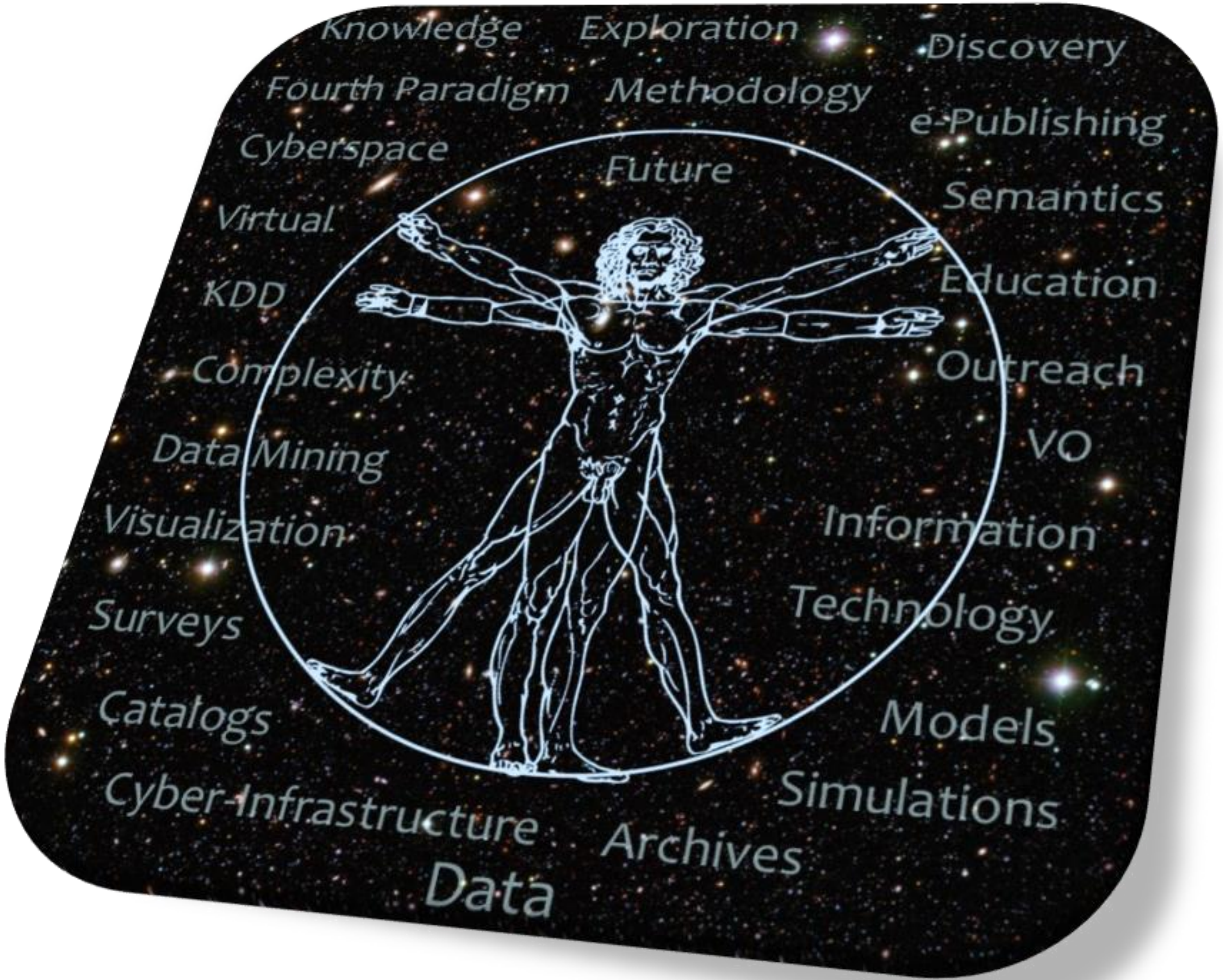


Astroinformatics in the data-driven Astronomy

Massimo Brescia

ISTITUTO NAZIONALE DI ASTROFISICA INAF

2017 ICT Workshop

Astronomy vs Astroinformatics

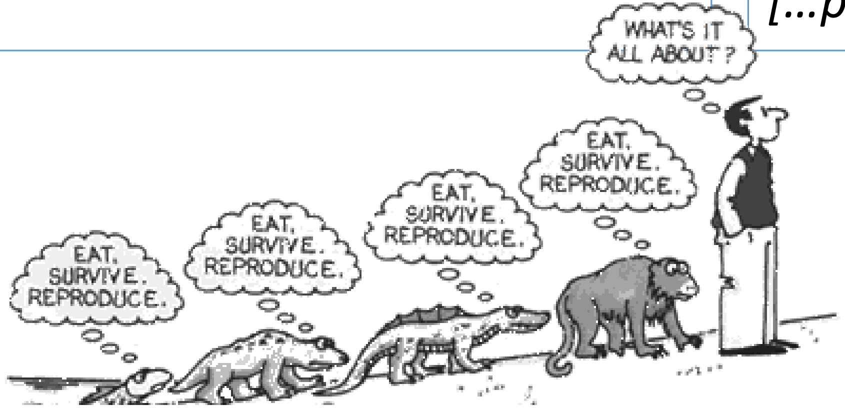
Most of the initial time has been spent to find a common language among communities…

How astronomers see astroinformaticians

How astroinformaticians see astronomers

…with doubtful but promising results

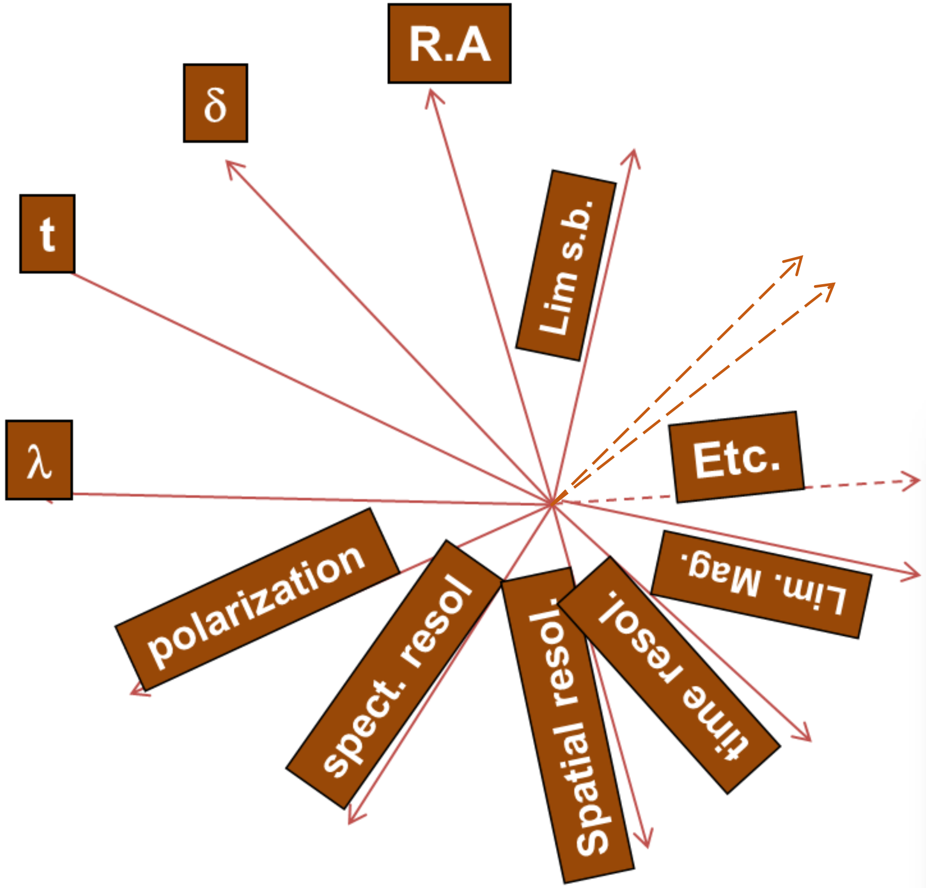
What is NOT Astroinformatics	What IS Astroinformatics
Look up sky object coordinates in an archive	Search for sky objects in an archive to find protometric similarities
Query a database search engine for information about «magnitude type»	Predict nature of sky objects in different catalogues, based on their physical features
Monitor the number of accesses to an astronomical database	Correlate accesses to an astronomical database with visualized information
Configure, improve and maintain the employee’s server infrastructure	Evaluate statistical speedup/data analytics tests about the server infrastructure
Perform electronic payment of the salaries of astronomers	Compare salaries of astronomers with their work production *[…please, don’t ask such service!!!]*



From where the word Astroinformatics come from?

If we collect a complete set of parameters (high-dimensional data) for a complete set of items within our domain of study, then we would have a *perfect* statistical model for that domain.

In other words Big Data becomes the model for a domain X → we call this **X-Informatics**



Anything we want to know about that domain is specified and encoded within the data (i.e. data-driven).

