Astronomy and Airlines: Expanding our horizons with Newton and GCRF

James Mullaney
The University of Sheffield, UK

Krzysztof Ulaczyk
Warwick University, UK

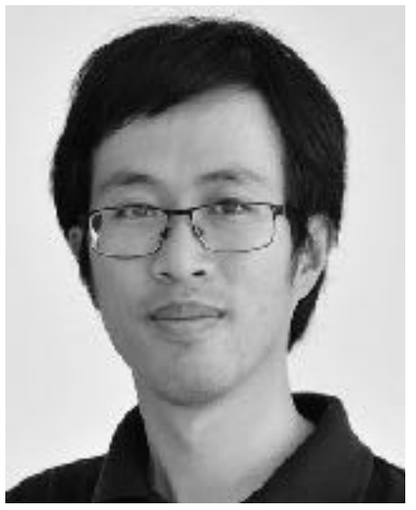
Tossapon Boongoen

MFU

Natthakan Iam-On

Utane Sawangawit

NARIT



Supachai Awiphan

Our Newton and GCRF “Big Data” projects

Since 2017:

- STFC/NARIT-Newton: Training Thai students and researchers in Big Data Analytics and Management.
- GCRF: Working with Thai businesses and organisations to address their Big Data needs.

Science & Technology Facilities Council

NARIT

GCRF Global Challenges Research Fund

Newton Fund

THANAPIRIYA PUBLIC COMPANY LIMITED

AirAsia

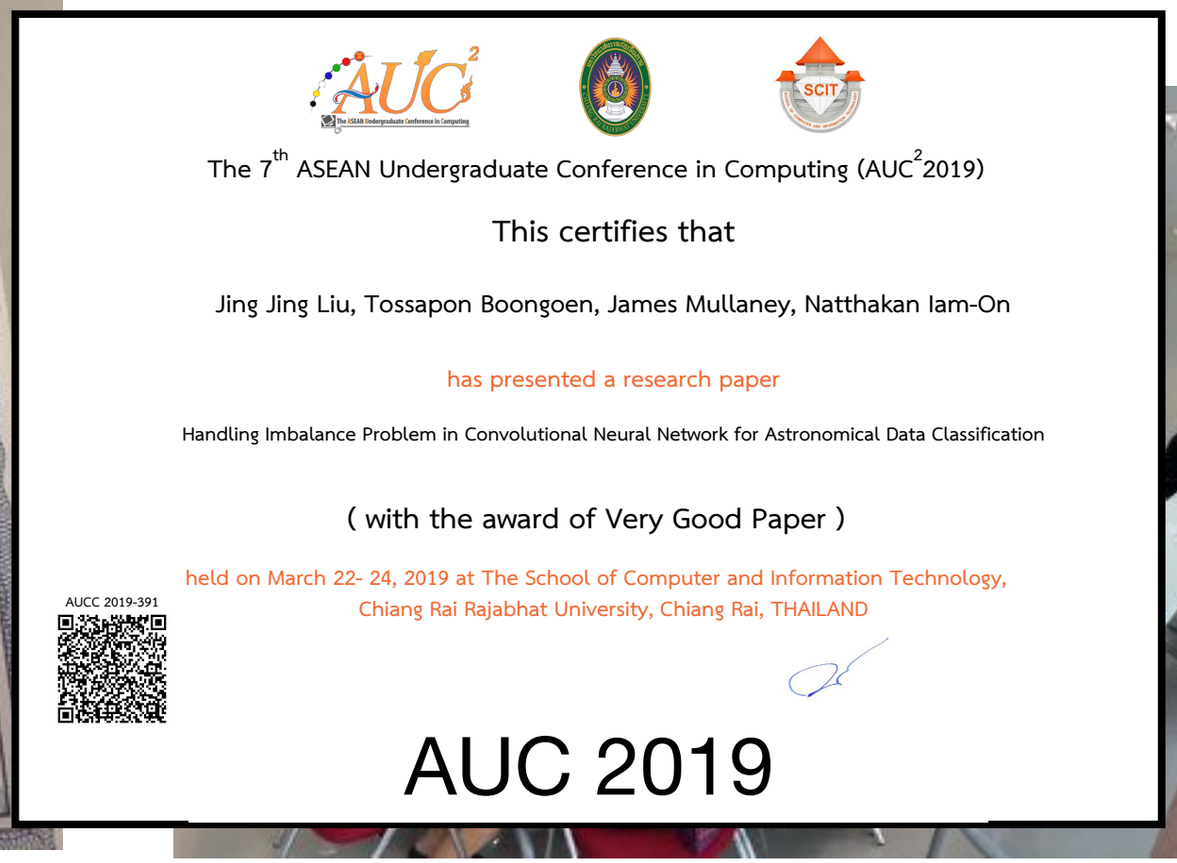
The University Of Sheffield.

Our Newton and GCRF “Big Data” projects

Since 2017:

- STFC/NARIT-Newton: Training Thai students and researchers in Big Data Analytics and Management.
- GCRF: Working with Thai businesses and organisations to address their Big Data needs.

ICMLC 2018



AUC 2019

Student involvement

Aireen Tabacolde

Rattanapong Yoyponsan

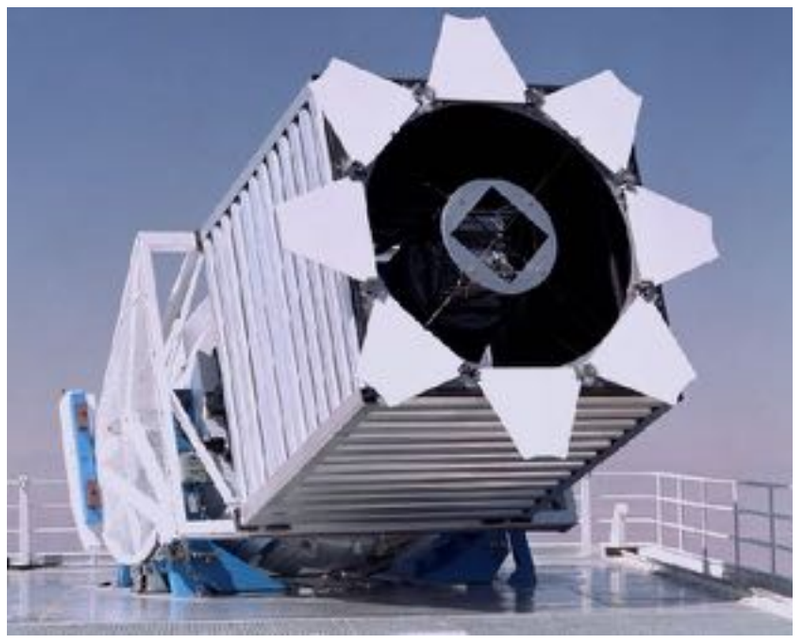
Terry Cortez

Jingjing Liu

My experience with Big Data

Prior to 2017:

- User of galaxy survey data, including SDSS.
- Correlations within datasets, typically consisting of 10^4-10^6 “rows”.
- Statistical techniques, incl. Maximum Likelihood, Bayesian, MCMC.



SDSS telescope

Post-2017:

- “Generator” of astronomy survey data;
- Consideration of data management;
- Machine learning techniques;
- Recently, cloud computing (AWS).

What is an astronomical survey?

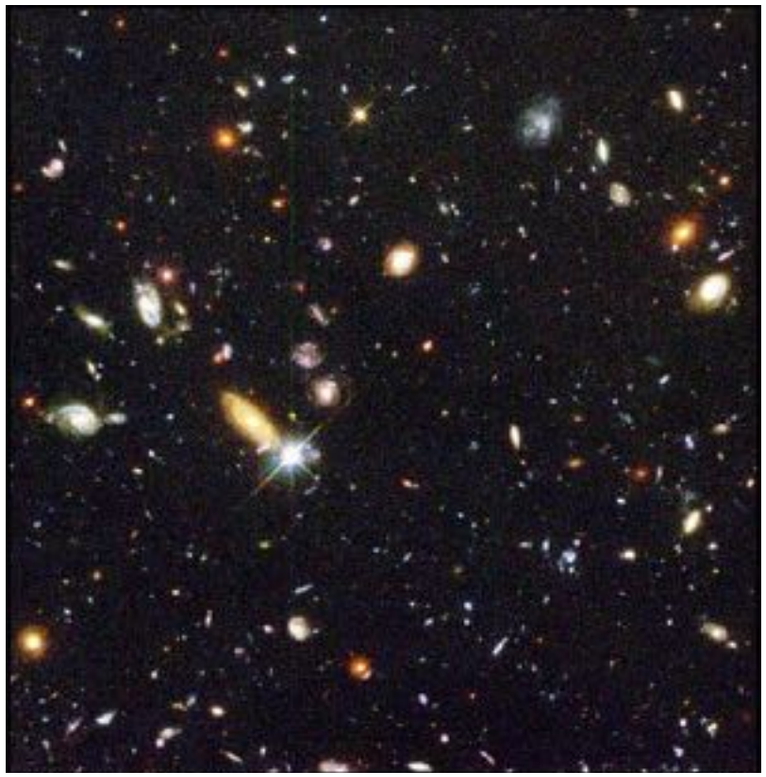
Most loosely:

