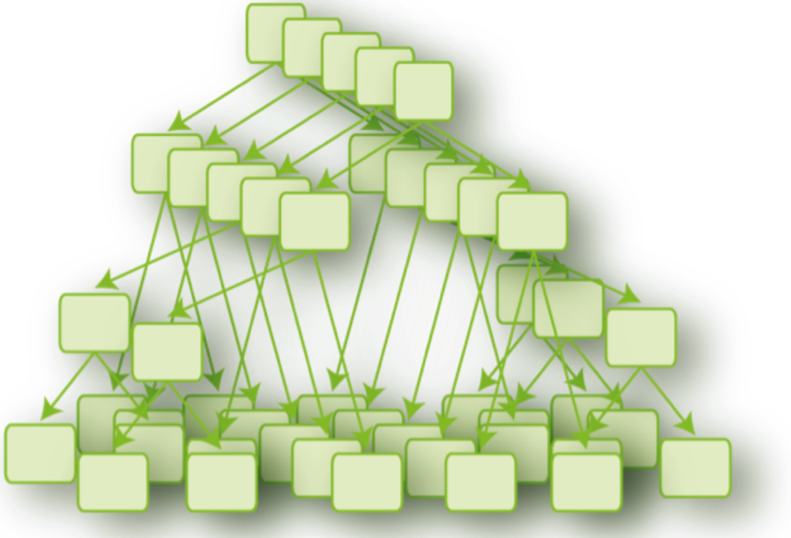


Introduction to Machine Learning in Astroparticle Physics

Tim Ruhe, TU Dortmund University

PHYSTAT-Gamma 09/27/22

tim.ruhe@tu-dortmund.de

Motivation

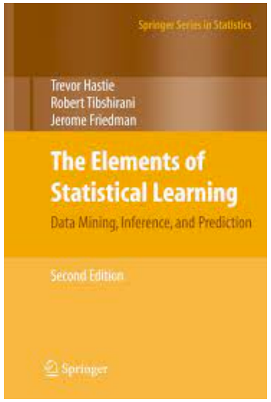
Von Banffy - Eigenes Werk, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=11657709

Source: CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=577886

Machine Learning provides tools to accomplish an analysis task faster and more accurately (when used correctly).

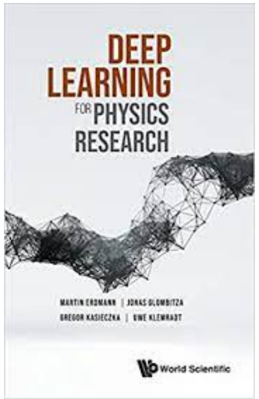
Source: Von smial (talk) - Eigenes Werk, FAL, https://commons.wikimedia.org/w/index.php?curid=6028669

Further Reading


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

- General Introduction to Statistical Learning
- Good start to get an overview
- A lot of extra material: https://hastie.su.domains/ElemStatLearn/
- (I believe you can also download the pdf there….)

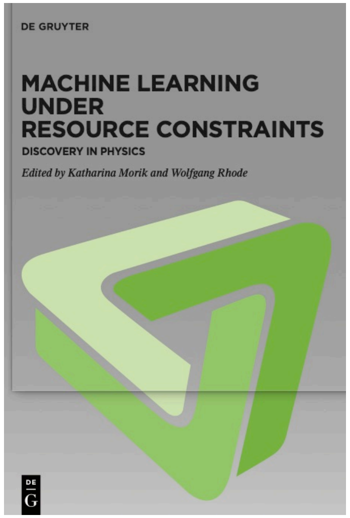
Source: https://hastie.su.domains/ElemStatLearn/


DEEP LEARNING FOR PHYSICS RESEARCH
MARTIN ERDMANN | JONAS GLOMBITZA
GREGOR KASIECZKA | UWE KLEMRADT
World Scientific

- Focus on Deep Learning and Neural Networks
- Nice pedagogic approach

Source: amazon

Further Reading



DE GRUYTER

MACHINE LEARNING UNDER RESOURCE CONSTRAINTS

DISCOVERY IN PHYSICS

Edited by Katharina Morik and Wolfgang Rhode

- Focus on astroparticle and particle physics
- Contains a lot of topics also covered in this talk
- Open access
- To be published by the end of 2022

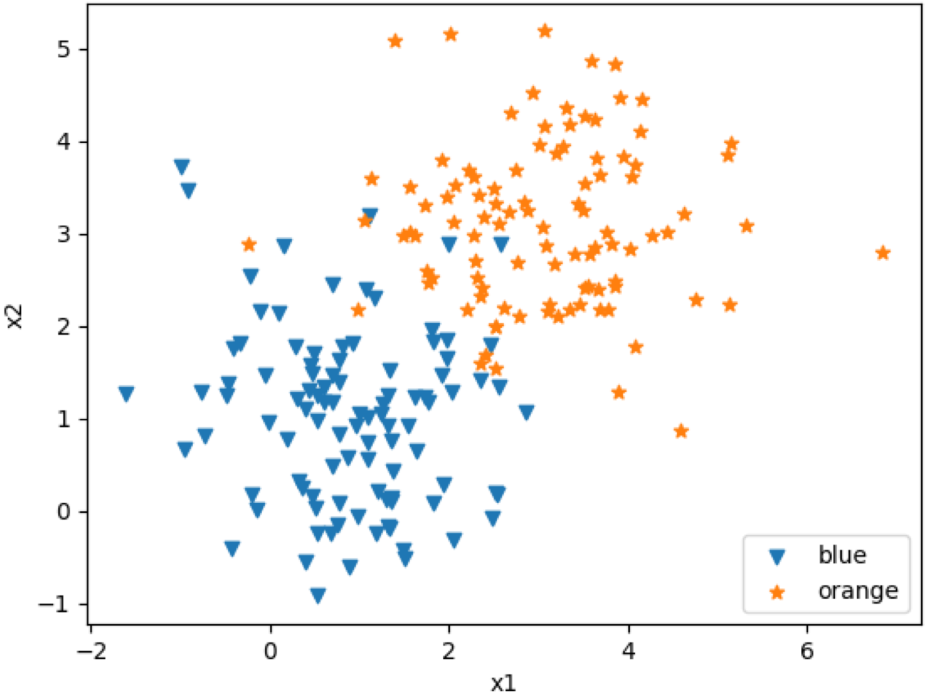
scikit learn

Source: Von The scikit-learn developers - github.com/scikit-learn/scikit-learn/blob/master/doc/logos/scikit-learn-logo.svg, BSD, https://commons.wikimedia.org/w/index.php?curid=71445288

Outline

- Nomenclature and stuff
- Feature Selection
- Selected Algorithms
- Neural Networks and Deep Learning

Nomenclature



N $(\vec{X}, y)$ pairs are referred to as training set
Or annotated data

Events (Examples) are characterized by a feature vector:

$$\vec{X} = (x_1 \ \ldots \ x_n)$$

In this example

$$\vec{X} = (x_1, x_2)$$

And a class variable

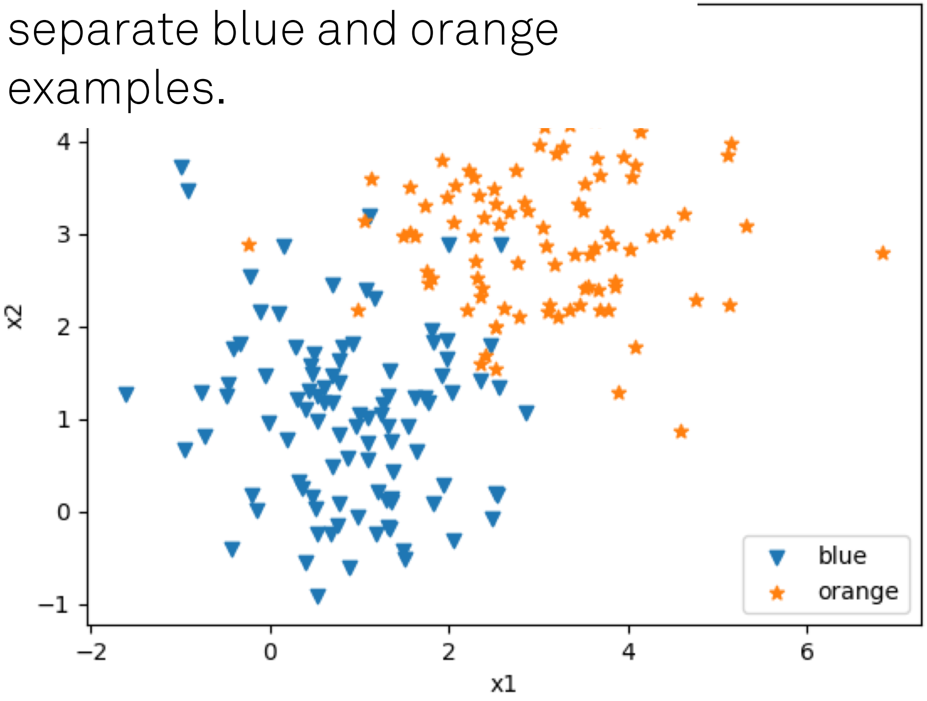
$$y \ \in \ [y_1 \ \ldots \ y_n]$$

In this example

$$y \ \in [blue, orange]$$

Nomenclature

Task: Build a model to separate blue and orange examples.



Events (Examples) are characterized by a feature vector:

$$\vec{X} = (x_1 \ \dots \ x_n)$$

In this example

$$\vec{X} = (x_1, x_2)$$

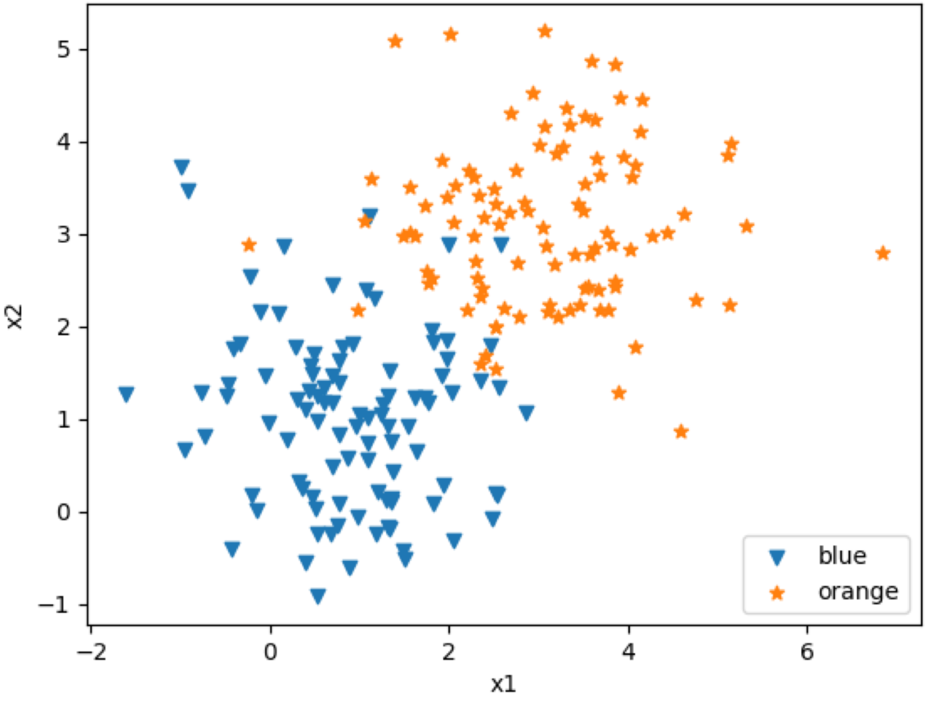
And a class variable

$$y \ \in \ [y_1 \ \dots \ y_n]$$

In this example

$$y \ \in [blue, orange]$$

The Linear Model



$$\hat{y} = \beta_0 + \sum_{i=1}^{p} x_i \beta_i$$

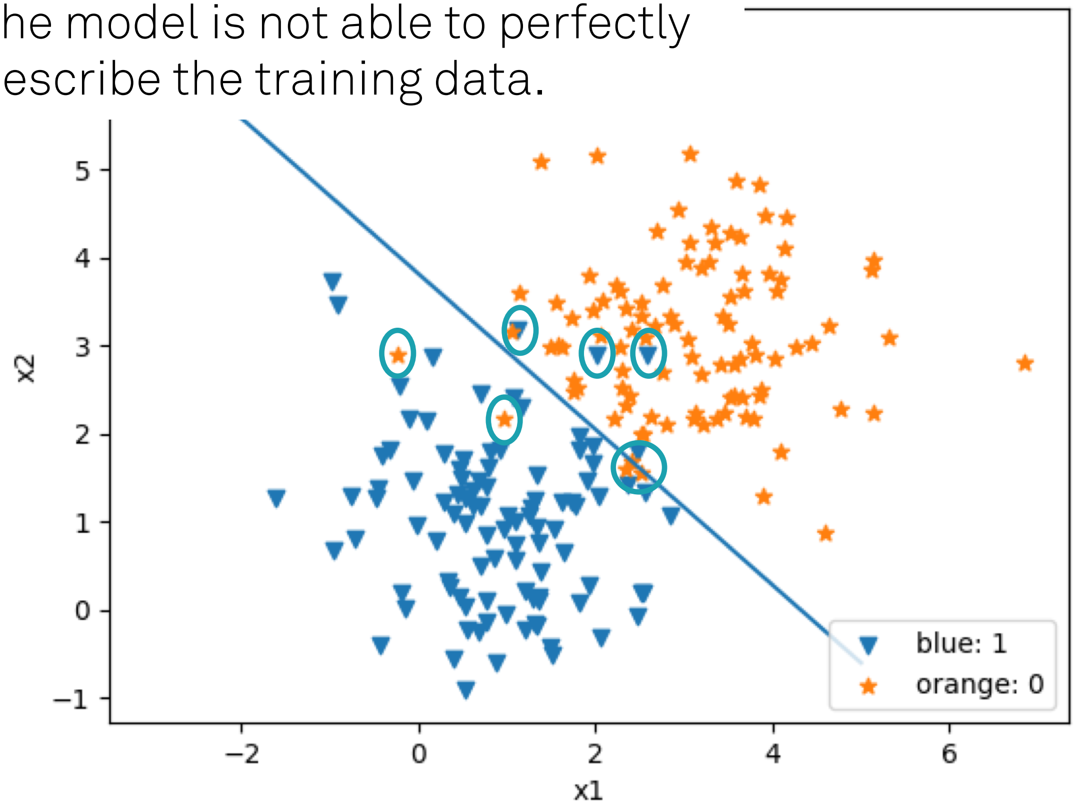
$$\hat{y} = \begin{cases} \text{orange: } 0 \\ \text{blue: } 1 \end{cases}$$

