Astronomical archives: Serving up the Universe

1.
Science Archives

Science Archives
Rationale

- Archival research
- Multi-wavelength astronomy
- Proposing
- Reproducibility
- Time variability
- Support of developing countries
- Citizen-Science
- Outreach

Science Archives
Photons

Position

Energy

Time

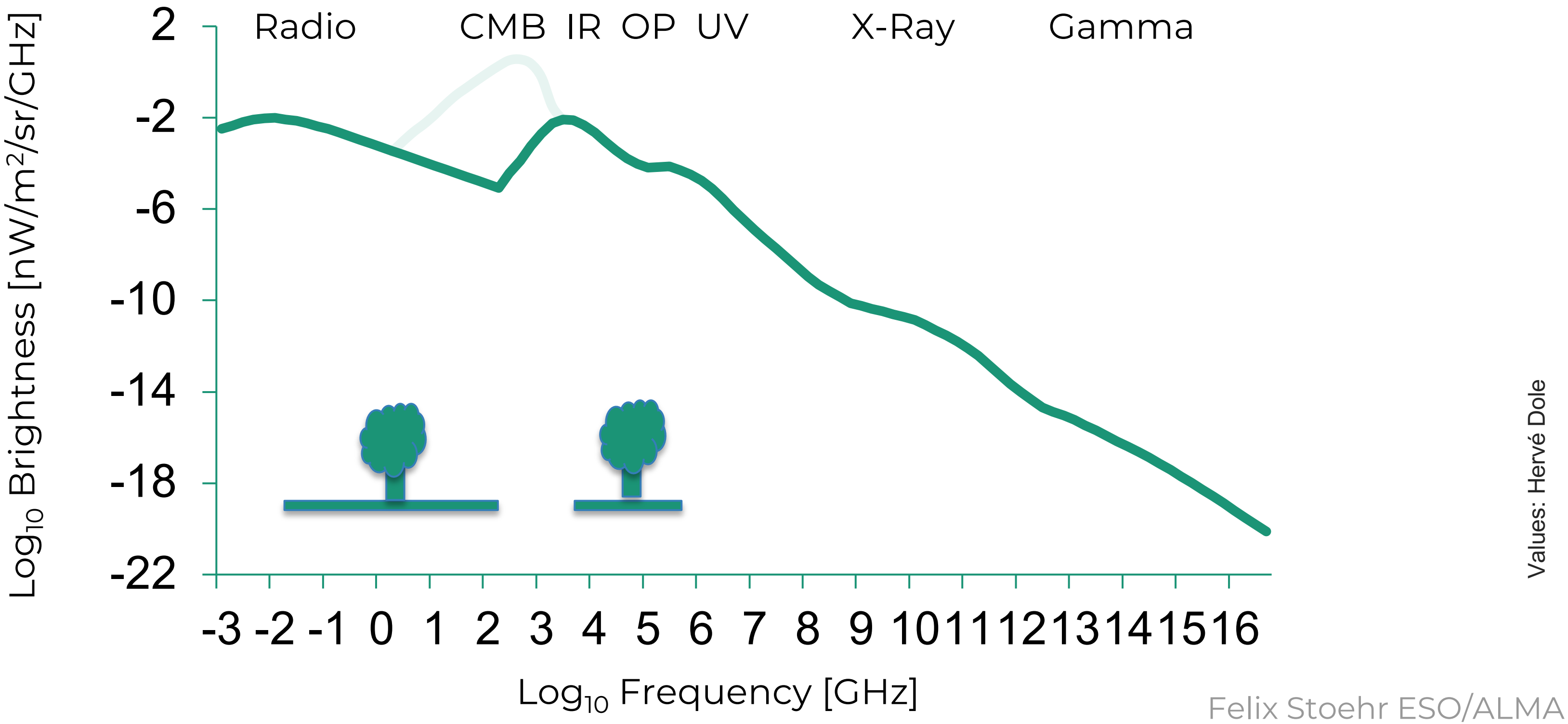
Polarisation

Science Archives
6D hypercubes

Pos1	Pos2	Energy	Time	Pol.	Quantity	
1	1	N	1	1	1	Spectrum
1	1	1	N	1	1	Time-series
N	N	1	1	1	N	Image with error map
N	N	N	1	N	1	Data cube with polarisations