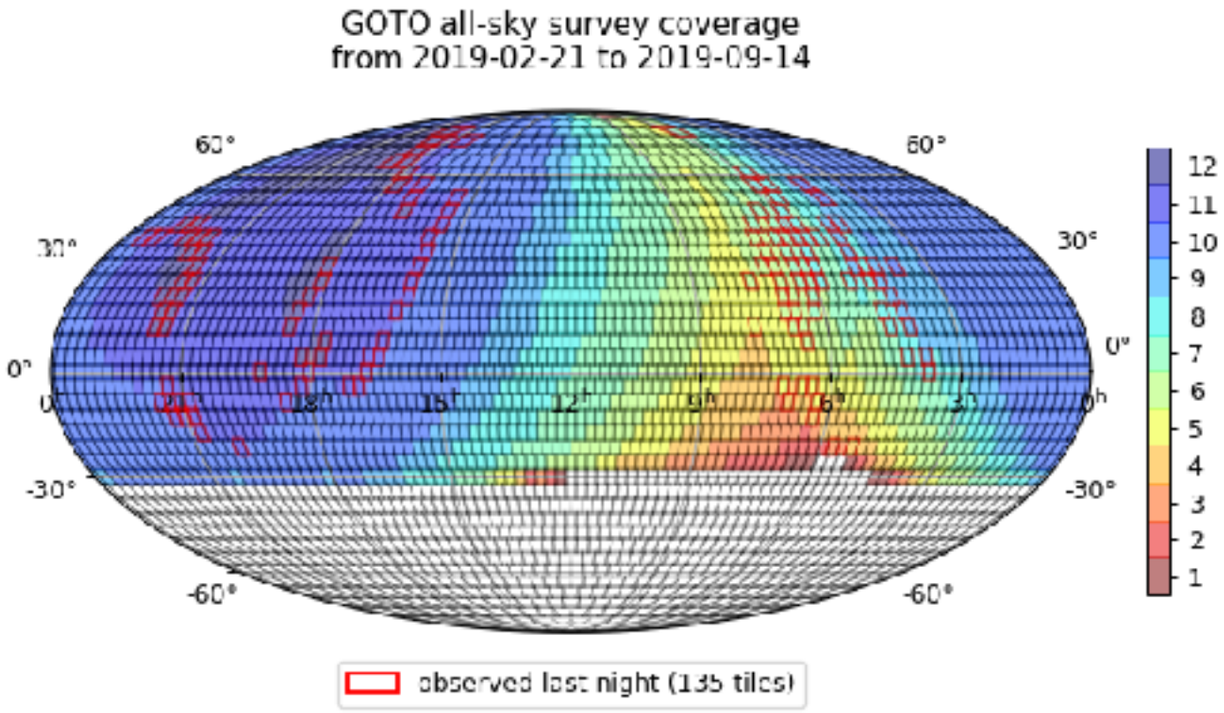
Any set of observations of a sample of astronomical sources (stars/galaxies/planets) that have some common feature.

For the context of this talk:

A set of observations of a contiguous area of sky, measuring and cataloguing every detectable source of “light” within that area.

Hubble deep field



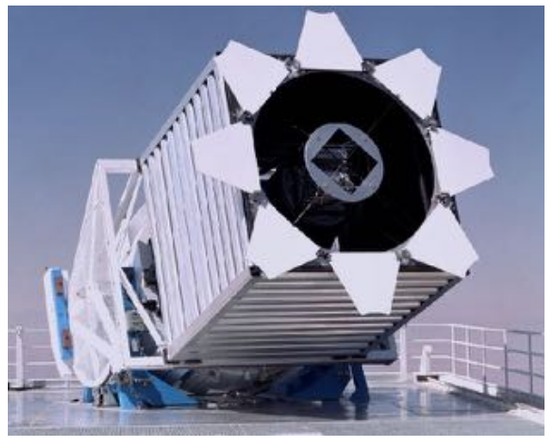


Optical astronomical surveys: the state of the art

Sloan Digital Sky Survey (2000-today):

Largely single-epoch coverage of the sky

10^6 digital images, catalogues 1 billion sources.



PanSTARRS (2008-today):

Largely single-epoch coverage of the sky

Catalogue of 3 billion (DR1) sources.

PTF and ZTF (2009-today):

Repeated coverage of the sky

$3x10^6$ images, catalogue of 22 billion detections

Large Synaptic Survey Telescope (2020-):

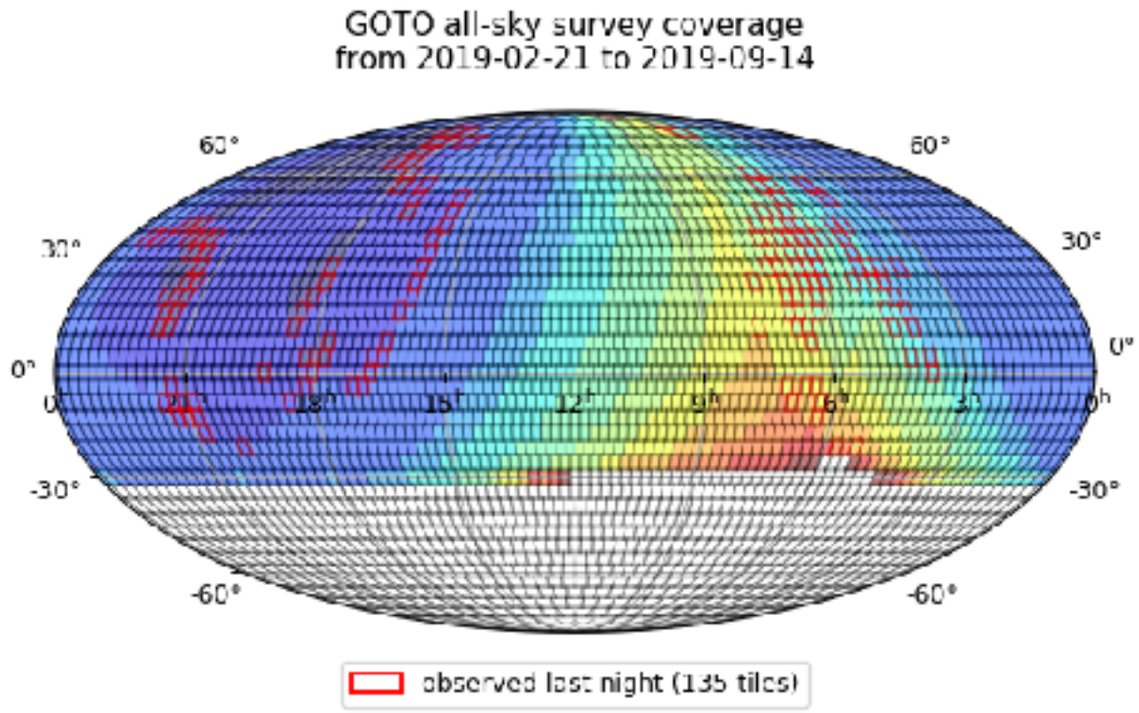
Repeated coverage of the sky

$6x10^6$ images, catalogue of 7 trillion detections

GOTO survey

Gravitational-wave Optical Transient Observer

Surveys the entire observable night sky roughly once every 2-3 weeks.



Collaboration of: Warwick, Monash, Sheffield, Leicester, Armagh, NARIT, Turku, Manchester

GOTO's 3Vs of Big Data

Delivers around 1000x50MB = 50GB of images *per night*.

Volume

Each image contains around 3×10^4 sources; corresponds to 10^7 measurements *per night*.

Velocity

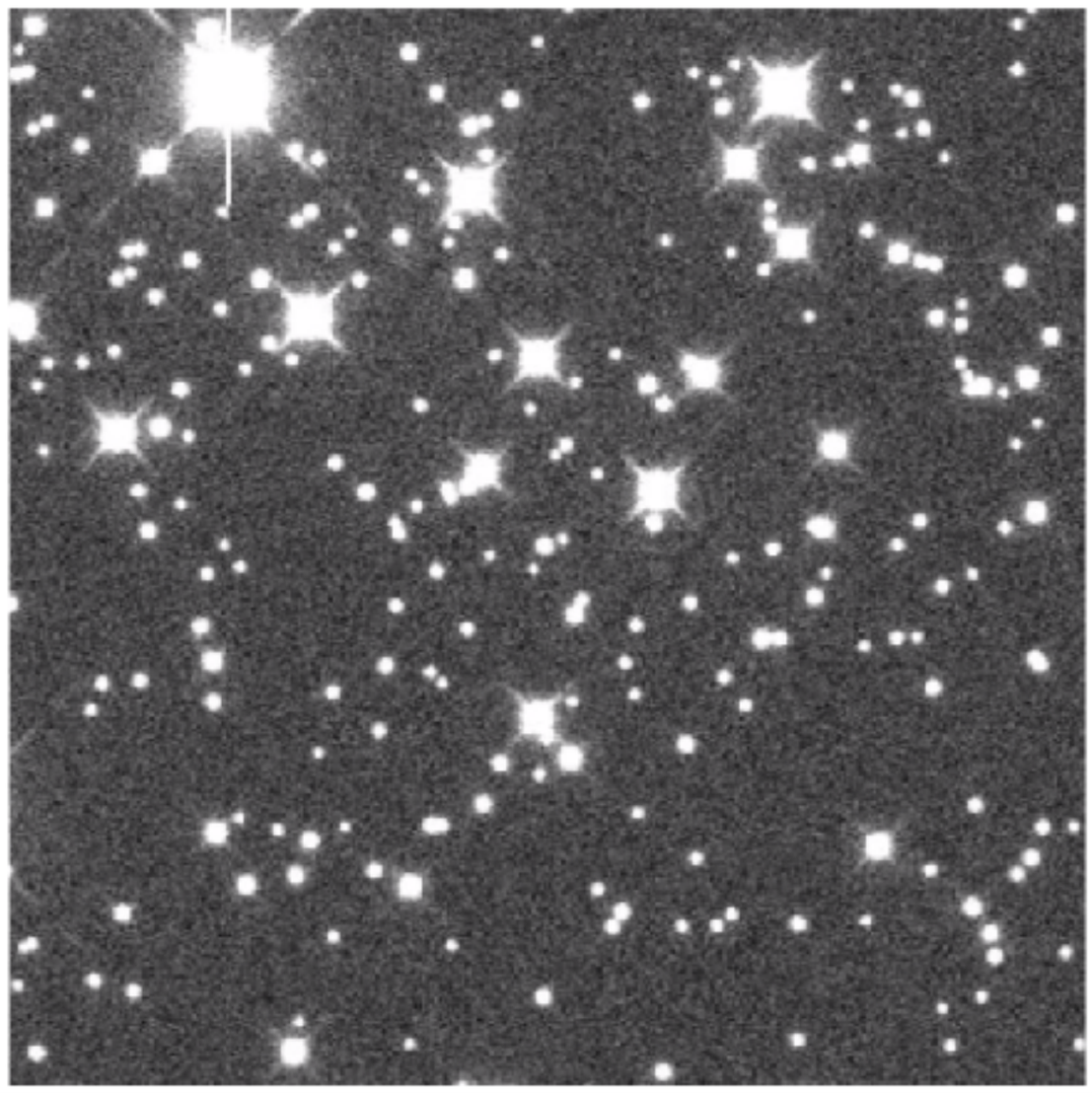
Each source could be a galaxy, a star, an asteroid, a planet, or something spurious (CCD defect, cosmic ray, etc.)

