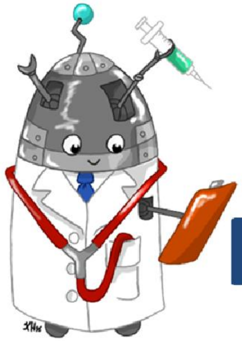
Andreas Holzinger

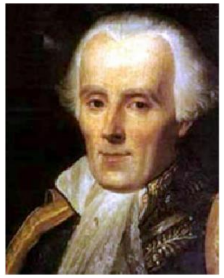
185.A83 Machine Learning for Health Informatics

2018S, VU, 2.0 h, 3.0 ECTS

Lecture 09 - Module 07 – Week 23 – 05.06.2018



From Clinical Decision Support to explainable AI (ex-AI)



a.holzinger@hci-kdd.org

http://hci-kdd.org/machine-learning-for-health-informatics-course

Keywords

- Decision support system (DSS)
- MYCIN – Rule Based Expert System
- GAMUTS in Radiology
- Reasoning under uncertainty
- Example: Radiotherapy planning
- Example: Case-Based Reasoning
- Explainable Artificial intelligence
- Re-trace > Understand > Explain
- Transparency > Trust > Acceptance
- Fairness > Transparency > Accountability
- Methods of Explainable AI

Advance Organizer (1/2)

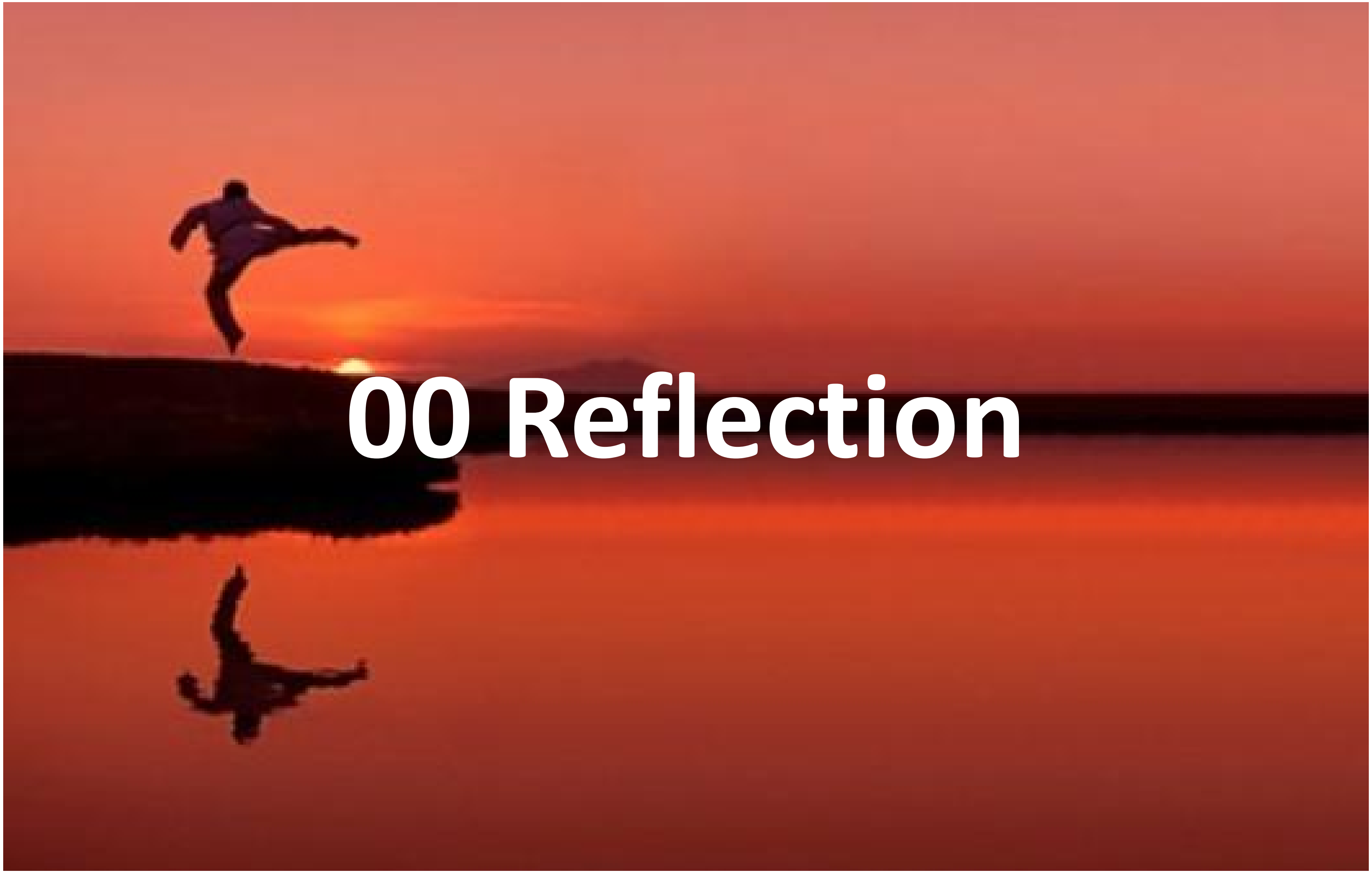
- **Case-based reasoning (CBR)** = process of solving new problems based on the solutions of similar past problems;
- **Certainty factor model (CF)** = a method for managing uncertainty in rule-based systems;
- **CLARION =** Connectionist Learning with Adaptive Rule Induction ON-line (CLARION) is a cognitive architecture that incorporates the distinction between implicit and explicit processes and focuses on capturing the interaction between these two types of processes. By focusing on this distinction, CLARION has been used to simulate several tasks in cognitive psychology and social psychology. CLARION has also been used to implement intelligent systems in artificial intelligence applications.
- **Clinical decision support (CDS)** = process for enhancing health-related decisions and actions with pertinent, organized clinical knowledge and patient information to improve health delivery;
- **Clinical Decision Support System (CDSS)** = expert system that provides support to certain reasoning tasks, in the context of a clinical decision;
- **Collective Intelligence** = shared group (symbolic) intelligence, emerging from cooperation/competition of many individuals, e.g. for consensus decision making;
- **Crowdsourcing** = a combination of "crowd" and "outsourcing" coined by Jeff Howe (2006), and describes a distributed problem-solving model; example for crowdsourcing is a public software beta-test;
- **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;

Advance Organizer (2/2)

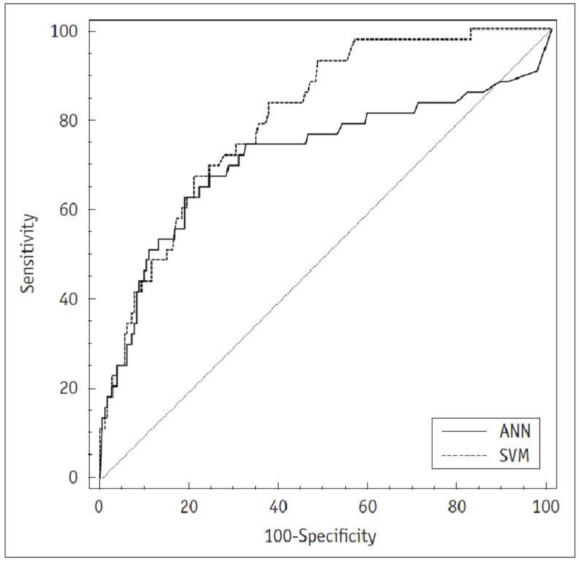
- **DXplain** = a DSS from the Harvard Medical School, to assist making a diagnosis (clinical consultation), and also as an instructional instrument (education); provides a description of diseases, etiology, pathology, prognosis and up to 10 references for each disease;
- Etiology = in medicine (many) factors coming together to cause an illness (see causality)
- Explainable AI = Explainability = upcoming fundamental topic within recent AI; answering e.g. **why** a decision has been made
- **Expert-System** = emulates the decision making processes of a human expert to solve complex problems;
- **GAMUTS** in Radiology = Computer-Supported list of common/uncommon differential diagnoses;
- **ILIAD** = medical expert system, developed by the University of Utah, used as a teaching and testing tool for medical students in problem solving. Fields include Pediatrics, Internal Medicine, Oncology, Infectious Diseases, Gynecology, Pulmonology etc.
- Interpretability = there is no formal technical definition yet, but it is considered as a prerequisite for trust
- **MYCIN** = one of the early medical expert systems (Shortliffe (1970), Stanford) to identify bacteria causing severe infections, such as bacteremia and meningitis, and to recommend antibiotics, with the dosage adjusted for patient's body weight;
- **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

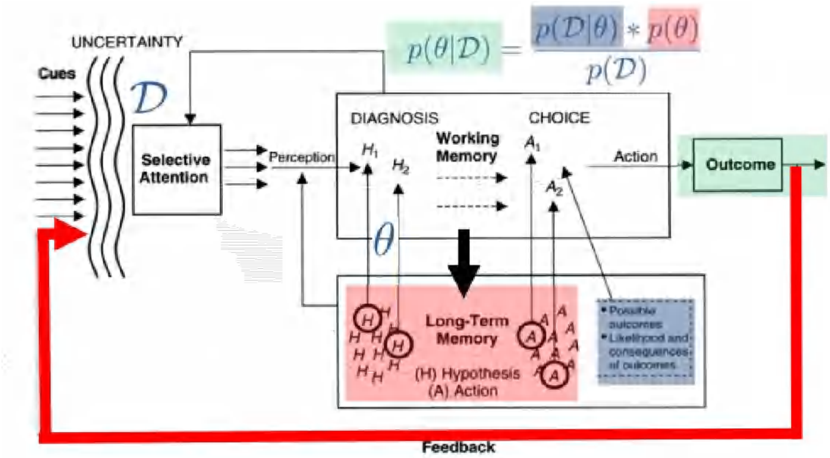
- … can apply your knowledge gained in the previous lectures to example systems of decision support;
- … have an overview about the core principles and architecture of decision support systems;
- … are familiar with the certainty factors as e.g. used in MYCIN;
- … are aware of some design principles of DSS;
- … have seen similarities between DSS and KDD on the example of computational methods in cancer detection;
- … have seen basics of CBR systems;

- 00 Reflection – follow-up from last lecture
- 01 Decision Support Systems (DSS)
- 02 Computers help making better decisions?
- 03 History of DSS = History of AI
- 04 Example: Towards Personalized Medicine
- 05 Example: Case Based Reasoning (CBR)
- 06 Towards Explainable AI
- 07 Some methods of explainable AI

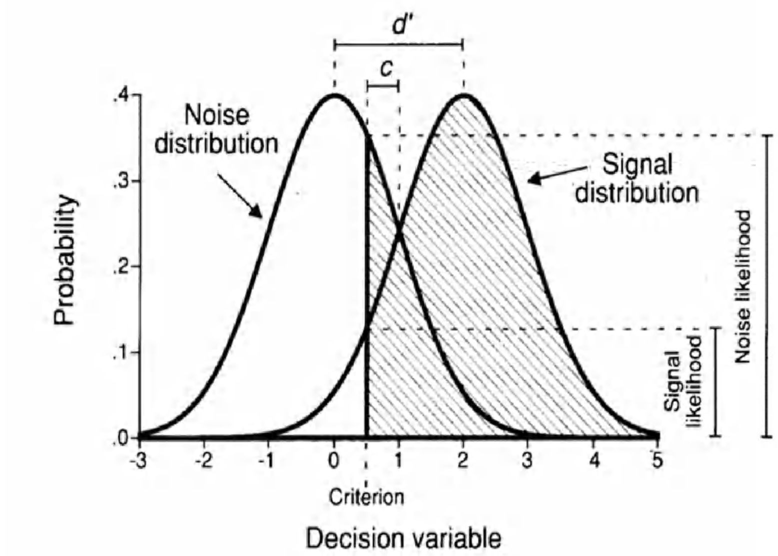

00 Reflection

Reflection from last lectures

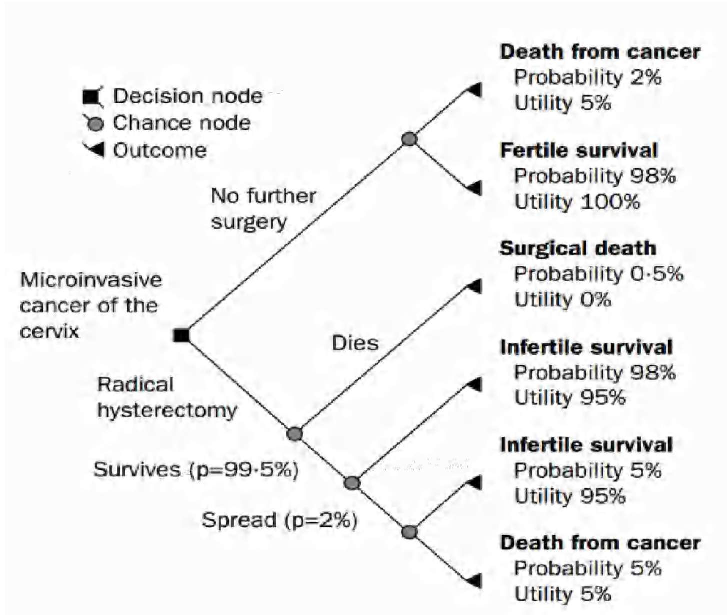




2

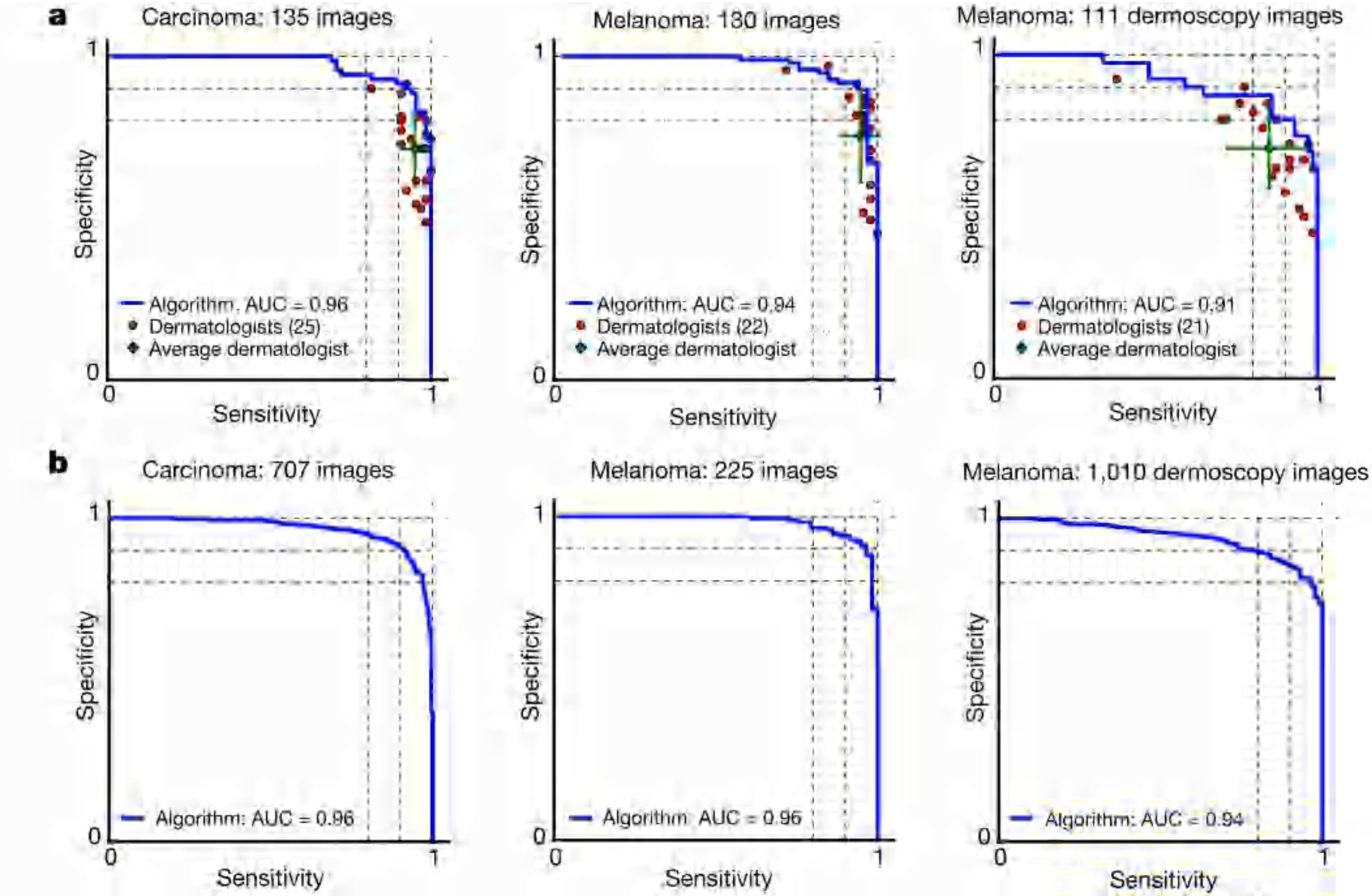


3



4

How do you explain this …

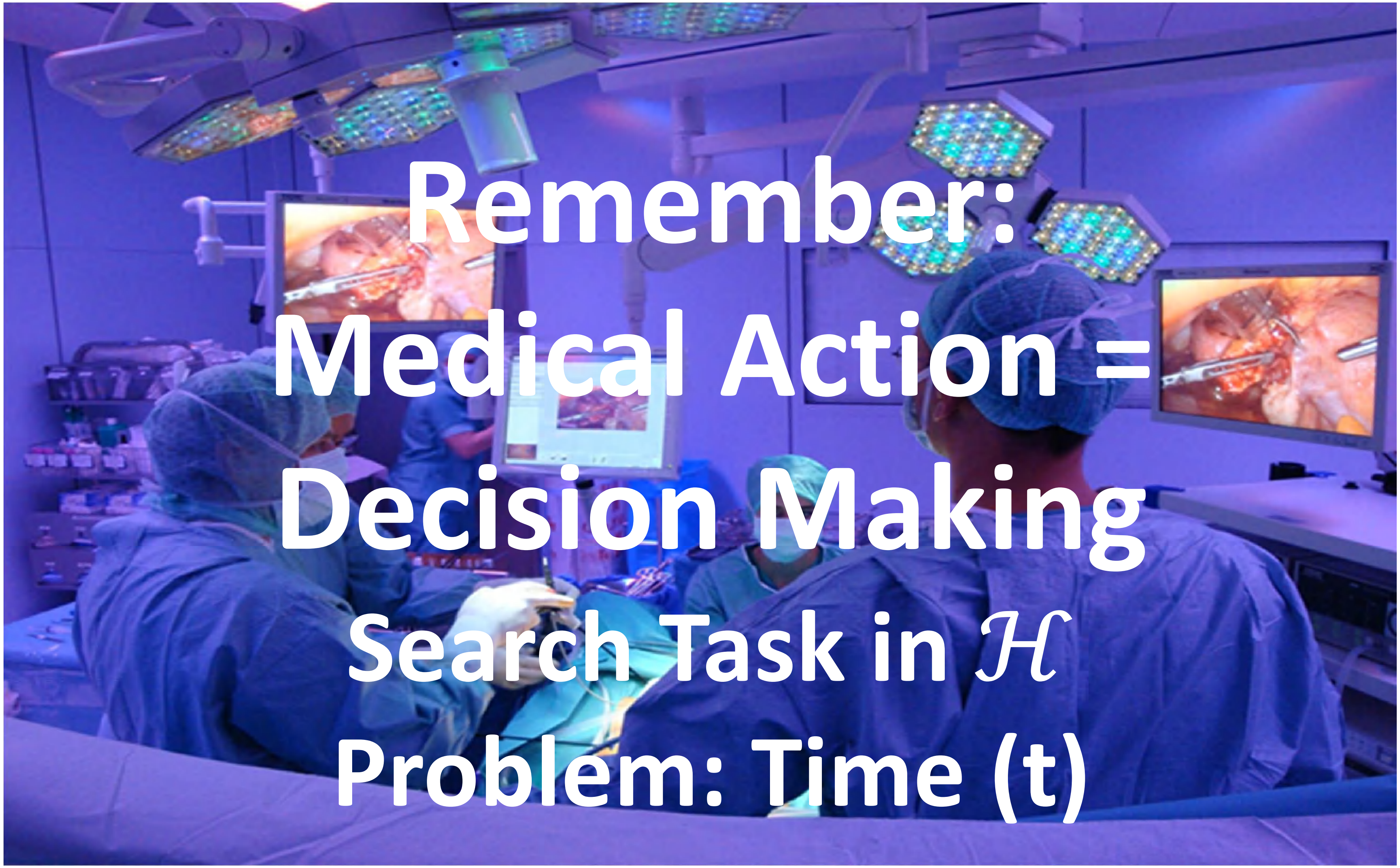


Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118

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 Decision Support Systems



Remember:
Medical Action = Decision Making
Search Task in $\mathcal{H}$
Problem: Time (t)

Search in an arbitrarily high-dimensional space < 5 min.!

Decision Making is central in any (medical) work

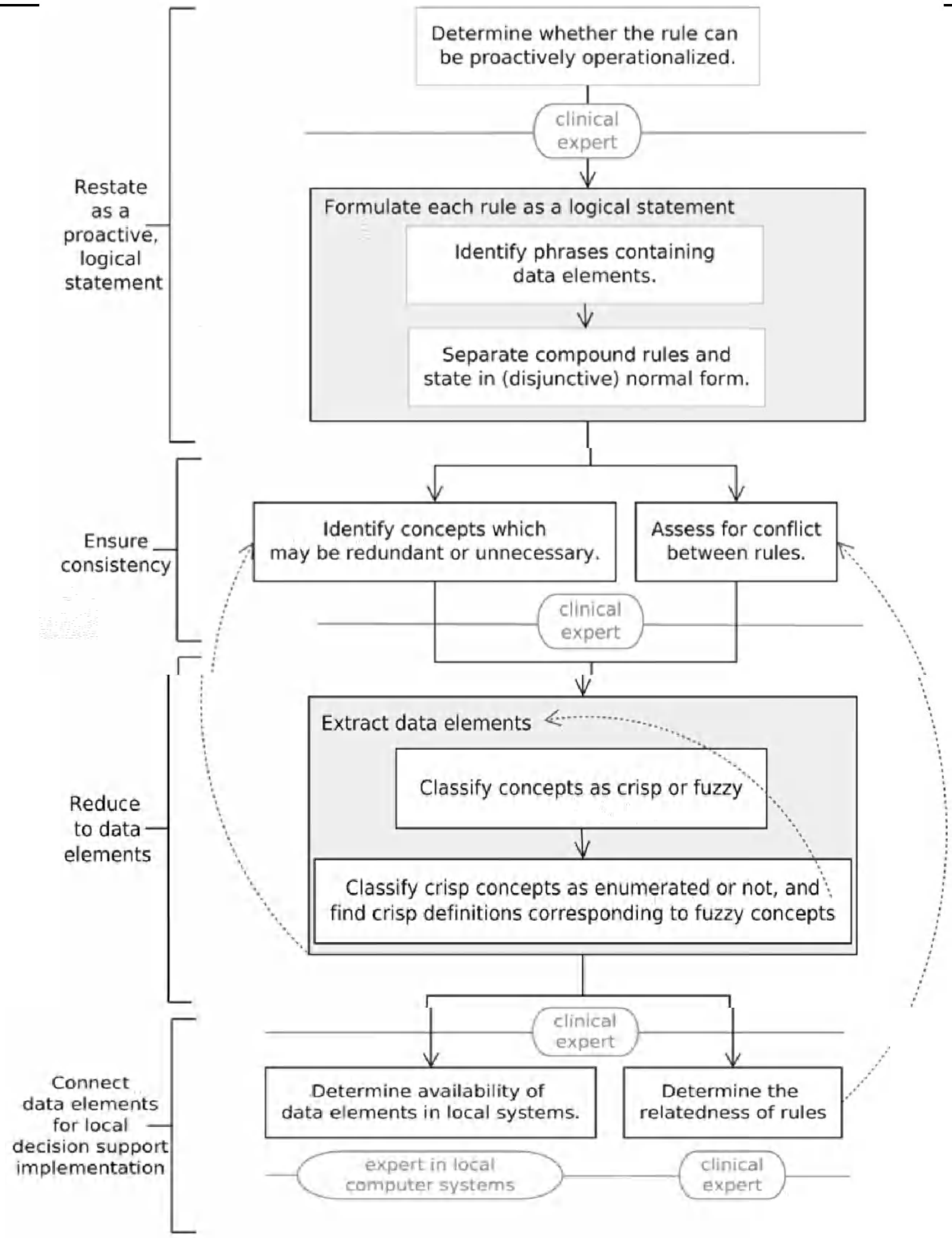
Source: Cisco (2008). Cisco Health Presence Trial at Aberdeen Royal Infirmary in Scotland

- 400 BC Hippocrates (460-370 BC), father of western medicine:
  - A medical record should accurately reflect the course of a disease
  - A medical record should indicate the probable cause of a disease
- **1890** William Osler (1849-1919), father of modern western medicine
  - **Medicine is a science of uncertainty and an art of probabilistic decision making**
- Today
  - Prediction models are based on data features, patient health status is modelled as high-dimensional feature vectors …

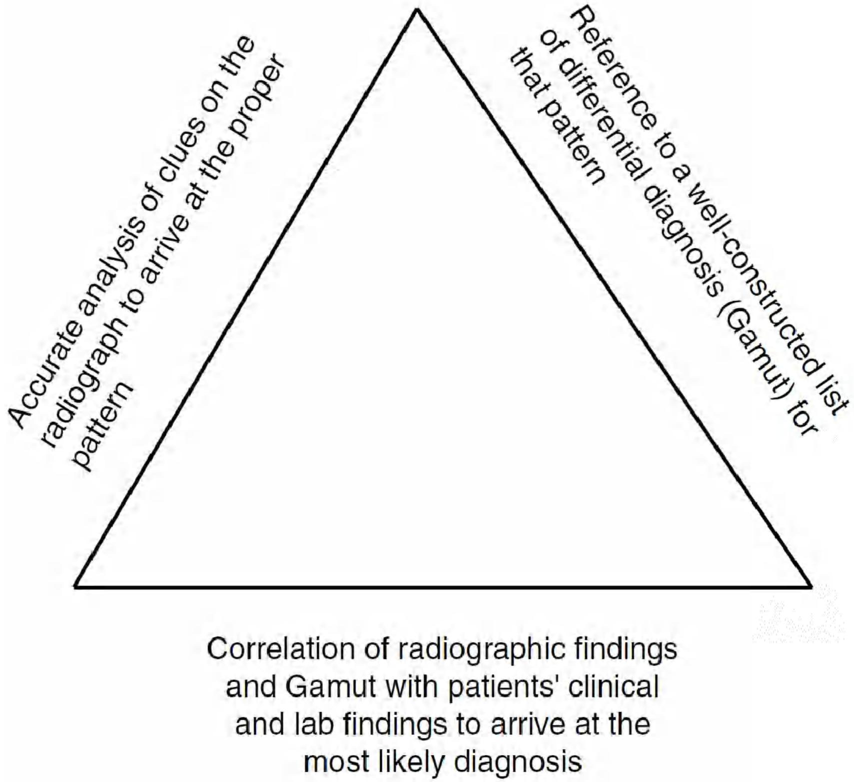
- Clinical guidelines are **systematically** developed documents to assist doctors and patient decisions about appropriate care;
- In order to build DS, based on a guideline, it is **formalized** (transformed from natural language to a logical algorithm), and
- **implemented** (using the algorithm to program a DSS);
- To increase the quality of care, they must be linked to a process of care, for example:
  - "80% of diabetic patients should have an HbA1c below 7.0" could be linked to processes such as:
  - "All diabetic patients should have an annual HbA1c test" and
  - "Patients with values over 7.0 should be rechecked within 2 months."
- **Condition-action rules** specify one or a few conditions which are linked to a specific action, in contrast to narrative guidelines which describe a series of branching or iterative decisions unfolding over time.
- Narrative guidelines and clinical rules are two ends of a continuum of clinical care standards.

Example: Clinical Guidelines

Medlock, S., Opondo, D., Eslami, S., Askari, M., Wierenga, P., de Rooij, S. E. & Abu-Hanna, A. (2011) LERM (Logical Elements Rule Method): A method for assessing and formalizing clinical rules for decision support. *International Journal of Medical Informatics, 80, 4, 286-295.*



Example: Triangulation to find diagnoses



Reeder, M. M. & Felson, B. 2003. *Reeder and Felson's gamuts in radiology: comprehensive lists of roentgen differential diagnosis, New York, Springer Verlag.*

Gamut F-137

PHRENIC NERVE PARALYSIS OR DYSFUNCTION

COMMON

1. Iatrogenic (eg, surgical injury; chest tube; therapeutic avulsion or injection; subclavian vein puncture)
2. Infection (eg, tuberculosis; fungus disease; abscess)
3. Neoplastic invasion or compression (esp. carcinoma of lung)

UNCOMMON

1. Aneurysm$_g$, aortic or other
2. Birth trauma (Erb's palsy)
3. Herpes zoster
4. Neuritis, peripheral (eg, diabetic neuropathy)
5. Neurologic disease$_g$ (eg, hemiplegia; encephalitis; polio; Guillain-Barré S.)
6. Pneumonia
7. Trauma

Reference

1. Prasad S, Athreya BH: Transient paralysis of the phrenic nerve associated with head injury. JAMA 1976;236:2532–2533

Example - Gamuts in Radiology

REEDER AND FELSON'S
GAMUTS IN RADIOLOGY

GAMUT G-25
EROSIVE GASTRITIS*

COMMON
1. Acute gastritis (eg, alcohol abuse)
2. Crohn's disease
3. Drugs (eg, aspirin, NSAID; steroids)
4. *Helicobacter pylori* infection
5. Idiopathic
6. [Normal areae gastricae]
7. Peptic ulcer; hyperacidity

UNCOMMON
1. Corrosive gastritis
2. *Cryptosporidium* antritis
3. [Lymphoma]
4. Opportunistic infection (eg, candidiasis {moniliasis}; herpes simplex; cytomegalovirus)
5. Postoperative gastritis
6. Radiation therapy
7. Zollinger-Ellison S., multiple endocrine neoplasia (MEN) S.

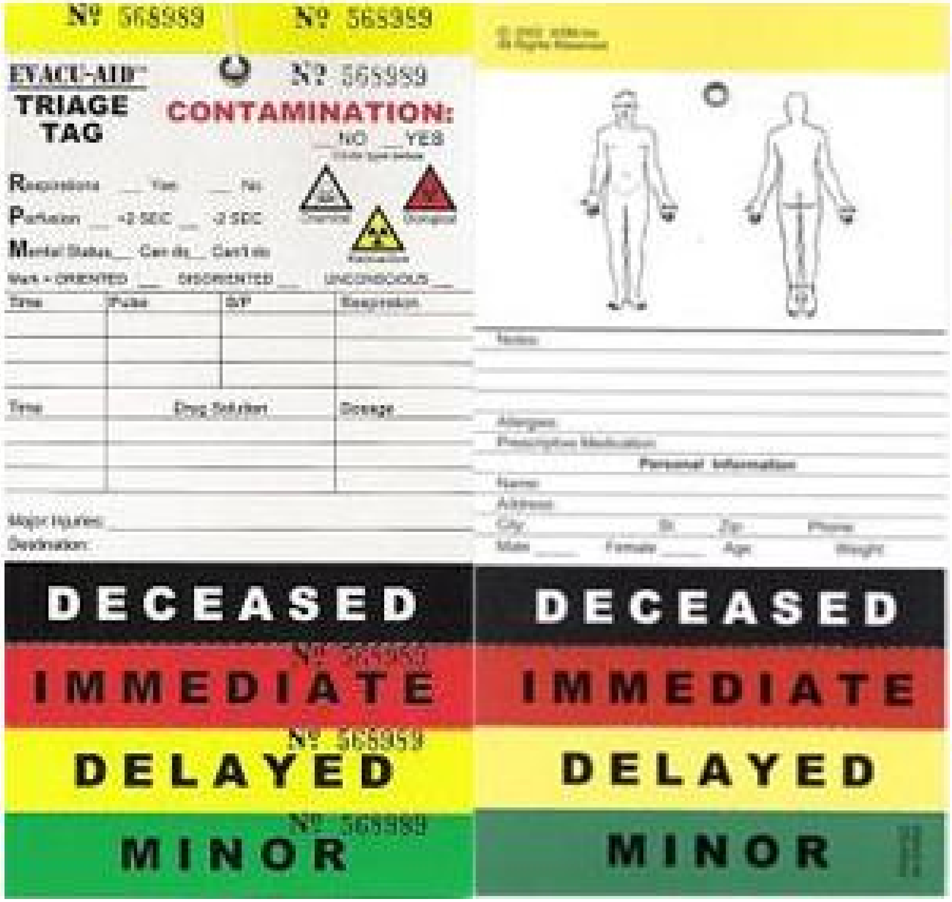
* Superficial erosions or aphthoid ulcerations seen especially with double contrast technique.

[ ] This condition does not actually cause the gamuted imaging finding, but can produce imaging changes that simulate it.