Therefore, the final goal of any Big Data Science is to find those encodings, patterns and knowledge nuggets.

But Big Data Science means:

HPC demanding
Efficient cross-matching
Self-adaptive learning machines

Some examples in next slides

Example: Multi-Messenger Astrophysics

Combination of gravitational wave, interferometer and electromagnetic telescopes to better know the Cosmos.

Our understanding of the Universe is very different from the one that depended only on optical telescopes. 90% of the Universe is dark, emitting no electromagnetic radiation. But now we know that it interacts also gravitationally, by emitting GWs.

~220 TB per year (LIGO+GEO)
+ data from other multi-messenger counterparts

Real-time matched filtering requires ~ 100 Gflops

Example: LSST

Large Synoptic Survey Telescope

3-Gigapixel camera

A **6 GB** image every 20 seconds

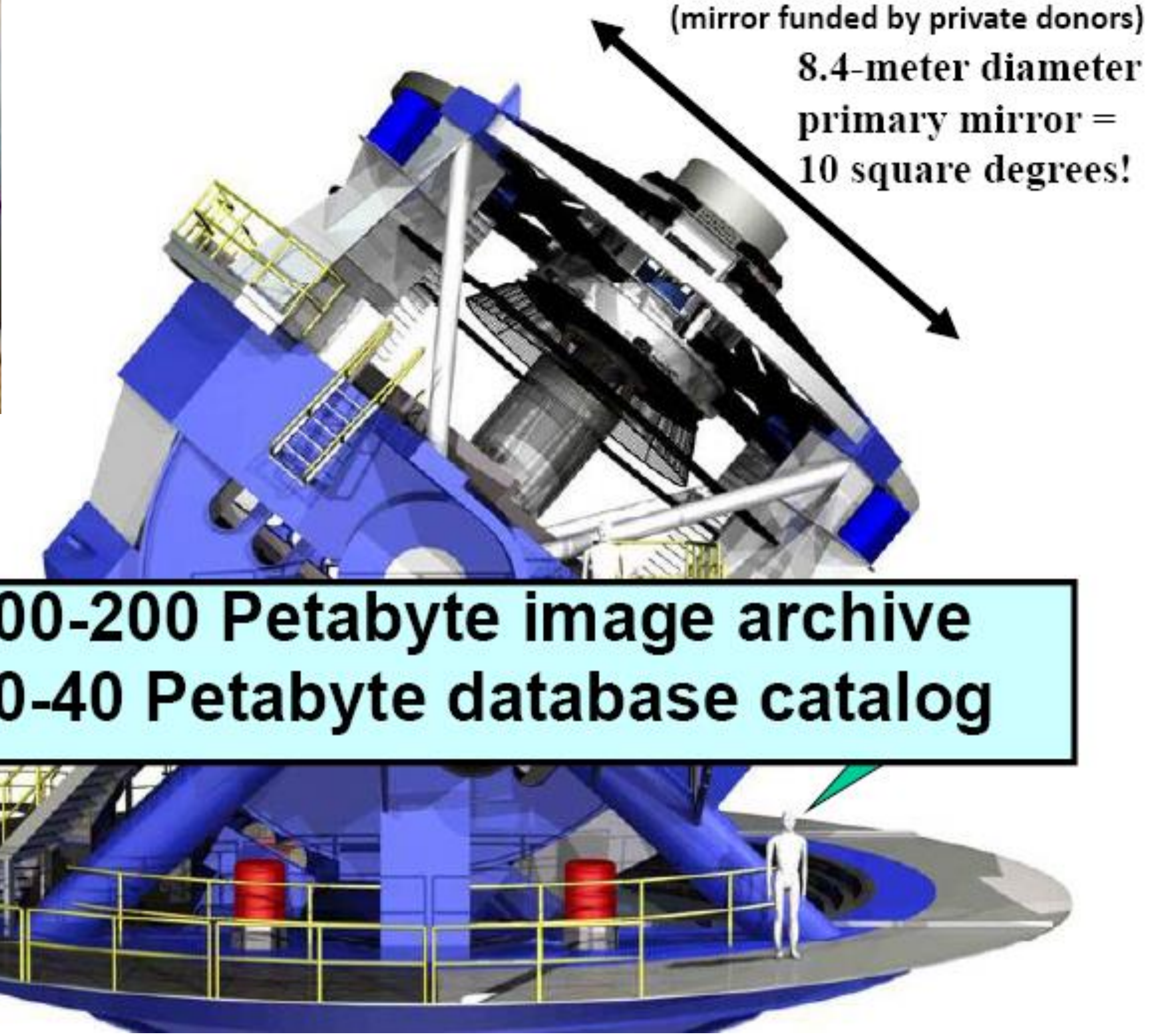
30 TB every night for 10 years

100 PB final image data archive (all public)

20 PB final science catalogue database

50 billion object database

Real-time event mining: **~10million events** per night X 10 years (and for most of them a follow-up observation is required…!!!)



Example: SKA

Square Kilometer Array

~100 PB data gathered every day

Data collected in one single day would take **~2million years** to playback on an ipod

So sensitive to detect an airport radar on a planet at **~30 light years** far from Earth

SKA central computer will have the processing power of about **100 million PCs**

antennas will produce **100 times** the global internet traffic

SKA will use enough optical fibre to wrap **twice** around the Earth

So, what is Astroinformatics?

Astroinformatics arises from the **X-Informatics** paradigm, also known as fourth paradigm of Science

After Theory, Experiments, Simulations, the 4th paradigm is **data-driven Science** = Scientific Knowledge Discovery in Databases

Astroinformatics (Knowledge Discovery in Astrophysical Databases):

- Characterize the known
  - Feature selection, Parameter space analysis
- Assign the new from the known
  - Regression, classification, supervised learning
- Explore the unknown
  - Clustering, unsupervised learning
- Discover the unknown
  - Outlier detection and analytics (serendipity)
- Benefits of very large datasets:
  - Best statistics of “typical” events, cross-correlation, automated search for “rare” events

Basic astronomical knowledge problems #1

The clustering problem:

Finding clusters of objects within a data set

What is the significance of the clusters (statistically and scientifically)?

What is the optimal algorithm for finding friends-of-friends or nearest neighbors?

N is $>10^{10}$, so what is the most efficient way to sort?

Number of dimensions ~ 1000 – therefore, we have an enormous subspace search problem

Are there pair-wise (2-point) or higher-order (N-way) correlations?

N is $>10^{10}$, so what is the most efficient way to do an N-point correlation?

algorithms that scale as $N^2 \log N$ won't get us there

Unsupervised Machine Learning Methods:

- need little or none a-priori knowledge;
- do not reproduce biases present in the Knowledge Base;
- require more complex error evaluation (through complex statistics);
- are computationally intensive;
- are not user friendly *(… more an art than a science; i.e. lot of experience required)*

"a blind man in a dark room - looking for a black cat - which may be not there"
Charles Bowen

Basic astronomical knowledge problems #2

Outlier detection: (unknown unknowns)

Finding the objects and events that are outside the bounds of our expectations (outside known clusters)

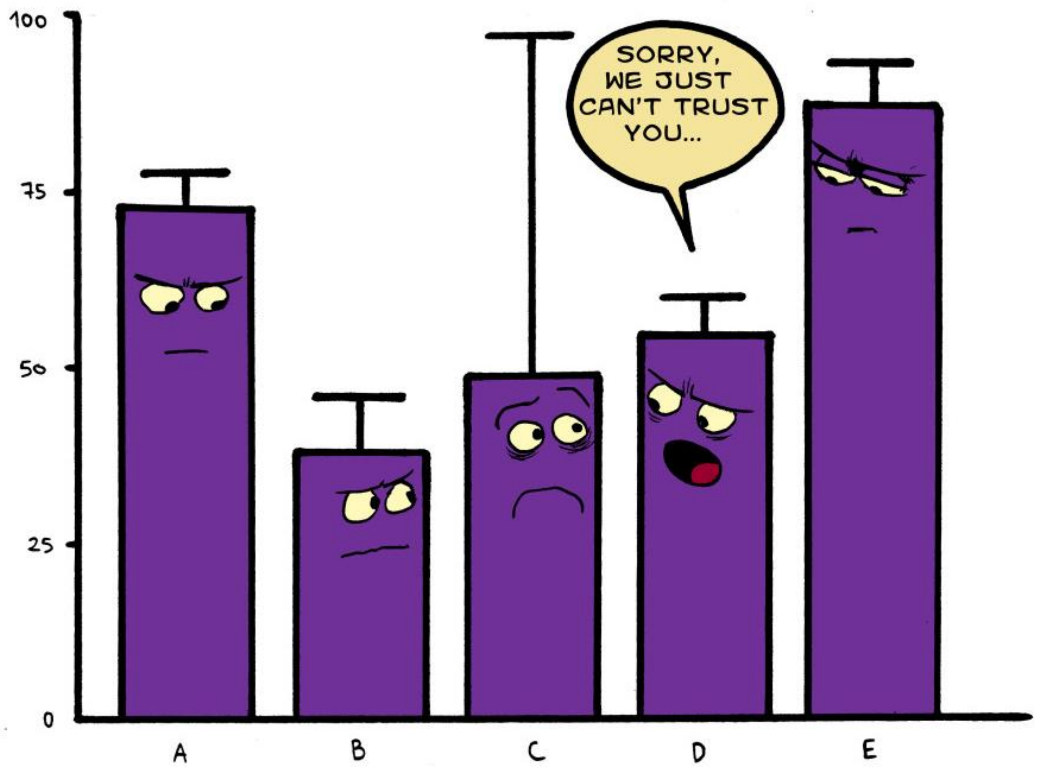
These may be real scientific discoveries (serendipity) or garbage

