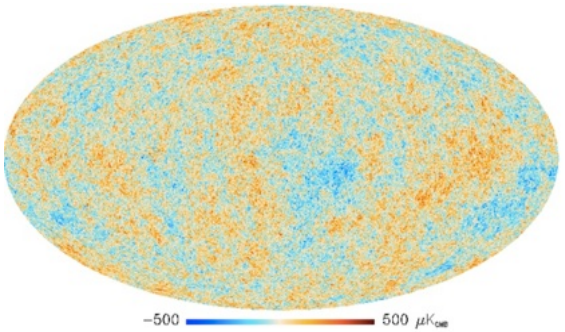
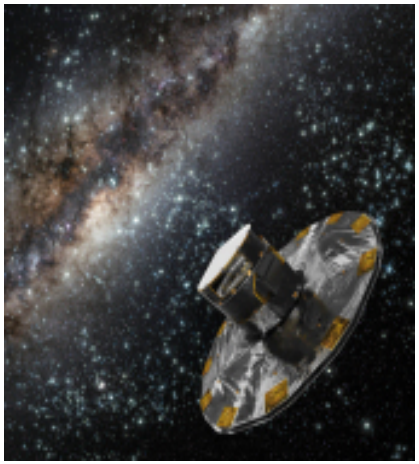
From the Big Bang to Big Data

Ofer Lahav (UCL)





Outline

- What is ‘Big Data’?
- What does it mean to computer scientists vs physicists?
- The Alan Turing Institute
- Machine learning examples from Astronomy
- The next big projects
- Challenges

What is ‘Big Data’?

- Wikipedia’s definition: “data sets that are so large or complex that TRADITIONAL data processing applications are inadequate to deal with them”.
- Clearly, this is a ‘moving target’.
- "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.” (Gartner)

The Alan Turing Institute

- Founding universities:

  Cambridge, Edinburgh, Oxford, UCL, Warwick
- Based at the British Library

- See summary of an ATI summit held in Jan 2016 at the RS on "Big Data in Physical Sciences":

  https://indico.cern.ch/event/449964/overview

Astrophysics and HEP examples

- Google: 3.5 Billion Google searches per day
- LHC: 600 Million collisions per second
  (only 100 per second are ‘interesting’)

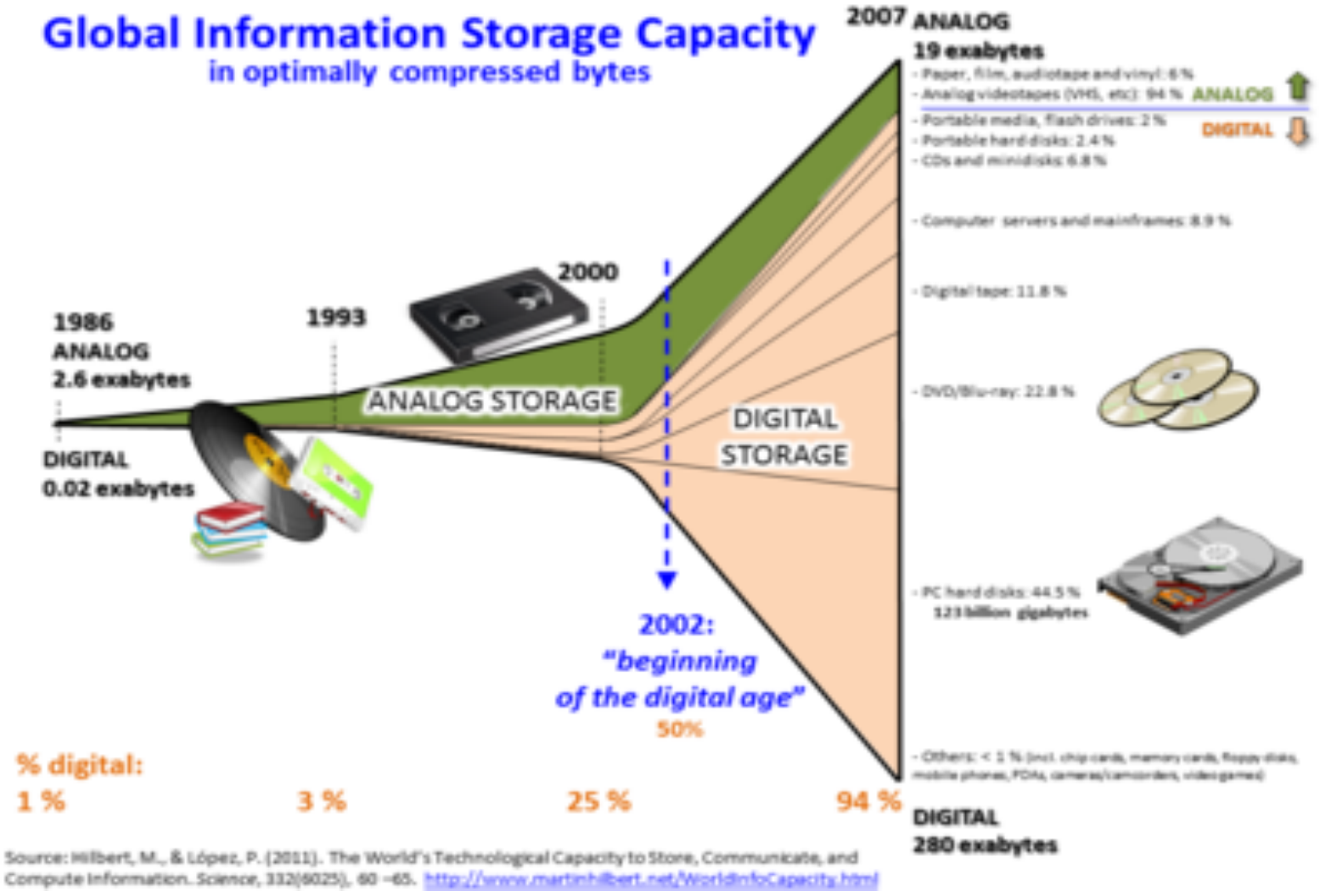
- SDSS: 200 Giga(10^9) Bytes per night
- DES: 1 Tera (10^12) Bytes per night
- LSST: 15 Tera Bytes per night
- SKA: 1 Peta (10^15) Bytes per day

