Ay 119 – Spring 2026

Some Introductory Comments

S. George Djorgovski

Caltech

Welcome to Ay 119 – Some Class Logistics

Class website: **https://sites.astro.caltech.edu/~george/ay119/**

Everything will be posted there. Please read the **Syllabus**

Lectures/Discussion: Tuesday 4-5 pm in 304 Cahill – attendance is ***mandatory***

and please ***ask questions***

This class is a very, very broad-brush introduction to some topics in Astroinformatics = Data Science applied to astronomy, but also broader

To really learn about it, you need to do some follow-up study

Weekly thematic lectures, assignments due next Tuesday

Attendance + assignments is all you need for P/F

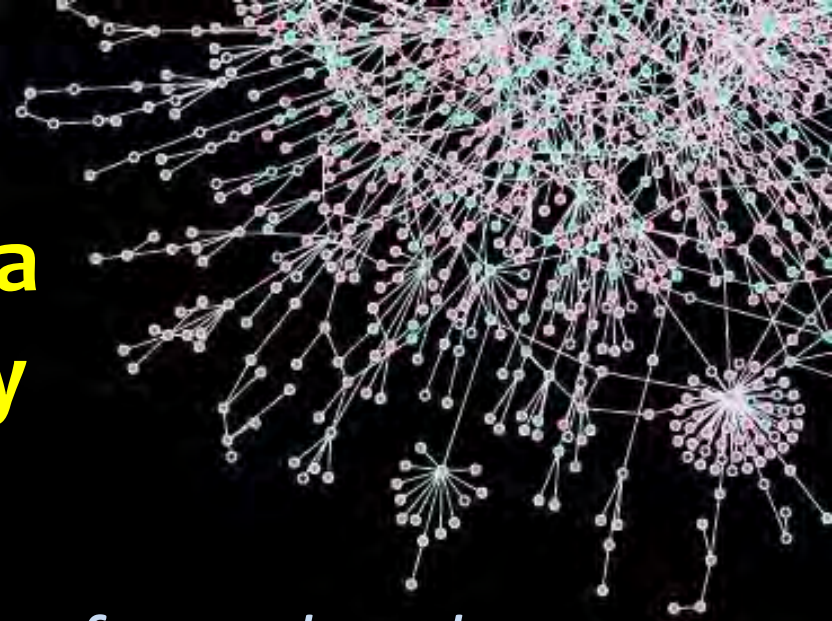
If you want to get a letter grade, you will also need to complete a class project.

See the Syllabus for details

Exponential Growth of Data Volumes

on Moore's law time scales

… and Data Complexity



Understanding of complex phenomena requires complex data!

From data poverty to data glut
From data sets to data streams
From static to dynamic, evolving data
From anytime to real-time analysis and discovery
From centralized to distributed resources
From ownership of data to ownership of expertise

These Challenges are universal

Djorgovski

Big Data is Not About the Data

It is about **knowledge discovery** in the data

The ***information content*** of modern data is so high as to enable a profitable **data mining**

That is where the **Machine Learning and AI** come in

Data fusion reveals new knowledge that was present, but was ***not recognizable*** in the individual data sets, adding to the data complexity

Data complexity is the key challenge (scalability, visualization) and finding and interpreting complex structures in data may exceed the cognitive capacity of a human mind

Data Science is Not a Science

... in the traditional sense: it does not have its own application domain

It is **a collection of methodologies** derived from the computer science and engineering, statistics, etc., applied in other fields

Its **universality** and the universality of data challenges makes it a new part of the **scientific method** for the 21st century

It is the **new universal language of science** and other quantitative fields. It plays the same role that mathematics did since the 17th century

It is thus an **interdisciplinary glue/lubricant**

- Many important problems (e.g., climate change, sustainability, energy, brain...) are inherently inter/multi-disciplinary
- Discoveries are often made at the field boundaries

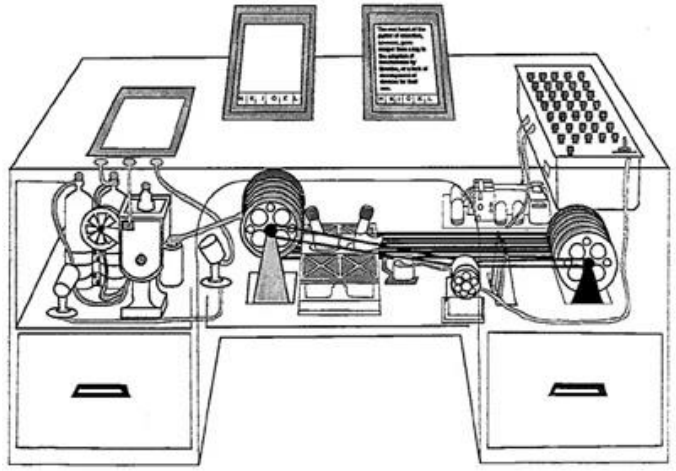
History of AI: the Visionaries

Vannevar Bush, The Atlantic, July 1945

A SCIENTIST OF THE FUTURE RECORDS EXPERIMENTS WITH A TINY CAMERA FITTED WITH UNIVERSAL-FOCUS LENS. THE SMALL SQUARE IN THE EYEGLASS AT THE LEFT SIGHTS THE OBJECT

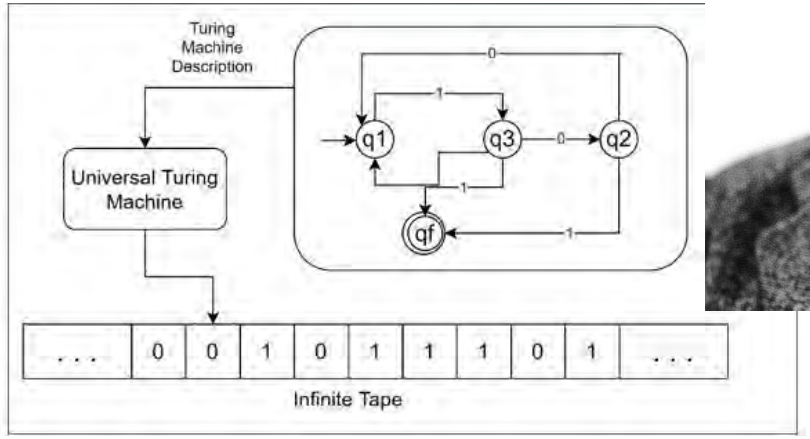
AS WE MAY THINK

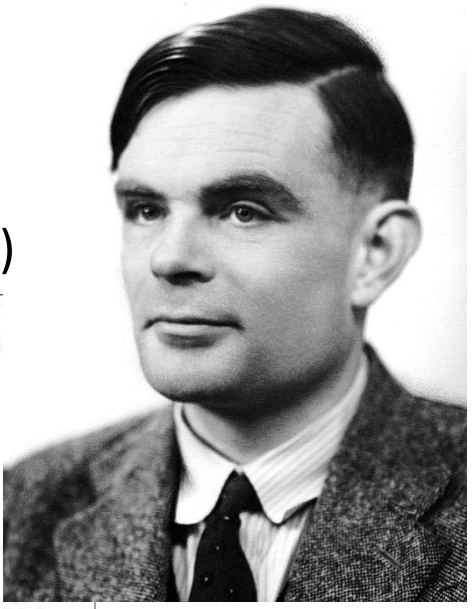
A TOP U. S. SCIENTIST FORESEES A POSSIBLE FUTURE WORLD IN WHICH MAN-MADE MACHINES WILL START TO THINK



Memex, a foundational concept for a mechanized, personal library that stores, indexes, and links documents, books, and records for rapid retrieval

Universal Turing Machine (1936)





I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

Vol. lix. No. 236. October, 1950

The Birth of AI: the Dartmouth Workshop (1956)

The Dartmouth Summer Research Project on Artificial Intelligence, organised by John McCarthy, Marvin Minsky, Nathaniel Rochester, and Claude Shannon, was a 1956 summer workshop widely considered to be the founding event of artificial intelligence as a field.

McCarthy coined the term "Artificial Intelligence"

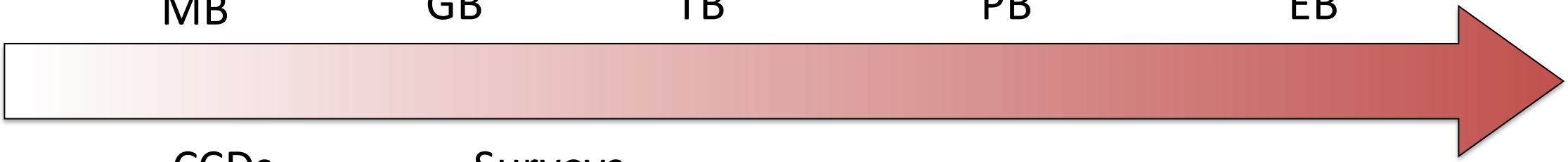
The Evolving Data-Driven Astronomy

Enabled and driven by the information technology

Dark(room) ages

1980 1990 2000 2010 2020

MB GB TB PB EB



CCDs

Image Proc.

