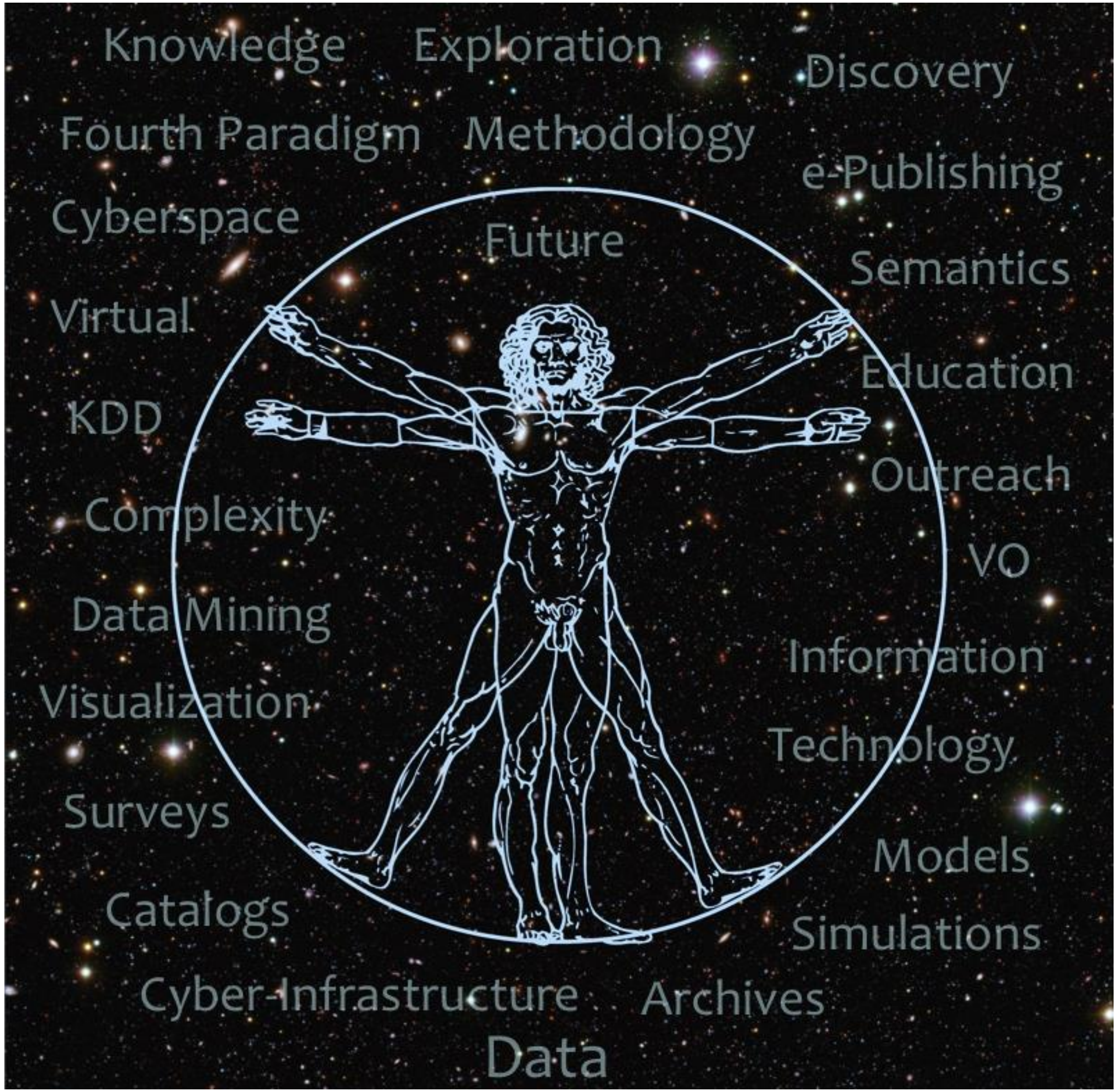


Astroinformatics in a nutshell

Massimo Brescia

Astronomy vs Astroinformatics

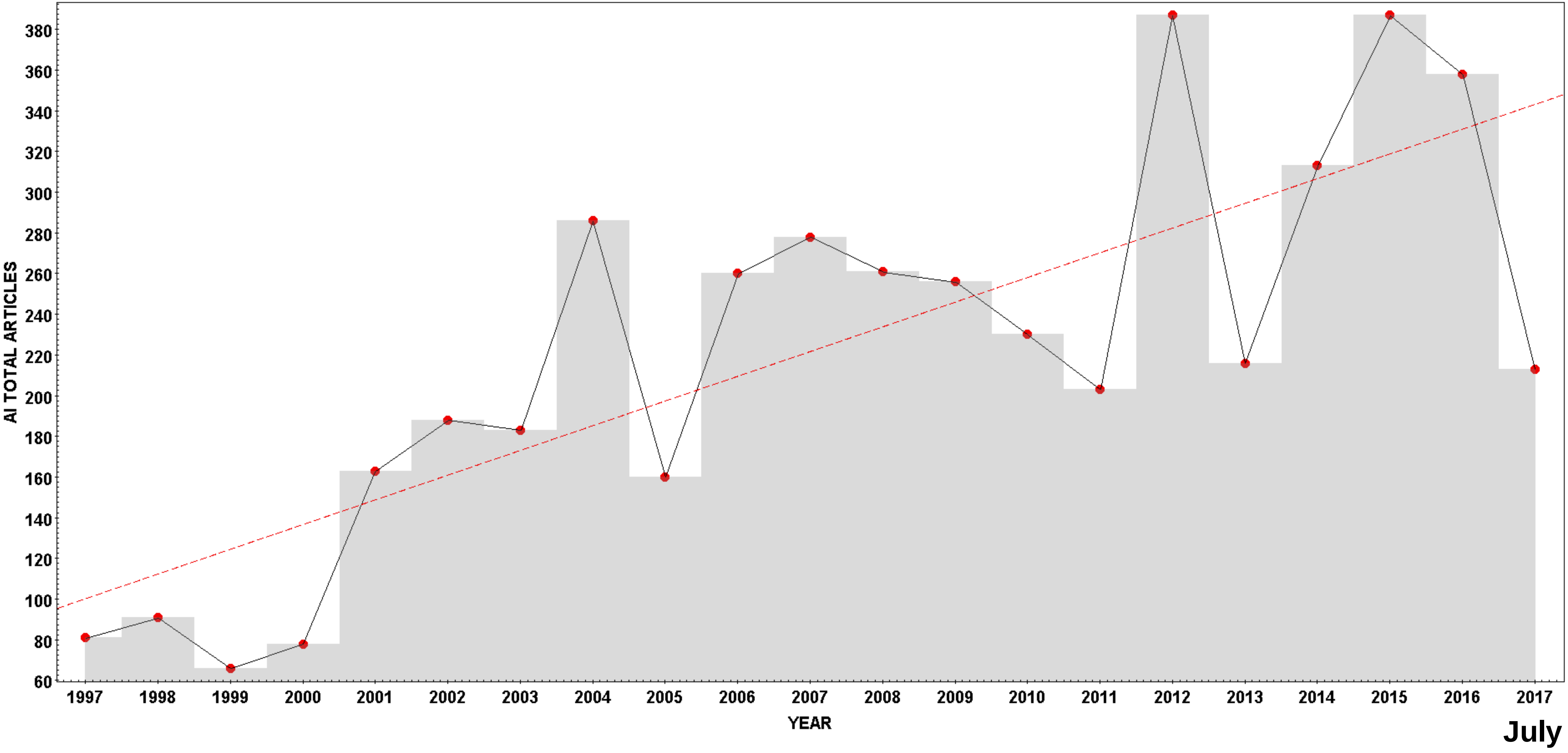
Most of the time has been spent to find a common language among communities…