Big numbers

- Exo-planets: 9-1 (+1?) +2000+…
- Gaia: 1B stars
- DES: 300M galaxies
- Euclid/LSST: 1B galaxies
- Simulations: N-body, Hydro

(many times the data)

Information storage

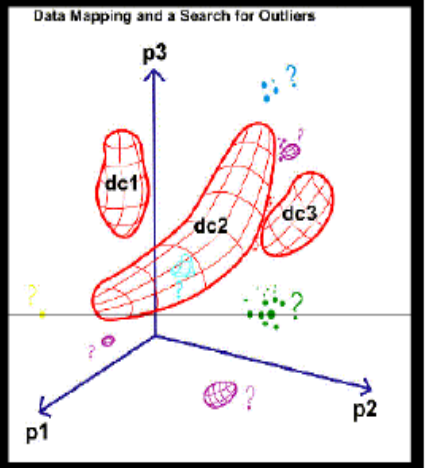


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

Big Data Science:

Scientific KDD (Knowledge Discovery from Data)

- Characterize the known (clustering, unsupervised learning)
- Assign the new (classification, supervised learning)
- Discover the unknown (outlier detection, semi-supervised learning)



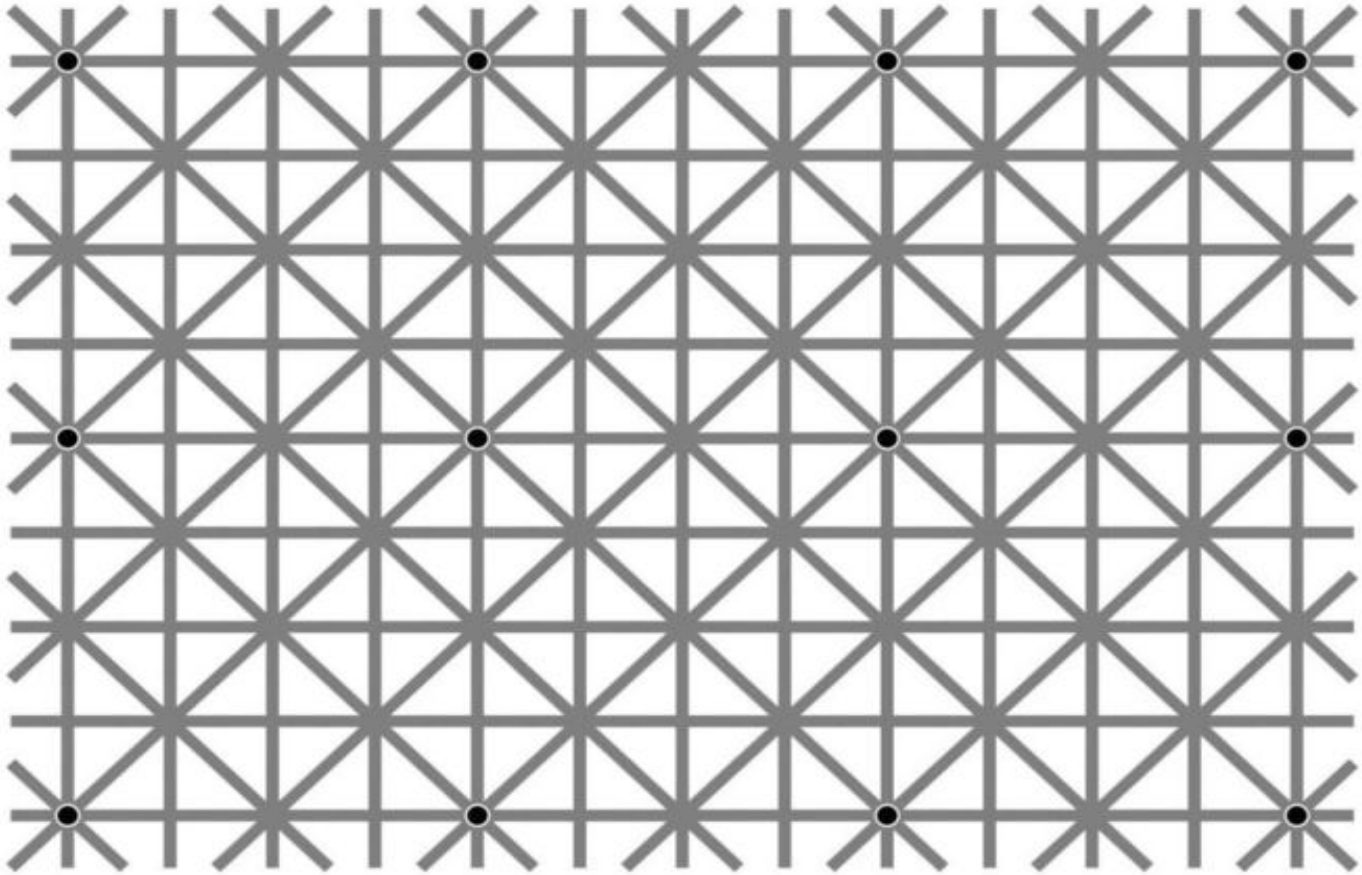
Graphic from S. G. Djorgovski

- Benefits of very large datasets:
  - best statistical analysis of "typical" events
  - automated search for "rare" events

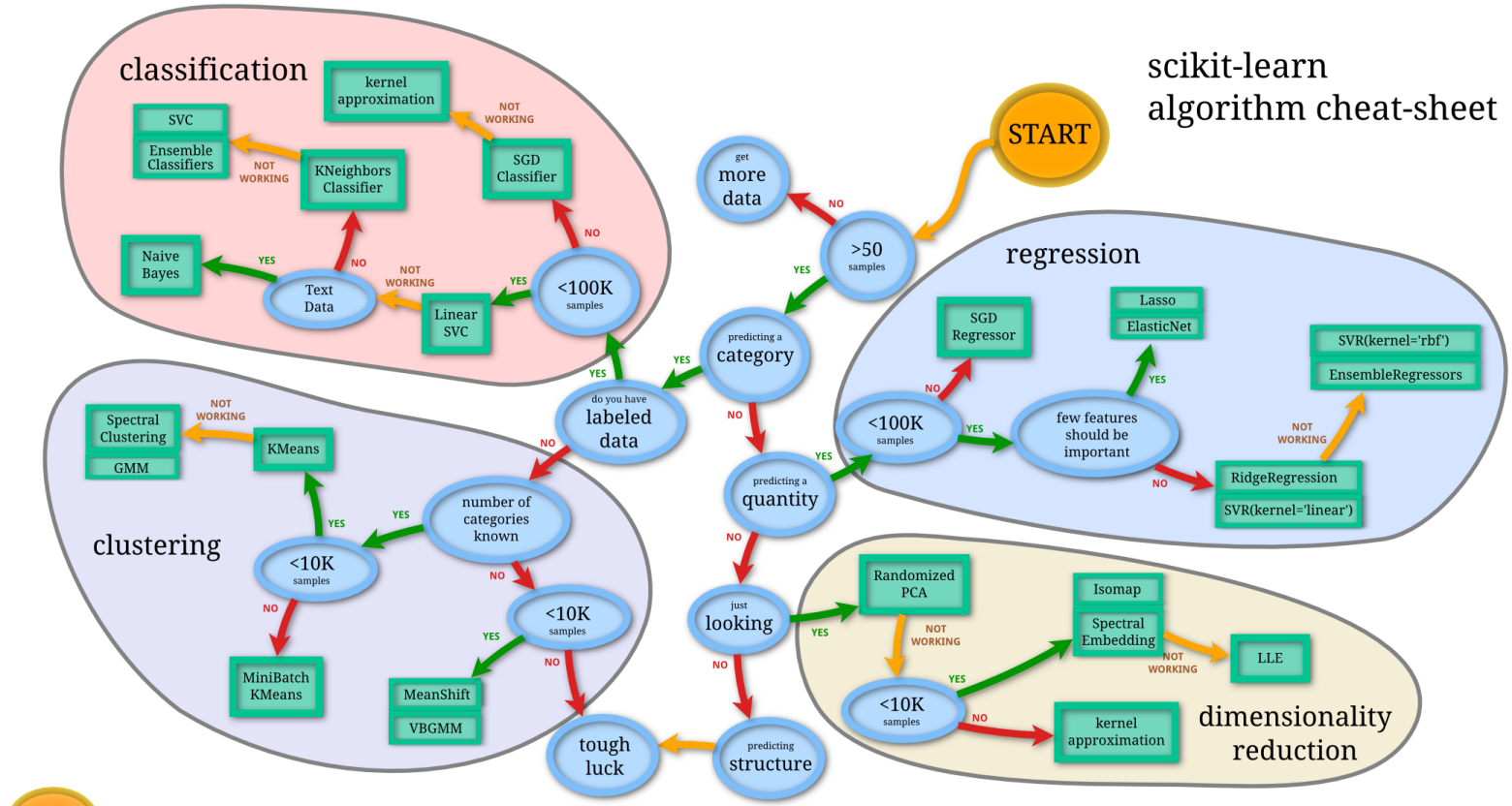
Slide from Kirk Borne

Can we trust just the human brain?

(can you see 12 black dots at once?)