Variety

Part 1:
Needles in Haystacks

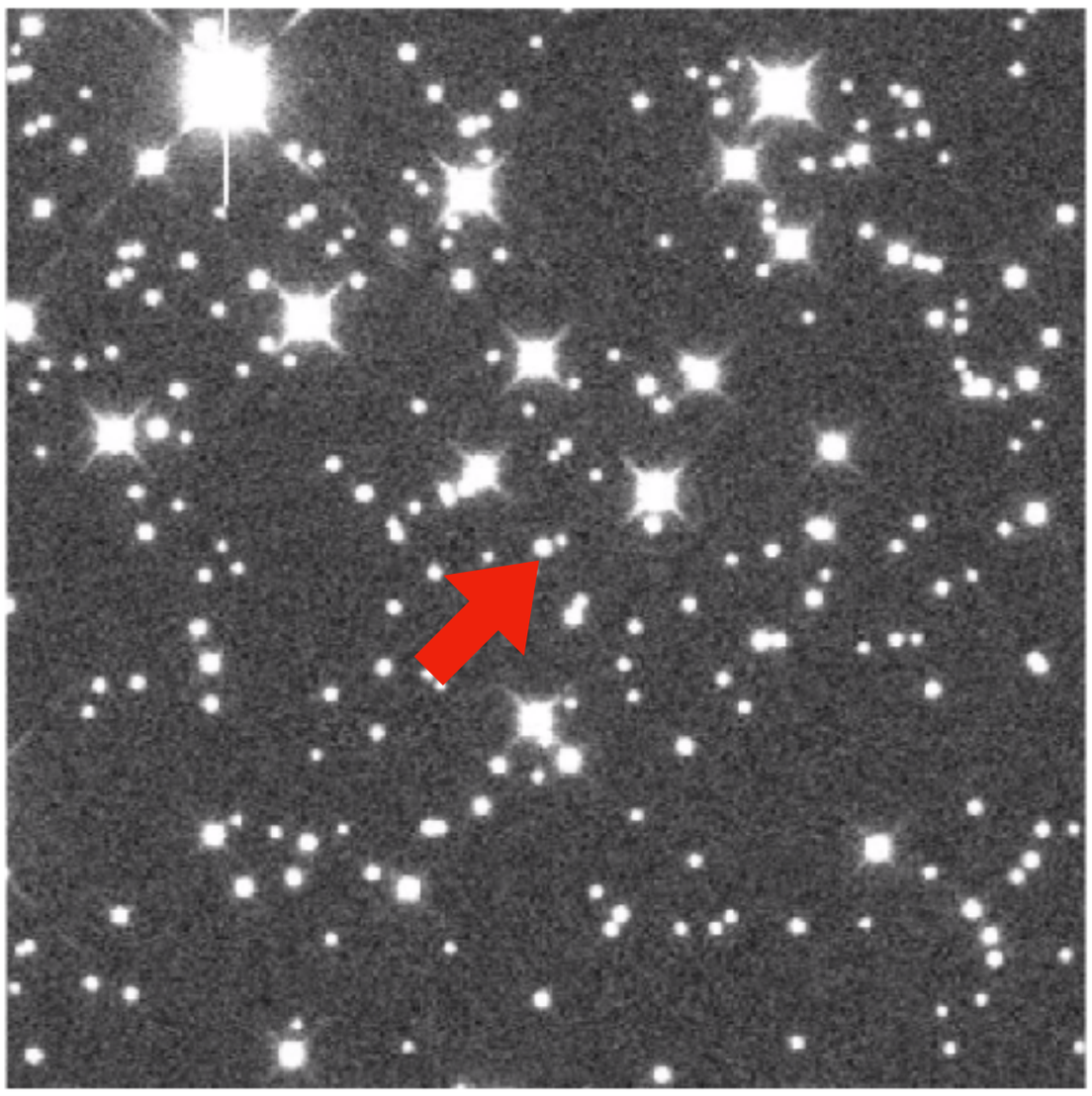
Spot the difference…

One of GOTO's key science goals is to identify astronomical sources that have changed in brightness…



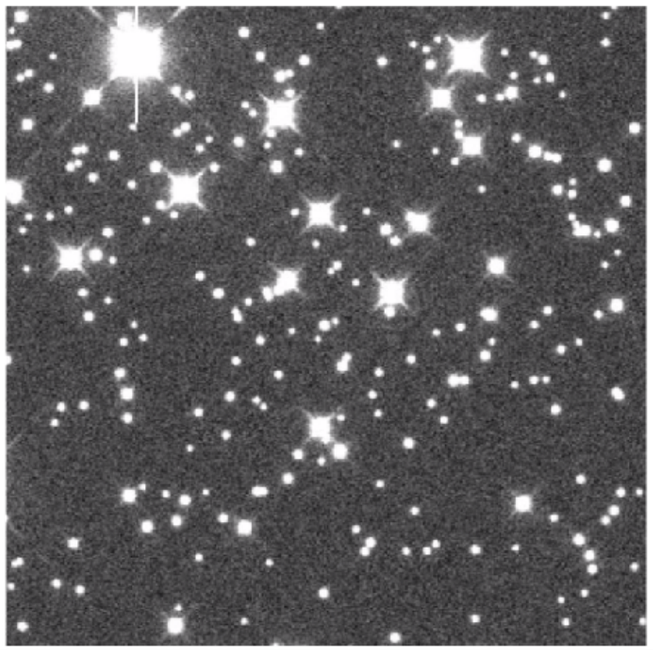
Spot the difference…

One of GOTO's key science goals is to identify astronomical sources that have changed in brightness…