Reeder, M. M. & Felson, B. (2003) *Reeder and Felson's gamuts in radiology: comprehensive lists of roentgen differential diagnosis. New York, Springer*

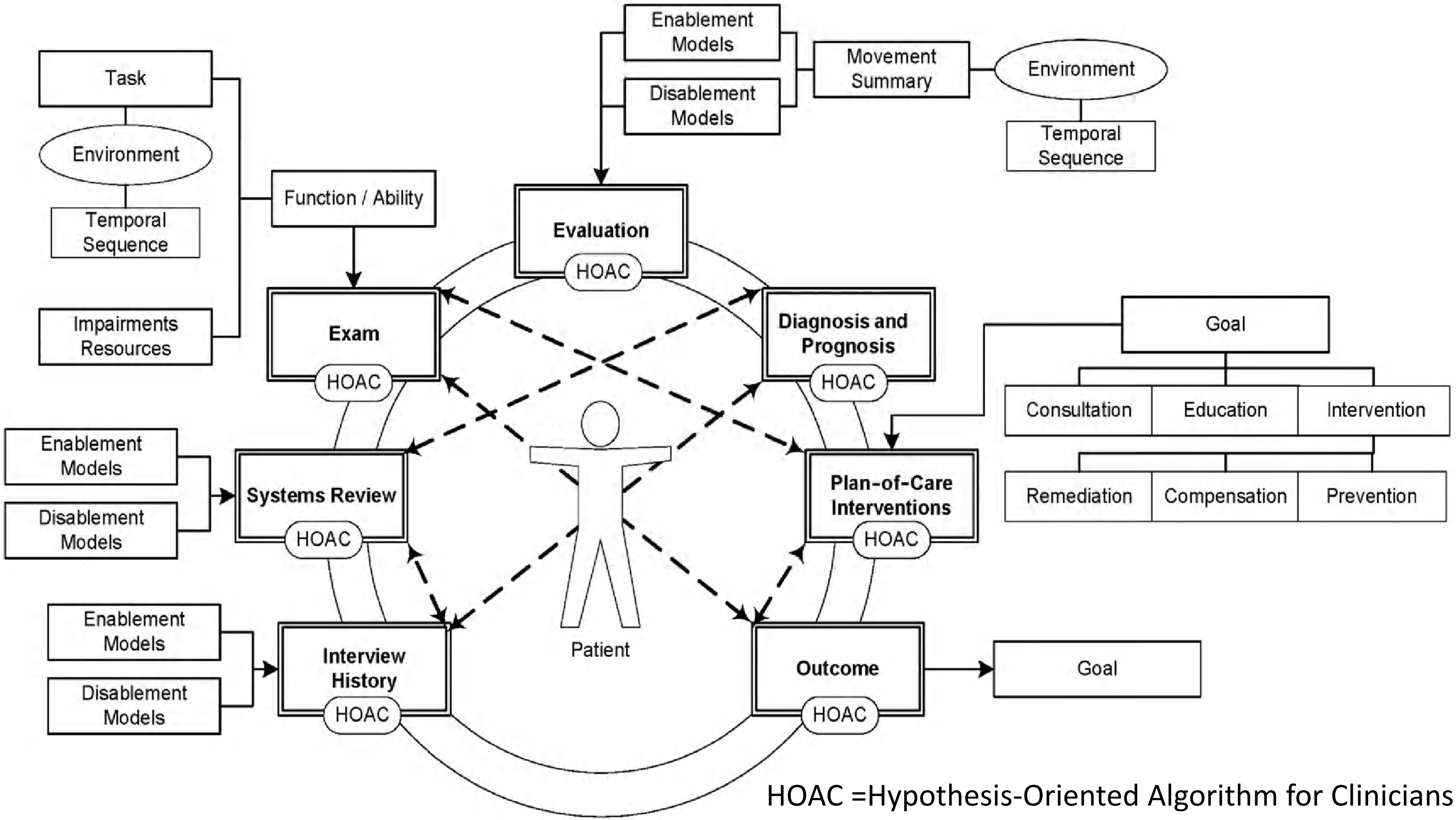
http://rfs.acr.org/gamuts/data/G-25.htm

Example: Triage Tags - International Triage Tags



Iserson, K. V. & Moskop, J. C. 2007. Triage in Medicine, Part I: Concept, History, and Types. Annals of Emergency Medicine, 49, (3), 275-281.

Image Source: http://store.gomed-tech.com



HOAC =Hypothesis-Oriented Algorithm for Clinicians

Schenkman, M., Deutsch, J. E. & Gill-Body, K. M. (2006) An Integrated Framework for Decision Making in Neurologic Physical Therapist Practice. *Physical Therapy, 86, 12, 1681-1702.*

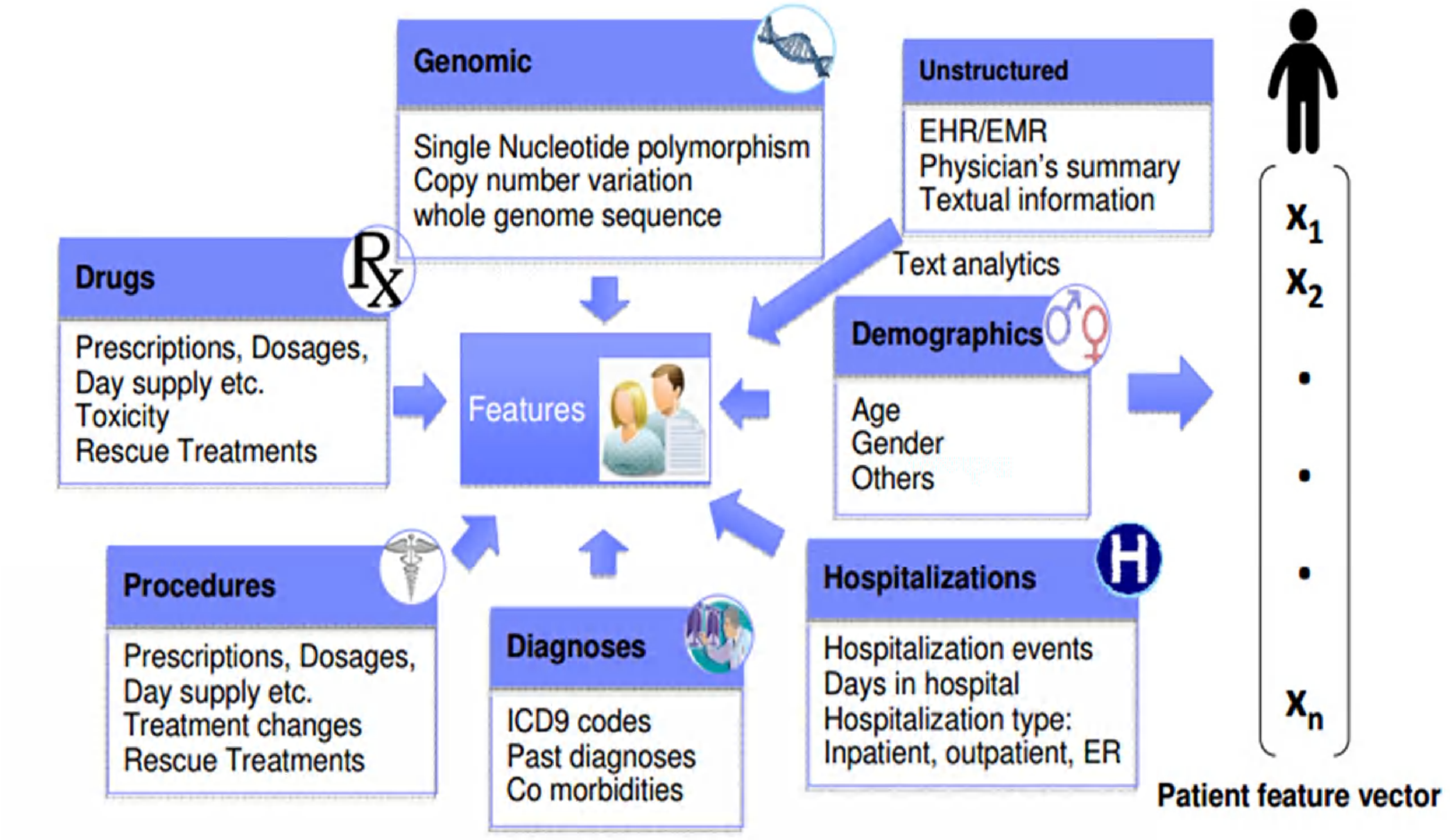
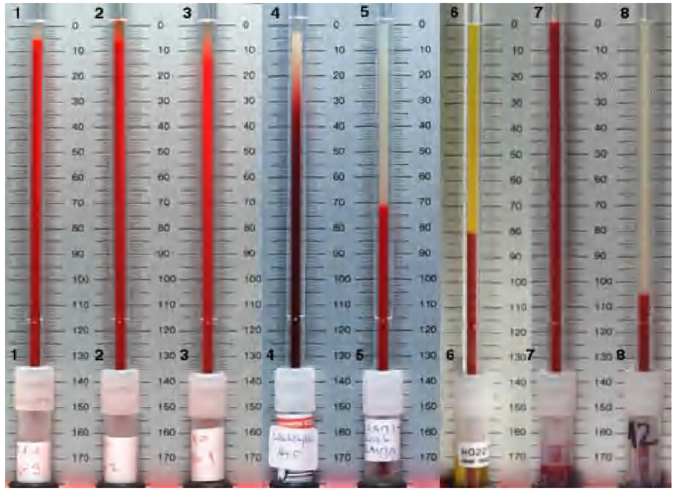


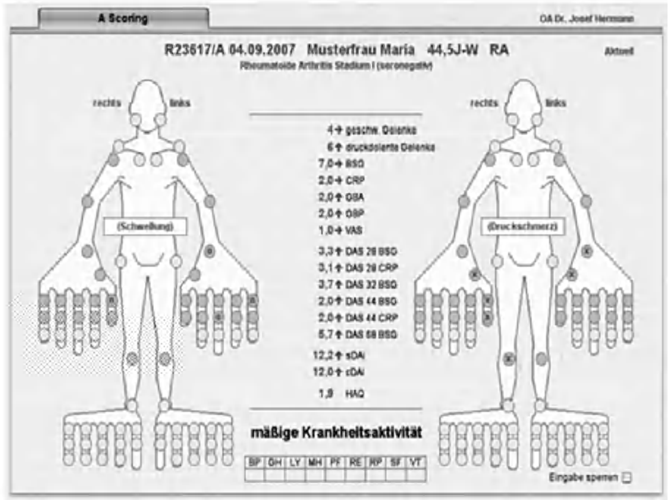
Image credit to Michal Rosen-Zvi

Chao, J., Parker, B. A. & Zvaifler, N. J. (2009) Accelerated Cutaneous Nodulosis Associated with Aromatase Inhibitor Therapy in a Patient with Rheumatoid Arthritis. *The Journal of Rheumatology, 36,* ***5, 1087-1088.***

Ikari, K. & Momohara, S. (2005) Bone Changes in Rheumatoid Arthritis. *New England Journal of Medicine, 353,* ***15, e13.***

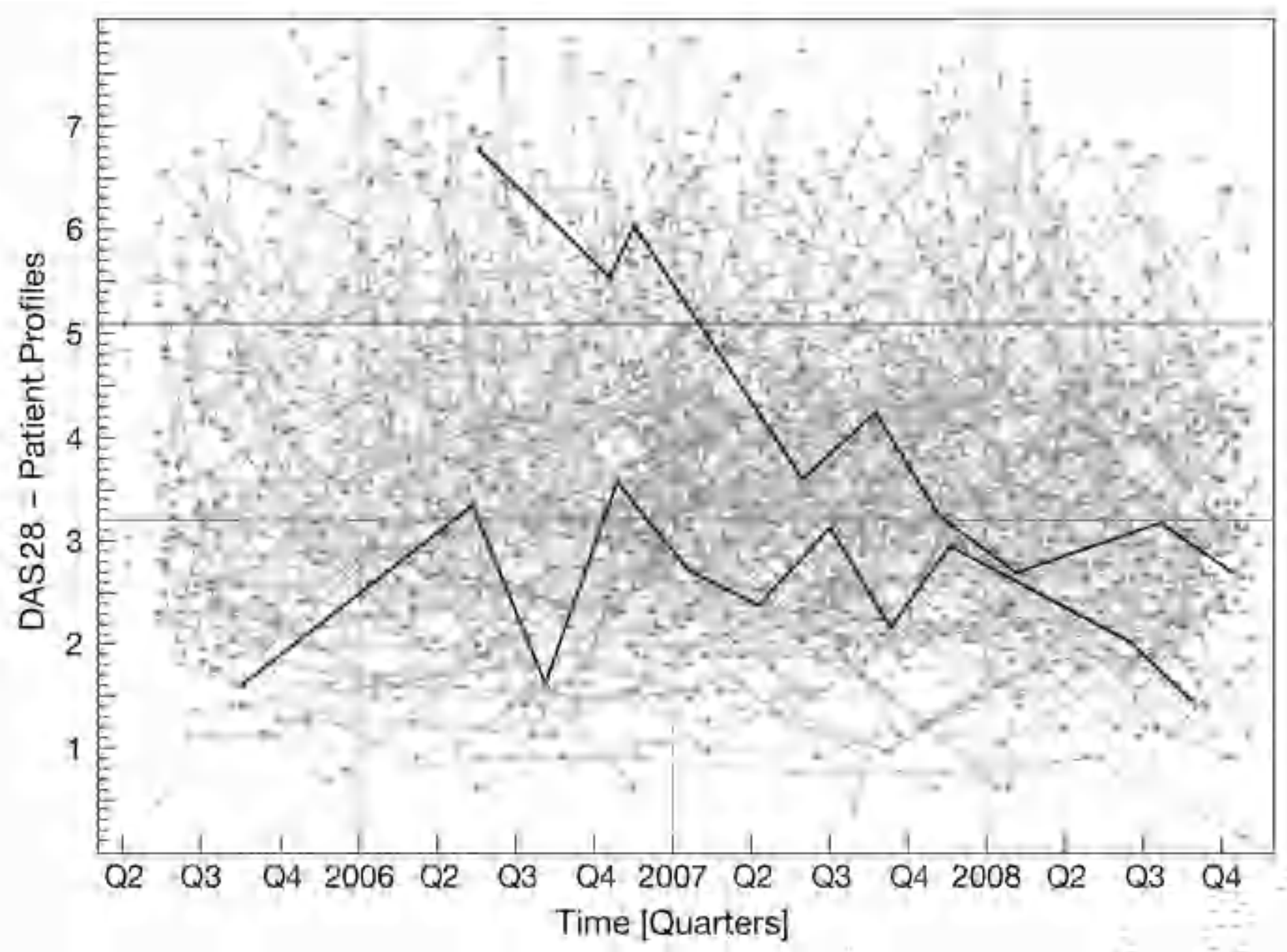
- 50+ Patients per day ~ 5000 data points per day …
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ **time-series data**





Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011). *Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, IEEE, 550-554.*

Gaining out Knowledge of time-series data



Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011). *Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, IEEE, 550-554.*

02 Can Computers help doctors to make better decisions?

Computers to help human doctors to make better decisions

"If you want a second opinion, I'll ask my computer."

http://biomedicalcomputationreview.org/content/clinical-decision-support-providing-quality-healthcare-help-computer

Augmenting Human Capabilities …

Two types of decisions (Diagnosis vs. Therapy)

- **Type 1 Decisions:** related to the **diagnosis,** i.e. computers are used to assist in diagnosing a disease on the basis of the individual patient data. Questions include:
  - What is the probability that this patient has a myocardial infarction on the basis of given data (patient history, ECG, …)?
  - What is the probability that this patient has acute appendices, given the signs and symptoms concerning abdominal pain?

- **Type 2 Decisions:** related to **therapy,** i.e. computers are used to select the best therapy on the basis of clinical evidence, e.g.:
  - What is the best therapy for patients of age x and risks y, if an obstruction of more than z % is seen in the left coronary artery?
  - What amount of insulin should be prescribed for a patient during the next 5 days, given the blood sugar levels and the amount of insulin taken during the recent weeks?

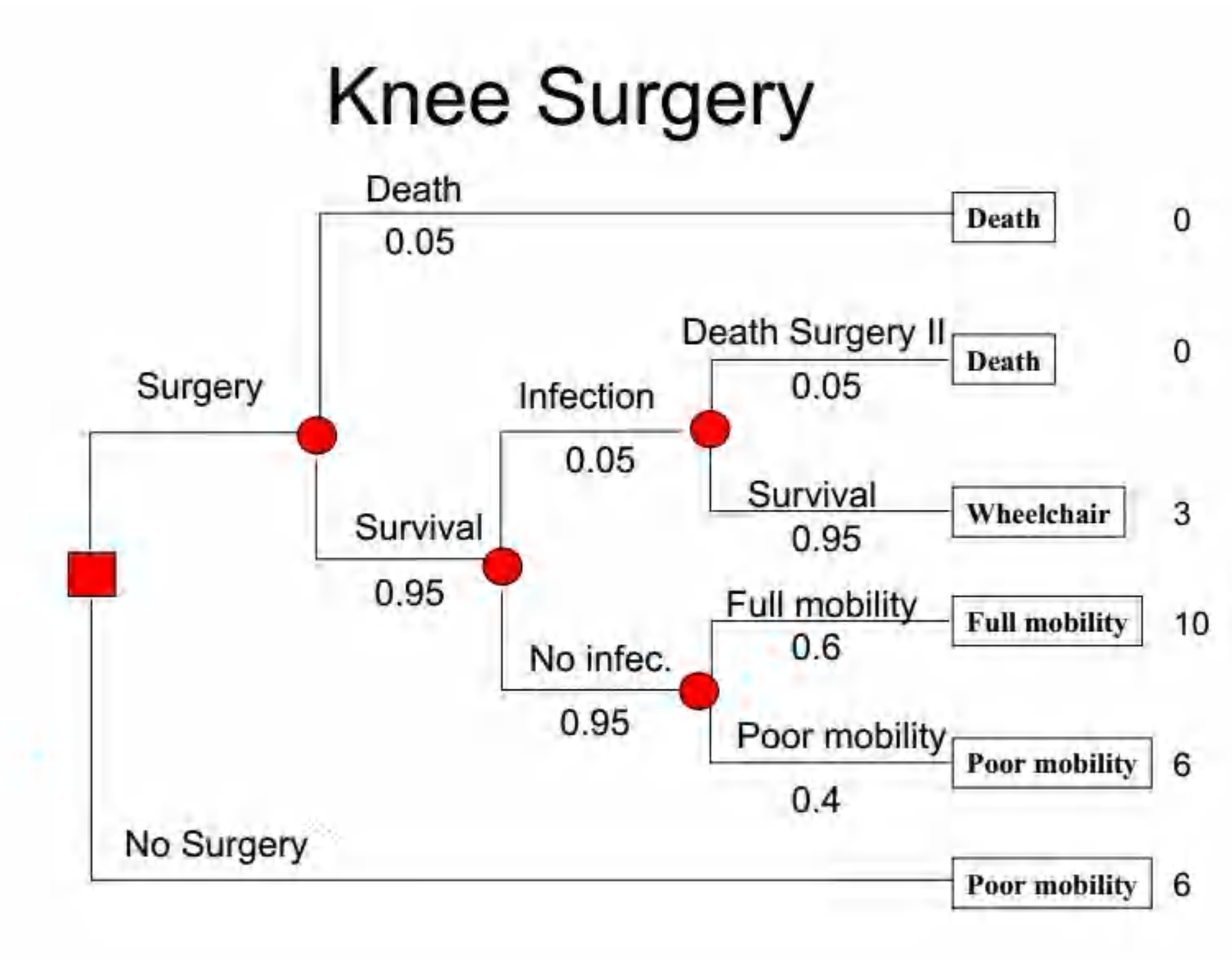
Bemmel, J. H. V. & Musen, M. A. 1997. *Handbook of Medical Informatics, Heidelberg, Springer.*

Example: Knee Surgery of a Soccer Player

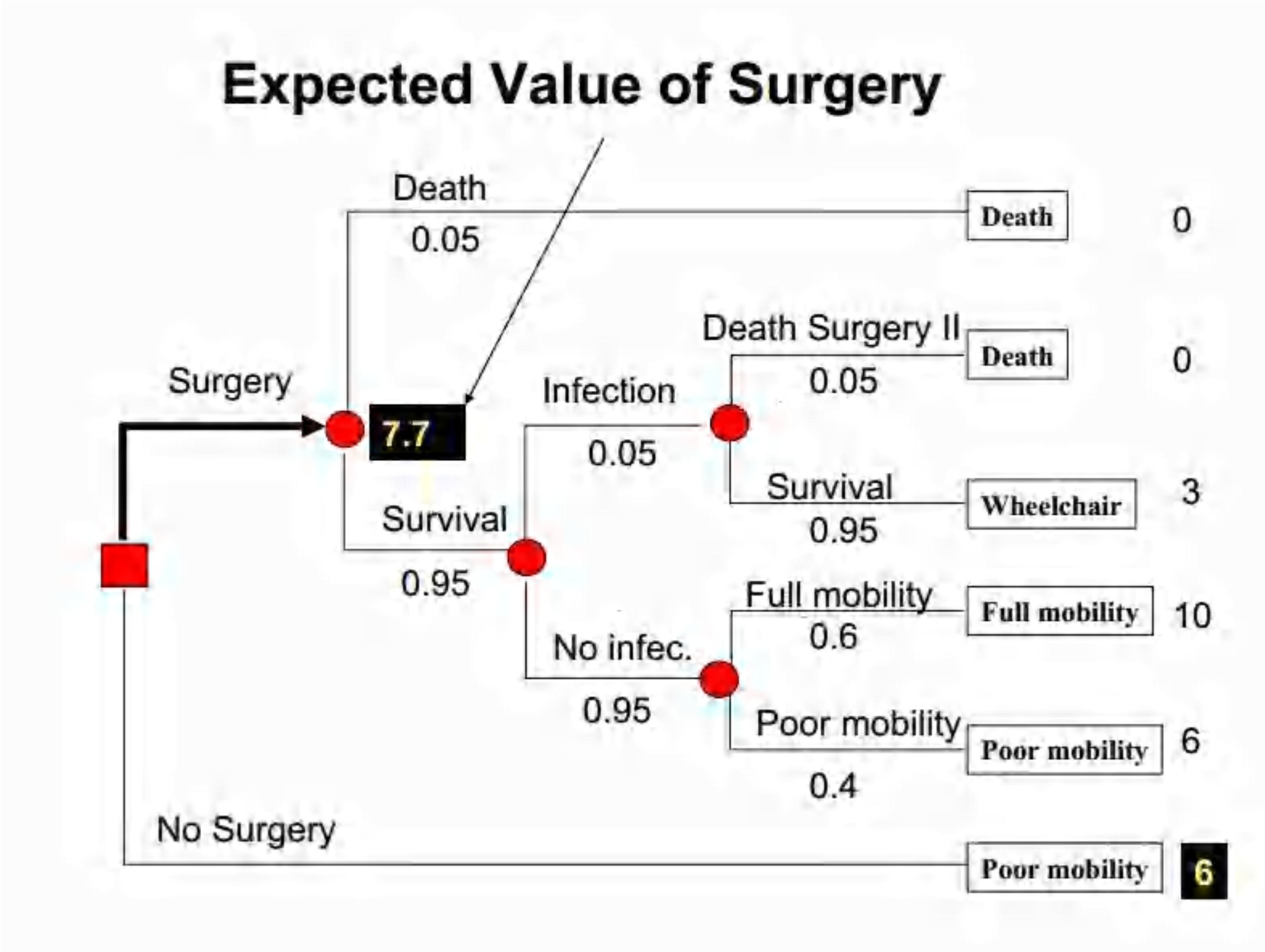
- Example of a Decision Problem
- Soccer player considering knee surgery
- Uncertainties:
- Success: recovering full mobility
- Risks: infection in surgery (if so, needs another surgery and may loose more mobility)
- Survival chances of surgery

Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support, Fall 2005
Instructors: Professor Lucila Ohno-Machado and Professor Staal Vinterbo

Decision Tree (this is known since Hippocrates!)



Helps to make rational decisions (risks vs. success)



Remember: Expected Utility Theory $E\ (U|d)$

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

The expected utility of decision $D = d$ is

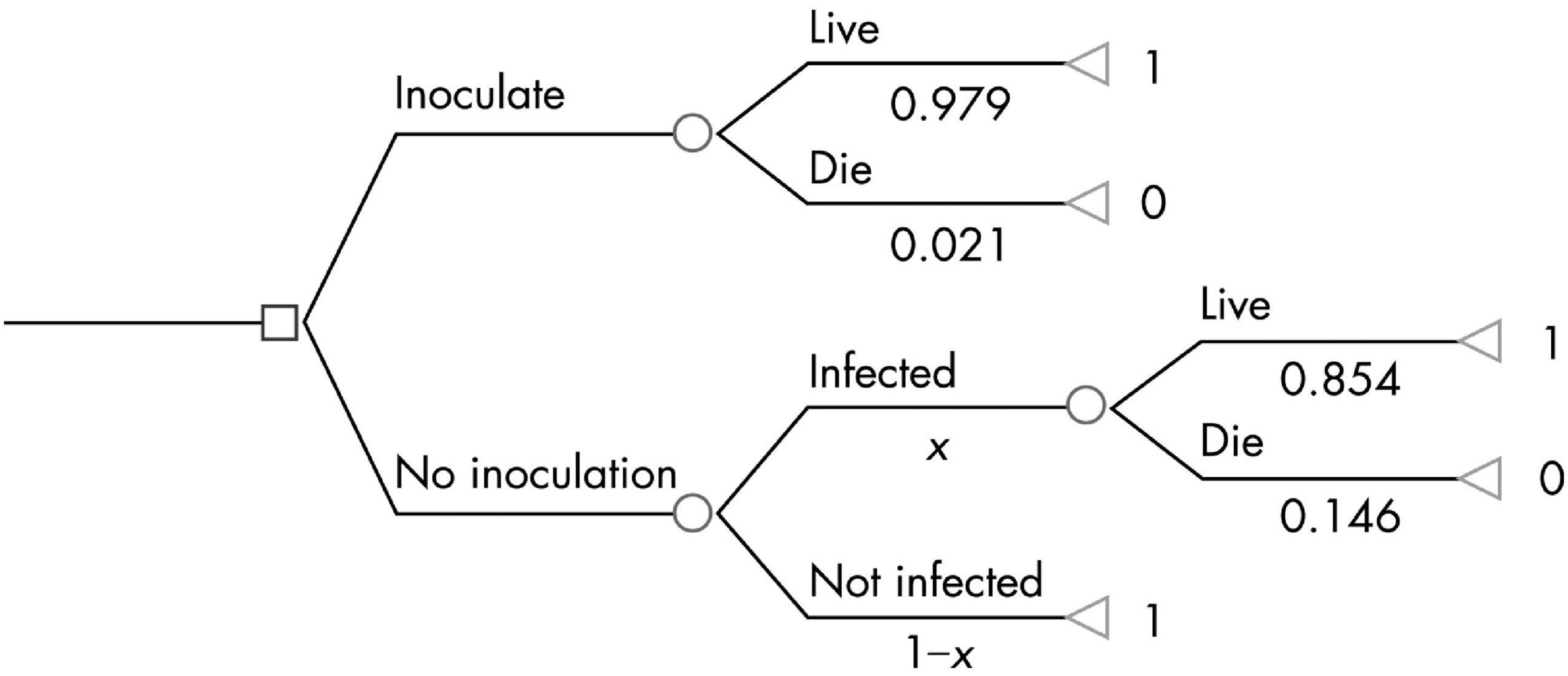
http://www.eoht.info/page/Oskar+Morgenstern

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

An optimal single decision is the decision $D = dmax$ whose expected utility is maximal:

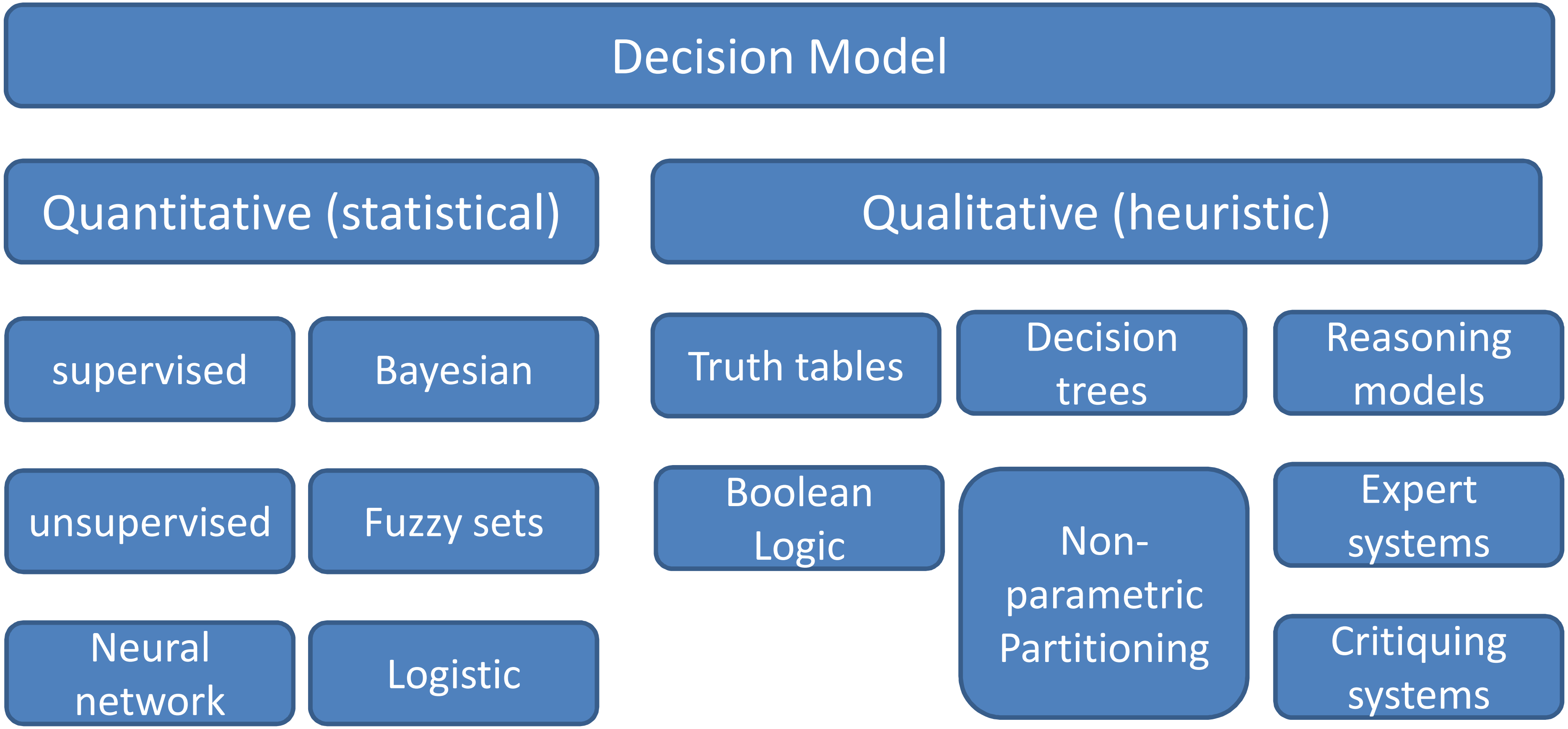
$$d_{\max} = \arg \max_{d \in \mathrm{dom}(D)} E(U \mid d)$$

Von Neumann, J. & Morgenstern, O. 1947. Theory of games and economic behavior, Princeton university press.