Solve e.g. by least squares fit

The Linear Model: Graphical Representation of the Model

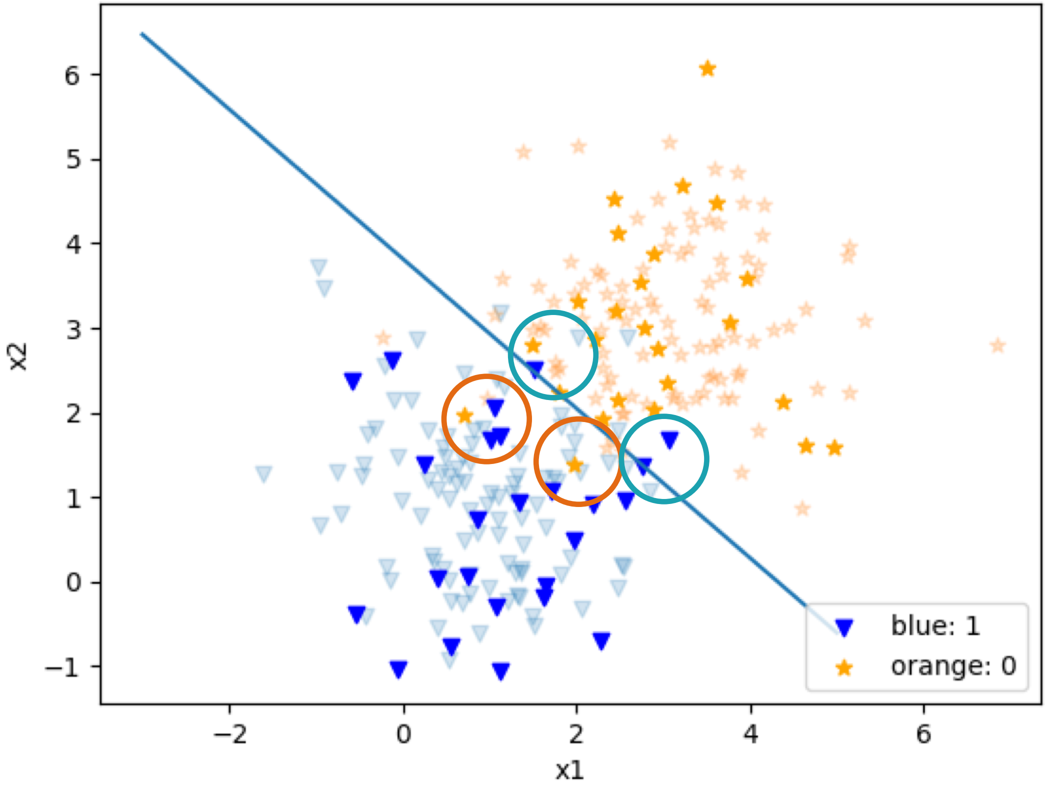
The model is not able to perfectly describe the training data.

Above line:
Classify as orange



Below line:
Classify as blue

Application to Unseen Data

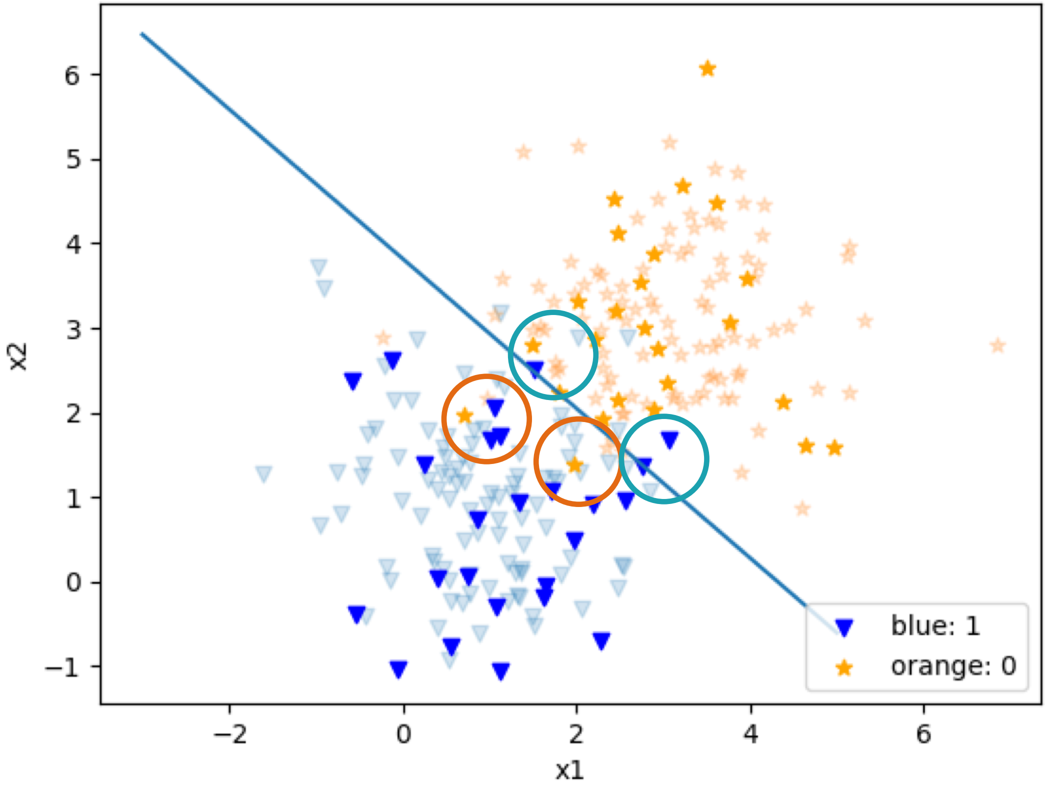


Confusion Matrix

	Blue (+)	Orange (-)
Blue (+)	22	2
Orange (-)	3	22

*I defined that border will be part of the orange class.

True and False Negatives and Postives



True Positives (TP)

False Positives (FP)

Confusion Matrix

	Blue (+)	Orange (-)
Blue (+)	22	2
Orange (-)	3	22

False Negatives (FN)

True Negatives (TN)

*I defined that border will be part of the orange class.

Quality Measures

Accuracy:

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

Precision:

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

Recall:

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

* These measures can sometimes have different names

True Positives (TP)

False Positives (FP)

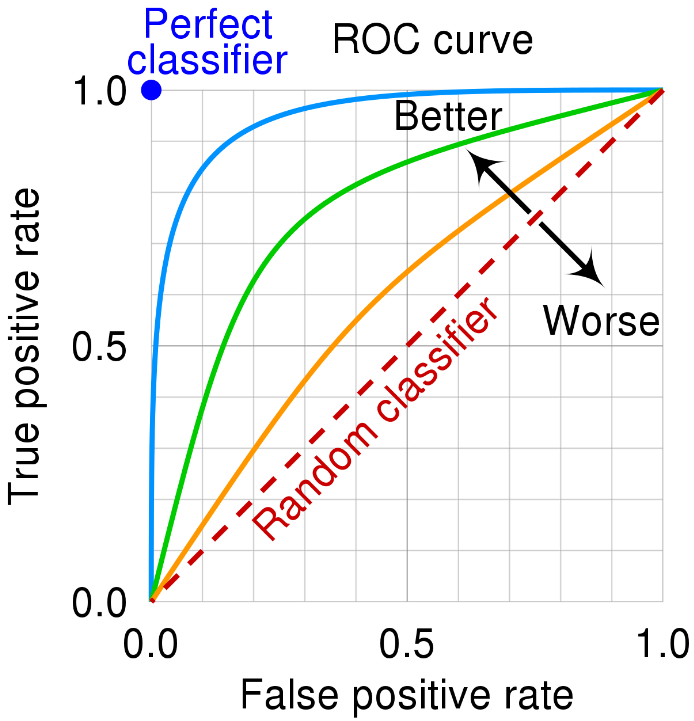
Confusion Matrix

	Blue (+)	Orange (-)
Blue (+)	22	2
Orange (-)	3	22

False Negatives (FN)

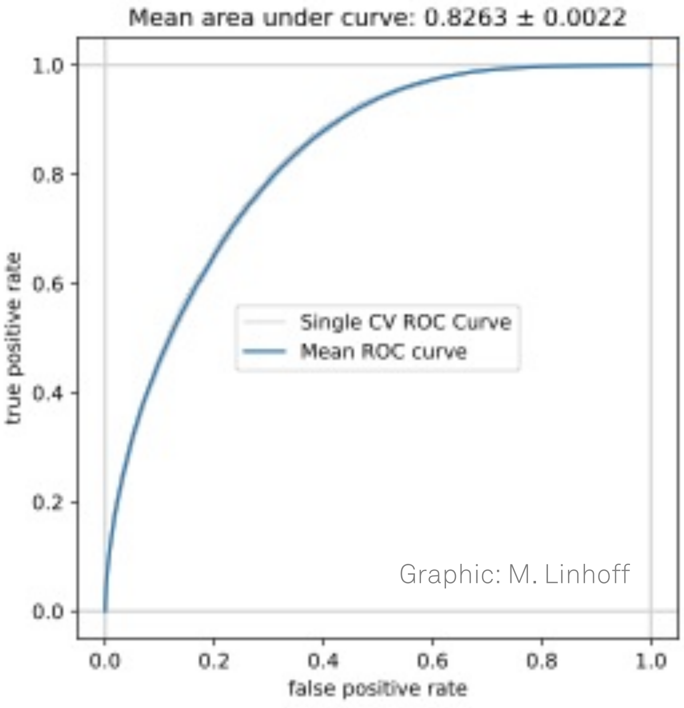
True Negatives (TN)

Area Under Curve



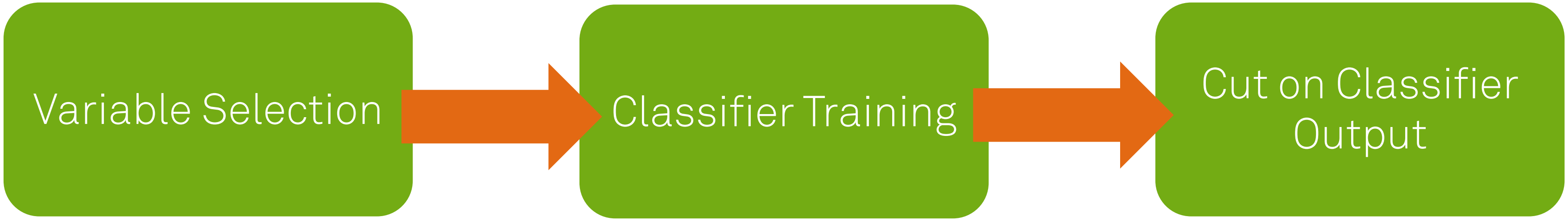
Source: By cmglee, MartinThoma - Roc-draft-xkcd-style.svg, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=109730045

Graphics: M. Linhoff [Learning Under Resource Constraints – Discovery in Physics] (in preparation)



