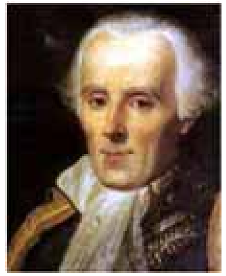
Andreas Holzinger

185.A83 Machine Learning for Health Informatics

2019S, VU, 2.0 h, 3.0 ECTS

Lecture 02 – Dienstag, 19.03.2019



From Clinical Decision Support to Causal Reasoning and explainable AI

andreas.holzinger AT tuwien.ac.at

https://human-centered.ai/machine-learning-for-health-informatics-class-2019

- 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
- Causality > Causability
- (Some) Methods of Explainable AI

Advance Organizer (1/2)

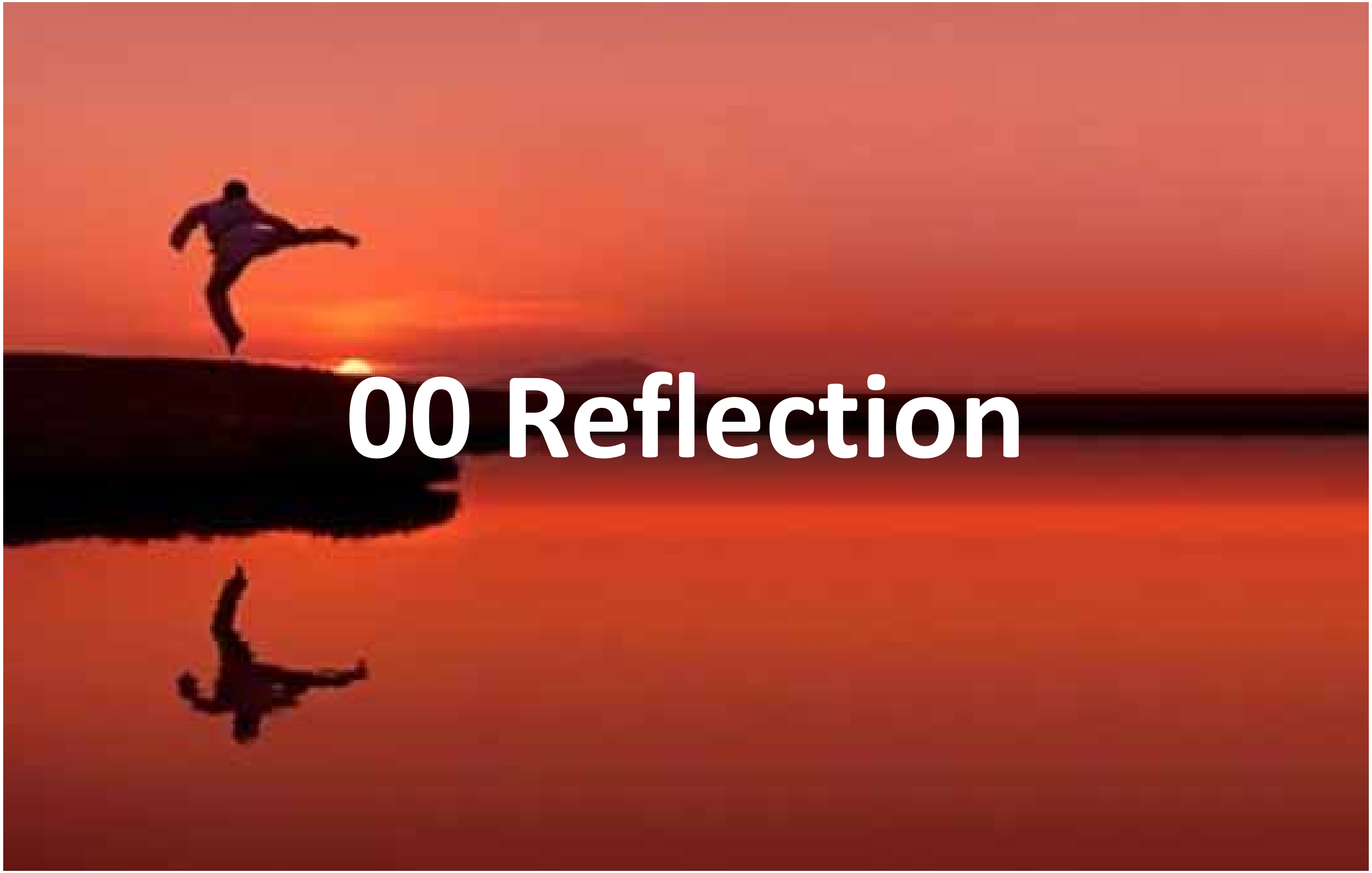
- **Causality** = fundamental relationship between cause and effect
- Causability = similar to the concept of usability the property of a human explanation
- **Case-based reasoning (CBR)** = process of solving new problems based on the solutions of similar past problems;
- **Certainty factor model (CF)** = 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;
- Counterfactual = relating to or expressing what has not happened or is not the case
- **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)

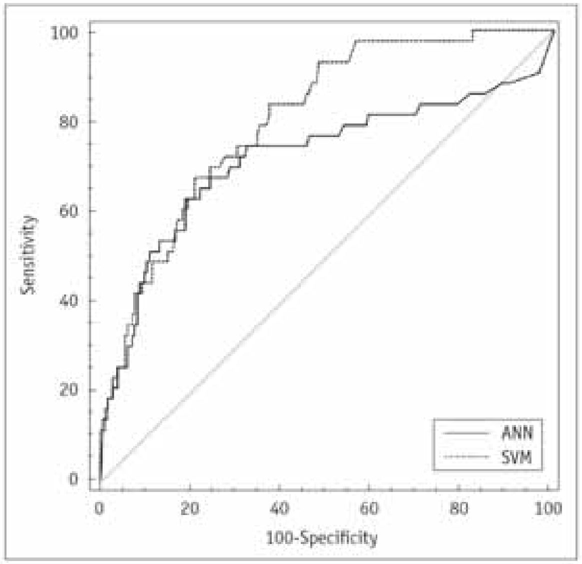
- **DXplain** = a DSS from the Harvard Medical School, to assist making a diagnosis (clinical consultation), and also as an instructional instrument (education); provides a description of diseases, etiology, pathology, prognosis and up to 10 references for each disease;
- Etiology = in medicine (many) 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

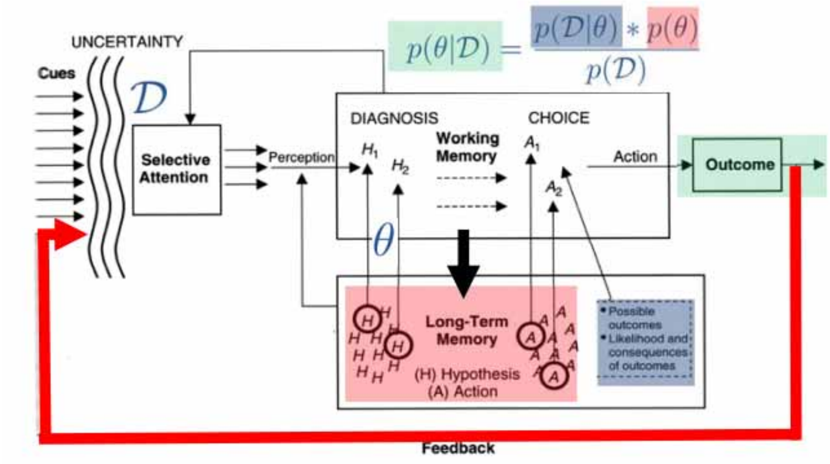
- 00 Reflection – follow-up from last lecture
- 01 Decision Support Systems (DSS)
- 02 History of DSS = History of AI
- 03 Example: Towards Personalized Medicine
- 04 Example: Case Based Reasoning (CBR)
- 05 Causal Reasoning
- 06 Explainability – Causability
- 07 (Some) Methods of Explainable AI

00 Reflection

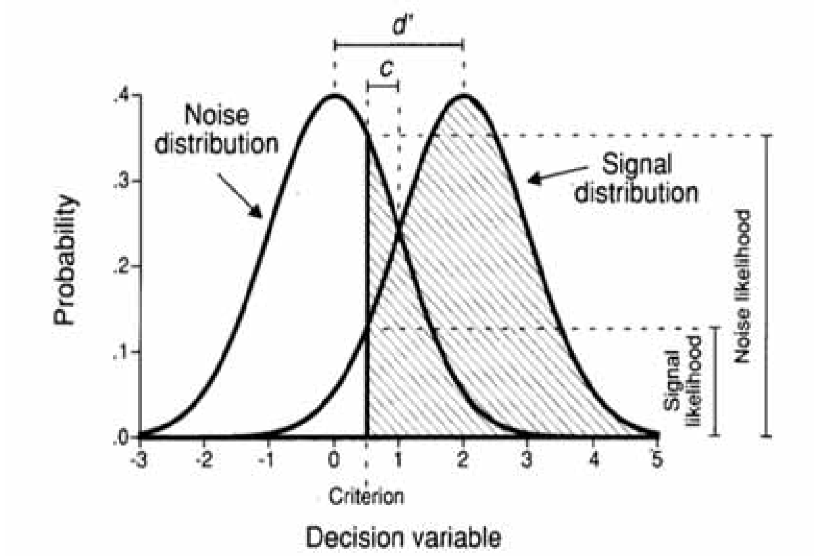


Reflection from last lecture

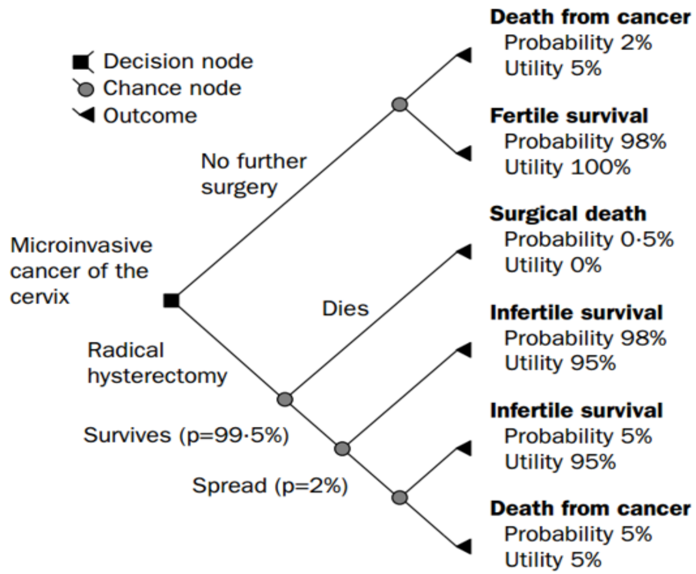




2

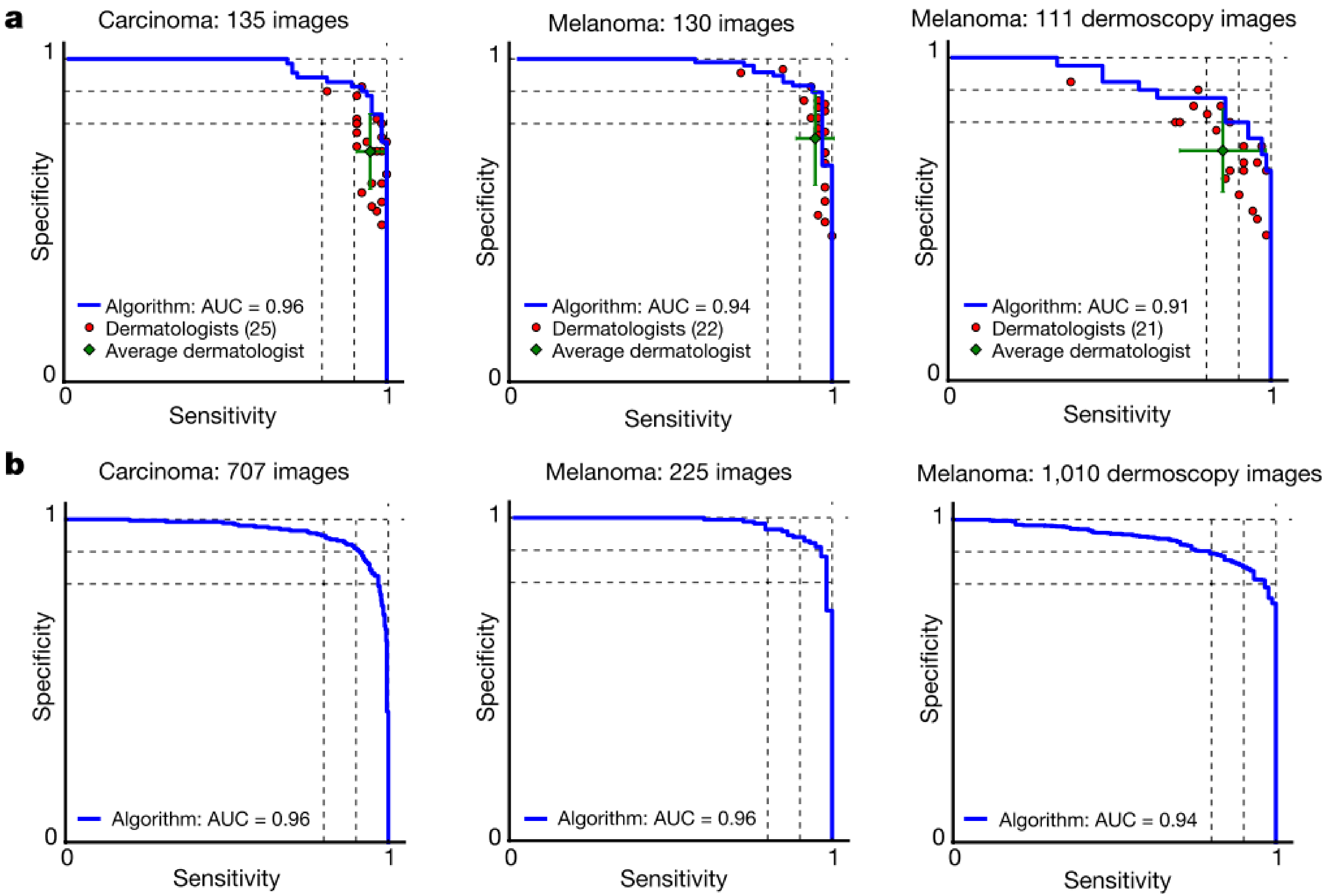


3



4

How do you explain this …

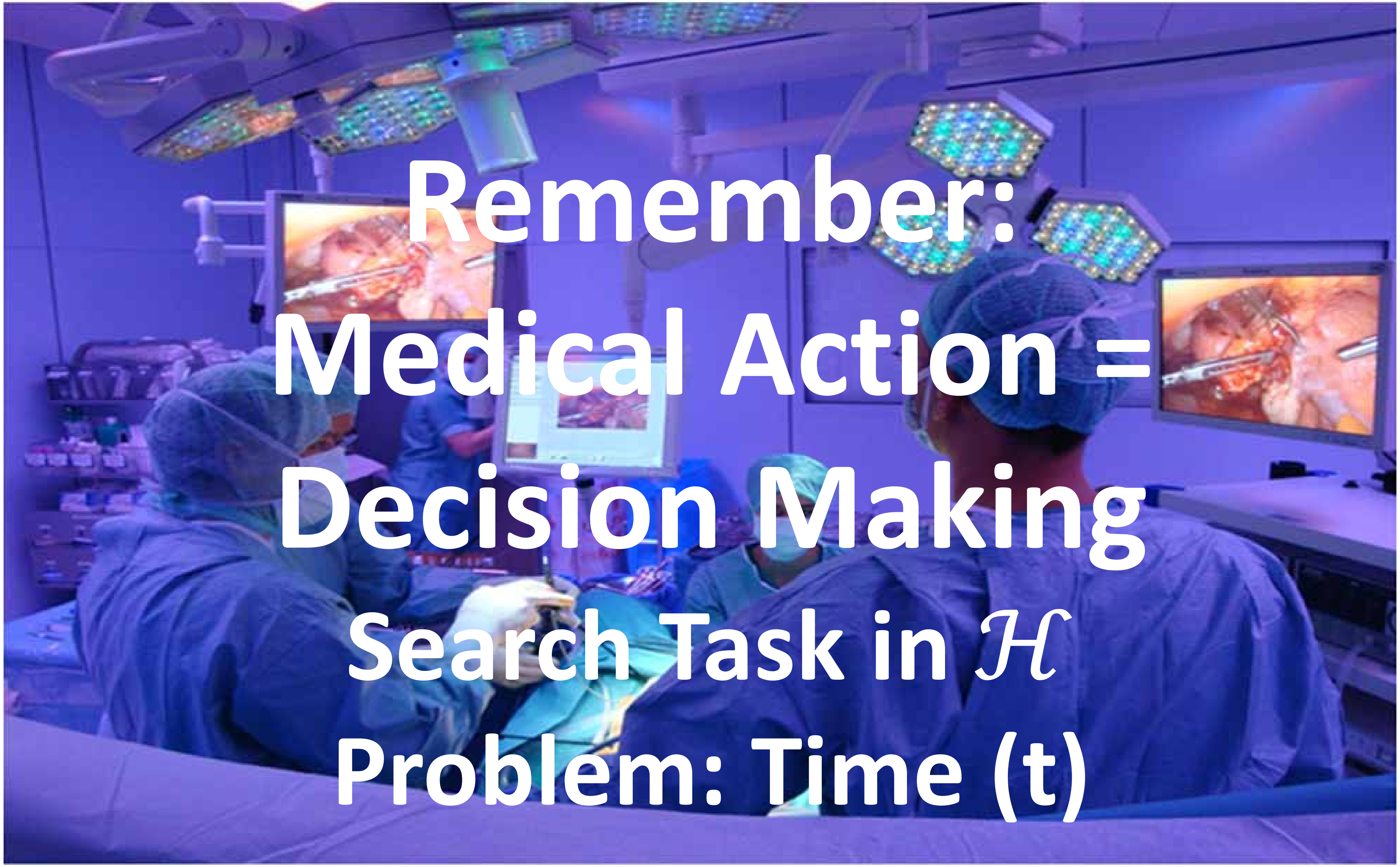


Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542, (7639), 115-118

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-ai interface and ai-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.!

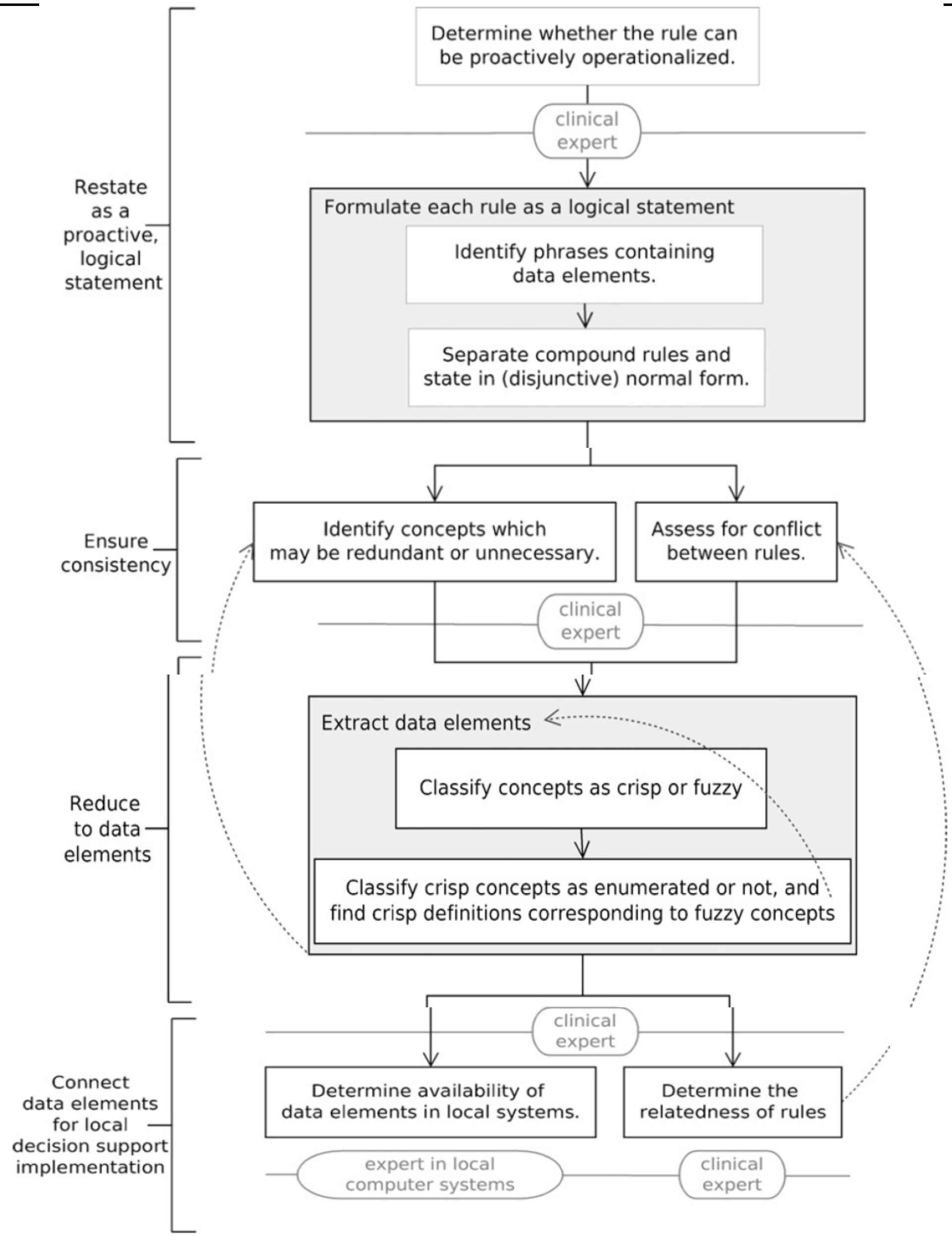
Decision Making is central in any (medical) work

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 …

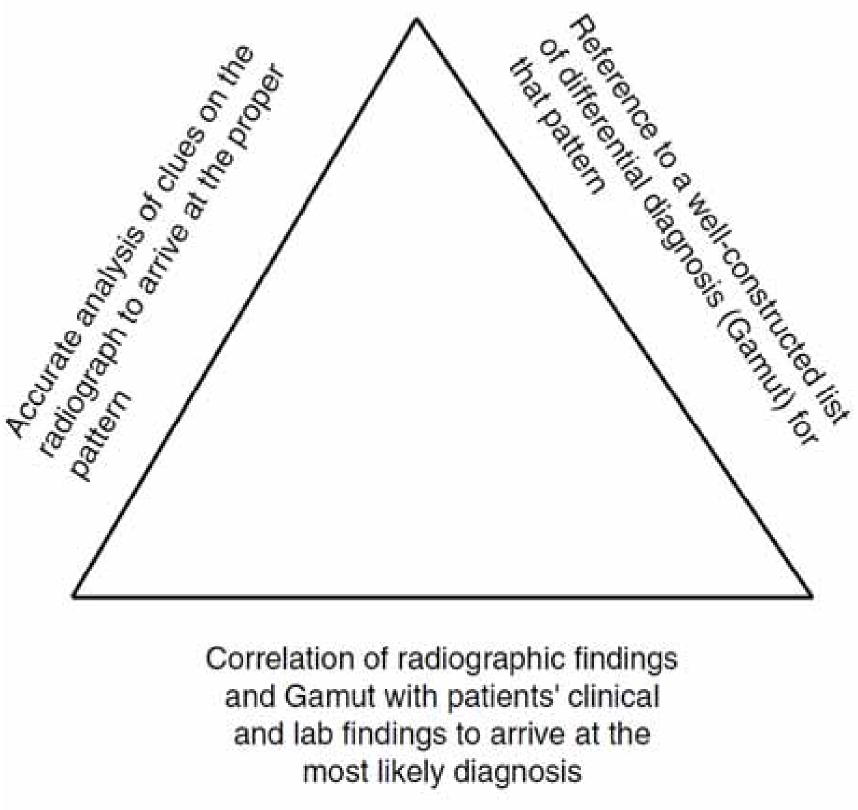
- Clinical guidelines are **systematically** developed documents to assist doctors and patient decisions about appropriate care;
- In order to build DS, based on a guideline, it is **formalized** (transformed from natural language to a logical algorithm), and
- **implemented** (using the algorithm to 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

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

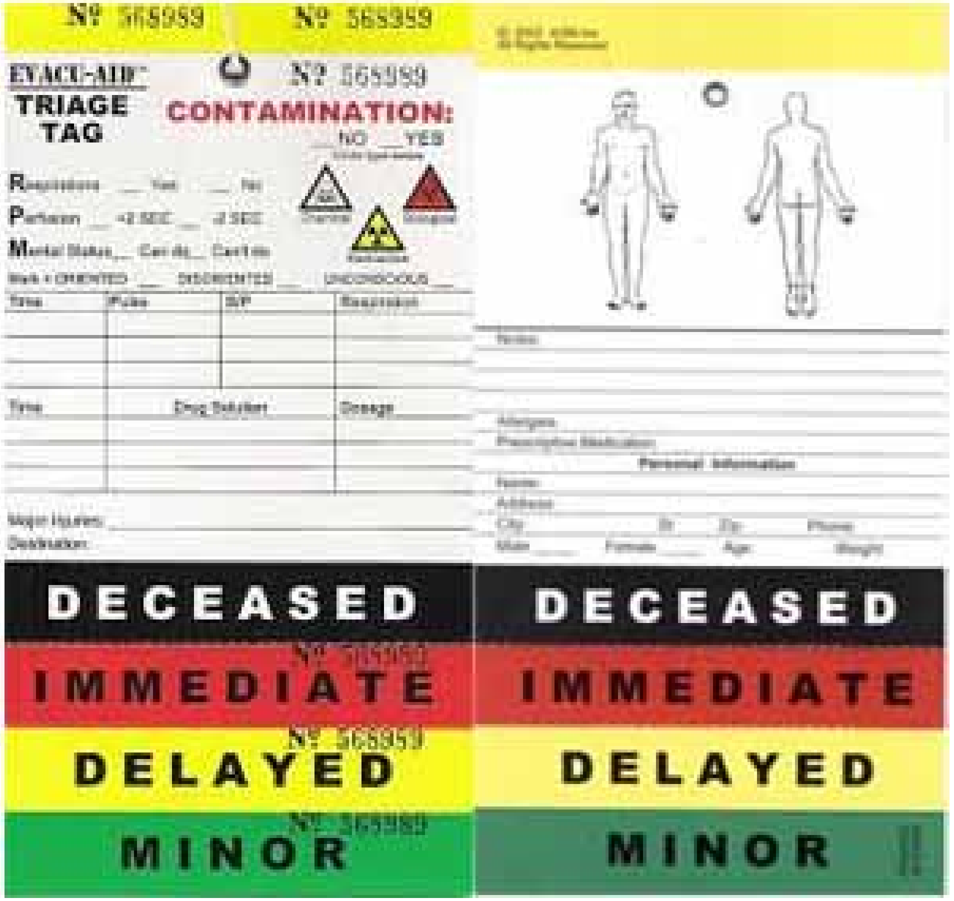
UNCOMMON
1. Corrosive gastritis
2. *Cryptosporidium* antritis
3. [Lymphoma]
4. Opportunistic infection (eg, candidiasis {moniliasis} ; herpes simplex; cytomegalovirus)
5. Postoperative gastritis
6. Radiation therapy
7. Zollinger-Ellison S. ; multiple endocrine neoplasia (MEN) S.

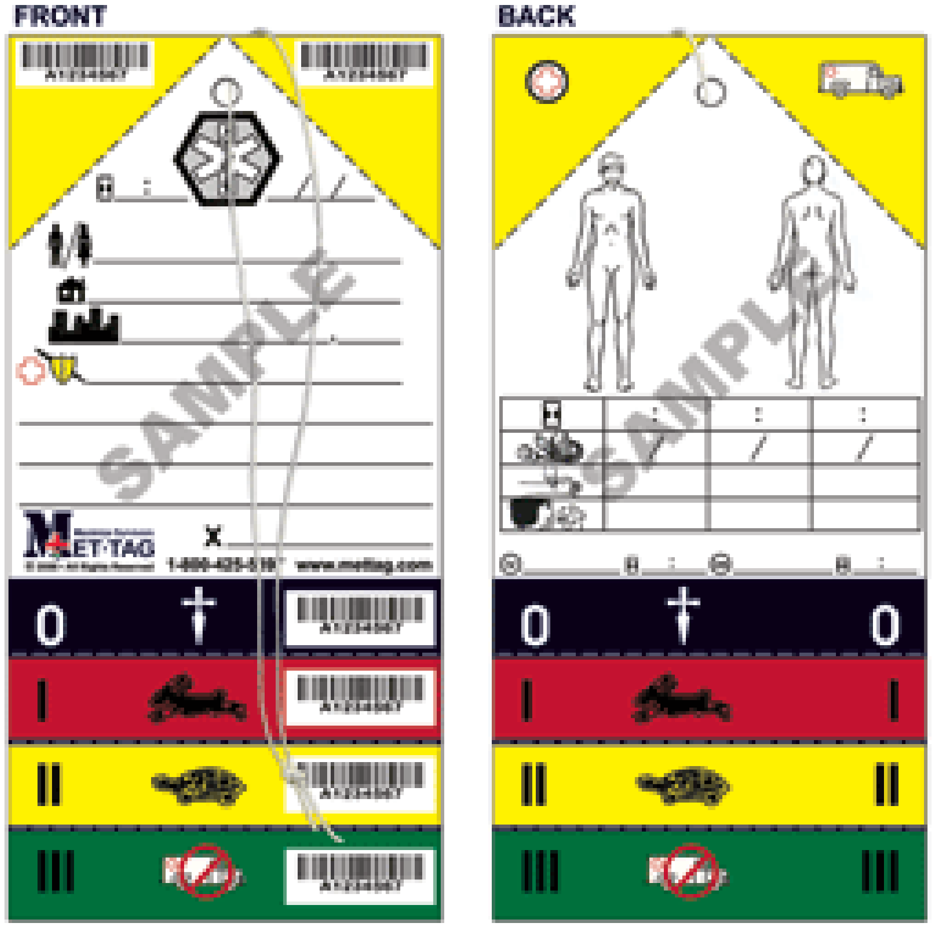
* Superficial erosions or aphthoid ulcerations seen especially with double contrast technique.

[ ] 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*

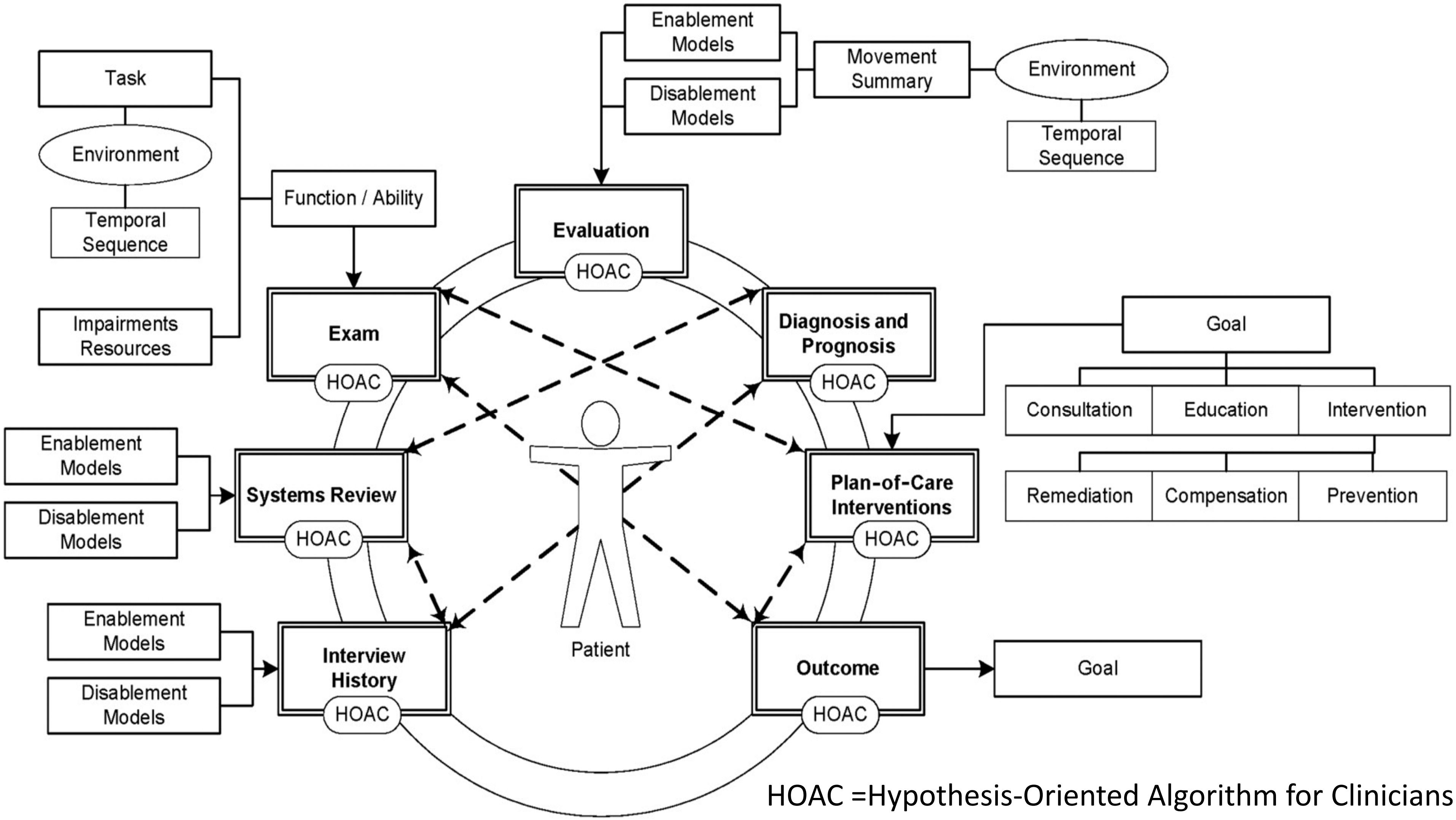
http://rfs.acr.org/gamuts/data/G-25.htm





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.*

Example Prediction Models > Feature Generation

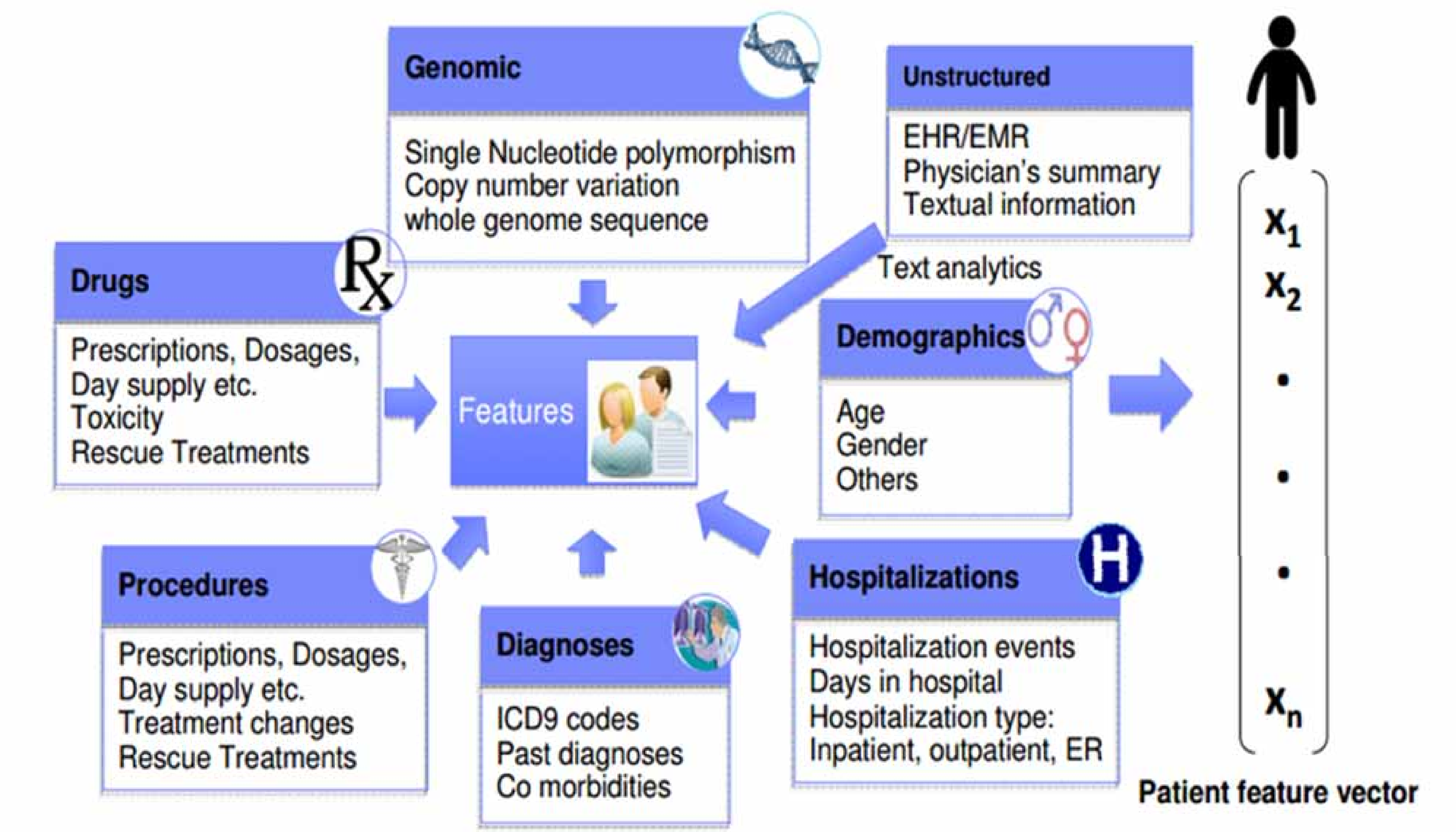


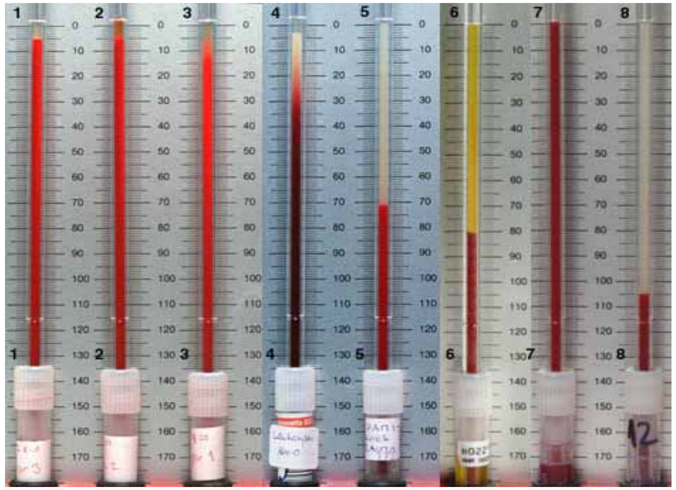
Image credit to Michal Rosen-Zvi

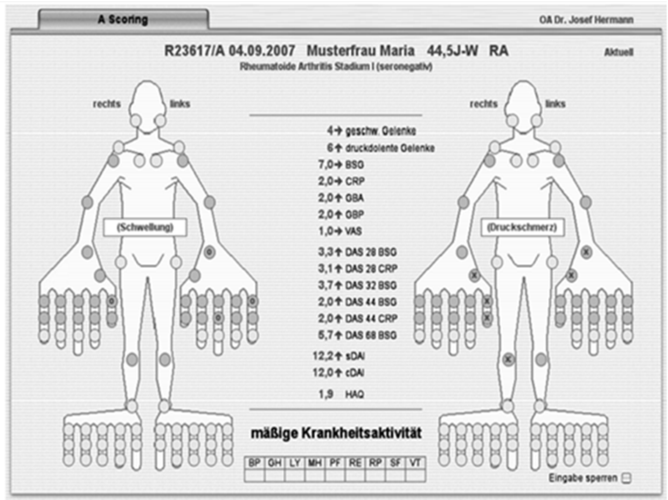
Chao, J., Parker, B. A. & Zvaifler, N. J. (2009) Accelerated Cutaneous Nodulosis Associated with Aromatase Inhibitor Therapy in a Patient with Rheumatoid Arthritis. *The Journal of Rheumatology, 36,* ***5, 1087-1088.***

Bone Changes …

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

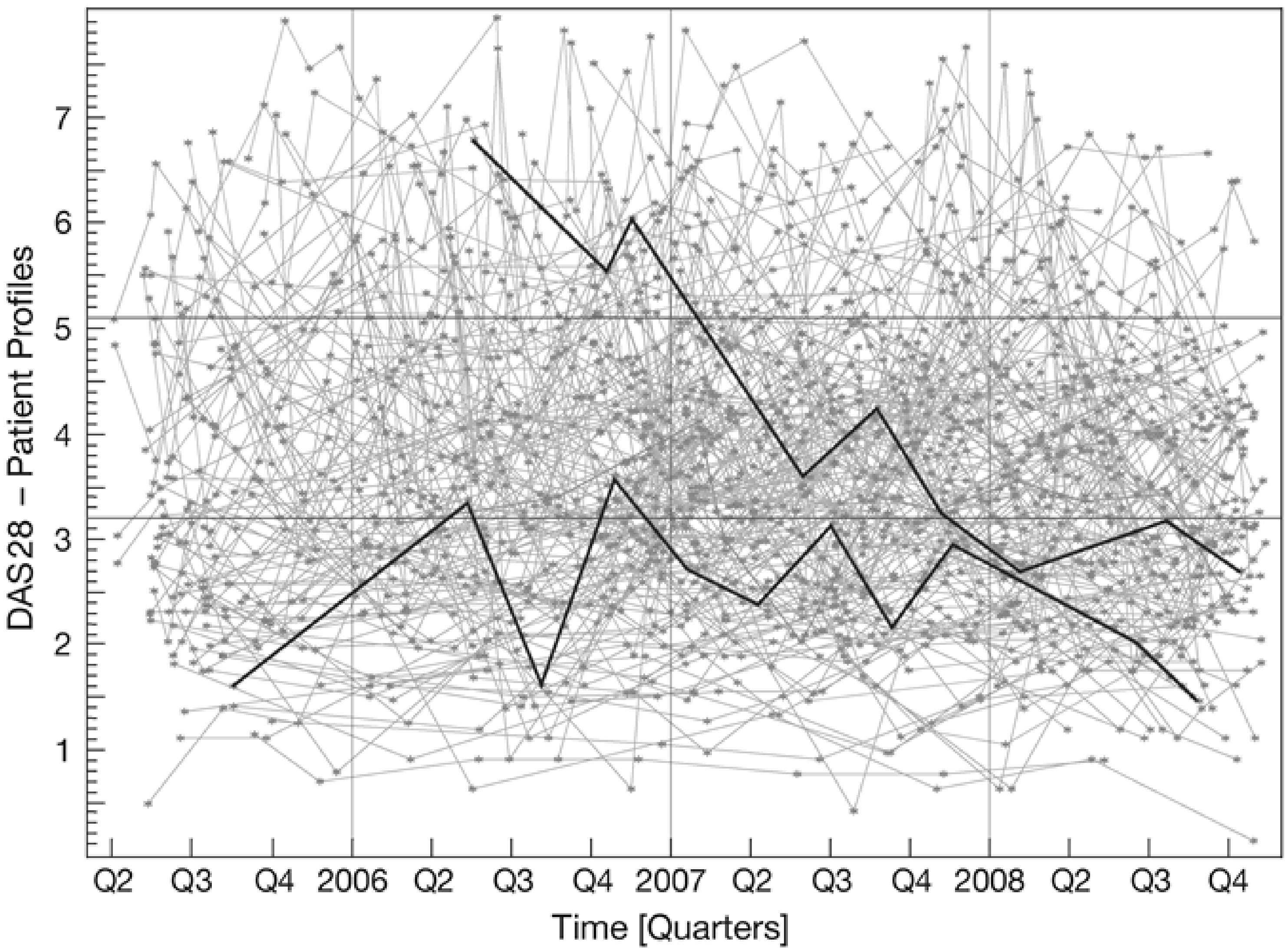
- 50+ Patients per day ~ 5000 data points per day …
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ **time-series data**





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.*

Can Computers help doctors to make better decisions?