Surveys

Pipelines

Databases

Virtual Observatory

Machine Learning

AstroInformatics

Deep Learning

AI

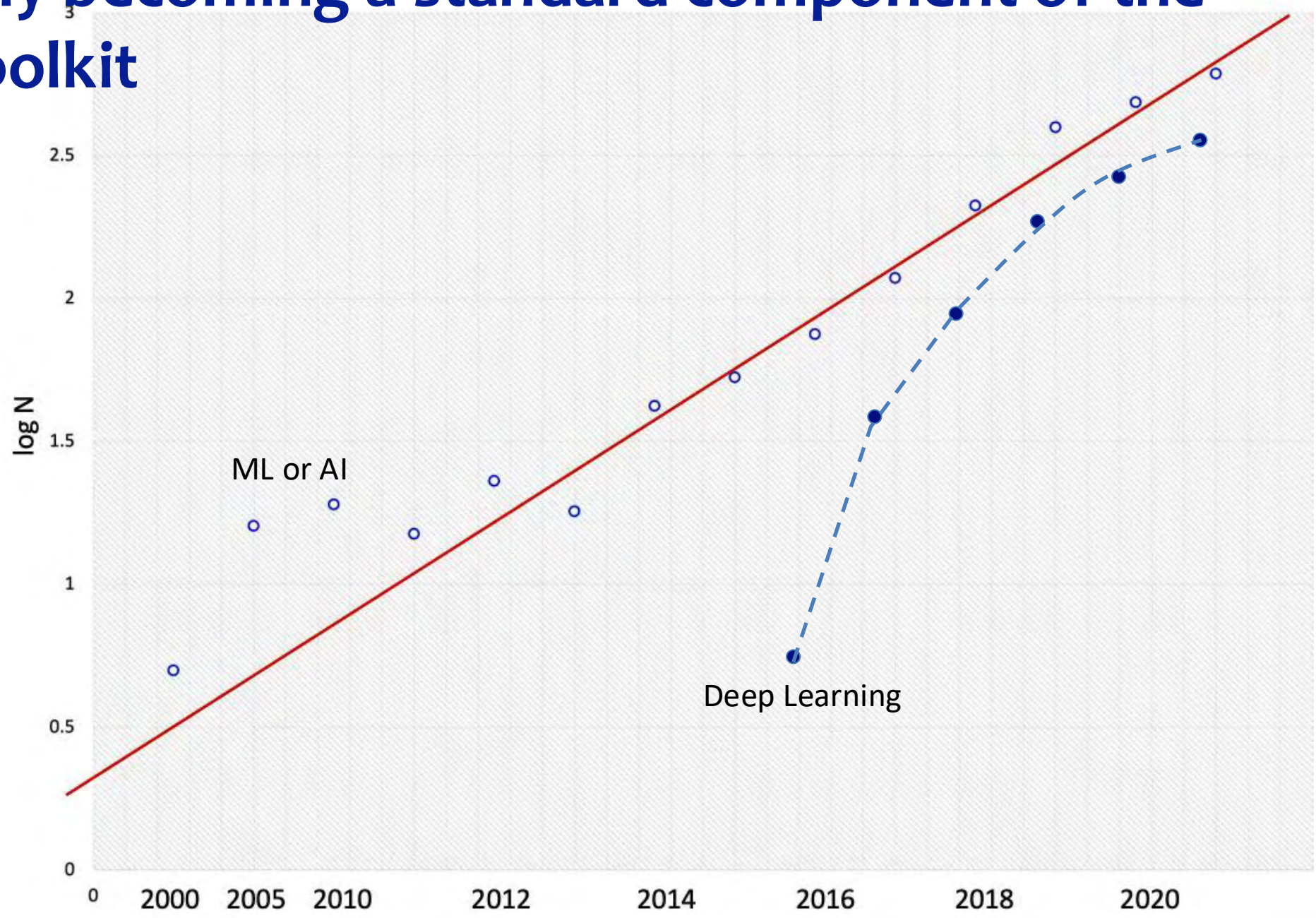
Astronomy was always
on the cutting edge of the
detector and data processing technologies

ML/AI are rapidly becoming a standard component of the astronomical toolkit

Number of papers with ML or AI keywords in their title or the abstract

An **exponential growth** with a doubling time of ~20 months

(Data from ADS)

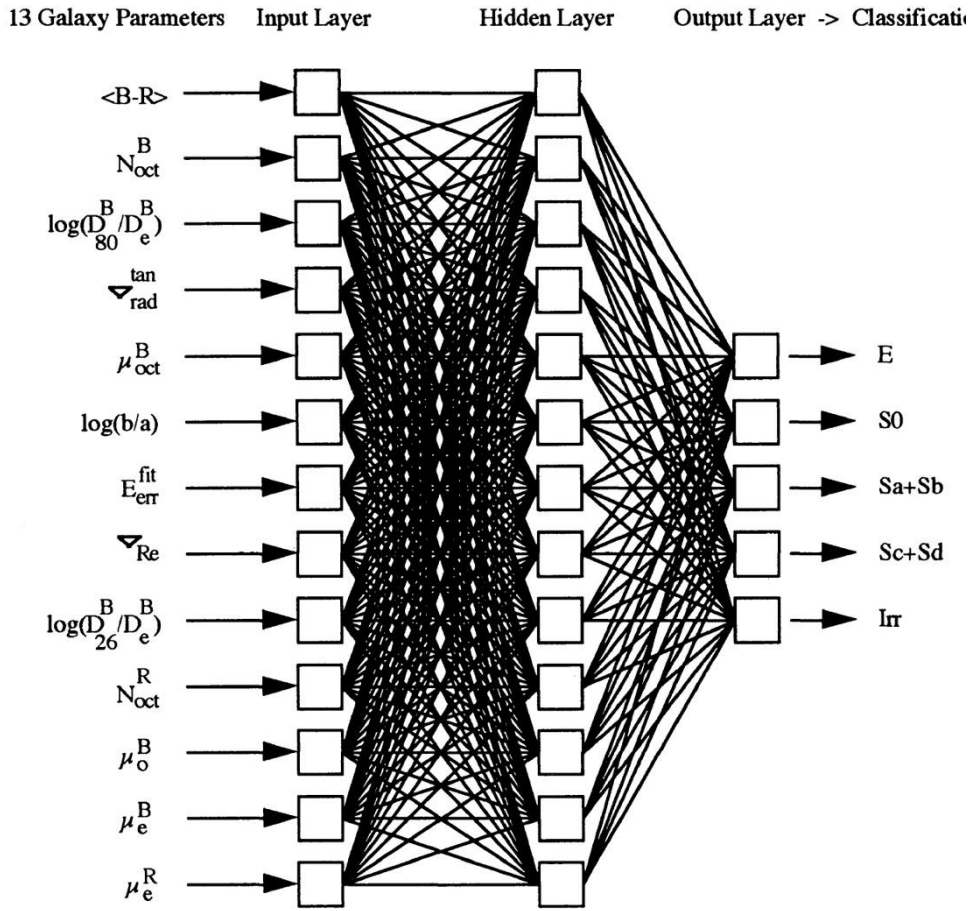


The Early Days (early/mid 1990's): Galaxy Morphology

Morphological classification of galaxies by Artificial Neural Networks

M. C. Storrie-Lombardi,[1] O. Lahav,[1] L. Sodré, Jr[2,3] and L. J. Storrie-Lombardi[1]

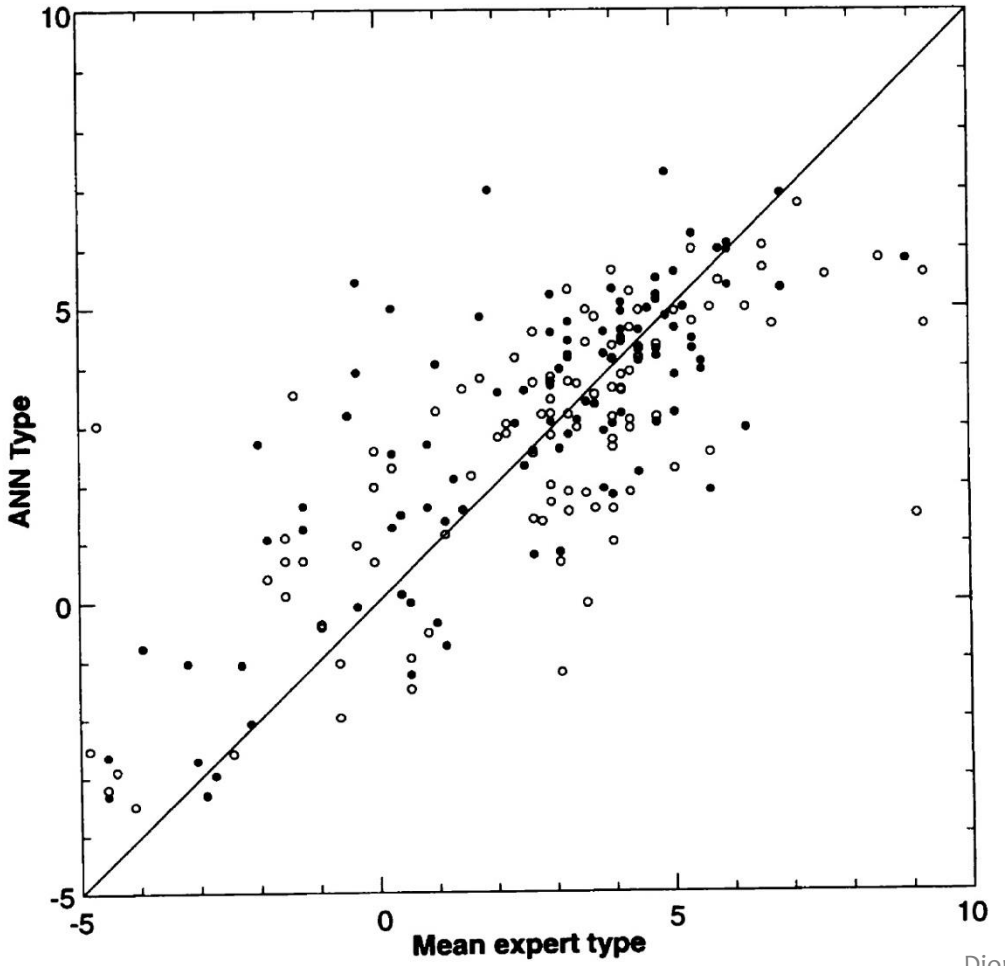
Mon. Not. R. Astron. Soc. (1992) **259**, Short Communication, 8P–12P



