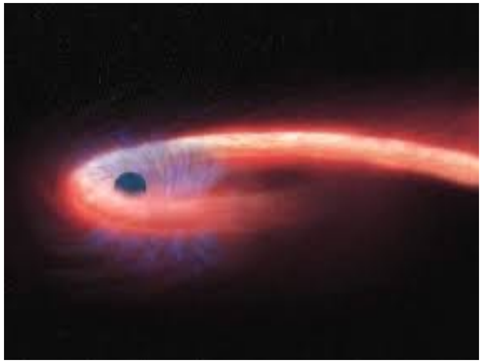
STFC Astronomy and AI Summer School
Canterbury (hybrid)
July 10th, 2024

Big Data in Astronomy

Contemplating a potential future

Matthew J. Graham
Research Professor
mjg@caltech.edu

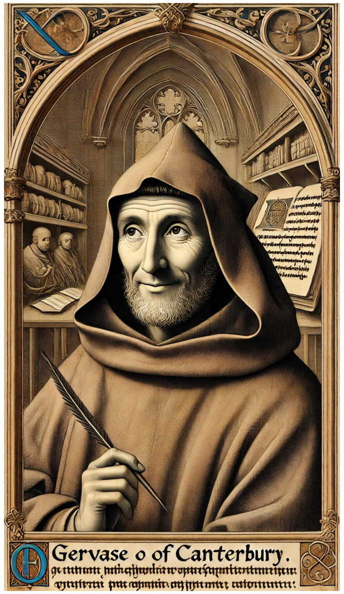


What is the impact of GenAI on astronomy?

- What is our current methodological approach?
- What are the type of problems we are facing, particularly the data deluge?
- What are our current techniques for dealing with these including the use of ML?
- What is wrong with this?
  - Are we limiting ourselves? What are the unexplored avenues?
  - Will the advent of GenAI (LLMs, etc.) make a difference?

An interesting event in Canterbury: 18/6/1178

"This year, on the Sunday before the Nativity of Saint John the Baptist, after sunset, at the first appearance of the moon, a miraculous sign appeared, witnessed by five or more men sitting opposite. For the new moon was bright, extending its horns to the east as is its nature; and behold, suddenly the upper horn was divided into two. From the middle of this division shot forth a burning torch, casting flames, coals, and sparks far and wide. Meanwhile, the lower part of the moon was twisted as if in distress, and, to use the words of those who reported this to me and saw it with their own eyes, the moon writhed like a struck snake. After this, it returned to its normal state. This change occurred twelve times or more, in such a way that the moon, as mentioned before, endured various torments of fire and then returned to its prior state. After these changes, from horn to horn, that is, along its length, it became half-blackened. Those men who saw this with their own eyes and reported it to me, who am writing this, were ready to pledge their faith or swear an oath that they added nothing false to the above account."

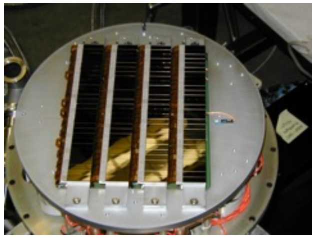
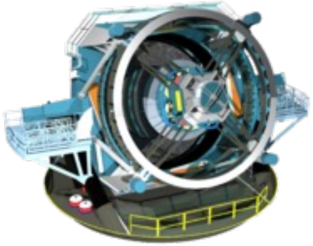


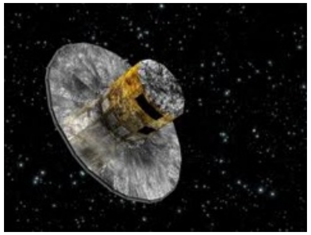
Billions of observations

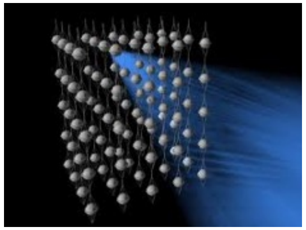
- Palomar-Quest Synoptic Sky Survey
- SDSS (Stripe 82)
- Catalina Real-time Transient Survey
- Palomar Transient Factory
- Zwicky Transient Factory
- Pan-STARRs
- SkyMapper
- ASKAP
- ThunderKat (MeerKAT)
- KEPLER
- GAIA
- LIGO
- IceCUBE
- LOFAR
- LSST
- SKA
- TESS
- ASAS-SN
- MASTER
- DES
- ATLAS
- BlackGEM

- GoTo
- MeerKAT
- ASKAP
- WISE
- OGLE
- DESI
- SDSS-V
- LAMOST

…





Billions of observations

- Palomar-Quest Synoptic Sky Survey
- SDSS (Stripe 82)
- Catalina Real-time
- Palomar Transient
- Zwicky Transient Fa
- Pan-STARRs
- SkyMapper
- ASKAP
- ThunderKat (Meer
- KEPLER
- GAIA
- LIGO
- IceCUBE
- LOFAR
- LSST
- SKA
- TESS
- ASAS-SN
- MASTER
- DES
- ATLAS
- BlackGEM

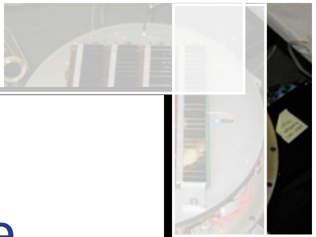
- SDSS-V
- LAMOST

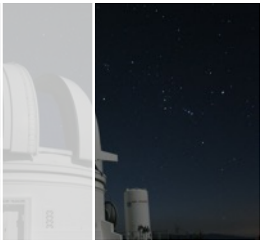
…

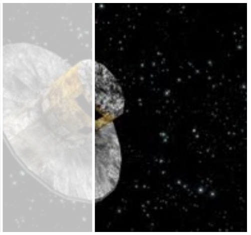
Multiple observations of the same astronomical source at different times and at different wavelengths

