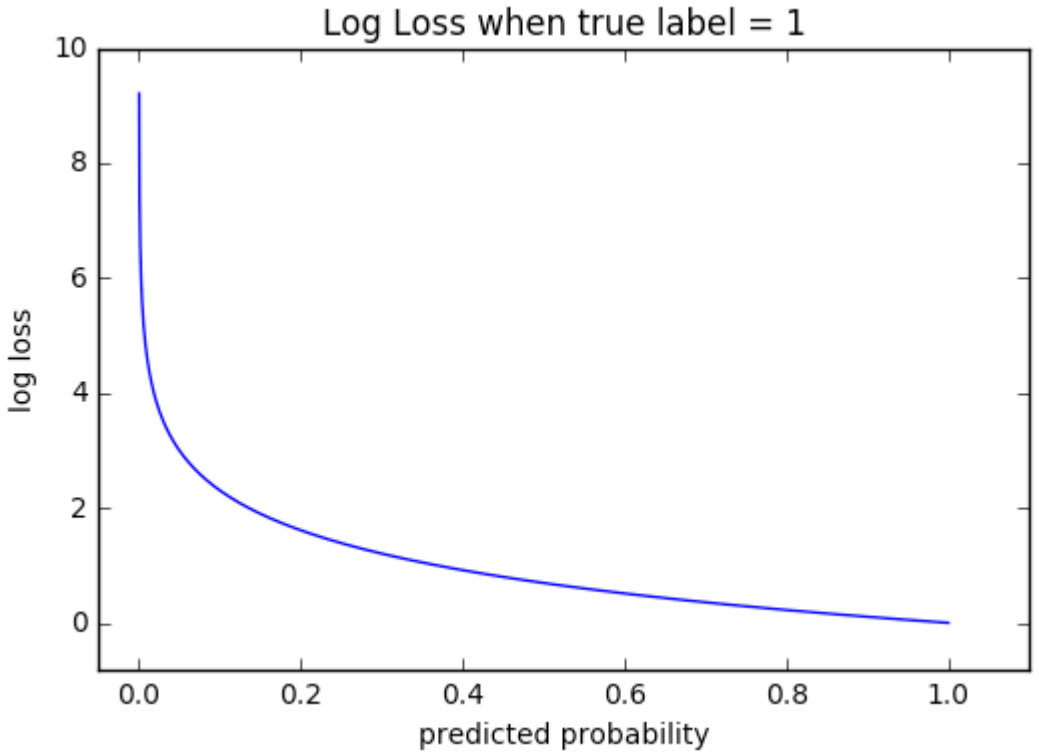
$$LogLoss_{MultiClass} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{i,j}\log(p_{i,j})$$

$N = number\ of\ instances$
$M = number\ of\ classes, i.e.\ labels$
$y_{i,j} = binary\ indicator\ if\ label\ j\ is\ correctly\ predicted\ for\ instance\ i\ [0/1]$
$p_{i,j} = probability\ that\ j\ is\ the\ correct\ label\ for\ i$

Alice Zheng, “Evaluating Machine Learning Models” (O’Reilly Media, Inc, 2015).

http://wiki.fast.ai/images/4/43/Log_loss_graph.png

Example

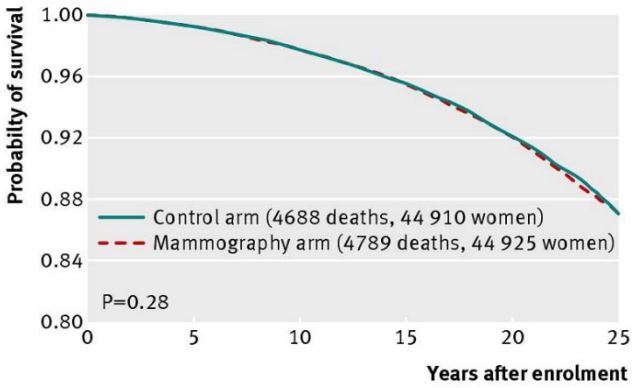


True Label	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
p_i	0.00000001	0.000001	0.0001	0.001	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.999
LogLoss	8	6	4	3	2	1	0.7	0.5	0.4	0.3	0.2	0.15	0.10	0.05	0.0004

True Label	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
p_i	0.001	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.999	0.99999	0.999999	0.99999999
LogLoss	0.0004	0.0044	0.05	0.10	0.15	0.22	0.3	0.4	0.5	0.7	1	3	5	6	8

Update: More Details on Metrics in Medicine

- **https://andrewgelman.com/2018/09/10/38592/**
- **https://andrewgelman.com/2017/09/02/cause-breast-cancer-specific-mortality-assignment-mammography-control/**
- **https://www.bmj.com/content/362/bmj.k3702**
- **https://www.nytimes.com/2009/11/17/health/17cancer.html**
- **JAMA https://www.health.harvard.edu/blog/new-mammography-guidelines-call-for-starting-later-and-screening-less-often-201510218466**
- **https://jamanetwork.com/journals/jama/fullarticle/2680553**
- **https://jamanetwork.com/journals/jama/fullarticle/2679928**



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Use Blackboard´s forum if a question may be relevant to other students, too. Email always both joeran.beel@scss.tcd and doug.leith@scss.tcd.ie. Give a meaningful subject, starting with "*[ML1819]*". No file attachments.

Week 04: Machine Learning Training & Evaluation

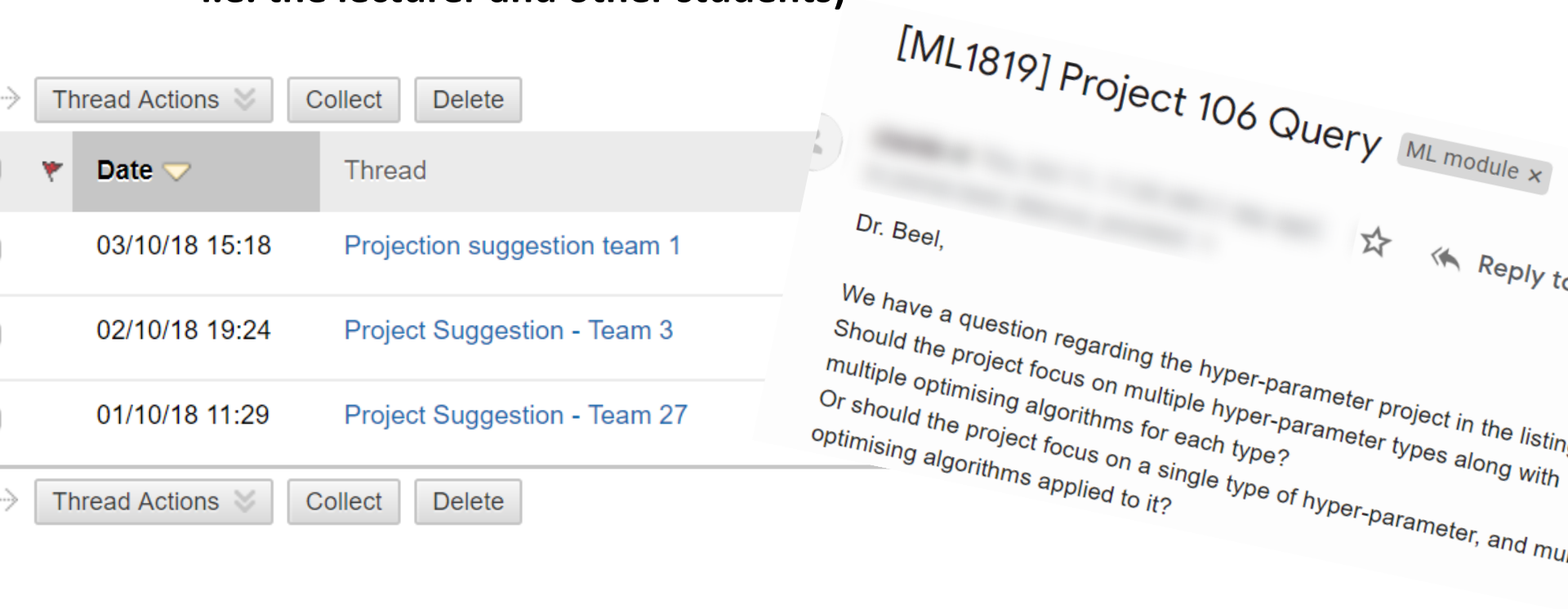
CS7CS4/CS4404 Machine Learning
v2 2018-10-08

Dr Joeran Beel
Assistant Professor in Intelligent Systems
Department of Computer Science and Statistics
Trinity College Dublin, Ireland

Dr Douglas Leith
Professor in Computer Systems
Department of Computer Science and Statistics
Trinity College Dublin, Ireland

Meaningful subjects

- **Unique and unambiguous (could not be send by another team or for another issue)**
- **Meaningful (covers the relevant issue – relevant for the receipient, i.e. the lecturer and other students)**



Ground Truth

Ground Truth

- **„Real truth“ can rarely be measured**
- **Ground truth is inferred/approximated**
- **The best possible measure available**
- **Examples**
  - Witnesses in a trial to find “truth”
  - User ratings to express user satisfaction
  - Purchase history to express user satisfaction
  - Citations to express impact of researchers
- **Often very difficult to find**

http://tmslboa.org/wp-content/uploads/2013/01/DSC_0003-1024x680.jpg

How to win a war?

- **In World War II, planes were a key resource. Often, US/UK planes heading to Germany were attacked, and many pilots did not return. The British were faced with the question whether to add extra armour to their planes. The only available information they had were the returning planes and bullet holes in the planes.**
- **Would you put extra armour to the planes and if so, where?**

1. No additional armour
2. Additional armour where many bullet holes are
3. Additional armour where no/few bullet holes are
4. Additional armour everywhere
5. No answer possible with the given information

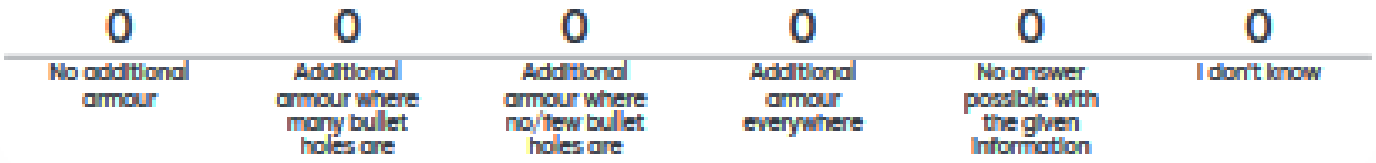
https://medium.com/@penguinpress/an-excerpt-from-how-not-to-be-wrong-by-jordan-ellenberg-664e708cfc3d

https://i.pinimg.com/originals/12/f6/26/12f62624ad000dc5887af8ac6df622f5.jpg

Go to **www.menti.com** and use the code **51 13 86**

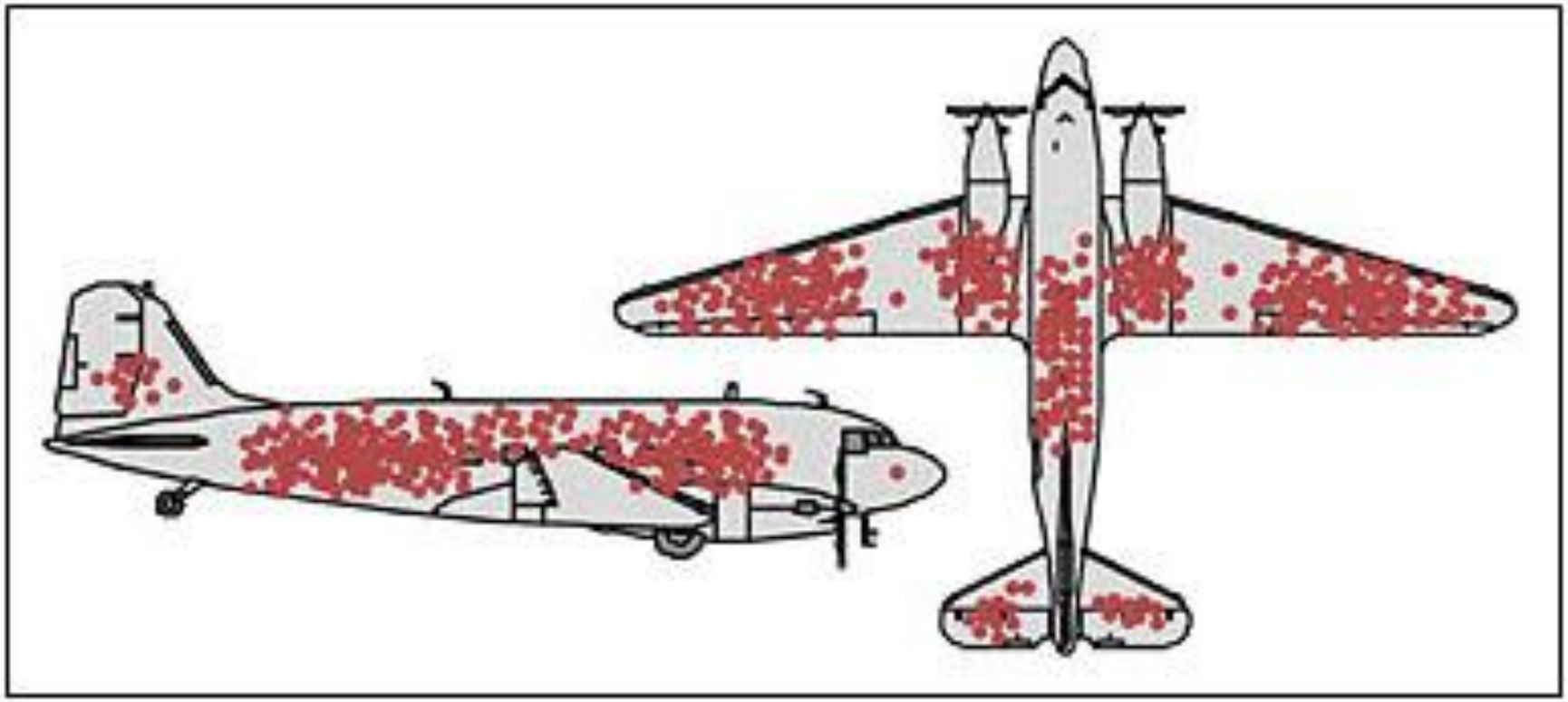
Would you add armour to the planes? If yes, where?

Mentimeter



Slide is not active **Activate**

0



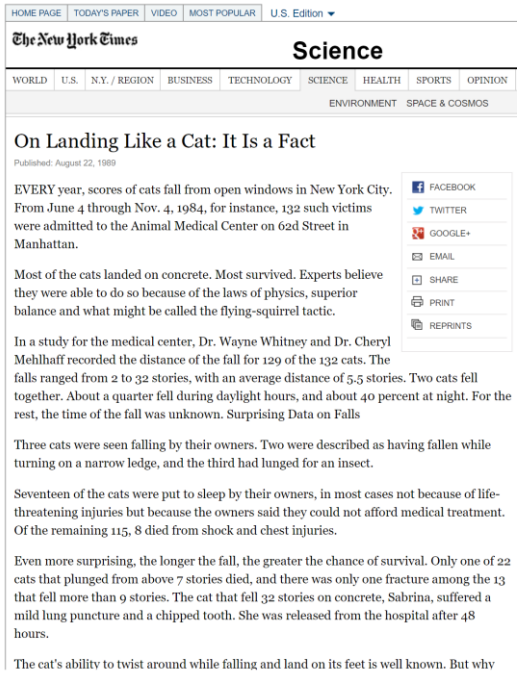
https://www.motherjones.com/wp-content/uploads/images/blog_raf_bullet_holes_0.jpg

Case Study (Homework)

Read this news article:

EVERY year, scores of cats fall from open windows in New York City. From June 4 through Nov. 4, 1984, for instance, 132 such victims were admitted to the Animal Medical Center on 62d Street in Manhattan. Most of the cats landed on concrete. Most survived. Experts believe they were able to do so because of the laws of physics, superior balance and what might be called the flying-squirrel tactic. In a study for the medical center, Dr. Wayne Whitney and Dr. Cheryl Mehlhaff recorded the distance of the fall for 129 of the 132 cats. The falls ranged from 2 to 32 stories, with an average distance of 5.5 stories. Two cats fell together. About a quarter fell during daylight hours, and about 40 percent at night. For the rest, the time of the fall was unknown. Three cats were seen falling by their owners. Two were described as having fallen while turning on a narrow ledge, and the third had lunged for an insect. Seventeen of the cats were put to sleep by their owners, in most cases not because of life-threatening injuries but because the owners said they could not afford medical treatment. Of the remaining 115, 8 died from shock and chest injuries. Even more surprising, the longer the fall, the greater the chance of survival. Only one of 22 cats that plunged from above 7 stories died, and there was only one fracture among the 13 that fell more than 9 stories. The cat that fell 32 stories on concrete, Sabrina, suffered a mild lung puncture and a chipped tooth. She was released from the hospital after 48 hours. The cat's ability to twist around while falling and land on its feet is well known. But why did cats from higher floors fare better than those on lower ones? One explanation is that the speed of the fall does not increase beyond a certain point, Dr. Mehlhaff and Dr. Whitney said in the December 1987 issue of The Journal of the American Veterinary Medical Association. This point, ''terminal velocity,'' is reached relatively quickly in the case of cats. Terminal velocity for a cat is 60 miles per hour; for an adult human, 120 m.p.h. Until a cat reaches terminal velocity, the two speculated, the cat reacts to acceleration by reflexively extending its legs, making it more prone to injury. But after terminal velocity is reached, they said, the cat might relax and stretch its legs out like a flying squirrel, increasing air resistance and helping to distribute the impact more evenly. ''Cats may be behaving like well-trained paratroopers,'' Dr. Jared Diamond, who teaches physiology at the University of California at Los Angeles Medical School, wrote in the August issue of the magazine Natural History.