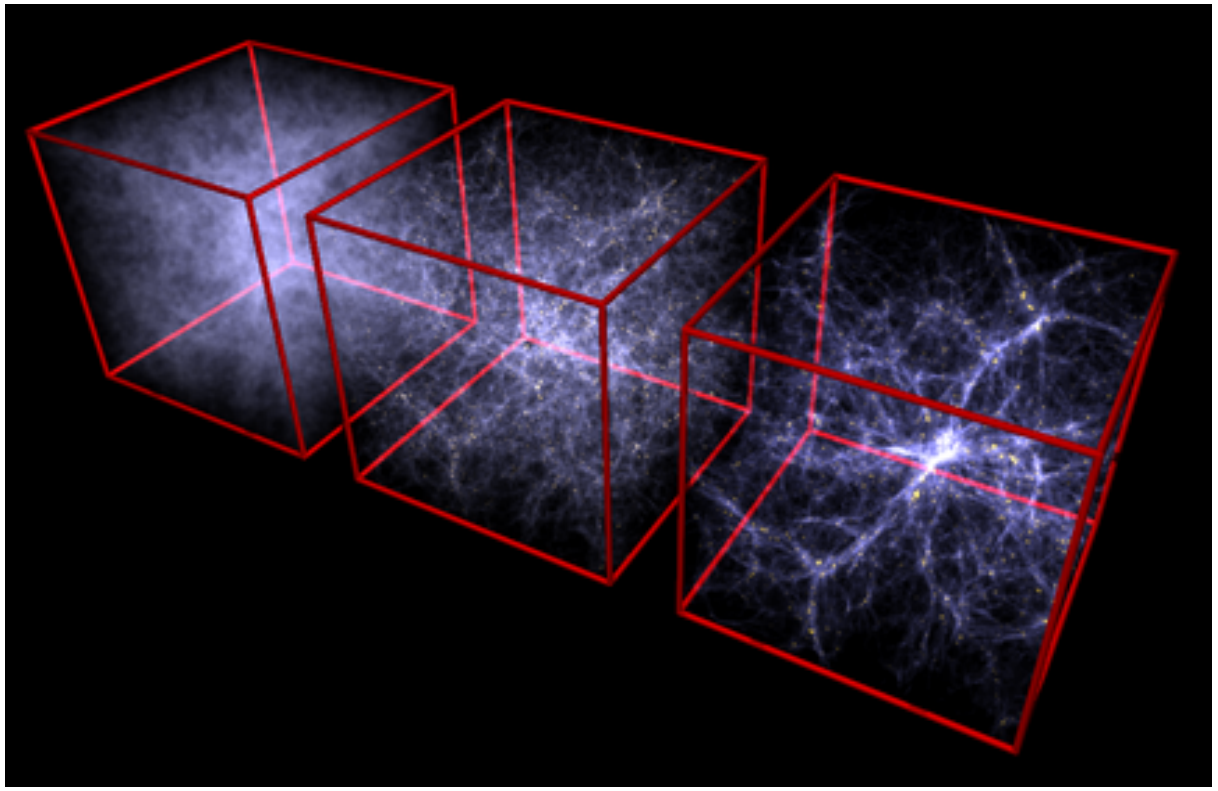


Can we trust Machine Learning?



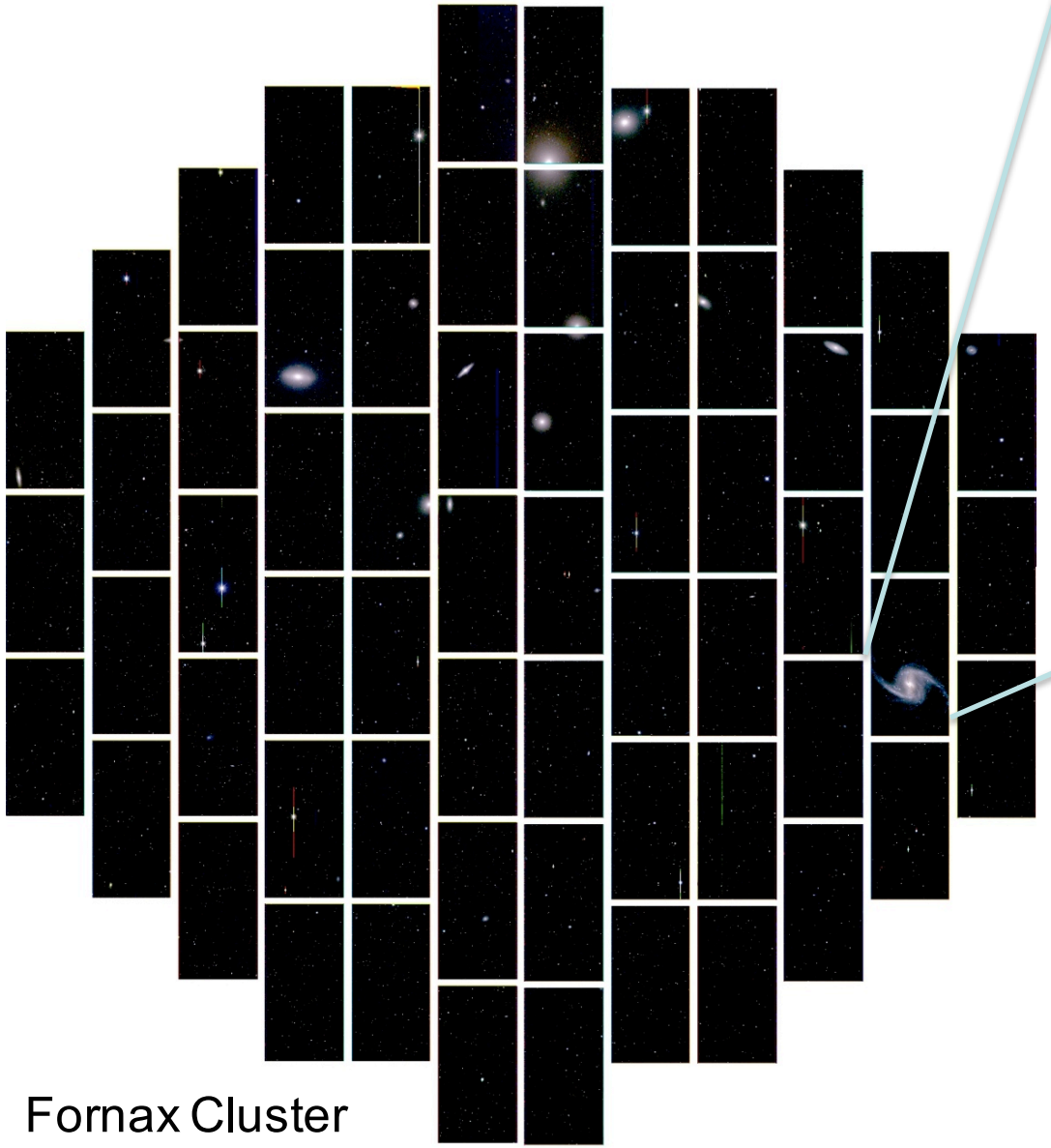
Big Data in mock universes: How to contrast with the data?