How astronomers see astroinformaticians

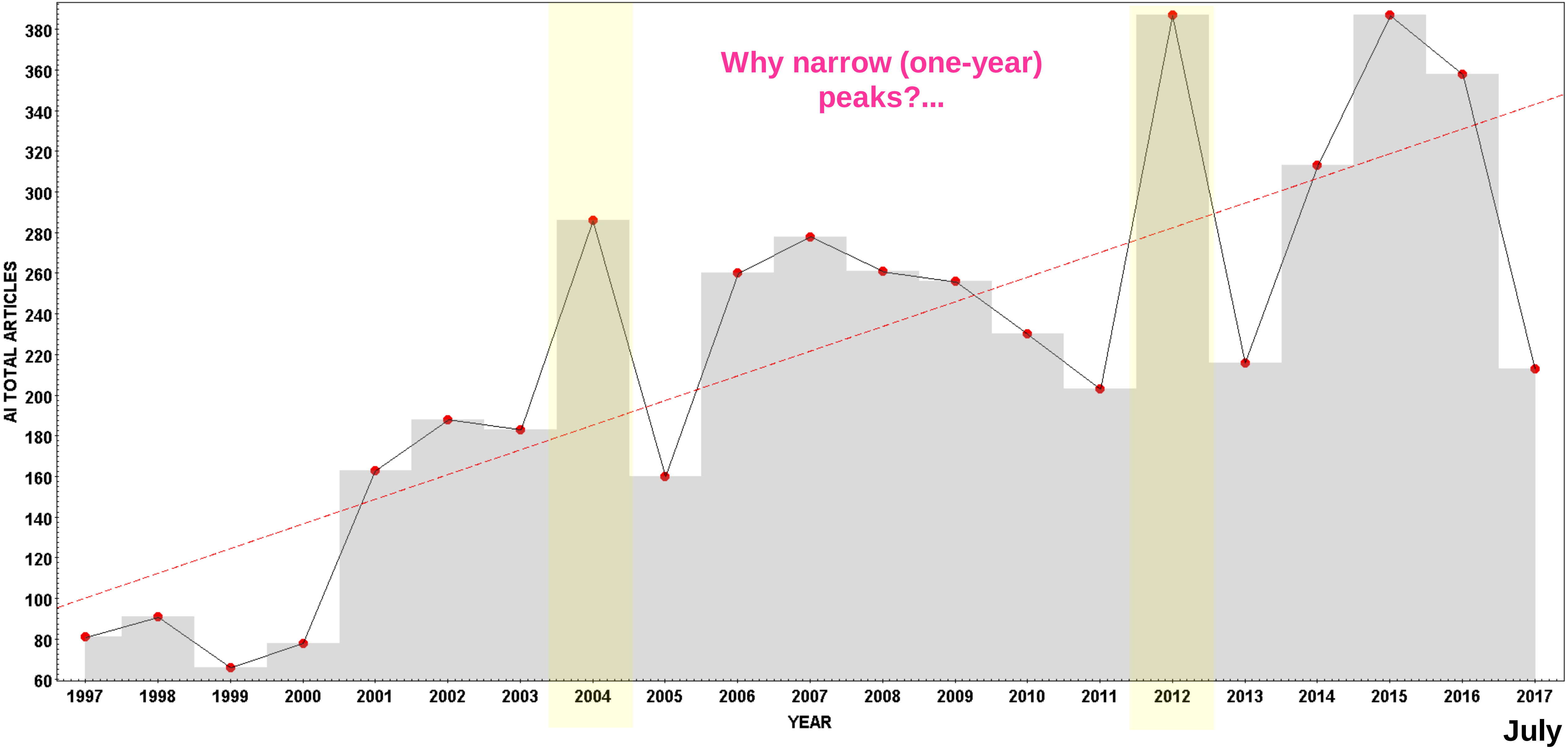
How astroinformaticians see astronomers

…with doubtful but promising results

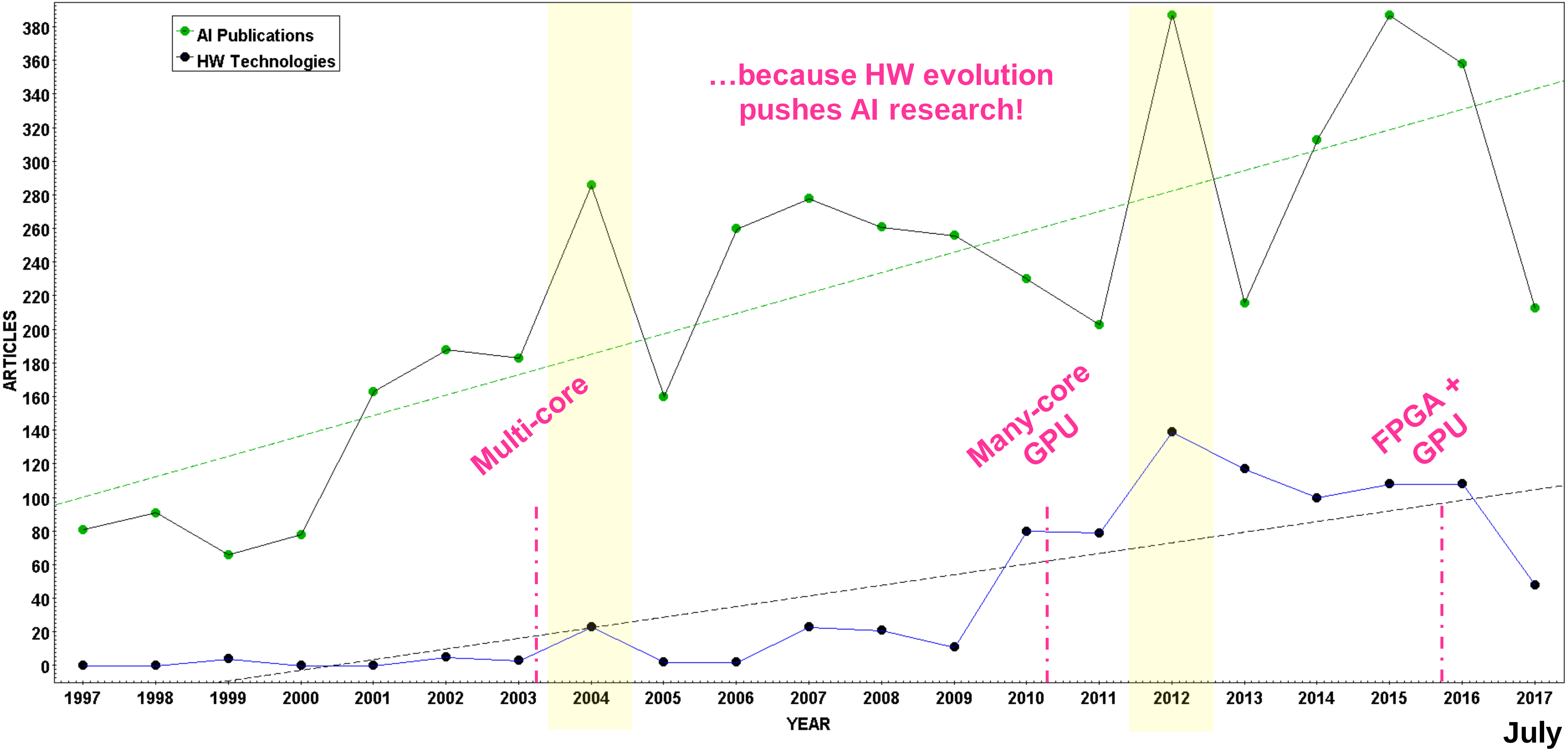
2 decades of Astroinformatics production