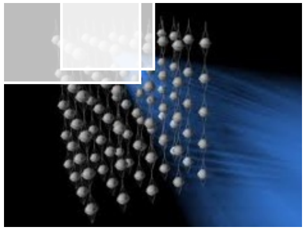
⇒ sparse multivariate time series

⇒ high volume, high complexity



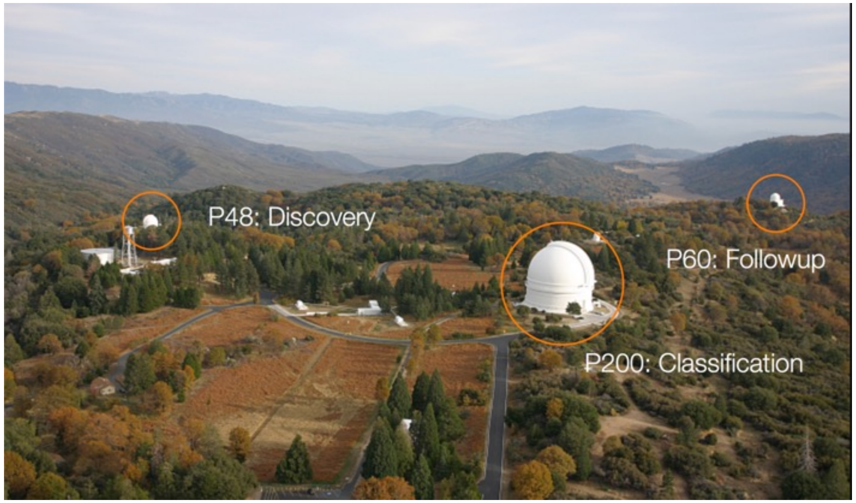


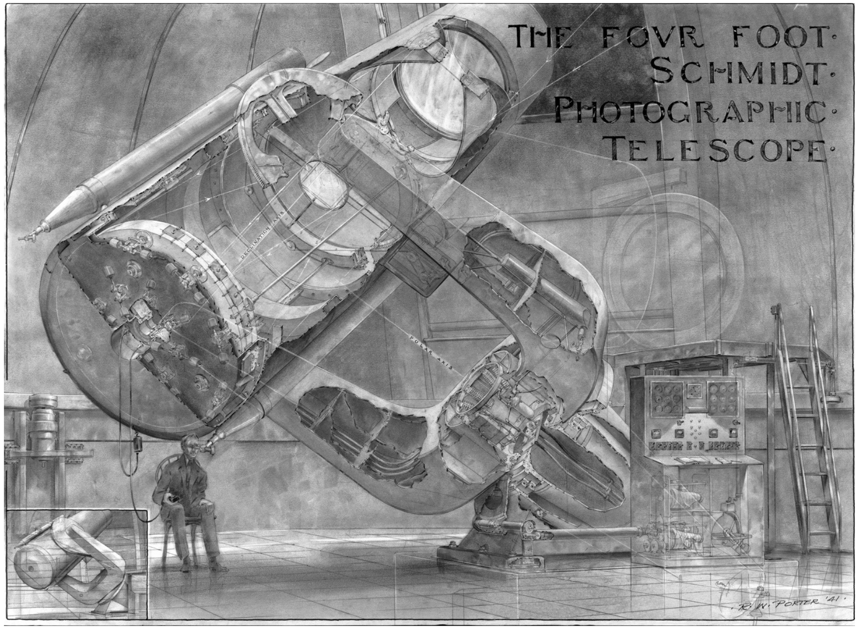




ZTF: the first industrial astronomical stream

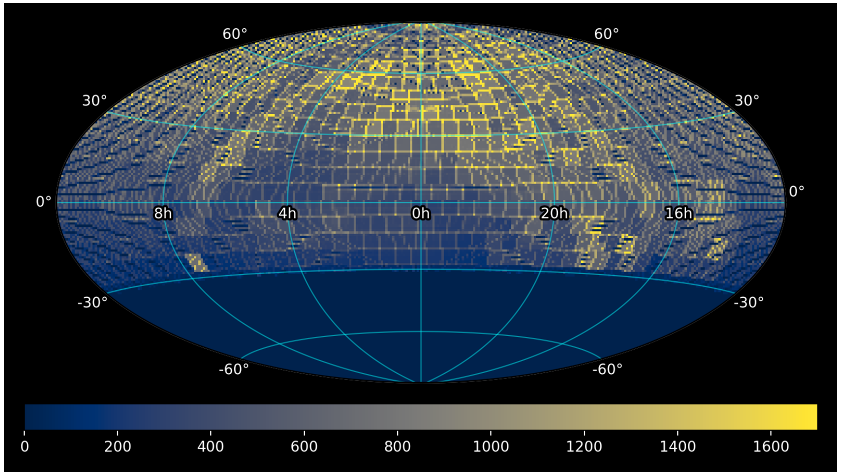
- The Palomar Oschin 48” telescope took its first image on Sept 30, 1948
- ZTF is its latest instrument: a 47 deg^2 field of view camera with > 660M pixels
- It can cover 3750 deg^2 / hr to 20.5-21 mag (30s exposures)
- It carries out a full northern sky every two nights in *g, r* (and *i*)
- First light was Oct 2017; survey started Mar 2018; first alerts issued Jun 2018
- Supports ToO programs for MMA
- ZTF is ~10% of Rubin LSST





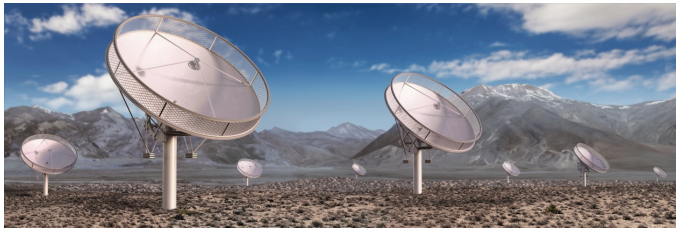
ZTF by the numbers

- 1.4 TB (compressed) of image data per night
- 6 PB over the past six years
- More than 1 million exposures taken – over 1 year of open shutter time
- Total sky area covered is 47 million deg^2 – 1.5 – 2x LSST 10-year total
- >787 billion photometric measurements for over 4.72 billion sources
- Over 1 billion sources have more than 50 data points in g and r
- Up to 1 million transient alerts per night
- Over 700 million alerts (56 TB) published



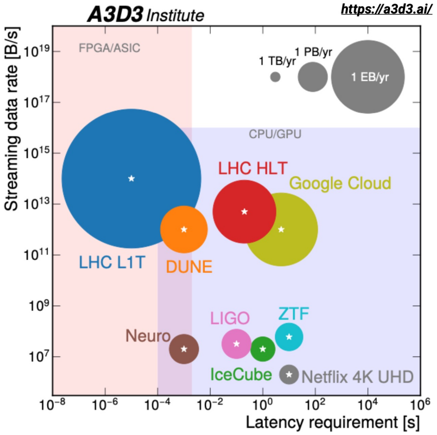
The next generation of surveys and facilities

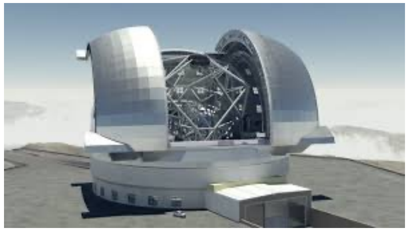
- LSST (2025): ~20 TB/day => 10 PB/yr
- CSST (2026): ~TBs/day => 10s PB/yr
- Roman Space Telescope (2027): 20 TB/day => 7.3 PB/yr
- ngVLA (2030s): 20 TB/day => 7.3 PB/yr
- SKA (2030s): ~1 PB/day => 300 PB/yr
- ELT (~2030): ~PBs/yr
- DSA-2000 (2028): 3.5 PB/day => 1.3 EB/yr (26 LHCs)
- DUNE (2028): => ~TBs/s => 1.8 EB/yr



For comparison:

- HL-LHC (2029): 700 TB/s => 1 EB/yr
- Facebook: 300 PB of data





What do we do with a billion time series?

