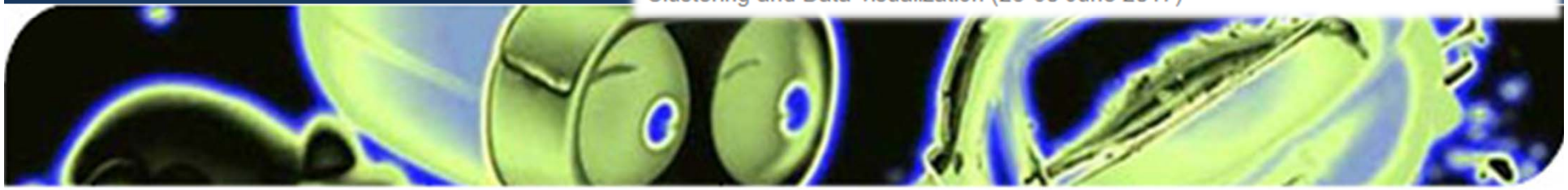
Alfredo Vellido

The eye of the beholder :

Visualization and interpretability in practical applications

WSOM+ 2017

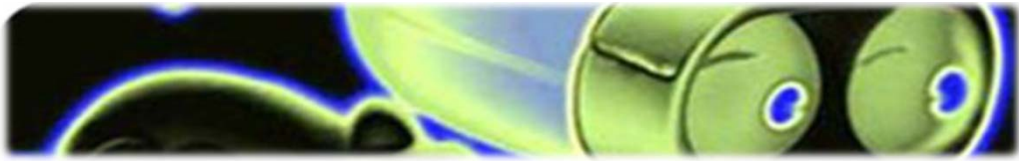
12th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (28-30 June 2017)



Soft Computing Research Group

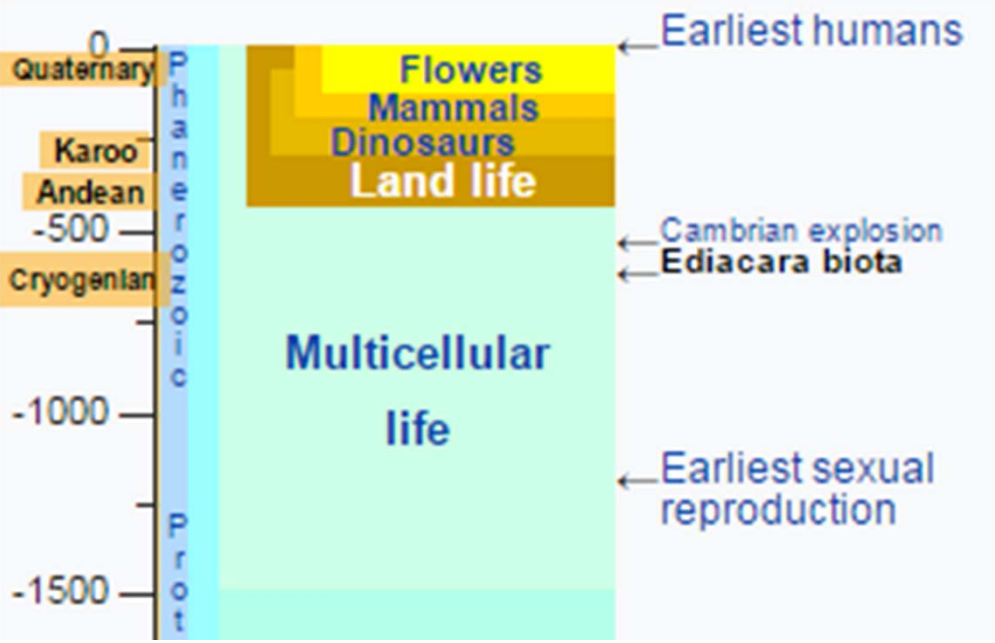
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

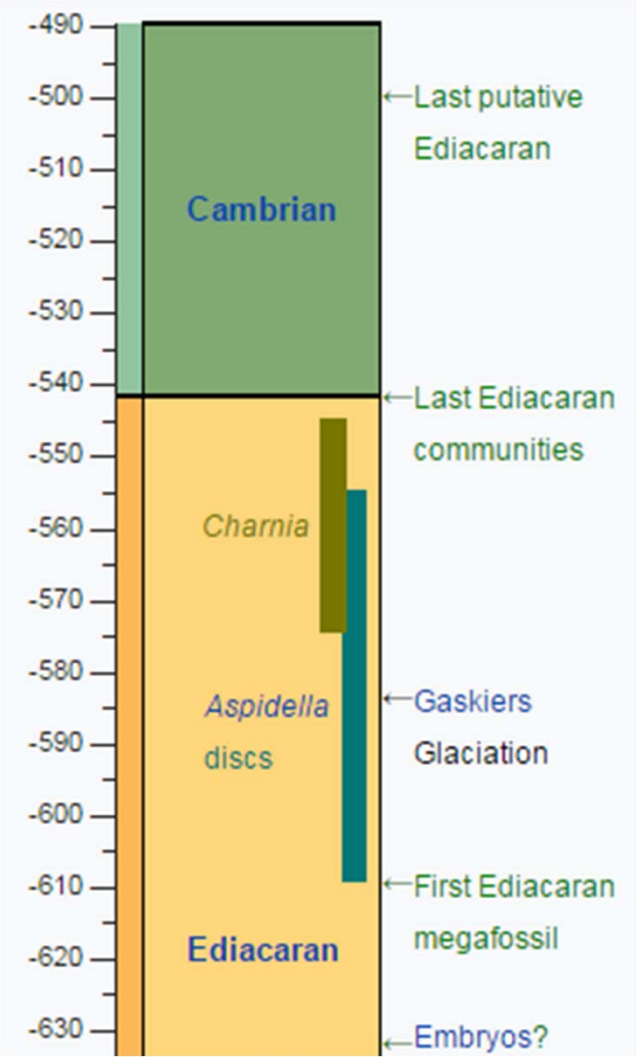
The eye of the beholder



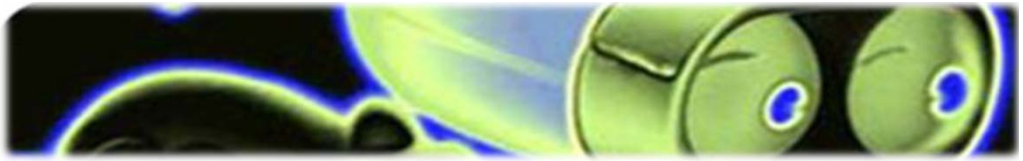
Context as (pre)history

Enter the *ediacaran*





The eye of the beholder



Context as (pre)history

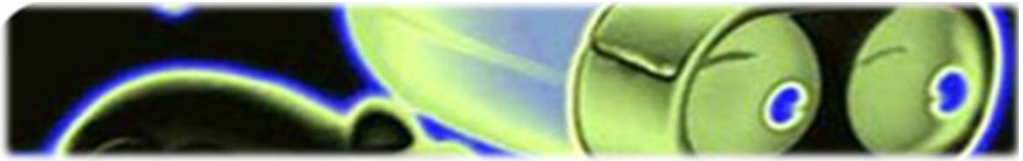
Enter the *ediacaran*

Meet *Kimberella*

Bilaterian: beyond jellyfish-like radial structures. Bilateral structure as a driver towards more complex nervous systems.

Early eyes?: some early evidence of photoreceptors associated to nervous systems. Still unresolved: did the visual system developed independently from the neuromotor control system?

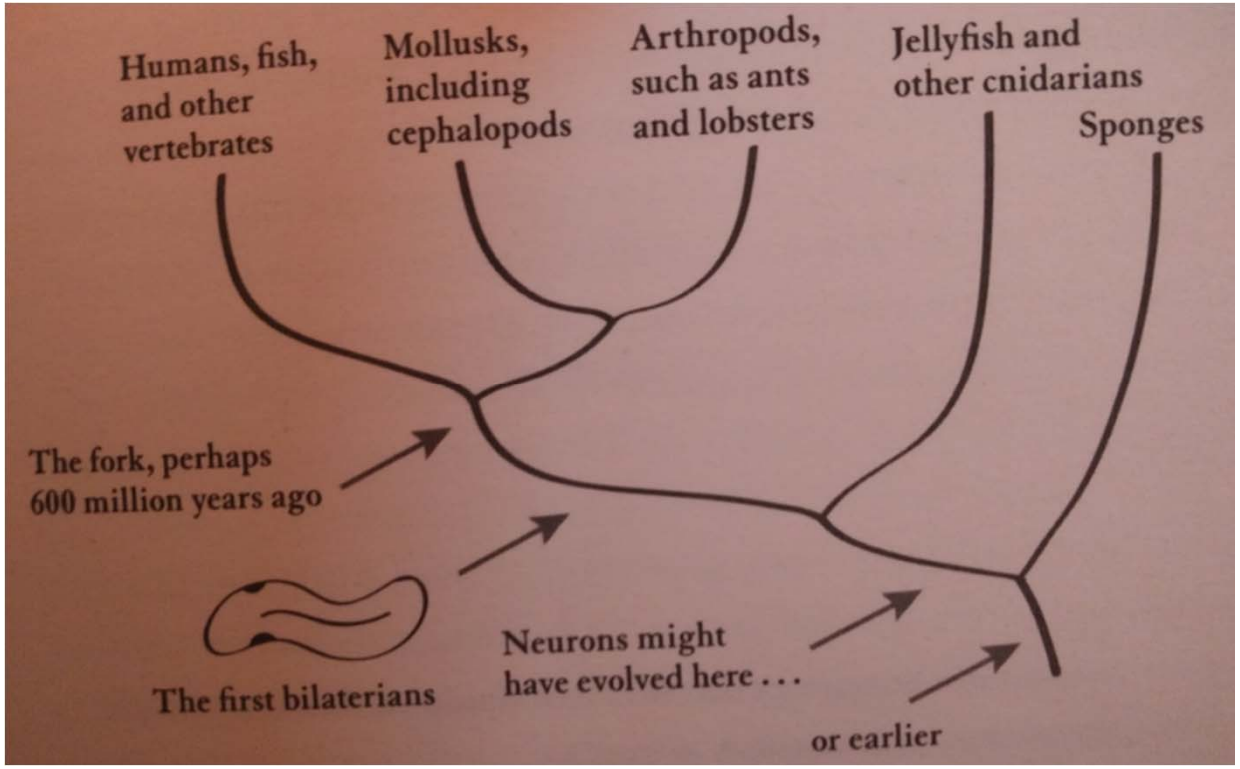
The eye of the beholder



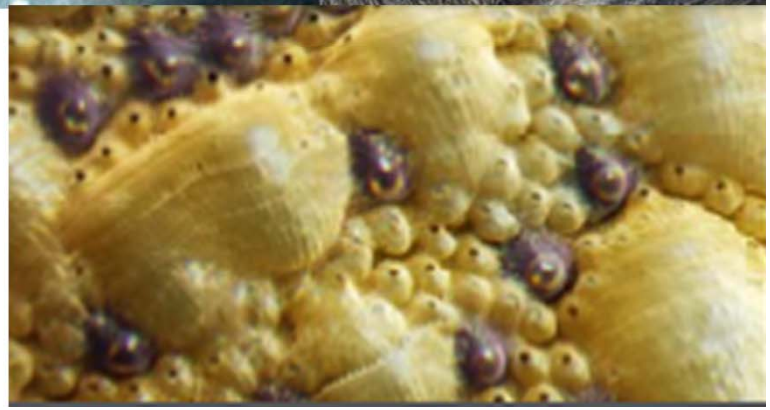
Context as (pre)history

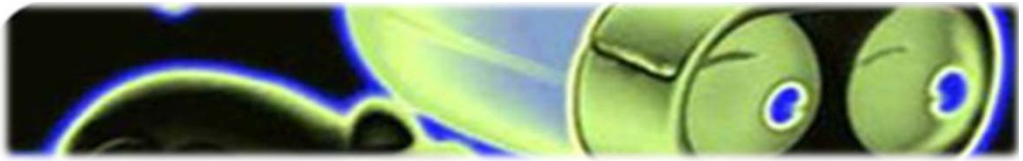
Enter the *Cambrian:* A time of weapons and sensors

"From this point on, the mind evolved in response to other minds …"



Eyes and visual systems as an extremelly succesful evolutionary solution that has emerged in very diverse forms through the tree of life.