2 decades of Astroinformatics production



2 decades of Astroinformatics production



What is NOT Astroinformatics

Look up sky object coordinates in an archive

Query a database search engine for information about «magnitude type»

Monitor the number of accesses to an astronomical database

Configure, improve and maintain the employee’s server infrastructure

Perform electronic payment of the salaries of astronomers

What IS Astroinformatics

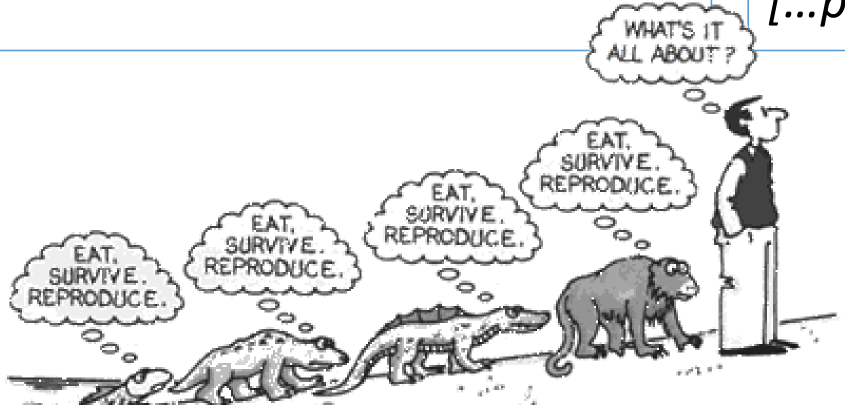
Search for sky objects in an archive to find protometric similarities

Predict nature of sky objects in different catalogues, based on their physical features

Correlate accesses to an astronomical database with visualized information

Evaluate statistical speedup/data analytics tests about the server infrastructure

Compare salaries of astronomers with their work production *[…please, don’t ask such service!!!]*



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 extraction and 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, semi-supervised learning
- Benefits of very large datasets:
  - Best statistics of “typical” events, 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:

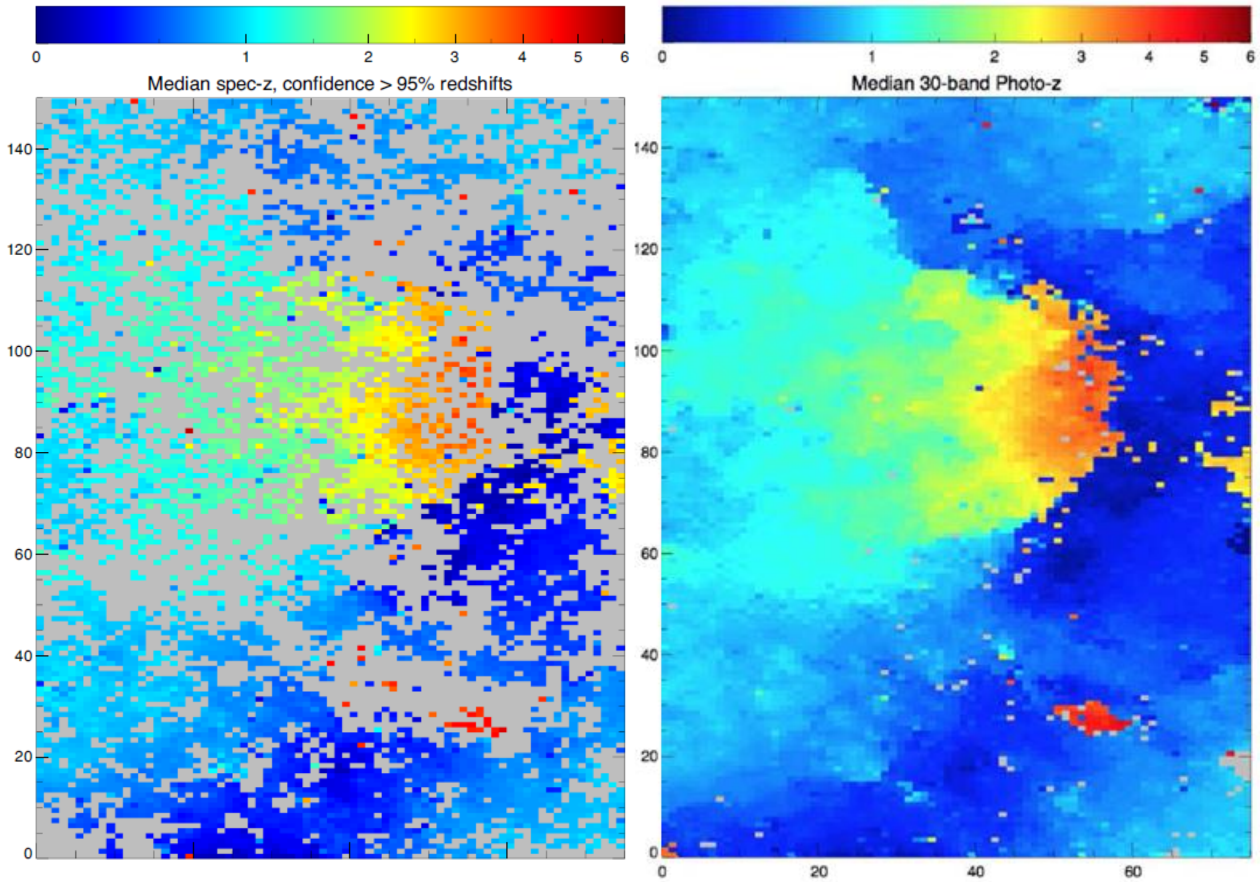
- need little or none a-priori knowledge;
- do not reproduce biases present in the Knowledge Base;
- require more complex error evaluation (through complex statistics);
- are computationally intensive;
- are not user friendly *(… more an art than a science; i.e. lot of experience required)*

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

Charles Bowen

#1 - Clustering with Astroinformatics

(EUCLID photo-z) – *Masters et al. 2015, ApJ, 813, 1*



Projection on the SOM neural surface of Cosmos galaxy photometric distributions after self-adaptive learning
Left: zspec density sampling in colour cells.
Right: photo-z as median of 30-band Cosmos photo-z of galaxies associated with each SOM cell

SOM clustering applied to existing photometric data from the COSMOS survey selected to approximate the anticipated Euclid weak lensing sample.

