

Mini Course

# From Data Science to interpretable AI

in medical decision support



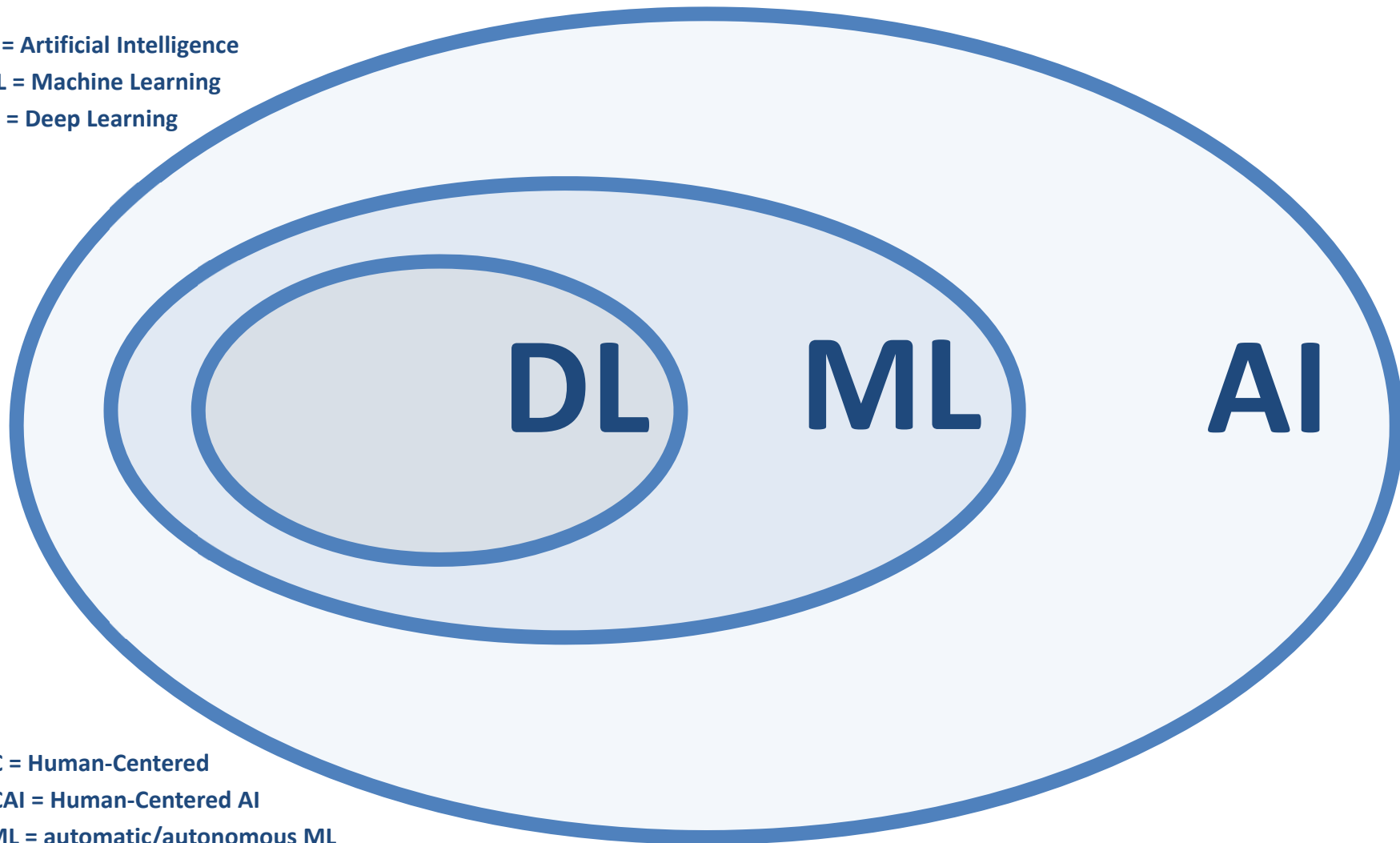
Assoc. Prof. Dr. Andreas HOLZINGER (Medical University Graz)

## Primer on Probability, Information & Learning from Data for AI/Machine Learning

**AI = Artificial Intelligence**

**ML = Machine Learning**

**DL = Deep Learning**



**HC = Human-Centered**

**HCAI = Human-Centered AI**

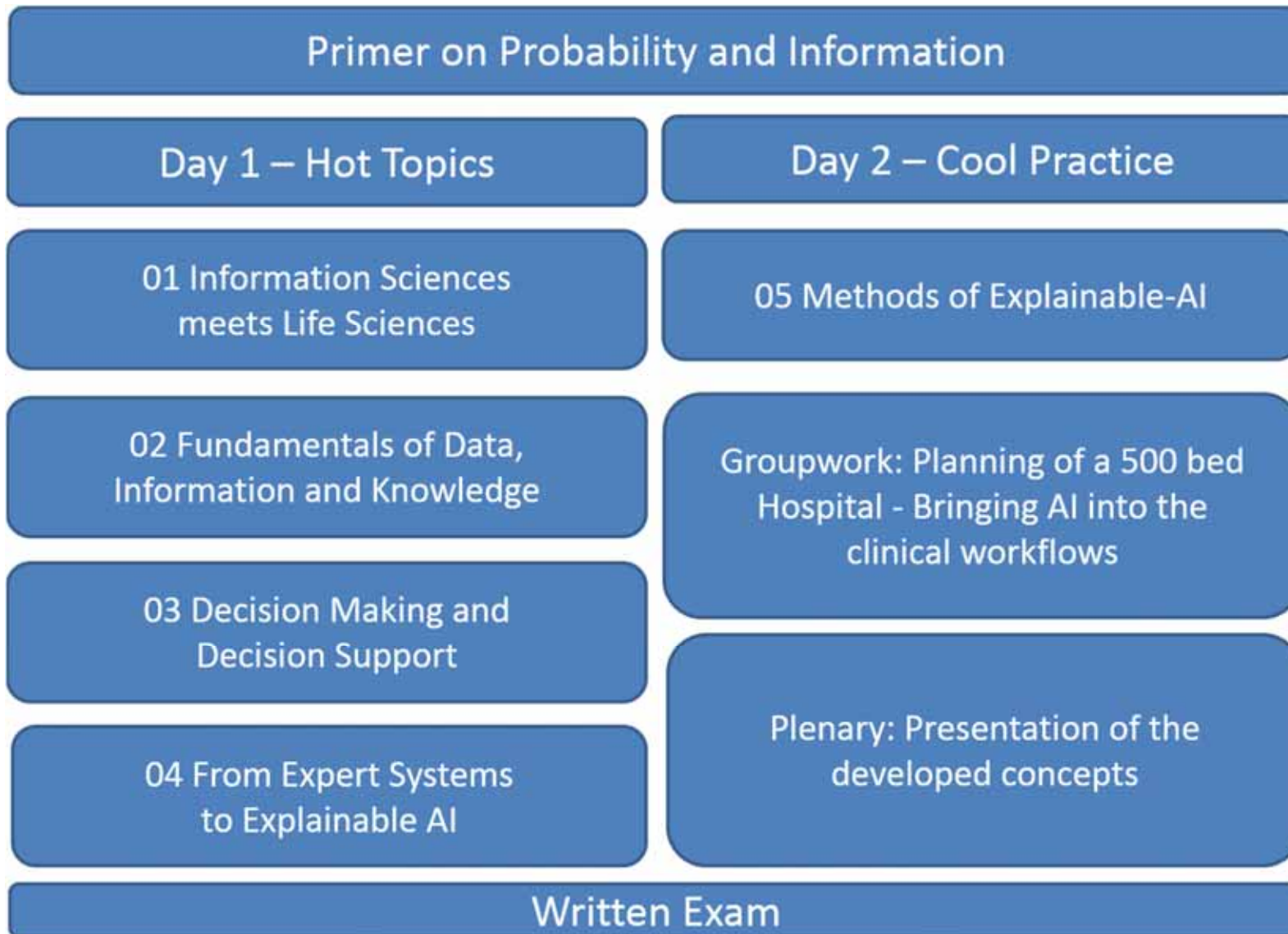
**aML = automatic/autonomous ML**

**iML = interactive ML, interpretable ML**

**KDD = Knowledge Discovery from Data**

**ExAI = explainable AI**

Andreas Holzinger, Peter Kieseberg, Edgar Weippl & A Min Tjoa 2018. Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI. Springer Lecture Notes in Computer Science LNCS 11015. Cham: Springer, pp. 1-8, doi:[10.1007/978-3-319-99740-7\\_1](https://doi.org/10.1007/978-3-319-99740-7_1)

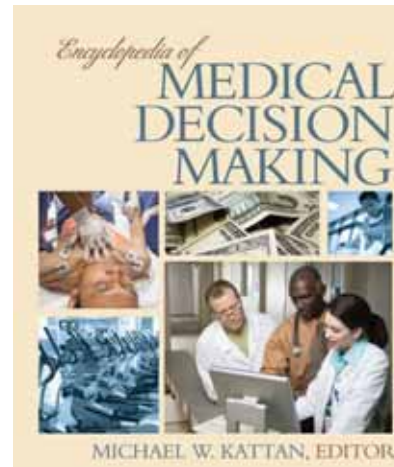


- Probability Distribution
- Probability Density
- Frequentist/Bayesian
- Continuous/Discrete
- Independent/Dependent
- Identical/Non-Identical
- Correlation/Causation
- Joint Probability
- Conditional Probability

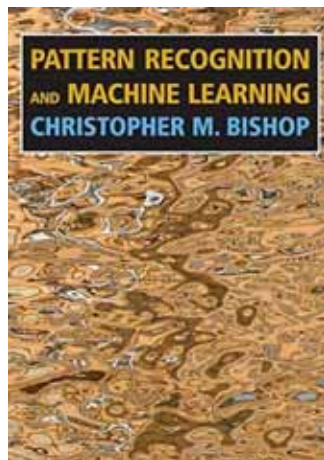
- Information Entropy
- Mutual Information
- Kullback-Leibler Divergence
- Maximum Likelihood Estimation
- Maximum a Posteriori Estimate
- Bayesian Estimate
- Bayesian Learning
- Fisher Information
- Marginal Likelihood

- **01 Probability**
- **02 Expectation and Expected Utility Theory**
- **03 Bayes and Laplace**
- **04 Information Theory & Entropy**
- **05 Cross Entropy & Kullback-Leibler Divergence**
- **06 Optimization**

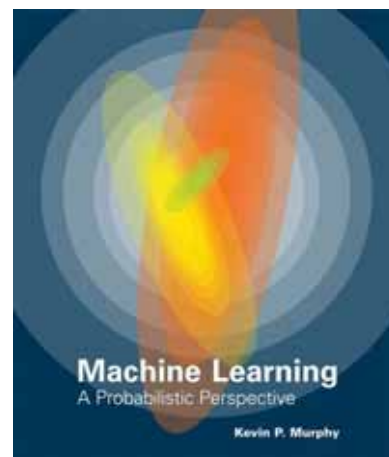
# Book recommendations (selection)



Michael W. Kattan (ed.) 2009.  
Encyclopedia of medical decision making, London: Sage.



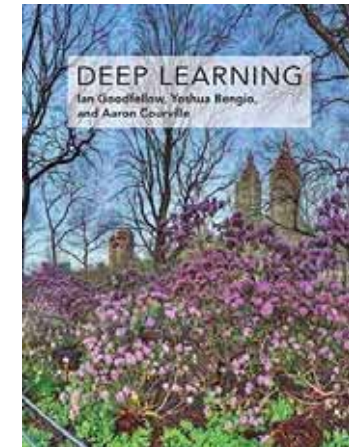
Christopher M. Bishop  
2006. Pattern Recognition  
and Machine Learning,  
New York, Springer.



Kevin P. Murphy 2012.  
Machine learning: a  
probabilistic perspective,  
Cambridge (MA), MIT press.



David Barber 2012. Bayesian  
reasoning and machine  
learning, Cambridge,  
Cambridge University Press.



Ian Goodfellow, Yoshua  
Bengio & Aaron Courville  
2016. Deep Learning,  
Cambridge (MA), MIT Press.