Difference imaging

One way to achieve this is via difference imaging…

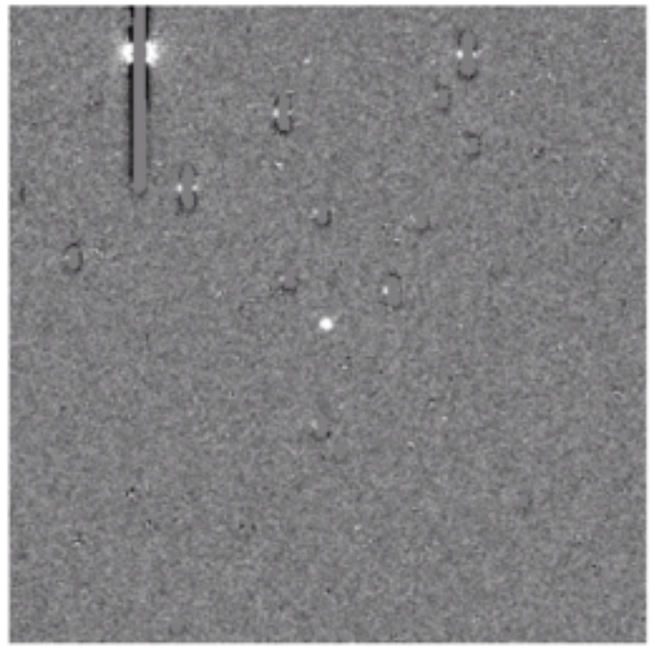


Input

−

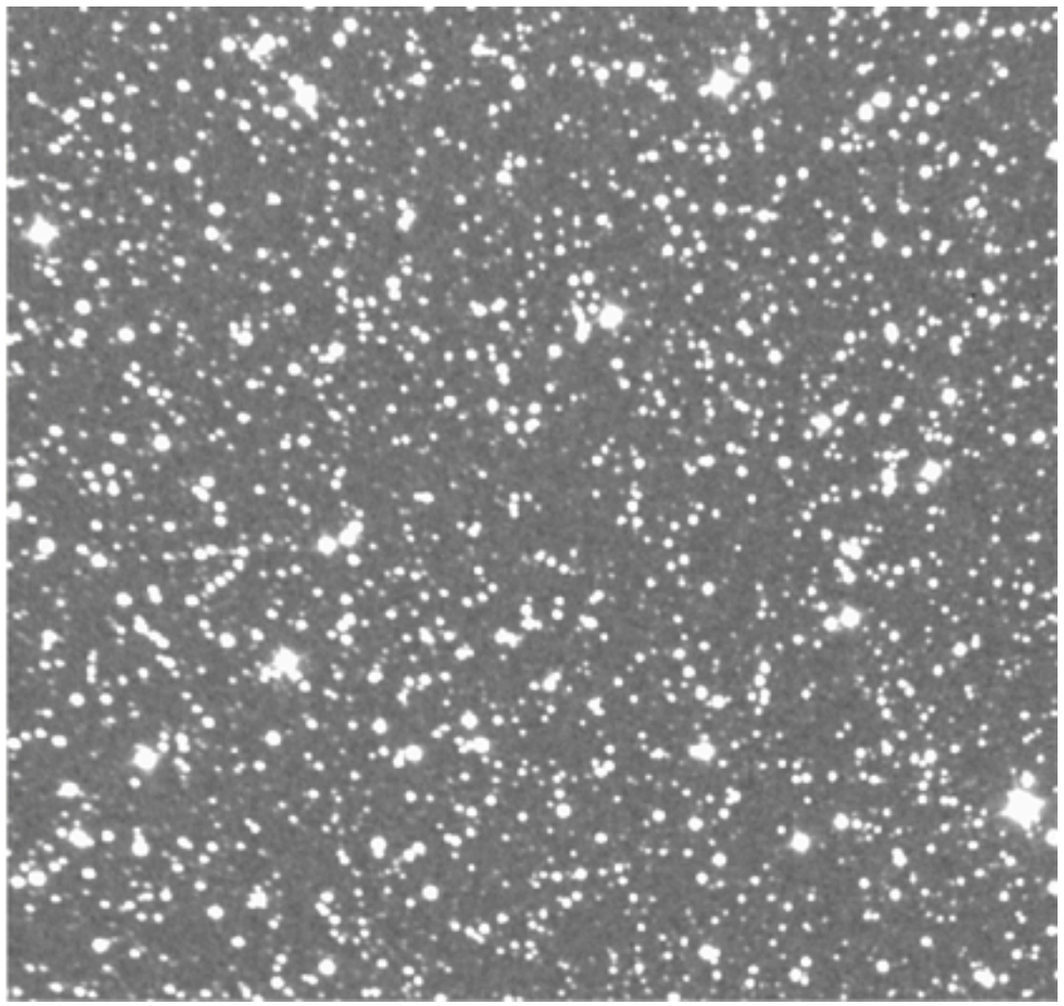
Reference

=

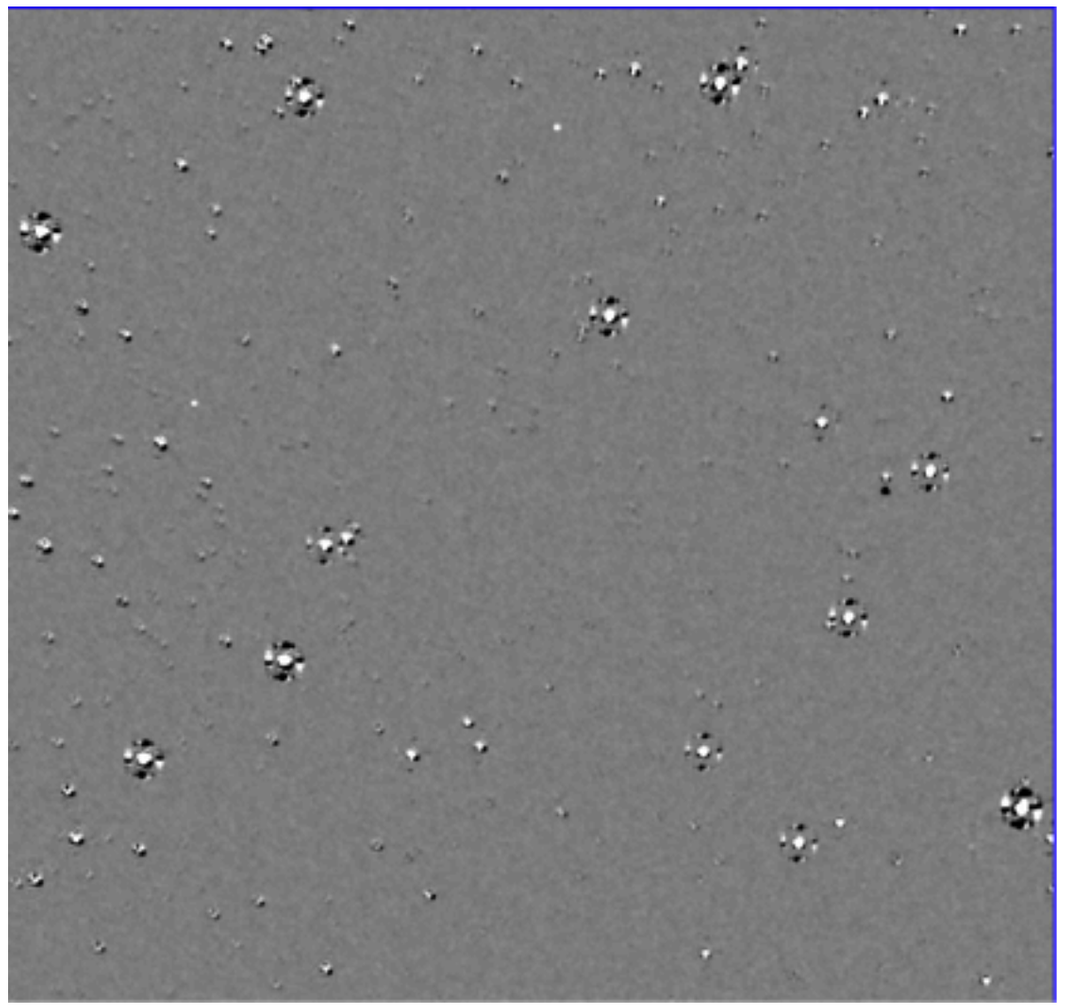


Difference

A typical difference image



Original



Difference

Typically, difference images contain far more artefacts than genuine transient sources; >1000:1 not uncommon.

A nice ML/AI problem??

The unbalanced problem

Supervised Machine Learning works best when there are similar numbers in each category.

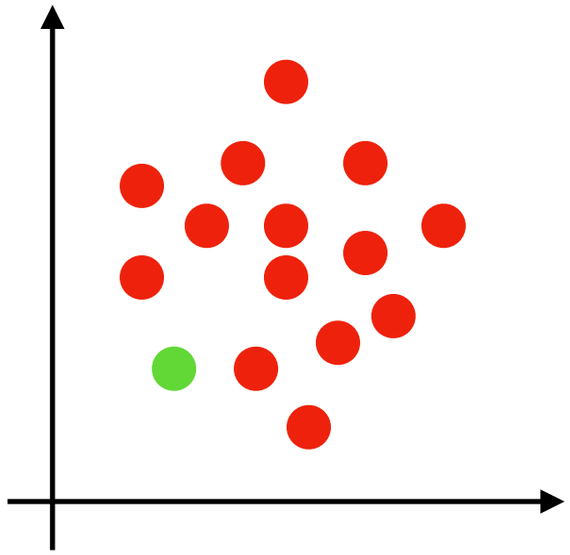
However, in extremely unbalanced scenarios…

the model “learns” that the best strategy is to simply assume everything belongs to the dominant class.

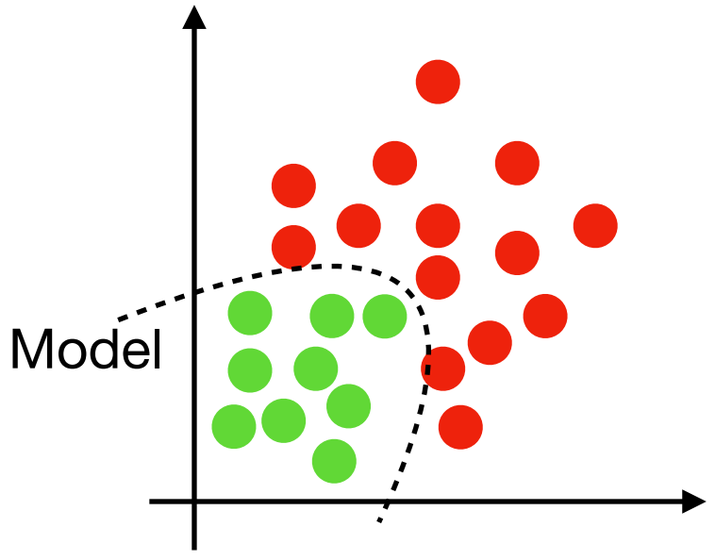
“I get >>99% accuracy when I assume everything is a defect”!