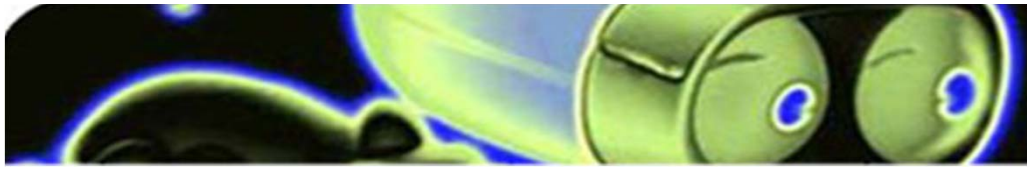


Context as (pre)history

Meet the octopus

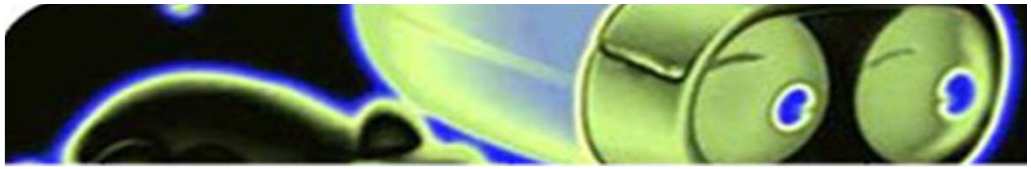
Departed from our branch in the evolutionary tree over 600M years ago and still it has evolved **eyes** not too dissimilar from ours (arguably better 'designed') and a **large brain** (Highest invertebrate brain-to-body mass ratio), **not fully centralized** (2/3 of neurons outside central brain; arms have "brains" of their own, with 10K+ neurons per suction cup) There is evidence that some of their 'relatives' are able to see through their skin (cuttlefish)

The eye of the beholder



Bear that octopus in mind

The eye of the beholder



Evento

EuroVis Invited Talk: Martin Wattenberg & Fernanda Viégas (Google)

Fecha+hora

Martes, 13 de junio de 2017
desde las 14:15 hasta las 15:40
(CEST)

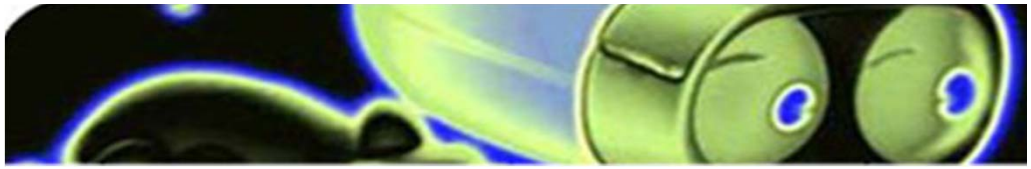
Ubicación

Open to members of UPC
Auditori Vèrtex, UPC
Campus Nord
Barcelona
España

Estado del pago

Pedido gratuito

The eye of the beholder



Wattenber and Viégas belong to the Big Picture (Data Visualization) Group

@ Google Brain

Their introductory talk title:

Visualization: The Secret Weapon of Machine Learning

Topics:

The TensorFlow Playground (see next slide)

How to scale visualization to DL?

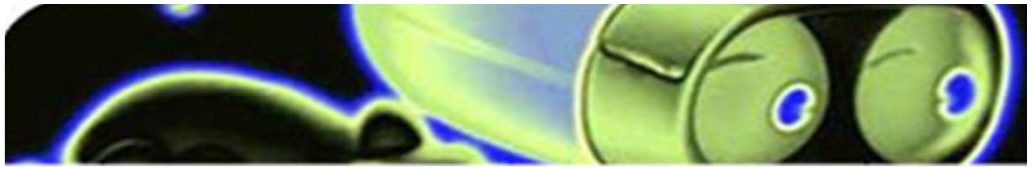
How to explain ANN decisions to end users?

Different end users: ← ML expert … Final user →

Can we automate visualization to show relevant features?

Right level of complexity to show?

The eye of the beholder



playground.tensorflow.org/#activation=tanh&batchSize=10&dataset=circle&regDataset=reg-plane&learningRate=0.03&regularizationRate=0&noise=0&networkShape=4,2&see

Bookmarks viajes uni freerange eBIB

Tinker With a **Neural Network** Right Here in Your Browser.
Don't Worry, You Can't Break It. We Promise.

Epoch 000,000 | Learning rate 0.03 | Activation Tanh | Regularization None | Regularization rate 0 | Problem type Classification

DATA
Which dataset do you want to use?

Ratio of training to test data: 50%

Noise: 0

Batch size: 10

FEATURES
Which properties do you want to feed in?

X_1
X_2
X_1^2
X_2^2
X_1X_2

2 HIDDEN LAYERS

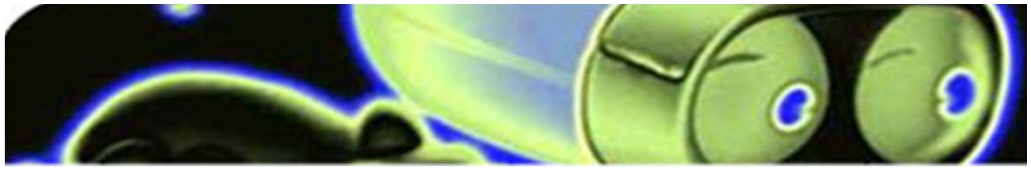
4 neurons | 2 neurons

*The outputs are mixed with varying **weights**, shown by the thickness of the lines.*

This is the output from one neuron

OUTPUT
Test loss 0.522
Training loss 0.512

The eye of the beholder



The Big Picture (Data Visualization) Group

@ Google Brain

Visualization: The Secret Weapon of Machine Learning

Their unsupervised weapon (see next slide):

projector.tensorflow.org

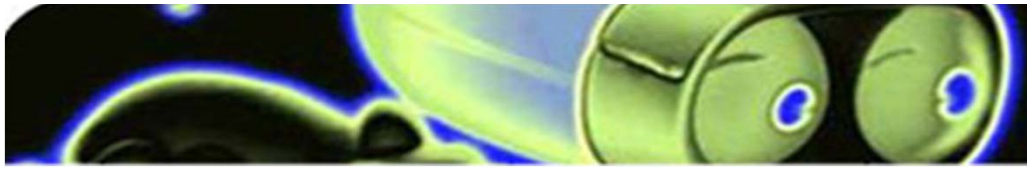
High dim spaces?

Right model to project?

PCA & t-SNE

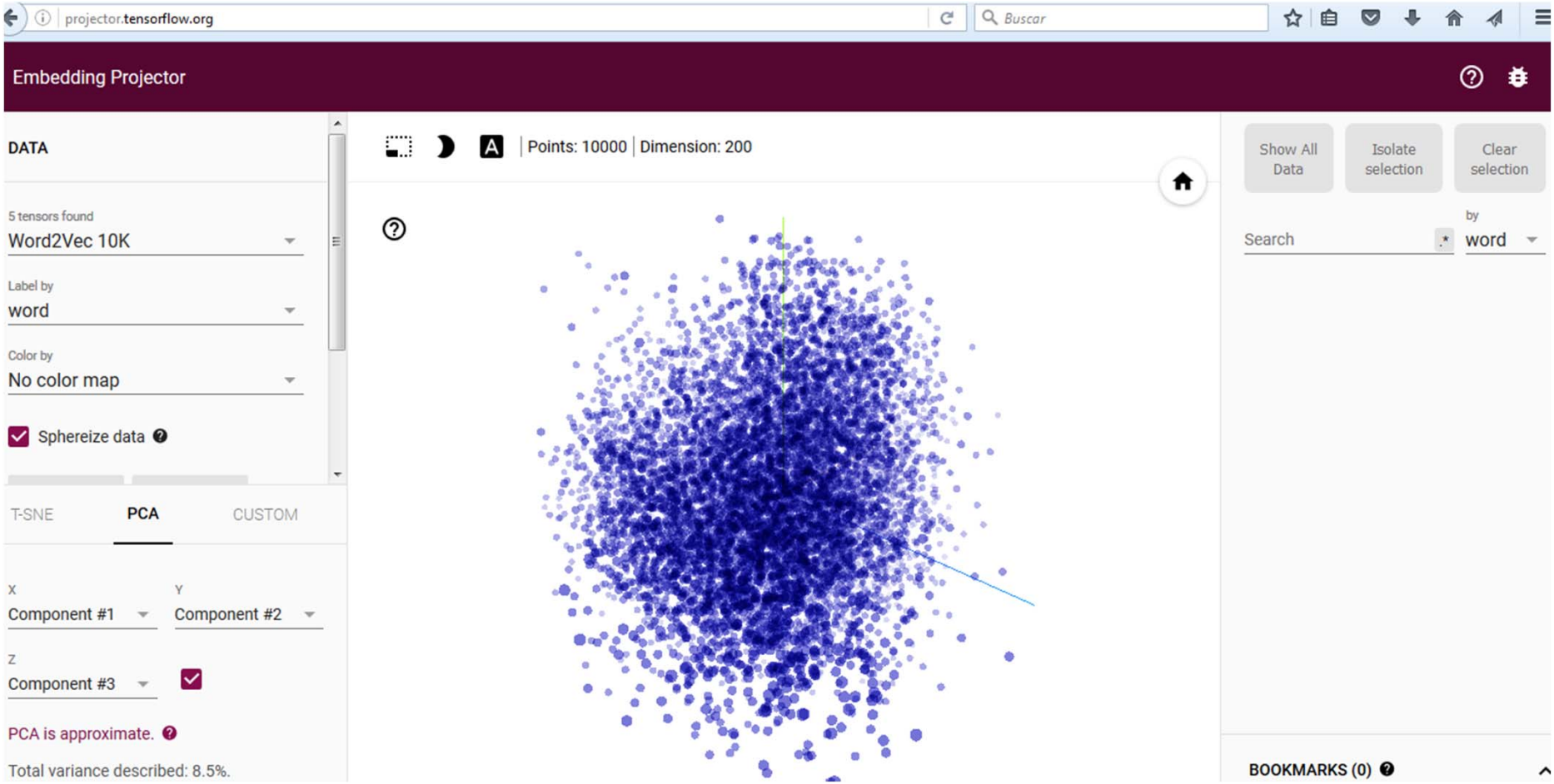
Too much to show?

The eye of the beholder

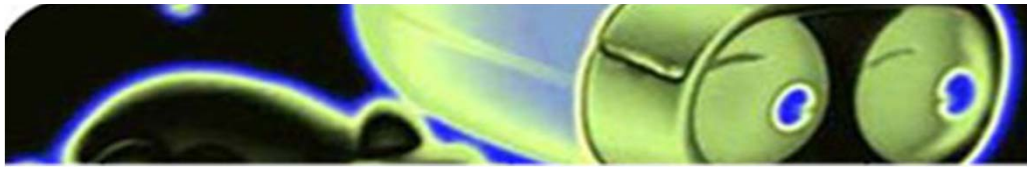


The Big Picture (Data Visualization) Group

@ Google Brain



The eye of the beholder



Interpretability just in recent conferences in BCN

NIPS, ICANN, EuroVIS

Interpretable ML for Complex Systems
NIPS 2016 Workshop

About | Call for Papers | Invited Talks | Organizers | Program Committee | Schedule

Interpreting the structure and predictions of complex models

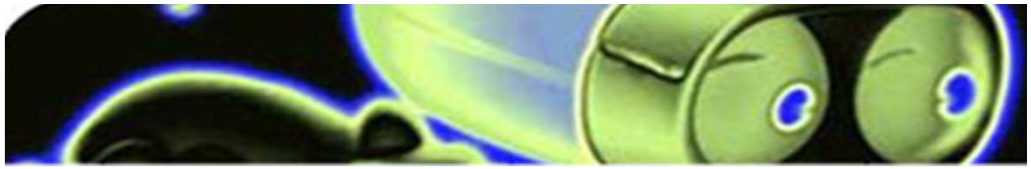
Complex machine learning models, such as deep neural networks, have recently achieved great predictive successes for visual object recognition, speech perception, language modeling, and information retrieval. These predictive successes are enabled by automatically learning expressive features from the data. Typically, these learned features are a priori unknown, difficult to engineer by hand, and hard to interpret. This workshop is about interpreting the structure and predictions of these complex models.