Population behaviors

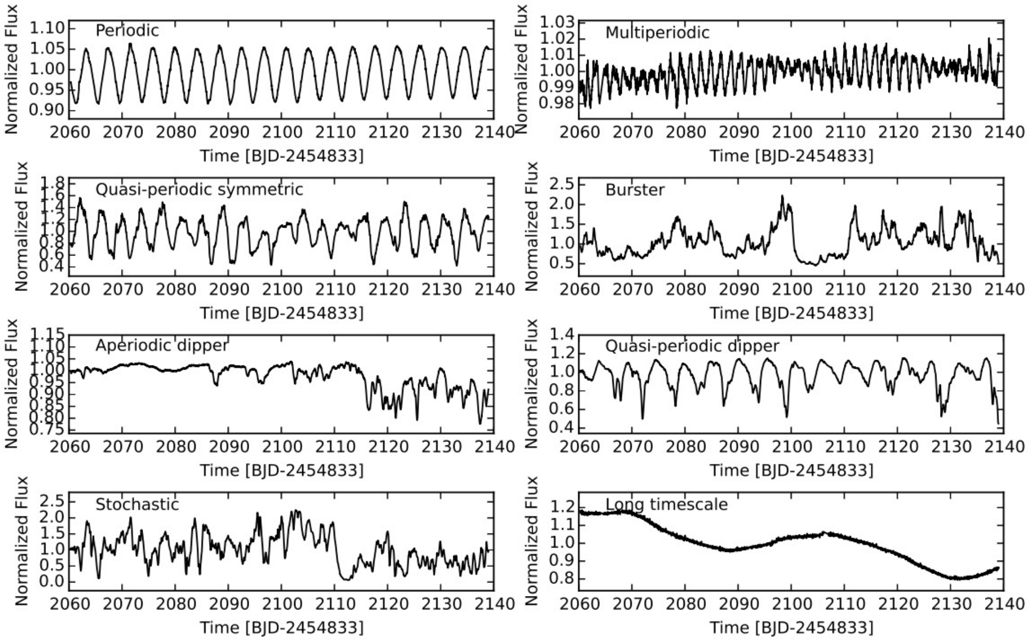
- Characterize
- Categorize
- Classify

Outliers

- Extreme sources
- Changes of behavior

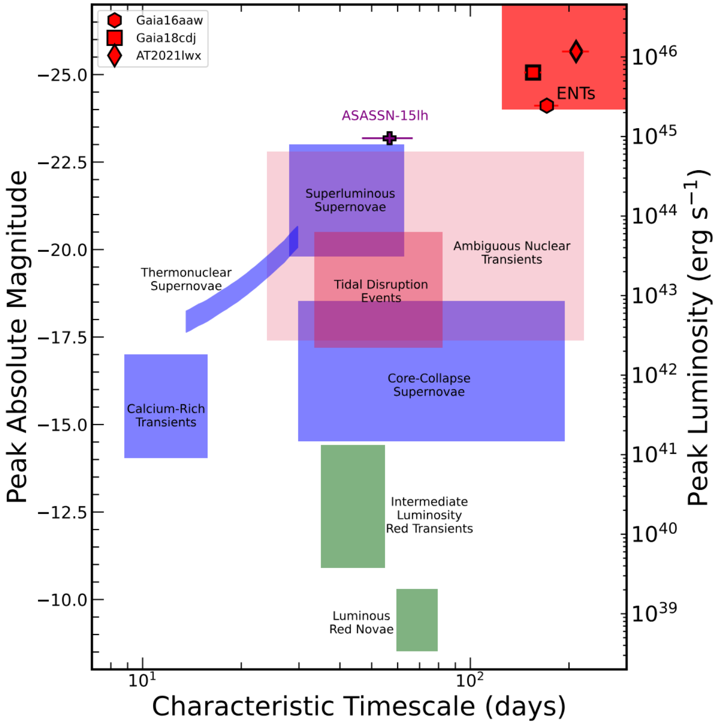
Models (physical/statistical)

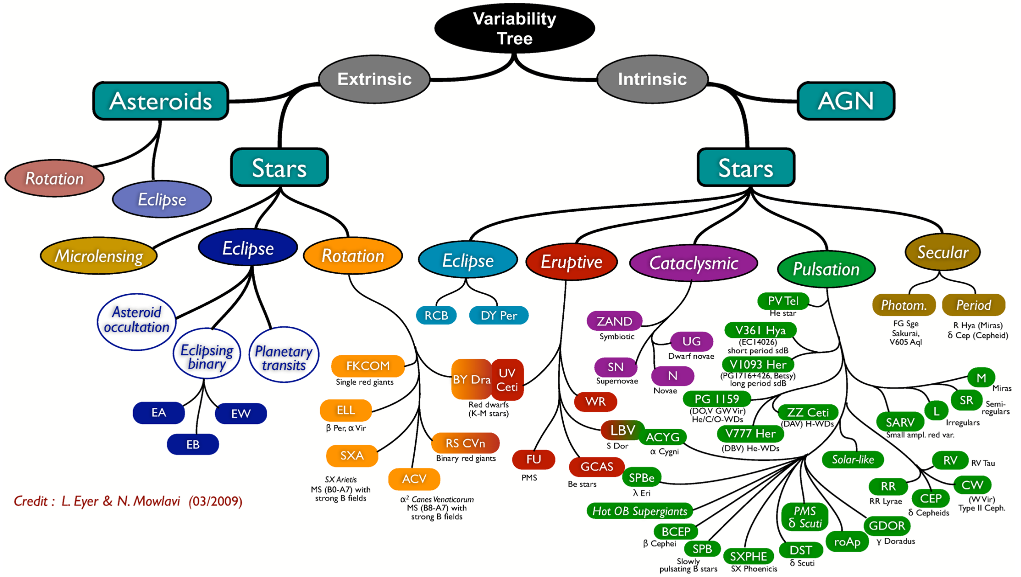
- Interpolation
- Forecasting



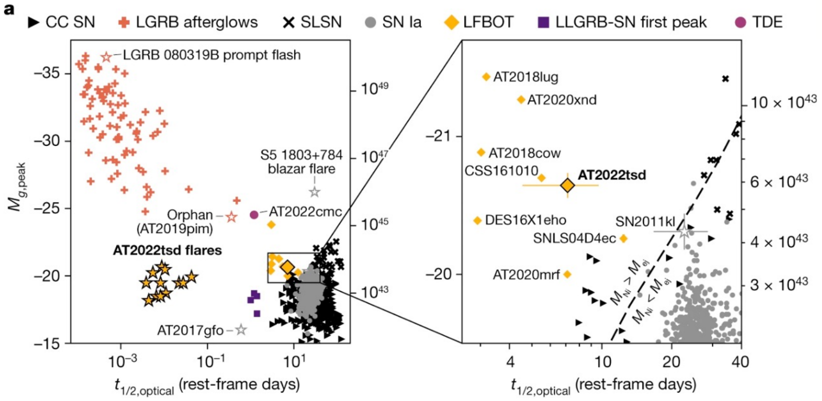
(Cody & Hillenbrand 2018)

