# MAKE Decisions
## Medical Information Science for Decision Support

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

https://hci-kdd.org/mini-course-make-decisions-practice

Day 2 > Part 5 > 20.09.2018

# Methods of explainable AI
**concise overview (details in LV 706.315 TU Graz)**

# Science is to test ideas – Engineering is putting it into practice

## Day 1 - Hot Ideas

- 01 Information Sciences meets Life Sciences
- 02 Data, Information and Knowledge
- 03 Decision Making and Decision Support
- 04 DSS: from Expert Systems to explainable AI

## Day 2 - Cool Practice

- **05 Methods of Explainable-AI**
- Groupwork: Planning of a 500 bed Hospital - Bringing AI into the workflows
- Plenary: Presenting of the developed concepts

## Keywords

- Explainable Artificial intelligence
- Re-trace > Understand > Explain
- Transparency > Trust > Acceptance
- Fairness > Transparency > Accountability
- Methods of Explainable AI
- Transparent Machine Learning
- Intelligible > Reliability > Trustworthy
- Interpretability > Explicable > Accountable
- Interpretable Deep Learning

# Advance Organizer (1/2)

- **Ante-hoc Explainability (AHE)** := such models are interpretable by design, e.g. glass-box approaches; typical examples include linear regression, decision trees/lists, random forests, Naive Bayes and fuzzy inference systems; or GAMs, Stochastic AOGs, and deep symbolic networks; they have a long tradition and can be designed from expert knowledge or from data and are useful as framework for the interaction between human knowledge and hidden knowledge in the data.
- **Decision Making** = central cognitive process in every medical activity, resulting in the selection of a final choice of action out of several alternatives;
- **Decision Support System (DSS)** = is an IS including knowledge based systems to interactively support decision-making activities, i.e. making data useful;
- **Explainability** := motivated by the opaqueness of so called “black-box” approaches it is the ability to provide an explanation on why a machine decision has been reached (e.g. why is it a cat what the deep network recognized). Finding an appropriate explanation is difficult, because this needs understanding the context and providing a description of causality and consequences of a given fact. (German: Erklärbarkeit; siehe auch: Verstehbarkeit, Nachvollziehbarkeit, Zurückverfolgbarkeit, Transparenz)
- **Explanation** := set of statements to describe a given set of facts to clarify causality, context and consequences thereof and is a core topic of knowledge discovery involving “why” questionss (“Why is this a cat?”). (German: Erklärung, Begründung)
- **Explanatory power** := is the ability of a set hypothesis to effectively explain the subject matter it pertains to (opposite: explanatory impotence).
- **European General Data Protection Regulation** (EU GDPR) := Regulation EU 2016/679 – see the EUR-Lex 32016R0679 , will make black-box approaches difficult to use, because they often are not able to explain why a decision has been made (see explainable AI).

# Advance Organizer (2/2)

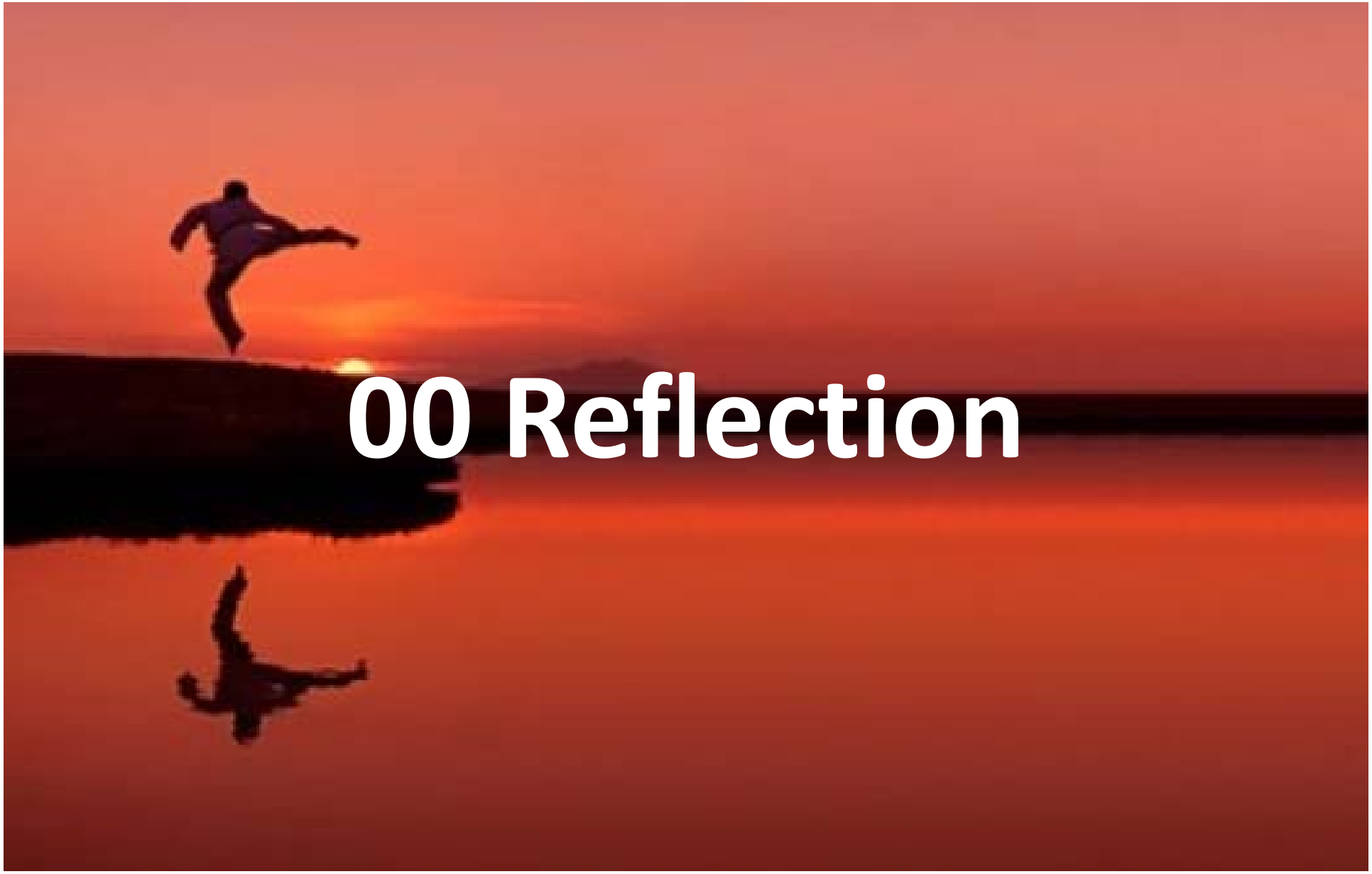
- **Interactive Machine Learning (iML)** := machine learning algorithms which can interact with – partly human – agents and can optimize its learning behaviour trough this interaction. Holzinger, A. 2016. Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics (BRIN), 3, (2), 119-131.
- **Inverse Probability** := an older term for the probability distribution of an unobserved variable, and was described by De Morgan 1837, in reference to Laplace's (1774) method of probability.
- **Post-hoc Explainability (PHE)** := such models are designed for interpreting black-box models and provide local explanations for a specific decision and re-enact on request, typical examples include LIME, BETA, LRP, or Local Gradient Explanation Vectors, prediction decomposition or simply feature selection.
- **Reasoning** = cognitive (thought) processes involved in making medical decisions (clinical reasoning, medical problem solving, diagnostic reasoning;
- **Transparency** = opposite of opacity of black-box approaches, and connotes the ability to understand how a model works (that does not mean that it should always be understood, but that – in the case of necessity – it can be re-enacted

## Learning Goals: At the end of this lecture you …

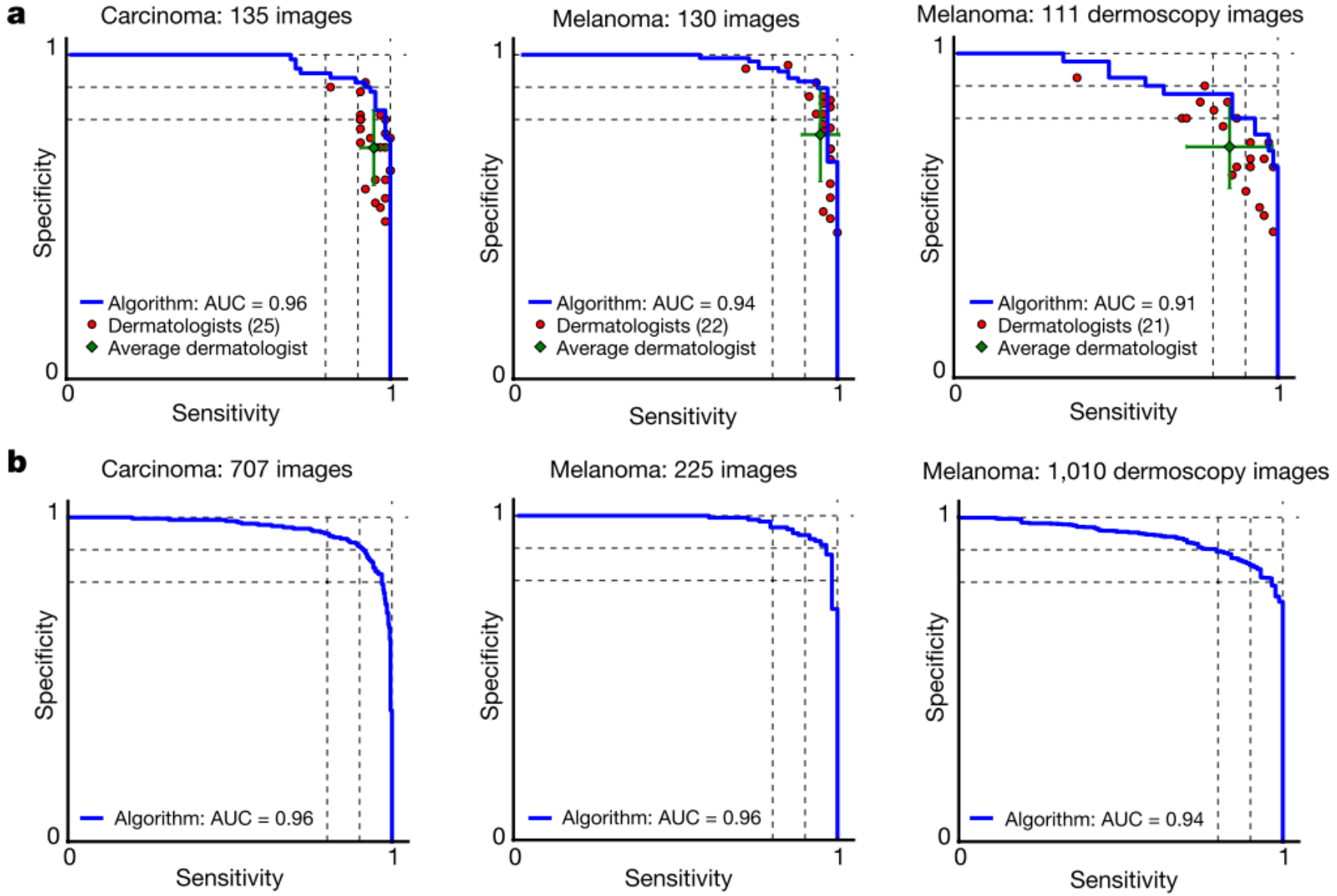
- … appreciate why this thematic is of increasing importance for the international community
- … understand some basic concepts of what explainability, transparency, understandable is
- … are aware of some methods of ex-AI
- … comprehend the complexity of context and the grand goal of context adaptive systems
- … see how important causal reasoning is
- … get a feeling on how important this field is for the medical domain for building trust, acceptance and reliability ….

# Agenda

- 00 Reflection – follow-up from last lecture
- 01 Towards explainable AI
- 02 Causal Reasoning
- 03 Tradeoff: Explainability vs. Accuracy
- 04 (Some) Current State of the Art Methods
- 05 Stochastic And-Or-Graphs


00 Reflection

# How do you explain this …




Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118

## Key Challenges

- Remember: Medicine is an complex application domain – dealing most of the time with **probable information!**
- Some challenges include:
- (a) defining hospital system architectures in terms of generic tasks such as diagnosis, therapy planning and monitoring to be executed for (b) medical reasoning in (a);
- (c) patient information management with (d) minimum uncertainty.
- Other challenges include: (e) knowledge acquisition and encoding, (f) human-computer interface and interaction; and (g) system integration into existing clinical legacy and proprietary environments, e.g. the enterprise hospital information system; to mention only a few.

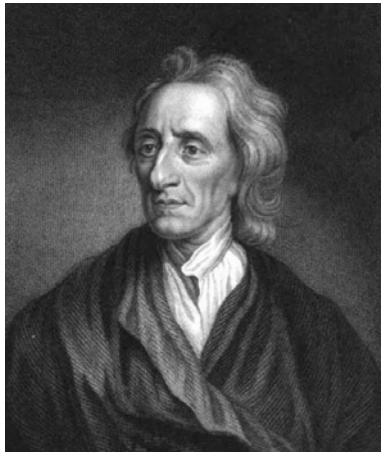
# 01 Towards Explainable AI

# Why explainability?

- 1) Wrong decisions can be very costly …
- 2) Insights from new perspectives …
- 3) Raising trust, acceptance, responsibility, …
- 4) Compliance to legislation …

## What is understanding ?

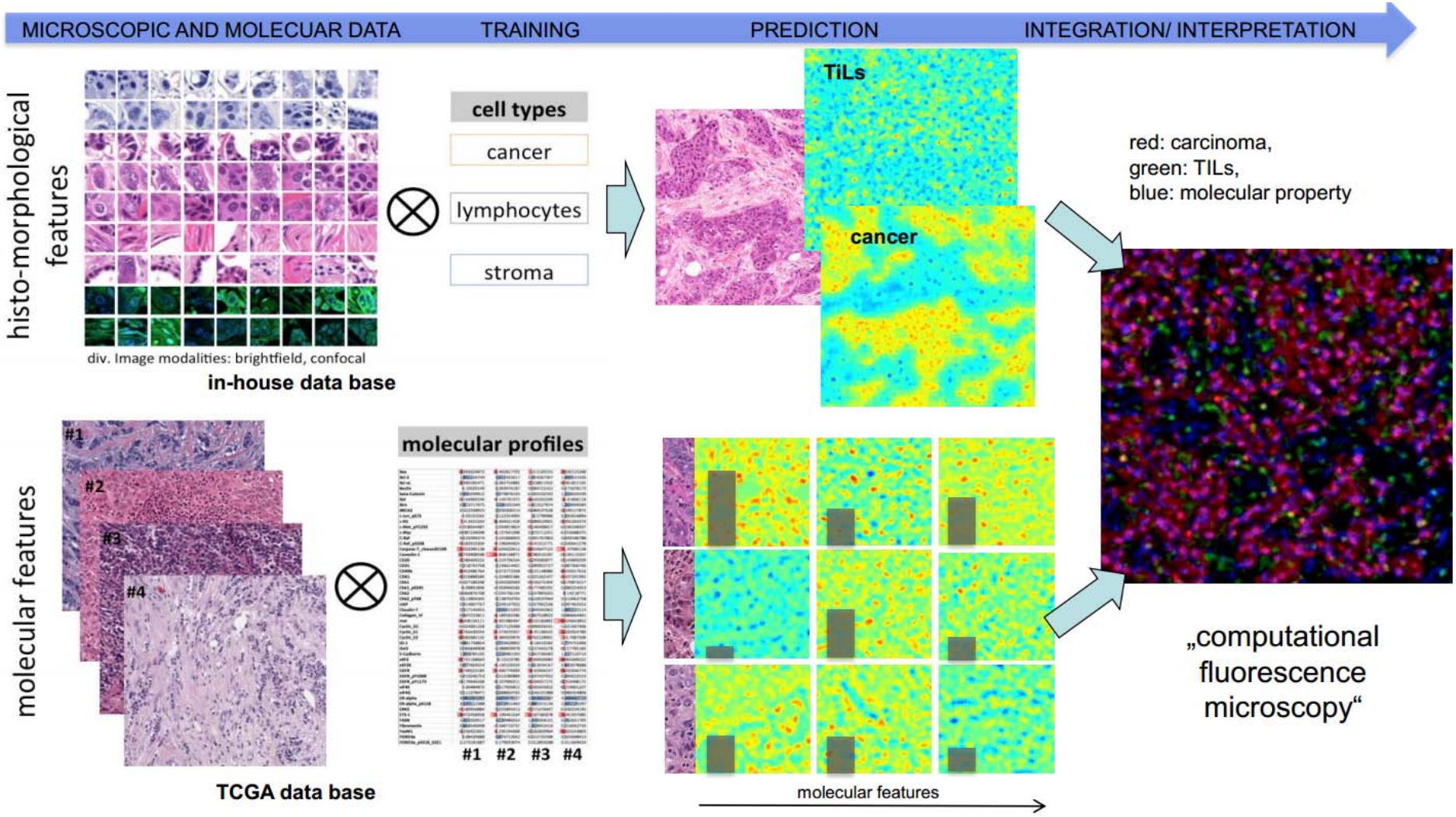
- Understanding … psychological concept correlated with the ability to make inferences.
- Mechanistic understanding
  - fMRI – which area is stimulated if one moves a finger
- Functional understanding $f : \mathbb{R}^d \rightarrow \mathbb{R}$
  - what is the relation between input and output
- Causal understanding
  - what is cause and effect (-> Judea Pearl)



John Locke 1841. An essay concerning human understanding, Teaside, Thomas TEGG.

# “It is not a human move … I’ve never seen such a move …”

# Integration and Interpretation




Alexander Binder, Michael Bockmayr, Miriam Hägele, Stephan Wienert, Daniel Heim, Katharina Hellweg, Albrecht Stenzinger, Laura Parlow, Jan Budczies & Benjamin Goeppert 2018. Towards computational fluorescence microscopy: Machine learning-based integrated prediction of morphological and molecular tumor profiles. arXiv:1805.11178vl.

# Context understanding is difficult

# Deep Learning Context recognition state-of-the-art

a woman riding a horse on a dirt road

an airplane is parked on the tarmac at an airport

a group of people standing on top of a beach

Andrej Karpathy & Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3128-3137.

Image Captions by deep learning : github.com/karpathy/neuraltalk2

Image Source: Gabriel Villena Fernandez; Agence France-Press, Dave Martin (left to right)

# 02 Causal Reasoning

# Example: Discovery of causal relationships from data …

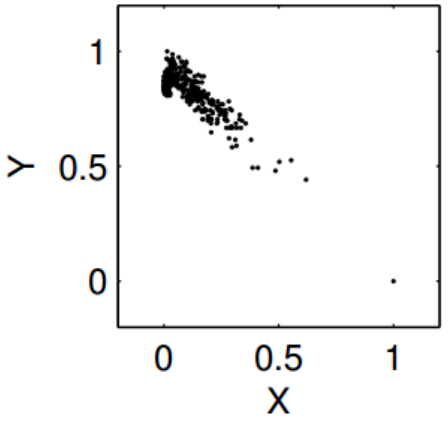


Hans Holbein d.J., 1533,
The Ambassadors,
London: National Gallery

Lopez-Paz, D., Muandet, K., Schölkopf, B. & Tolstikhin, I. 2015. Towards a learning theory of cause-effect inference. Proceedings of the 32nd International Conference on Machine Learning, JMLR, Lille, France.