# 01 Probability

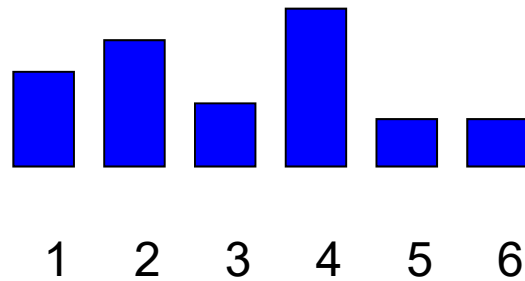


- *Probability*  $p(x)$  is the formal study of laws of chance and managing uncertainty; allows to measure (many) events
  - Frequentist\* view: coin toss
  - Bayesian\* view: probability as a measure of belief (this is what made machine learning successful)
  - $p(x) = 1$  means that all events occur for certain
  - Information is a measure for the reduction of uncertainty
  - If something is 100 % certain its uncertainty = 0
  - Uncertainty is max. if all choices are equally probable (I.I.D = independent and identically distributed)
  - Uncertainty (as information) sums up for independent sources:  $\sum_x p(x = X) = 1$

\*) Bayesian vs. Frequentist - please watch the excellent video of Kristin Lennox (2016): <https://www.youtube.com/watch?v=eDMGDhyDxuY>

# Probability Density Function PDF vs. Cumulative Density Function CDF

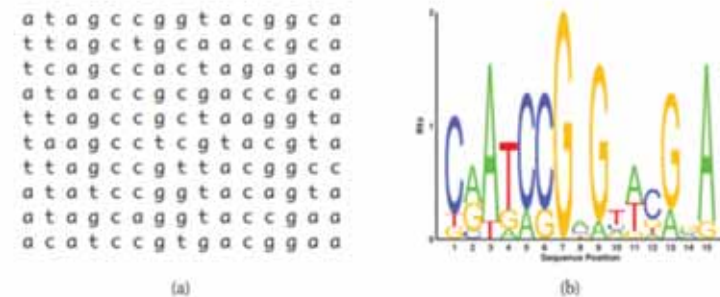
- Discrete distributions:



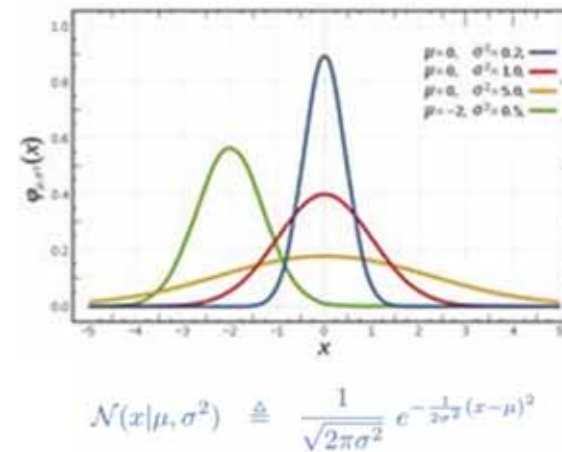
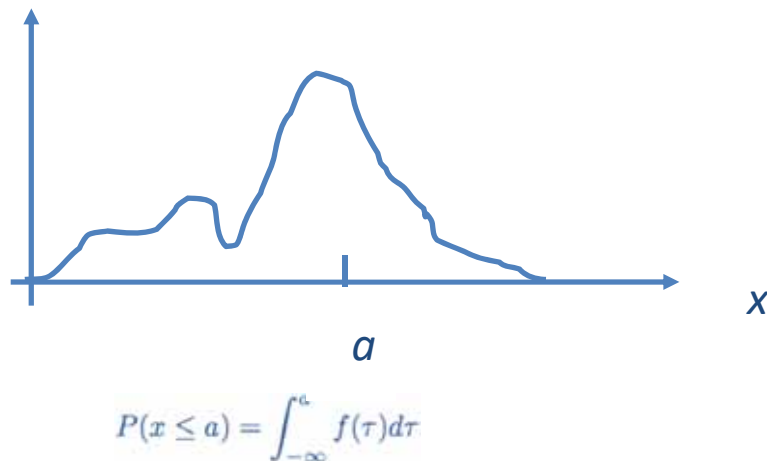
$$\sum_i P(X = x_i) = 1$$

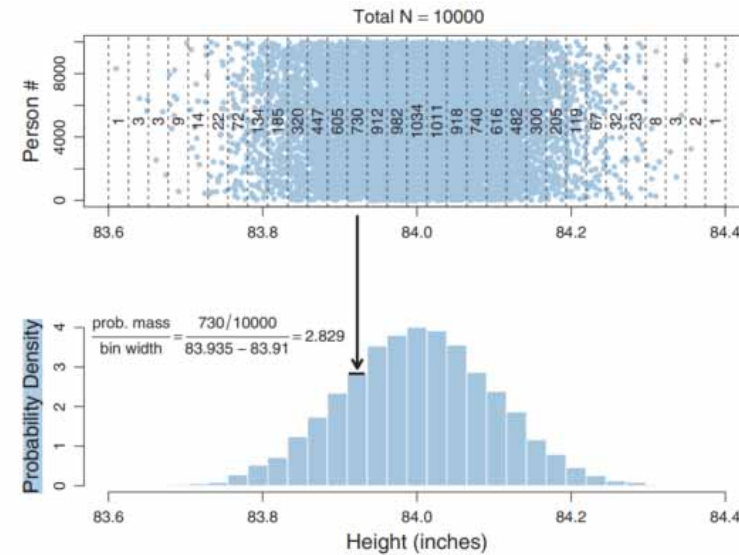
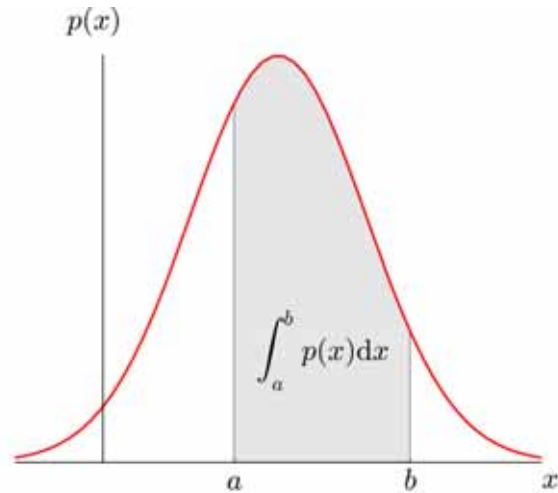
Name	$n$	$K$	$x$
Multinomial	-	-	$x \in \{0, 1, \dots, n\}^K, \sum_{k=1}^K x_k = n$
Multinoulli	1	-	$x \in \{0, 1\}^K, \sum_{k=1}^K x_k = 1$ (1-of- $K$ encoding)
Binomial	-	1	$x \in \{0, 1, \dots, n\}$
Bernoulli	1	1	$x \in \{0, 1\}$

Table 2.1 Summary of the multinomial and related distributions.

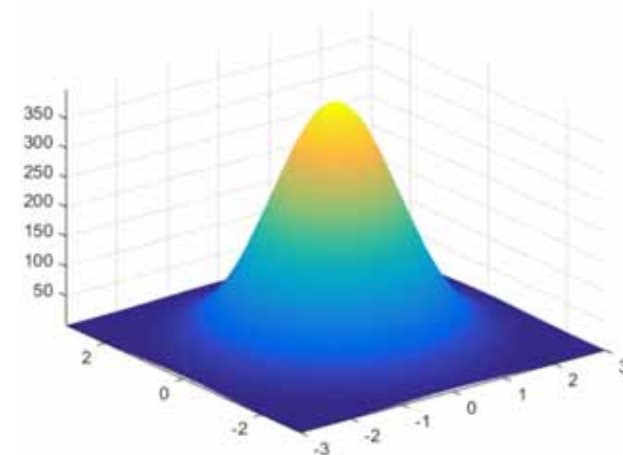
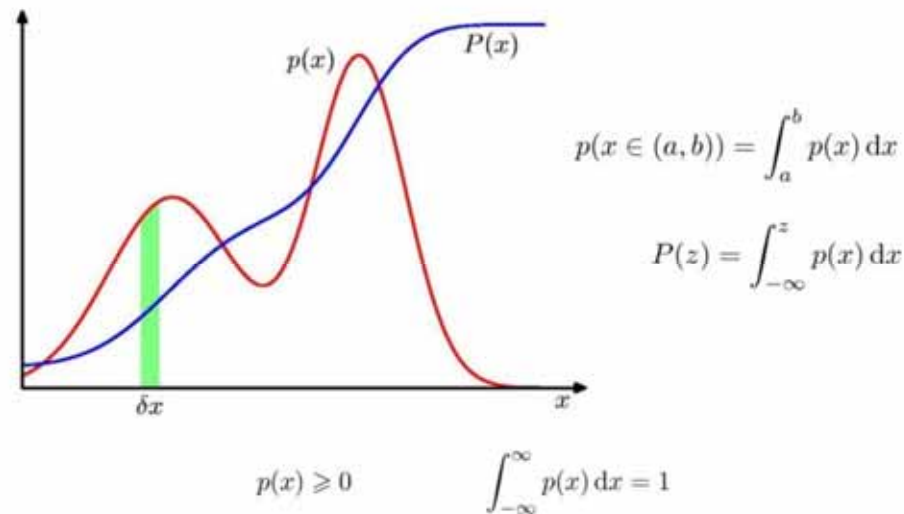


- Continuous: Probability density function (PDF) vs Cumulative Density Function (CDF):





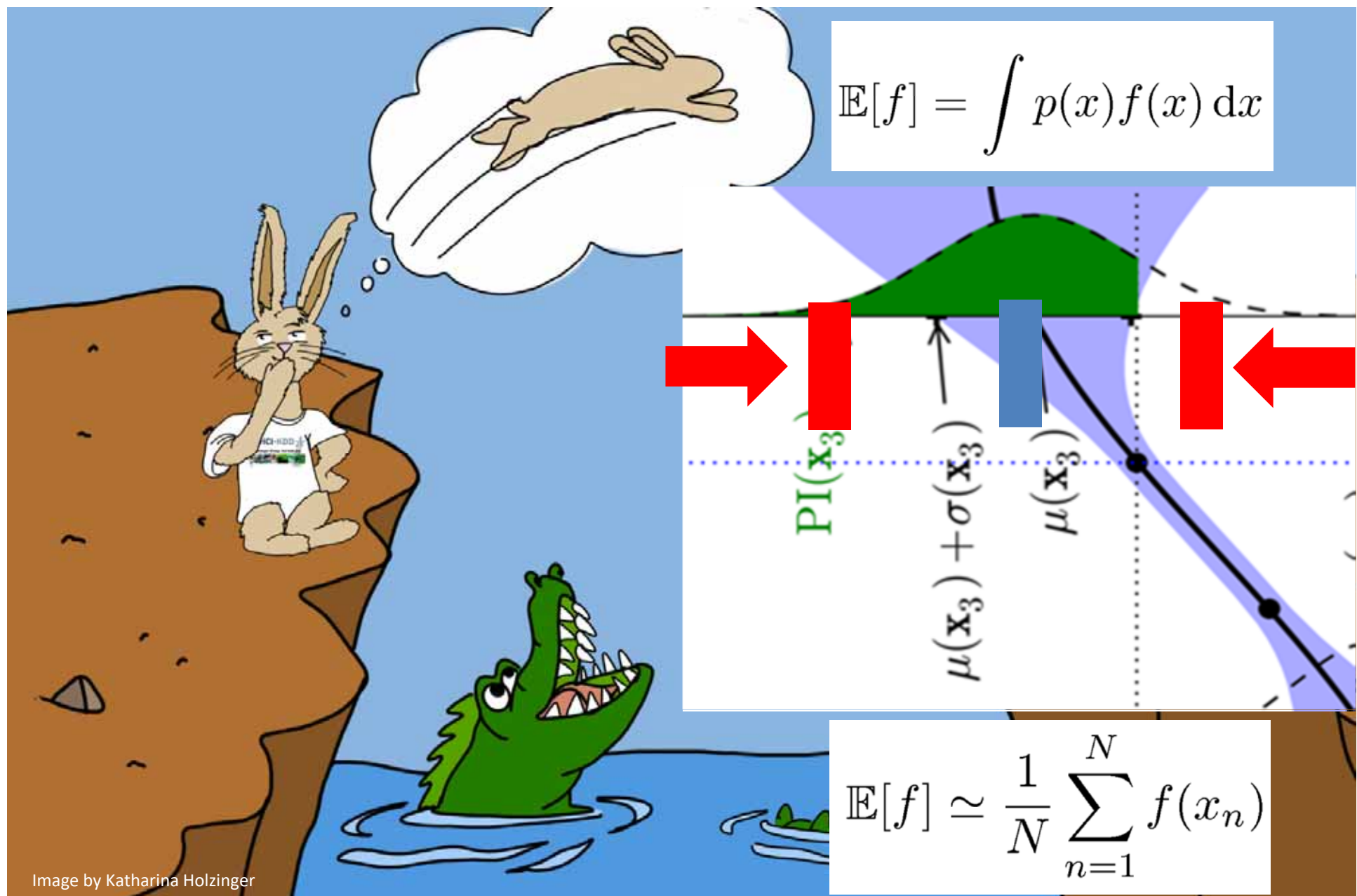
John Kruschke 2014. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, Amsterdam et al., Academic Press.



<https://brilliant.org/wiki/multivariate-normal-distribution>

# 02 Expectation and Expected Utility Theory

# Estimate Confidence Interval: Uncertainty matters !



For a single decision variable an agent can select  $D = d$  for any  $d \in \text{dom}(D)$ .

The expected utility of decision  $D = d$  is



<http://www.eoht.info/page/Oskar+Morgenstern>

$$E(U \mid d) = \sum_{x_1, \dots, x_n} P(x_1, \dots, x_n \mid d) U(x_1, \dots, x_n, d)$$

An optimal single decision is the decision  $D = d_{\max}$  whose expected utility is maximal:

$$d_{\max} = \arg \max_{d \in \text{dom}(D)} E(U \mid d)$$

John Von Neumann & Oskar Morgenstern 1944. Theory of games and economic behavior, Princeton university press.

# 03 Joint Probability Conditional Probability



We say that two events are **independent** if their joint probability equals the product of their individual probabilities,

$$p(A, B) = p(A)p(B).$$

In this case we use the notation  $A \perp B$ . Two random variables are independent if this is true for all values that the random variables can take.

By using the product, we see that two variables are independent iff

$$p(A)p(B) = p(A, B) = p(A|B)p(B),$$

or equivalently that

$$p(A|B) = p(A).$$

This means that knowing that  $B$  happened tells us nothing about the probability of  $A$  happening, and vice versa.

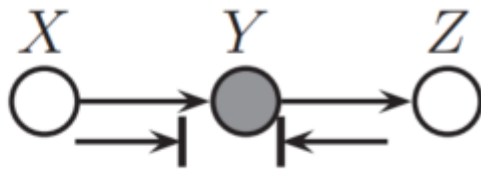
A generalization of independence is **conditional independence**, where we consider independence given that we know a third event  $C$  occurred,

$$p(A, B|C) = p(A|C)p(B|C),$$

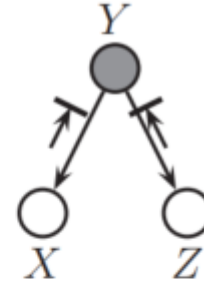
and in this case we use the notation  $A \perp B | C$ . Conditional independence is much weaker than marginal independence, and we often make use of it to model high-dimensional probability distributions.



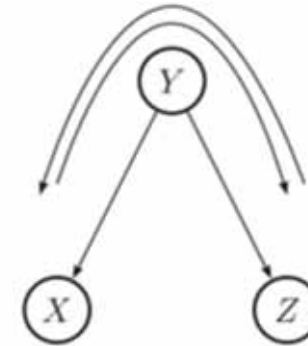
# When we condition on y, are x and z independent?