Interpreting the learned features and the outputs of complex systems allows us to more fundamentally understand our data and predictions, and to build more effective models. For example, we may build a complex model to predict long range crime activity. But by interpreting the learned structure of the model, we can gain new insights into the processing driving crime events, enabling us to develop more effective public policy. Moreover, if we learn, for example, that the model is making good predictions by discovering how the geometry of clusters of crime events affect future activity, we can use this knowledge to design

Key Dates

Workshop: 10 Dec 2016

The eye of the beholder



Interpretability just in recent conferences in BCN

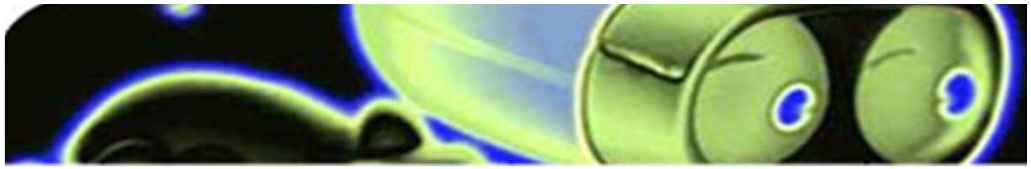
ICANN16

25th International Conference on Artificial Neural Networks

Workshop on Machine Learning and Interpretability

Tuesday, 6 September 2016 at **BarcelonaTech** Campus Nord, Spain

The eye of the beholder



CEX 2017

Comprehensibility and Explanation in AI and ML

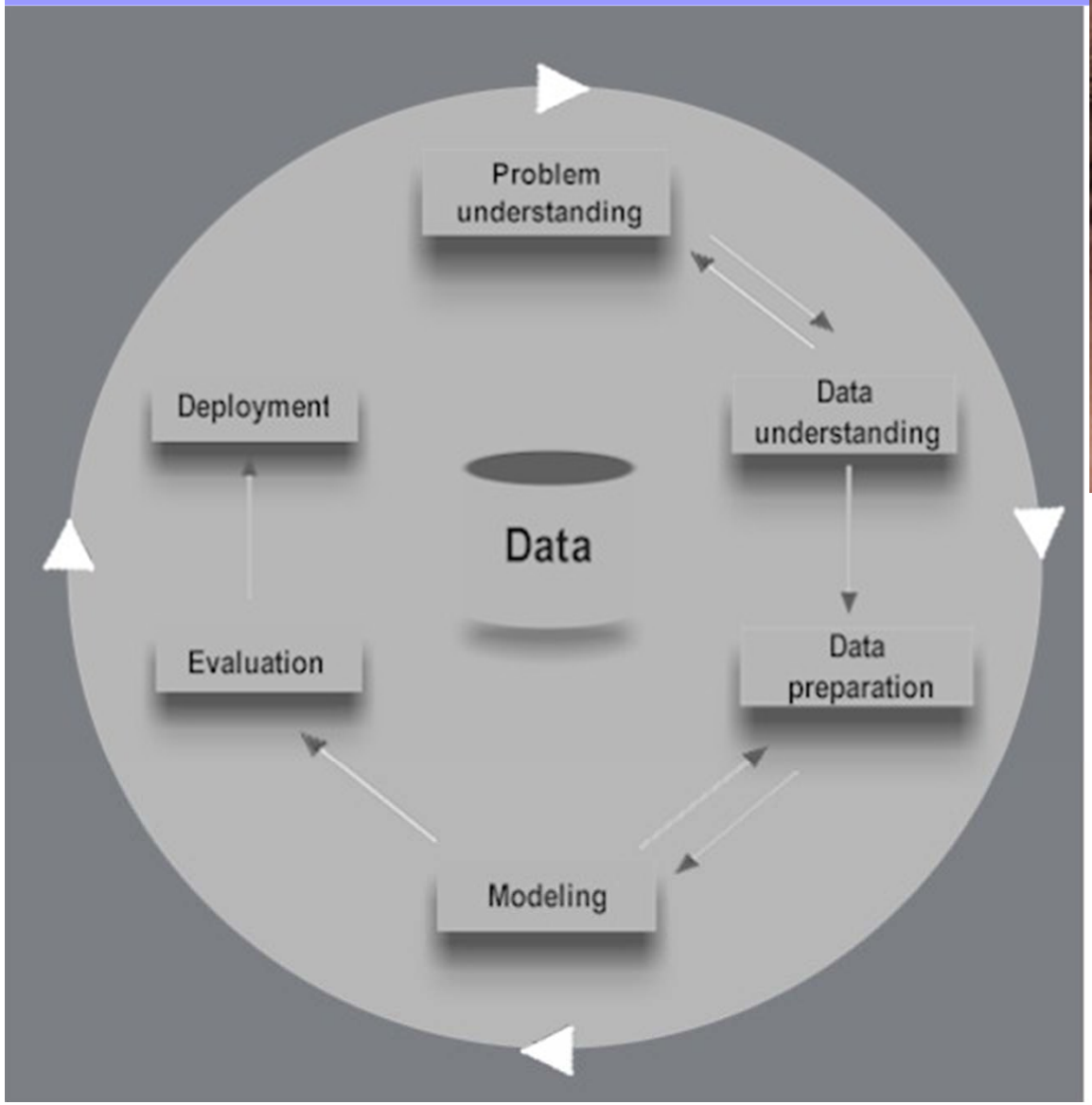
cex 2017

Quite frequently demands for better comprehensible and explainable Artificial Intelligence (AI) and Machine Learning (ML) systems are being put forward. The **cex** workshop sheds light on notions such as "comprehensibility" and "explanation" in the context of AI and ML, working towards a better understanding of what an explanation is when talking about intelligent systems, what it means to comprehend a system and its behaviour, and how human-machine interaction can take these dimensions into account.

The cex workshop will be held at the 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017).

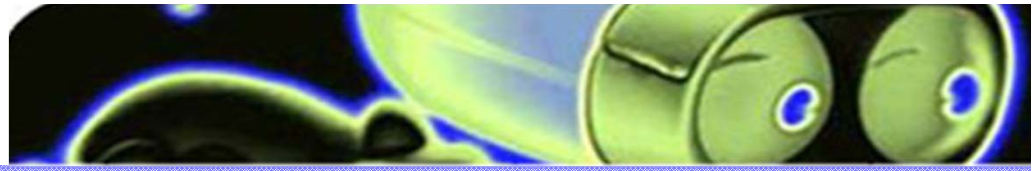
The eye of the beholder

Information visualization



Arguably, we can locate info/data visualization and visual analytics in different places of the data analysis cycle: **data understanding** (through visual exploration), **data preparation** (feature extraction, for instance, can be visualization oriented), and/or **data modeling** (linear/nonlinear data visualization models)

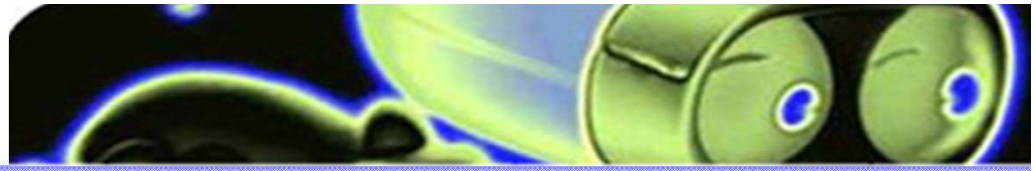
The eye of the beholder



Information visualization

- We aim discover the main characteristics of usually complex **multivariate data sets**, helping bring important aspects of the data into focus (think **attention**) for study in subsequent phases of the analysis.
- The task of **data visualization** is central to **both** data exploration and **model interpretation**.
- Adequate data visualizations can help us to gain insights into a problem without the frame of a conjecture: **Out of a deductive model of research to reap the benefits of a more inductive one**.
- The problem of **knowledge generation through data visualization**, is not circumscribed to data science *per se*. It can be addressed from the viewpoints of both **artificial pattern recognition (APR) and natural pattern recognition (NPR).**

The eye of the beholder



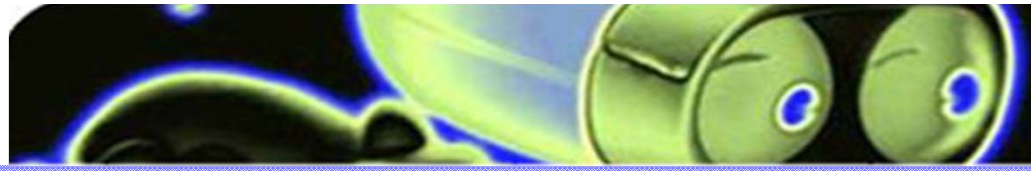
Information visualization

Artificial pattern recognition (APR):
Through the definition of methods /techniques /algorithms for data pre-processing and visualization.

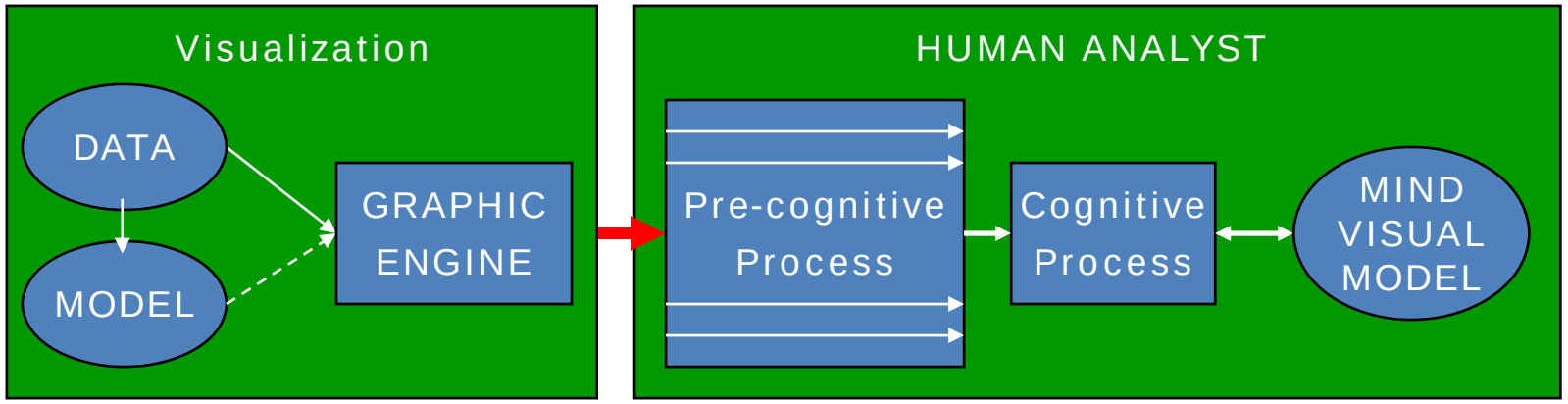
Natural pattern recognition (NPR):
Understanding visualization as the **cognitive processing of visual stimuli conducted by the human brain**
Humans are equipped with visual NPR as a tool to understand the patterns of their natural environment and operate upon it.