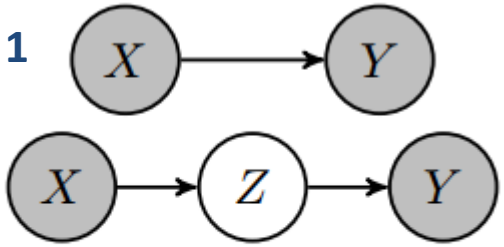
https://www.youtube.com/watch?v=9KiVNIUMmCc

# Decide if X → Y, or Y → X using only observed data



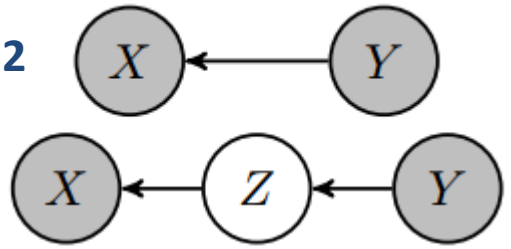

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf 2016. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17, (1), 1103-1204.

**1**




$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

**2**




$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

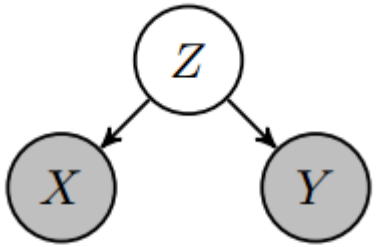
**3**


$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

**4**

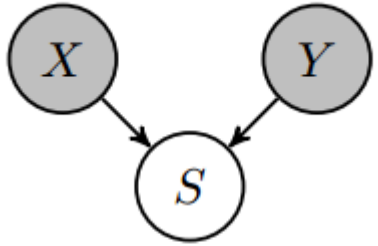

$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

**5**




$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

**6**




$$\mathbb{P}_{Y\,|\,s} \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x),s} = \mathbb{P}_{Y\,|\,x,s}$$
$$\mathbb{P}_{X\,|\,s} \neq \mathbb{P}_{X\,|\,\mathrm{do}(y),s} = \mathbb{P}_{X\,|\,y,s}$$

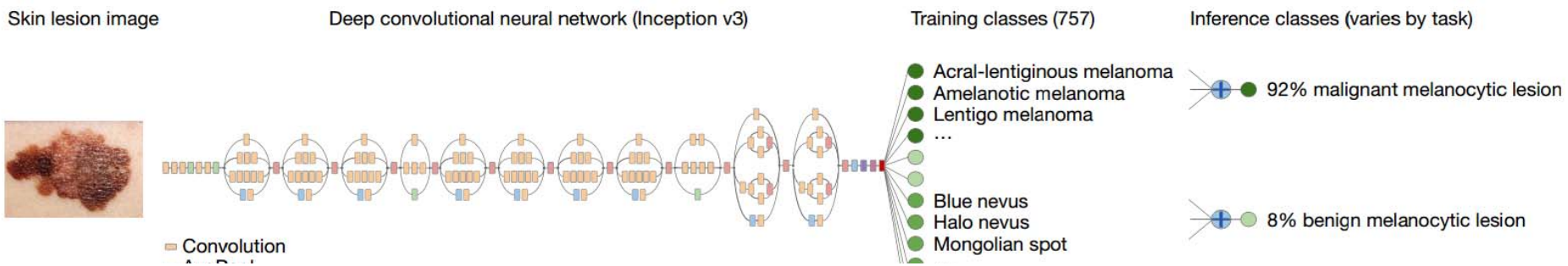
- **"How do humans generalize from few examples?"**
  - Learning relevant representations
  - Disentangling the explanatory factors
  - Finding the shared underlying explanatory factors, in particular between $P(x)$ and $P(Y|X)$, with a causal link between $Y \rightarrow X$

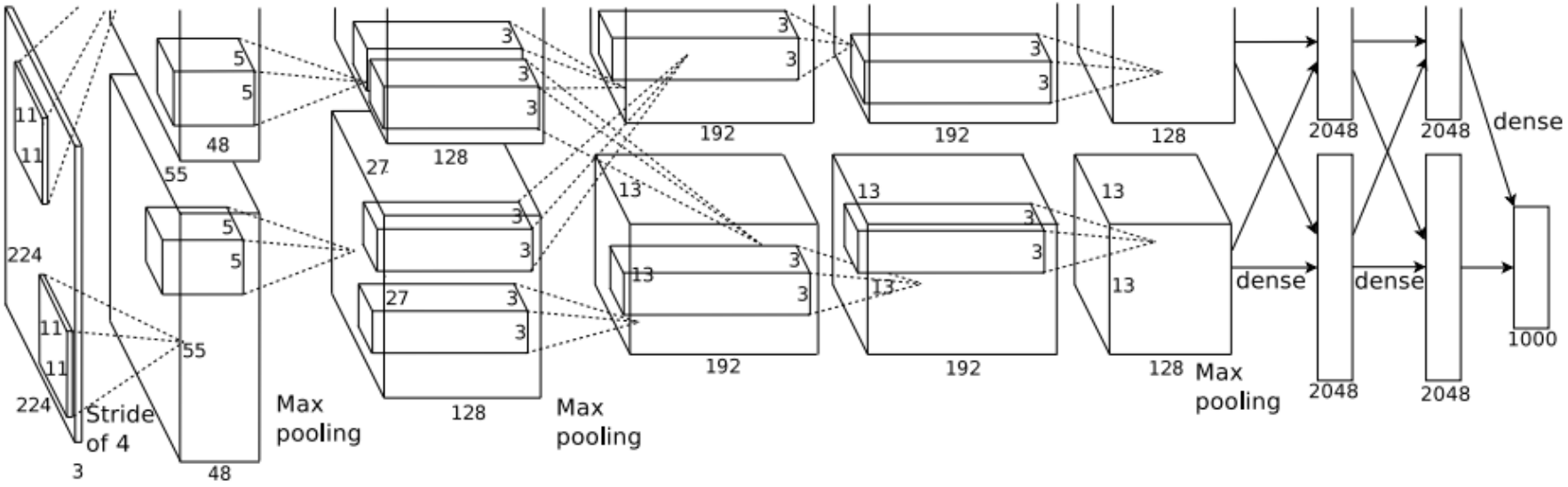
Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.

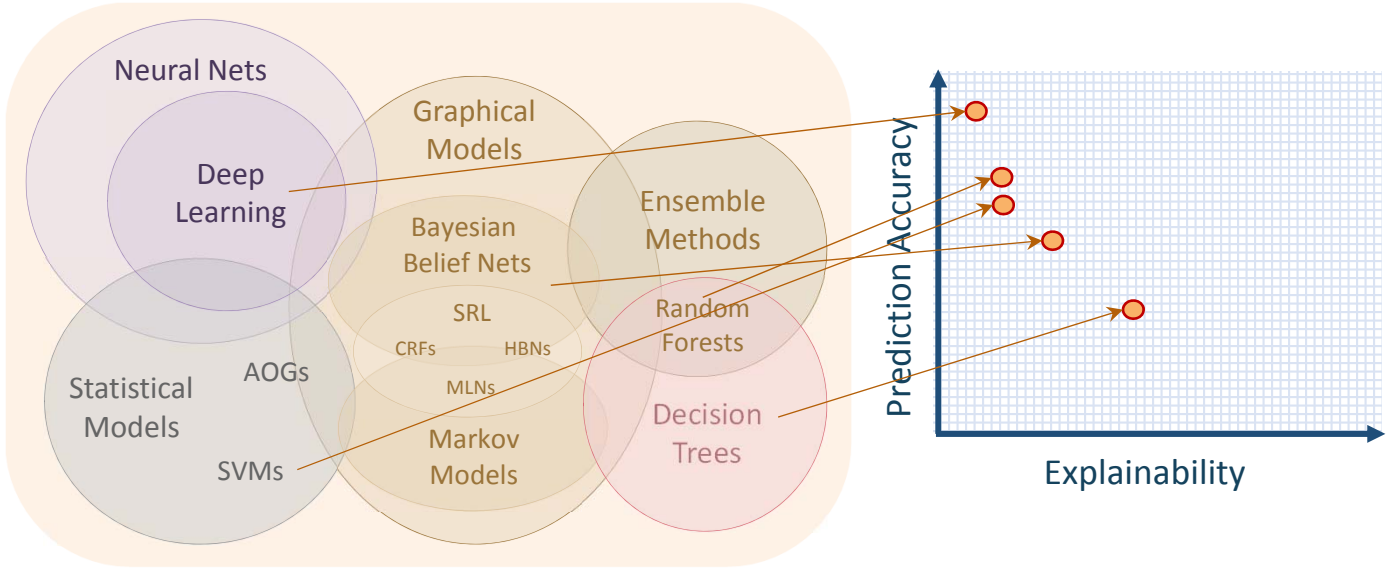
# 03 Tradeoff: Explainability vs. Accuracy

# Deep Convolutional Neural Network Pipeline

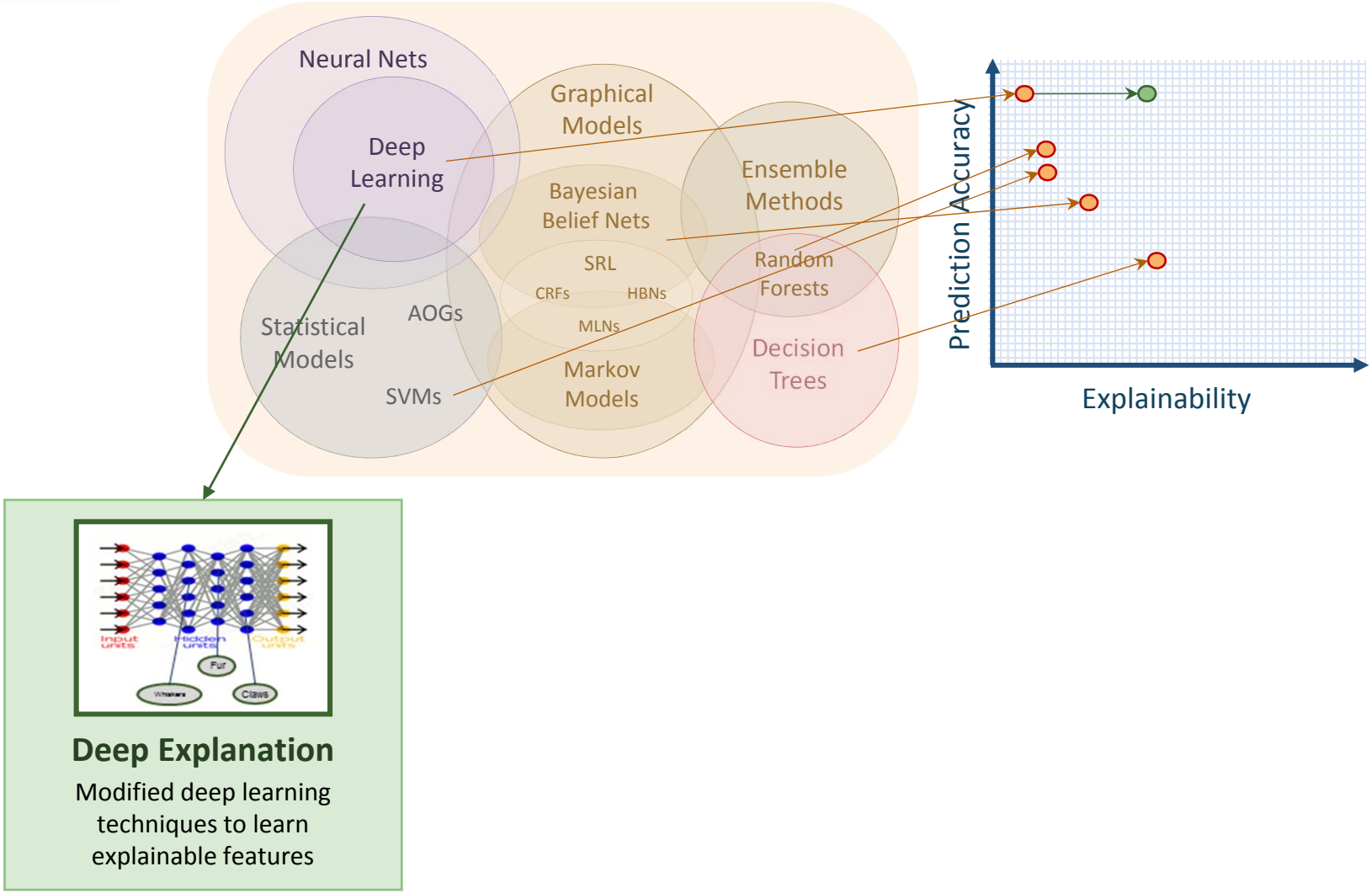


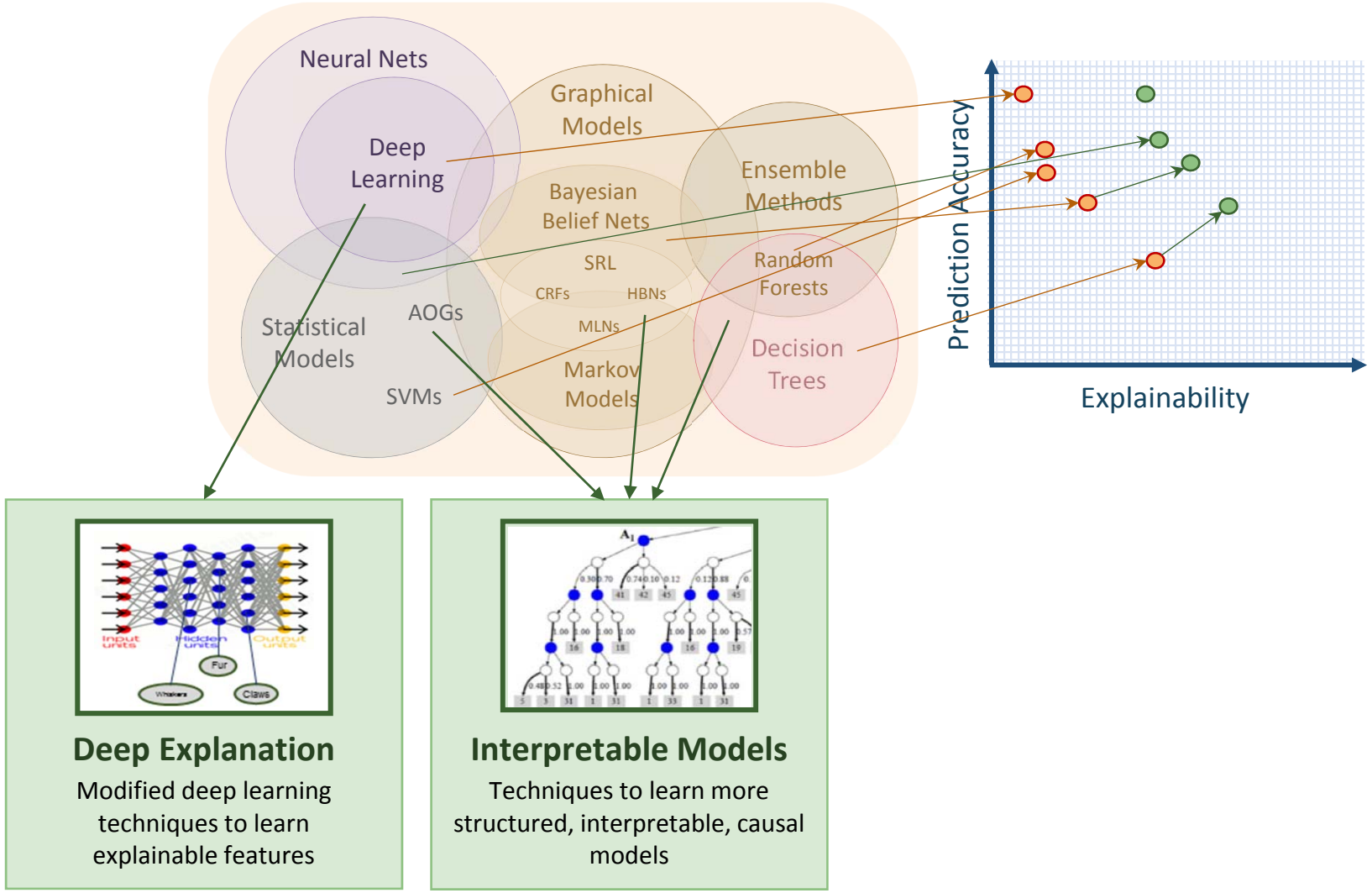

Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.



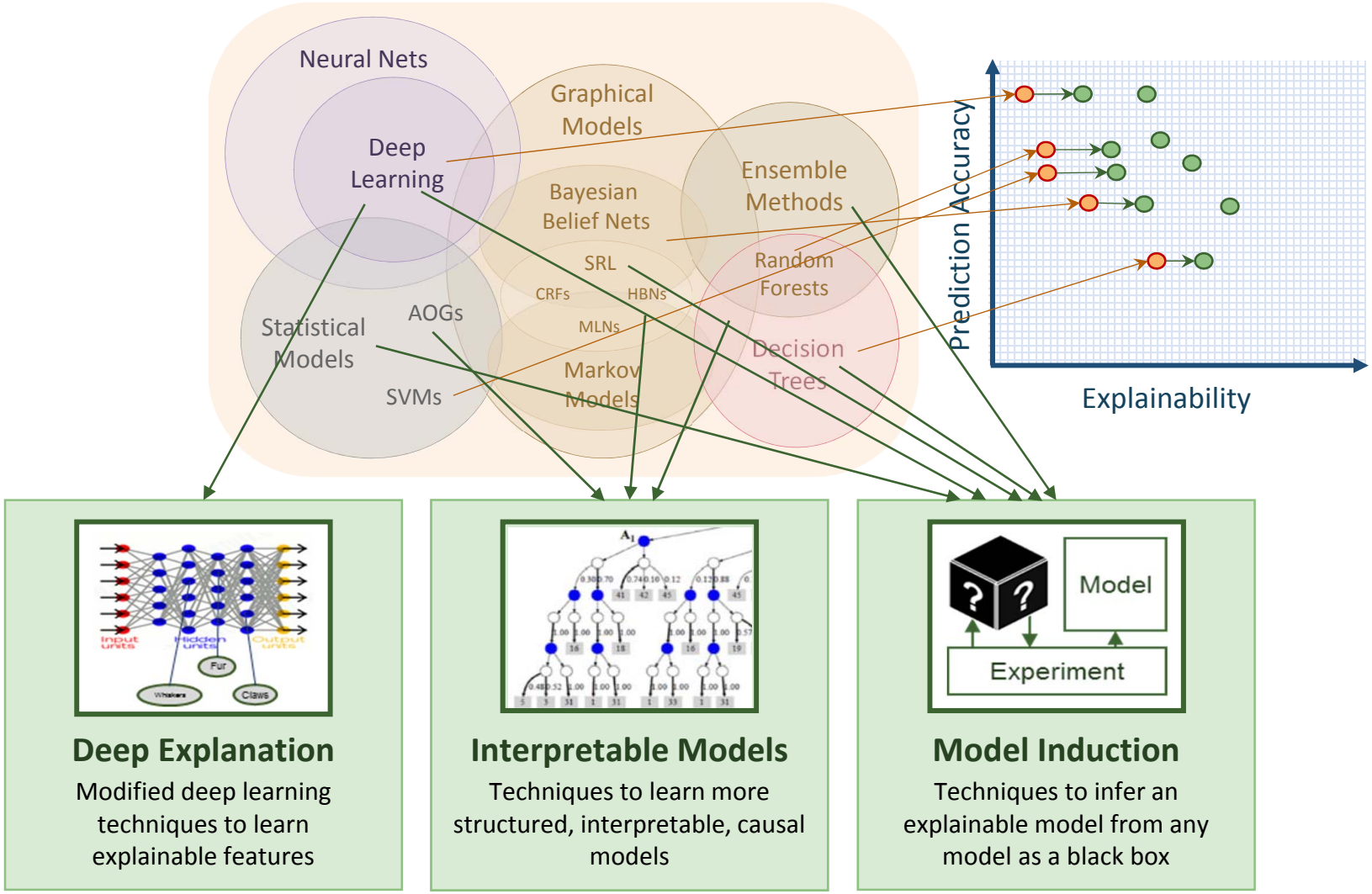

David Gunning 2016. Explainable artificial intelligence (XAI): Technical Report Defense Advanced Research Projects Agency DARPA-BAA-16-53, Arlington, USA, DARPA.