Galaxies, Human Eyes, and Artificial Neural Networks

O. Lahav et al.

SCIENCE • VOL. 267 • 10 FEBRUARY 1995



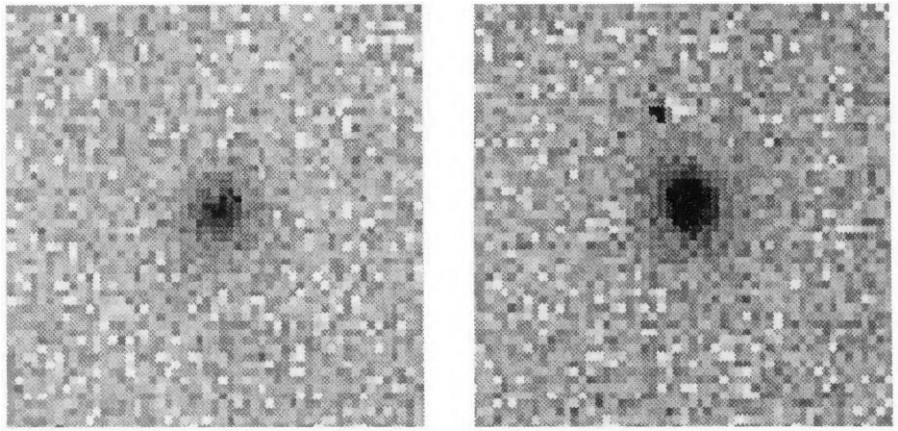
Djorgovski

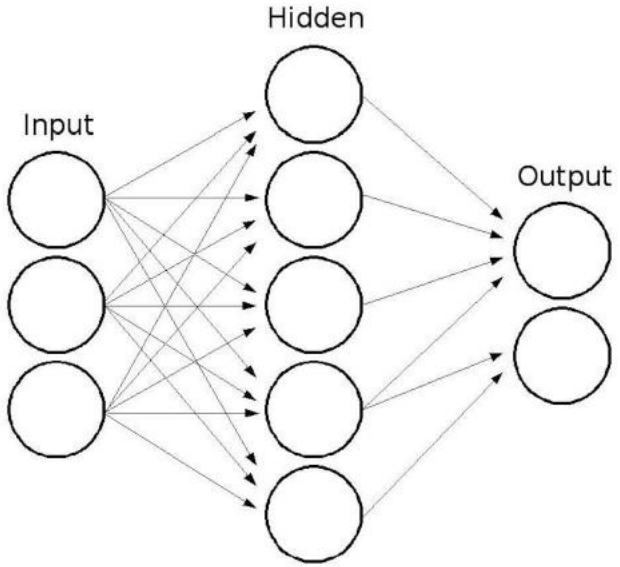
The Early Days (early/mid 1990's): Star-Galaxy Separation

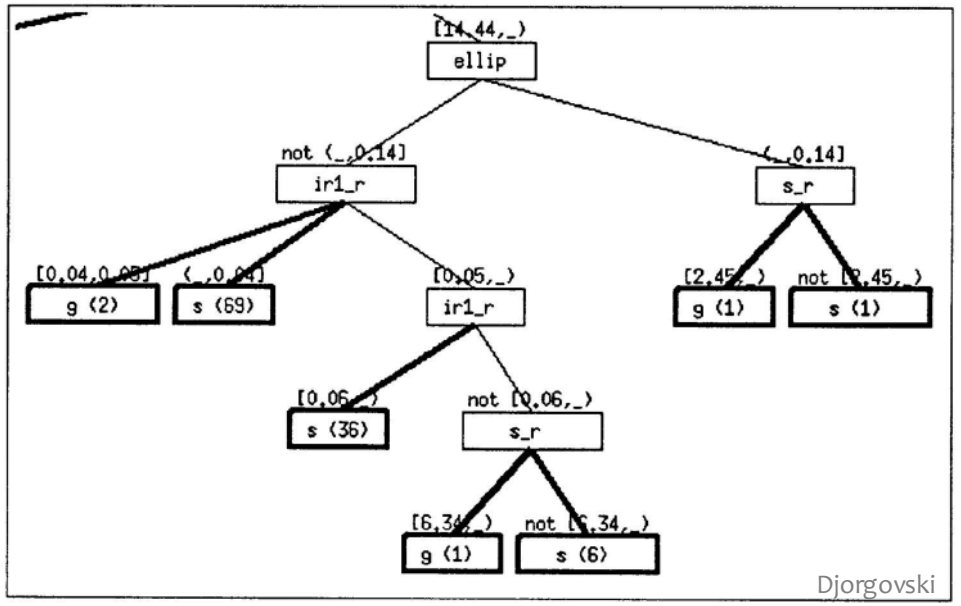
The initial large scale ML applications were for the star-galaxy separation in the first panoramic digital sky surveys (DPOSS, 2MASS, SDSS, etc.), using supervised classifiers, e.g., Decision Trees (Weir et al. 1995) or Artificial Neural Nets (Odewahn et al. 2004). Higher resolution CCD images provided labeled training, validation, and testing data sets.

ML was used to replace humans doing tedious, repetitive tasks, and doing it consistently and objectively

Examples from DPOSS

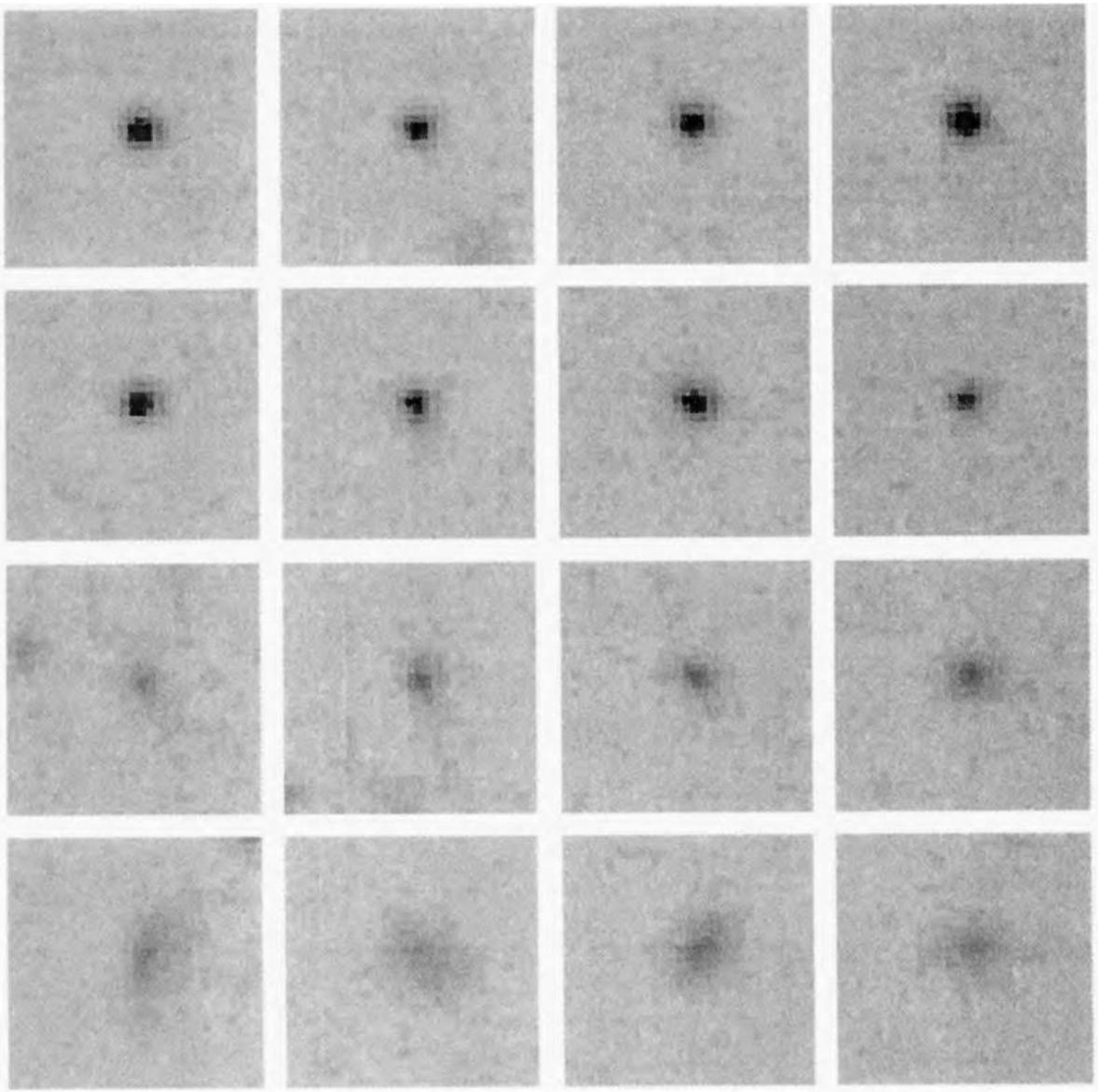






The Early Days (early/mid 1990's): Object Classification

An early application of **unsupervised clustering** (*Autoclass*) applied for a morphological classification of sources detected in DPOSS (Weir et al. 1995)