Re-balancing the problem

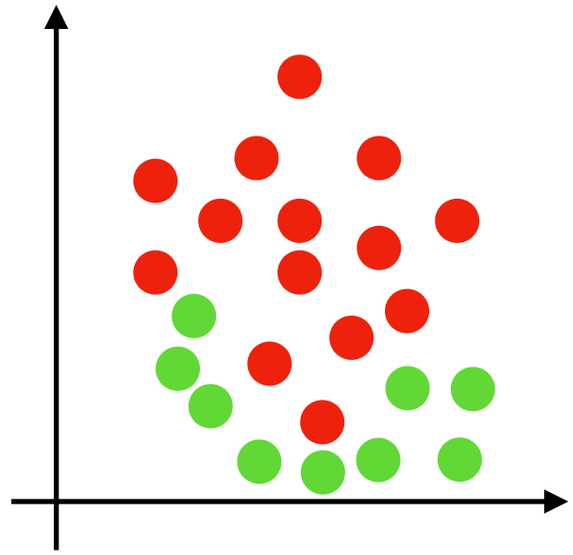
One way to address this is by up-sampling the minority class…



Input



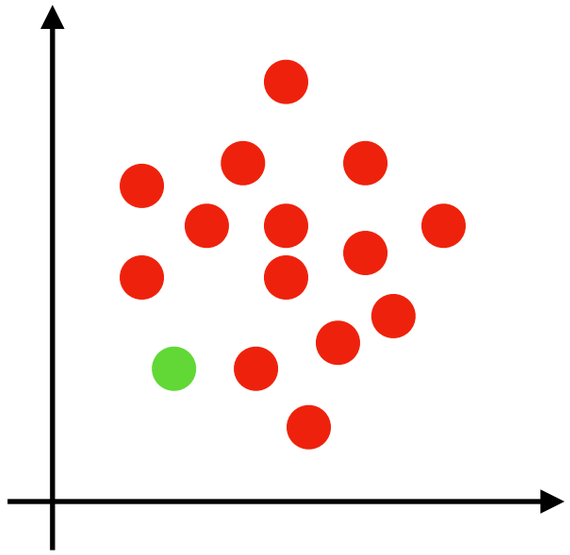
Upsampled



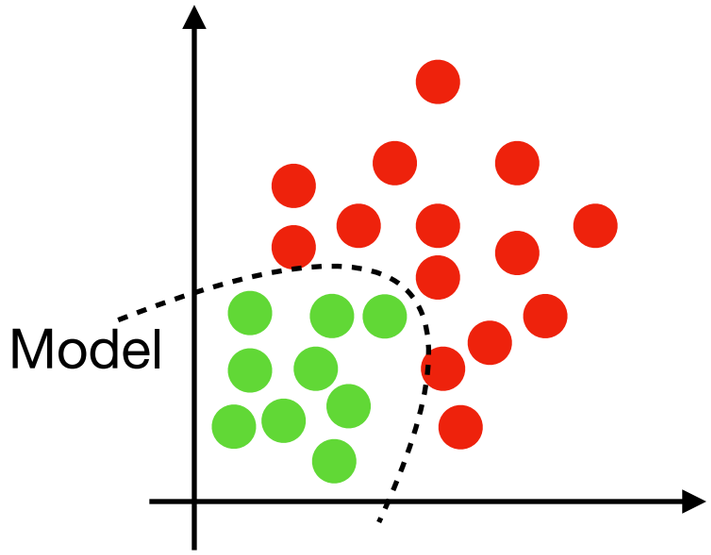
“Reality”

Re-balancing the problem

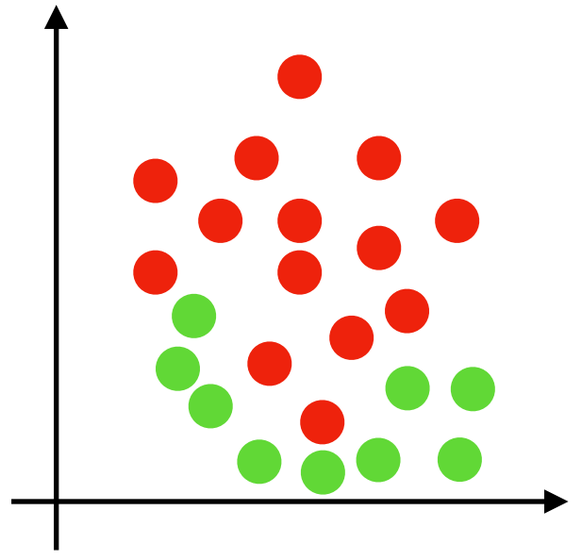
One way to address this is by up-sampling the minority class…



Input

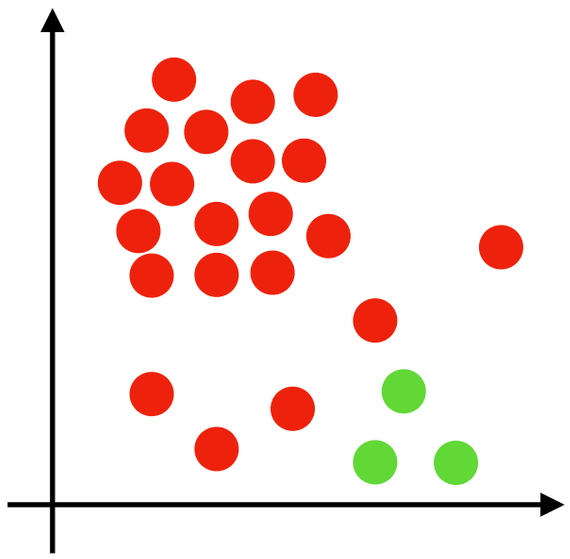


Upsampled

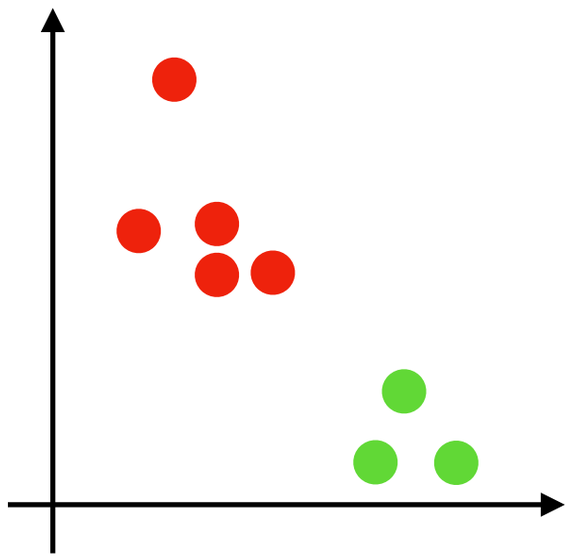


“Reality”

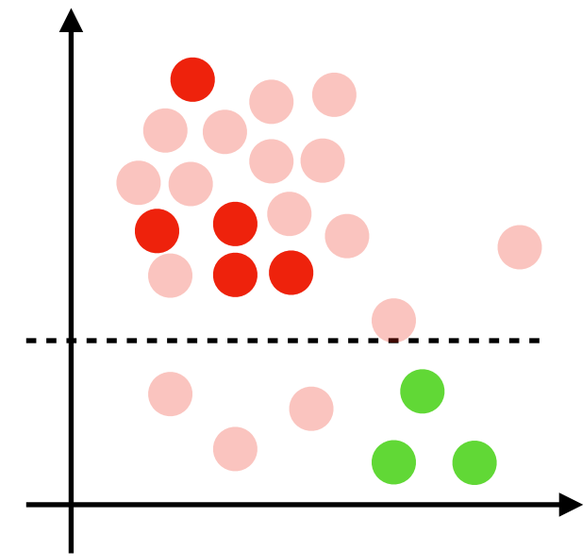
Or one could downsample the dominant class…



Input



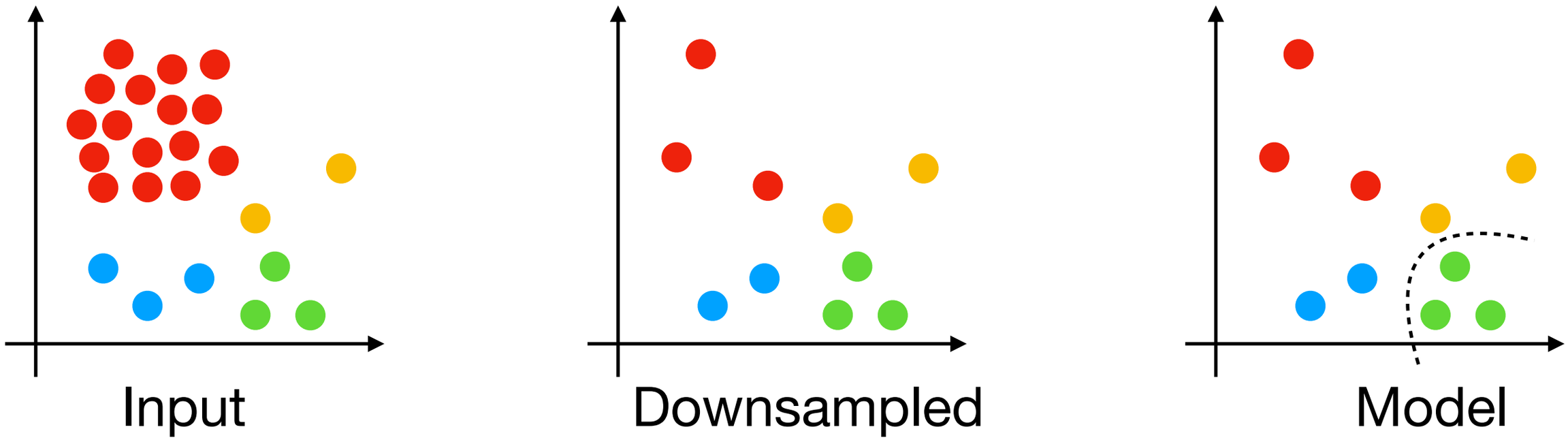
Downsampled



Model

Using clustering analysis to downsample

Try to get representative samples from all areas of parameter space…



see Tabacolde et al. (2018)

The “Unknown Unknowns”

“And then there’s the…things we don’t know we don’t know.”