Outlier detection is therefore useful for:

- Novelty Discovery – *is my Nobel prize waiting?*
- Anomaly Detection – *is the detector system working?*
- Data Quality Assurance – *is the data pipeline working?*

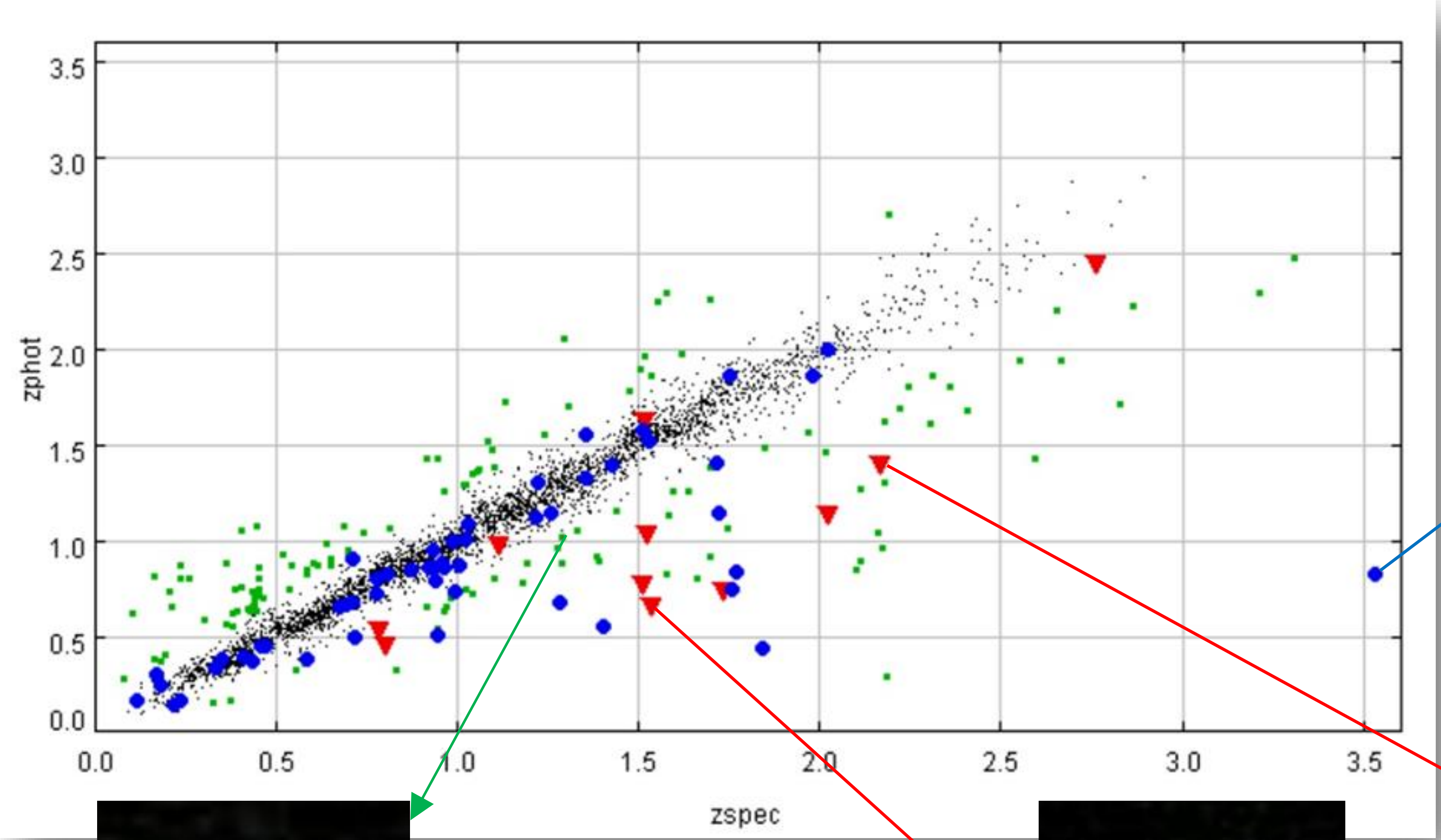
How does one optimally find outliers in 10^3-D parameter space? or in interesting subspaces (in lower dimensions)?

How do we measure their "interestingness"?

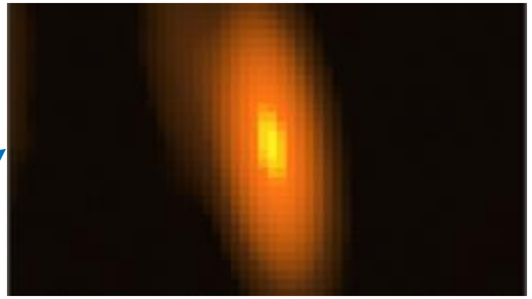


#2 - Catastrophic outliers as peculiar objects

(photo-z for GALEX+SDSS+UKIDSS+WISE QSOs) – *Brescia et al. 2013, ApJ, 772, 2*



- **Blu dots: blazars**
- **Green dots: unknown**
- **Red triangles: gravitationally lensed QSOs**

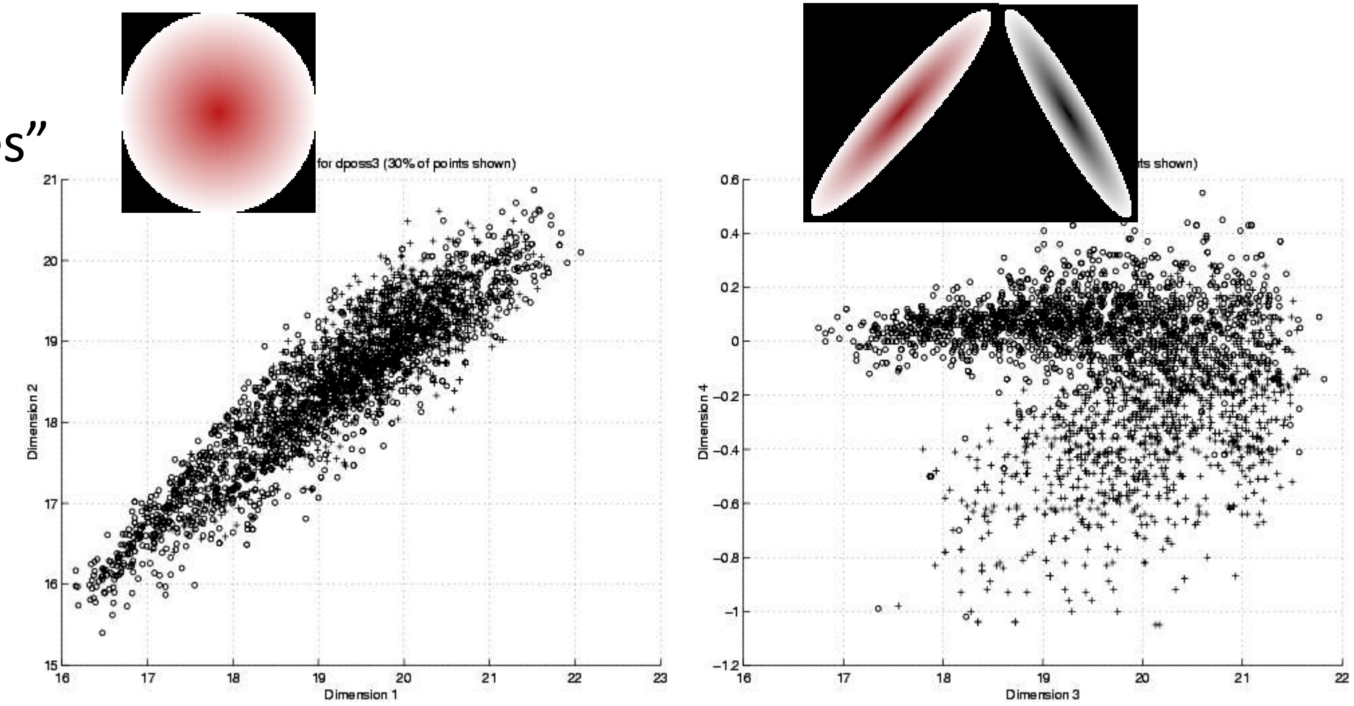


Basic astronomical knowledge problems #3

The dimension reduction problem:

