KANDINSKY Patterns as IQ-Test for machine learning

Andreas Holzinger[1,3], Michael Kickmeier-Rust[2], Heimo Müller[1]

[1]Medical University Graz, Austria
[2]University of Teacher Education, St. Gallen, Switzerland
[3]Alberta Machine Intelligence Institute, Edmonton, Canada

@aholzin #KandinskyPatterns

In this paper we …

- 1) provide some background on testing intelligence,
- 2) report on preliminary results from 271 participants of our online study on explainability, and
- 3) introduce Kandinsky Patterns [1] as an IQ-Test for machines

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

“Solve intelligence – then solve everything else”

https://youtu.be/XAbLn66iHcQ?t=1h28m54s

Demis Hassabis, 22 May 2015

The Royal Society,
Future Directions of Machine Learning Part 2

Google DeepMind

To reach a level of usable intelligence we need to …

- 1) learn from prior data
- 2) extract knowledge
- 3) generalize, i.e. guessing where a probability mass function concentrates
- 4) fight the curse of dimensionality
- 5) disentangle **underlying explanatory factors of data**, i.e.
- 6) **understand** the data in the **context** of an application domain

Why is measuring intelligence important?

"If you can't measure it,

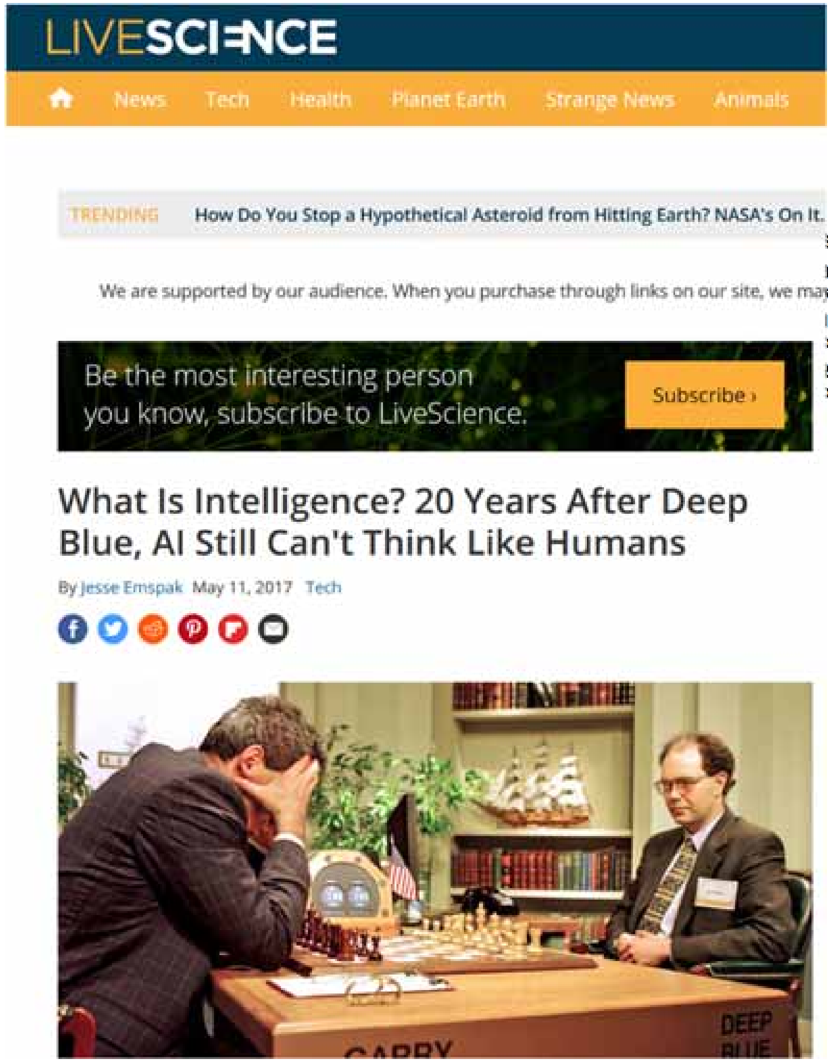
nor assign it an exact numerical value, nor express it in numbers,

then your knowledge is of a meager and unsatisfactory kind"

(attributed to William Thomson (1824-1907), aka Lord Kelvin)

Intelligence is not just about pattern recognition! Are we measuring the right things in AI?

Success-sample 1: IBM Deep Blue vs. Garry Kasparov



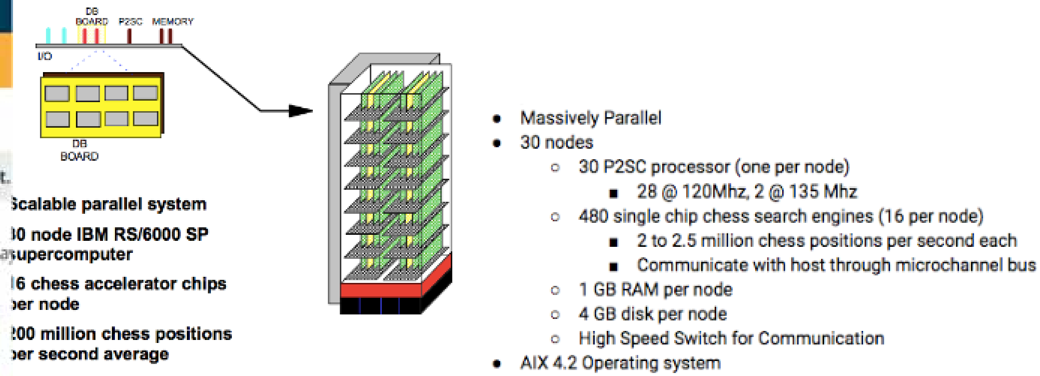


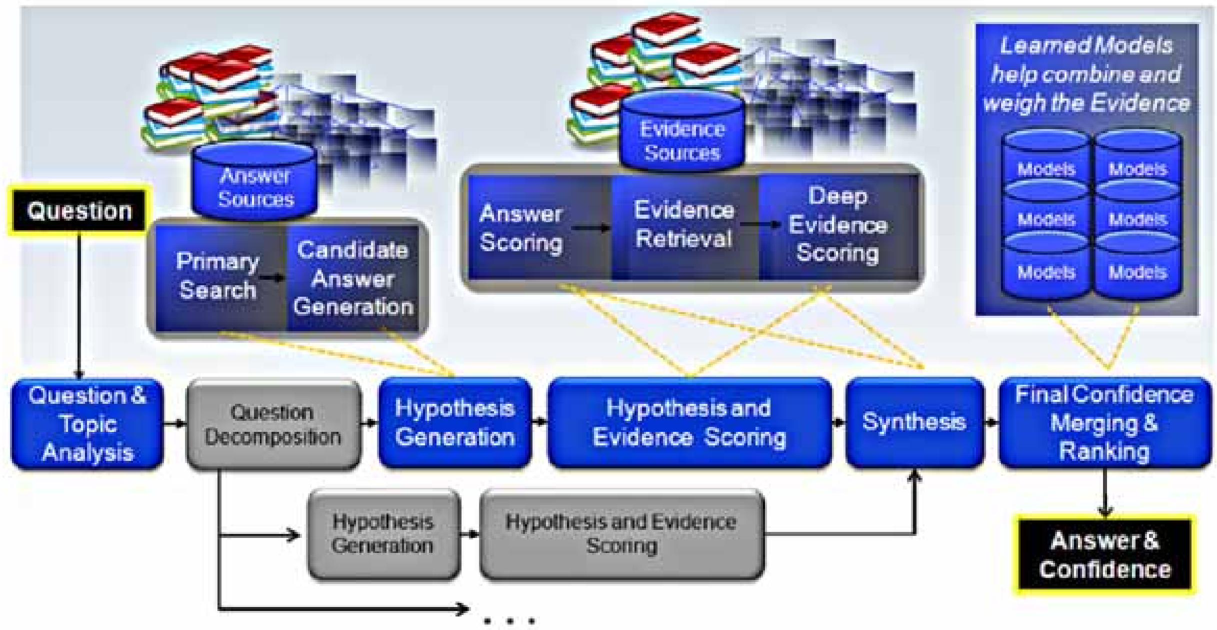
Image Source: https://hackernoon.com/hn-images/0*EsGv18e_FYOlQ_VI.jpg

https://www.livescience.com/59068-deep-blue-beats-kasparov-progress-of-ai.html

Murray Campbell, A. Joseph Hoane Jr & Feng-Hsiung Hsu 2002. Deep blue. Artificial intelligence, 134, (1-2), 57-83, doi:10.1016/S0004-3702(01)00129-1.

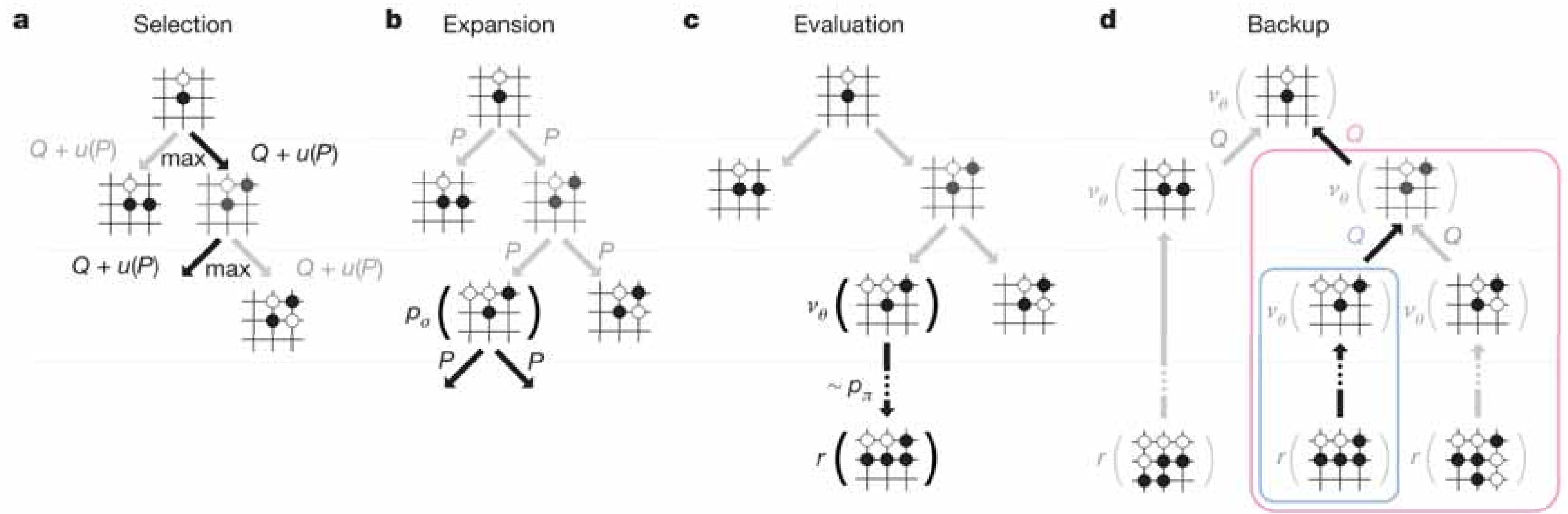
Success-sample 2: IBM Watson vs. Rutter & Jennings

https://img.purch.com/jeopardy-watson-ibm/w/630/aHR0cDovL21lZGlhLmJlc3RvZm1pY3JvLmNvbS8zL0IvMjgxMzUxL29yaWdpbmFsL1dhdHNvbi1hdS1KZW9wYXJkeS1fLTR1Z2Vlay5qcGc=



David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek & Erik T Mueller 2013. Watson: beyond jeopardy! Artificial Intelligence, 199, 93-105, doi:10.1016/j.artint.2012.06.009.

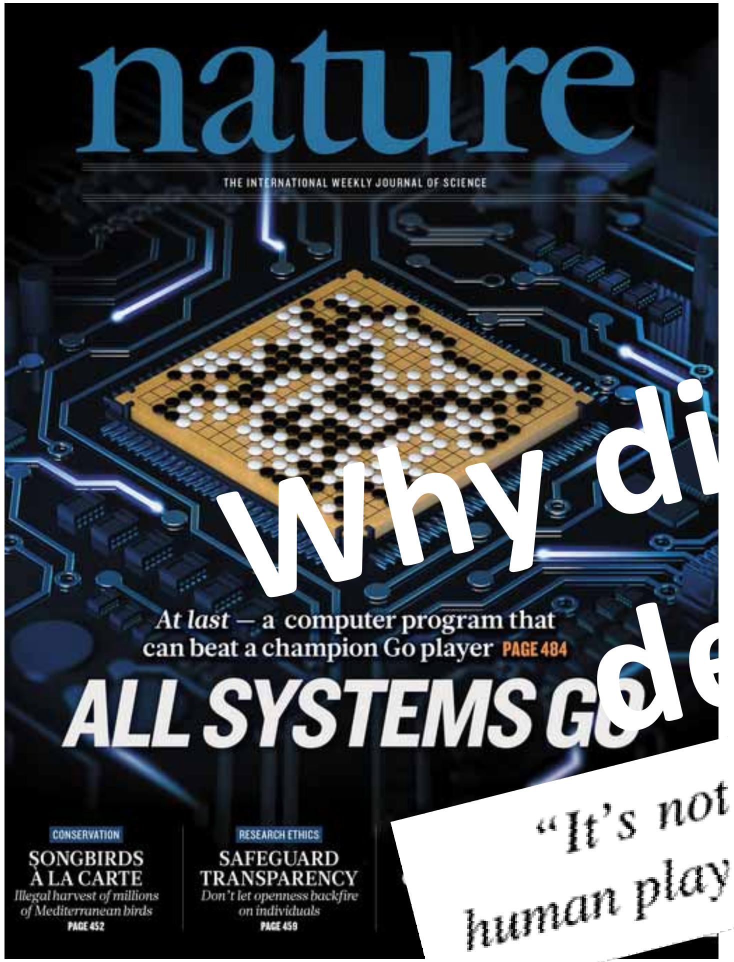
Success 3: Deep Mind's AlphaGo vs. Hui & Sedol



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

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.

Correlation is not Causation



Why did it make this decision???

"It's not a human move. I've never seen a human play this move." (Fan Hui, 2016).