- **IV/VA** defines processes that unify DR algorithms and visual user (interactive) interfaces to allow us to explore data and interpret models using **graphical metaphors**, in a way that helps to **circumvent some of the inherent limitations of human vision**.

The eye of the beholder



Information visualization

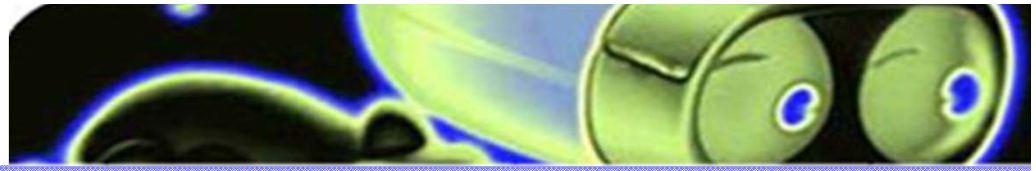


Pre-cognitive processing: jigsaw pieces, or the interaction of visual elements put together by the brain's own data pre-processing.

Visualization vs. mental model: visual patterns, illusions and *Gestalt* laws.

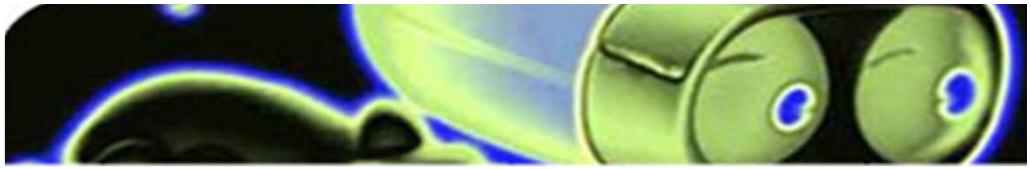
Perception vs. sensation: *"It may often be rather hard to say how much from perceptions as derived from the sense of sight is due directly to sensation, and how much of them, on the other hand, is due to experience and training"*
Hermann von Helmholtz, 1860 (Pollen, 1999, *Cerebral Cortex*)

The eye of the beholder

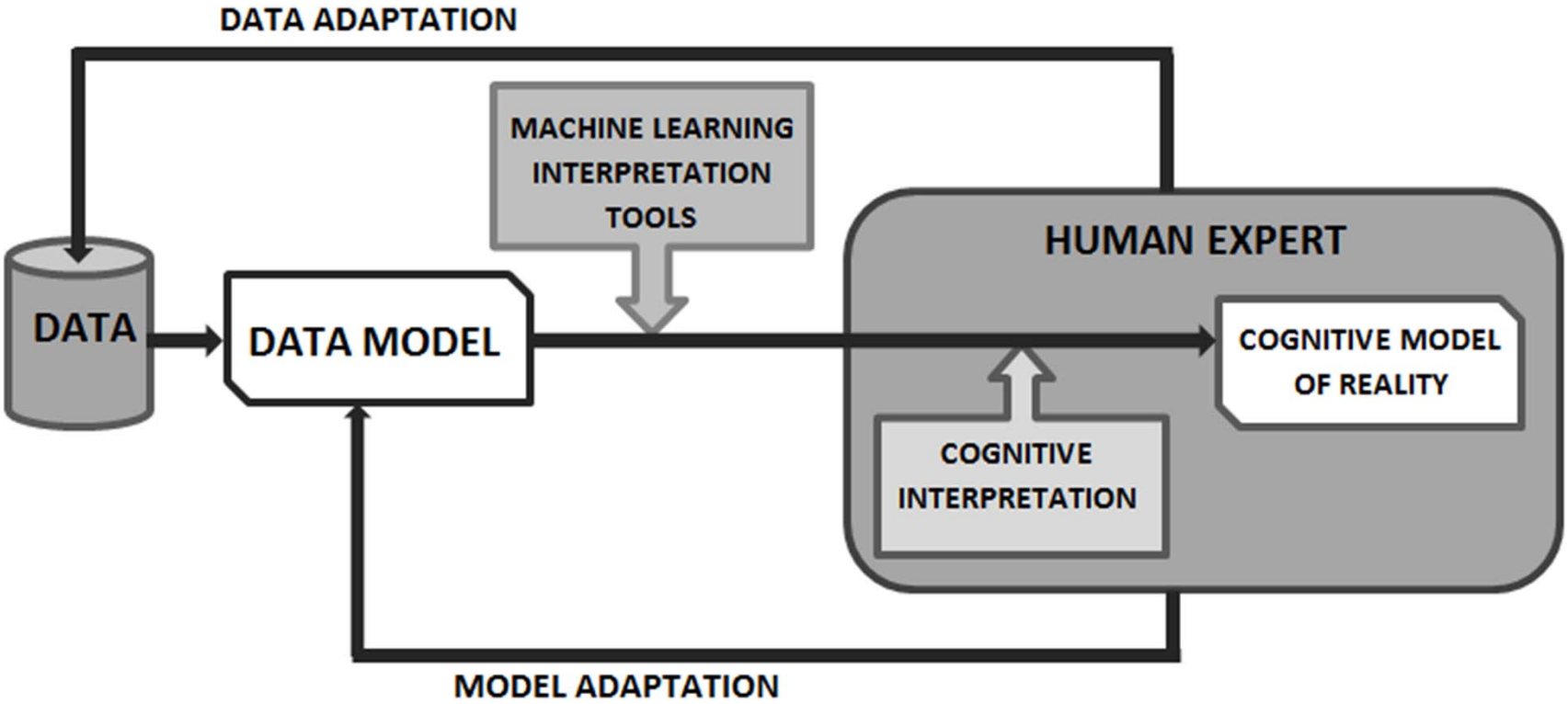


Information visualization

The eye of the beholder

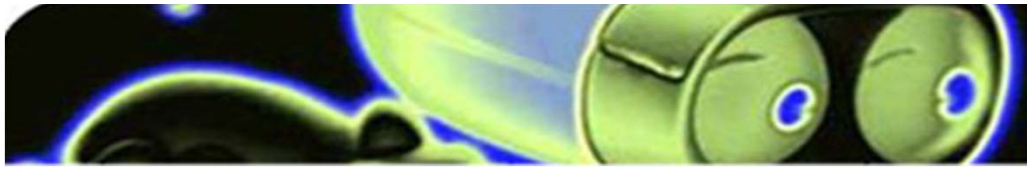


The human factor



- In the end, interpretability is a paramount quality that ML and related methods should aim to achieve if they are **to be applied in practice**.

The eye of the beholder



So we need a formal framework for visualization

Lucky us, help is on its way

ELSEVIER

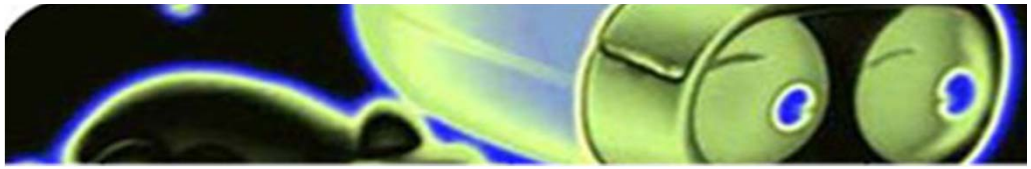
Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

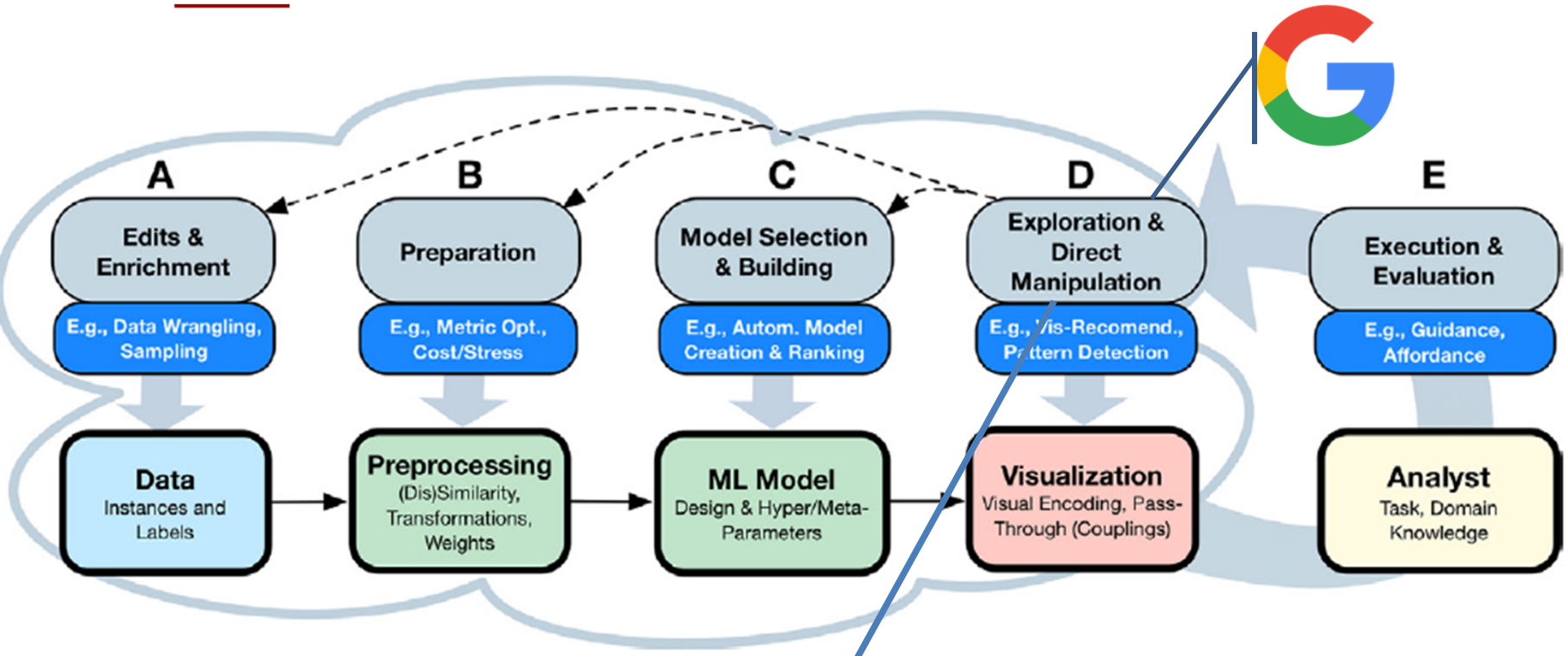
What you see is what you can change: Human-centered machine learning by interactive visualization

Dominik Sacha[a,*], Michael Sedlmair[b], Leishi Zhang[c], John A. Lee[d], Jaakko Peltonen[e], Daniel Weiskopf[f], Stephen C. North[g], Daniel A. Keim[a]