For reading and discussion: Michael Duerr-Specht, Randy Goebel & Andreas Holzinger 2015. Medicine and Health Care as a Data Problem: Will Computers become better medical doctors? In: Holzinger, Andreas, Roecker, Carsten & Ziefle, Martina (eds.) Smart Health, State-of-the-Art SOTA Lecture Notes in Computer Science LNCS 8700. Heidelberg, Berlin, New York: Springer, pp. 21-40, doi:10.1007/978-3-319-16226-3_2.

Computers help human doctors to make better decisions?

Reasoning Process	Human	Computer
Abductive Hypothesis generation	Uniquely capable of complex pattern recognition and creative thought. "the whole is greater than the sum of its parts"	Matches multiple individual correlations from extensive data banks based on preconceived algorithms. Secondary construction of relationships. "the whole equals the sum of its parts"
Inductive Symptom → Disease	Limited database. Subject to biases - Anchoring bias - Confirmation bias - Premature closure	Extensive database. Probability based on Bayesian statistics, no significant bias. Limitation based on available data.
Deductive Disease → Symptoms, Treatment	Limited database. Personal intuition and experience affect decision making.	Extensive database. Application of rules of evidence based medicine with potential biases.

Michael Duerr-Specht, Randy Goebel & Andreas Holzinger 2015. Medicine and Health Care as a Data Problem: Will Computers become better medical doctors? In: Holzinger, Andreas, Roecker, Carsten & Ziefle, Martina (eds.) Smart Health, State-of-the-Art SOTA Lecture Notes in Computer Science LNCS 8700. Heidelberg, Berlin, New York: Springer, pp. 21-40, doi:10.1007/978-3-319-16226-3_2.

monday afternoon

december 9

3:45 p.m. / arena

Chairman:
DR. D. C. ENGELBART
Stanford Research Institute
Menlo Park, California

a research center for augmenting human intellect

This session is entirely devoted to a presentation by Dr. Engelbart on a computer-based, interactive, multiconsole display system which is being developed at Stanford Research Institute under the sponsorship of ARPA, NASA and RADC. The system is being used as an experimental laboratory for investigating principles by which interactive computer aids can augment intellectual capability. The techniques which are being described will, themselves, be used to augment the presentation.
The session will use an on-line, closed circuit television hook-up to the SRI computing system in Menlo Park.
Following the presentation remote terminals to the system, in operation, may be viewed during the remainder of the conference in a special room set aside for that purpose.

COMPUTER SCIENCE

Superhuman AI for heads-up no-limit poker: Libratus beats top professionals

Noam Brown and Tuomas Sandholm*

No-limit Texas hold'em is the most popular form of poker. Despite artificial intelligence (AI) successes in perfect-information games, the private information and massive game tree have made no-limit poker difficult to tackle. We present Libratus, an AI that, in a 120,000-hand competition, defeated four top human specialist professionals in heads-up no-limit Texas hold'em, the leading benchmark and long-standing challenge problem in imperfect-information game solving. Our game-theoretic approach features application-independent techniques: an algorithm for computing a blueprint for the overall strategy, an algorithm that fleshes out the details of the strategy for subgames that are reached during play, and a self-improver algorithm that fixes potential weaknesses that opponents have identified in the blueprint strategy.

Noam Brown & Tuomas Sandholm 2018. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. Science, 359, (6374), 418-424, doi:10.1126/science.aao1733.

Source: https://web.stanford.edu/dept/SUL/library/extra4/sloan/mousesite/dce1968conferenceannouncement.jpg

Computer Science > Artificial Intelligence

Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology

Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Müller, Robert Reihs, Kurt Zatloukal

(Submitted on 19 Dec 2017)

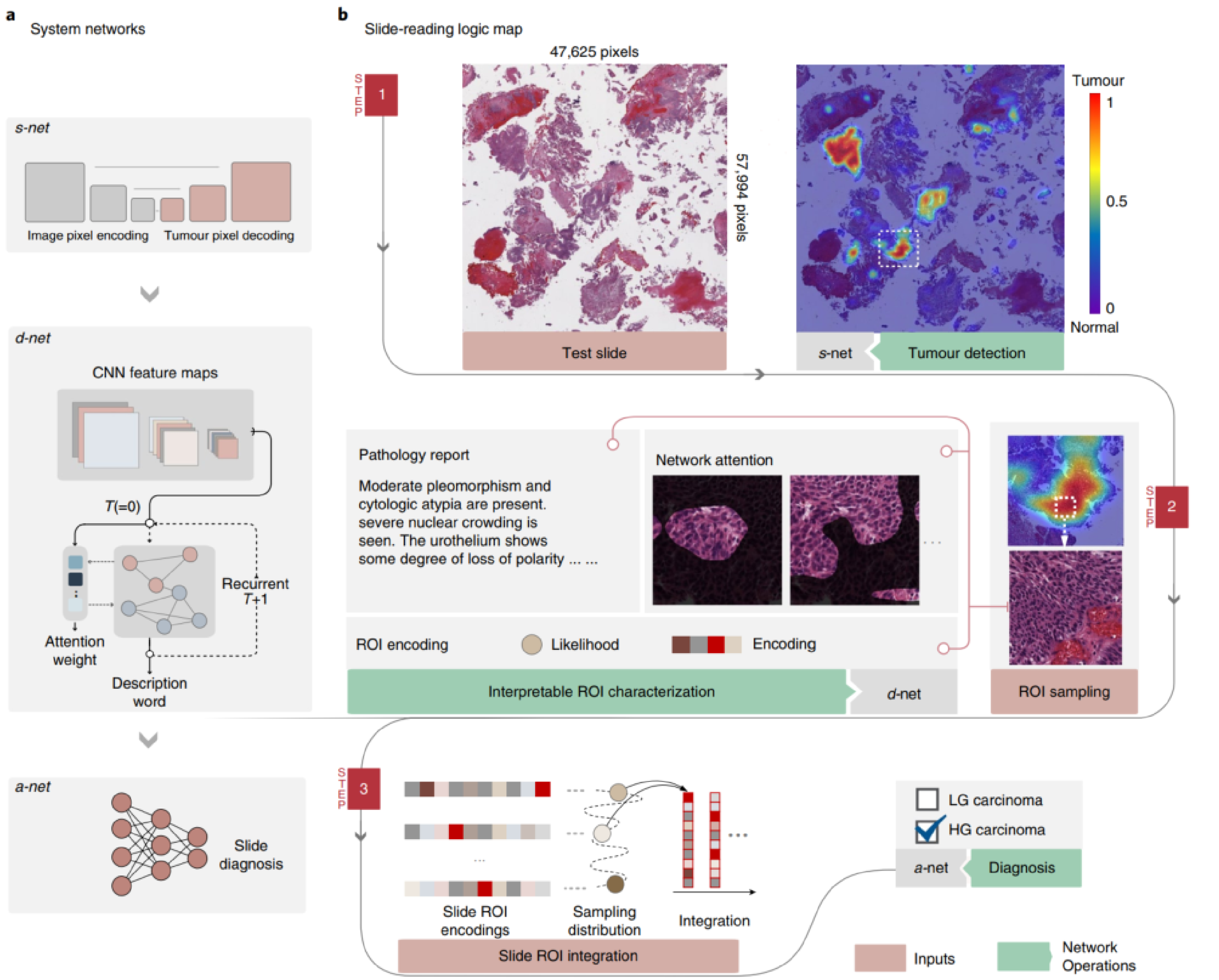
Digital pathology is not only one of the most promising fields of diagnostic medicine, but at the same time a hot topic for fundamental research. Digital pathology is not just the transfer of histopathological slides into digital representations. The combination of different data sources (images, patient records, and *omics data) together with current advances in artificial intelligence/machine learning enable to make novel information accessible and quantifiable to a human expert, which is not yet available and not exploited in current medical settings. The grand goal is to reach a level of usable intelligence to understand the data in the context of an application task, thereby making machine decisions transparent, interpretable and explainable. The foundation of such an "augmented pathologist" needs an integrated approach. While machine learning algorithms require many thousands of training examples, a human expert is often confronted with only a few data points. Interestingly, humans can learn from such few examples and are able to instantly interpret complex patterns. Consequently, the grand goal is to combine the possibilities of artificial intelligence with human intelligence and to find a well-suited balance between them to enable what neither of them could do on their own. This can raise the quality of education, diagnosis, prognosis and prediction of cancer and other diseases. In this paper we describe some (incomplete) research issues which we believe should be addressed in an integrated and concerted effort for paving the way towards the augmented pathologist.

Subjects: **Artificial Intelligence (cs.AI)**; Machine Learning (stat.ML)
Cite as: arXiv:1712.06657 [cs.AI]

Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Müller, Robert Reihs & Kurt Zatloukal 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. arXiv:1712.06657.

Pathologist level interpretable whole-slide diagnosis

Zizhao Zhang, Pingjun Chen, Mason Mcgough, Fuyong Xing, Chunbao Wang,
Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui & Jasreman Dhillon 2019.
Pathologist-level interpretable whole-slide cancer diagnosis with deep learning.
Nature Machine Intelligence, 1, (5), 236-245, doi:10.1038/s42256-019-0052-1.

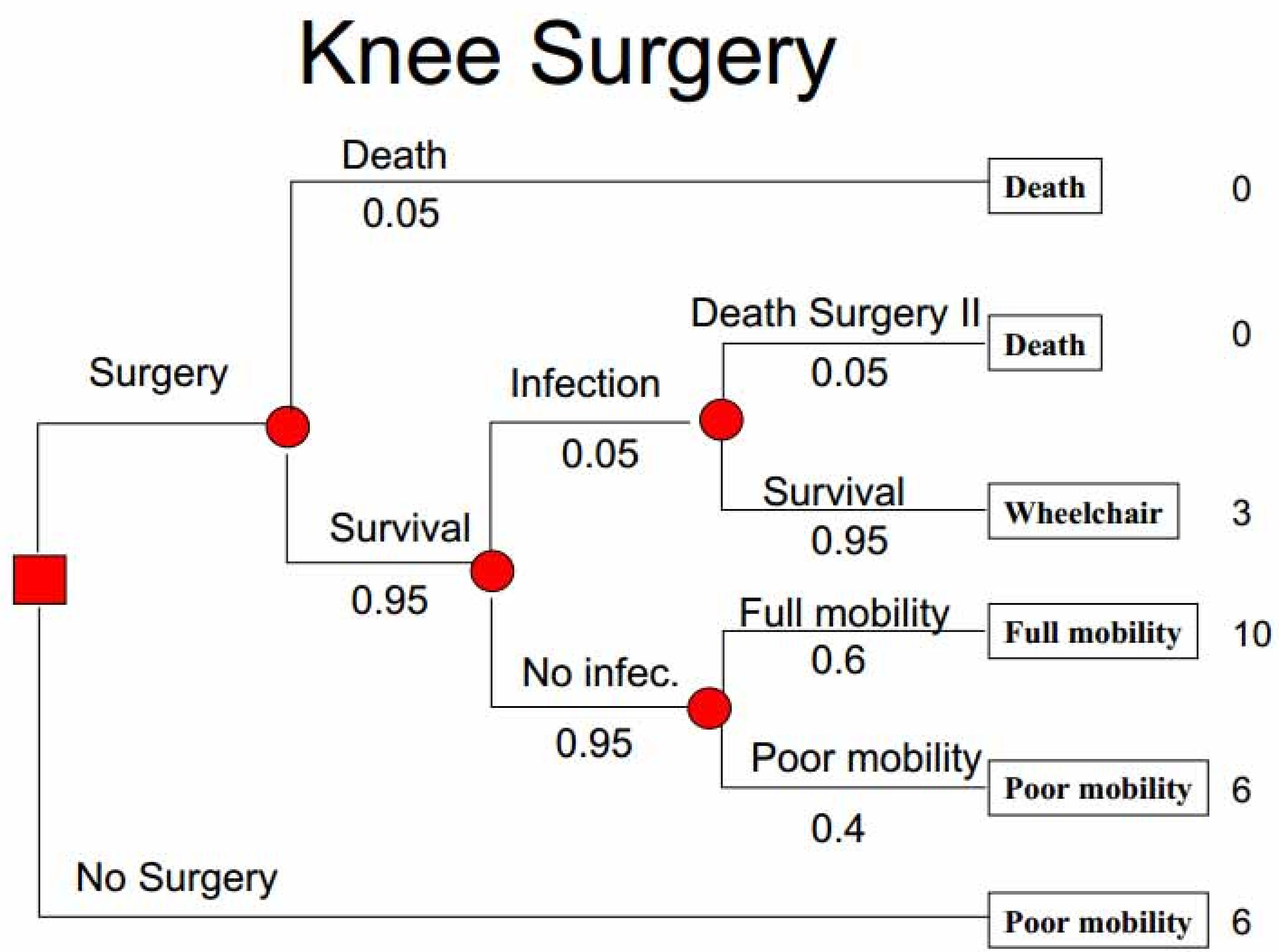


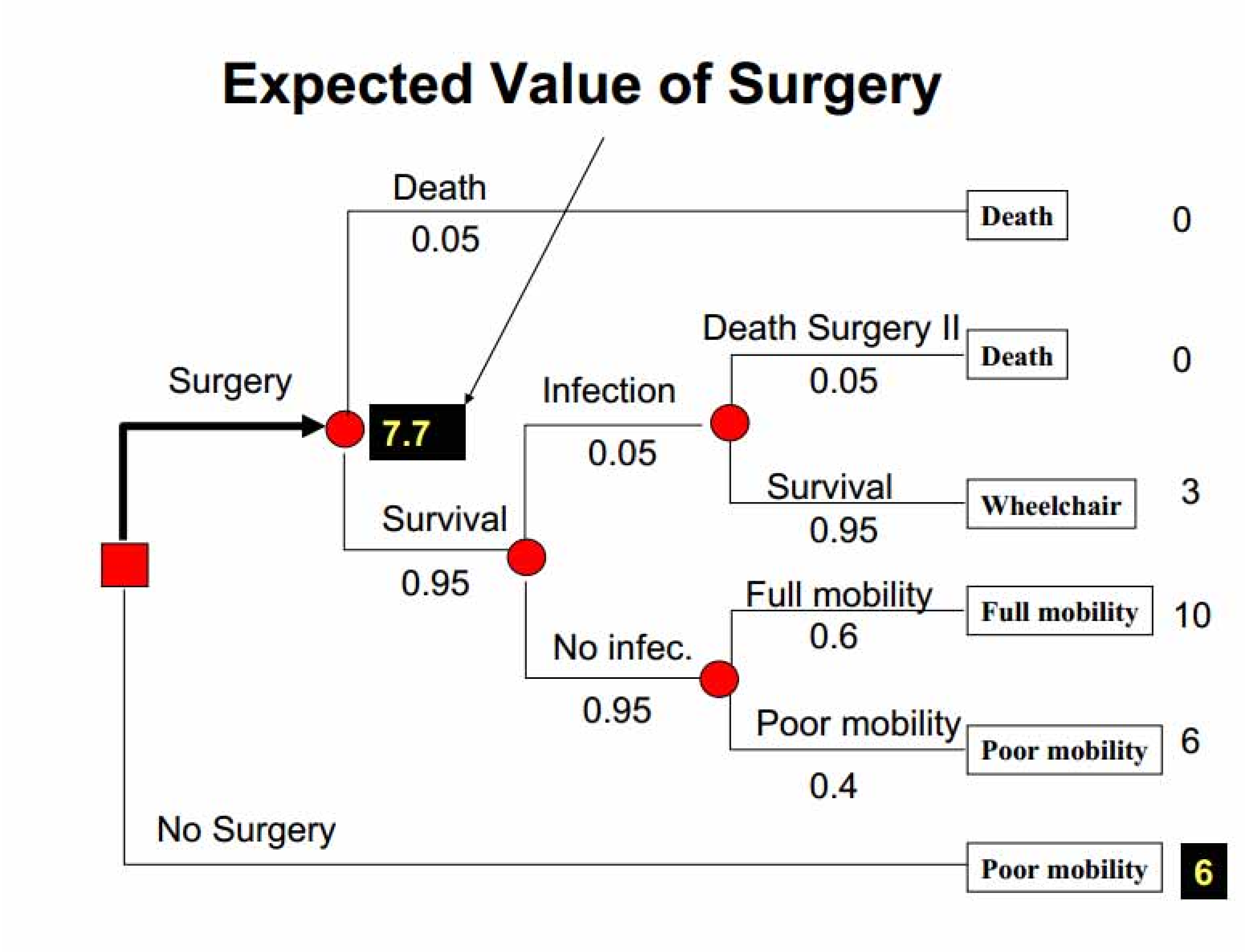
- **Type 1 Decisions:** related to the **diagnosis**, i.e. AI/ML is used to assist in diagnosing a disease on the basis of the individual patient data. Questions include:
  - What is the probability that this patient has a myocardial infarction on the basis of given data (patient history, ECG, …)?
  - 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. AI/ML is 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?

Jan H. Van Bemmel & Mark A. Musen 1997. Handbook of Medical Informatics, Heidelberg, Springer.

- 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


Knee Surgery
Death
0.05
Death
0
Death Surgery II
0.05
Death
0
Surgery
Infection
0.05
Survival
0.95
Wheelchair
3
Survival
0.95
Full mobility
0.6
Full mobility
10
No infec.
0.95
Poor mobility
0.4
Poor mobility
6
No Surgery
Poor mobility
6



Expected Value of Surgery
Death
0.05
Death
0
Death Surgery II
0.05
Death
0
Surgery
7.7
Infection
0.05
Survival
0.95
Wheelchair
3
Survival
0.95
Full mobility
0.6
Full mobility
10
No infec.
0.95
Poor mobility
0.4
Poor mobility
6
No Surgery
Poor mobility
6

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

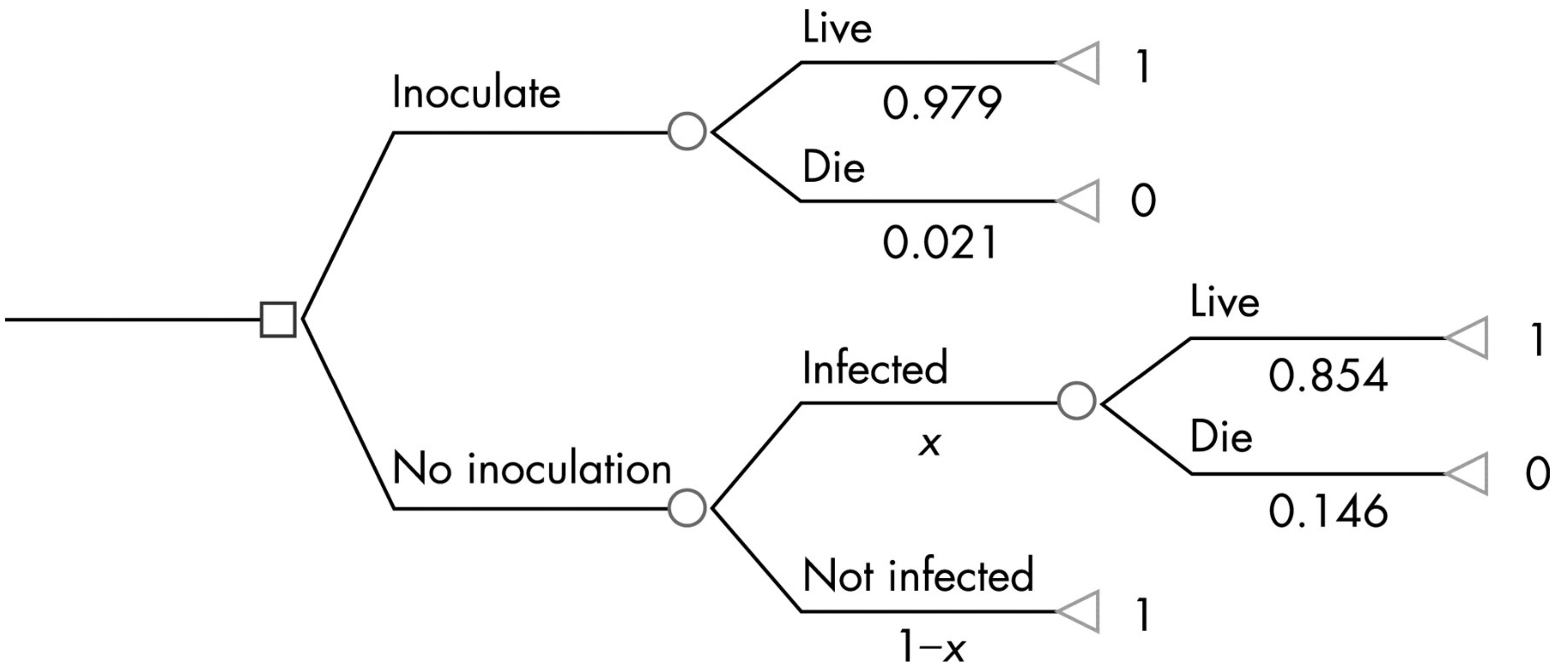
$$E(U \mid d) = \sum_{x_1,\ldots,x_n} P(x_1,\ldots,x_n \mid d) U(x_1,\ldots,x_n,d)$$

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

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

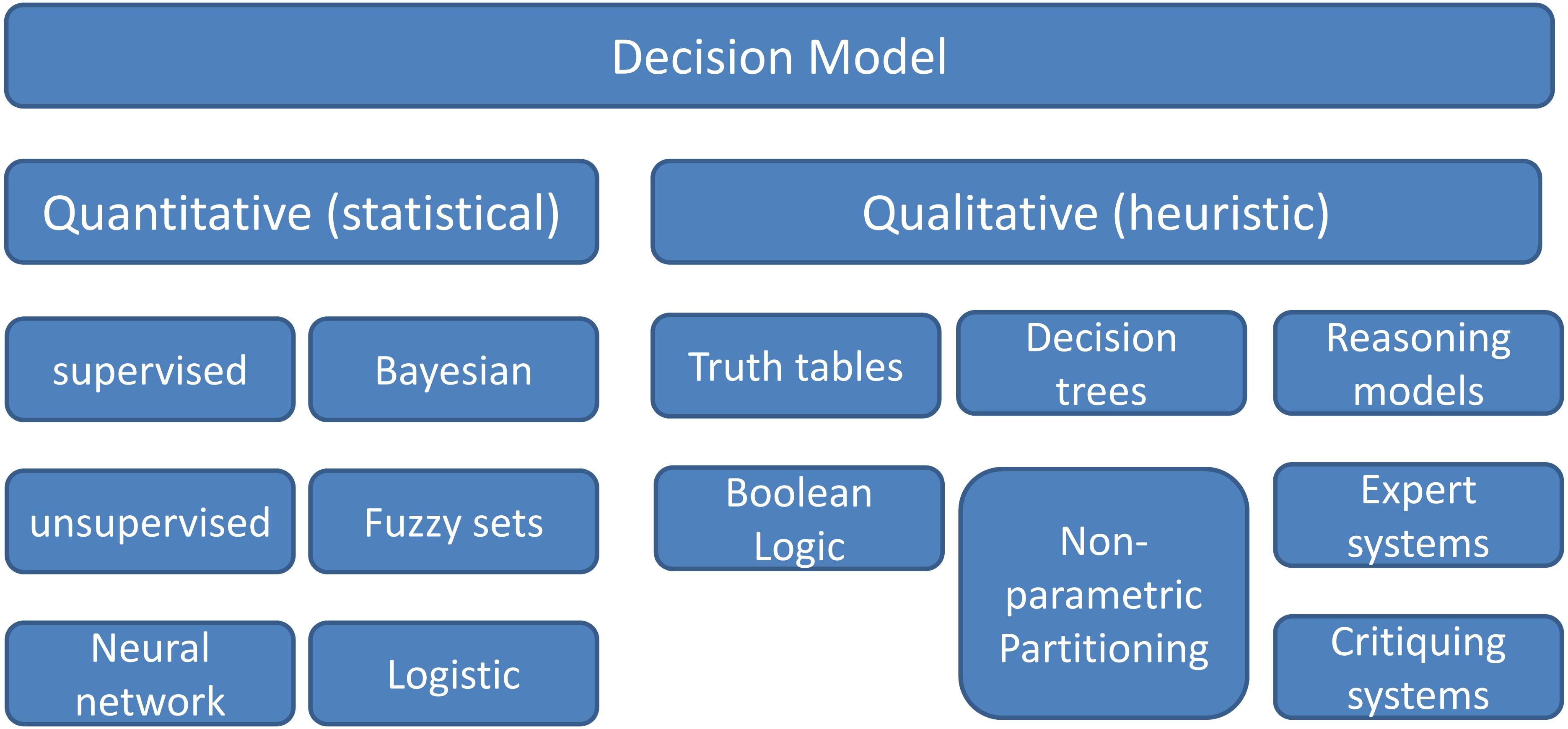
John Von Neumann & Oskar Morgenstern 1944. *Theory of games and economic behavior,* Princeton, Princeton university press.

Clinical Decision Tree (CDT) is still state-of-the-art



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.*

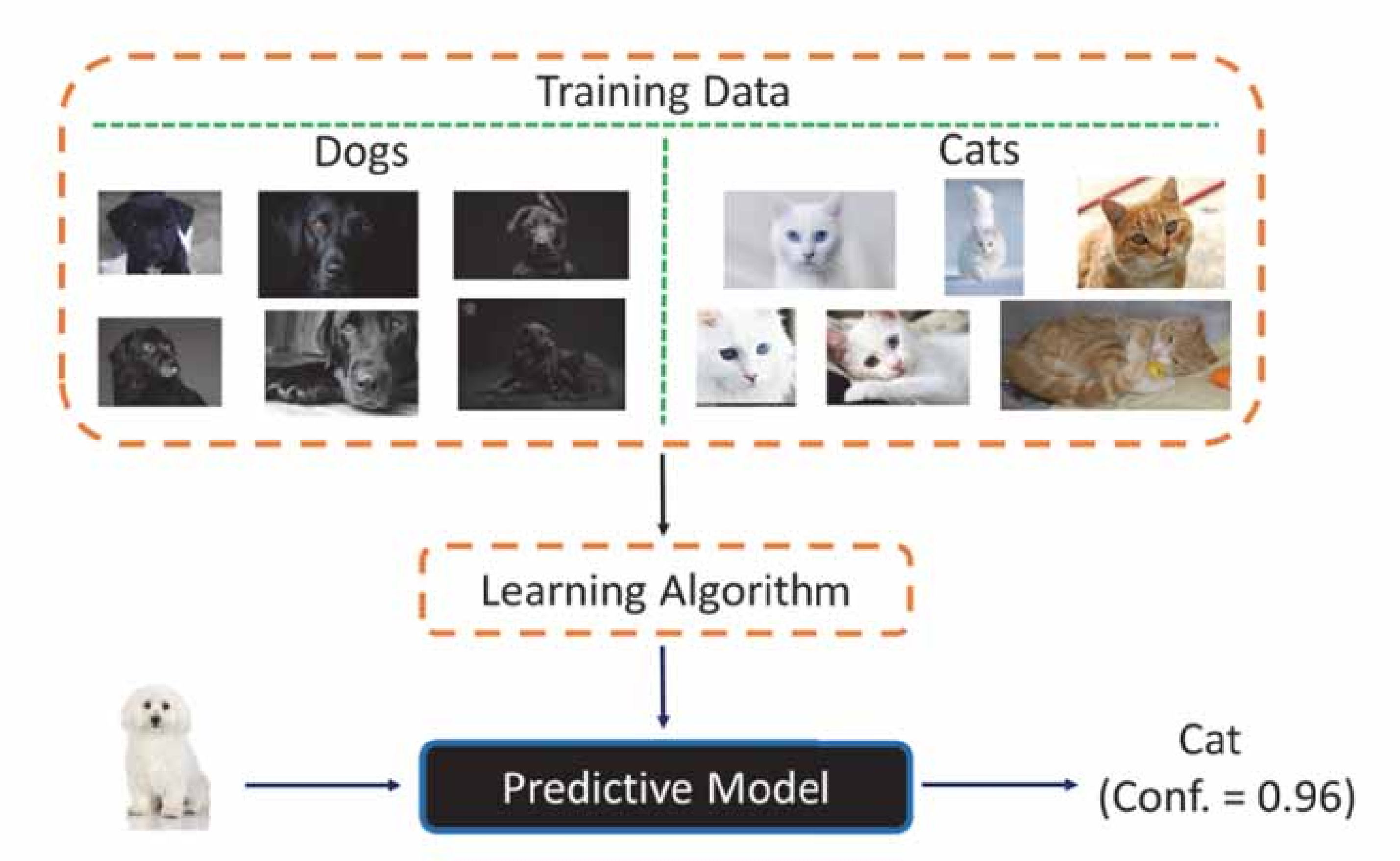
Taxonomy of Decision Support Models



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

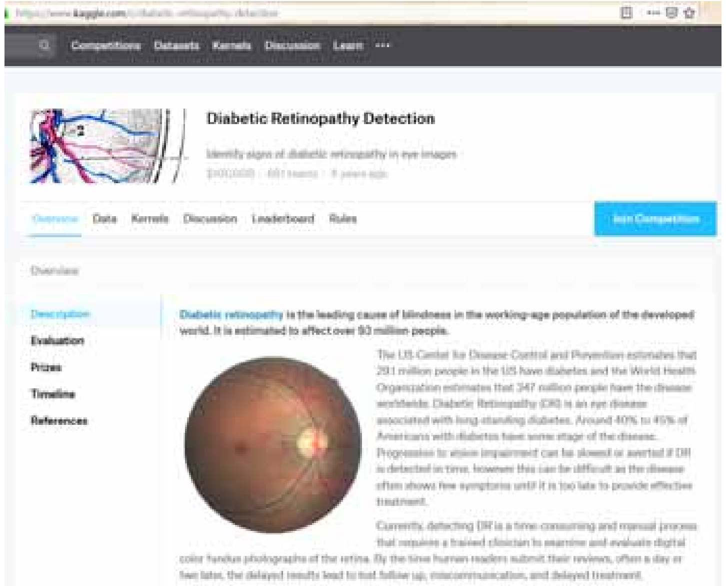
- Need for robust algorithms
- Need for trustworthy, fair and accountable algorithms
- Augmenting the doctor – not replacing them, but let “Chimpanzee”-Work do by algorithms
- Focus of the doctors to cognitively high-end demanding, challenging work
- Double-Check (“look at this corner, maybe there is something relevant)”
- Many of the questions of medical doctors need causal explanations “the why” !!

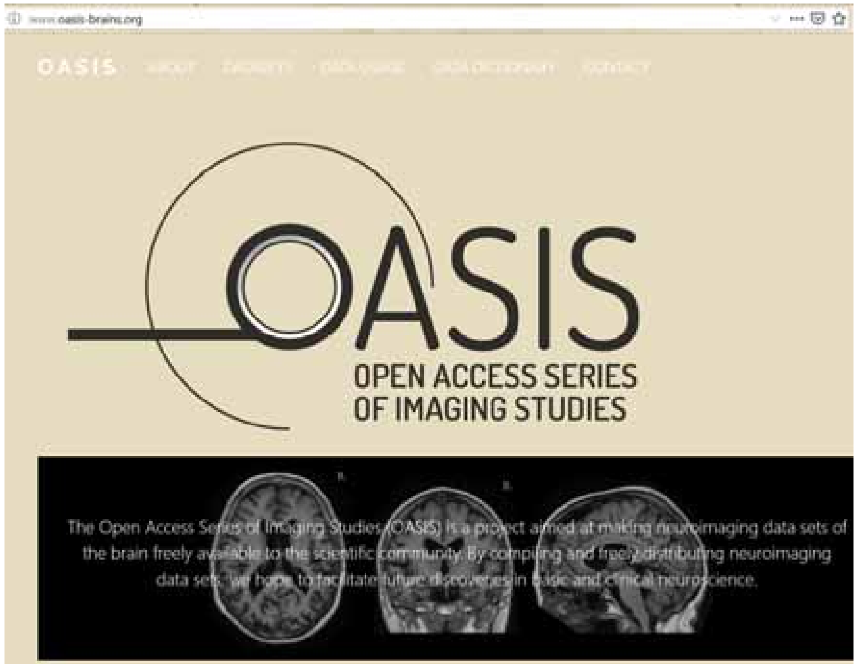
Big chance for medicine Identifying Unknown Unknowns

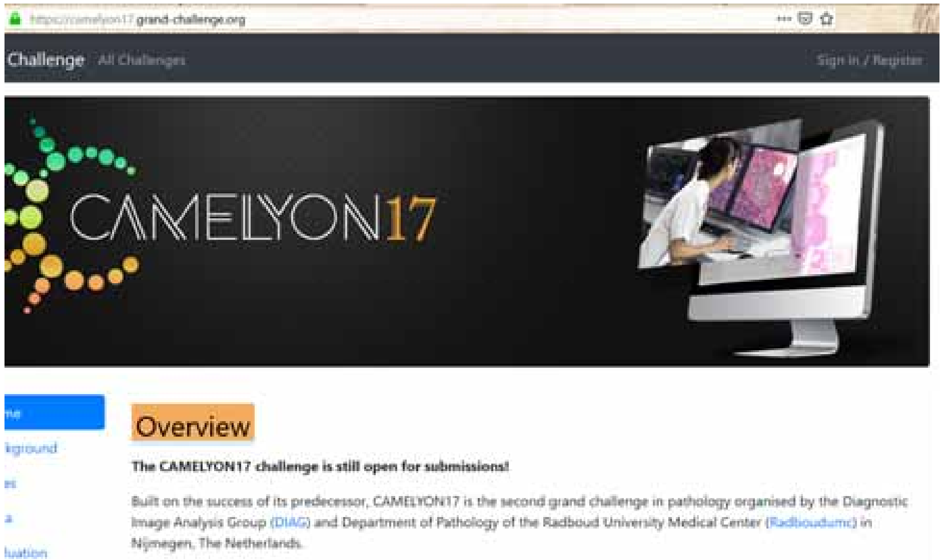


Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Eric Horvitz. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. Thirty-First AAAI Conference on Artificial Intelligence, 2017.