(a)



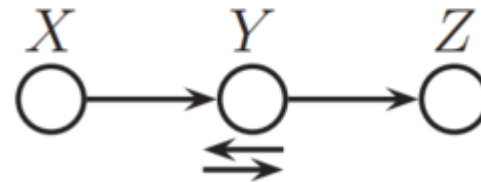
(b)



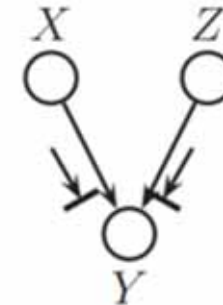
(c)



(d)



(e)



(f)

$$p(x, z|y) = \frac{p(x)p(y|x)p(z|y)}{p(y)} = \frac{p(x, y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

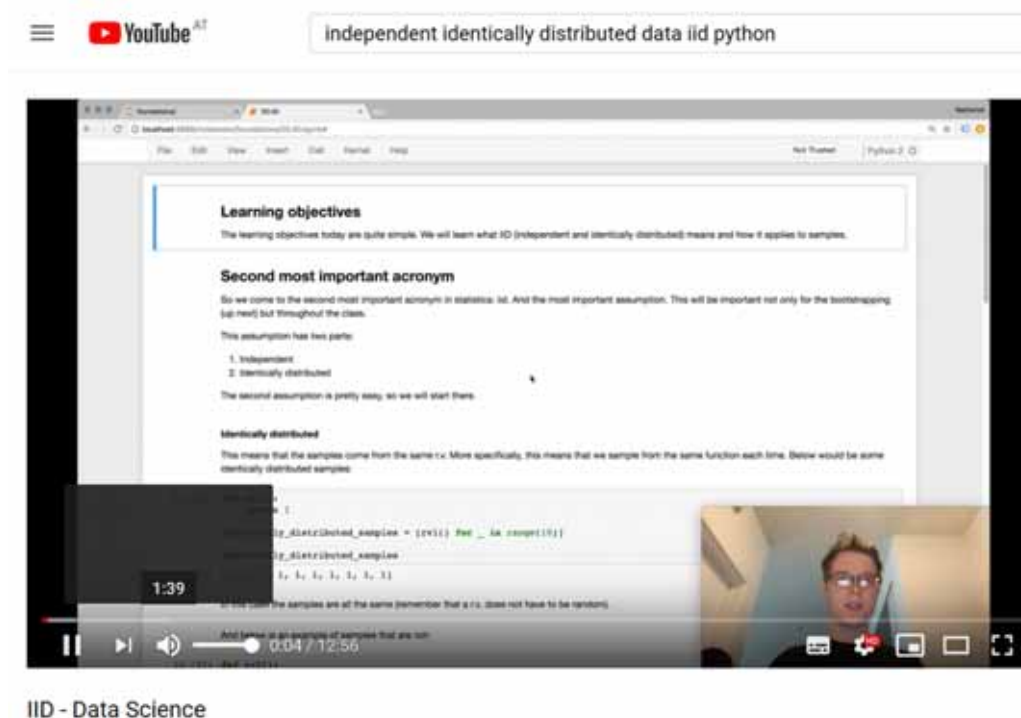
$$x \perp z|y \quad X \leftarrow Y \rightarrow Z \quad p(x, y, z) = p(y)p(x|y)p(z|y)$$

# 04 Independent Identically distributed (IID, i.i.d.)

**Definition 8.32** (Independent and Identically distributed). For a variable  $x$ , and a set of i.i.d. observations,  $x^1, \dots, x^N$ , conditioned on  $\theta$ , we assume there is no dependence between the observations

$$p(x^1, \dots, x^N | \theta) = \prod_{n=1}^N p(x^n | \theta) \quad (8.6.7)$$

Watch these examples: <https://www.youtube.com/watch?v=lhznecgCXeo>



# 05 Bayes and Laplace

# Bayes and Laplace on one single slide

$P(x)$  probability of  $x$   
 $P(x|\theta)$  conditional probability of  $x$  given  $\theta$   
 $P(x, \theta)$  joint probability of  $x$  and  $\theta$

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}|\theta)$  likelihood of  $\theta$   
 $P(\theta)$  prior probability of  $\theta$   
 $P(\theta|\mathcal{D})$  posterior of  $\theta$  given  $\mathcal{D}$

$$P(x, \theta) = P(x)P(\theta|x) = P(\theta)P(x|\theta)$$

**Bayes Rule:**

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

**Marginalization**

$$P(x) = \int P(x, \theta) d\theta$$

**Model Comparison:**

$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m) d\theta$$

**Prediction:**

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

$$P(x|\mathcal{D}, m) = \int P(x|\theta)P(\theta|\mathcal{D}, m)d\theta \quad (\text{for many models})$$



Note: This is probably not Bayes, but this image is heavily in use



Pierre Simon de Laplace (1749-1827)

Please refer to the excellent Lectures of Zoubin Ghahramani for profound details, <http://mlg.eng.cam.ac.uk/zoubin>

# 06 Information Theory & Entropy

- Information is a measure for the reduction of uncertainty
- If something is 100 % certain its uncertainty = 0
- Uncertainty is max. if all choices are equally probable (I.I.D)
- Uncertainty (as information) sums up for independent sources