How suitable is this data to train a machine learning algorithm to predict the survival of actually falling cats (i.e. “out-of-sample” cats) when given the time of the day, ground material, height (number of stories), and reason for the fall (e.g. lunged for an insect)? Ignore that the sample size is rather small.


HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

The New York Times — Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

ENVIRONMENT | SPACE & COSMOS

On Landing Like a Cat: It Is a Fact

Published: August 22, 1989

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

EVERY year, scores of cats fall from open windows in New York City. From June 4 through Nov. 4, 1984, for instance, 132 such victims were admitted to the Animal Medical Center on 62d Street in Manhattan.

Most of the cats landed on concrete. Most survived. Experts believe they were able to do so because of the laws of physics, superior balance and what might be called the flying-squirrel tactic.

In a study for the medical center, Dr. Wayne Whitney and Dr. Cheryl Mehlhaff recorded the distance of the fall for 129 of the 132 cats. The falls ranged from 2 to 32 stories, with an average distance of 5.5 stories. Two cats fell together. About a quarter fell during daylight hours, and about 40 percent at night. For the rest, the time of the fall was unknown. Surprising Data on Falls

Three cats were seen falling by their owners. Two were described as having fallen while turning on a narrow ledge, and the third had lunged for an insect.

Seventeen of the cats were put to sleep by their owners, in most cases not because of life-threatening injuries but because the owners said they could not afford medical treatment. Of the remaining 115, 8 died from shock and chest injuries.

Even more surprising, the longer the fall, the greater the chance of survival. Only one of 22 cats that plunged from above 7 stories died, and there was only one fracture among the 13 that fell more than 9 stories. The cat that fell 32 stories on concrete, Sabrina, suffered a mild lung puncture and a chipped tooth. She was released from the hospital after 48 hours.

The cat's ability to twist around while falling and land on its feet is well known. But why

http://www.nytimes.com/1989/08/22/science/on-landing-like-a-cat-it-is-a-fact.html

Ground Truth in Machine Learning

- **The training (and validation) labels are the ground truth**
- **The label is typically something being measured**
- **They may be wrong, biased, sparse, with noise …**

Example of a Poor Ground-Truth

Recommender System for Researchers

Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia

Joeran Beel
Trinity College Dublin
ADAPT Centre, Ireland
j@beel.org

Akiko Aizawa
National Institute of Informatics
Tokyo, Japan
aizawa@nii.ac.jp

Corinna Breitinger
University of Konstanz
Germany
corinna.breitinger@uni.kn

Bela Gipp
University of Konstanz
Germany
bela.gipp@uni.kn

ABSTRACT

Only few digital libraries and reference managers offer recommender systems, although such systems could assist users facing information overload. In this paper, we introduce Mr. DLib's recommendations-as-a-service, which allows third parties to easily integrate a recommender system into their products. We explain the recommender approaches implemented in Mr. DLib (content-based filtering among others), and present details on 57 million recommendations, which Mr. DLib delivered to its partner GESIS Sowiport. Finally, we outline our plans for future development, including integration into JabRef, establishing a living lab, and providing personalized recommendations.

KEYWORDS

recommender system, digital library, reference management software, recommendation-as-a-service, RaaS, API, web service

1 INTRODUCTION

Recommender systems in academia help researchers and scientists overcome information overload. However, only a few operators of academic services, such as digital libraries and reference managers, offer recommender systems to their users. The lack of recommender systems is possibly due to the high costs of implementing and maintaining such systems. Furthermore, operators may lack the expertise to design recommender systems. In this paper, we introduce Mr. DLib's recommendations-as-a-service (RaaS), which allows operators of academic services to easily integrate recommendations in their system. For the operator, the development and maintenance effort is minimal and no expertise in designing recommender systems is required.

Only a few other companies and organizations offer recommender-systems as-a-service for academia. BibTip [15] and bX [10] are commercial RaaS provider that offer co-occurrence-based recommendations. This approach is a generic recommendation approach applicable to a variety of items (documents, movies etc.), and suitable for rather large systems with many users [3]. Additionally, coverage is rather low, i.e. the approach recommends only a fraction of the documents in a library's catalogue. CORE [13, 14] and Babel [17] offer RaaS through an API, JavaScript client, and browser plug-in. Both services recommend open-access documents (CORE indexed approx. 68 million documents, Babel approx. 40 million), and Babel states that they are welcoming researchers to evaluate novel algorithms in Babel.

While these services have their strengths and weaknesses, only Mr. DLib offers all the following features: open source; not-for-profit; supportive of research, i.e. information about the RaaS, its architecture, and effectiveness is published [5, 7, 8, 11, 16]; integration into reference managers *and* digital libraries (of any size); recommendations for open-access articles *and* the option for partners to manage private document collections; capable of recommending all documents in a corpus (100% coverage).

2 MR. DLIB'S RECOMMENDER SYSTEM

Mr. DLib was originally developed as a Machine-readable Digital Library at the University of California, Berkeley and introduced at JCDL 2011 [2]. Since September 2016, Mr. DLib provides recommendations-as-a-service. The concept is illustrated in Figure 1. A user browses a web site of Mr. DLib's partner, e.g. a digital library. (1) When the user looks at a specific article's detail page, the web site requests a list of related articles from Mr. DLib's RESTful Web Service as a HTTP GET request:

GET /v1/documents/{document_id}/related_documents/ [1]

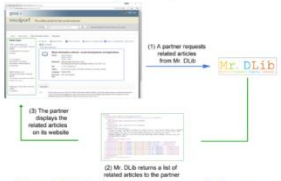


Figure 1: Illustration of the recommendation process

(2) Upon receiving the request, Mr. DLib computes a list of related articles, and returns the list as XML. (3) The partner website converts the XML to HTML and displays the recommendations on its web page (or mobile app, or desktop application).[2] Before this process begins, Mr. DLib indexes the metadata of the partner's documents (title, authors, abstract, venue, keywords). Mr. DLib uses Apache Lucene/Solr's *More-Like-This* function to calculate document relatedness. We are also experimenting with alternative recommendation approaches, such as stereotype and most-popular recommendations [5]. Additionally, we are

[1] Example: https://api.mr-dlib.org/v1/documents/gesis-bib-136994/related_documents/

[2] Example: http://sowiport.gesis.org/search/id/gesis-bib-136994

experimenting with key-phrase extraction and a bibliometric re-ranking based on readership data from Mendeley [7, 8, 16].

The production system api.mr-dlib.org and development system api-dev.mr-dlib.org each run on a dedicated server (i7-6700k, 32GB RAM, SSD hard drives). The development system is also used for resource intensive tasks, including document indexing, key-phrase extraction, and the calculation of bibliometrics. The uptime of the servers is constantly monitored[3], and the average response time to deliver recommendations is 682ms. We further run a beta system api-beta.mr-dlib.org on a virtual machine (4 cores, 14 GB RAM).

3 MR. DLIB'S DIGITAL LIBRARY PARTNER

Mr. DLib's first partner is the digital library Sowiport[4], which is Germany's largest social science repository, operated by the GESIS institute [12]. Mr. DLib has indexed around 10 million documents from Sowiport. While GESIS agreed to let their documents be recommended to all partners of Mr. DLib, GESIS chose to recommend only its own content on Sowiport.

Between September 18th 2016 and February 11th 2017, Mr. DLib delivered 57,435,086 recommendations to Sowiport. Users clicked on 77,468 recommendations. This equates to an overall click-through rate (CTR) of 0.13%. This CTR is rather low, and, as shown in Figure 2, there was a notable variance (e.g. 0.19% in September and 0.10% in December). The variance may be caused by different algorithms used. In addition, recommendations are delivered when web spiders, such as the Google Bot, crawl the Sowiport website. In contrast, clicks are logged with JavaScript, which is usually not executed by web spiders.

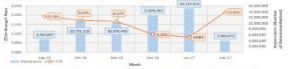


Figure 2: Recommendations and CTR by Month

4 LICENSE AND POLICY

Mr. DLib advocates an 'open culture' and publishes its code as open source on GitHub[5]. Project details are described in a public WIKI (Confluence)[6], and issues managed in a public ticket tracker (JIRA) [7]. Data from our research is published on Harvard's Dataverse[8] (if we can ensure privacy and copyrights of our partners). We invite other researchers to evaluate their recommendation algorithms with Mr. DLib and our partners[9].

5 OUTLOOK

In the future, we plan to add more partners (e.g. JabRef [11] and Docear [1, 4, 6]), import more documents (e.g. CORE [14]), improve the recommendation quality (e.g. cross-language recommendations [18]), and make them more personalized. We will also implement a JavaScript client that allows an easier integration of Mr. DLib's recommender system into partner websites, and monitors the number of delivered and clicked recommendations more reliably. In the long term, we will extend the scope of Mr. DLib to recommend not only related articles, but also other items relevant in academia, such as calls for papers and research grants, as well as recommending researchers as potential collaborators. In addition, we plan to establish an interactive evaluation task [9] that allows other researchers to evaluate new recommendation approaches in Mr. DLib's recommender system.

5 ACKNOWLEDGEMENTS

This work was supported by a fellowship within the FITweltweit programme of the German Academic Exchange Service (DAAD) and a scholarship of the Carl Zeiss Foundation. This publication has also emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 13/RC/2106. We are further grateful for the support provided by Zeljko Carevic, Philipp Mayr, Siddharth Dinesh, Sophie Siebert, Stefan Feyer, Sara Mahmoud, Gabor Neusch, and Felix Beierle.

REFERENCES

[1] Beel, J. et al. 2011. Docear: An Academic Literature Suite for Searching, Organizing and Creating Academic Literature. *11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (2011), 465–466.

[2] Beel, J. et al. 2011. Introducing Mr. DLib, a Machine-readable Digital Library. *Proceedings of the 11th ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)* (2011), 463–464.

[3] Beel, J. et al. 2016. Research Paper Recommender Systems: A Literature Survey. *International Journal on Digital Libraries*. 4 (2016), 305–338.

[4] Beel, J. et al. 2009. SciPlore MindMapping' - A Tool for Creating Mind Maps Combined with PDF and Reference Management. *D-Lib Magazine*. 15, 11.

[5] Beel, J. et al. 2017. Stereotype and Most-Popular Recommendations in the Digital Library Sowiport. *Proceedings of the 15th International Symposium of Information Science (ISI)* (2017).

[6] Beel, J. et al. 2014. The Architecture and Datasets of Docear's Research Paper Recommender System. *D-Lib Magazine*. 20, 11/12 (2014).

[7] Beel, J. and Dinesh, S. 2017. Real-World Recommender Systems for Academia: The Gain and Pain in Developing, Operating, and Researching them. *Fifth Workshop on Bibliometric-enhanced Information Retrieval (BIR)*. 6–17.

[8] Beierle, F. et al. 2017. Exploring Choice Overload in Related-Article Recommendations in Digital Libraries. *5th International Workshop on Bibliometric-enhanced Information Retrieval (BIR)*.

[9] Bogers, T. et al. 2017. Lessons Learned from the CHiC and SBS Interactive Tracks: A Wishlist for Interactive IR Evaluation. *CHIIR 2017 Second Workshop on Supporting Complex Search Tasks (SCST 2017)* (2017), 11–14.

[10] ExLibris 2013. bX Usage-Based Services transform your discovery experience! *Web page*. http://www.exlibrisgroup.com/category/bXUsageBasedServices.

[11] Feyer, S. et al. 2017. Integration of the Scientific Recommender System Mr. DLib into the Reference Manager JabRef. *Proceedings of the 39th European Conference on Information Retrieval (ECIR)* (2017).

[12] Hienert, D. et al. 2015. Digital Library Research in Action – Supporting Information Retrieval in Sowiport. *D-Lib Magazine*. 21, 3/4 (2015).

[13] Knoth, P. et al. 2017. Towards effective research recommender systems for repositories. *Proceedings of the Open Repositories Conference* (2017).

[14] Knoth, P. and Zdrahal, Z. 2012. CORE: three access levels to underpin open access. *D-Lib Magazine*. 18, 11/12 (2012).

[15] Monnich, M. and Spiering, M. 2008. Einsatz von BibTip als Recommendersystem in Bibliothekskatalog. *Bibliotheksdienst*. 42, 1 (2008).

[16] Siebert, S. et al. 2017. Extending a Research Paper Recommendation System with Bibliometric Measures. *5th International Workshop on Bibliometric-enhanced Information Retrieval (BIR)*.

[17] Wesley-Smith, I. and West, J.D. 2016. Babel: A Platform for Facilitating Research in Scholarly Article Discovery. *Proceedings of the 25th International Conference Companion on World Wide Web* (2016), 389–394.

[18] Zhou, D. et al. 2016. A study of user profile representation for personalized cross-language information retrieval. *Aslib Journal of Information Management*. 68, 4 (2016), 448–477.

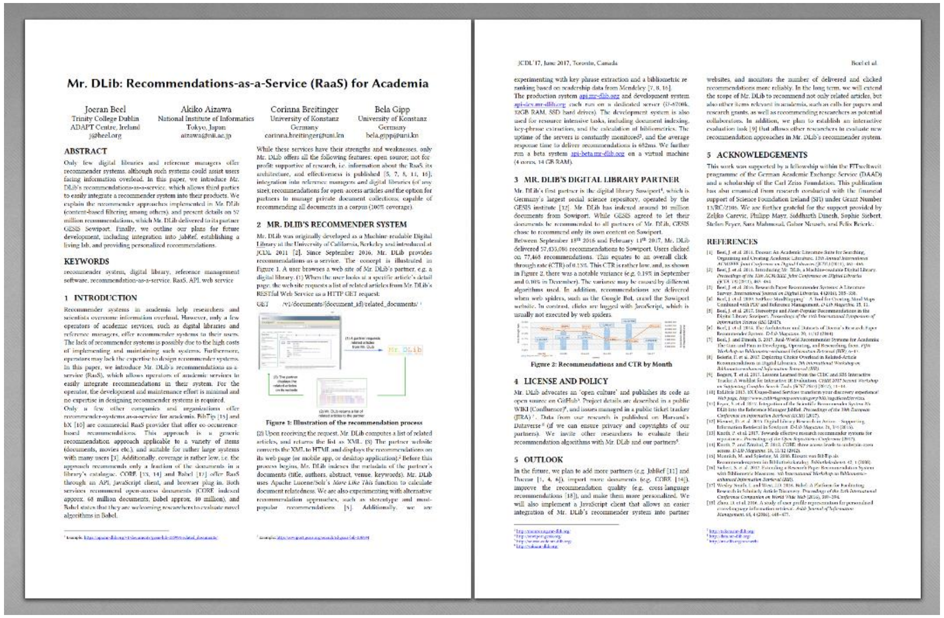
[3] http://monitoring.mr-dlib.org/
[4] http://sowiport.gesis.org/
[5] http://source-code.mr-dlib.org/
[6] http://wiki.mr-dlib.org/
[7] http://tickets.mr-dlib.org/
[8] http://data.mr-dlib.org/
[9] http://mr-dlib.org/research/

Problems

The assumption that existing reference lists are a good ground truth implies that authors know the literature very well and cited the best papers. This is not correct.