e.g. intensive Bayesian + MCMC approaches



Springel et al.

The Dark Energy Survey



Fornax Cluster

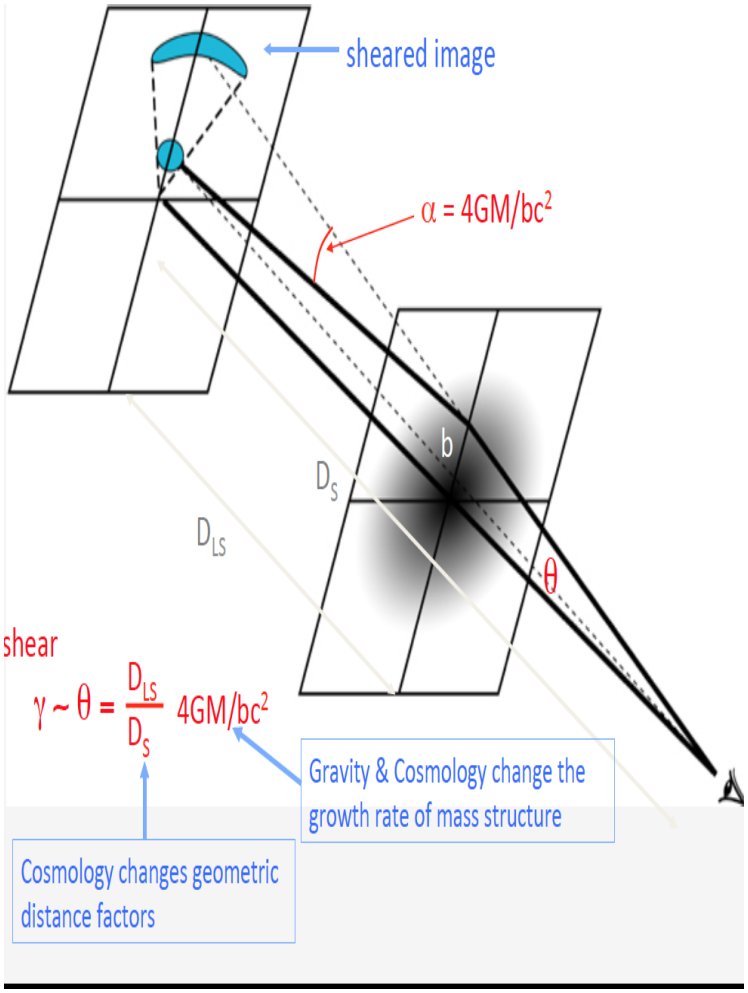
NGC 1365

Objects	As of Dec 2015	Expected from full 5yr DES
Galaxies with photo-z (> 10 sigma)	7M (SV), 100M (Y1+Y2),	300M
Galaxies with shapes	3M (SV), 80M (Y1+Y2)	200M
Galaxy clusters (lambda>5)	150K (Y1+Y2)	380K
SN Ia SLSN	1000 2 + confirmed + candidates	Thousands 15-20
New Milky Way companions	17	25
QSO's at z> 6 Lensed QSO's	1 + confirmed + candidates 2 + candidates	375 100 (i<21)
Stars (> 10 sigma)	2M (SV), 30M (Y1+Y2)	100M
Solar System: Trans Neptunian Objects Jupiter Trojans Main Belt asteroids Kuiper Belt Objects	32 in SN fields + 2 in the WF 19 300K (Y1+Y2)	50 + many more in the wide field 500-1000