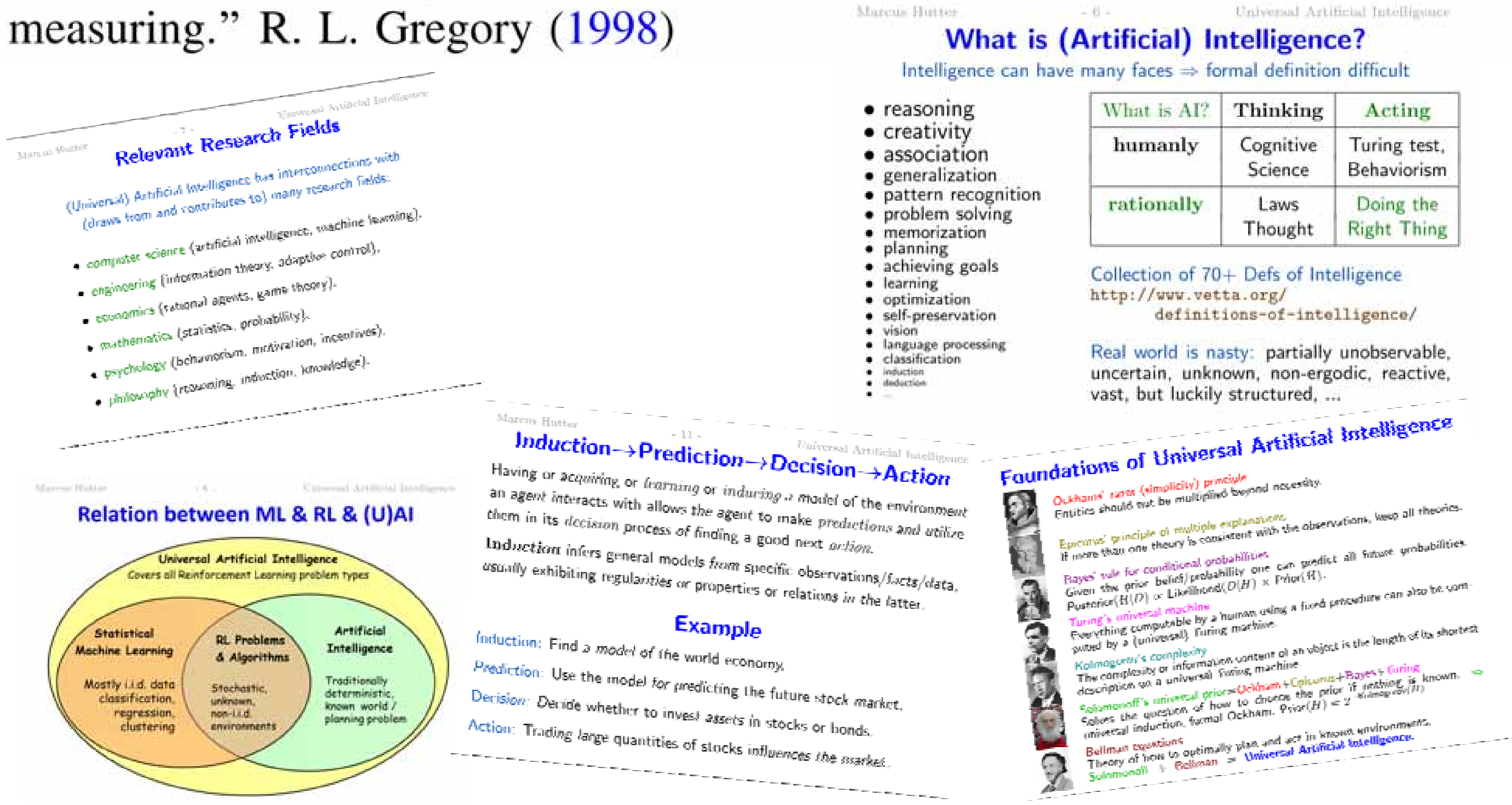
...ilver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, 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.

Can human intelligence be measured as a very special case of machine intelligence?

Tarek Besold, José Hernández-Orallo & Ute Schmid 2015. Can machine intelligence be measured in the same way as human intelligence? KI-Künstliche Intelligenz, 29, (3), 291-297, doi:10.1007/s13218-015-0361-4.

What is (artificial) intelligence?

"Innumerable tests are available for measuring intelligence, yet no one is quite certain of what intelligence is, or even just what it is that the available tests are measuring." R. L. Gregory (1998)



Shane Legg & Marcus Hutter 2007. Universal intelligence: A definition of machine intelligence. Minds and machines, 17, (4), 391-444, doi:10.1007/s11023-007-9079-x.

Alan Turing (1950)

Vol. LIX. No. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. *The Imitation Game.*

I propose to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair?

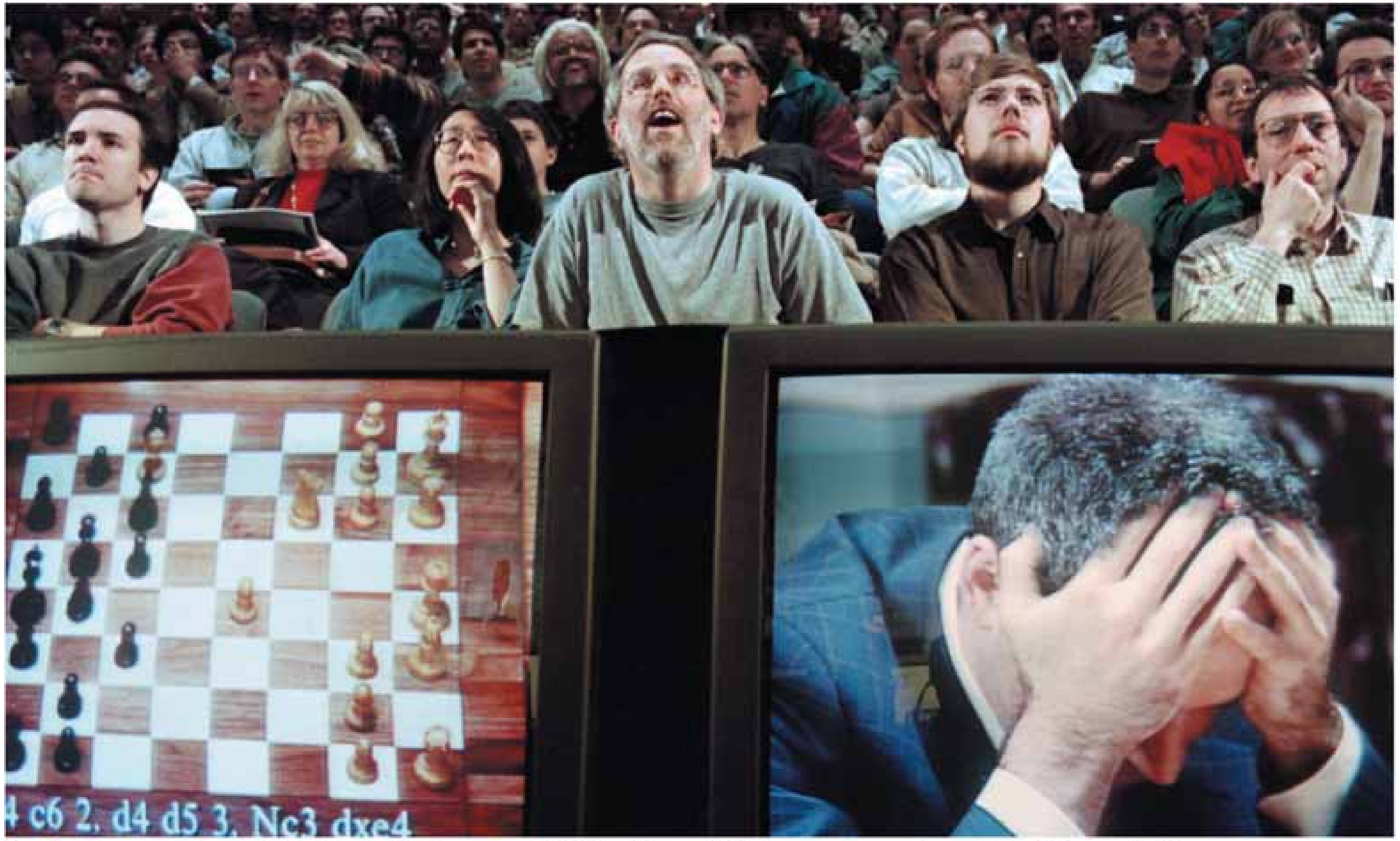
Now suppose X is actually A, then A must answer. It is A's

28 433

- **An algorithm can be considered intelligent for a certain task - if it can solve this task.**
- **Task-centric approach**
- **requires a-priori knowledge**

Humans are not centered on only one single task

Demis Hassabis 2017. Artificial Intelligence: Chess match of the century. Nature, 544, 413, doi:10.1038/544413a.



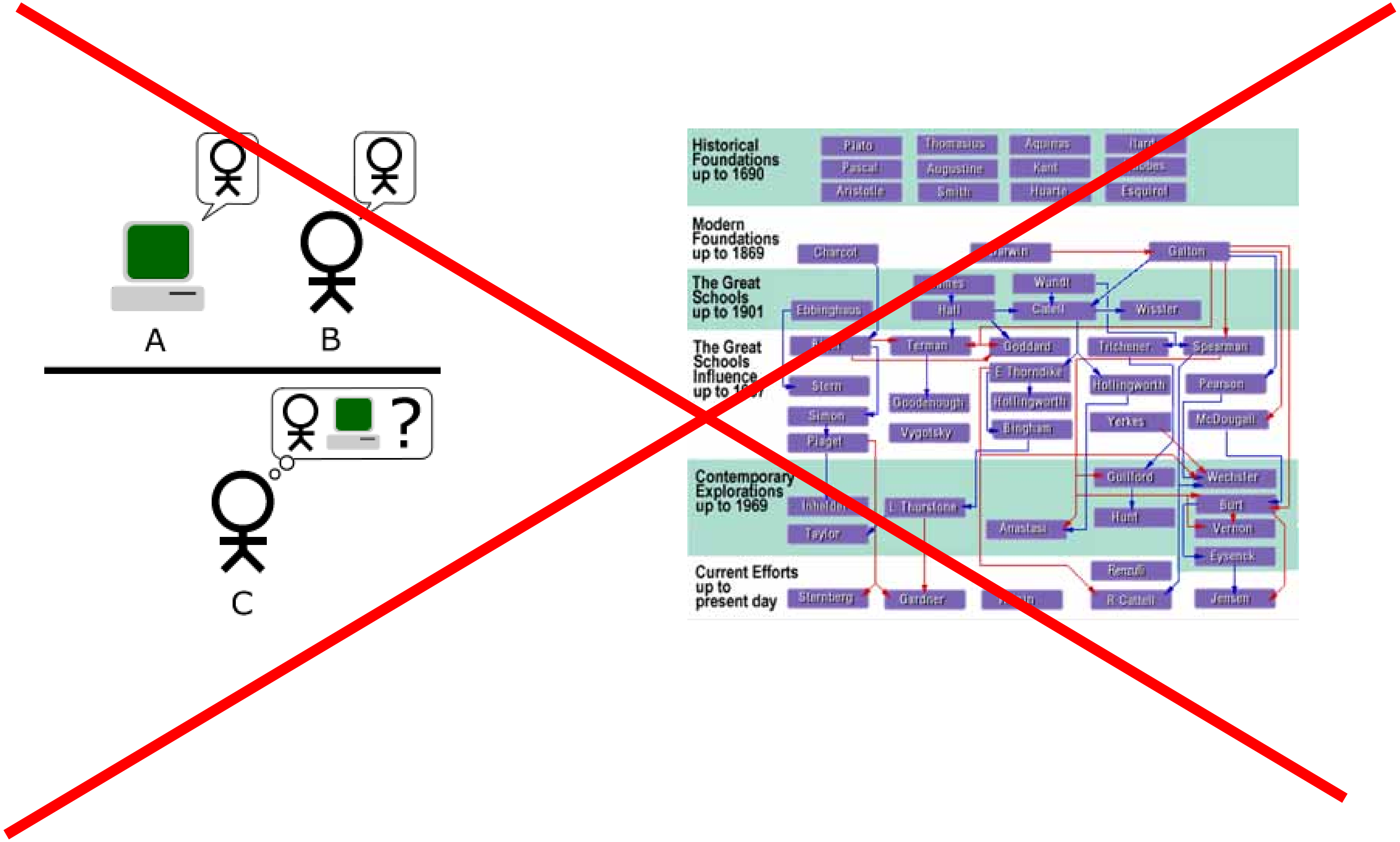
Grandmaster Garry Kasparov during the last of six games against Deep Blue in 1997; the computer won the match by 3.5 games to 2.5.

ARTIFICIAL INTELLIGENCE

Chess match of the century

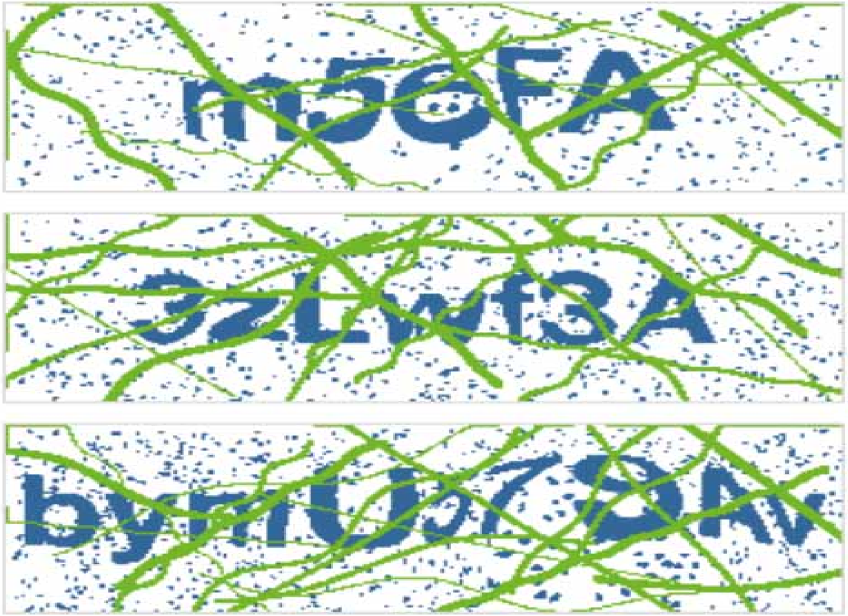
Demis Hassabis lauds Garry Kasparov's account of battling supercomputer Deep Blue.