—D. Rumsfeld

How do we train a model to identify the things we’ve never seen before?

The “Unknown Unknowns”

“And then there’s the…things we don’t know we don’t know.”

—D. Rumsfeld

How do we train a model to identify the things we’ve never seen before?

There’s huge potential for unsupervised machine learning within astronomy over the coming years to decades…

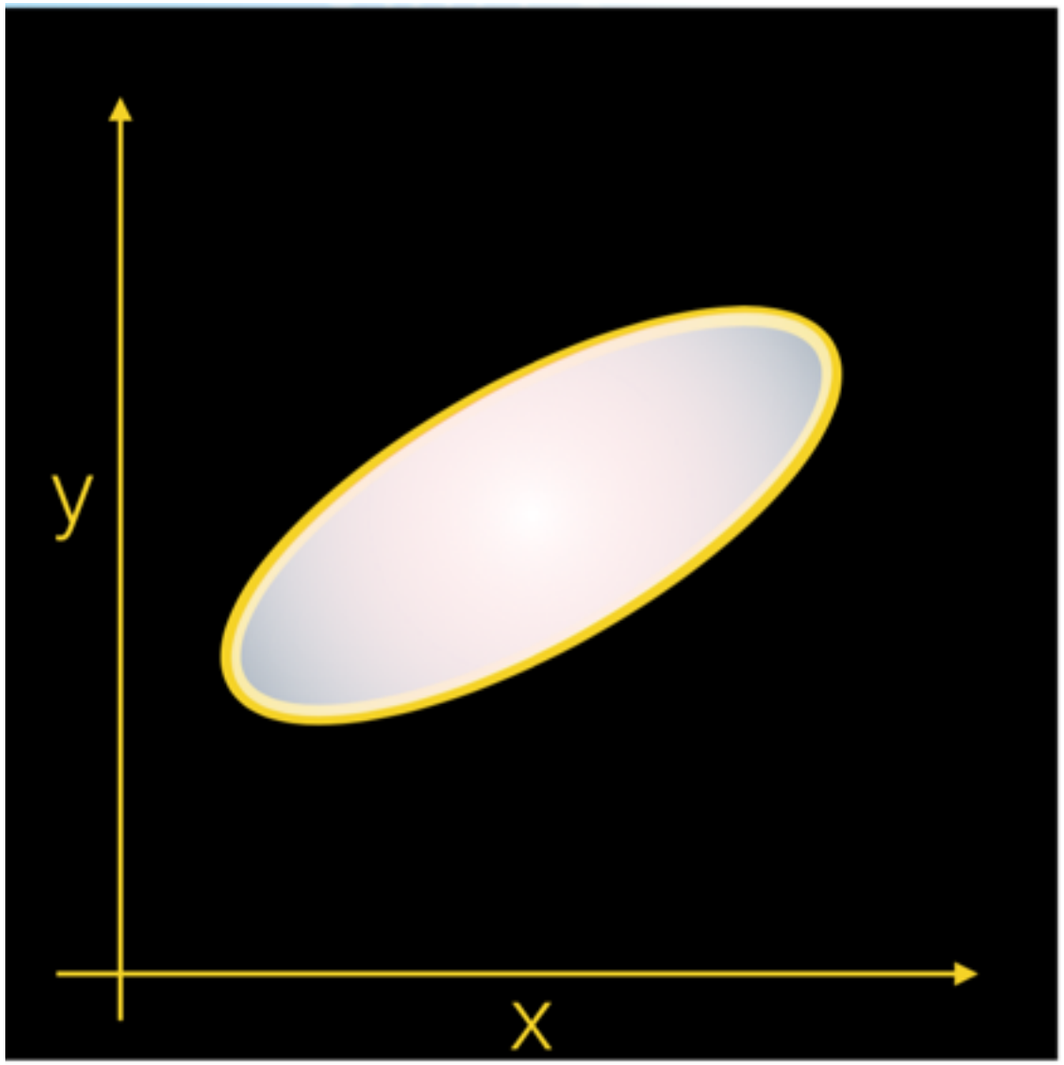
…but in an unsupervised system, how do we know our model is catching all the interesting “events”.

Part 2:
“Seeing” the big picture

Feature engineering in astronomy

Ultimately, in observational astronomy, *only* have images

Measure
Position
Shape
Brightness
Colours
Changes

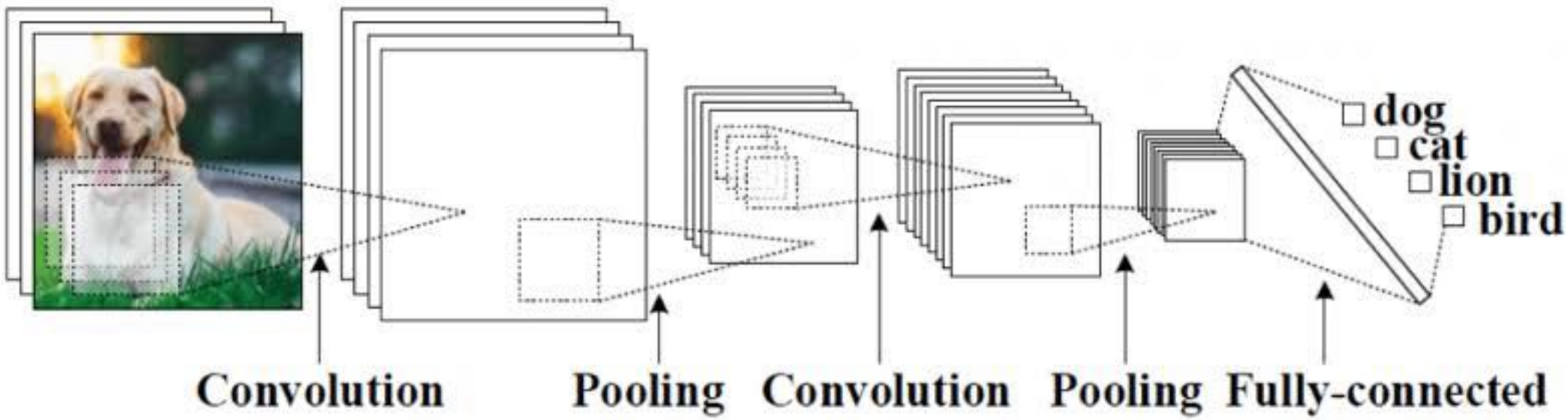


Derive
Distances
Energy output
Temperatures
Ages
Make-up
Densities
Masses
Velocities
etc.

Is right to boil our data down into a limited set of parameters?
Should we instead be using every pixel?

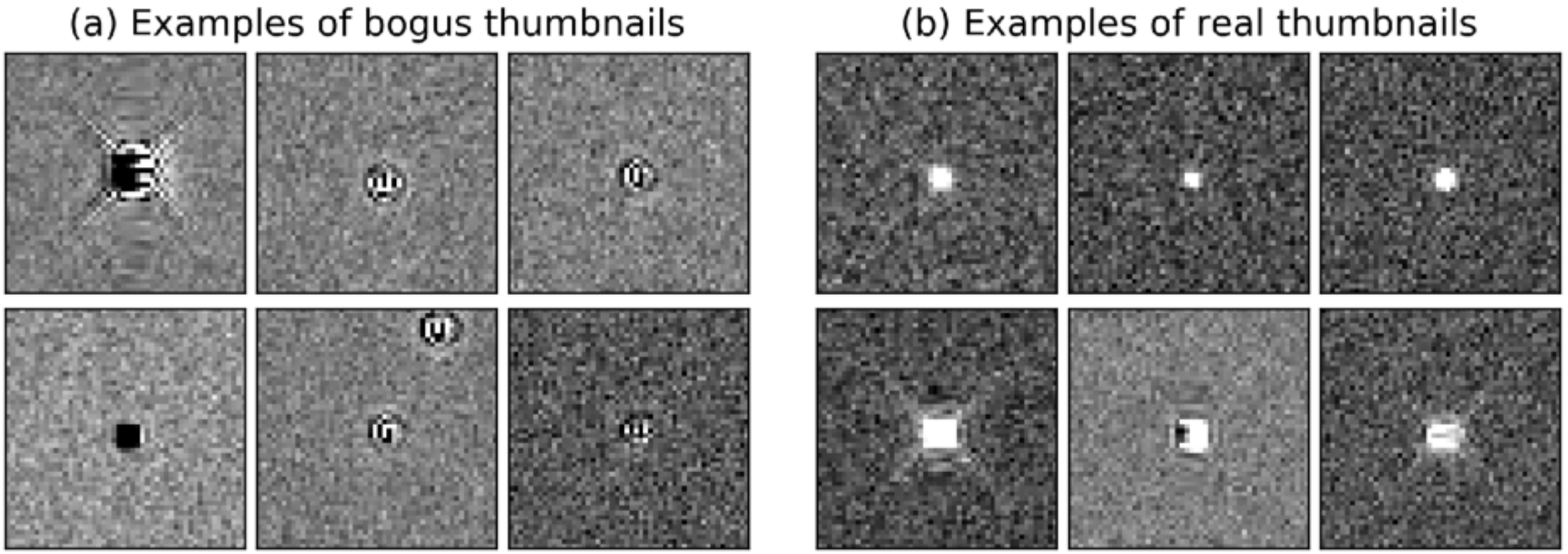
Convolutional Neural Networks

Convolutional Neural Networks (CNNs) - which work using filters to pick-out features in raw pixel data - are widely used to classify images.