Conceptual bases/biases





Credit : L. Eyer & N. Mowlavi (03/2009)

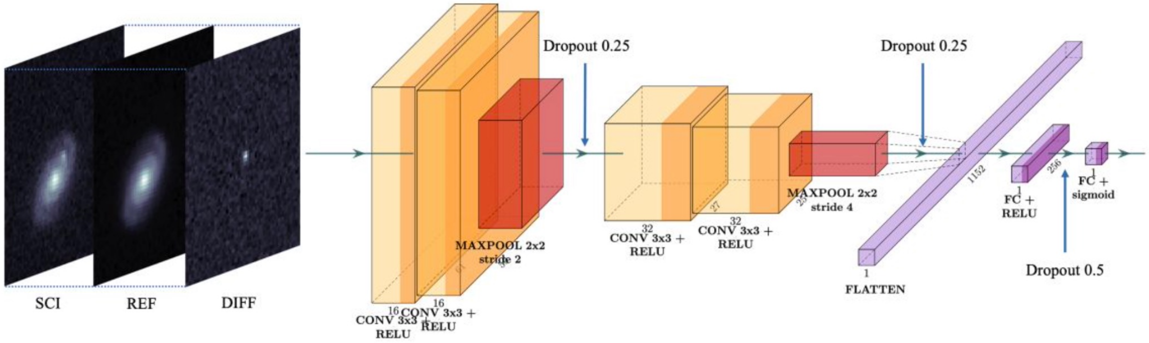


Make assumptions about the statistical nature of the data and the underlying physical processes that generate it

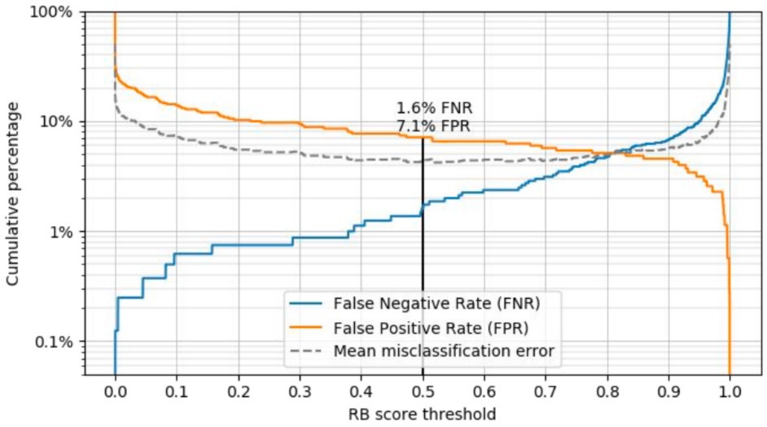
Our first human replacement: real/bogus

braai (Duev+ 19):

- Using 3 x 63 x 63 32-bit alert thumbnails: science, reference, difference
- VGG6 (312k parameter CNN) model for real-bogus classification



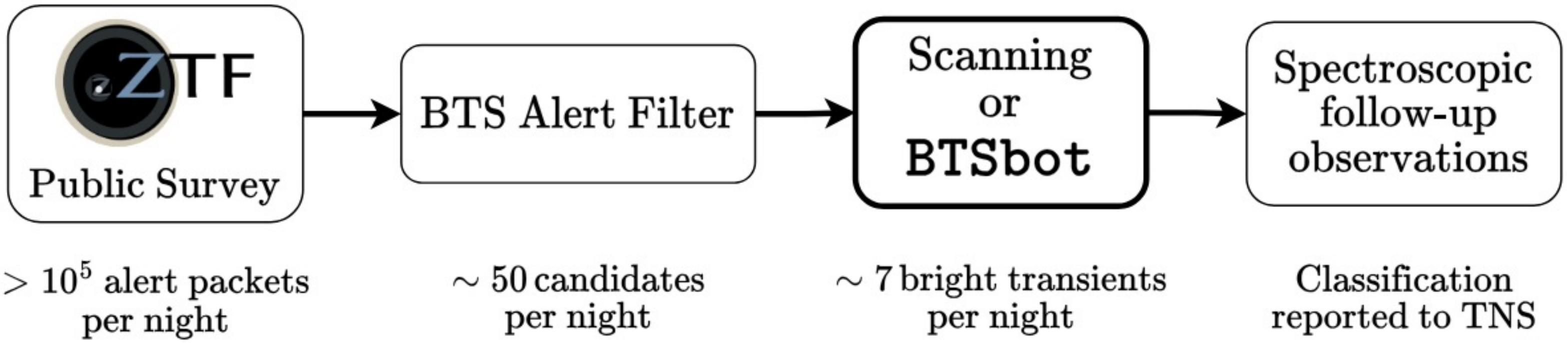
- State-of-the-art performance:



First end-to-end automation

BTSbot (Rehumtulla+ 24):

- Automatically submits reports of spectroscopically classified SN Ias to Transient Name Server (TNS):



ZTF Public Survey → BTS Alert Filter → Scanning or BTSbot → Spectroscopic follow-up observations

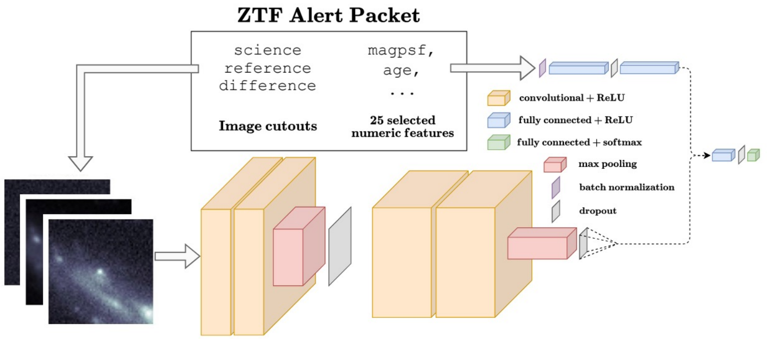
$> 10^5$ alert packets per night

$\sim$ 50 candidates per night

$\sim$ 7 bright transients per night

Classification reported to TNS

- >1000 sources saved by BTSbot
- >700 SEDM triggers sent
- >100 fully autonomously classified SN Ia
- A significant boost in survey efficiency



Don't let humans work with data

- The human brain is an amazing piece of bioengineering: connected to a 1 Gb/s network (nervous system), it offers an exaflop of computing power with 2.5 PB of storage with just 20 W of power
- The creative power is proven:

- However, our brains evolved for efficient tool-based survival in dry arid grasslands and not the 21st century data landscape
- Human decision theory is based on fight-fright-flight response
- The measured processing speed is ~60 bits/sec (mental arithmetic)
- And our own writings agree with us (24 km of text with 750 billion tokens)

Why are humans bad at decision making?