Ferrando, A., Pagano, E., Scaglione, L., Petrinco, M., Gregori, D. & Ciccone, G. (2009) A decision-tree model to estimate the impact on cost-effectiveness of a venous thromboembolism prophylaxis guideline. *Quality and Safety in Health Care, 18, 4, 309-313.*

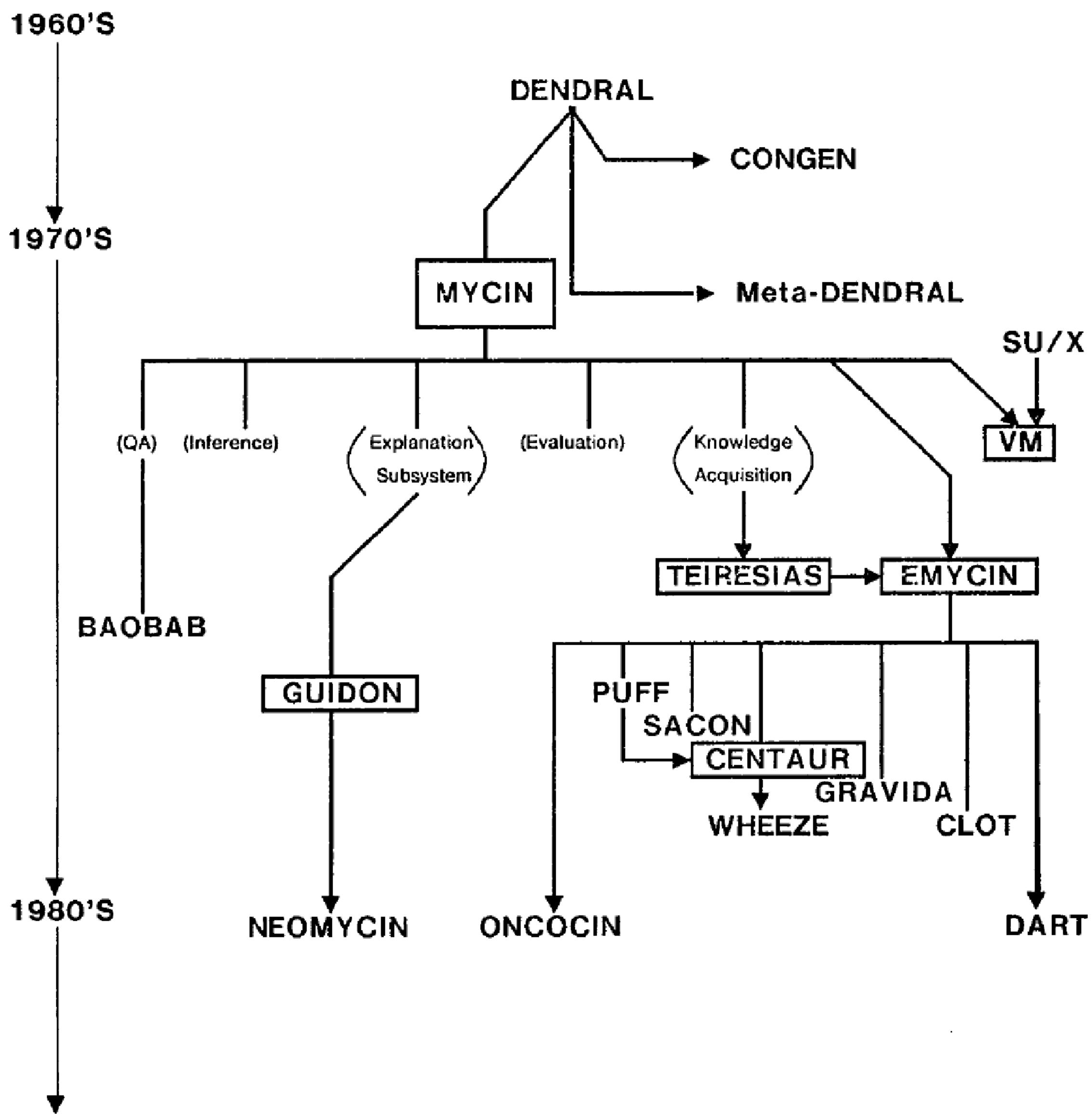
Taxonomy of Decision Support Models



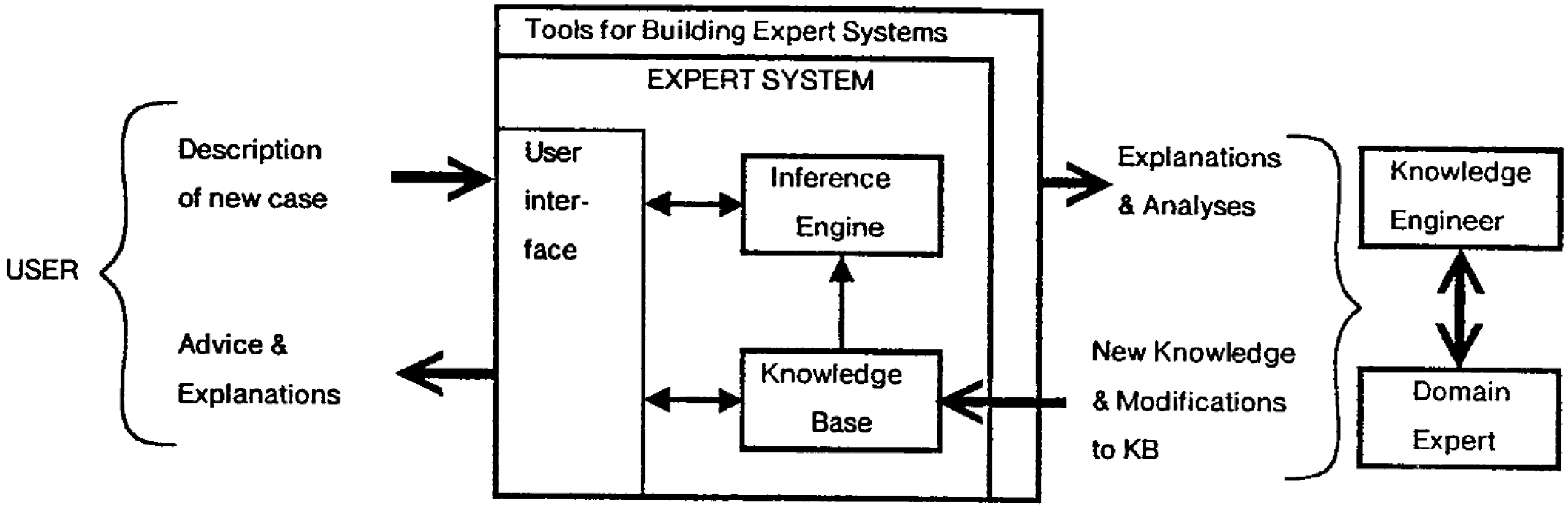
Extended by A. Holzinger after: Bemmel, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*

03 History of DSS = History of AI

- **1943** McCulloch, W.S. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biology, 5, (4), 115-133, doi:10.1007/BF02459570.
- **1950** Turing, A.M. Computing machinery and intelligence. Mind, 59, (236), 433-460.
- **1959** Samuel, A.L. Some studies in machine learning using the game of checkers. IBM Journal of research and development, 3, (3), 210-229, doi:10.1147/rd.33.0210.
- **1975** Shortliffe, E.H. & Buchanan, B.G. 1975. A model of inexact reasoning in medicine. Mathematical biosciences, 23, (3-4), 351-379, doi:10.1016/0025-5564(75)90047-4.

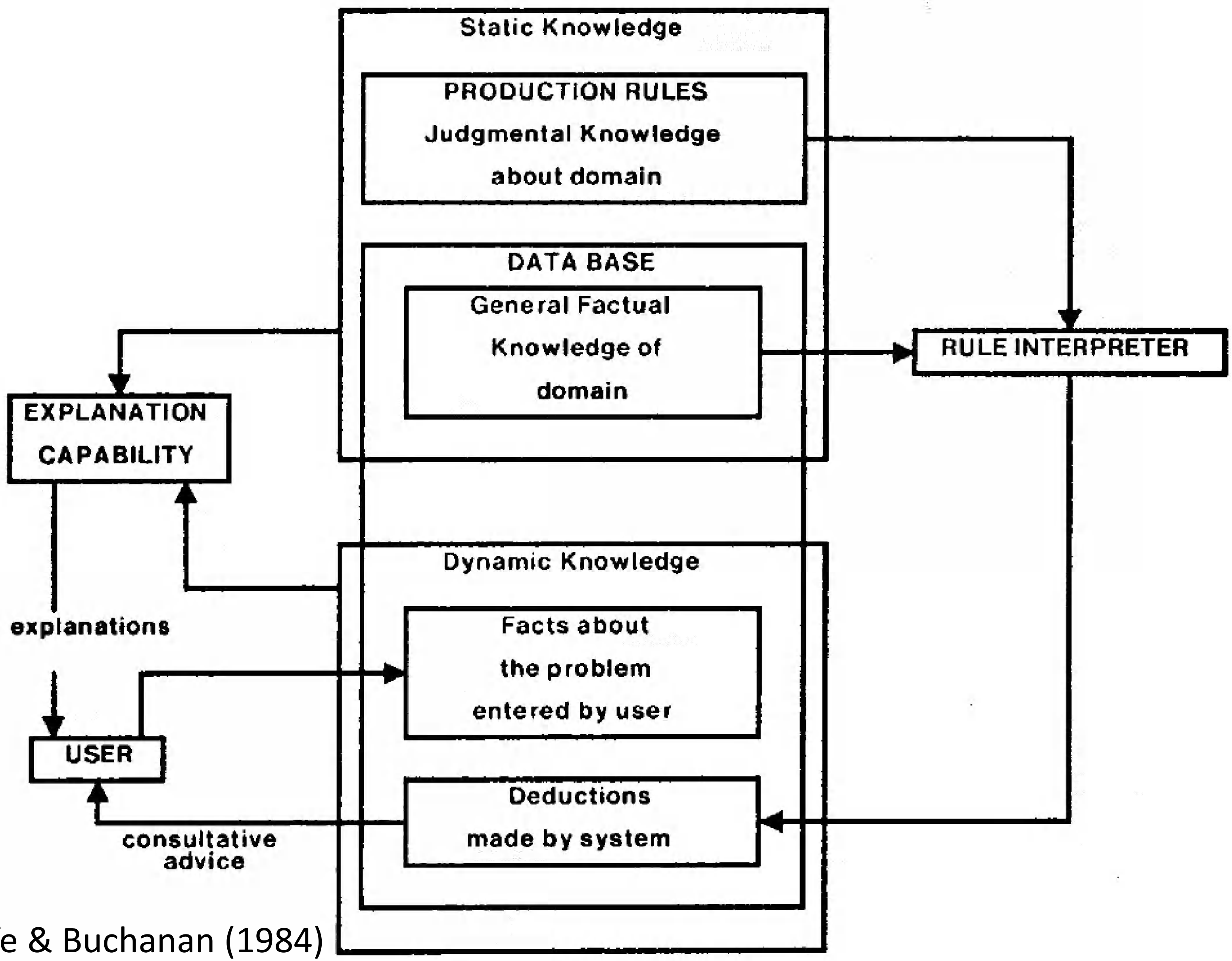


Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*



Shortliffe, T. & Davis, R. (1975) Some considerations for the implementation of knowledge-based expert systems *ACM SIGART Bulletin, 55, 9-12.*

Static Knowledge versus dynamic knowledge



Shortliffe & Buchanan (1984)

- The information available to humans is often imperfect – imprecise - uncertain.
- This is especially in the medical domain the case.
- An **human agent** can cope with deficiencies.
- Classical logic permits only **exact reasoning**:
- IF A is true THEN A is non-false and
  IF B is false THEN B is non-true
- Most real-world problems do not provide this exact information, mostly it is inexact, incomplete, uncertain and/or **un-measurable!**

- MYCIN is a rule-based Expert System, which is used for therapy planning for patients with bacterial infections
- Goal oriented strategy ("Rückwärtsverkettung")
- To every rule and every entry a certainty factor (CF) is assigned, which is between 0 und 1
- Two measures are derived:
- MB: measure of belief
- MD: measure of disbelief
- Certainty factor – CF of an element is calculated by:

$$CF[h] = MB[h] - MD[h]$$

- CF is positive, if more evidence is given for a hypothesis, otherwise CF is negative
- CF[h] = +1 -> h is 100 % true
- CF[h] = –1 -> h is 100% false

Original Example from MYCIN

h_1 = The identity of ORGANISM-1 is streptococcus
h_2 = PATIENT-1 is febrile
h_3 = The name of PATIENT-1 is John Jones

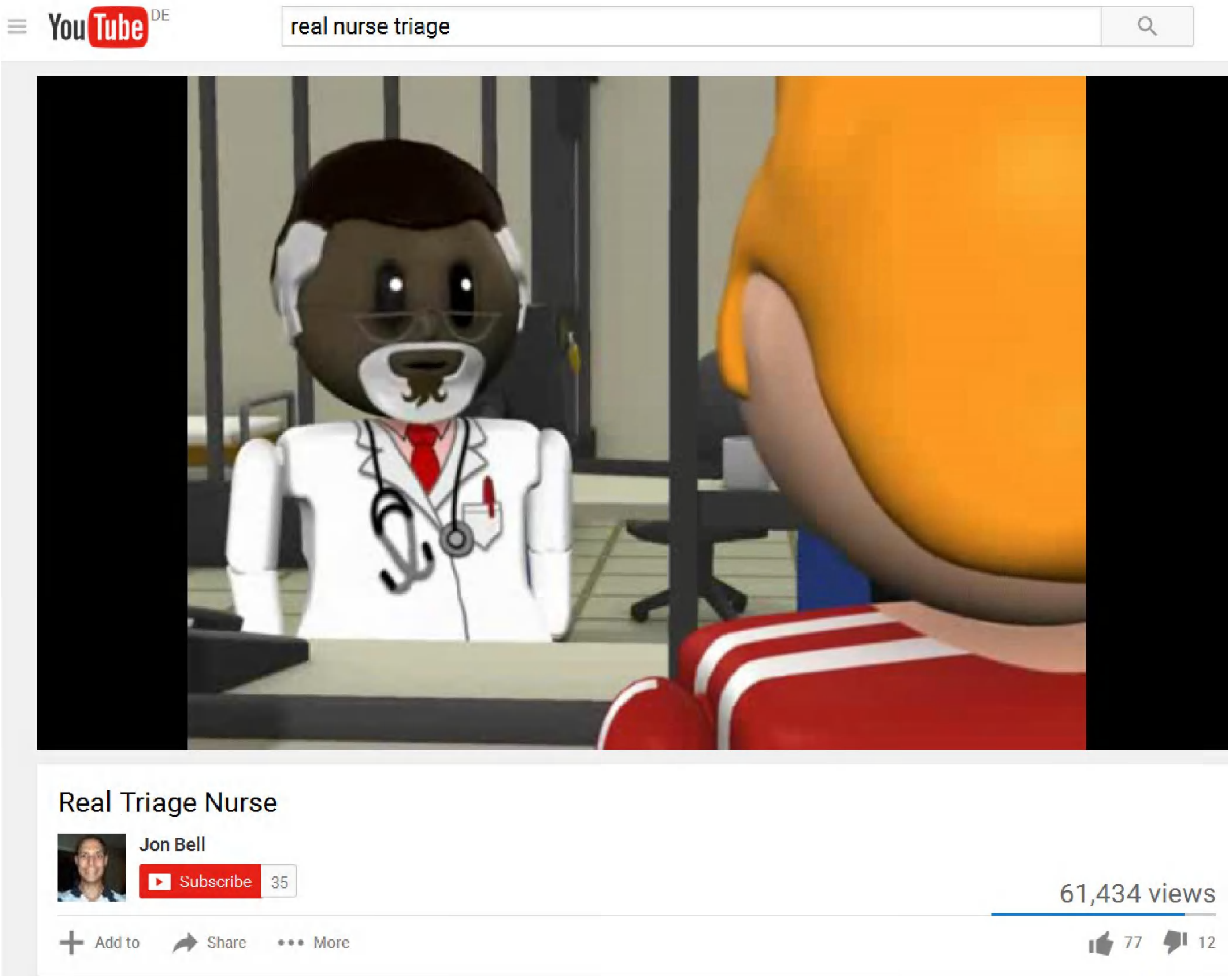
$CF[h_1,E] = .8$: There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus

$CF[h_2,E] = -.3$: There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile

$CF[h_3,E] = +1$: It is definite (1) that the name of PATIENT-1 is John Jones

Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*

MYCIN was *no* success in the clinical routine



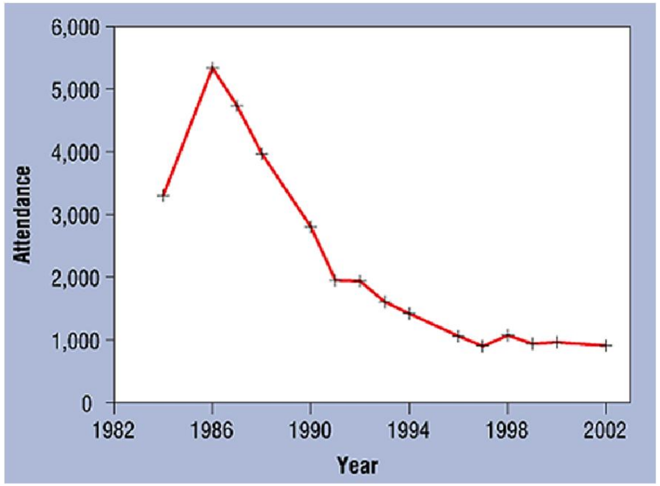
However, AI was extremely popular in the 1970ies

Image credit to Bernhard Schölkopf

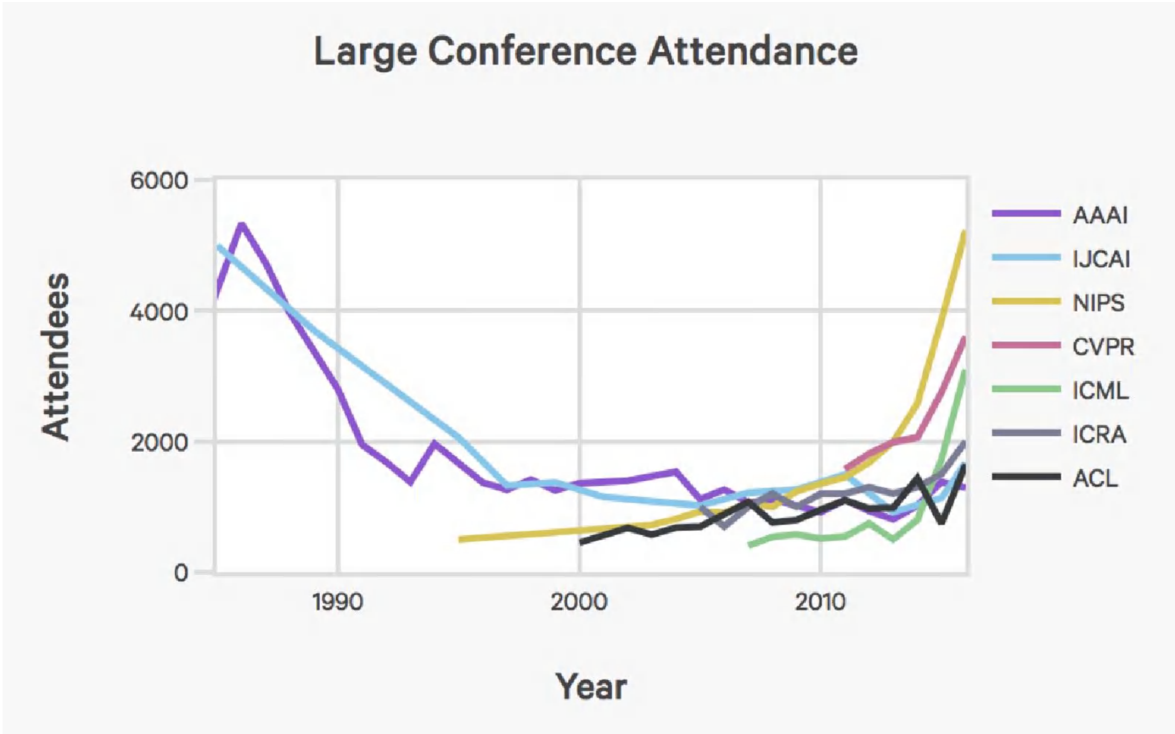
Cybernetics was praised as the solution for everything

Image credit to Bernhard Schölkopf

https://blogs.dxc.technology/2017/04/25/are-we-heading-toward-an-ai-winter/



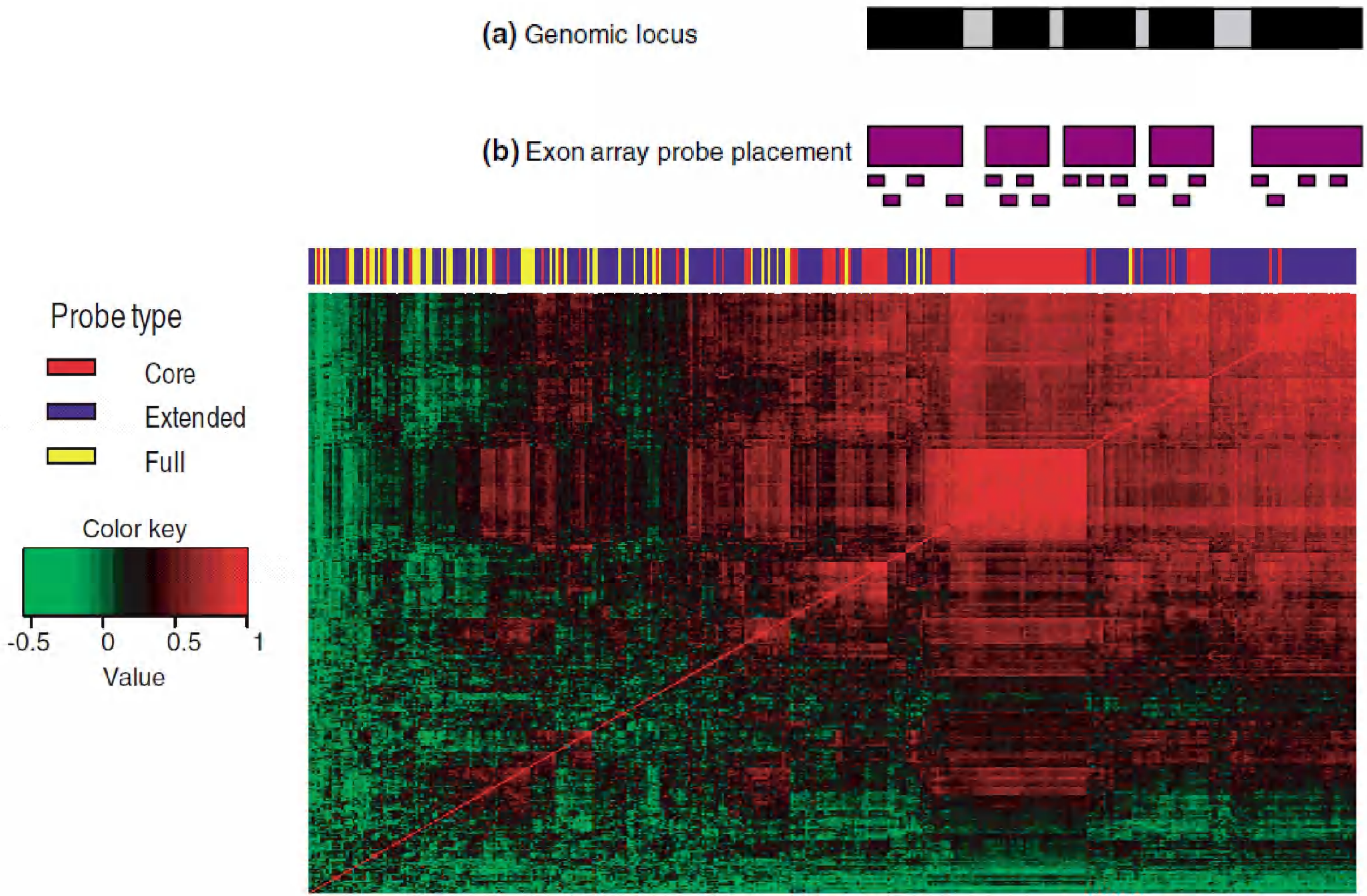
https://www.computer.org/csl/mags/ex/2003/03/x3018.html



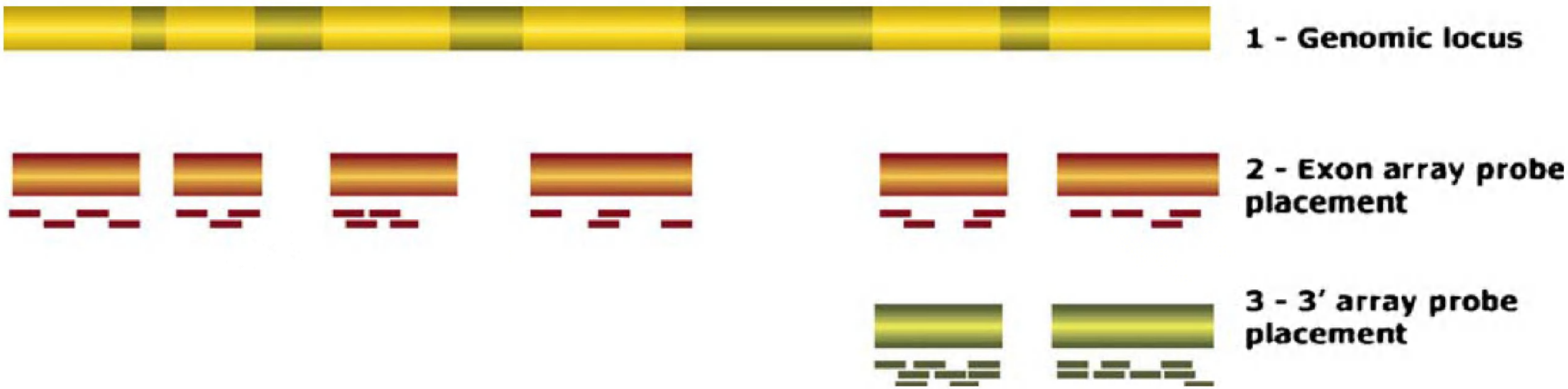
https://medium.com/machine-learning-in-practice/nips-accepted-papers-stats-26f124843aa0

04 Example: P4-Medicine

Slide 8-22 Example: Exon Arrays



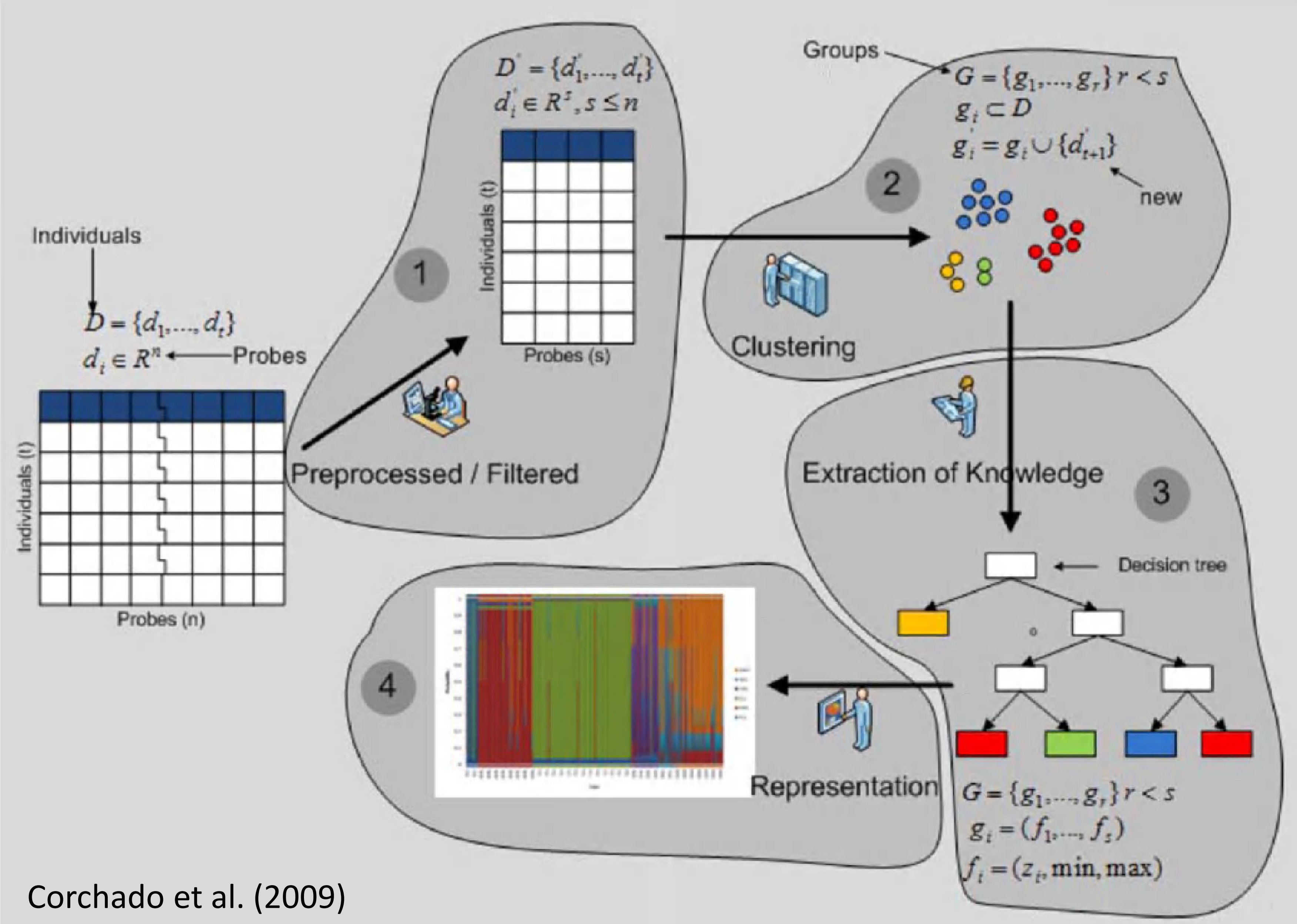
Kapur, K., Xing, Y., Ouyang, Z. & Wong, W. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biology, 8, 5, R82.*



Exon array structure. Probe design of exon arrays. (1) Exon—intron structure of a gene. Gray boxes represent introns, rest represent exons. Introns are not drawn to scale. (2) Probe design of exon arrays. Four probes target each putative exon. (3) Probe design of 30expression arrays. Probe target the 30end of mRNA sequence.

Corchado, J. M., De Paz, J. F., Rodriguez, S. & Bajo, J. (2009) Model of experts for decision support in the diagnosis of leukemia patients. *Artificial Intelligence in Medicine, 46, 3, 179-200.*

Slide 8-24 Computational leukemia cancer detection 2/6

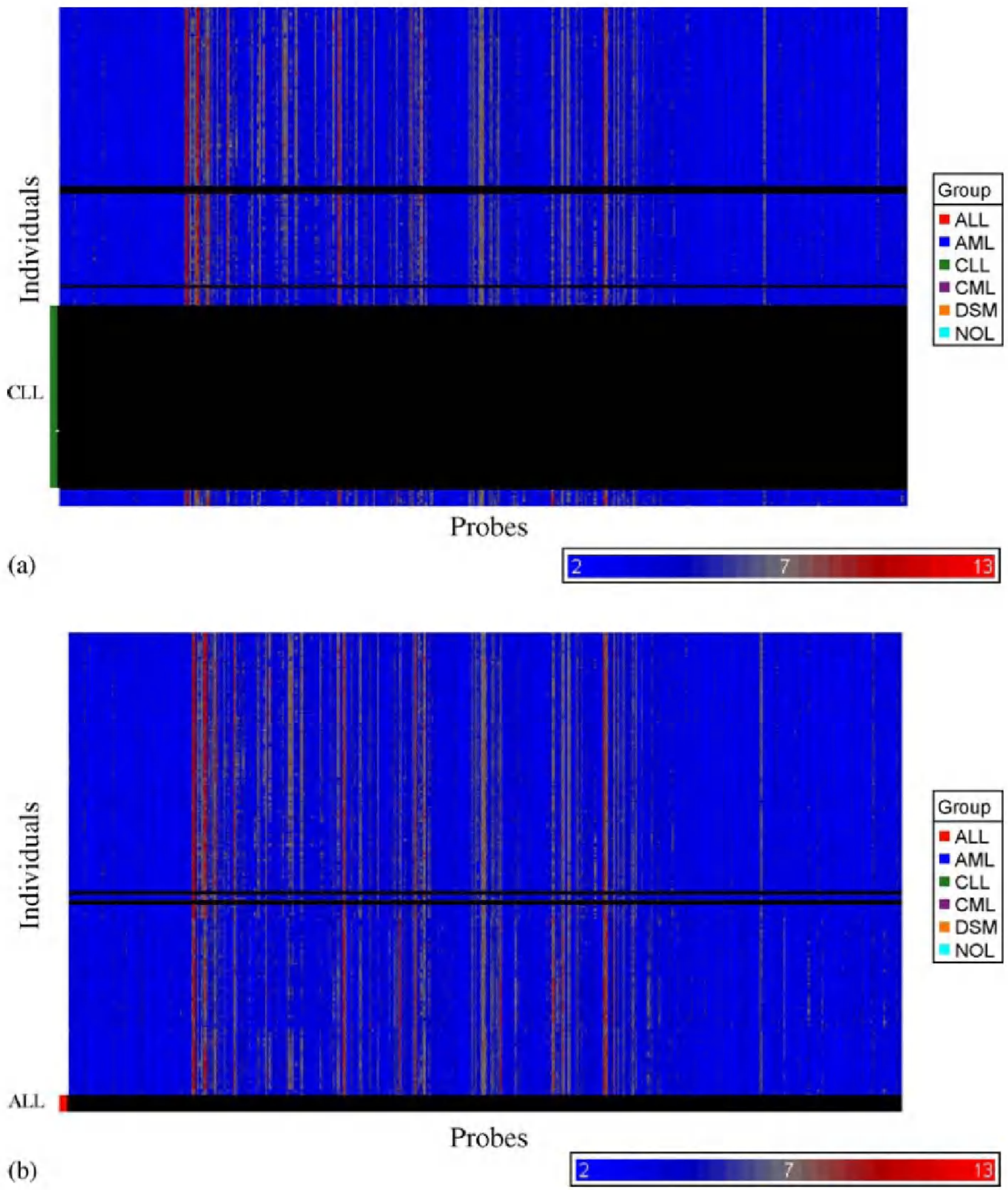


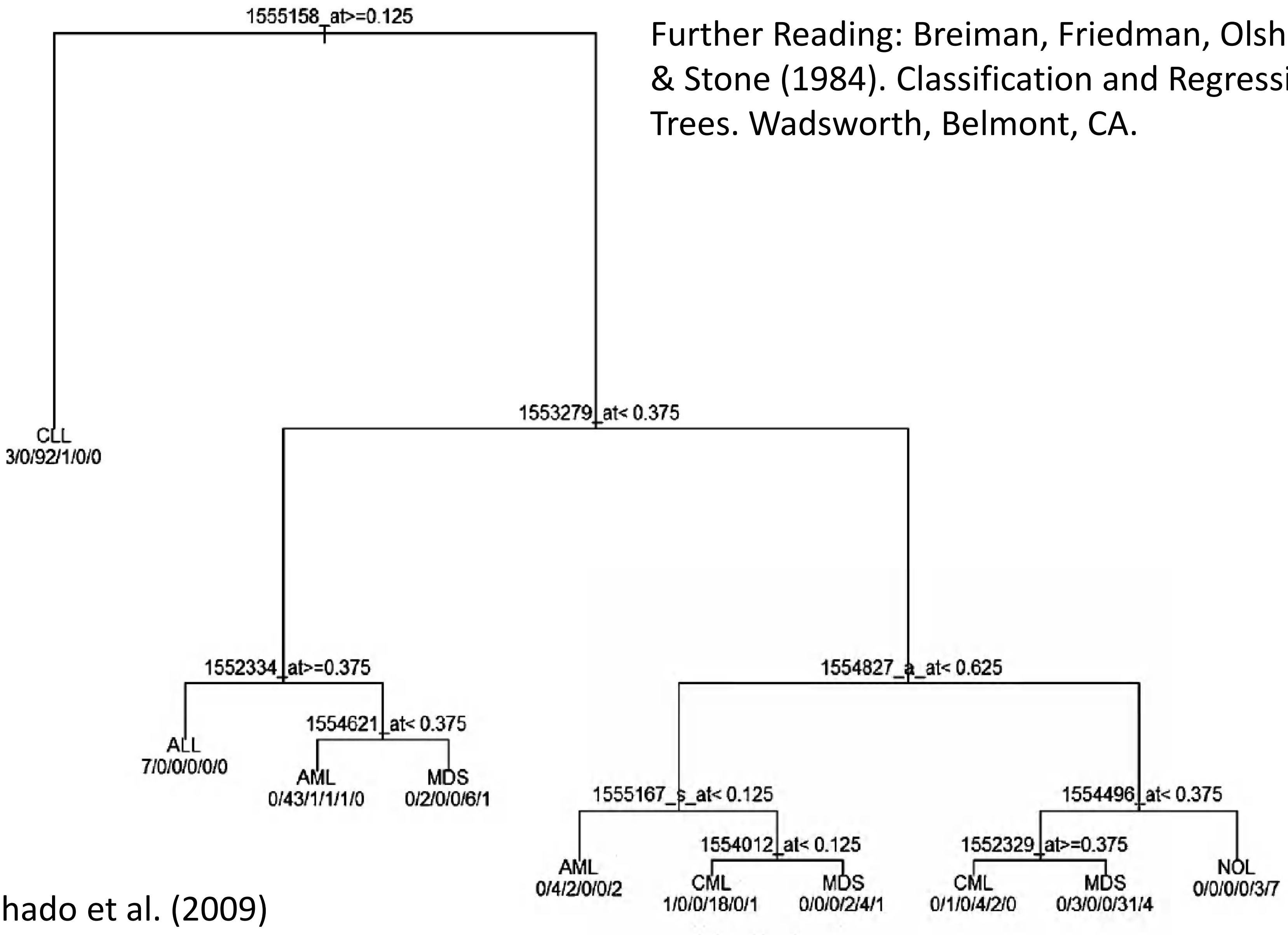
Corchado et al. (2009)