Example medical data sets openly available

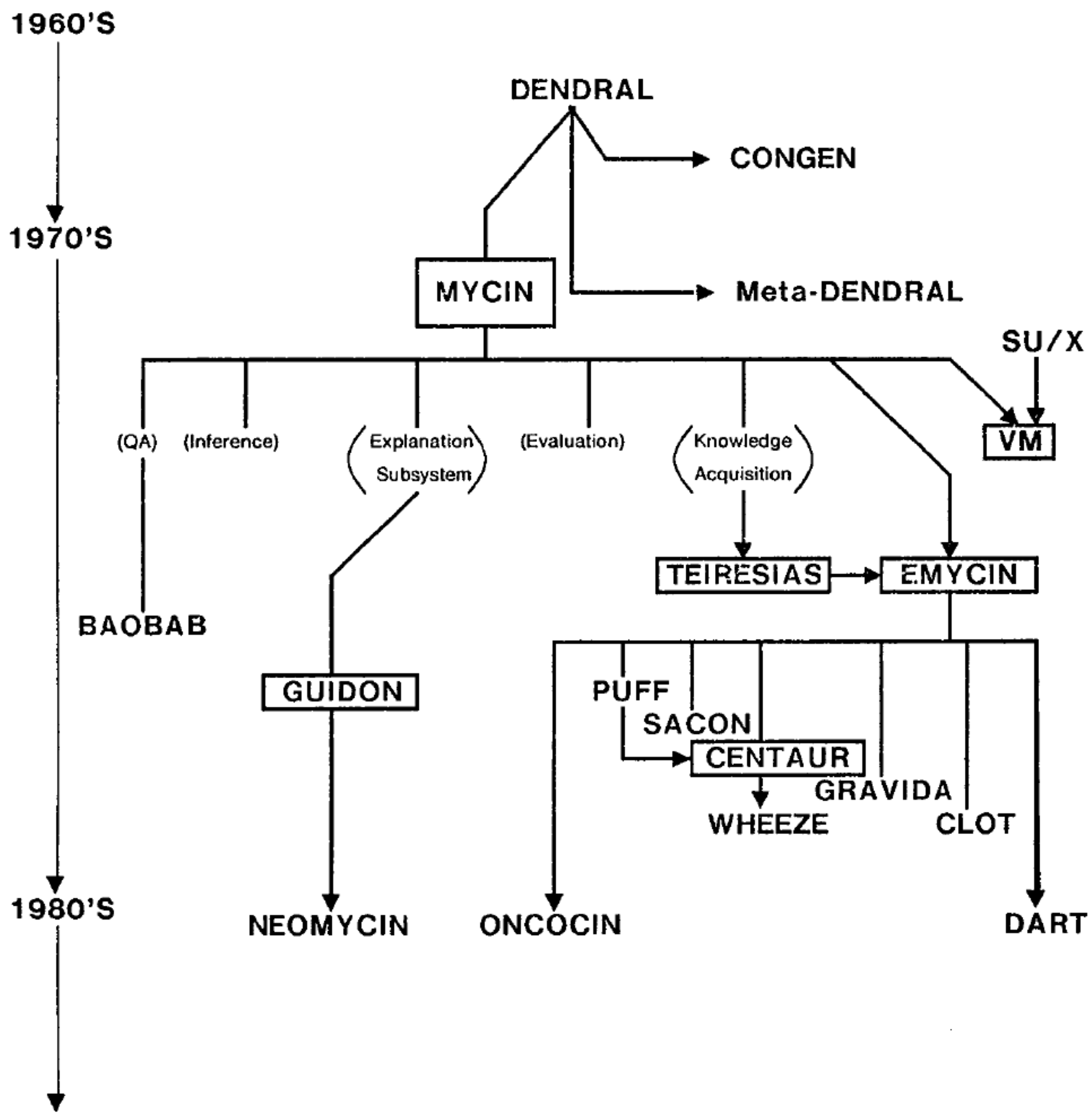






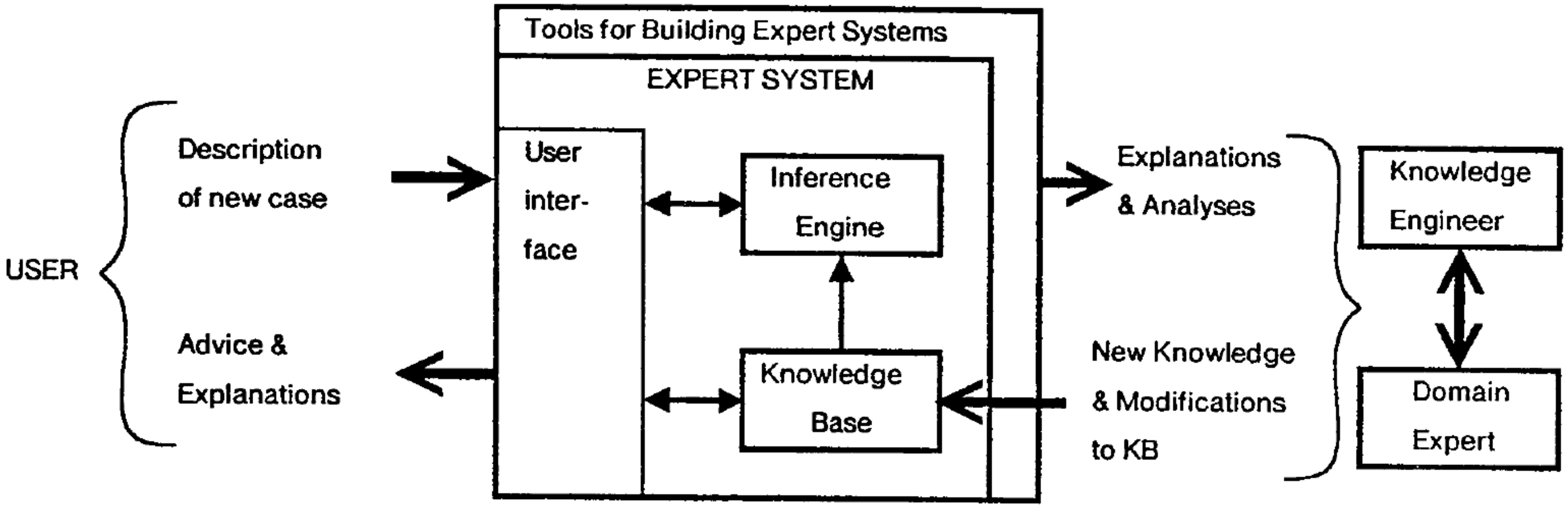
02 History of DSS = History of AI

- **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.
- 1958 John McCarthy Advice Taker: programs with common sense
- **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.
- 1978 Bellman, R. Can Computers Think? Automation of Thinking, problem solving, decision-making …



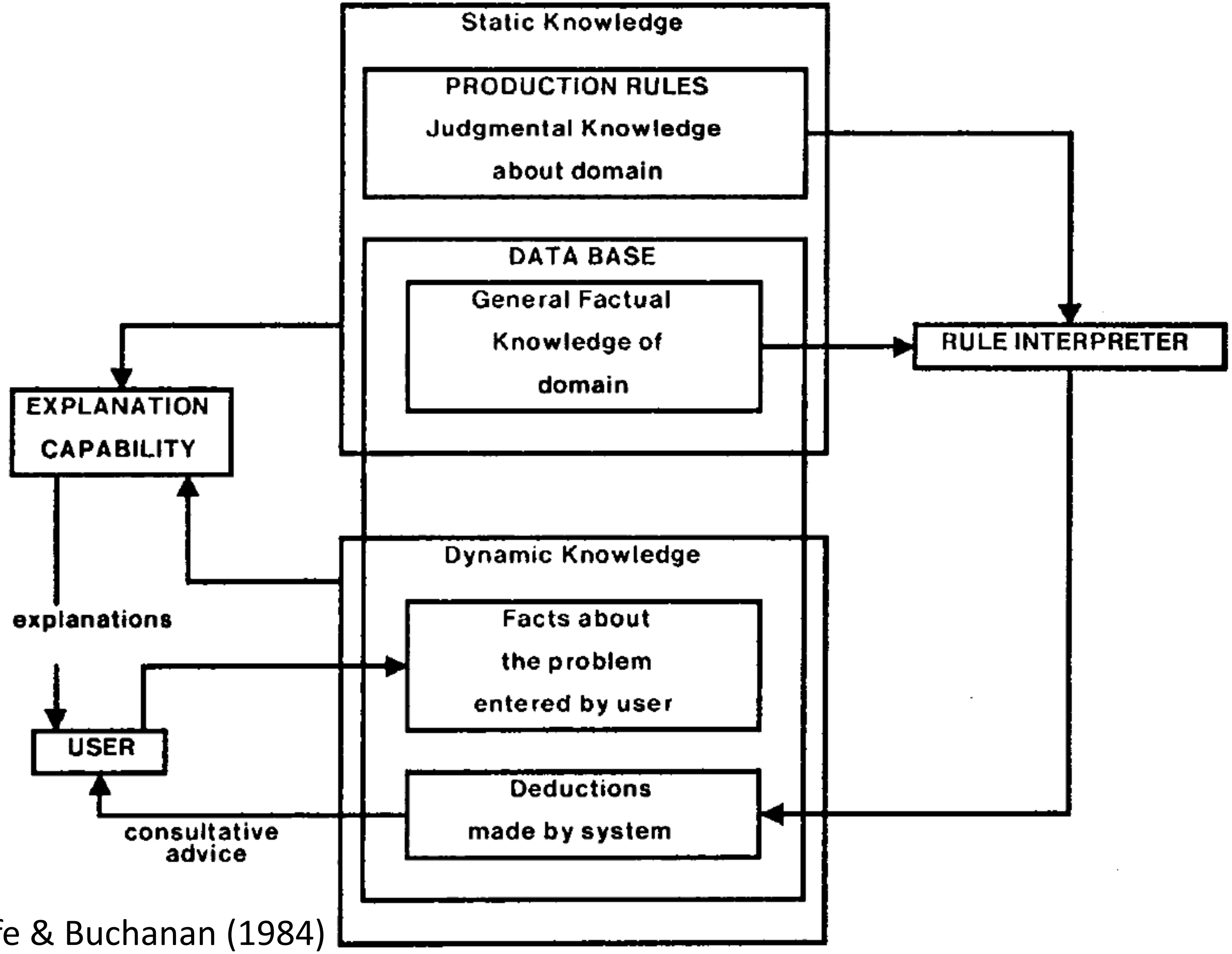
Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*

Early Knowledge Based System Architecture



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.*

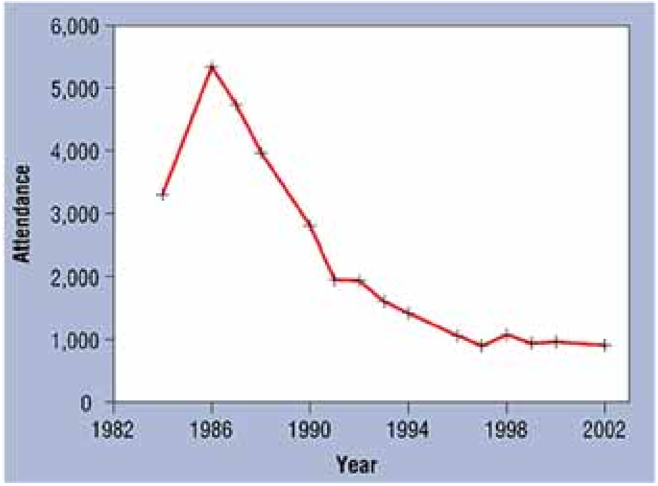
MYCIN was *no* success in the clinical routine

Image credit to Bernhard Schölkopf

Image credit to Bernhard Schölkopf

The AI winter was bitter cold …

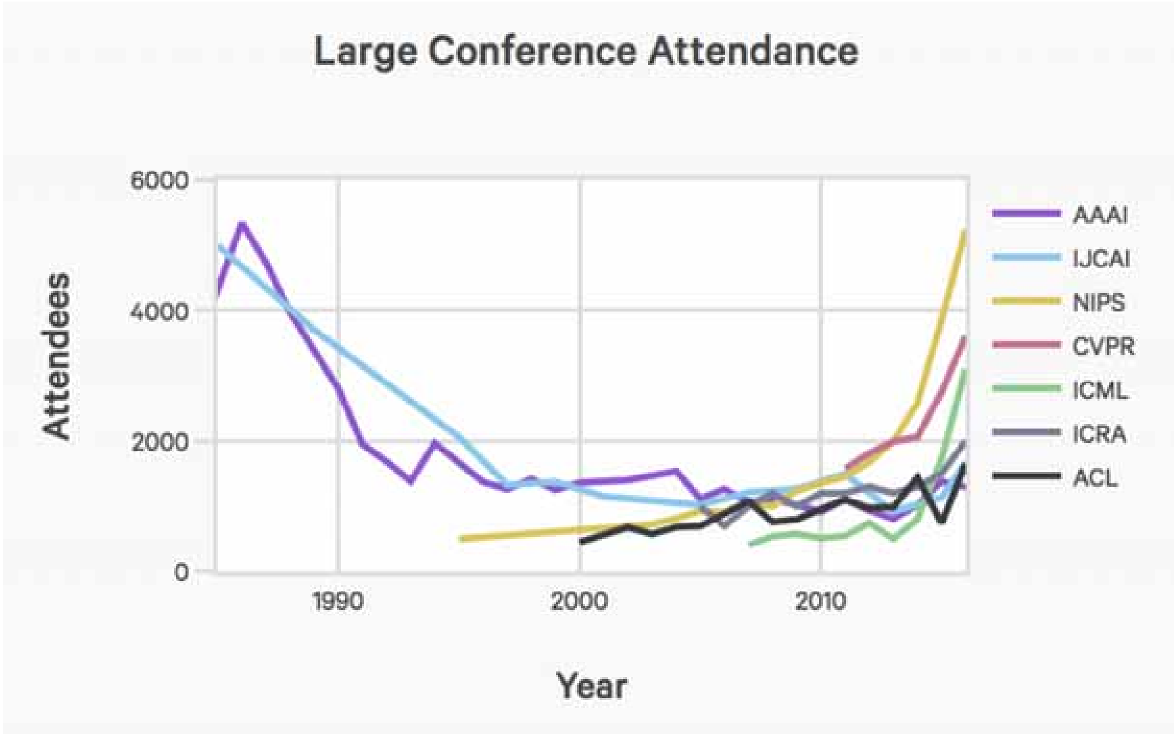
James Hendler 2008. Avoiding another AI winter. *IEEE Intelligent Systems,* 23, (2), 2-4, doi:10.1109/MIS.2008.20.



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

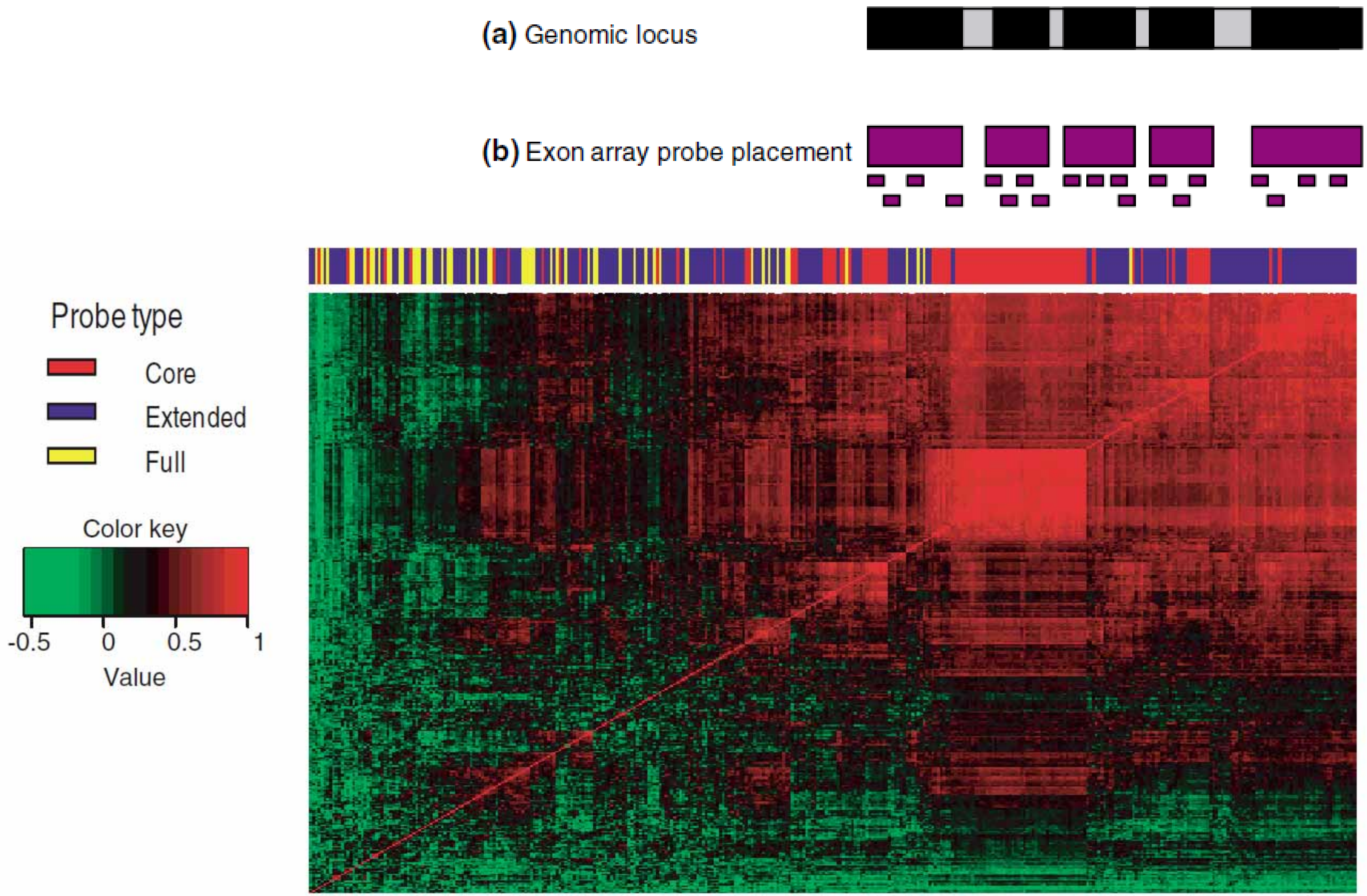
AAAI = AAAI Conference on Artificial Intelligence: https://aaai.org/Conferences/AAAI-20/

International Joint Conference on Artificial Intelligence: https://ijcai20.org/



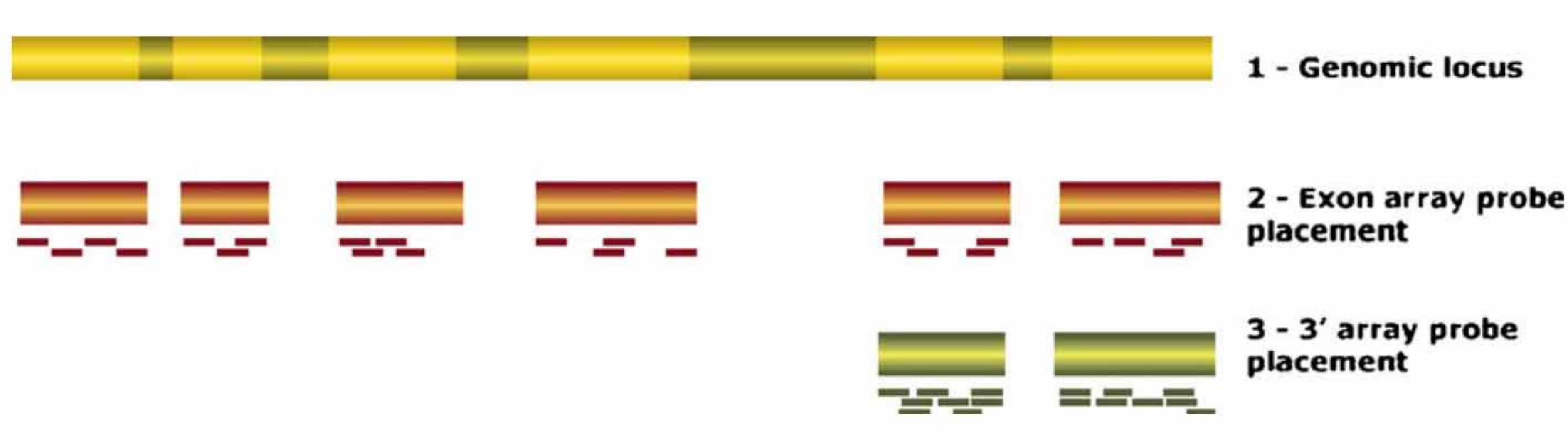
https://medium.com/machine-learning-in-practice/nips-accepted-papers-stats-26f124843aa0

03 Example: P4-Medicine



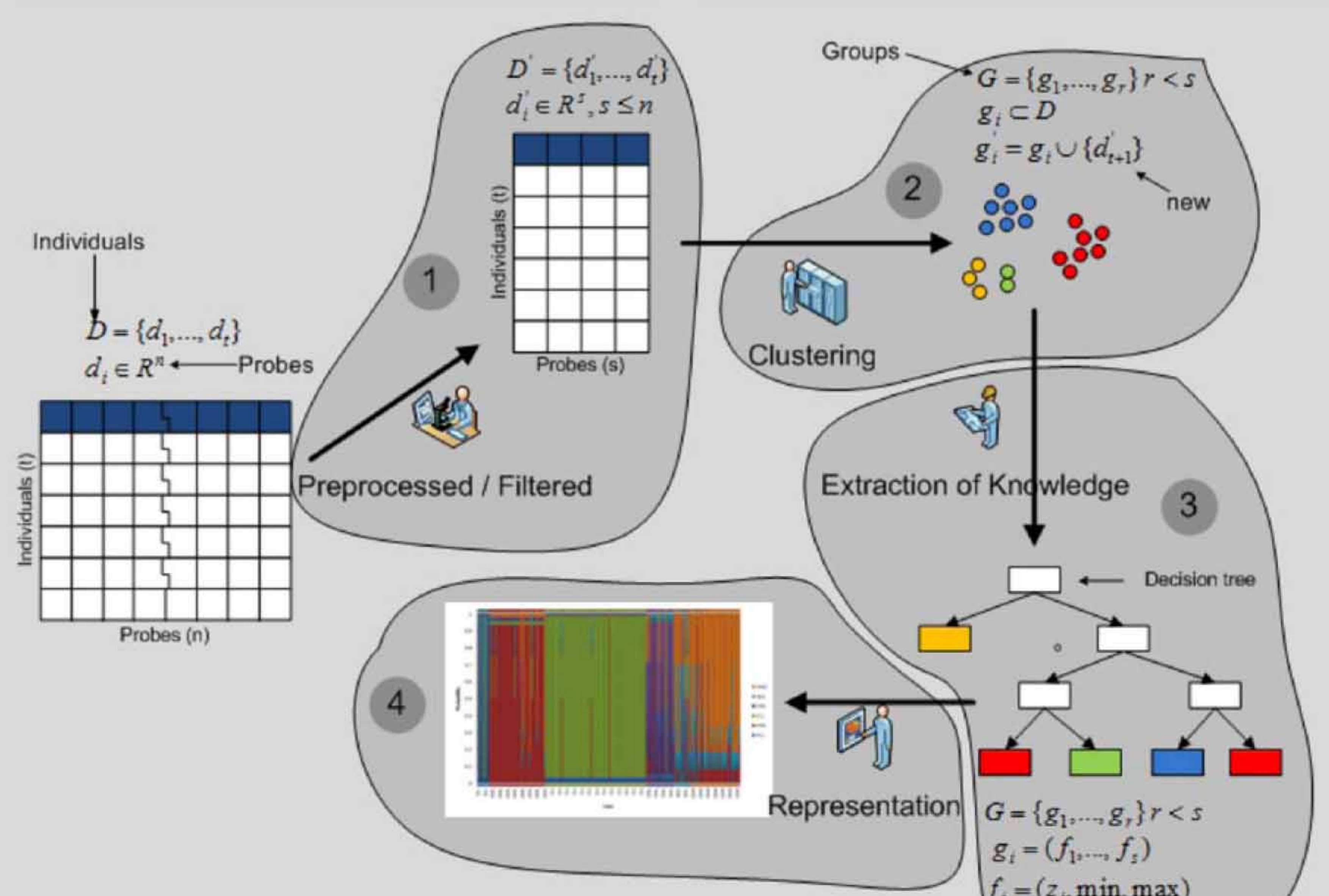
Kapur, K., Xing, Y., Ouyang, Z. & Wong, W. (2007) Exon arrays provide accurate assessments of gene expression. *Genome Biology, 8, 5, R82.*

Slide 8-23 Computational leukemia cancer detection 1/6



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.*



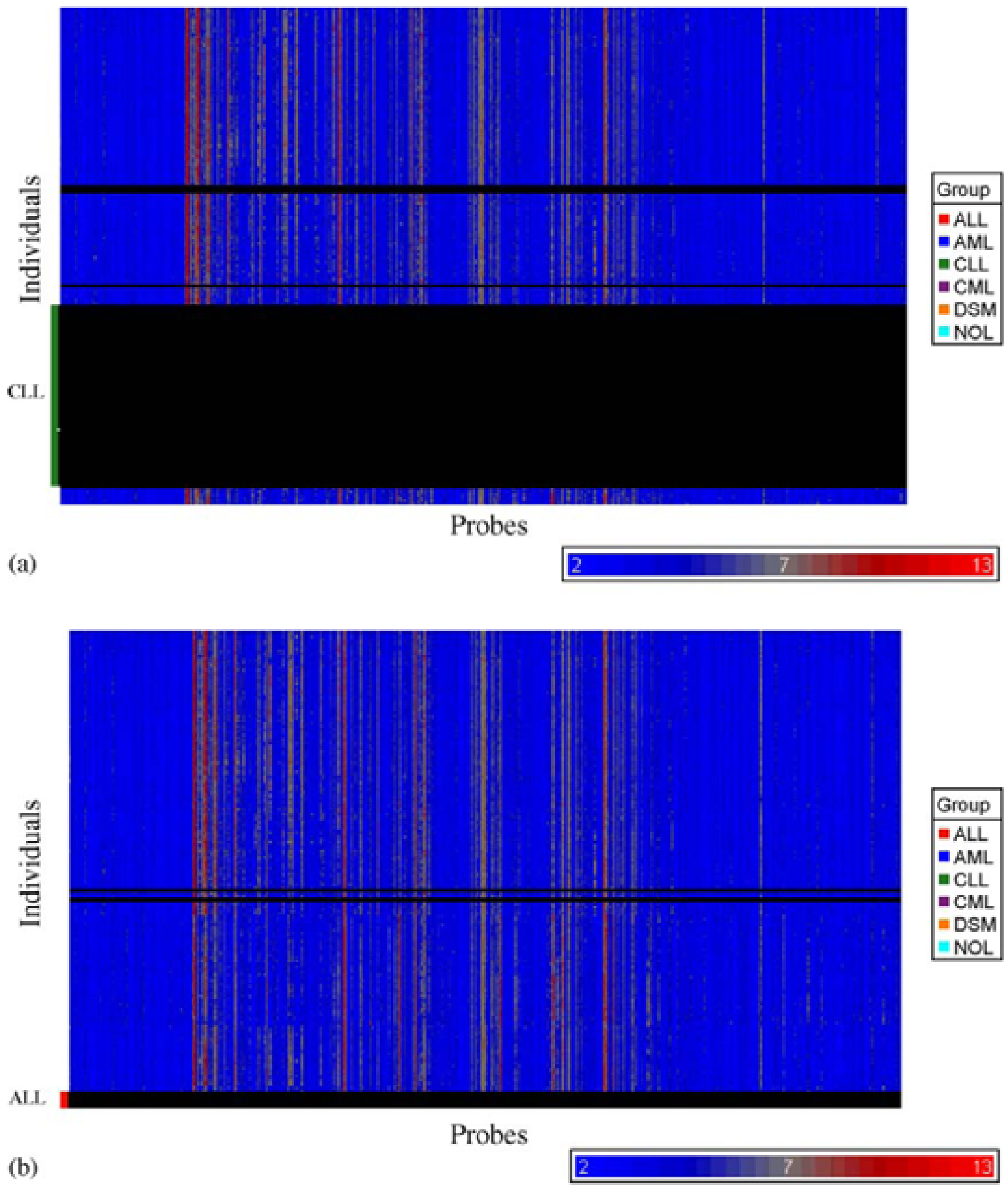
Corchado et al. (2009)

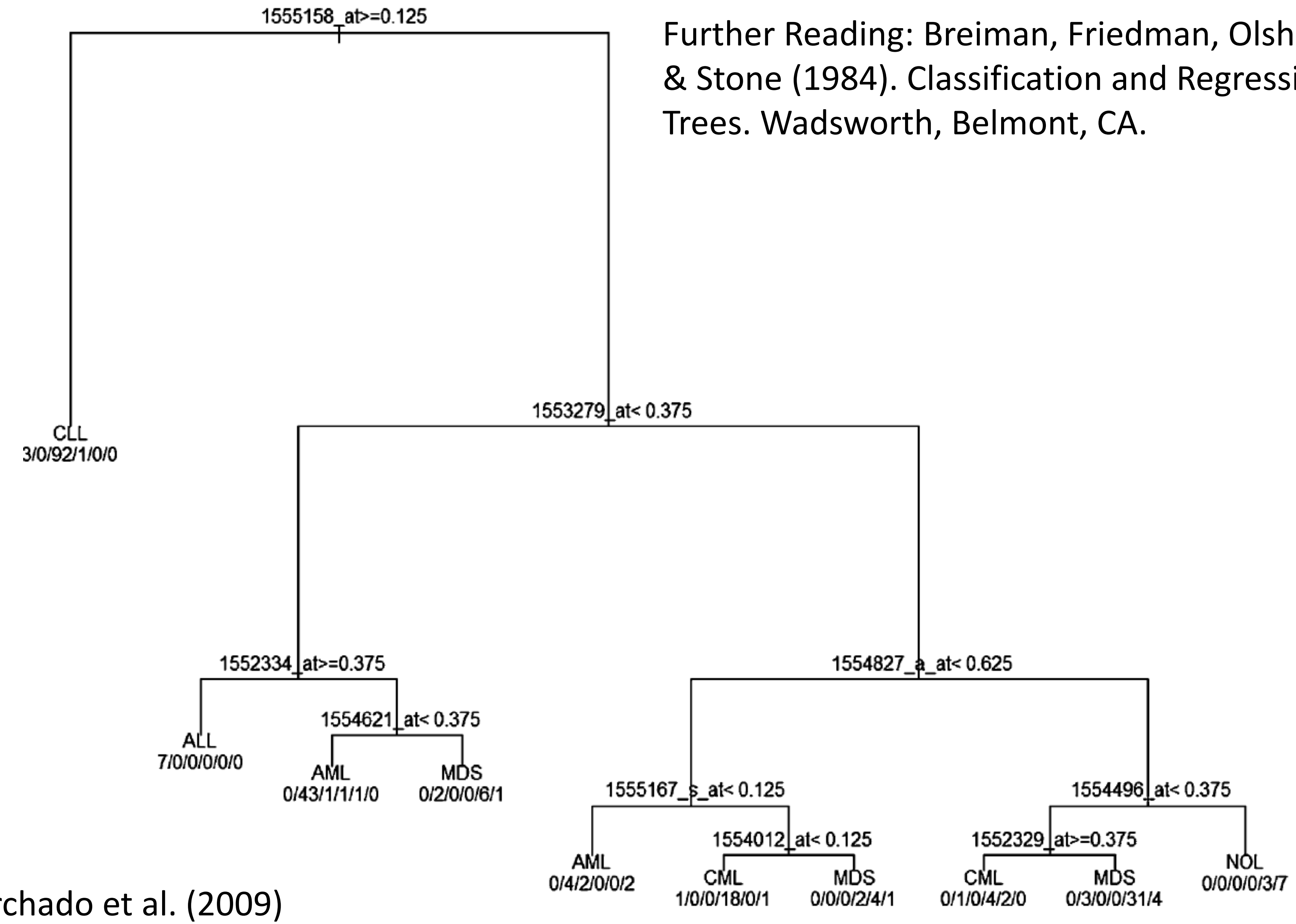
Slide 8-25 Computational leukemia cancer detection 3/6

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)

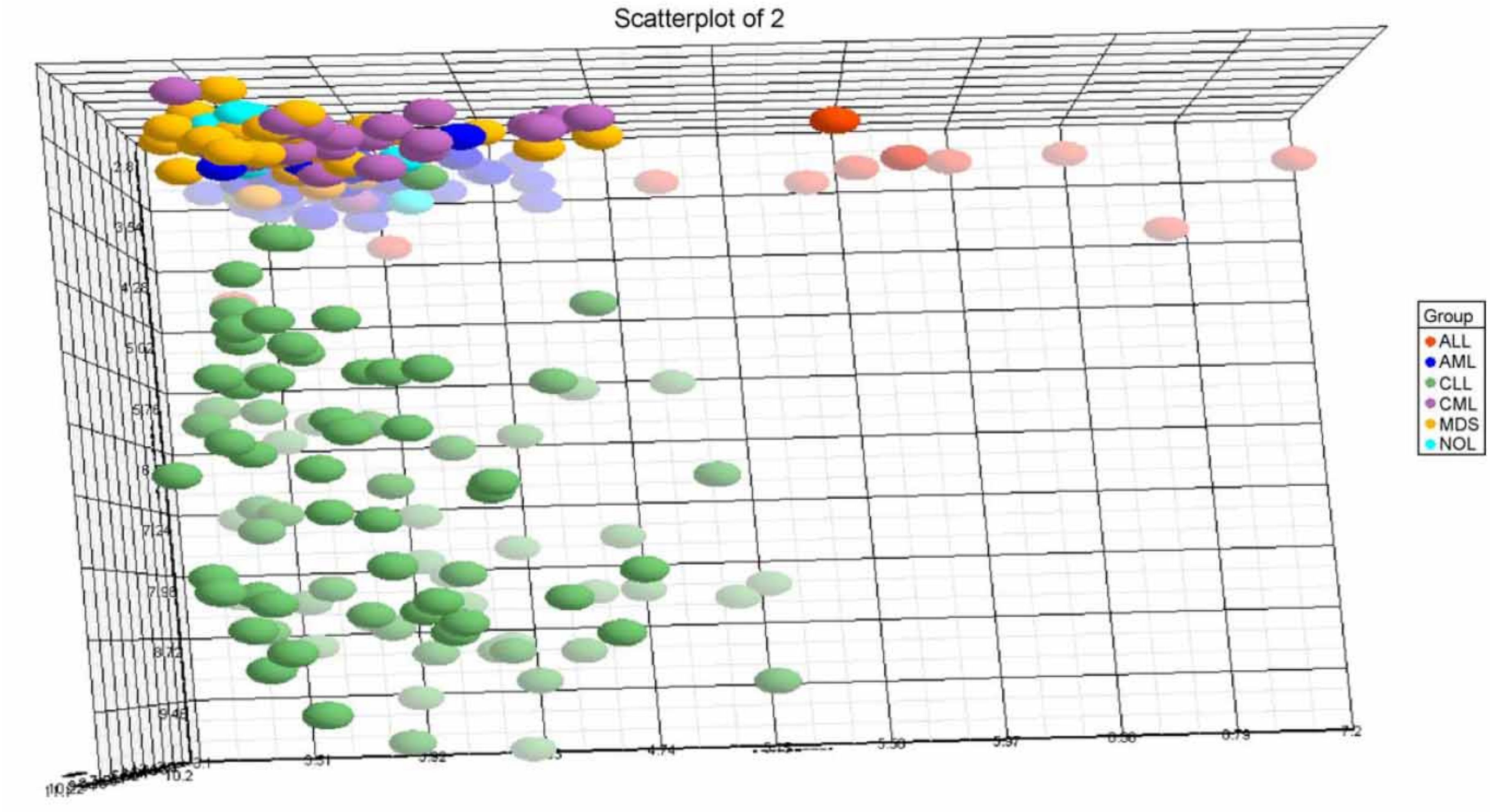




Further Reading: Breiman, Friedman, Olshen, & Stone (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.

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

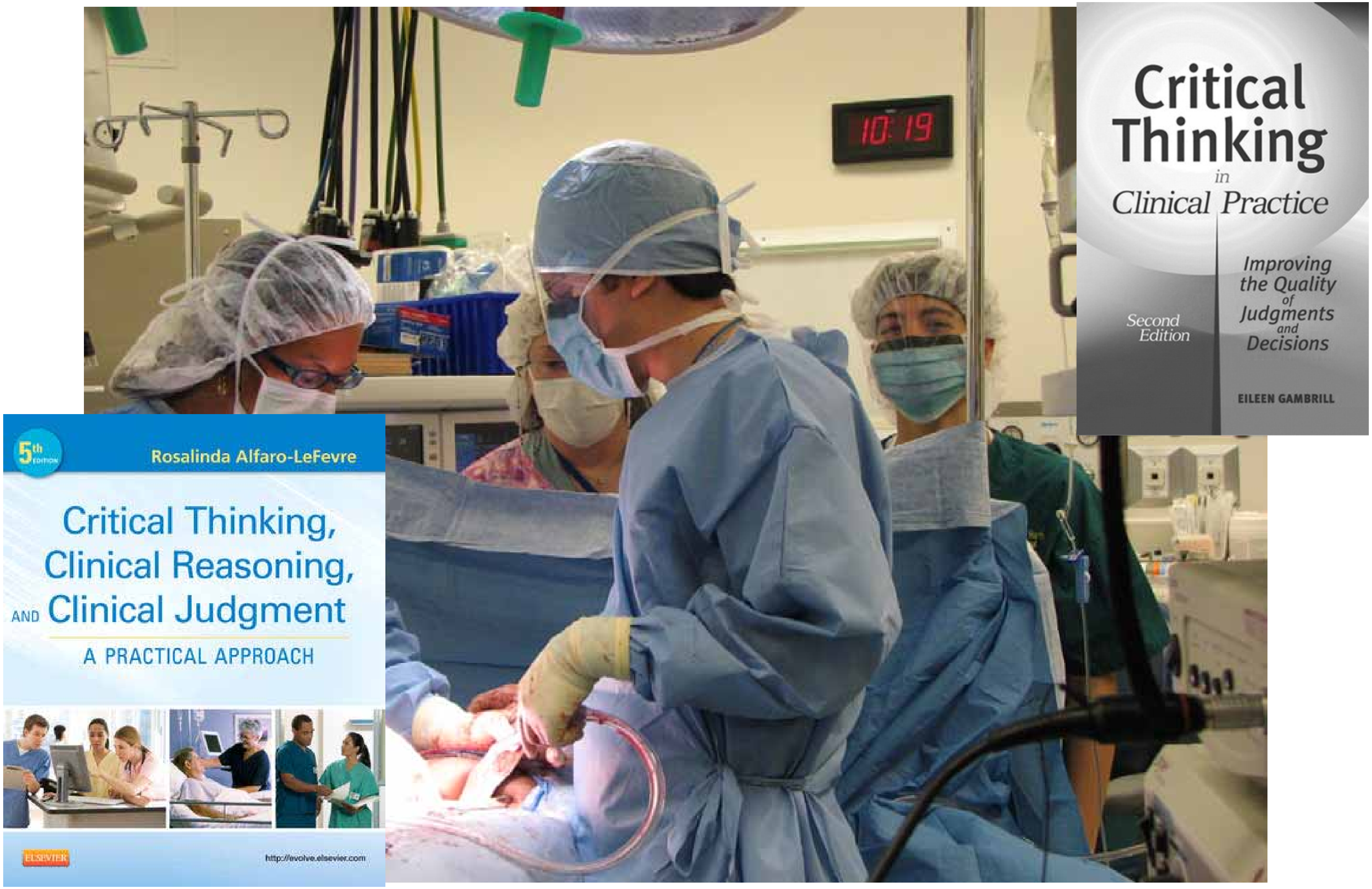


Corchado et al. (2009)

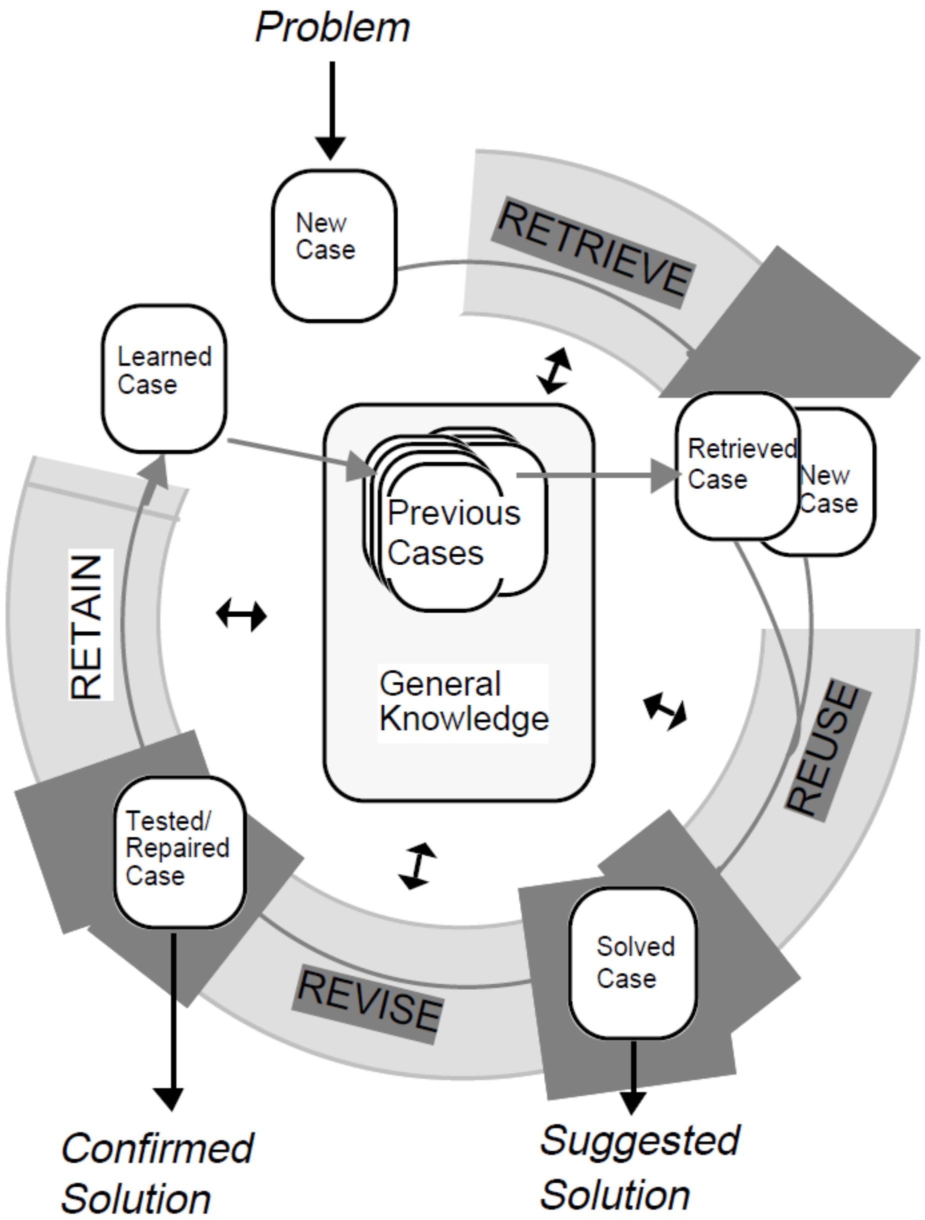
- The model of Corchado et al. (2009) combines:
- 1) methods to **reduce the dimensionality** of the original data set;
- 2) pre-processing and data filtering techniques;
- 3) a clustering method to classify patients; and
- 4) extraction of knowledge techniques
- The system reflects how human experts 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