“Humans are not inherently `bad’ at decision-making, but there are several cognitive biases, limitations, and challenges that can sometimes lead to *less-than-optimal* decisions”:

- Cognitive Bias
- Emotional Influence
- Limited Information
- Time Constraints
- Heuristics
- Overconfidence
- Loss Aversion
- Groupthink
- Framing Effects
- Sunk Cost Fallacy

HM Government
UK TRANSITION
UK'S
NEW START
LET'S GET
GOING
The UK has left the EU. As we transition to our new relationship, we all need to take action to prepare.
Sign up now on gov.uk/transition
Check
Change
Go

Why should humans be taken out of the loop?

ChatGPT says that “there are several reasons for advocating for this”:

- Efficiency
- Safety
- Eliminating Bias
- Scalability
- Consistency
- Cost Reduction
- Rapid Decision Making

Why should humans be taken out of the loop?

ChatGPT says that “there are several reasons for advocating for this”:

- Efficiency
- Safety
- Eliminating Bias
- Scalability
- Consistency
- Cost Reduction
- Rapid Decision Making

A brief history of automated astronomy

1985 *Microcomputer Control of Telescopes* by Trueblood and Genet

1999 ROTSE detects first simultaneous GRB optical image

2002 RAPTOR is first fully autonomous closed loop robotic telescope

2006 VOEventNet + P48/PAIRITEL: the first (carefully) automated followup observation of a generic transient

2007 RoboNet + eSTAR

2008 CRTS begins – primary source of VOEvents

2012 LCOGT begins

2018 ZTF begins – era of industrial transient astronomy

2019 *Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning* by Ishida et al.

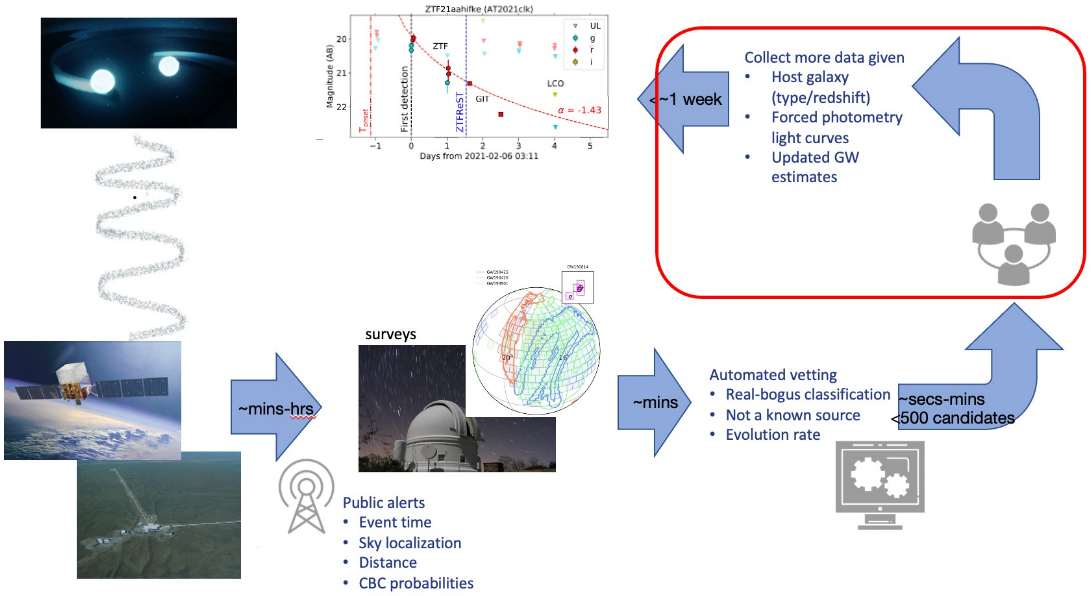
2022 1000th SNe detected with P48, spectra with SEDM, classified with SNIaScore, submitted to TNS => no humans in loop

2023 ZTF passes 600 million alerts

A brief introduction to reinforcement learning

- **State** – world observed by the agent: *ZTF transient light curves*
- **Action** – choices presented to the agent: *obtain follow-up observation*
- **Reward** – score the agent receives: *utility of follow-up observation*
- **Policy** – rule specifying action to take: *take observation with maximum reward*
- Goal – to learn Q – state-action value of policy π – or π

- Process needs to be:
  - Free from bias
  - Low latency
  - Scalable



Collect more data given
- Host galaxy (type/redshift)
- Forced photometry light curves
- Updated GW estimates

Public alerts
- Event time
- Sky localization
- Distance
- CBC probabilities

Automated vetting
- Real-bogus classification
- Not a known source
- Evolution rate

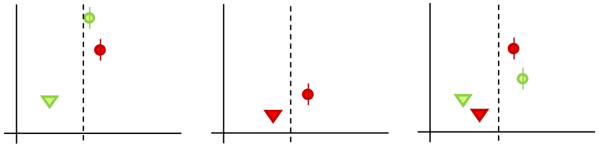
Pythia: a toy kilonova follow-up agent

RL agent that strategizes follow-up to identify kilonovae:

- Learns to evaluate the explore/exploit tradeoff
- Solves the credit assignment problem form any delayed consequences
- Adapts to new information from its own actions or other sources

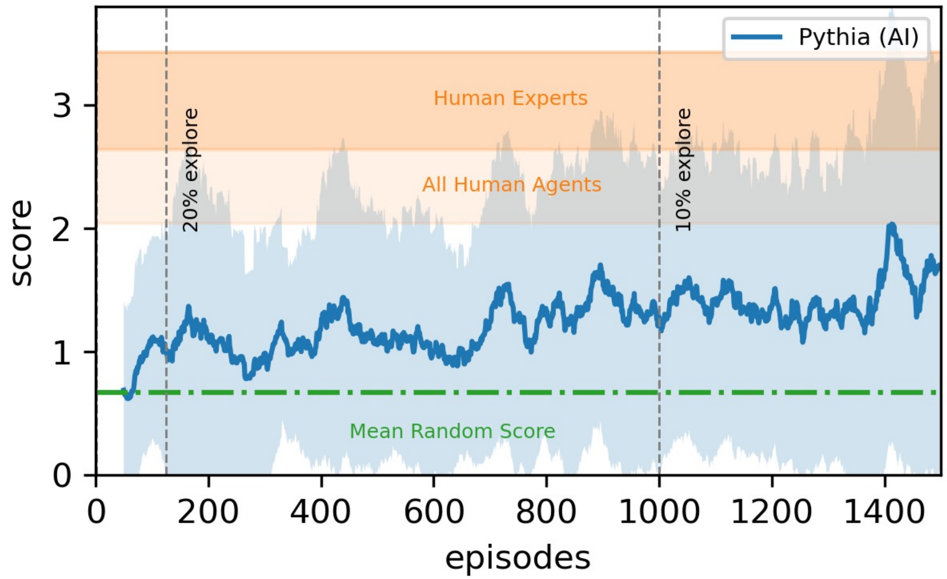
Toy sequential decision making under uncertainty problem:

- 9 transients, one of which (always) is true kilonovae (min photometry = 1)
  - Contaminants are SNe, unassociated GRB afterglows, shock breakout (do not include observation significance)
- Followup in ZTF g, r, or I (300s exposure) per day
  - Finite horizon – 6 days (no action on day 1)
- Reward 1 if agent adds data to the kilonova else 0
  - Maximize the number of followup to the true kilonova (non-model specific objective with the expectation that more data ~ better constraints)



(Sravan+ 2023)

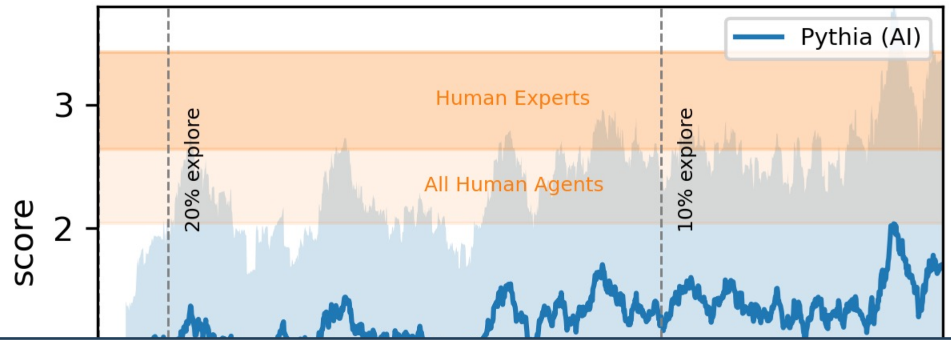
Pythia vs humans



Sravan+ 2023

agent	score	frac KN > 1 follow-up
Pythia	**1.84**	**0.81**
Non-expert 1	2.04	0.54
Non-expert 2	3.15	0.86
Expert 1	2.64	0.76
Expert 2	2.74	0.78
Expert 3	2.94	0.72
Expert 4	**3.43**	**0.9**

Pythia vs humans

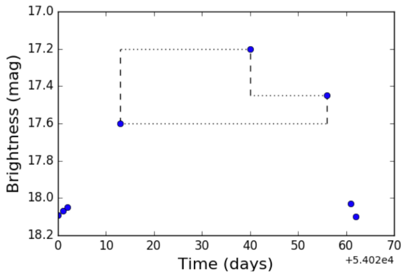


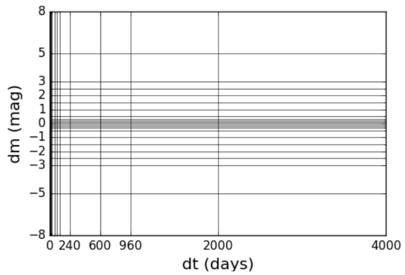
Optimized follow-up is a learnable problem

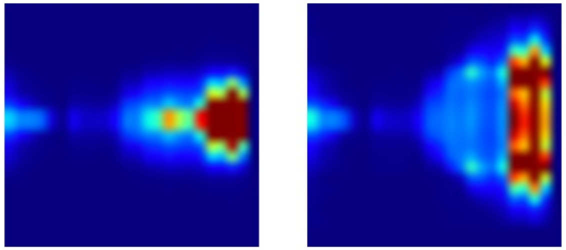
Non-expert 2	3.15	0.86
Expert 1	2.64	0.76
Expert 2	2.74	0.78
Expert 3	2.94	0.72
Expert 4	**3.43**	**0.9**

Baby steps to automated discovery

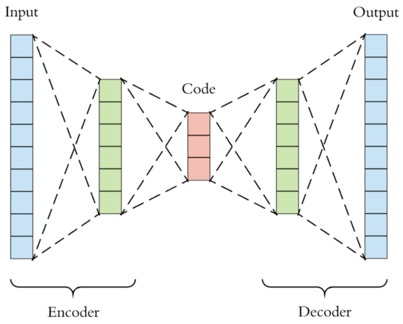
- Alternate data representations – are these more optimal?



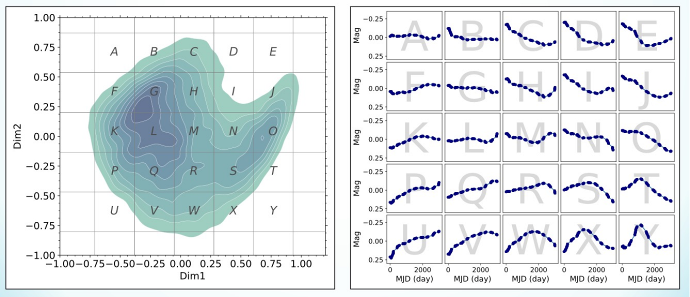




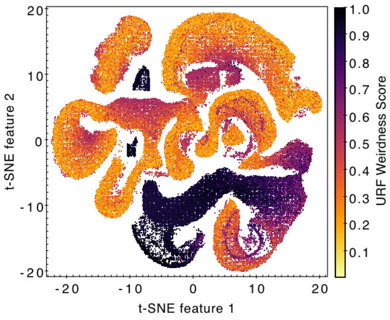
- Dimensionality reduction – learnt representations

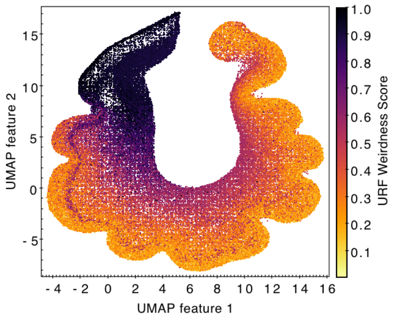


The latent space gives a learnt lower-dimension version of the input data



- Unsupervised categorization
  - T-SNE and UMAP are dimensional reduction techniques that provide low dimensional mappings of high dimensional data whilst retaining topological information





The promise of multi-modal foundation models

- Large models pre-trained on vast amounts of data in a self-supervised manner
- Natural language interfaces for queries, explanations, writing and coding assistance
- Current astronomical application to solve the representation challenge:
  - Large images with varying dynamic ranges and complex multi-variate time series are reduced to a lower dimensional representation (token) that can then be processed by a downstream model (transformer-based architecture)
  - Vector embeddings of different data modalities: alerts, images, spectra, time series – that allow cross-modal analysis
  - Fine tuning for specific science cases

What about automated scientific discovery?