- **Authors do not always cite for „honest“ reasons**
- **Authors do not all relevant literature (otherwise, they wouldn’t need a recommender system)**

Recommender system can only be as good as the authors; never better


Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia

Gold Standard

- **Analog to monetary gold standard that allows comparing the value of currencies (Rudd, 1979)**
- **Best method *or* data (under reasonable conditions)**
  - Data: (Largest) Dataset with the most accurate ground-truth
  - Method: Best performing method under reasonable circumstances
- **Medical Example (Method): Cancer Screening**
  - Ideal method: Autopsy
  - Gold Standard Method: e.g. x-ray (best alternative method that keeps patient alive)
- **Machine Learning Example (Data)**
  - Many datasets (ground-truths) that annotate the relevance of documents and search queries
  - One best dataset (the gold standard)

https://www.ncbi.nlm.nih.gov/pubmed/443963

Beyond Accuracy (also in Academia)

- **Minimize harm**
- **Serendipity**
- **Diversity**
- **Novelty**
- **Coverage**

http://gaus.ee/wp-content/uploads/2016/09/image-1044369-breitwandaufmacher-xpfa-1044369.jpg

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Key Performance Indicators (KPIs)

Goals In the business world

- **Maximize profits**
- **Maximize Revenues**
- **Minimize Costs (labour, hardware, licenses, legal…)**
- **Gain as many users as possible (and retain as many existing users as possible)**
- **Maximize User Satisfaction**
- **Have the best product**
  - The most effective one
  - The cheapest (with acceptable effectiveness)
  - The best value for money

Costs

- **Labor**
- **Servers**
- **Legal / Licenses**
- **…**
- **Costs are almost never considered in Academia**

WIRED Netflix Never Used Its $1 Million Algorithm Due T

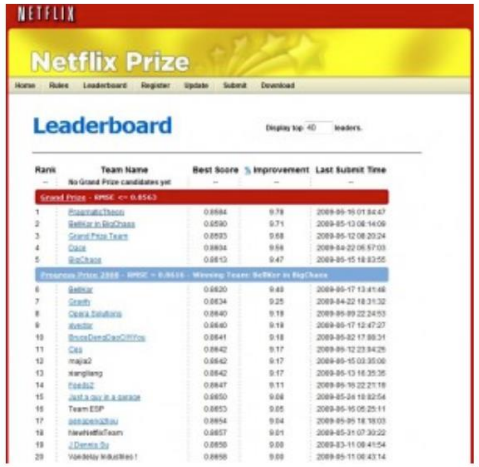
BUSINESS | CULTURE | DESIGN | GEAR | SCI

SHARE

11

CASEY JOHNSTON, ARS TECHNICA BUSINESS 04.16.12 08:20 AM

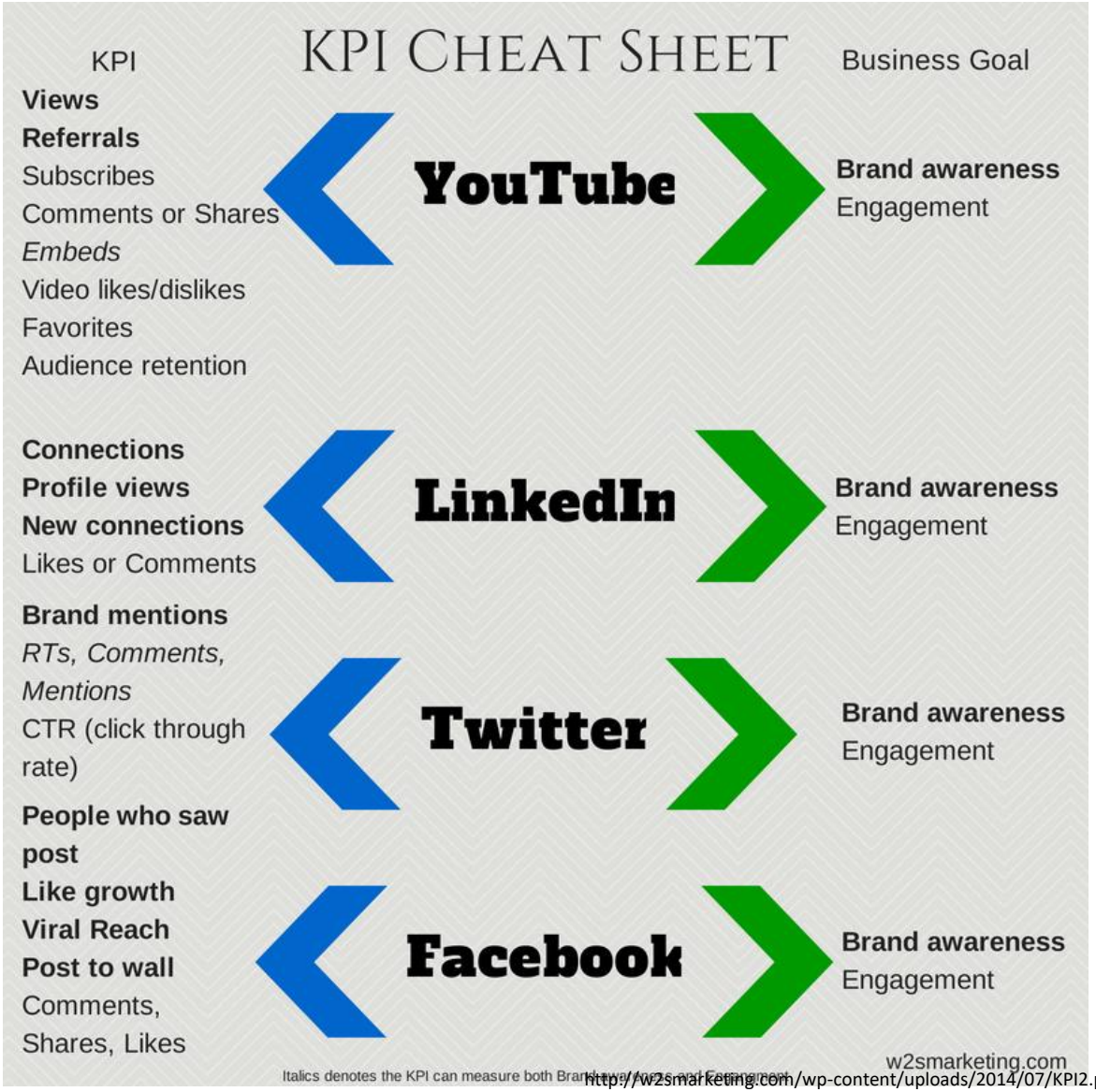
NETFLIX NEVER USED ITS $1 MILLION ALGORITHM DUE TO ENGINEERING COSTS



Netflix awarded a $1 million prize to a developer team in 2009 for an algorithm that increased the accuracy of the company's recommendation engine by 10 percent. But it doesn't use the million-dollar code, and has no

https://www.wired.com/2012/04/netflix-prize-costs/

Key Performance Indicator (KPI)



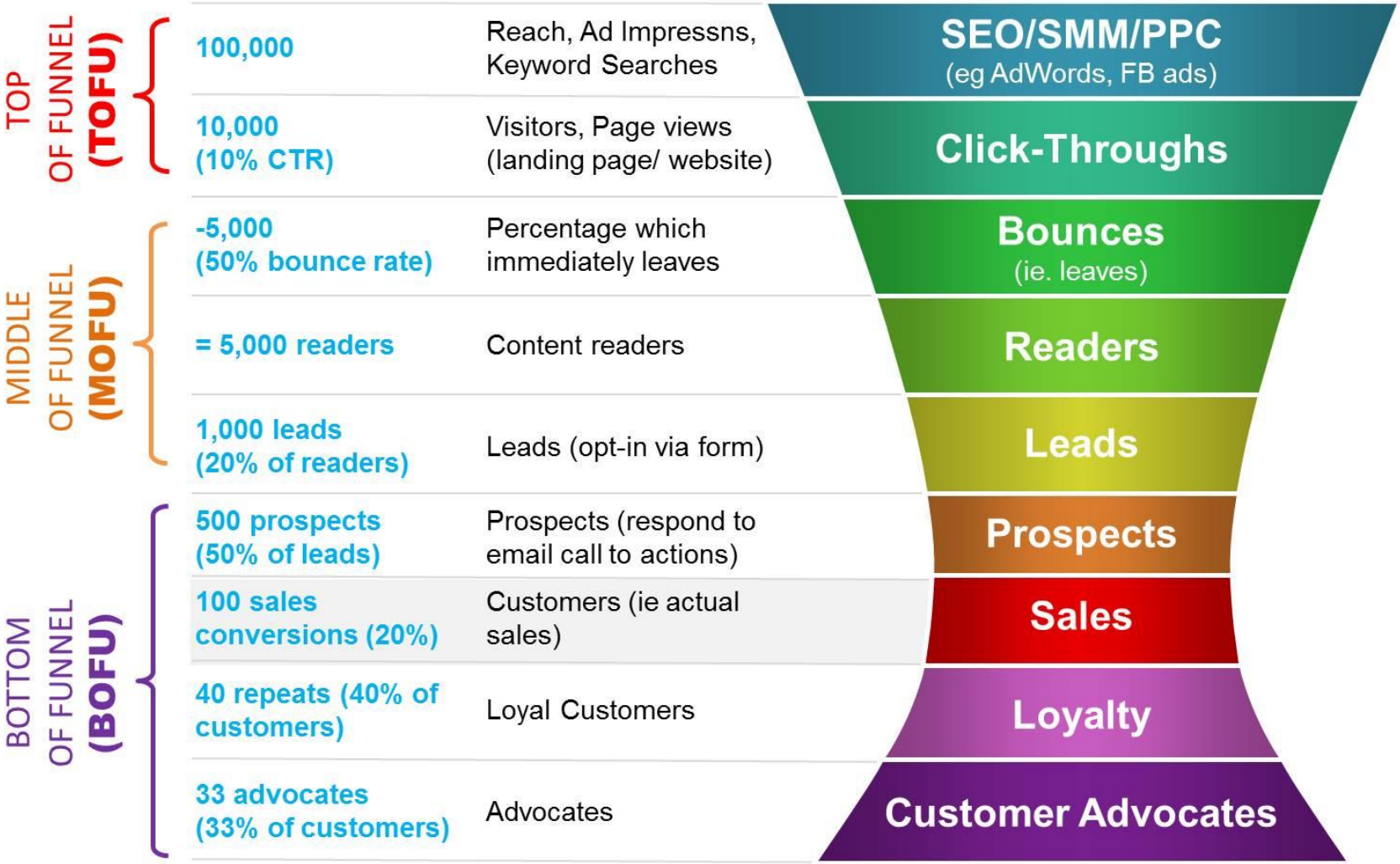
http://w2smarketing.com/wp-content/uploads/2014/07/KPI2.png

KPIs

Business Objective	Common KPIs	Customer Touchpoints Ask customers for feedback here
 Increase Revenues	Increase conversion rate	On your product recommendations
	Lower shopping cart abandonment rate	In your product pages
	Increase average order value	Throughout the checkout process
 Improve Customer Satisfaction	Improve Net Promoter Score	Whenever the customer is asked to complete a task (ex. log in)
	Lower Customer Effort Score	Following a customer service interaction
	Improve task completion rate	Across all of your other customer touchpoints
	Increase time on site	
 Increase Customer Loyalty	Improve Net Promoter Score / Customer Satisfaction Rating	After they have logged into your site or mobile app
	Increase total customer interaction with the brand (across channels)	In your customer retention emails
	Increase click through rate on personalized content	In your social media pages
	Improve support response time	Following a customer service interaction
 Lower Bounce Rate	Increase click through rate	In your support pages
	Improve campaign performance	On pages in your website with high bounce rates
		In your customer forums
 Improve Customer Service	Improve contact center satisfaction ratings	In your support and forum pages
	Decrease customers support request rate through self-support and automation	In your social media pages
 Improve In-Branch Sales	Decrease time from session start to product pages on mobile devices	In your mobile site and app
	Improve product recommendations on mobile devices	In your retail stores
	Improve product recommendations on mobile devices	
	Improve in-store customer satisfaction ratings	

https://fonolo.com/wp-content/uploads/2015/11/Customer-Experience-Cheat-Sheet2.png

Digital Marketing Funnel (Analytics)



http://coolerinsights.com/wp-content/uploads/2016/04/Digital-Marketing-Funnel-Analytics-.jpg

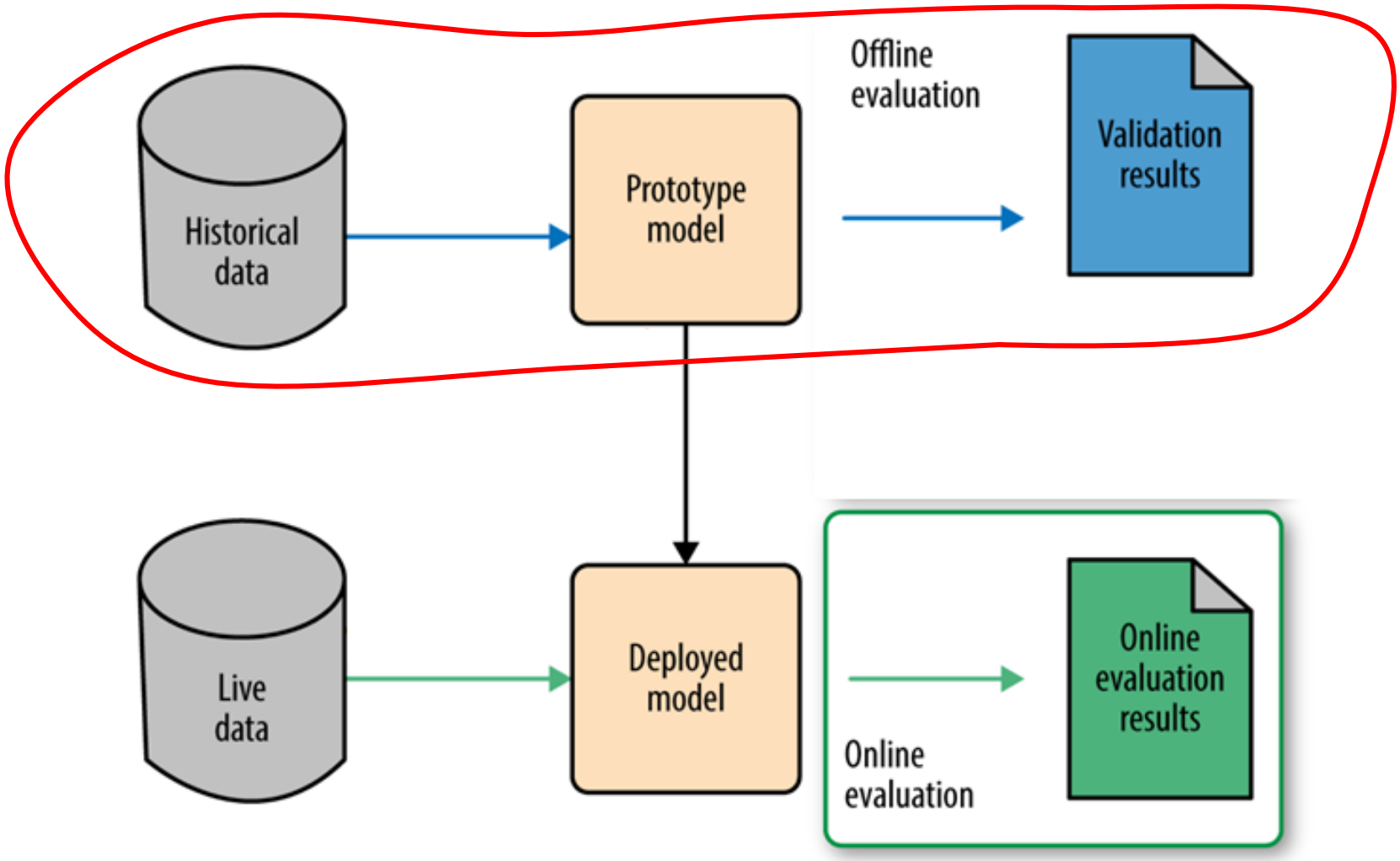
https://blog.awaremanager.com/wp-content/uploads/2017/07/cheatsheet.png

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Online Evaluations

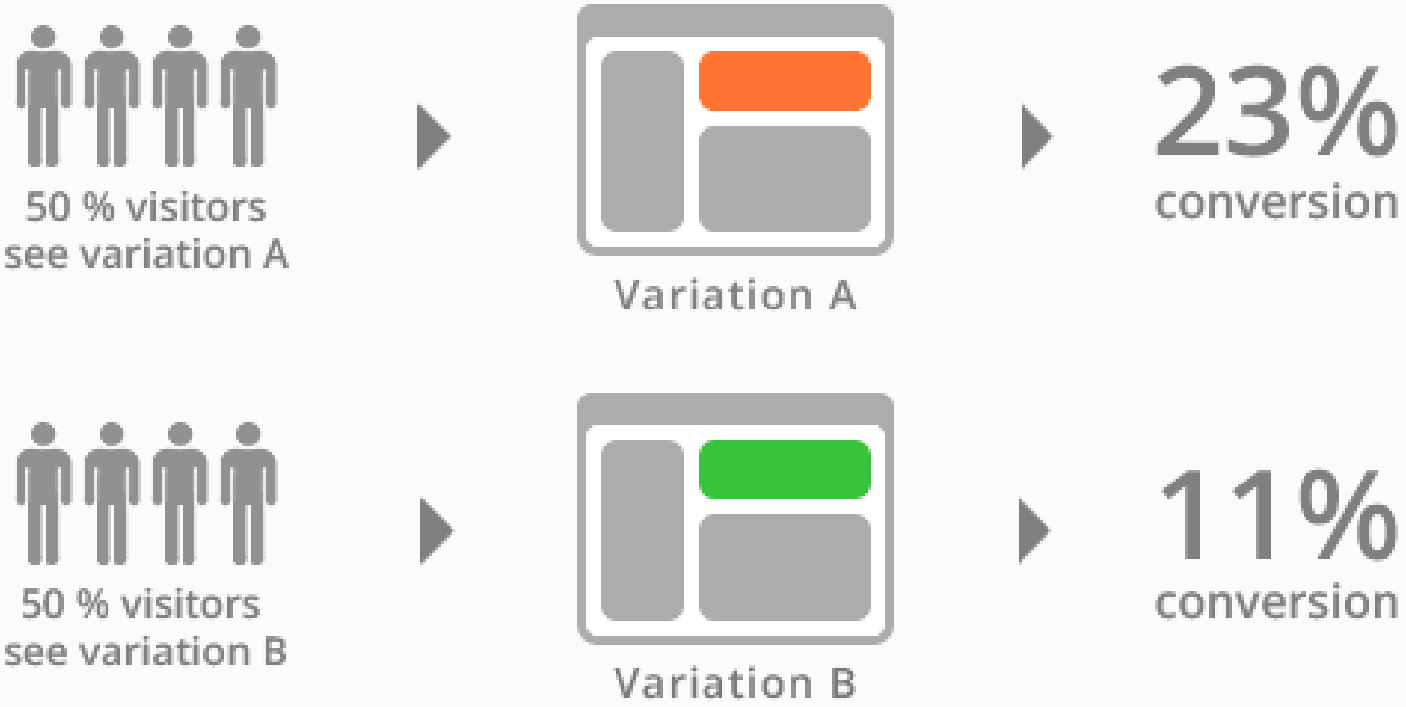
Evaluation Types: Offline and Online

Typical Focus of ML Lectures and Books



Alice Zheng, "Evaluating Machine Learning Models" (O'Reilly Media, Inc, 2015).