The eye of the beholder

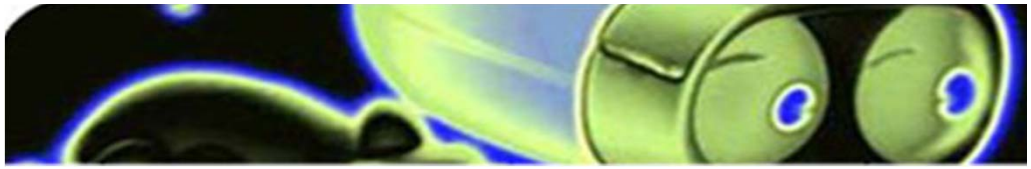


So we need a formal framework for visualization



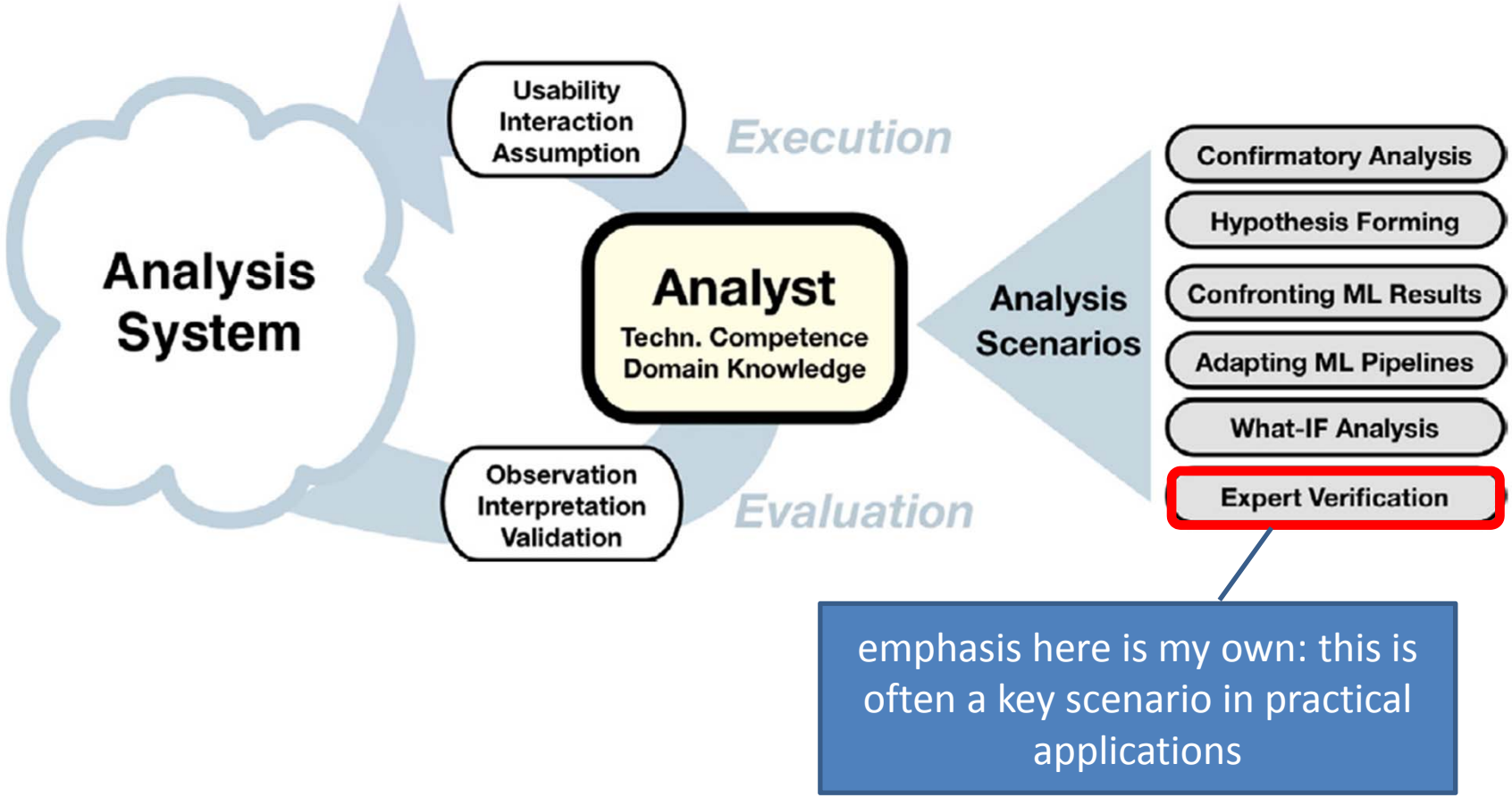
Aupetit M, Sedlmair M (2016) ***SepMe: 2002 new visual separation measures***, *In Procs. of the IEEE Pacific Visualization Symposium*, 1–8.

The eye of the beholder



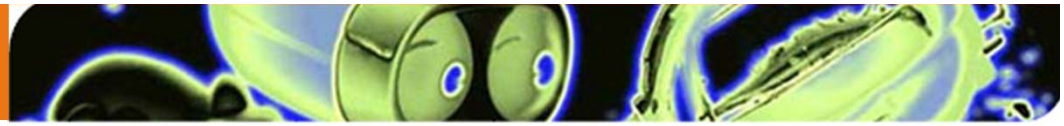
So we need a formal framework for visualization

Lucky us, help is on its way



HUMANS, INTERPRETABILITY, BIO- & HEALTH

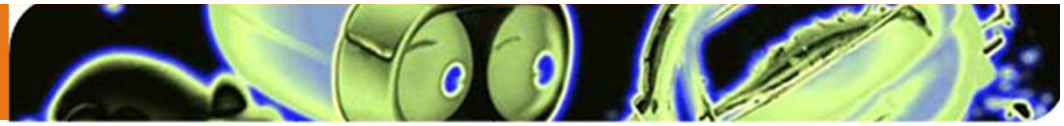
BIG DATA



How big is yours? … (2013, KDNuggets)

What was the largest dataset you analyzed / data mined? [322 votes]	
less than 1 MB (12)	3.7%
1.1 to 10 MB (8)	2.5%
11 to 100 MB (14)	4.3%
101 MB to 1 GB (50)	15.5%
1.1 to 10 GB (59)	18%
11 to 100 GB (52)	16%
101 GB to 1 Terabyte (59)	18%
1.1 to 10 TB (39)	12%
11 to 100 TB (15)	4.7%
101 TB to 1 Petabyte (6)	1.9%
1.1 to 10 PB (2)	0.6%
11 to 100 PB (0)	0%
over 100 PB (6)	1.9%

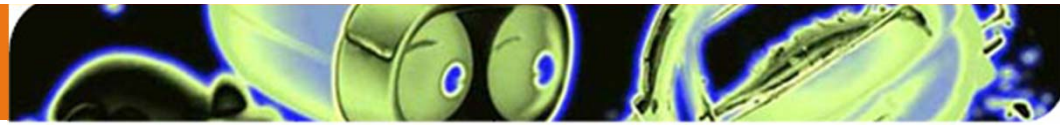
BIG & SMALL



Some fun facts:

- Google processes over **20 PB** worth of data **every day.**
- Back in December 2007, YouTube generated **27 PB** of traffic.
- The CERN Large Hadron Collider (HLC) generetes about **20 PB** of usable data **per year**.
- The volume of **global annual data traffic** is expected exceed 60,000 PB in 2016, from 8,000 petabytes in 2011
- In the next decade, astronomers expect to be processing **10 PB of data every hour** from the Square Kilometre Array (SKA) telescope ► **one exabyte every four days**.

The Big Data Interpretation Challenge



OK, but this is about the big ITCorps, la US NSA and the like, isn't it? … or maybe not …

Nov 14

FOCUS ON BIG DATA

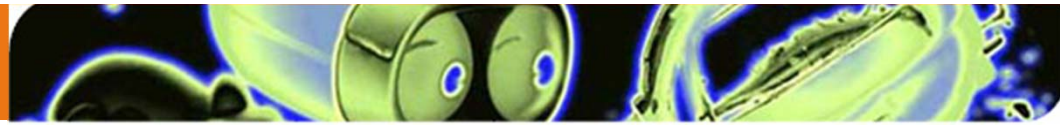
EDITORIAL

nature neuroscience

Focus on big data

Nature Neuroscience **presents a special focus issue highlighting big data efforts under way in the field.**

The number of big data projects in neuroscience, such as the BRAIN initiative in the United States or the Human Brain Mapping initiative in Europe, has been increasing in recent years. Will such big data efforts become the modus operandi in neuroscience, replacing smaller scale, hypothesis-driven science? How much insight will be gained from such projects? What are the best ways to go about conducting such studies of the connectivity and activity of neurons seek to understand the relationship between these neural data and behavior. On page 1455, Alex Gomez-Marin and colleagues review technological advances that have accelerated the collection and analysis of big behavioral data, but argue that substantial challenges remain in interpreting the results of such efforts. They conclude that large-scale quantitative and open behavioral



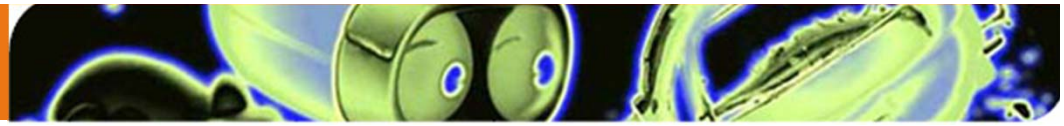
OK, but this is about the big ITCorps, la US NSA and the like, isn't it? … or maybe not …

NATURE | TECHNOLOGY FEATURE

Biology: The big challenges of big data

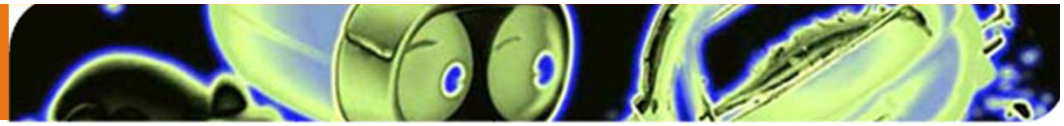
Vivien Marx

Nature **498**, 255–260 (13 June 2013) doi:10.1038/498255a
Published online 12 June 2013



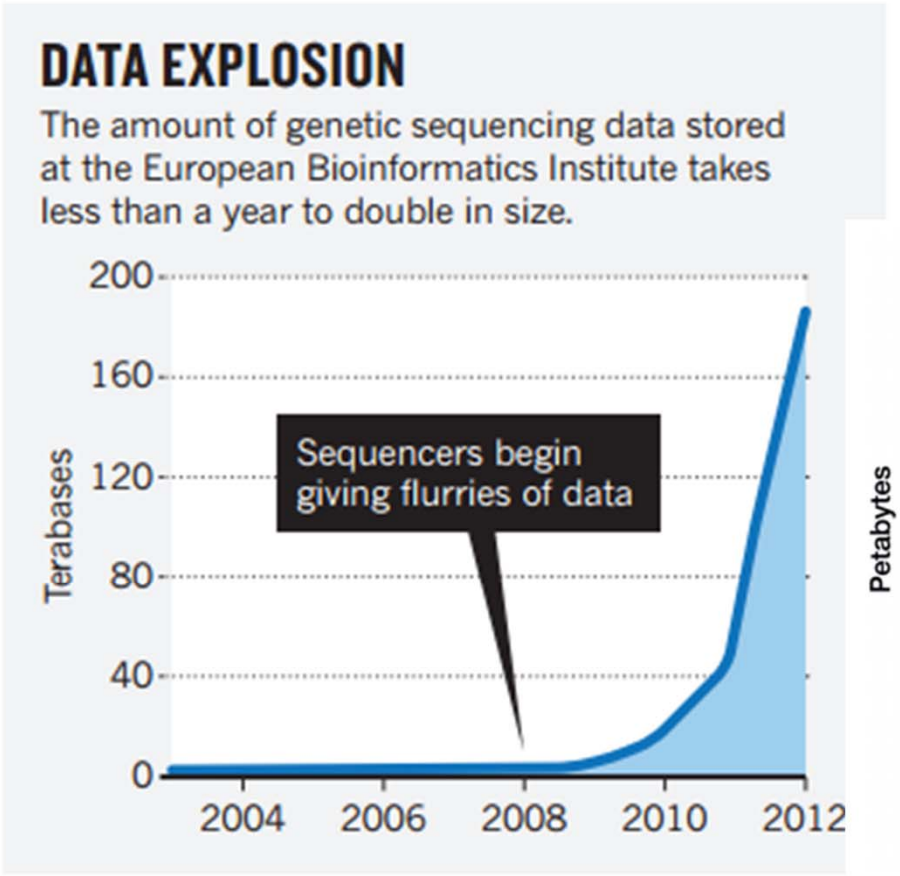
OK, but this is about the big ITCorps, la US NSA and the like, isn't it? … or maybe not …