Finding correlations and “fundamental planes” of features in the parameter space

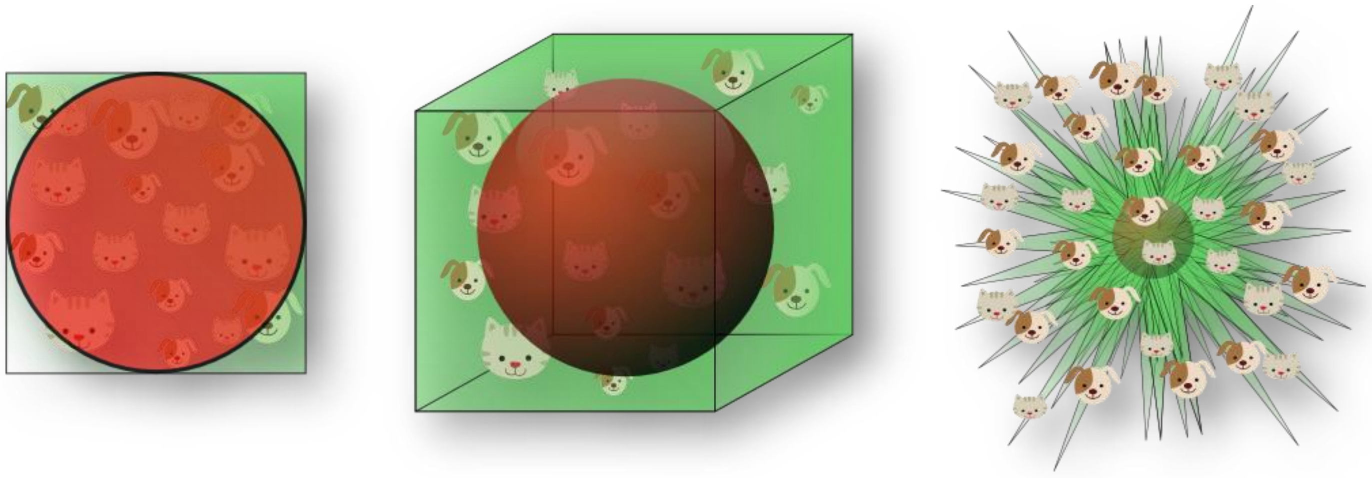
– Number of attributes can be hundreds or thousands, therefore clusters (classes) and correlations may exist/separate in some parameter subspaces, but not in others



- **The Curse of High Dimensionality !**

– Are there combinations (linear or non-linear functions) of observational parameters that correlate strongly with one another?

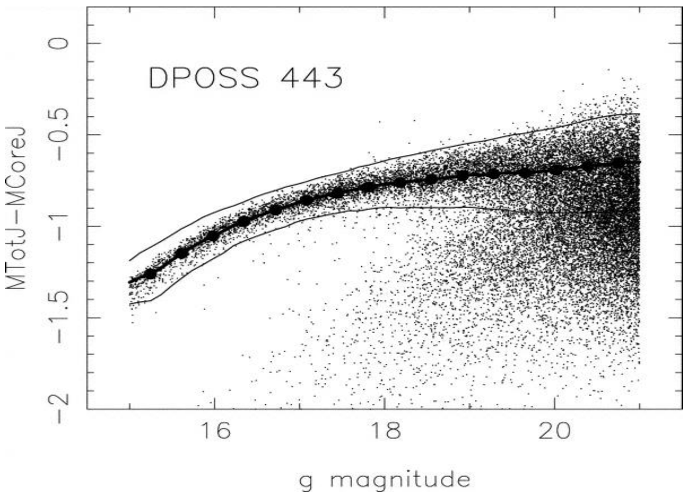
– Are there eigenvectors or condensed representations (e.g., basis sets) that represent the full set of properties?

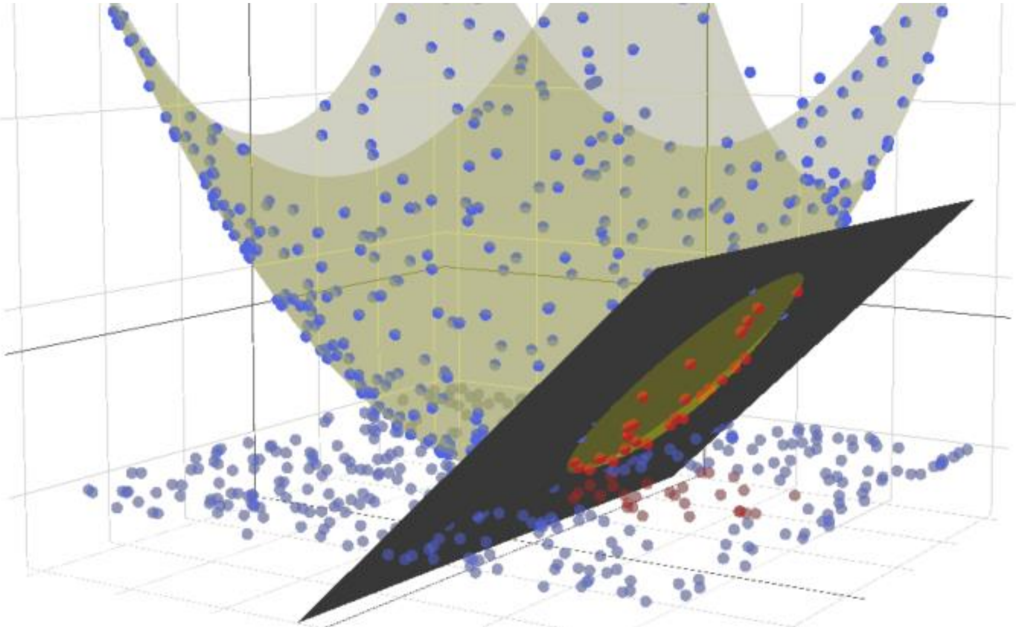


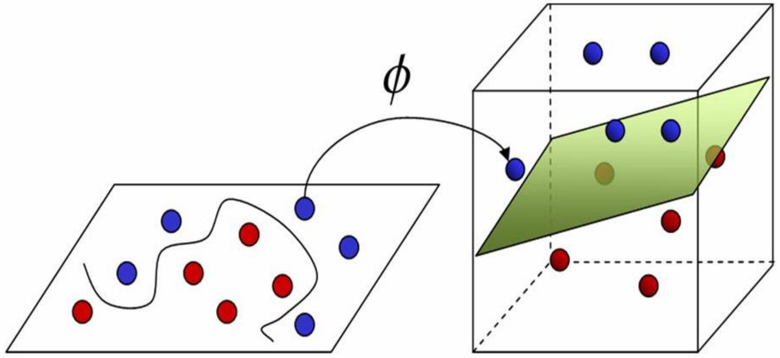
Basic astronomical knowledge problems #4

The superposition / decomposition problem:

Finding distinct clusters (Classes) among objects that overlap in parameter space







What if there are 10^{10} objects that overlap in a 10^3-D parameter space?

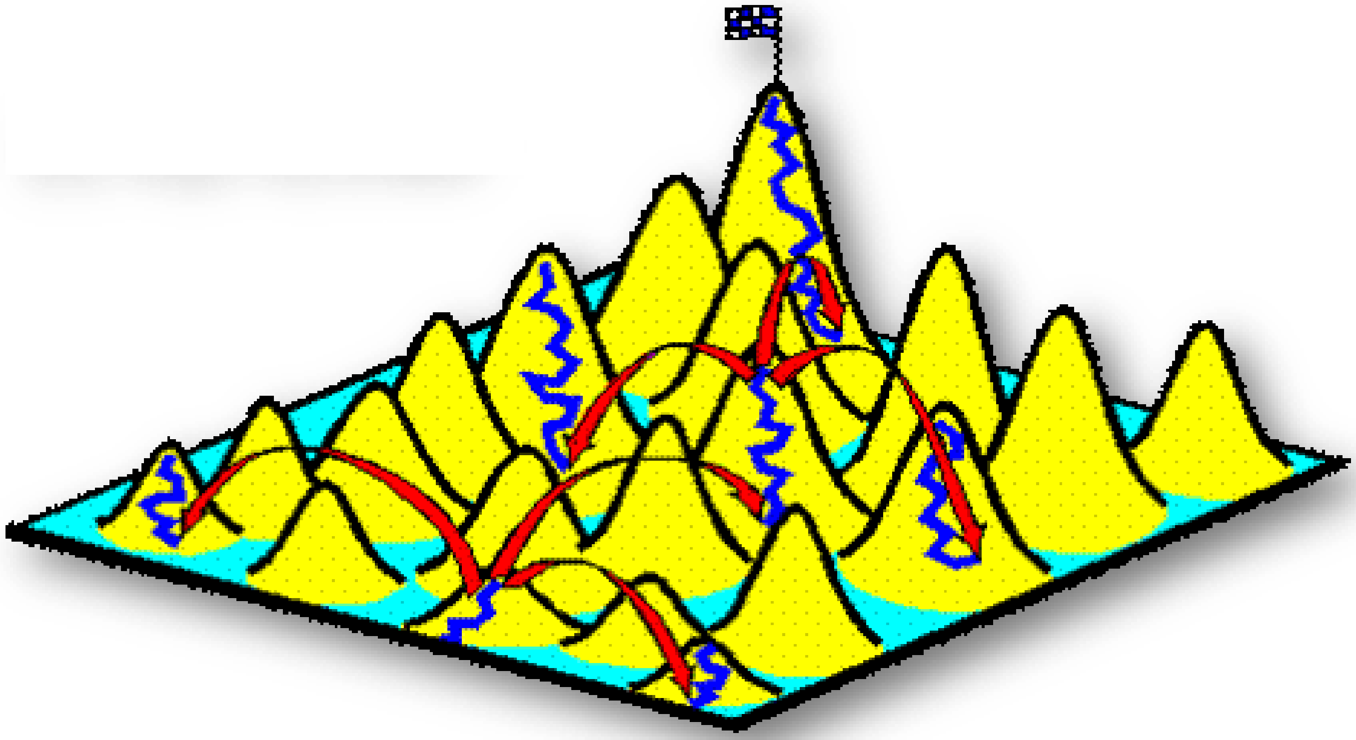
What is the optimal way to separate and extract the different unique classes of objects?