ROC characteristic for the FACT Open Crab data set

Exemplary Data Analysis Pipeline

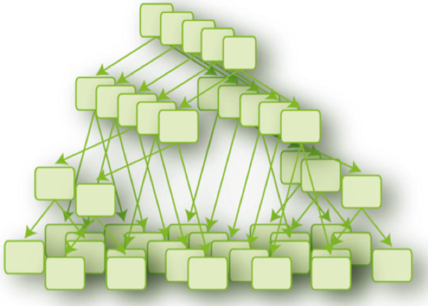


Variable Selection

Classifier Training

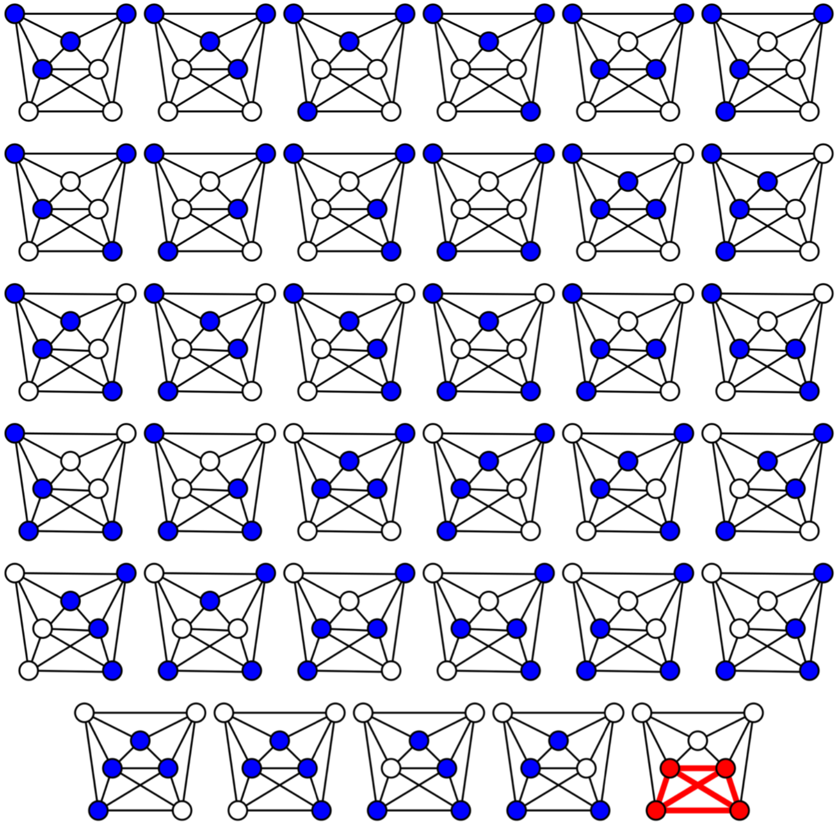
Cut on Classifier Output

Source: https://www.pinterest.com/pin/550354016946043419/



Picture: CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=14260

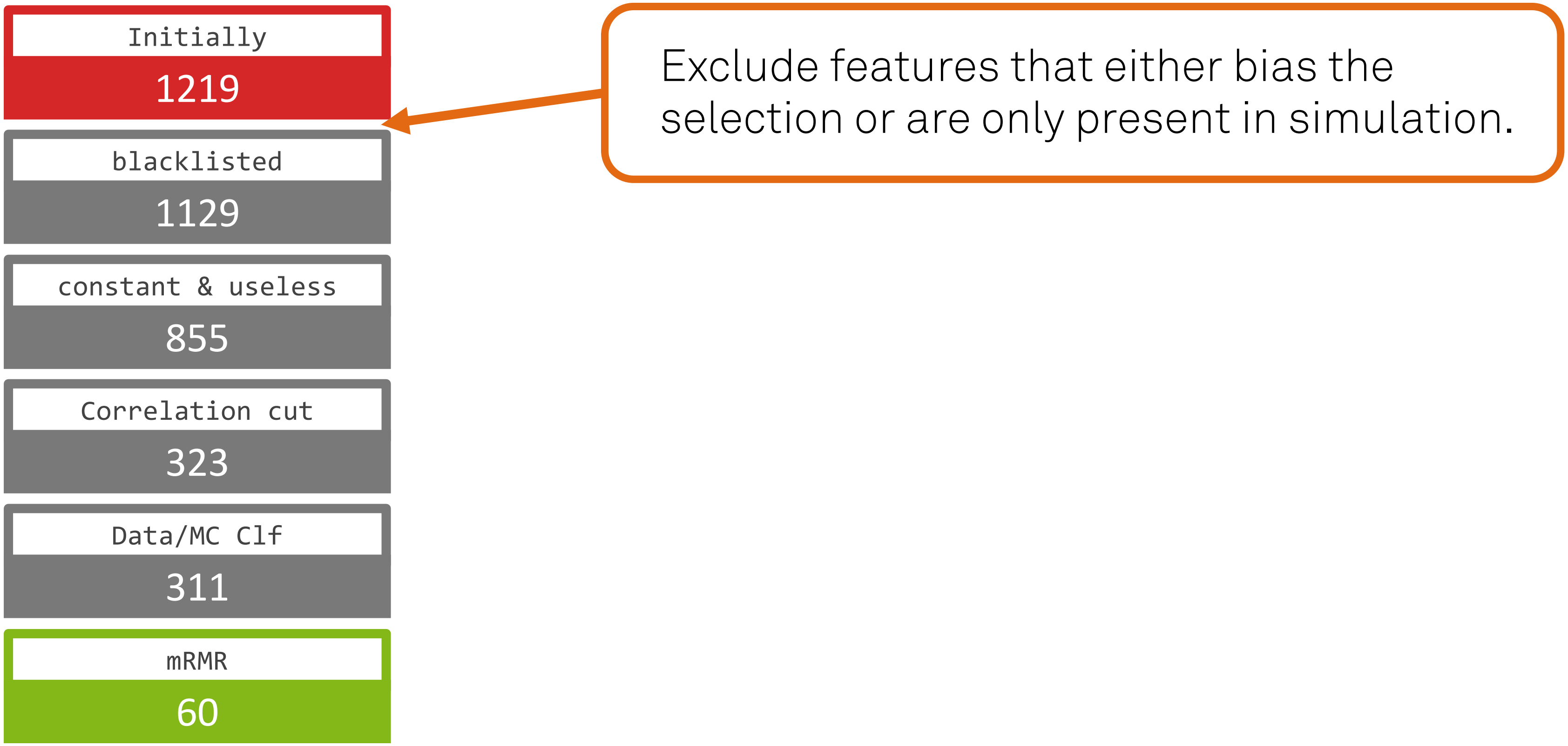
Variable Selection: Try possible combinations



Yes, but...

Source: By Thore Husfeldt at English Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=31823619

Feature Selection



Initially

1219

blacklisted

1129

constant & useless

855

Correlation cut

323

Data/MC Clf

311

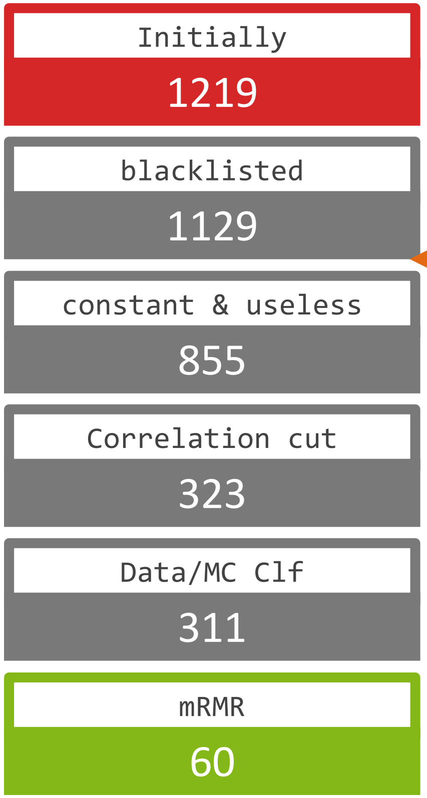
mRMR

60

Exclude features that either bias the selection or are only present in simulation.

M. Börner, PhD thesis (2018)

Feature Selection



Initially
1219

blacklisted
1129

constant & useless
855

Correlation cut
323

Data/MC Clf
311

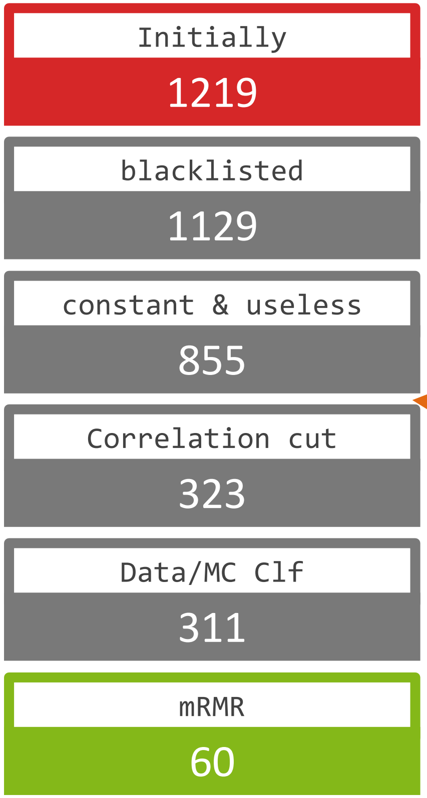
mRMR
60

Exclude features that either bias the selection or are only present in simulation.

Constant features do not carry information.

M. Börner, PhD thesis (2018)

Feature Selection



Initially

1219

blacklisted

1129

constant & useless

855

Correlation cut

323

Data/MC Clf

311

mRMR

60

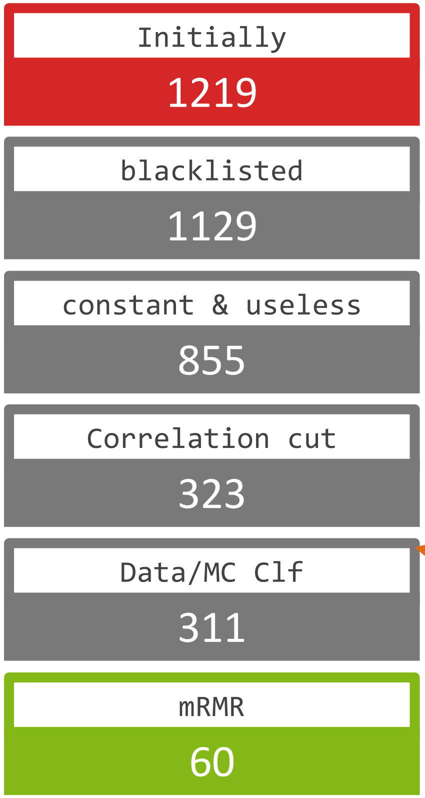
M. Börner, PhD thesis (2018)

Exclude features that either bias the selection or are only present in simulation.

Constant features do not carry information.

Strongly correlated features do not contain new information (or only very little)

Feature Selection



Exclude features that either bias the selection or are only present in simulation.

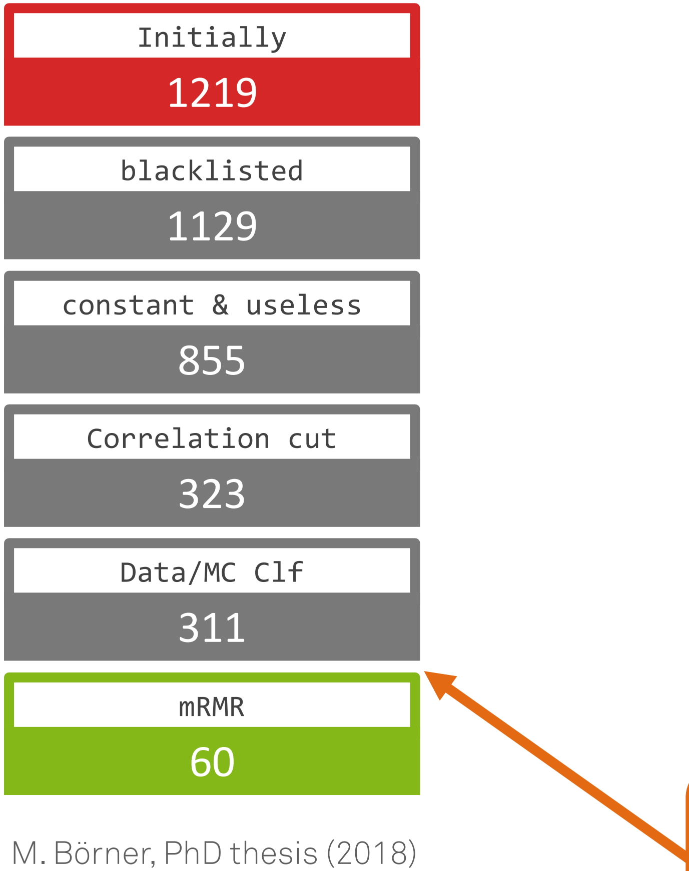
Constant features do not carry information.

Strongly correlated features do not contain new information (or only very little)

Simulated and experimental data should agree to not bias the result.

M. Börner, PhD thesis (2018)

Feature Selection



M. Börner, PhD thesis (2018)

Exclude features that either bias the selection or are only present in simulation.

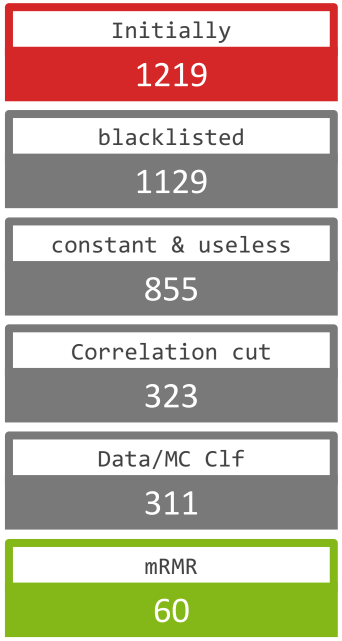
Constant features do not carry information.

Strongly correlated features do not contain new information (or only very little)

Simulated and experimental data should agree to not bias the result.

Automated selection by a feature selection algorithm.

Minimum Redundancy Maximum Relevance



Initially
1219

blacklisted
1129

constant & useless
855

Correlation cut
323

Data/MC Clf
311

mRMR
60

M. Börner, PhD thesis (2018)

- Select features according to relevance and redundancy

- Feature set is built by iteratively adding features that fulfill the following criterion

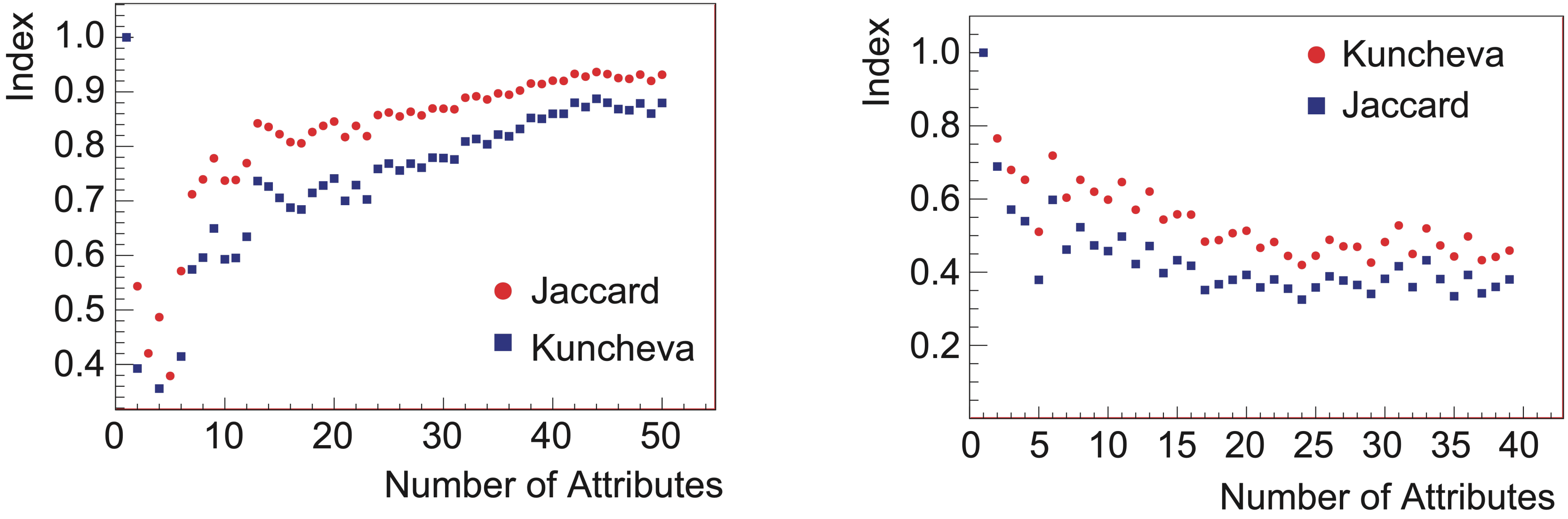
$$\max_{x_j \,\epsilon\, X - S_{m-1}} \left[I(x_j, c) - \frac{1}{m-1} \sum_{x_i \,\epsilon\, S_{m-1}} I(x_i, x_j) \right]$$

Peng, H.C., Long, F., and Ding, C., IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8, pp. 1226–1238, 2005.