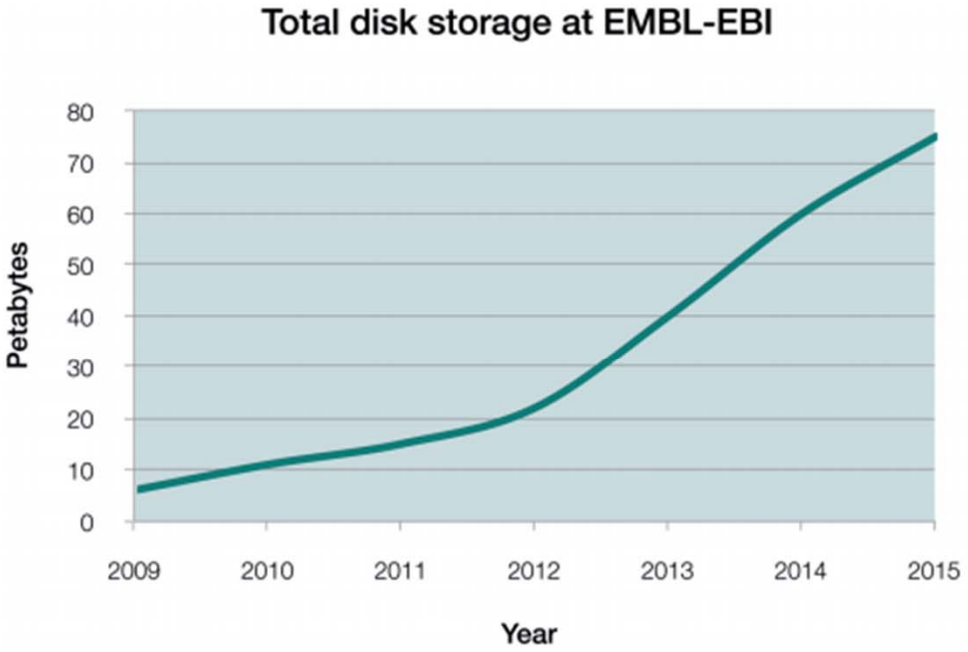
- Some extracts from the latter reference:
- "Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with **handling**, **processing** and **moving** information that were once the domain of astronomers and high-energy physicists."
- "The **European Bioinformatics Institute (EBI)** in Hinxton, UK, […] one of the world's largest biology-data repositories, currently stores **20 petabytes** […] Genomic data account for 2 Pb of that, a **number that more than doubles every year.**"

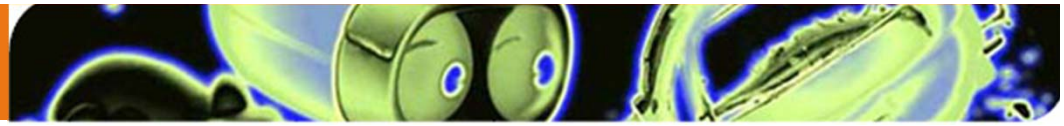


OK, but this is about the big ITCorps, la US NSA and the like, isn't it? …

SOURCE: EMBL-EBI







OK, but this is about the big ITCorps, la US NSA and the like, isn't it? …

- "**Most researchers tend to download remote data to local hardware for analysis. But this method is "backward"**, says Andreas Sundquist, chief technology officer of DNAnexus. "The data are so much larger than the tools, it makes no sense to be doing that." The alternative is to **use the cloud for both data storage and computing** […] There's no reason to move data outside the cloud. You can do analysis right there …"

which would fit some of these guys' proposals …

Evento

EuroVis Invited Talk: Martin Wattenberg & Fernanda Viégas (Google)

Fecha+hora

Martes, 13 de junio de 2017
desde las 14:15 hasta las 15:40
(CEST)

Ubicación

Open to members of UPC
Auditori Vèrtex, UPC
Campus Nord
Barcelona
España

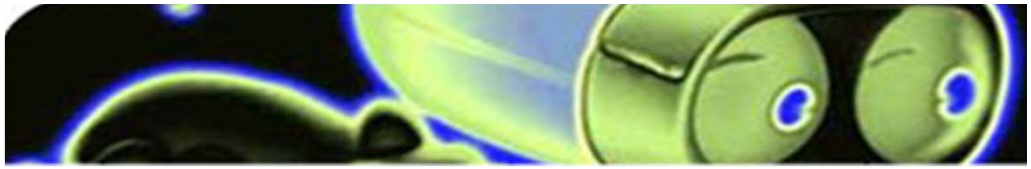
Estado del pago

Pedido gratuito

SMALL DATA

(more common than not)

The eye of the beholder



HUMANS, INTERPRETABILITY, BIO- & HEALTH

Data keep coming, which challenges attached:

Extreme data heterogeneity

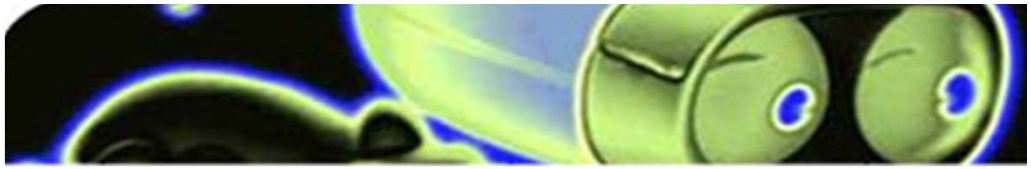
Multiple scales and multi-modality

Imbalance sample size ← → dimensionality

SM Reza, Transforming Big Data into Computational Models for Personalized Medicine and Health Care, Dialogues in Clinical Neuroscience, ***2016****, 18(3): 339-343*

"Health care systems generate a huge volume of different types of data. Due to the complexity and challenges inherent in studying medical information, it is not yet possible to create a comprehensive model capable of considering all the aspects of health care systems. There are different points of view regarding what the most efficient approaches toward utilization of this data would be."

The eye of the beholder



HUMANS, INTERPRETABILITY, BIO- & HEALTH

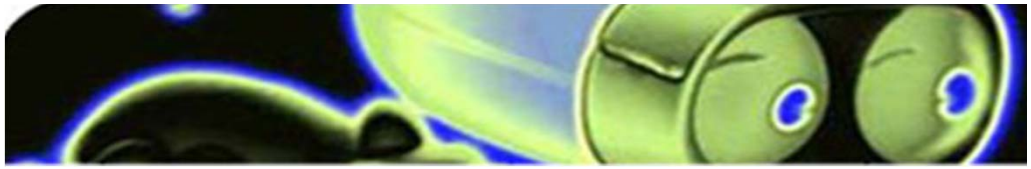
Massive gaps between info / patterns / knowledge / decision making

Machines learn, doctors decide

> ***"Machine Learning employed in healthcare will act as a tool to aid and refine specific tasks performed by human professionals"***

MJ Reid, ***Black-Box Machine Learning: Implications for Healthcare****, Polygeia, April 6, 2017*

The eye of the beholder



HUMANS, INTERPRETABILITY, BIO- & HEALTH

Biomedical data analysis in translational research

Bhanot, Gyan; Biehl, Michael; Villmann, Thomas; Zühlke, Dietlind

Published in:
25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2017

Authors highlight research questions & problems in Biomedicine & ML:

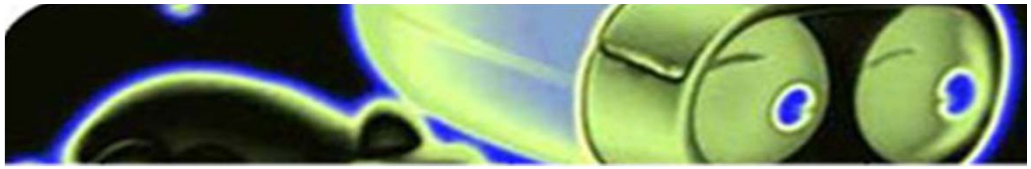
Analysis of structured, inhomogeneous and multimodal data

Feature selection and biomarker detection

Diagnosis and classification

...

The eye of the beholder



HUMANS, INTERPRETABILITY, BIO- & HEALTH

Biomedical data analysis in translational research

Bhanot, Gyan; Biehl, Michael; Villmann, Thomas; Zühlke, Dietlind

Visualization & visual analytics (!)

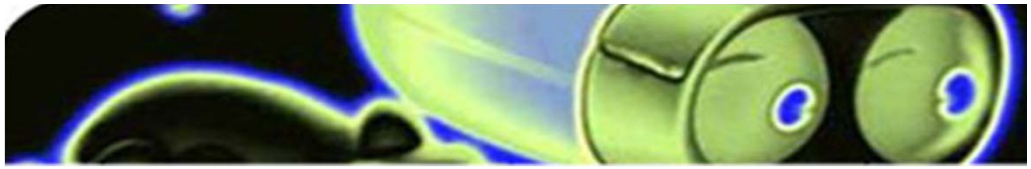
(a) Which visualization techniques are most suitable for heterogeneous and structured data in the biomedical context? How can we visualize relevant features in order to facilitate their interpretation by human experts?

(c) How can one design feedback loops that can improve visualization models by integrating requirements/limitations of domain experts with the model?

(d) How can we effectively visualize patient monitoring and critical events in a longitudinal assessment of patients?

(e) How can we usefully display the errors/uncertainties resulting e.g. from embedding high-dimensional data into low dimensional spaces? Can these be visualized in an intuitive way?

The eye of the beholder

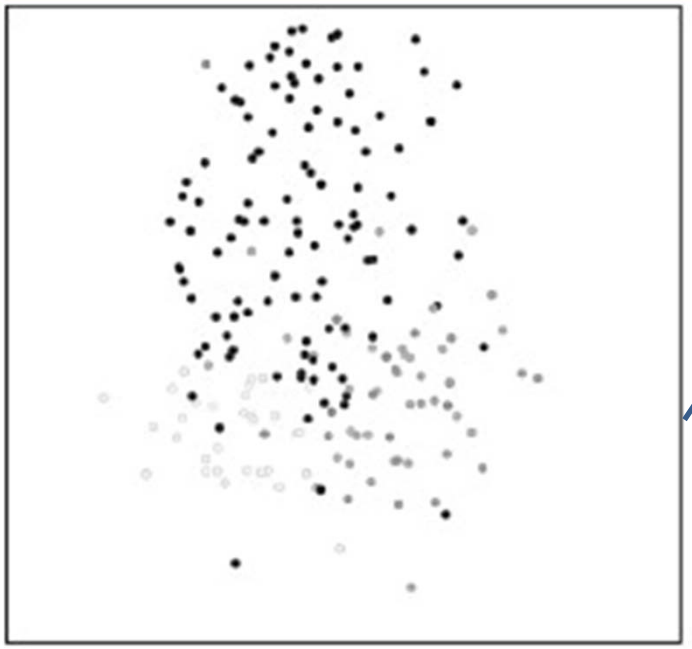


Some examples from my own research: Who is the customer?

- **An example in the medical domain**. Interpretation can also be **a matter of shared language**. We could be talking in a different language that domain experts might not be interested to talk.

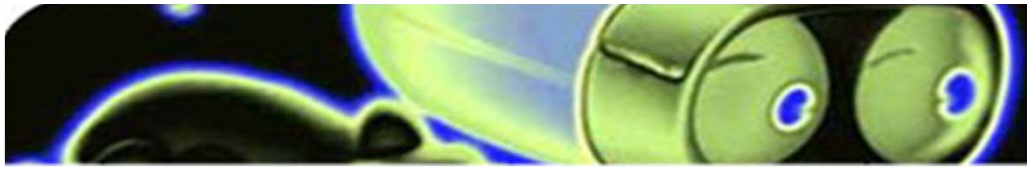
Brain tumour diagnosis for radiologists:

Vellido, A., Romero, E., González-Navarro, F.F., Belanche-Muñoz, Ll., Julià-Sapé, M., Arús, C. (2009) Outlier exploration and diagnostic classification of a multi-centre 1H-MRS brain tumour database. ***Neurocomputing***, 72(13-15), 3085-3097.