DES Non—DE Overview arXiv:1601.00329

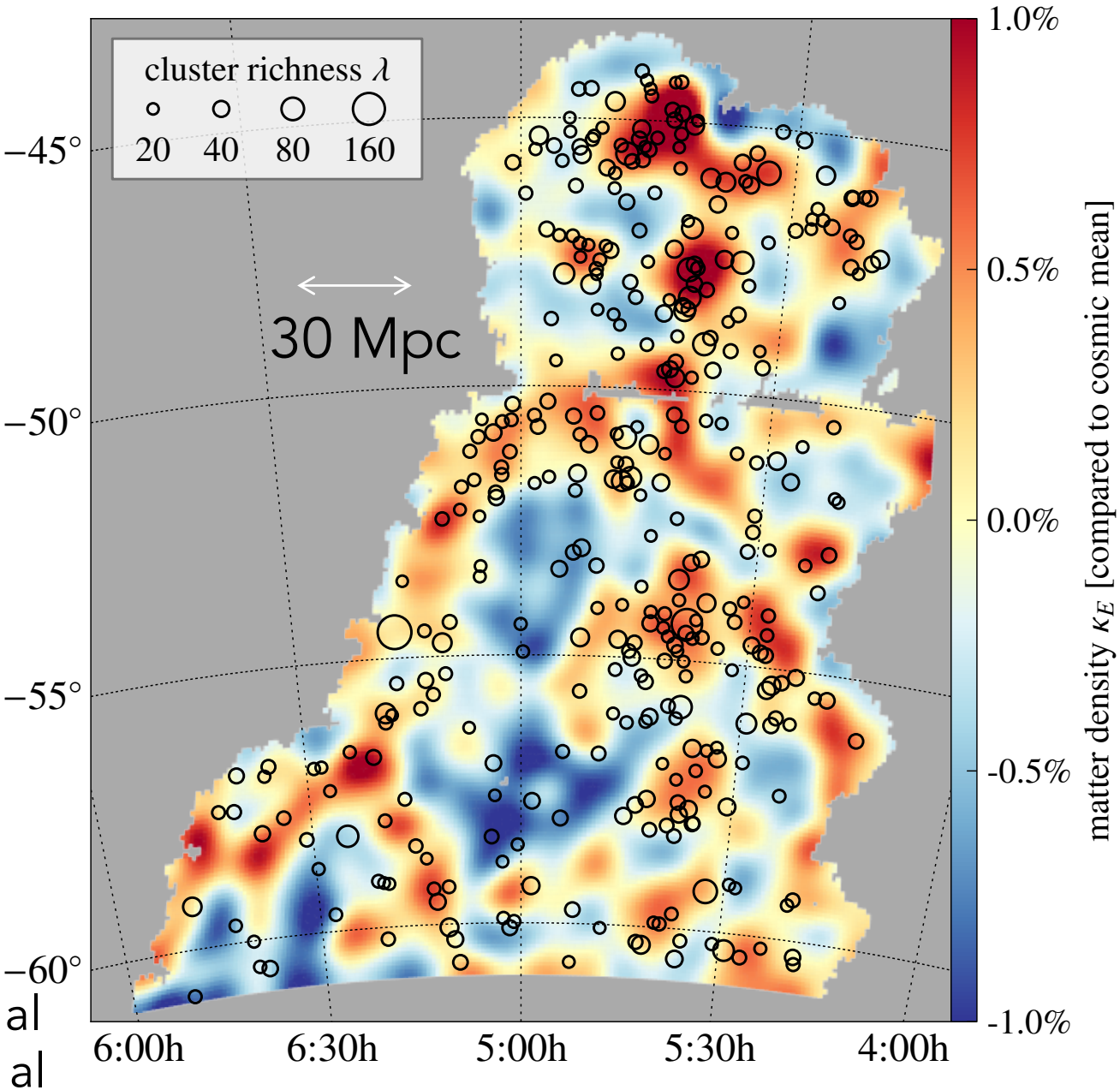
Gravitational Lensing: Weak and Strong

HST CLASH cluster MACS1206



DARK ENERGY SURVEY

DES Mass Map from Weak Lensing