We can robustly map the empirical distribution of galaxies in the 30-D colour space defined by the expected Euclid filters. Galaxies within a SOM cell have the same SED by definition. SOM is basically a map of the observed SEDs in the universe.

Crucially, the method lets us determine whether a spectroscopic training sample is representative of the full photometric space occupied by the galaxies in a survey.

Mapping this colour distribution lets us determine where - in galaxy colour space - redshifts from current spectroscopic surveys exist and where they are systematically missing.

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)

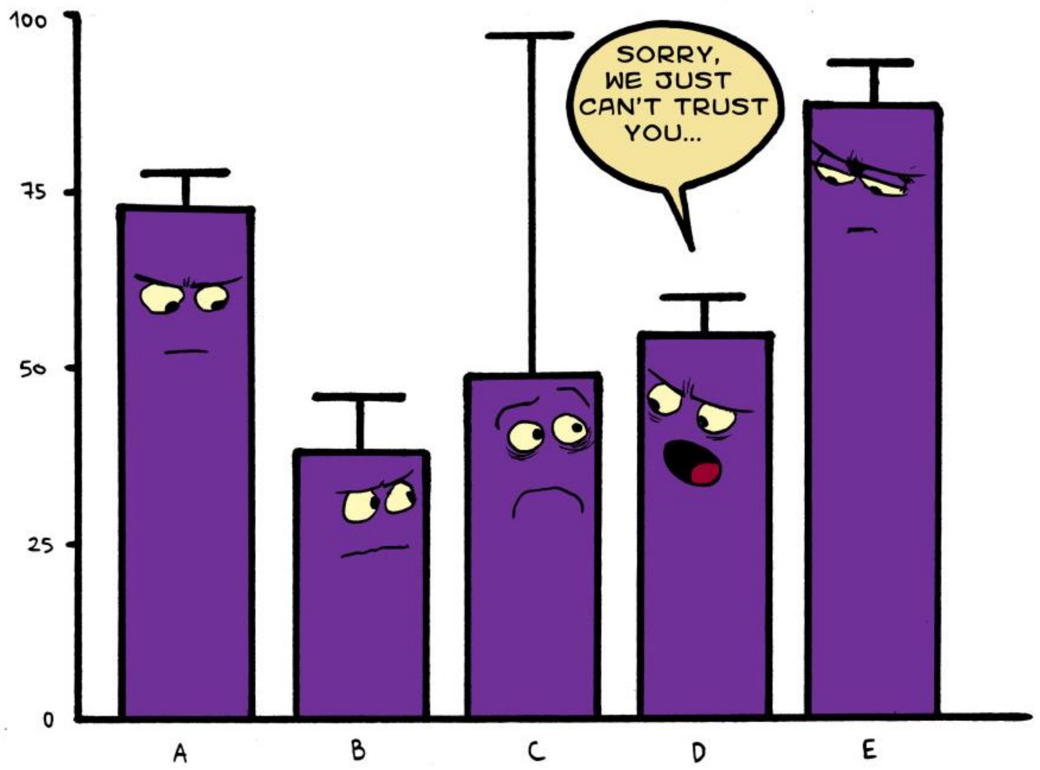
These may be real scientific discoveries or garbage

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?*

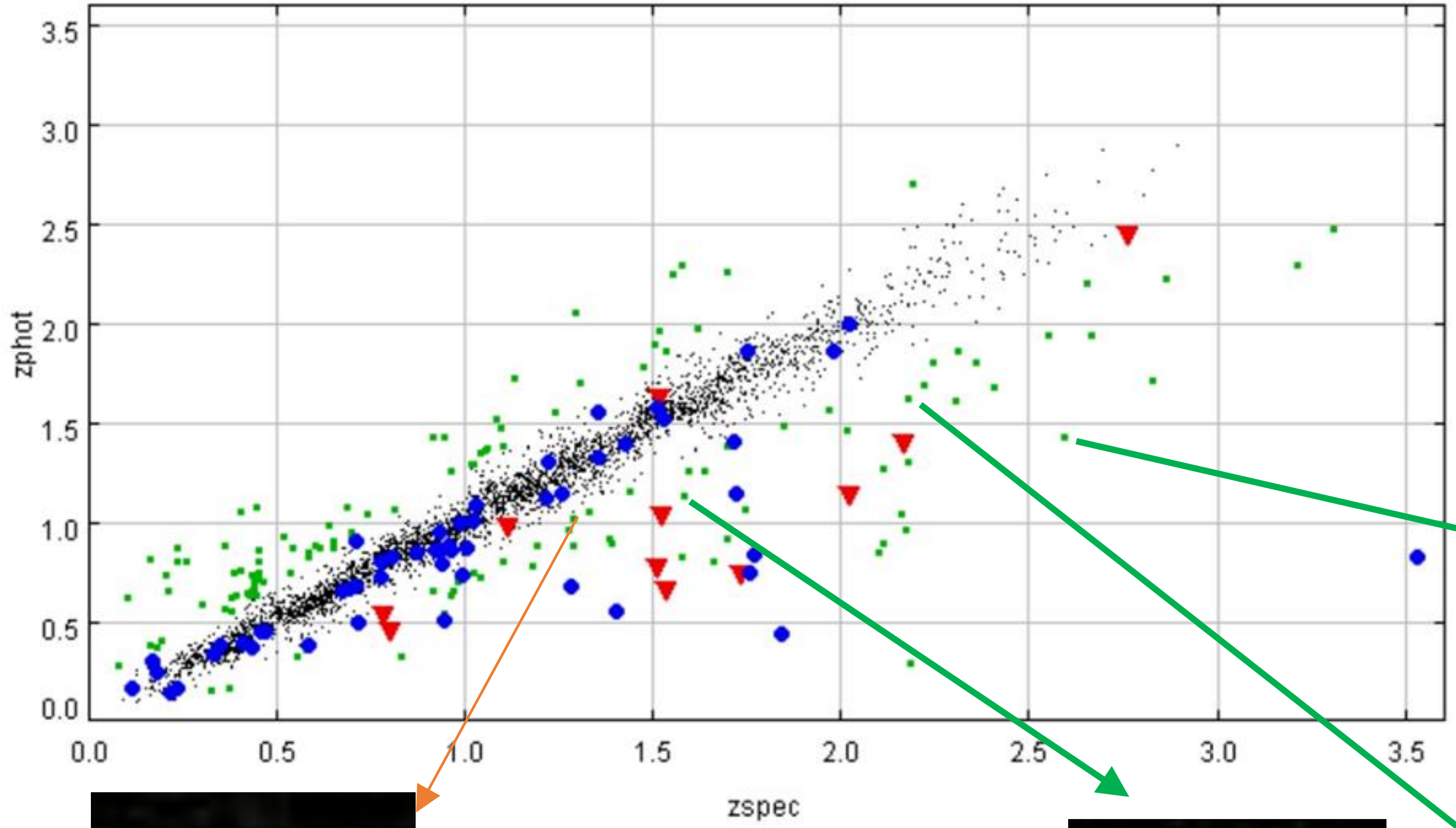
How does one optimally find outliers in 10^3-D parameter space? or in interesting subspaces (in lower dimensions)?