Ding, C., & Peng, H., *Journal of bioinformatics and computational biology*, *3*(02), 185-205. (2005)

Feature Selection Stability

Ludmila I. Kuncheva: A STABILITY INDEX FOR FEATURE SELECTION



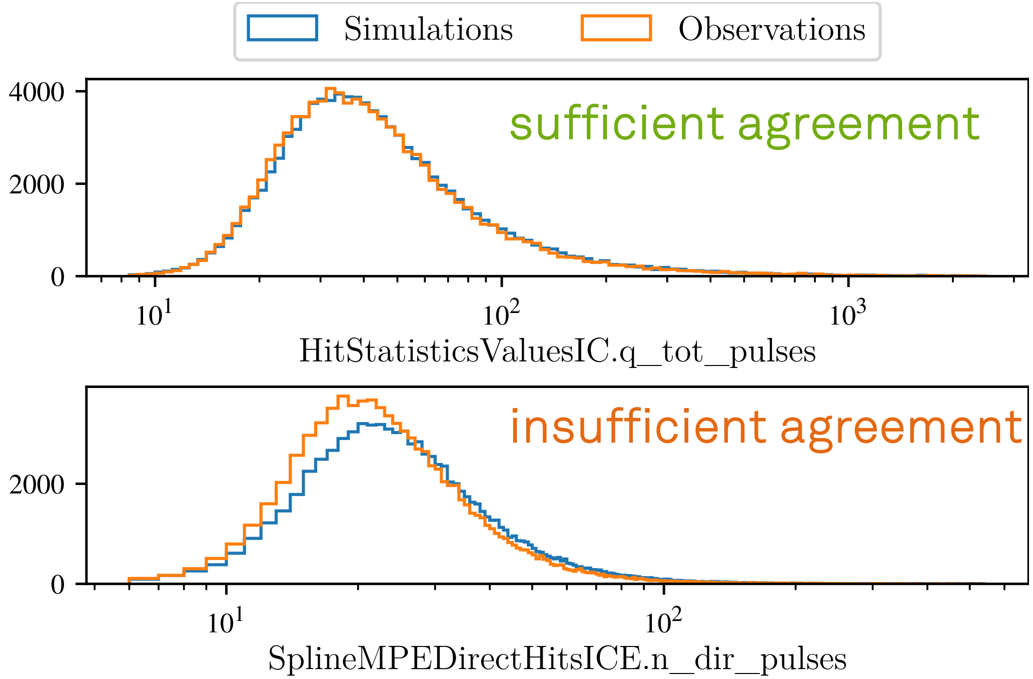
$$k = |A| = |B|$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \qquad I_C(A,B) = \frac{rn - k^2}{k(n-k)}$$

$$r = |A \cap B|$$

n: number of features

Detection of Data/Simulation Disagreements

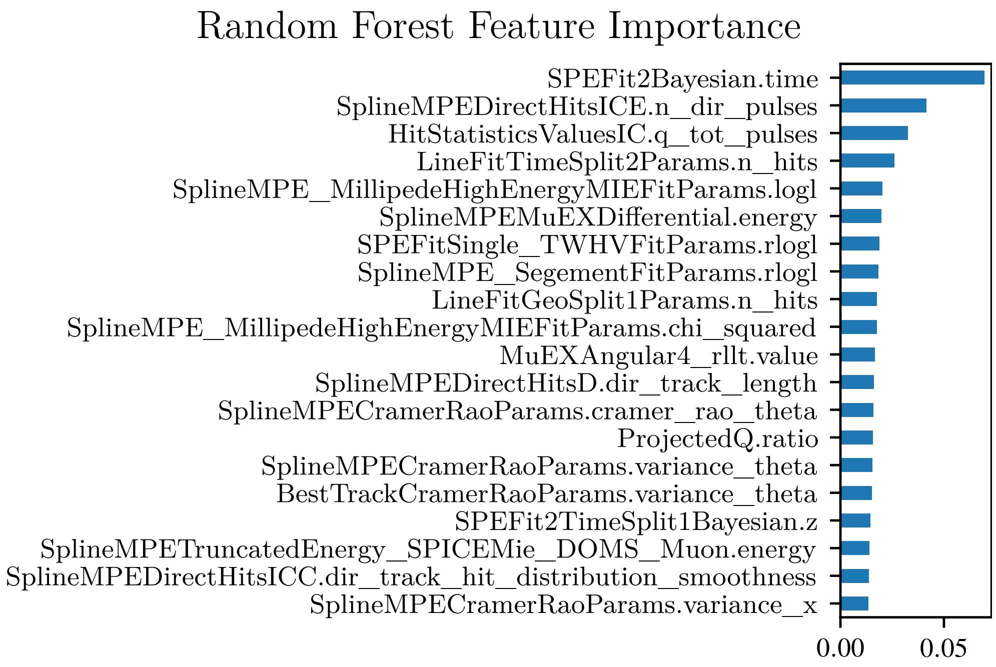


Graphics: M. Linhoff [Learning Under Resource Constraints – Discovery in Physics] (in preparation)

Challenges when inspecting distributions by eye:

- only looking at one-dimensional distributions
- Systematic errors in simulation will also affect correlations between features
- Which metric ???
- Which threshold ???

Detection of Data/Simulation Disagreements

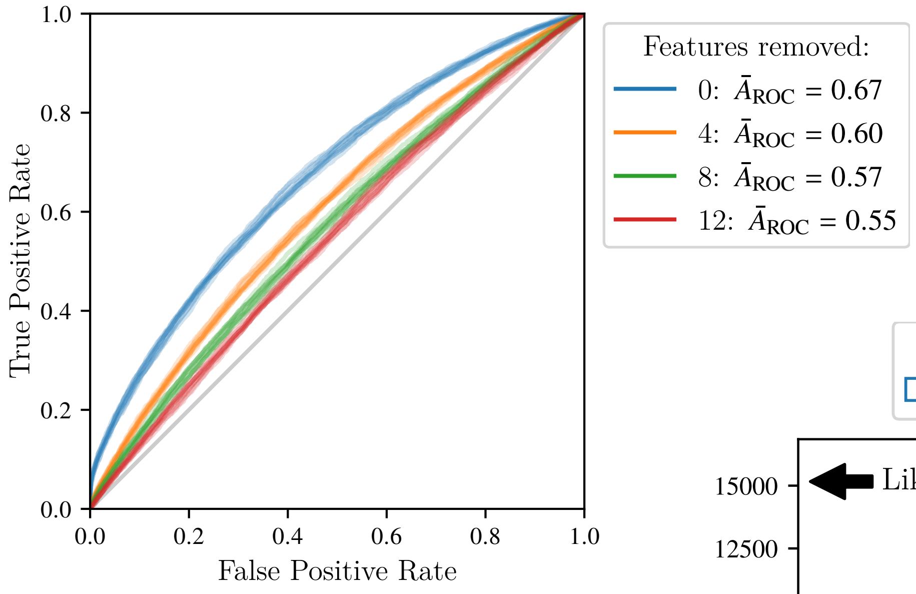


Graphics: M. Linhoff [Learning Under Resource Constraints – Discovery in Physics] (in preparation)

General Idea:

- Train classifier to distinguish simulated and experimental data
- Hard to impossible for a perfect agreement
- Sort features according to their importance
- Discard top n features
- Advantage: Extent to which mismatches can be tolerated is set by the classifier

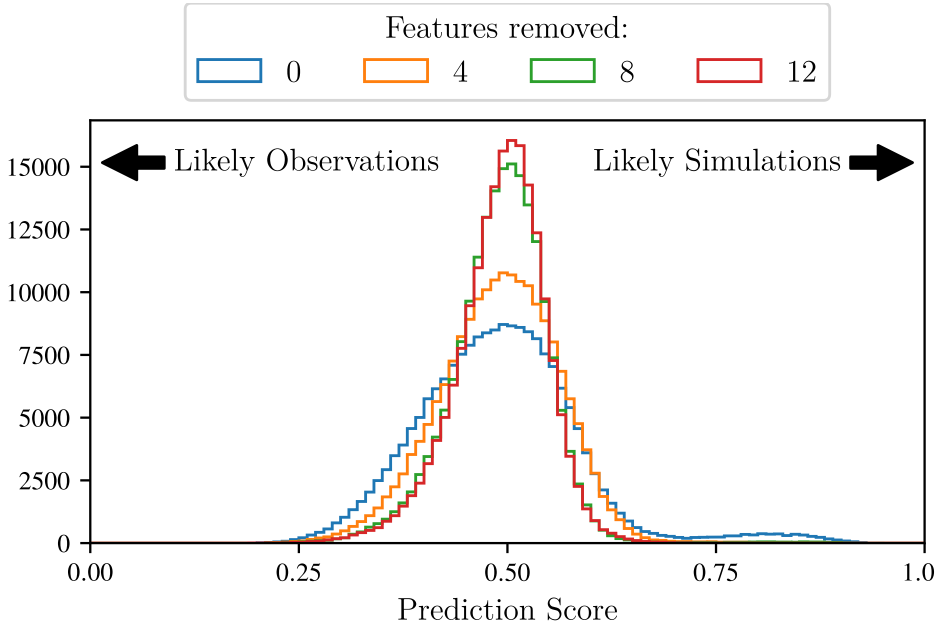
Detection of Data/Simulation Disagreements



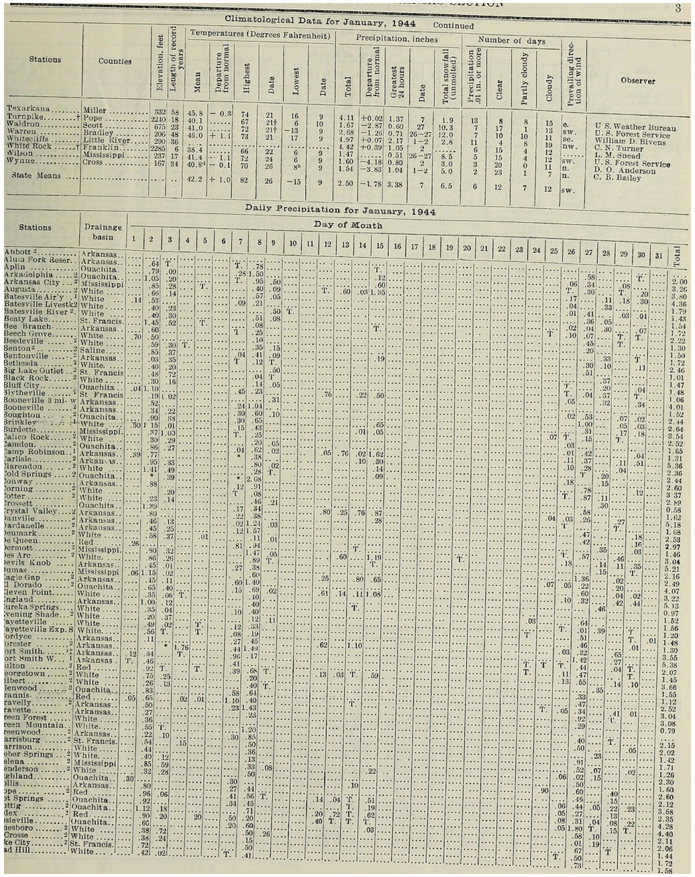
Graphics: M. Linhoff [Learning Under Resource Constraints – Discovery in Physics] (in preparation)

Area under Curve is close to 0.5 (close to random guess).

Prediction score centered around 0.5 (close to random guess).



Replacing Missing Values



SECTION 3

Climatological Data for January, 1944 Continued

Stations	Counties	Elevation, feet	Length of record years	Temperatures (Degrees Fahrenheit)						Precipitation, inches					Number of days				Prevailing direction of wind	Observer
				Mean	Departure from normal	Highest	Date	Lowest	Date	Total	Departure from normal	Greatest 24 hours	Date	Total snowfall (unmelted)	Precipitation .01 in. or more	Clear	Partly cloudy	Cloudy		
Texarkana	Miller	332	58	45.8	− 0.3	74	21	16	9	4.11	+0.02	1.37	7	1.9	13	8	8	15	e.	U S. Weather Bureau
Turnpike†	Pope	2240	18	40.1	......	67	21†	6	10	1.67	−2.87	0.60	27	10.3	7	17	1	13	sw.	U. S. Forest Service
Waldron	Scott	675	23	41.0	......	72	21†	−13	9	2.68	−1.26	0.71	26−27	12.0	7	10	10	11	se.	William D. Bivens
Warren	Bradley	206	48	46.0	+ 1.1	73	21	17	9	4.97	+0.07	2.17	1−2	2.8	11	4	8	19	nw.	C. N. Turner
Whitecliffs	Little River	290	36	.....	......	......	......	......	......	4.42	+0.39	1.05	2	......	6	15	4	12	......	L. M. Snead
White Rock†	Franklin	2285	6	38.4	......	66	22	6	9	1.47	......	0.51	26−27	8.5	5	15	4	12	sw.	U. S. Forest Service
Wilson	Mississippi	237	17	41.4	− 1.1	72	24	6	9	1.60	−4.18	0.80	2	3.0	3	20	0	11	n.	D. O. Anderson
Wynne	Cross	167	34	40.8[d]	− 0.1	70	26	8[h]	9	1.54	−3.83	1.04	1−2	5.0	2	23	1	7	n.	C. B. Bailey
State Means	....	....	..	42.2	+ 1.0	82	26	−15	9	2.50	−1.78	3.38	7	6.5	6	12	7	12	sw.	