Quantity: flux, counts, errors, weights, …

Science Archives
Photons in the Universe



Science Archives
Best practices

- Physical quantities
- Unscoped search
- Observations, Proposals, Publications
- Target-list upload
- Previews
- Modern user-experience
- Programmatic access (VO)
- Metadata are public
- Result table + SkyView
- Science-grade products + PL
- Anonymous downloads
- Self-describing FITS files
- Parallel downloads
- Authors must cite data-use
- Frequent Reprocessing

Science Archives
Usage

fraction	cumulative	Search Field
26.9%	26.9%	Source Name (Resolver)
25.5%	52.4%	Project Code
11.4%	63.8%	Ra Dec
8.0%	71.8%	Source Name (ALMA)
7.8%	79.7%	PI Name
3.7%	83.4%	Band
3.7%	87.1%	Public Data
2.0%	89.1%	Frequency
1.4%	90.5%	Start Date
1.1%	91.6%	
1.1%	92.8%	Spatial Resolution
1.0%	93.7%	Project Abstract
0.9%	94.6%	Science Keyword
0.8%	95.4%	Project Title
0.8%	96.2%	Galactic
0.7%	96.9%	Targetlist
0.5%	97.4%	Proposal Authors
0.5%	97.9%	Spectral Resolution
0.4%	98.4%	Integration Time
0.2%	98.6%	Continuum Sensitivity
0.2%	98.8%	FOV
0.2%	99.0%	Polarisation Type
0.2%	99.1%	First Author
0.1%	99.3%	Water Vapour
0.1%	99.4%	Spatial Scale Max
0.1%	99.5%	Line Sensitivity
0.1%	99.6%	Authors
0.1%	99.7%	Publication Year
0.1%	99.7%	Publication Count
0.1%	99.8%	Publication Title
0.1%	99.9%	Publication Abstract
0.1%	100.0%	Bandwidth
0.0%	100.0%	Bibcode

Science Archives
No time to talk about:

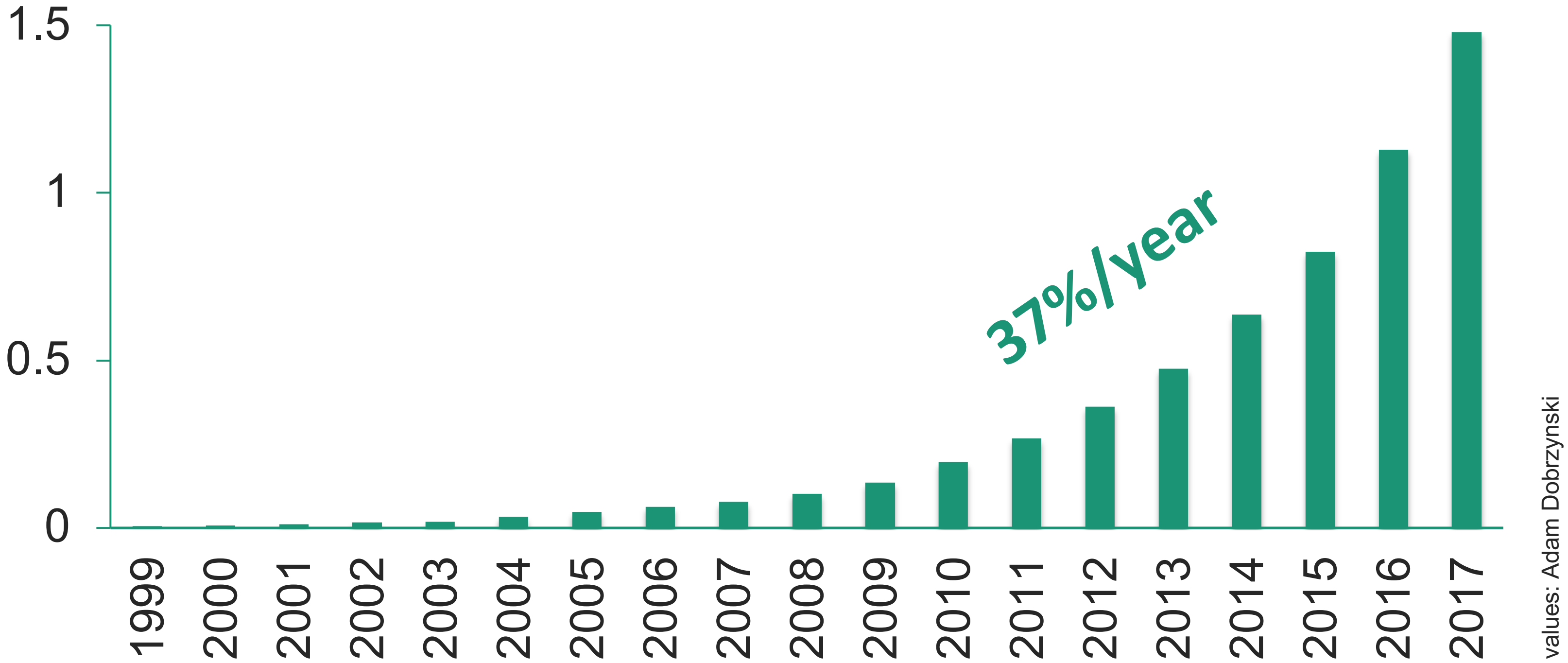
- FITS
- Code reuse
- Keeping data alive