04 Example: Case Based Reasoning (CBR)

Slide 8-29 Thinking – Reasoning – Deciding – Acting

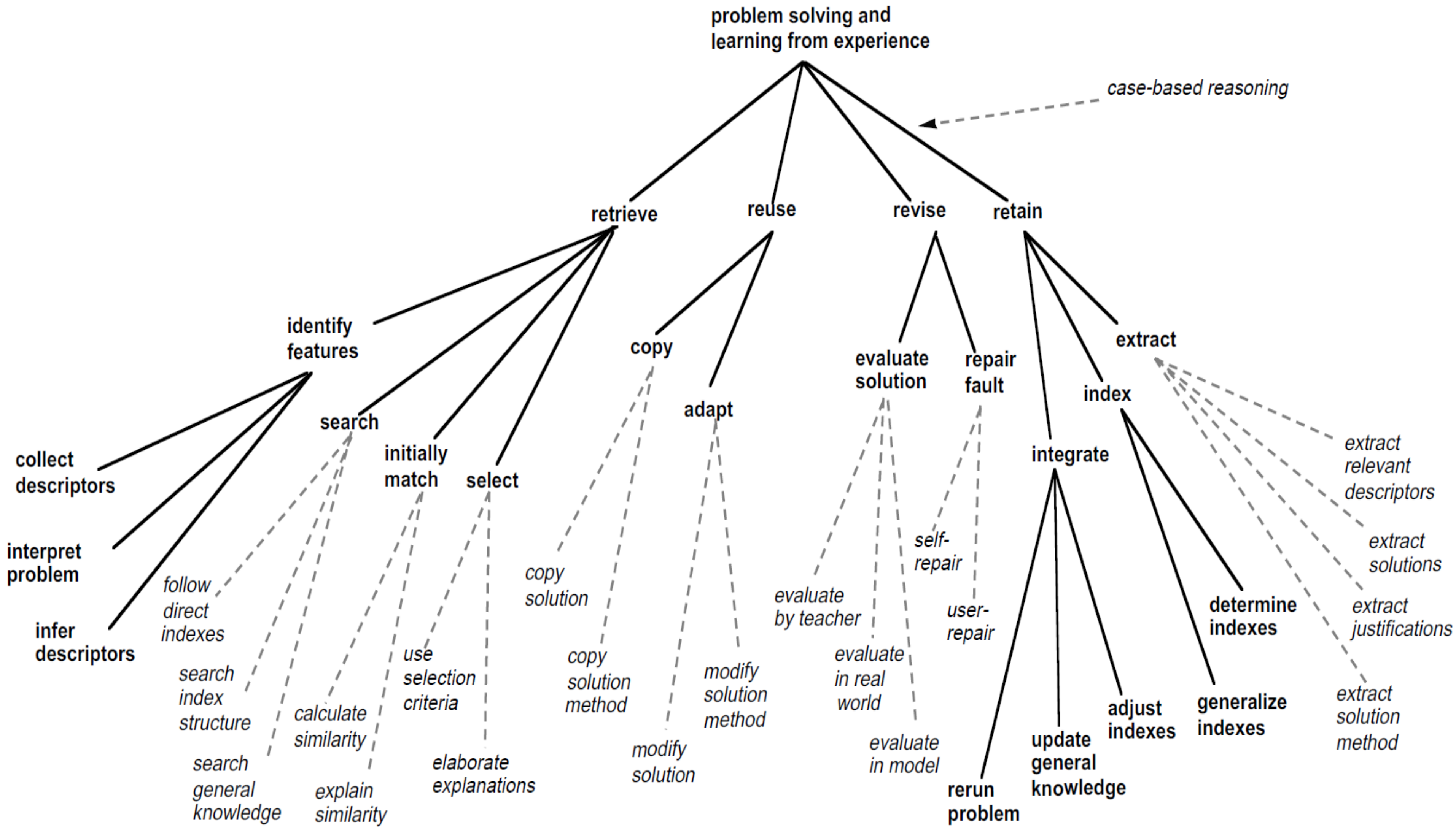


Slide 8-30 Case Based Reasoning (CBR) Basic principle



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

Slide 8-31 The task-method decomposition of CBR

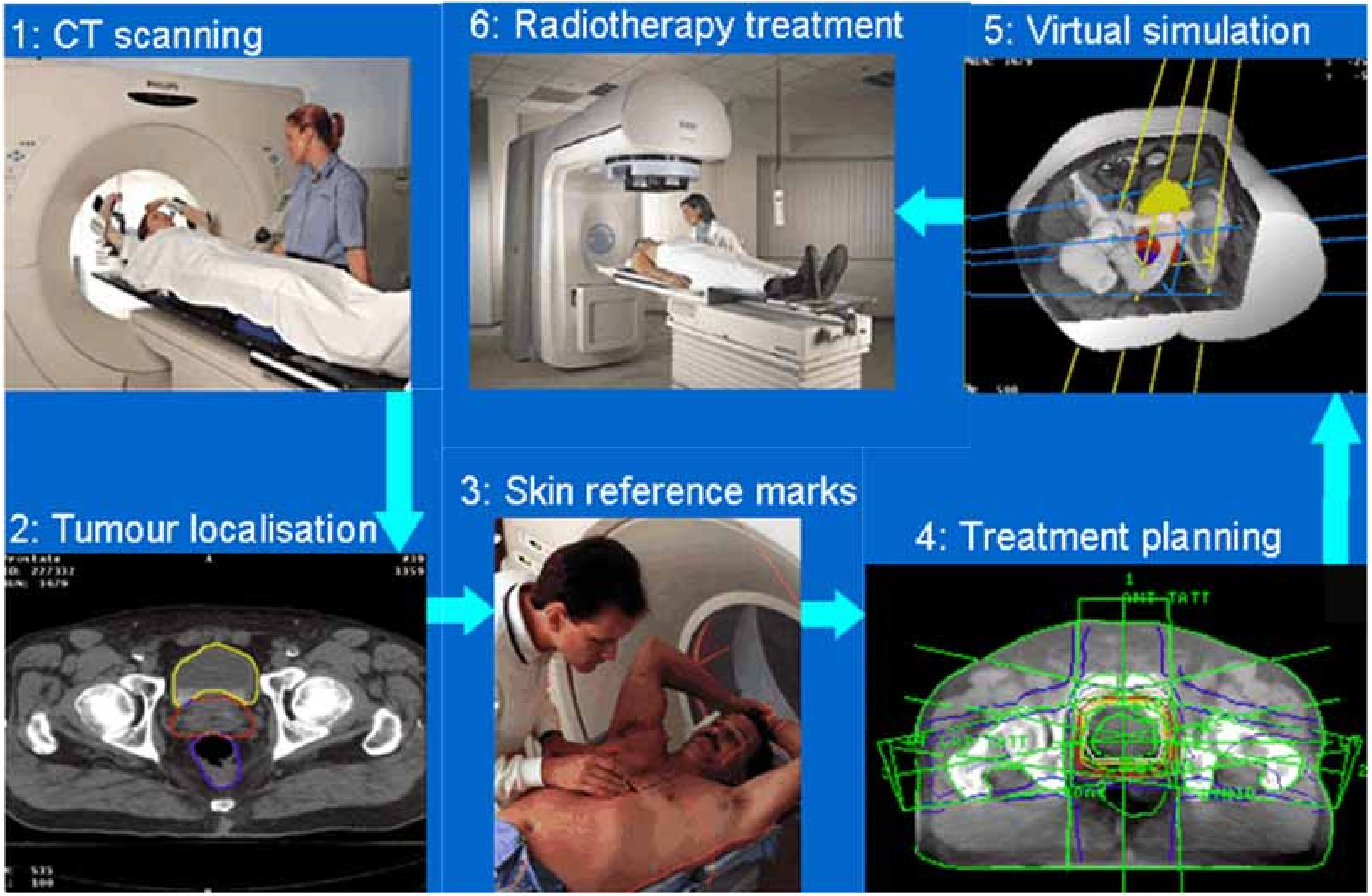


Aamodt & Plaza (1994)

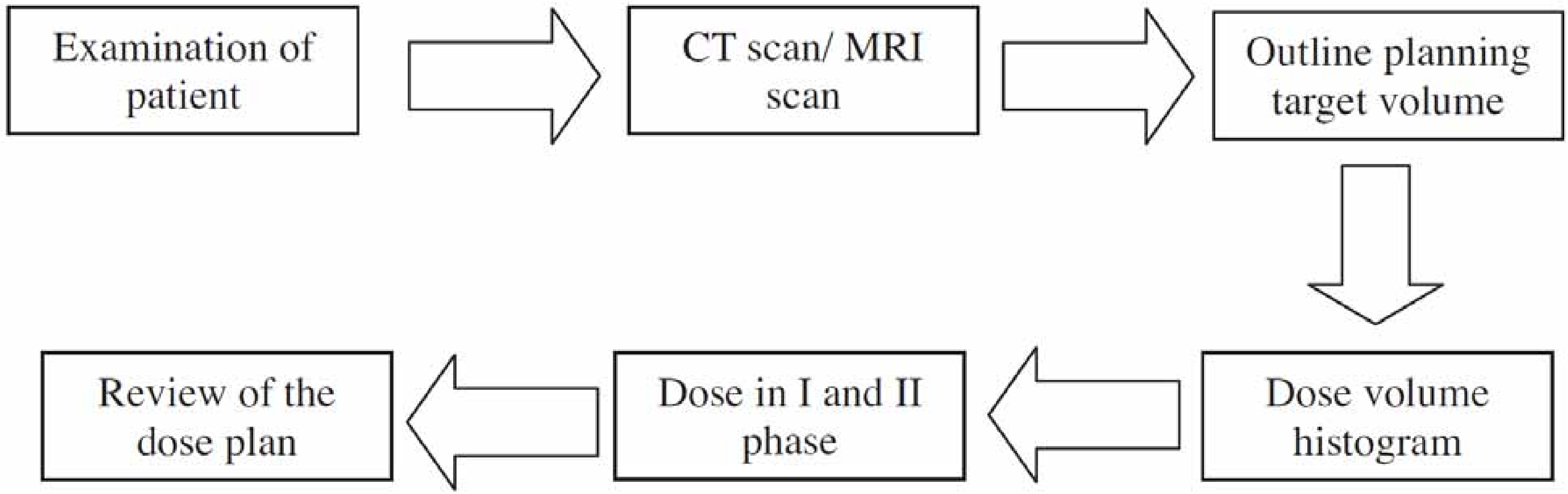
Slide 8-32 CBR Example: Radiotherapy Planning 1/6

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

Slide 8-33 CBR Example: Radiotherapy Planning 2/6



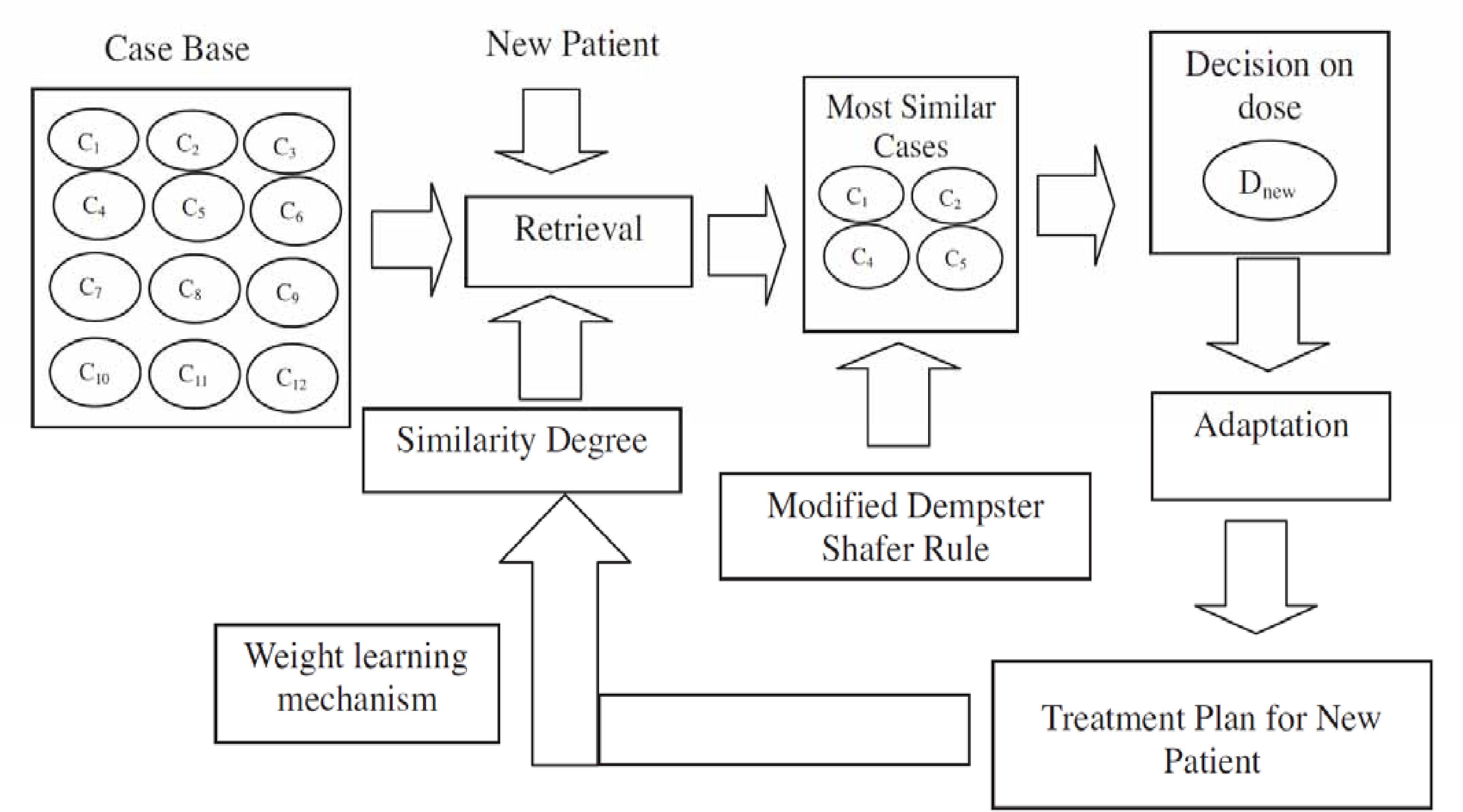
Source: Imaging Performance Assessment of CT Scanners Group, http://www.impactscan.org



Measures:

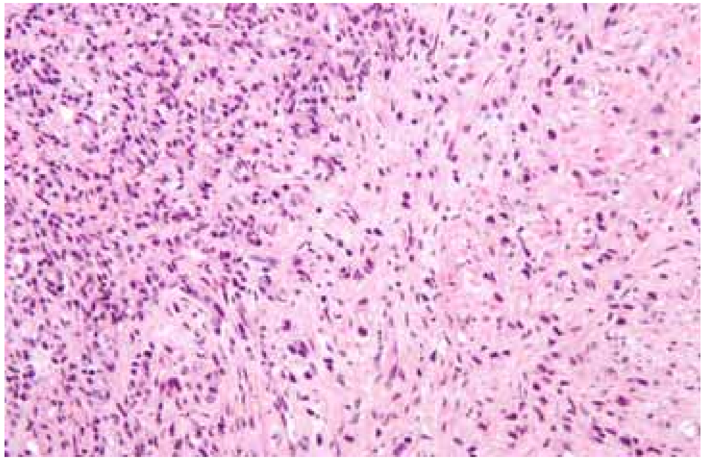
1) Clinical Stage = a labelling system
2) Gleason Score = grade of prostate cancer = integer between 1 to 10; and
3) Prostate Specific Antigen (PSA) value between 1 to 40
4) Dose Volume Histogram (DVH) = pot. risk to the rectum (66, 50, 25, 10 %)

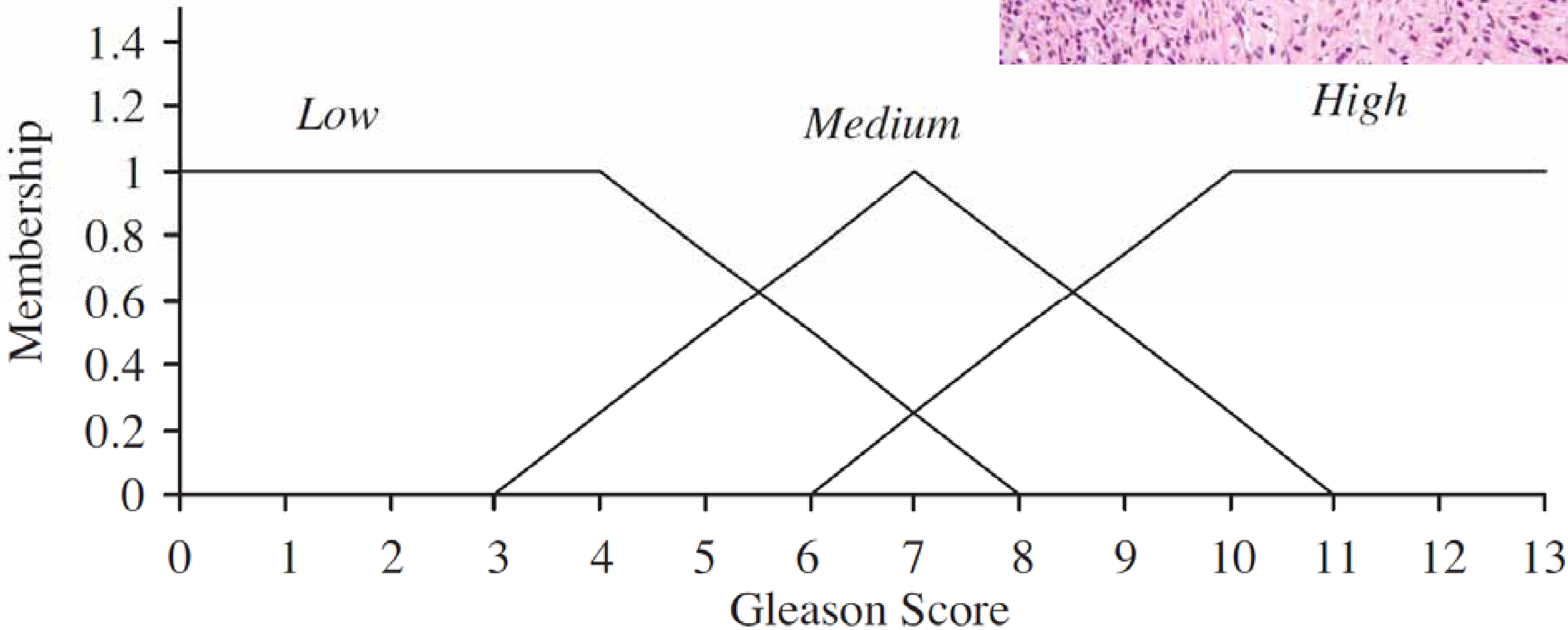
Petrovic, S., Mishra, N. & Sundar, S. (2011) A novel case based reasoning approach to radiotherapy planning. *Expert Systems With Applications, 38, 9, 10759-10769.*



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

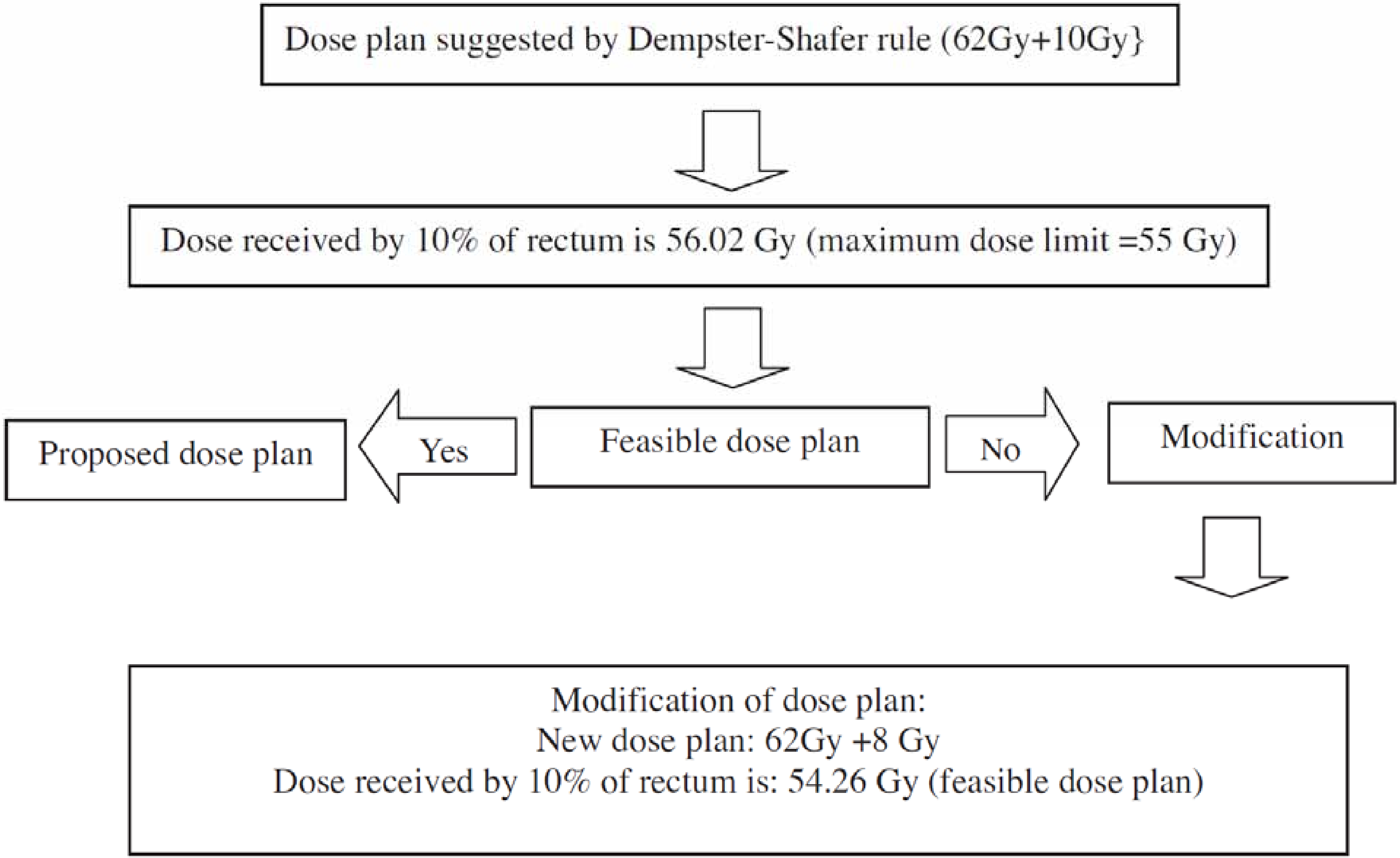




Petrovic, S., Mishra, N. & Sundar, S. (2011) A novel case based reasoning approach to radiotherapy planning. *Expert Systems With Applications, 38, 9, 10759-10769.*

Slide 8-37 Case Based Reasoning 6/6

Petrovic et al. (2011)



05 Causal Reasoning

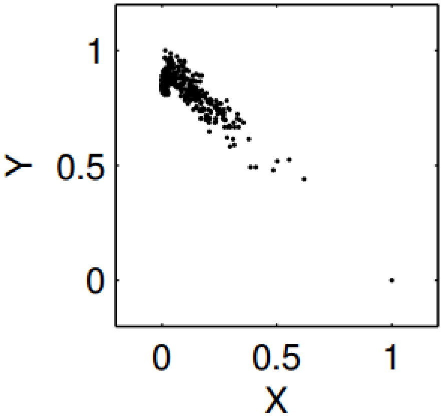
15b

- **“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.

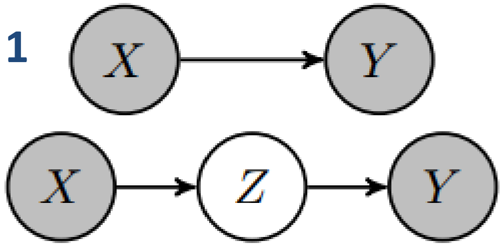
Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285, doi:10.1126/science.1192788.

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



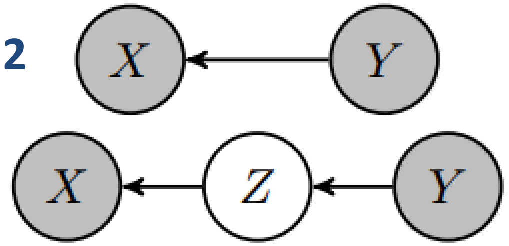
Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler & Bernhard Schölkopf 2016. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, 17, (1), 1103-1204.

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

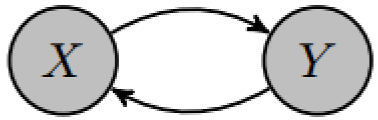


$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

3

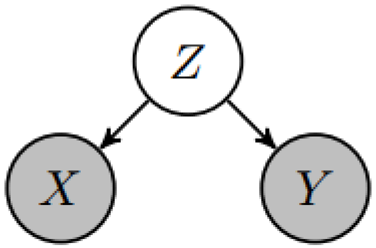
$$\mathbb{P}_Y = \mathbb{P}_{Y\,|\,\mathrm{do}(x)} = \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X = \mathbb{P}_{X\,|\,\mathrm{do}(y)} = \mathbb{P}_{X\,|\,y}$$

4



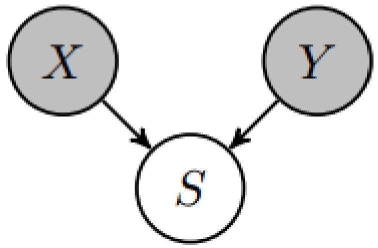
$$\mathbb{P}_Y \neq \mathbb{P}_{Y\,|\,\mathrm{do}(x)} \neq \mathbb{P}_{Y\,|\,x}$$
$$\mathbb{P}_X \neq \mathbb{P}_{X\,|\,\mathrm{do}(y)} \neq \mathbb{P}_{X\,|\,y}$$

5



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

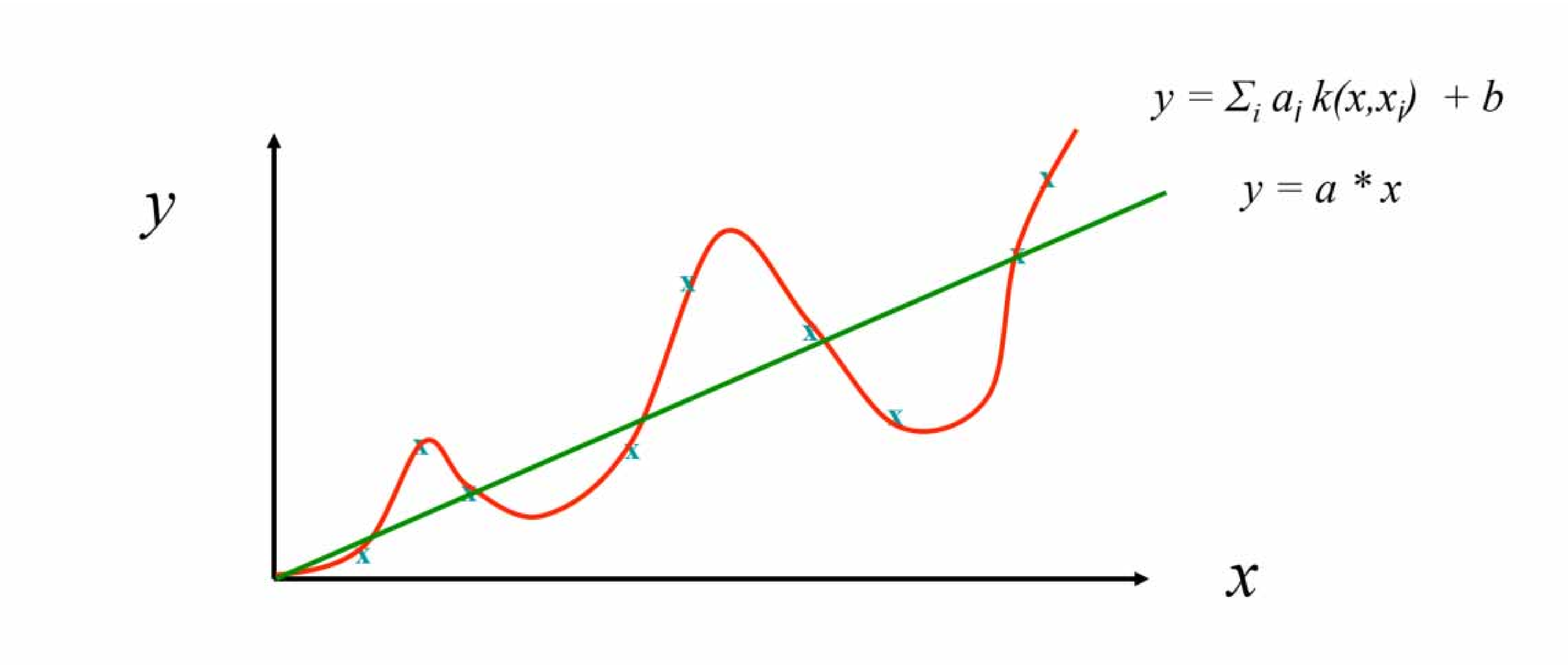
6



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

- **Deductive Reasoning** = Hypothesis > Observations > Logical Conclusions
  - DANGER: Hypothesis must be correct! DR defines whether the truth of a conclusion can be determined for that rule, based on the truth of premises: A=B, B=C, therefore A=C
- **Inductive reasoning** = makes broad generalizations from specific observations
  - DANGER: allows a conclusion to be false if the premises are true
  - generate hypotheses and use DR for answering specific questions
- **Abductive reasoning** = inference = to get the best explanation from an incomplete set of preconditions.
  - Given a true conclusion and a rule, it attempts to select some possible premises that, if true also, may support the conclusion, though not uniquely.
  - Example: "When it rains, the grass gets wet. The grass is wet. Therefore, it might have rained." This kind of reasoning can be used to develop a hypothesis, which in turn can be tested by additional reasoning or data.

- := information provided by direct observation (empirical evidence) in contrast to information provided by inference
  - Empirical evidence = information acquired by observation or by experimentation in order to verify the truth (fit to reality) or falsify (non-fit to reality).
  - Empirical inference = drawing conclusions from empirical data (observations, measurements)
  - Causal inference = drawing a conclusion about a causal connection based on the conditions of the occurrence of an effect.
    - Causal inference is an example of causal reasoning.



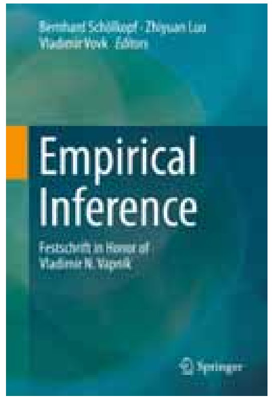
Gottfried W. Leibniz (1646-1716)

Hermann Weyl (1885-1955)

Vladimir Vapnik (1936-)

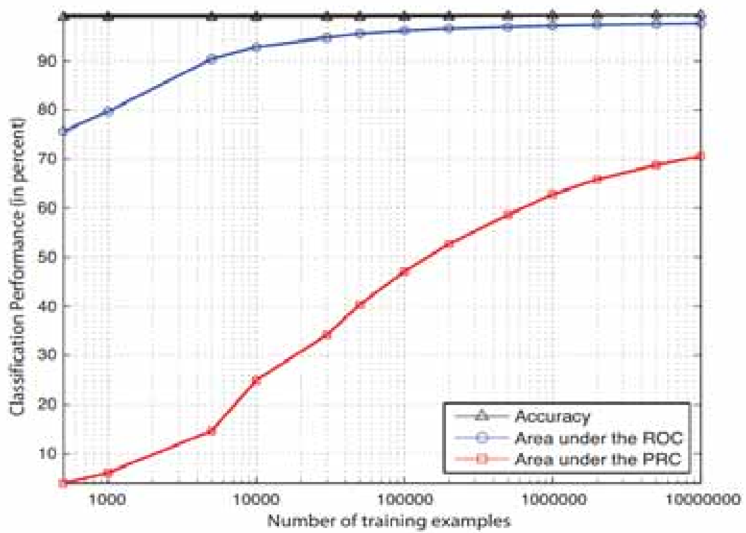
Alexey Chervonenkis (1938-2014)

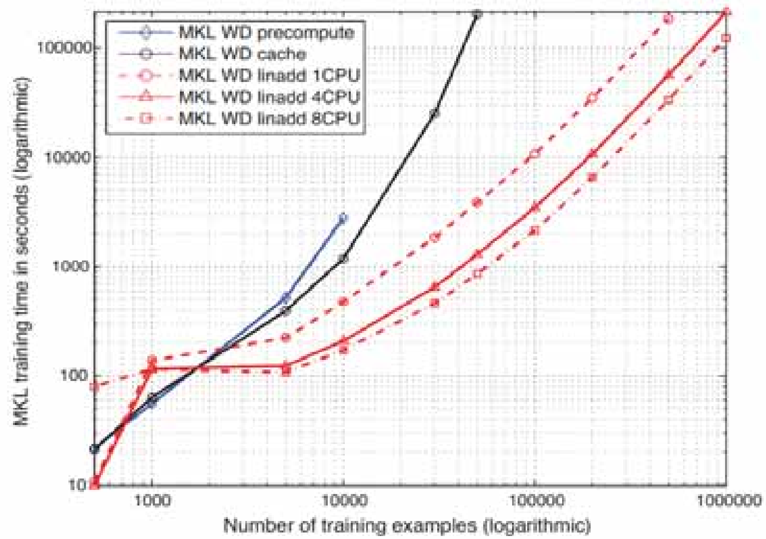
Gregory Chaitin (1947-)



Remember: hard inference problems

- High dimensionality (curse of dim., many factors contribute)
- Complexity (real-world is non-linear, non-stationary, non-IID *)
- Need of large top-quality data sets
- Little prior data (no mechanistic models of the data)
  - *) = Def.: a sequence or collection of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent





Sören Sonnenburg, Gunnar Rätsch, Christin Schaefer & Bernhard Schölkopf 2006. Large scale multiple kernel learning. Journal of Machine Learning Research, 7, (7), 1531-1565.

06 Explainability

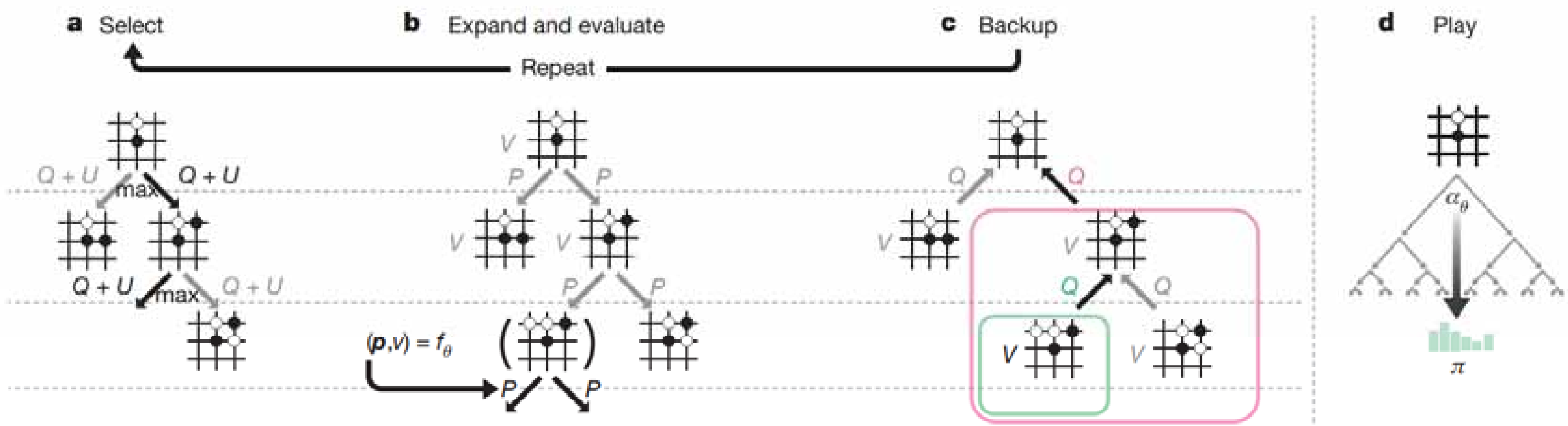
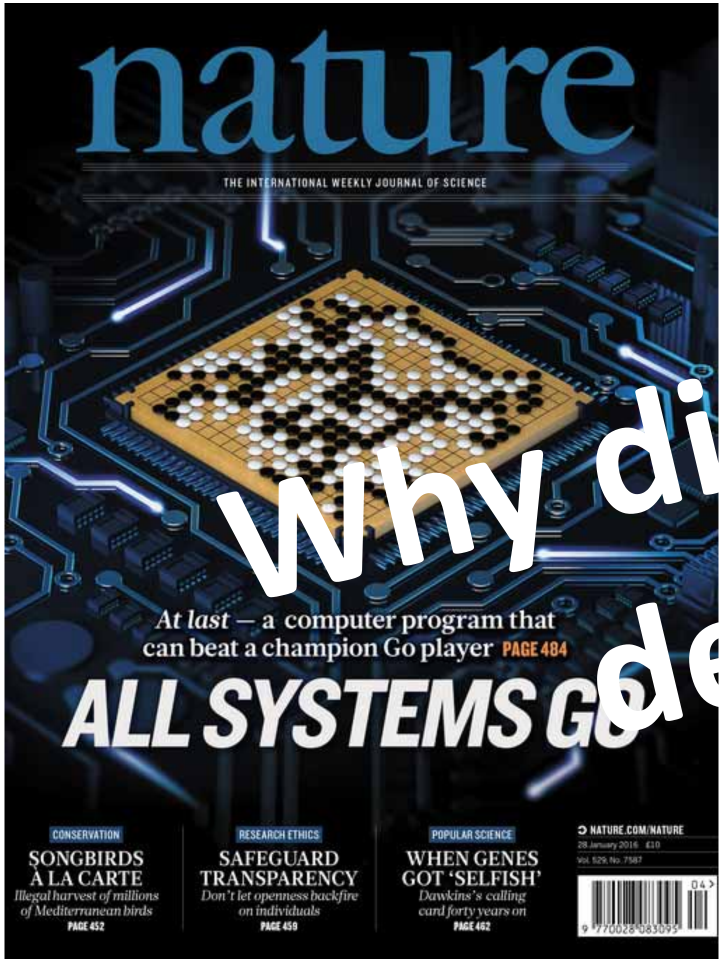


Figure 2 | MCTS in AlphaGo Zero. **a**, Each simulation traverses the tree by selecting the edge with maximum action value Q, plus an upper confidence bound U that depends on a stored prior probability P and visit count N for that edge (which is incremented once traversed). **b**, The leaf node is expanded 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 $\boldsymbol{\pi}$ 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.



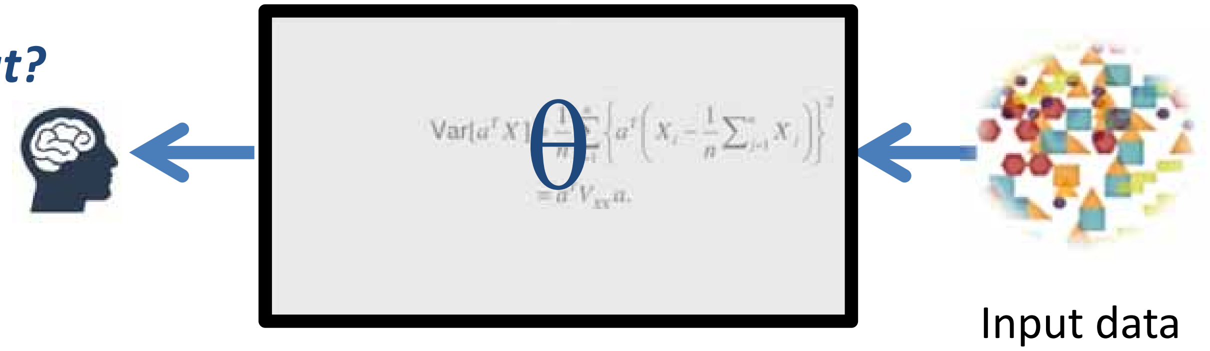
Why did it make this decision???

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.