A/B Tests

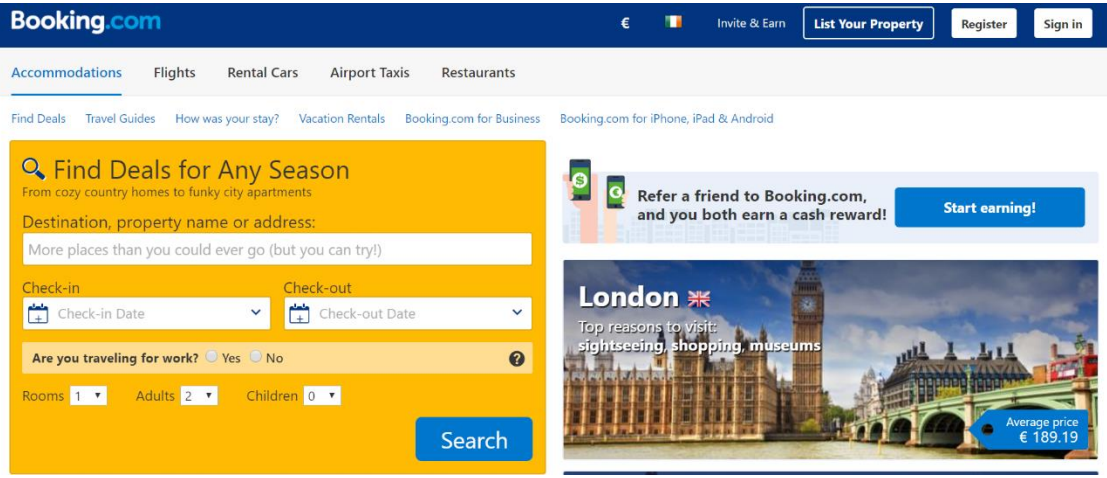


https://d5bygqdtbohob.cloudfront.net/wp-content/themes/vwo/images/ab-test-guide/what-ab-test.png

http://coolerinsights.com/wp-content/uploads/2016/04/Digital-Marketing-Funnel-Analytics-.jpg

A, B, C, D … Test

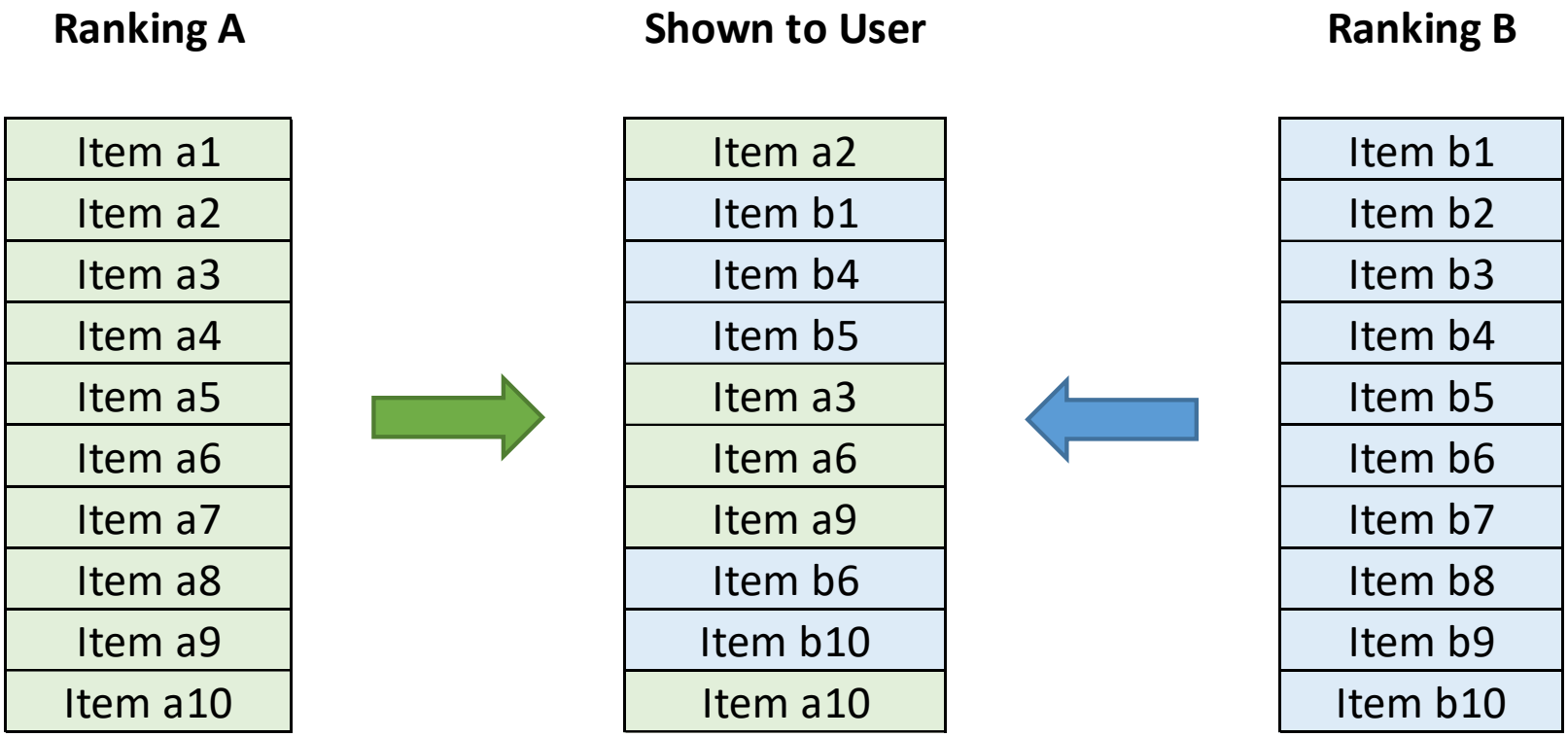
"In the testing phase, there are 1000 experiments running at any time. To put it into perspective, there are more versions of the Booking.com website live than there are humans that have ever lived."



https://blog.taplytics.com/how-booking-com-a-b-tests-like-nobodys-business-8158fd75d6b6

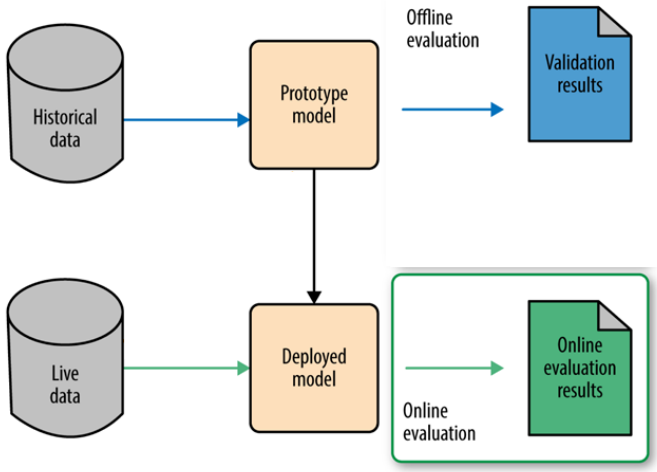
Interleaving

- **Rankings are mixed**
- **All kind of variations**
  - Random mix
  - Top k mix
  - Fixed amount mix



Offline vs. Online Evaluation

- **Offline evaluation**
  - Measures success on historical data
  - Uses metrics like accuracy, RMSE, precision, …
- **Online evaluation**
  - Measures success on live data
  - Often uses metrics like click-through rate, conversion rate, likes
- **(User Studies)**
- **Live data (and metrics) can also be used for offline evaluations**



Alice Zheng, "Evaluating Machine Learning Models" (O'Reilly Media, Inc, 2015).

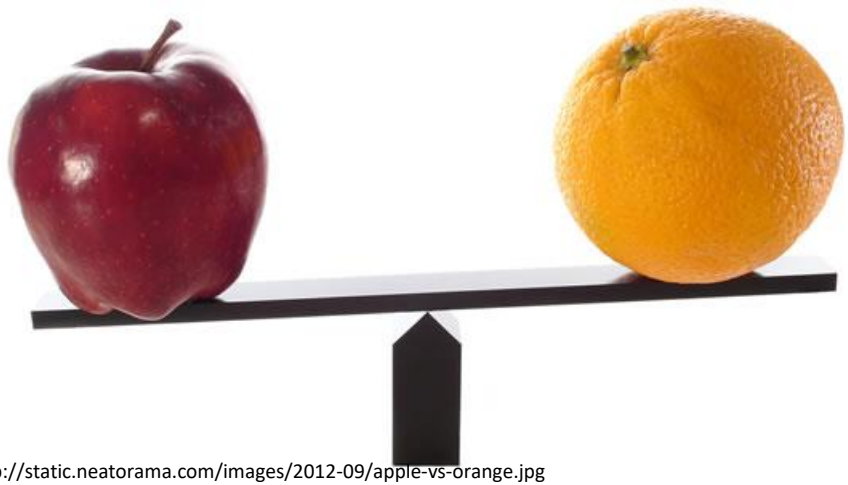
Correlation of Offline and Online Metrics

Correlation of Ratings and …	
Ratings	--
CTR	0.78
CTR (Set)	0.78
DTR	0.65
ATR	0.61
CiTR	0.52
P@3	0.62
P@10	0.65
MRR	0.55
nDCG	0.67

Correlation of CTR and …	
Ratings	0.78
CTR	--
CTR (Set)	0.97
DTR	0.73
ATR	0.53
CiTR	0.42
P@3	0.41
P@10	0.48
MRR	0.30
nDCG	0.28

Correlation of …	
P@3 and P@10	0.92
P@10 and MRR	0.56
P@10 and nDCG	0.55
nDCG and MRR	0.71

Joeran Beel, "Towards Effective Research-Paper Recommender Systems and User Modeling based on Mind Maps," PhD Thesis. Otto-von-Guericke Universität Magdeburg (2015).



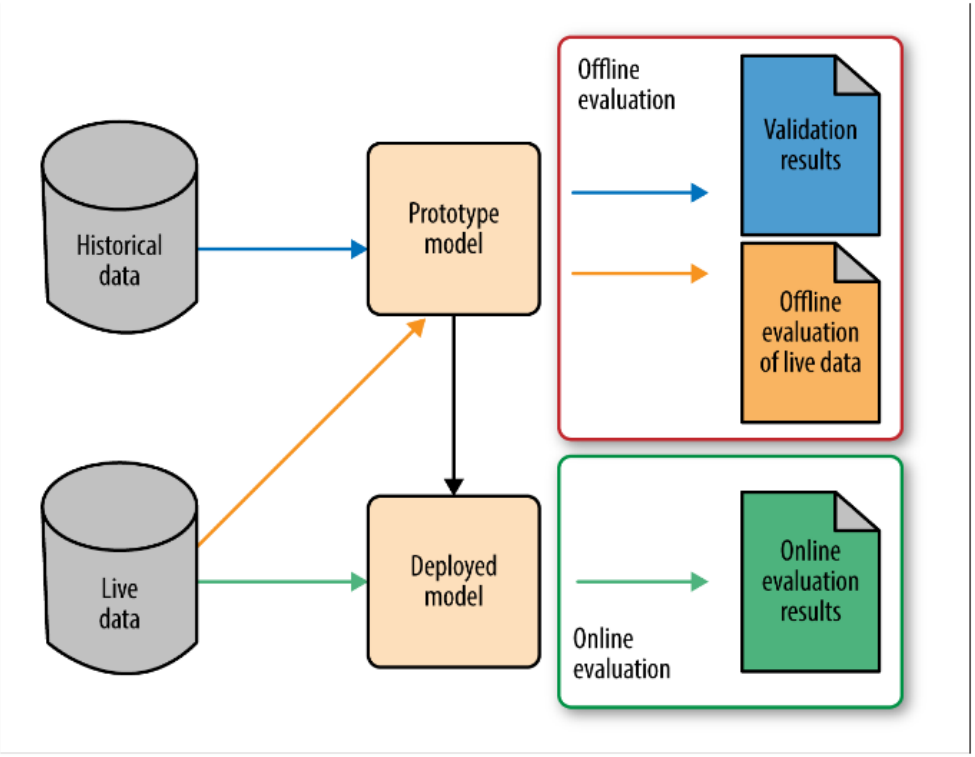
http://static.neatorama.com/images/2012-09/apple-vs-orange.jpg

Pros & Cons Of Online Evaluations

- **Pros**
  - Typically most relevant evaluation type
- **Cons**
  - Takes lots of time to run
  - Implementation effort is high
  - Only limited number of tests possible

Distribution Shift

- **Assumption of offline evaluation:**
  - Data is stationary (not changing over time)
  - Model that performs well in offline evaluation will perform well in live system
- **Often, not realistic**
- **Particularly a problem for researchers**
  - They work with old data
  - They assume their findings will generalize to environments with different data
- **Can be measured by comparing difference between offline evaluation and online evaluation performance**
- **If there is a big discrepancy -> update offline data, retrain, and/or re-engineer**



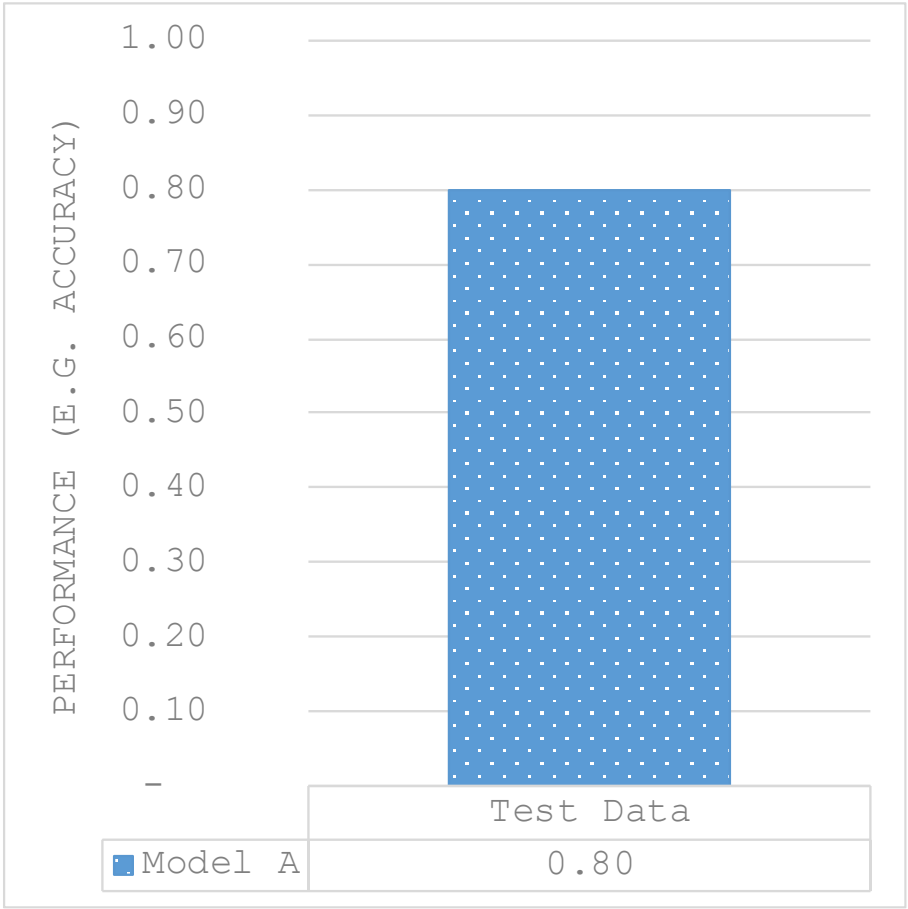
Alice Zheng, "Evaluating Machine Learning Models" (O'Reilly Media, Inc, 2015).

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Statistical Significance & Confidence Intervals

How reliable are results?