David Gunning 2016. Explainable artificial intelligence (XAI): Technical Report Defense Advanced Research Projects Agency DARPA-BAA-16-53, Arlington, USA, DARPA.




David Gunning 2016. Explainable artificial intelligence (XAI): Technical Report Defense Advanced Research Projects Agency DARPA-BAA-16-53, Arlington, USA, DARPA.




David Gunning 2016. Explainable artificial intelligence (XAI): Technical Report Defense Advanced Research Projects Agency DARPA-BAA-16-53, Arlington, USA, DARPA.

# Post-hoc vs. Ante-hoc

Post-hoc: Select a model and develop a technique to make it transparent

$f(x) = \text{DeepNet}(x)$

Ante-hoc: Select a model that is already transparent and optimize it

$$f(x) = \sum_{i=1}^{d} g_i(x_i)$$

contribution of $i$th variable

Different dimensions of "interpretability"

prediction

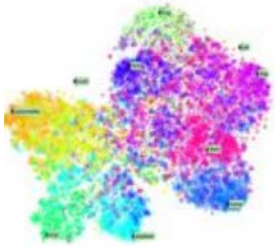
*"Explain why a certain pattern x has been classified in a certain way f(x)."*

model

*"What would a pattern belonging to a certain category typically look like according to the model."*

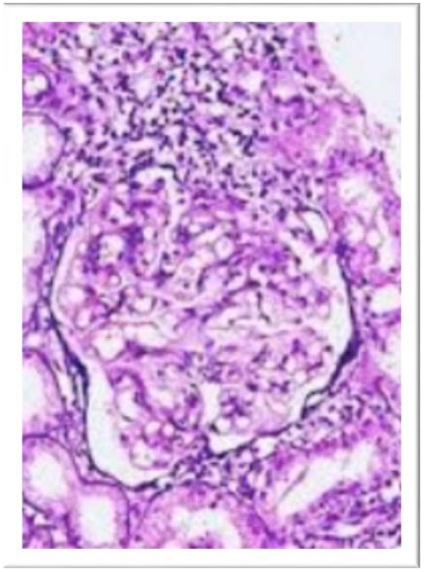
data

*"Which dimensions of the data are most relevant for the task."*



Montavon, G., Samek, W. & Müller, K.-R. 2017. Methods for interpreting and understanding deep neural networks. arXiv:1706.07979.

# Model/Data improvement: Reducing general errors



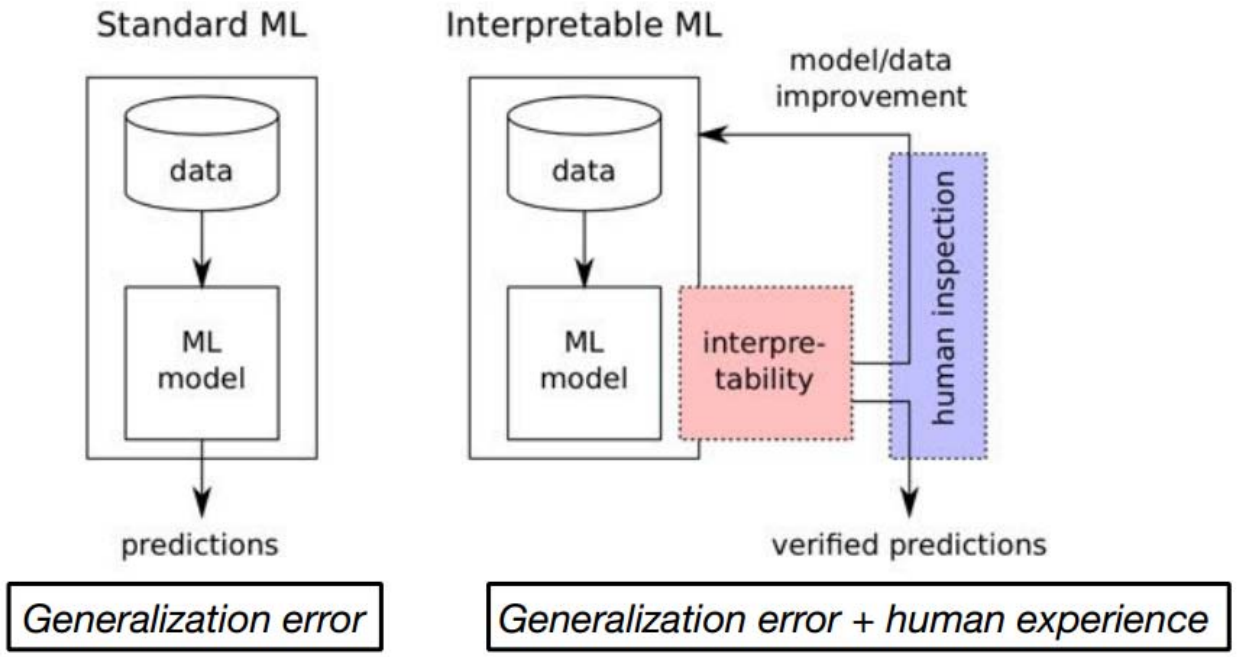


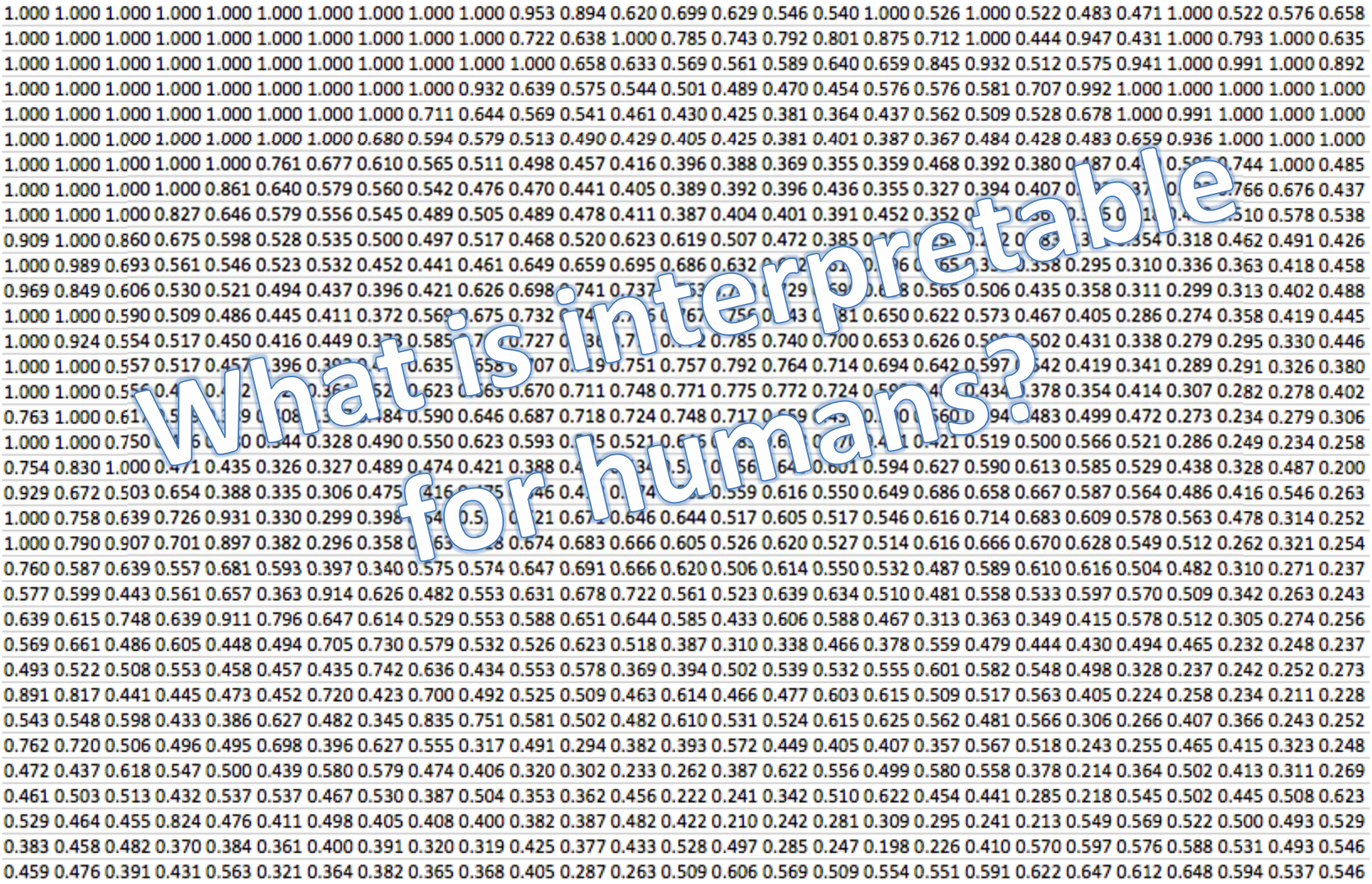

Image credit to: Samek, Montavon & Müller Tutorial at ICASSEP 2017

- Wrong decisions can be costly and dangerous!
- Verify that classifier works as expected
- Improve classifier continuously
- Human learning inspired by machine learning

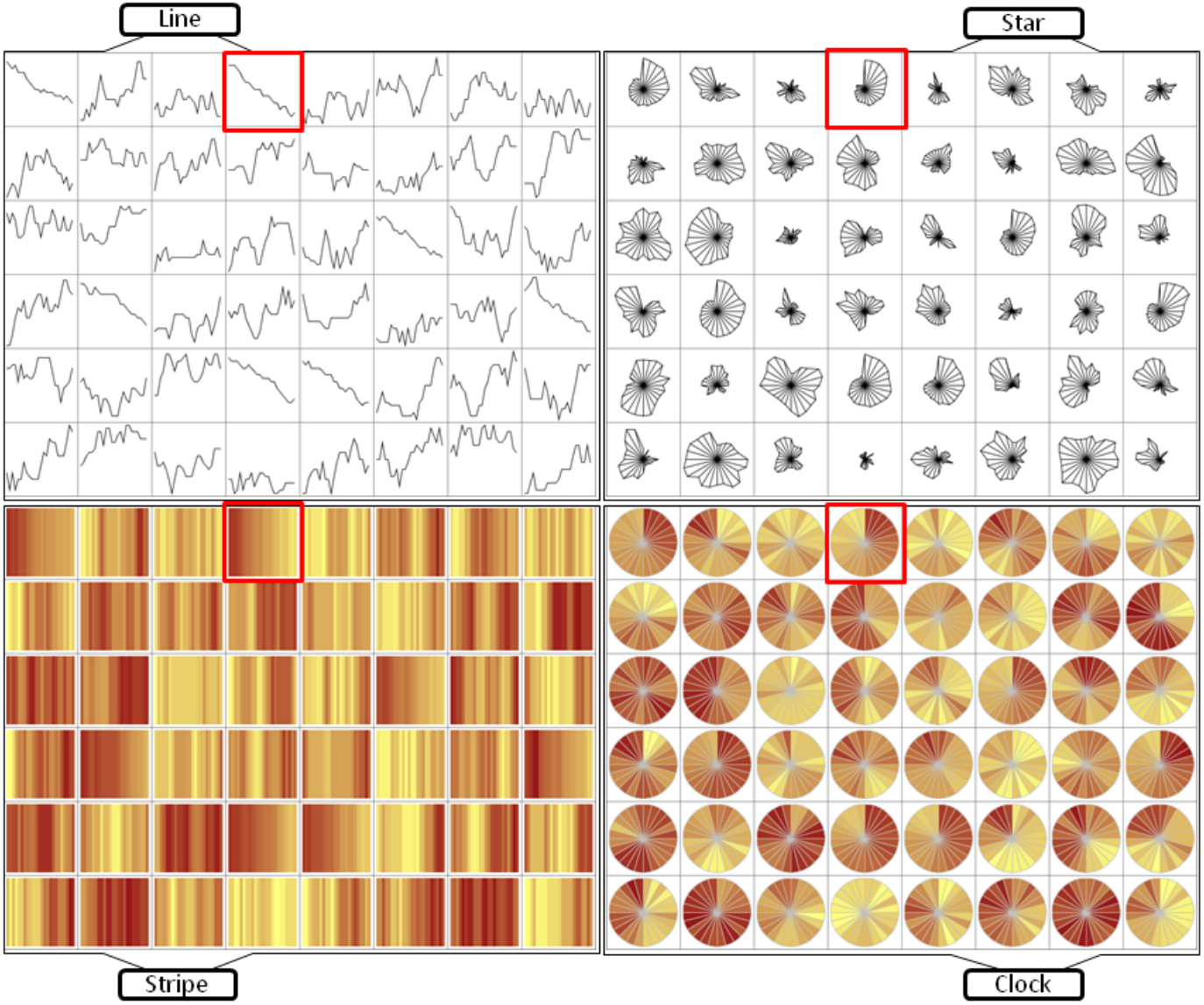
## Research and Development in Interpretability …

- Interpretability as a novel kind for supporting teaching, learning and knowledge discovery,
- Particularly in abstract fields (informatics)
- Compliance to European Law “the right of explanation”
- Check for bias in machine learning results
- Fostering trust, acceptance, making clear the reliability … “can you trust your results”

Katharina Holzinger, Klaus Mak, Peter Kieseberg & Andreas Holzinger 2018. Can we trust Machine Learning Results? Artificial Intelligence in Safety-Critical decision Support. ERCIM News, 112, (1), 42-43.


What is interpretable
for humans?

# What is understandable, interpretable, intelligible?



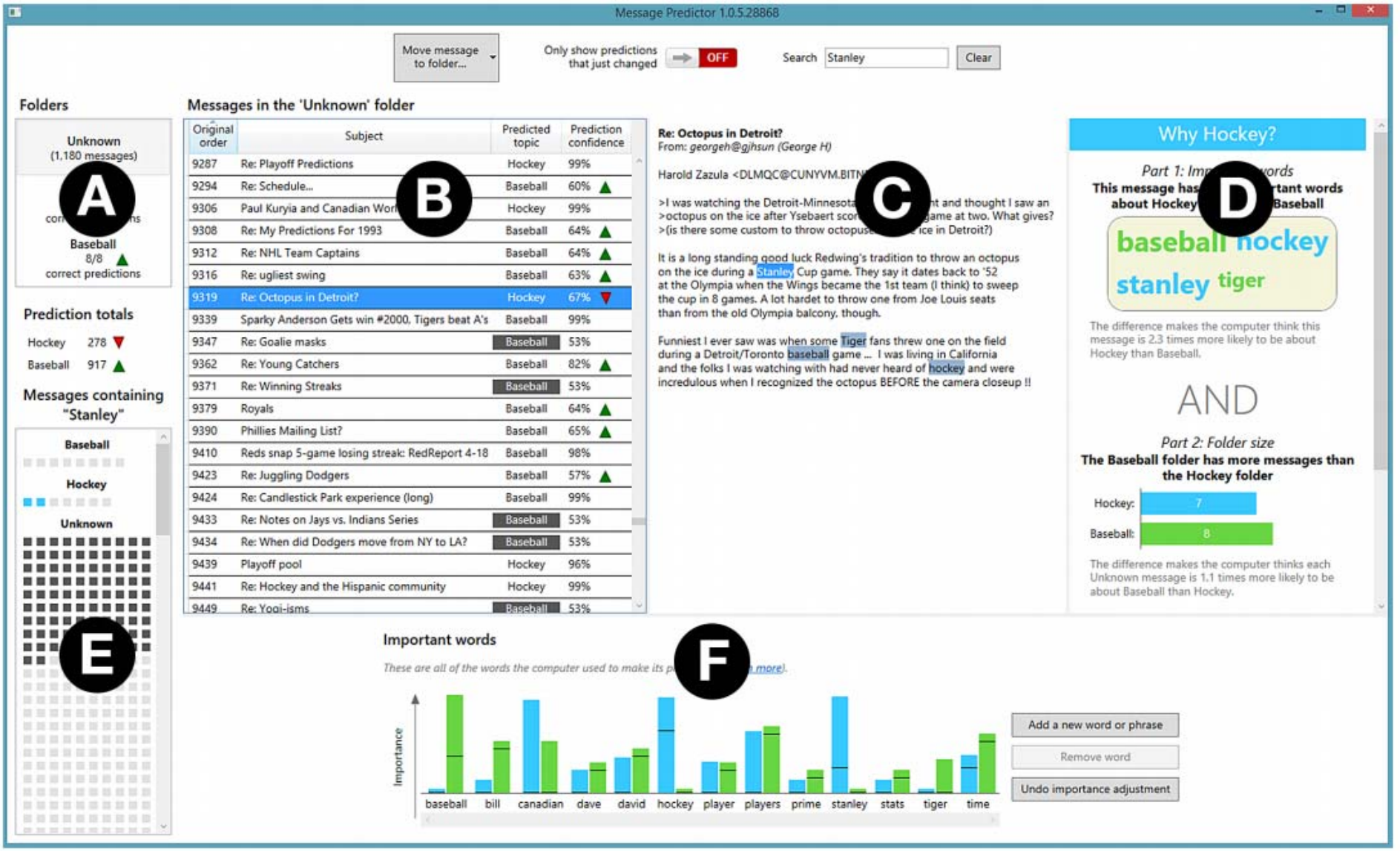

https://www.vis.uni-konstanz.de/en/members/fuchs/

# Explainable AI is a huge challenge for visualization


Michael Hund, Dominic Boehm, Werner Sturm, Michael Sedlmair, Tobias Schreck, Torsten Ullrich, Daniel A. Keim, Ljiljana Majnaric & Andreas Holzinger 2016. Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. Brain Informatics, 3, (4), 233-247, doi:10.1007/s40708-016-0043-5.