2.
Observatories

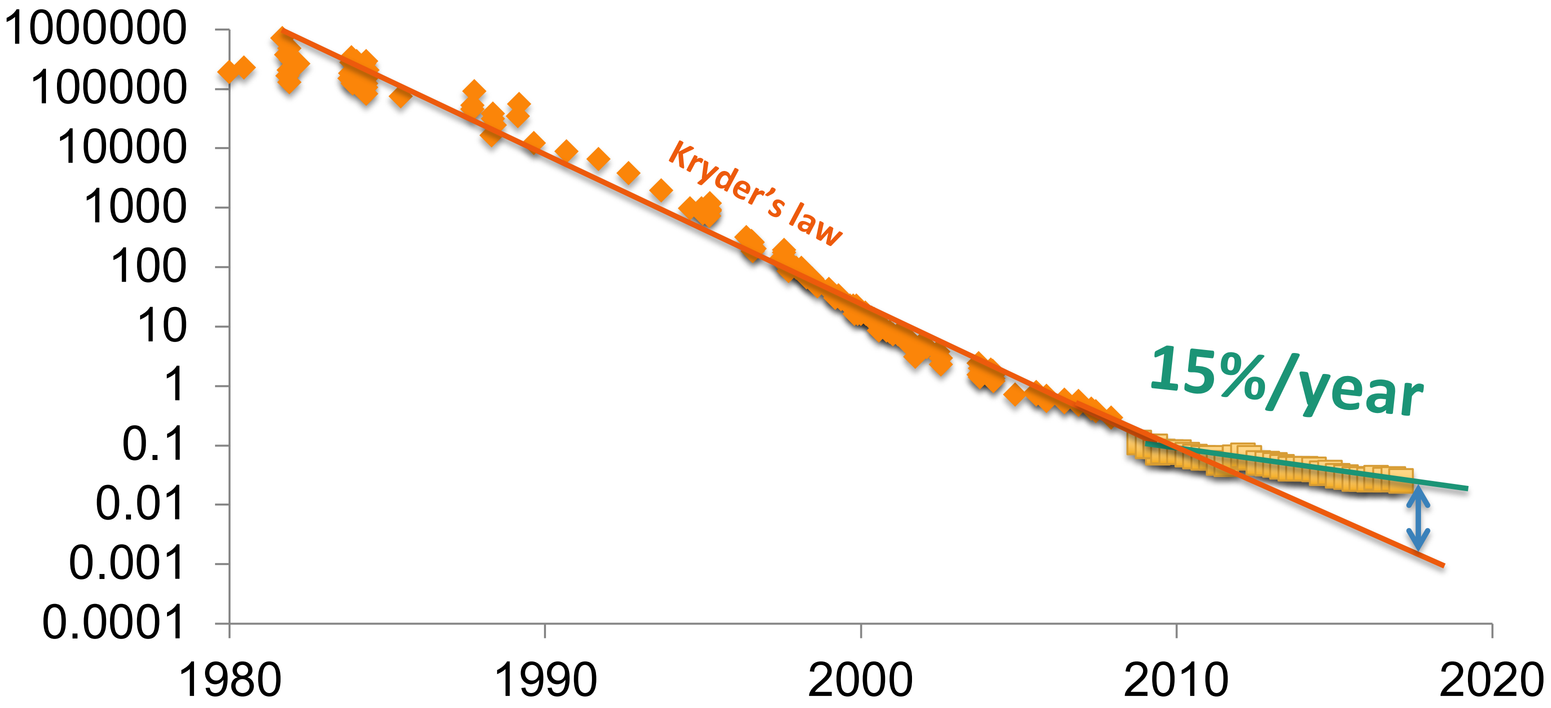
Observatories
Trends

- From experiments to observatories
- From single archives to data-portals
- Increased use of VO standards and protocols
- Science-grade data approach universally accepted
- Massive data-sets