A = acute, C = chronic,
L = lymphocytic, M = myeloid

- **ALL** = cancer of the blood AND bone marrow caused by an abnormal proliferation of lymphocytes.
- **AML** = cancer in the bone marrow characterized by the proliferation of myeloblasts, red blood cells or abnormal platelets.
- **CLL** = cancer characterized by a proliferation of lymphocytes in the bone marrow.
- **CML** = caused by a proliferation of white blood cells in the bone marrow.
- **MDS** (Myelodysplastic Syndromes) = a group of diseases of the blood and bone marrow in which the bone marrow does not produce a sufficient amount of healthy cells.
- **NOL** (Normal) = No leukemias

Corchado et al. (2009)

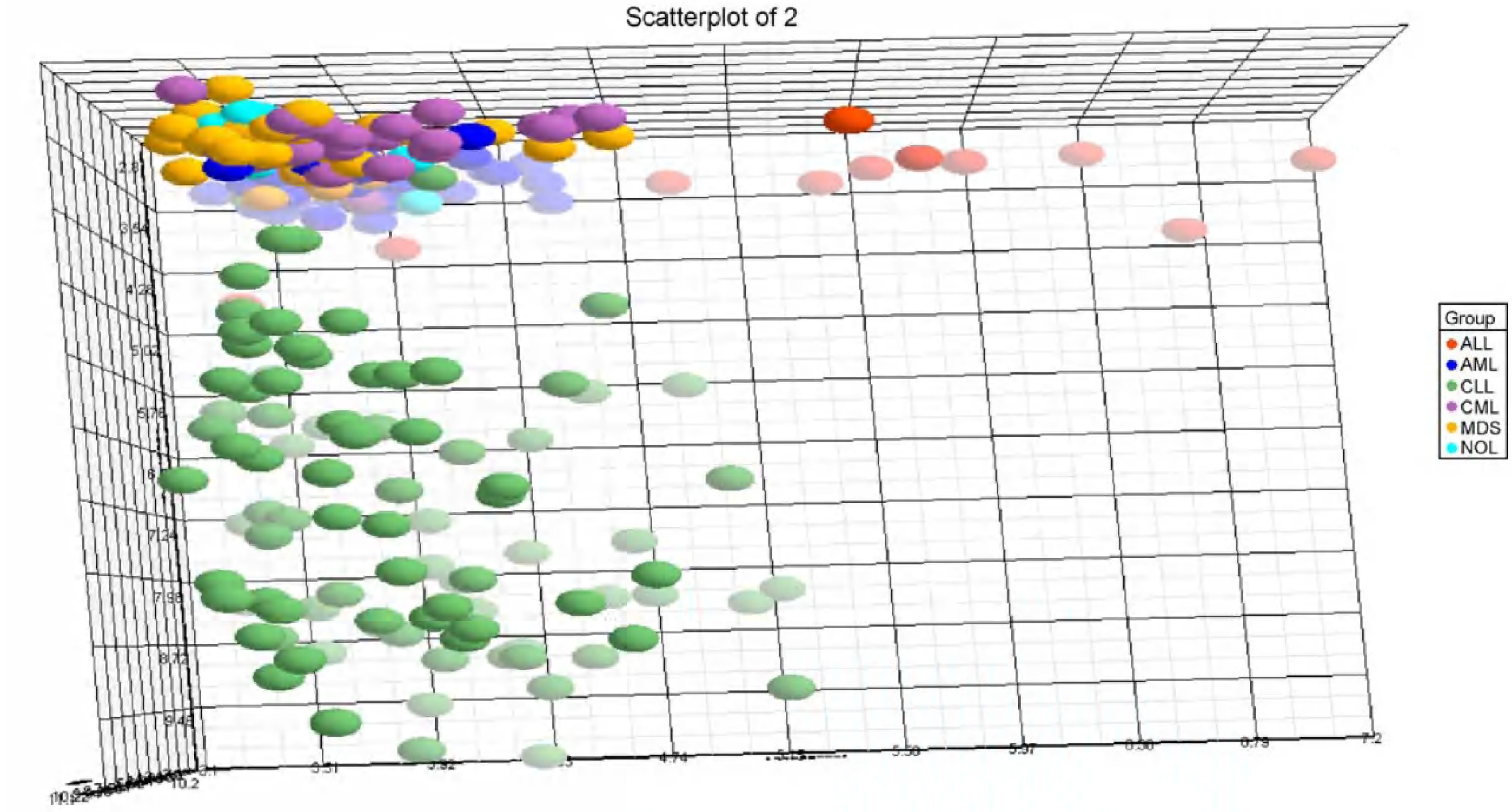




Further Reading: Breiman, Friedman, Olshen, & Stone (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.

Corchado et al. (2009)

Classification CLL—ALL. Representation of the probes of the decision tree which classify the CLL and ALL to 1555158_at, 1553279_at and 1552334_at



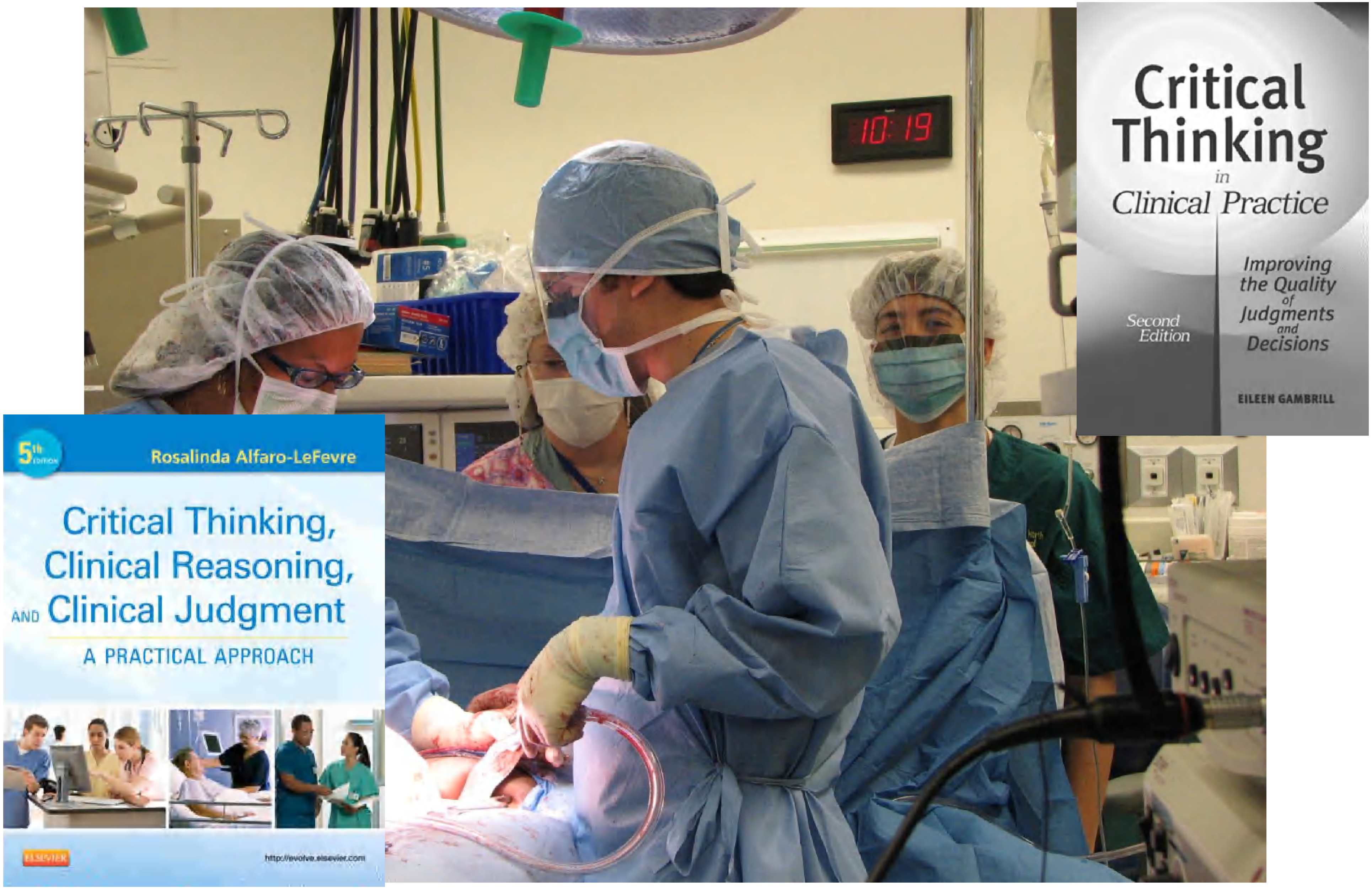
Corchado et al. (2009)

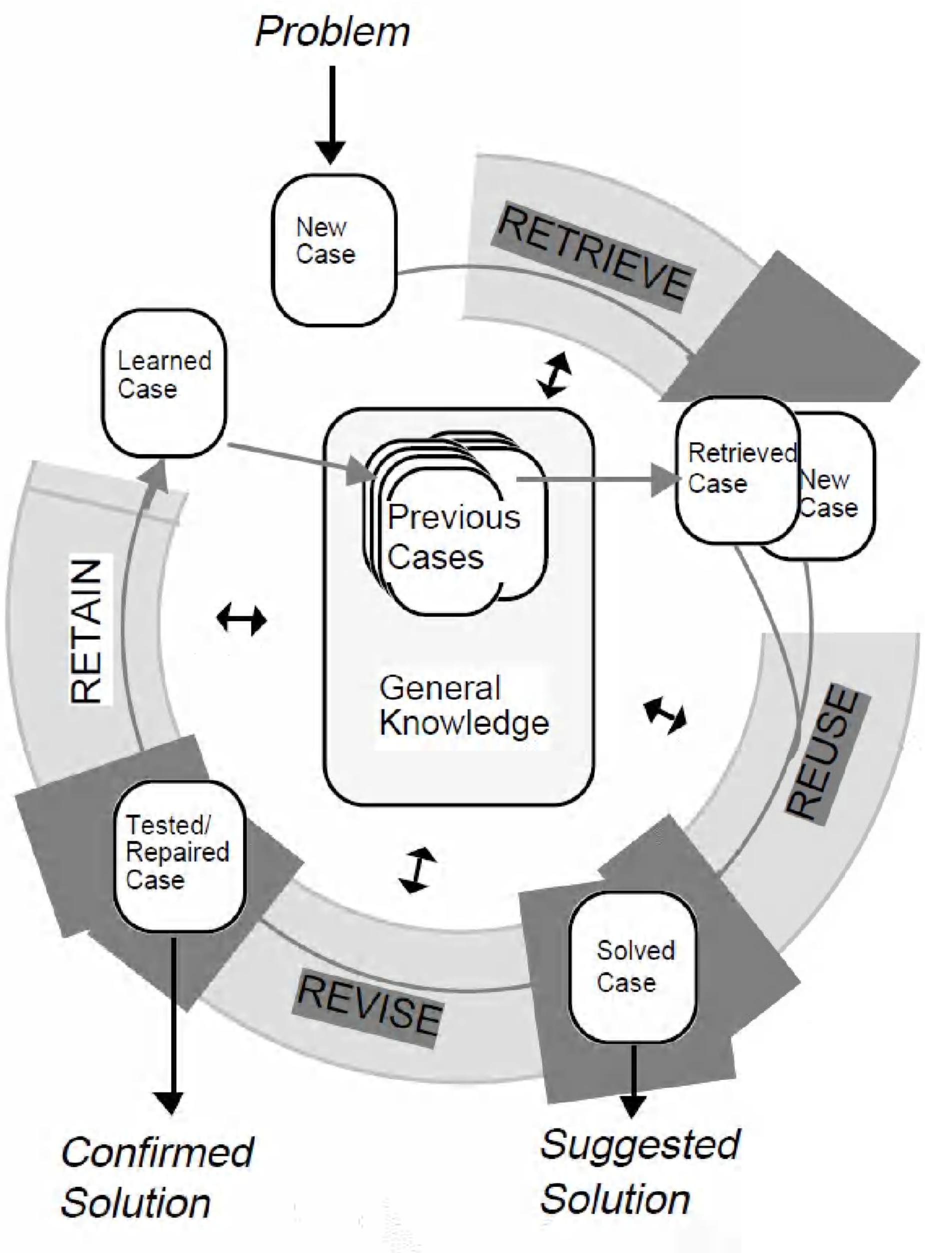
Computational leukemia cancer detection 6/6

- The model of Corchado et al. (2009) combines:
- 1) methods to **reduce the dimensionality** of the original data set;
- 2) pre-processing and data filtering techniques;
- 3) a clustering method to classify patients; and
- 4) extraction of knowledge techniques
- The system reflects how human experts work in a lab, but
- 1) **reduces the time** for making predictions;
- 2) **reduces the rate of human error;** and
- 3) **works with high-dimensional data** from exon arrays

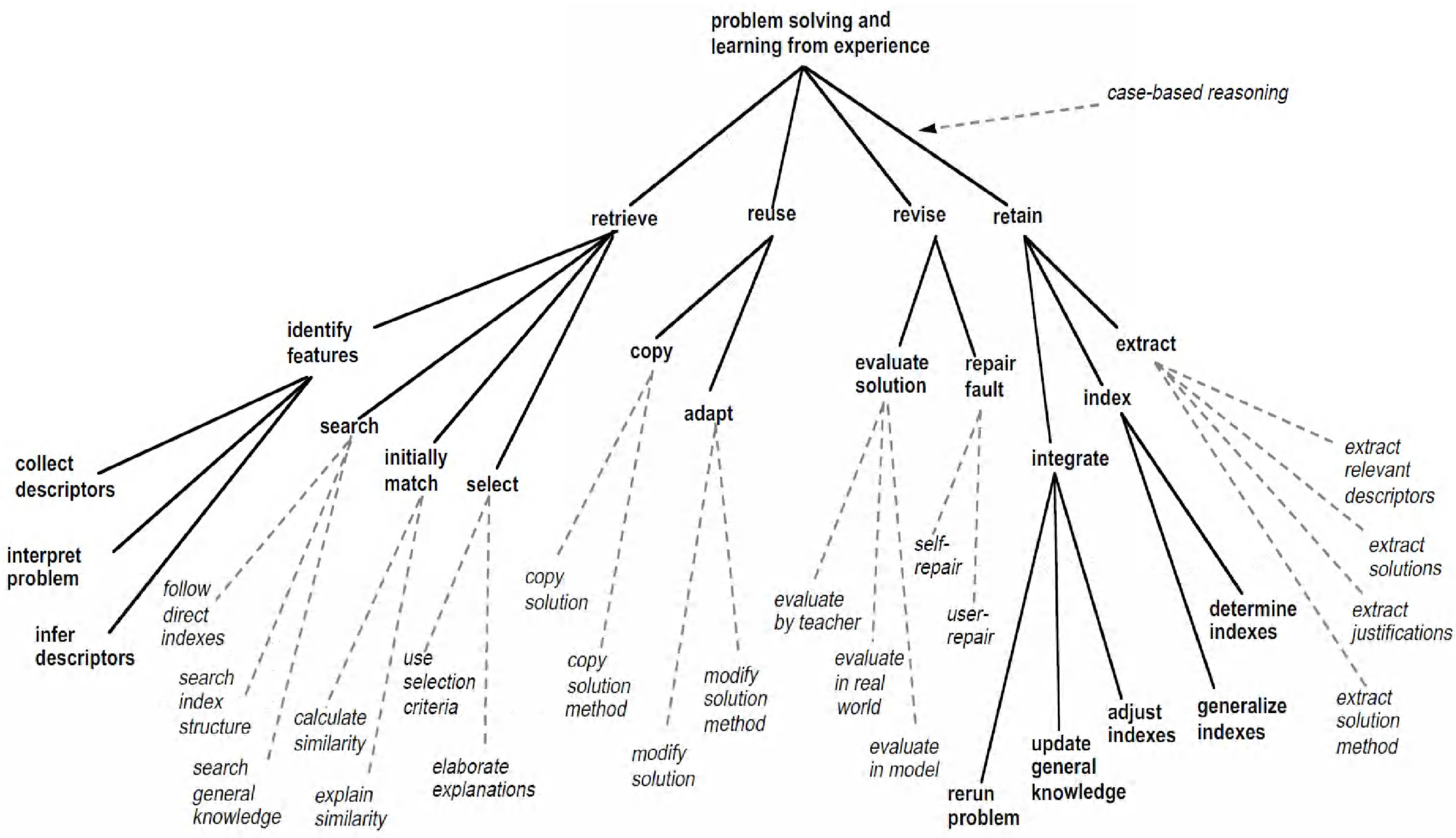
05 Example: Case Based Reasoning (CBR)

Slide 8-29 Thinking – Reasoning – Deciding – Acting





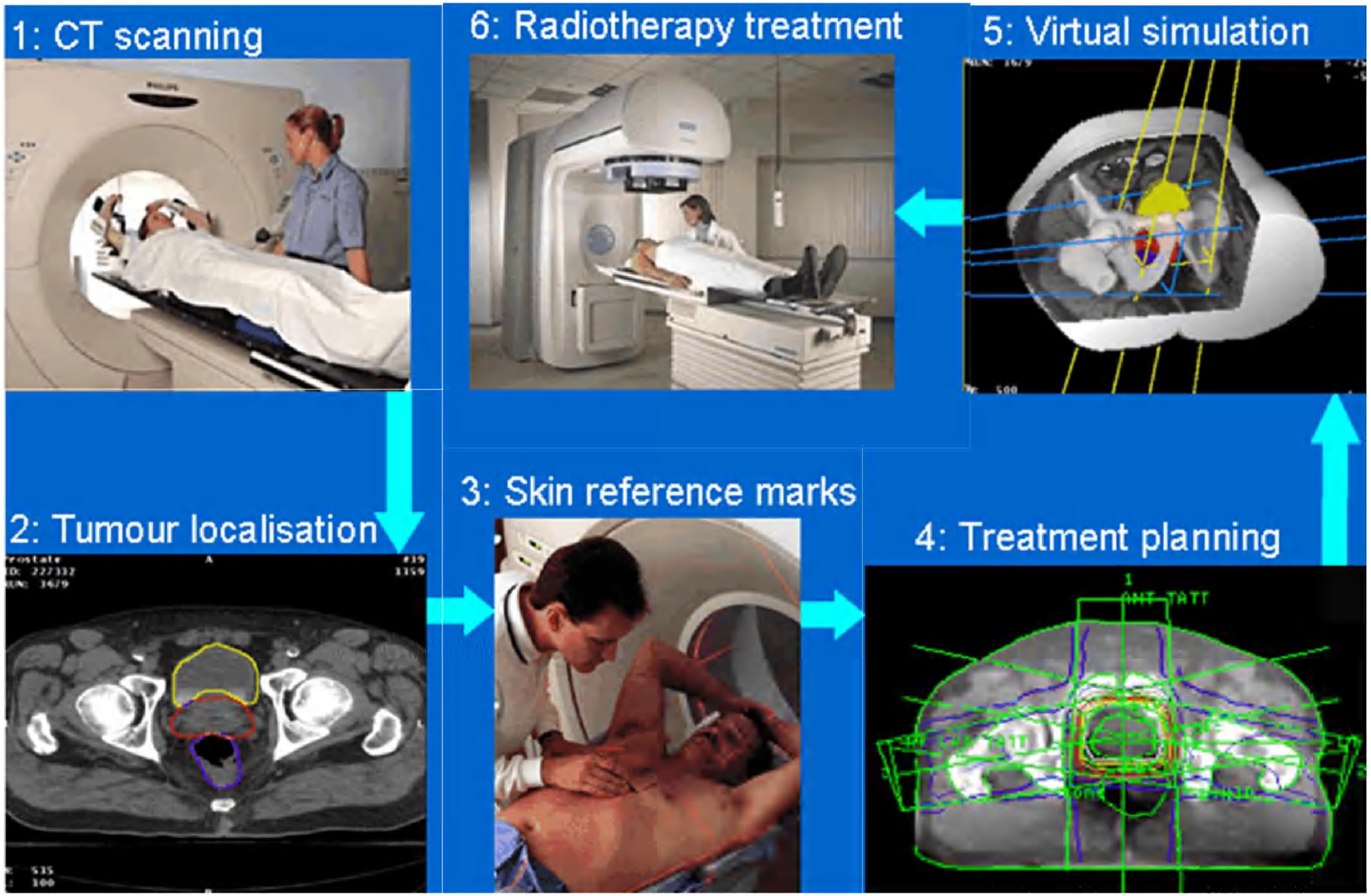
Aamodt, A. & Plaza, E. (1994) Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Communications, 7, 1, 39-59.*



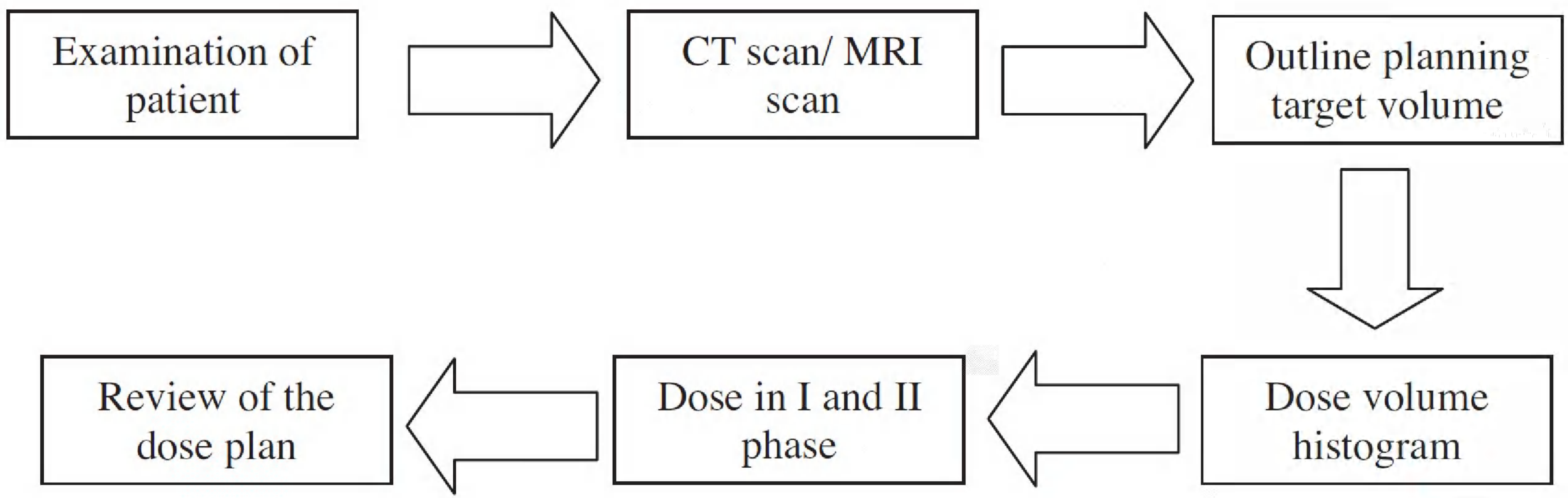
Aamodt & Plaza (1994)

Slide 8-32 CBR Example: Radiotherapy Planning 1/6

Source: http://www.teachingmedicalphysics.org.uk



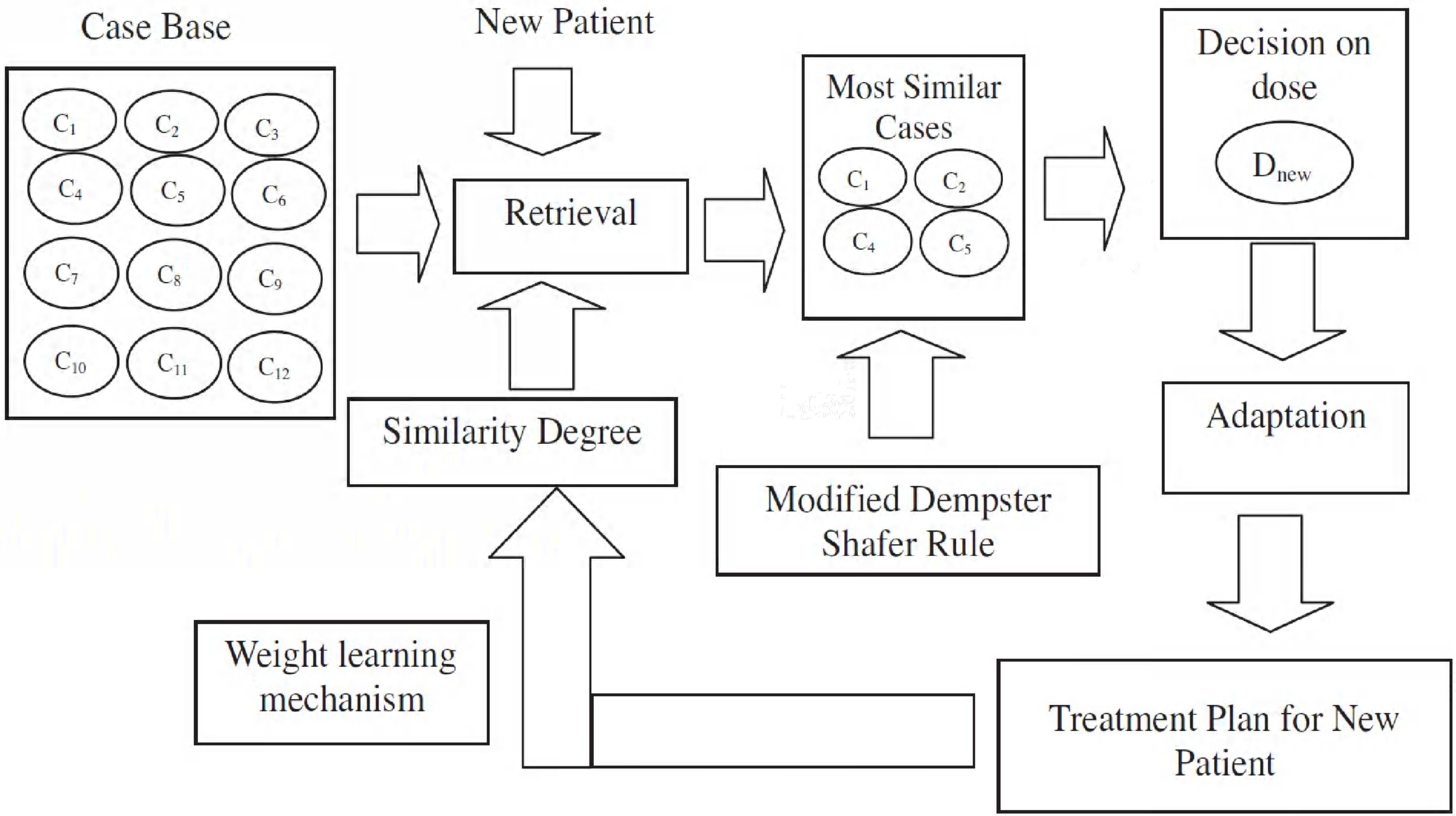
Source: Imaging Performance Assessment of CT Scanners Group, http://www.impactscan.org



Measures:

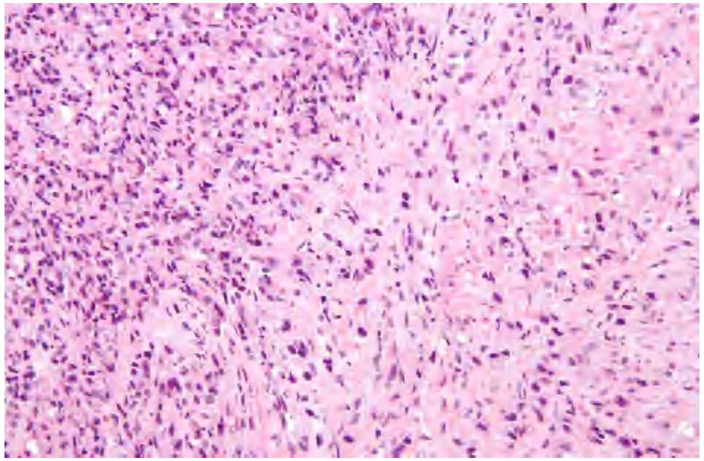
1) Clinical Stage = a labelling system
2) Gleason Score = grade of prostate cancer = integer between 1 to 10; and
3) Prostate Specific Antigen (PSA) value between 1 to 40
4) Dose Volume Histogram (DVH) = pot. risk to the rectum (66, 50, 25, 10 %)

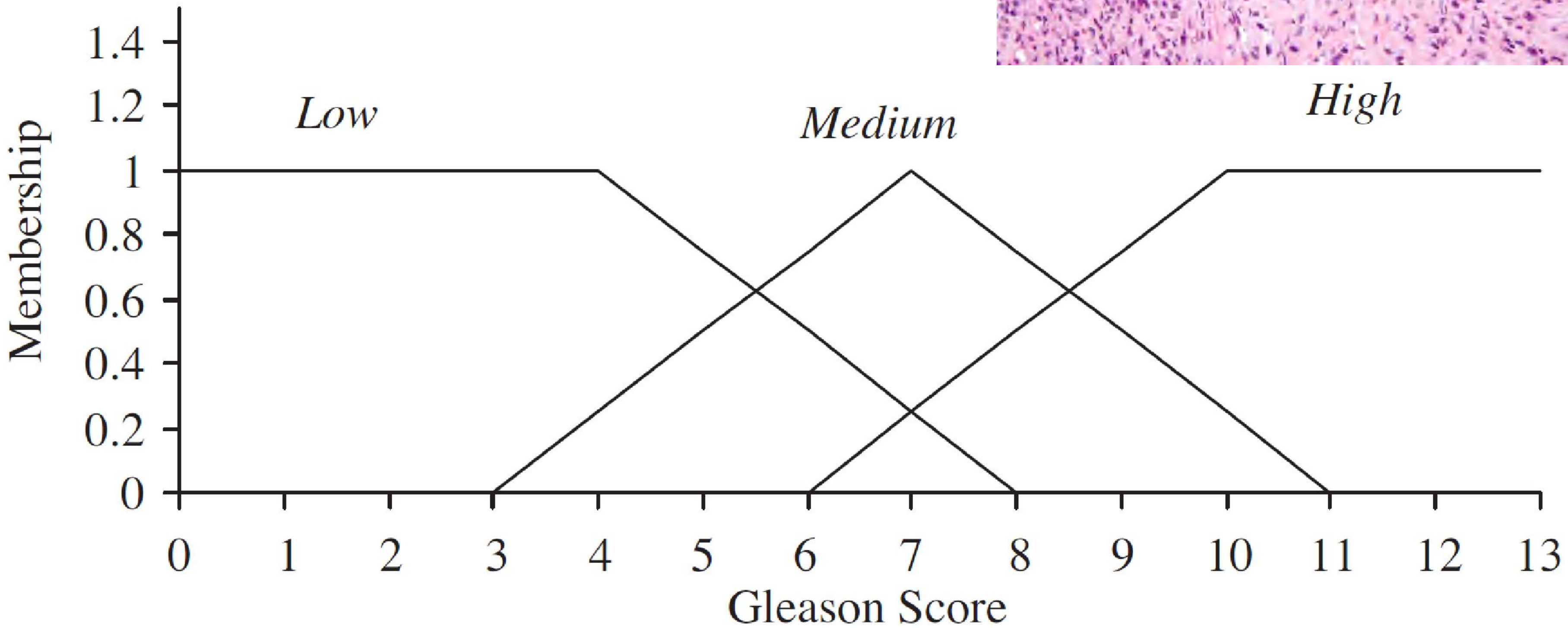
Petrovic, S., Mishra, N. & Sundar, S. (2011) A novel case based reasoning approach to radiotherapy planning. *Expert Systems With Applications, 38, 9, 10759-10769.*