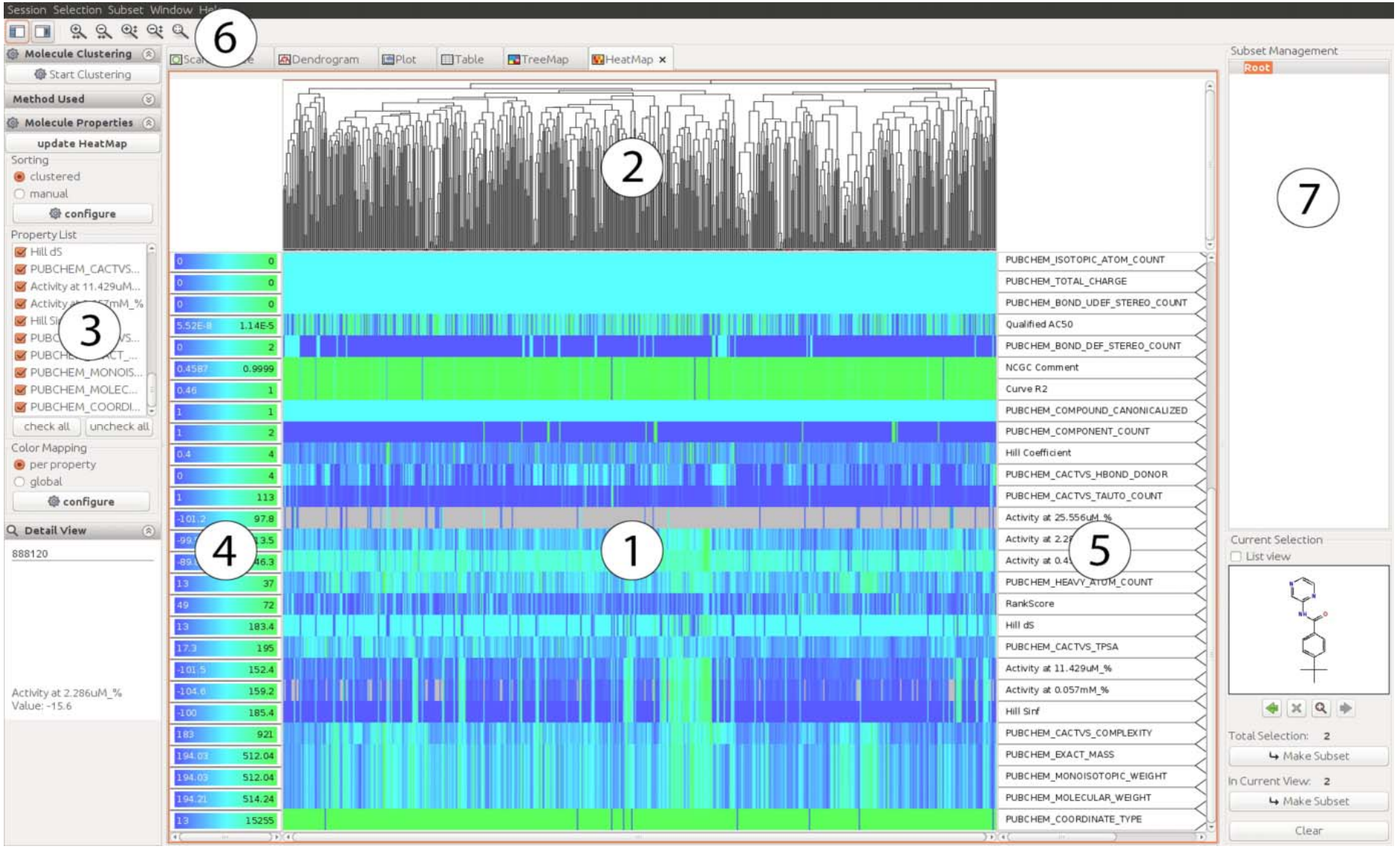
# 04 Current State of the Art

# Example for an Explanation Interface



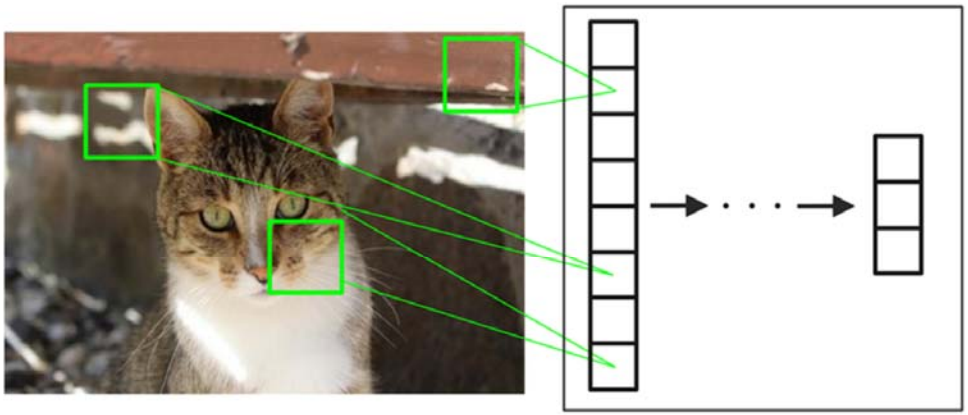

Todd Kulesza, Margaret Burnett, Weng-Keen Wong & Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI 2015), 2015 Atlanta. ACM, 126-137, doi:10.1145/2678025.2701399.

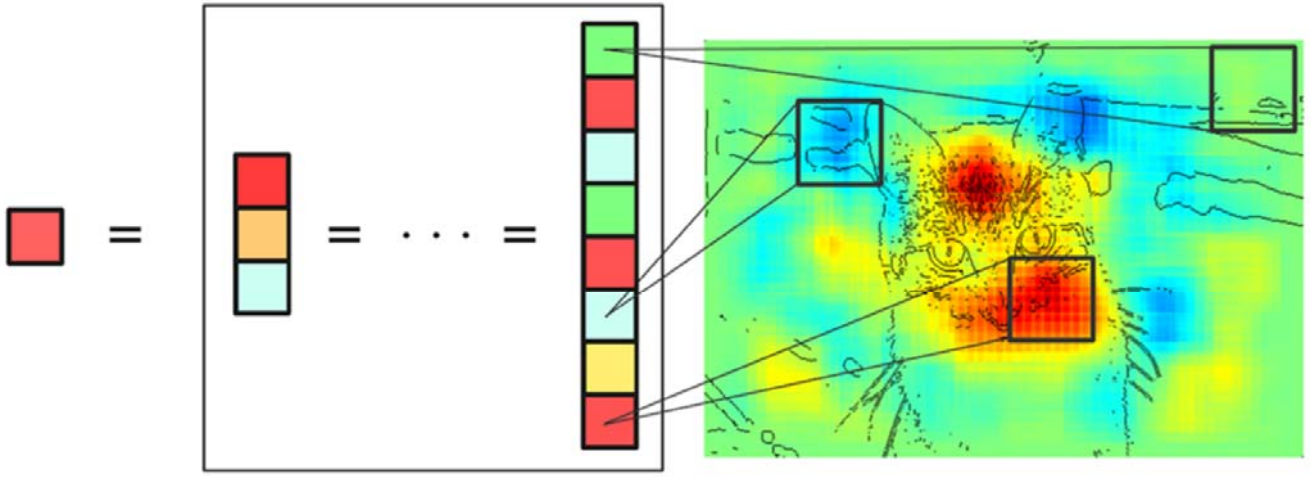
# Example for an Explanation Interface - open work ☺




Werner Sturm, Till Schaefer, Tobias Schreck, Andeas Holzinger & Torsten Ullrich. Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps In: Borgo, Rita & Turkay, Cagatay, eds. EG UK Computer Graphics & Visual Computing CGVC 2015, 2015 University College London (UCL). Euro Graphics (EG), 77-84, doi:10.2312/cgvc.20151247.

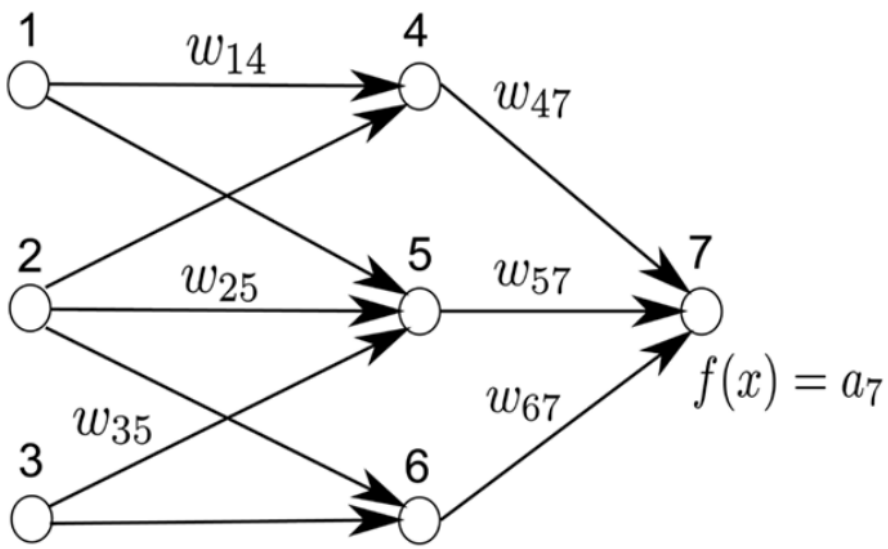
# LRP Layer-Wise Relevance Propagation

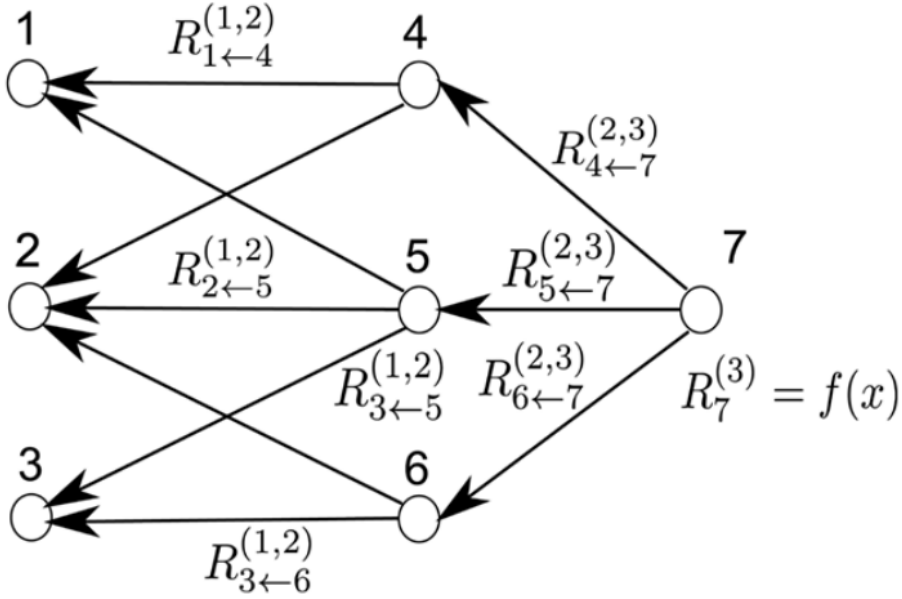



Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

# A NN-classifier during prediction time

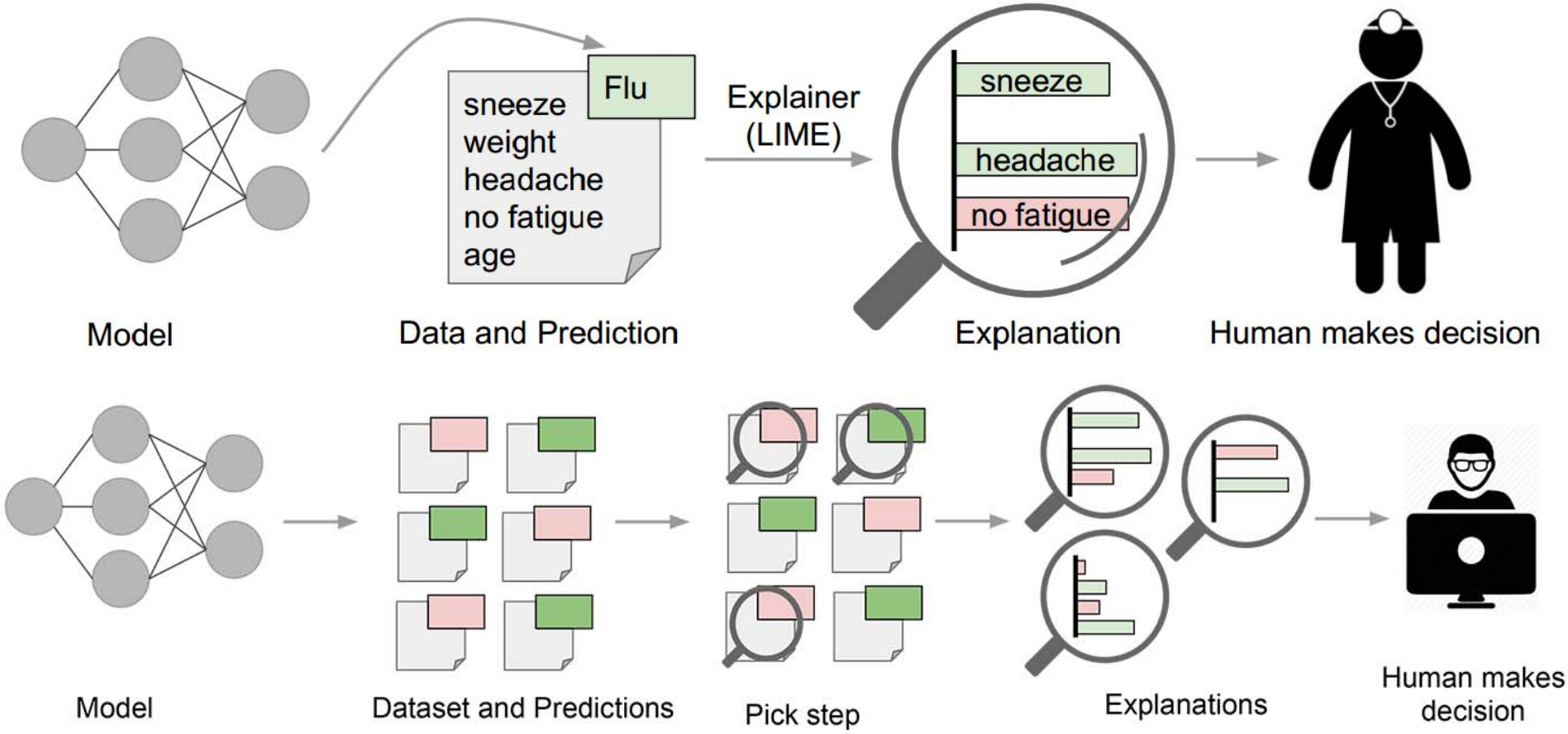



$$f(x) = \cdots = \sum_{d\in l+1} R_d^{(l+1)} = \sum_{d\in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)}$$

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.

# LIME – Local Interpretable Model Agnostic Explanations




Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. ACM, 1135-1144, doi:10.1145/2939672.2939778.

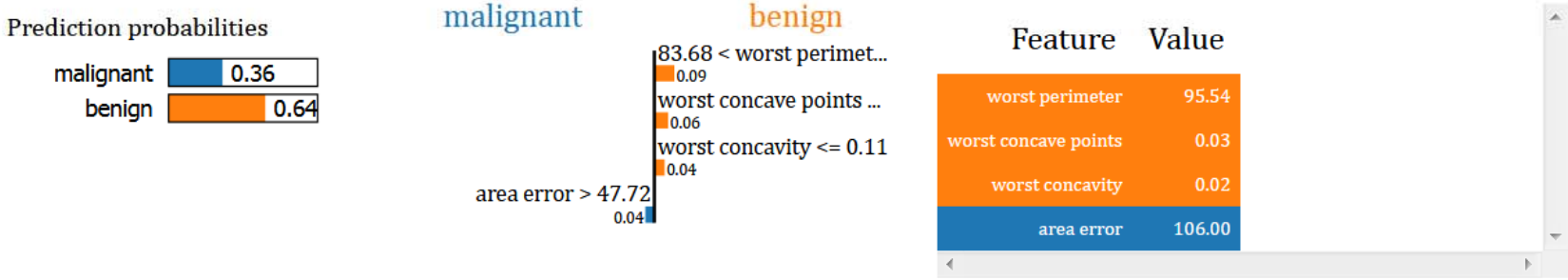
# Example LIME – Model Agnostic Explanation

In [12]:
```
explainer = lime.lime_tabular.LimeTabularExplainer(X_train, feature_names=breast.feature_names, class_names=breast.targe
```

Here we will take a sample from the test set (in this case the sample at index 76) and create an explainer instance for this sample. This will let us see why the algorithm made its prediction visually.