CNNs in astronomy

Work on CNNs - including our own - suggest that they can outperform feature-based ANNs in some classification tasks.



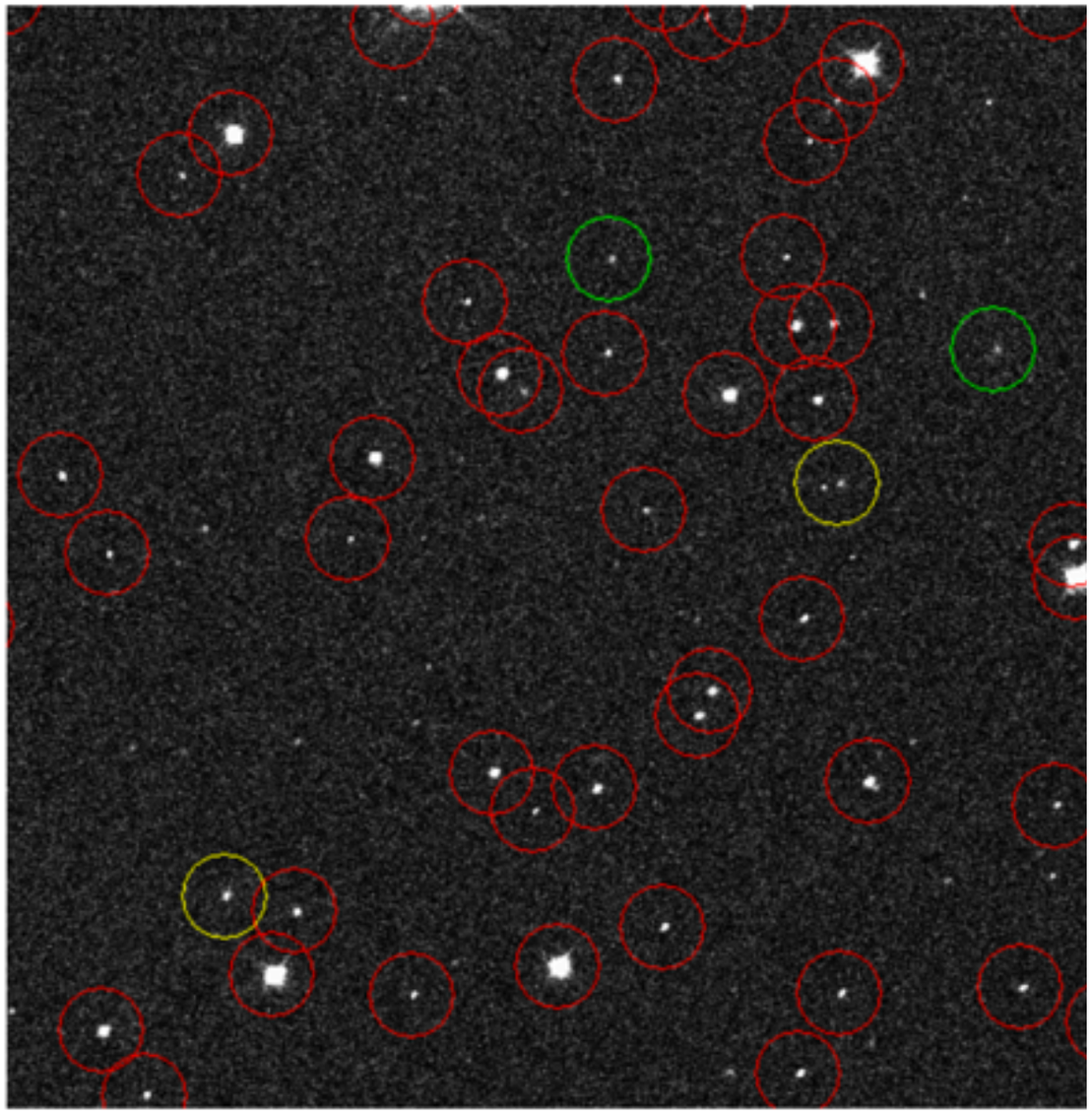
But…CNNs are more data intensive and more “black-boxy” than feature-based systems.

JingJing Liu et al 2018

Beyond the postage-stamps

Much CNN work to date in astronomy surveys first requires a source-detection and cutout-step…

But what if we could skip past that step and do detection and classification in one step.

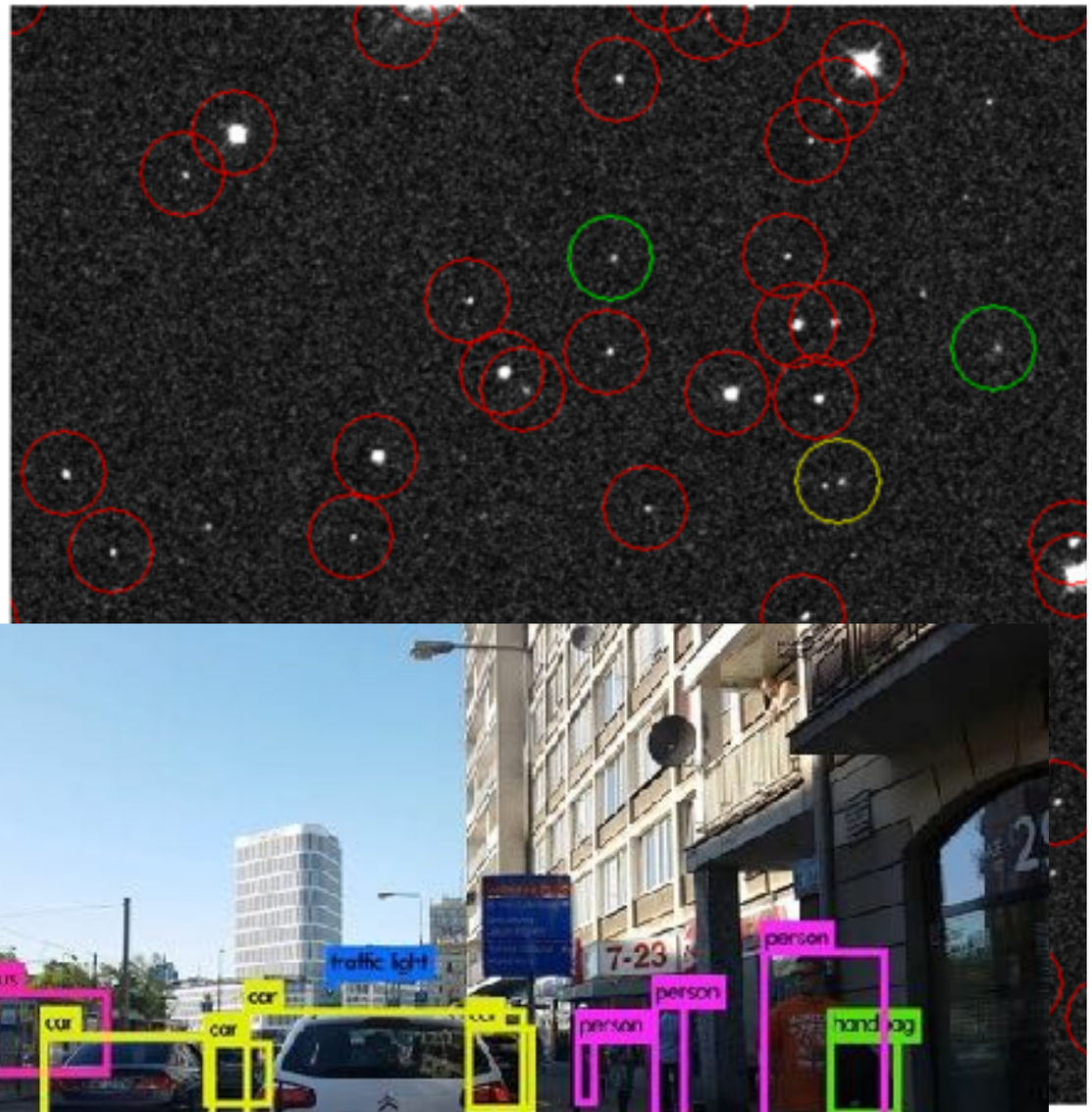


Beyond the postage-stamps

Much CNN work to date in astronomy surveys first requires a source-detection and cutout-step…

But what if we could skip past that step and do detection and classification in one step.

e.g., “You Only Look Once” algorithms.



Part 3:
The What, Where, and How of Data Storage

Big Data Storage in astronomy

Post-2000, astronomical surveys have largely used relational (SQL) databases for their data management.

If designed well, offers extremely efficient data searches over potentially billions of entries.

Big Data Storage in astronomy

Post-2000, astronomical surveys have largely used relational (SQL) databases for their data management.

If designed well, offers extremely efficient data searches over potentially billions of entries.

We use a Postgresql database to store GOTO data. Currently holding hundreds of millions of entries.

Very good at delivering data for very specific requests.