Petrovic, S., Mishra, N. & Sundar, S. (2011) A novel case based reasoning approach to radiotherapy planning. *Expert Systems With Applications, 38, 9, 10759-10769.*

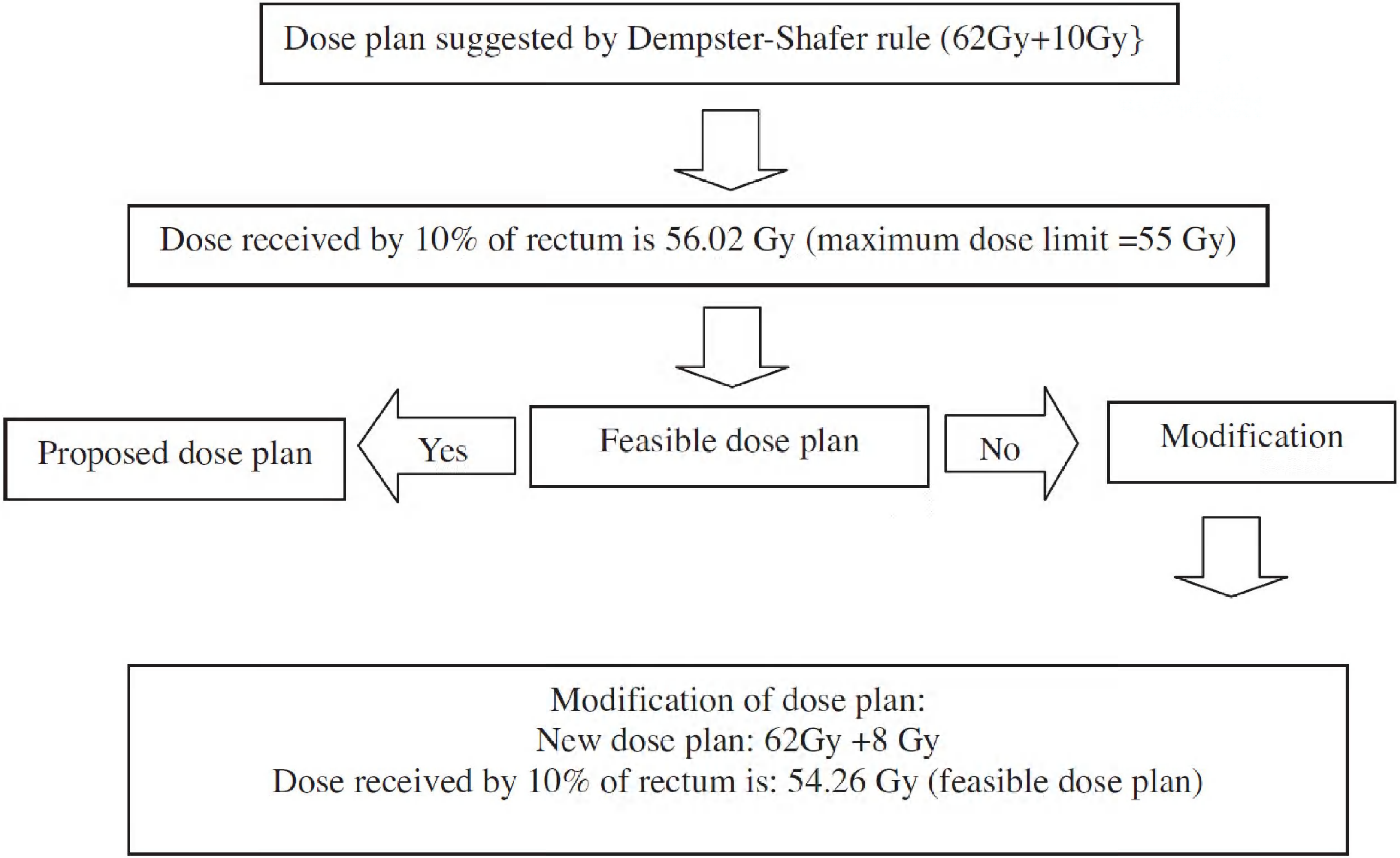
Gleason score evaluates the grade of prostate cancer. Values: integer within the range





Petrovic, S., Mishra, N. & Sundar, S. (2011) A novel case based reasoning approach to radiotherapy planning. *Expert Systems With Applications, 38, 9, 10759-10769.*

Petrovic et al. (2011)



06 Towards Explainable AI

Mastering the game of Go without human knowledge

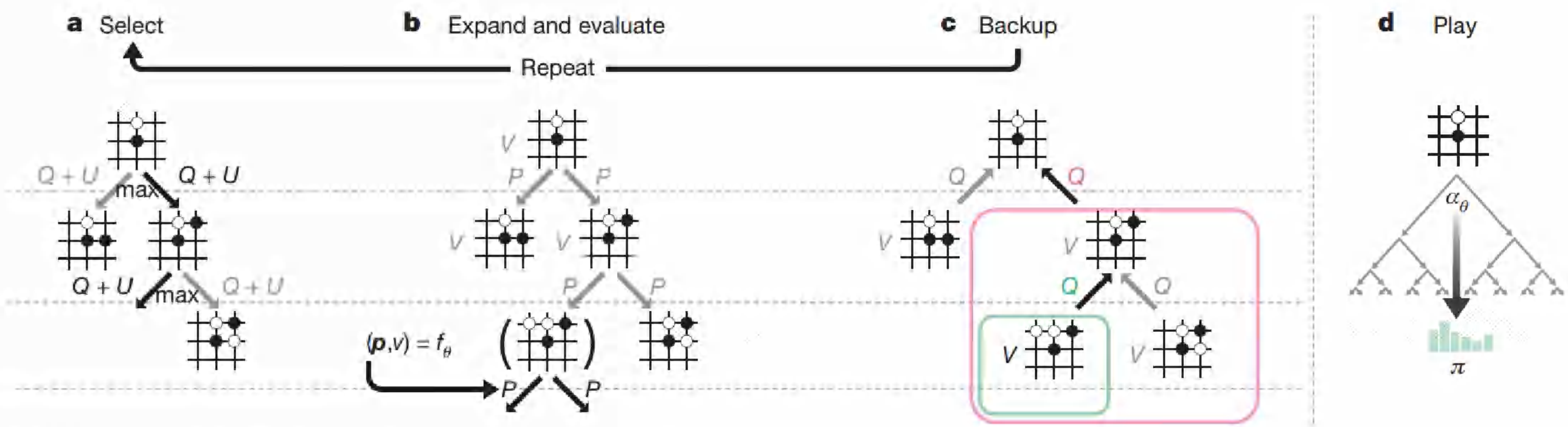


Figure 2 | MCTS in AlphaGo Zero. a, Each simulation traverses the tree by selecting the edge with maximum action value *Q*, plus an upper confidence bound *U* that depends on a stored prior probability *P* and visit count *N* for that edge (which is incremented once traversed). **b**, The leaf node is expanded and the associated position *s* is evaluated by the neural network $(P(s,\cdot),V(s))=f_\theta(s)$; the vector of *P* values are stored in the outgoing edges from *s*. **c**, Action value *Q* is updated to track the mean of all evaluations *V* in the subtree below that action. **d**, Once the search is complete, search probabilities π are returned, proportional to $N^{1/\tau}$, where *N* is the visit count of each move from the root state and τ is a parameter controlling temperature.

19 OCTOBER 2017 | VOL 550 | NATURE | 355

$$(\boldsymbol{p}, v) = f_\theta(s) \quad \text{and} \quad l = (z - v)^2 - \boldsymbol{\pi}^{\mathrm{T}} \log \boldsymbol{p} + c\|\theta\|^2$$

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Van Den Driessche, Thore Graepel & Demis Hassabis 2017. Mastering the game of go without human knowledge. Nature, 550, (7676), 354-359, doi:doi:10.1038/nature24270.

Why did it make this decision???

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel & Demis Hassabis 2016. Mastering the game of Go with deep neural networks and tree search. Nature, 529, (7587), 484-489, doi:10.1038/nature16961.

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)

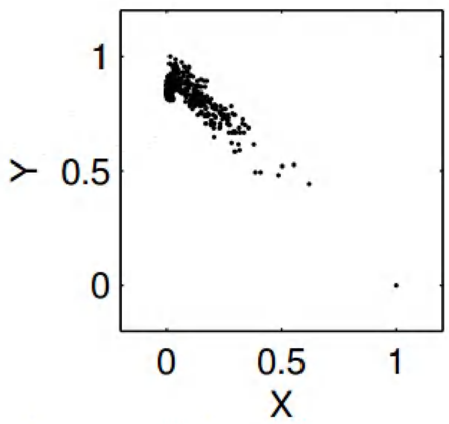
Example: Discovery of causal relationships from data ...

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

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

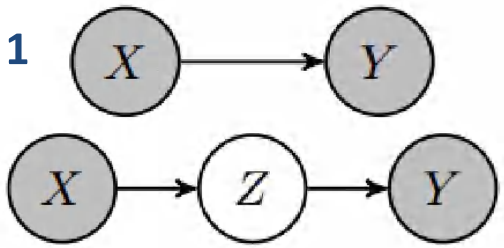
Lopez-Paz, D., Muandet, K., Schölkopf, B. & Tolstikhin, I. 2015. Towards a learning theory of cause-effect inference. Proceedings of the 32nd International Conference on Machine Learning, JMLR, Lille, France.

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



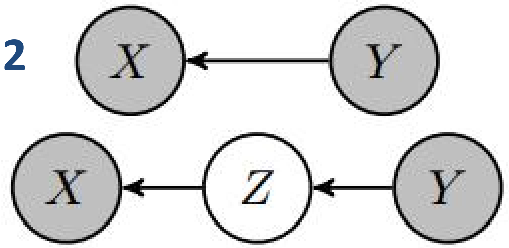
Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf 2016. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17, (1), 1103-1204.

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

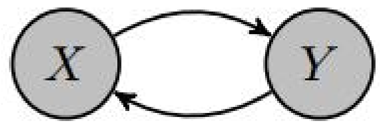


$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

3

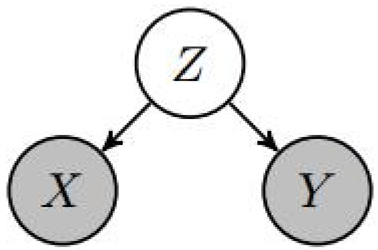
$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

4



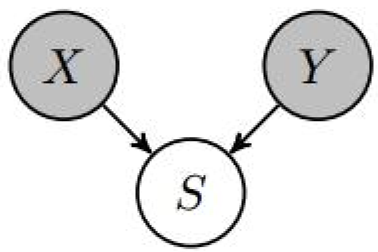
$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

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



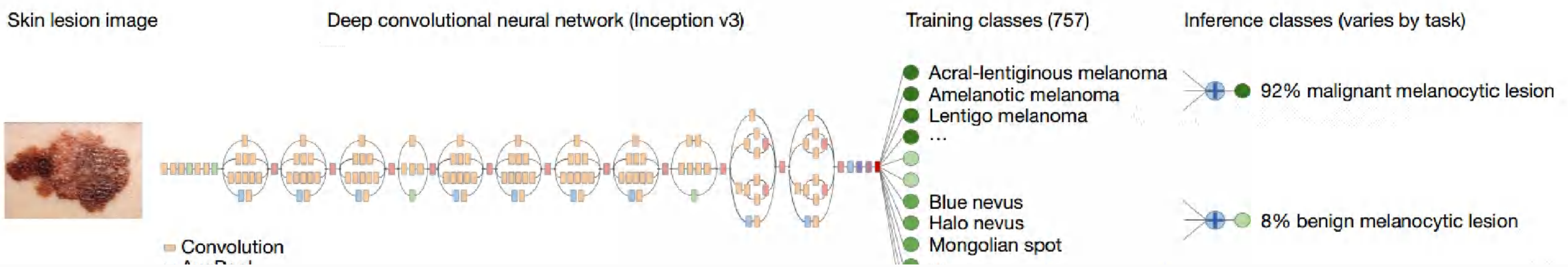
$$\mathbb{P}_{Y\,|\,s} \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x),s} = \mathbb{P}_{Y\,|\,x,s}$$
$$\mathbb{P}_{X\,|\,s} \neq \mathbb{P}_{X\,|\,\mathrm{do}(y),s} = \mathbb{P}_{X\,|\,y,s}$$

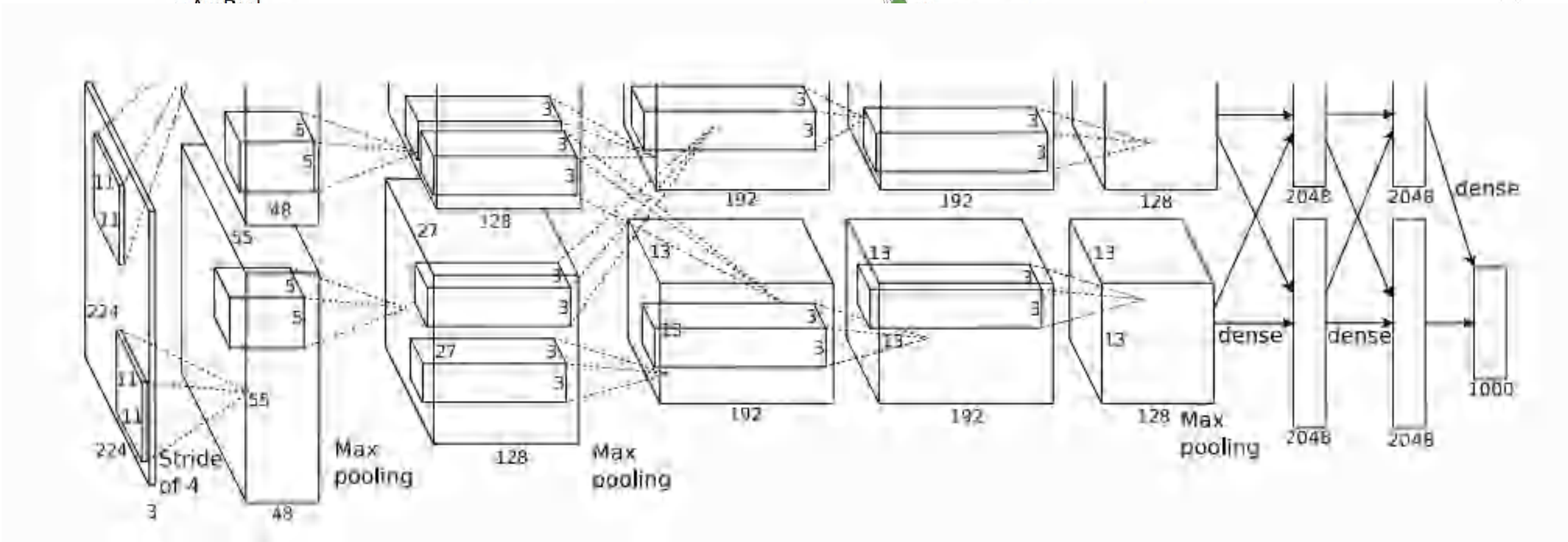
15b

- **"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 $Y \rightarrow X$

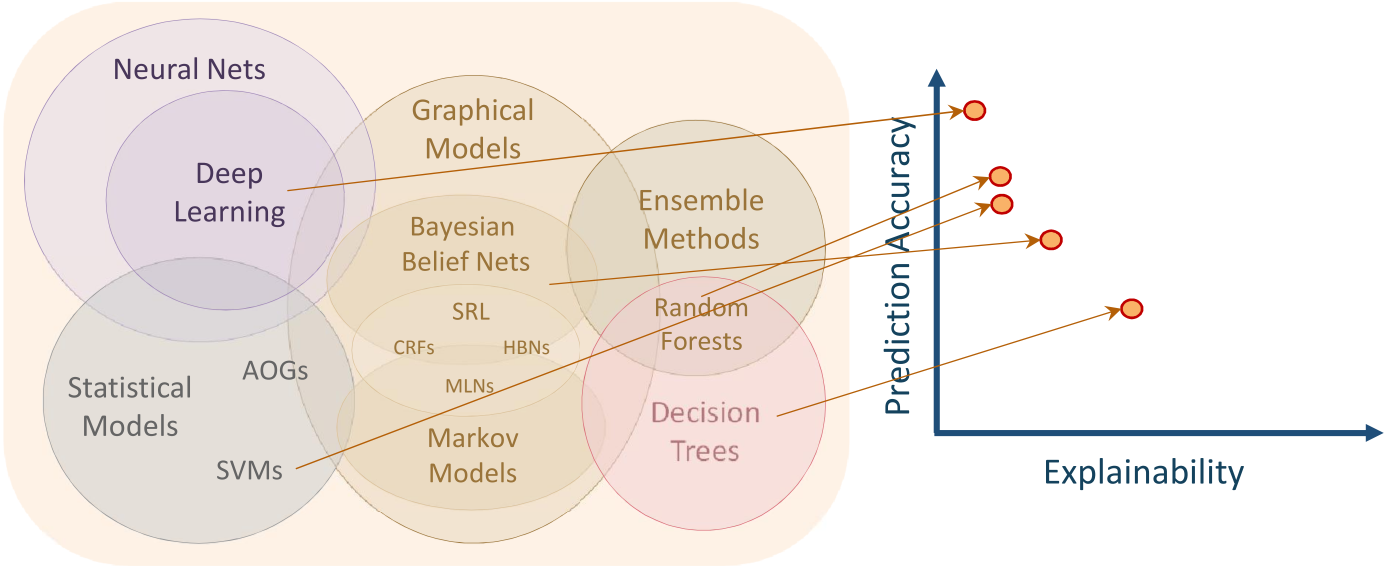
Bengio, Y., Courville, A. & Vincent, P. 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35, (8), 1798-1828, doi:10.1109/TPAMI.2013.50.

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.

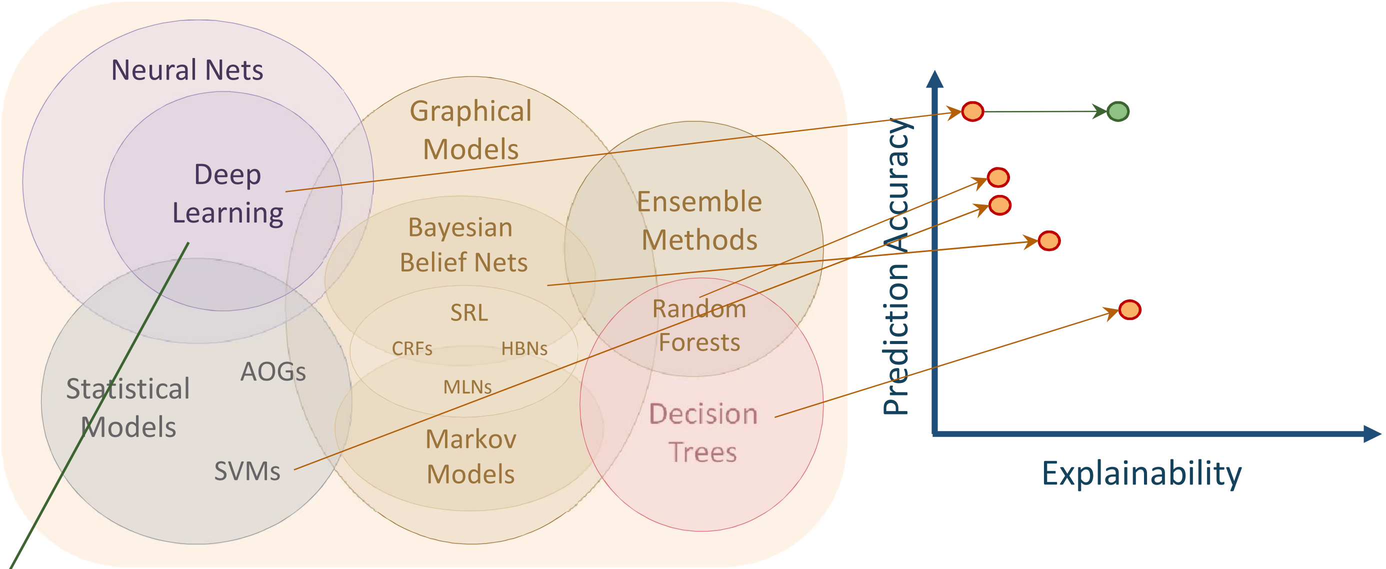


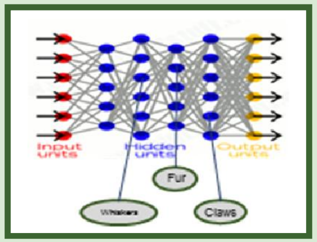


Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q., eds. Advances in neural information processing systems (NIPS 2012), 2012 Lake Tahoe. 1097-1105.



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

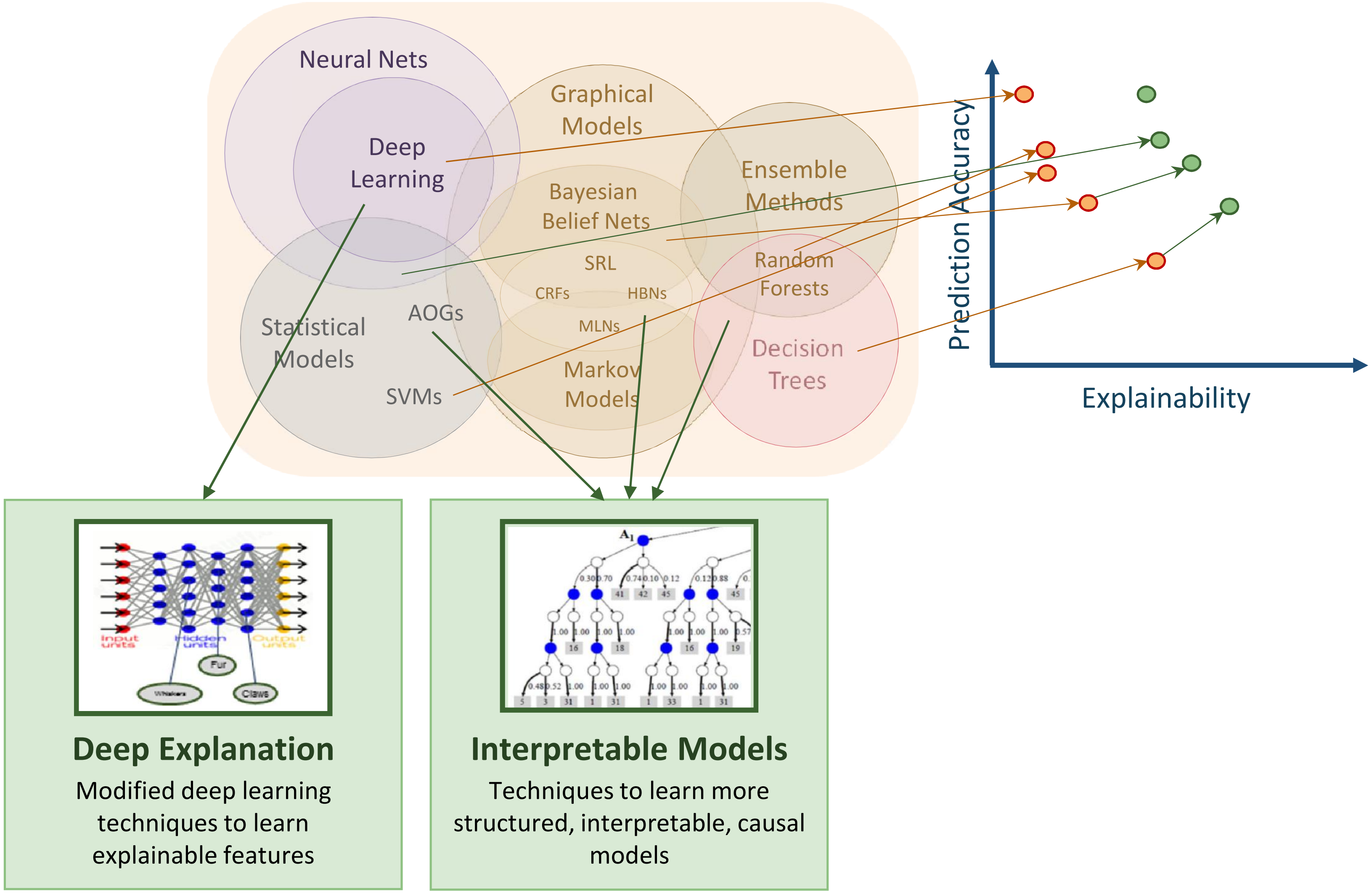




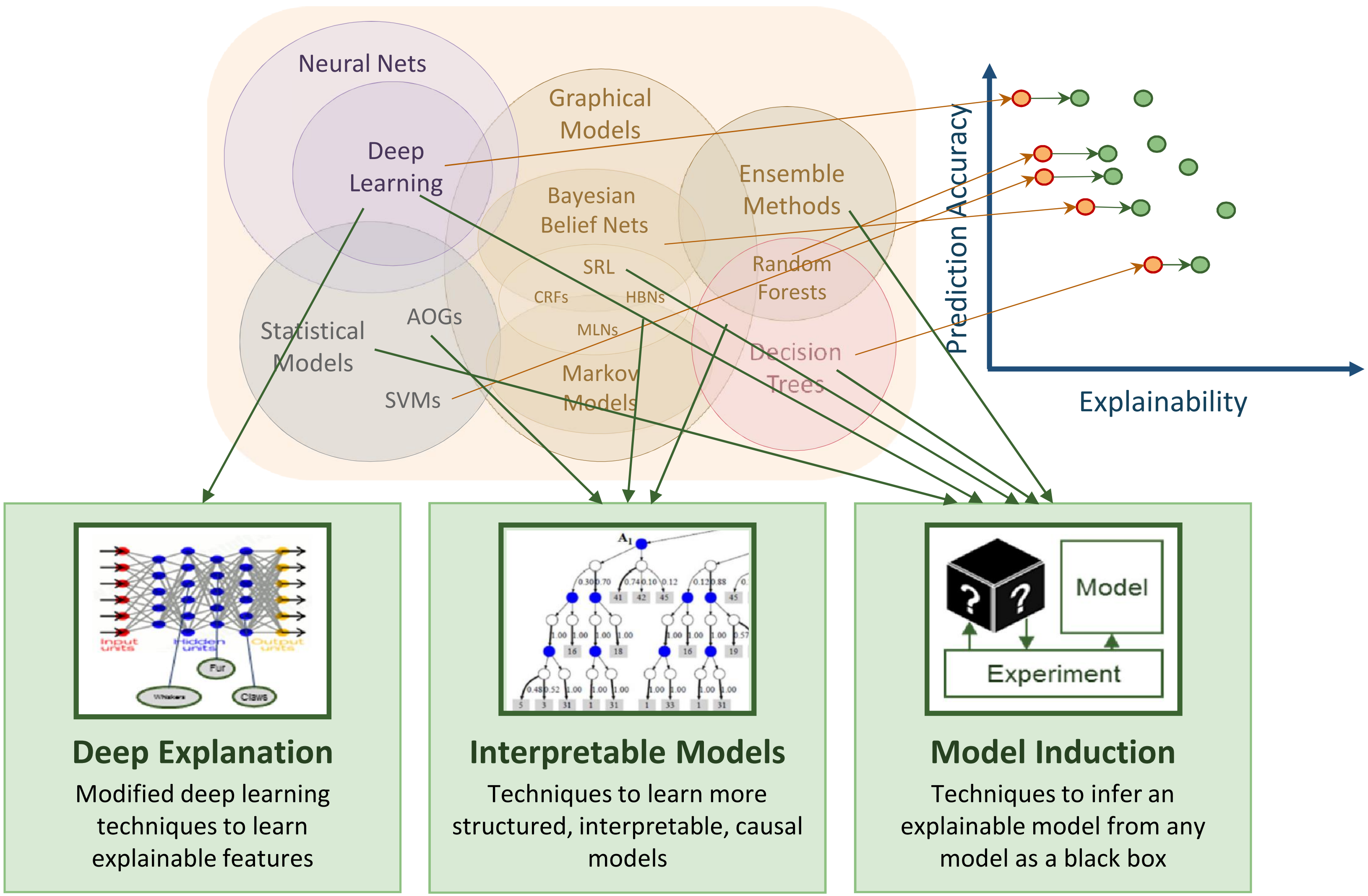
Deep Explanation

Modified deep learning techniques to learn explainable features

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

$$f(\boldsymbol{x}) = \mathrm{DeepNet}(\boldsymbol{x})$$

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

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} \overbrace{g_i(x_i)}^{\text{contribution of } i\text{th variable}}$$

Different dimensions of "interpretability"

prediction

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

model

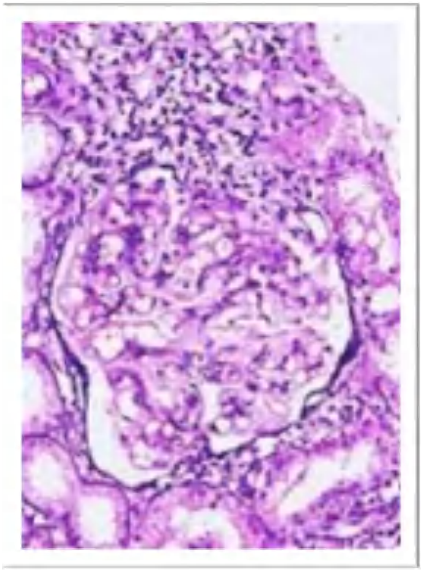
"What would a pattern belonging to a certain category typically look like according to the model."