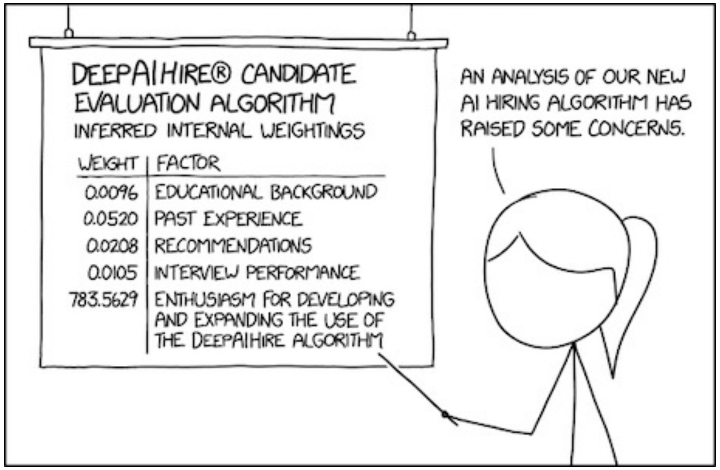
- Neural networks learn mappings between input and output data sets:

Universal approximation theorem (Hornik, Stinchcombe & White 1989):

A neural network with a single hidden layer and non-linear activation functions can represent *any* borel-measurable function

- These have traditionally been black boxes and explainable AI attempts to tell us what is going on

- What if the system could take two data sets, derive an analytical expression that links the two, and then explain it in natural language?


DEEPAIHIRE® CANDIDATE EVALUATION ALGORITHM
INFERRED INTERNAL WEIGHTINGS
WEIGHT | FACTOR
0.0096 | EDUCATIONAL BACKGROUND
0.0520 | PAST EXPERIENCE
0.0208 | RECOMMENDATIONS
0.0105 | INTERVIEW PERFORMANCE
783.5629 | ENTHUSIASM FOR DEVELOPING AND EXPANDING THE USE OF THE DEEPAIHIRE ALGORITHM
AN ANALYSIS OF OUR NEW AI HIRING ALGORITHM HAS RAISED SOME CONCERNS.

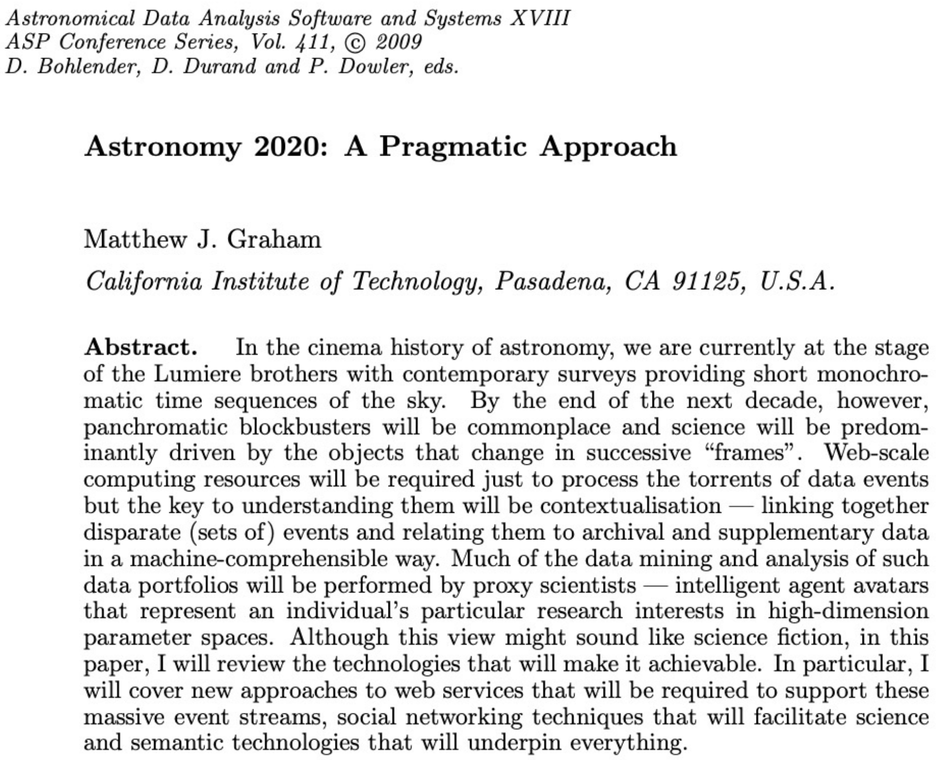
What about automated scientific discovery?

- Symbolic regression is a technique that derives the optimal analytical expression(s) for a data set (see Graham+ 2013, Cranmer+ 2020 for astronomy application; Udrescu+ 2020 for more physics)
- Consider mapping from data $\boldsymbol{x}_i$ to some variable z and then model it as:

$$z = f\left(\sum g(\boldsymbol{x}_i)\right)$$

  where f and g are trained neural networks.
- We can then fit g and f using SR (and with a much smaller subset of $\boldsymbol{x}_i$ than NN training)
- Data can be subset to test out of content applicability/generality
- Finally we can pass the analytical expressions to a LLM for description

A dream of the future (from 2009)



Astronomical Data Analysis Software and Systems XVIII
ASP Conference Series, Vol. 411, © 2009
D. Bohlender, D. Durand and P. Dowler, eds.

Astronomy 2020: A Pragmatic Approach

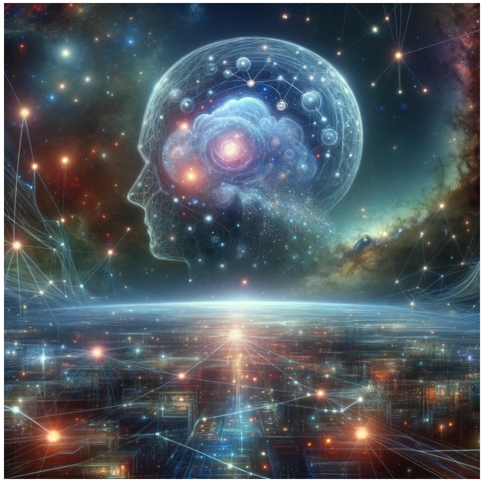
Matthew J. Graham

California Institute of Technology, Pasadena, CA 91125, U.S.A.