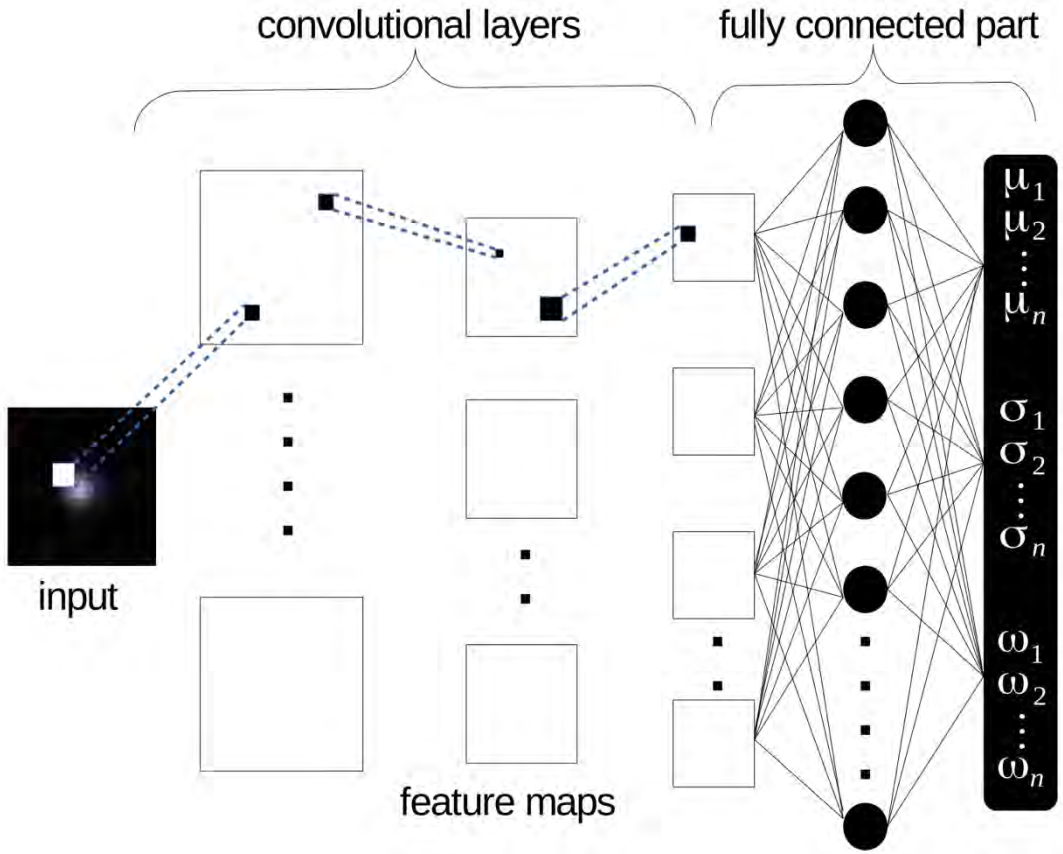
Stars

Three classes of galaxies, depending on their concentration

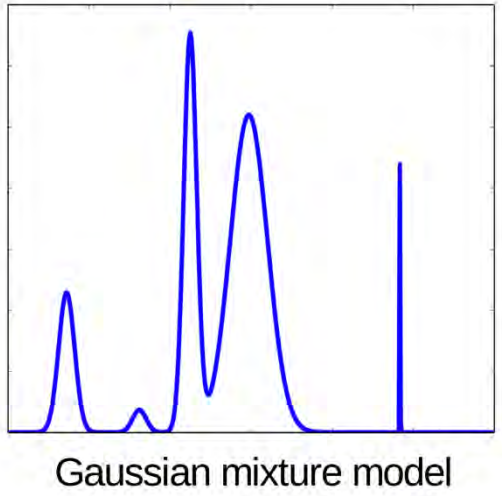
Early Applications (late 1990's - present): Photometric Redshifts

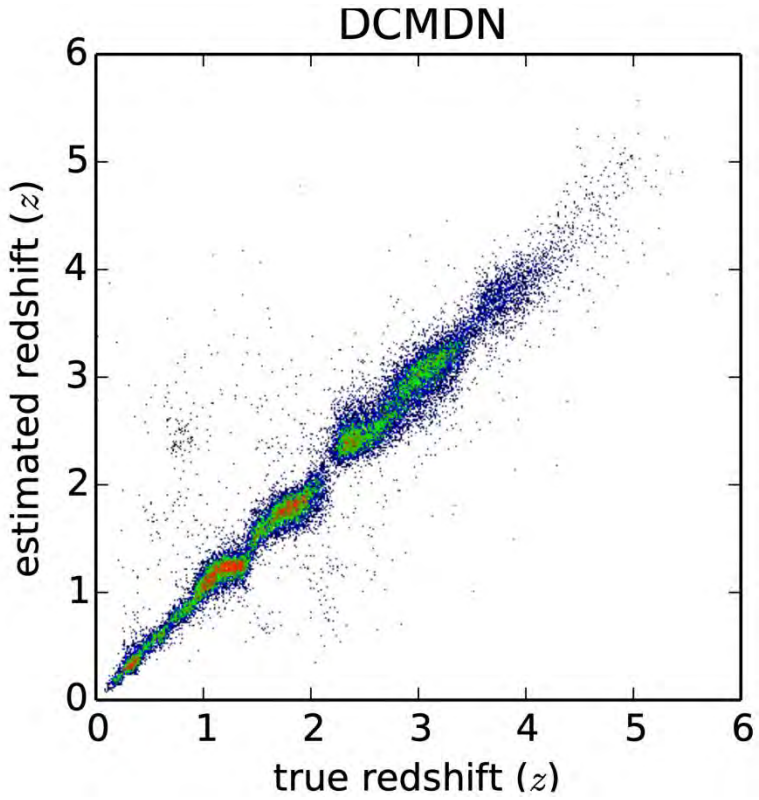
- Many methods have been used, combining various ML tools
- **A key insight:** photo-z's as ***probability density distributions*** (which are rarely simple gaussians, often multimodal), rather than a single number – the error bars are ill-defined

Great savings of the large telescope time for spectroscopy, given a sufficient accuracy and reliability



D'Isanto & Polsterer (2018)



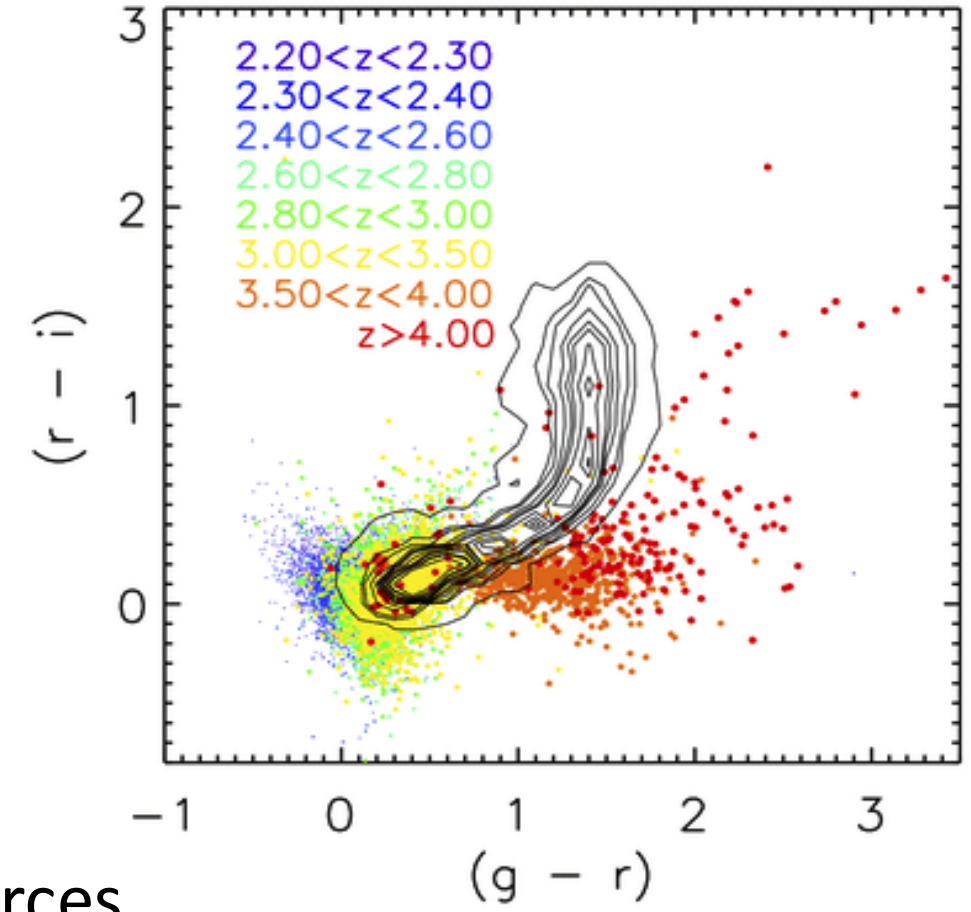


Model-Based Outlier Search and Surprises (mid-1990's to the present)