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



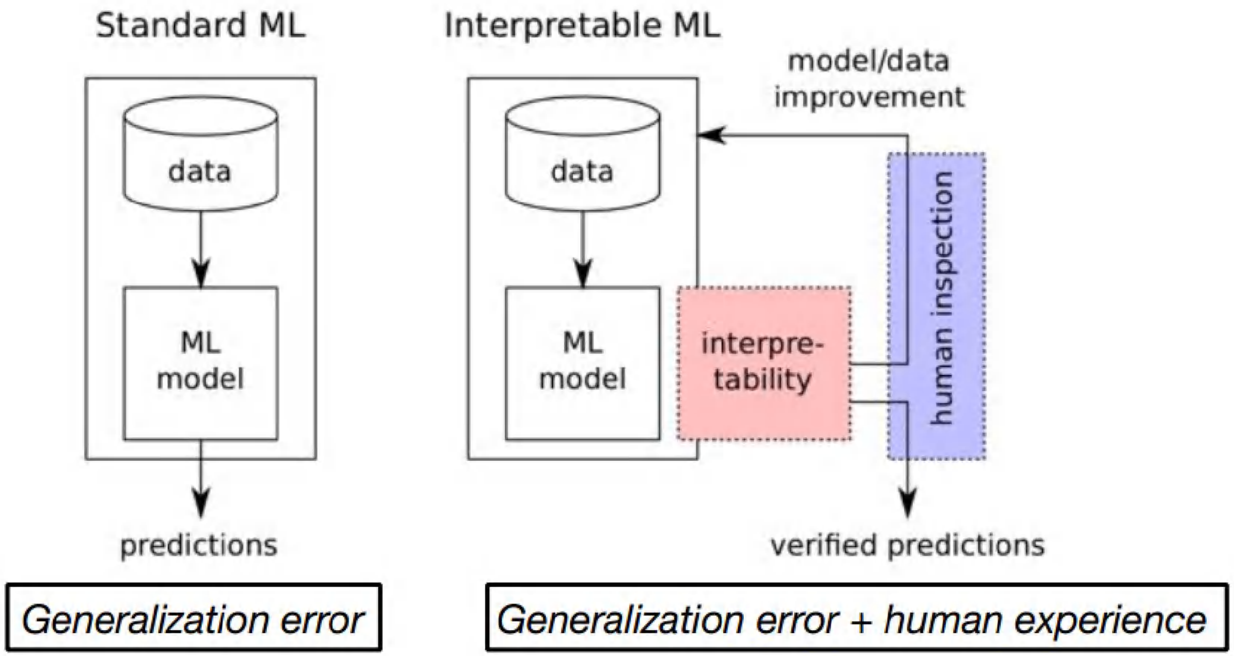


Image credit to: Samek, Montavon & Müller Tutorial at ICASSEP 2017

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

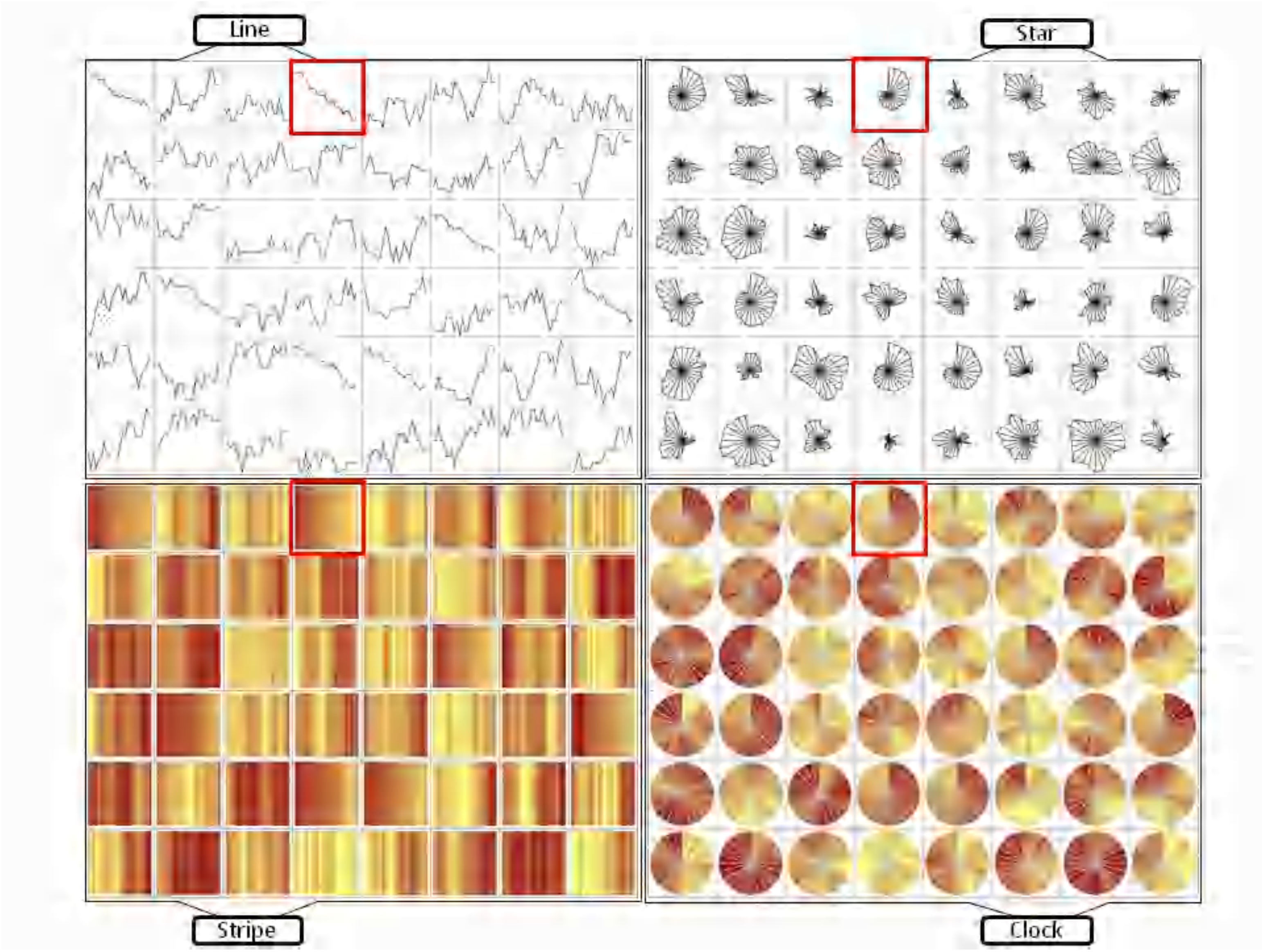
Research and Development in Interpretability …

- 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

Andreas Holzinger 2018. Explainable AI (ex-AI). Informatik-Spektrum, 41, (2), 138-143, doi:10.1007/s00287-018-1102-5.

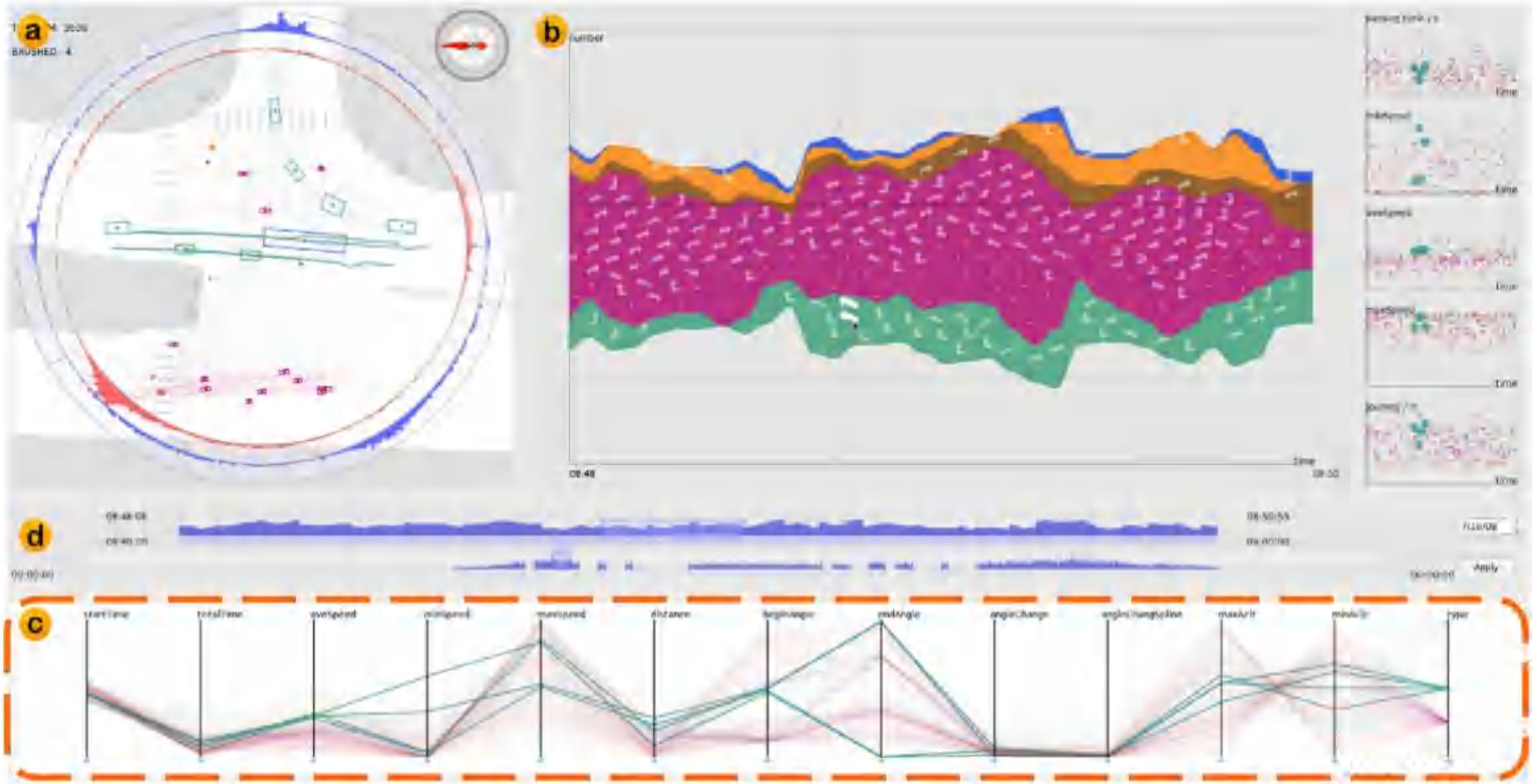

What is interpretable for humans?

What is understandable, interpretable, intelligible?



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

Explainable AI is a huge challenge for visualization

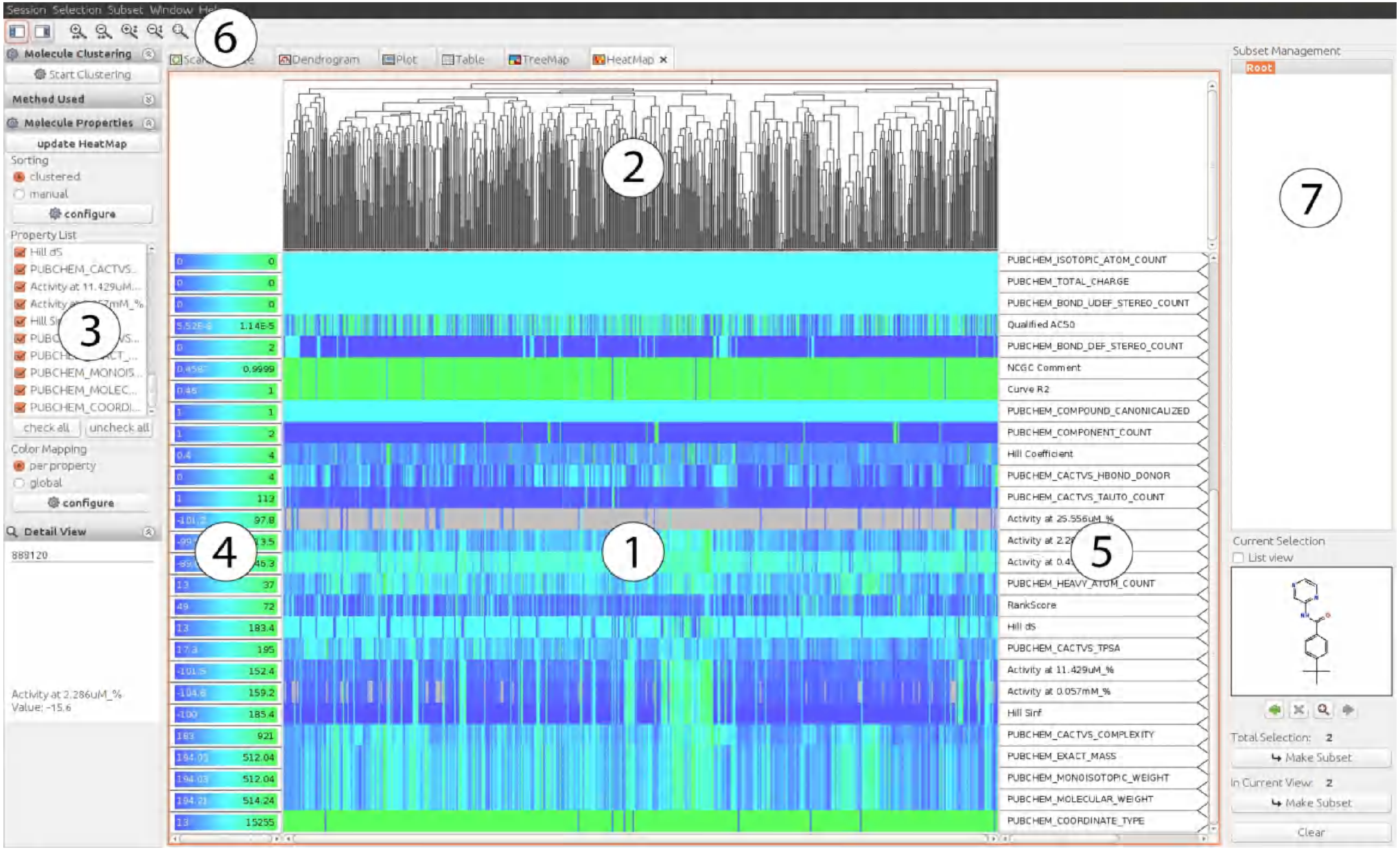


07 Methods of Explainable AI

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.

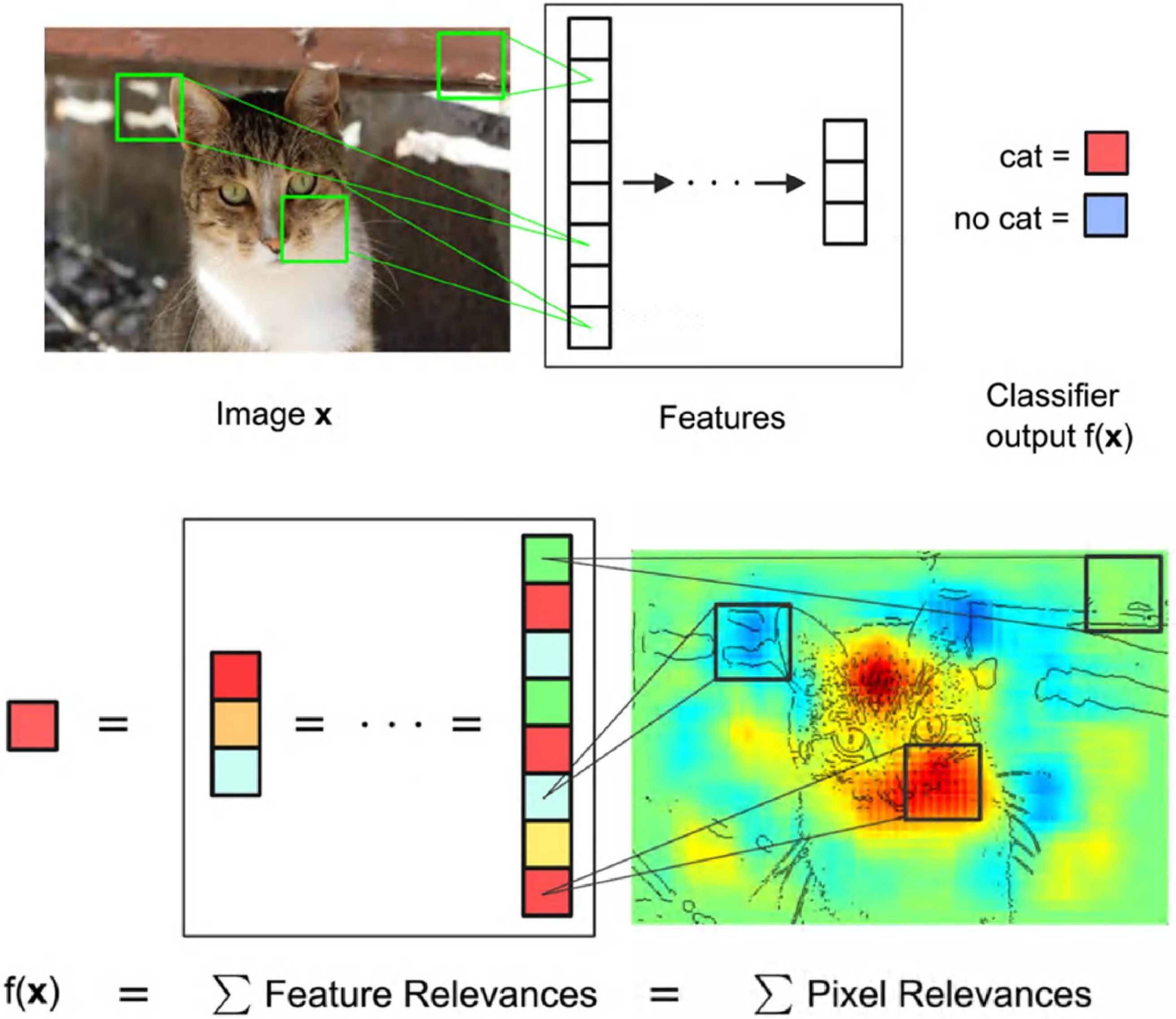
Example for an Explanation Interface - open work ☺



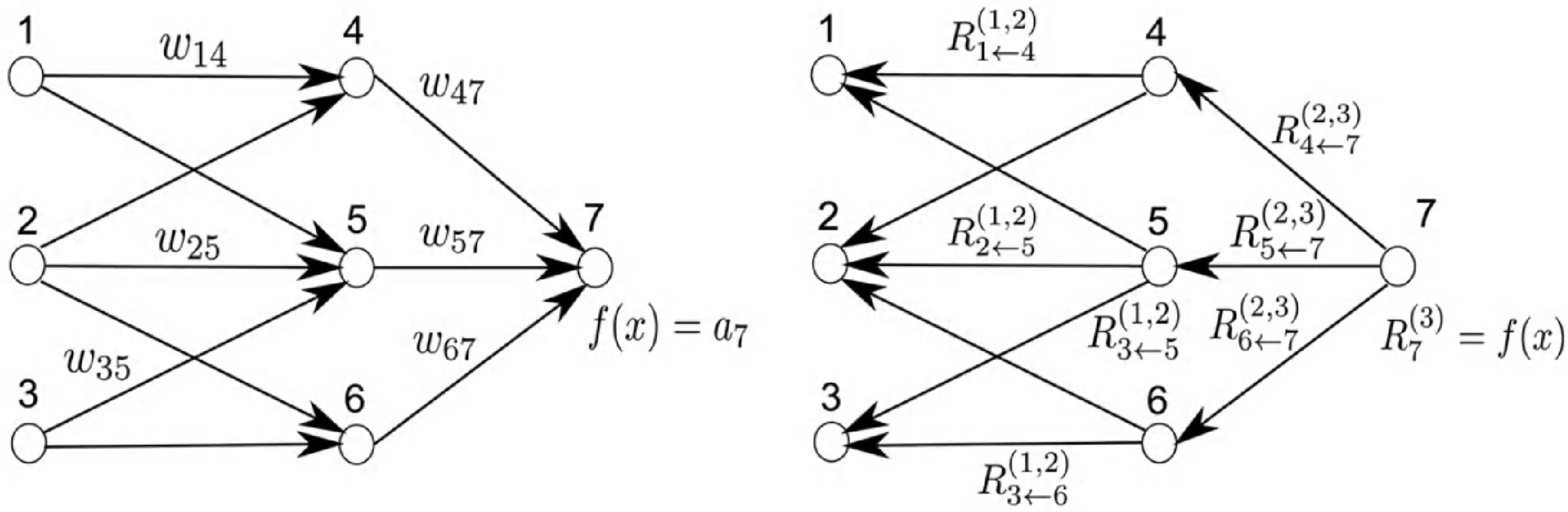
Werner Sturm, Till Schaefer, Tobias Schreck, Andeas Holzinger & Torsten Ullrich. Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps In: Borgo, Rita & Turkay, Cagatay, eds. EG UK Computer Graphics & Visual Computing CGVC 2015, 2015 University College London (UCL). Euro Graphics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



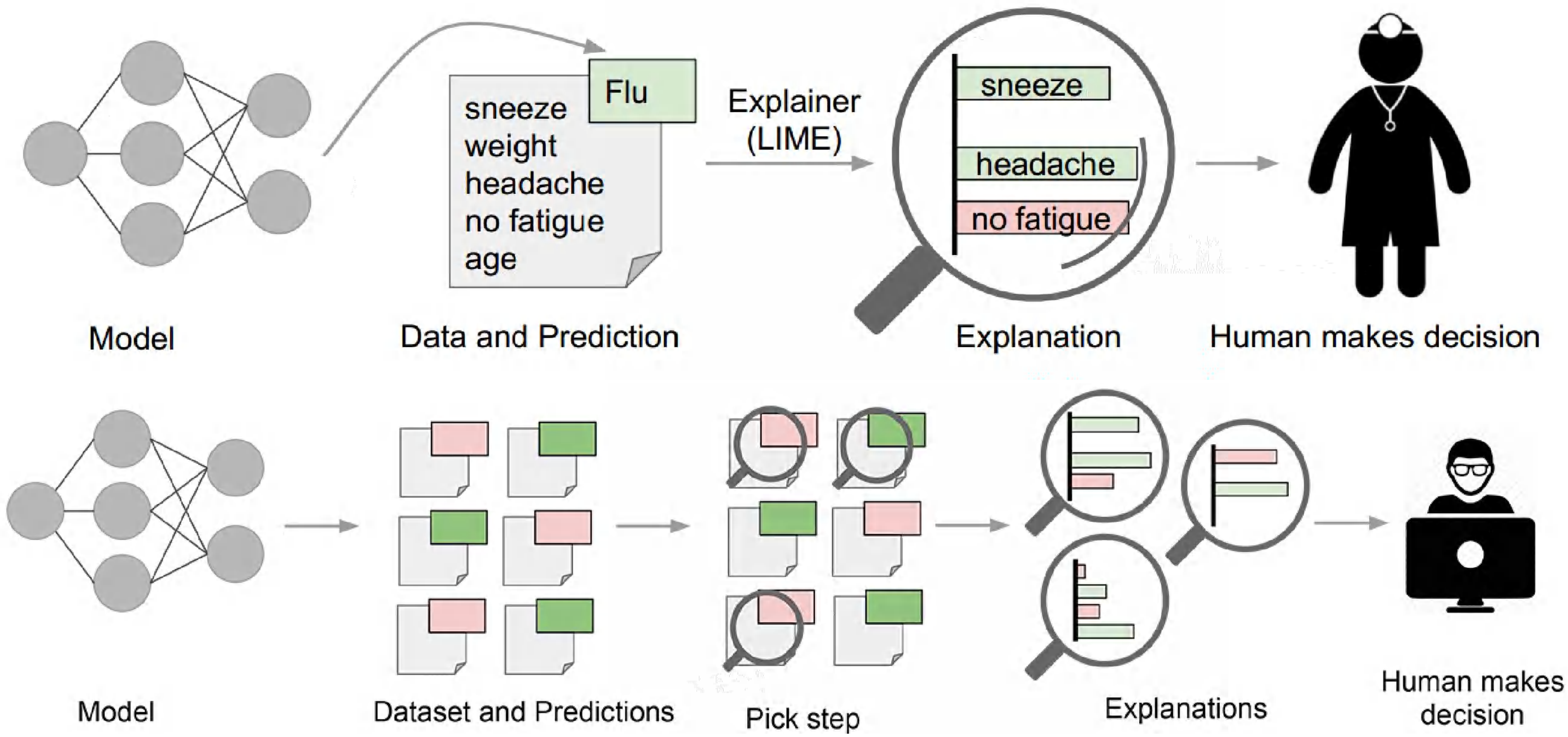
A NN-classifier during prediction time



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

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



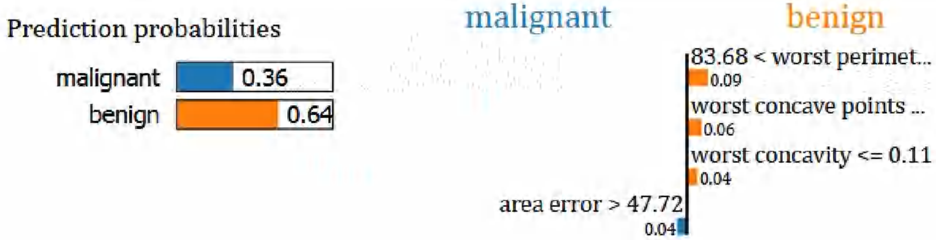
Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. ACM, 1135-1144, doi:10.1145/2939672.2939778.