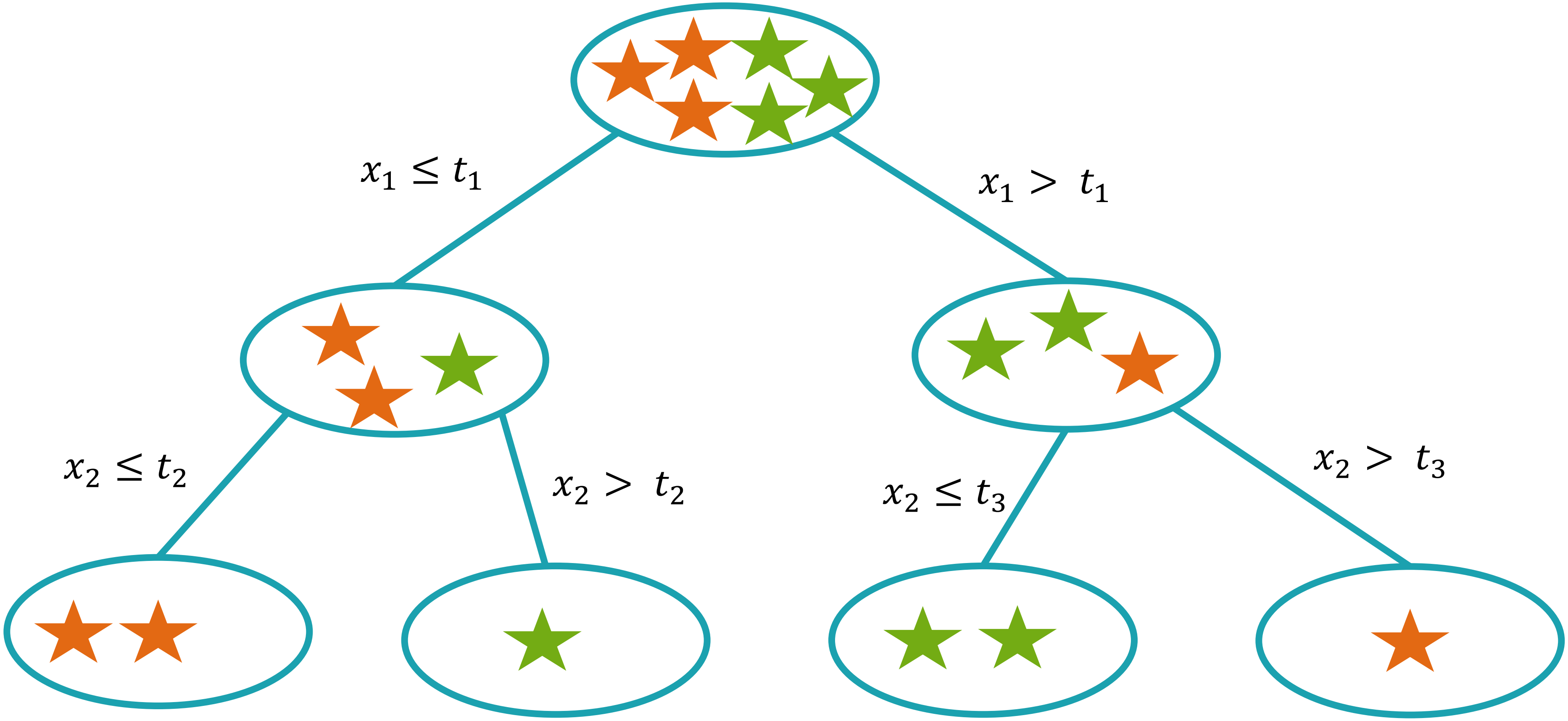
Daily Precipitation for January, 1944

Stations	Drainage basin	Day of Month																															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	Total
Abbott [2]	Arkansas																																....
Alum Fork Reser.	Arkansas		.64	T.				T.	.78							T.												.58			T.		2.00
Aplin	Ouachita		.79	.09				.28	1.50							.12											.06	.34		.08			3.26
Arkadelphia [2]	Ouachita		1.05	.20				T.	.95	.50						.60											T.	.30		T.	.20		3.80
Arkansas City [2]	Mississippi		.85	.28		T.			.40	.09			T.	.60	.03	1.35											.17		.11	.18	.30		4.36
Augusta [2]	White		.66	.14					.57	.05																	.04		.33				1.79
Batesville Air'y [1]	White	.14	.53					.09	.21																		.01	.41		.03	.01		1.43
Batesville Livestk [2]	White		.40	.23						.50	T.																	.36	.05				1.54
Batesville River [2]	White		.40	.30					.51	.08																	.02	.04	.30		.07		1.72
Beaty Lake	St. Francis		1.45	.52		T.			.08							T.										T	.10	.07		T.	T.		2.22
Bee Branch	Arkansas		.60					T	.25																			.45		T.			1.30
Beech Grove	White	.70	.50						.10																			.20					1.50
Beedeville [2]	White		.59	.30	T.				.35	.15																			.33		T		1.72
Benton [2] [2]	Saline		.85	.37				.04	.41	.09						.19												.30	.10		.11		2.46
Bentonville [2]	Arkansas		.03	.35				T	.12	T.																		.51					1.01
Bethesda [2]	White.		.40	.20						.50																			.37				1.47
Big Lake Outlet [2]	St. Francis		.48	.72					.04	T																	T.		.20		.04		1.48
Black Rock [2]	White		.30	.16					.14	.05																	T.	.04	.37		T.		1.06
Bluff City	Ouachita	.04	1.10					.45	.23				.76		.22	.50											.05		.32		.34		4.01
Blytheville [2]	St Francis		.19	1.02						.31																							1.52
Booneville 3 mi- w	Arkansas		.52					.24	1.04																		.02	.53		.07	.02		2.44
Booneville [2]	Arkansas		.34	.22				.30	.60	.10																		1.00		.05	.03		2.64
Boughton [2]	Ouachita		.90	.38				.30	.65							.65												.31		.17	.18		3.54
Brinkley [1]	White	.50	1 15	.01				.15	.43						.01	.05										.07	T.	.15		T.			2.52
Burdette	Mississippi.		.37	1.00				T.	.25																		.03						1.65
Calico Rock [2]	White		.30	.29					.20	.05																	.01	.42			.04		1.31
Camden [2]	Ouachita		.86	.27				.04	.62	.02			.05	.76	.02	1.62											.11	.37		.11	.51		5.36
Camp Robinson [1]	Arkansas	.39	.77					*	.38						.10	.30											.10	.28		.04			2.36
Carlisle [2]	Arkansas		.95	.33					.80	.02						.14												T	.20				2.44
Clarendon [2]	White		1.41	.49					.28	T.						.09											.18		.15				2.60
Cold Springs [2]	Ouachita		*	.39				*	2.08																			.78			.12		3.37
Conway	Arkansas		.88					.12	.91																		T.	.87	.11				2.89
Corning [2]	White			.20				T	.08																				.30				0.58
Cotter [2]	White		.23	.14					.46	.21																		.58					1.62
Crossett	Ouachita		1.39					.17	.34				.80	.25	.76	.87										.04	.03	.26		.27			5.18
Crystal Valley [2]	Arkansas		.80					.22	.38							.28											T			T.			1.68
Danville [2]	Arkansas		.46	.13				.02	1.24	.03																		.47			.18		2.53
Dardanelle [2]	Arkansas		.45	.25				.12	1.57																			.42			.16		2.97
Denmark [2]	White		.58	.37		.01			.11	.01																			.35		.03		1.46
De Queen	Red	.26						.81	.94						T.												T.	.57		.46			3.04
Dermott [2]	Mississippi.		.80	.32					1.47	.05				.60		1.19					T.						.18		.14	11	.35		5.21
Des Arc [2]	White.		.86	.26					.89	T.						T.													.15		T.		2.16
Devils Knob	Arkansas		.45	.01				.27	.38																			1.36		.02			2.49
Dumas	Mississippi	.06	1.15	.02					.60				.25		.80	.65										.07	.05	.22		.20			4.07
Eagle Gap [2]	Arkansas		.45	.11				.60	1.40																			.60		.04	.02		3.22
El Dorado [2]	Ouachita		.65	.40				.15	.69	.02			.61	.14	.11	1.08											.10	.32		.42	.44		5.13
Eleven Point	White		.35	.06	T.				.10																				.46				0.97
England	Arkansas		1.00	.12					.40						T.																		1.52
Eureka Springs	White		.35	.04				.10	.40																.03			.64					1.56
Evening Shade [2]	White		.20	.37					.12	.11																	T.	.01	.39		T		1.20
Fayetteville	White		.49	.02		T.		.12	.33																T			.51			T.	.01	1.48
Fayetteville Exp. S	White.		.56	T.		T.		.08	.19																			.46			.01		1.30
Fordyce	Arkansas		.11					.27	.45				.62		1.10												.03	.32		.65			3.55
Forester [2]	Arkansas			*	1.76			.44	1.49																			1.42		.27			5.38
Fort Smith [1]‡	Arkansas	.12	.34		T.			.96	.17																T.	T	T.	.44		.04	T.		2.07
Fort Smith W. [1]	Arkansas	T.	.46					.41																	T.		.11	.47			T.		1.45
ulton [2]	Red		.92	T.		T.		.39	.68	T.			.13	.03	T.	.59											.13	.55		.14	.10		3.66
eorgetown [2]	White		.75	.25					.20																				.35				1.55
ilbert [2]	White		.26	.13					.40	T.																		.33					1.12
lenwood [2]	Ouachita		.83					.58	.64																			.47					2.52
rannis	Red	.05	.65		.02	.01		1.10	.40						T.											T.	.05	.34		.41	01		3.04
ravelly [2]	Arkansas		.50					.23	1.43																			.92		T.			3.08
ravette	Arkansas		.27						.23																			.29					0.79
reen Forest	White		.36																														....
reen Mountain	White.		.55	T.					1.20																			.40		T.			2.15
reenwood [2]	Arkansas		.22	.10				.30	.85																			.50			.05		2.02
arrisburg [2]	St. Francis		.54		.15				.50																				.23				1.42
arrison	White		.44						.36																			.91					1.71
eber Springs [2]	White.		.40	.12					.13																			.52	.07		.02		1.26
elena [2]	Mississippi		.85	.59					.33	.08						.22											.06	.02	.15				2.30
enderson [2]	White		.32	.28					.50																			.50					1.60
ighland	Ouachita	.30						.30							.10											.90		.60		.40			2.60
ollis	Arkansas		.80					.27	.44																			.46		.15			2.12
ope [2]	Red		.96	.06				.41	.56	T.			.14	.04	T.	.51											.06	.44	.05	.22	.23		3.68
ot Springs	Ouachita		.92					.34	.45						T.	.19											.05	.27		.13			2.35
uttig [2]	Ouachita		1.12	.18					.71				.20	.72	T.	.62											.08	.31	.04	.08	.22		4.28
dex [2]	Red		.90	.20		20		.50	.20				.40	T.	T.	T.											.05	1.80	T	.15	T.		4.40
ssieville	Ouachita		.60					.20	.60							.03												.58	.10				2.11
nesboro [2]	White		.38	.72					.50	.26																		.01	.19				2.06
Crosse [2]	White		.38	.24					.15																			.67		T			1.44
ke City [2]	St. Francis		.72						.50																		T.	.50					1.72
ad Hill	White		.42	.02				T.	.41																			.73					1.58

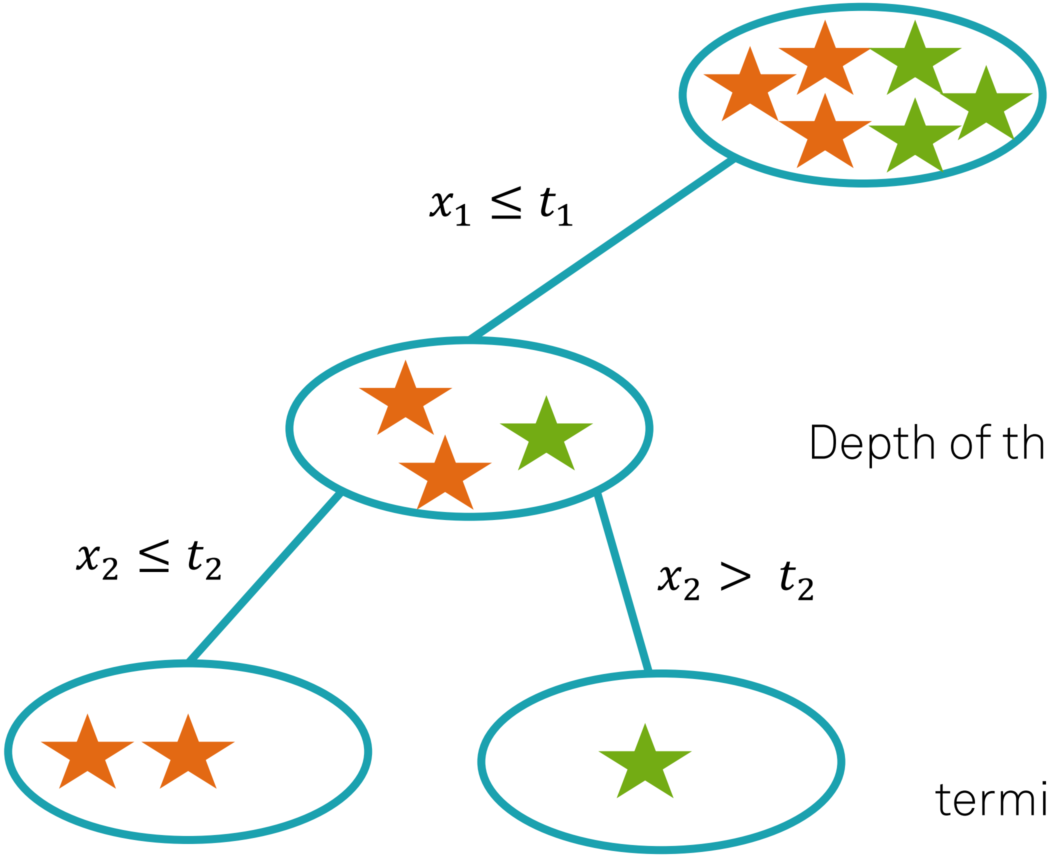
- Some algorithms can only handle numerical values
- Missing values (nans, infs, …) can be replaced
  - Average
  - Median
  - Constant
  - …
- Features with too many missing values can be excluded
- Missing value can actually provide valuable information, e.g. reconstruction algorithm failed because this is an event with poor information