Visualization as part of a quest for anomalies and outliers

The eye of the beholder



Some examples from my own research: Who is the customer?

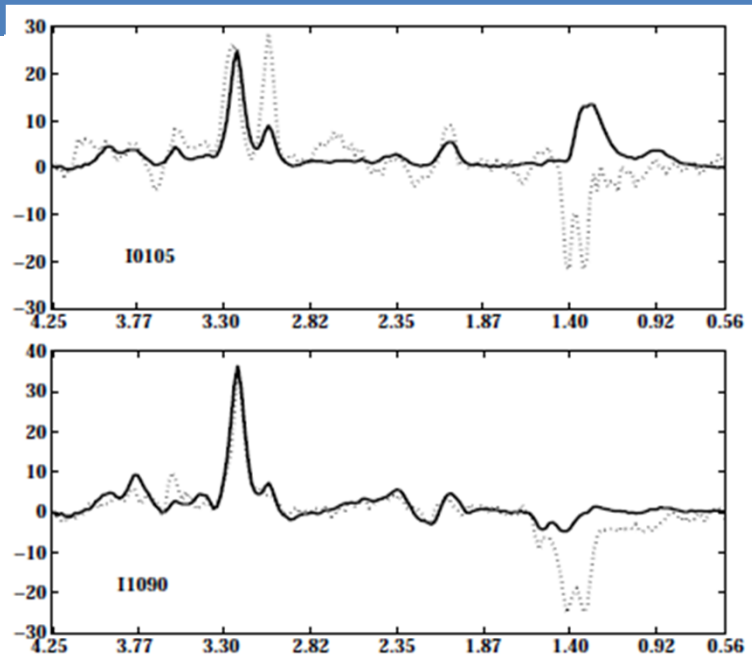
- **An example in the medical domain**. Radiologists did not consider as outliers many of the cases that our machine learning considered as such. Why? because as humans they were still capable of identifying and characterizing such cases without ambiguity.

Brain tumour diagnosis for radiologists:

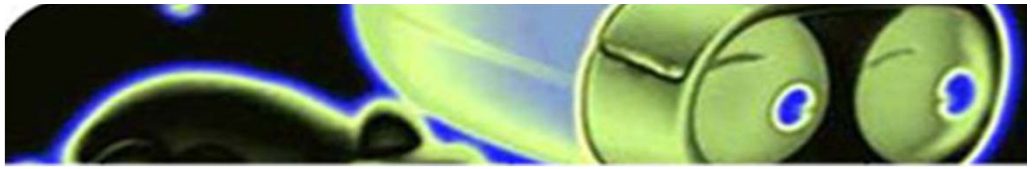
Vellido, A., Romero, E., González-Navarro, F.F., Belanche-Muñoz, Ll., Julià-Sapé, M., Arús, C. (2009) Outlier exploration and diagnostic classification of a multi-centre 1H-MRS brain tumour database. ***Neurocomputing***, 72(13-15), 3085-3097.

- **Beware of prior expert knowledge**

Id	*Tum*	*Dis*	*Artifact-relat. outl.*					
			noi	wat	ali	bas	pol	edd
I1061	G1(a2)				X			
I0062*	G2(gl)		X	X		X		
I0105*	G2(gl)	X						
I0172	G2(gl)			X		X		
I0175*	G2(gl)		X				X	
I0354*	G2(gl)			X			X	
I0428*	G2(gl)			X			X	



The eye of the beholder



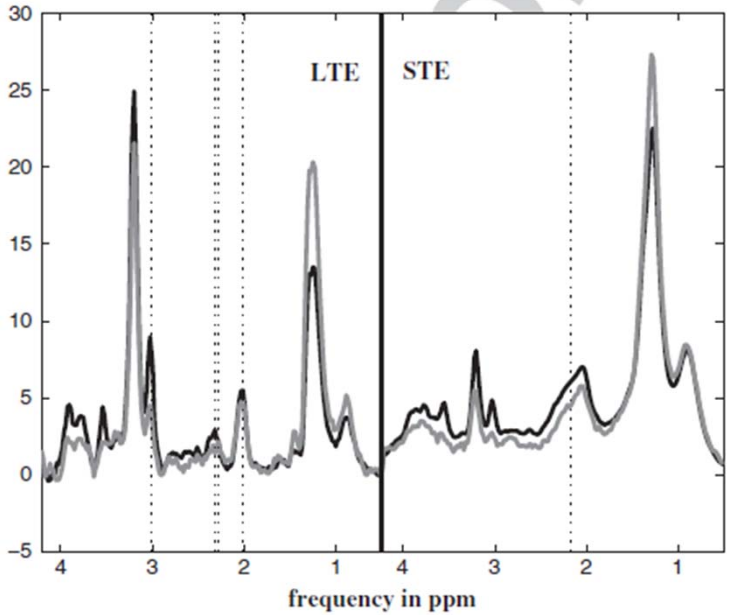
Some examples from my own research: Who is the customer? (b)

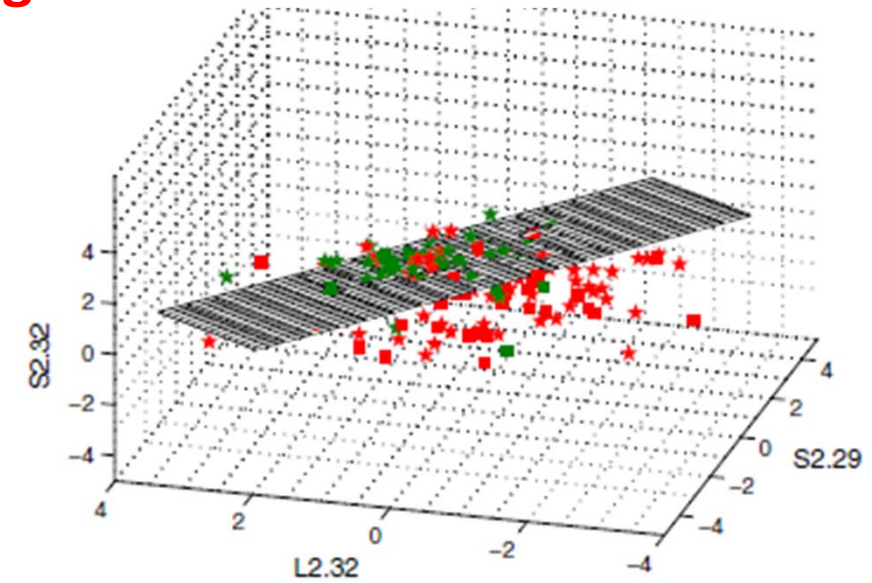
- **An example in the medical domain**. Only the most straightforward 3-D visualization of a difficult classification problem finally convinced the medical experts of the relevance of our feature selection + classification results.

- **Brain tumour diagnosis for radiologists:**

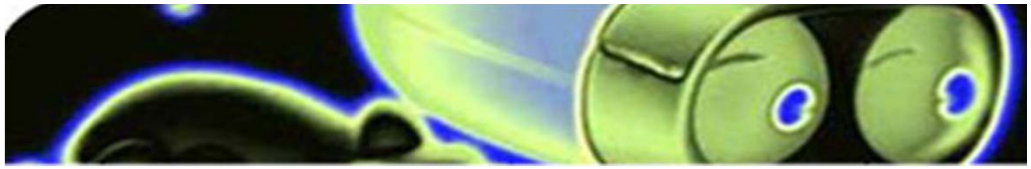
Vellido, A., Romero, E., Julià-Sapé, M., Majós, C., Moreno-Torres, À., and Arús, C. (2012) Robust discrimination of glioblastomas from metastatic brain tumors on the basis of single-voxel proton MRS. ***NMR in Biomedicine***, 25(6), 819-828.

- **Beware of prior expert knowledge**





The eye of the beholder



Some examples from my own research: Who is the customer?

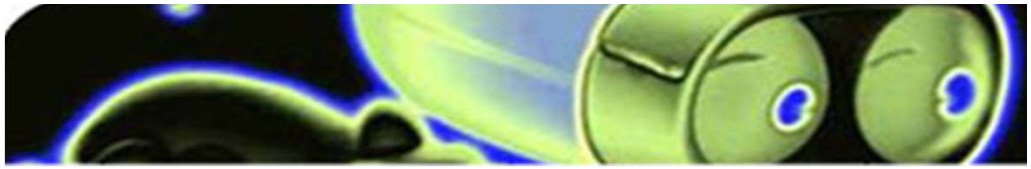
- **Another example from a different medical domain**. Medical experts may only accept a parsimonious outcome from a ML method, as **they require an explainable basis for their decision making** that **complies with their standard operational guidelines**, often based on simple and rigid attribute scores.

Mortality prediction due to sepsis at the ICU:

Ribas, V.J., Vellido, A., Ruiz-Rodríguez, J.C., Rello, J. (2012) Severe sepsis mortality prediction with logistic regression over latent factors. ***Expert Systems with Applications***, 39(2), 1937-1943.

- **Only simple well-trodden models were accepted, despite better results being achieved with more sophisticated, less known models.**
- **Logistic Regression + Factor Analysis:** as complex as it gets in an application context in which **standard scores** (SOFA, APACHE) are routinely used.
- Regardless success in prediction, end-user adoption of alternative methods should not be expected.
- **Beware of existing methods of interpretation …**

The eye of the beholder



Some examples from my own research: Who is the customer?

- **Another example from a different medical domain**. Medical experts may only accept a parsimonious outcome from a ML method, as **they require an explainable basis for their decision making** that **complies with their standard operational guidelines**, <u>often based on simple and rigid attribute scores</u>.

Mortality prediction due to sepsis at the ICU:

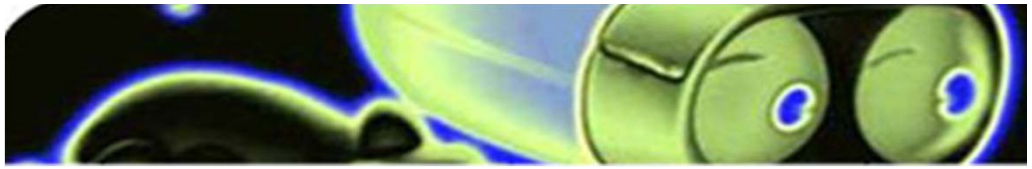
Ribas, V.J., Vellido, A., Ruiz-Rodríguez, J.C., Rello, J. (2012) Severe sepsis mortality prediction with logistic regression over latent factors. ***Expert Systems with Applications***, 39(2), 1937-1943.