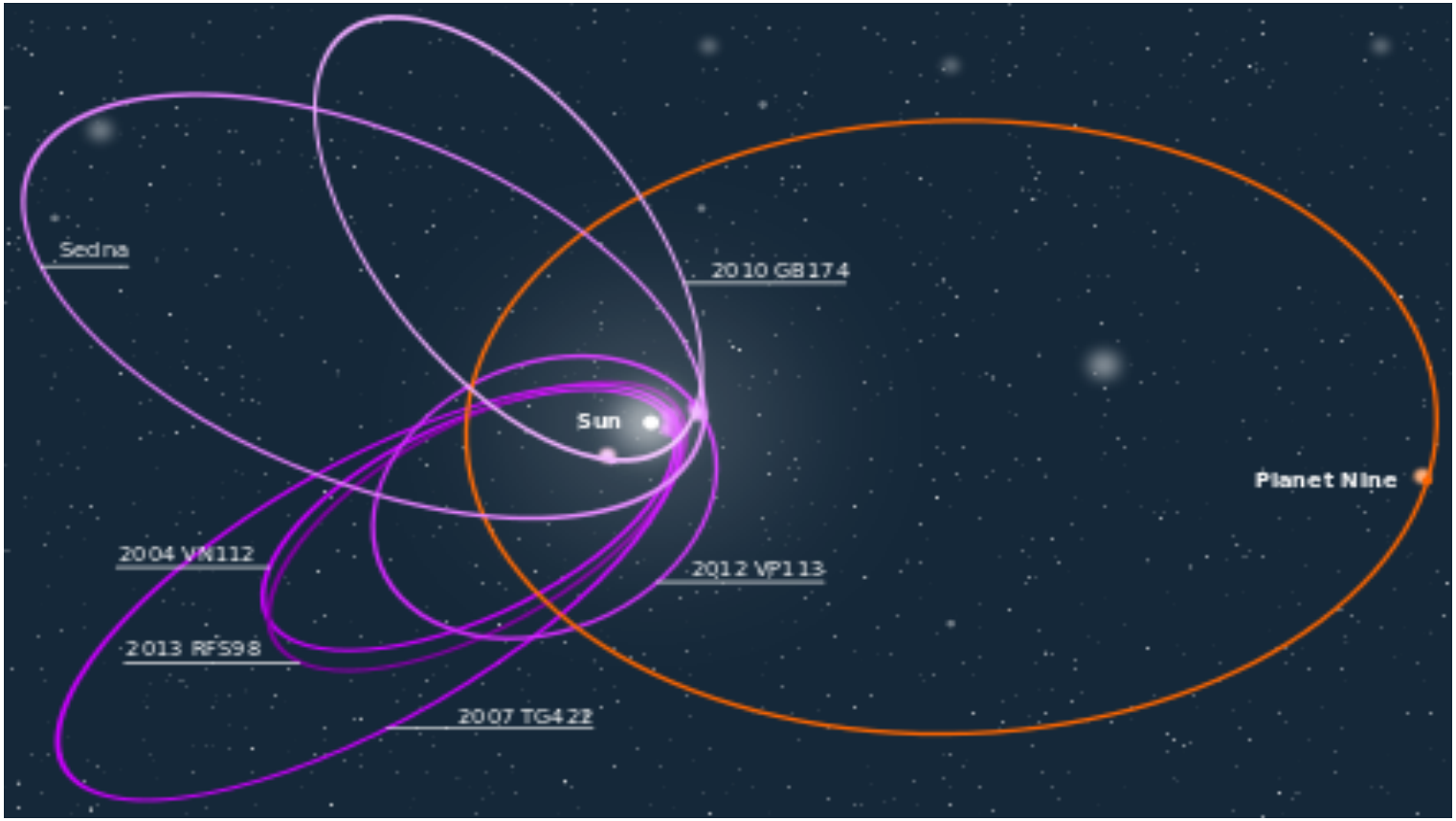


Galaxy clusters

Chang, et al
Vikram, et al

The search for Planet 9

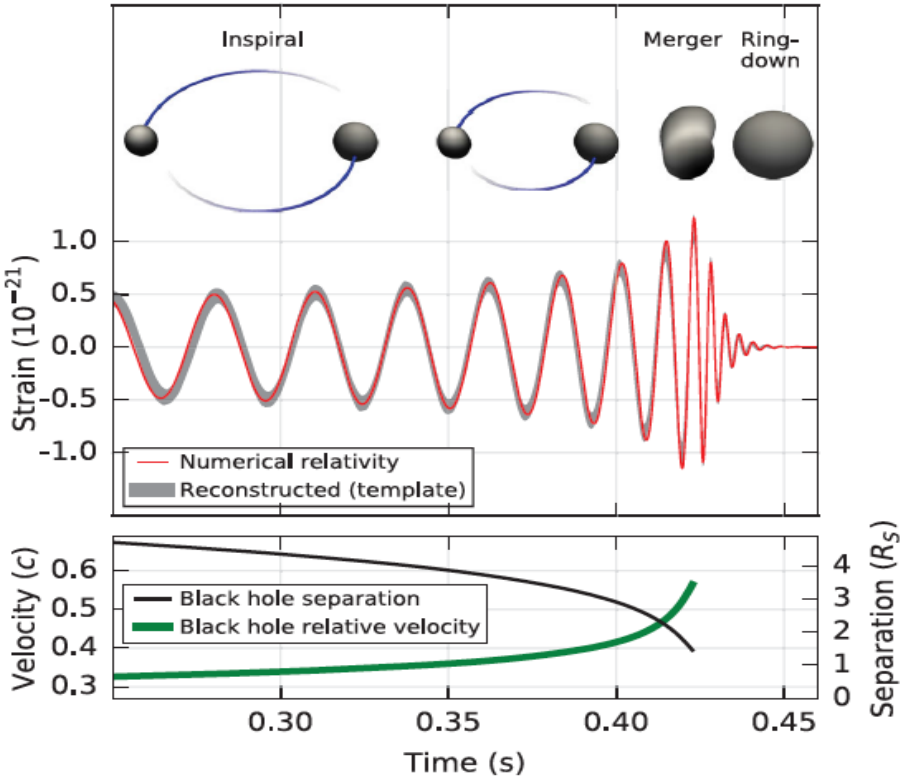
(one of the 6 minor planets discovered by DES)