Abstract. In the cinema history of astronomy, we are currently at the stage of the Lumiere brothers with contemporary surveys providing short monochromatic time sequences of the sky. By the end of the next decade, however, panchromatic blockbusters will be commonplace and science will be predominantly driven by the objects that change in successive "frames". Web-scale computing resources will be required just to process the torrents of data events but the key to understanding them will be contextualisation — linking together disparate (sets of) events and relating them to archival and supplementary data in a machine-comprehensible way. Much of the data mining and analysis of such data portfolios will be performed by proxy scientists — intelligent agent avatars that represent an individual's particular research interests in high-dimension parameter spaces. Although this view might sound like science fiction, in this paper, I will review the technologies that will make it achievable. In particular, I will cover new approaches to web services that will be required to support these massive event streams, social networking techniques that will facilitate science and semantic technologies that will underpin everything.

2014: LSST will produce 100 GB/night
2020: SKA

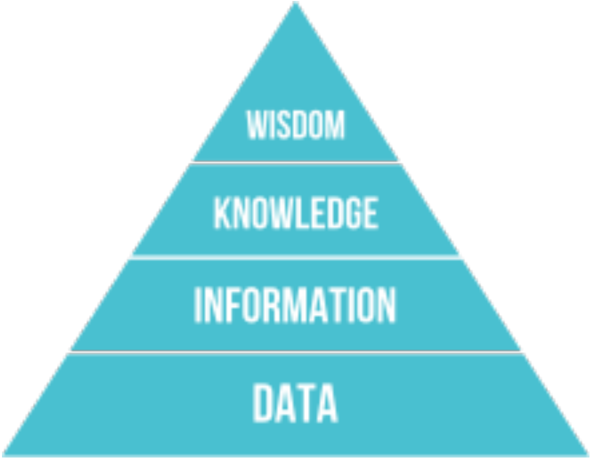
"We will wake to the Universe Today, summarizing the changes in position, flux/spectra, and new observations of billions of objects within the past 24 hours"



"Data exploration, visualization, and analysis will occur in virtual spaces with [agent systems] mediating between us and the data…through textual, verbal and gestural communication"

A dream of the future: the 2024 version

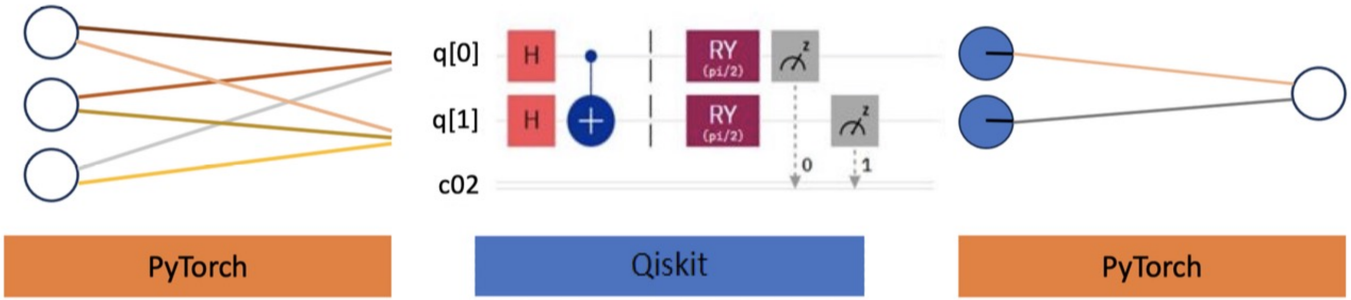
- TBs to PBs of data per night produced by facilities with fast low-latency high throughput inferencing models (embedded ML) driving control and decision systems
- Information extracted (optimized representations) and followup decisions made according to a teleological learnt strategy
- Patterns and relationships identified and put into context with other
- If science is defined by continuous differentiable relationships then automated discovery becomes increasingly more effective



=> “You will wake up and your smartphone will ~~tell~~ explain to you what **it** discovered last night”

Venturing into non-classical realms

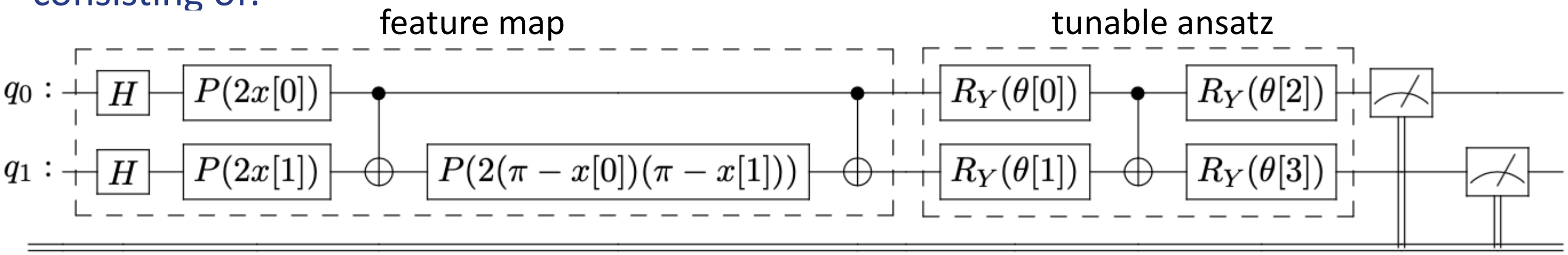
- Qubits ($|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$) operate in a high dimensional Hilbert space: n qubits describe a 2^n space
- Superposition and entanglement operations have no classical equivalence: more efficient exploration and more complex correlations
- Forget quantum supremacy, *quantum advantage* is performing tasks faster or more efficiently than classical computers
- Current Noisy Intermediate Scale Quantum (NISQ) computing involves systems with few to moderate (<500) qubits
- NISQ systems are also typically hybrid classical-quantum with classical computer handling optimization and measurement



Quantum Machine Learning (AstroQML)

qbraai (Abani+, in prep.)

- Variational Quantum Circuits (VQCs) are the equivalent of a traditional neural network consisting of:



- Classic optimization minimizes cost function/expectation value based on ansatz parameters

- Challenge to find optimal VQC architecture (quantum kernel)

Model	Input data	Training epochs	Training time (s)	Accuracy
braai	63x63x3	5	89	79.8%
VQC	63x63x3	5	23	73.4%
braai	63x63x3	100	6950	96.7%
VQC	63x63x3	100	1757	69.9%
braai	28x28x3	5	192	77.9%
VQC	**28x28x3**	**5**	**20**	**95.6%**

Quantum advantage erat demonstrandum!

Speculation

- Is it all hype? There is still no astronomical discovery that could not have been made without machine learning
- Would we trust/believe an artificial discovery?
  - Would it be subject to stricter tests/controls than human discovery?
- Is discovery constrained to what we can understand?
  - Evolutionary circuit design that relied on amplified radio signals from nearby PCs that were stable over the 2 ms sampling period
- Is there a new/alternate maths waiting to be discovered that makes better sense of the universe as a scientific language?
- Can AI identify other intelligences via technosignatures or unrecognized signals?