# Entropy as measure for disorder



low entropy  
low complexity



medium entropy  
high complexity

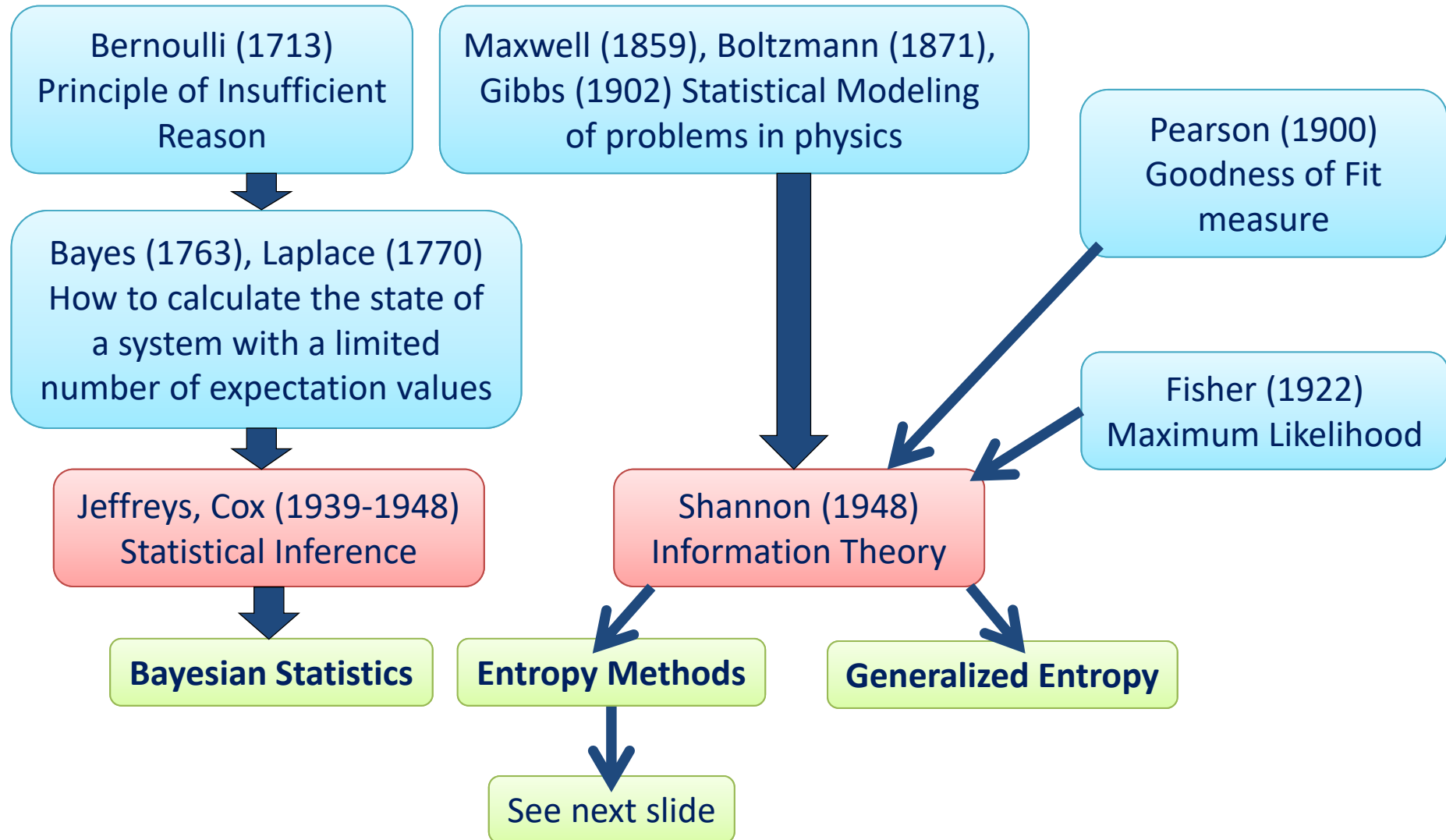


high entropy  
low complexity

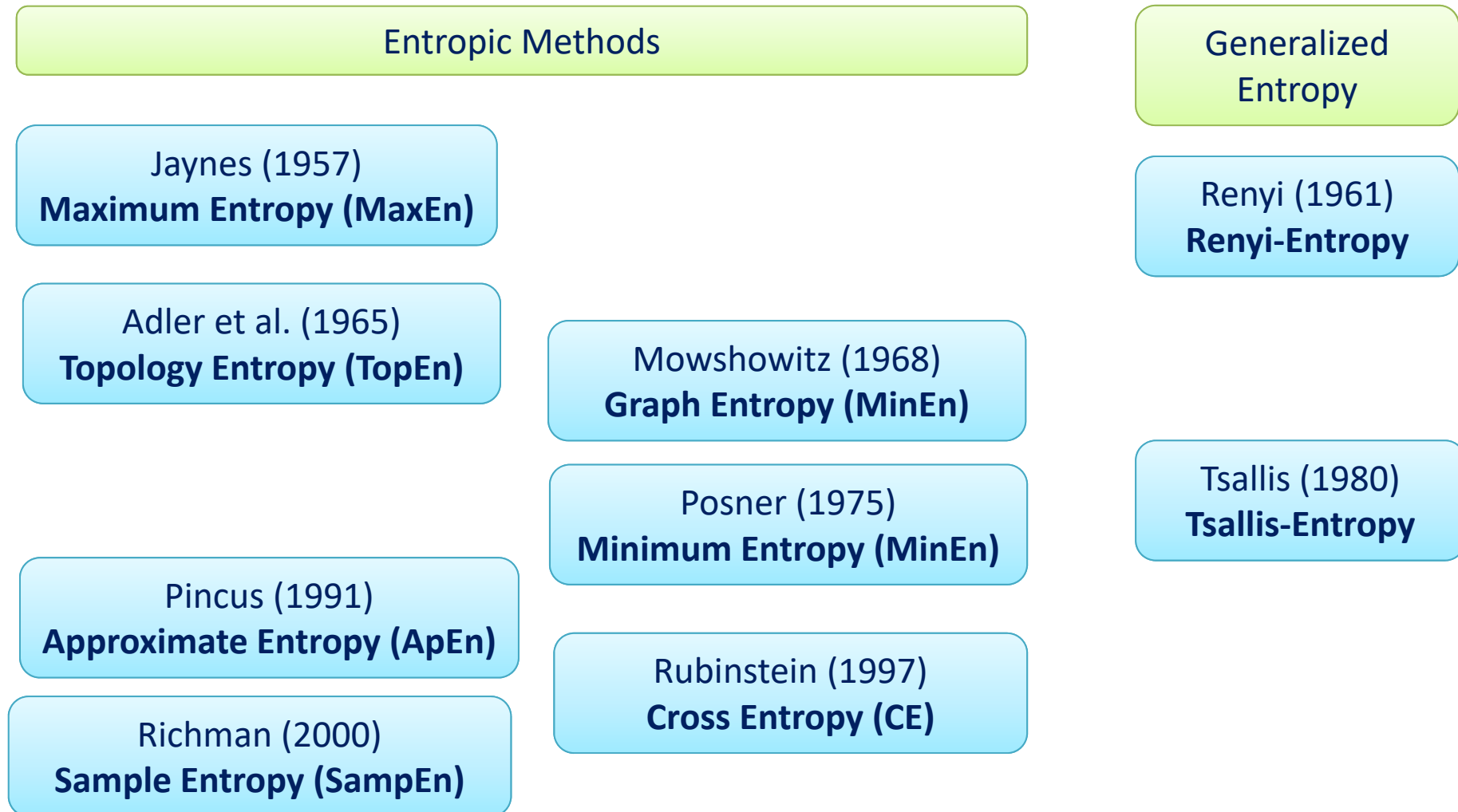
<http://www.scottaaronson.com>



# An overview on the History of Entropy

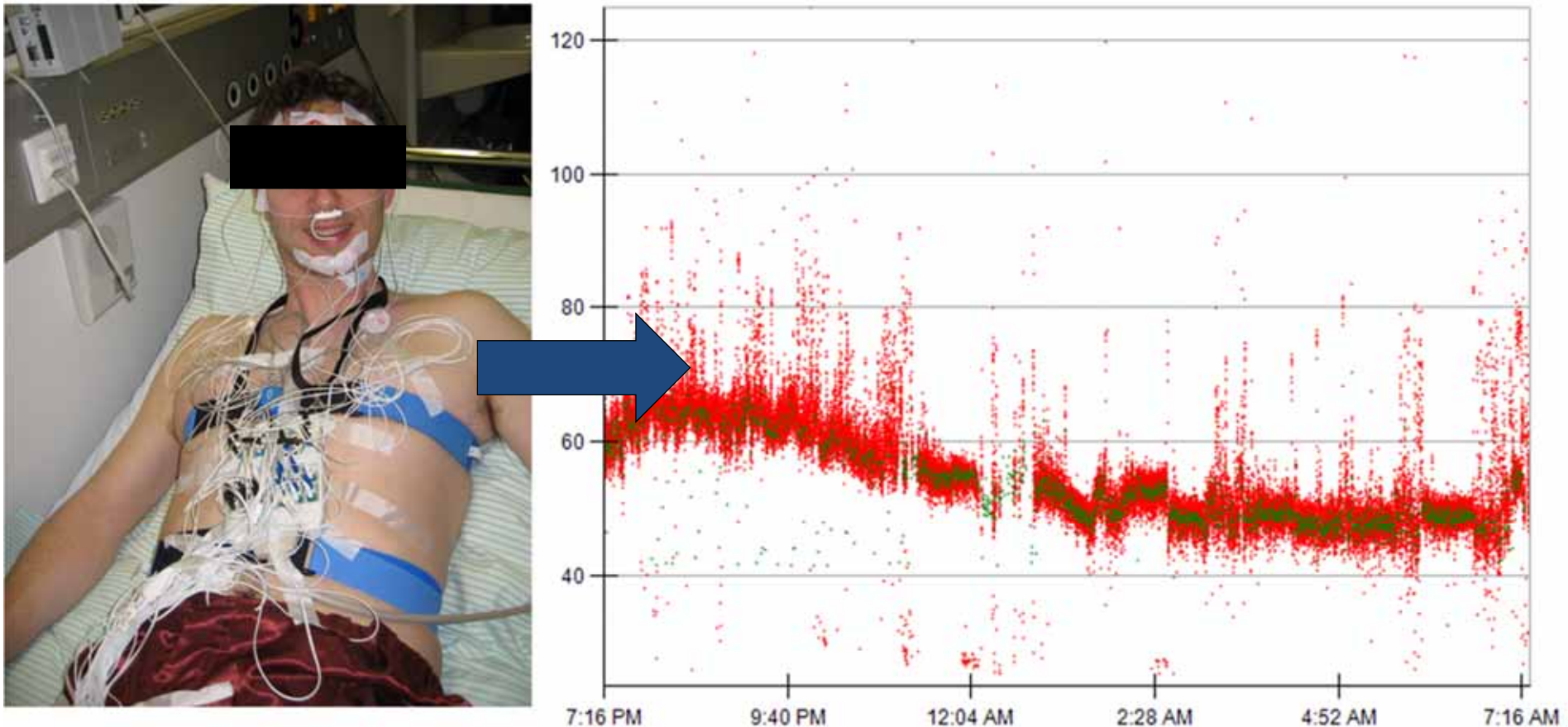


confer also with: Golan, A. (2008) Information and Entropy Econometric: A Review and Synthesis. *Foundations and Trends in Econometrics*, 2, 1-2, 1-145.



Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A. & Koslicki, D. 2014. On Entropy-Based Data Mining. In: Holzinger, A. & Jurisica, I. (eds.) Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 209-226.

## Example of the usefulness of ApEn (1/3)



Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H. & Fred, A. 2012. On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. *In: Huang, R., Ghorbani, A., Pasi, G., Yamaguchi, T., Yen, N. & Jin, B. (eds.) Active Media Technology, Lecture Notes in Computer Science, LNCS 7669. Berlin Heidelberg: Springer, pp. 646-657.*

EU Project EMERGE (2007-2010)

Let:  $\langle x_n \rangle = \{x_1, x_2, \dots, x_N\}$

$$\vec{X}_i = (x_i, x_{(i+1)}, \dots, x_{(i+m-1)})$$

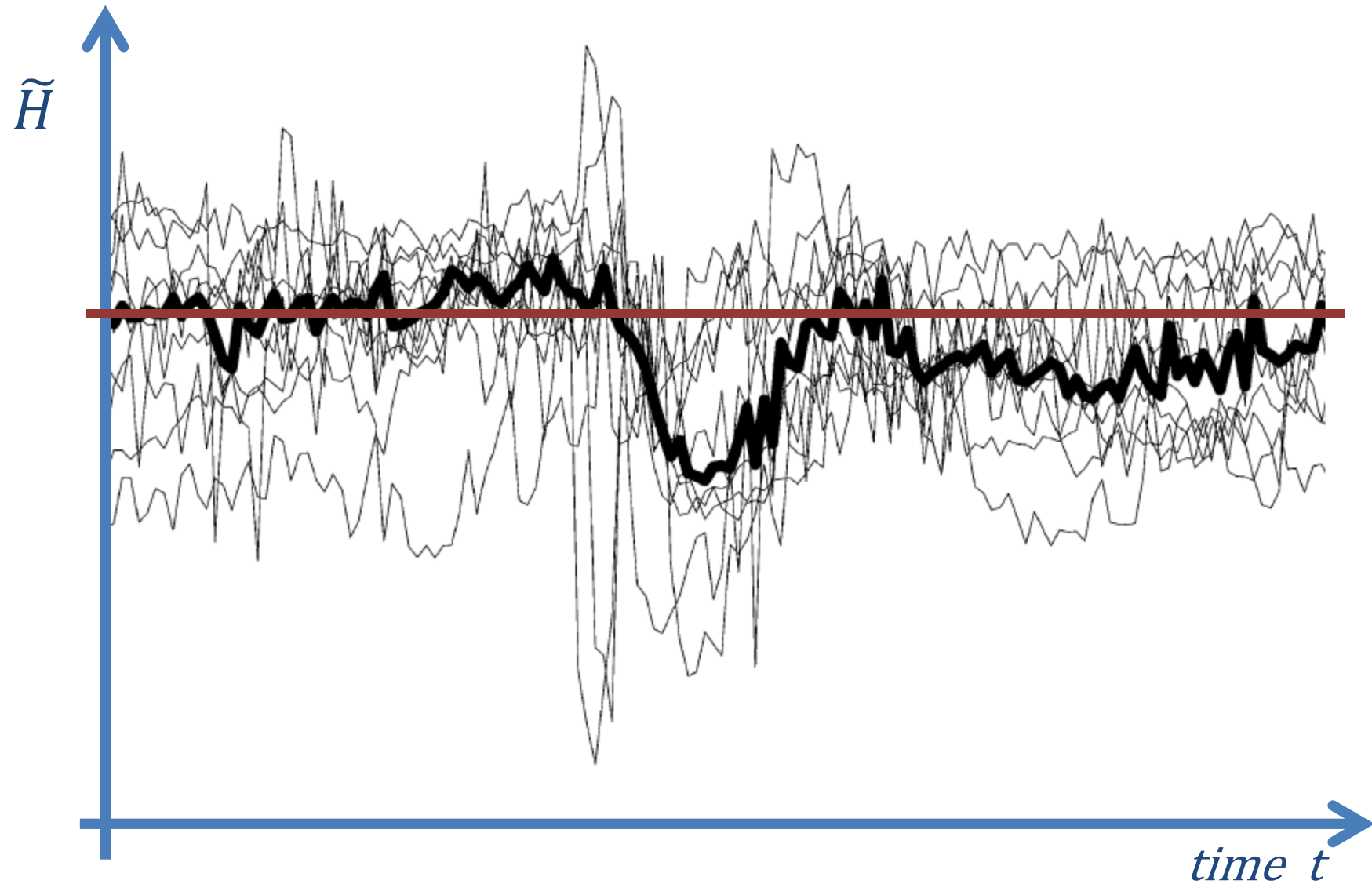
$$\|\vec{X}_i, \vec{X}_j\| = \max_{k=1,2,\dots,m} (|x_{(i+k-1)} - x_{(j+k-1)}|)$$

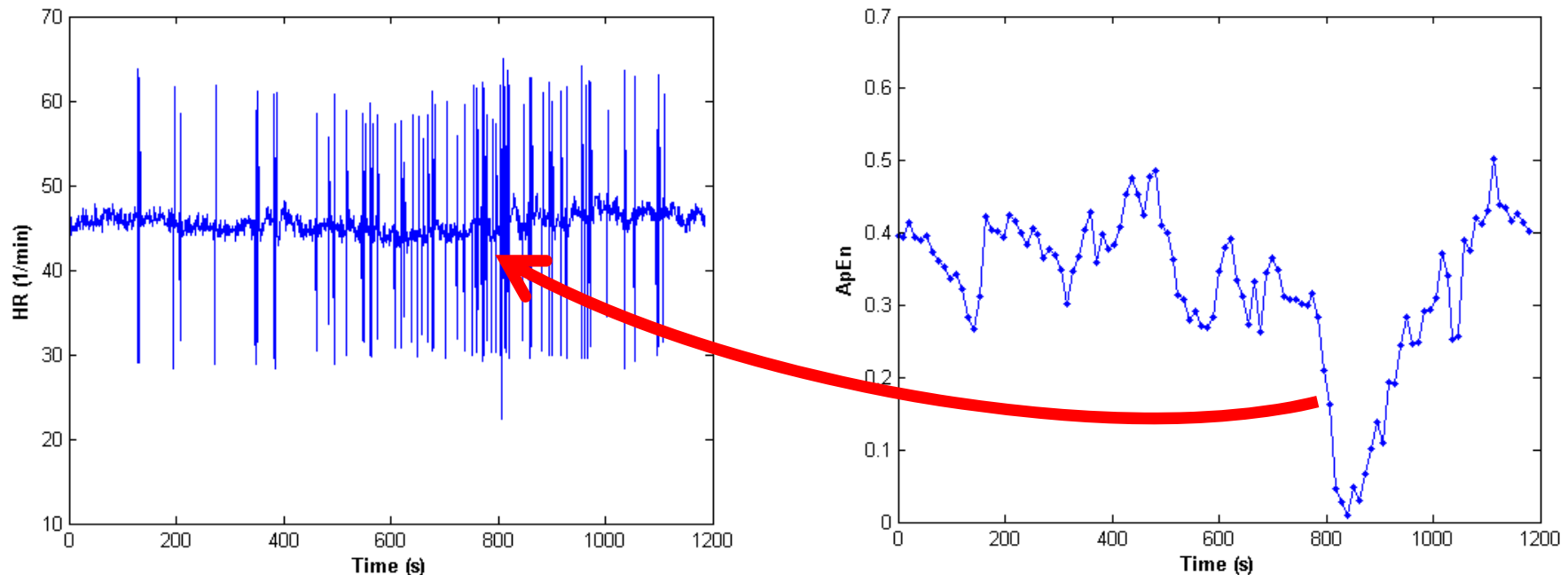
$$\tilde{H}(m, r) = \lim_{N \rightarrow \infty} [\phi^m(r) - \phi^{m+1}(r)]$$

$$C_r^m(i) = \frac{N^m(i)}{N - m + 1} \quad \phi^m(r) = \frac{1}{N - m + 1} \sum_{t=1}^{N-m+1} \ln C_r^m(i)$$

Pincus, S. M. (1991) Approximate Entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 88, 6, 2297-2301.

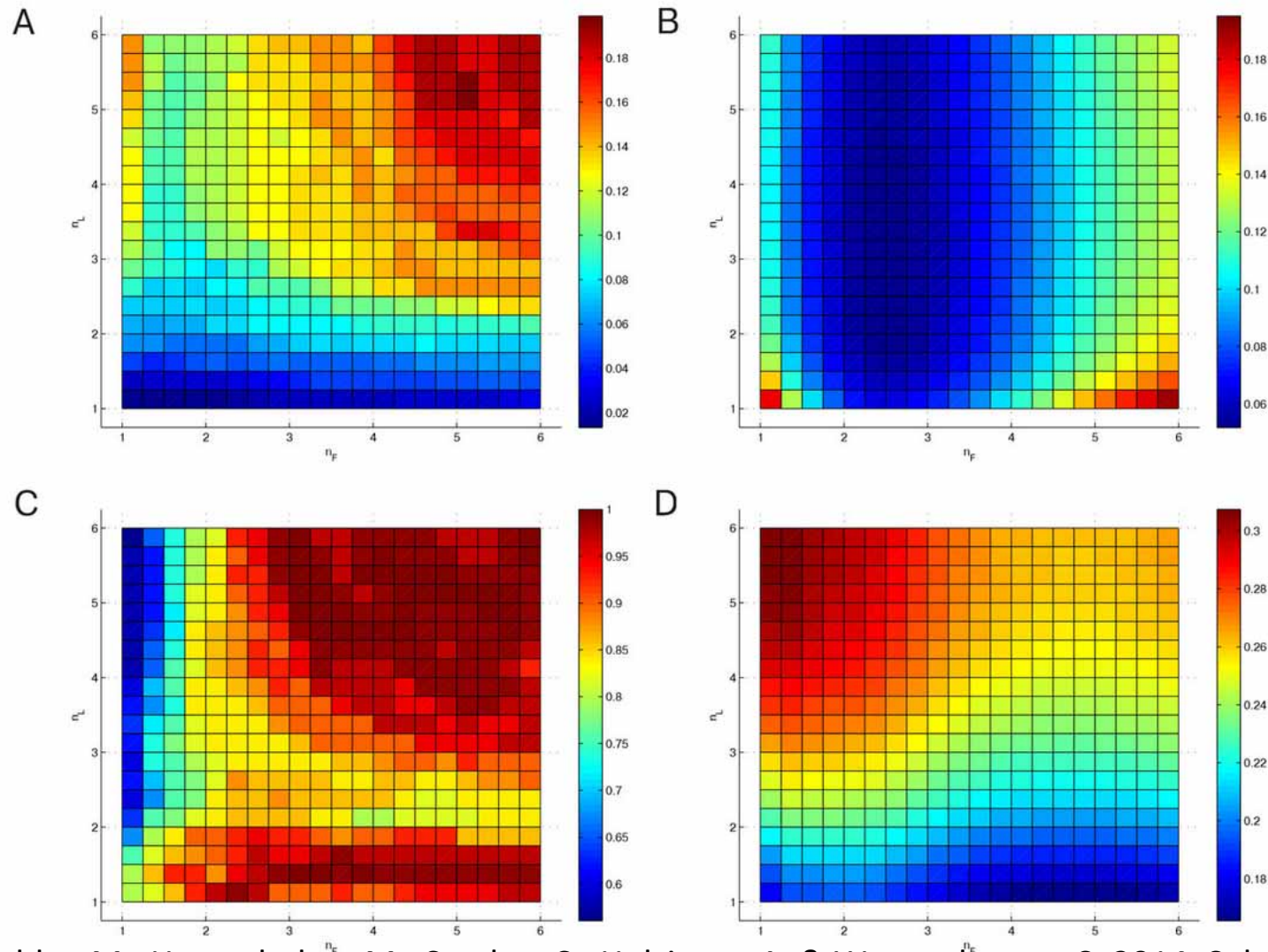
## Example: ApEn (2)