How are constraints applied?

Basic astronomical knowledge problems #5

The optimization problem:

Finding the optimal (best-fit, global maximum likelihood) solution to complex multivariate functions over very high-dimensional spaces



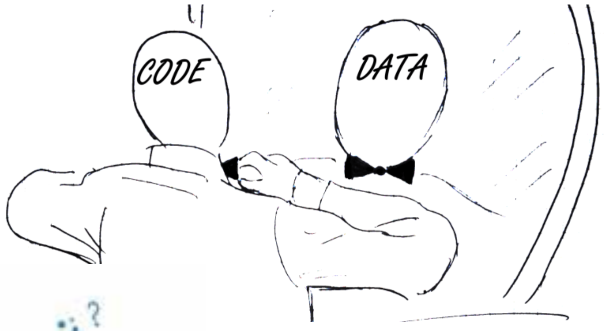
Astroinformatics methodologies

Bayesian classification
Mixture of Gaussians
Error Gradient descent
Error Hessian approximation
Neural Networks
Decision Trees
Genetic Algorithms
Softmax
Cross-entropy
…

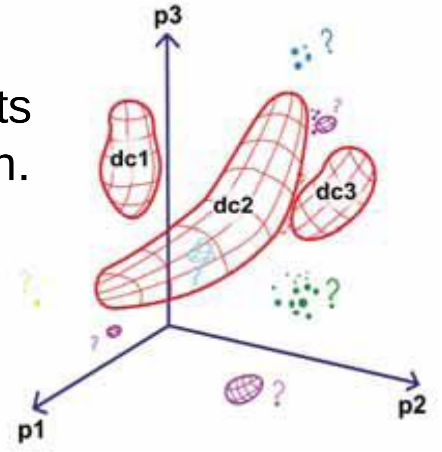
All HPC demanding!!

The changing landscape of astronomical research

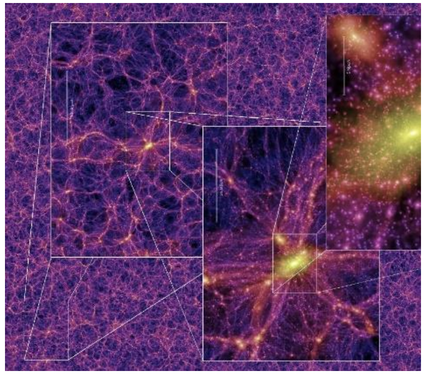
- **Past:** 100's to 1000's of independent distributed heterogeneous data/metadata repositories.
- **Today:** astronomical data are now accessible uniformly from <u>federated</u> distributed heterogeneous sources = **Virtual Observatory**.
- **Future:** astronomy is and will become even more data-intensive in the coming decade with the growth of massive data-producing sky surveys.



Challenge #1: it will be prohibitively difficult to transport the data to the user application. Therefore … ***SHIP THE CODE TO THE DATA !***
We need Distributed Data Mining methodology…



Challenge #2: surveys are useful to measure and collect data from all objects present in large regions of sky, in a systematic, controlled, repeatable fashion. But … **AUTOMATIC SELF-ADAPTIVE METHODS ARE REQUIRED TO EXPLORE AND CROSS-CORRELATE THEIR DATA!**



Challenge #3: we must be ready when huge of data will come. Mock data must be provided to ensure that data analytics methods will be compliant, efficient and scalable. Therefore … ***IMPROVE SIMULATIONS AND INFRASTRUCTURES TO MAKE INTENSIVE TESTS ON YOUR CODE!***

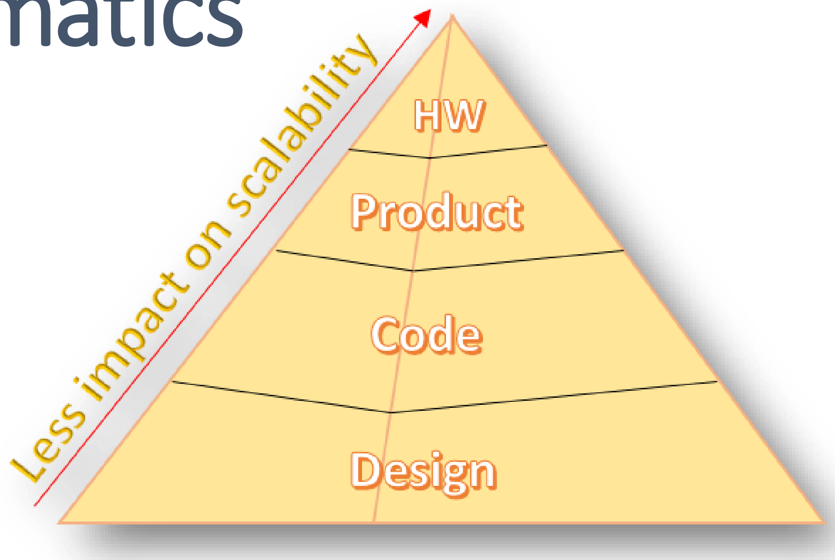
General Challenges in Astronomy over next decade addressable by Astroinformatics

Scalability of statistical, computational & data mining algorithms to peta- and exa- scales

Algorithms to optimize of simultaneous multi-point fitting across massive **multi-dimensional data cubes**

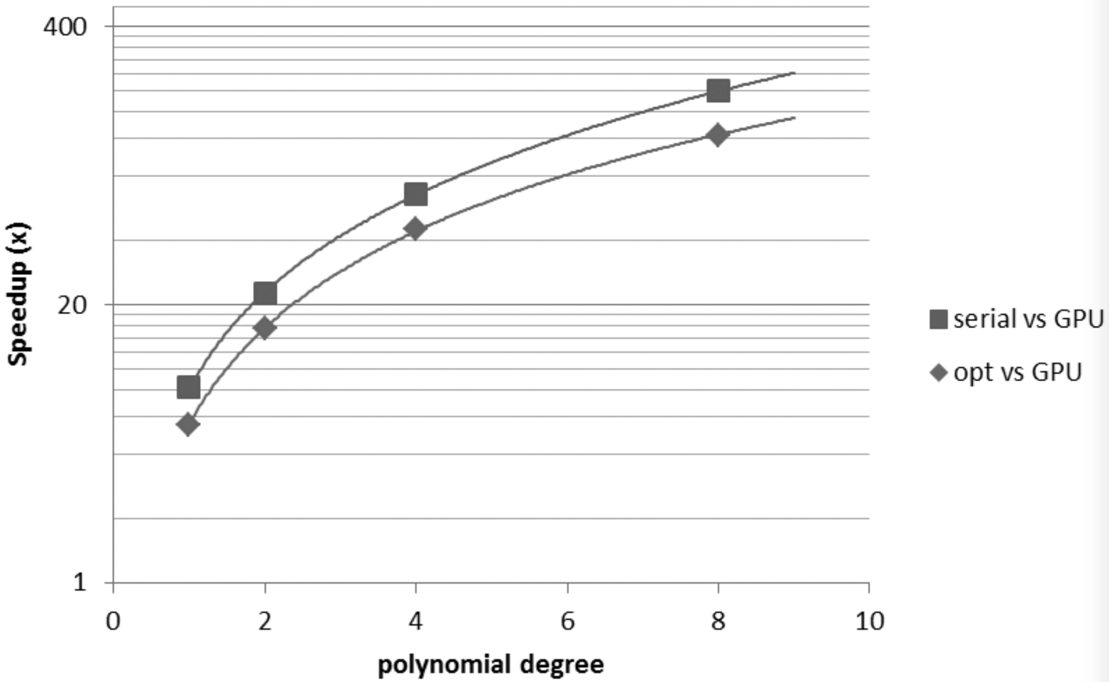
Petascale analytics for **visual data analysis** of massive databases (including feature detection, pattern discovery, clustering, class discovery, dimension reduction)