Observatories
Petabytes at ESO

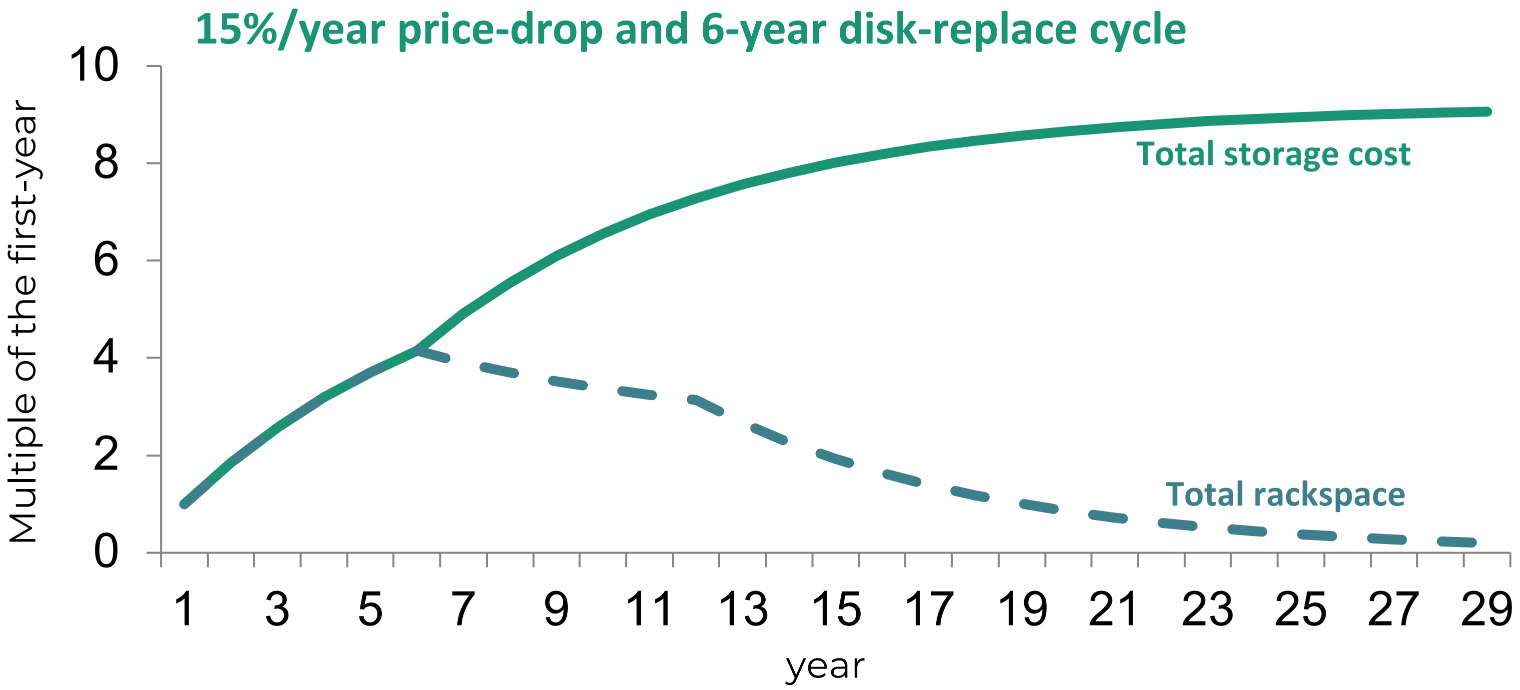


Observatories
Hard-disks: USD per Gigabyte



values: mkomo.com, blackblaze.com

Observatories
Linear data intake



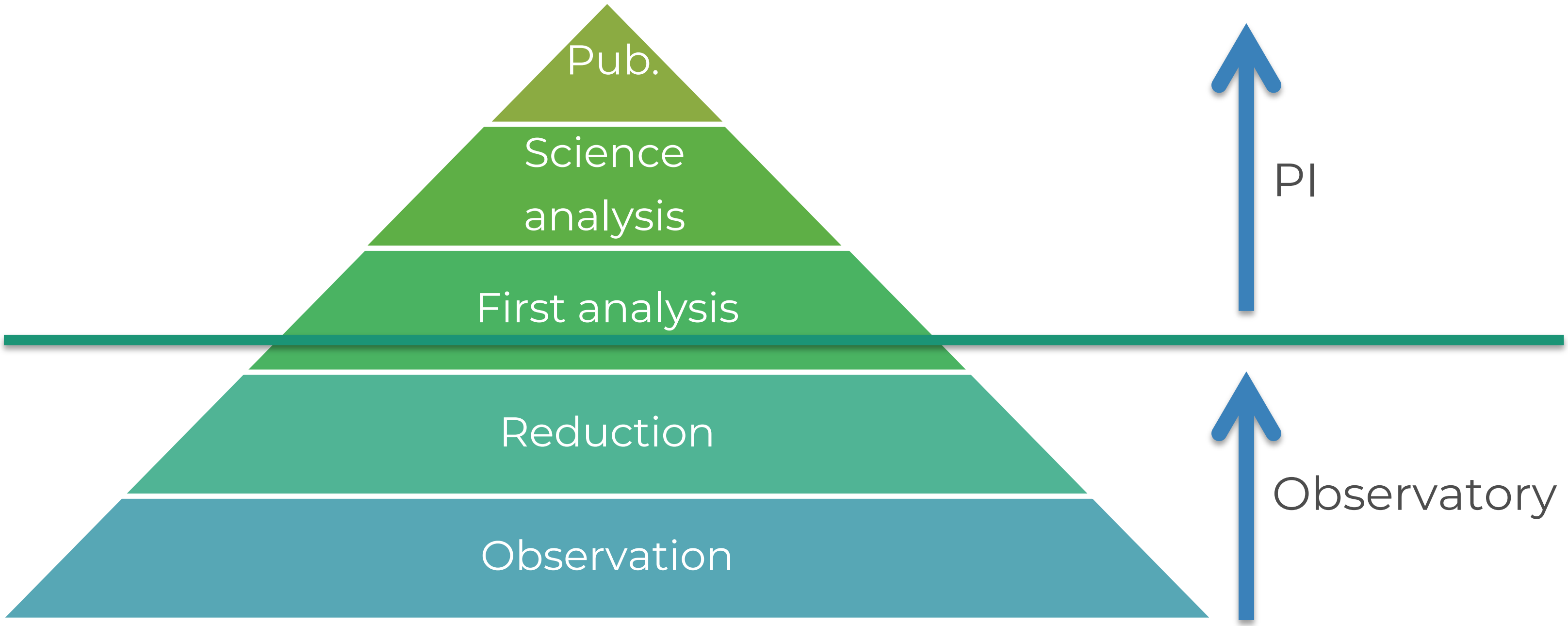
Observatories
So much data

- Today
  - VLT + ALMA + Magic 70GB/year/astronomer
  - WMA 350GB/year/astronomer
- 2030
  - VLT + ELT + ALMA + CTA 1TB/year/astronomer
  - SKA 200TB/year/astronomer
- Astronomers don't scale: they will be the rare resource