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)
```



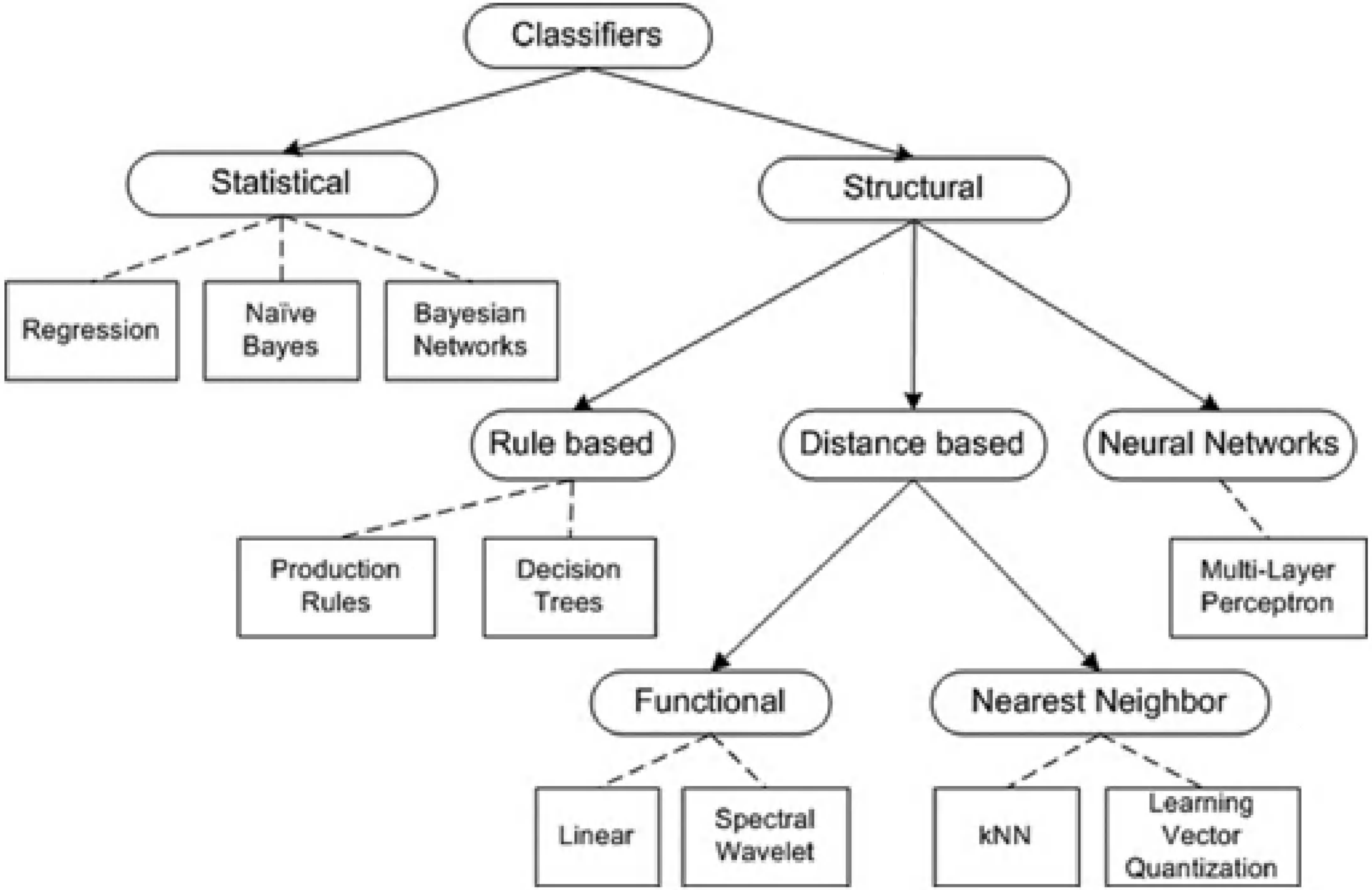
Feature	Value
worst perimeter	95.54
worst concave points	0.03
worst concavity	0.02
area error	106.00

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

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 ≥ 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:

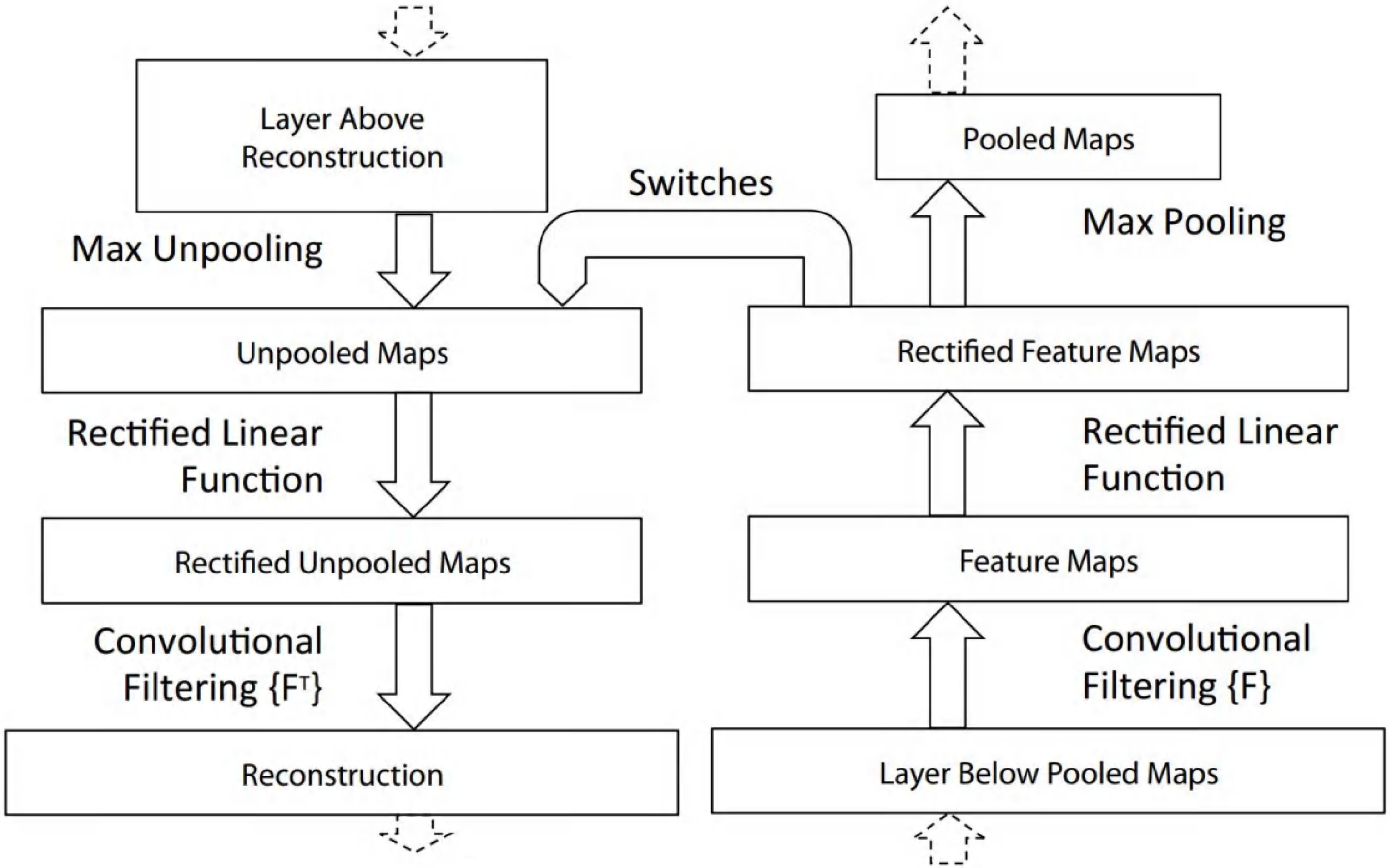
If Past-Depression =Yes **and** Tiredness =No **and** Exercise =No **and** Insomnia =Yes, **then** Depression

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

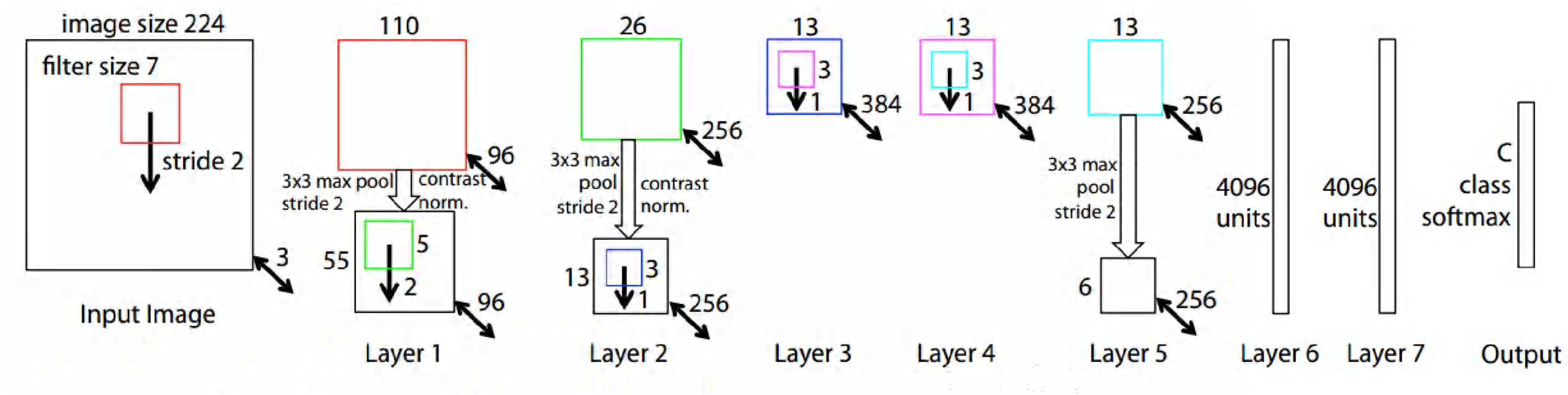
Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Jure Leskovec 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.

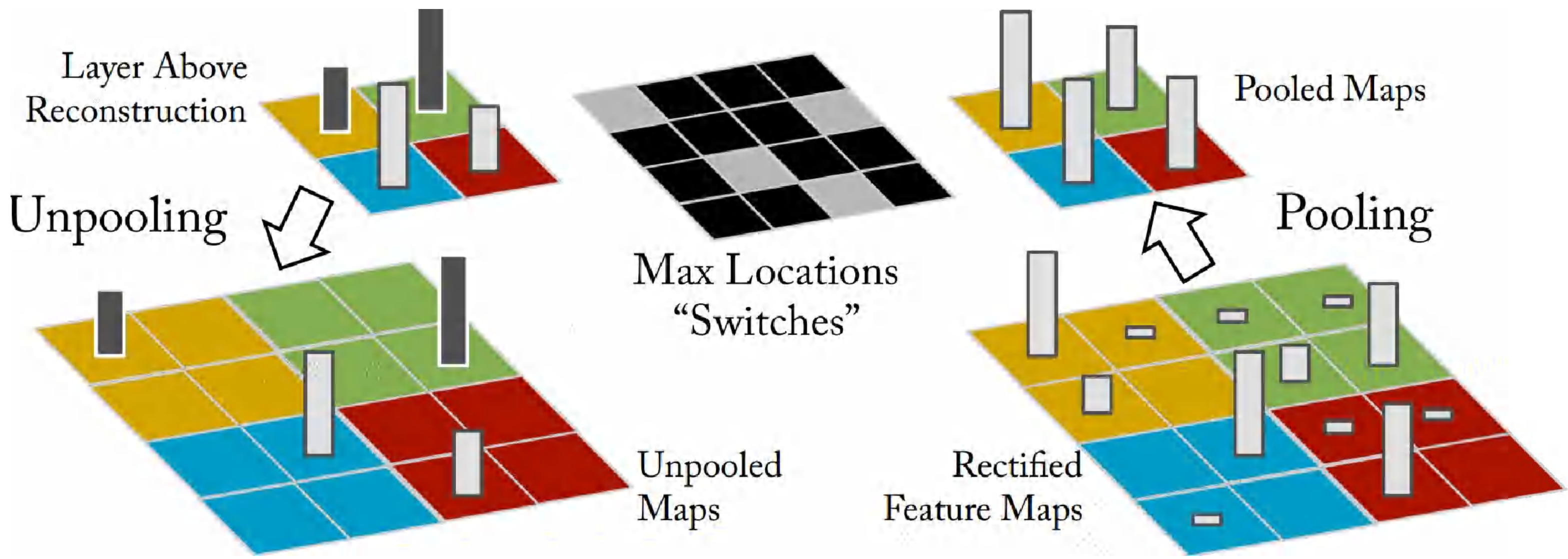
Example: Interpretable Deep Learning Model



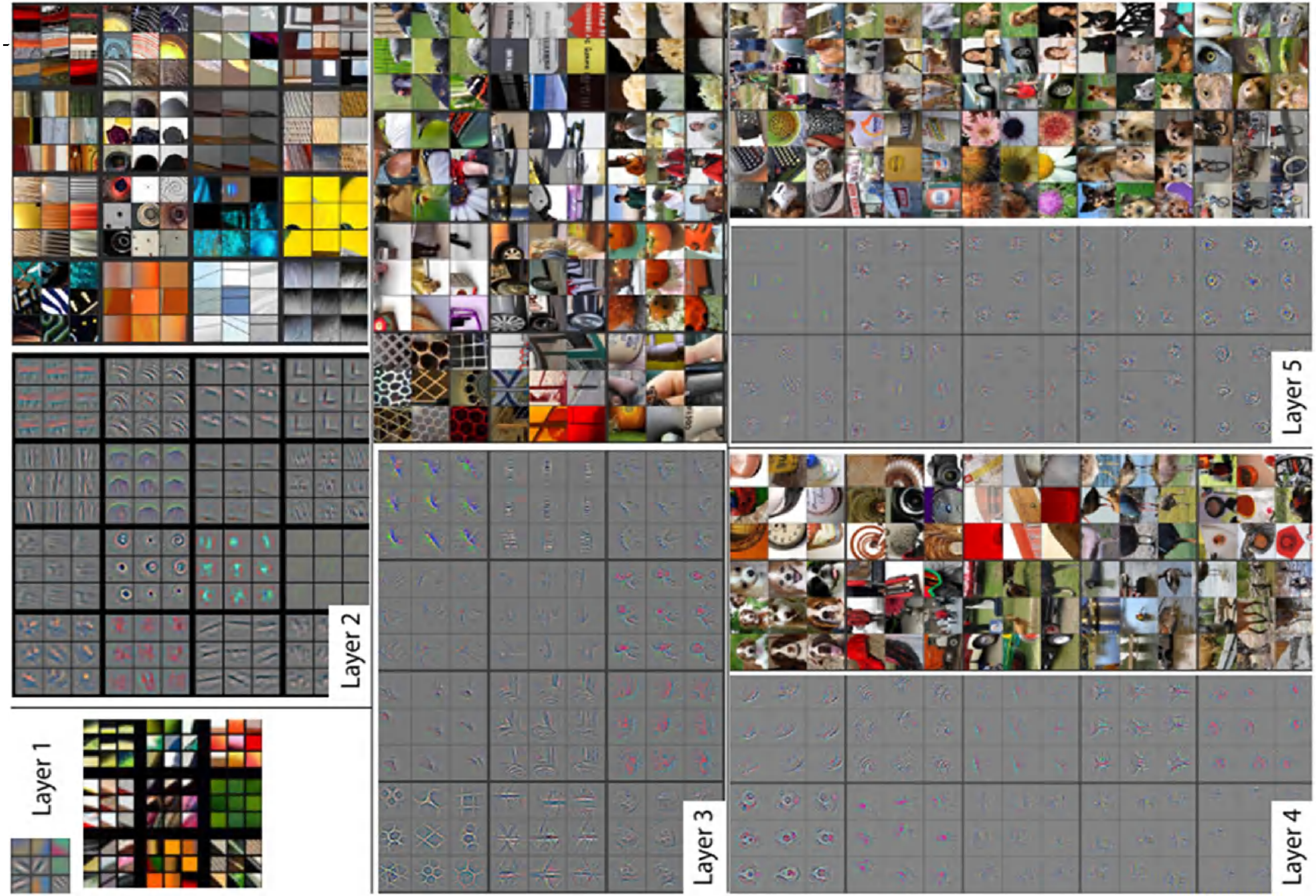
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

Visualizing a Conv Net with a De-Conv Net

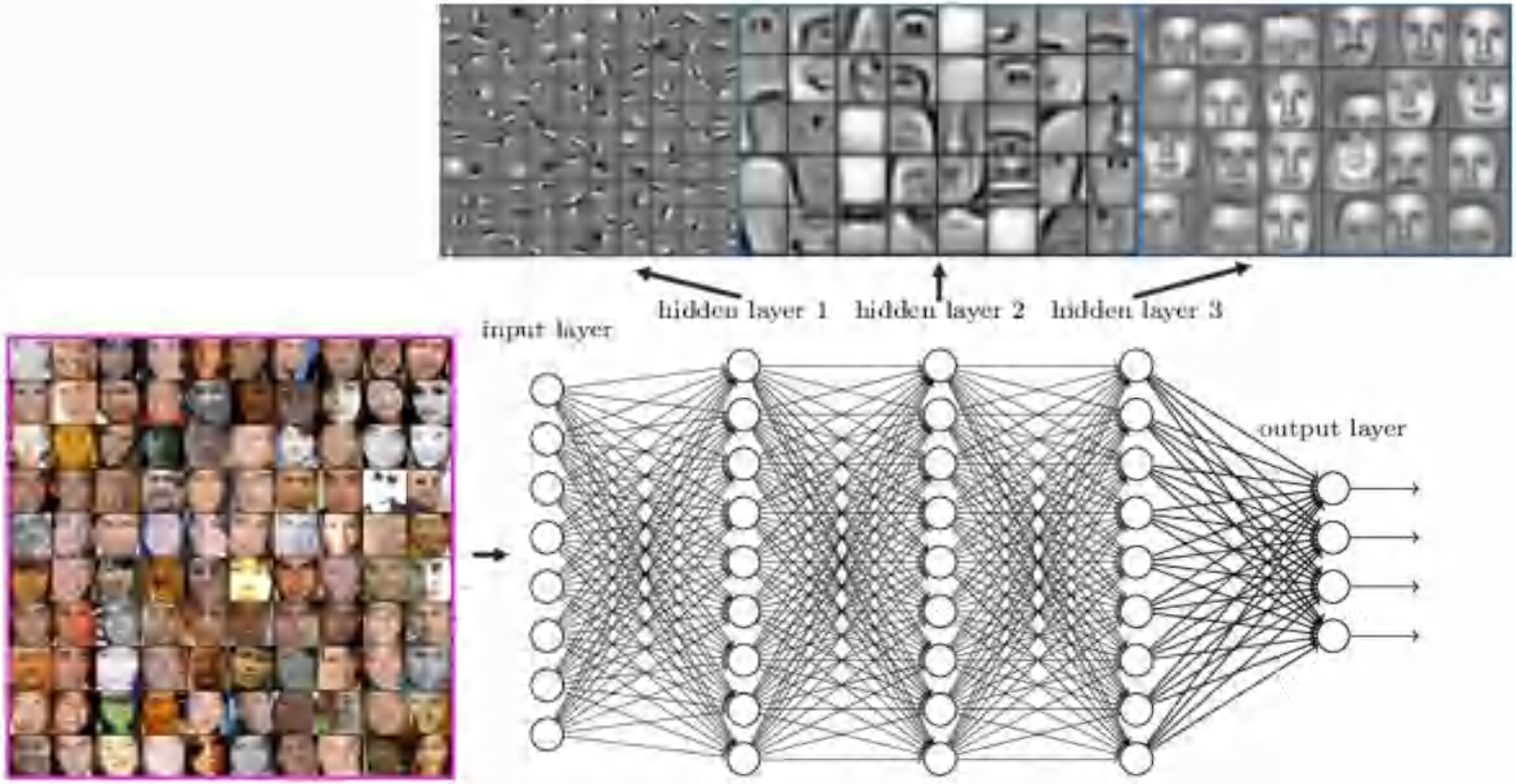




Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.

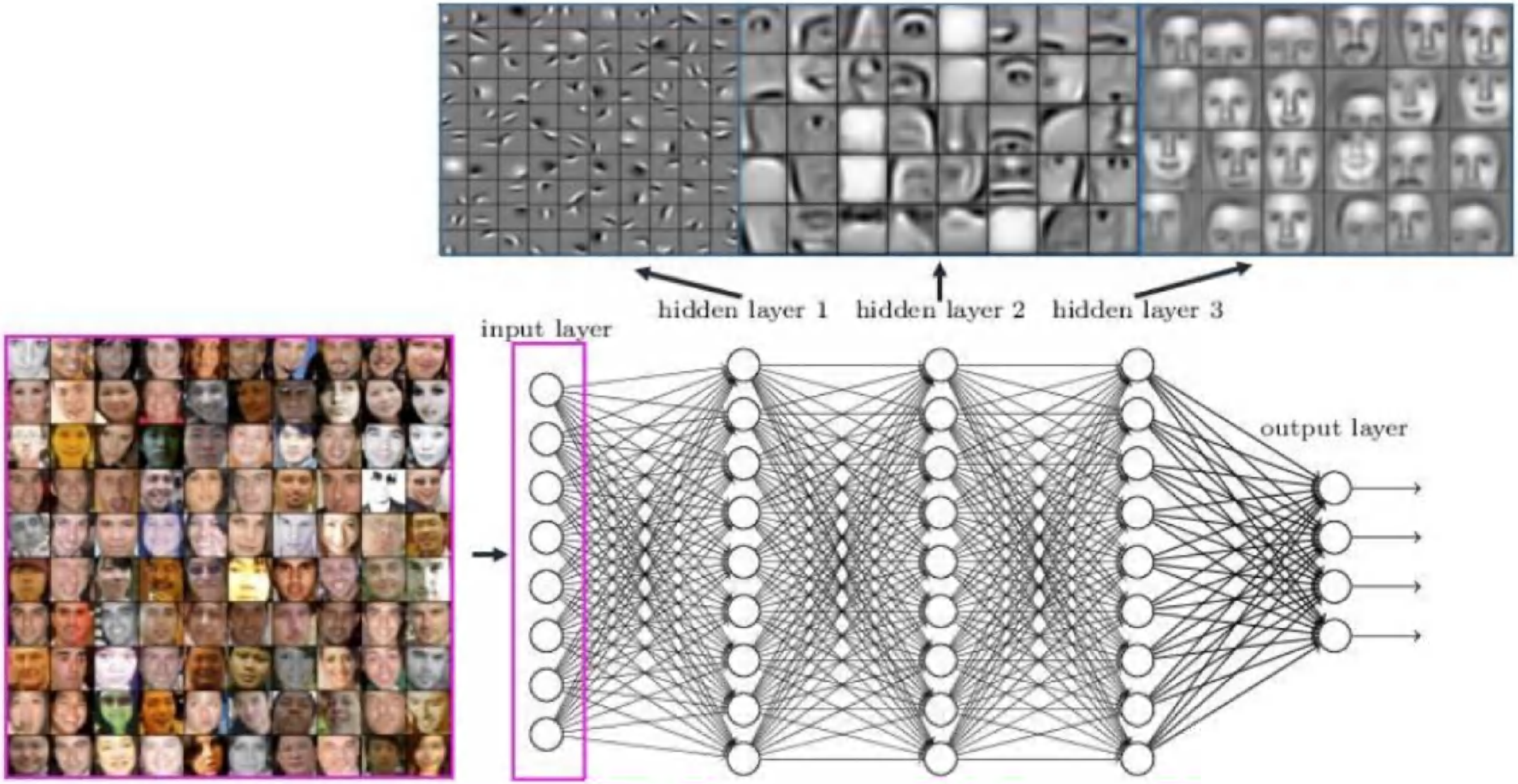


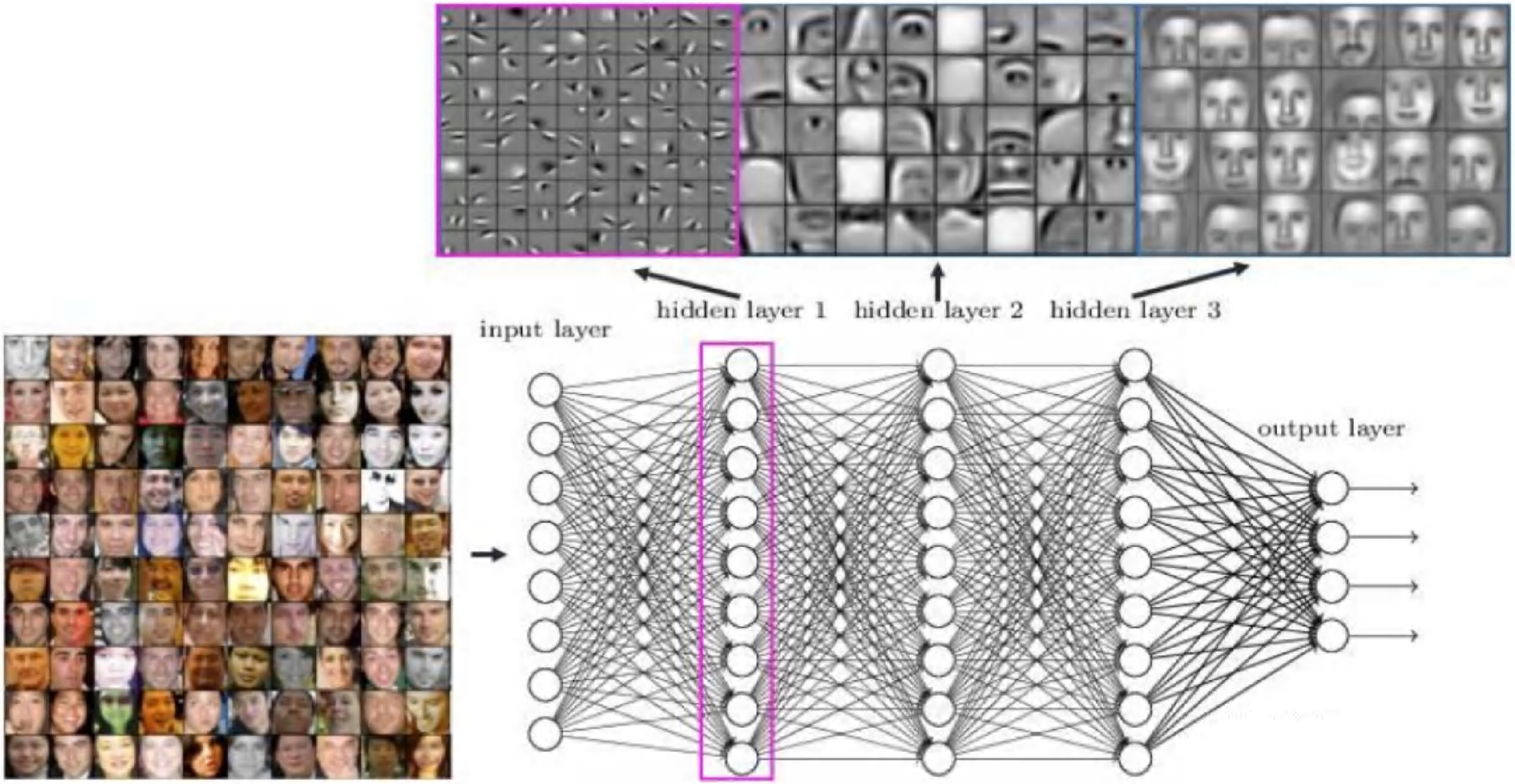
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

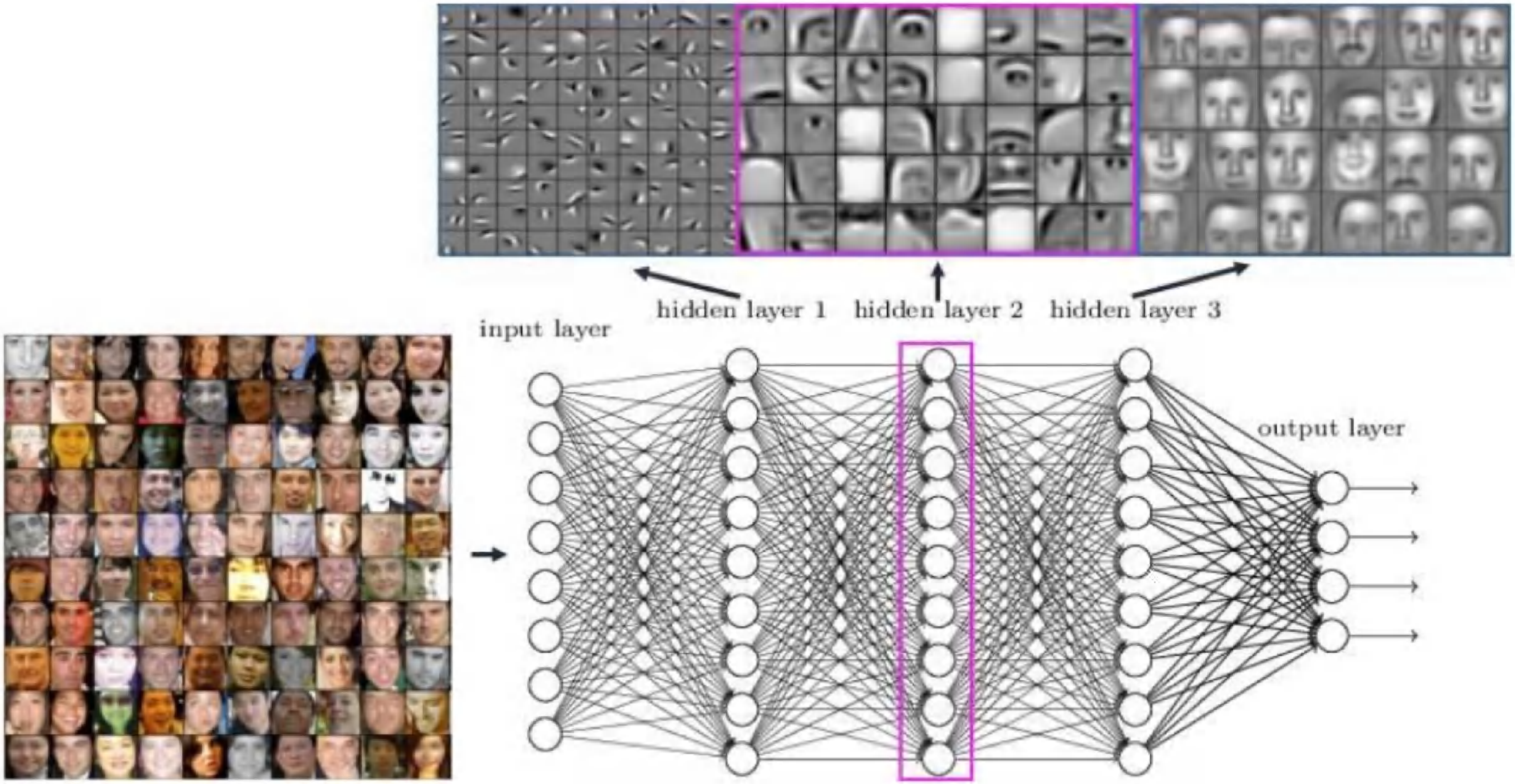


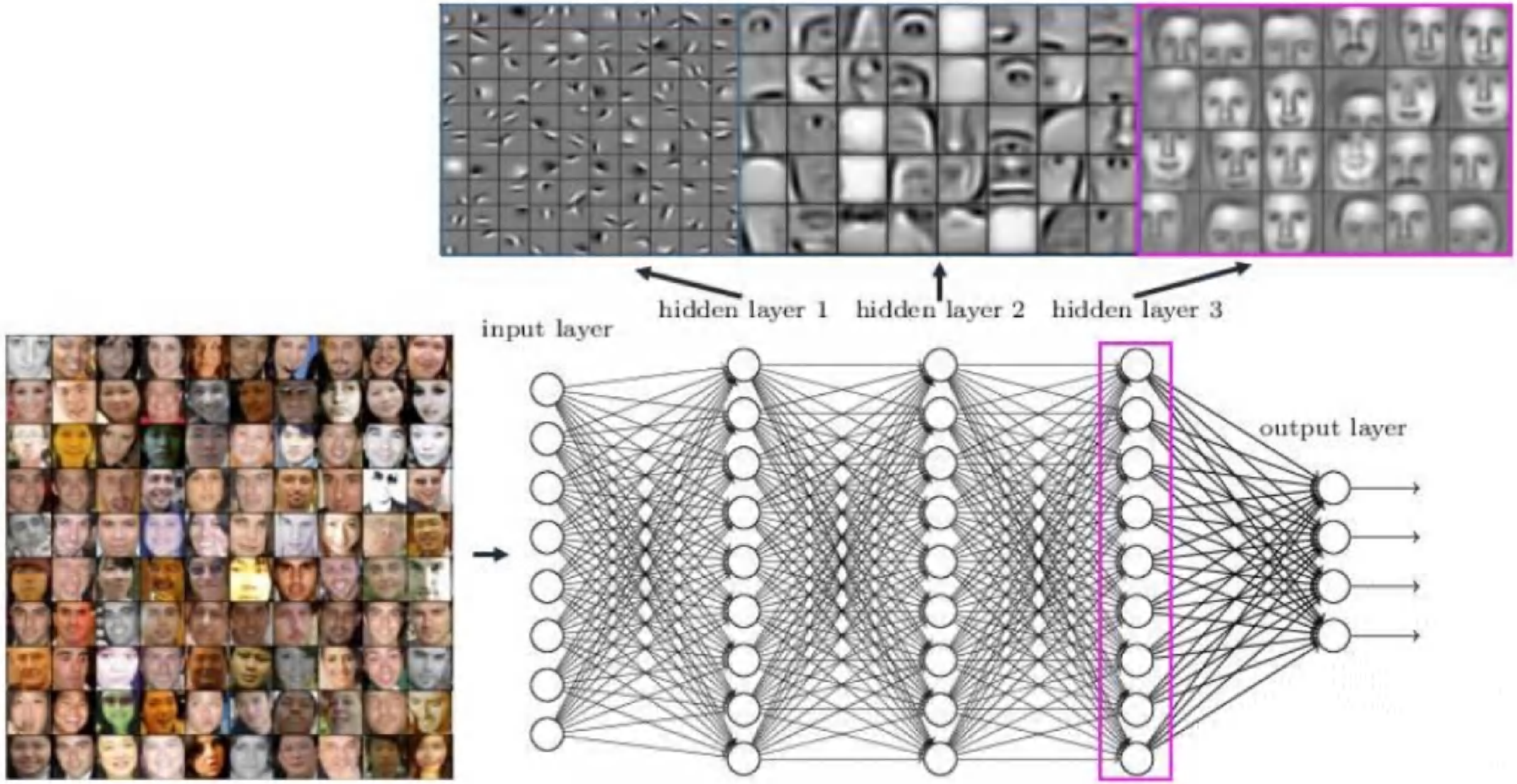
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

The world is compositional (Yann LeCun)

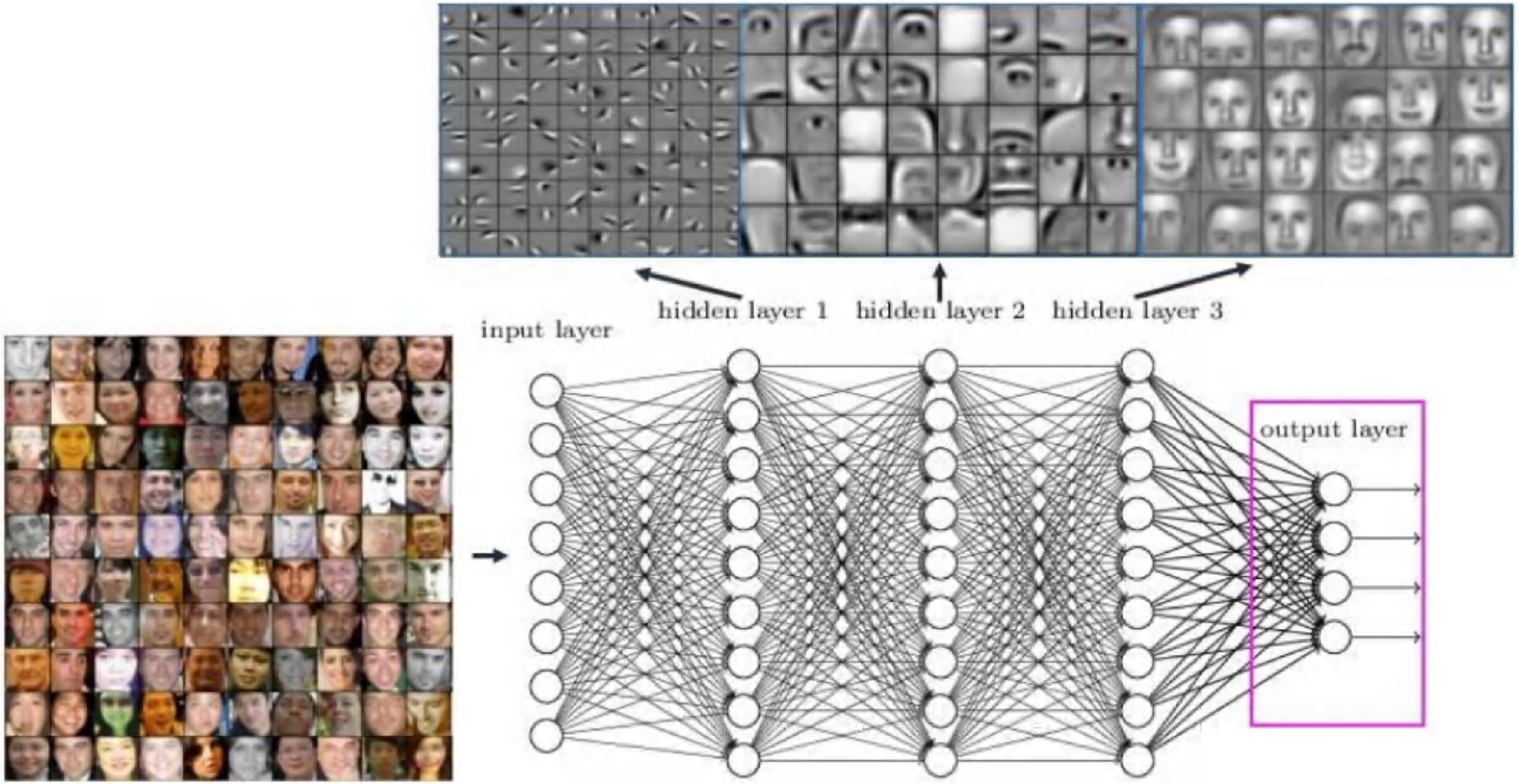


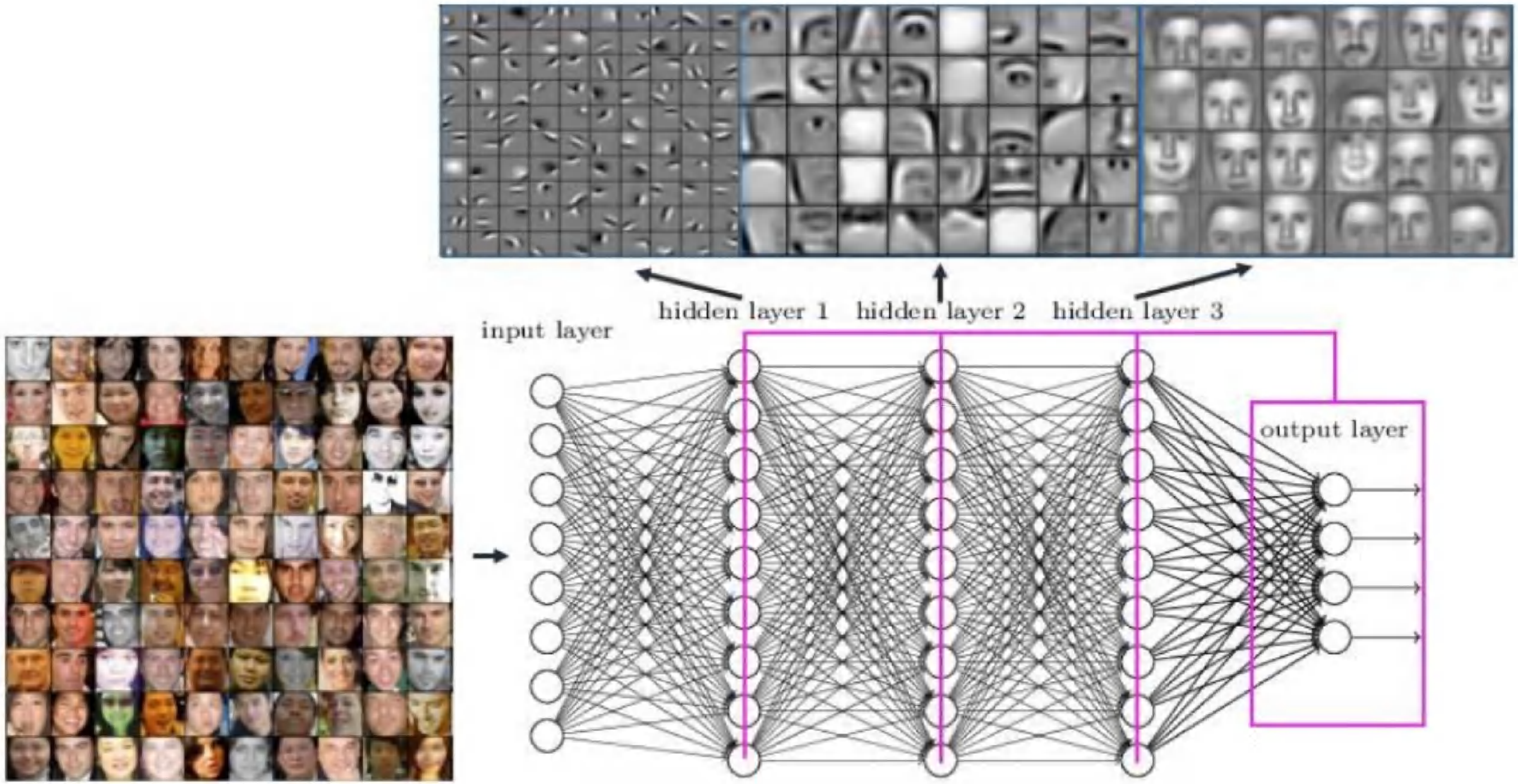

input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer


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

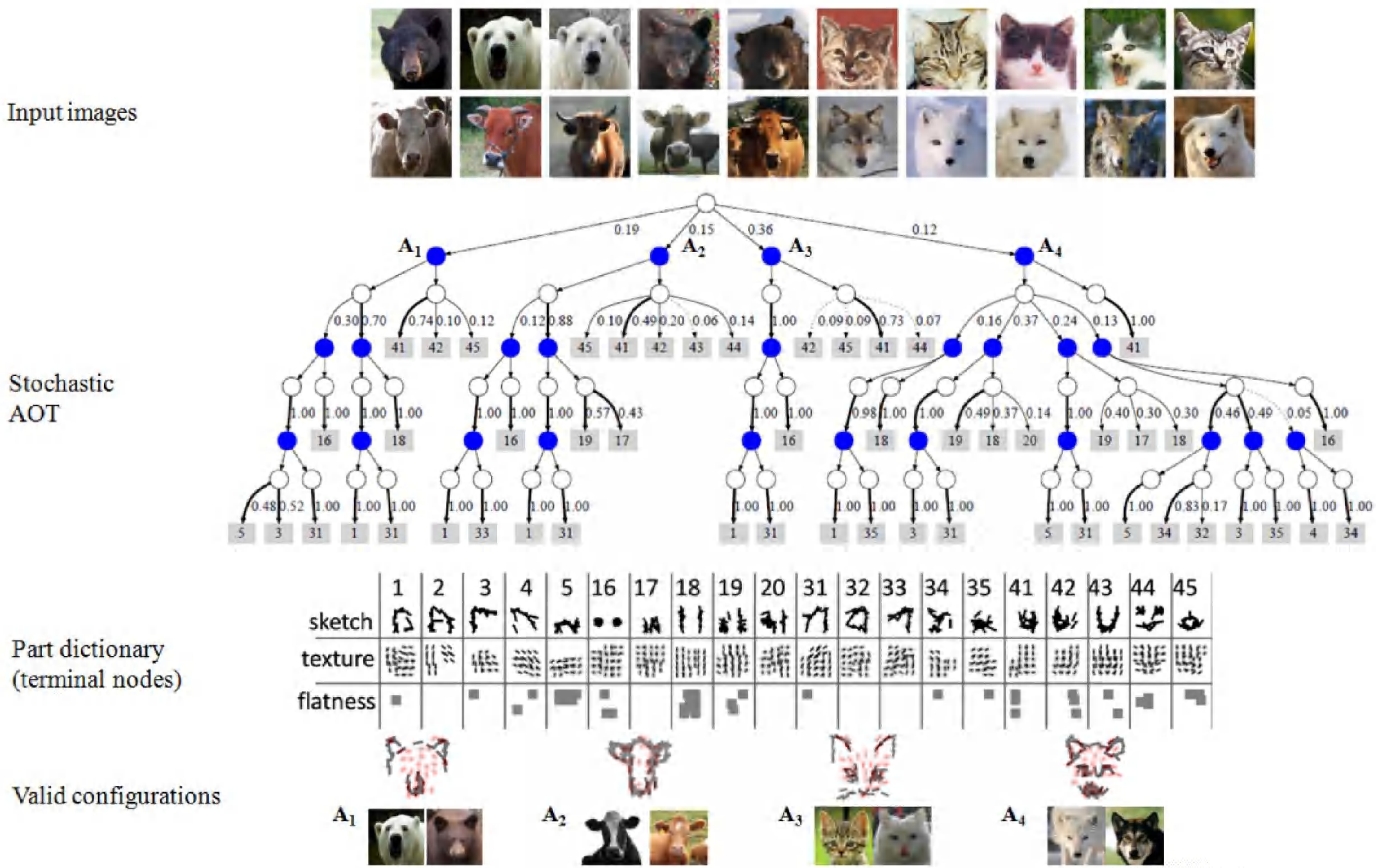


input layer
hidden layer 1
hidden layer 2
hidden layer 3
output layer

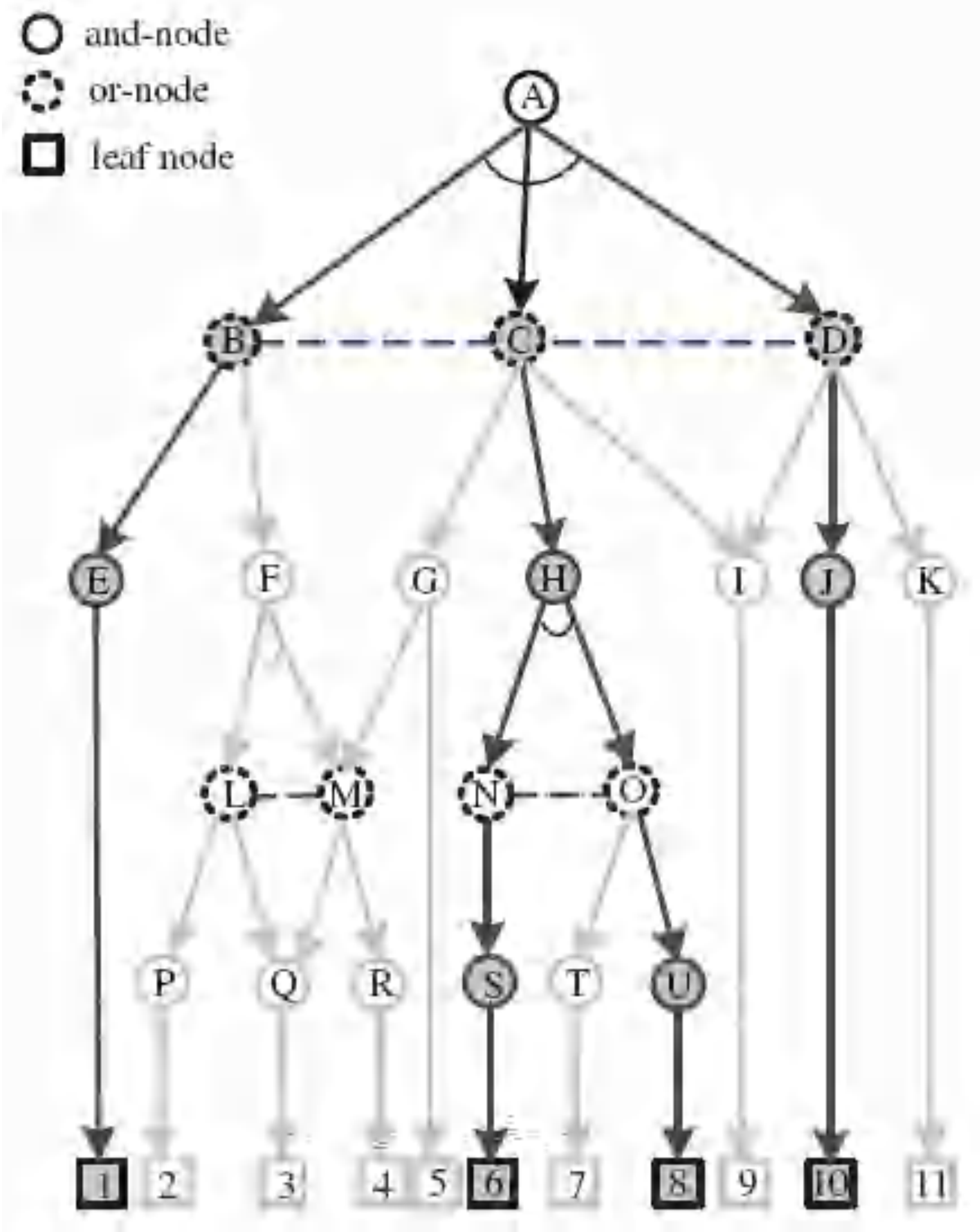

hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer


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

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.



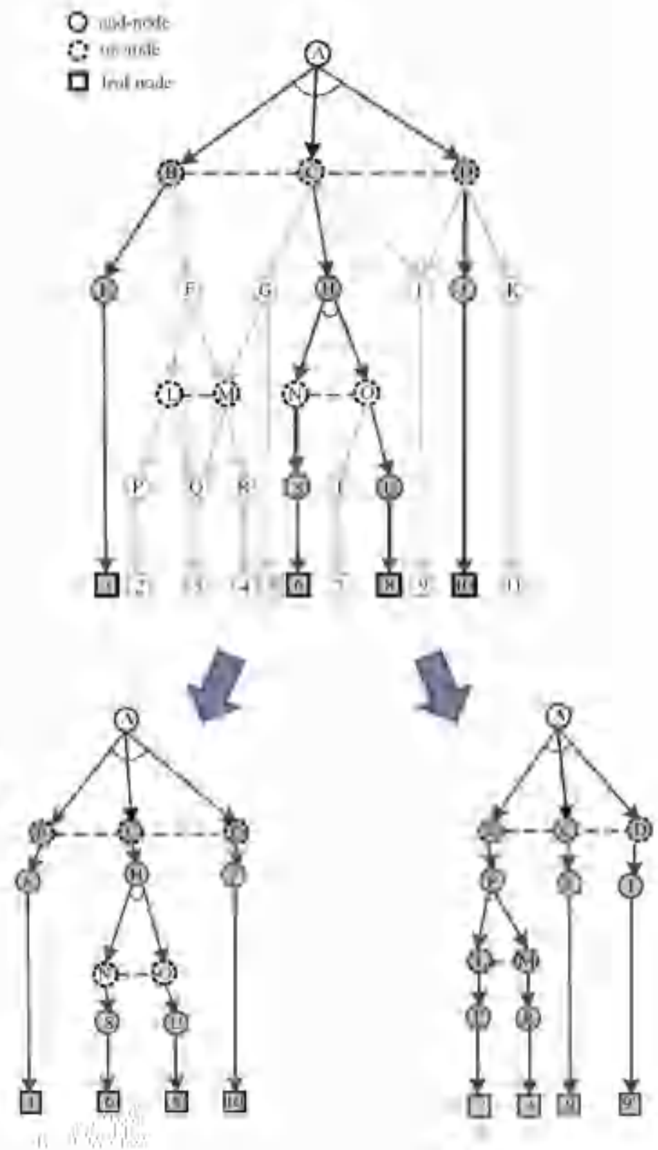
- Algorithm for this framework
  - Top-down/bottom-up computation
- Generalization of small sample
  - Use Monte Carlos simulation to synthesis more configurations
- Fill **semantic gap**

Images credit to Zhaoyin Jia (2009)

- 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: $\alpha(t)$

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

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



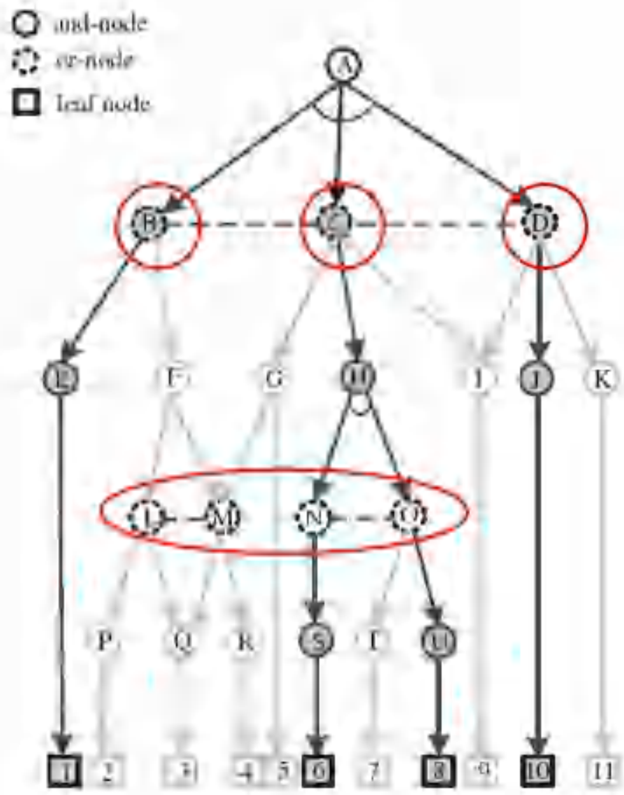
- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

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

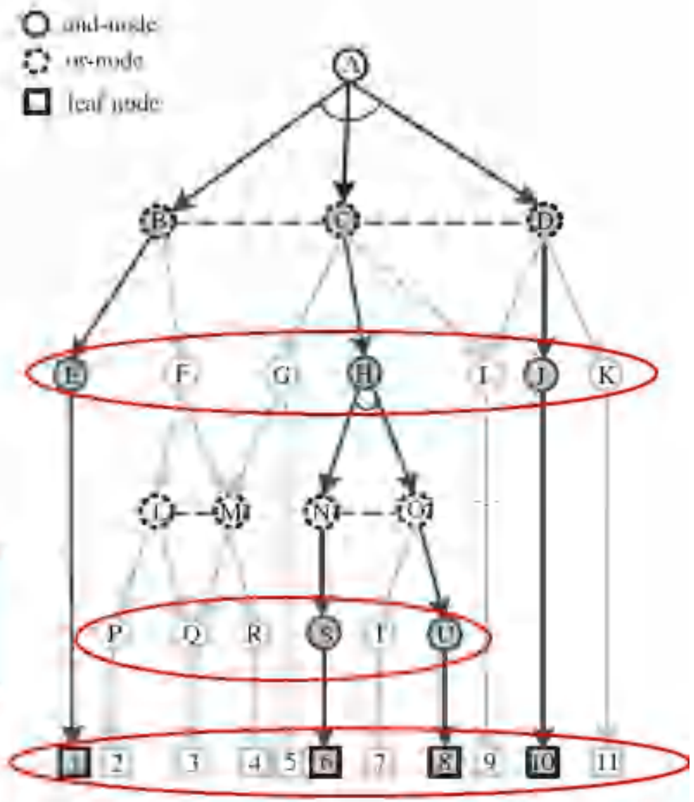
$$\xi(pg)=\boxed{\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))}+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

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

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



- 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: $\alpha(t)$



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

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

Weigh the local compatibility of primitives (geometric and appearance)

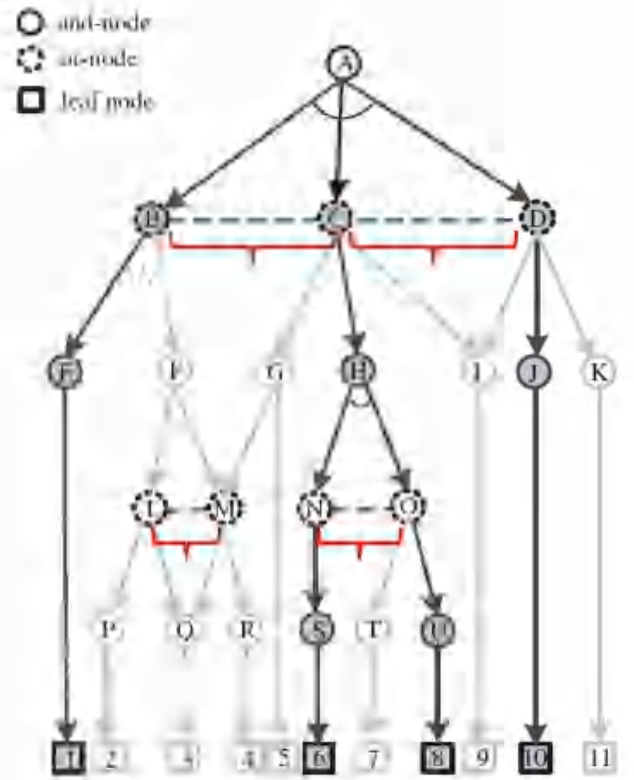
- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

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

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

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

Spatial and appearance between primitives (parts or objects)

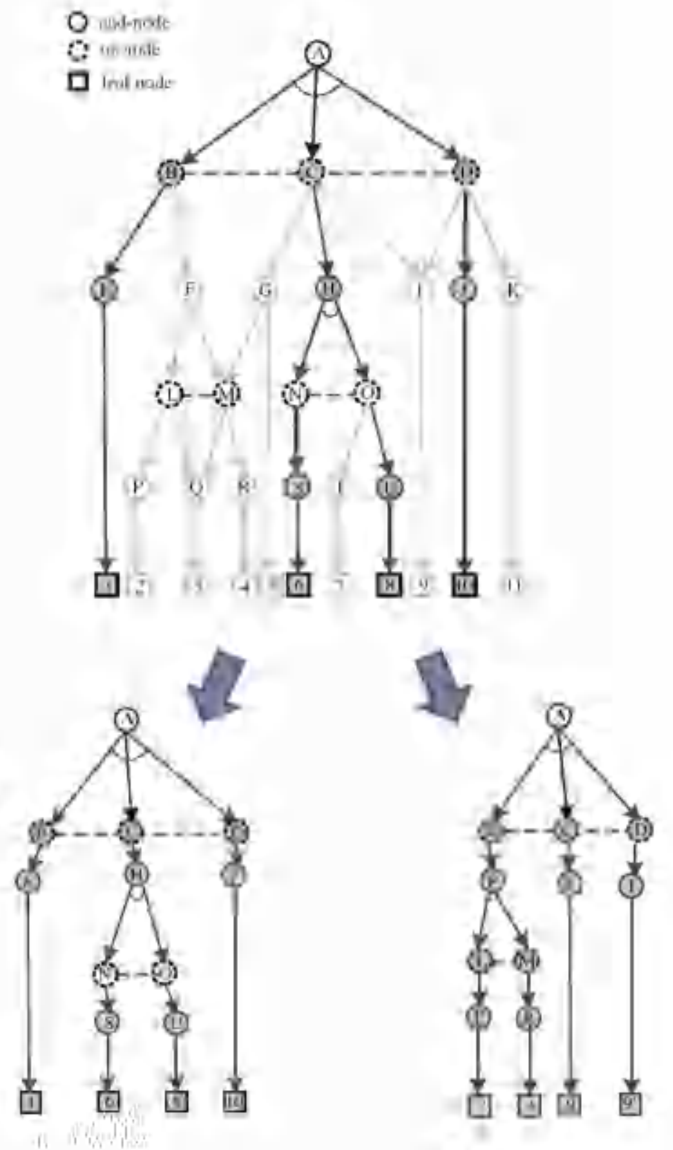


- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

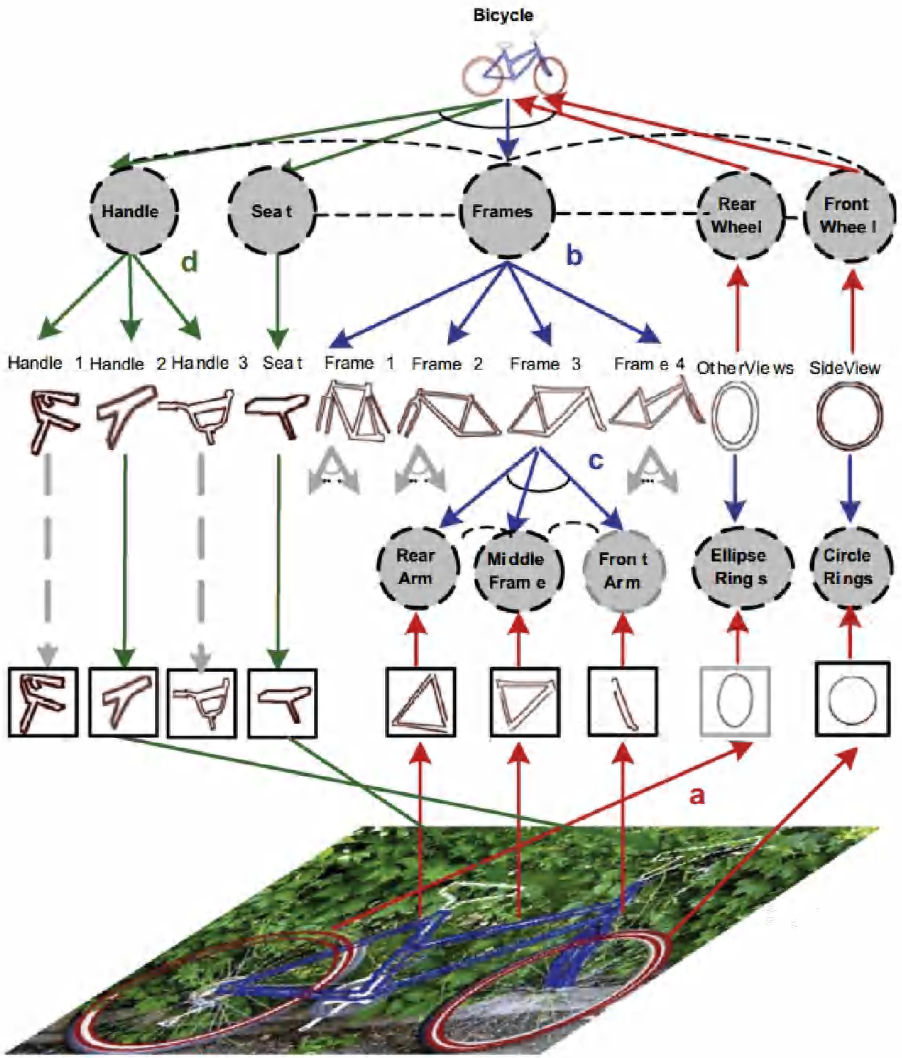