Alternatives to SQL

SQL efficiency depends on design, if we search off-index, searches become painfully slow.

Need to know likely queries during the design phase.

SQL — by its nature — only works with highly structured data and isn’t optimised for analytics.

Alternatives to SQL

SQL efficiency depends on design, if we search off-index, searches become painfully slow.

Need to know likely queries during the design phase.

SQL — by its nature — only works with highly structured data and isn't optimised for analytics.

Hadoop
Distributed storage framework that can work with structured and unstructured data.

Open to far more complex analytics than SQL.

We've attempted to use Hadoop for GOTO data, but there's a considerable learning curve.

Part 4:
Clouds are on the horizon

Part 5:
The Airline Part

What does Cloud Computing offer?

Service like Amazon Web Services, Google Cloud, Microsoft Azure offer:

- Almost unlimited compute instances;
- Almost unlimited storage;
- Relational databases;
- Hadoop-like infrastructure;
- Web servers;
- etc etc.

Why we're using AWS for GOTO data?

As well as the nightly processing of data, every ~6 months we want to combine all data to reach higher sensitivities.

It takes about 2 weeks on a 52-core machine to process the 6 monthly-data. We don't want this to interrupt our nightly processing.

Option 1: Buy a second machine: £10,000 investment;

Why we're using AWS for GOTO data?

As well as the nightly processing of data, every ~6 months we want to combine all data to reach higher sensitivities.

It takes about 2 weeks on a 52-core machine to process the 6 monthly-data. We don't want this to interrupt our nightly processing.

Option 1: Buy a second machine: £10,000 investment;

Option 2: Use University Central Computing Systems;

Why we're using AWS for GOTO data?

As well as the nightly processing of data, every ~6 months we want to combine all data to reach higher sensitivities.

It takes about 2 weeks on a 52-core machine to process the 6 monthly-data. We don't want this to interrupt our nightly processing.

Option 1: Buy a second machine: £10,000 investment;

Option 2: Use University Central Computing Systems;

Option 3: Rent cloud computing:

- 20,000 core-hours @ £0.013 per core-hour = £260;
- 4TB storage for 1 week = £50
- 1TB data extraction from AWS: £50
- Total: ~£400 (and definitely below £1000).

AWS PCluster: HPC Cluster available to anyone

Originally, AWS not really set up to mimic science-grade HPC clusters.

In 2018, developed PCluster, which automates the launch and maintenance of multiple cores.

Includes familiar cluster management software, such as SLURM, MPIEXEC etc.

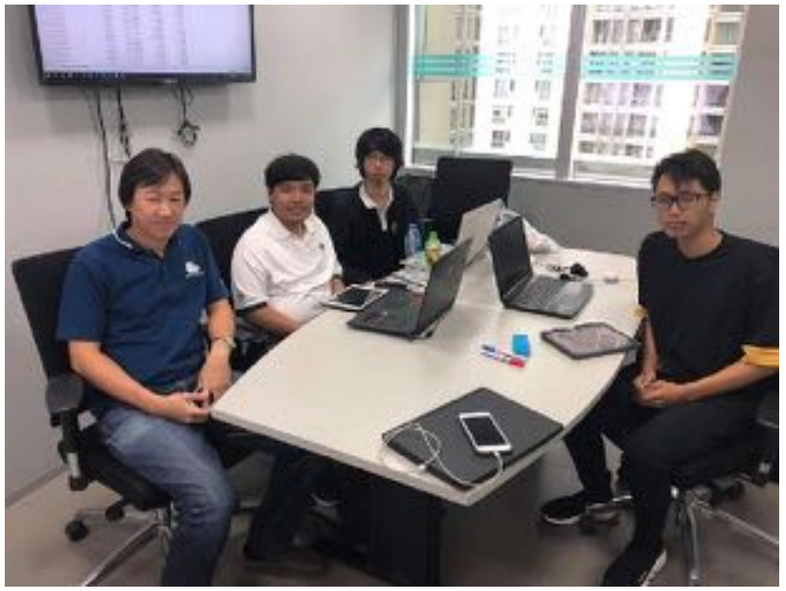
You can take your regular MPI-enabled software, install it on a PCluster, and launch as many nodes as you like.

No initial outlay — ideal for sporadic or one-off jobs.

…and STFC Newton/GCRF were happy to fund it!

Part 5:
The Airline Part

GCRF Project: Working with external partners

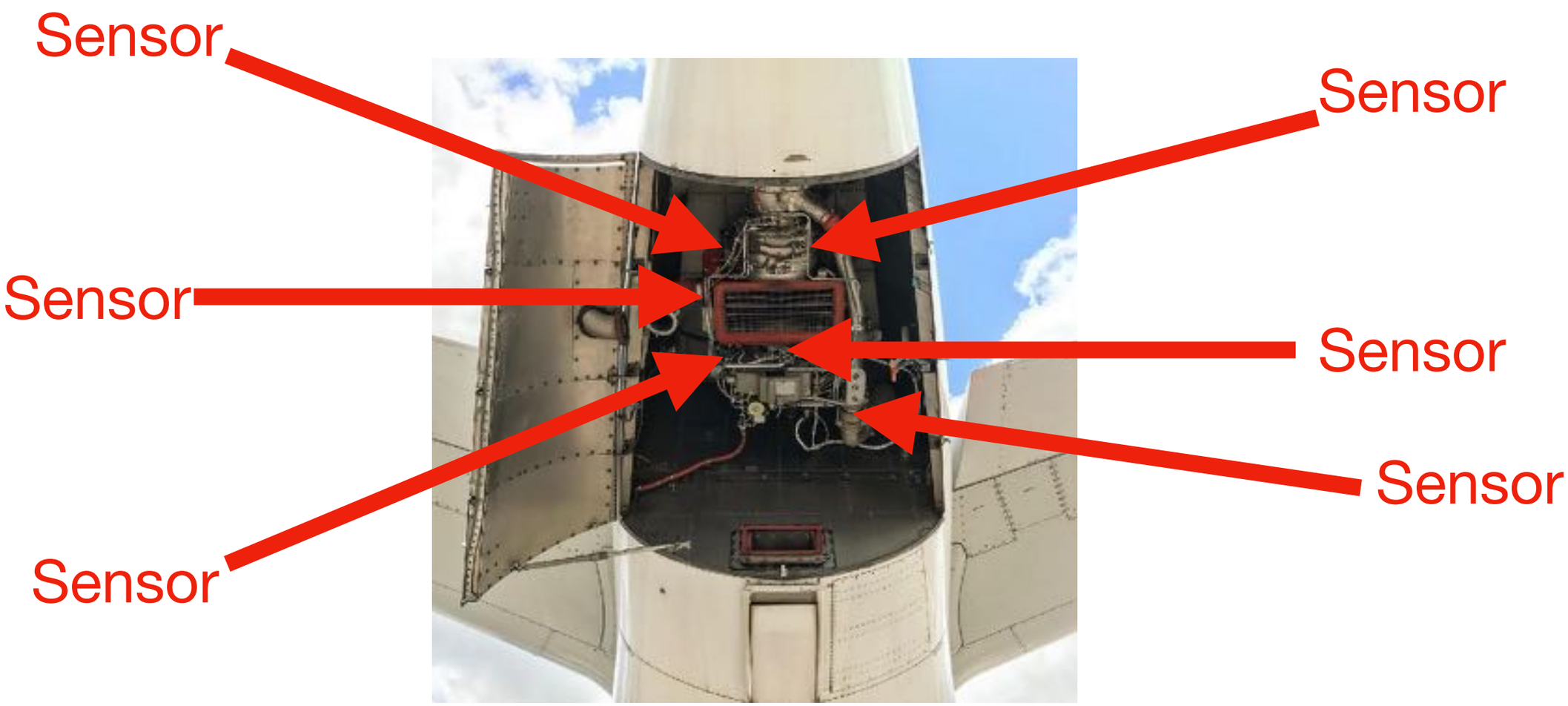


Thai AirAsia Project

Auxiliary Power Unit (APU)

Task: To develop a model that takes information from all the various sensors to predict when the APU will fail.

Thai AirAsia Project



Auxiliary Power Unit (APU)

Task: To develop a model that takes information from all the various sensors to predict when the APU will fail.

Thai AirAsia Project

Thai AirAsia Project

- External partners describe problem to students and provide relevant datasets.

Thai AirAsia Project

- External partners describe problem to students and provide relevant datasets.
- Students work on developing a first draft of a model and feed back to external partners.

Thai AirAsia Project

- External partners describe problem to students and provide relevant datasets.
- Students work on developing a first draft of a model and feed back to external partners.
- After further iterations, students spend two weeks based at the HQ of the external partners, working with them on further improvements.

Thai AirAsia Project

- External partners describe problem to students and provide relevant datasets.
- Students work on developing a first draft of a model and feed back to external partners.
- After further iterations, students spend two weeks based at the HQ of the external partners, working with them on further improvements.
- Once the final model is developed, students move to working on a front-end system to enable partner employees to use the system easily.
- Students write a final report on their experience and product.

Summary

- Optical astronomy is very much within the Big Data era (and arguably has been for many years)

- Methods of analysing this data took a while to catch-up, but progress is increasing rapidly.

- Interesting problems still be be fully addressed - unbalanced data, unknown unknowns, post-SQL analytics, on-image analysis.

- Commercial Cloud Computing mimicking Science HPC is a viable option for one-off or sporadic processing.

- If you know how to reach-out, there are still many external businesses and organisations to work with.