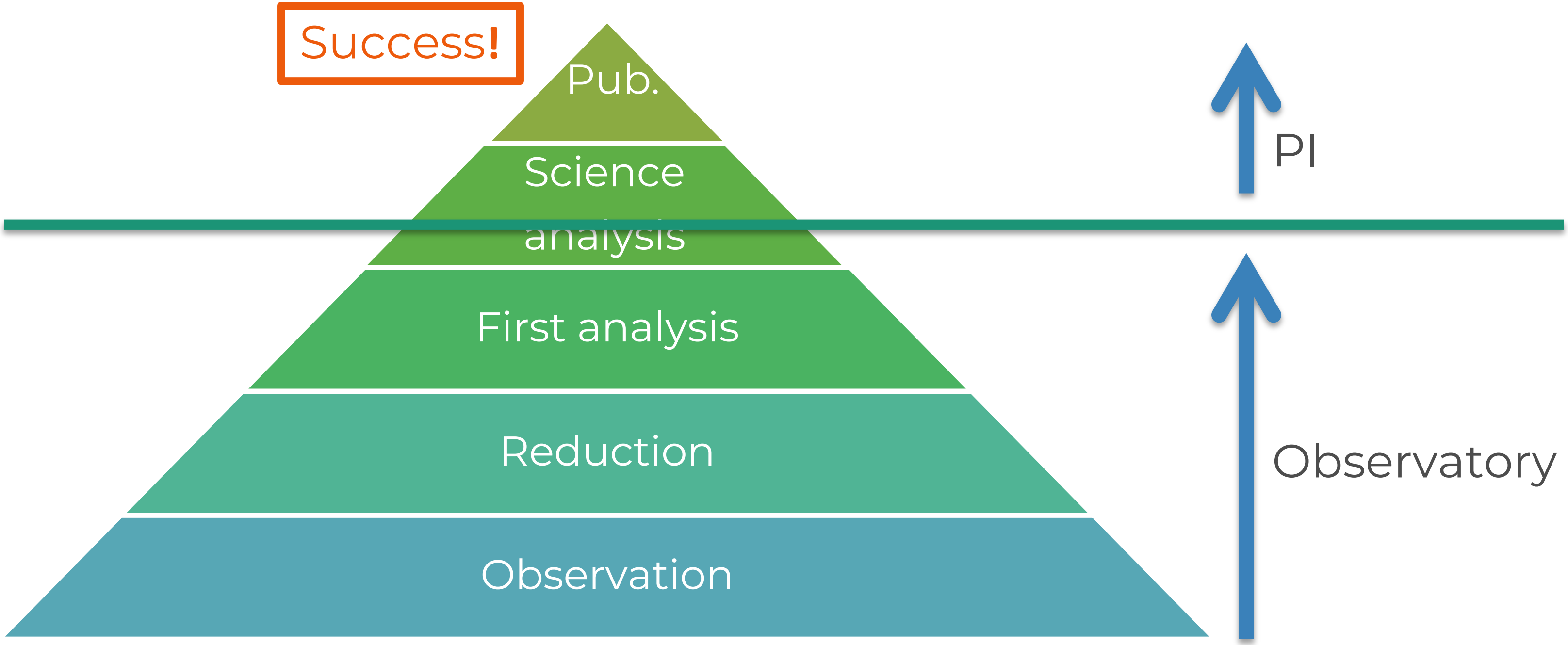
Observatories
So much data: solutions

- “Think of taking less data” (Alex Szalay)
- Process data to higher levels
- Machines do astronomy