In [18]:
```
# For this demonstration, let's take the same sample each time, in this case sample index 86
i = 76
# For a random sample uncomment out the following line
# i = np.random.randint(0, X_test.shape[0])

exp = explainer.explain_instance(X_test[i], random_forest.predict_proba, num_features=4)
exp.show_in_notebook(show_table=True, show_all=False)
```

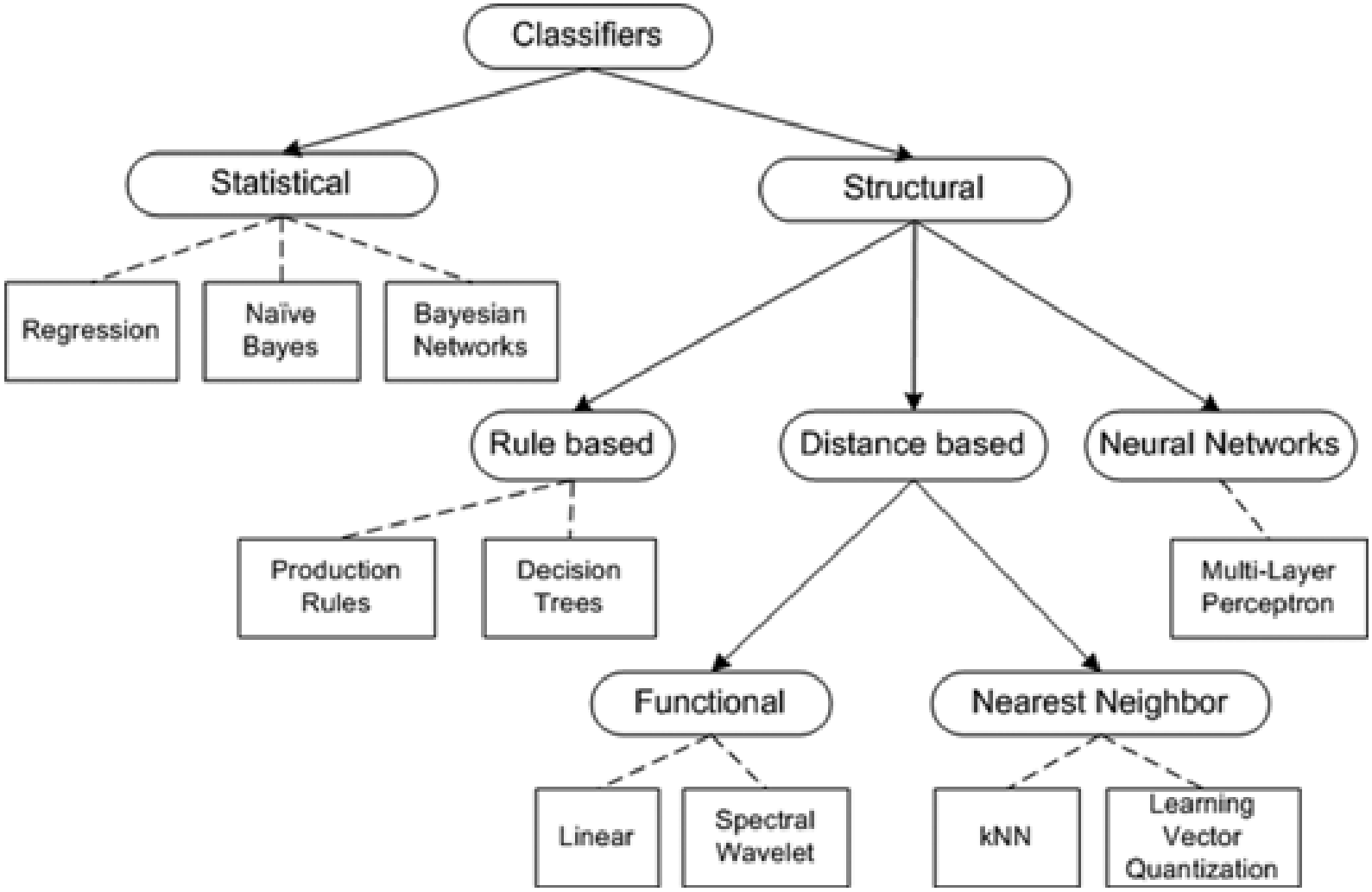



As you can see, the random forest algorithm has predicted with a probability of 0.64 that the sample at index 76 in the test set is malignant.

When using the explainer, we set the `num_features`parameter to 4, meaning the explainer shows the top 4 features that contributed to the prediction probabilities.

We chose 76 as it was a borderline decision. For example sample 86 is much more clear (this will we will set the `num_features` parameter to include all features so that we see each feature's contribution to the probability):

# Remember: there are myriads of classifiers …




https://stats.stackexchange.com/questions/271247/machine-learning-statistical-vs-structural-classifiers

# Black Box Explanations through Transparent Approximations

**If** Age $<50$ **and** Male =Yes:

**If** Past-Depression =Yes **and** Insomnia =No **and** Melancholy =No, **then** Healthy

**If** Past-Depression =Yes **and** Insomnia =Yes **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

**If** Age $\geq 50$ **and** Male =No:

**If** Family-Depression =Yes **and** Insomnia =No **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

**If** Family-Depression =No **and** Insomnia =No **and** Melancholy =No **and** Tiredness =No, **then** Healthy

**Default**:

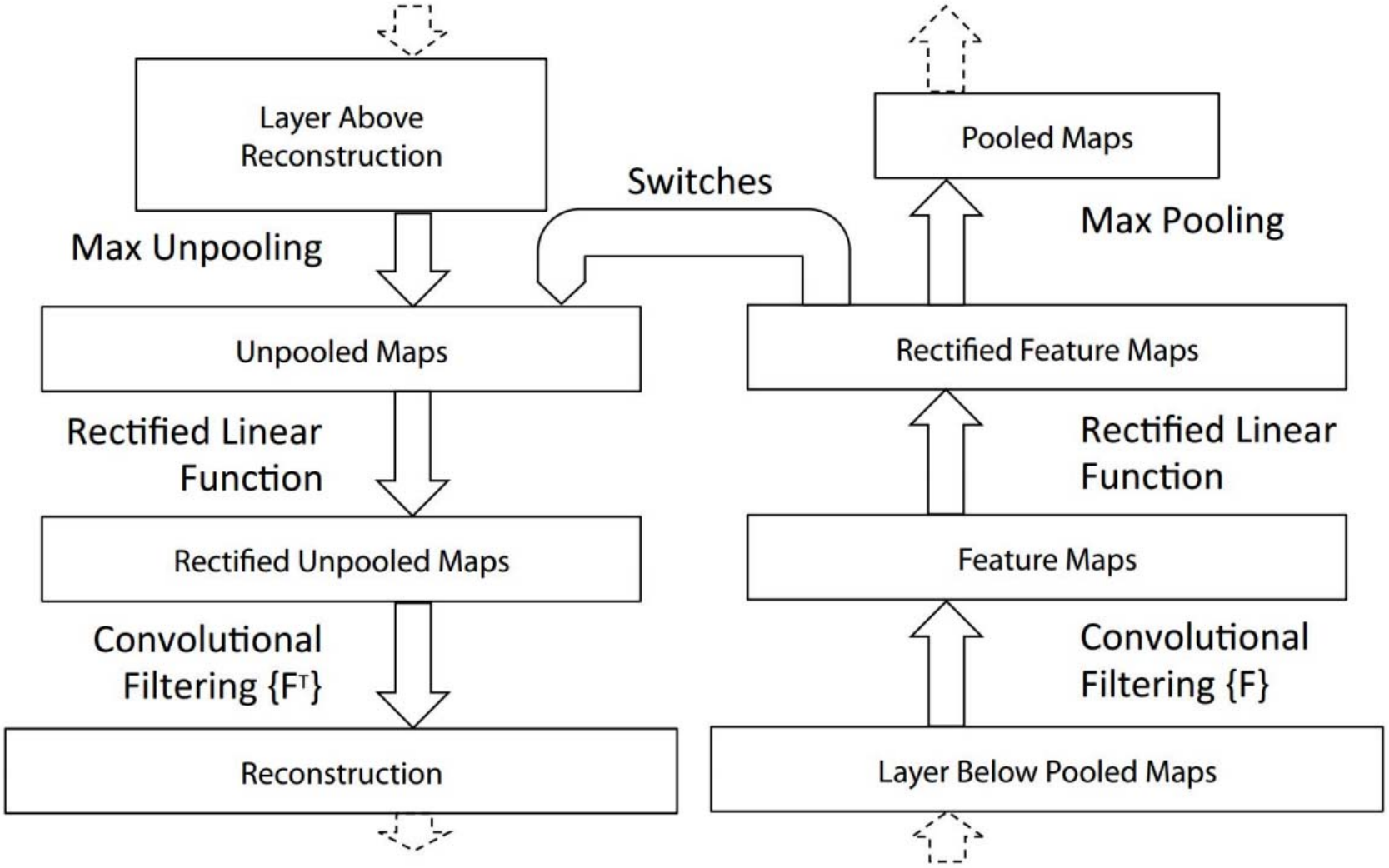
**If** Past-Depression =Yes **and** Tiredness =No **and** Exercise =No **and** Insomnia =Yes, **then** Depression

**If** Past-Depression =No **and** Weight-Gain =Yes **and** Tiredness =Yes **and** Melancholy =Yes, **then** Depression

**If** Family-Depression =Yes **and** Insomnia =Yes **and** Melancholy =Yes **and** Tiredness =Yes, **then** Depression

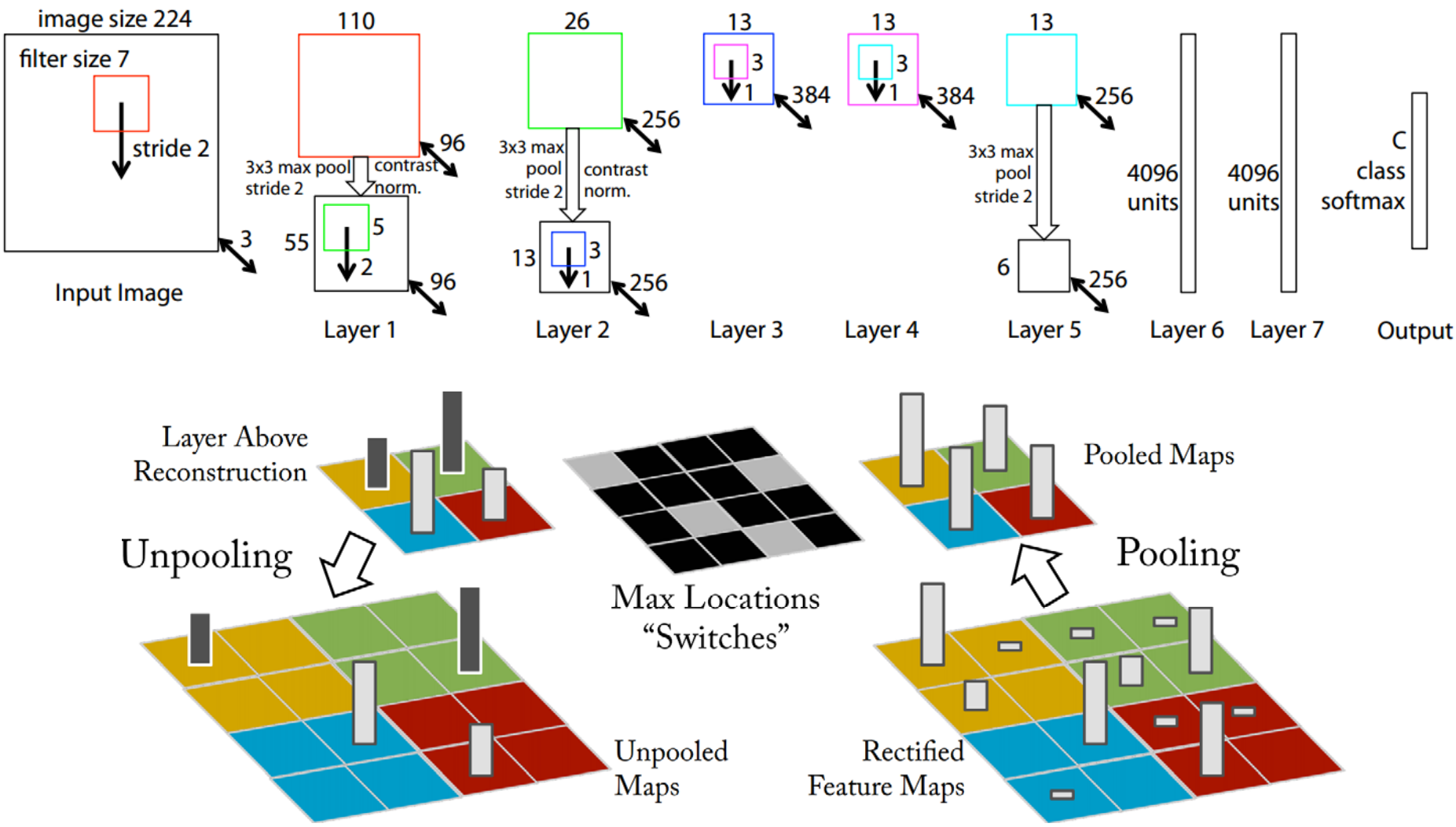
Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Jure Leskovec 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.

# Example: Interpretable Deep Learning Model

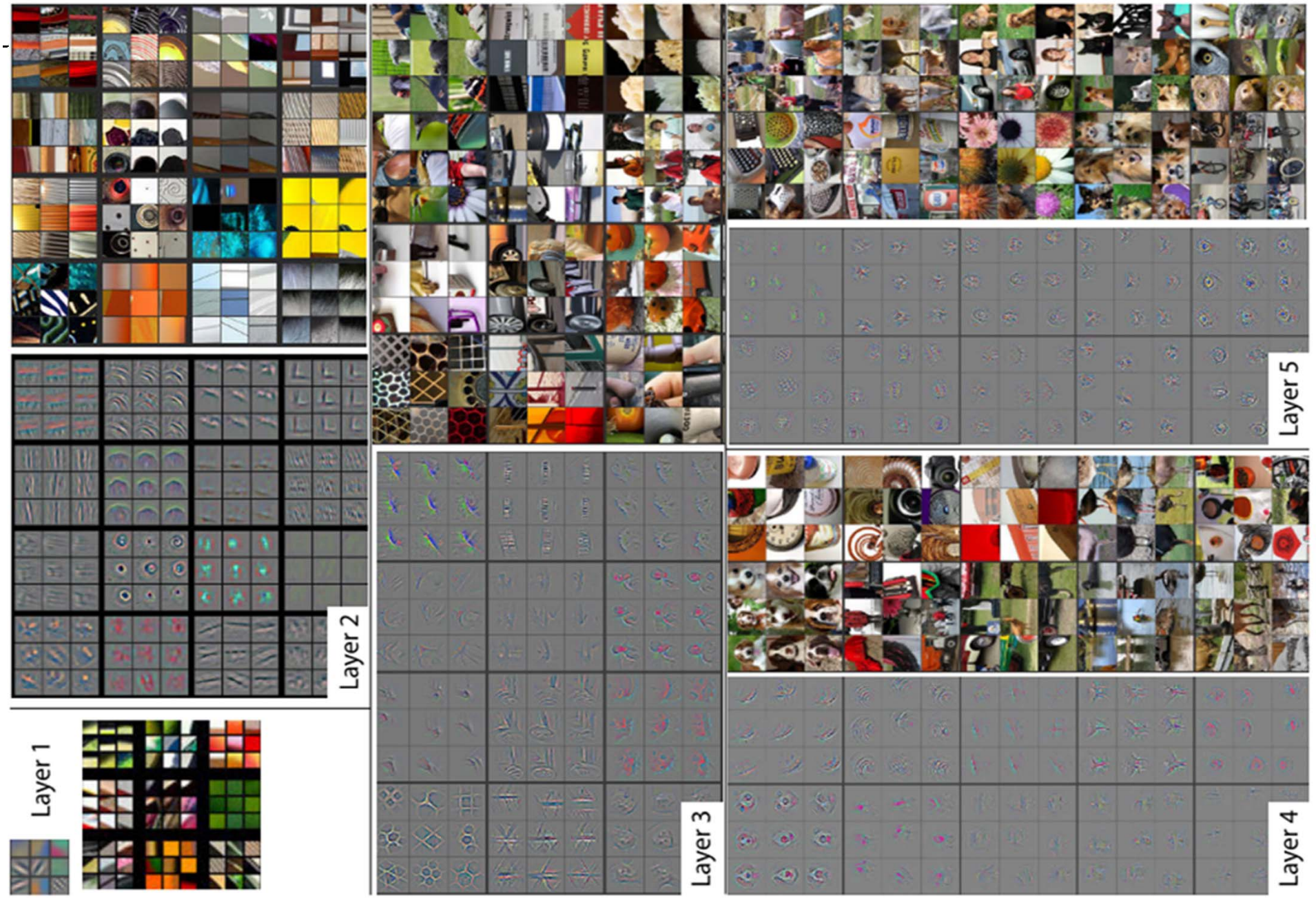



Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

# Visualizing a Conv Net with a De-Conv Net

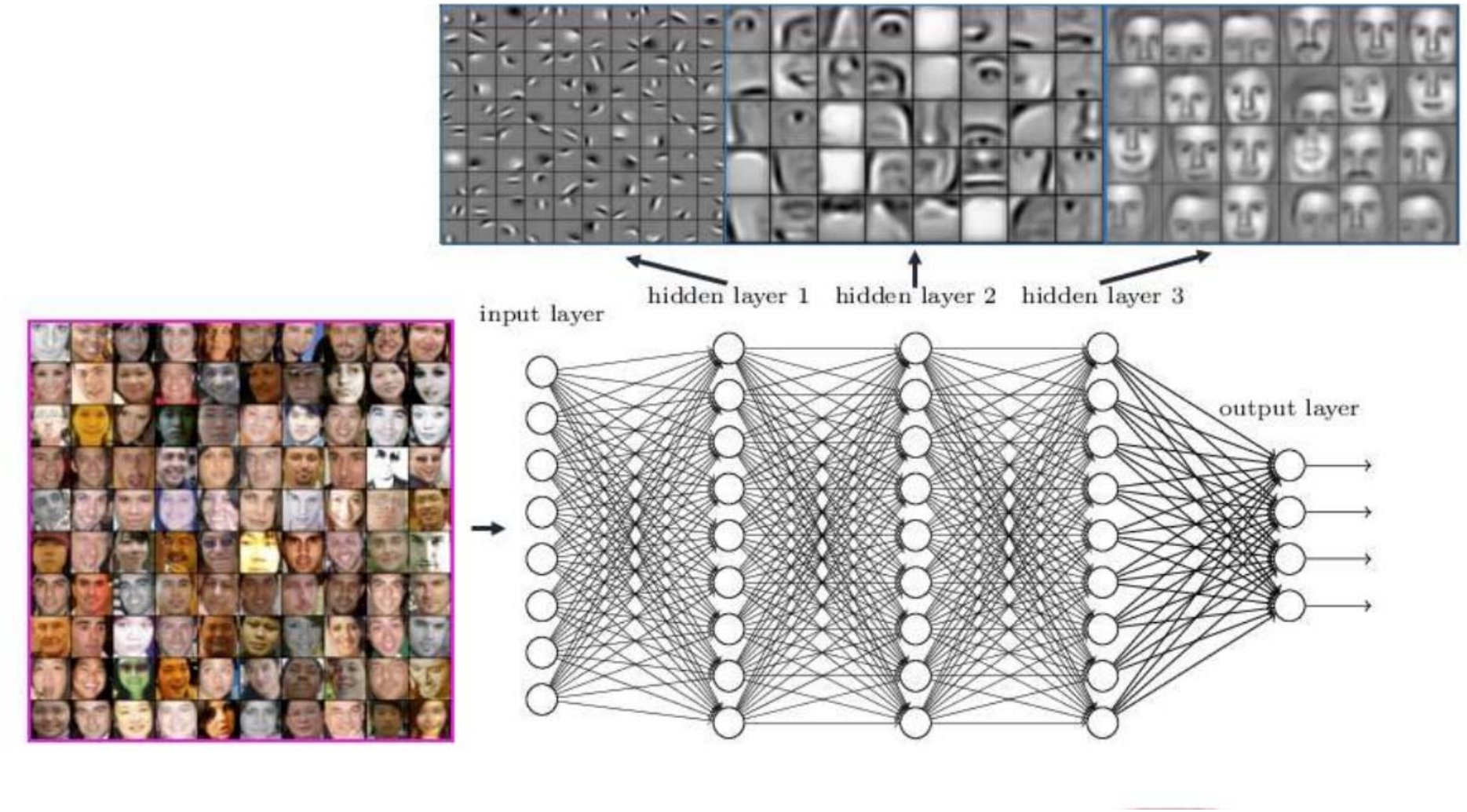



Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.