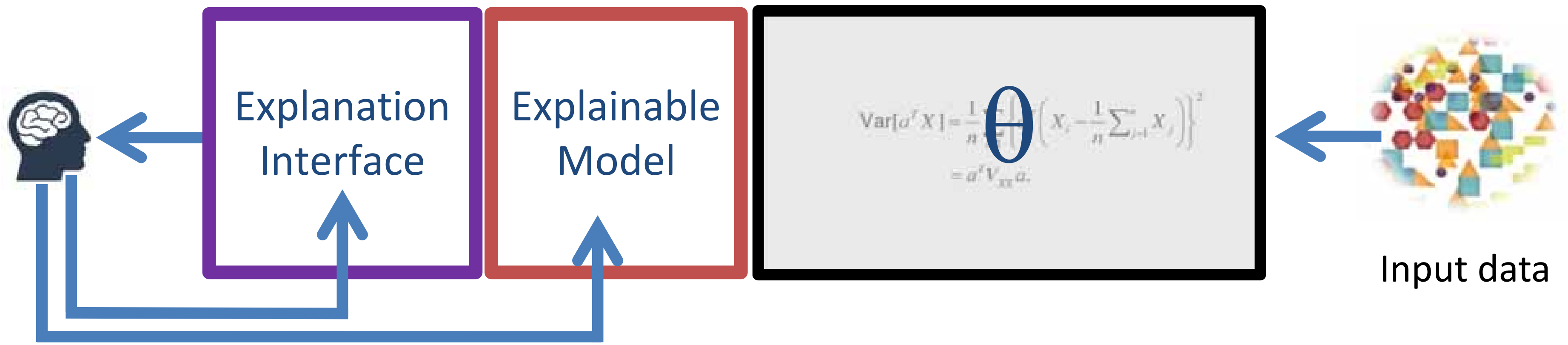
Why did the algorithm do that?

Can I trust these results?

How can I correct an error?



A possible solution



The domain expert can understand why …

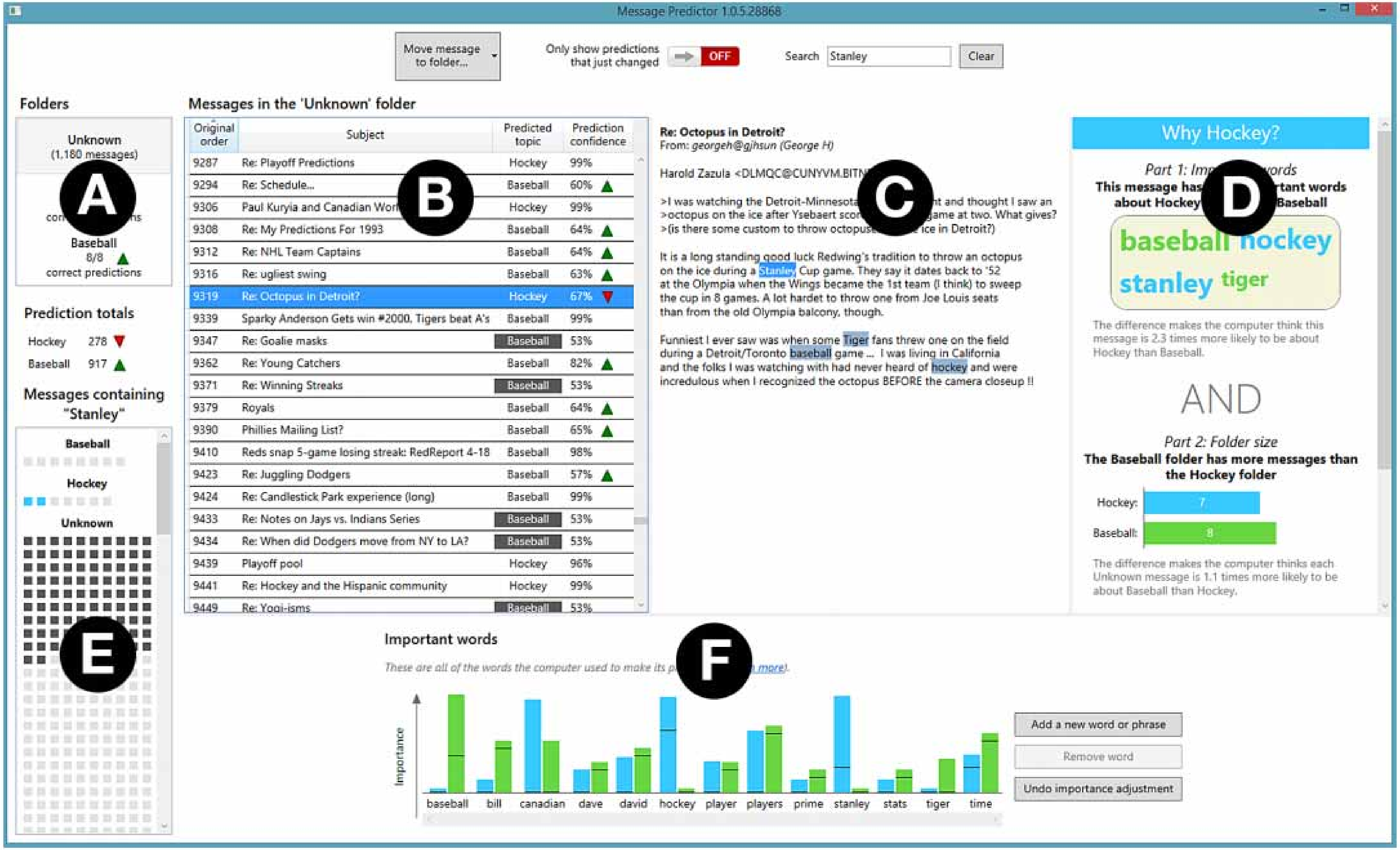
The domain expert can learn and correct errors …

The domain expert can re-enact on demand …

What is interpretable for humans?

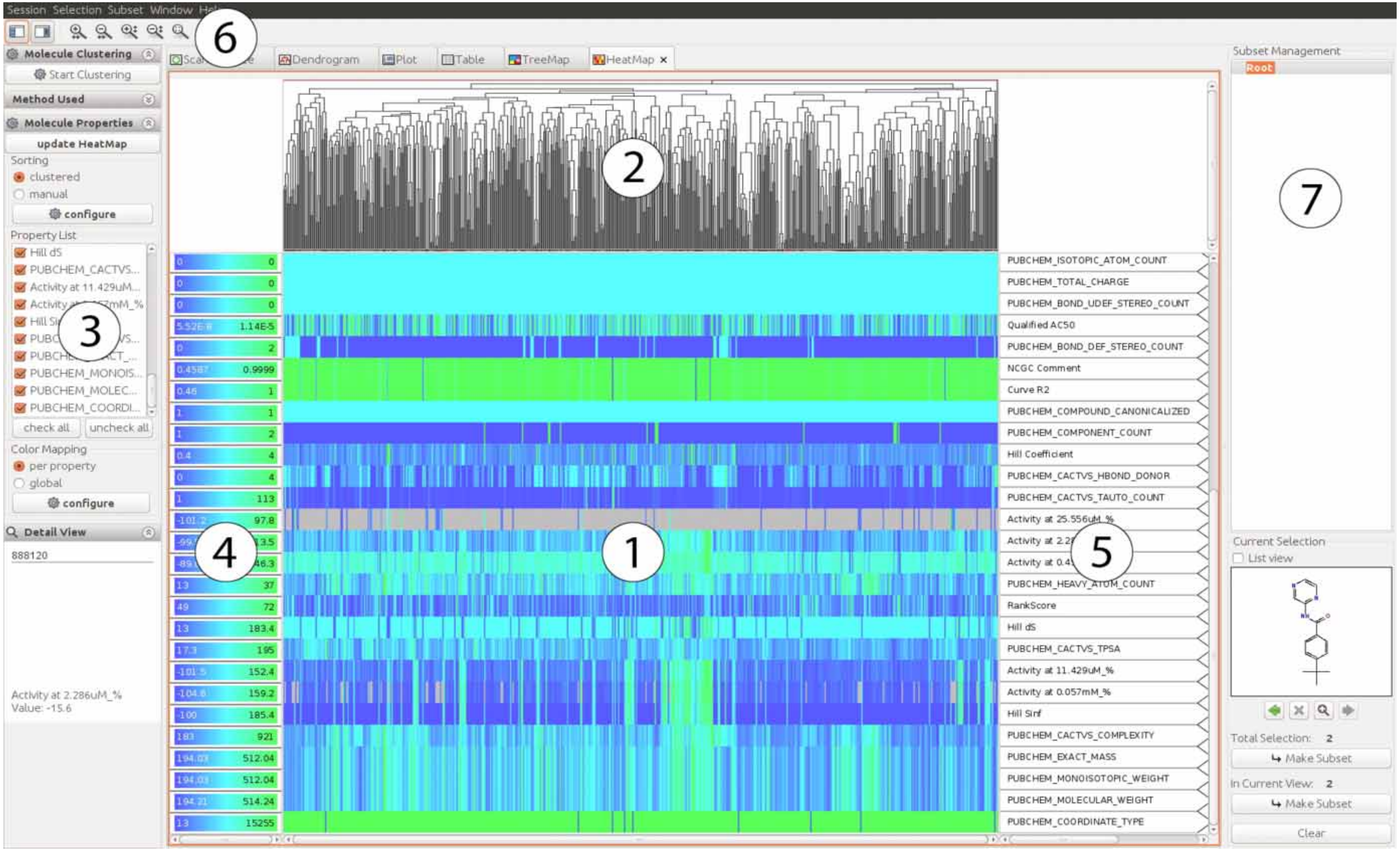
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.953	0.894	0.620	0.699	0.629	0.546	0.540	1.000	0.526	1.000	0.522	0.483	0.471	1.000	0.522	0.576	0.658
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.722	0.638	1.000	0.785	0.743	0.792	0.801	0.875	0.712	1.000	0.444	0.947	0.431	1.000	0.793	1.000	0.635
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.658	0.633	0.569	0.561	0.589	0.640	0.659	0.845	0.932	0.512	0.575	0.941	1.000	0.991	1.000	0.892
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.932	0.639	0.575	0.544	0.501	0.489	0.470	0.454	0.576	0.576	0.581	0.707	0.992	1.000	1.000	1.000	1.000	1.000
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.711	0.644	0.569	0.541	0.461	0.430	0.425	0.381	0.364	0.437	0.562	0.509	0.528	0.678	1.000	0.991	1.000	1.000	1.000
1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.680	0.594	0.579	0.513	0.490	0.429	0.405	0.425	0.381	0.401	0.387	0.367	0.484	0.428	0.483	0.659	0.936	1.000	1.000	1.000
1.000	1.000	1.000	1.000	1.000	0.761	0.677	0.610	0.565	0.511	0.498	0.457	0.416	0.396	0.388	0.369	0.355	0.359	0.468	0.392	0.380	[illegible]	[illegible]	[illegible]	0.744	1.000	0.485
1.000	1.000	1.000	1.000	0.861	0.640	0.579	0.560	0.542	0.476	0.470	0.441	0.405	0.389	0.392	0.396	0.436	0.355	0.327	0.394	0.407	[illegible]	[illegible]	[illegible]	0.766	0.676	0.437
1.000	1.000	1.000	0.827	0.646	0.579	0.556	0.545	0.489	0.505	0.489	0.478	0.411	0.387	0.404	0.401	0.391	0.452	0.352	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.510	0.578	0.538
0.909	1.000	0.860	0.675	0.598	0.528	0.535	0.500	0.497	0.517	0.468	0.520	0.623	0.619	0.507	0.472	0.385	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.354	0.318	0.462	0.491	0.426
1.000	0.989	0.693	0.561	0.546	0.523	0.532	0.452	0.441	0.461	0.649	0.659	0.695	0.686	0.632	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.358	0.295	0.310	0.336	0.363	0.418	0.458
0.969	0.849	0.606	0.530	0.521	0.494	0.437	0.396	0.421	0.626	0.698	0.741	0.737	[illegible]	[illegible]	0.729	[illegible]	[illegible]	0.565	0.506	0.435	0.358	0.311	0.299	0.313	0.402	0.488
1.000	1.000	0.590	0.509	0.486	0.445	0.411	0.372	0.569	0.675	0.732	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.650	0.622	0.573	0.467	0.405	0.286	0.274	0.358	0.419	0.445
1.000	0.924	0.554	0.517	0.450	0.416	0.449	[illegible]	0.585	[illegible]	0.727	[illegible]	[illegible]	[illegible]	0.785	0.740	0.700	0.653	0.626	[illegible]	0.502	0.431	0.338	0.279	0.295	0.330	0.446
1.000	1.000	0.557	0.517	0.457	0.396	[illegible]	[illegible]	0.635	0.658	0.707	[illegible]	0.751	0.757	0.792	0.764	0.714	0.694	0.642	0.597	0.542	0.419	0.341	0.289	0.291	0.326	0.380
1.000	1.000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.623	[illegible]	0.670	0.711	0.748	0.771	0.775	0.772	0.724	[illegible]	[illegible]	0.434	0.378	0.354	0.414	0.307	0.282	0.278	0.402
0.763	1.000	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.484	0.590	0.646	0.687	0.718	0.724	0.748	0.717	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.483	0.499	0.472	0.273	0.234	0.279	0.306
1.000	1.000	0.750	[illegible]	[illegible]	0.344	0.328	0.490	0.550	0.623	0.593	[illegible]	0.521	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.421	0.519	0.500	0.566	0.521	0.286	0.249	0.234	0.258
0.754	0.830	1.000	0.471	0.435	0.326	0.327	0.489	0.474	0.421	0.388	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.601	0.594	0.627	0.590	0.613	0.585	0.529	0.438	0.328	0.487	0.200
0.929	0.672	0.503	0.654	0.388	0.335	0.306	0.475	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.559	0.616	0.550	0.649	0.686	0.658	0.667	0.587	0.564	0.486	0.416	0.546	0.263
1.000	0.758	0.639	0.726	0.931	0.330	0.299	0.398	[illegible]	[illegible]	[illegible]	[illegible]	0.646	0.644	0.517	0.605	0.517	0.546	0.616	0.714	0.683	0.609	0.578	0.563	0.478	0.314	0.252
1.000	0.790	0.907	0.701	0.897	0.382	0.296	0.358	[illegible]	[illegible]	0.674	0.683	0.666	0.605	0.526	0.620	0.527	0.514	0.616	0.666	0.670	0.628	0.549	0.512	0.262	0.321	0.254
0.760	0.587	0.639	0.557	0.681	0.593	0.397	0.340	0.575	0.574	0.647	0.691	0.666	0.620	0.506	0.614	0.550	0.532	0.487	0.589	0.610	0.616	0.504	0.482	0.310	0.271	0.237
0.577	0.599	0.443	0.561	0.657	0.363	0.914	0.626	0.482	0.553	0.631	0.678	0.722	0.561	0.523	0.639	0.634	0.510	0.481	0.558	0.533	0.597	0.570	0.509	0.342	0.263	0.243
0.639	0.615	0.748	0.639	0.911	0.796	0.647	0.614	0.529	0.553	0.588	0.651	0.644	0.585	0.433	0.606	0.588	0.467	0.313	0.363	0.349	0.415	0.578	0.512	0.305	0.274	0.256
0.569	0.661	0.486	0.605	0.448	0.494	0.705	0.730	0.579	0.532	0.526	0.623	0.518	0.387	0.310	0.338	0.466	0.378	0.559	0.479	0.444	0.430	0.494	0.465	0.232	0.248	0.237
0.493	0.522	0.508	0.553	0.458	0.457	0.435	0.742	0.636	0.434	0.553	0.578	0.369	0.394	0.502	0.539	0.532	0.555	0.601	0.582	0.548	0.498	0.328	0.237	0.242	0.252	0.273
0.891	0.817	0.441	0.445	0.473	0.452	0.720	0.423	0.700	0.492	0.525	0.509	0.463	0.614	0.466	0.477	0.603	0.615	0.509	0.517	0.563	0.405	0.224	0.258	0.234	0.211	0.228
0.543	0.548	0.598	0.433	0.386	0.627	0.482	0.345	0.835	0.751	0.581	0.502	0.482	0.610	0.531	0.524	0.615	0.625	0.562	0.481	0.566	0.306	0.266	0.407	0.366	0.243	0.252
0.762	0.720	0.506	0.496	0.495	0.698	0.396	0.627	0.555	0.317	0.491	0.294	0.382	0.393	0.572	0.449	0.405	0.407	0.357	0.567	0.518	0.243	0.255	0.465	0.415	0.323	0.248
0.472	0.437	0.618	0.547	0.500	0.439	0.580	0.579	0.474	0.406	0.320	0.302	0.233	0.262	0.387	0.622	0.556	0.499	0.580	0.558	0.378	0.214	0.364	0.502	0.413	0.311	0.269
0.461	0.503	0.513	0.432	0.537	0.537	0.467	0.530	0.387	0.504	0.353	0.362	0.456	0.222	0.241	0.342	0.510	0.622	0.454	0.441	0.285	0.218	0.545	0.502	0.445	0.508	0.623
0.529	0.464	0.455	0.824	0.476	0.411	0.498	0.405	0.408	0.400	0.382	0.387	0.482	0.422	0.210	0.242	0.281	0.309	0.295	0.241	0.213	0.549	0.569	0.522	0.500	0.493	0.529
0.383	0.458	0.482	0.370	0.384	0.361	0.400	0.391	0.320	0.319	0.425	0.377	0.433	0.528	0.497	0.285	0.247	0.198	0.226	0.410	0.570	0.597	0.576	0.588	0.531	0.493	0.546
0.459	0.476	0.391	0.431	0.563	0.321	0.364	0.382	0.365	0.368	0.405	0.287	0.263	0.509	0.606	0.569	0.509	0.554	0.551	0.591	0.622	0.647	0.612	0.648	0.594	0.537	0.546

Example for an Explanation Interface



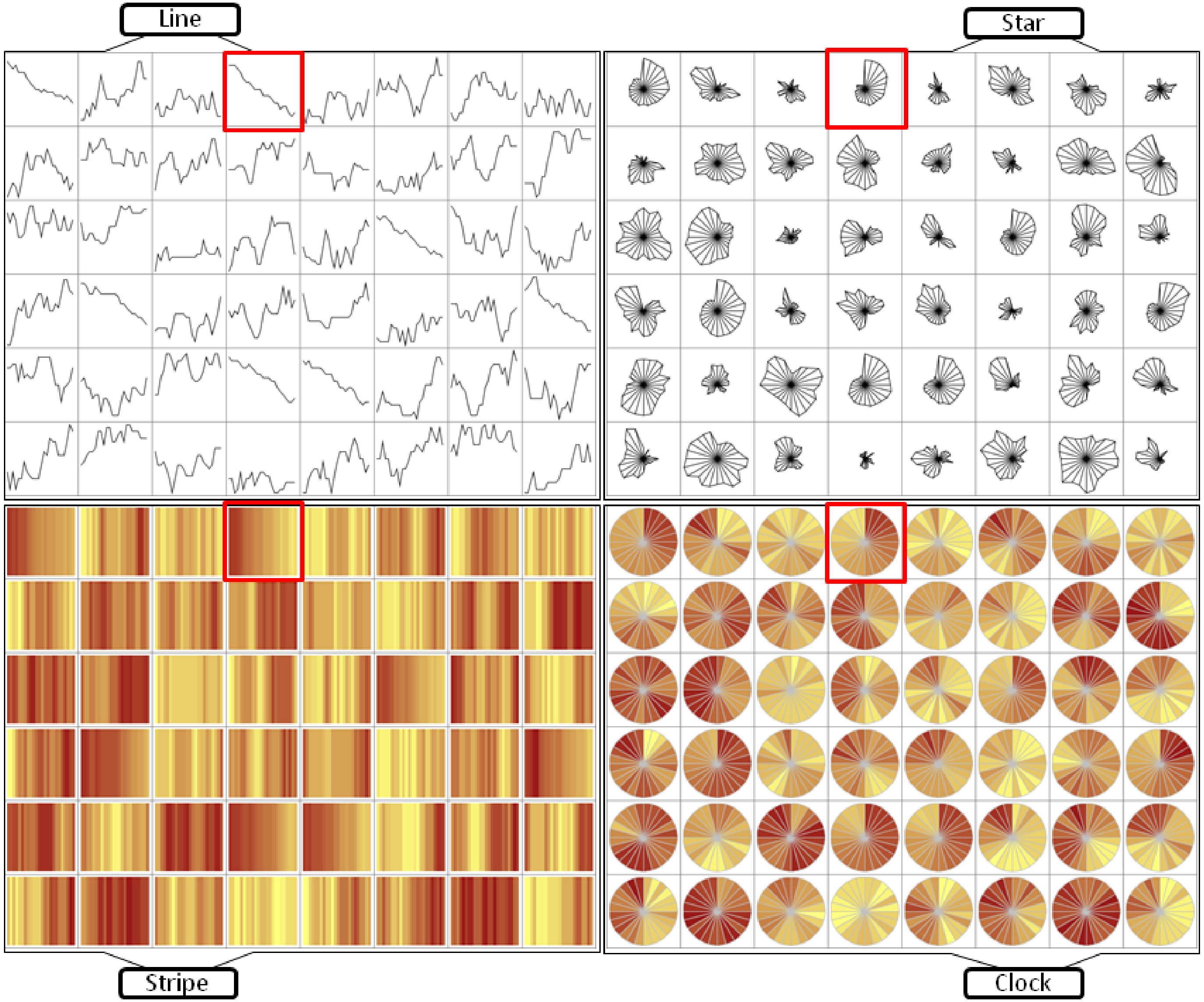
Todd Kulesza, Margaret Burnett, Weng-Keen Wong & Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI 2015), 2015 Atlanta. ACM, 126-137, doi:10.1145/2678025.2701399.

Example for an Explanation Interface - open work ☺



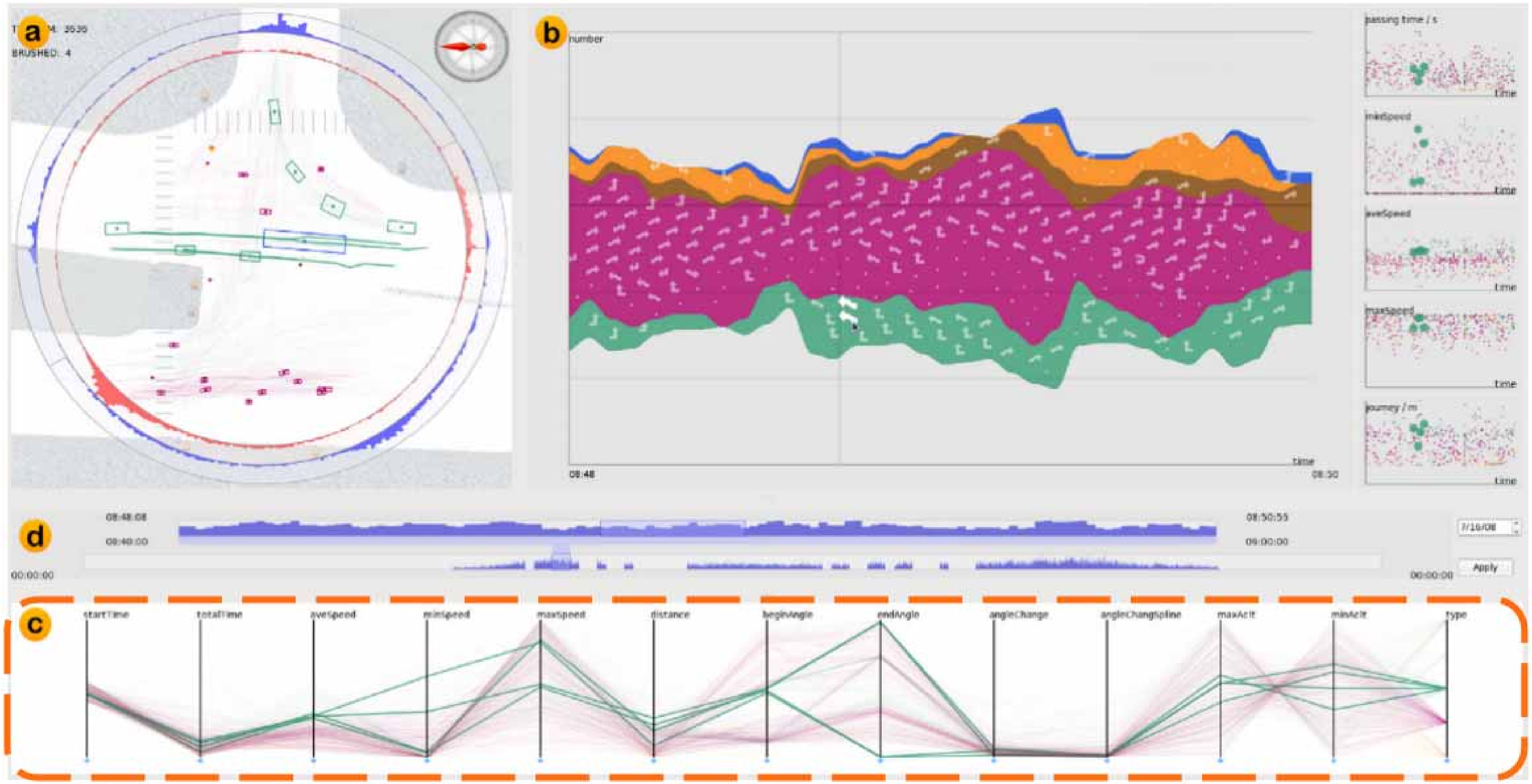
Werner Sturm, Till Schaefer, Tobias Schreck, Andeas Holzinger & Torsten Ullrich. Extending the Scaffold Hunter Visualization Toolkit with Interactive Heatmaps In: Borgo, Rita & Turkay, Cagatay, eds. EG UK Computer Graphics & Visual Computing CGVC 2015, 2015 University College London (UCL). Euro Graphics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

What is understandable, interpretable, intelligible?

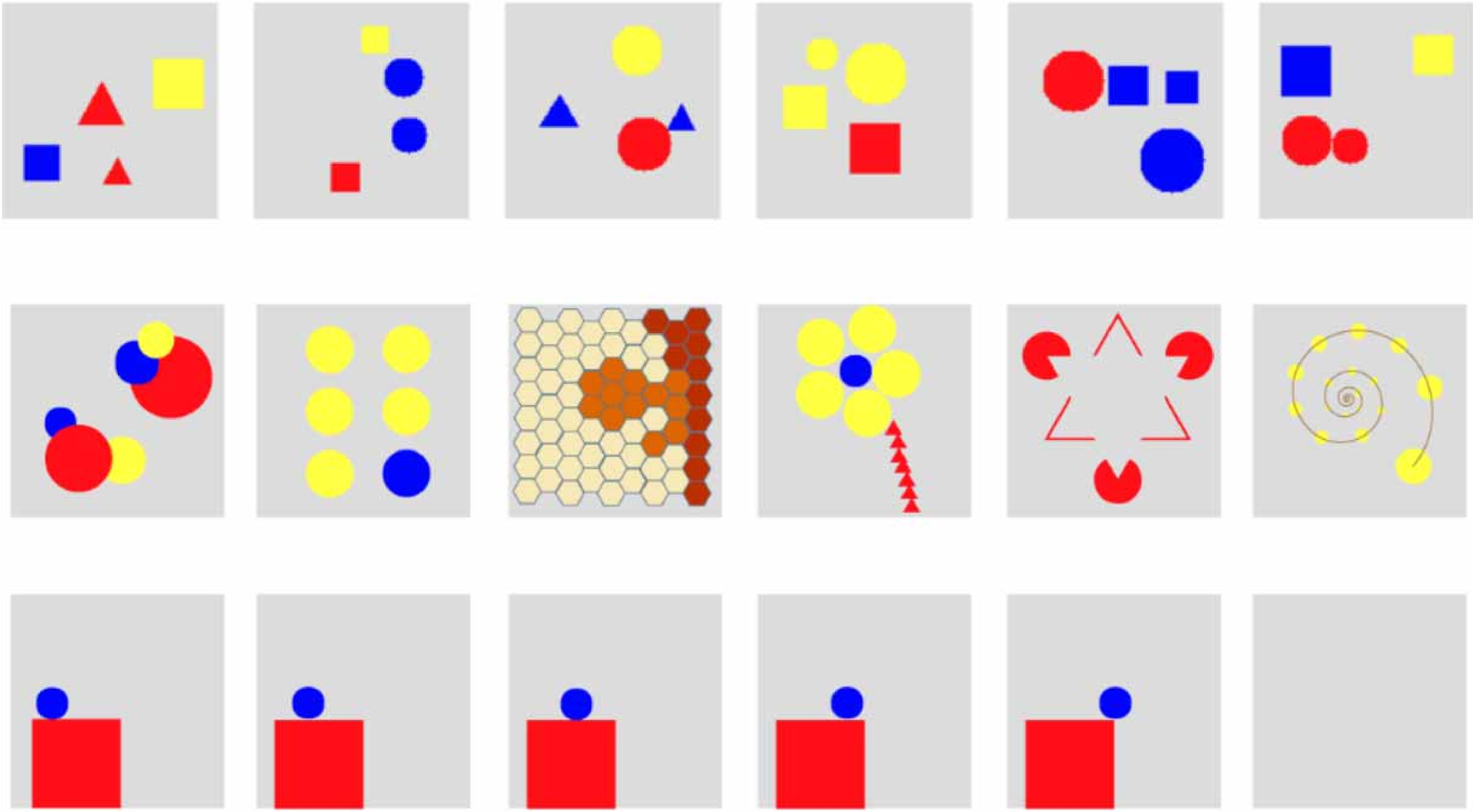


https://www.vis.uni-konstanz.de/en/members/fuchs/

Explainable AI is a huge challenge for visualization



Kandinsky Patterns: A possible IQ-Test for machines …



Heimo Müller & Andreas Holzinger 2019. Kandinsky Patterns. arXiv:1906.00657.

07 Methods of Explainable AI

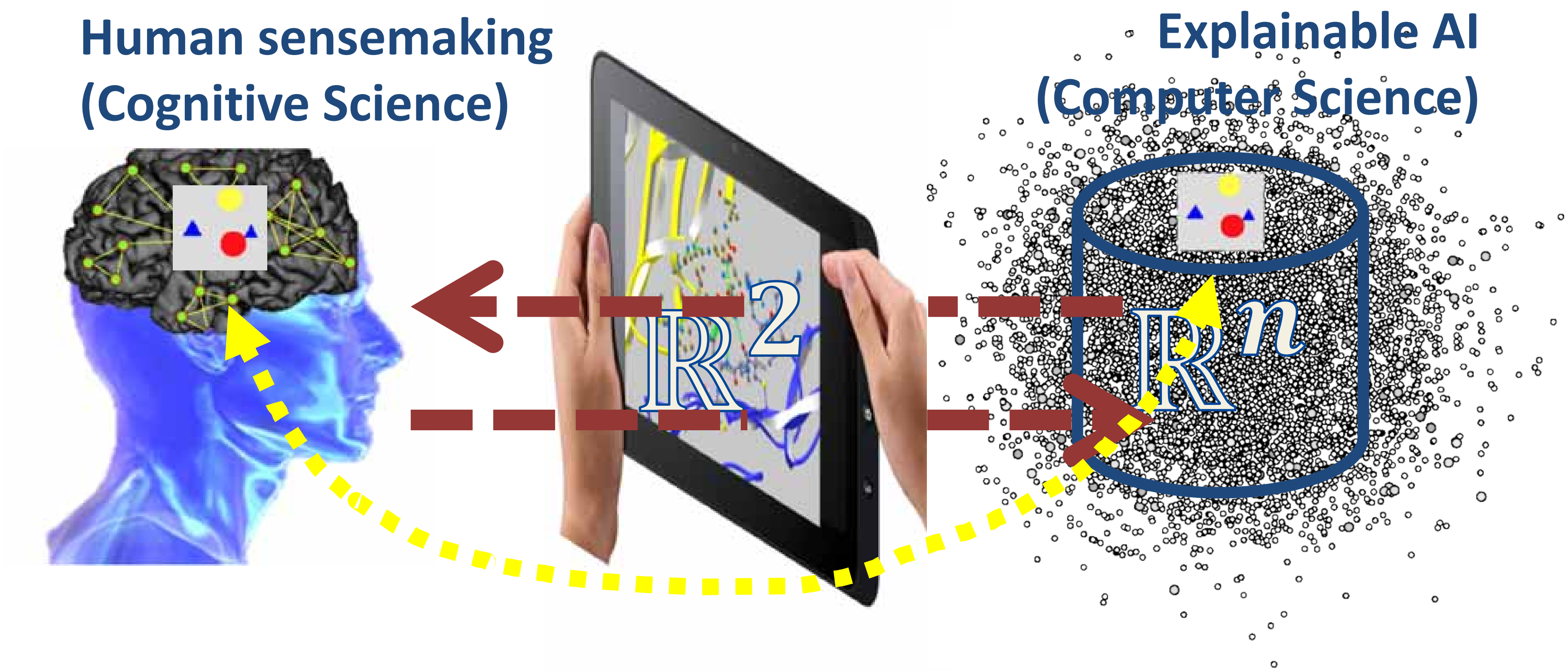
- 1) Gradients
- 2) Sensitivity Analysis
- 3) Decomposition Relevance Propagation (Pixel-RP, Layer-RP, Deep Taylor Decomposition, …)
- 4) Optimization (Local-IME – model agnostic, BETA transparent approximation, …)
- 5) Deconvolution and Guided Backpropagation
- 6) Model Understanding
  - Feature visualization, Inverting CNN
  - Qualitative Testing with Concept Activation Vectors TCAV
  - Network Dissection

Andreas Holzinger LV 706.315 From explainable AI to Causability, 3 ECTS course at Graz University of Technology https://human-centered.ai/explainable-ai-causability-2019 (course given since 2016)

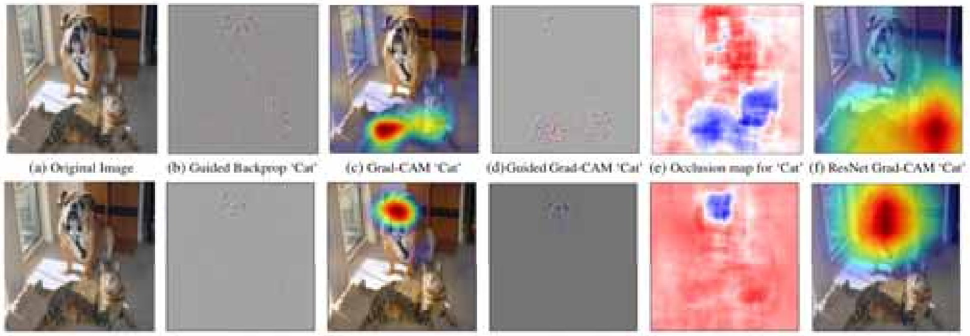
Explainability :=
a property of a system
(“the AI explanation)
Causability :=
a property of a person
(“the Human explanation)

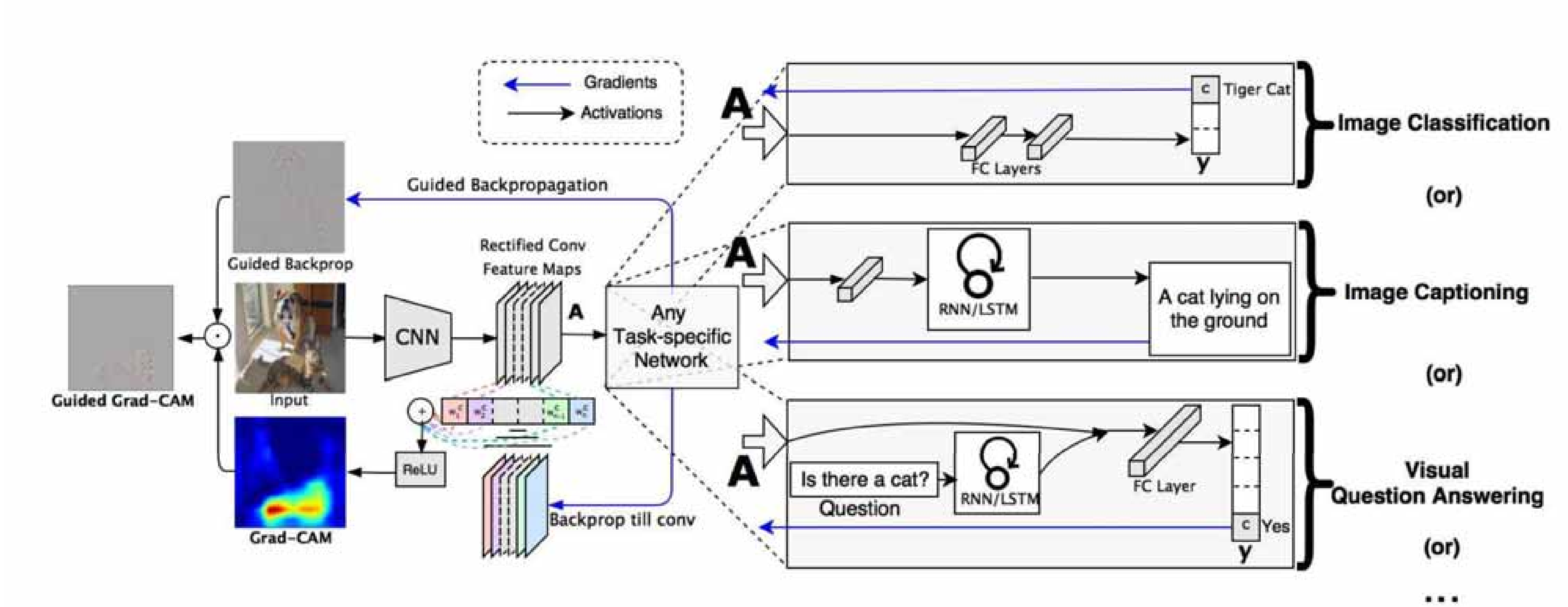
Andreas Holzinger et al. 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

- **Causability := a property of a person (Human)**
- **Explainability := a property of a system (Computer)**

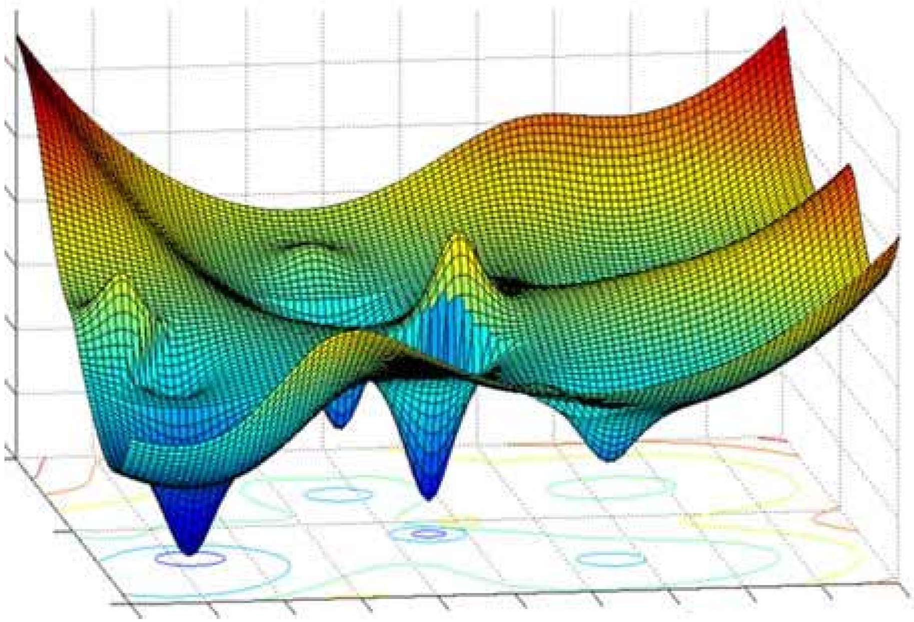


Gradients

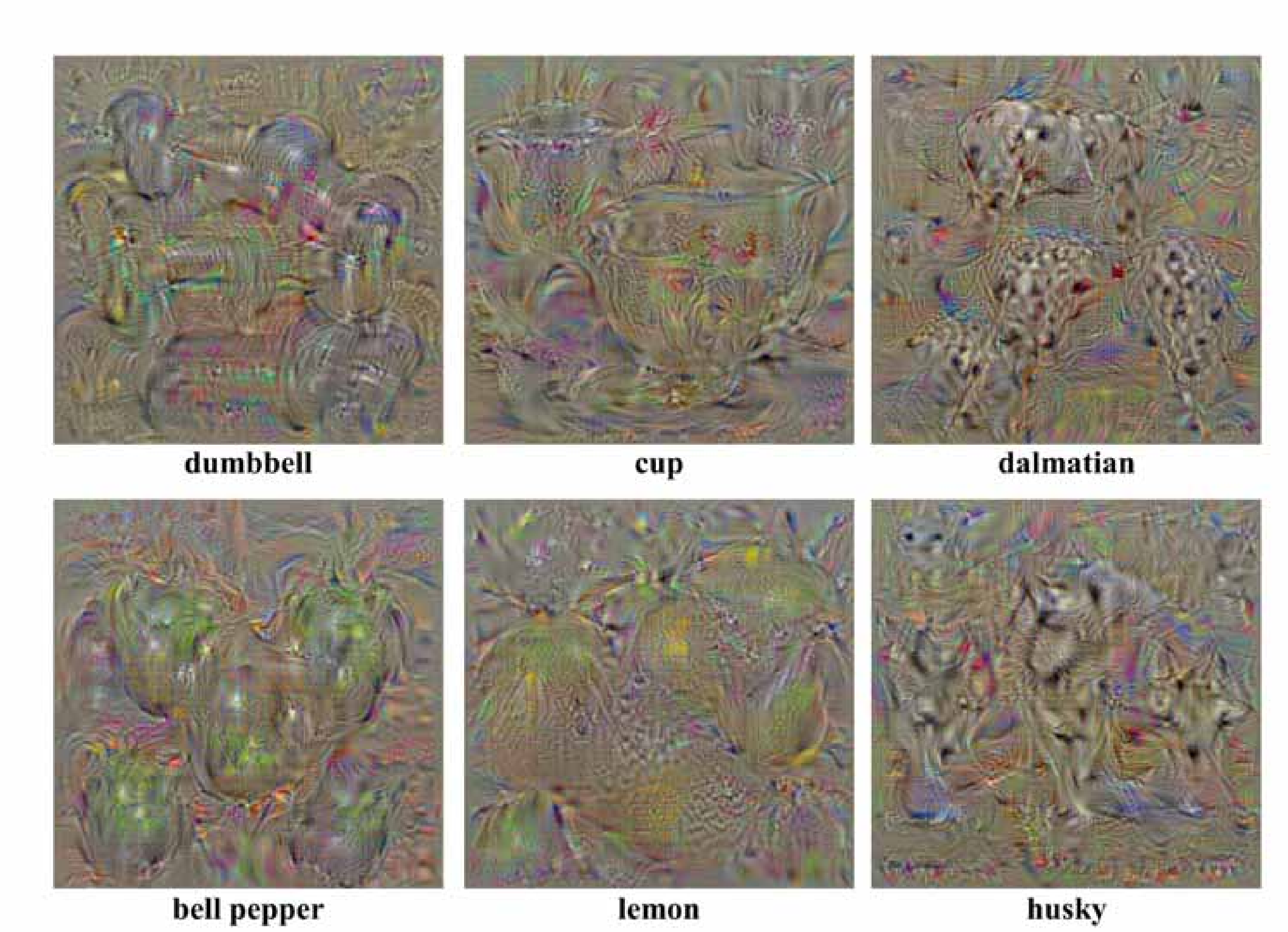




Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh & Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. ICCV, 2017. 618-626.