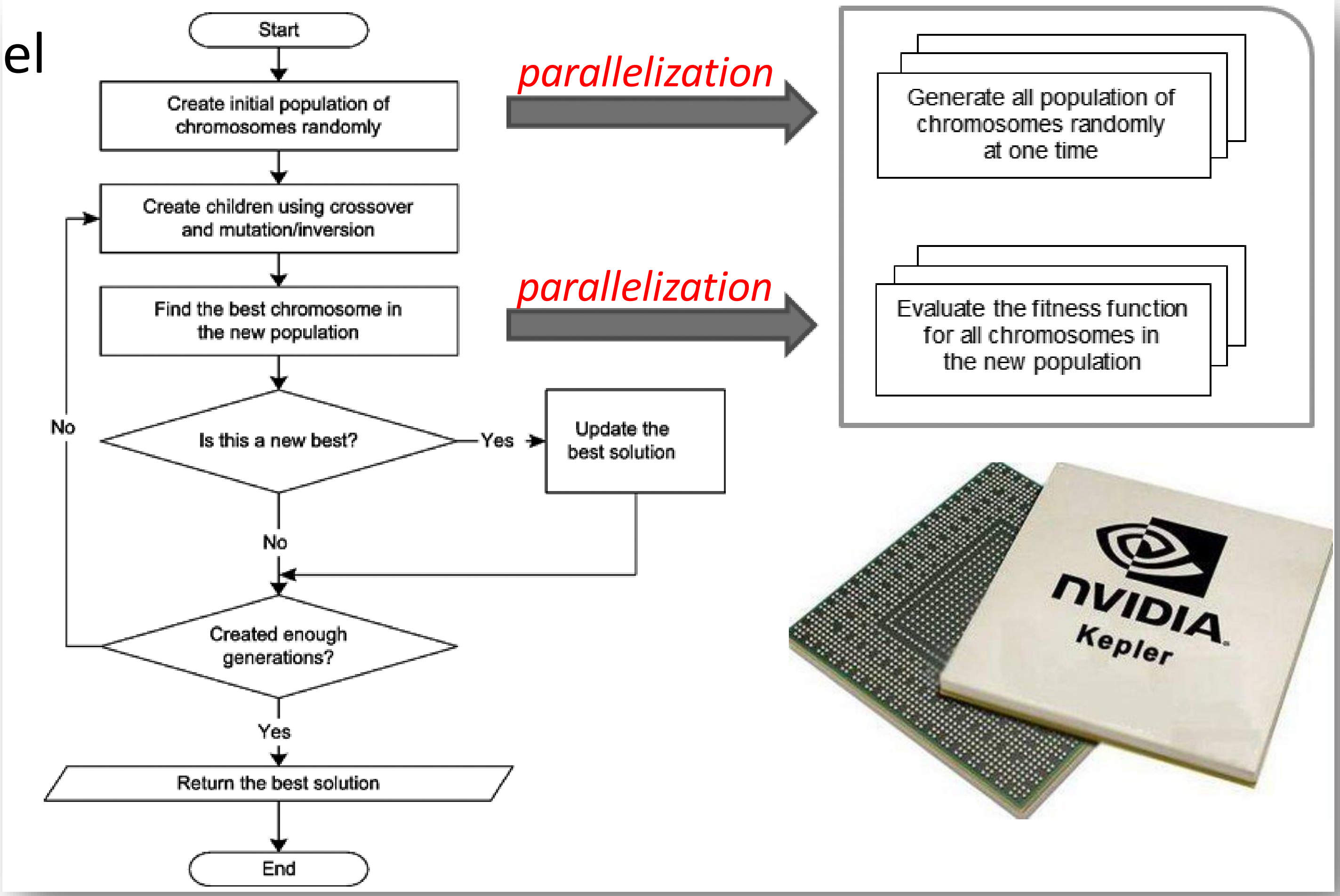
Rapid query, **cross-matching** and search algorithms for highly-dimensional petabyte databases



Virtuous Astroinformatics: Genetic Algorithm with GPU

Typical time-critical machine learning model





GPU outperforms CPU performance by a factor from 8x up to 200x.
It enables intensive use of the algorithm, previously impossible to be achieved with a CPU

Cavuoti et al. 2014, New Astronomy, 26, 12

http://dame.dsf.unina.it/dameware.html

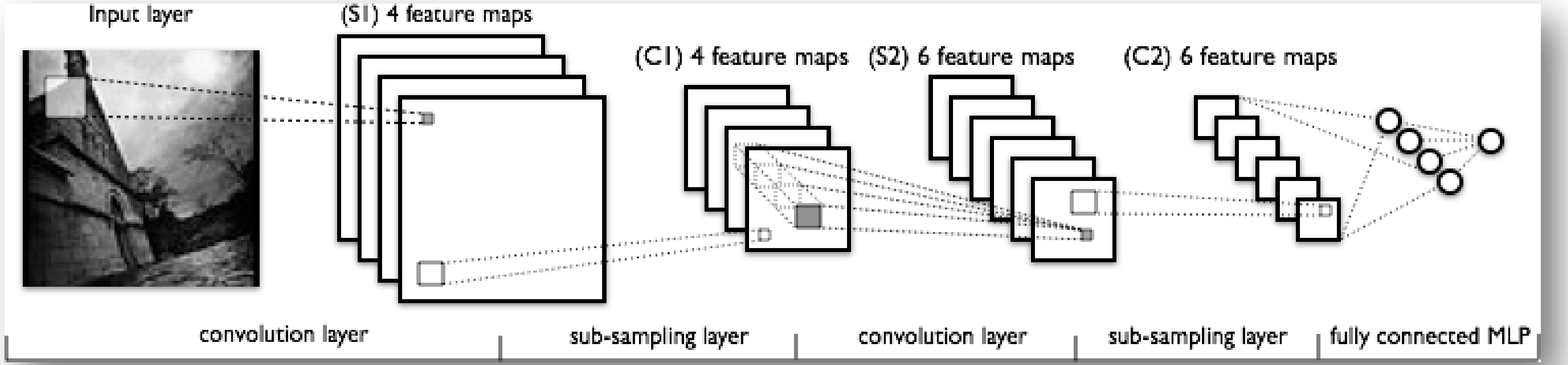
Fitness function based on a polynomial expansion of pattern parameters

$$Out(pat_k) \cong a_0 + \sum_{i=1}^{m}\sum_{j=1}^{d} a_j \cos(j\, f_i) + \sum_{i=1}^{m}\sum_{j=1}^{d} b_j \sin(j\, f_i)$$

- m = number of parameters d = polynomial degree

Virtuous Astroinformatics:

Deep Learning



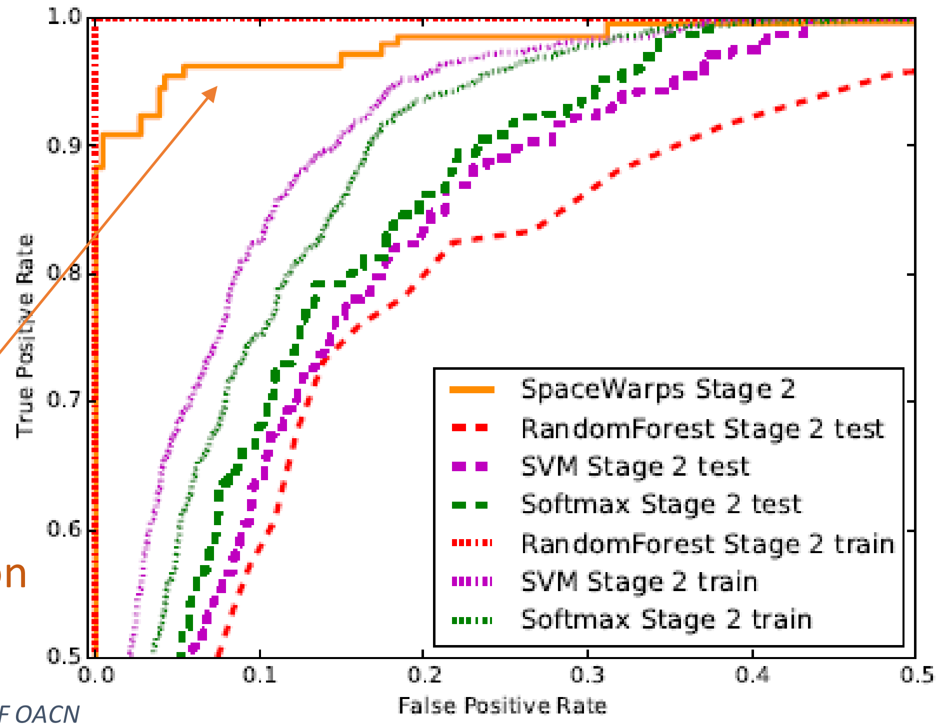
Example of use case: Strong Lensing with CNN