Turing test is no IQ-Test and IQ-Tests are not for AI



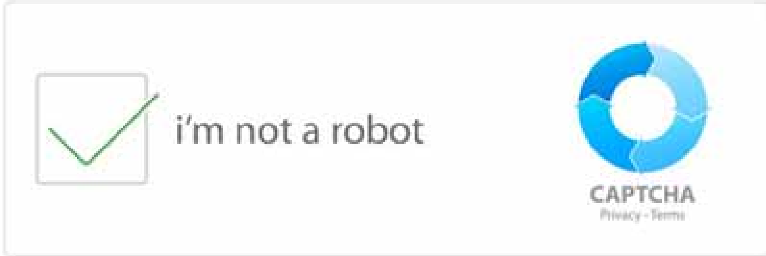
David L. Dowe & José Hernández-Orallo 2012. IQ tests are not for machines, yet. Intelligence, 40, (2), 77-81, doi:10.1016/j.intell.2011.12.001.

By-product: AI Problems as Security Primitives



$$\min_m \left\{ \exists k : \sum_{i=k+1}^{m} \binom{m}{i} \beta^i (1-\beta)^{m-i} \geq 1-\epsilon \;\&\; \sum_{i=0}^{k} \binom{m}{i} \eta^i (1-\eta)^{m-i} \leq \epsilon \right\}$$

Luis Von Ahn, Manuel Blum, Nicholas J. Hopper & John Langford 2003. CAPTCHA: Using hard AI problems for security. In: Bilham, E. (ed.) Advances in Cryptology-Eurocrypt 2003. Berlin: Springer-Verlag Berlin, pp. 294-311, doi:10.1007/3-540-39200-9_18.



Marios Belk, Panagiotis Germanakos, Christos Fidas, Andreas Holzinger & George Samaras 2013. Towards the Personalization of CAPTCHA Mechanisms Based on Individual Differences in Cognitive Processing. Human Factors in Computing and Informatics, Lecture Notes in Computer Science, LNCS 7946. Berlin Heidelberg: Springer, pp. 409-426, doi:10.1007/978-3-642-39062-3_26.

SCIENCE ROBOTICS | FOCUS

ARTIFICIAL INTELLIGENCE

Parallel testing of vehicle intelligence via virtual-real interaction

Li Li[1]*, Xiao Wang[2,3]*, Kunfeng Wang[2,3]*, Yilun Lin[2,3,4]*, Jingmin Xin[5]*, Long Chen[6,7], Linhai Xu[5], Bin Tian[2,7], Yunfeng Ai[4,7], Jian Wang[7,8], Dongpu Cao[7,8,9], Yuehu Liu[5], Chenghong Wang[10,11], Nanning Zheng[5†], Fei-Yue Wang[2†]

A self-driven closed-loop parallel testing system implements more challenging tests to accelerate evaluation and development of autonomous vehicles.

Although researchers and automobile manufacturers have built several proving grounds (*1*) and testing datasets (*2*) dedicated to autonomous driving, tests for intelligent vehicles remain time-consuming, inefficient, and sometimes dangerous for people who use the same roads.

According to Turing (*3*), a system could be said to be intelligent enough for special narios tested by simulations should also be re-evaluated and verified in field tests to validate the effectiveness of the simulation systems and the reliability of the hardware of autonomous vehicles.

Therefore, a human-in-loop simulation system is useful to evaluate the performance of vehicles (*5*–*7*) efficiently. A human expert can first vaguely define the tasks and per-

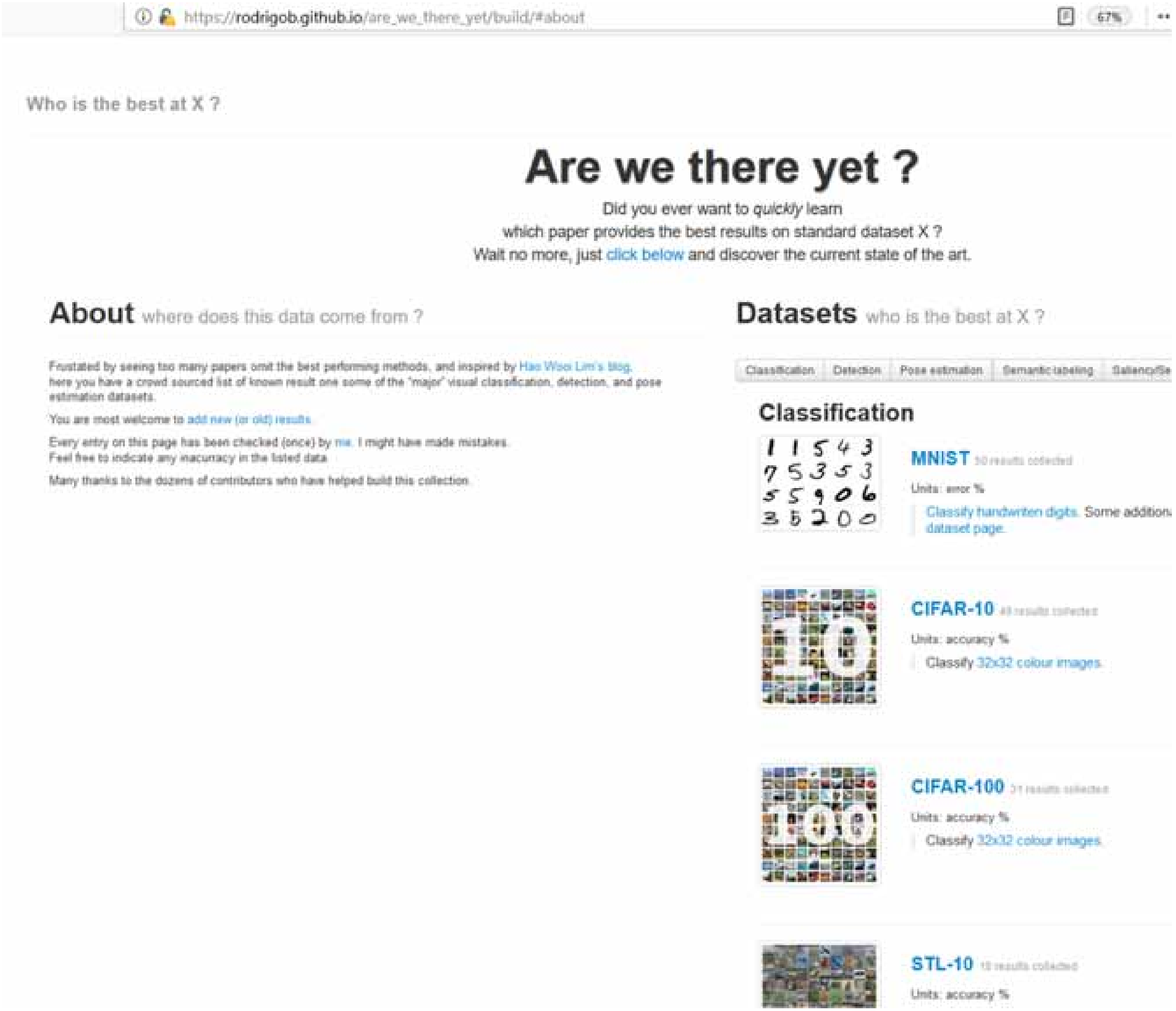
We can rearrange the spatiotemporal ranges of task rectangles to sample different driving scenarios that belong to the same category so as to ensure that autonomous vehicles could work for these driving scenarios. Adding more semantic task atoms over time permits us to increase the complexity of tests and expand the range of scenarios in which autonomous vehicles can be

Li Li, Xiao Wang, Kunfeng Wang, Yilun Lin, Jingmin Xin, Long Chen, Linhai Xu, Bin Tian, Yunfeng Ai & Jian Wang 2019. Parallel testing of vehicle intelligence via virtual-real interaction. Science Robotics, 4, (eaaw4106), 1-3.

- Human intelligence was not seen as the capability to solve one particular task, such as a pure classification task, it was considered being a much wider construct.
- Moreover, human intelligence generally was not measured in an isolated way but always in relation to an underlying population.
- By the example of the self-driving cars, the question would be whether one car can drive better against all the other cars, or even whether and to what extent the car does better than human drivers.

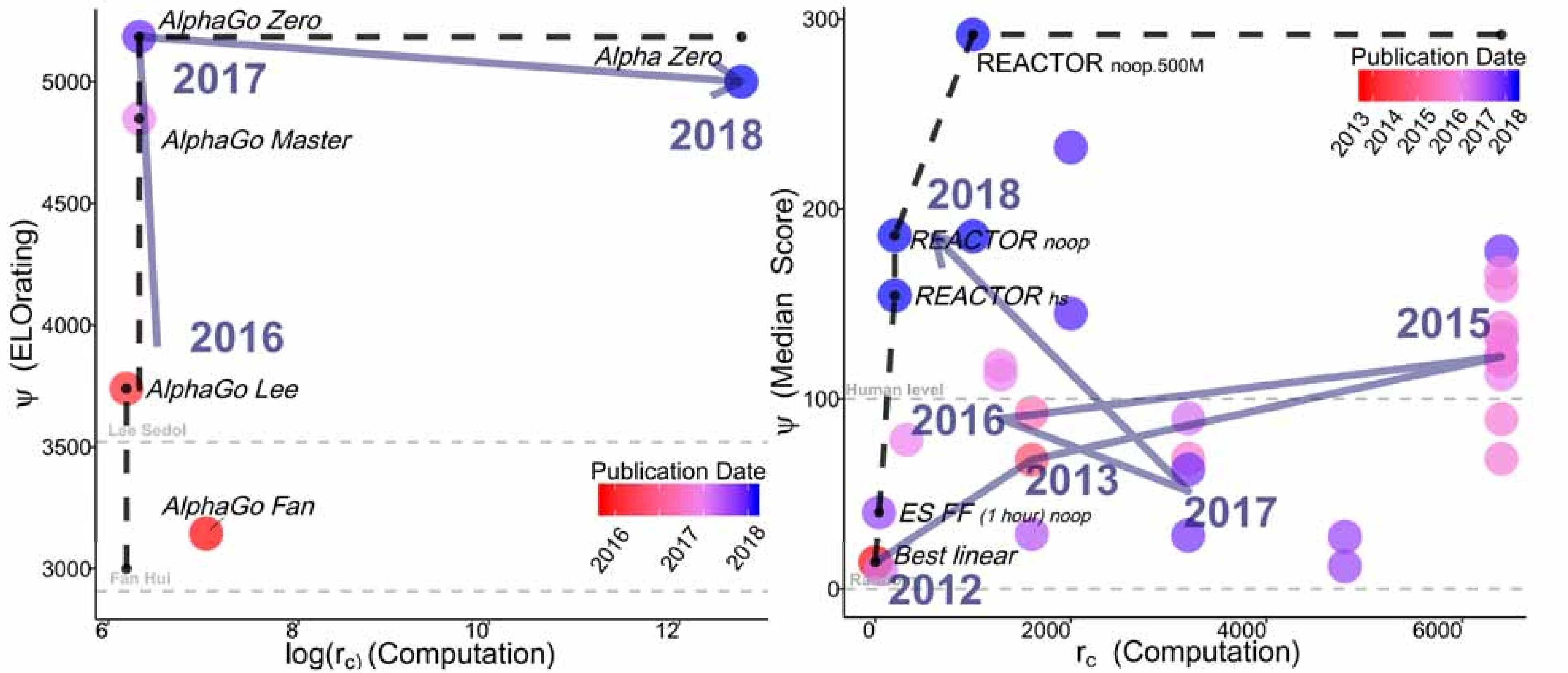
John Raven 2000. The Raven's progressive matrices: change and stability over culture and time. Cognitive psychology, 41, (1), 1-48, doi:10.1006/cogp.1999.0735.

CD-MAKE All these benchmarks have nothing to do with IQ



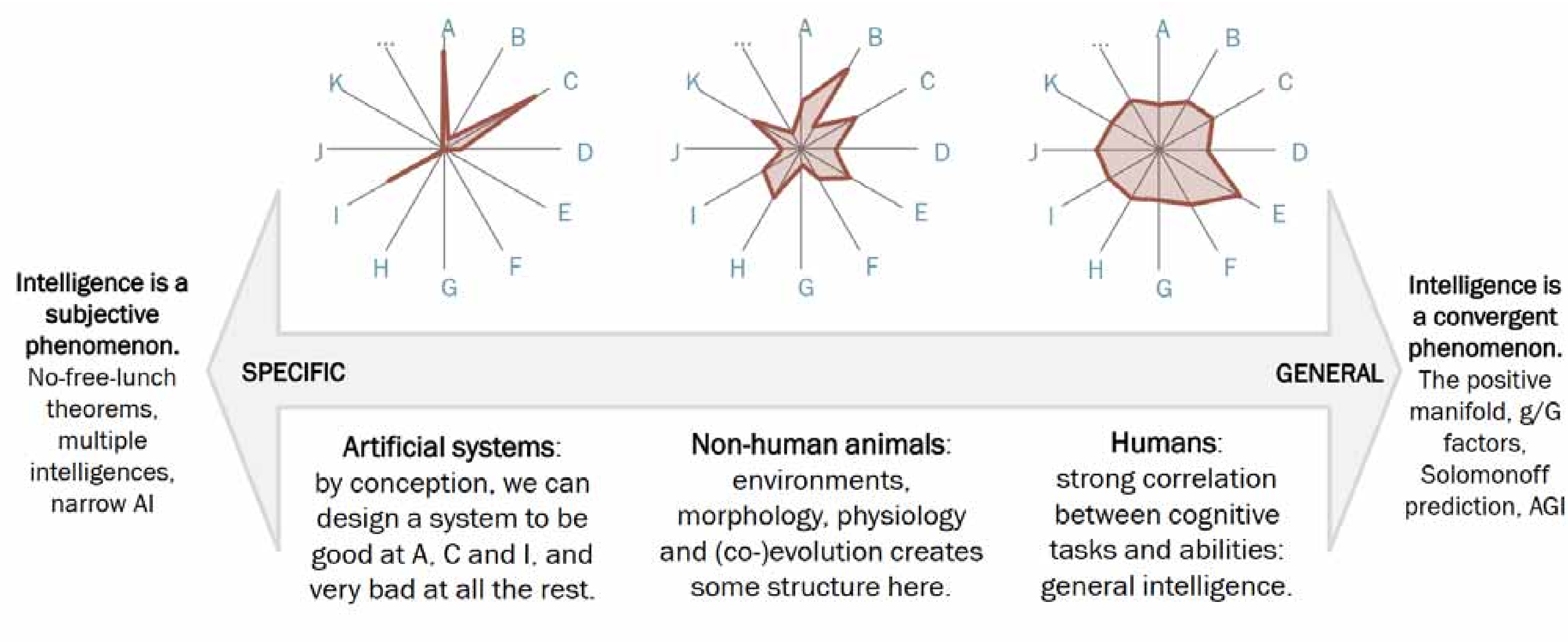
https://www.eff.org/ai/metrics

Multidim utility space for Alpha & ALE (Google)