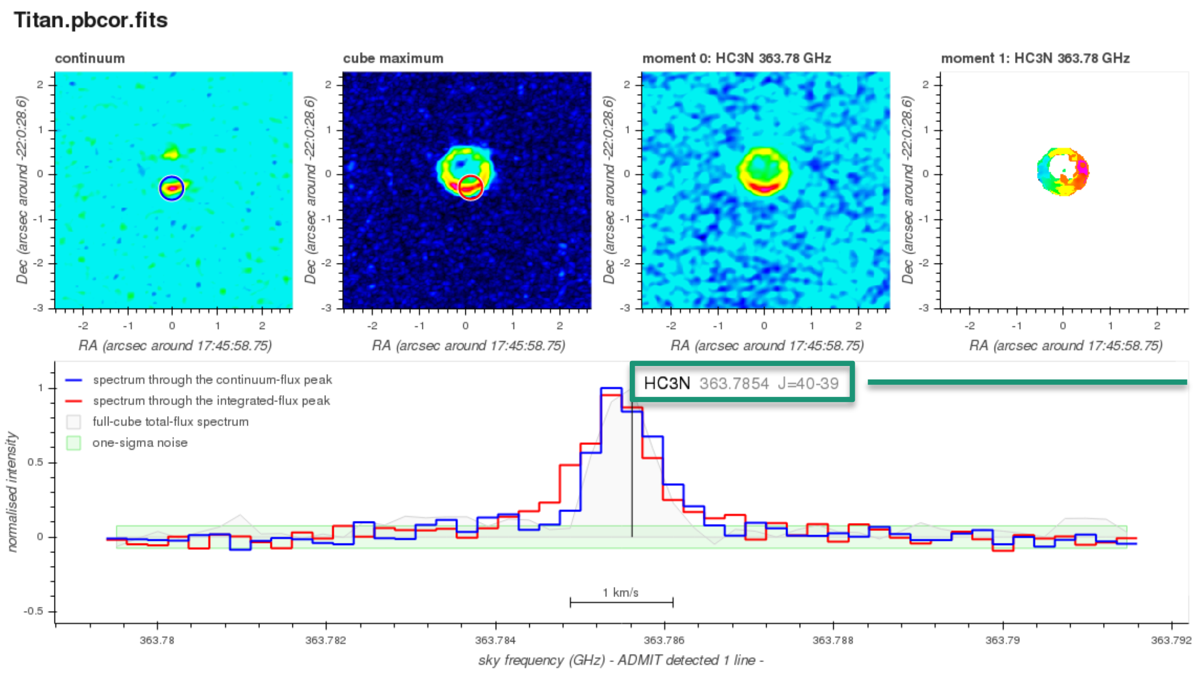
Observatories
More responsibility



Observatories
More responsibility



Observatories
First analysis



ALMA Data Mining Toolkit (ADMIT)

Observatories
3D

- ESO: VLT (MUSE, KMOS, SINFONI), ELT (HARMONI)
- ALMA
- SKA
- LOFAR
- WMA
- ATHENA
- Keck (ESI)
- JWST (MIRI, NIRSpec)
- …

3.
Machine Learning

Machine Learning
Will be needed

- in data processing
- in quality control
- in source-extraction
- in source-classification

Relevant for Science Archives

Machine Learning
explosion

1810.07857: Multiband galaxy morphologies for CLASH: a convolutional neural network transferred from CANDELS

1810.07703: A Deep Learning Approach to Galaxy Cluster X-ray Masses

1810.01483: DeepCMB: Lensing Reconstruction of the Cosmic Microwave Background with Deep Neural Networks

1810.00592: Extracting gamma-ray information from images with convolutional neural network methods on simulated Cherenkov Telescope Array data

1810.07888: Classifying Lensed Gravitational Waves in the Geometrical Optics Limit with Machine Learning

1809.09622: Convolutional Neural Networks for Spectroscopic Redshift Estimation on Euclid Data

1809.05748: Segmentation of coronal holes in solar disk images with a convolutional neural network

1809.03043: Fast Radio Burst 121102 Pulse Detection and Periodicity: A Machine Learning Approach

1809.03315: Deep Learning Based Detection of Cosmological Diffuse Radio Source

1809.09722: TSARDI: a Machine Learning data rejection algorithm for transiting exoplanet light curves

1809.02154: From FATS to feets: Further improvements to an astronomical feature extraction tool based on machine learning

1809.01934: Towards online triggering for the radio detection of air showers using deep neural networks

1809.01691: Galaxy detection and identification using deep learning and data augmentation

Machine Learning explosion

1808.09955: QuasarNET: Human-level spectral classification and redshifting with Deep Neural Networks

1808.09739: Detecting Radio Frequency Interference in radio-antenna arrays with the Recurrent Neural Network algorithm