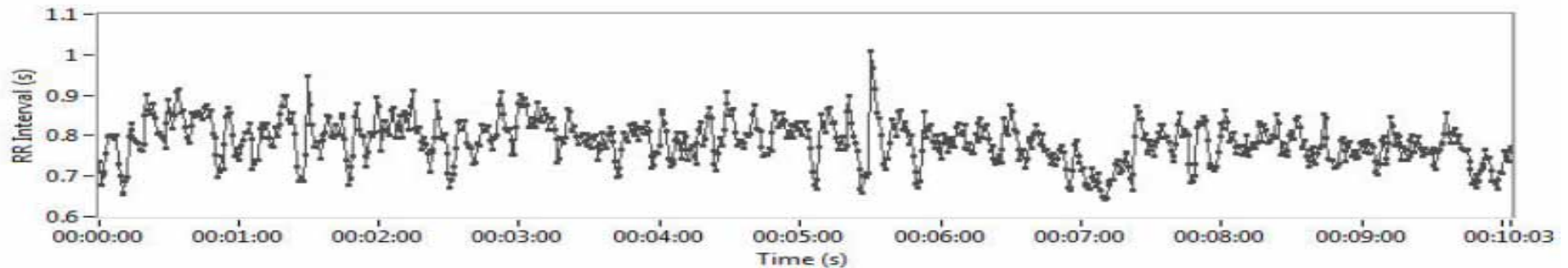
Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A. & Koslicki, D. 2014. On Entropy-Based Data Mining. In: Holzinger, A. & Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 209-226.





Mayer, C., Bachler, M., Hortenhuber, M., Stocker, C., Holzinger, A. & Wassertheurer, S. 2014. Selection of entropy-measure parameters for knowledge discovery in heart rate variability data. *BMC Bioinformatics*, 15, (Suppl 6), S2, doi:doi:10.1186/1471-2105-15-S6-S2.

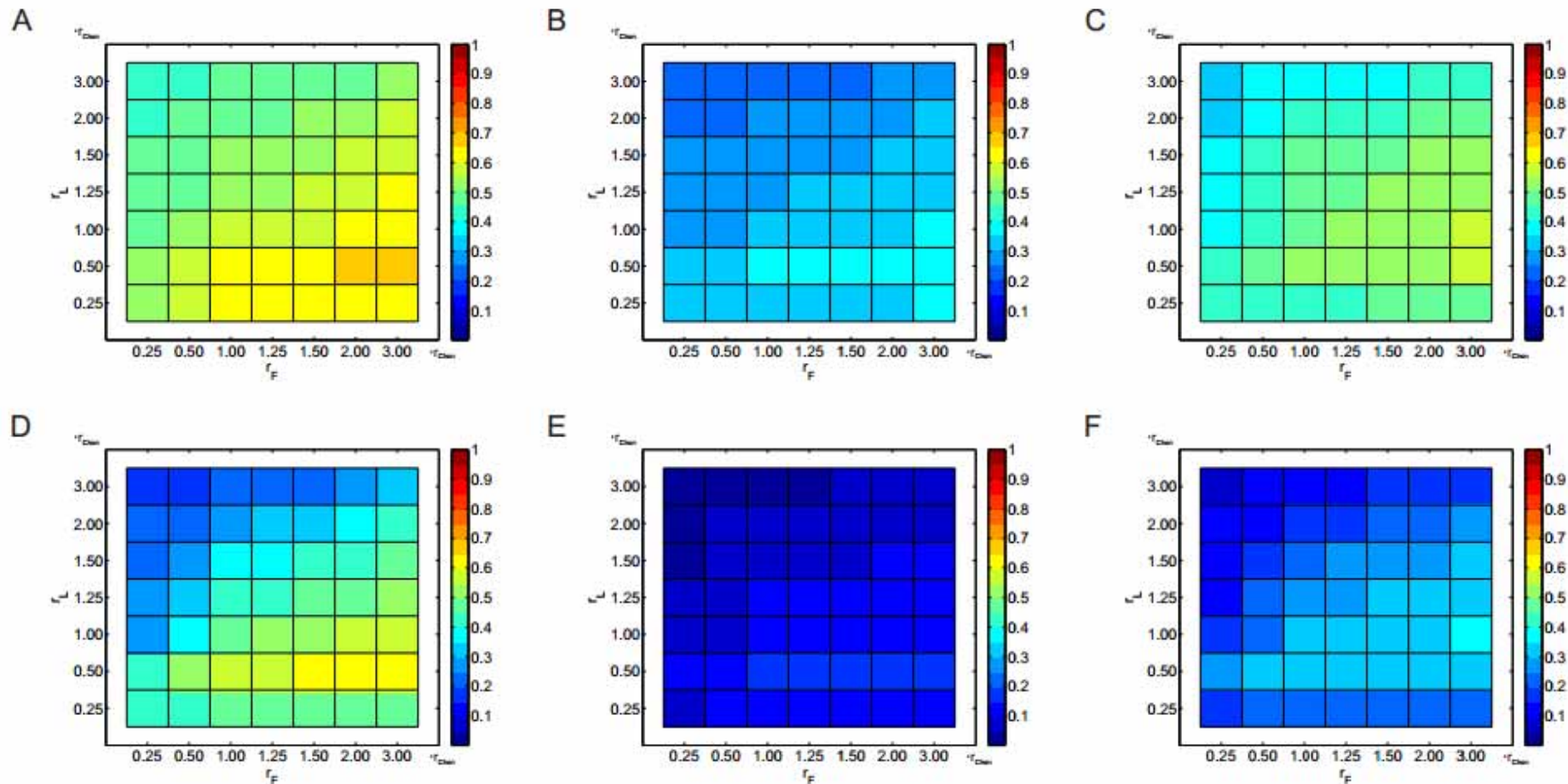
## Summary: Example Heart Rate Variability



- Heart Rate Variability (HRV) can be used as a marker of cardiovascular health status.
- Entropy measures represent a family of new methods to quantify the variability of the heart rate.
- Promising approach, due to ability to discover certain patterns and shifts in the "apparent ensemble amount of randomness" of stochastic processes,
- measure randomness and **predictability of processes.**

Mayer, C., Bachler, M., Holzinger, A., Stein, P. K. & Wassertheurer, S. 2016. The Effect of Threshold Values and Weighting Factors on the Association between Entropy Measures and Mortality after Myocardial Infarction in the Cardiac Arrhythmia Suppression Trial (CAST). *Entropy*, 18, (4), 129, doi::10.3390/e18040129.





Mayer, C., Bachler, M., Holzinger, A., Stein, P. K. & Wassertheurer, S. 2016. The Effect of Threshold Values and Weighting Factors on the Association between Entropy Measures and Mortality after Myocardial Infarction in the Cardiac Arrhythmia Suppression Trial (CAST). *Entropy*, 18, (4), 129, doi::10.3390/e18040129.

# 07 Cross-Entropy Kullback-Leibler Divergence

- Entropy:
  - Measure for the **uncertainty** of random variables
- Kullback-Leibler divergence:
  - **comparing two distributions**
- Mutual Information:
  - measuring the **correlation** of two random variables

## ON INFORMATION AND SUFFICIENCY

BY S. KULLBACK AND R. A. LEIBLER

*The George Washington University and Washington, D. C.*

**1. Introduction.** This note generalizes to the abstract case Shannon's definition of information [15], [16]. Wiener's information (p. 75 of [18]) is essentially the same as Shannon's although their motivation was different (cf. footnote 1, p. 95 of [16]) and Shannon apparently has investigated the concept more completely. R. A. Fisher's definition of information (intrinsic accuracy) is well known (p. 709 of [6]). However, his concept is quite different from that of Shannon and Wiener, and hence ours, although the two are not unrelated as is shown in paragraph 2.

R. A. Fisher, in his original introduction of the *criterion of sufficiency*, required "that the statistic chosen should summarize the whole of the relevant information supplied by the sample," (p. 316 of [5]). Halmos and Savage in a recent paper, one of the main results of which is a generalization of the well known Fisher-Neyman theorem on sufficient statistics to the abstract case, conclude, "We think that confusion has from time to time been thrown on the subject by . . . , and (c) the assumption that a sufficient statistic contains all the information in only the technical sense of 'information' as measured by variance," (p. 241 of [8]). It is shown in this note that the information in a sample as defined herein, that is, in the Shannon-Wiener sense cannot be increased by any statistical operations and is invariant (not decreased) if and only if sufficient statistics are employed. For a similar property of Fisher's information see p. 717 of [6], Doob [19].

We are also concerned with the statistical problem of discrimination ([3], [17]), by considering a measure of the "distance" or "divergence" between statistical populations ([1], [2], [13]) in terms of our measure of information. For the statistician two populations differ more or less according as to how difficult it is to discriminate between them with the best test [14]. The particular measure of divergence we use has been considered by Jeffreys ([10], [11]) in another connection. He is primarily concerned with its use in providing an invariant density of *a priori* probability. A special case of this divergence is Mahalanobis' generalized distance [13].