How do we measure their "interestingness"?



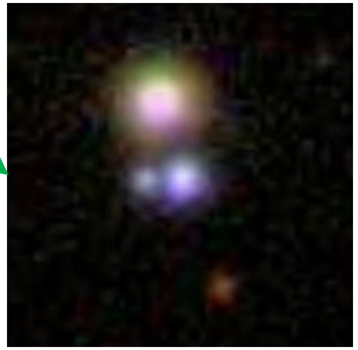
#2 - Catastrophic outliers as peculiar objects

(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**

Peculiar objects

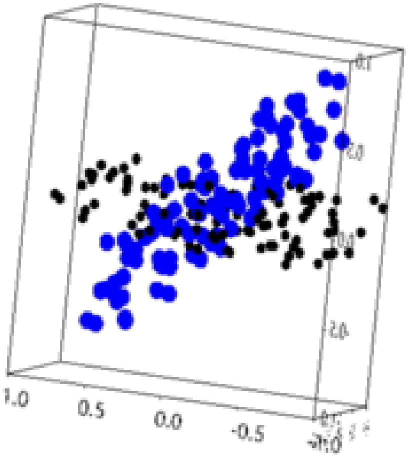


Gravitational lens candidates

Basic astronomical knowledge problems #3

The dimension reduction problem:

Finding correlations and "fundamental planes" of features in the parameter space

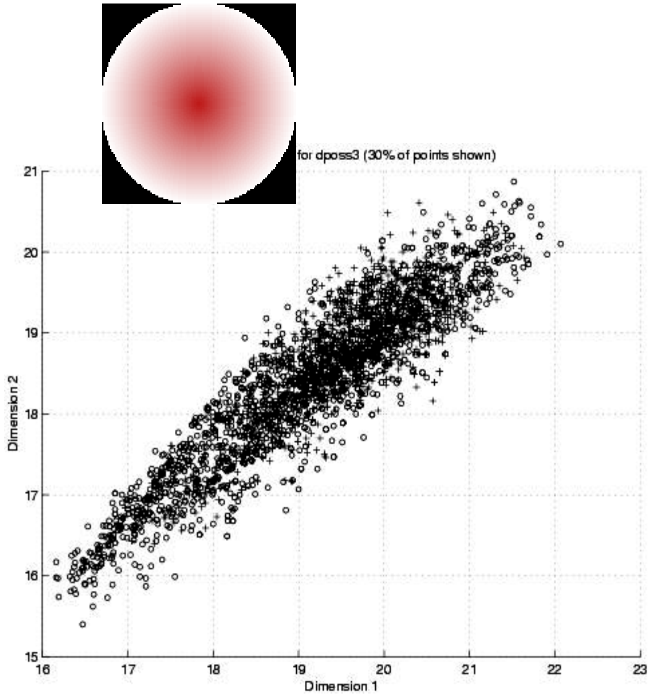


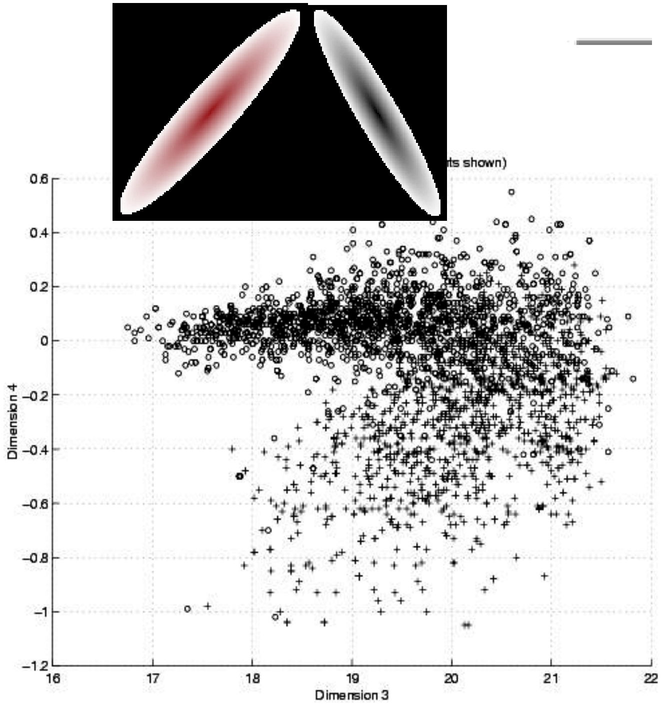
– Number of attributes can be hundreds or thousands, therefore clusters (classes) and correlations may exist/separate in some parameter subspaces, but not in others

- **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?





#3 – Feature Analytics

Lesson to be learned

Features which carry most of the information are not those usually selected by the astronomer on his/her personal experience….