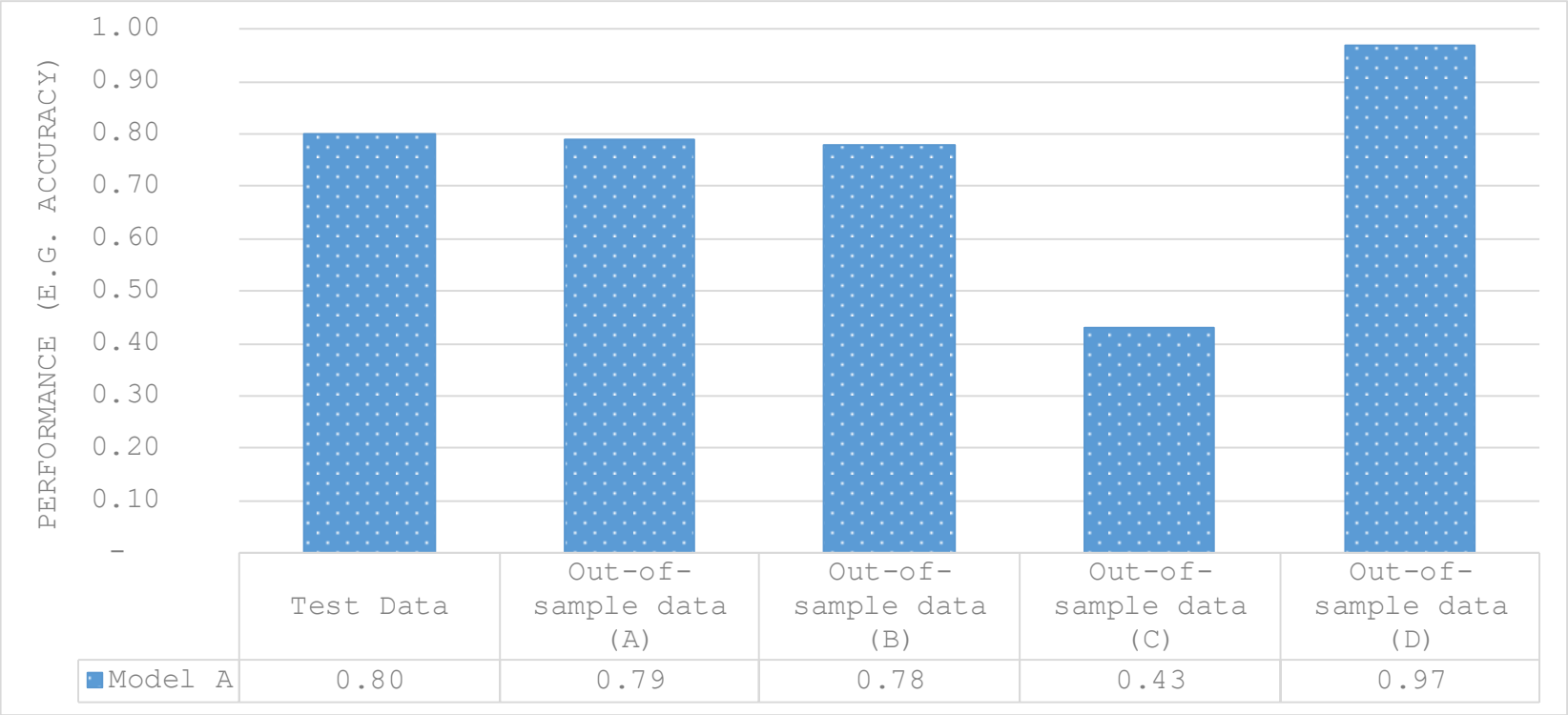
- **Given the results on the test data, what performance would be „expectable“ on the out-of-sample data?**
- **Assumption: Test data is representative for out-of-sample data**



What could we expect? (1)

A, B, C, D?

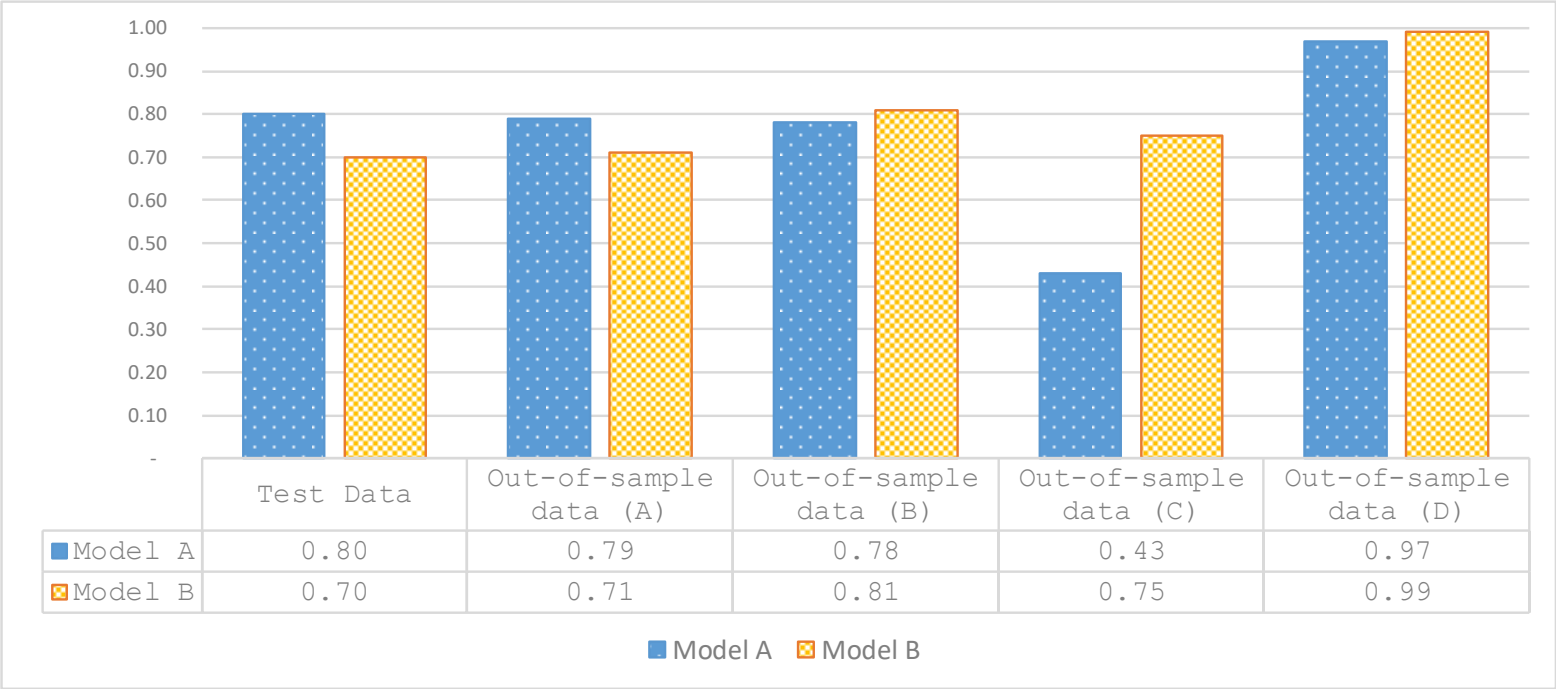
- **Given the results on the test data, would you rather expect outcome A, B, C, or D when applying the model to out-of-sample data?**



	Test Data	Out-of-sample data (A)	Out-of-sample data (B)	Out-of-sample data (C)	Out-of-sample data (D)
Model A	0.80	0.79	0.78	0.43	0.97

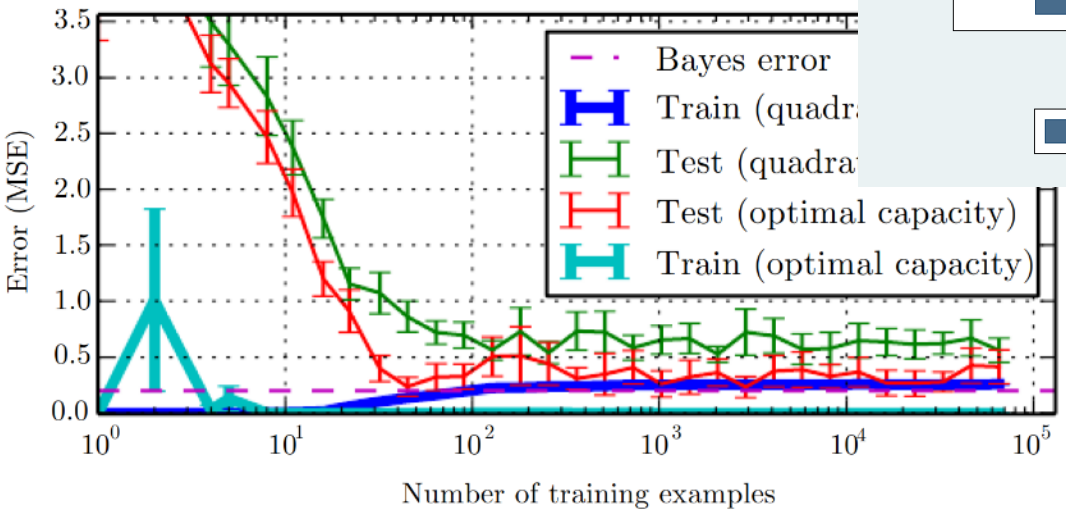
What could we expect? (2)

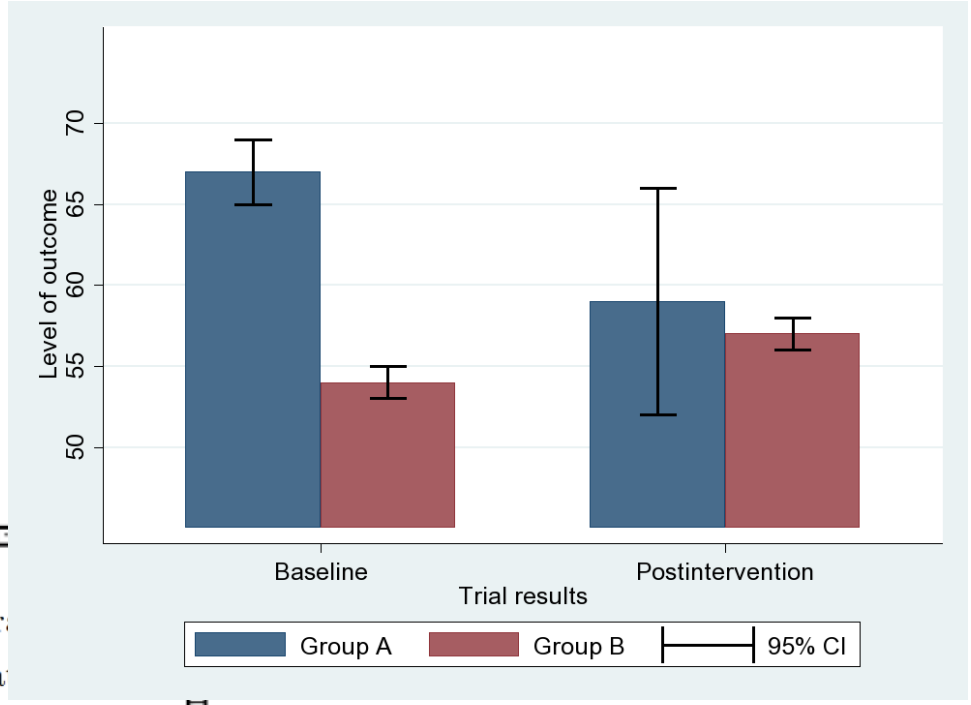
- **Given the results on the test data, would you rather expect outcome A, B, C, or D when applying the model to out-of-sample data?**



	Test Data	Out-of-sample data (A)	Out-of-sample data (B)	Out-of-sample data (C)	Out-of-sample data (D)
Model A	0.80	0.79	0.78	0.43	0.97
Model B	0.70	0.71	0.81	0.75	0.99

Confidence Intervals

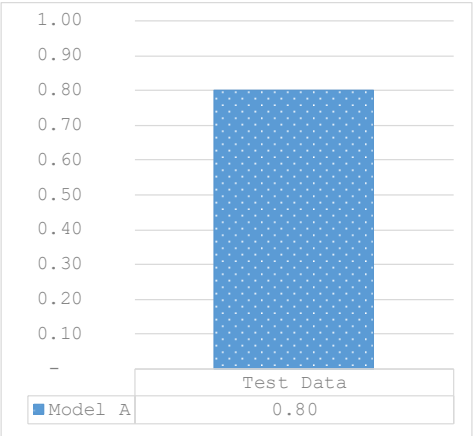


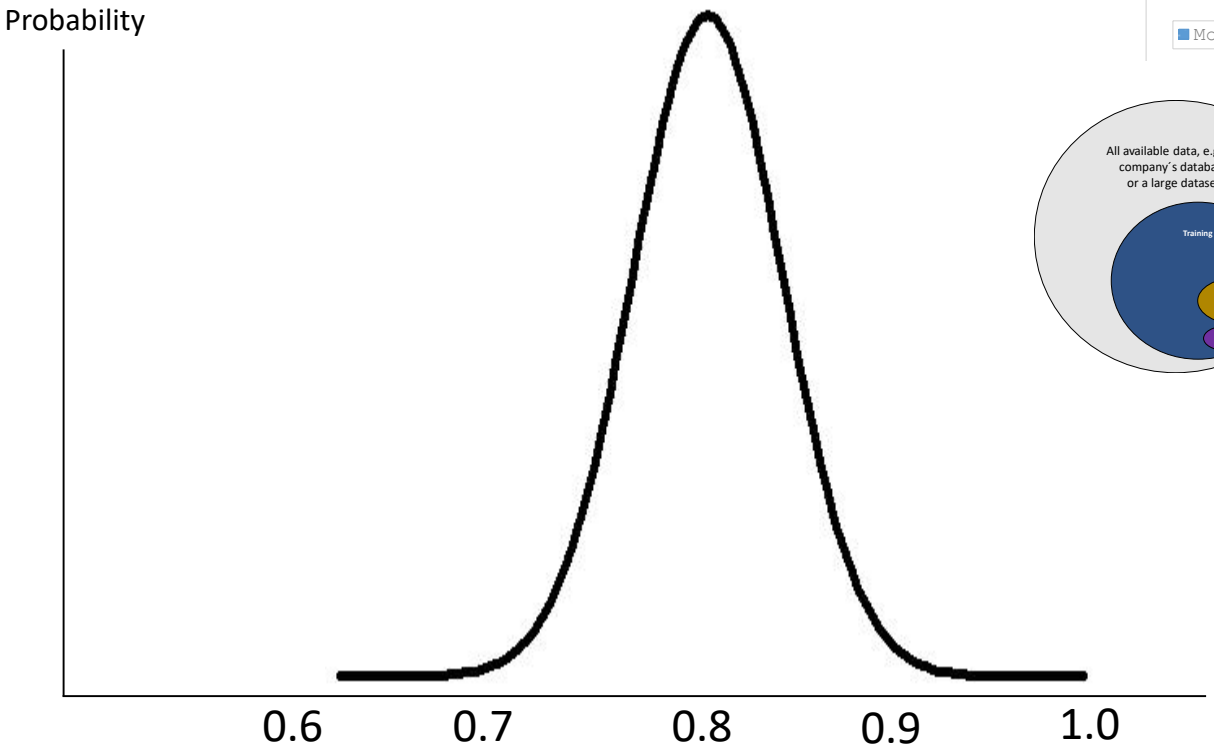


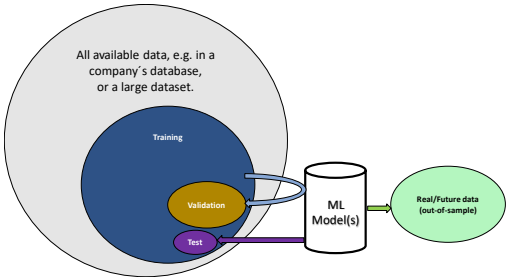
Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep learning* (MIT press, 2016).

https://theesspreckelsen.files.wordpress.com/2014/08/alex2.png

Confidence (Normal Distribution)







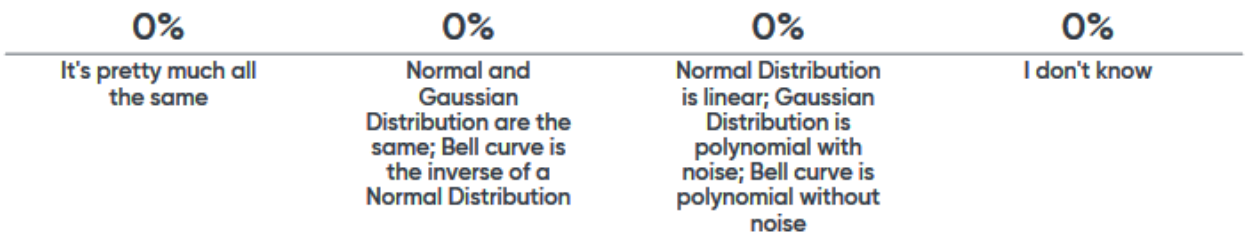
http://blogs.discovermagazine.com/gnxp/files/2012/05/norm.jpg

Survey: Normal vs Gauss vs. Bell

Go to **www.menti.com** and use the code **51 13 86**

What is the difference between a Normal Distribution, the Gaussian Distribution, and a Bell Curve Distribution?

Mentimeter



Slide is not active **Activate**

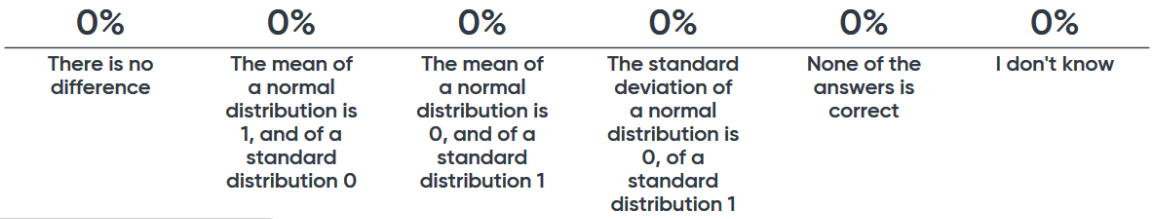
0

Survey: Standard vs. Normal Distribution

Go to **www.menti.com** and use the code **51 13 86**

What is the difference between a standard and a normal distribution?