This is despite messages such as this:

"First, machine learning will dramatically improve prognosis. Current prognostic models (e.g., the Acute Physiology and Chronic Health Evaluation [APACHE] score and the Sequential Organ Failure Assessment [SOFA] score) are restricted to only a handful of variables, because humans must enter and tally the scores. But data could instead be drawn directly from EHRs or claims databases, allowing models to use thousands of rich predictor variables. Does doing so lead to better predictions? "

Z Obermeyer, EJ Emanuel (2016) New England Journal of Medicine, 375(13)

The eye of the beholder

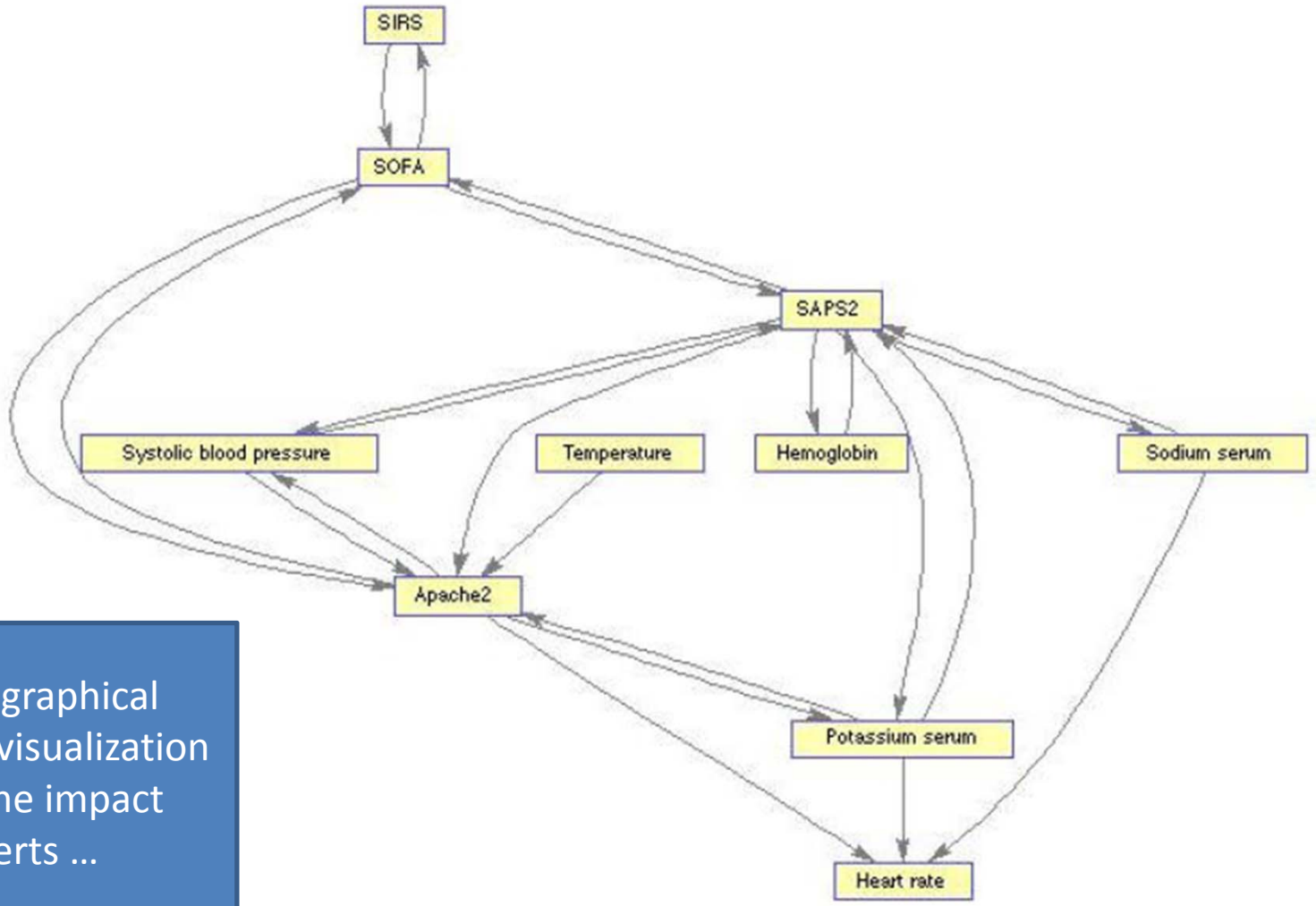


Who is the customer?

- **Another** parsimon **decision** simple an

Mortality

Ribas, V.J., V prediction w 39(2), 1937-



More complex graphical models and their visualization achieved half the impact among experts …

SIMPLE MODELS x SMALL DATA

- **This experience at least partially fits the viewpoint of "politically incorrect" Prof. David J. Hand**

Statist. Sci.
Volume 21, Number 1 (2006), 1-14.

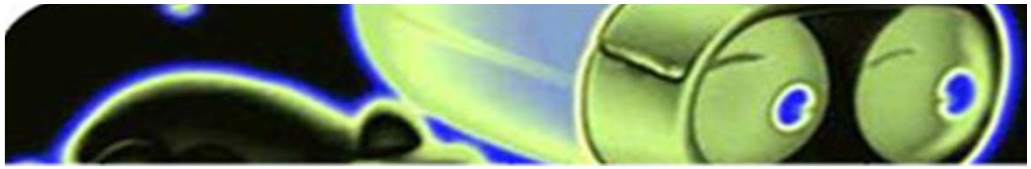
Classifier Technology and the Illusion of Progress

David J. Hand

Abstract

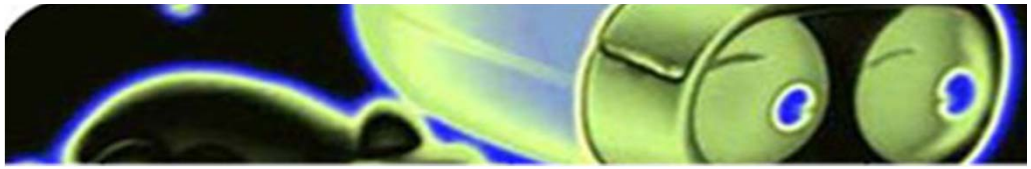
A great many tools have been developed for supervised classification, ranging from early methods such as linear discriminant analysis through to modern developments such as neural networks and support vector machines. A large number of comparative studies have been conducted in attempts to establish the relative superiority of these methods. **This paper argues that these comparisons often fail to take into account important aspects of real problems, so that the apparent superiority of more sophisticated methods may be something of an illusion**. In particular, simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.

The eye of the beholder



Some general comments on DR for Interpretation

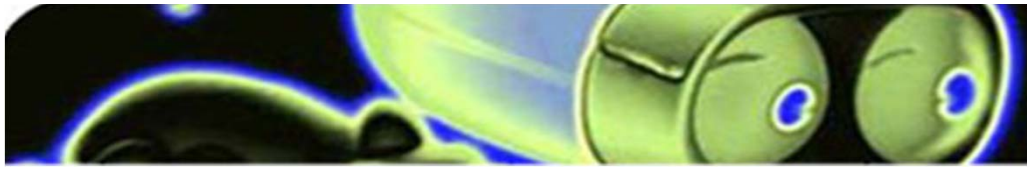
- Problems of **high and very high dimensionality** are **becoming commonplace** in bio-fields such as, for instance, bioinformatics …
- Almost no problem is interpretable in practice if all data attributes are retained and used to provide an outcome. Furthermore, **data of very high-dimensionality are bound to show unexpected geometrical properties** that might affect their modeling and bias the interpretation of results.
- **Two main DR approaches**: **feature selection** (supervised and unsupervised) and **feature extraction**, in which new non-observable features are created on the basis of the observed ones ...
- **NOTE:** some of the most popular DR techniques in real-world applications are precisely some of the simplest ones (e.g., the ubiquitous PCA).



Some general comments on (NL)DR for Interpretation

- Many relevant ML contributions to the problem of multivariate data DR have stemmed from the **field of NLDR**.
- The **challenge of interpretability is very explicit here**: NLDR methods rarely provide an easy interpretation of the outcome in terms of the original data features, because such **outcome is usually a non-trivial nonlinear function of these features**.
- NLDR techniques usually attempt to **minimize** the **distortion** they introduce in the **mapping of the data from the observed space onto lower-dimensional spaces**.

The eye of the beholder

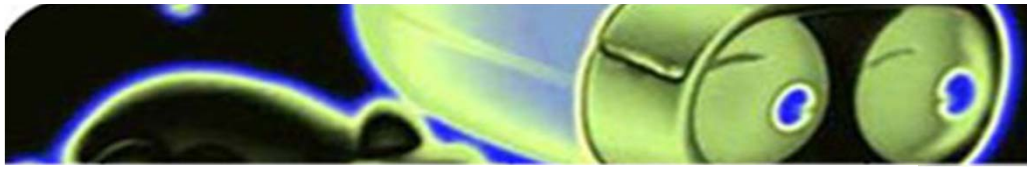


(Yet) more help on the way …

IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS VOL. 23, NO. 1, JANUARY 2017

Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis

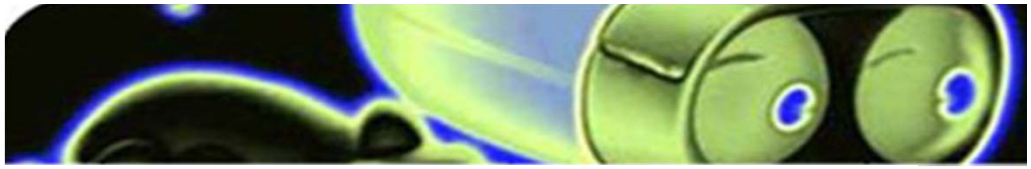
Dominik Sacha, Leishi Zhang, Michael Sedlmair, John A. Lee, Jaakko Peltonen, Daniel Weiskopf, Stephen C. North, Daniel A. Keim



Wrap up

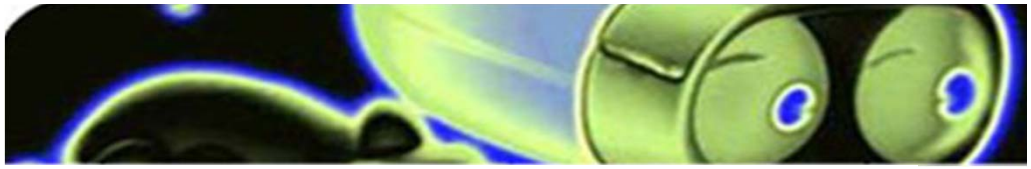
- The journey from data to knowledge is treacherous and **knowledge discovery processes require model interpretation**.
- **ML & related models' users** should aspire not only to integrate interpretation strategies … but also to ***negotiate*** with end-users (very specially those in the medical and heath domains) the terms of that interpretability.
- **NLDR models** (and nonlinear models in general) are not likely to become mainstream in many app fields unless they are designed with interpretability in mind.

- **Graphical metaphors** (visualization, graphical models) are naturally suited as interpretability tools.
- When handling interpretability, we can only ignore **NPR** at our own peril and adding interactivity to the VA process could be (the?) key to succesful application.



Some further reading

- Vellido, A, Martín-Guerrero, JD, & Lisboa, PJ (2012). Making machine learning models interpretable. In *ESANN 2012*, pp. 163-172
- Van Belle, V, Lisboa, PJG (2013) Research directions in interpretable machine learning models. In *ESANN 2013* pp. 533-541
- Turner, R (2015) A Model Explanation System. In Black Box Learning and Inference *NIPS Workshop*
- Schulz, A, Gisbrecht, A, & Hammer, B (2015) Using discriminative dimensionality reduction to visualize classifiers. *Neural Processing Letters*, 42(1), 27-54.
- Condry, N (2016). Meaningful models: utilizing conceptual structure to improve Machine Learning interpretability. *arXiv* preprint arXiv:1607.00279.
- Doshi-Velez, F & Kim, B (2017) Towards a rigorous science of interpretable Machine Learning. arXiv:1702.08608v2
- Lipton, ZC (2017) The mythos of model interpretability. arXiv:1606.03490v3
- Goodman, B & Flaxman, S (2016) EU regulations on algorithmic decision-making and a "right to explanation". *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY. arXiv* preprint arXiv:1606.08813.



Some reading

- Goodman, B & Flaxman, S (2016) EU regulations on algorithmic decision-making and a "right to explanation". *arXiv* preprint arXiv:1606.08813.

Abstract

We summarize the potential impact that the **European Union's new General Data Protection Regulation** will have on the routine use of machine learning algorithms. Slated to **take effect as law across the EU in 2018**, it will restrict automated individual decision-making (that is, algorithms that make decisions based on user-level predictors) which "significantly affect" users. The law will also effectively create a "**right to explanation**", whereby a user can ask for an explanation of an algorithmic decision that was made about them. We argue that while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation.

The eye of the beholder