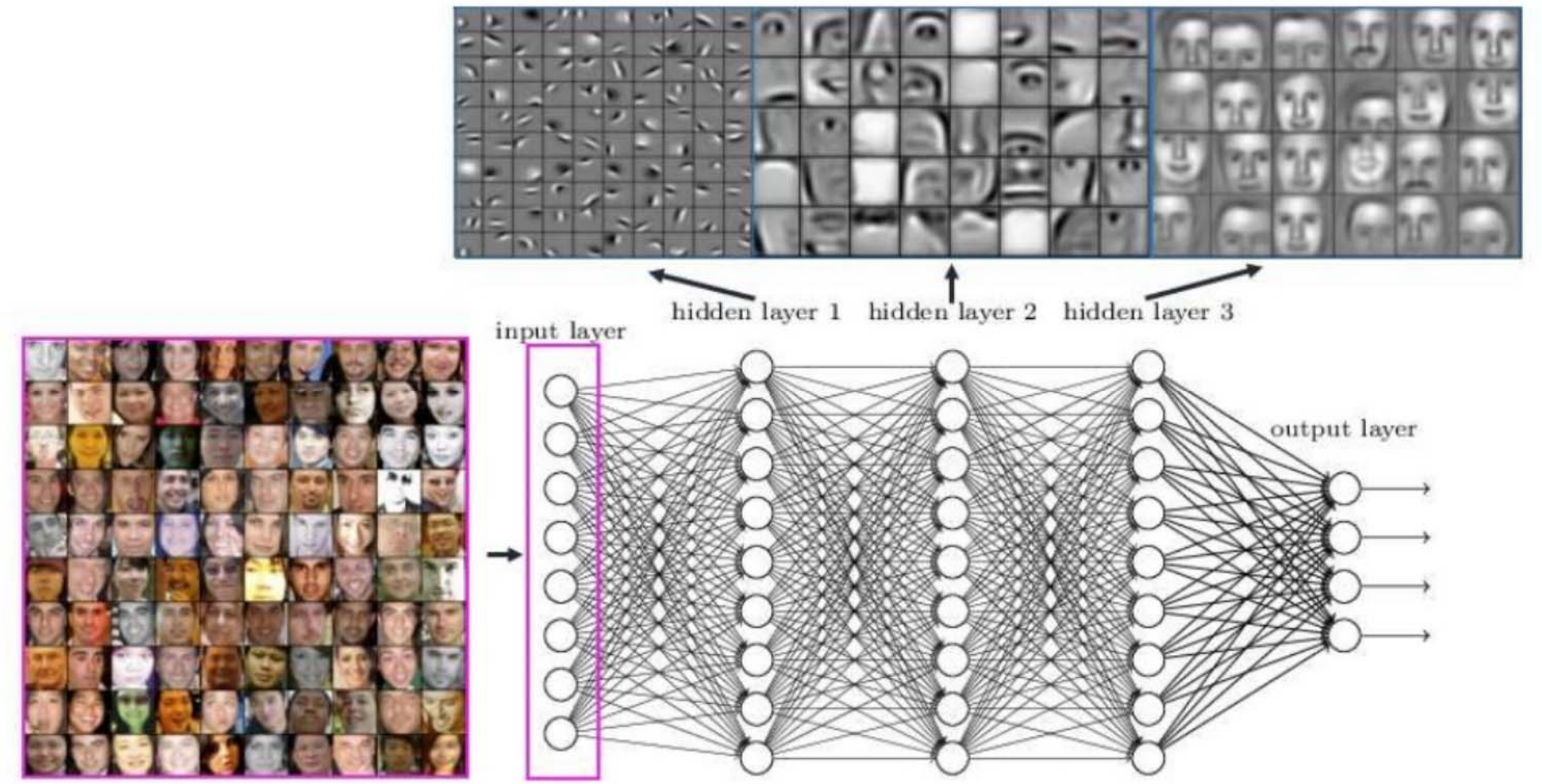

Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

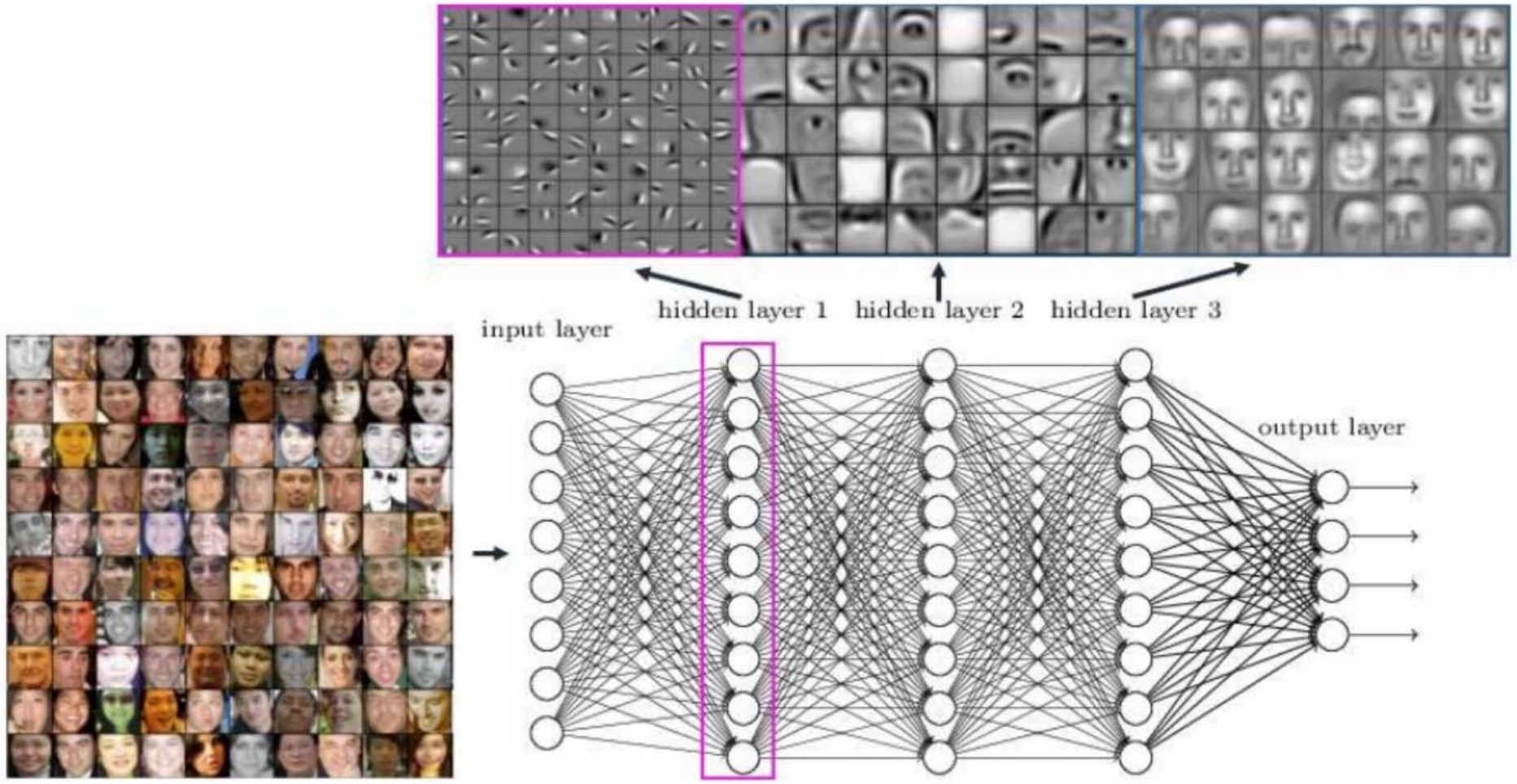
# The world is compositional (Yann LeCun)

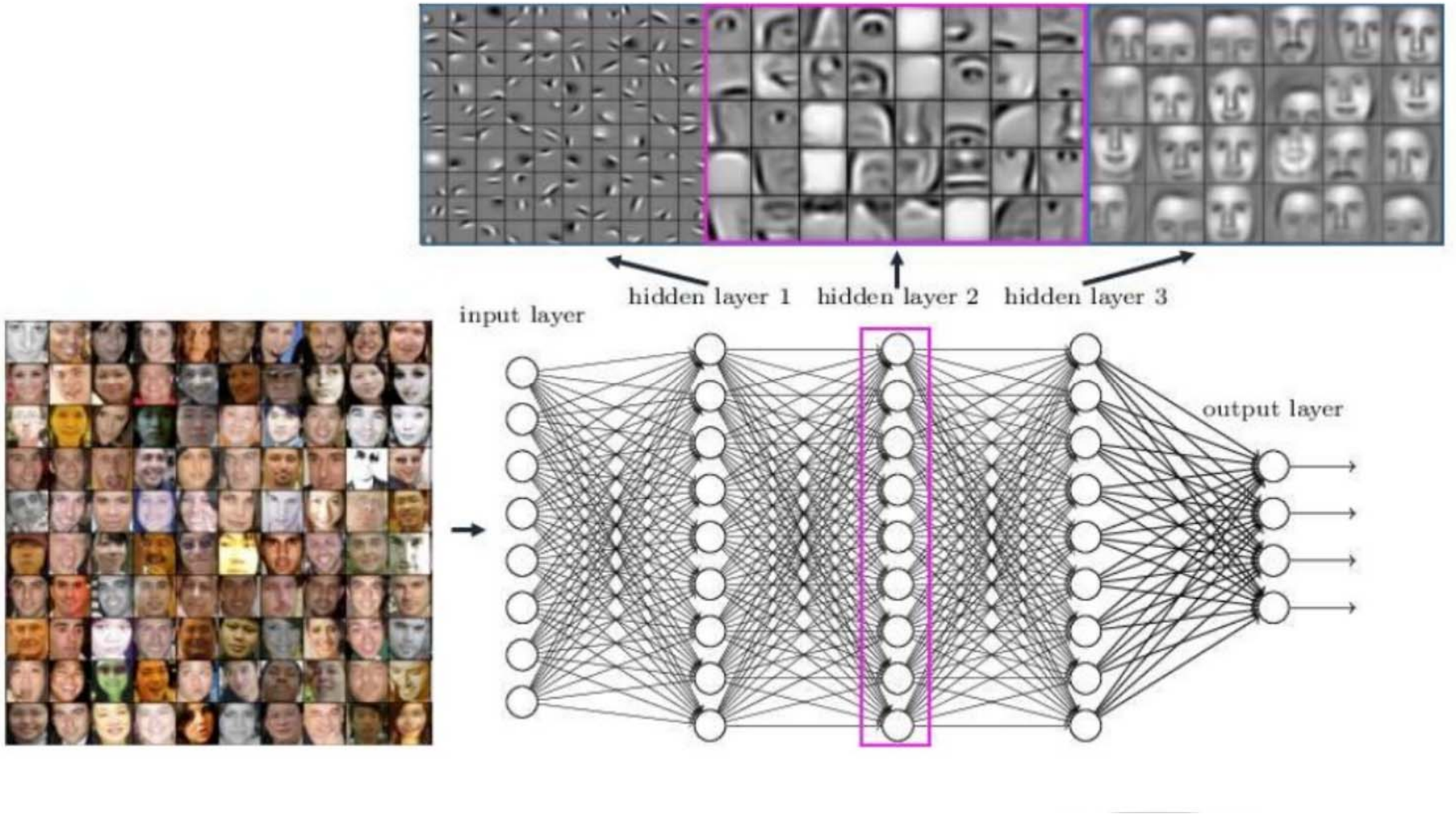


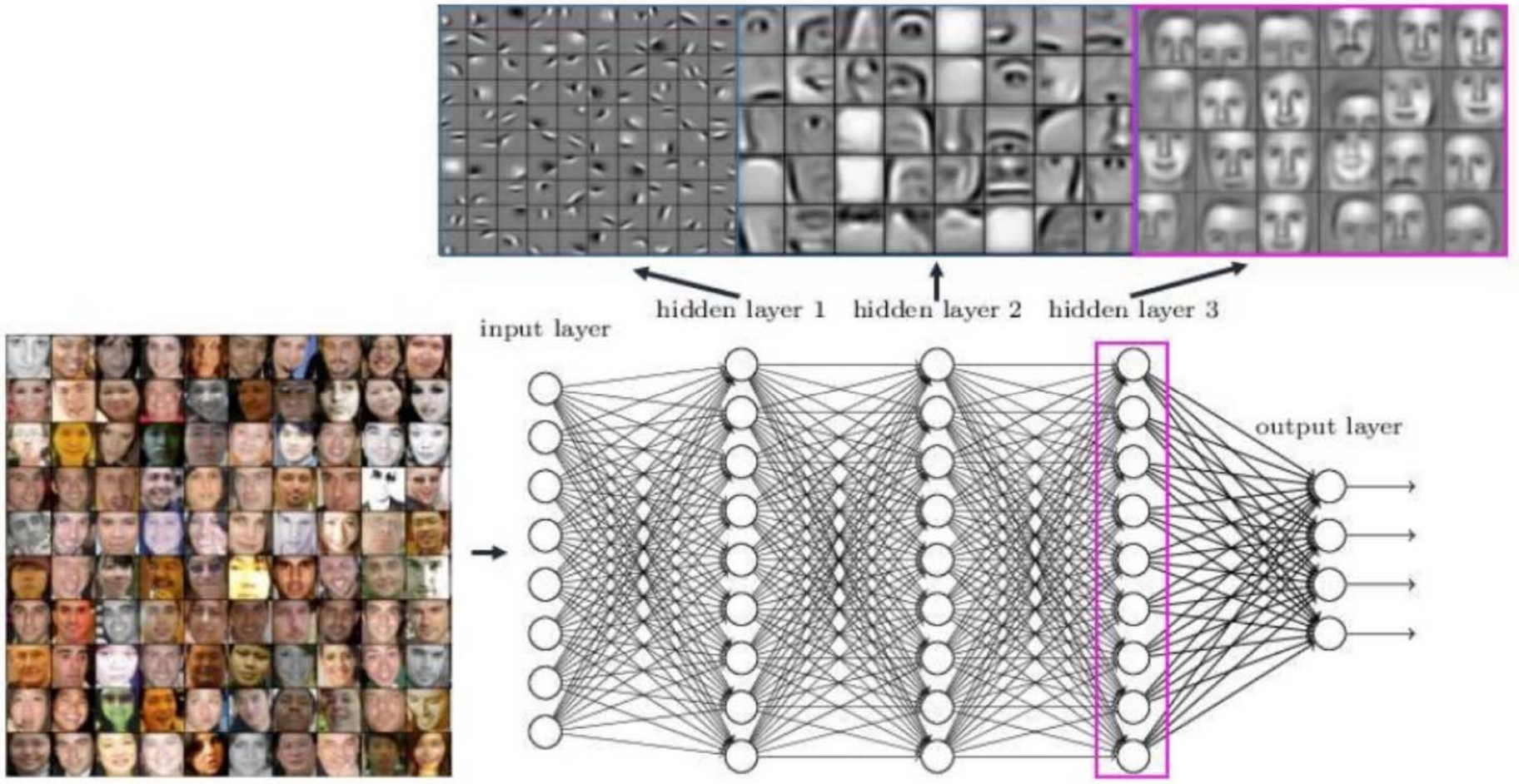

Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

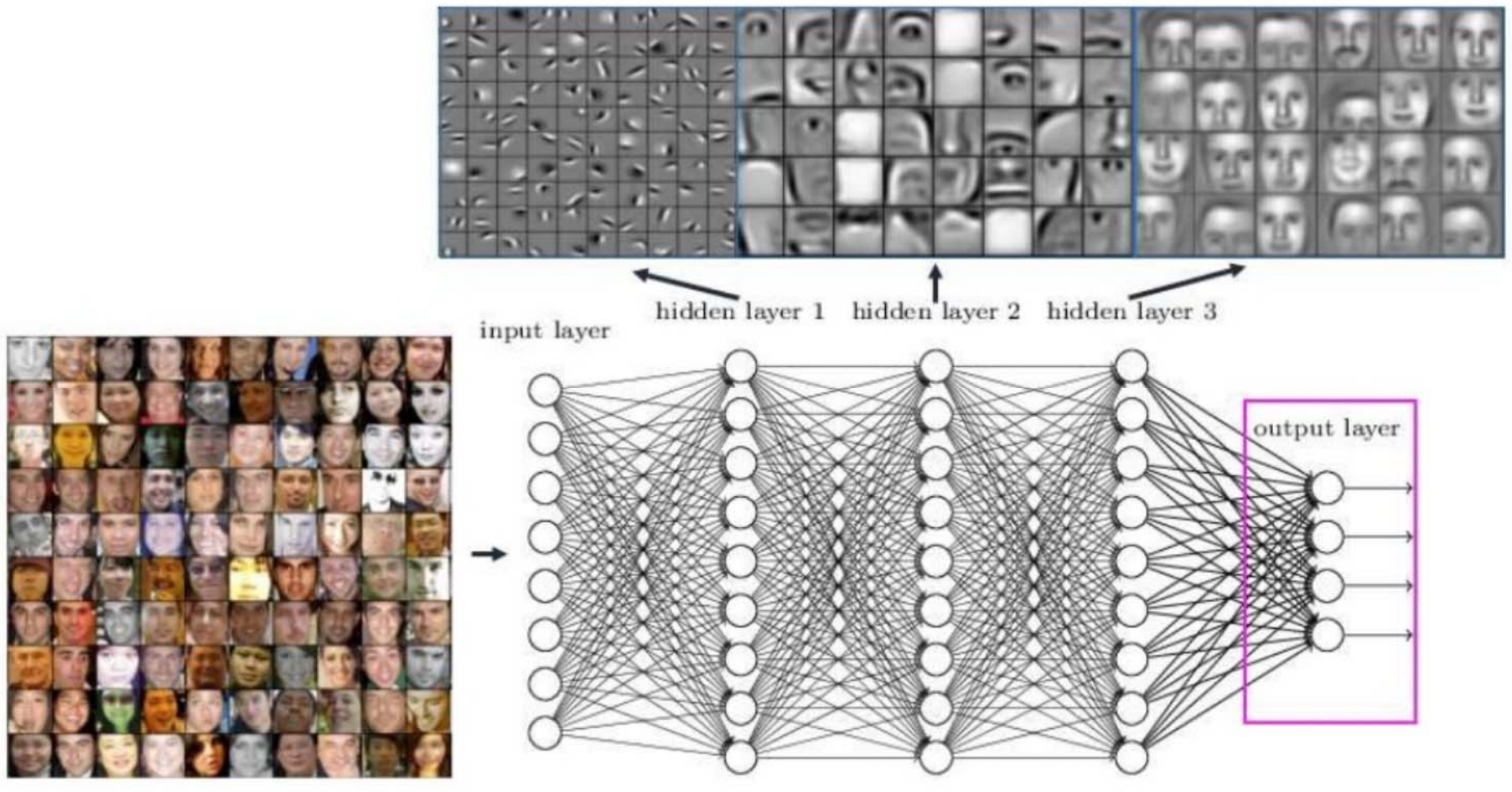
# The world is compositional (Yann LeCun)