Source: By Internet Archive Book Images - https://www.flickr.com/photos/internetarchivebookimages/14590021147/Source book page: https://archive.org/stream/climatologicalda4753unit/climatologicalda4753unit#page/n168/mode/1up, No restrictions, https://commons.wikimedia.org/w/index.php?curid=41910405

Decision Trees



Decision Trees Nomenclature



$x_1 \leq t_1$

$x_2 \leq t_2$

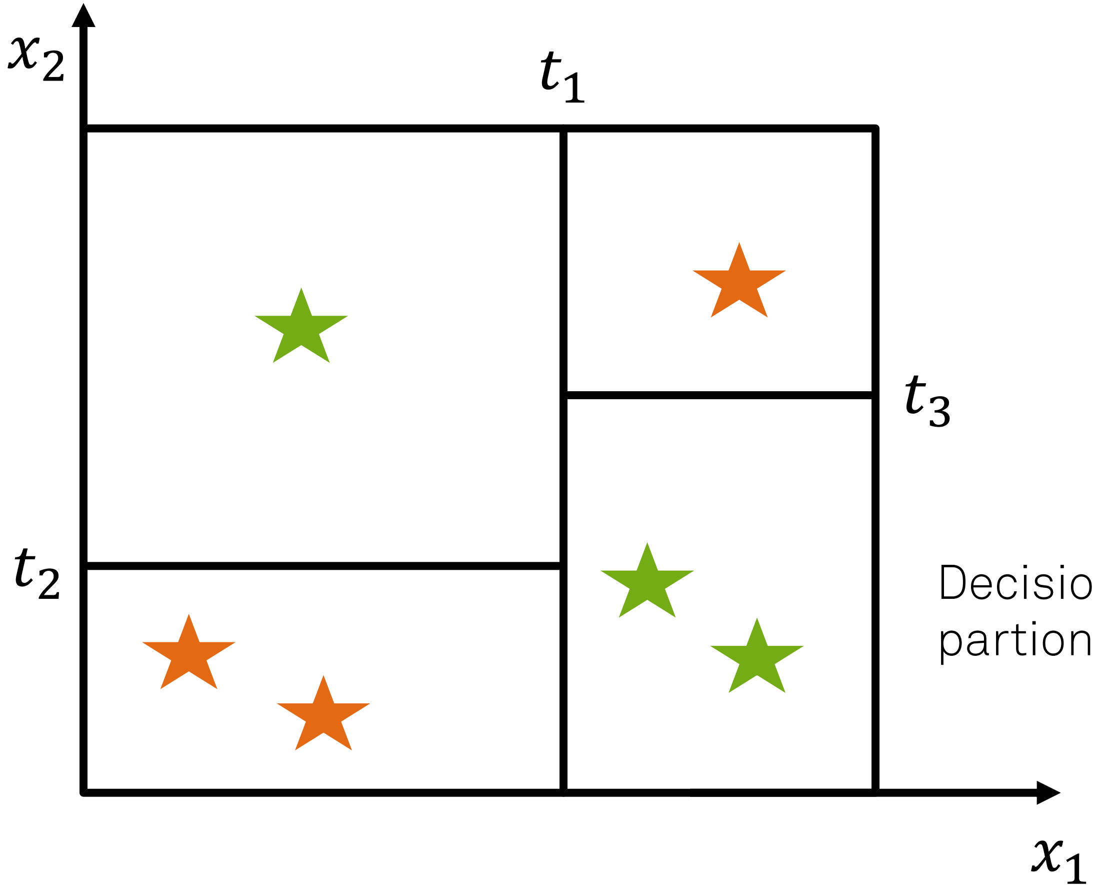
$x_2 > \ t_2$

root node

Depth of the tree: 3

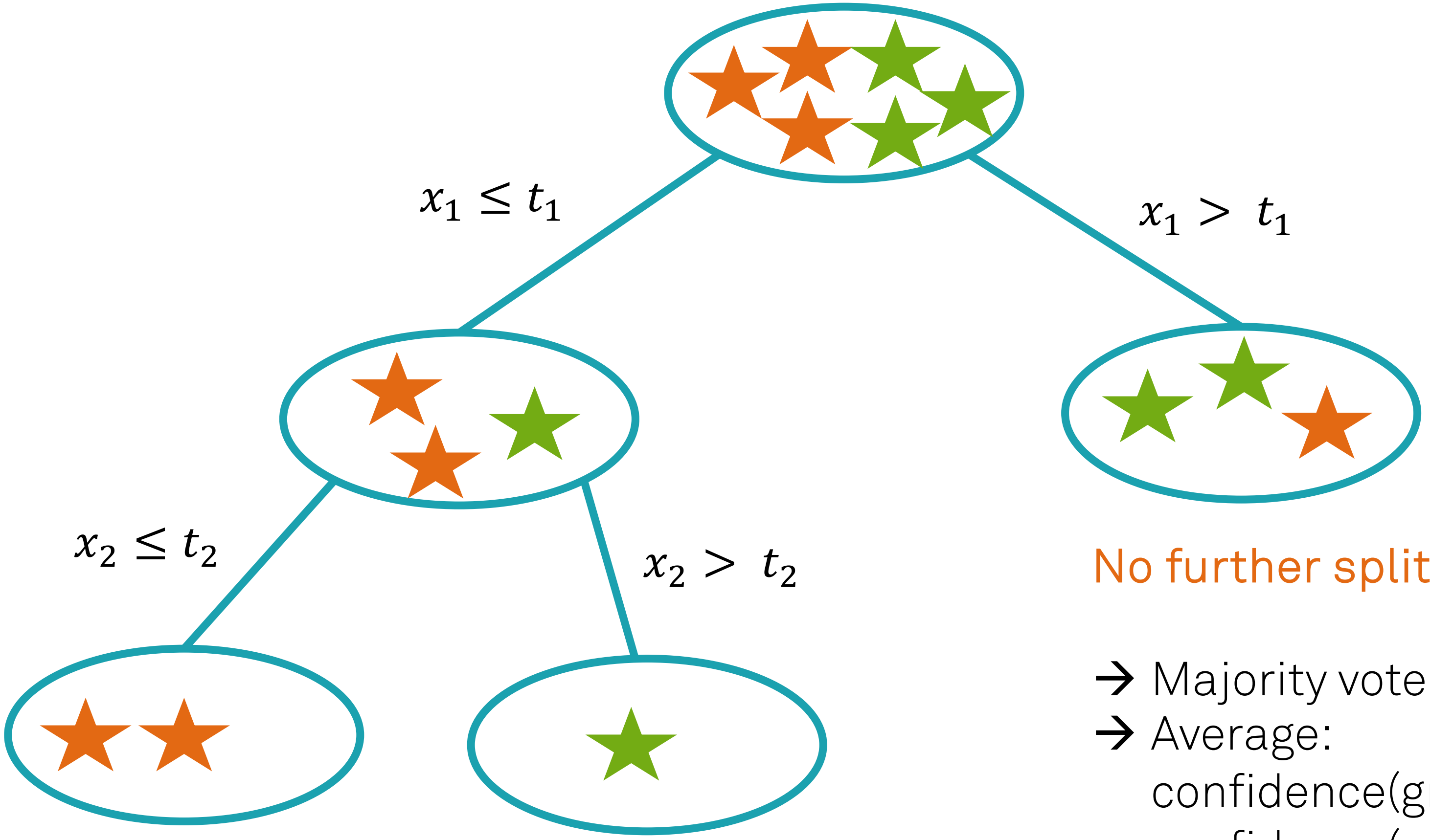
terminal nodes/leaves

Decision Trees



Decision Trees achieve a binary partioning of the feature space.

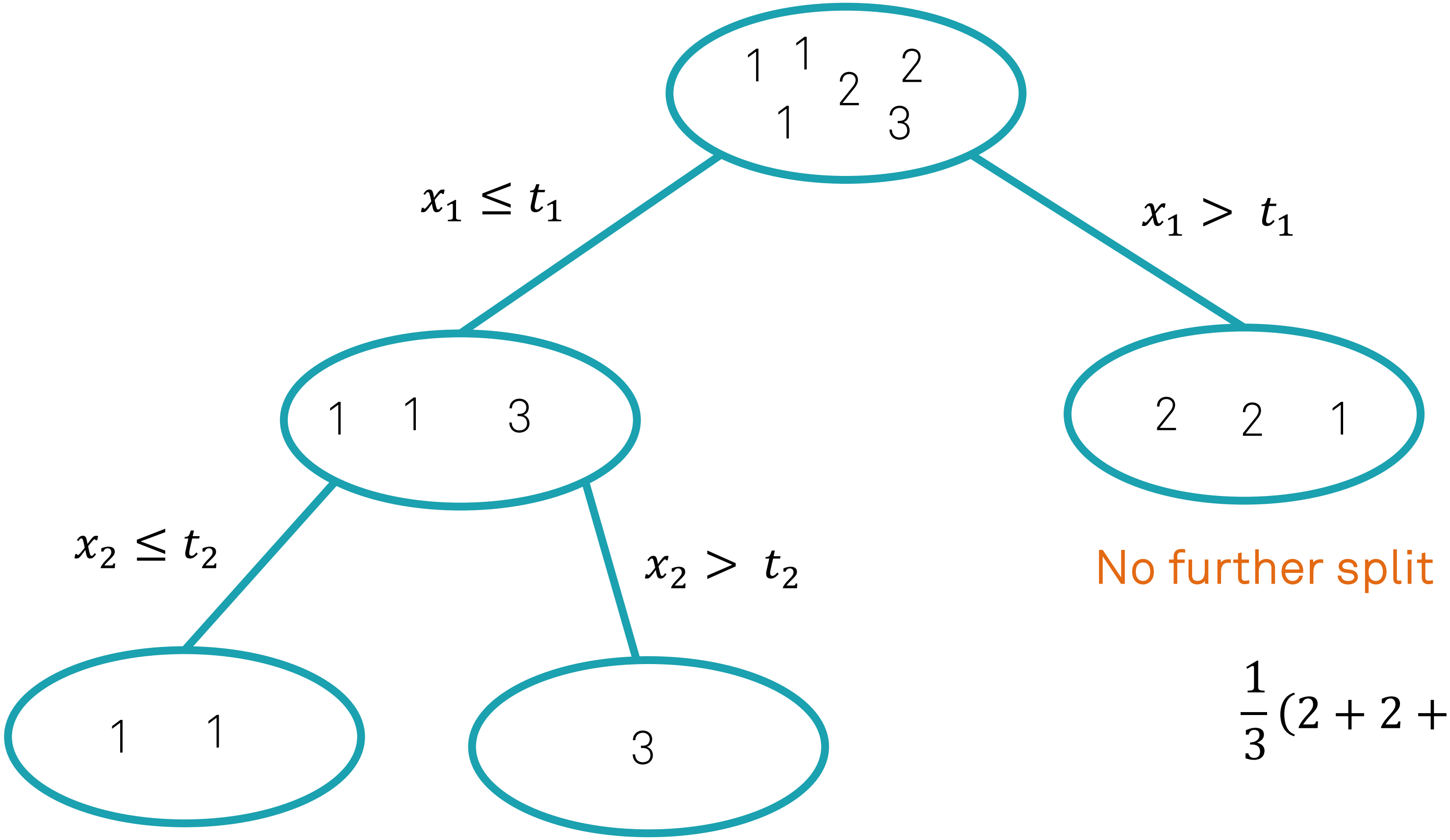
Decision Trees



No further split possible

→ Majority vote: green
→ Average:
confidence(green)= 2/3,
confidence(orange) = 1/3

Decision Trees for Regression



No further split possible

$$\frac{1}{3}(2+2+1)=\frac{5}{3}$$

How to decide where to split?

Impurity Measures:

$P(\omega_j)$: Fraction of patterns at node N in class ω_j

Cross entropy:

$$i(N) = -\sum_{j=1}^{K} P(\omega_j) log_2(\omega_j)$$

Gini Impurity:

$$i(N) = \sum_{i \neq j} P(\omega_i) P\left(\omega_j\right) = \frac{1}{2}\left[1 - \sum_{j} P^2\,(\omega_j)\right]$$

How to decide where to split?