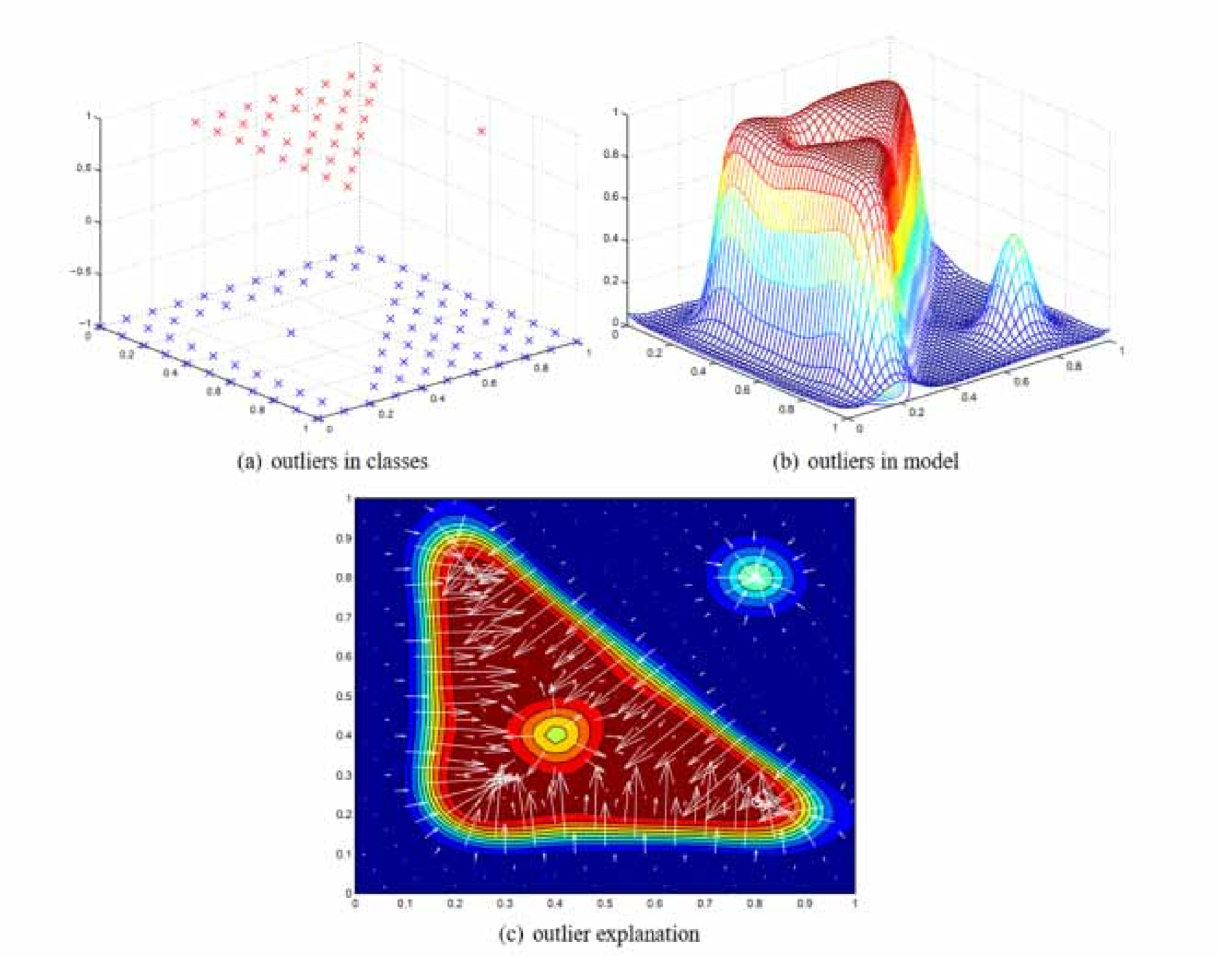


https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivative-and-gradient-articles/a/the-gradient



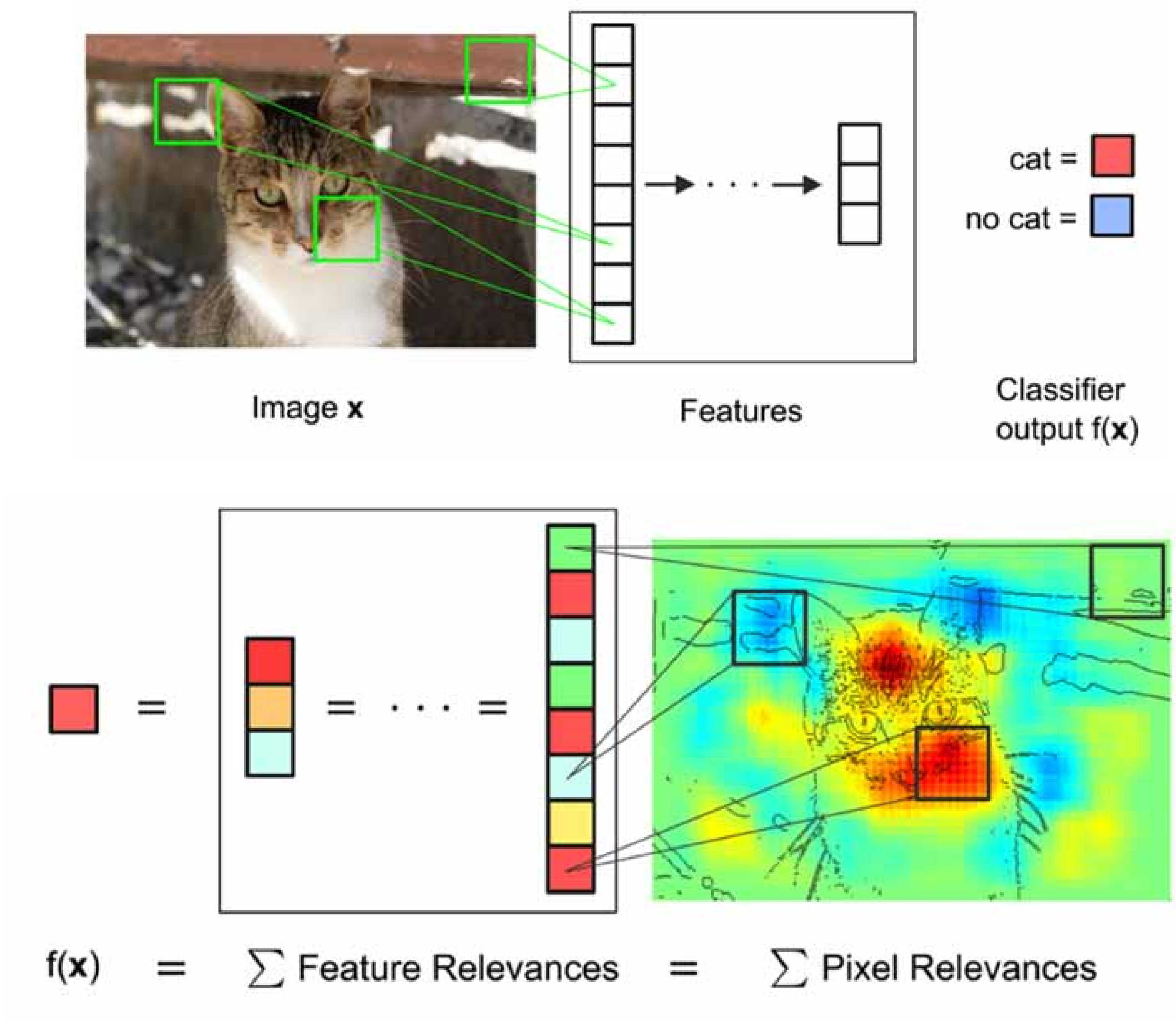
Karen Simonyan, Andrea Vedaldi & Andrew Zisserman 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv:1312.6034.

Gradients

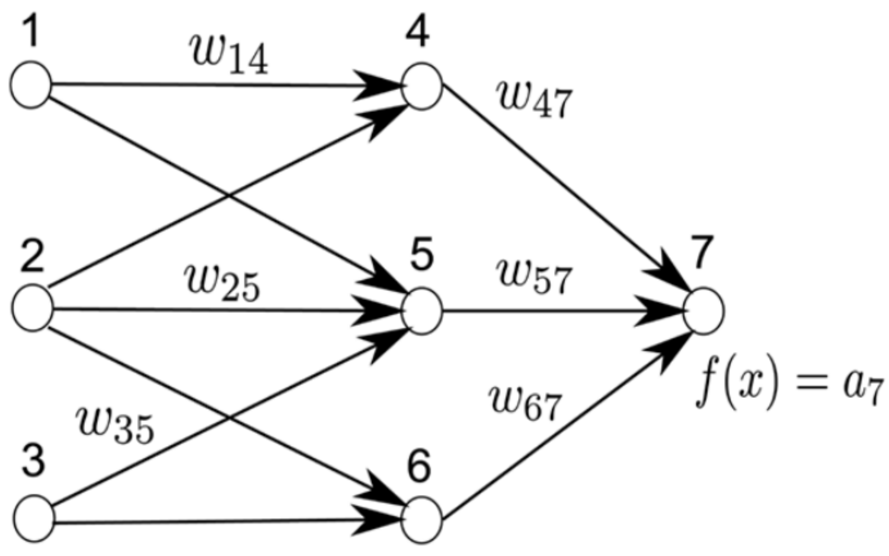


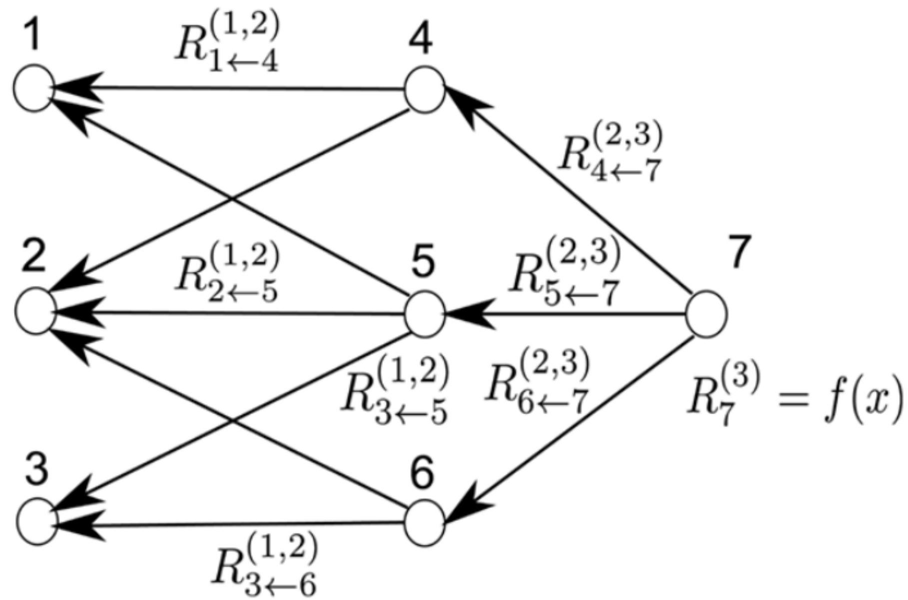
David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen & Klaus-Robert Mueller 2010. How to explain individual classification decisions. Journal of machine learning research (JMLR), 11, (6), 1803-1831.

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.

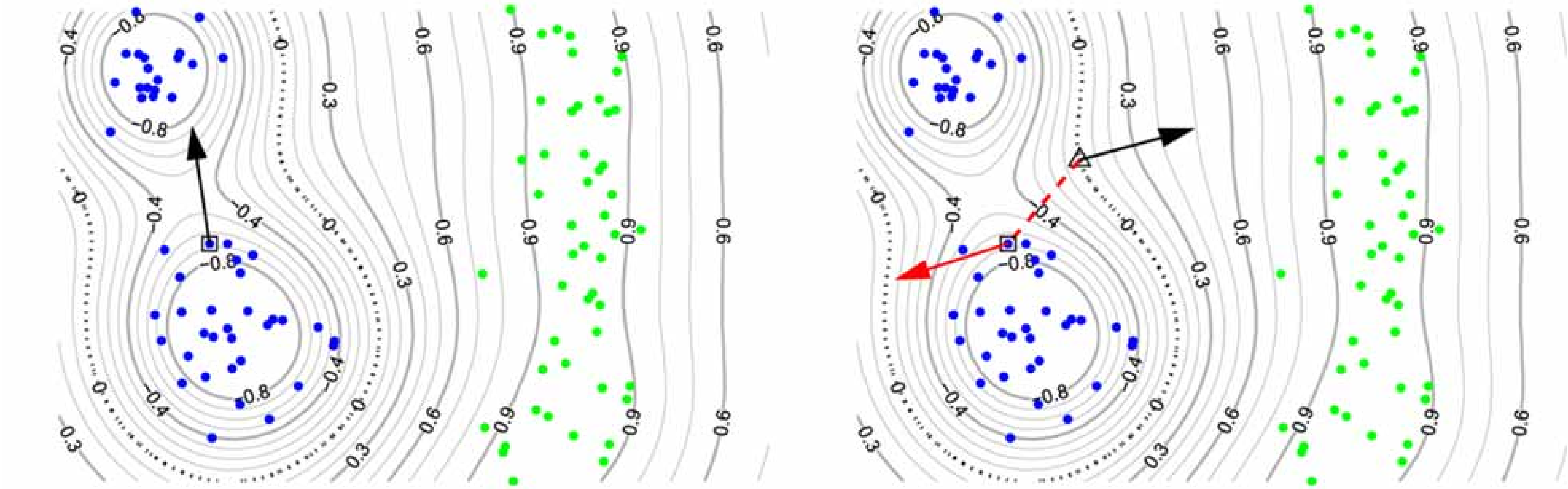




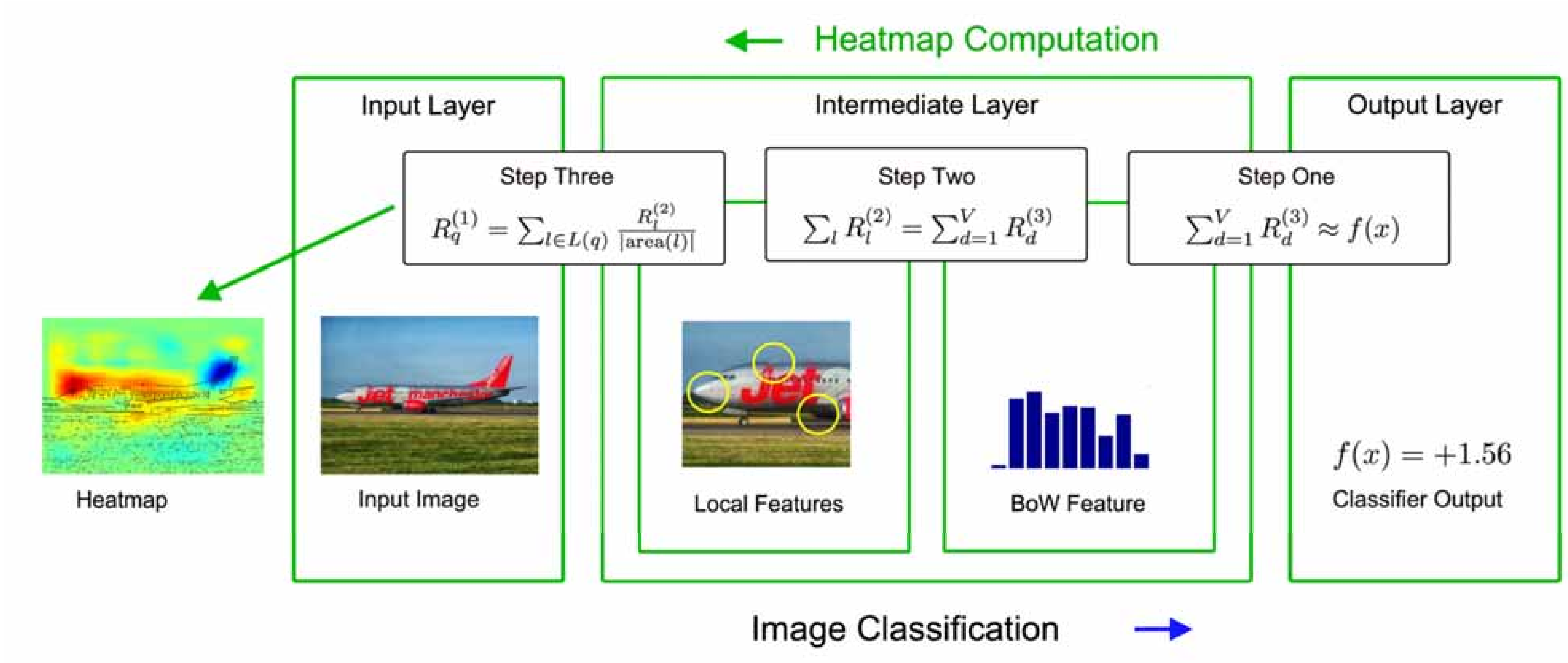
$$f(x) = \cdots = \sum_{d\in l+1} R_d^{(l+1)} = \sum_{d\in l} R_d^{(l)} = \cdots = \sum_{d} R_d^{(1)}$$

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.

Example Taylor Decomposition

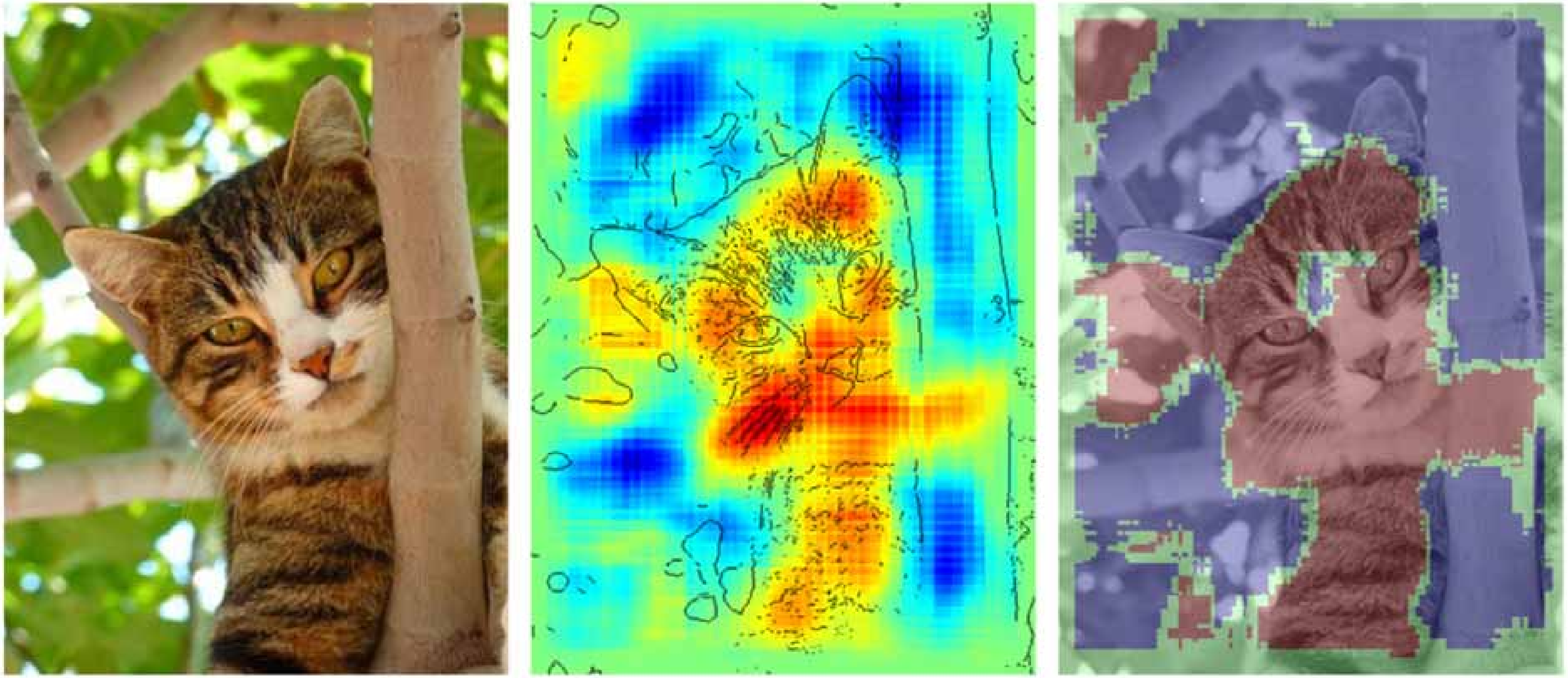


Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller & Wojciech Samek 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi:10.1371/journal.pone.0130140.



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.

Pixel-wise decomposition for bag-of-words features



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.

Definition 1. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *conservative* if the sum of assigned relevances in the pixel space corresponds to the total relevance detected by the model:

$$\forall\, \boldsymbol{x}: f(\boldsymbol{x}) = \sum_p R_p(\boldsymbol{x}).$$

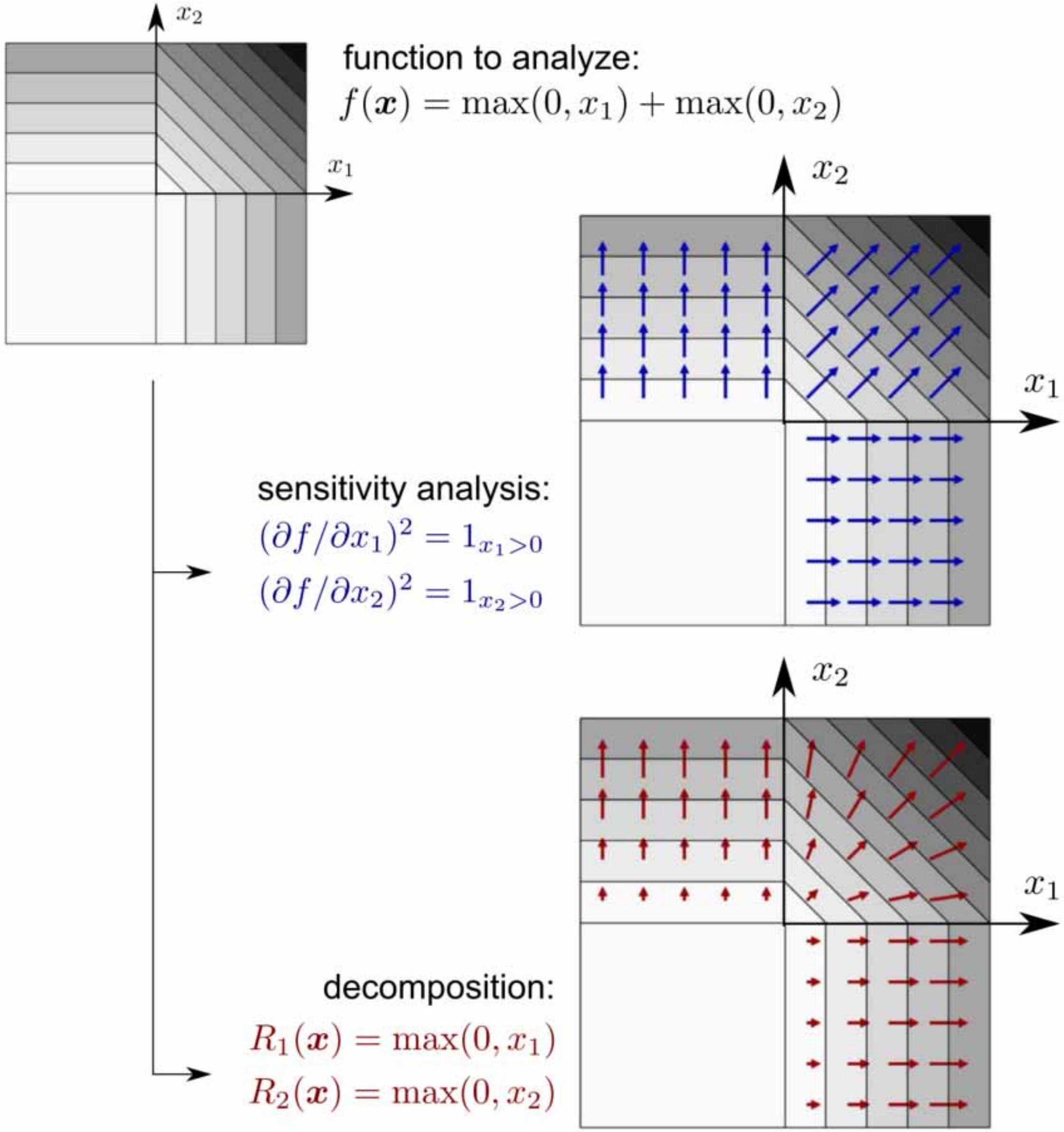
Definition 2. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *positive* if all values forming the heatmap are greater or equal to zero, that is:

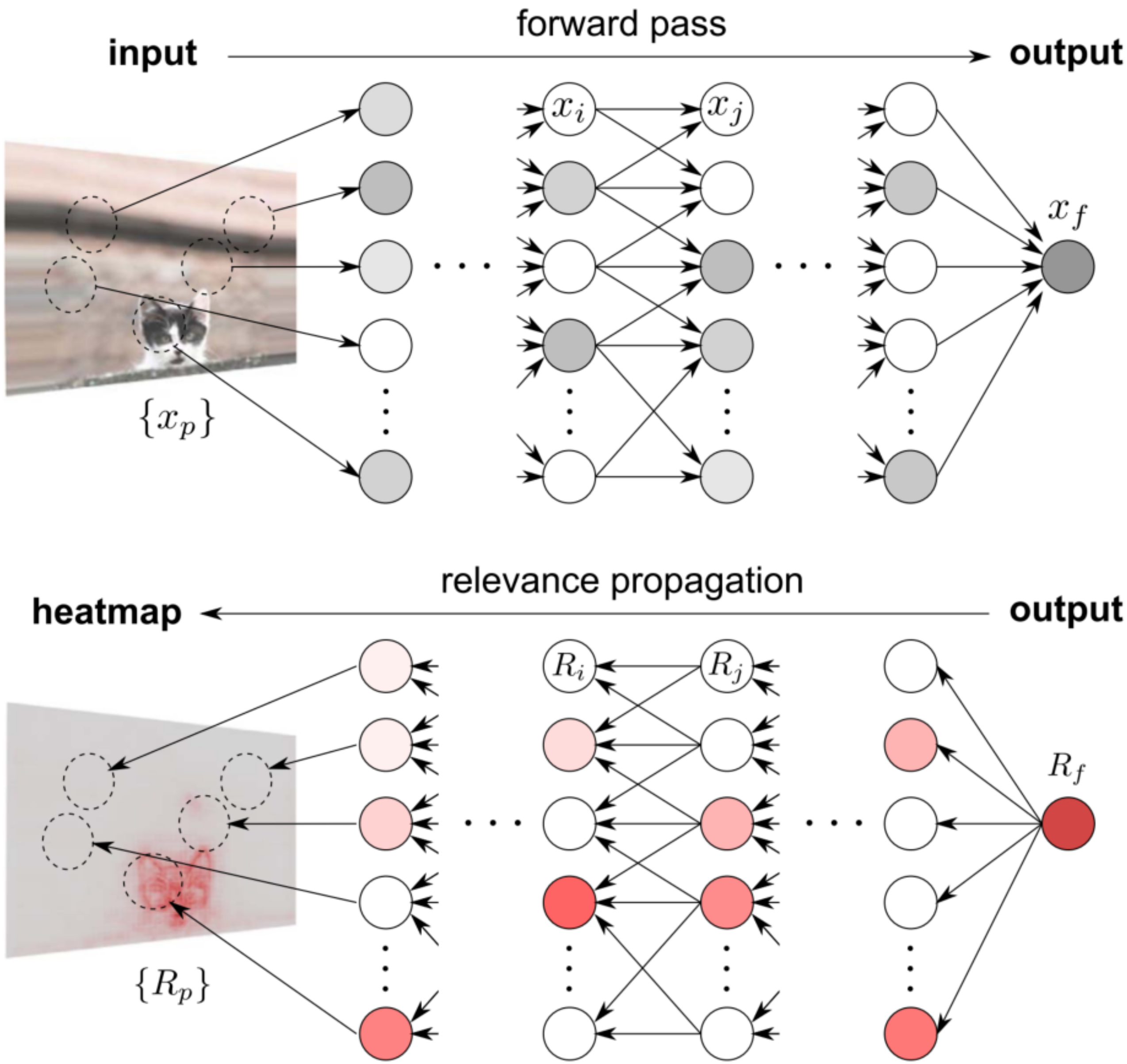
$$\forall\, \boldsymbol{x}, p: R_p(\boldsymbol{x}) \geq 0$$

Definition 3. A heatmapping $\boldsymbol{R}(\boldsymbol{x})$ is *consistent* if it is conservative *and* positive. That is, it is consistent if it complies with Definitions 1 and 2.

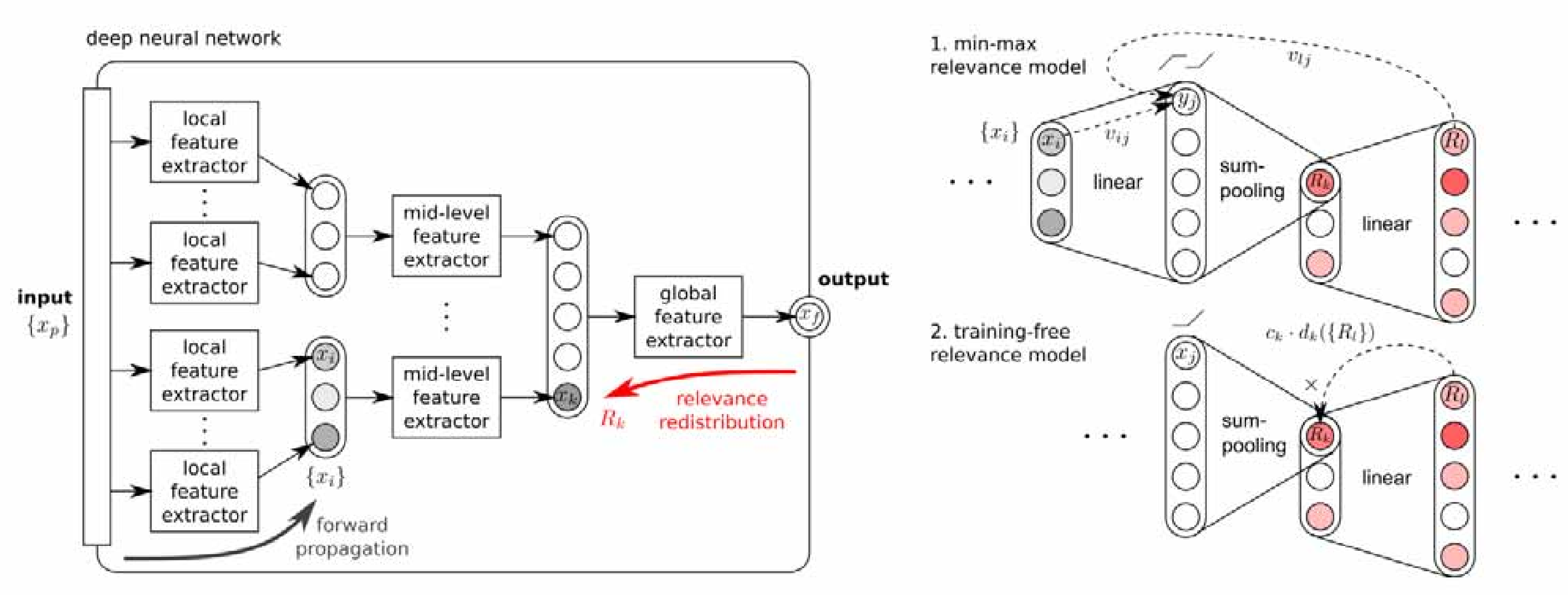
Gregoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek & Klaus-Robert Müller 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition, 65, 211-222, doi:10.1016/j.patcog.2016.11.008.

Sensitivity Analysis vs. Decomposition





forward pass
input
output
x_i
x_j
x_f
$\{x_p\}$
relevance propagation
heatmap
output
R_i
R_j
R_f
$\{R_p\}$



deep neural network
input
local feature extractor
local feature extractor
local feature extractor
local feature extractor
mid-level feature extractor
mid-level feature extractor
global feature extractor
output
relevance redistribution
forward propagation
1. min-max relevance model
linear
sum-pooling
linear
2. training-free relevance model
sum-pooling
linear

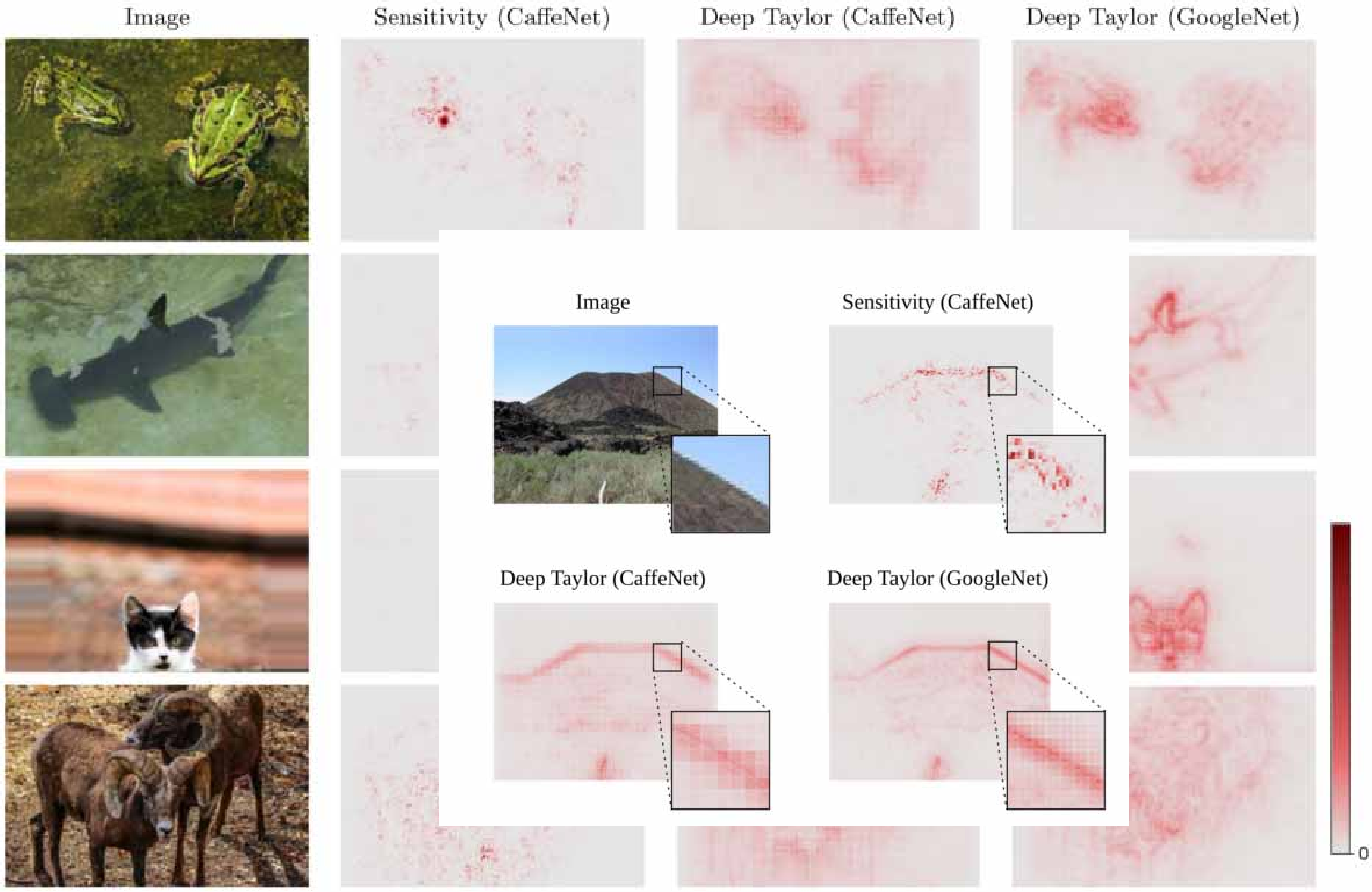
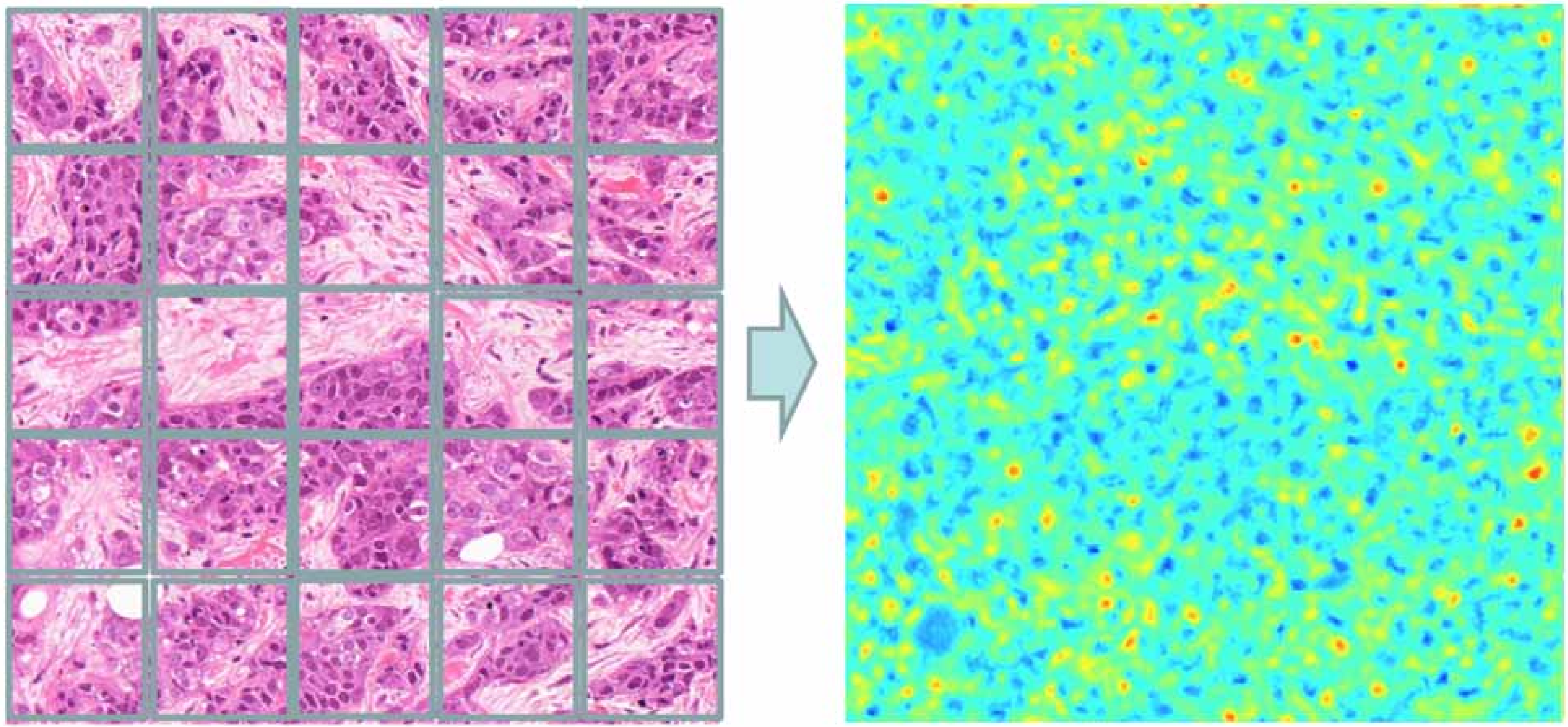
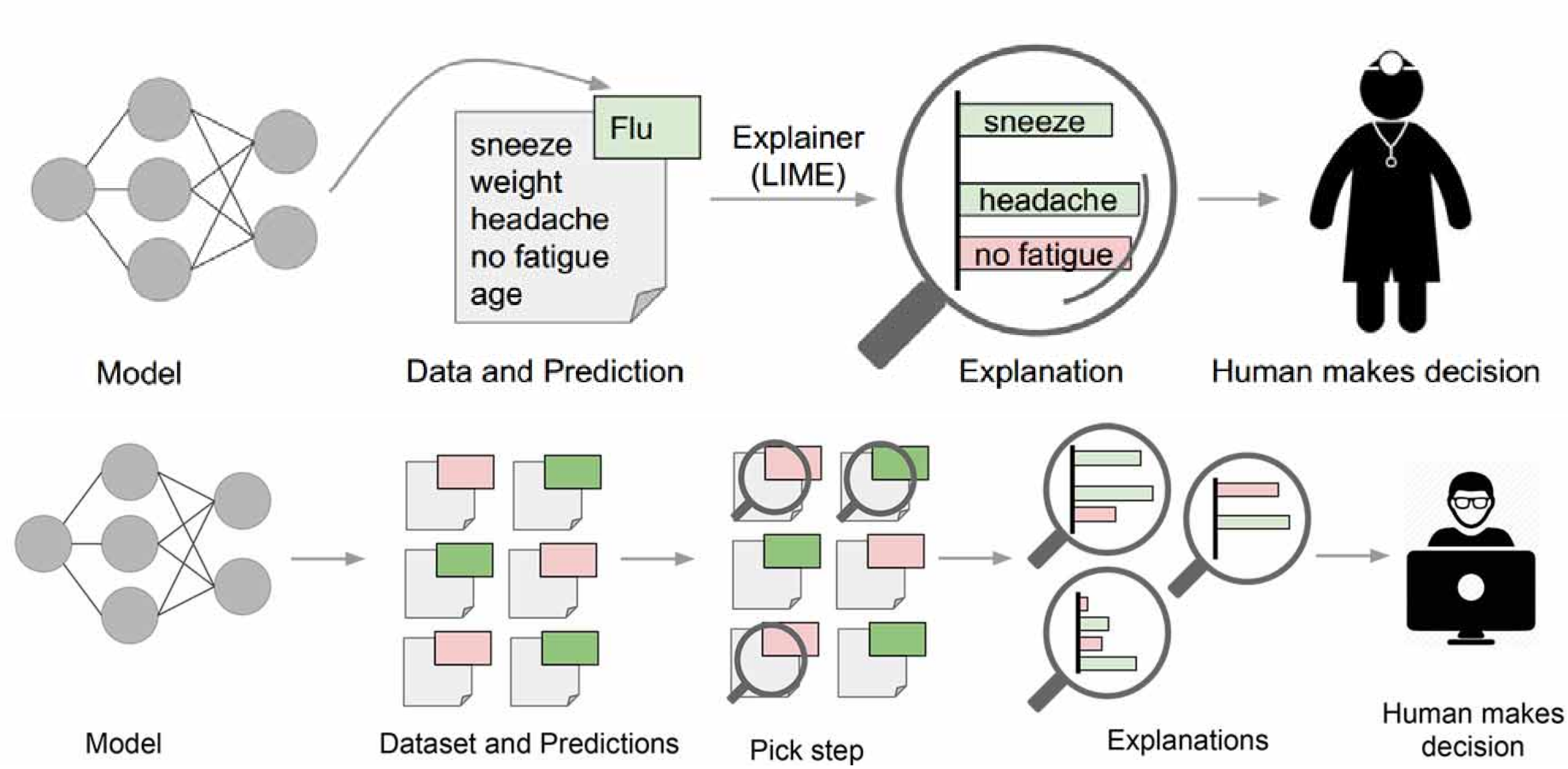


Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
Image
Sensitivity (CaffeNet)
Deep Taylor (CaffeNet)
Deep Taylor (GoogleNet)
0

Example 2 Histopathology



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.