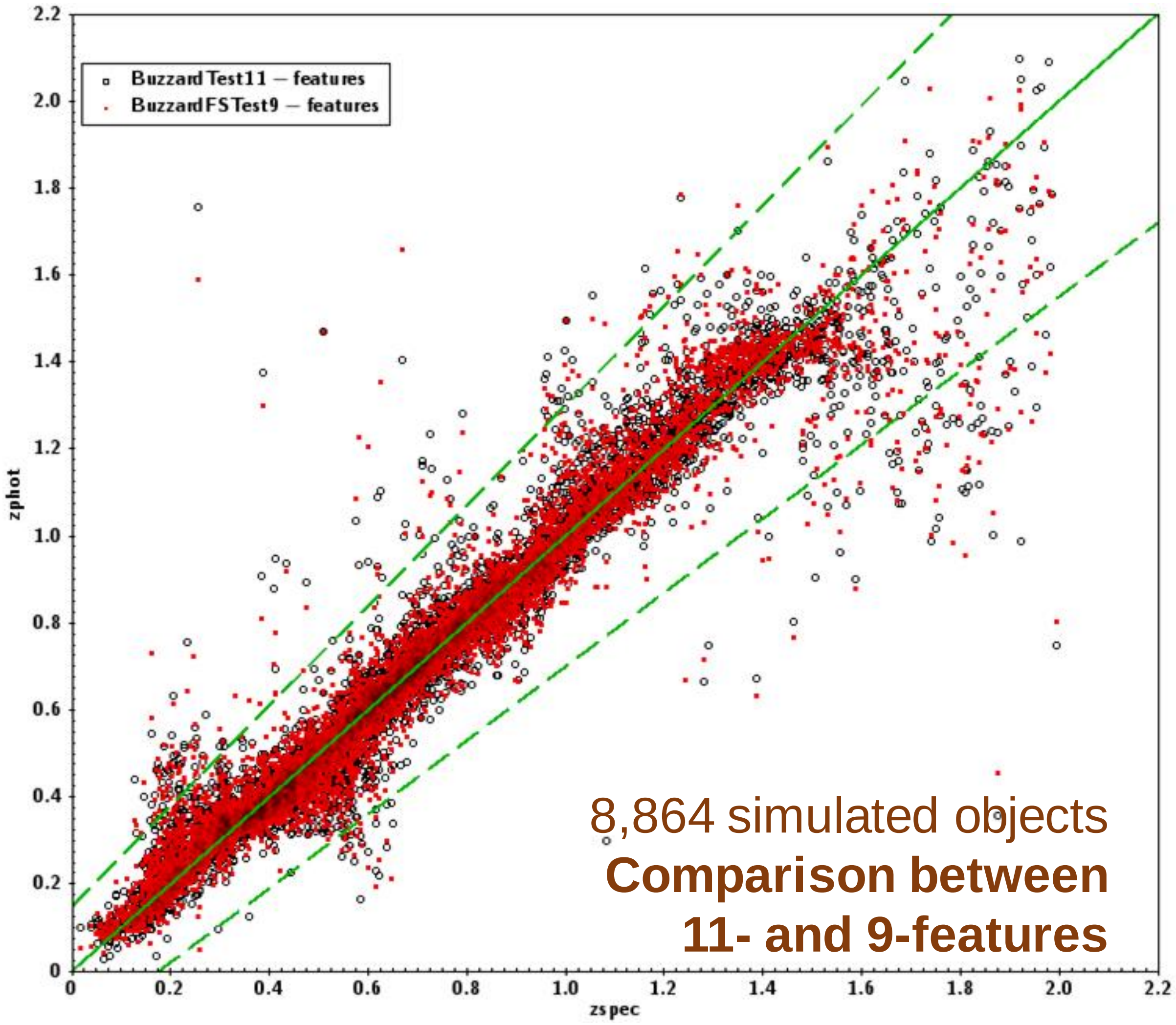
Let the data speak for themselves ?

Time consuming, very demanding in HW and computing power

Feature Selection
with PHiLAB
(*Brescia et al., in prep.*)

Statistics	11-features	9-features
bias	-0.002028	-0.002427
σ	0.050	0.049
NMAD	0.023	0.022
η>0.15	2.12%	2.12%

(LSST photo-z) – *LSST Challenge Collab. 2017, in prep.*

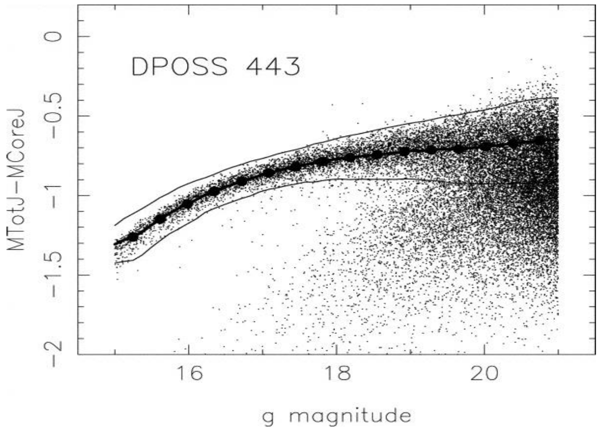


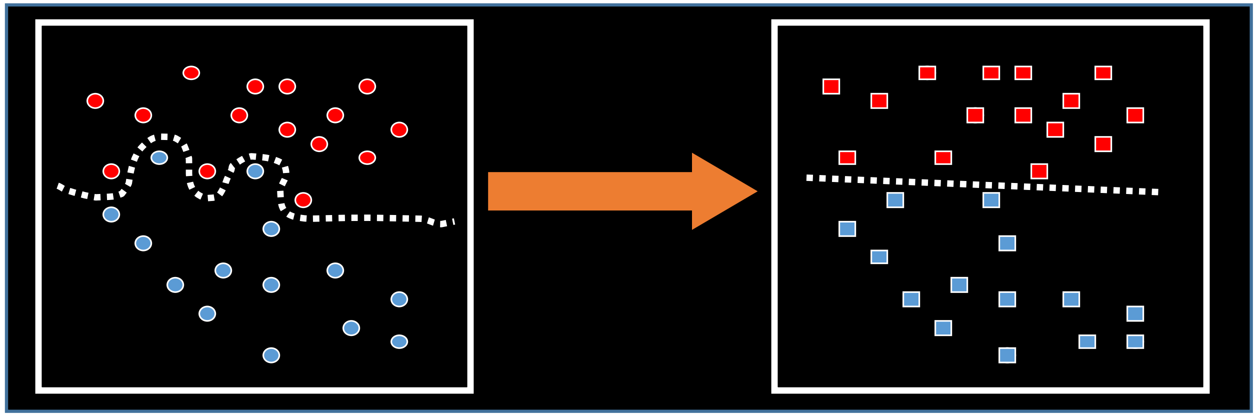
(rejected i and z magnitudes)

Basic astronomical knowledge problems #4

The superposition / decomposition problem:

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





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

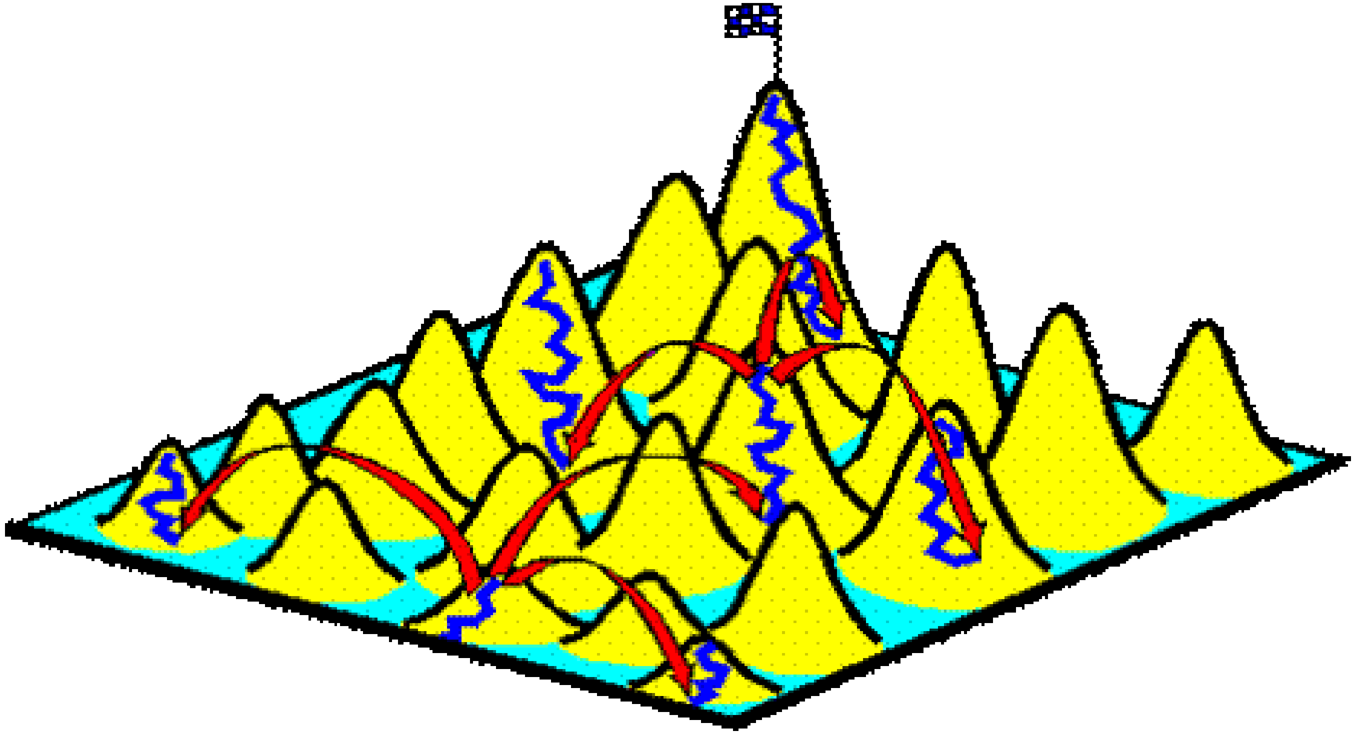
What is the optimal way to separate and extract the different unique classes of objects?

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 methodology

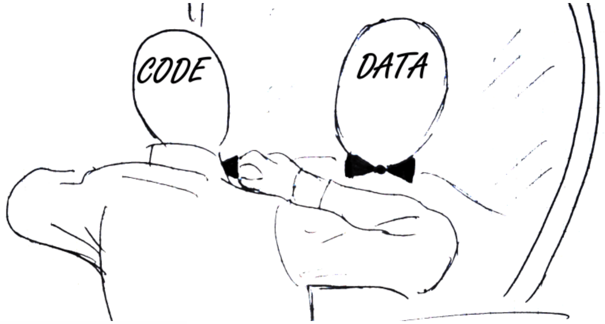
Bayesian classification
Mixture of Gaussians
Error Gradient descent
Error Hessian approximation
Neural Networks
Genetic Algorithms
Softmax
Cross-entropy

Summary of key Astronomy problems where Astroinformatics may help

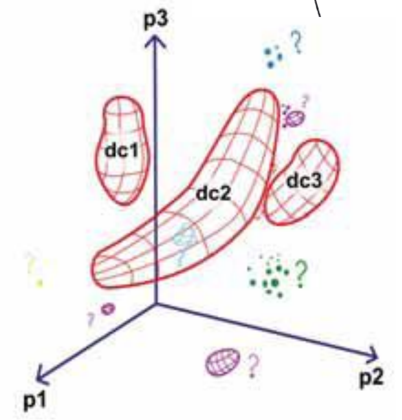
- Efficient Cross-Matching of objects from different catalogues
- The distance problem (*e.g.,* Photometric Redshift estimators)
- Star-Galaxy separation ; QSO-Star separation
- Cosmic-Ray Detection in images
- Supernova Detection and Classification
- Morphological Classification (galaxies, AGN, gravitational lenses, ...)
- Class and Subclass Discovery (brown dwarfs, methane dwarfs, ...)
- Weak and strong lensing detection
- Dimension Reduction = Correlation Discovery
- Learning Rules for improved classifiers
- Classification of massive radio data streams
- Real-time Classification of Astronomical Events
- Clustering of massive data collections
- Novelty, Anomaly, Outlier Detection in massive databases

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.

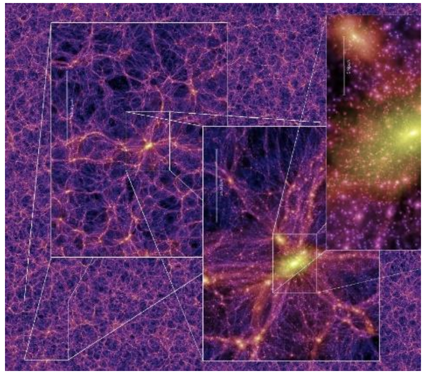


Challenge #1: it will be prohibitively difficult to transport the data to the user application. Therefore … ***SHIP THE CODE TO THE DATA !***
We need Distributed Data Mining methodology…

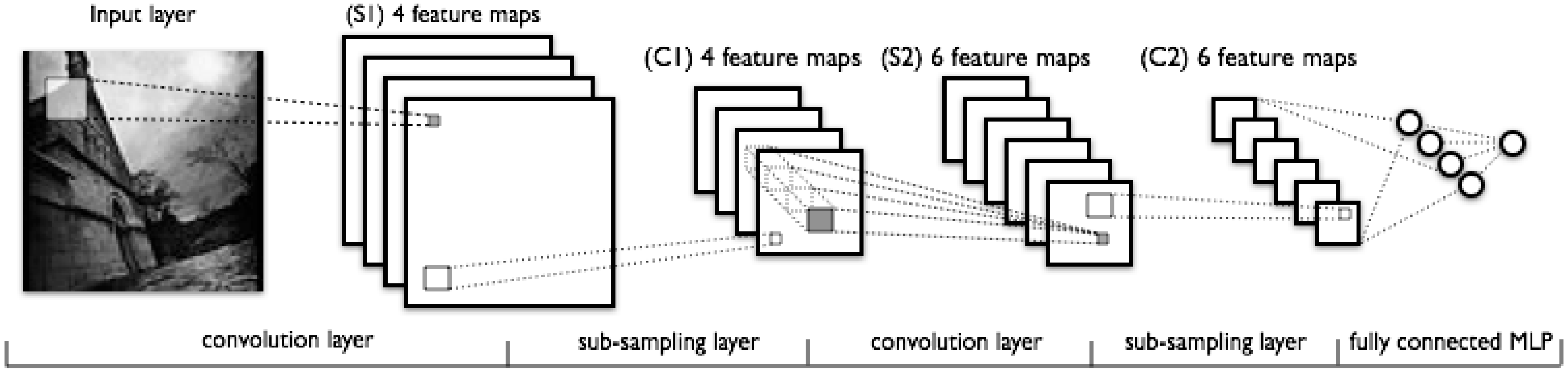


Challenge #2: surveys are useful to to measure and collect data from all objects present in large regions of sky, in a systematic, controlled, repeatable fashion. But … **AUTOMATIC SELF-ADAPTIVE METHODS ARE REQUIRED TO EXPLORE AND CROSS-CORRELATE THEIR DATA!**