1808.08371: Protostellar classification using supervised machine learning algorithms

1808.07491: Weak lensing shear estimation beyond the shape-noise limit: a machine learning approach

1808.06977: Searching for Sub-Second Stellar Variability with Wide-Field Star Trails and Deep Learning

1808.05728: Machine Learning Classification of Gaia Data Release 2

1808.04728: CosmoFlow: Using Deep Learning to Learn the Universe at Scale

1808.04428: Deep learning of multi-element abundances from high-resolution spectroscopic data

1808.03626: Enhanced Rotational Invariant Convolutional Neural Network for Supernovae Detection

1808.00011: Analyzing interferometric observations of strong gravitational lenses with recurrent and convolutional neural networks

Machine Learning
Explainable AI

- Dan Jacobsen (ORNL)
- 2.36 Exaops (17 million GPU cores)
- Explainable AI (iterative Random Forests)

Machine Learning
Catalogue of the Universe

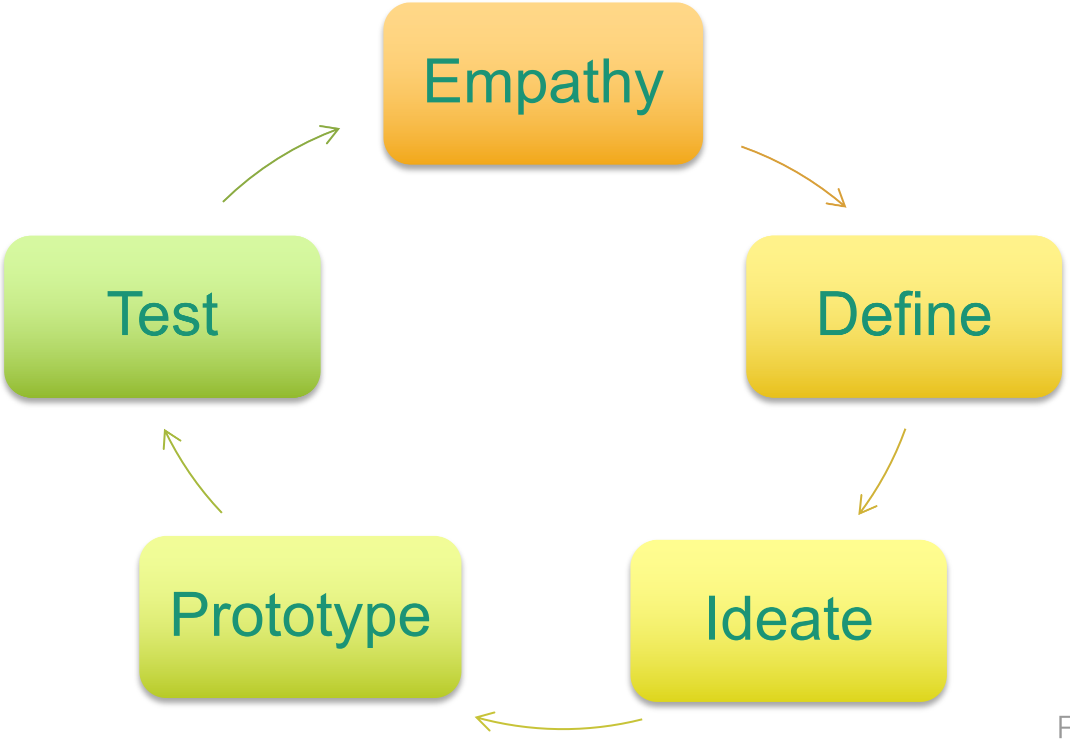
- AI combines and cross-matches all surveys and catalogues
- AI goes through all science-grade data of all observatories
- Use SIMBAD + NED for learning
- Hierarchical

4.
User Experience

User-experience
2007

“ *User-experience is how it **feels***

User-experience
Design Thinking



open.sap.com/courses/dt1

5.
Conclusions

Conclusions
Summary

- Photons are simple
- Science Archive best practices have emerged
- Huge challenges ahead for observatories (and Archives)
- AI comes to the rescue
- Embrace user-centric design