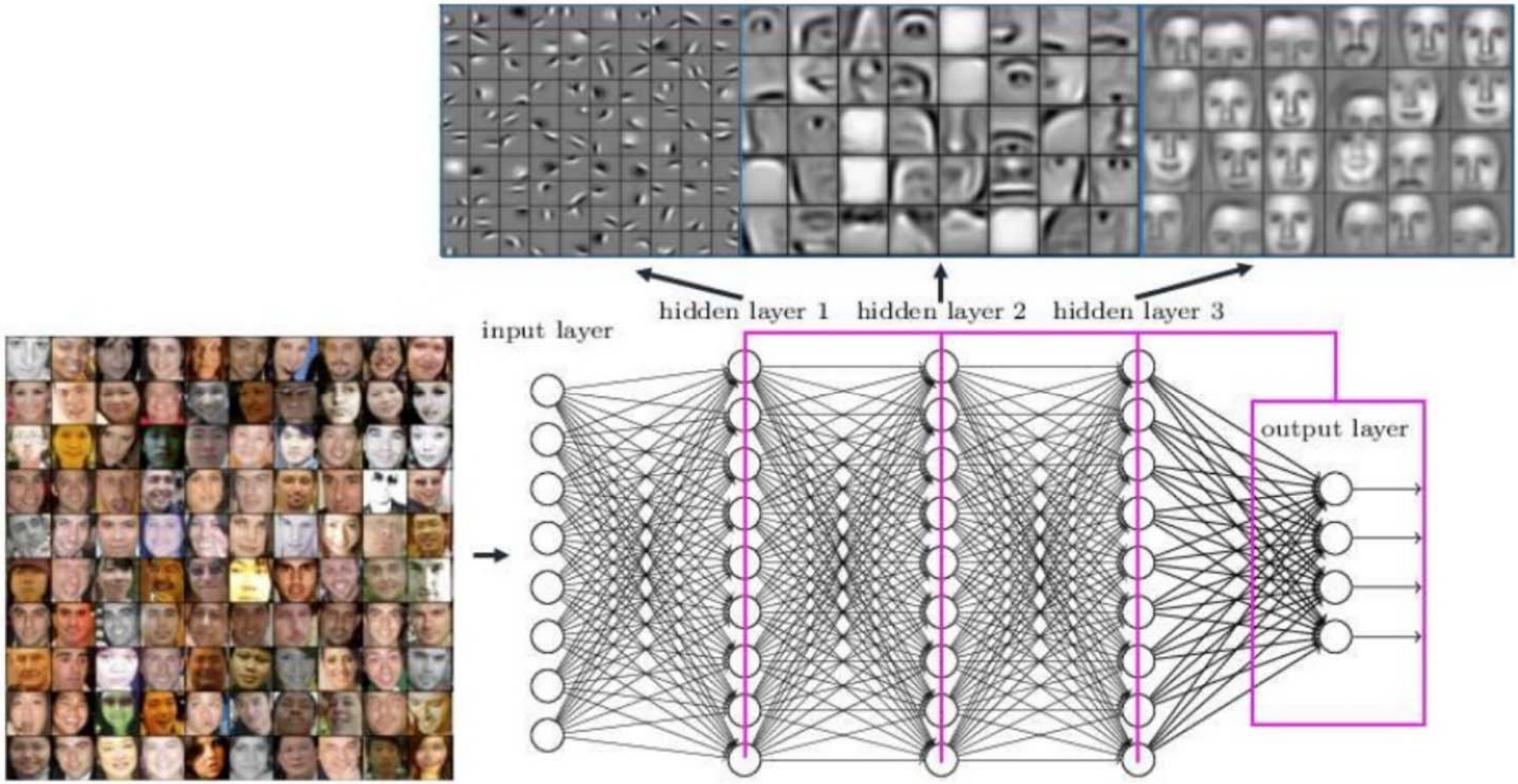



hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer


hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer

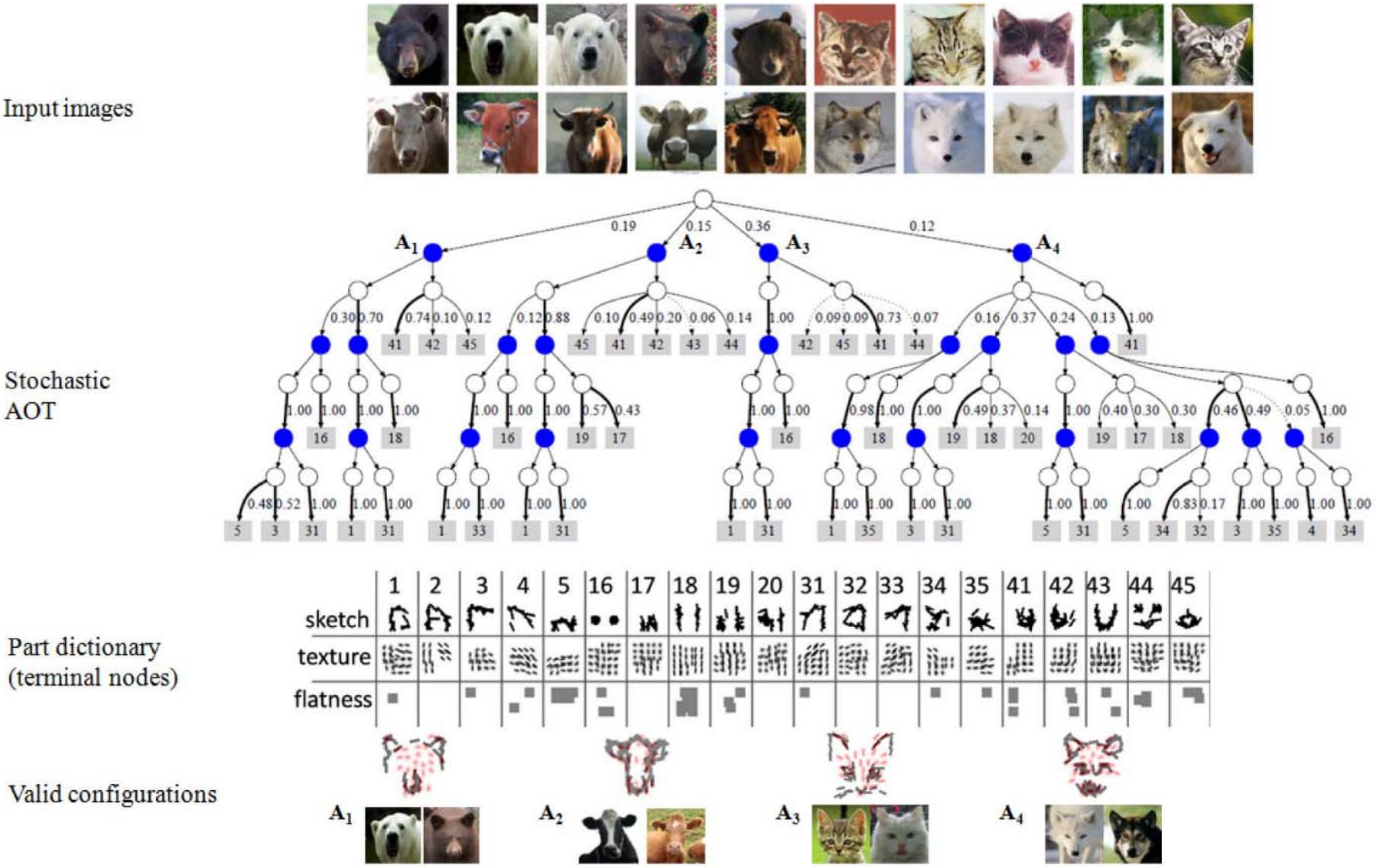

hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer


hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer


hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer

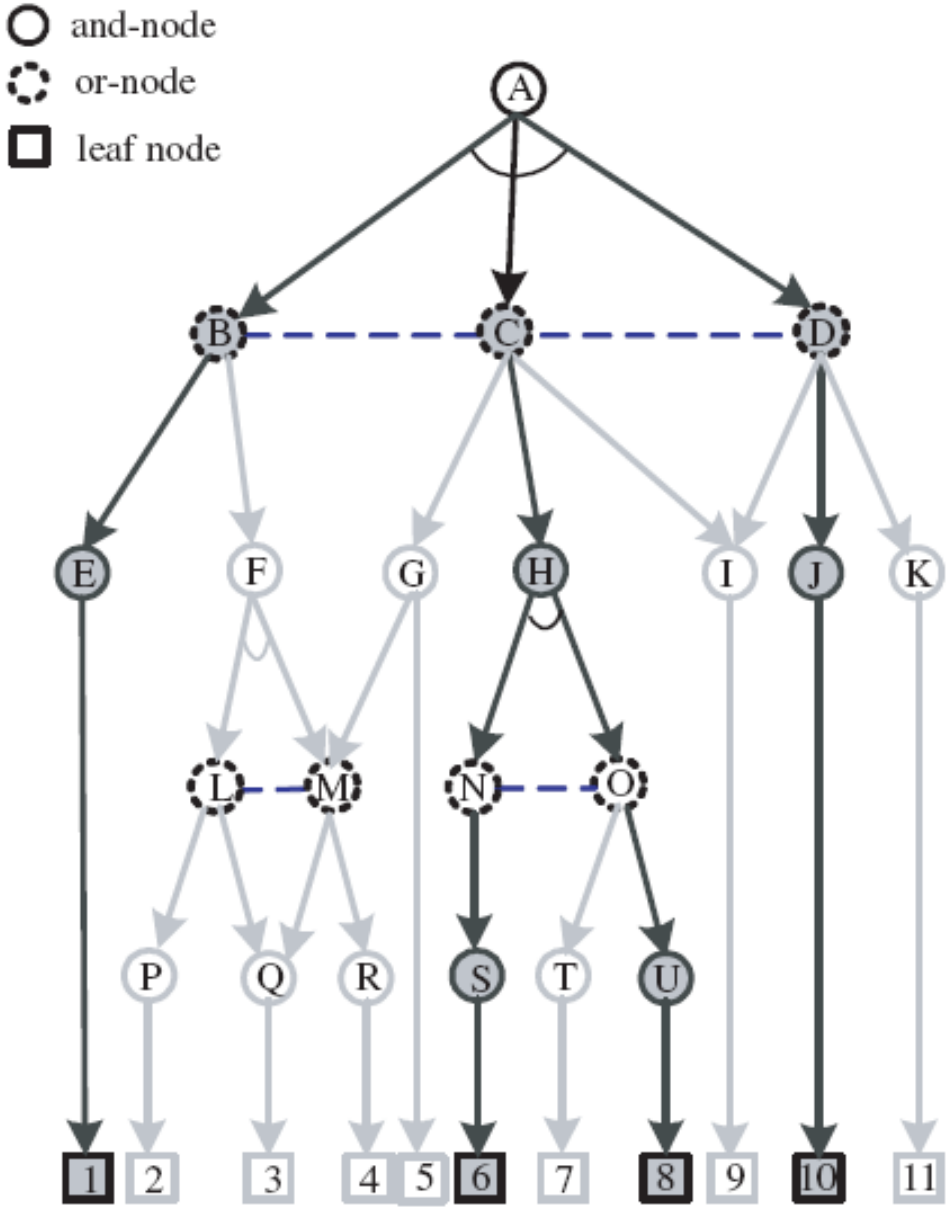
# 05 Stochastic AOG

# Stochastic AND-OR Templates for visual objects




Zhangzhang Si & Song-Chun Zhu 2013. Learning and-or templates for object recognition and detection. IEEE transactions on pattern analysis and machine intelligence, 35, (9), 2189-2205, doi:10.1109/TPAMI.2013.35.

# Framework for vision: AND-OR Graphs




- Algorithm for this framework
  - Top-down/bottom-up computation
- Generalization of small sample
  - Use Monte Carlos simulation to synthesis more configurations
- Fill **semantic gap**

Images credit to Zhaoyin Jia (2009)

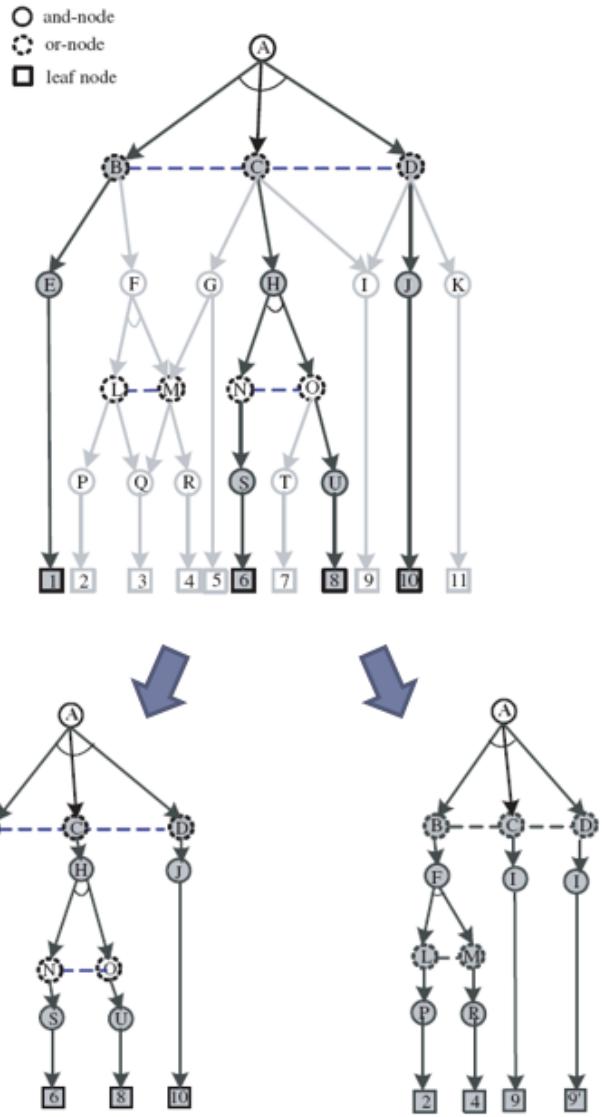
# Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$

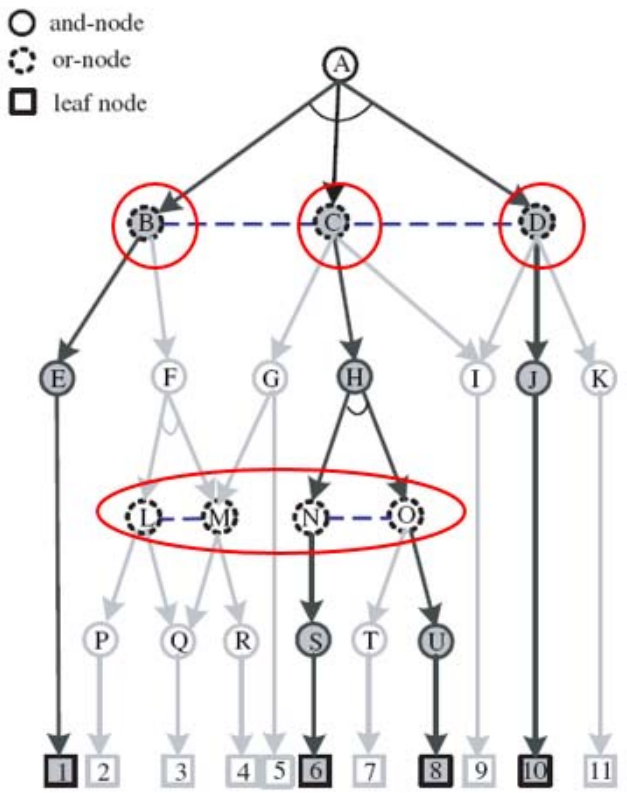


# Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$




SCFG: weigh the frequency at the children of or-nodes

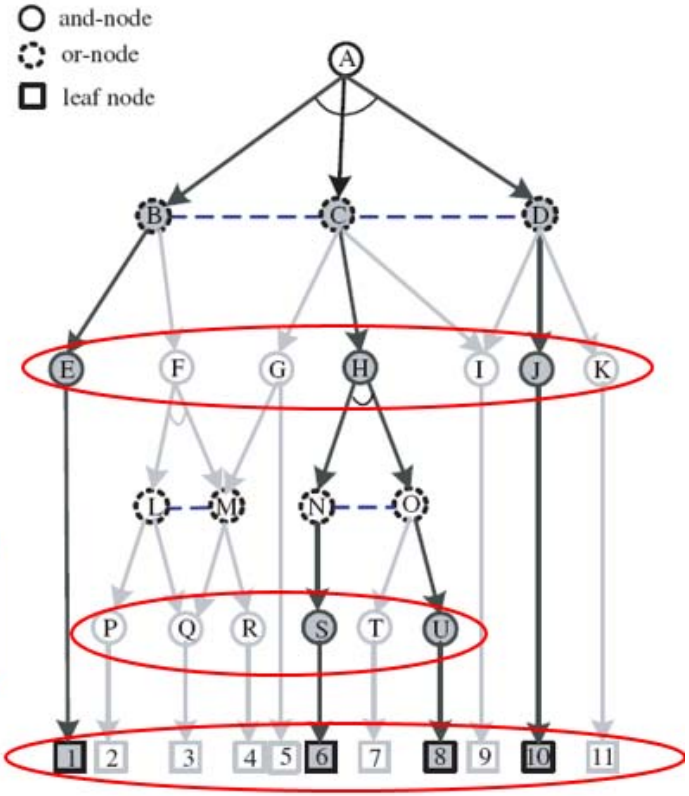
# Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$




Weigh the local compatibility of primitives (geometric and appearance)

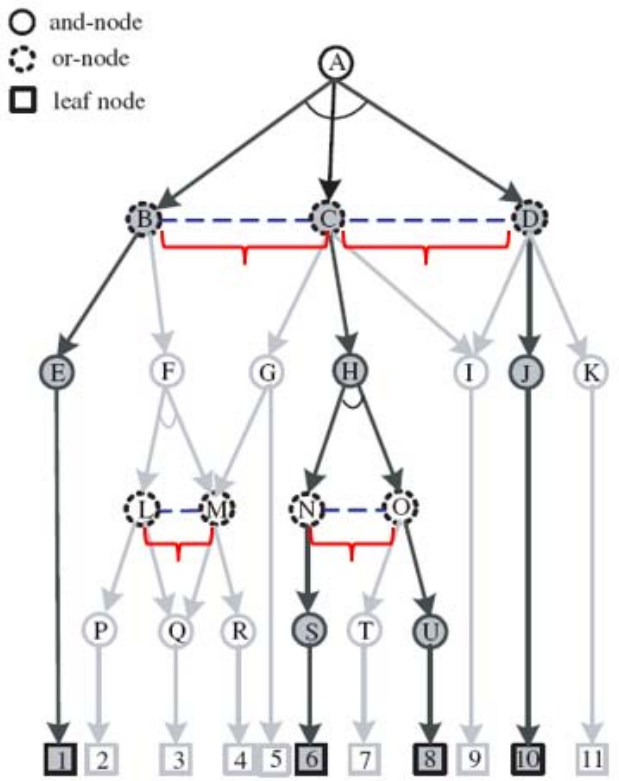
# Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$