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!***



Promising Astroinformatics: Deep Learning



Example of CNN use case: Strong Lensing

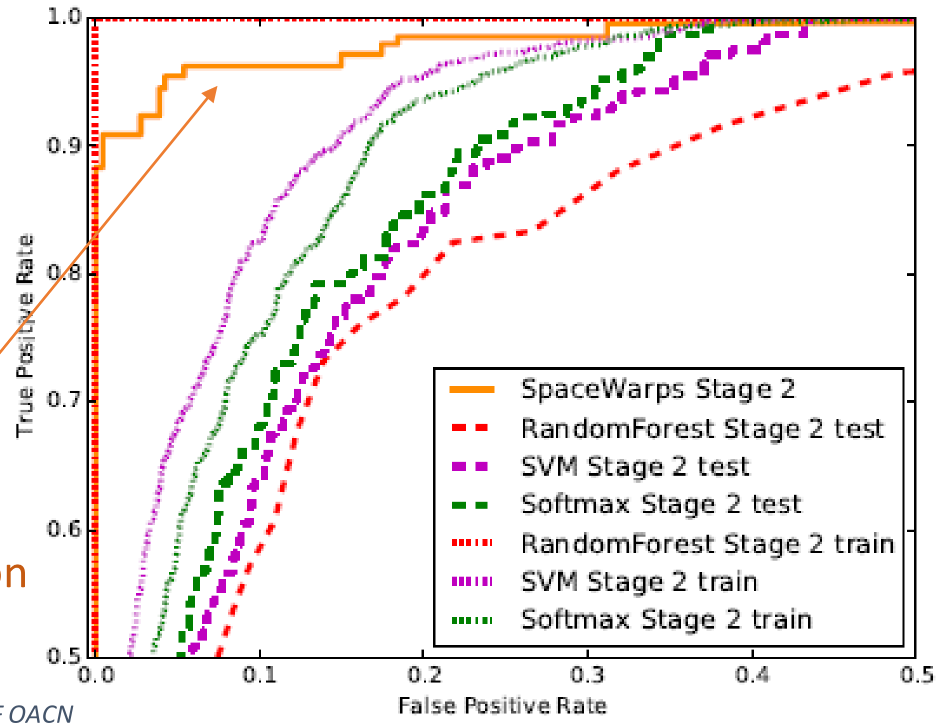
Containing simulated strong lenses

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

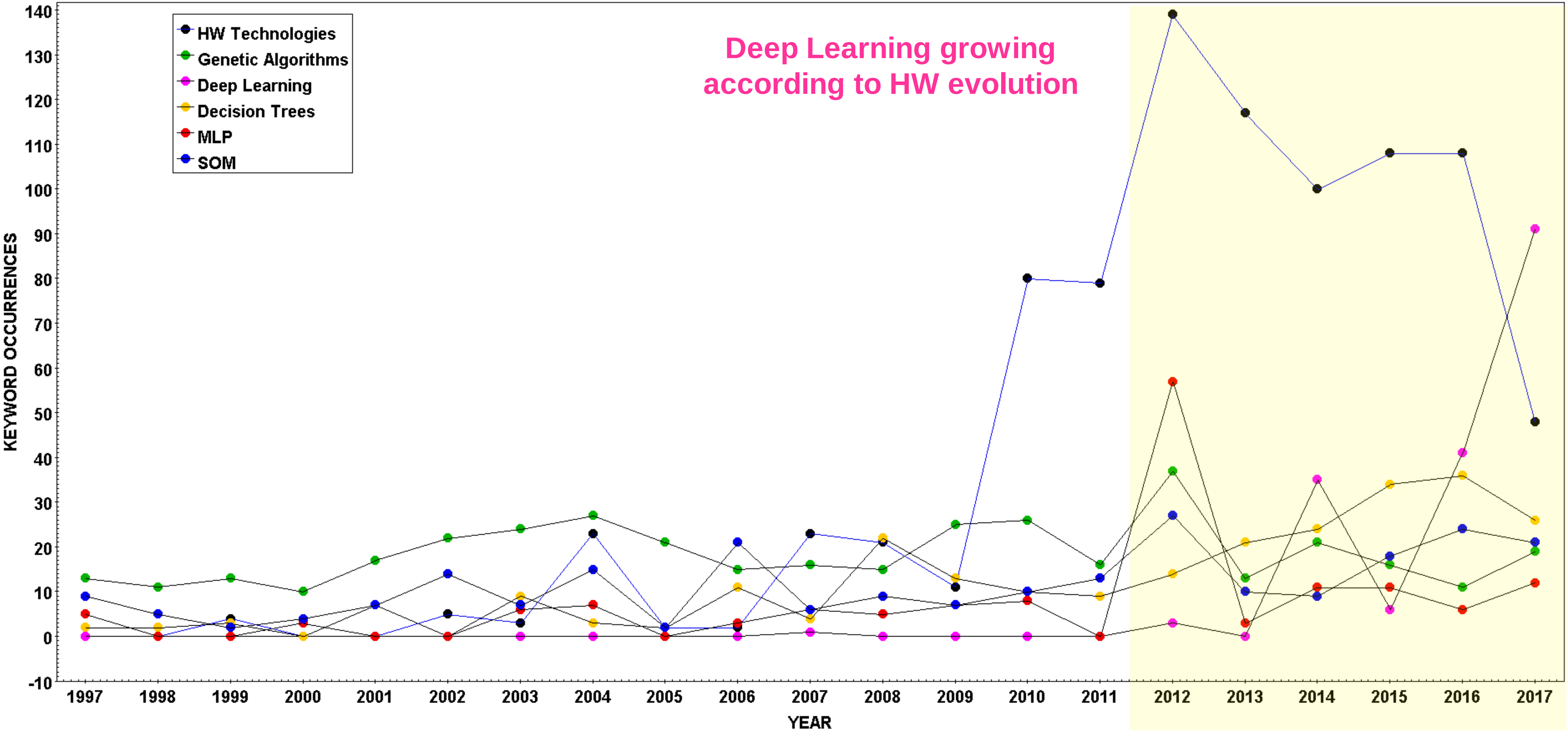
Containing no lenses

fields with cutouts on the corner

User classification by eyes



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?)

Link back to: arXiv, form interface, contact.

Download:

- PDF
- Other formats

(license)

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

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

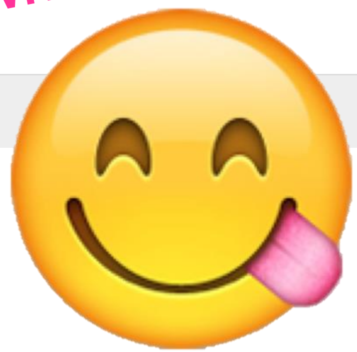
References & Citations
- NASA ADS

1 blog link (what is this?)

DBLP - CS Bibliography
listing | bibtex
Mohamed Moustafa

Bookmark (what is this?)

Deep Learning growing EVERYWHERE…!



Interested?...follow us @

Astroinformatics 2017

November 7-10

Cape Town, South Africa

...or maybe not...ehm!