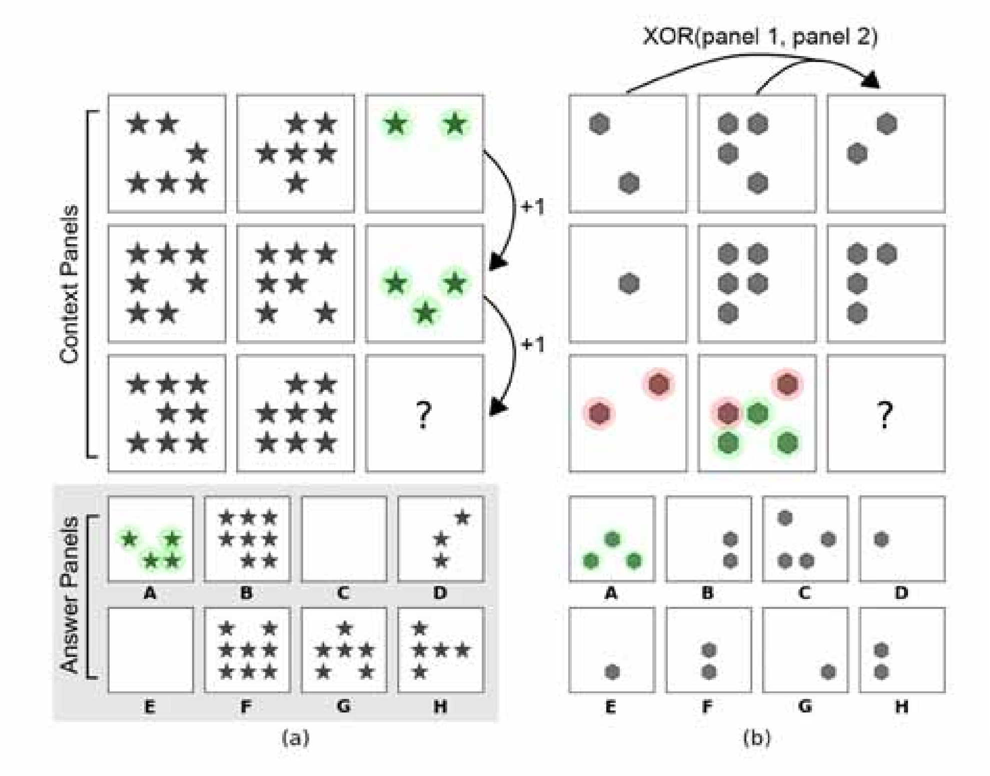
Fernando Martínez-Plumed, Shahar Avin, Miles Brundage, Allan Dafoe, Sean Ó Héigeartaigh & José Hernández-Orallo 2018. Accounting for the neglected dimensions of ai progress. arXiv preprint arXiv:1806.00610.

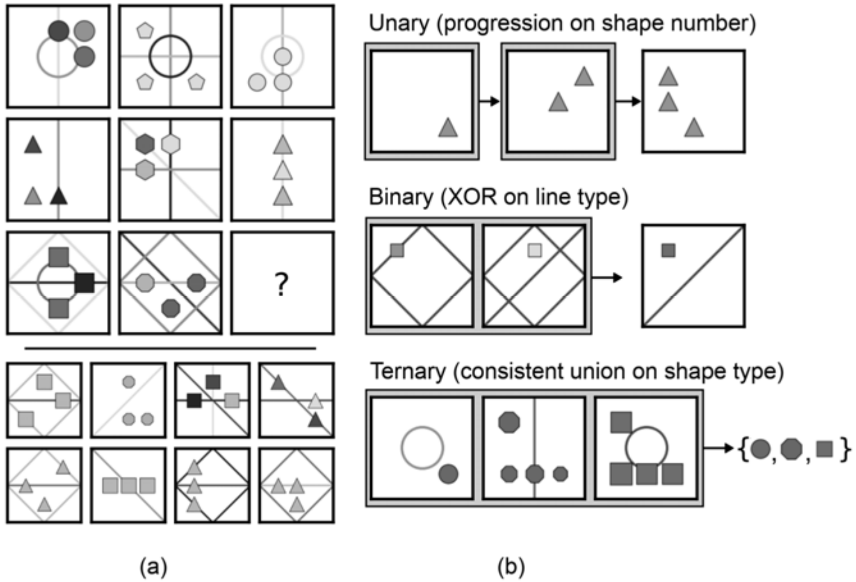
Intelligence = subjective, no clear definitions …



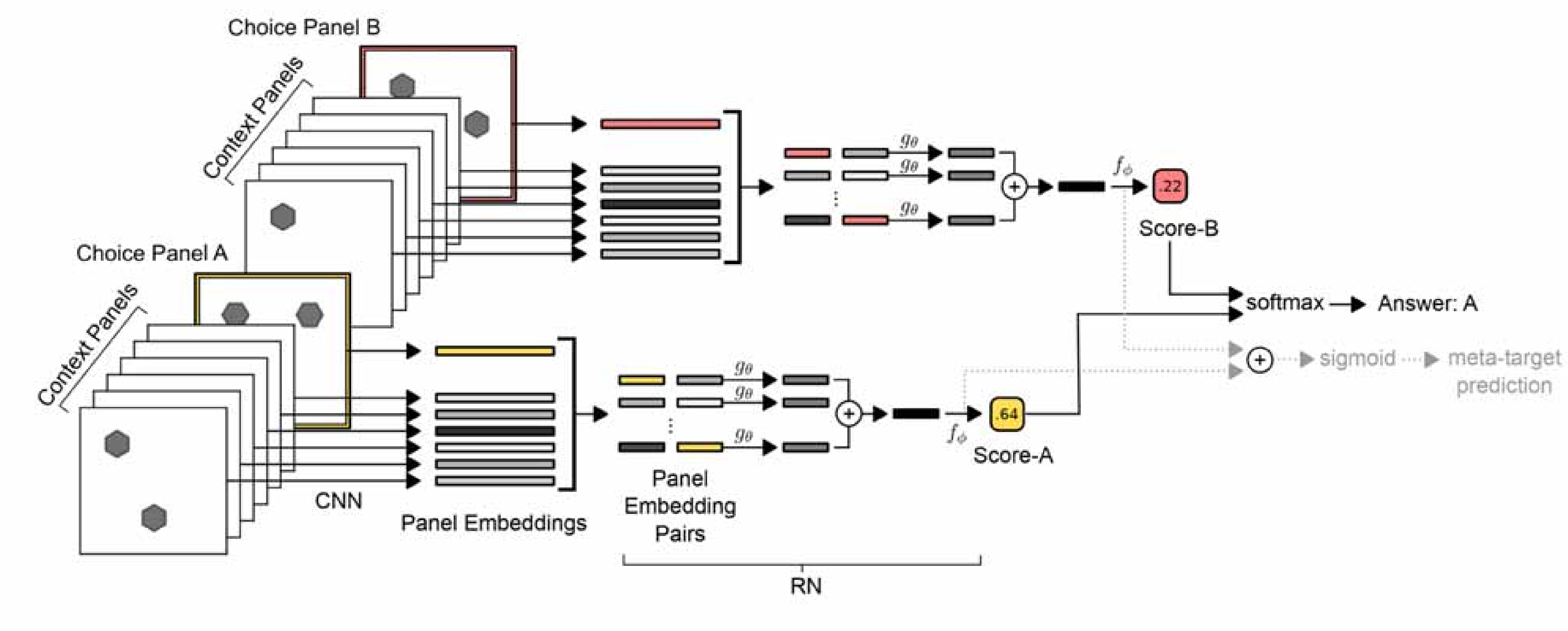
José Hernández-Orallo 2017. The measure of all minds: evaluating natural and artificial intelligence, Cambridge University Press, doi:10.1017/9781316594179.

Raven-style Progressive Matrices



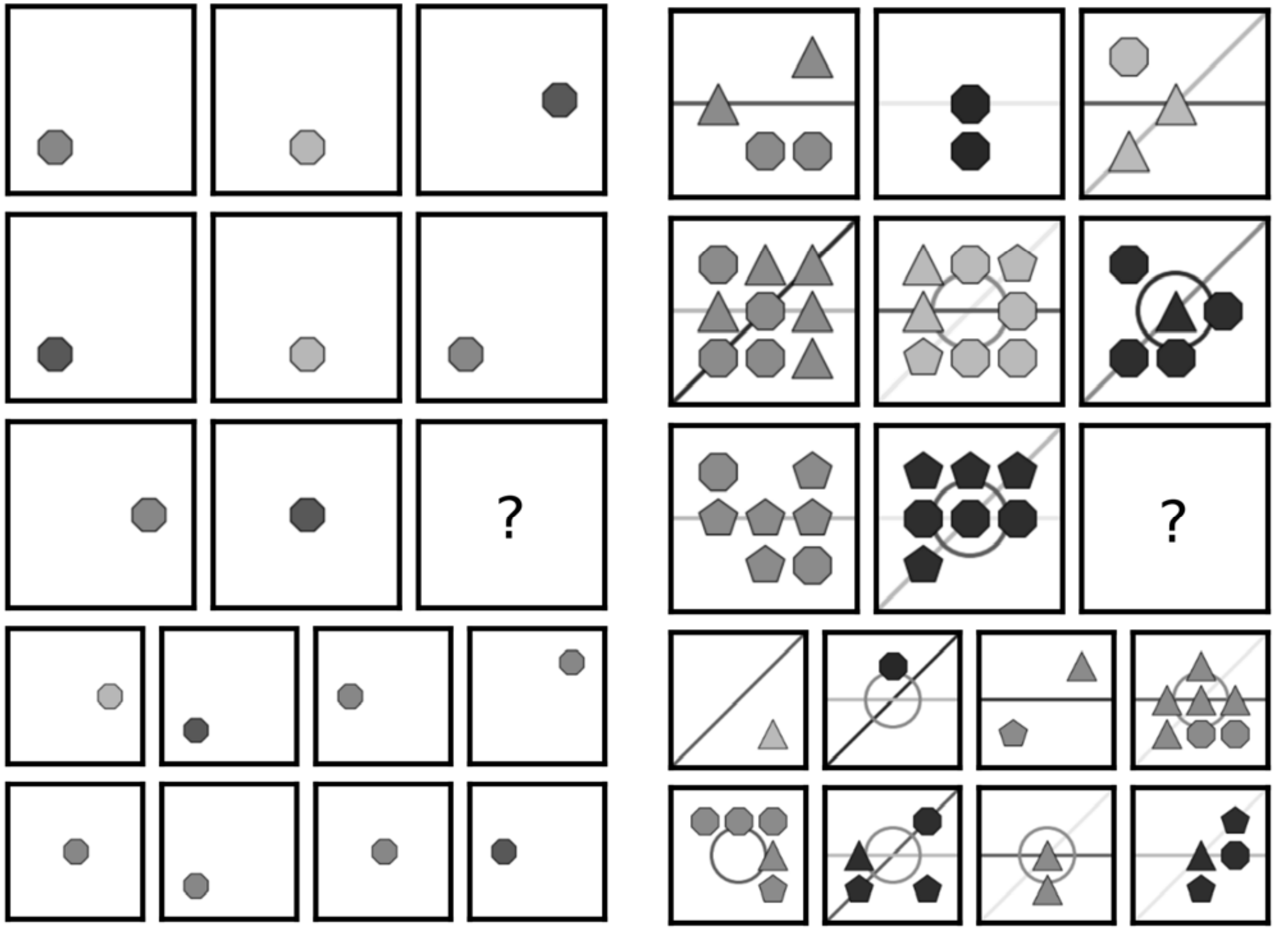


Adam Santoro, Felix Hill, David Barrett, Ari Morcos & Timothy Lillicrap 2018. Measuring abstract reasoning in neural networks. 35th International Conference on Machine Learning. Stockholm: PMLR, pp. 4477-4486.