Examples: searches for quasars or brown dwarfs in a color space, i.e., **physical classes** of known, interesting types of objects ("known knowns")

Based on an *empirical model* of how the type of objects of interest would manifest in a particular observable parameter space

ML-based selection of targets for a follow-up spectroscopy – optimizing telescope use



Doing a systematic, unbiased classification of sources using **unsupervised clustering**, and possibly discovering **new classes of objects** ("unknown unknowns") remains a major goal

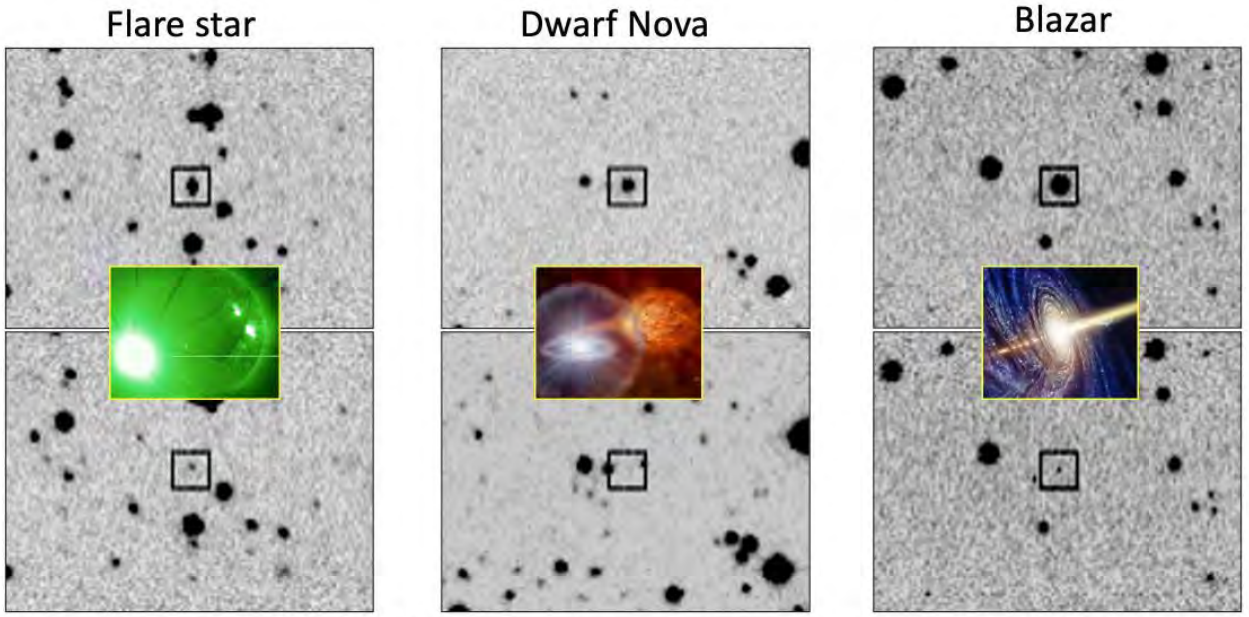
Djorgovski

Time Domain Astronomy: Growing Challenges
(2000's to the present – and the future)

All the challenges from the panoramic sky surveys, plus the variability

A major new growth area of astrophysics, driven by the new generation of large digital synoptic sky surveys

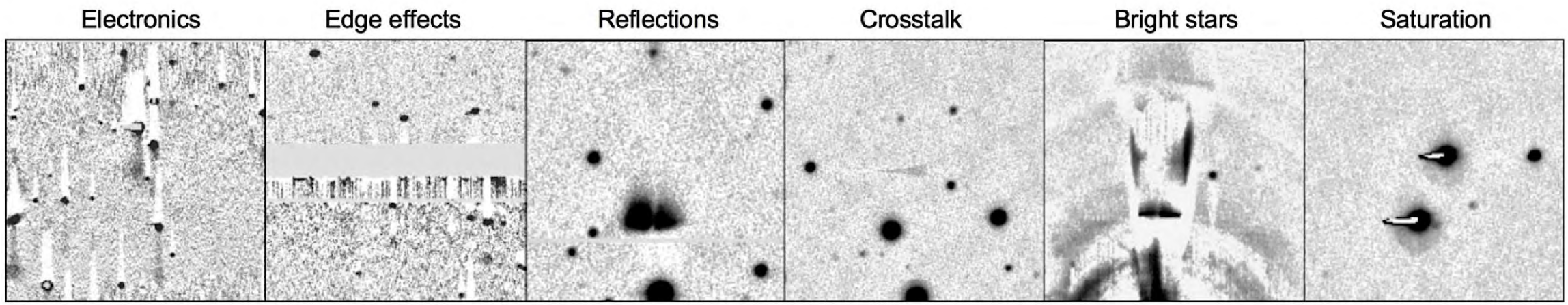
Rich phenomenology, from the Solar system to cosmology and extreme relativistic physics. For some phenomena, time domain information is a key to the physical understanding

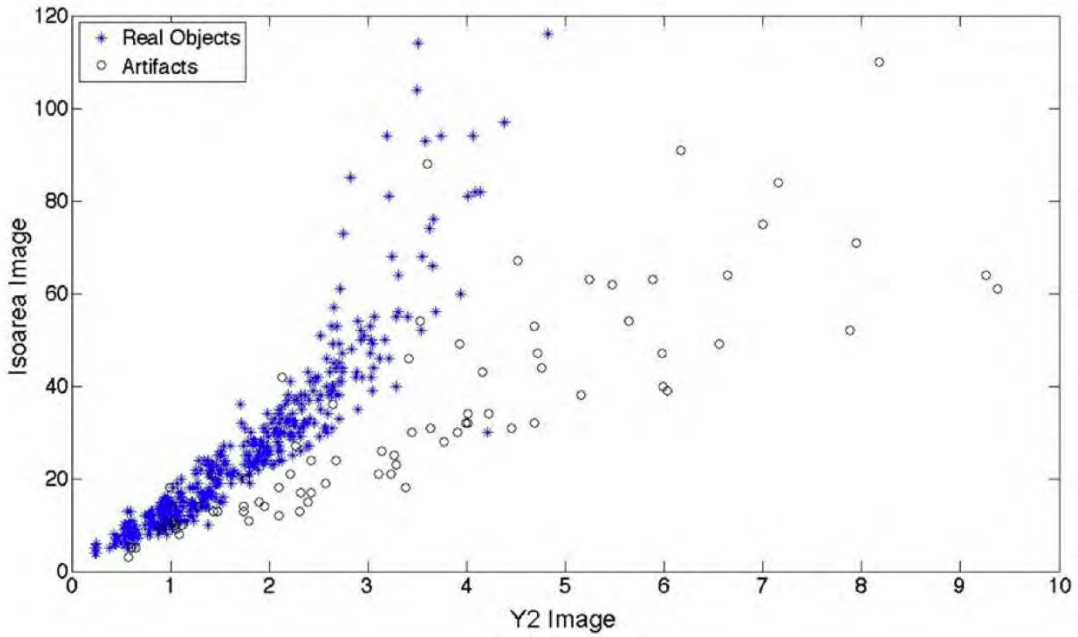


Vastly different physical phenomena, and yet they look the same!
Which ones are the most interesting and worthy of follow-up?

The key challenge: ***physical classification*** of variable or transient sources, in real time, using their light curves (irregularly sampled, sparse) and the available contextual information: spatial, temporal, and multi-wavelength (heterogeneous, missing)

Time Domain Surveys: Automated Filtering of Artifacts





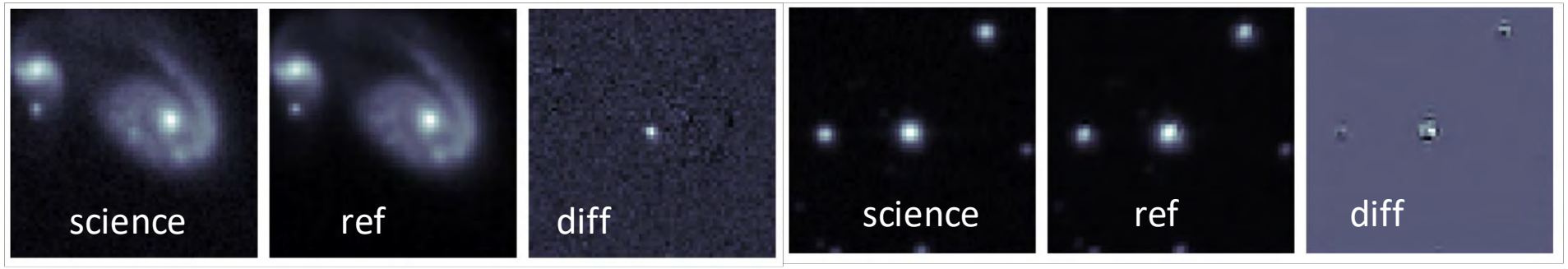
Especially critical for the surveys that use image subtraction ("real/bogus")

← Automated classification and rejection of artifacts masquerading as transient events in the PQ survey pipeline, using a Multi-Layer Perceptron ANN; accuracy > 95%