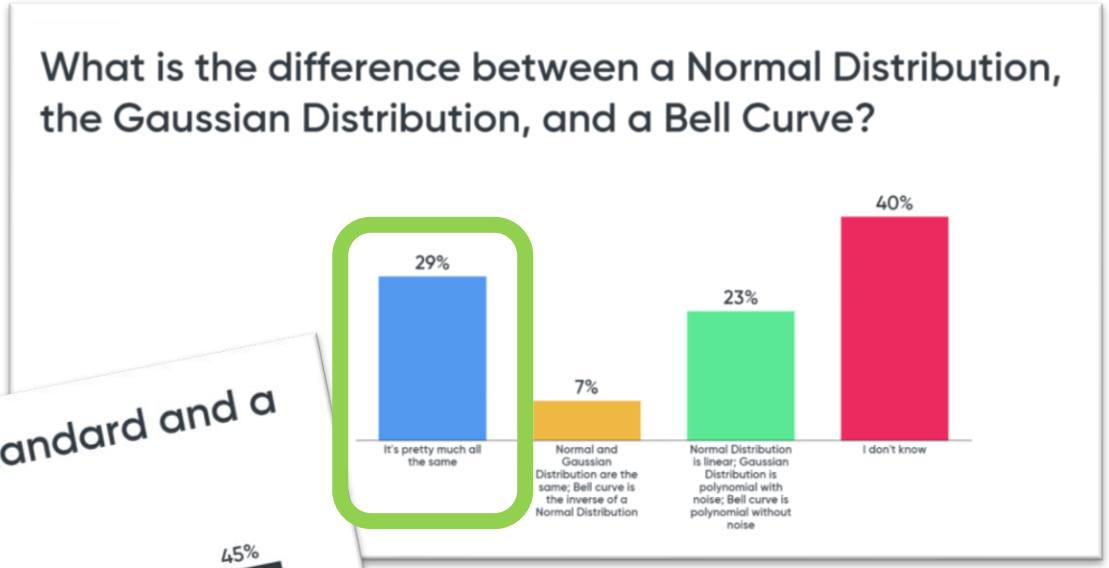
Mentimeter

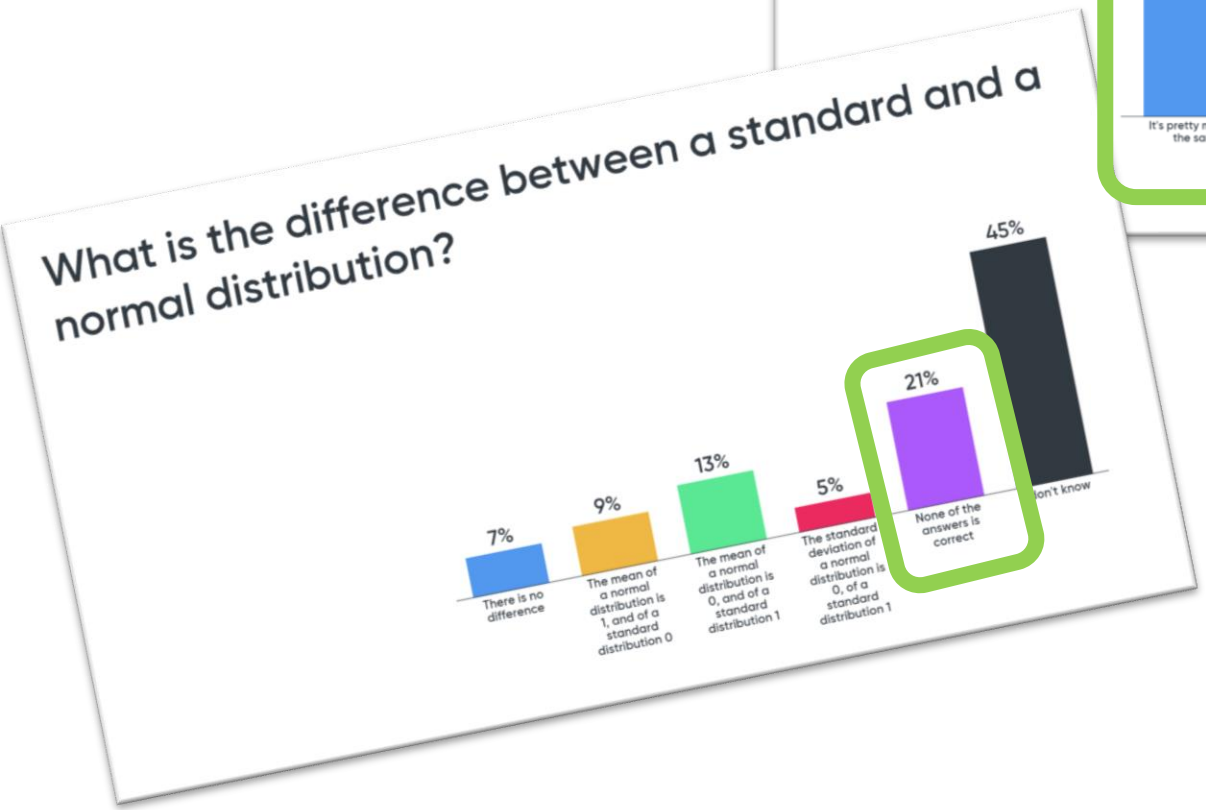


Slide is not active **Activate**

0

Last year‘s results





Standard Distribution (also called „Standard Normal Distribution“) is a special type of a normal distribution with a mean=0 and std deviation = 1.

Basic Formulas

Mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

Variance $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Standard Deviation (SD) $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$

Standard Error of the Mean $SEM = \frac{\sigma}{\sqrt{N}}$

SEM is always smaller than SD

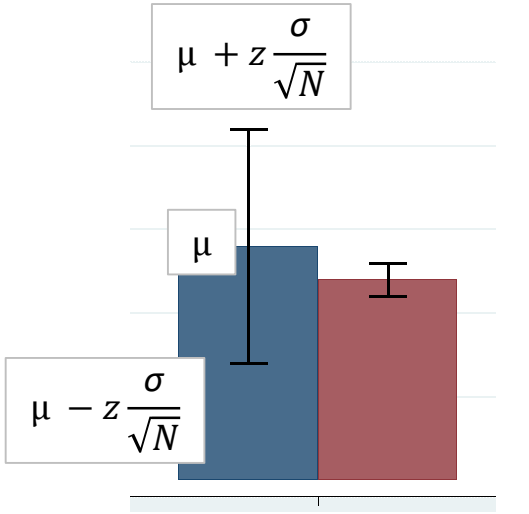
Bessel's Correction

$$\sigma^2_{sample} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\sigma_{sample} = \sqrt{\sigma^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2}$$

http://www.mathsisfun.com/data/standard-deviation.html
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3148365/

Confidence interval $[\mu - z\frac{\sigma}{\sqrt{N}}; \mu + z\frac{\sigma}{\sqrt{N}}]$ **with the lower bound of expected performance being** $\mu - z\frac{\sigma}{\sqrt{N}}$**, the upper bound being** $\mu + z\frac{\sigma}{\sqrt{N}}$**, centred on mean** μ **with confidence level** z**, and standard deviation** σ ***and sample size*** N



https://theesspreckelsen.files.wordpress.com/2014/08/alex2.png

Confidence Level	z
99%	2.58
98%	2.33
95%	1.96
90%	1.65

Approximation

- $\sigma = SEM$ **(if true population is at least 20 times larger than sample)**
- p **= Sample proportion = Performance (e.g. accuracy) on test set**
- **Expected classification accuracy: p + - SE**

$$\sigma = SEM = \sqrt{\frac{p(1-p)}{n}}$$

$np \geq 10$
$n(1-p) \geq 10$

Smallest sample size is needed for p=0.5

- **For instance, 100 samples, 0.85 accuracy → Expected accuracy is 85±3.6% → 95% confidence interval [0.78;0.92]**

$$SEM = \sqrt{\frac{0.85(1-0.85)}{100}} = 0.036$$

$$[\mu - z\,\sigma;\ \mu + z\,\sigma]$$

Kubat, An Introduction to Machine Learning (Springer, 2015). Chapter 12 (page 235)

http://stattrek.com/estimation/confidence-interval-proportion.aspx?Tutorial=AP

More

- **Kubat, An Introduction to Machine Learning (Springer, 2015). Chapter 12 (page 235)**
- **https://pdfs.semanticscholar.org/presentation/cd8b/d562e656e25f215f8ce906baee4642124a9d.pdf**
- **https://people.cs.pitt.edu/~milos/courses/cs2750-Spring04/lectures/class4.pdf**
- **https://ocw.mit.edu/courses/mathematics/18-05-introduction-to-probability-and-statistics-spring-2014/readings/MIT18_05S14_Reading24.pdf**
- **https://www.youtube.com/watch?v=A82brFpdr9g**
- **https://www.kdnuggets.com/2016/08/central-limit-theorem-data-science-part-2.html**

Bootstrap confidence intervals
Class 24, 18.05
Jeremy Orloff and Jonathan Bloom

1 Learning Goals

1. Be able to construct and sample from the empirical distribution of data.
2. Be able to explain the bootstrap principle.
3. Be able to design and run an empirical bootstrap to compute confidence intervals.
4. Be able to design and run a parametric bootstrap to compute confidence intervals.

2 Introduction

The empirical bootstrap is a statistical technique popularized by Bradley Efron in 1979. Though remarkably simple to implement, the bootstrap would not be feasible without modern computing power. The key idea is to perform computations on the data itself to estimate the variation of statistics that are themselves computed from the same data. That is, the data is 'pulling itself up by its own bootstrap.' (A google search of 'by ones own bootstraps' will give you the etymology of this metaphor.) Such techniques existed before 1979, but Efron widened their applicability and demonstrated how to implement the bootstrap effectively using computers. He also coined the term 'bootstrap' [1].

Our main application of the bootstrap will be to estimate the variation of point estimates; that is, to estimate confidence intervals. An example will make our goal clear.

Example 1. Suppose we have data

$$x_1, x_2, \ldots, x_n$$

If we knew the data was drawn from $N(\mu, \sigma^2)$ with the unknown mean μ and known variance σ^2 then we have seen that

$$\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

Now suppose the data is drawn from some completely unknown distribution. To have a name we'll call this distribution F and its (unknown) mean μ. We can still use the sample mean $\overline{x}$ as a point estimate of μ. But how can we find a confidence interval for μ around $\overline{x}$? Our answer will be to use the bootstrap!

In fact, we'll see that the bootstrap handles other statistics as easily as it handles the mean. For example: the median, other percentiles or the trimmed mean. These are statistics where, even for normal distributions, it can be difficult to compute a confidence interval from theory alone.

[1] Paraphrased from Dekking et al. *A Modern Introduction to Probabilty and Statistics*, Springer, 2005, page 275.

1

P value – the holy grail

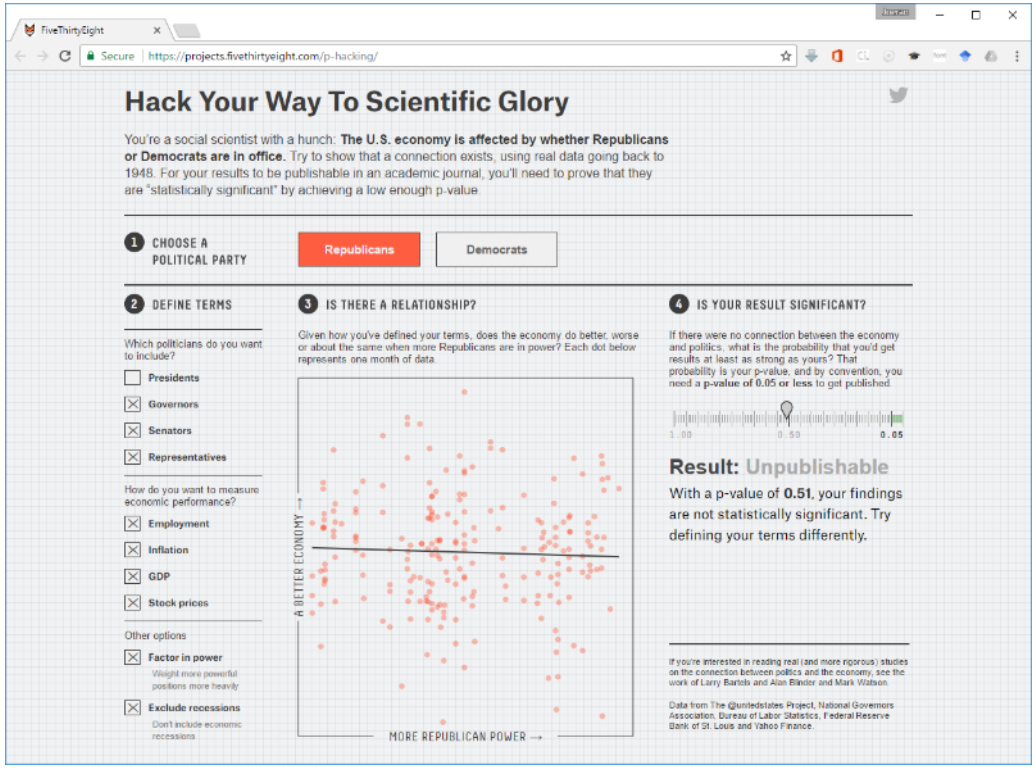
Statistical Significance

- **Describes the probability that an observed difference is caused by chance**
- **Typical p value should be smaller than .05 or .01**
- **Statistically insignificant results are almost always of little value**
- **Statistically significant results can still be false, or practically insignificant**
- **“Experimental data yielding a P value of .05 means that there is only a 5 percent chance of obtaining the observed (or more extreme) result if no real effect exists”**
- **“The p-value gives information about the probability of obtaining evidence. It doesn’t quantify the strength of the evidence.”**

https://www.sciencenews.org/article/odds-are-its-wrong
http://asq.org/quality-progress/2011/08/statistics-roundtable/not-significant-but-important.html

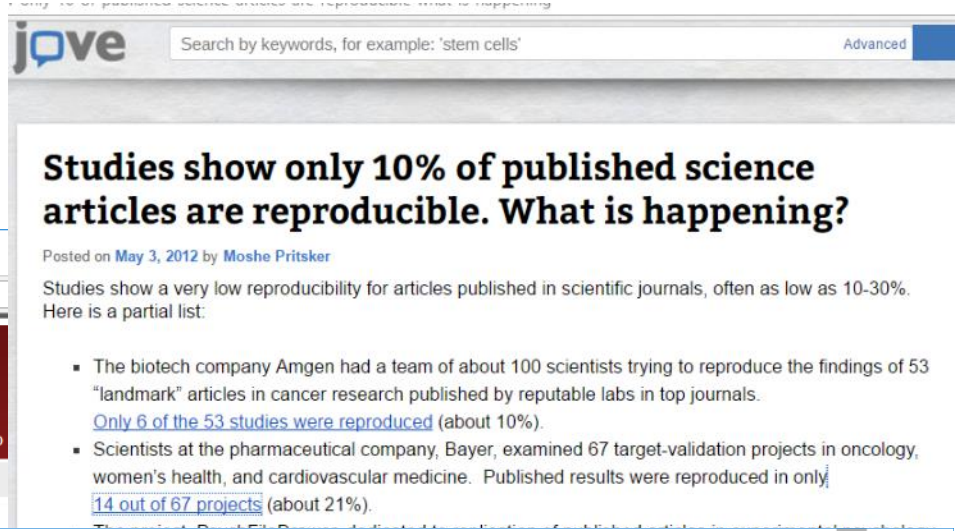
P-hacking

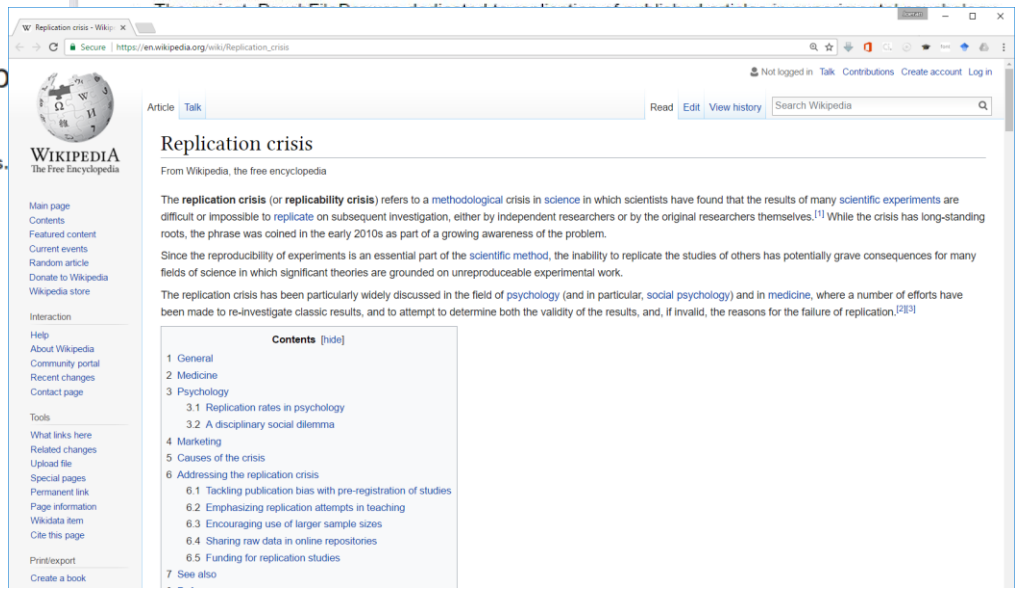
"If you torture your data long enough, it will confess"