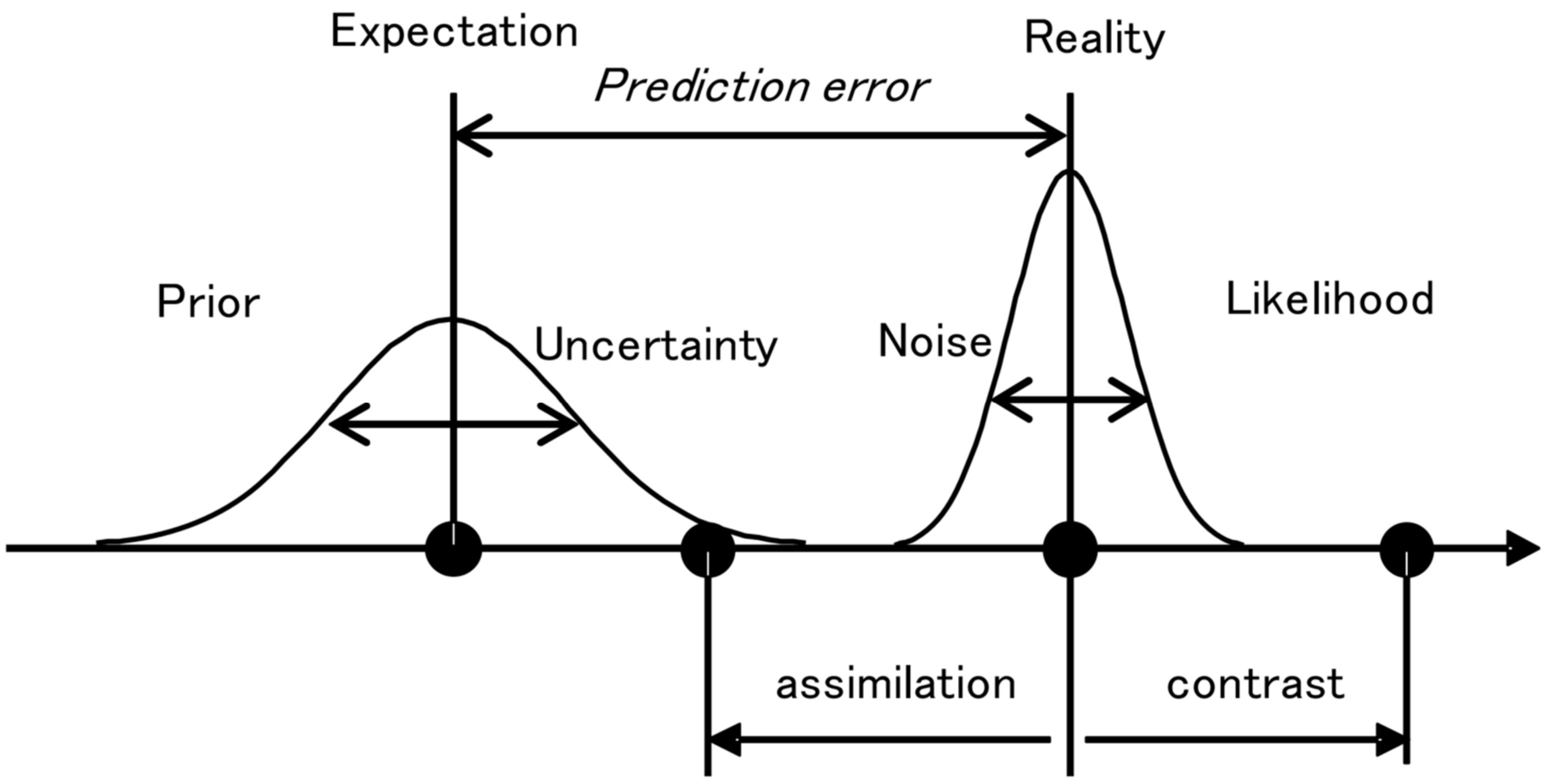
Adam Santoro, Felix Hill, David Barrett, Ari Morcos & Timothy Lillicrap 2018. Measuring abstract reasoning in neural networks. 35th International Conference on Machine Learning. Stockholm: PMLR, pp. 4477-4486.

The effect of distraction

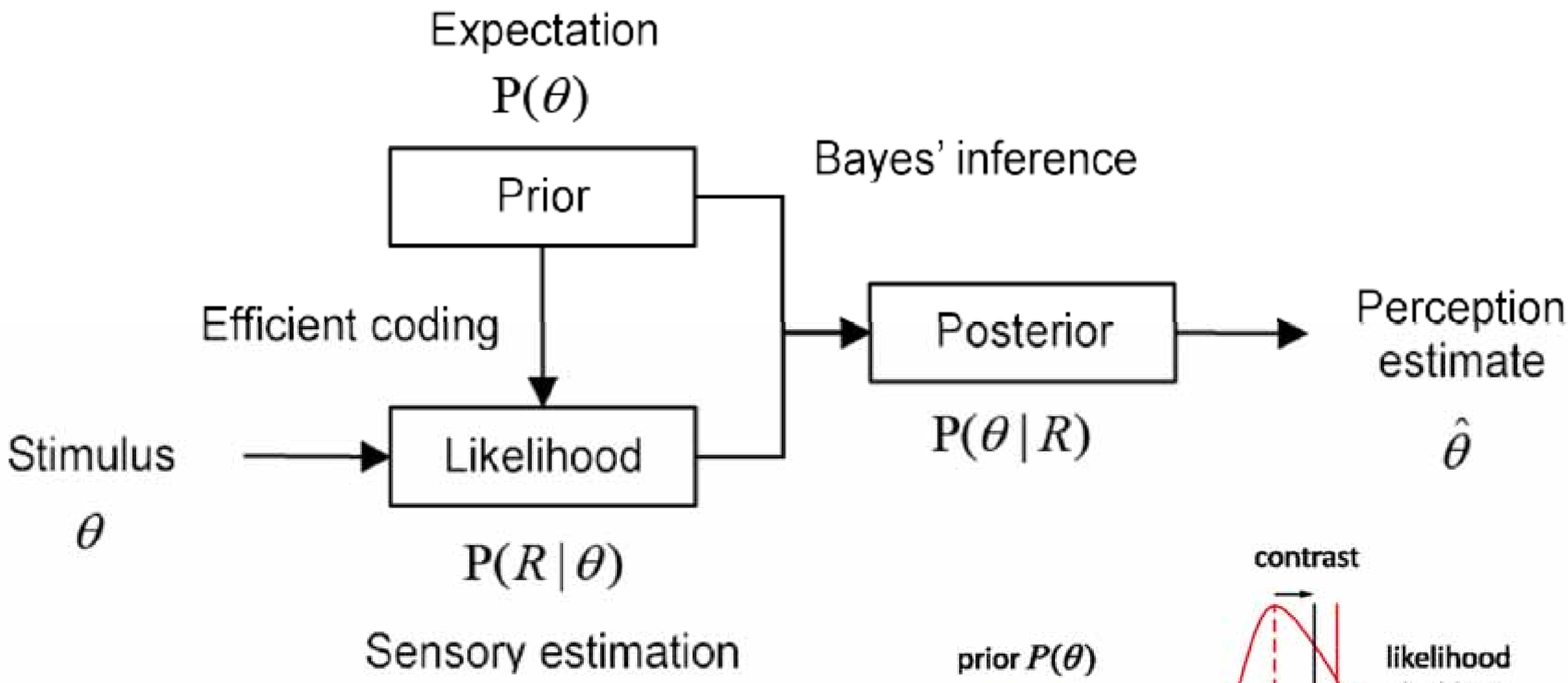


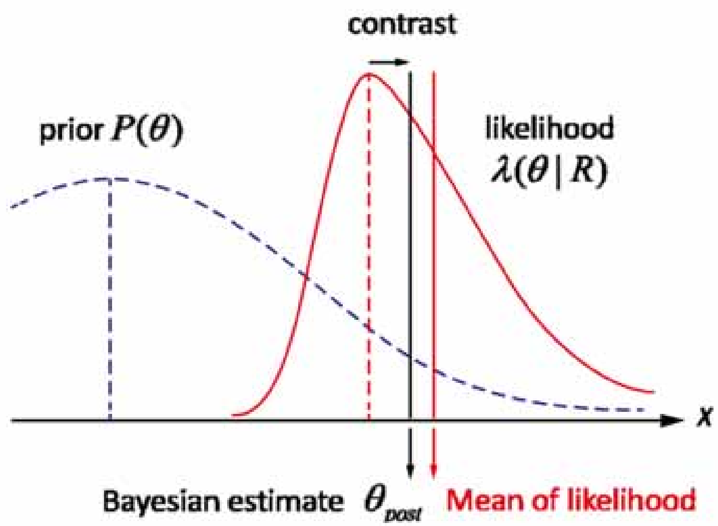
(a) (b)

Adam Santoro, Felix Hill, David Barrett, Ari Morcos & Timothy Lillicrap 2018. Measuring abstract reasoning in neural networks. 35th International Conference on Machine Learning. Stockholm: PMLR, pp. 4477-4486.



Hideyoshi Yanagisawa. How does expectation affect sensory experience? A theory of relativity in perception. 5th International Symposium on Affective Science and Engineering ISASE 2019, 2019 Kogakuin. Japan Society of Kansei Engineering, doi:10.5057/isase.2019-C000014.





Hideyoshi Yanagisawa. How does expectation affect sensory experience? A theory of relativity in perception. 5th International Symposium on Affective Science and Engineering ISASE 2019, 2019 Kogakuin. Japan Society of Kansei Engineering, doi:10.5057/isase.2019-C000014.

Humans can generalize from few examples, and …

- understand relevant representations,
- find concepts between $P(x)$ and $P(Y|X)$,
- with a causal link between $Y \rightarrow X$

even Children can make inferences from little, noisy, incomplete data …

This image is in the public domain, Source: freedesignfile.com

Brenden M. Lake, Ruslan Salakhutdinov & Joshua B. Tenenbaum 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050

Ok, this is also hard for humans …

See also: Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572. Images see: https://imgur.com/a/K4RWn

Adversarial Examples that Fool both Computer Vision and Time-Limited Humans

Gamaleldin F. Elsayed* Google Brain gamaleldin.elsayed@gmail.com

Shreya Shankar Stanford University

Brian Cheung UC Berkeley

Nicolas Papernot Pennsylvania State University

Alex Kurakin Google Brain

Ian Goodfellow Google Brain

Jascha Sohl-Dickstein Google Brain jaschasd@google.com

Abstract

Machine learning models are vulnerable to **adversarial examples**: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.

Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow & Jascha Sohl-Dickstein 2018. Adversarial Examples that Fool both Human and Computer Vision. arXiv:1802.08195.

CD-MAKE The state-of-the-art is terrible …

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

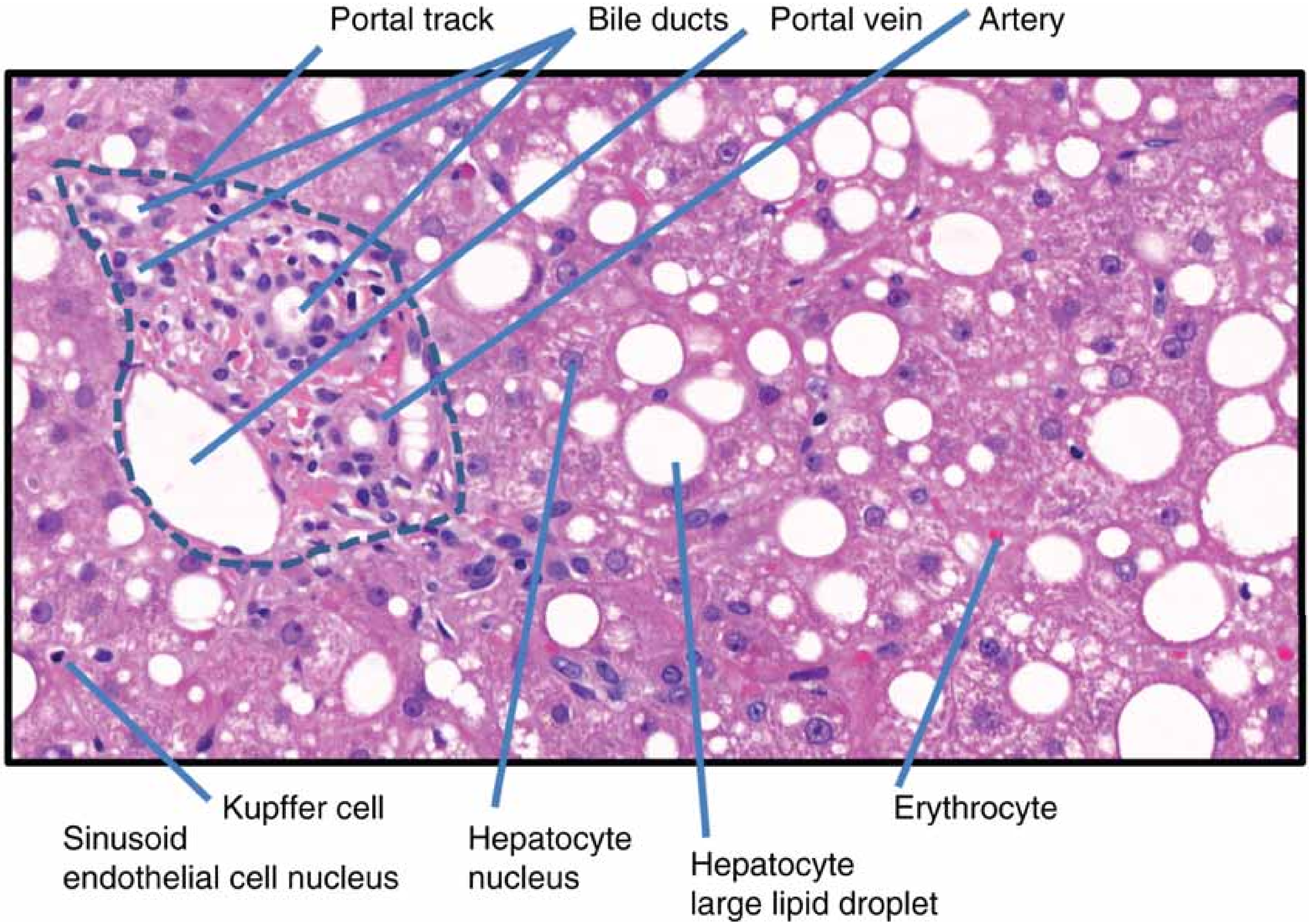
Andrej Karpathy & Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3128-3137.

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

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

Kandinsky Patterns

Motivation: Typical Task of a human Pathologist …



Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of AI in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, doi:10.1002/widm.1312.

The medical report is legally binding - *not* the image

Original Articles

Clinicians Are From Mars and Pathologists Are From Venus

Clinician Interpretation of Pathology Reports

Seth M. Powsner, MD; José Costa, MD; Robert J. Homer, MD, PhD

• *Context.*—Text reports convey critical medical information from pathologists, radiologists, and subspecialty consultants. These reports must be clear and comprehensible to avoid medical errors. Pathologists have paid much attention to report completeness but have ignored the corresponding issue of report comprehension. This situation presents an increasingly serious potential problem. As laboratories are consolidated and as reports are disseminated in new ways (eg, via the World Wide Web), the target audience becomes more diverse and less likely to have any contact with pathologists beyond the written reports them- … on these cases were administered open-book-examination style to surgical attending physicians and trainees during surgical conferences at an academic medical center.

Main Outcome Measures.—Scores from questionnaires.

Results.—Surgeons misunderstood pathologists' reports 30% of the time. Surgical experience reduced but did not eliminate the problem. Streamlined report formatting exacerbated the problem.