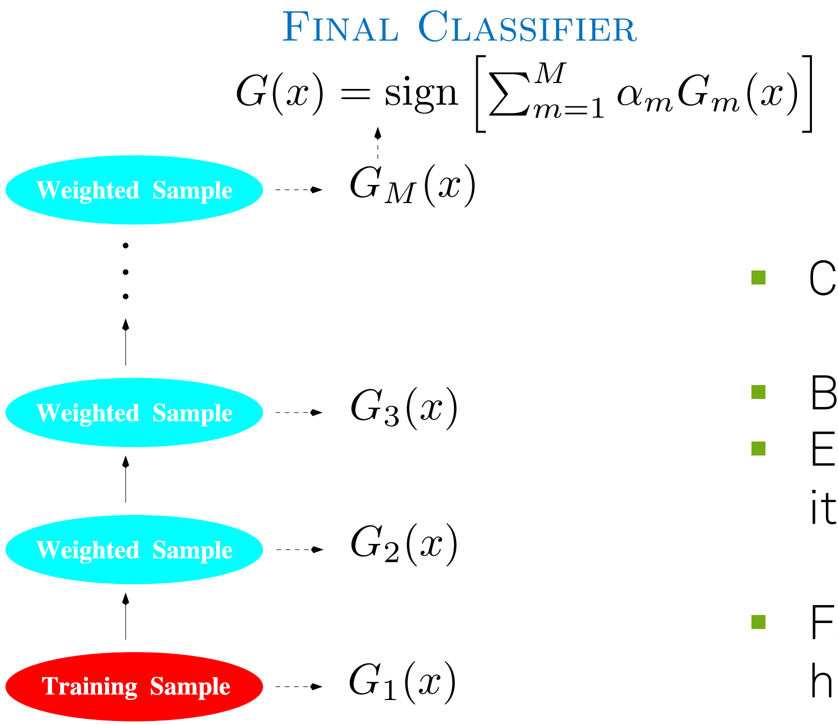
Misclassification impurity:

$$i(N) = 1 - max_j P(\omega_j)$$

To determine the optimal split consider the decrease in impurity:

$$\Delta i(N) = i(N) - P_L \cdot i(N_L) - (1 - P_L) \cdot i(N_R)$$

Boosted Decsion Trees



FINAL CLASSIFIER

$$G(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m G_m(x)\right]$$

Weighted Sample $G_M(x)$

Weighted Sample $G_3(x)$

Weighted Sample $G_2(x)$

Training Sample $G_1(x)$

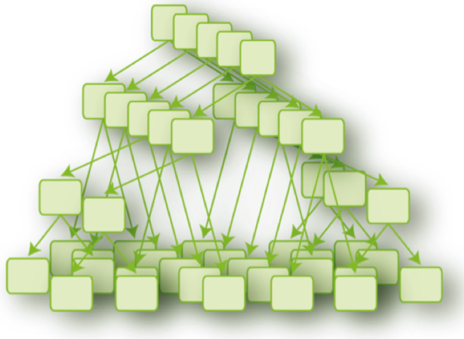
- Classifiers are weighted by
  $$\alpha_m = \log((1 - err_m)/err_m)$$
- Better classifiers obtain higher weights
- Example weights are updated in every iteration
  $$w_i \leftarrow w_i \cdot \exp(\alpha_m \cdot I(y_i \neq G(x_i))$$
- Falsely classified examples obtain higher weights in the next iteration

Source: Elements of Statistical Learning, Figure 10.1

Random Forest

$x_1 \leq t_1$

Random subset of examples to build each tree.



Random subset features to determine the optimal split

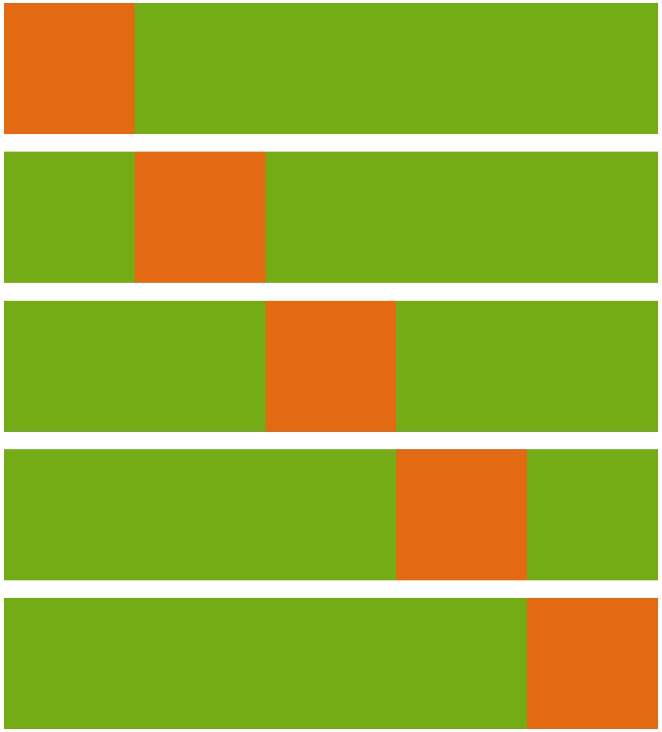
Random forests utilize an ensemble of independent weak classifiers (decision trees) to obtain a better classification.

Final classification is achieved via:

$$c_j = \frac{1}{n_{tress}} \sum_{i=1}^{n_{trees}} c_{ij}$$

c_{ij}: Classification for example j by tree i

Validation

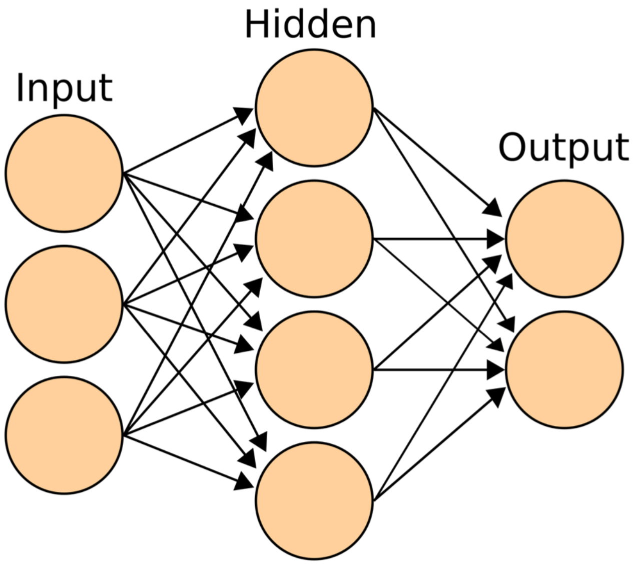


- In order to not optimize on statistical fluctuations in the test set it is advisable to use cross validation for parameter optimization
- It can also be useful to have an additional validation set to validate the performance of the optimized parameter settings

🟩 Use for training

🟧 Use for testing

Neural Networks



Source: By en:User:Cburnett - This W3C-unspecified vector image was created with Inkscape., CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1496812

Source: By Alvesgaspar - Top left: File:Cat August 2010-4.jpg by AlvesgasparTop middle: File:Gustav chocolate.jpg by Martin BahmannTop right: File:Orange tabby cat sitting on fallen leaves-Hisashi-01A.jpg by HisashiBottom left: File:Siam lilacpoint.jpg by Martin BahmannBottom middle: File:Felis catus-cat on snow.jpg by Von.grzankaBottom right: File:Sheba1.JPG by Dovenetel, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17960205

- Highly successful, e.g in image classification
- Hyped

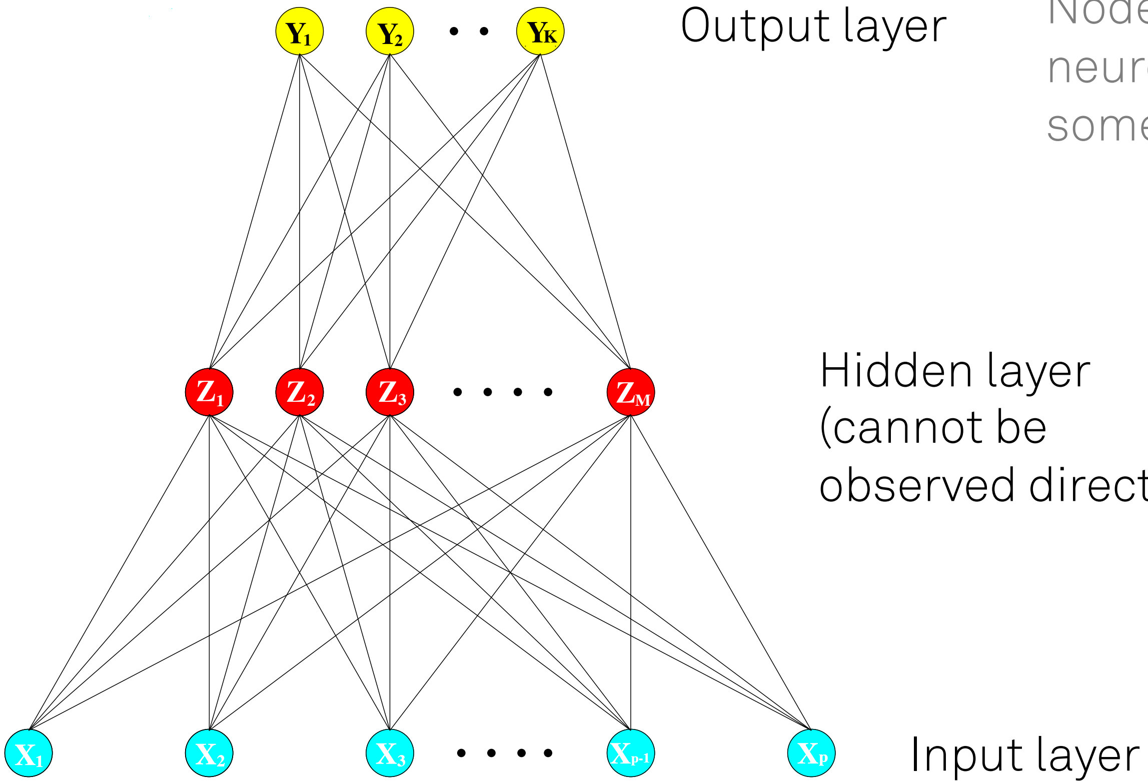
Reasons for Using Neural Networks (in IceCube)

- Improved reconstruction methods will lead to increased sensitivity for the detection of sources
- Hardware limitations at the South Pole
- Events need to be processed in a given time frame to prevent pileup
- Limitations call for robust method that can handle raw data in constant time
- Neural networks are compuattionally inexpensive once the network is trained
- Fixed amount of operations, runtime is (largely) independent of the input
- Translational invariance (position of the classified object does not impact the class)
- Physics of neutrino interaction is invariant in time and space

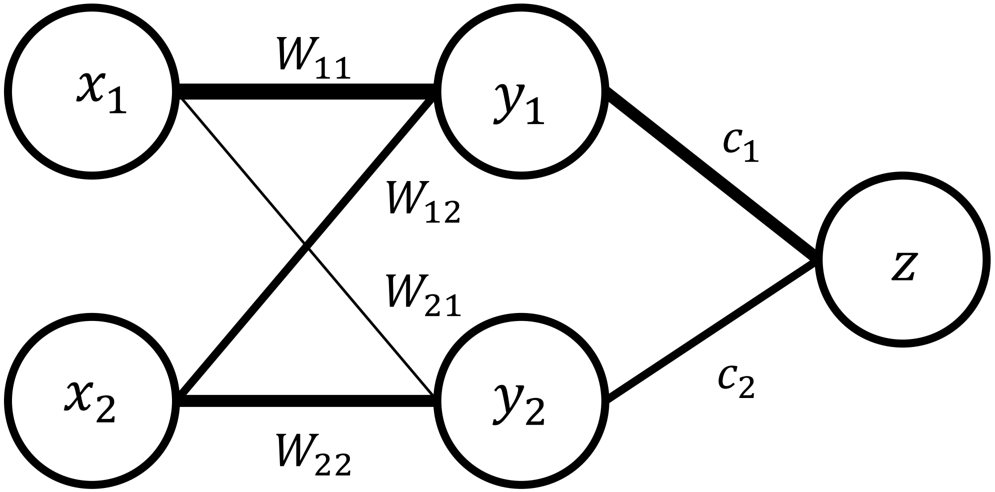
General Idea of Neural Networks

Originally developed to mimic the human brain.

Nodes are sometimes called neurons, and connections are sometimes called synapses.



General Idea of Neural Networks



Nodes are linear combinations of nodes from previous layers.

The task is to optimize the weights such that the estimated label z matches the true label $\hat{z}$.

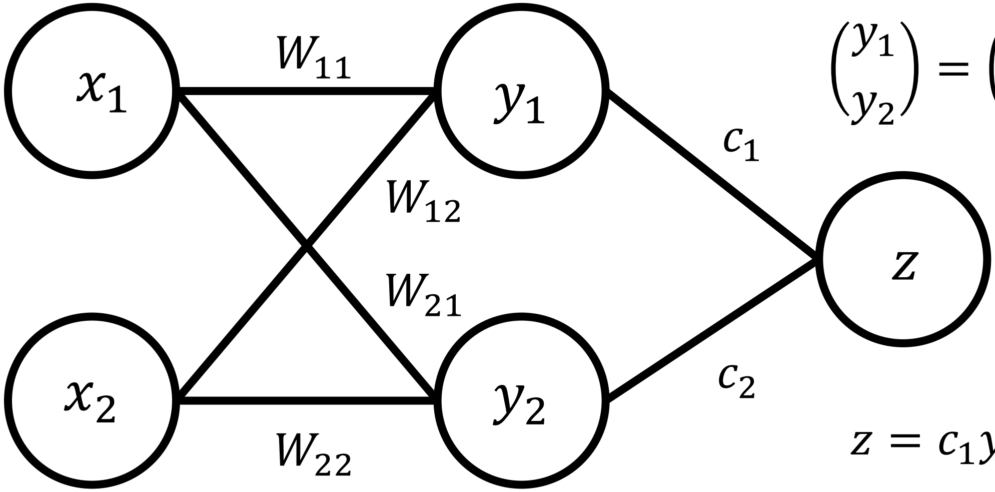
A feed-forward neural network with linear output and at least one hidden layer with a finite number of nodes can approximate any of the above* functions with arbitrary precision**.

*Continuous functions on closed bounded subsets of the Eucledian space $\mathbb{R}^n$.