- **https://projects.fivethirtyeight.com/p-hacking/**

Statistically Significant but Not Reproducible

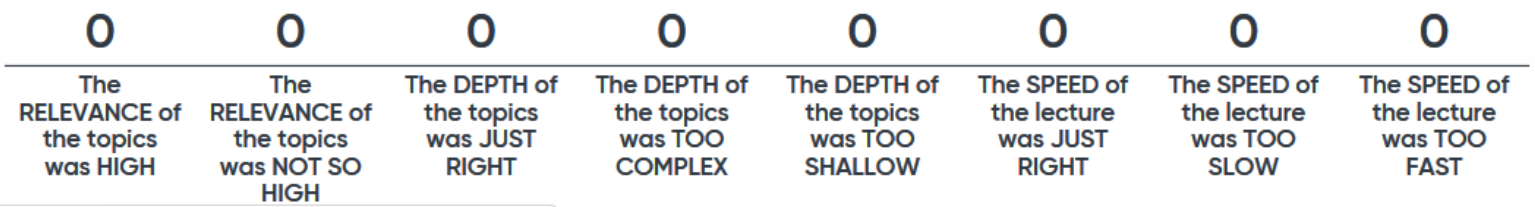




Go to **www.menti.com** and use the code **51 13 86**

Lecture Evaluation

Mentimeter



Slide is not active **Activate**

0

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

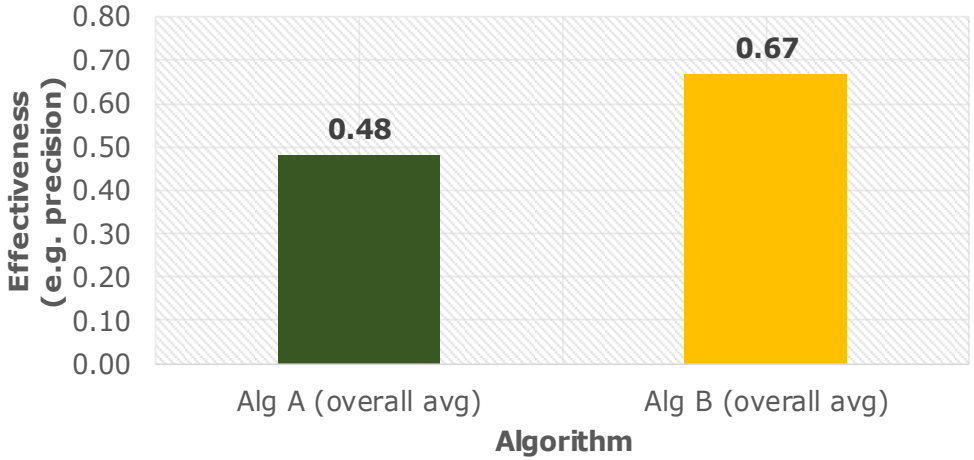
Outtakes

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

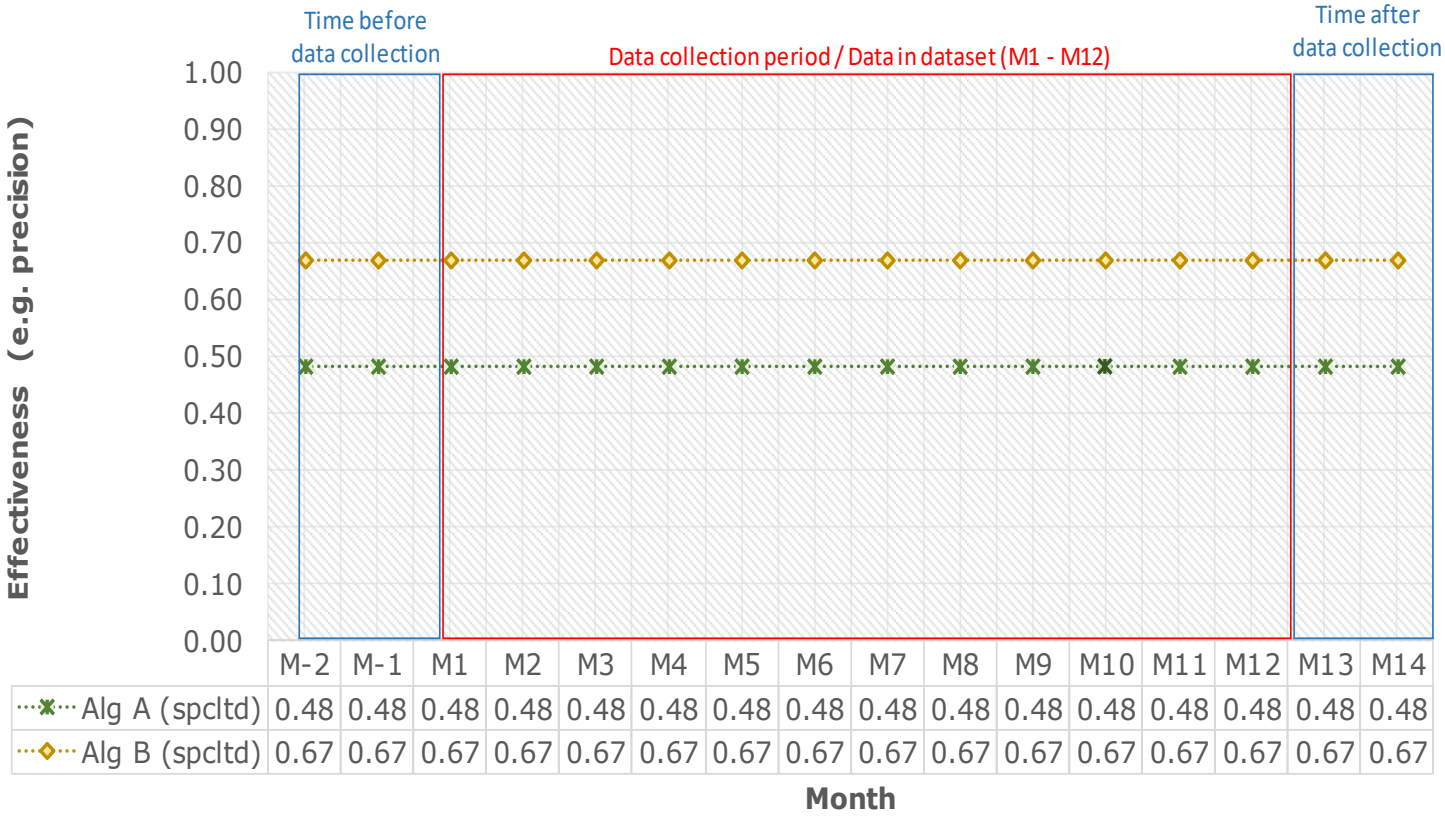
The importance of time

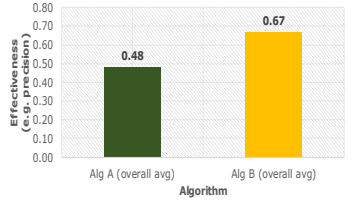
Reporting Performance over time

The Normal Way Of Reporting Performance

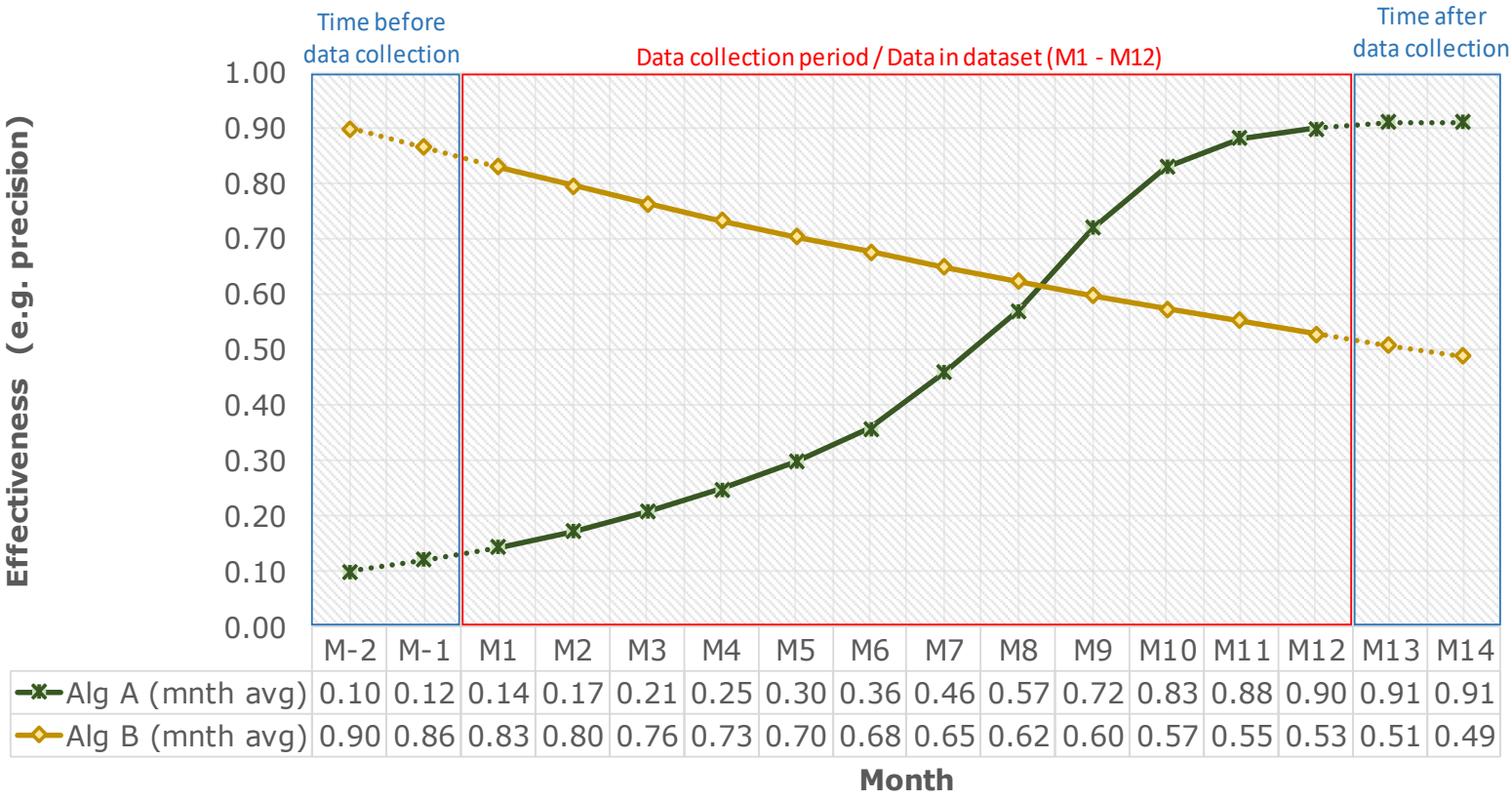


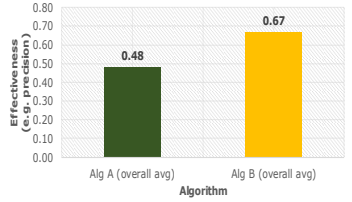
The Underlying (Stupid) Assumption



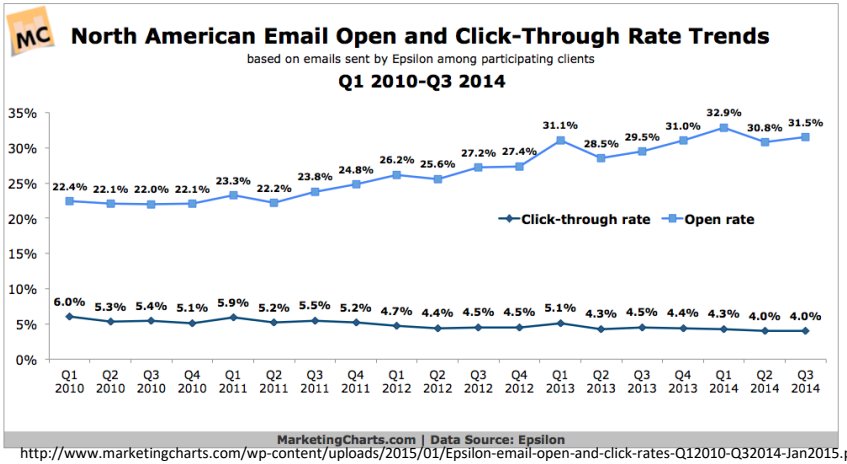


A better alternative

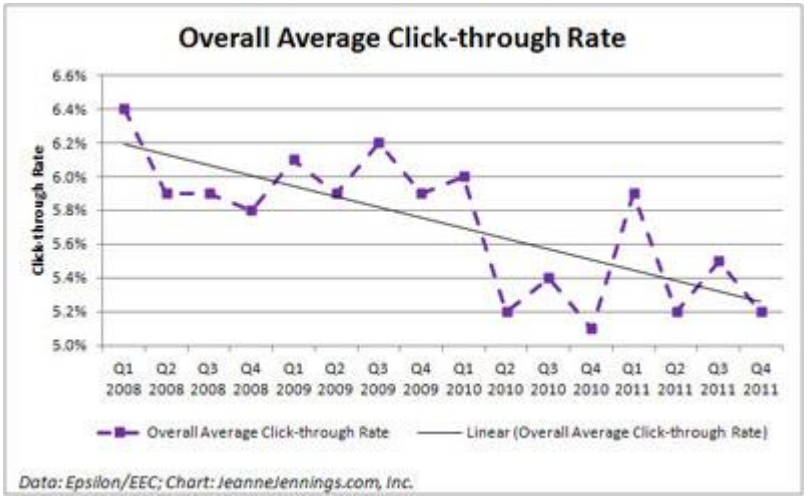




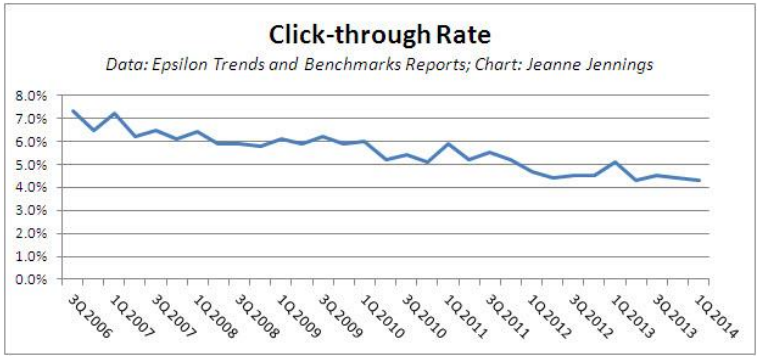
Changes over time



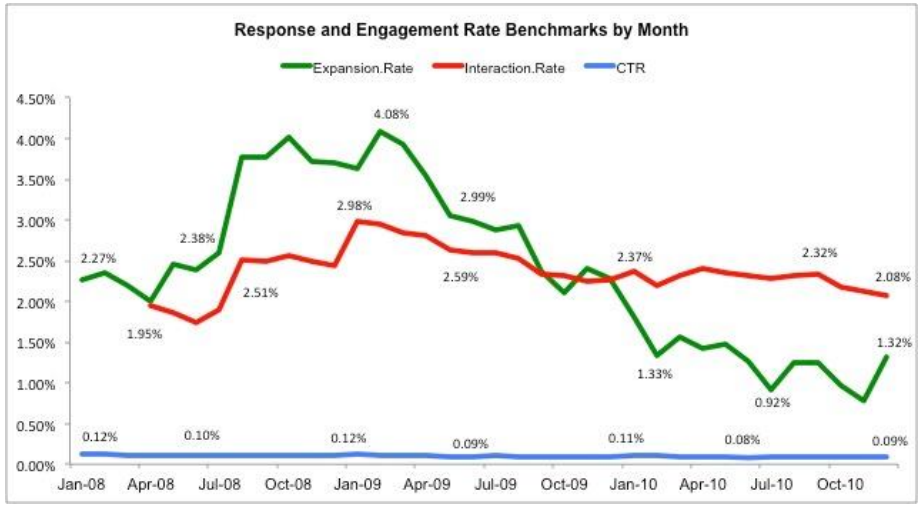
http://www.marketingcharts.com/wp-content/uploads/2015/01/Epsilon-email-open-and-click-rates-Q12010-Q32014-Jan2015.png