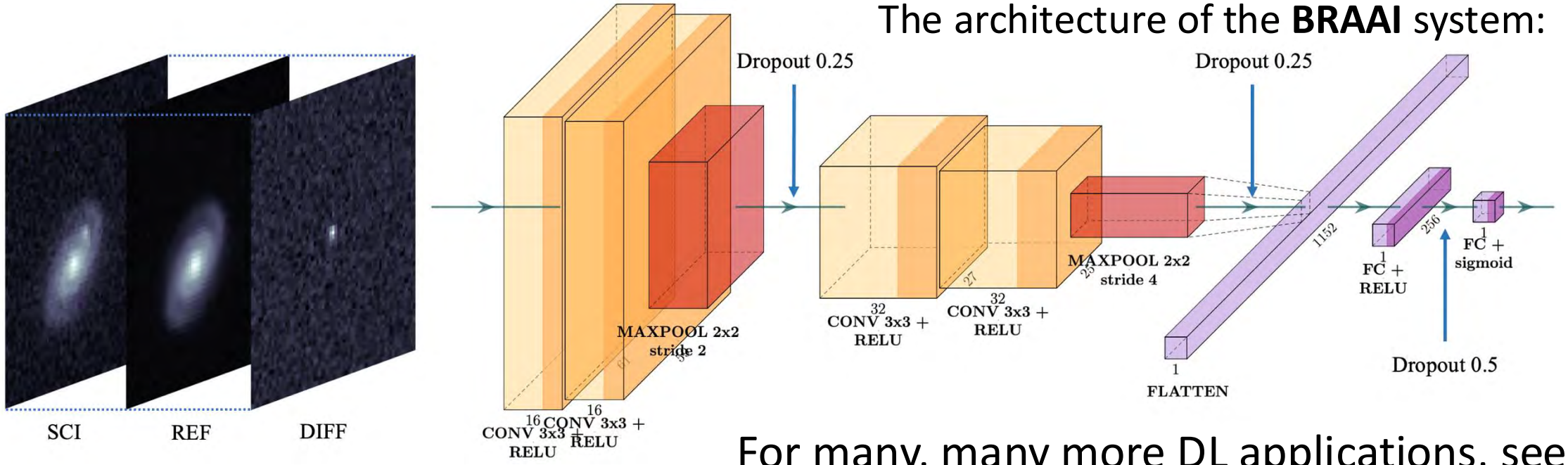
(Donalek et al. 2008)

Deep Learning Applications: Real/Bogus Separation

Automated removal of image subtraction artifacts for ZTF *(Duev, Mahabal, et al. 2019)*



The architecture of the **BRAAI** system:



For many, many more DL applications, see

Huertas-Company & Lanusse, PASA, 40, e001 (2023) | doi:10.1017/pasa.2022.55

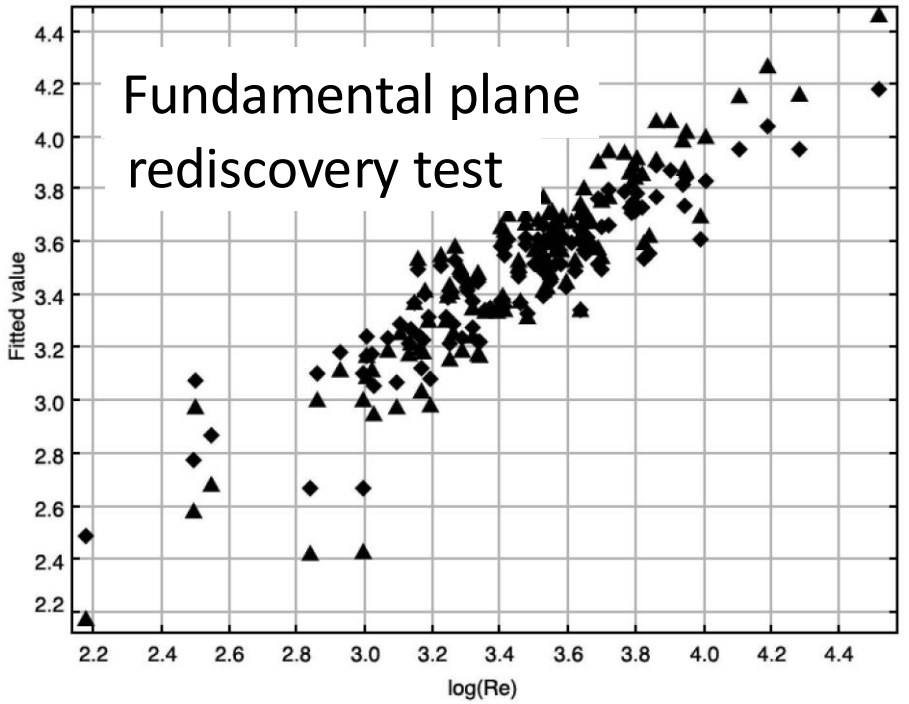
Machine Discovery of Analytical Relationships

(see Graham et al. 2013, MNRAS 431, 2371; also Tegmark et al. 2018 - 2022)

Traditionally we have been using ML to describe the ***geometry and statistics of data distributions*** in feature spaces, e.g., clustering, correlations, outliers, etc.

But can ML/AI help us discover ***analytical expressions that describe the relationships hidden in the data***?

- Employ **symbolic regression** to determine best-fitting functional form to data and its parameters simultaneously
- Specify building blocks to be used: algebraic operators, analytical functions, constants
- Using commercial package *Eureqa (Nutonian)*



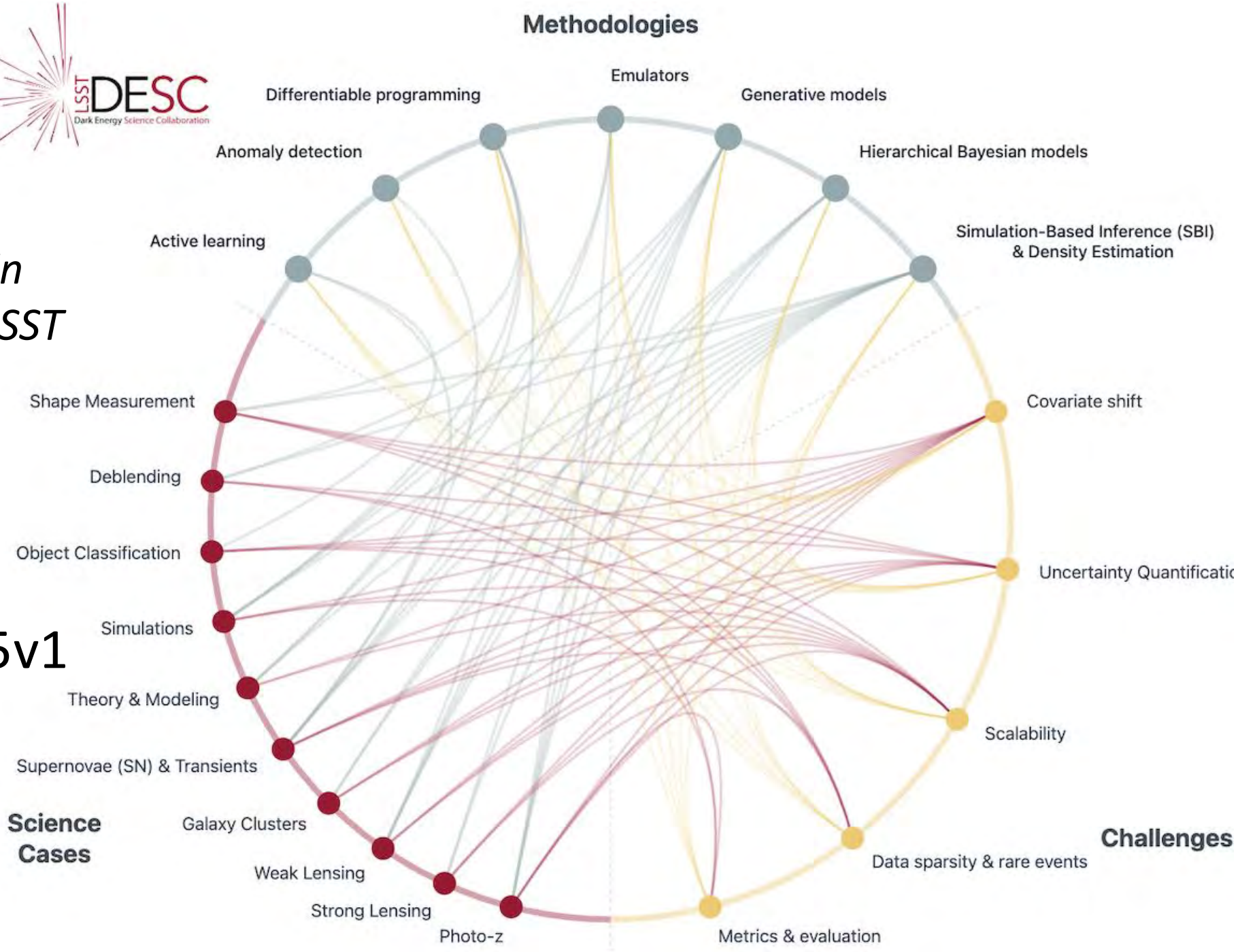
- Test: rediscover known astrophysical correlations (HRD, FP)