Containing simulated strong lenses

fields with cutouts on the corner

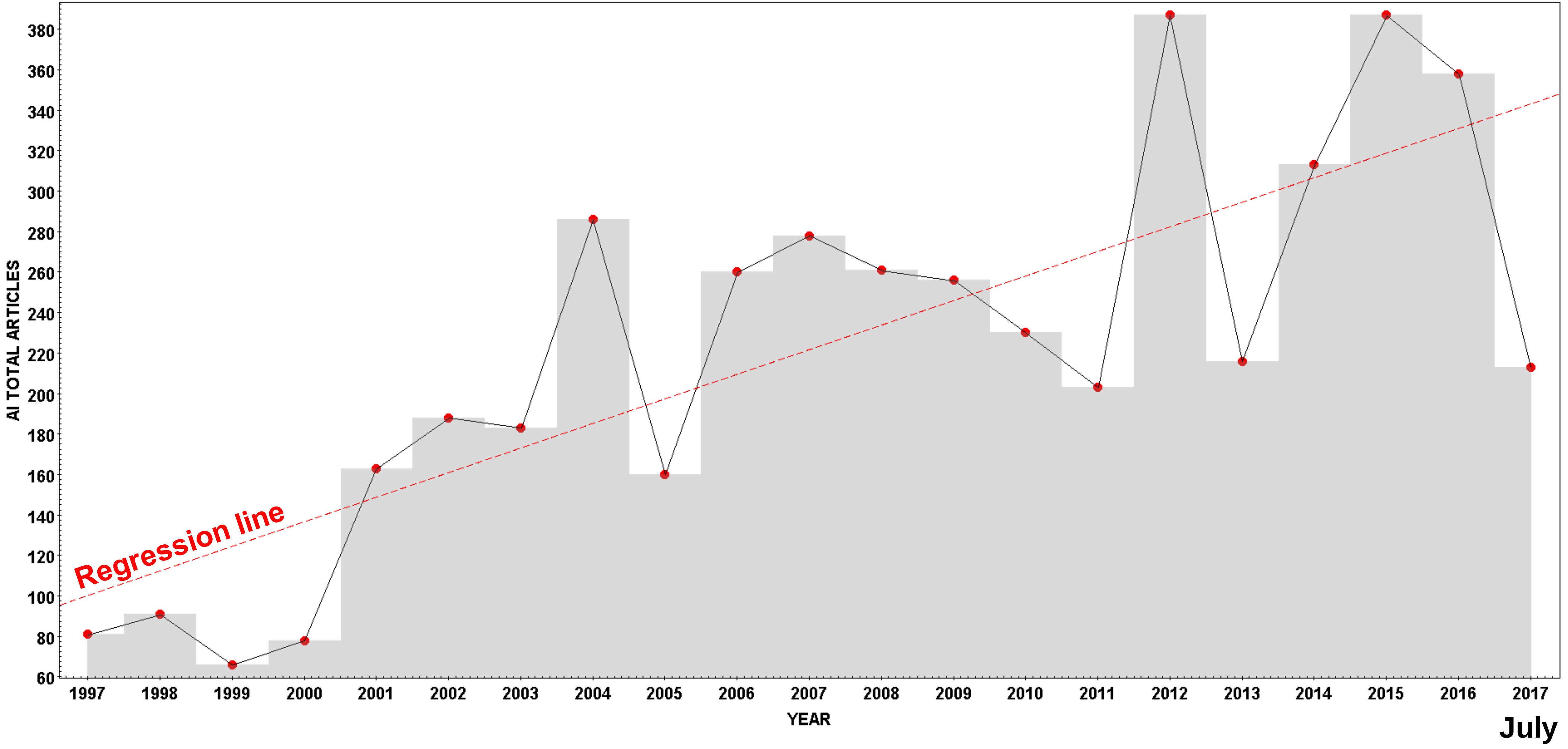
Containing no lenses

(CFHT Legacy Survey) – *More et al. 2016, MNRAS 455, 2*

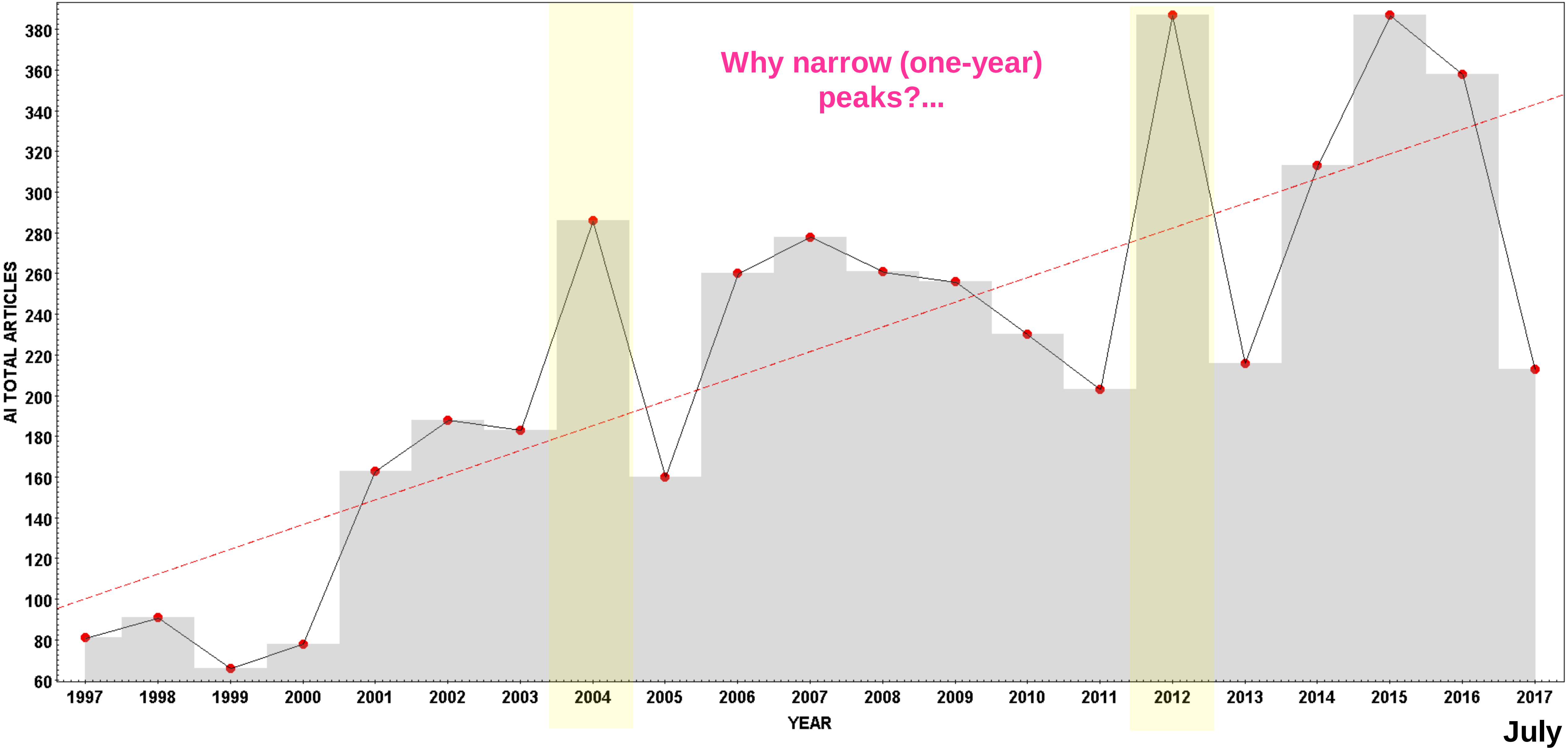


User classification by eyes

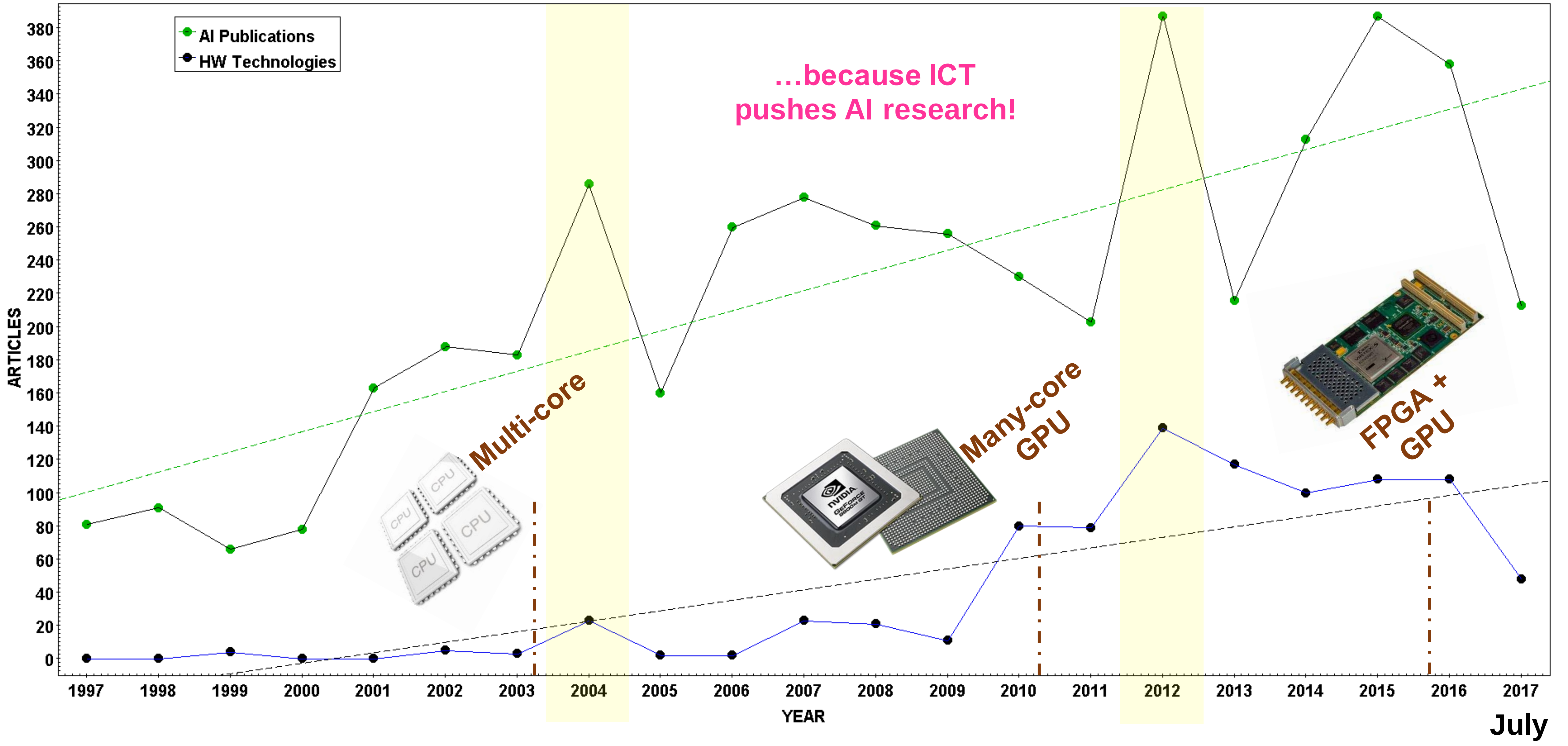
2 decades of Astroinformatics production



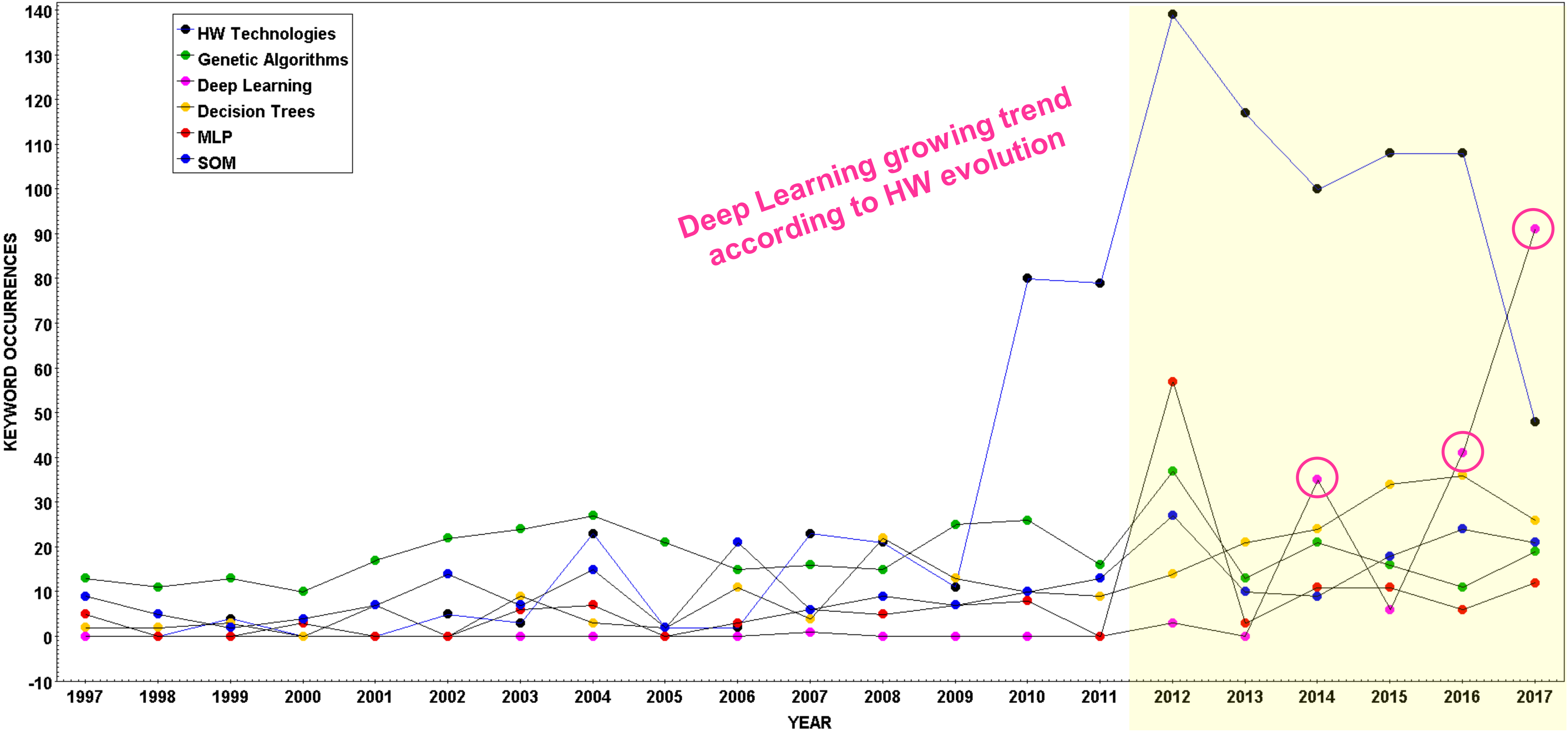
2 decades of Astroinformatics production



2 decades of Astroinformatics production



2 decades of Astroinformatics production



~~Astro~~ Pornoinformatics production

Cornell University Library

We gratefully acknowledge support from the Simons Foundation and member institutions

arXiv.org > cs > arXiv:1511.08899

Search or Article ID inside arXiv | All papers | Broaden your search using Semantic Scholar

(Help | Advanced search)

Computer Science > Computer Vision and Pattern Recognition

Applying deep learning to classify pornographic images and videos

Mohamed Moustafa

(Submitted on 28 Nov 2015)

It is no secret that pornographic material is now a one-click-away from everyone, including children and minors. General social media networks are striving to isolate adult images and videos from normal ones. Intelligent image analysis methods can help to automatically detect and isolate questionable images in media. Unfortunately, these methods require vast experience to design the classifier including one or more of the popular computer vision feature descriptors. We