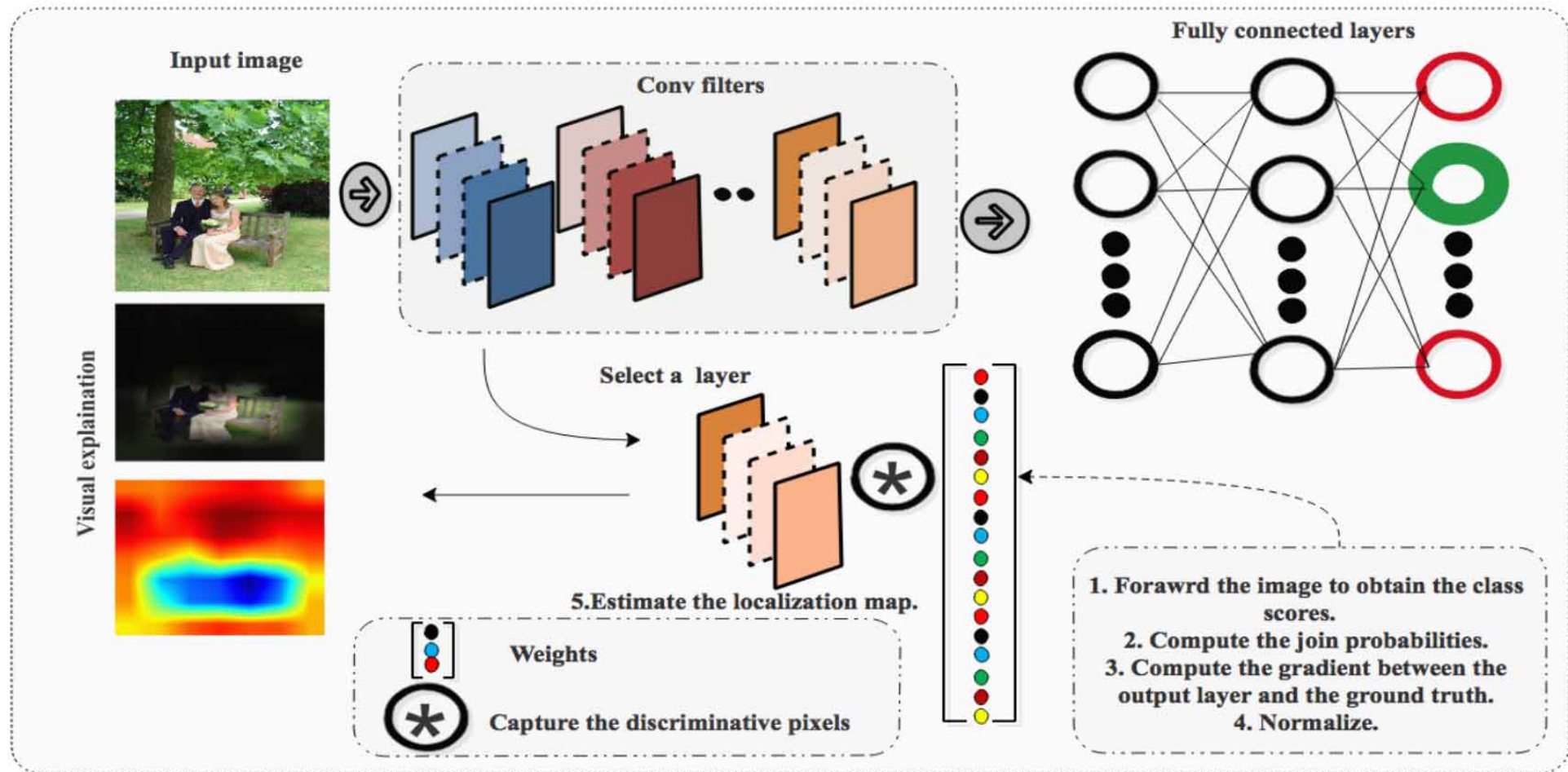
Solomon Kullback 1907-1994



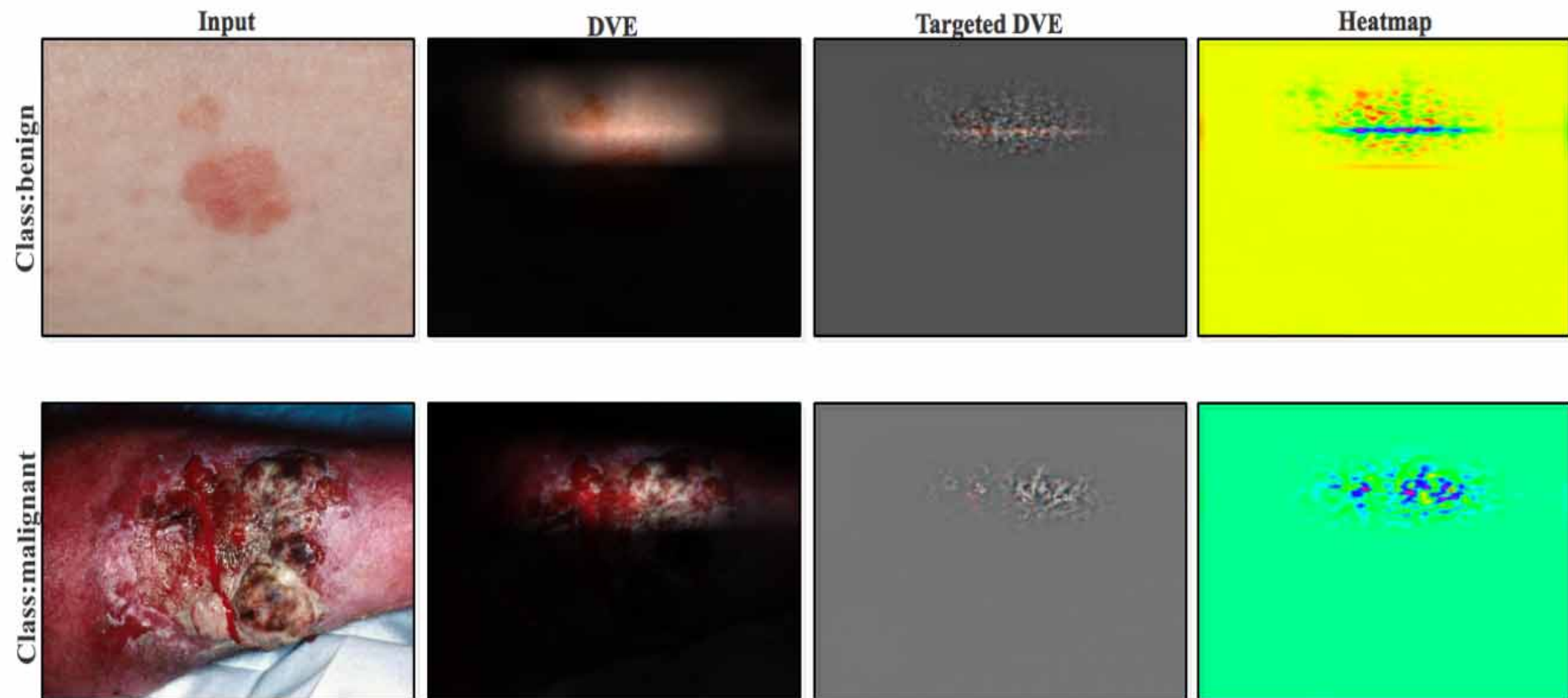
Richard Leibler 1914-2003

Kullback, S. & Leibler, R. A.  
1951. On information and  
sufficiency. The annals of  
mathematical statistics, 22, (1),  
79-86,  
[www.jstor.org/stable/2236703](http://www.jstor.org/stable/2236703)





Housam Khalifa Bashier Babiker & Randy Goebel 2017. Using KL-divergence to focus Deep Visual Explanation. arXiv preprint arXiv:1711.06431.



Housam Khalifa Bashier Babiker & Randy Goebel 2017. An Introduction to Deep Visual Explanation. arXiv preprint arXiv:1711.09482.

$$p_{ij} = \frac{(1 + \|k_i - k_j\|^2)^{-1}}{\sum_{u \neq v} (1 + \|k_u - k_v\|^2)^{-1}} \quad (1)$$

Here  $p_{ij}$  denotes the joint probabilities,  $k$  is the raw class scores before softmax,  $i$  indexes a neuron value and  $\sum_{u \neq v}$  combines all the values. For the ground truth we estimate the pairwise affinities with perplexity. We then compute the KL-divergence gradient i.e.  $\frac{\delta y'}{\delta y} \Rightarrow z$  derived here [6]. We also normalize the gradient to a zero mean and unit variance as follows:

$$\alpha = \frac{z - \mu}{\sigma z} \quad (2)$$

The obtained weights  $\alpha$  capture the relevant information in the feature maps acquired by the network. These weights are applied to every feature map  $x_i \in X$  as to identify the discriminative pixels which influence the final prediction output as follows:

$$E_{KL-divergence} = \sum_i \sum_j x_i * |\alpha_j| \quad (3)$$

Housam Khalifa Bashier Babiker & Randy Goebel 2017. Using KL-divergence to focus Deep Visual Explanation. arXiv preprint arXiv:1711.06431.

---

**Algorithm 1** Proposed approach

---

**Input:** image, ground truth  $y$

**Output:** Discriminative localization map  $\Rightarrow E_{KL-divergence}$

Apply a single forward-pass to estimate  $\Rightarrow y'$

Compute the joint probabilities for both  $y'$  and  $y$

Compute the gradient and normalize using (2)  $\Rightarrow \alpha$

initialize  $E_{KL-divergence}$  to zero

**for**  $i = 1$  **to**  $nFeatureMaps$  **do**

    Initialize temp to zero

**for**  $j = 1$  **to**  $sizeof \alpha$  **do**

$temp \leftarrow temp + (x_i * |\alpha_j|)$

**end for**

$E_{KL-divergence} \leftarrow E_{KL-divergence} + temp$

**end for**

---

Housam Khalifa Bashier Babiker & Randy Goebel 2017. Using KL-divergence to focus Deep Visual Explanation. arXiv preprint arXiv:1711.06431.



$$H[x] = - \sum_x p(x) \log_2 p(x)$$

Shannon, C. E. 1948. A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423.

Important quantity in

- coding theory
- statistical physics
- machine learning

$$H[\mathbf{y}|\mathbf{x}] = - \iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \, d\mathbf{x}$$

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}]$$

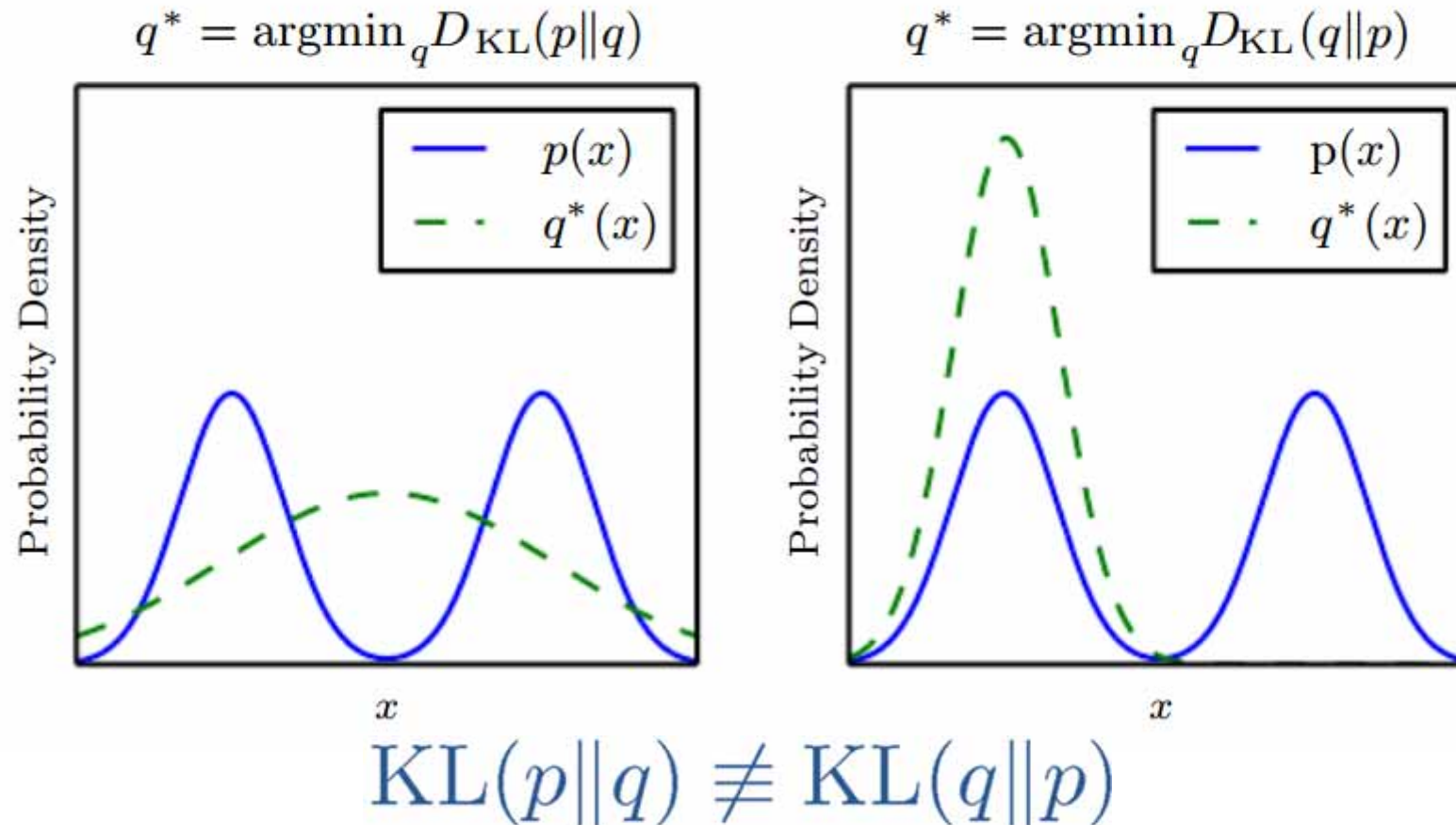
$$\begin{aligned}\text{KL}(p\|q) &= - \int p(\mathbf{x}) \ln q(\mathbf{x}) \, d\mathbf{x} - \left( - \int p(\mathbf{x}) \ln p(\mathbf{x}) \, d\mathbf{x} \right) \\ &= - \int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \, d\mathbf{x}\end{aligned}$$

$$\text{KL}(p\|q) \simeq \frac{1}{N} \sum_{n=1}^N \{ -\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n) \}$$

$$\text{KL}(p\|q) \geq 0$$

**KL-divergence is often used to measure the distance between two distributions**

## Note: KL is not symmetric!



Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep Learning, Cambridge (MA), MIT Press.

- ... are **robust** against noise;
- ... can be applied to **complex time series** with good replication;
- ... is **finite** for stochastic, noisy, composite processes;
- ... the values correspond directly to irregularities – good for detecting **anomalies**

# Mutual Information and Point Wise MI

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &\equiv \text{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= - \iint p(\mathbf{x}, \mathbf{y}) \ln \left( \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \end{aligned}$$

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]$$

- Measures how much reduction in uncertainty of  $X$  given the information about  $Y$
- Measures correlation between  $X$  and  $Y$
- Related to the “channel capacity” in the original Shannon information theory

Bishop, C. M. 2007. *Pattern Recognition and Machine Learning*, Heidelberg, Springer.



## Example: Disease-Disease Relationship

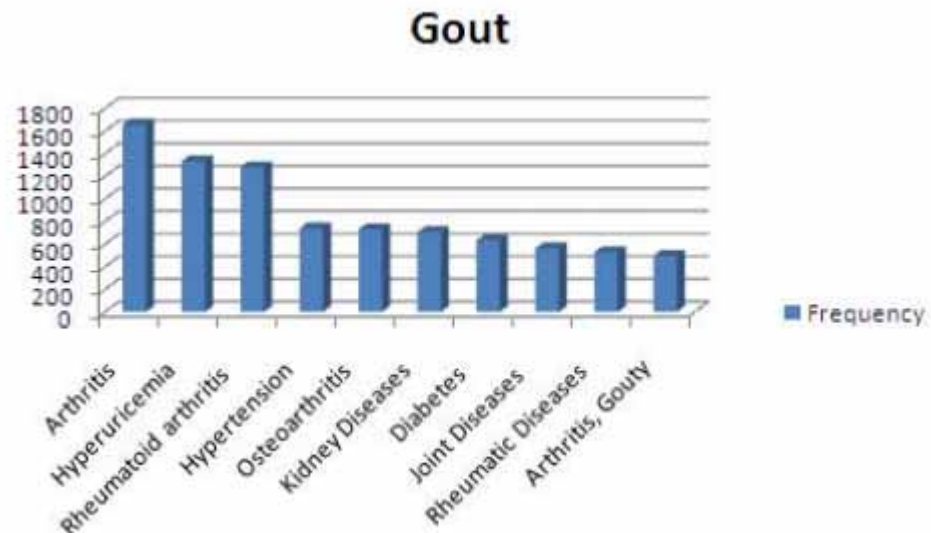
Let two words,  $w_i$  and  $w_j$ , have probabilities  $P(w_i)$  and  $P(w_j)$ . Then their mutual information  $PMI(w_i, w_j)$  is defined as:

$$PMI(w_i, w_j) = \log \left( \frac{P(w_i, w_j)}{P(w_i) P(w_j)} \right)$$

For  $w_i$  denoting *rheumatoid arthritis* and  $w_j$  representing *diffuse scleritis* the following simple calculation yields:

$$P(w_i) = \frac{94,834}{20,033,079}, \quad P(w_j) = \frac{74}{20,033,079}$$

$$P(w_i, w_j) = \frac{13}{94,834}, \quad PMI(w_i, w_j) = 7.7.$$



Holzinger, A., Simonic, K. M. & Yildirim, P. Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining and Knowledge Discovery to Assist Medical Decision Making. 36th Annual IEEE Computer Software and Applications Conference (COMPSAC), 16-20 July 2012 2012 Izmir. IEEE, 573-580, doi:10.1109/COMPSAC.2012.77.