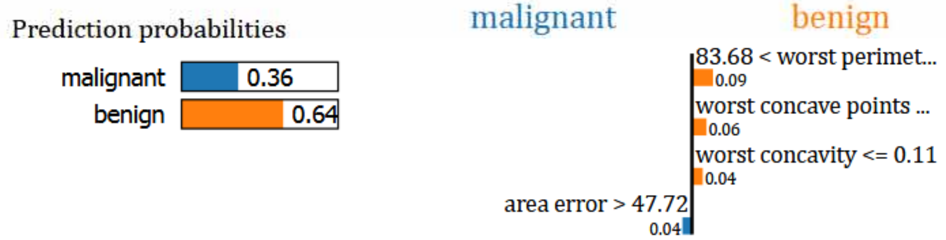
Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. ACM, 1135-1144, doi:10.1145/2939672.2939778.

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)
```

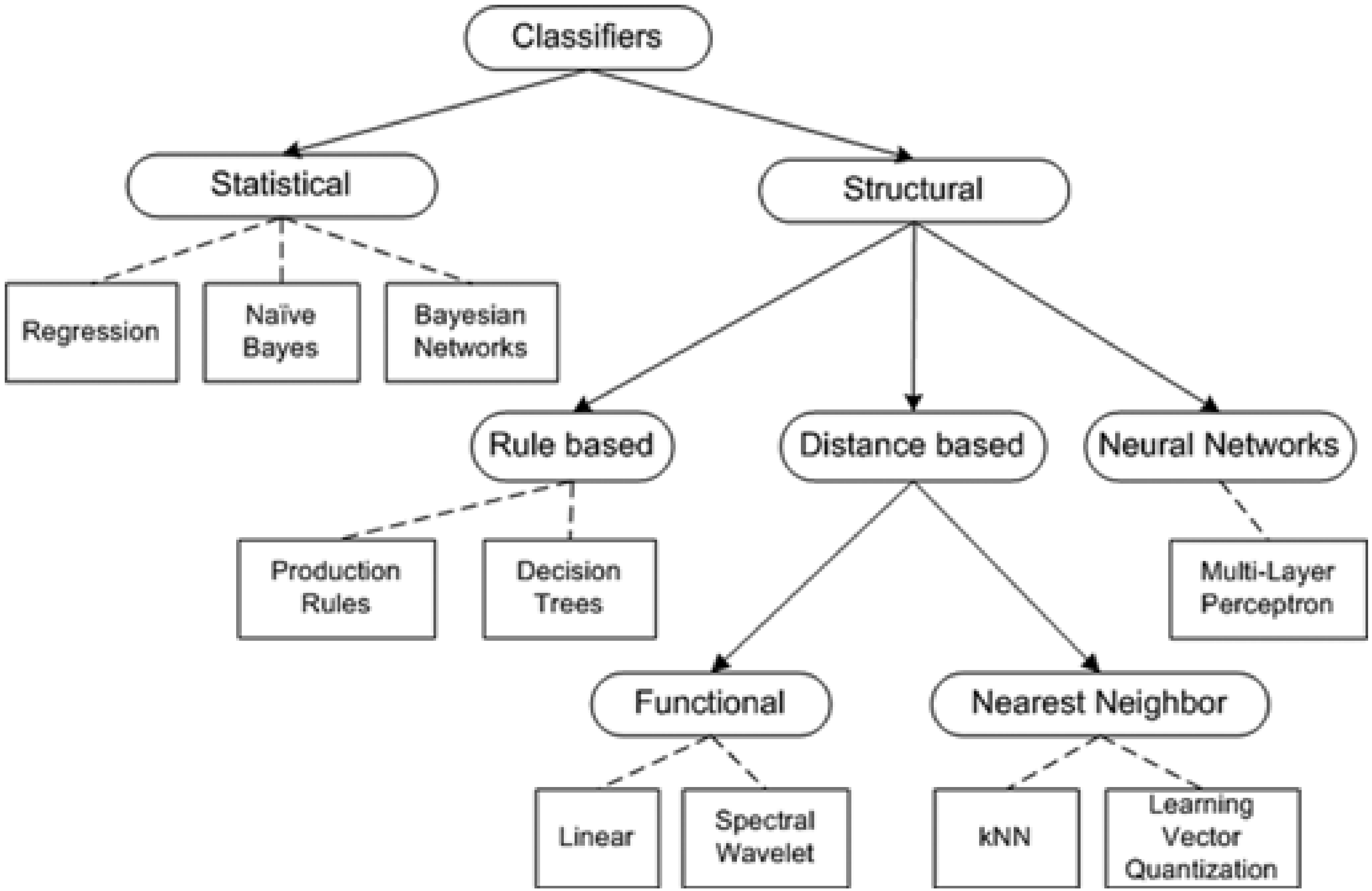


Feature	Value
worst perimeter	95.54
worst concave points	0.03
worst concavity	0.02
area error	106.00

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):



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:

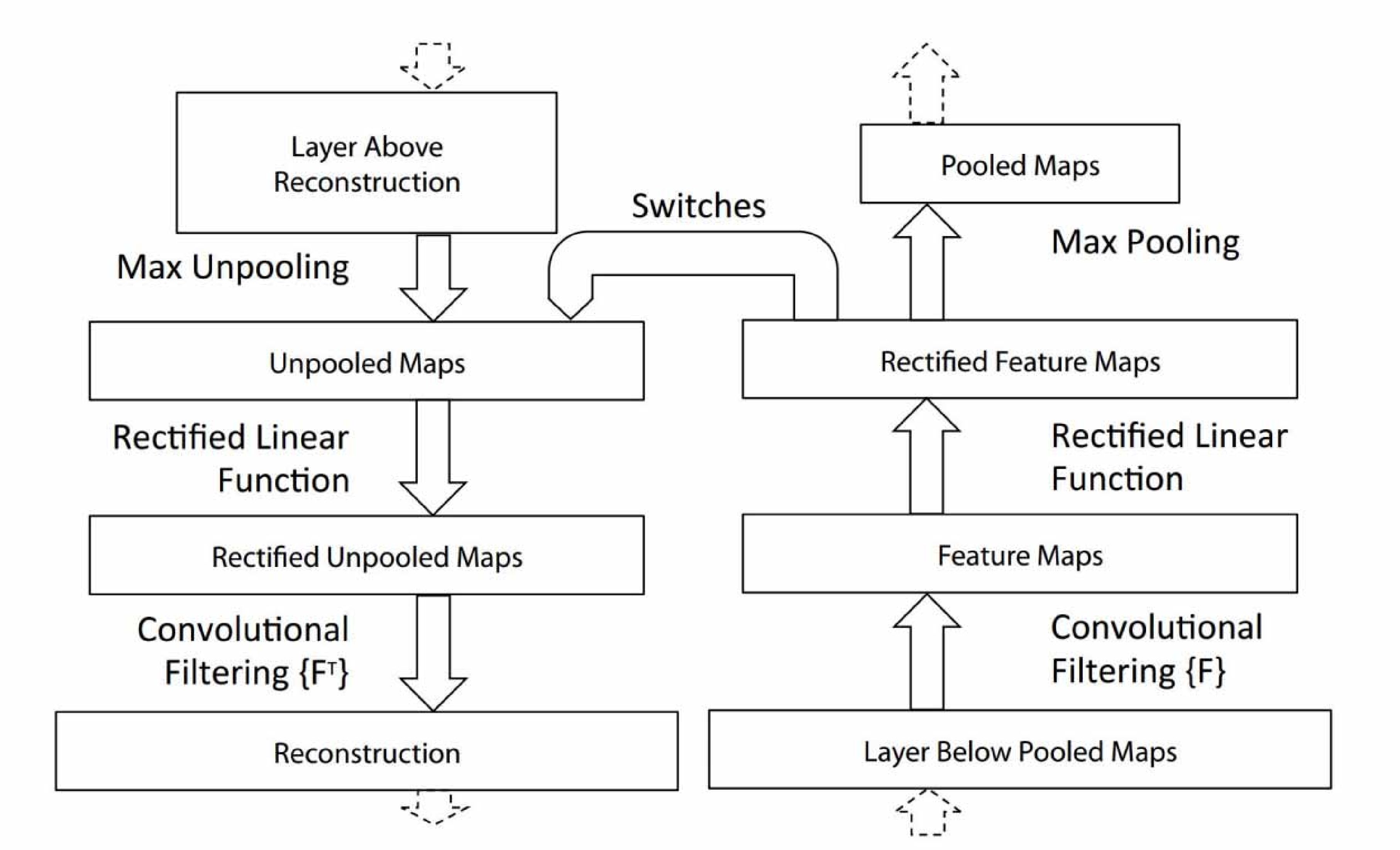
If Past-Depression = Yes **and** Tiredness = No **and** Exercise = No **and** Insomnia = Yes, **then** Depression

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

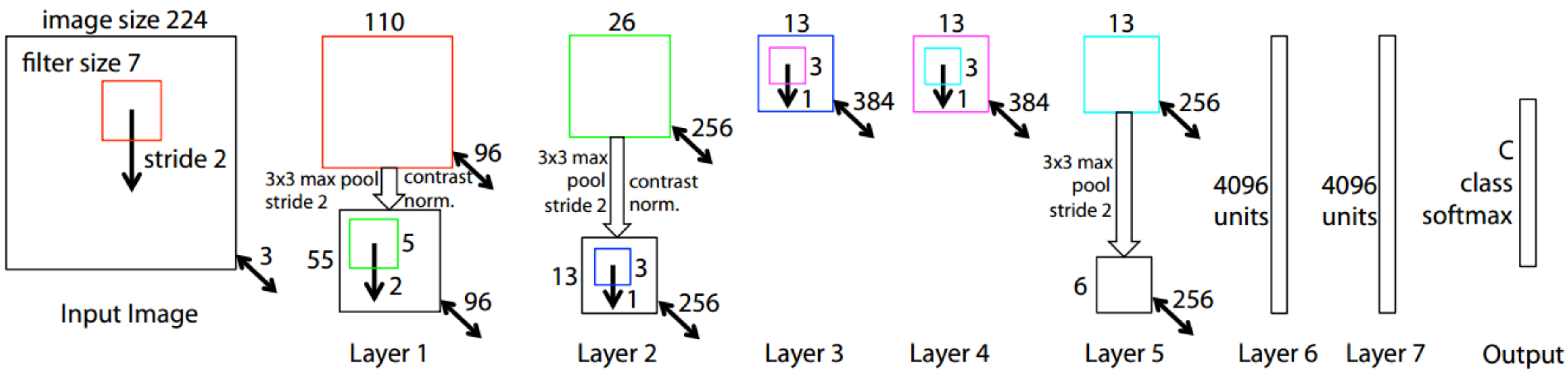
Himabindu Lakkaraju, Ece Kamar, Rich Caruana & Jure Leskovec 2017. Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.

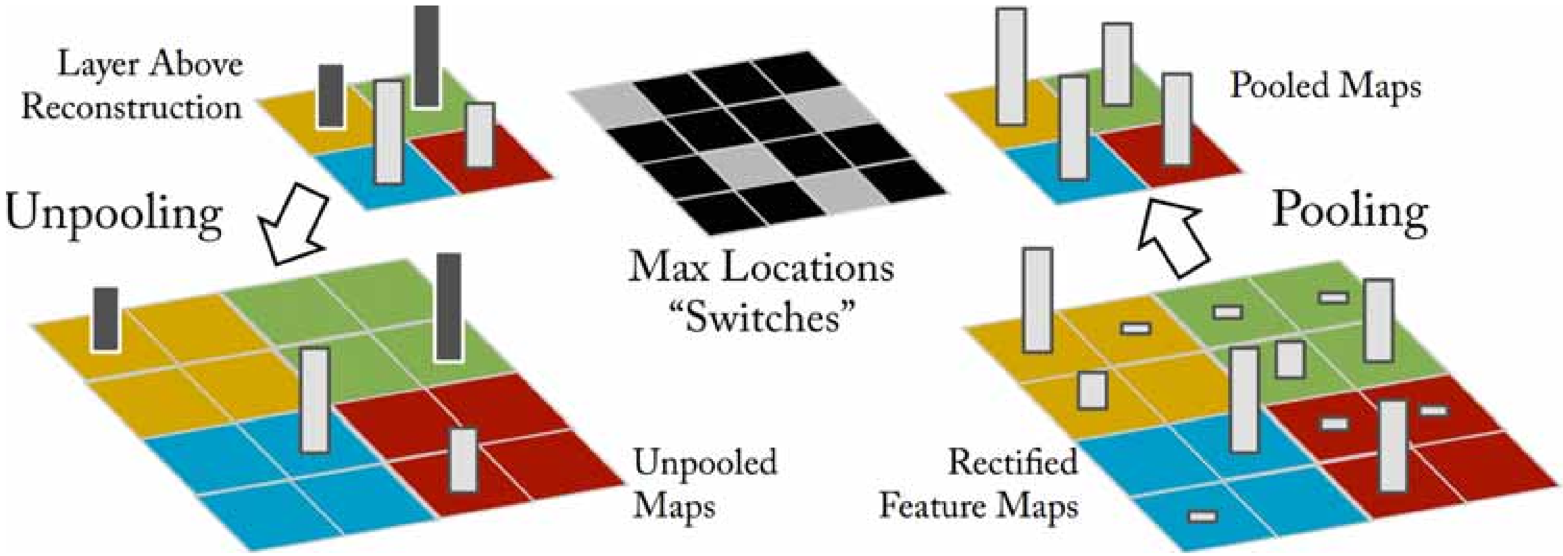
Example: Interpretable Deep Learning Model



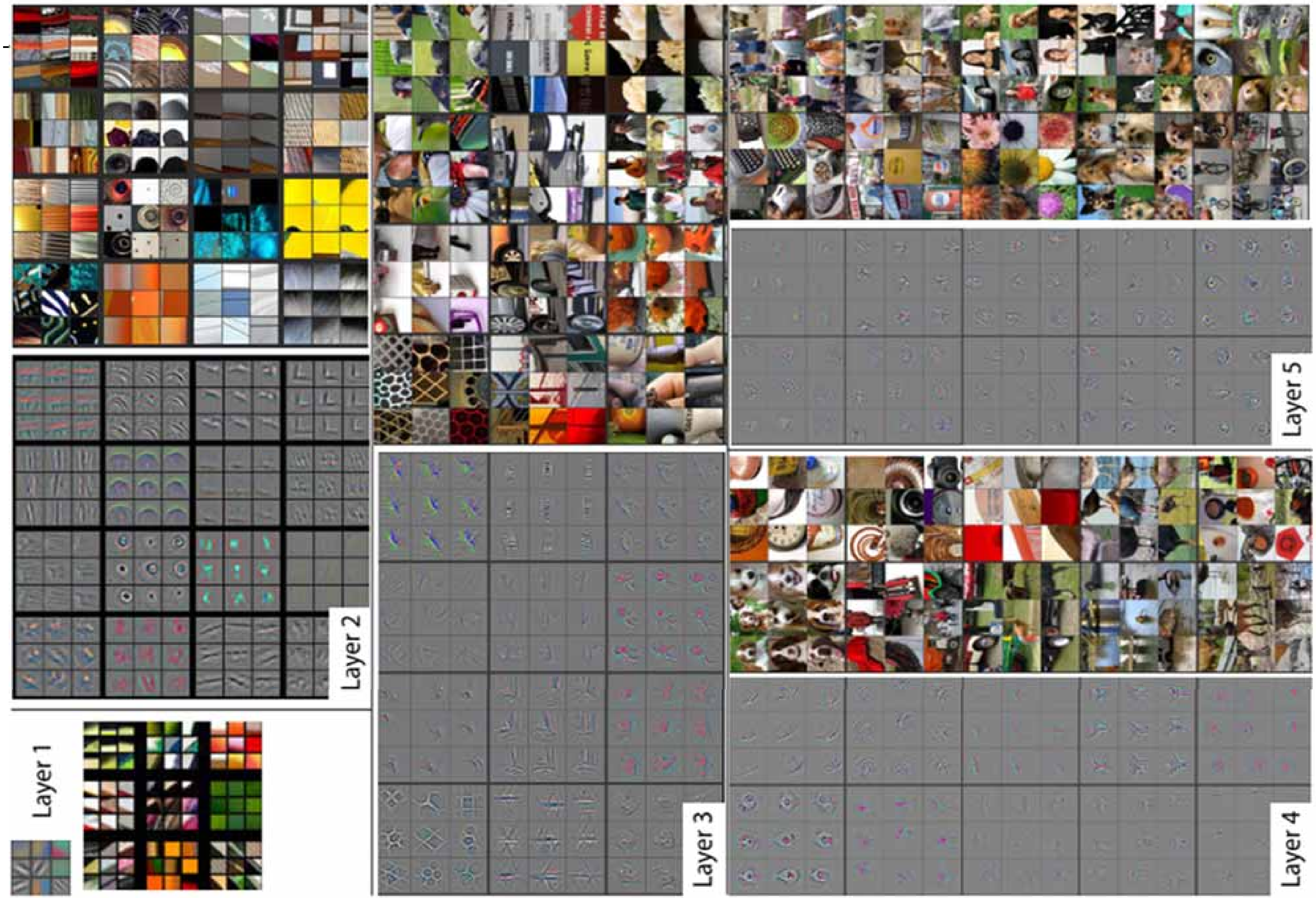
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

Visualizing a Conv Net with a De-Conv Net



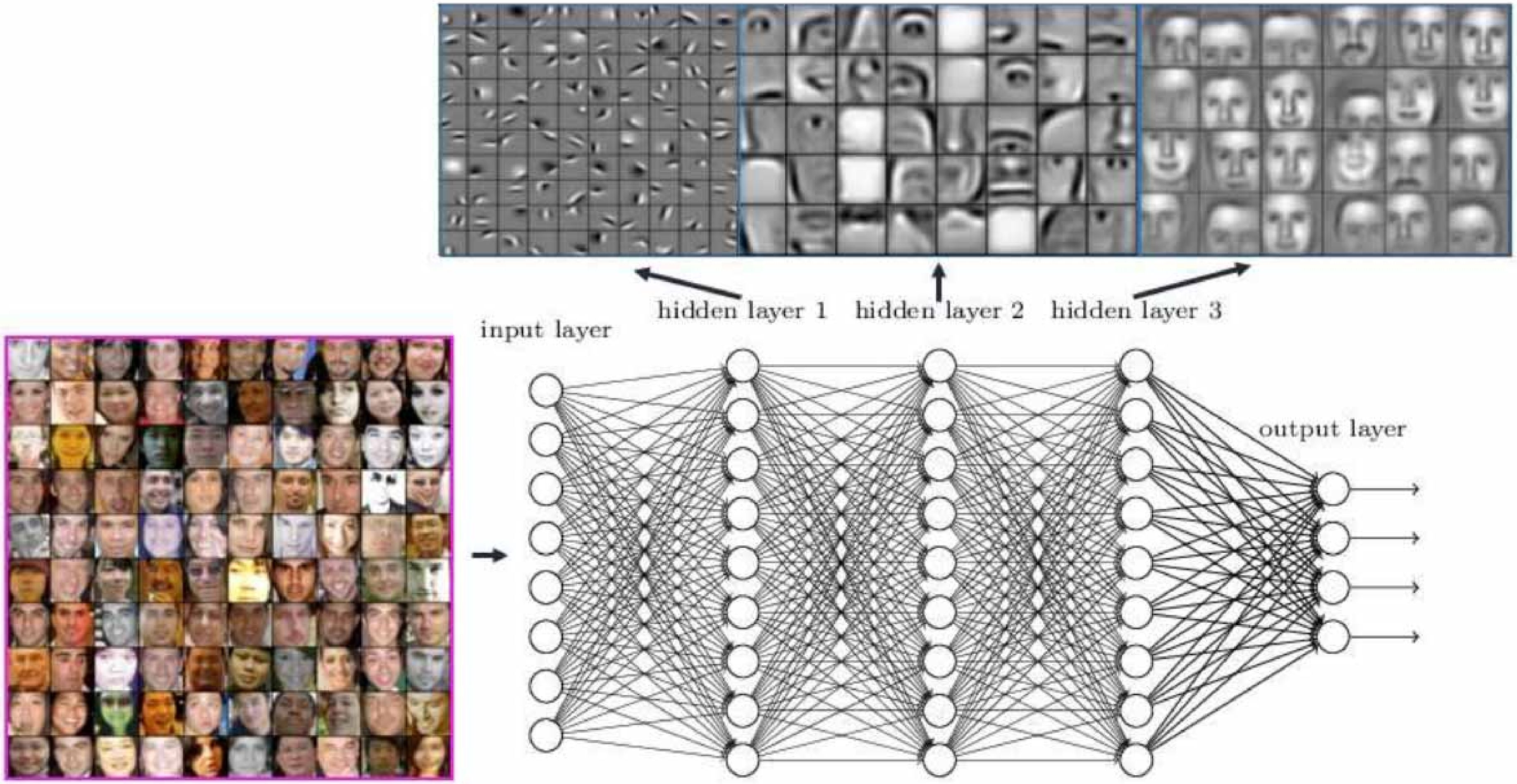


Matthew D. Zeiler & Rob Fergus 2014. Visualizing and understanding convolutional networks. In: D., Fleet, T., Pajdla, B., Schiele & T., Tuytelaars (eds.) ECCV, Lecture Notes in Computer Science LNCS 8689. Cham: Springer, pp. 818-833, doi:10.1007/978-3-319-10590-1_53.



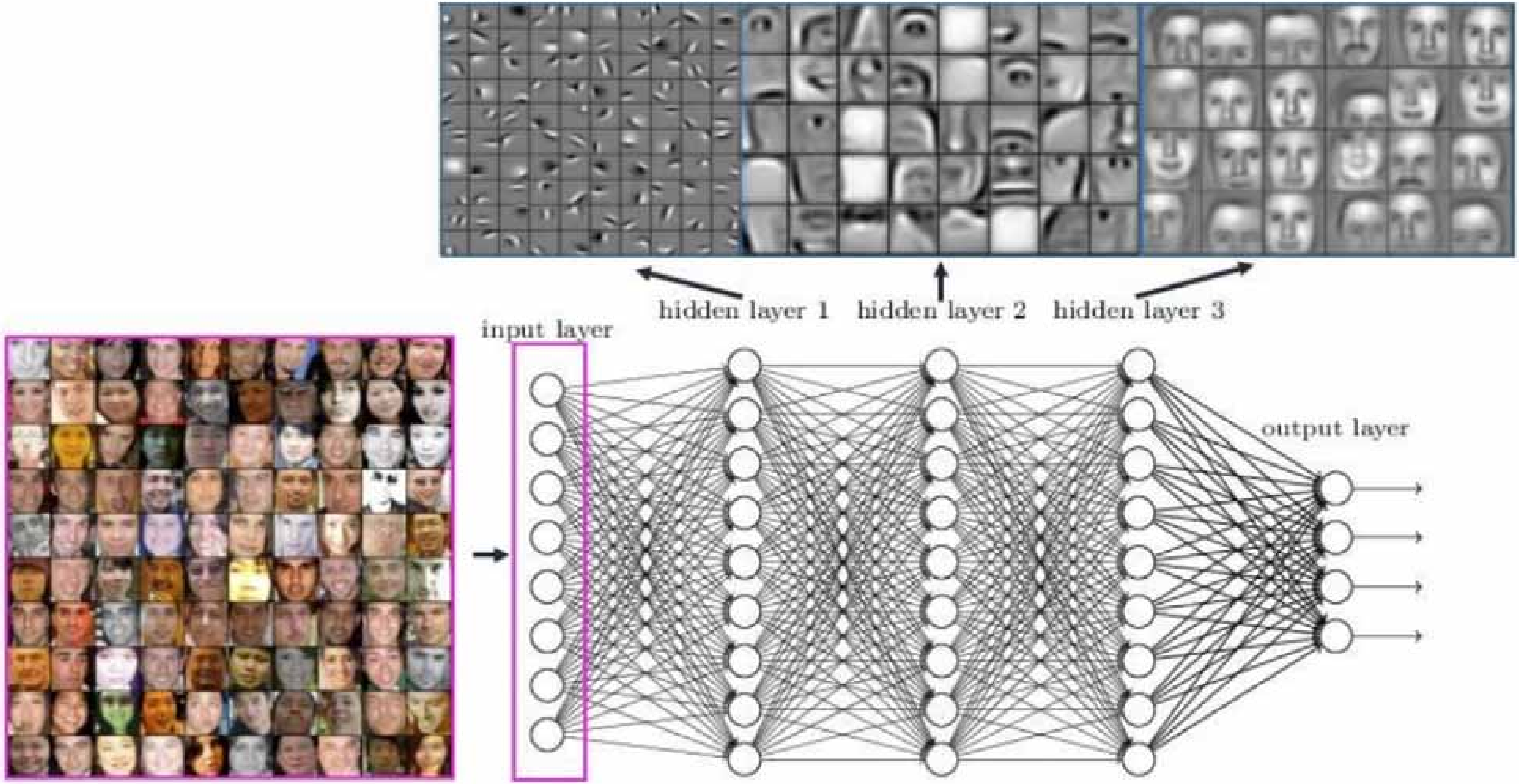
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901.

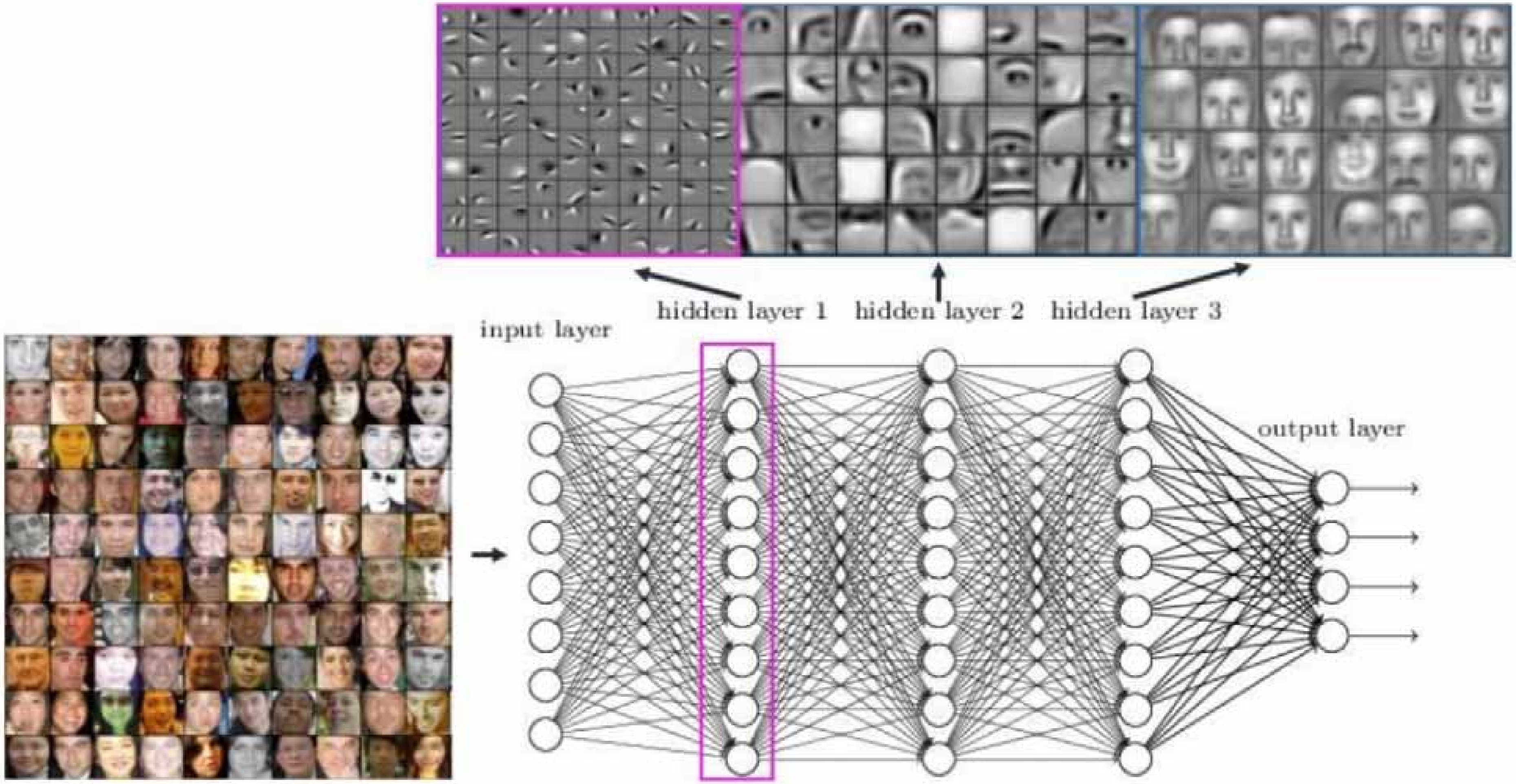
The world is compositional (Yann LeCun)

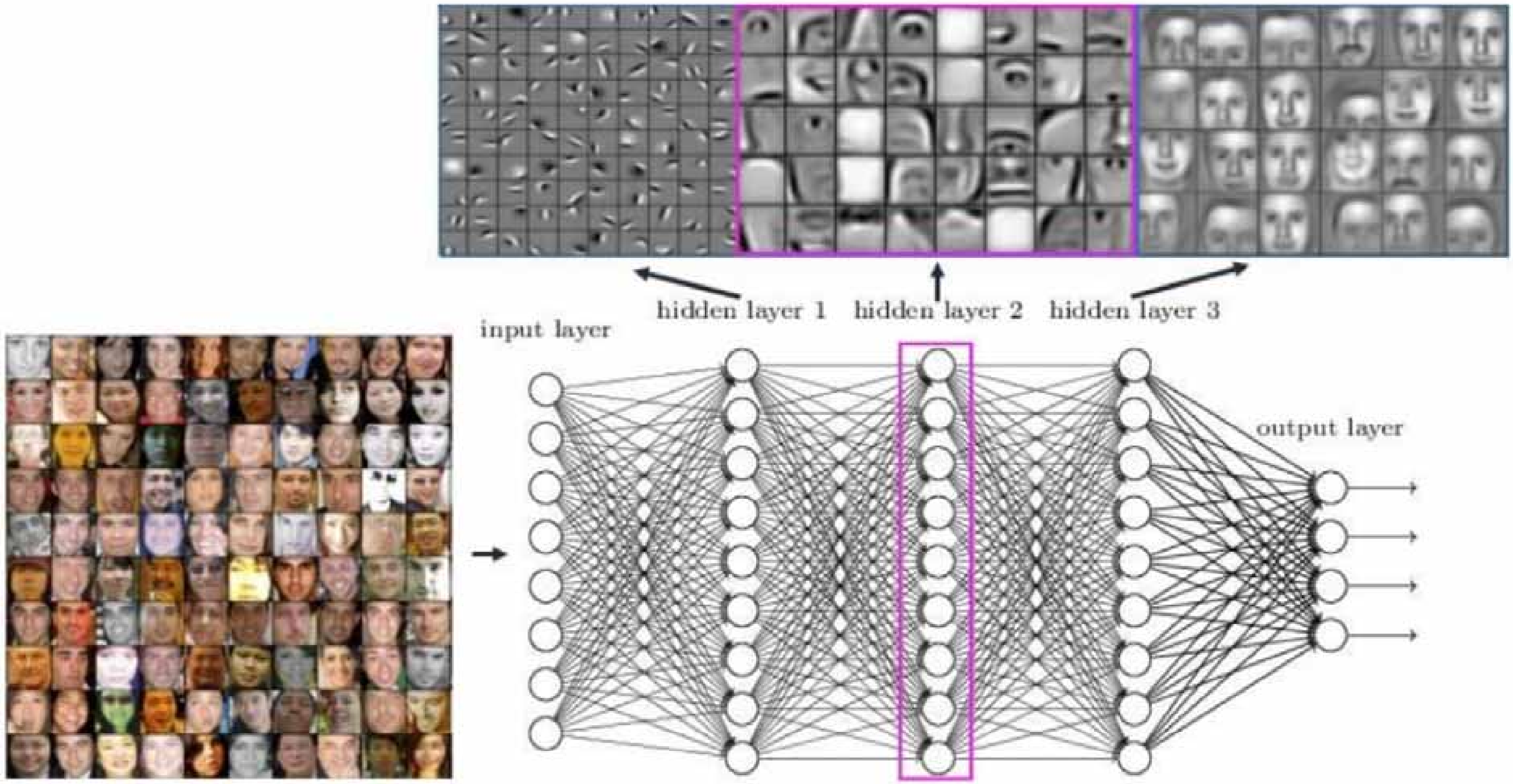


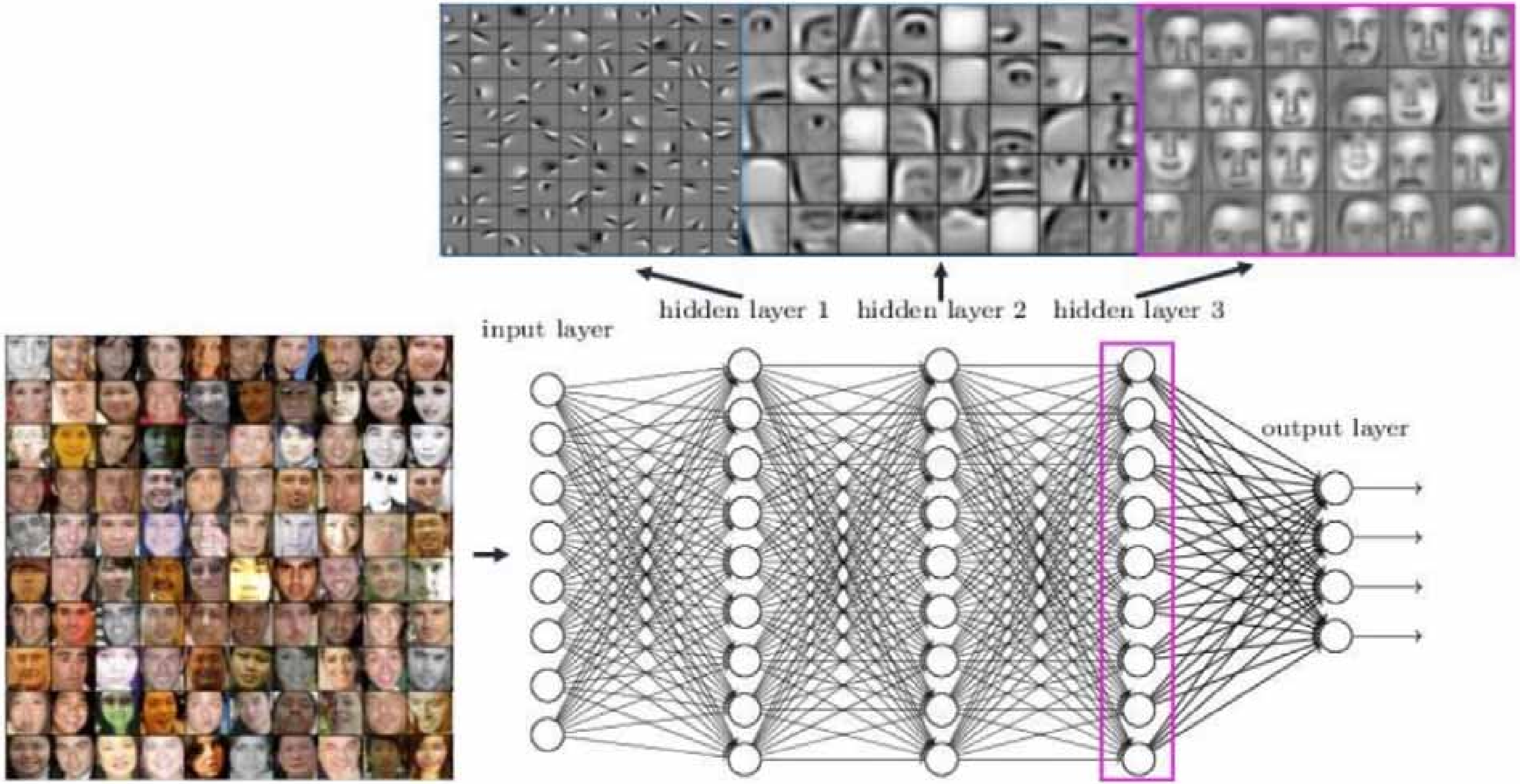
Matthew D. Zeiler & Rob Fergus 2013. Visualizing and Understanding Convolutional Networks. arXiv:1311.2901

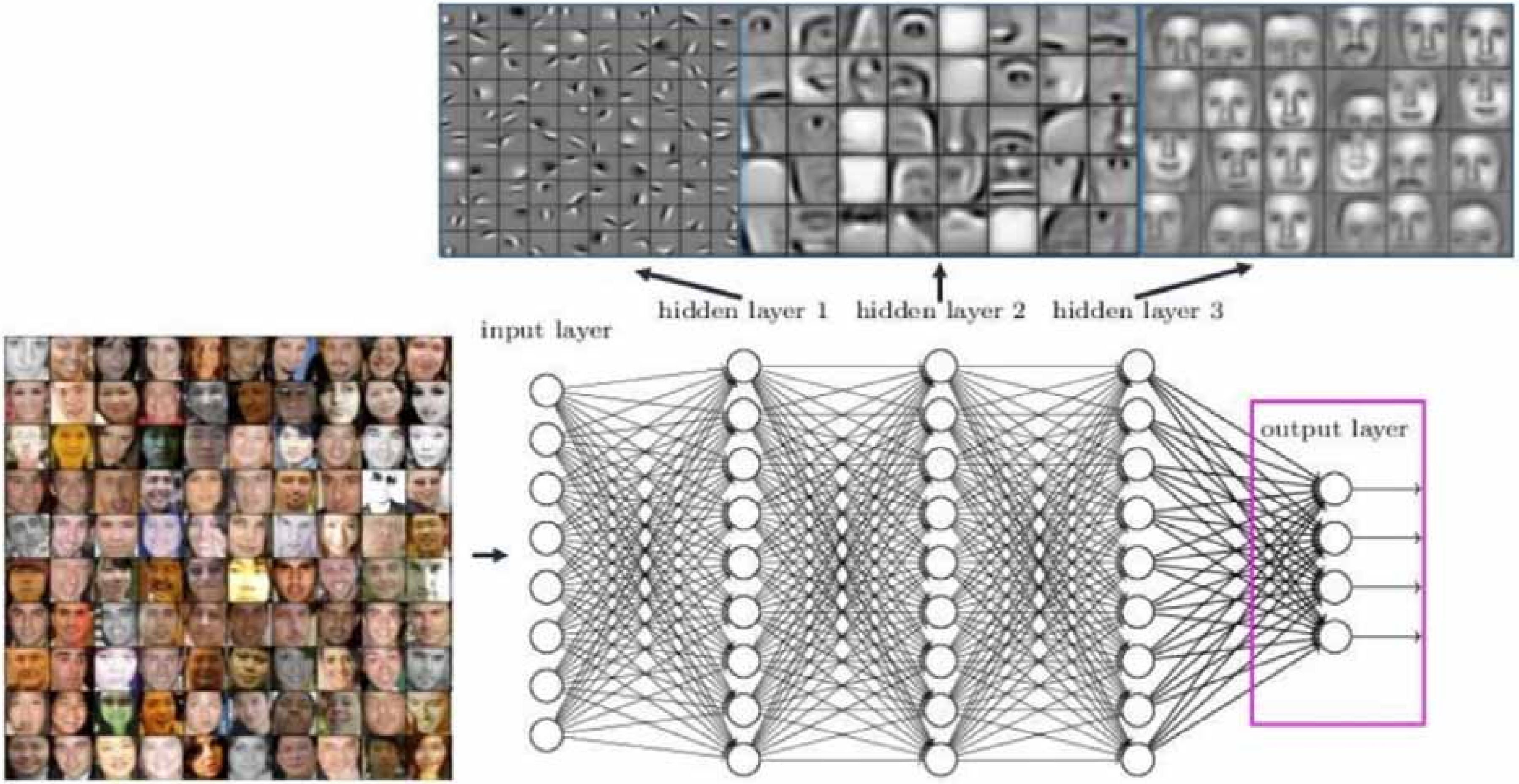
The world is compositional (Yann LeCun)



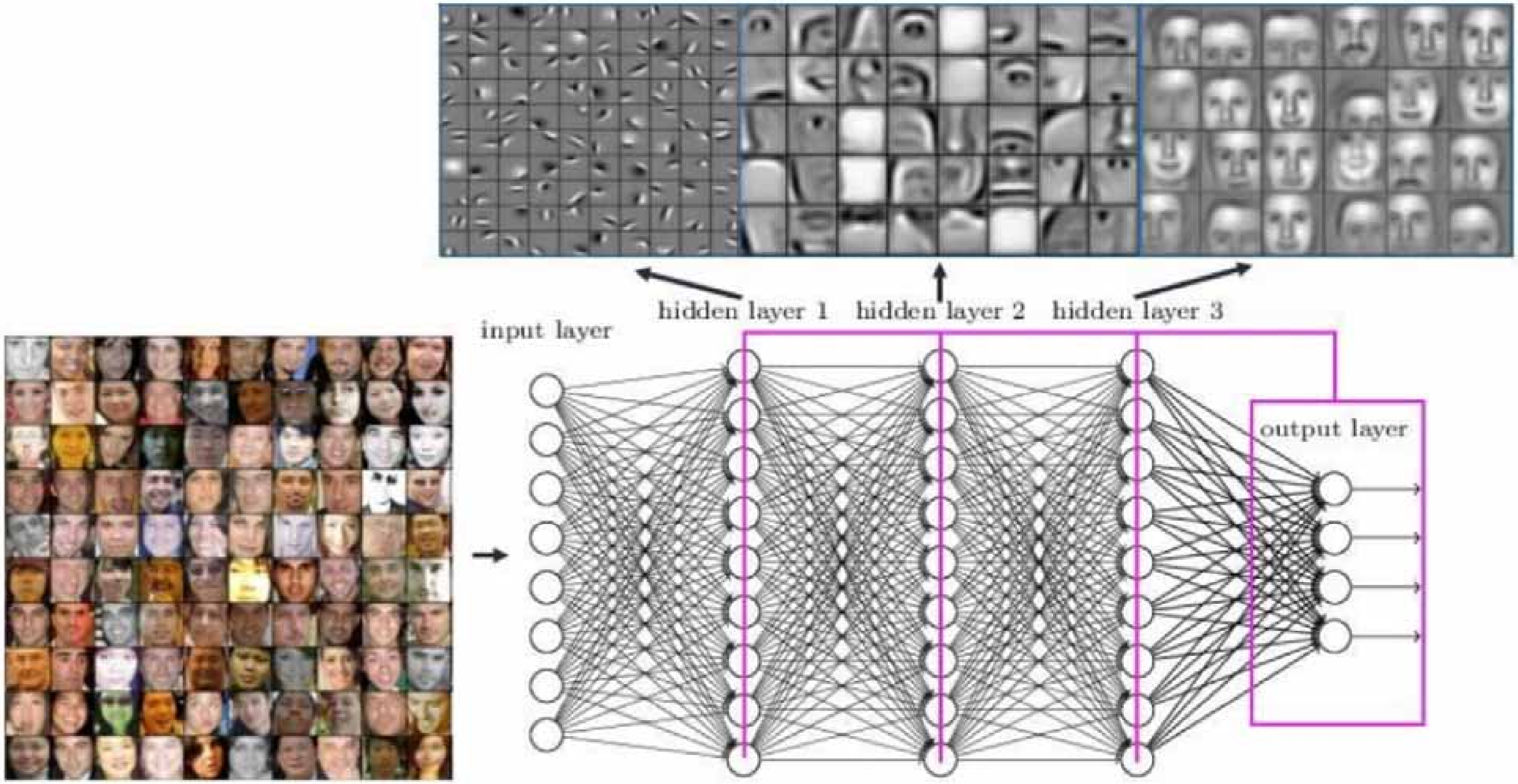

hidden layer 1
hidden layer 2
hidden layer 3
input layer
output layer


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


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

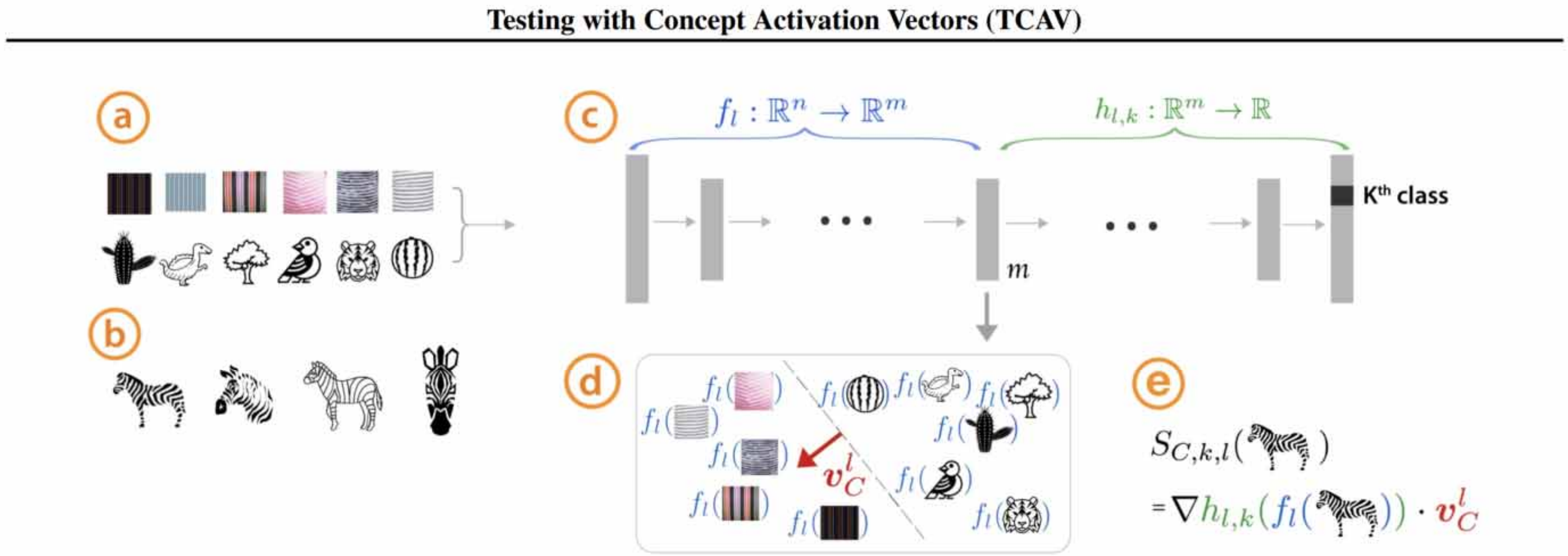
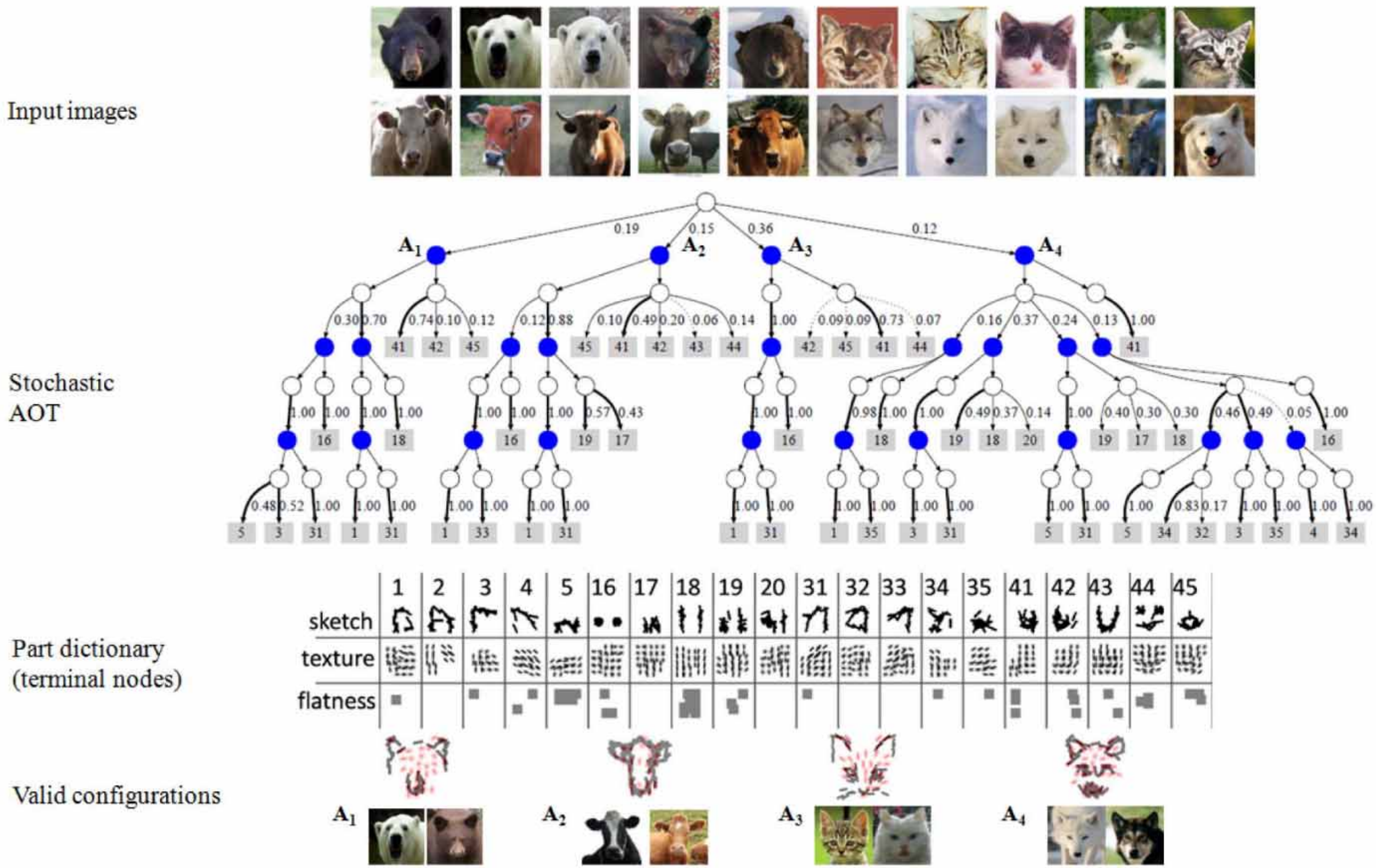


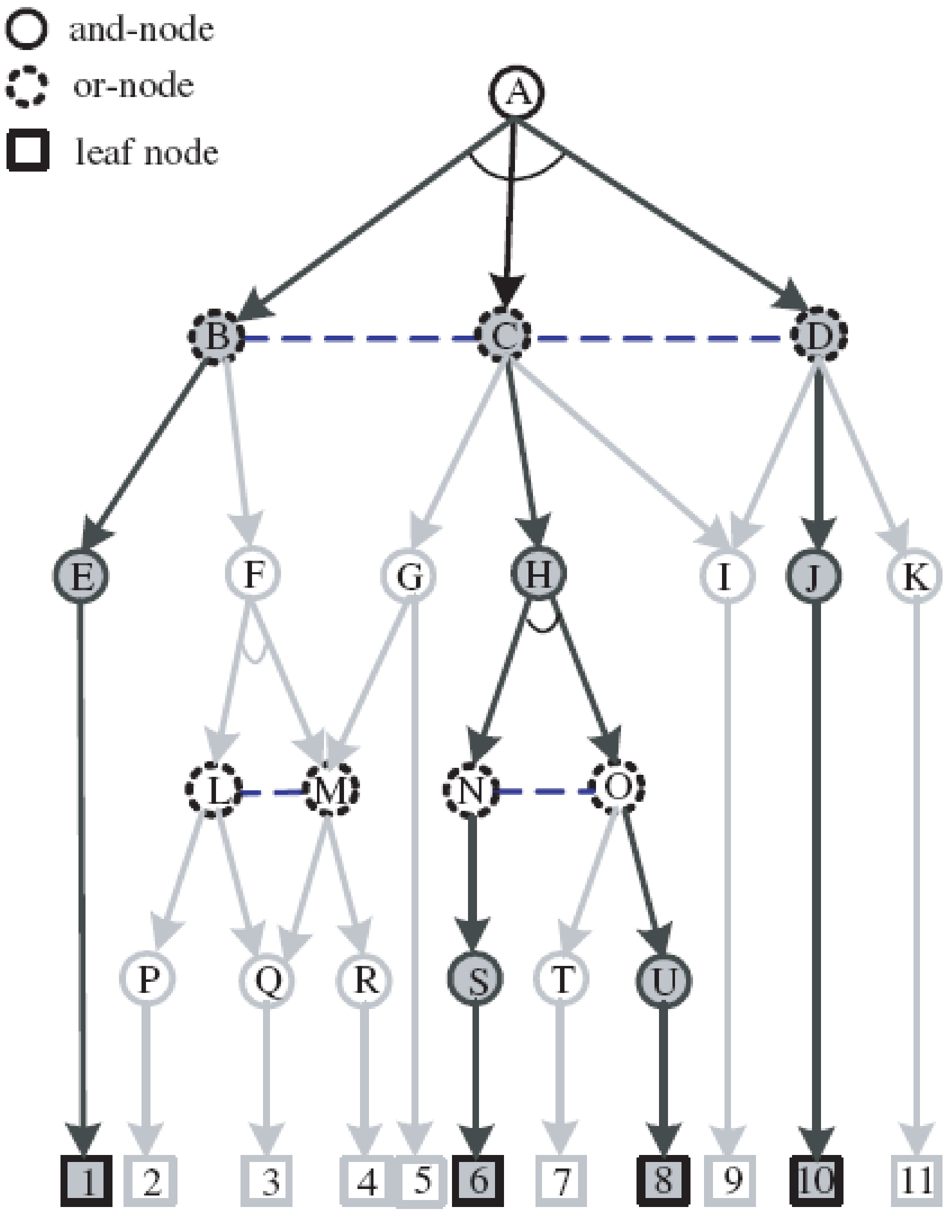
Figure 1. **Testing with Concept Activation Vectors:** Given a user-defined set of examples for a concept (e.g., 'striped'), and random examples ⓐ, labeled training-data examples for the studied class (zebras) ⓑ, and a trained network ⓒ, TCAV can quantify the model's sensitivity to the concept for that class. CAVs are learned by training a linear classifier to distinguish between the activations produced by a concept's examples and examples in any layer ⓓ. The CAV is the vector orthogonal to the classification boundary (v_C^l, red arrow). For the class of interest (zebras), TCAV uses the directional derivative $S_{C,k,l}(\boldsymbol{x})$ to quantify conceptual sensitivity ⓔ.

https://github.com/tensorflow/tcav

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler & Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). International Conference on Machine Learning (ICML), 2018. 2673-2682.



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**

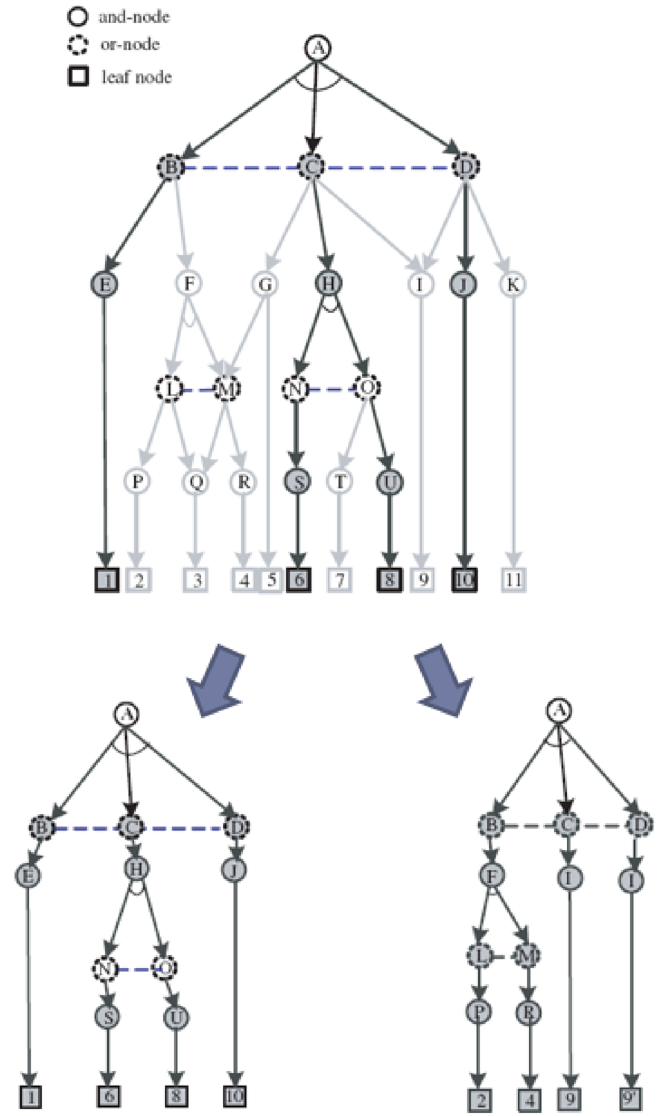