propose to build a classifier based on one of the recently flourishing deep learning techniques. Convolutional neural networks contain many layers for both automatic features extraction and classification. The benefit is an easier system to build (no need for hand-crafting features and classifiers). Additionally, our experiments show that it is even more accurate than the state of the art methods on the most recent benchmark dataset.

Comments: PSIVT 2015, the final publication is available at link.springer.com
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Multimedia (cs.MM); Neural and Evolutionary Computing (cs.NE)
Cite as: arXiv:1511.08899 **[cs.CV]**
(or arXiv:1511.08899v1 **[cs.CV]** for this version)

Submission history

From: Mohamed Moustafa [view email]
[v1] Sat, 28 Nov 2015 13:55:25 GMT (327kb,D)

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Download:

- PDF
- Other formats

(license)

Current browse context:
cs.CV
< prev | next >
new | recent | 1511

Change to browse by:
cs
cs.MM
cs.NE

References & Citations
- NASA ADS

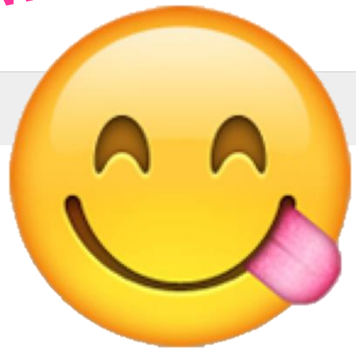
1 blog link (what is this?)

DBLP - CS Bibliography
listing | bibtex
Mohamed Moustafa

Bookmark (what is this?)

Link back to: arXiv, form interface, contact.

Deep Learning growing EVERYWHERE…!



Message for future generation scientists

The modern scientist must become like a Platypus, the most hybrid animal of the Planet (special evolution branch of Darwinian theory)