AI in Astronomy

From: *Opportunities in AI/ML for the Rubin LSST*

The LSST Dark Energy Science Collaboration (DESC), 2026

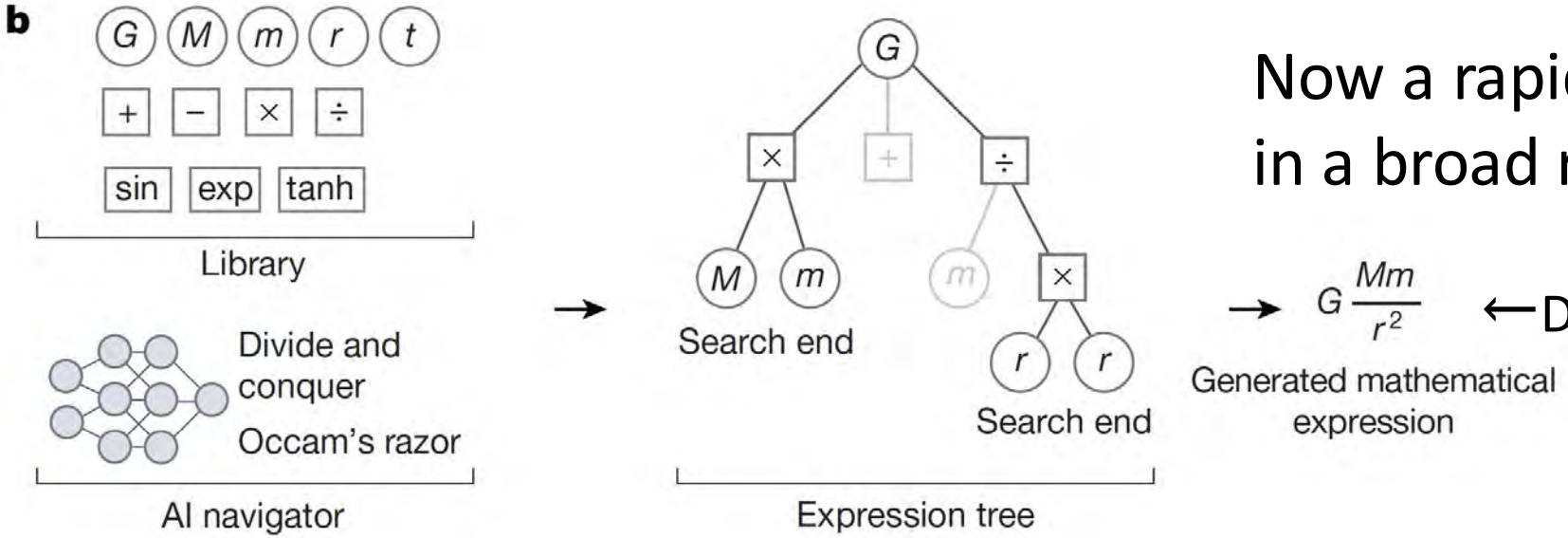
arXiv/2601.14235v1



AI for Science

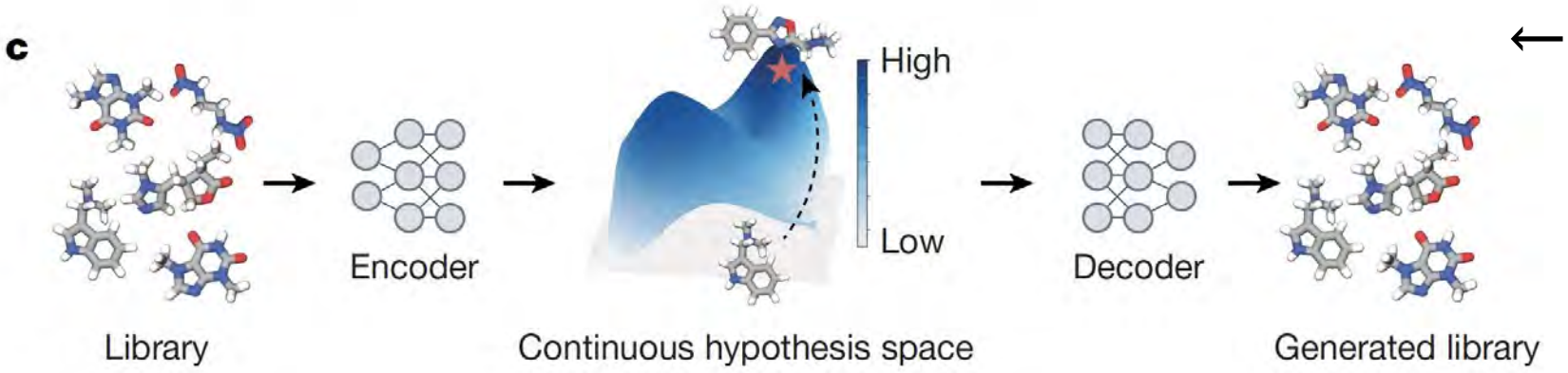
Scientific discovery in the age of artificial intelligence

Wang et al. Nature | Vol 620 | 3 August 2023 | **47** (many examples)



Now a rapidly growing methodology in a broad range of fields

←Discovery of analytical expressions, e.g, with symbolic regression



←New molecule/protein design in genomics, proteonomics, etc.