Spatial and appearance between primitives (parts or objects)

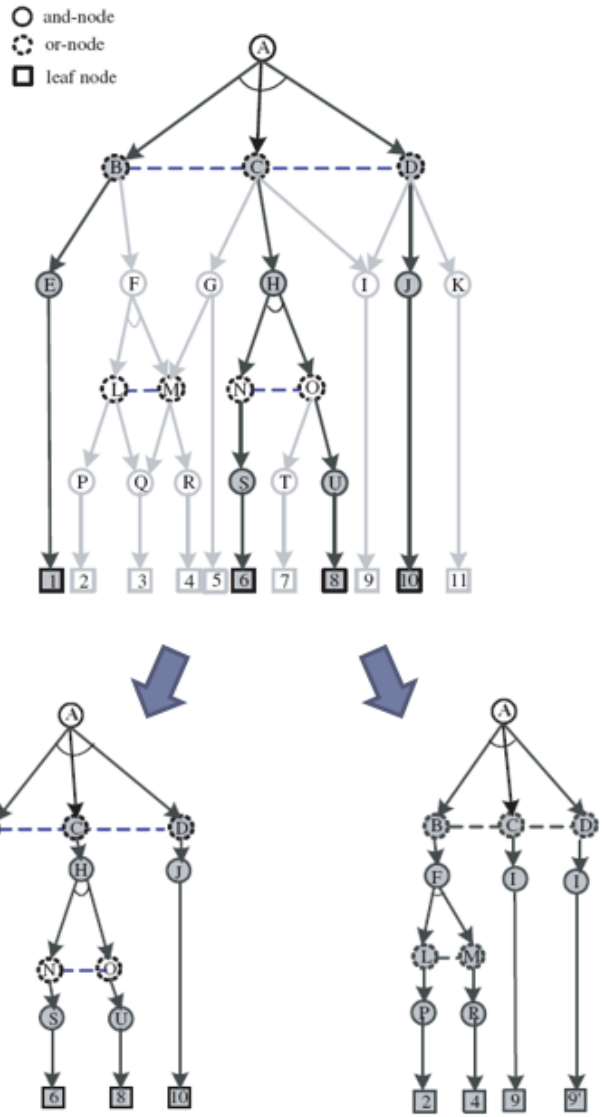
# Stochastic Model on AND-OR graph: Zhaoyin Jia (2009)

- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

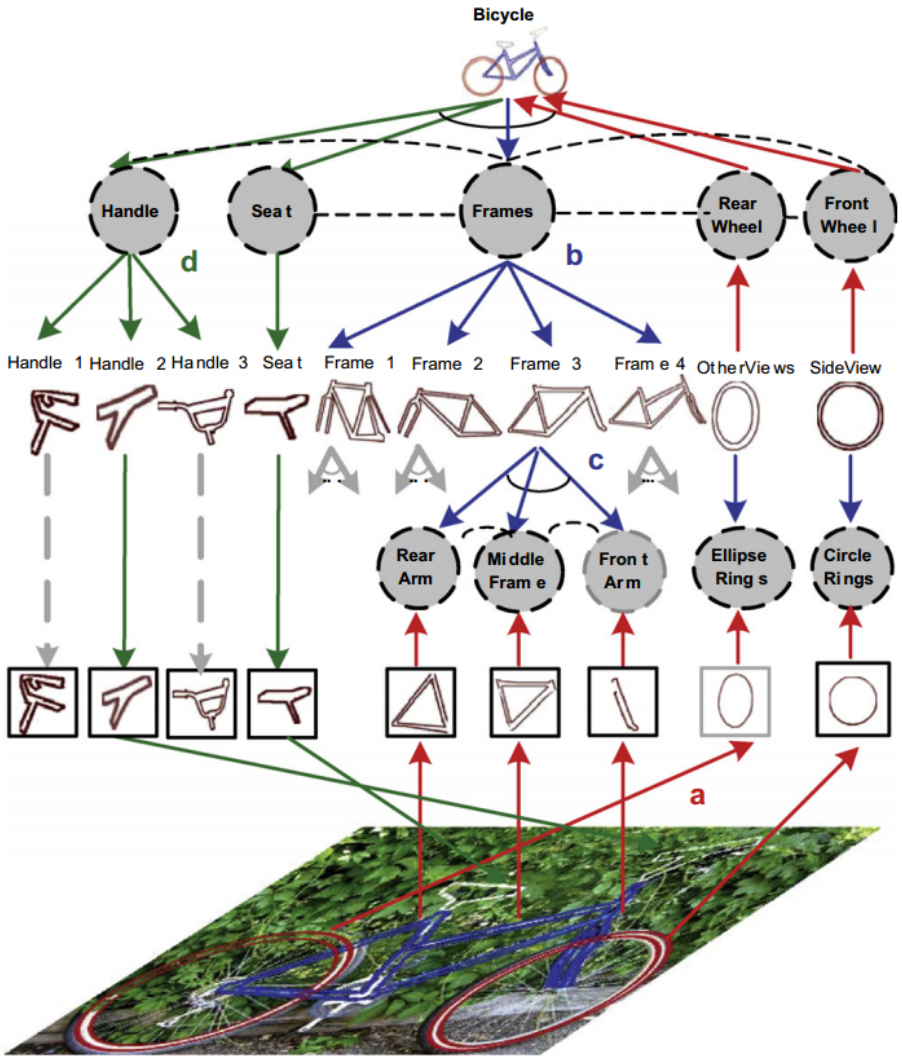
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

$$+\sum_{(i,j)\in E(pg)}\lambda_{ij}(v_i,v_j,\gamma_{ij},\rho_{ij})$$



# Stochastic graph grammar/comp. object representation




**Input:** an input image **I**, and a set of constructed And-Or graphs of compositional object categories.
**Output:** a parsing graph $pg_s$ of the scene that consists of the parsing graphs of detected objects.

- Repeat the following steps

1 Schedule the next node A to visit from the candidate parts.

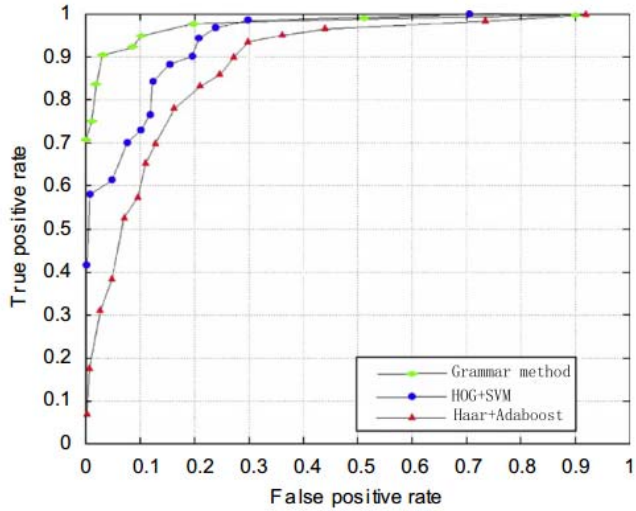
2 Call Bottom-up(A) to update A's **open** list.

i Detect terminal instances of A from the image.
ii Bind non-terminal instances of A from its children's **open** or **closed** lists

3 Call Top-down(A) to update A's **open** or **closed** lists.

i Accept hypotheses from A's **open** list to its **closed** list.
ii Remove (or disassemble) hypotheses from A's **closed** list.
iii Update the **open** lists for particles that overlap with node A.

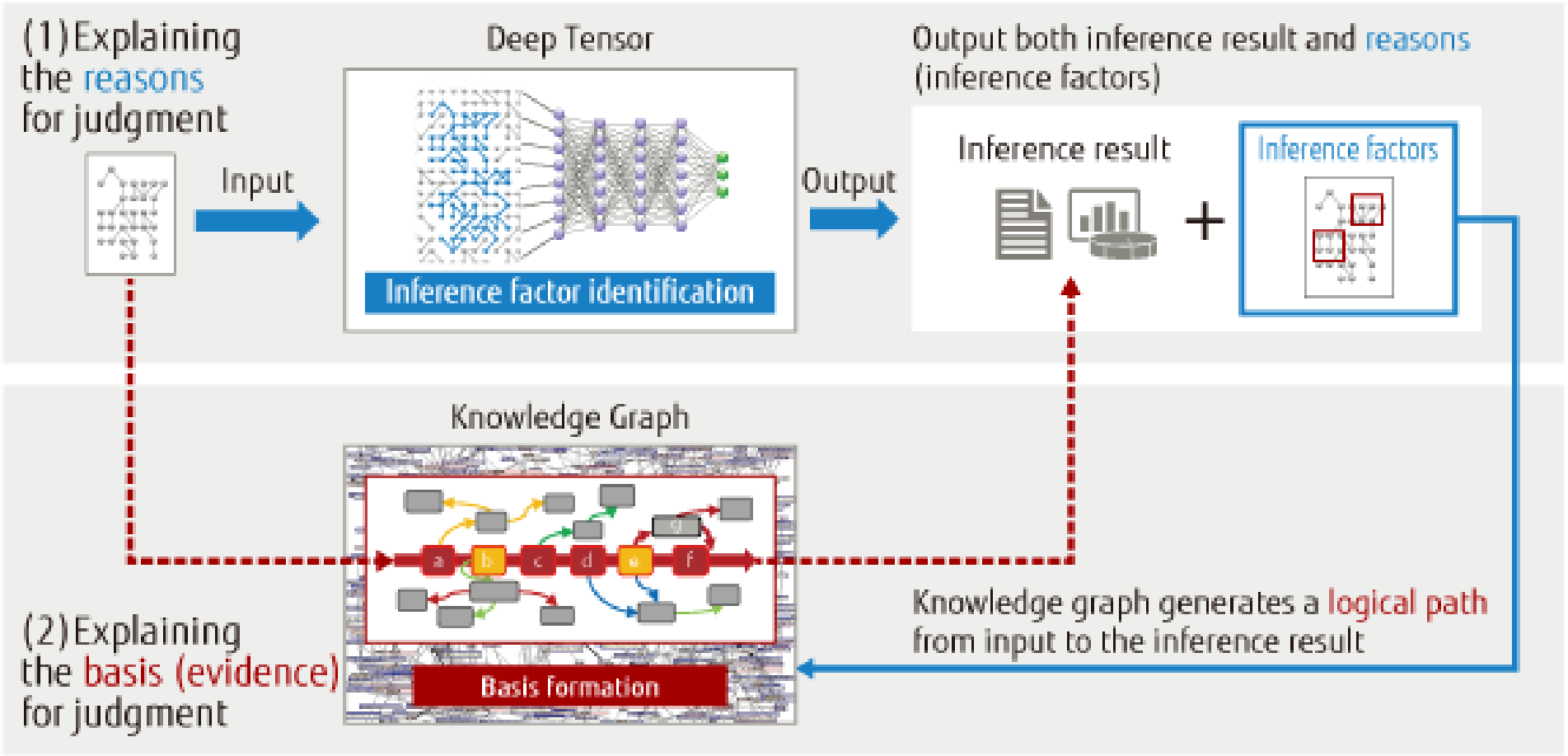
- Until the particles in **open** list with weights higher than the empirical threshold are exhausted. Output all parsing graphs whose root nodes are reached.




Liang Lin, Tianfu Wu, Jake Porway & Zijian Xu 2009. A stochastic graph grammar for compositional object representation and recognition. Pattern Recognition, 42, (7), 1297-1307, doi:10.1016/j.patcog.2008.10.033.

# Future Work

# Combination of Deep Learning with Ontologies




Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg & Andreas Holzinger. Explainable AI: the new 42? Springer Lecture Notes in Computer Science LNCS 11015, 2018 Cham. Springer, 295-303, doi:10.1007/978-3-319-99740-7_21.

## Seemingly trivial questions … ?

- When do we need explanations?
  - Of course ONLY in certain situations – most of the time we are happy with automatic approaches!
- What is a good explanation?
  - (obviously if the other did understand it)
  - Experiments needed!
- What is explainable/understandable/intelligible?
- When is it enough (Sättigungsgrad – you don't need more explanations – enough is enough)
- But how much is enough … ?

# Thank you!

# Appendix

# Explanations in Artificial Intelligence will be necessary


https://www.newyorker.com/cartoon/a19697

Noel C.F. Codella, Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei & Aleksandra Mojsilovic 2018. Teaching Meaningful Explanations. arXiv:1805.11648.



# Teaching Meaningful Explanations

**Noel C. F. Codella,* Michael Hind,* Karthikeyan Natesan Ramamurthy,***
**Murray Campbell, Amit Dhurandhar, Kush R. Varshney, Dennis Wei,**
**Aleksandra Mojsilović**
* These authors contributed equally.

IBM Research
Yorktown Heights, NY 10598
{nccodell,hindm,knatesa,mcam,adhuran,krvarshn,dwei,aleksand}@us.ibm.com

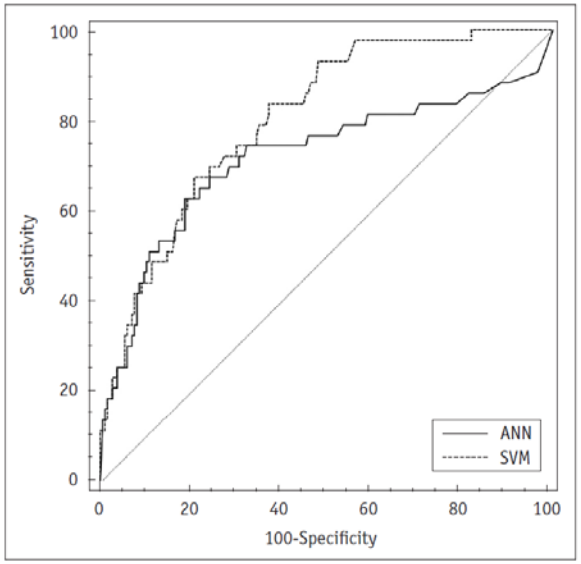
## Abstract

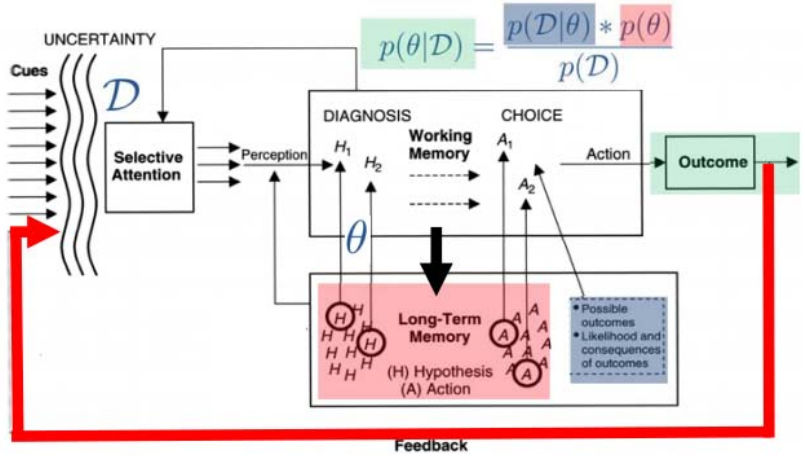
The adoption of machine learning in high-stakes applications such as healthcare and law has lagged in part because predictions are not accompanied by explanations comprehensible to the domain user, who often holds ultimate responsibility for decisions and outcomes. In this paper, we propose an approach to generate such explanations in which training data is augmented to include, in addition to features and labels, explanations elicited from domain users. A joint model is then learned to produce both labels and explanations from the input features. This simple idea ensures that explanations are tailored to the complexity expectations and domain knowledge of the consumer. Evaluation spans multiple modeling techniques on a simple game dataset, an image dataset, and a chemical odor dataset, showing that our approach is generalizable across domains and algorithms. Results demonstrate that meaningful explanations can be reliably taught to machine learning algorithms, and in some cases, improve modeling accuracy.

## 1 Introduction

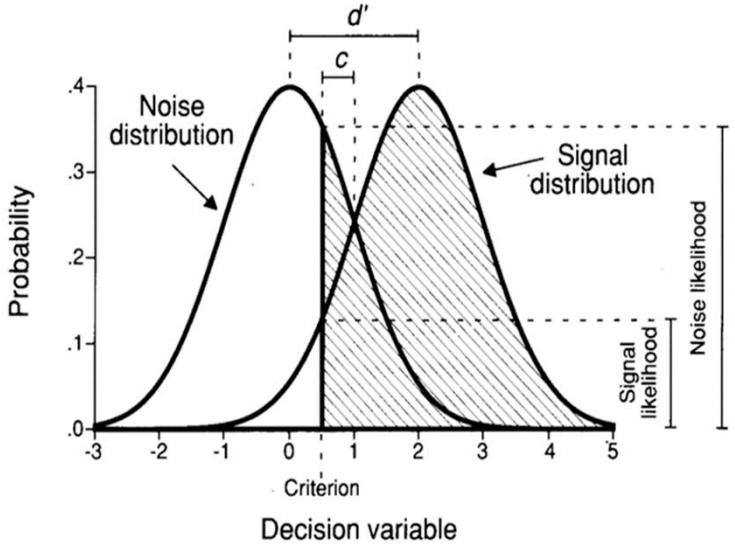
New regulations call for automated decision making systems to provide "meaningful information" on the logic used to reach conclusions [1–4]. Selbst and Powles interpret the concept of "meaningful information" as information that should be understandable to the audience (potentially individuals

# Reflection from last lectures

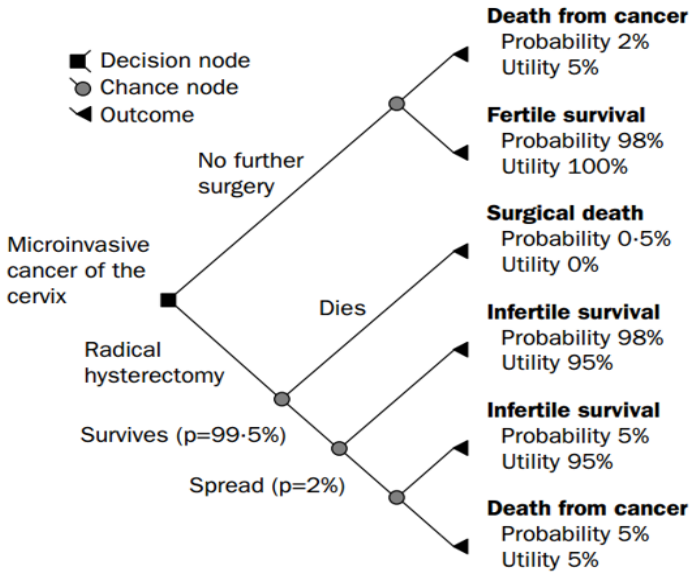



2




3




4