Conclusions.—A communication gap exists between pathologists and surgeons. Familiarity with report format and clinical experience help reduce this gap. Paradoxically, sty-

```
        Y A L E  -  N E W   H A V E N   H O S P I T A L
        A N A T O M I C   P A T H O L O G Y   R E P O R T

Patient: ________________          Service: UROLOGY DEPT(YNHH)
Hospital # 1234567                 Path # S94-12345
Birthdate: 02/08/28 (Age: 66)
Sex: M                   Physician: ________________
Accessioned on 08/30/94            UROLOGY 321 YPB
Reported on 09/01/94

Clinical Diagnosis and History:
CLINICAL IMPRESSION: PATIENT WITH BLADDER TUMOR AND
PROSTATIC NODULE.

Tissue Source:
Part 1: TUR TUMOR TRIGONE BLADDER
Part 2: TUR BASE BLADDER
Part 3: BIOPSY LEFT BASE PROSTATE  ...

DIAGNOSIS:==================================================

1) BLADDER, TRIGONE, TRANSURETHRAL RESECTION:
    -TRANSITIONAL CELL CARCINOMA, MODERATELY TO POORLY
     DIFFERENTIATED, GRADE III/IV
    -THE PATTERN OF GROWTH IS NODULAR AND PAPILLARY ...

2) BLADDER, BASE, TRANSURETHRAL RESECTION:
    -TRANSITIONAL CELL CARCINOMA, POORLY DIFFERENTIATED,
     GRADE IV/IV ...
```

Seth M. Powsner, José Costa & Robert J. Homer 2000. Clinicians are from Mars and pathologists are from Venus: clinician interpretation of pathology reports. Archives of pathology & laboratory medicine, 124, (7), 1040-1046.

The state-of-the-art is terrible …

a woman riding a horse on a dirt road

an airplane is parked on the tarmac at an airport

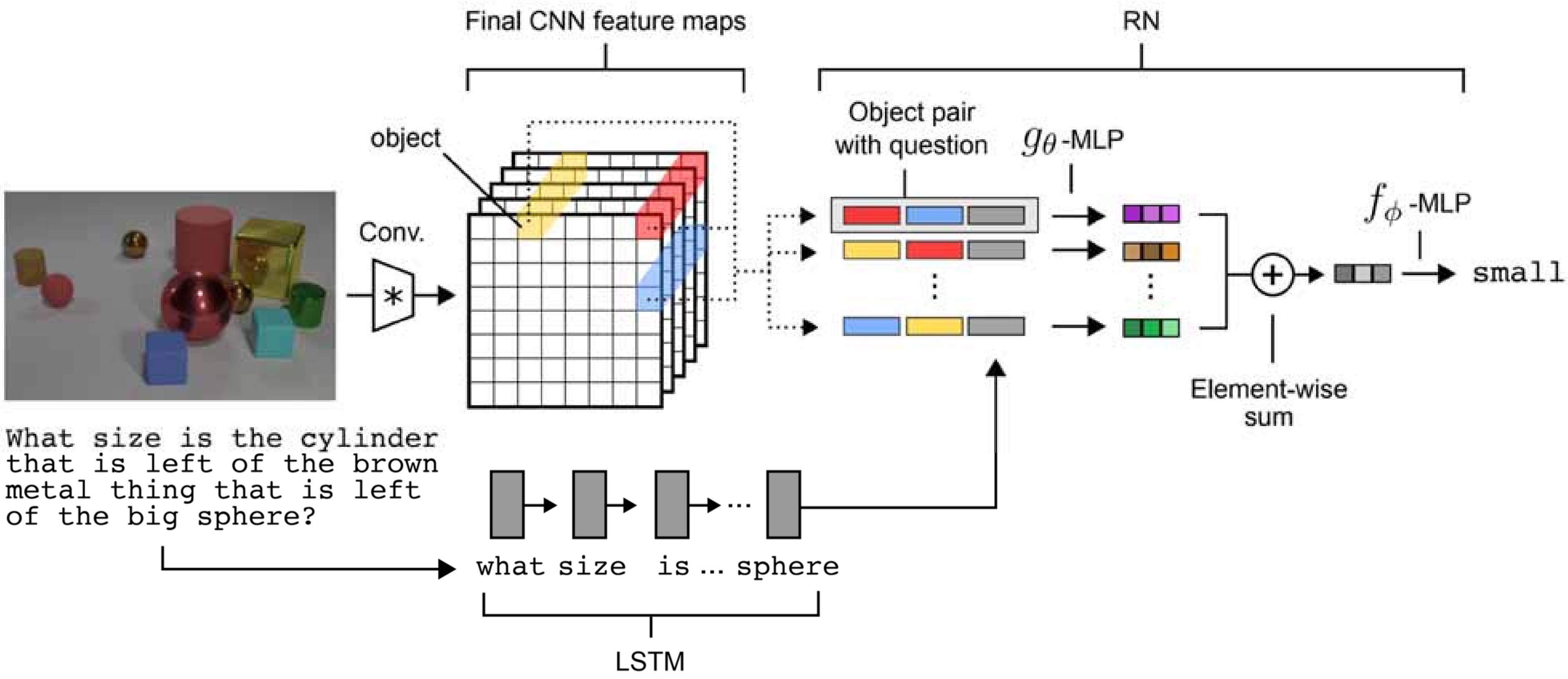
a group of people standing on top of a beach

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 dee learning : github.com/karpathy/neuraltalk2

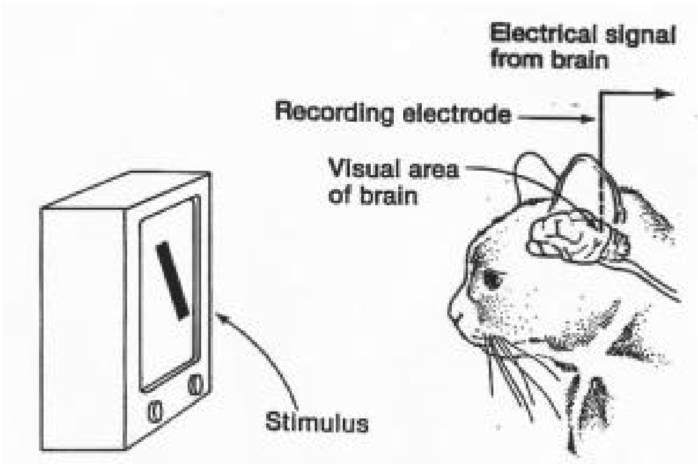
Image Source: Gabriel Villena Fernandez; Agence France-Press, Dave Martin (left to right)

Relational Reasoning

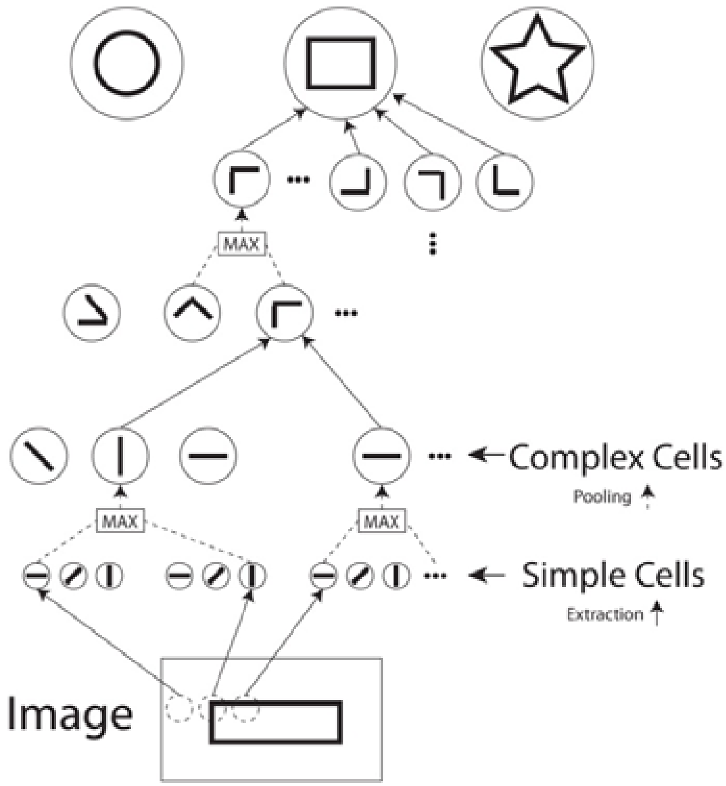


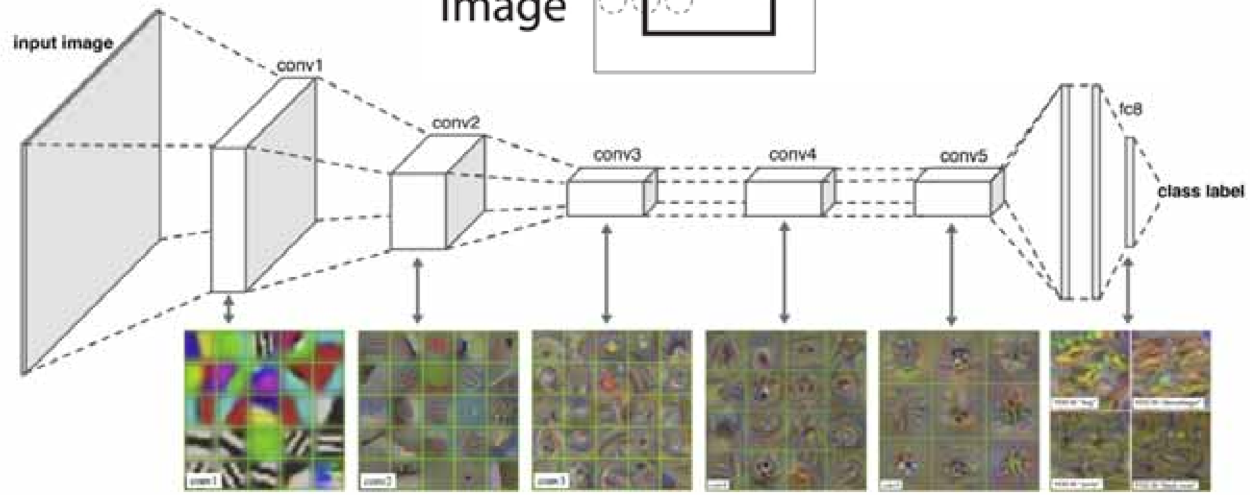
Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia & Timothy Lillicrap. A simple neural network module for relational reasoning. In: Guyon, Isabelle, Luxburg, Ulrike V., Bengio, Samy, Wallach, Hannah, Fergus, Rob, Vishwanathan, Svn & Garnett, Roman, eds. Advances in neural information processing systems (NIPS), 2017 Long Beach (CA). Neural Information Processing Society, 4967-4976.

Hubel & Wiesel (1962): Our world is compositional



David H. Hubel & Torsten N. Wiesel 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160, (1), 106-154, doi:10.1113/jphysiol.1962.sp006837.

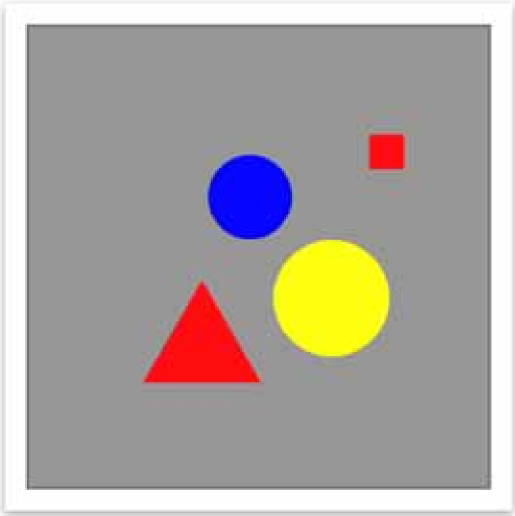




Wassily Kandinsky (1866 – 1944)

Komposition VIII, 1923, Solomon R. Guggenheim Museum, New York.Source: https://de.wikipedia.org/wiki/Wassily_Kandinsky
This images are in the public domain.

Definition 1: A Kandinsky Figure is …



- … a square image containing 1 to n geometric objects.
- Each object is characterized by its shape, color, size and position within this square.
- Objects do not overlap and are not cropped at the border.
- All objects must be easily recognizable and clearly distinguishable by a human observer.

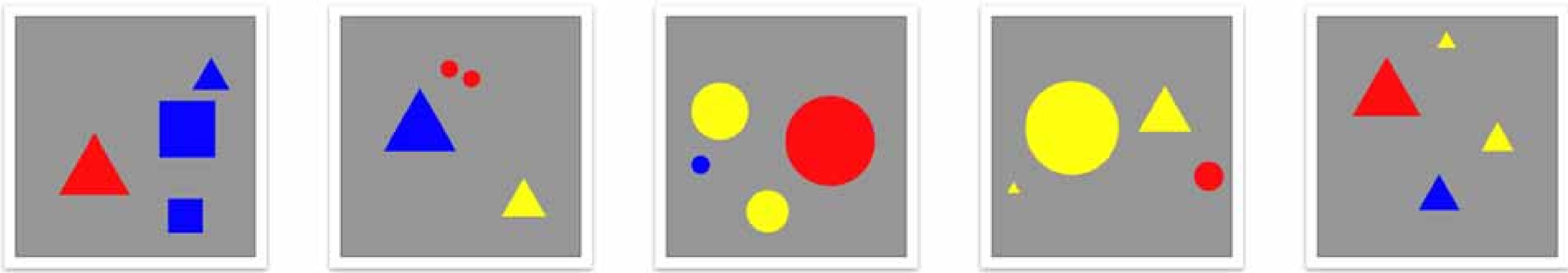
Definition 2 A statement $s(k)$

- about a Kandinsky Figure k is …
- either a mathematical function $s(k) \rightarrow B$; with B $(0,1)$
- or a $natural\ language\ statement$ which is true or false

- Remark: The evaluation of a natural language statement is always done in a specific context. In the followings examples we use **well known concepts from human perception** and linguistic theory.
- If $s(k)$ is given as an algorithm, it is essential that the function is a pure function, which is a computational analogue of a mathematical function.