Images credit to Zhaoyin Jia (2009)

- 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: $\alpha(t)$

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

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

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

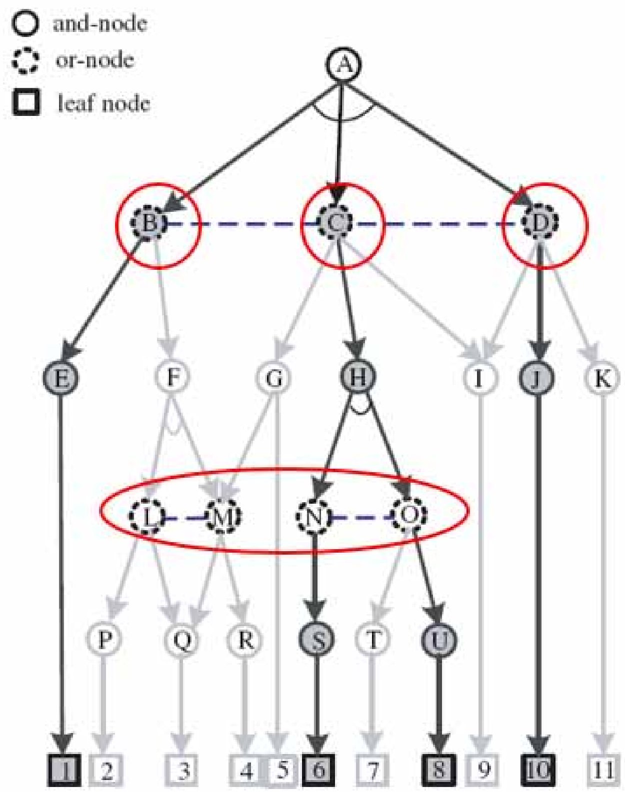


- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

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

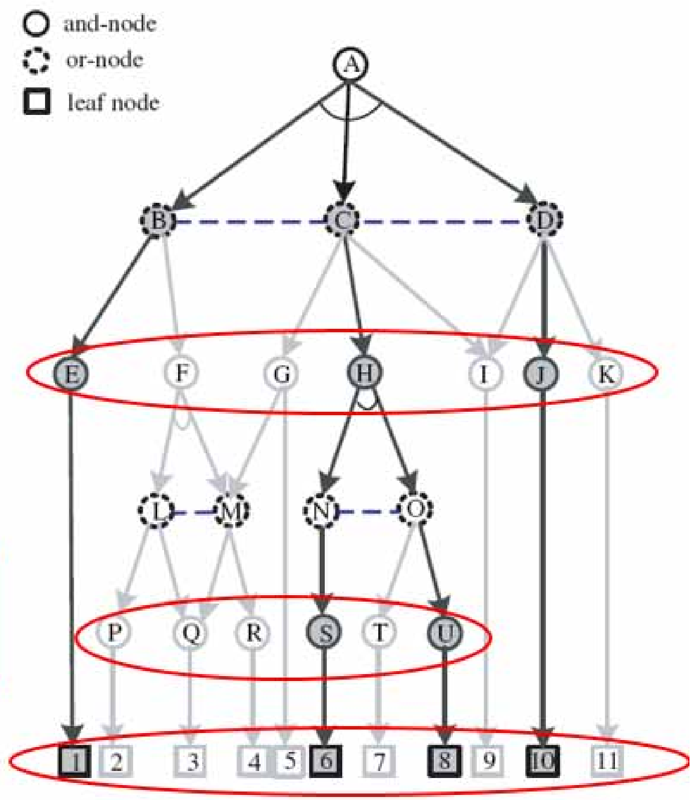
$$\xi(pg)=\sum_{v\in V^{Or}(pg)}\lambda_v(w(v))+\sum_{v\in V^{and}(pg)\cup T(pg)}\lambda_t(\alpha(t))$$

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



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: $\alpha(t)$



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

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

Weigh the local compatibility of primitives (geometric and appearance)

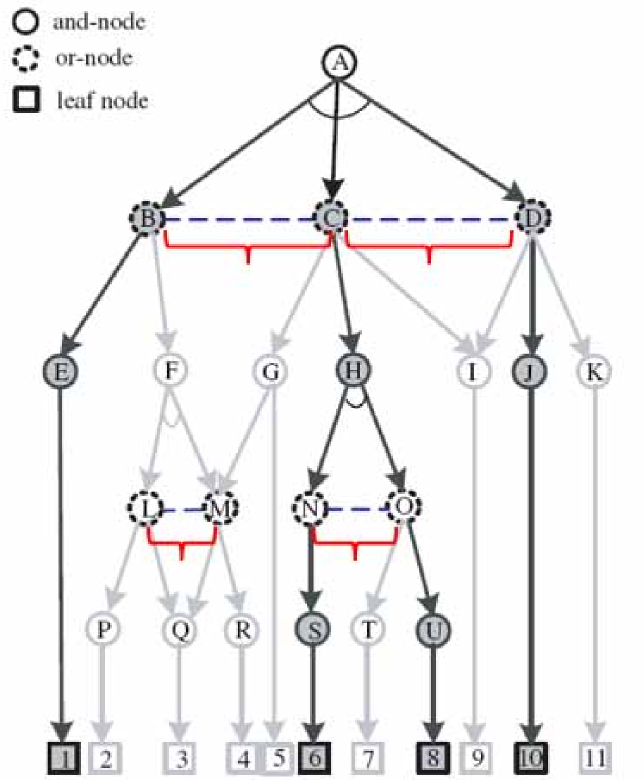
- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

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

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

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

Spatial and appearance between primitives (parts or objects)

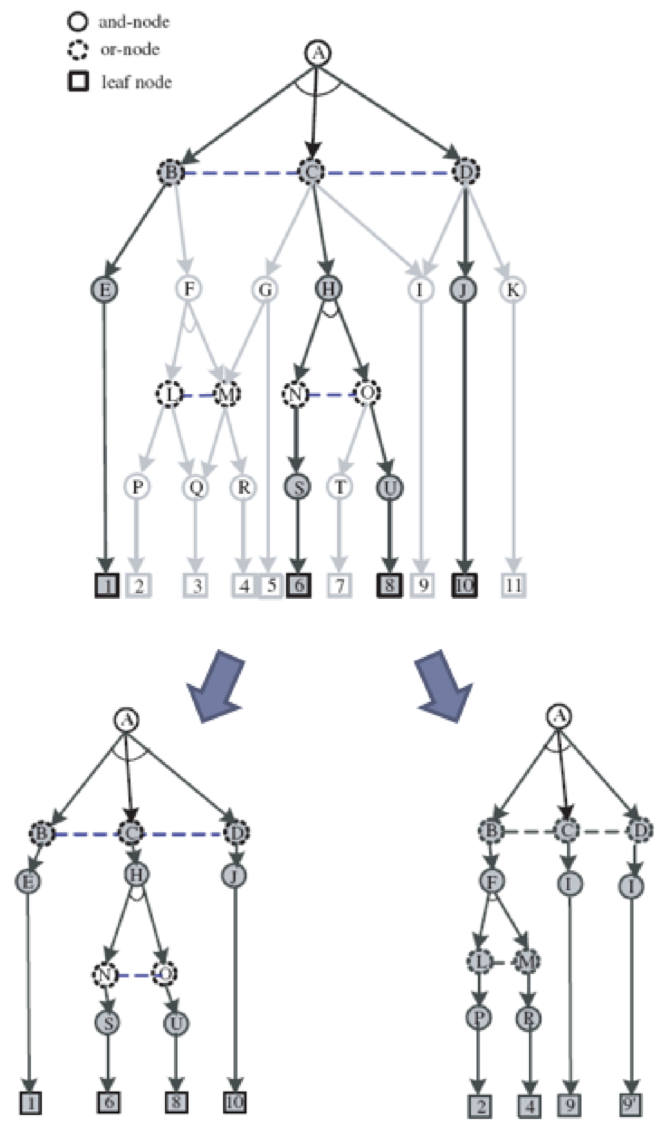


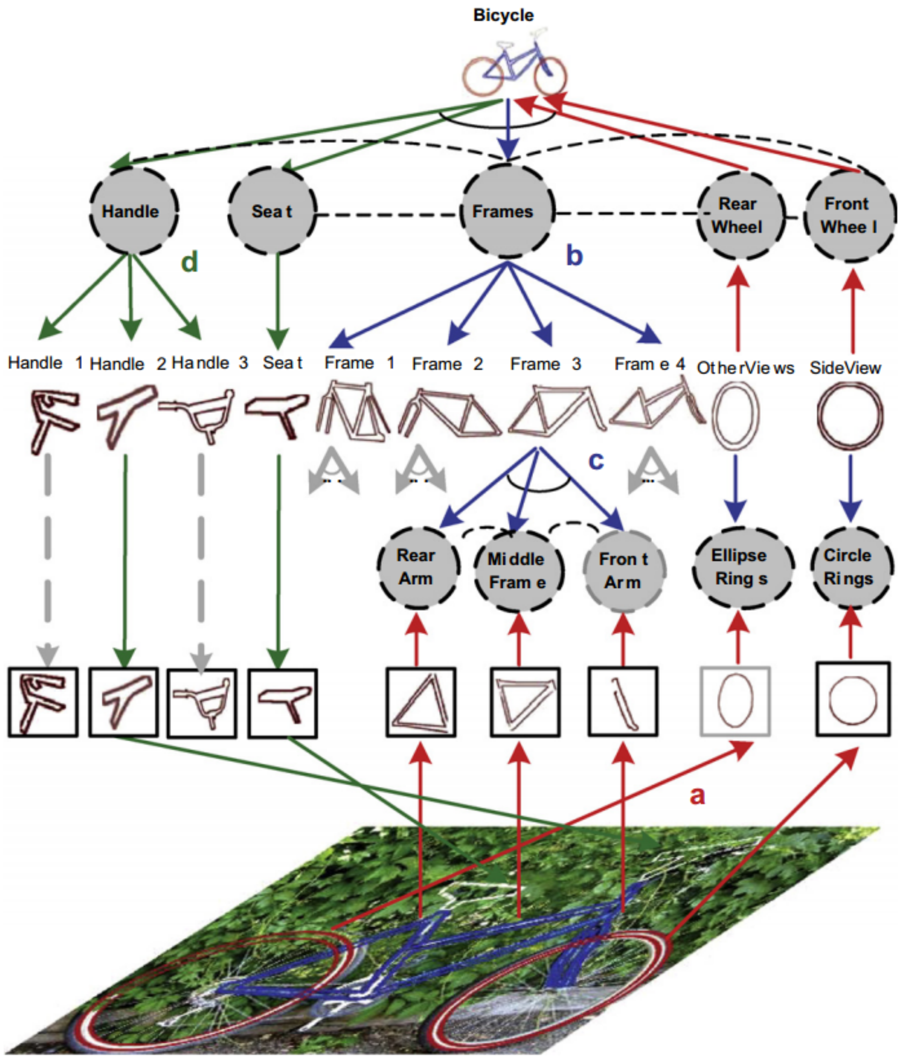
- Terminal (leaf) node: $T(pg)$
- And-Or node: $V^{or}(pg), V^{and}(pg)$
- Set of links: $E(pg)$
- Switch variable at Or-node: $w(t)$
- Attributes of primitives: $\alpha(t)$

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

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

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





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.

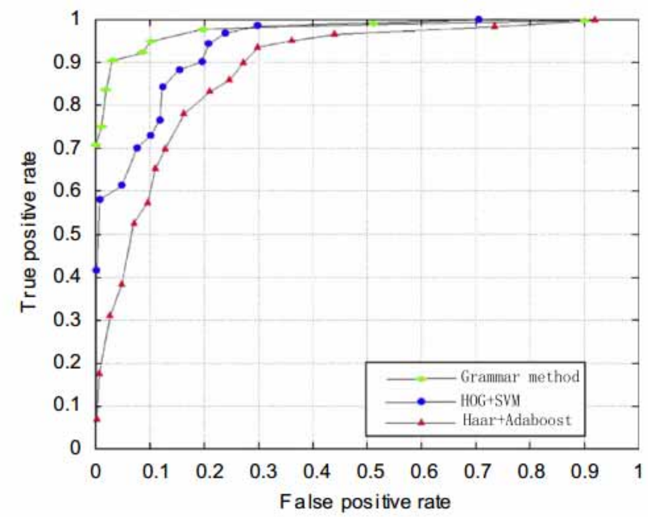
- Repeat the following steps

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

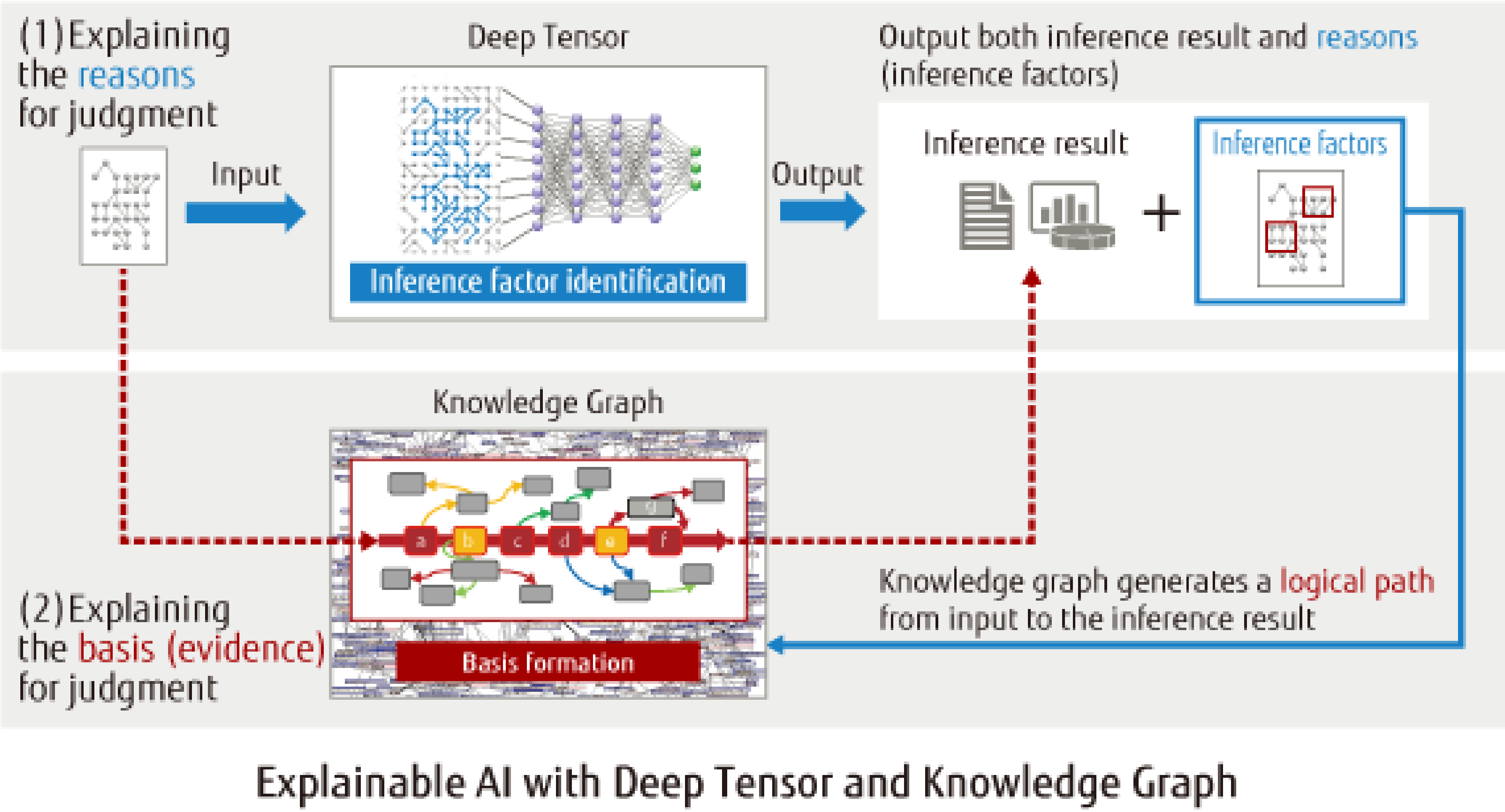
3 Call Top-down(A) to update A's **open** or **closed** lists.
 i Accept hypotheses from A's **open** list to its **closed** list.
 ii Remove (or disassemble) hypotheses from A's **closed** list.
 iii Update the **open** lists for particles that overlap with node A.

- 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



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 …

- Justification, Explanation and Causality
- Trust > scaffolded by justification of actions (why)
- Please always take into account the inherent uncertainty and incompleteness of medical data!

Alex John London 2019. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report, 49, (1), 15-21, doi:10.1002/hast.973.

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

arXiv:1805.11648v1 [cs.AI] 29 May 2018

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.


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 …

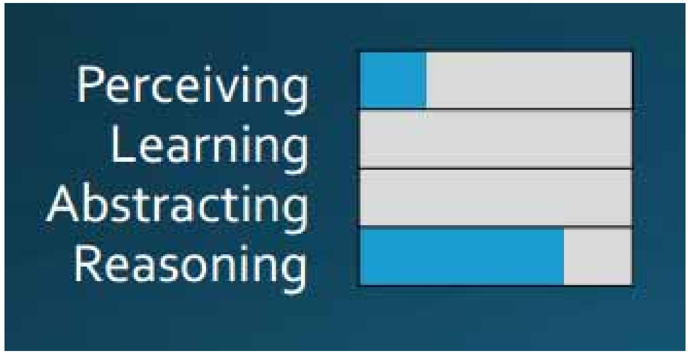


Image credit to John Launchbury

- 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 $p(x)$

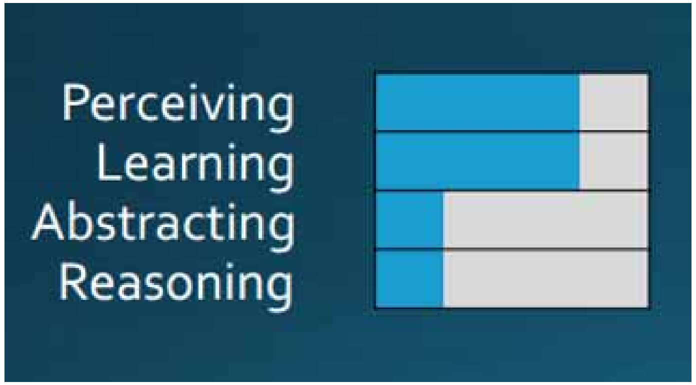


Image credit to John Launchbury

- 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

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

- 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 ...

Thank you!