$$SCP(x, y) = p(x|y) \cdot p(y|x) = \frac{p(x, y)}{p(y)} \cdot \frac{p(x, y)}{p(x)} = \frac{p(x, y)^2}{p(x) \cdot p(y)}$$

**Table 4** Comparison of FACTAs ranking of related concepts from the category Symptom for the query “rheumatoid arthritis” created by the methods co-occurrence frequency, PMI, and SCP

Frequency		PMI		SCP	
pain	5667	impaired body balance	7,8	swollen joints	0.002
Arthralgia	661	ASPIRIN INTOLERANCE	7,8	pain	0.001
fatigue	429	Epitrochlear lymphadenopathy	7,8	Arthralgia	0.001
diarrhea	301	swollen joints	7,4	fatigue	0.000
swollen joints	299	Joint tenderness	7	erythema	0.000
erythema	255	Occipital headache	6,2	splenomegaly	0.000
Back Pain	254	Neuromuscular excitation	6,2	Back Pain	0.000
headache	239	Restless sleep	5,8	polymyalgia	0.000
splenomegaly	228	joint crepitus	5,7	joint stiffness	0.000
Anesthesia	221	joint symptom	5,5	Joint tenderness	0.000
dyspnea	218	Painful feet	5,5	hip pain	0.000
weakness	210	feeling of malaise	5,5	metatarsalgia	0.000
nausea	199	Homan's sign	5,4	Skin Manifestations	0.000
Recovery of Function	193	Diffuse pain	5,2	neck pain	0.000
low back pain	167	Palmar erythema	5,2	Eye Manifestations	0.000
abdominal pain	141	Abnormal sensation	5,2	low back pain	0.000

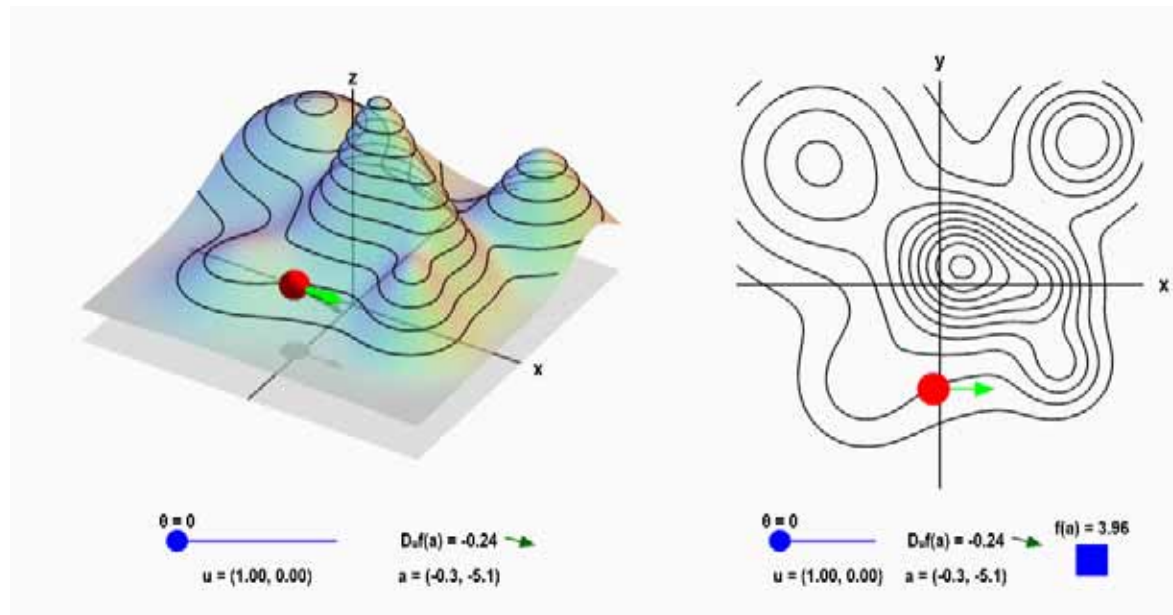
Holzinger, A., Yildirim, P., Geier, M. & Simonic, K.-M. 2013. Quality-Based Knowledge Discovery from Medical Text on the Web. In: Pasi, G., Bordogna, G. & Jain, L. C. (eds.) Quality Issues in the Management of Web Information, Intelligent Systems Reference Library, ISRL 50. Berlin Heidelberg: Springer, pp. 145-158, doi:10.1007/978-3-642-37688-7\_7.

- 1) Challenges include –omics data analysis, where KL divergence and related concepts could provide important **measures** for discovering biomarkers.
- 2) Hot topics are new entropy measures suitable for computations in the context of complex/uncertain data for ML algorithms.
- Inspiring is the abstract geometrical setting underlying ML main problems, e.g. Kernel functions can be completely understood in this perspective. Future work may include entropic concepts and geometrical settings.

- The case of higher order statistical structure in the data – nonlinear and hierarchical ?
- Outliers in the data – noise models?
- There are  $\frac{D(D+1)}{2}$  parameters in a multi-variate Gaussian model – what happens if  $D \gg ?$   
dimensionality reduction

# 08 Optimization

# Topographical maps, level curves, gradients



[https://mathinsight.org/applet/gradient\\_directional\\_derivative\\_mountain](https://mathinsight.org/applet/gradient_directional_derivative_mountain)





# Thank you!



# Appendix

# 00 Mathematical Notations

We denote random and fixed scalars by lower case, random and fixed vectors by bold lower case, and random and fixed matrices by bold upper case. Occasionally we use non-bold upper case to denote scalar random variables. Also, we use  $p()$  for both discrete and continuous random variables

$\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \dots$	"Calligraphic" font generally denotes sets or lists, e.g., data set $\mathcal{D} = x_1, \dots, x_n$
$x \in \mathcal{D}$	$x$ is an element of set $\mathcal{D}$
$x \notin \mathcal{D}$	$x$ is not an element of set $\mathcal{D}$
$\mathcal{D} \cup \mathcal{D}$	$x$ union of two sets, i.e., the set containing all elements of $\mathcal{D}$ and $\mathcal{D}$
$ \mathcal{D} $	cardinality of set $\mathcal{D}$ , i.e., the number of (possibly non-distinct) elements in it
$\max_x[\mathcal{D}]$	the maximum $x$ value in set $\mathcal{D}$
$\text{dom}(x)$	Domain of variable $x$
$x = x$	The variable $x$ is in the state $x$
$\text{dim}(x)$	For a discrete variable $x$ , this denotes the number of states $x$ can take
$x_{a:b}$	$x_a, x_{a+1}, \dots, x_b$
$\nless, \ngtr$	not less than; not greater than
$\neq$	not equal to
$\ll, \gg$	much less than; much greater than
$d/dx$	the derivative with respect to $x$
$\mathcal{M} \subset \mathcal{N}$	$\mathcal{M}$ is a subset of $\mathcal{N}$
$\mathcal{M} \supset \mathcal{N}$	$\mathcal{M}$ contains $\mathcal{N}$
$\mathcal{M} \cap \mathcal{N}$	intersection of $\mathcal{M}$ and $\mathcal{N}$
$\implies$	implies
$\iff$	equivalent to
$\exists$	there exists
$\forall$	for every

We use boldface lower-case to denote vectors, such as  $\vec{x}$ , and boldface upper-case to denote matrices, such as  $\vec{X}$ . We denote entries in a matrix by non-bold upper case letters, such as  $X_{ij}$ .

Vectors are assumed to be column vectors, unless noted otherwise. We use  $(x_1, \dots, x_D)$  to denote a column vector created by stacking  $D$  scalars. If we write  $\vec{X} = (\vec{x}_1, \dots, \vec{x}_n)$ , where the left hand side is a matrix, we mean to stack the  $\vec{x}_i$  along the columns, creating a matrix.

$$||\vec{x}|| = ||\vec{x}||_2 \quad \text{Euclidean or } \ell_2 \text{ norm} \quad \sqrt{\sum_{j=1}^d x_j^2}$$

$$||\vec{x}||_1 \quad \ell_1 \text{ norm} \quad \sum_{j=1}^d |x_j|$$

$$\vec{X}_{:,j} \quad \text{j'th column of matrix}$$

$$\vec{X}_{i,:} \quad \text{transpose of } i\text{'th row of matrix (a column vector)}$$

$$\vec{X}_{i,j} \quad \text{Element } (i, j) \text{ of matrix } \vec{X}$$

$X, Y$	Random variable
$P()$	Probability of a random event
$F()$	Cumulative distribution function(CDF), also called distribution function
$p(x)$	Probability mass function(PMF)
$f(x)$	probability density function(PDF)
$F(x, y)$	Joint CDF
$p(x, y)$	Joint PMF
$f(x, y)$	Joint PDF
$p(X Y)$	Conditional PMF, also called conditional probability
$f_{X Y}(x y)$	Conditional PDF
$X \perp Y$	X is independent of Y
$X \not\perp Y$	X is not independent of Y
$X \perp Y Z$	X is conditionally independent of Y given Z
$X \not\perp Y Z$	X is not conditionally independent of Y given Z
$X \sim p$	X is distributed according to distribution $p$
$\vec{\alpha}$	Parameters of a Beta or Dirichlet distribution
$\text{cov}[X]$	Covariance of X
$\mathbb{E}[X]$	Expected value of X
$\mathbb{E}_q[X]$	Expected value of X wrt distribution $q$
$\mathbb{H}(X)$ or $\mathbb{H}(p)$	Entropy of distribution $p(X)$
$\mathbb{I}(X; Y)$	Mutual information between X and Y
$\mathbb{KL}(p  q)$	KL divergence from distribution $p$ to $q$

$\ell(\vec{\theta})$	Log-likelihood function
$L(\theta, a)$	Loss function for taking action $a$ when true state of nature is $\theta$
$\lambda$	Precision (inverse variance) $\lambda = 1/\sigma^2$
$\Lambda$	Precision matrix $\Lambda = \Sigma^{-1}$
$\text{mode}[\vec{X}]$	Most probable value of $\vec{X}$
$\mu$	Mean of a scalar distribution
$\vec{\mu}$	Mean of a multivariate distribution
$\Phi$	cdf of standard normal
$\phi$	pdf of standard normal
$\vec{\pi}$	multinomial parameter vector, Stationary distribution of Markov chain
$\rho$	Correlation coefficient
$\text{sigm}(x)$	Sigmoid (logistic) function, $\frac{1}{1 + e^{-x}}$
$\sigma^2$	Variance
$\Sigma$	Covariance matrix



$\text{var}[x]$	Variance of $x$
$\nu$	Degrees of freedom parameter
$Z$	Normalization constant of a probability distribution
$\sim$	has the distribution, e.g., $p(x) \sim N(\mu, \sigma^2)$
$N(\mu, \sigma^2)$	multidimensional normal or Gaussian distribution with mean $\mu$ and variance $\sigma^2$
$O(h(x))$	big oh order of $h(x)$
$\Theta(h(x))$	big theta order of $h(x)$
$\Omega(h(x))$	big omega order of $h(x)$
$\sup_x f(x)$	the supremum value of $f(x)$ -the global maximum of $f(x)$ over all values of $x$
$p(x = tr)$	Probability of variable $x$ being in the state true
$p(x = fa)$	Probability of variable $x$ being in the state false
$p(x \cap y)$	Probability of $x$ and $y$
$p(x \cup y)$	Probability of $x$ or $y$
$p(x y)$	Probability of $x$ conditioned on $y$
$\langle f(x) \rangle_{g(x)}$	The average of the function $f(x)$ with respect to the distribution $p(x)$
$\sigma(x)$	The logistic sigmoid $\frac{1}{(1+\exp(-x))}$
$\text{erf}(x)$	The (Gaussian) error function



## Ghahramani

### Weather Prediction

Assume that the weather in London is independent and identically distributed across days. It can either rain (R) or be cloudy (C).