Definition 3 A Kandinsky Pattern K …

- … is defined as the subset of all possible Kandinsky Figures k with $s(k) \rightarrow 1$ or the natural language statement is true.
- $s(k)$ and a natural language statement are equivalent, if and only if the resulting Kandinsky Patterns contains the same Kandinsky Figures.
- $s(k)$ and the natural language statement are defined as the **Ground Truth** of a Kandinsky Pattern



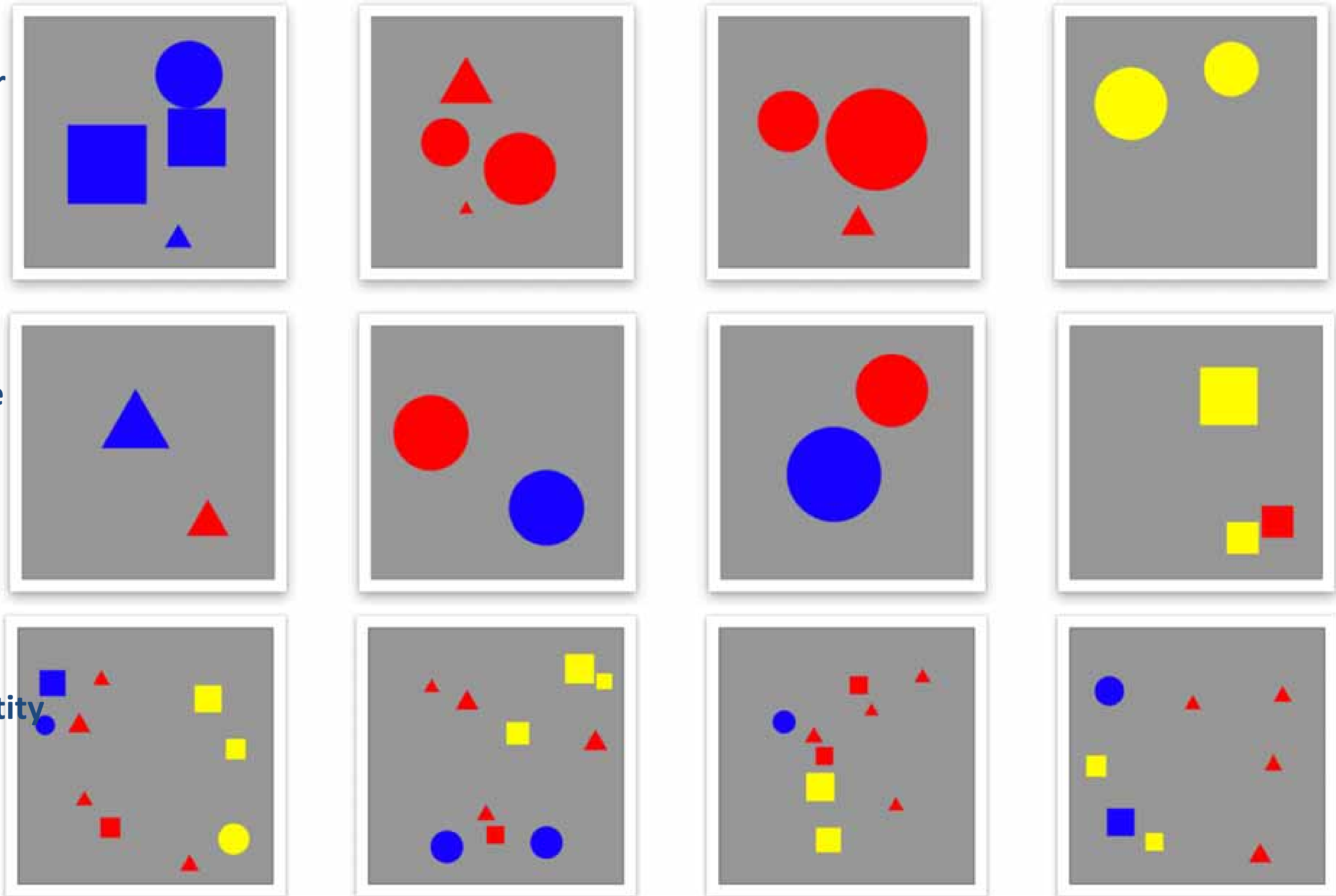
"… the Kandinsky Figure has two pairs of objects with the same shape, in one pair the objects have the same color, in the other pair different colors, two pairs are always disjunct, i.e. they don't share a object …".

How do humans explain? They use concepts!

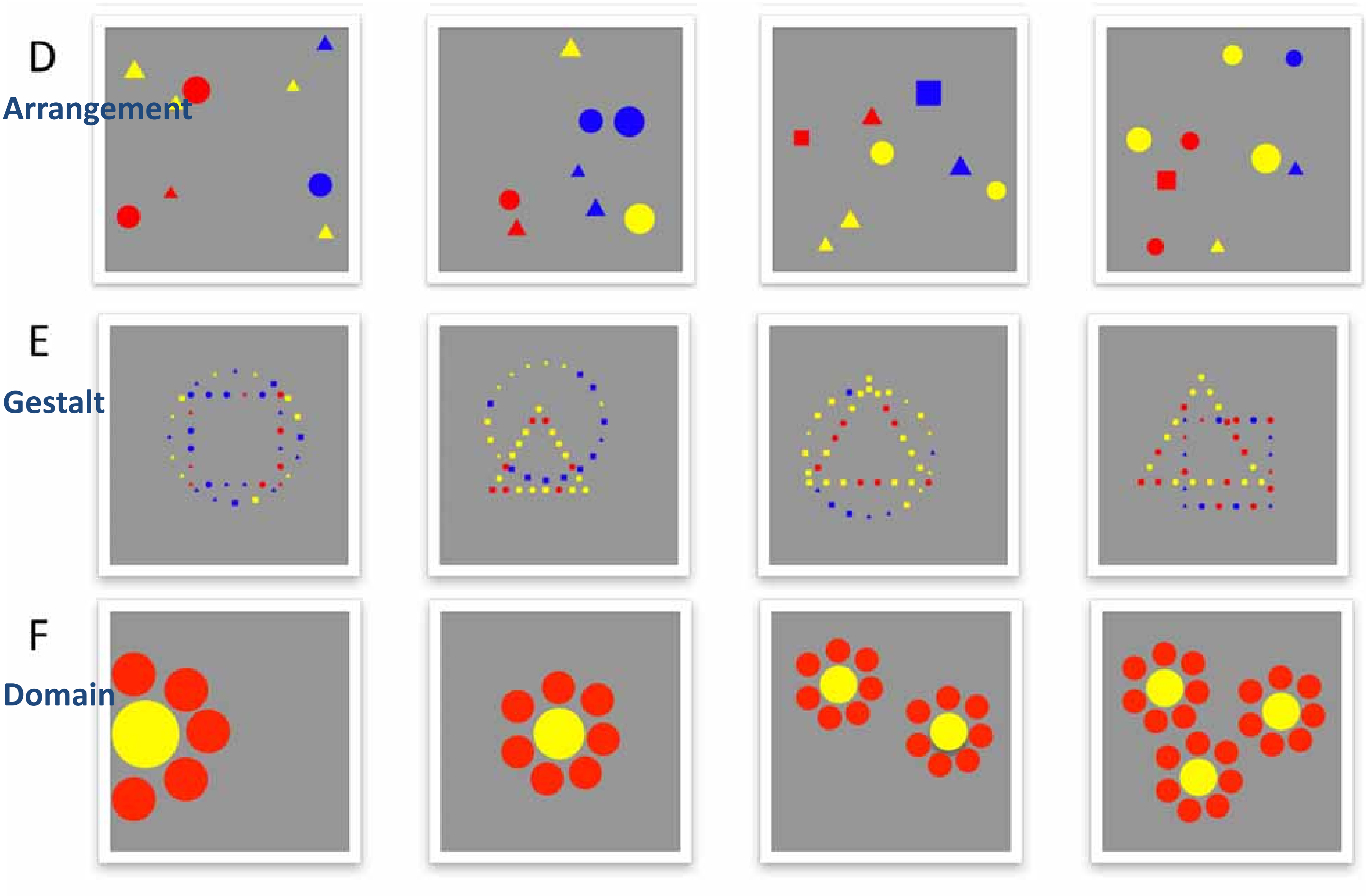
A
Colour

B
Shape

C
Quantity

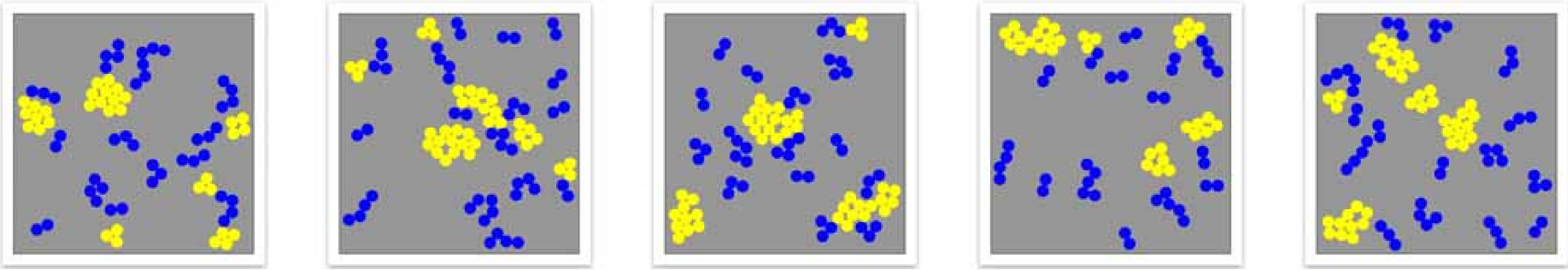


CD-MAKE How do humans explain? (2/2)