$$p(pg;\Theta,R,\Delta)=\frac{1}{Z(\Theta)}\exp(-\xi(pg))$$

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

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



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 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.

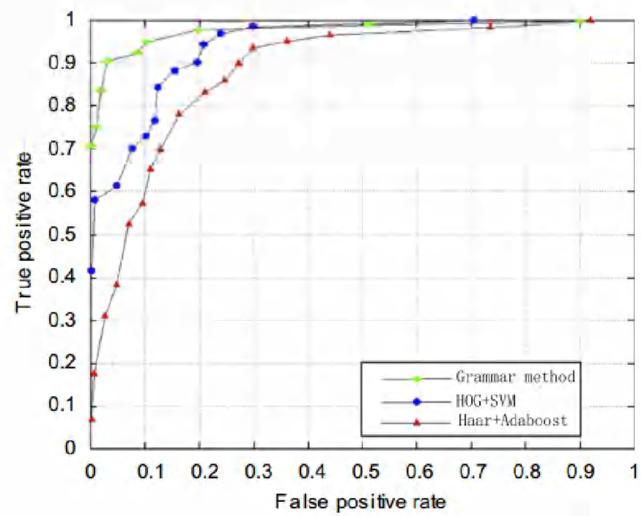
  2 Call Bottom-up(A) to update A's **open** list.

    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.

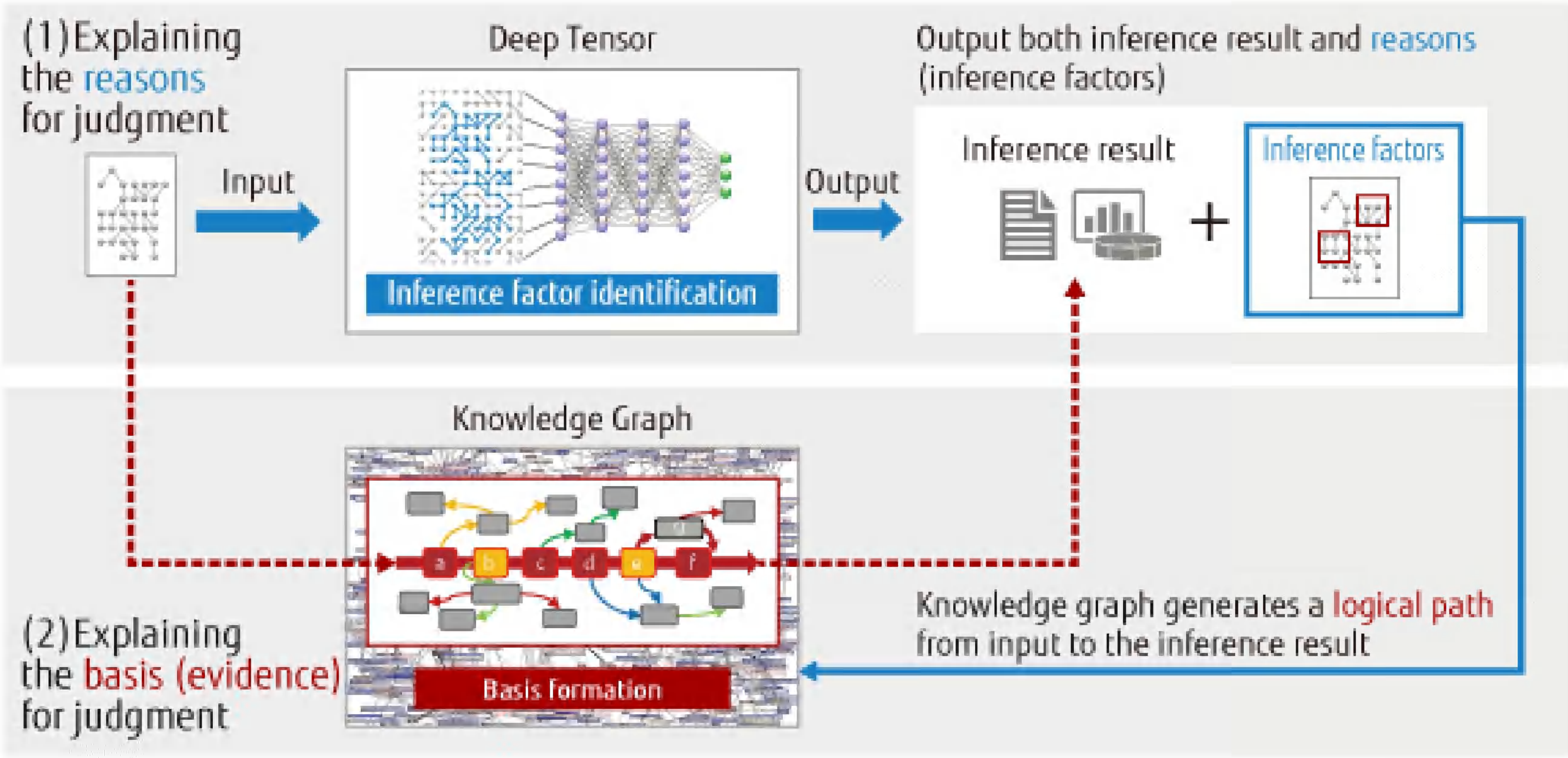
    i Accept hypotheses from A's **open** list to its **closed** list.
    ii Remove (or disassemble) hypotheses from A's **closed** list.
    iii Update the **open** lists for particles that overlap with node A.

- 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



Explainable AI with Deep Tensor and Knowledge Graph

http://www.fujitsu.com/jp/Images/artificial-intelligence-en_tcm102-3781779.png

- 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 it …

Explanations in Artificial Intelligence will be necessary

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

IBM is doing it now: teaching meaningful explanations

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.

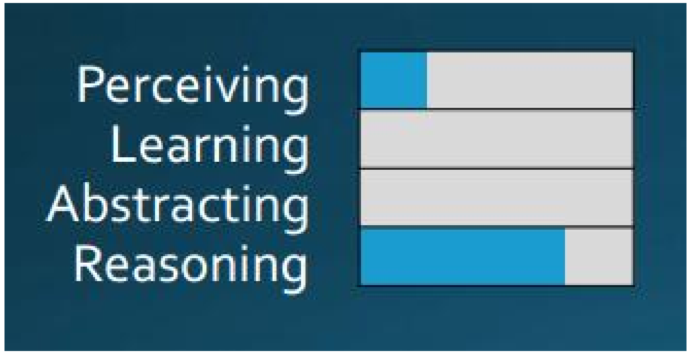
1 Introduction

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


Conclusion
This image is in the Public Domain

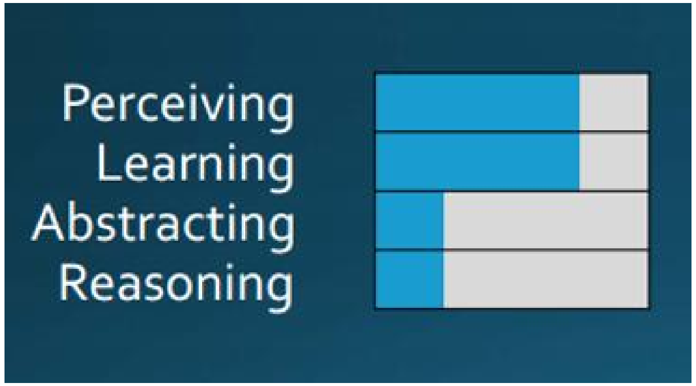
- Computational approaches can find in R^n what no human is able to see
- However, still there are many hard problems where a human expert in R^2 can understand the **context** and bring in experience, expertise, knowledge, intuition, …
- Black box approaches can not explain **WHY** a decision has been made …

The fist wave of AI (1943-1975): Handcrafted Knowledge



- Engineers create a set of logical rules to represent knowledge (Rule based Expert Systems)
- Advantage: works well in narrowly defined problems of well-defined domains
- Disadvantage: No adaptive learning behaviour and poor handling of $p(x)$

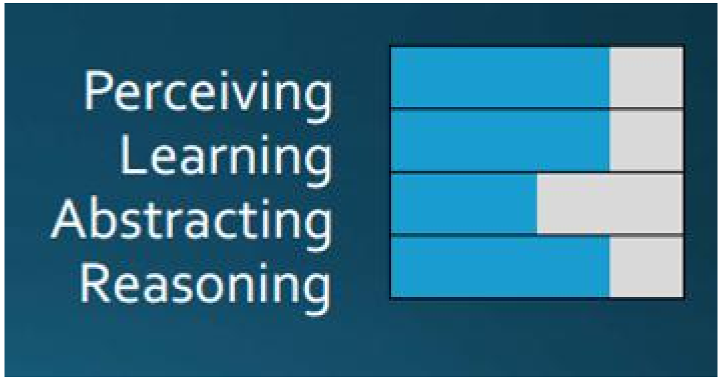
Image credit to John Launchbury



- Engineers create learning models for specific tasks and train them with "big data" (e.g. Deep Learning)
- Advantage: works well for standard classification tasks and has prediction capabilities
- Disadvantage: No contextual capabilities and minimal reasoning abilities

Image credit to John Launchbury

The third wave of AI (?): Adaptive Context Understanding



- A contextual model can perceive, learn and understand and abstract and reason
- Advantage: can use transfer learning for adaptation on unknown unknowns
- Disadvantage: Superintelligence …

Image credit to John Launchbury

- Myth 1a: Superintelligence by 2100 is inevitable!
- Myth 1b: Superintelligence by 2100 is impossible!
- **Fact: We simply don’t know it!**

- Myth 2: Robots are our main concern

  Fact: Cyberthreats are the main concern: it needs no body – only an Internet connection

- Myth 3: AI can never control us humans

  Fact: Intelligence is an enabler for control: We control tigers by being smarter …

Our goal: Contextual adaptations can explain decisions

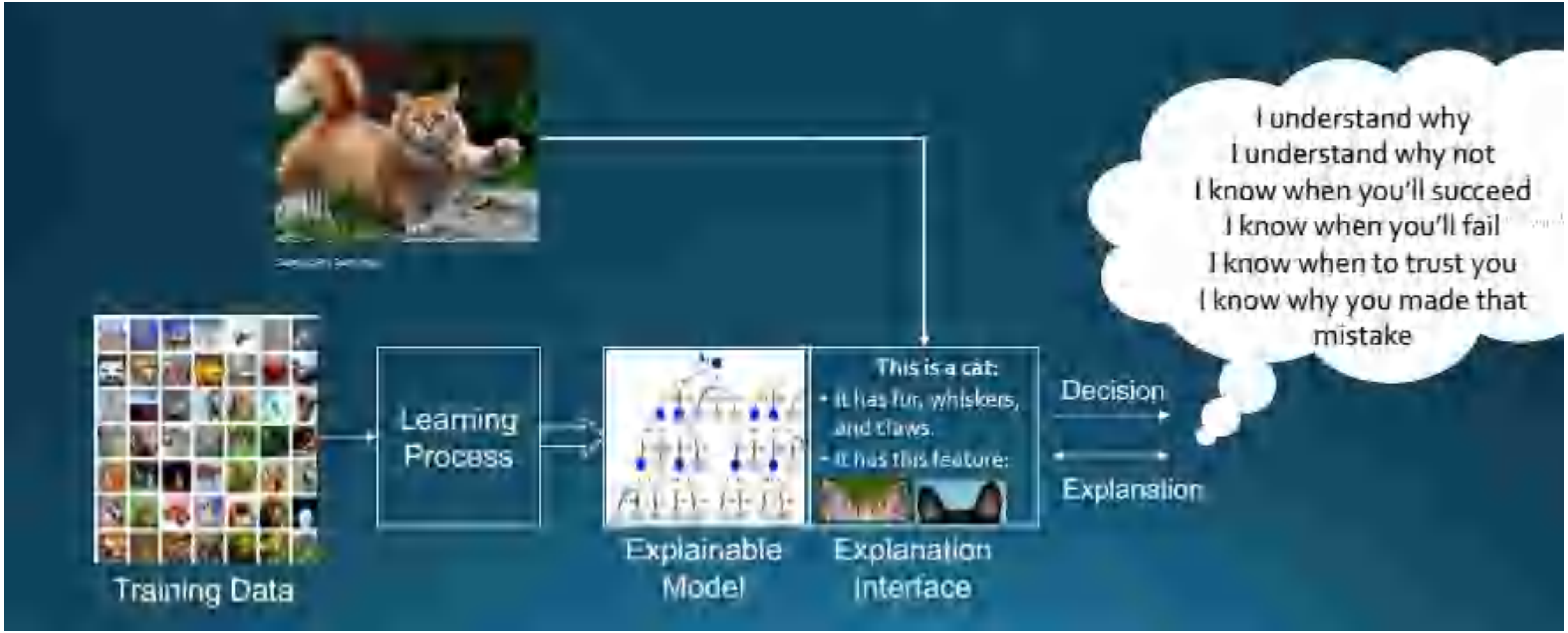
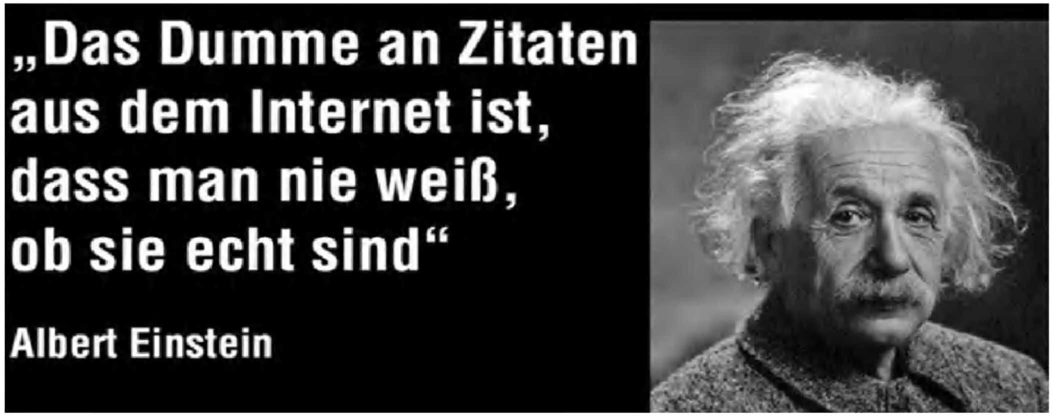


Image credit to John Launchbury

- Computers are incredibly fast, accurate and stupid,
- humans are incredibly slow, inaccurate and brilliant,
- together they are powerful beyond imagination

(Einstein never said that)

https://www.benshoemate.com/2008/11/30/einstein-never-said-that



Thank you!