**Figure and definition adopted from Erdmann et al.

More mathematically speaking

$$y_1 = W_{11}x_1 + W_{12}x_2 + b_1$$



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{11} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

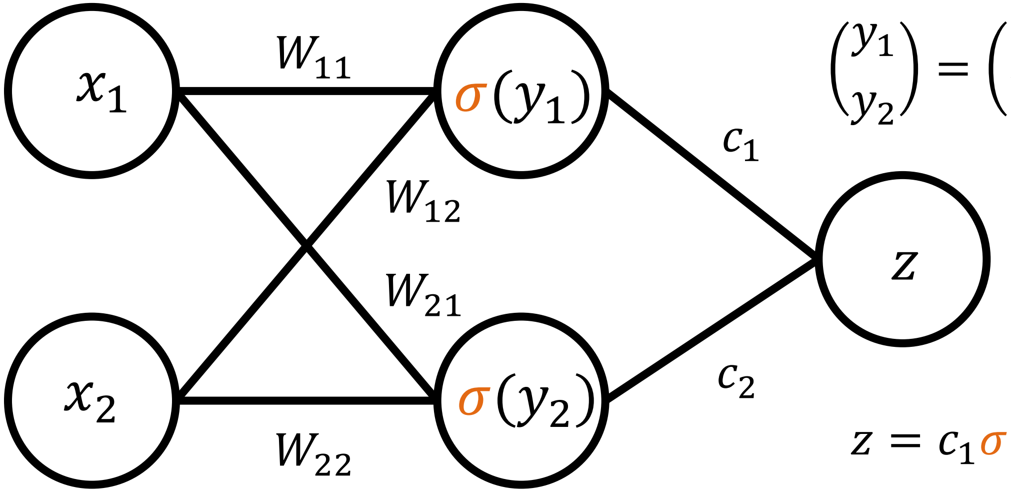
$$z = c_1 y_1 + c_2 y_2 + d$$

$$y_2 = W_{21}x_1 + W_{22}x_2 + b_2$$

**Figure adopted from Erdmann et al.

Adding non-linearity

$$y_1 = W_{11}x_1 + W_{12}x_2 + b_1$$



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{11} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

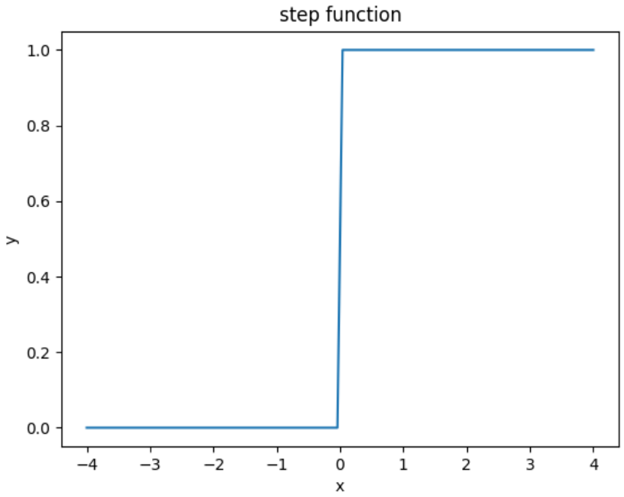
$$z = c_1\sigma(y_1) + c_2\sigma(y_2) + d$$

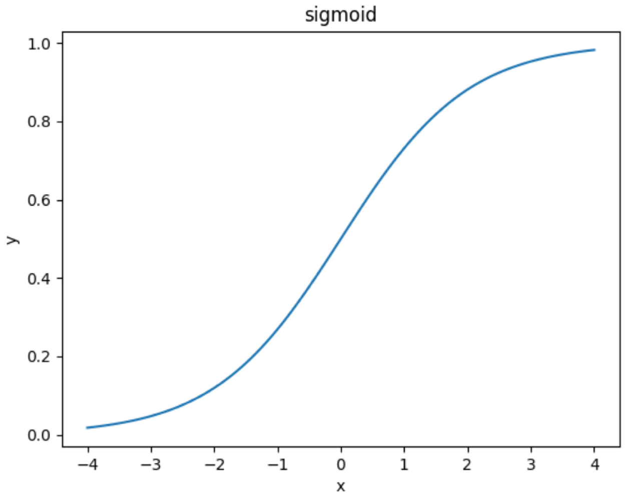
$$y_2 = W_{21}x_1 + W_{22}x_2 + b_2$$

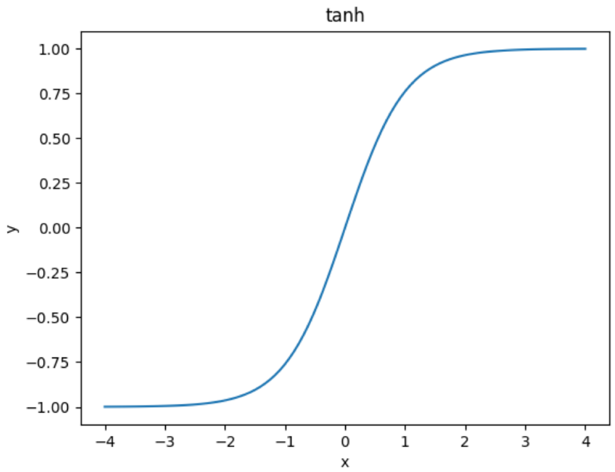
**Figure adopted from Erdmann et al.

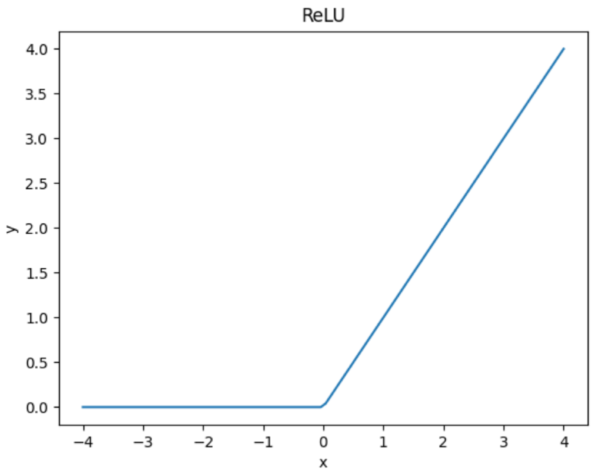
σ is generally referred to as the activation function.

Popular choices for σ









Input Preparation

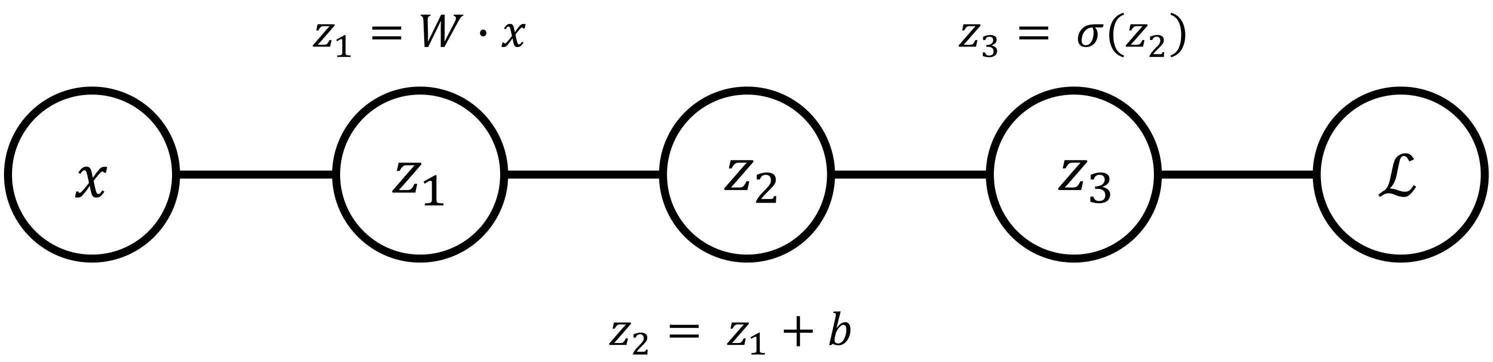
- **Zero-centered:** ReLU changes drastically around 0, $x_i - \langle x_i \rangle$ to include positive and negative values
- **Order of magnitude:** Large variables could be preferred in the network training $x'_i = \frac{x_i - \langle x_i \rangle}{\sigma_i}$
- **Logarithm** to achieve more evenly distributed data
- **Decorrelation:** highly correlated variables should be decorrelated

Epoch and (mini)batch

- Minibatch, Batch: Using all examples can be infeasible in case many parameters need to be optimized, instead random subsets (batches) of examples are used. The optimal size of the batch depends on the problem to be solved. Popular choices are 2^k

- Epoch: Complete use of all examples.

Weight Updates



$$z_1 = W \cdot x$$

$$z_3 = \sigma(z_2)$$

$$z_2 = z_1 + b$$

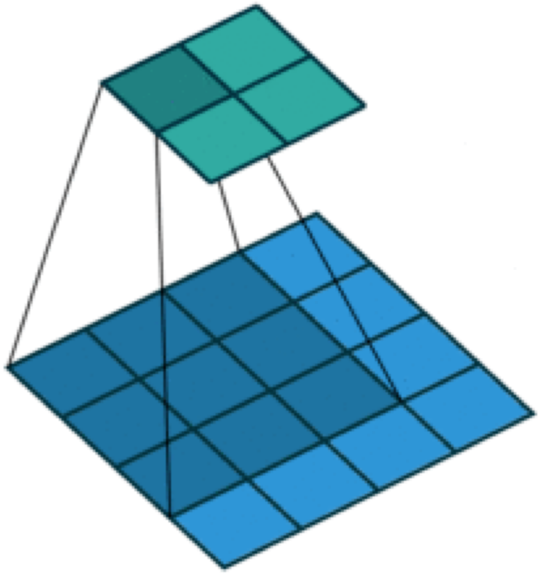
$$\frac{\partial \mathcal{L}}{\partial W} = \frac{\partial \mathcal{L}}{\partial z_3} \cdot \frac{\partial z_3}{\partial z_2} \cdot \frac{\partial z_2}{\partial z_1} \cdot \frac{\partial z_1}{\partial W}$$

$$W_{t+1} = W_t - \alpha \mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial W}\right]_t$$

$$\mathbb{E}\left[\frac{\partial \mathcal{L}}{\partial W}\right] = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{\partial \mathcal{L}}{\partial W}\right)_i$$

This is the basic idea, this will most likely be handled by an optimizer.

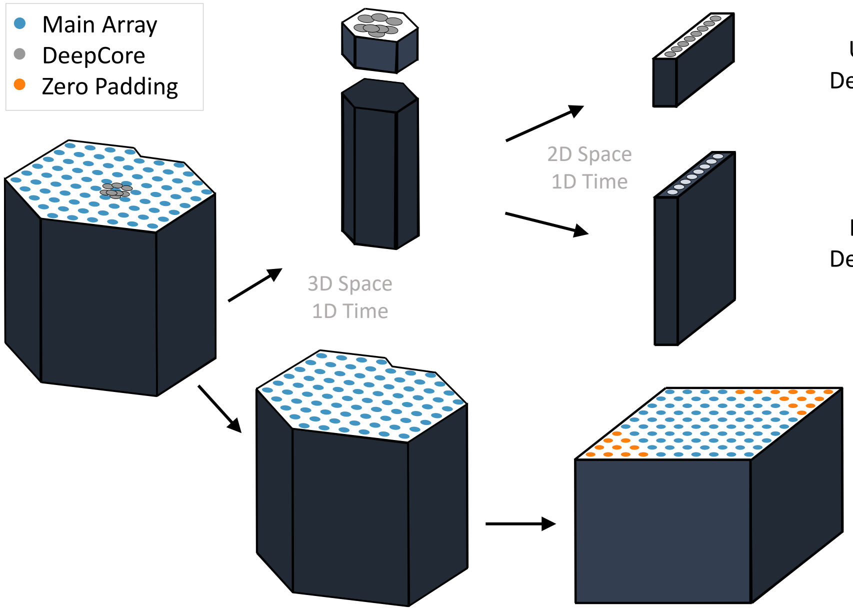
Convolutional Layers



Source: By Vincent Dumoulin, Francesco Visin - https://github.com/vdumoulin/conv_arithmetic, MIT, https://commons.wikimedia.org/w/index.php?curid =78003423

- Considering all pixels in an image in a fully connected network, results in too many parameters to be optimized
- The position of an object in an image should not alter the prediction (translational invariance)
- The convolutional operation exploits the neighbourhood of each pixel

Hexagonal Input Data



● Main Array
● DeepCore
● Zero Padding

3D Space
1D Time

2D Space
1D Time

Upper
DeepCore

(8 x 10 x n)

Lower
DeepCore

(8 x 50 x n)

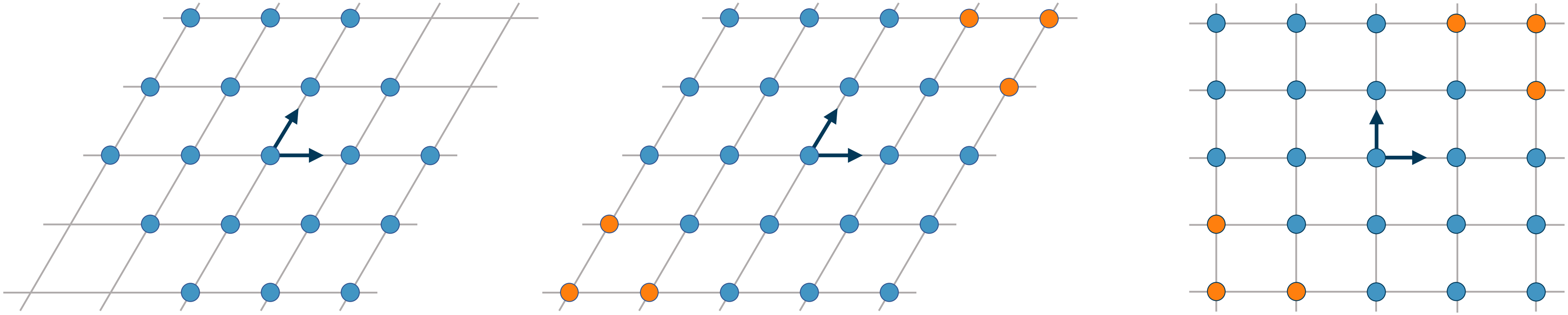
Main Array

3D Space
1D Time

(10 x 10 x 60 x n)

Abbasi et al., JINST, 16 (7) (2004).

Hexagonal Kernels



Abbasi et al., JINST, 16 (7) (2004).

Take-Away Messages

- Machine Learning and esp. Deep Learning is not magic
- Machine Learning and Deep Learning are tools that will help you to accomplish an analysis task faster and more accurately (when used correctly)
- The preprocessing of data is part of machine learning (and very important)
- Not every classifier is suited for every problem (consider runtime)
- If something fast and simple does the job: use it
- Make sure simulated and experimental data agree
- ...