… and *many, many* others

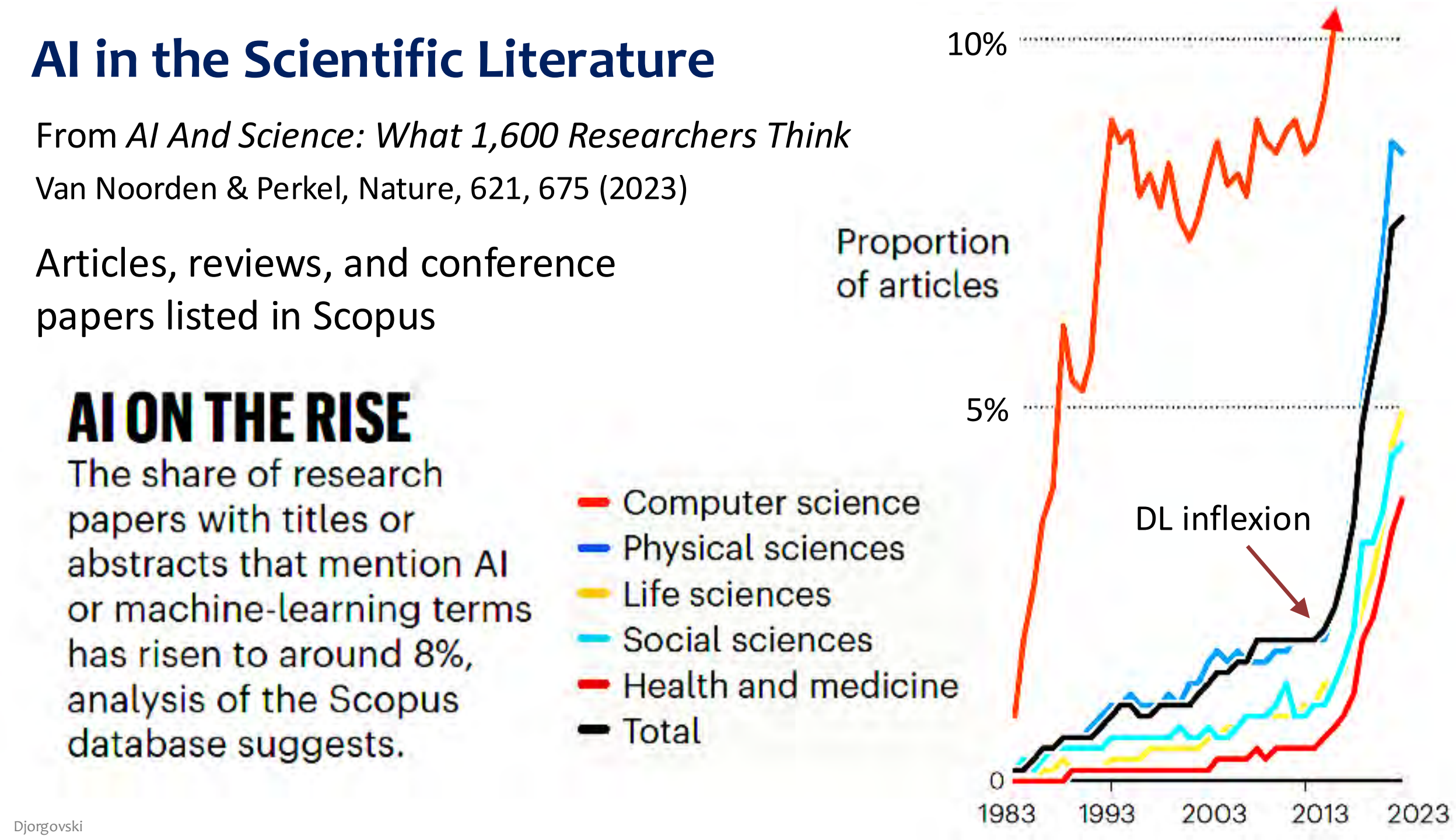
Djorgovski

AI in the Scientific Literature

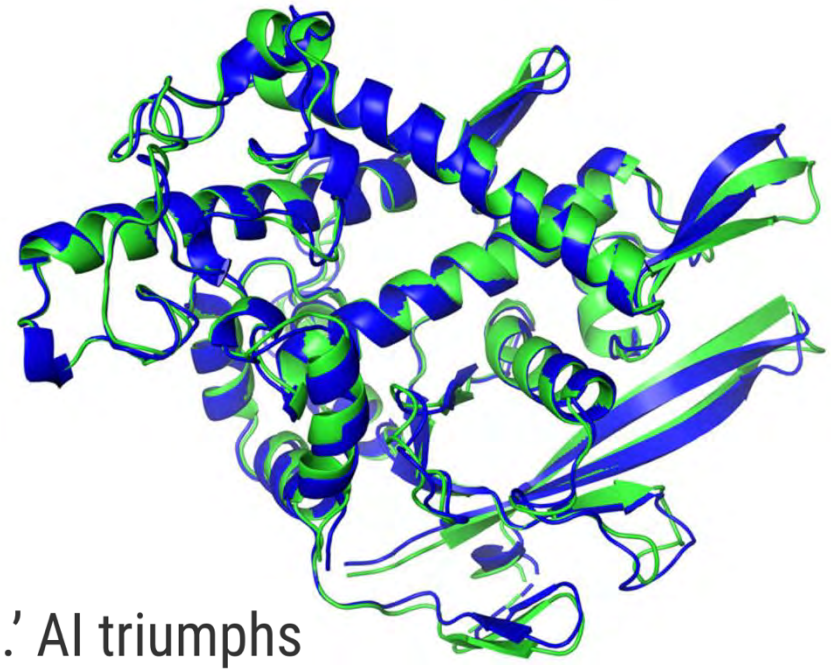
From *AI And Science: What 1,600 Researchers Think*

Van Noorden & Perkel, Nature, 621, 675 (2023)

Articles, reviews, and conference papers listed in Scopus



Now AI is making discoveries that *the humans could not*



DeepMind's AlphaFold

The New York Times Nov. 30, 2020

London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery

Researchers at DeepMind say they have solved "the protein folding problem," a task that has bedeviled scientists for more than 50 years.

Science 'The game has changed.' AI triumphs at solving protein structures

nature By Robert F. Service | Nov. 30, 2020 , 10:30 AM View all Nature Rese

Explore our content Journal information Subscribe

NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

Chemistry Nobel 2024

Demis Hasabis

John Jumper

Djorgovski

AI for Science Redux

Early science acceleration experiments with GPT-5

Sébastien Bubeck[1], Christian Coester[2], Ronen Eldan[1], Timothy Gowers[3], Yin Tat Lee[1], Alexandru Lupsasca[1,4], Mehtaab Sawhney[5], Robert Scherrer[4], Mark Sellke[1,6], Brian K. Spears[7], Derya Unutmaz[8], Kevin Weil[1], Steven Yin[1], Nikita Zhivotovskiy[9]

arXiv:2511.16072

AI models like GPT-5 are an increasingly valuable tool for scientists, but many remain unaware of the capabilities of frontier AI. We present a collection of short case studies in which GPT-5 produced new, concrete steps in ongoing research across mathematics, physics, astronomy, computer science, biology, and materials science. In these examples, the authors highlight how AI accelerated their work, and where it fell short; where expert time was saved, and where human input was still key. We document the interactions of the human authors with GPT-5, as guiding examples of fruitful collaboration with AI. Of note, this paper includes four new results in mathematics (carefully verified by the human authors), underscoring how GPT-5 can help human mathematicians settle previously unsolved problems. These contributions are modest in scope but profound in implication, given the rate at which frontier AI is progressing.

AI for Mathematics Research

AlphaGeometry, "Solving olympiad geometry without human demonstrations", Trinh et al., Nature 625, 476 (2024), approached performance of gold medalists at the International Mathematical Olympiad

AlphaGeometry2, "Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2", Chervonyi et al., arXiv:2502.03544, surpassed it.

Towards Autonomous Mathematics Research, Feng et al., arXiv:2602.10177

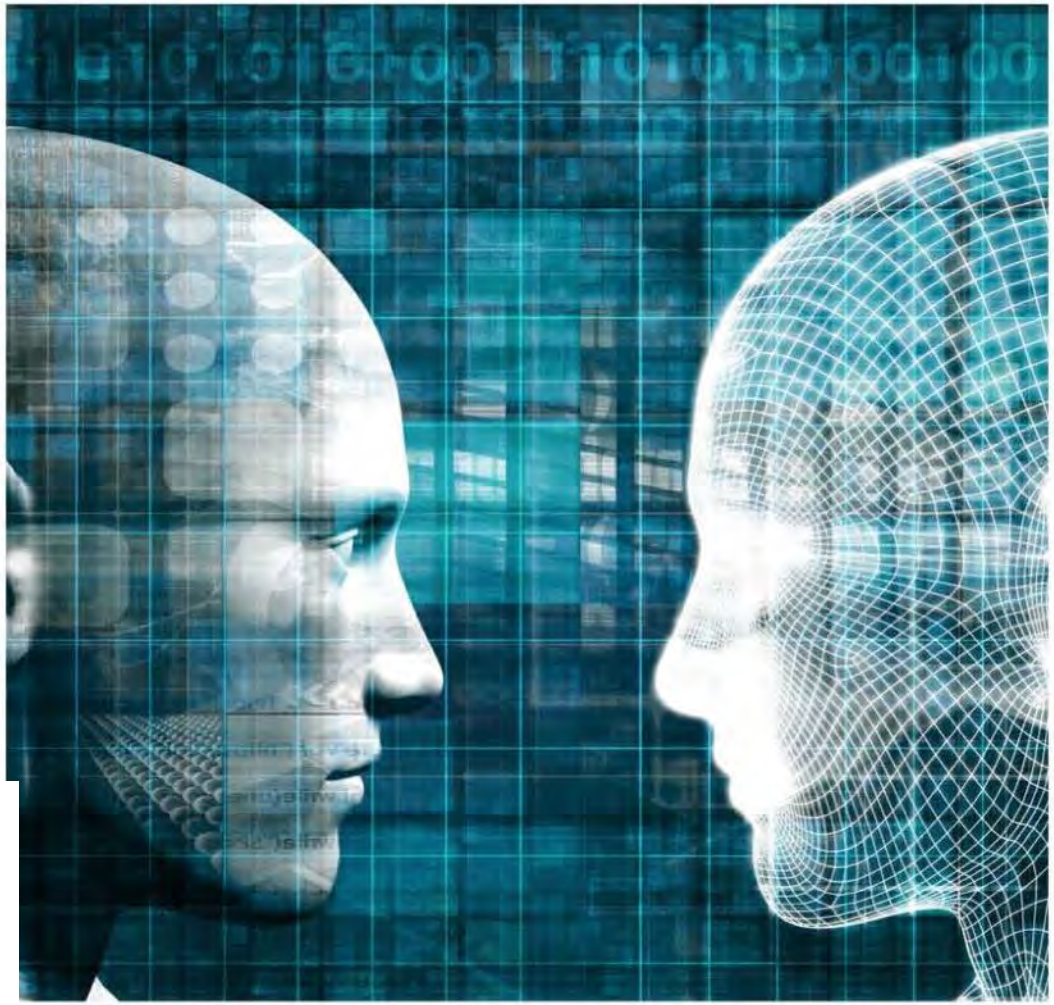
"… we introduce ***Aletheia***, a math research agent that iteratively generates, verifies, and revises solutions end-to-end in natural language … powered by an advanced version of Gemini Deep Think for challenging reasoning problems … intensive tool use to navigate the complexities of mathematical research. **Ultimately, we believe that AI will become a tool that enhances rather than replaces mathematicians** … formal verification systems are not yet capable of even formulating the questions of interest on most research frontiers … we have introduced specialized math reasoning agents, incorporating informal natural language verification…"

Carbon Brains Meet Silicon Brains

Like any other technology, Artificial Intelligence is **a tool that enhances human capabilities**

But it ***thinks differently***

It can help us explore and analyze highly complex, multidimensional data spaces, and discover new things



Machine Intelligence is essential for the exploration of complex data and concepts. It *augments our cognitive abilities*

We are entering the era of a Human-AI collaborative discovery

“Man-Computer Symbiosis”

J. C. R. Licklider (1915-1990), cognitive scientist, thinker, and DARPA manager to whom we owe the existence on the Internet

Man-Computer Symbiosis

J. C. R. LICKLIDER†

IRE Trans. on Human Factors in Electronics, v. HFE-1, pp. 4-11, 1960

"Man-computer symbiosis is an expected development in cooperative interaction between men and electronic computers.”

“It seems entirely possible that, in due course, electronic or chemical machines will outdo the human brain in most of the functions we now consider exclusively within its province … There will nevertheless be a fairly long interim during which the main intellectual advances will be made by men and computers working together in intimate association."

Matt Shumer, Feb. 9, 2026, https://shumer.dev/**something-big-is-happening**

"... in 2025, new techniques for building [AI] models unlocked a much faster pace of progress ... then it got even faster ... then faster again. Each new model ... was better by a wider margin, and the time between new model releases was shorter."

"On February 5th, OpenAI released GPT-5.3 Codex. In the technical documentation, they included this:

"GPT-5.3-Codex is ***our first model that was instrumental in creating itself***. The Codex team used early versions to debug its own training, manage its own deployment, and diagnose test results and evaluations."

Also on Feb. 5, Anthropic released Opus 4.6. Dario Amodei … says AI is now writing "much of the code" at his company, and ... we may be "only 1–2 years away from ***a point where the current generation of AI autonomously builds the next***."

"The AI helped build itself"