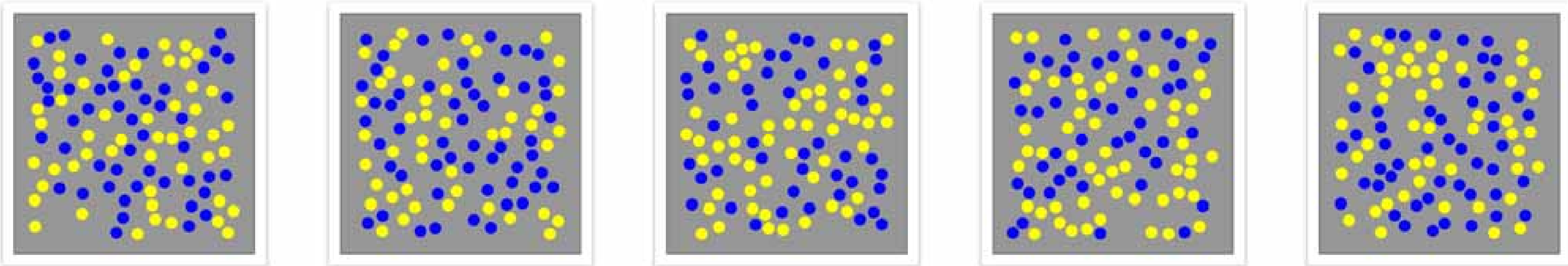


Domain Concept from Pathology

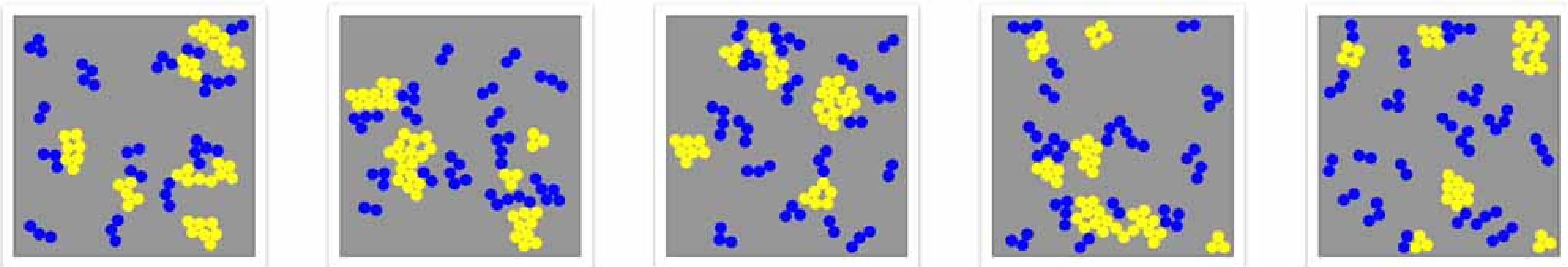
A) True



B) False

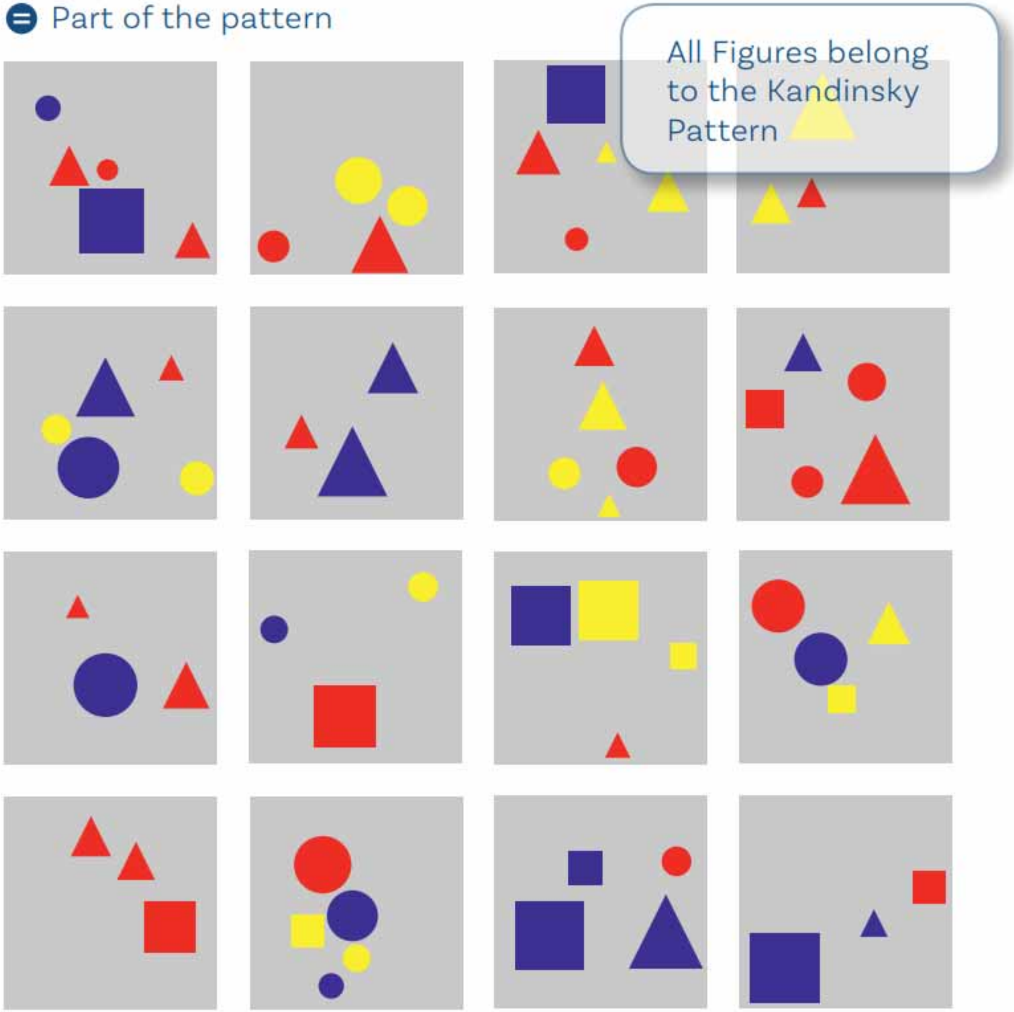


C) Counterfactual

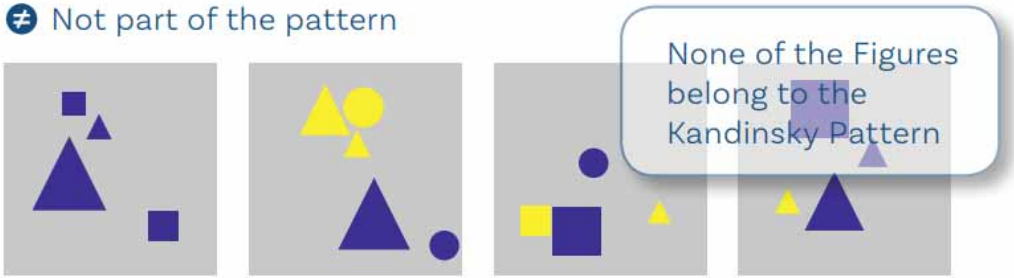


You find this booklet in your conference bag ...

Part of the pattern



Not part of the pattern



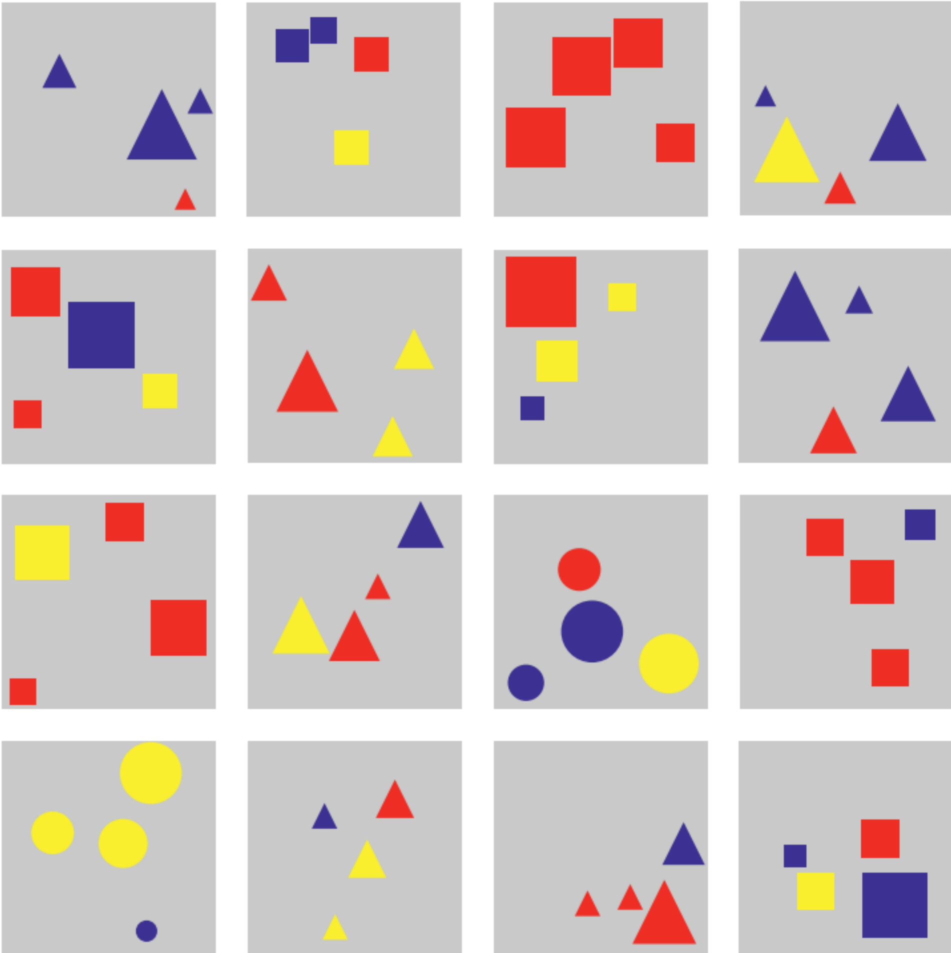
Hypothesis 1

It only contains circles and triangles.

Hypothesis 2

It contains at least a red object.

Part of the pattern



Not part of the pattern

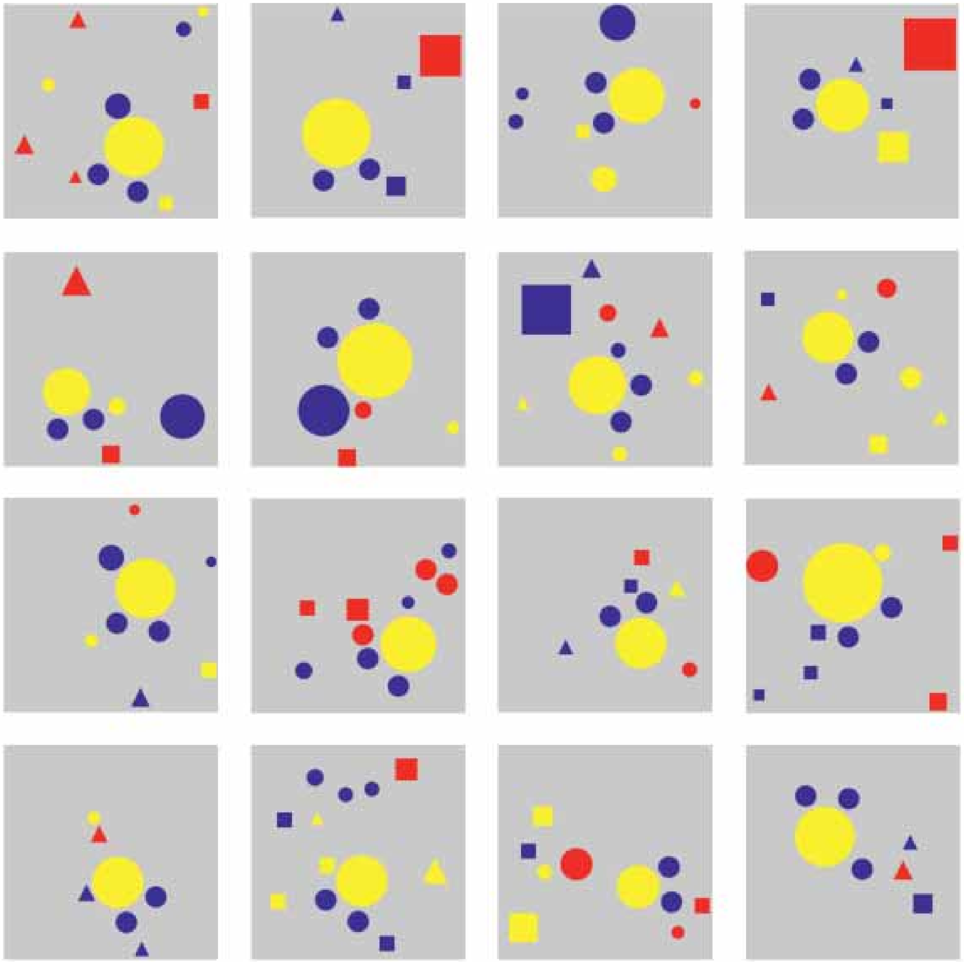
S2 Basic Pattern 2

Title: **All of Same Shape ->**

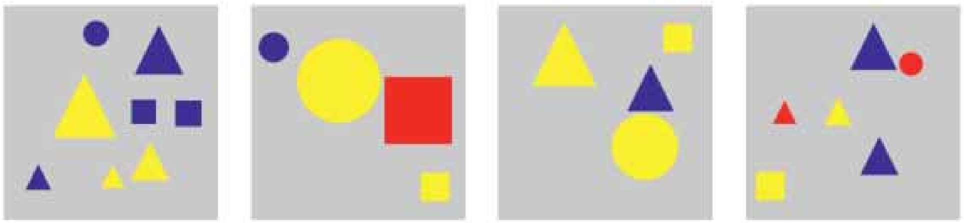
All objects have the same shape.

Hint: Don't be distracted by the colors

Part of the pattern



Not part of the pattern



S8 Basic Pattern 8

Title: **Mickey Mouse** ->

Every figure contains a pattern which is made out of a big yellow circle and two smaller blue ones and looks like a Mickey Mouse.

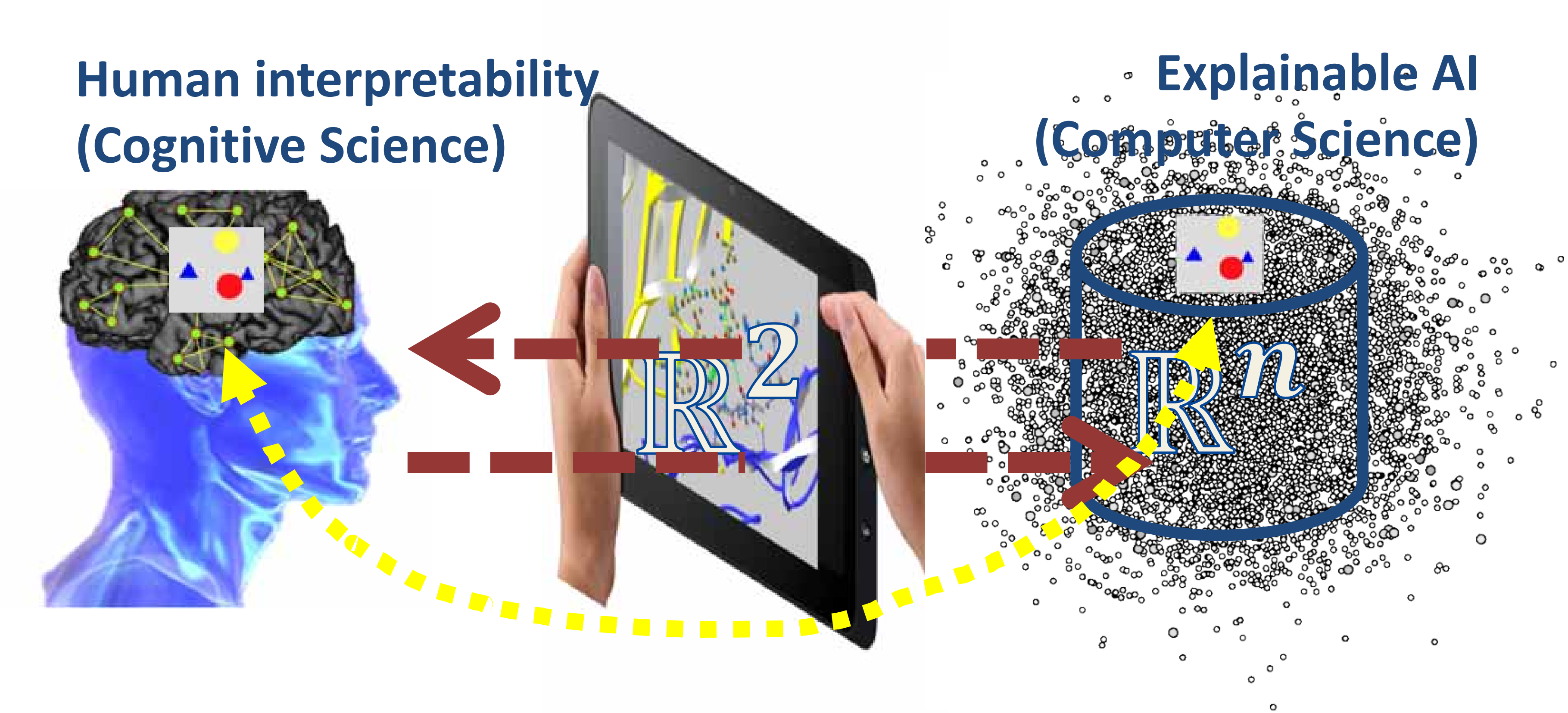
Conclusion

Results of our experiments with humans

- The results show that the majority of explanations was made based on the properties of individual elements in an image
- i.e., shape, color, size, …
- and the appearance of individual objects (number)
- Comparisons of elements (e.g., more, less, bigger, smaller, etc.) were significantly less likely, and
- the location of objects, interestingly, played almost no role in the explanation of the images.

- Although humans tend to make more errors, human intelligence is more reliable and robust against catastrophic errors, whereas
- AI is vulnerable against software, hardware and energy failures.
- Human intelligence develops based on infinite interactions with an infinite environment, while AI is limited to the small world of a particular task.
- The development of intelligence, therefore, is the result of the incremental interplay between challenge/task, a conceptual change (physiological as well as mentally) of the system, and the assessment of the effects of the conceptual change.
- To advance AI, specifically in the direction of explainable AI, we suggest bridging the human strength and the human assessment methods with those of AI.

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



Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal & Heimo Mueller 2019. Causability and Explainability of AI in Medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, doi:10.1002/widm.1312.

Thank you!
Now, take part in the explainable-AI challenge:

https://human-centered.ai/kandinksy-challenge