**Data:**  $\mathcal{D} = (R C C R R C R \dots)$

**Parameters:**  $\theta \stackrel{\text{def}}{=} \text{Probability of rain}$

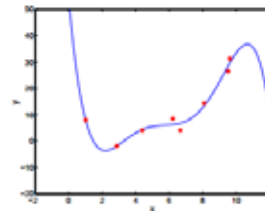
$$\begin{aligned} P(R|\theta) &= \theta \\ P(C|\theta) &= 1 - \theta \end{aligned}$$

**Goal:** To infer  $\theta$  from the data and predict future outcomes  $P(R|\mathcal{D})$ .

### Polynomial Regression

**Data:**  $\mathcal{D} = \{(x^{(n)}, y^{(n)})\}$  for  $n = 1, \dots, N$

$$\begin{aligned} x^{(n)} &\in \mathbb{R} \\ y^{(n)} &\in \mathbb{R} \end{aligned}$$



**Parameters:**  $\theta = (a_0, \dots, a_m, \sigma)$

**Model:**

$$y^{(n)} = a_0 + a_1 x^{(n)} + a_2 x^{(n)2} \dots + a_m x^{(n)m} + \epsilon$$

where

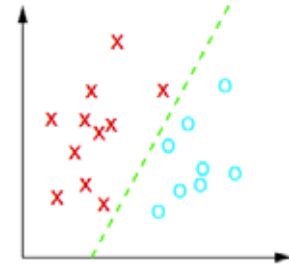
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

**Goal:** To infer  $\theta$  from the data and to predict future outputs  $P(y|\mathcal{D}, x, m)$

### Linear Classification

**Data:**  $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}$  for  $n = 1, \dots, N$   
data points

$$\begin{aligned} \mathbf{x}^{(n)} &\in \mathbb{R}^D \\ y^{(n)} &\in \{+1, -1\} \end{aligned}$$



**Parameters:**  $\theta \in \mathbb{R}^{D+1}$

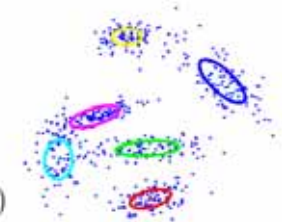
$$P(y^{(n)} = +1|\theta, \mathbf{x}^{(n)}) = \begin{cases} 1 & \text{if } \sum_{d=1}^D \theta_d x_d^{(n)} + \theta_0 \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

**Goal:** To infer  $\theta$  from the data and to predict future labels  $P(y|\mathcal{D}, \mathbf{x})$

### Clustering with Gaussian Mixtures (Density Estimation)

**Data:**  $\mathcal{D} = \{\mathbf{x}^{(n)}\}$  for  $n = 1, \dots, N$

$$\mathbf{x}^{(n)} \in \mathbb{R}^D$$



**Parameters:**  $\theta = ((\mu^{(1)}, \Sigma^{(1)}), \dots, (\mu^{(m)}, \Sigma^{(m)}), \pi)$

**Model:**

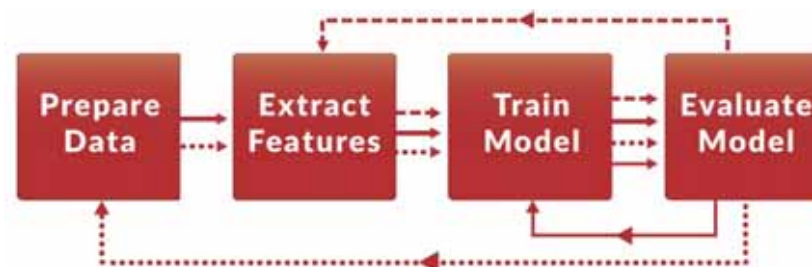
$$\mathbf{x}^{(n)} \sim \sum_{i=1}^m \pi_i p_i(\mathbf{x}^{(n)})$$

where

$$p_i(\mathbf{x}^{(n)}) = \mathcal{N}(\mu^{(i)}, \Sigma^{(i)})$$

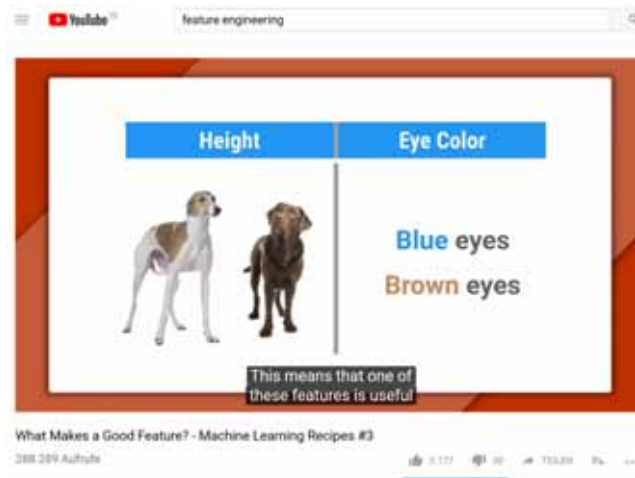
**Goal:** To infer  $\theta$  from the data and predict the density  $p(\mathbf{x}|\mathcal{D}, m)$

- 1) Linear algebra,
- 2) probability/statistics, and  $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$
- 3) optimization
- typical data organization is in arrays (matrices),
- rows represent the samples (data items)
- columns represent attributes (features, representations, covariates), ML = “feature engineering”, new trend “automatic learning”
- Simplest: each training input  $x_i$  is a d-dimensional vector of numbers, but  $x_i$  could be a complex object (image, sentence, graph, time series, molecular shape, etc.



See a typical ML-pipeline here:

<https://www.datasciencecentral.com/profiles/blogs/data-version-control-iterative-machine-learning>



<https://www.youtube.com/watch?v=N9fDIAfICMY>

**Definition 1** (*Relevant to the target*). A feature  $x_i$  is *relevant to a target concept  $c$*  if there exists a pair of examples  $A$  and  $B$  in the instance space such that  $A$  and  $B$  differ only in their assignment to  $x_i$  and  $c(A) \neq c(B)$ .

**Definition 3** (*Weakly relevant to the sample/distribution*). A feature  $x_i$  is *weakly relevant to sample  $S$*  (or to target  $c$  and distribution  $D$ ) if it is possible to remove a subset of the features so that  $x_i$  becomes strongly relevant.

Avrim L. Blum & Pat Langley 1997. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97, (1), 245-271, doi:10.1016/S0004-3702(97)00063-5.

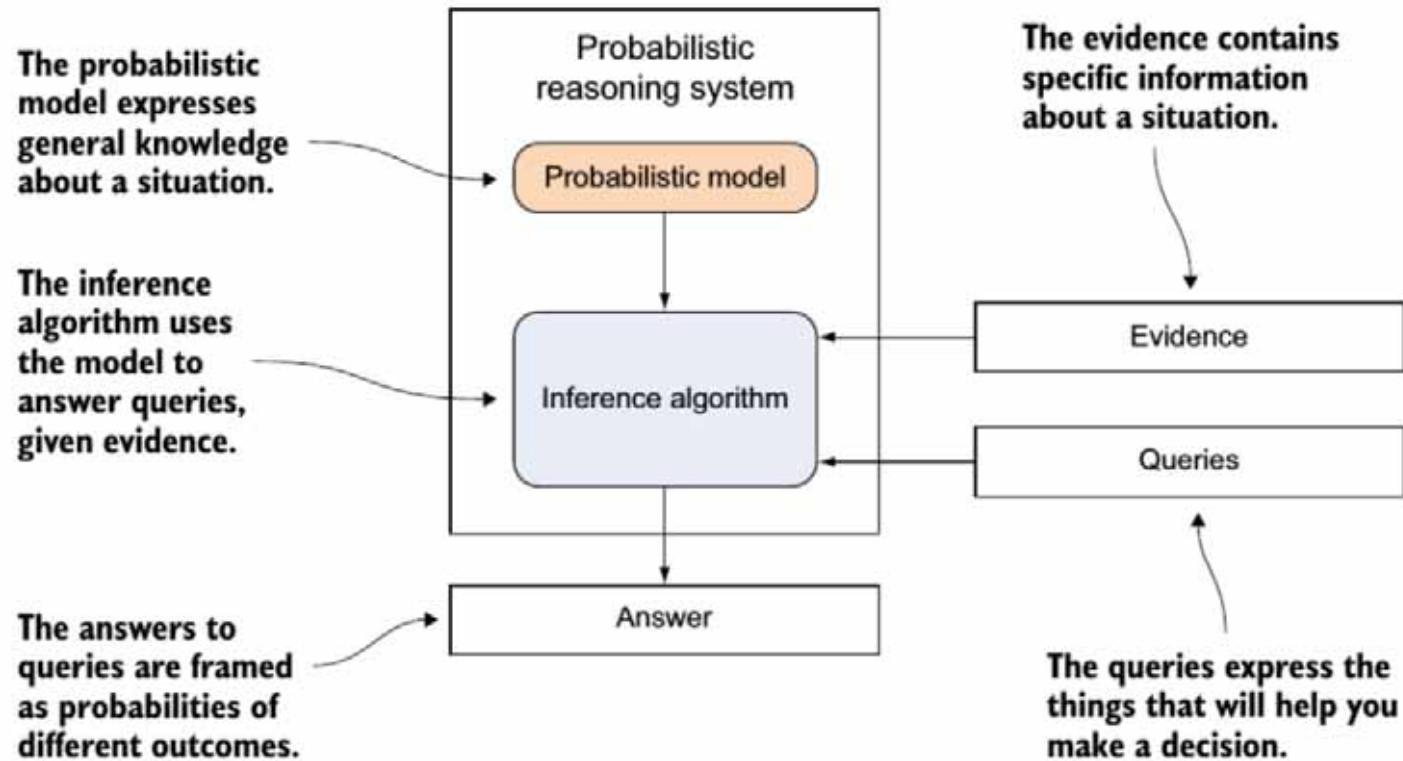
Thus the high-level features found in primates are not expected to occur also in simpler animals, such as the fish or frog. Across species, therefore, we find an enormous spectrum of features, especially if we include those specialized trigger patterns or “innate releasing mechanisms” reported by the ethologists (Thorpe, 1963; Tinbergen, 1951). Given this vast collection, it might seem unlikely that one could abstract away some principles that define “what makes a good feature?” However, here we attempt to do just that.

Our guiding hypothesis is that “seeing” is the inference of world properties from image elements—i.e. the various patterns of intensities on the retina. A “feature” is typically viewed as a measurement of image structure, at the level for example of Marr’s primal sketch (Marr, 1982). Clearly, many different kinds of measurements or “features” are possible. Intuitively, however, those most often sought after will point directly and reliably to a unique, *meaningful* event in the world. But the criterion that a feature be meaningful implies that the perceiver has some goal or context in mind. For example, for a baby gull the significance of a red spot in the image depends on whether it is seen in the context of a traffic light or as coloration on the beak of an adult gull (Figure 1). In the context of a beak, its salience is sufficient to trigger a feeding response. Somehow the gull is primed to immediately make the necessary inference. Hence we propose that “what makes a good feature” should include the property of having a ready explanation for its appearance (MacKay, 1978; MacKay, 1985).

Allan Jepson & Whitman Richards 1992. What makes a good feature. *Spatial vision in humans and robots*, 89-126.

# Practical Example: Probabilistic Reasoning System

Remark: compare with early Expert Systems (see module 3)



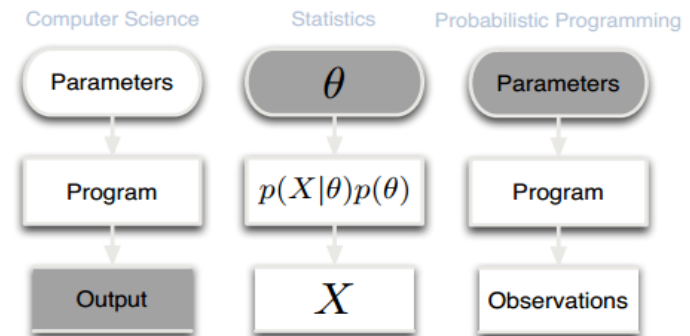
Avi Pfeffer 2016. Practical probabilistic programming, Shelter Island (NY), Manning Publications.

## Probabilistic Programming

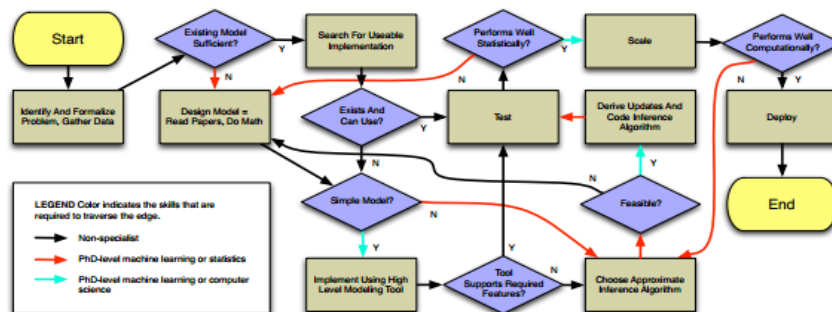
Frank Wood  
fwood@robots.ox.ac.uk



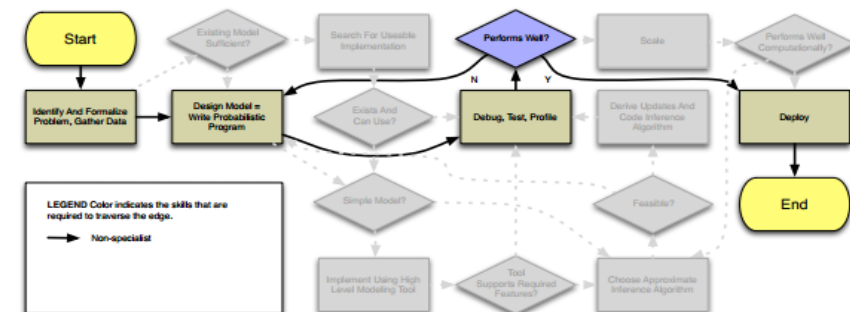
## What Is It?



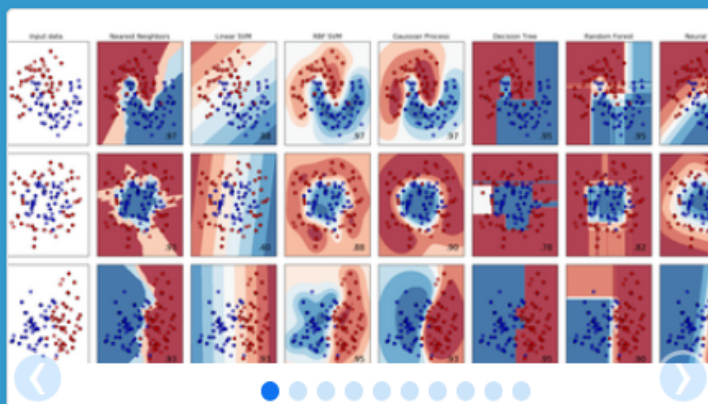
## The Way Machine Learning Is



## The Way Machine Learning Will Be







# scikit-learn

Machine Learning in Python

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

## Classification

Identifying to which category an object belongs to.

**Applications:** Spam detection, Image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, ... — Examples

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.  
**Algorithms:** SVR, ridge regression, Lasso, ... — Examples

## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes  
**Algorithms:** k-Means, spectral clustering, mean-shift, ... — Examples

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

**Algorithms:** PCA, feature selection, non-negative matrix factorization. — Examples

## Model selection

Comparing, validating and choosing parameters and models.

**Goal:** Improved accuracy via parameter tuning

**Modules:** grid search, cross validation, metrics. — Examples

## Preprocessing

Feature extraction and normalization.

**Application:** Transforming input data such as text for use with machine learning algorithms.

**Modules:** preprocessing, feature extraction. — Examples