http://www.flyerco.com/blog/wp-content/uploads/2014/08/041612-ctr.jpg



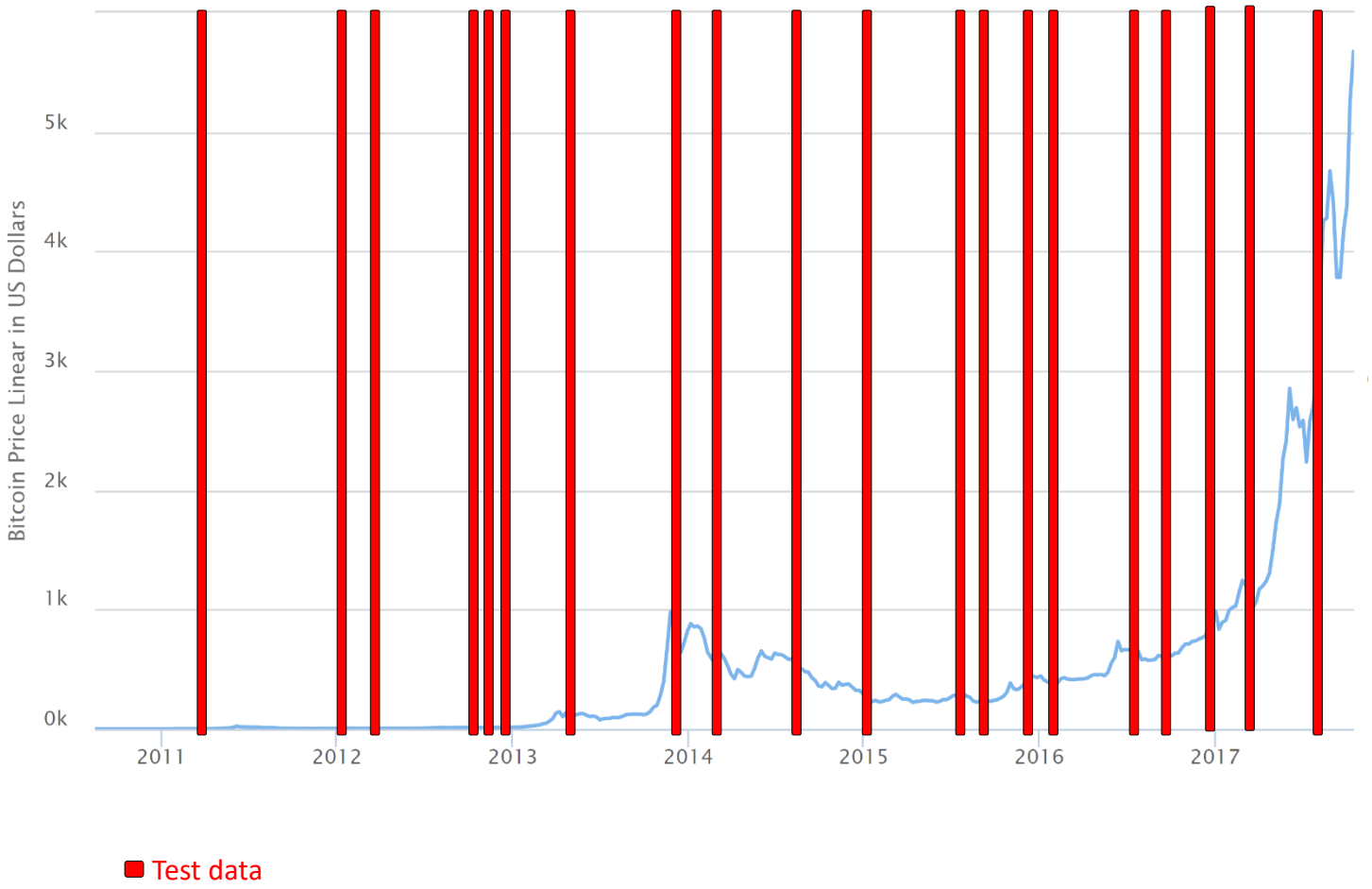
https://media.licdn.com/mpr/mpr/shrinknp_800_800/AAEAAQAAAAAAAAWvAAAAJDUzMWNjYzEzLTg2YTYtNDIwMy04ZDIxLTc2ZmUxYTJmZTdiNQ.jpg



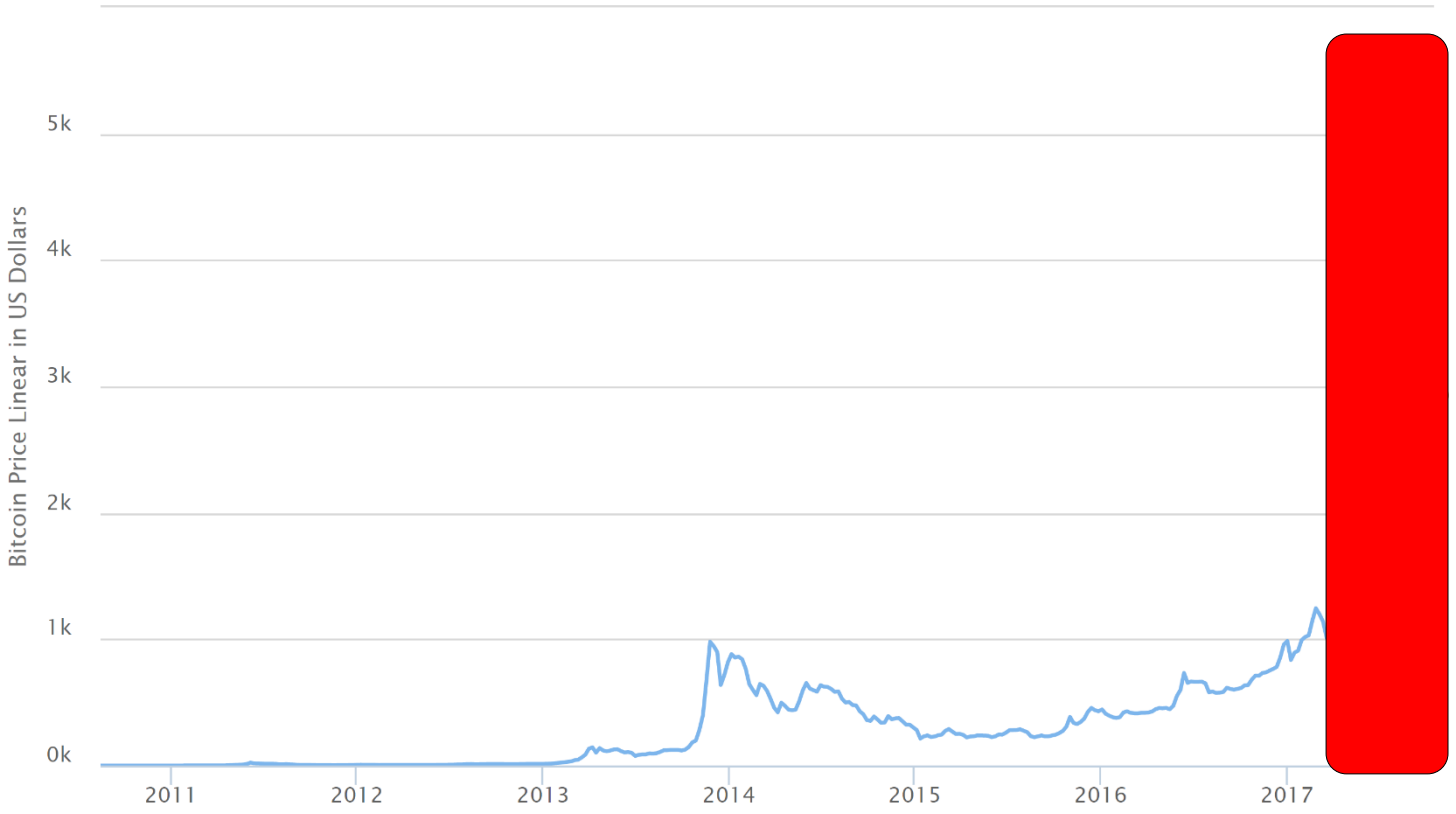
http://68.media.tumblr.com/tumblr_lq2up3LbB41qe9xbio1_1280.jpg

Consider time in training and testing

How difficult would it be to predict the values in the Test Set?



And now?

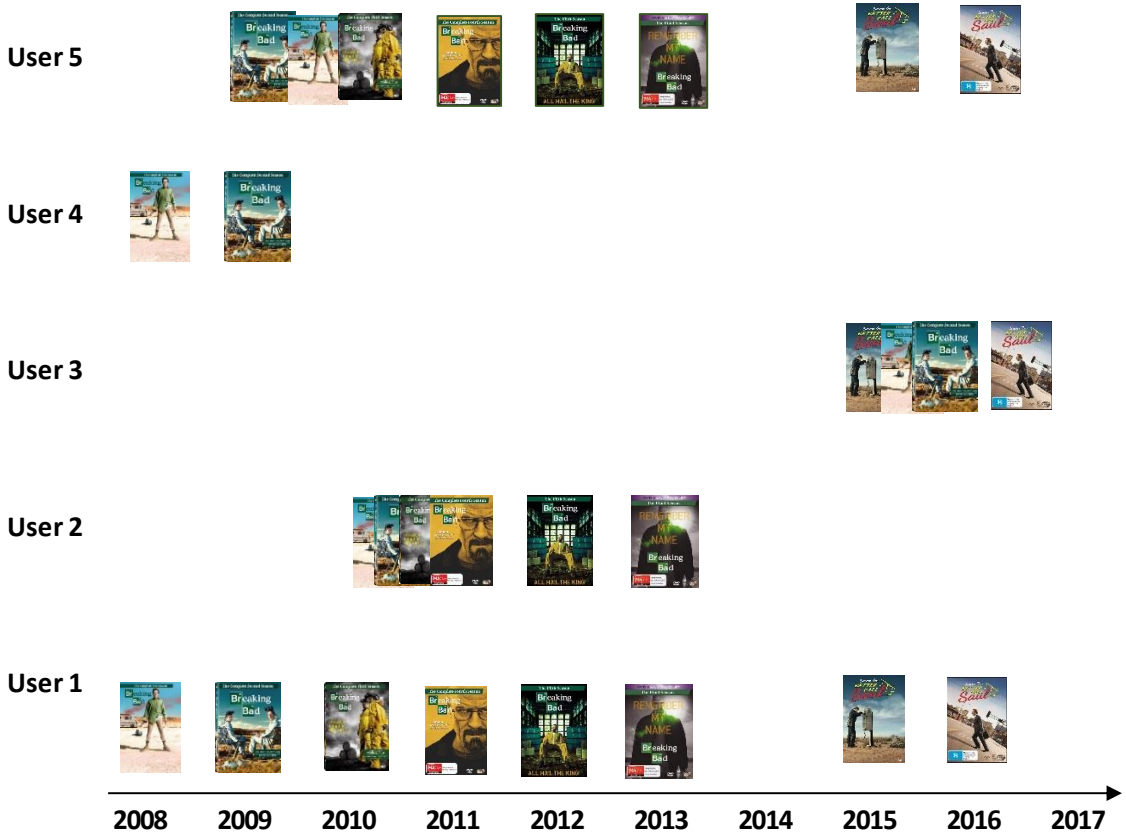

Bitcoin Price Linear in US Dollars
5k
4k
3k
2k
1k
0k
2011
2012
2013
2014
2015
2016
2017
Test data

Time-Series Training

- **Often, time matters**
- **Prediction for current value is easy if future is known, i.e. used for training**
- **For instance**
  - Stock Market
  - Climate Change
  - Navigation

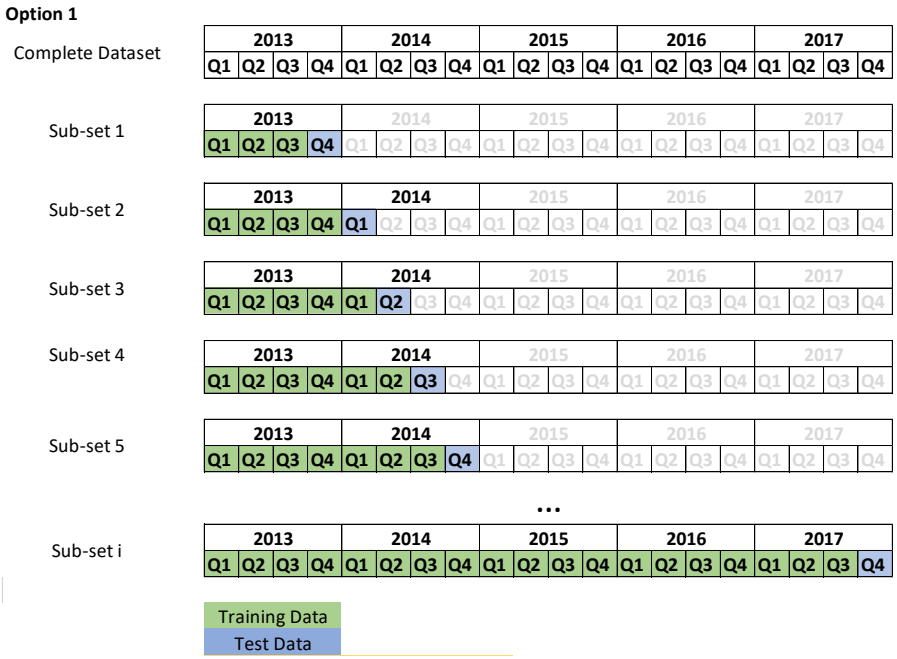
Time is not considered in most „normal“ applications

- **Movie recommendations**
- **Face recognition**
- **…**

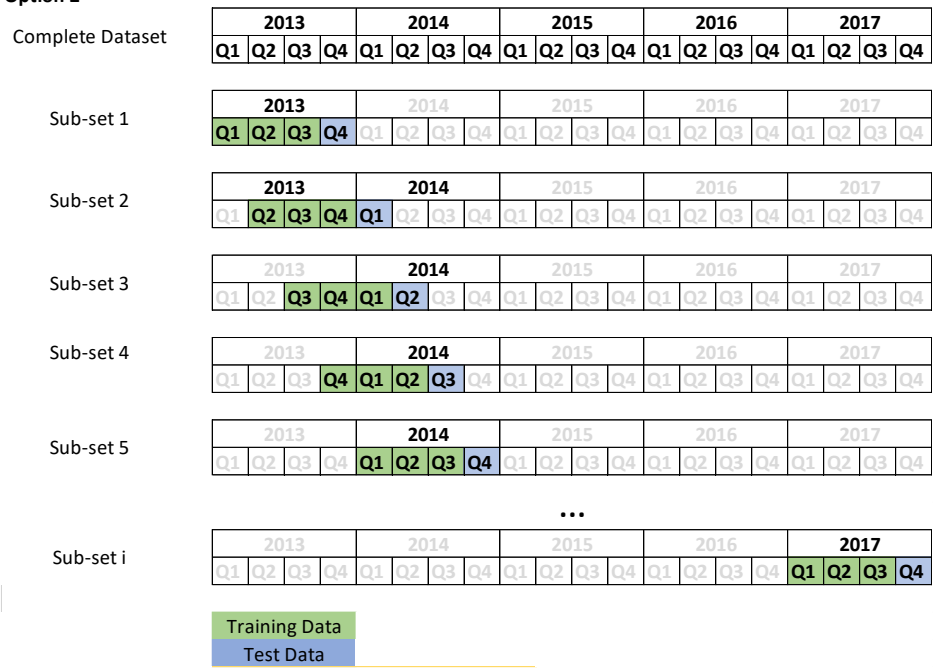


Time-based Splitting

Option 1



Option 2



Considering Time / Backtesting / Time-Series

- **http://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/**
- **https://en.wikipedia.org/wiki/Backtesting**
- **http://www.ulb.ac.be/di/map/gbonte/ftp/time_ser.pdf**
- **http://machinelearningmastery.com/time-series-datasets-for-machine-learning/**
- **http://machinelearningmastery.com/time-series-forecasting-supervised-learning/**

- **https://www.youtube.com/watch?v=0Rnq1NpHdmw**

7 Myths about Ground Truths (and Gold Standards)

1. **"One Truth: Most data collection efforts assume that there is one correct interpretation for every input example.**
2. **Disagreement Is Bad: To increase the quality of annotation data, disagreement among the annotators should be avoided or reduced.**
3. **Detailed Guidelines Help: When specific cases continuously cause disagreement, more instructions are added to limit interpretations.**
4. **One Is Enough: Most annotated examples are evaluated by one person.**
5. **Experts Are Better: Human annotators with domain knowledge provide better annotated data.**
6. **All Examples Are Created Equal: The mathematics of using ground truth treats every example the same; either you match the correct result or not.**
7. **Once Done, Forever Valid: Once human annotated data is collected for a task, it is used over and over with no update. New annotated data is not aligned with previous data."**

http://media.npr.org/assets/img/2017/03/21/learning-styles_custom-751056055fae1257b7f6dc964ad5c5cfb8824c12-s900-c85.jpg

https://www.aaai.org/ojs/index.php/aimagazine/article/view/2564/2468

Performance vs. Effectiveness vs. Efficiency

- **Effectiveness**
  - Doing the right things
- **Efficiency**
  - Doing things right
  - Achievement per unit of input (may be the “wrong” achievement)

- **Performance**
  - Sometimes used as synonym for effectiveness
  - Sometimes used as umbrella term

Domain-Specific Metrics

BLEU

- **Metric for Machine Translation**
- **Not relevant for this module**

Central Limit Theorem

- **Given/Assumed**
  - a large number of observations
  - Large random sample with *n* observations
  - Samples are “random“ in terms of independent from the previous observations
- **Mean (and sum) of samples follows normal distribution**
- **The larger n, the closer the mean and sum of the sample get to true values**