David Gerdes et al, DES TMO WG

The new window of Gravitational Waves:

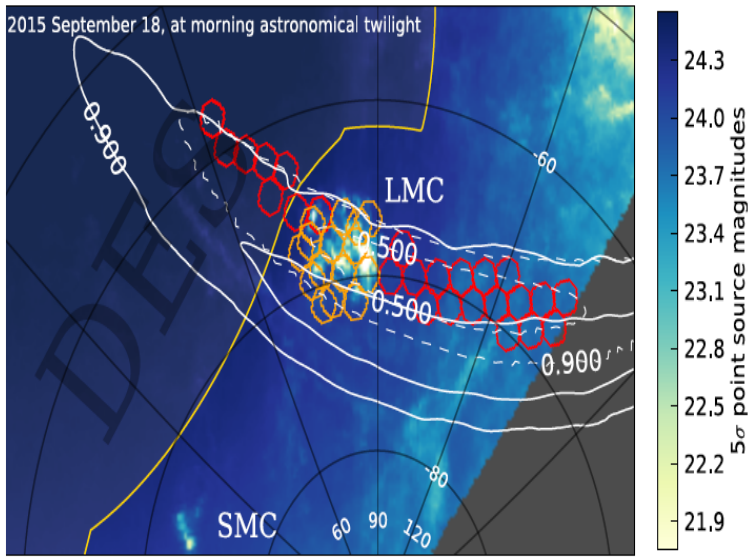
3 events detected so far; on per month expected;
and from 2019 a few per week



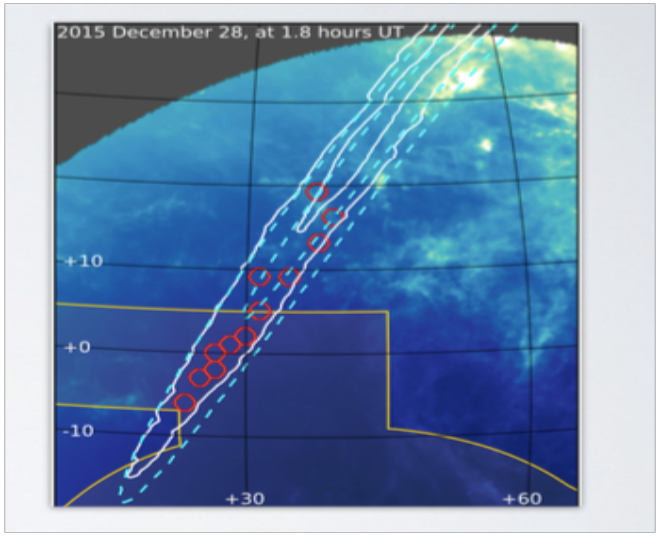
LIGO collaboration 2016

DES LIGO GW follow ups

GW150914



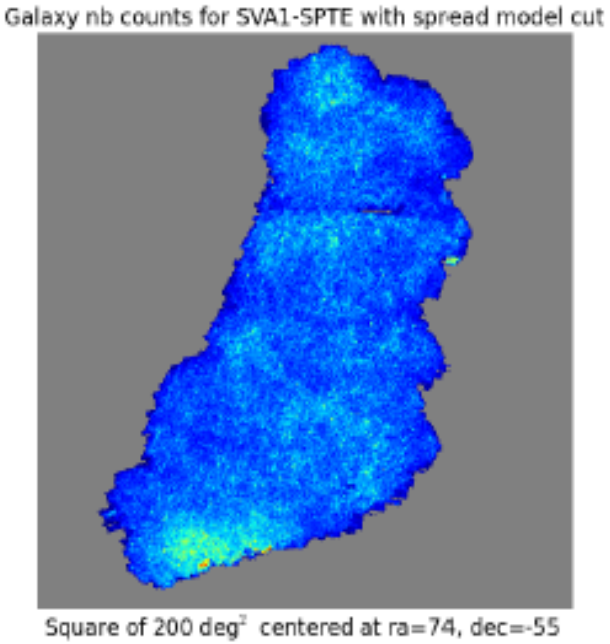
GW151226

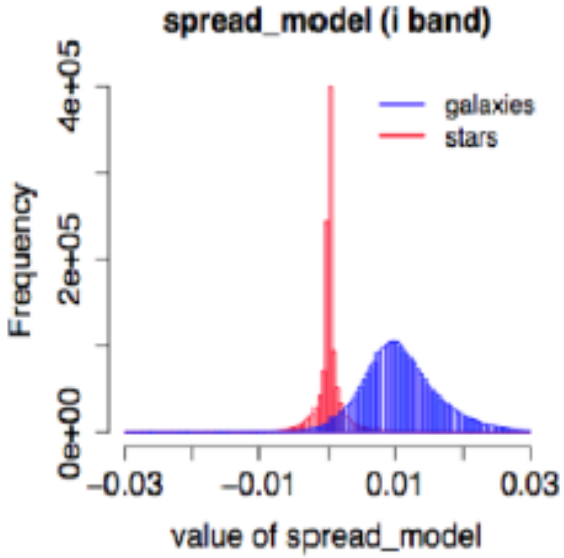


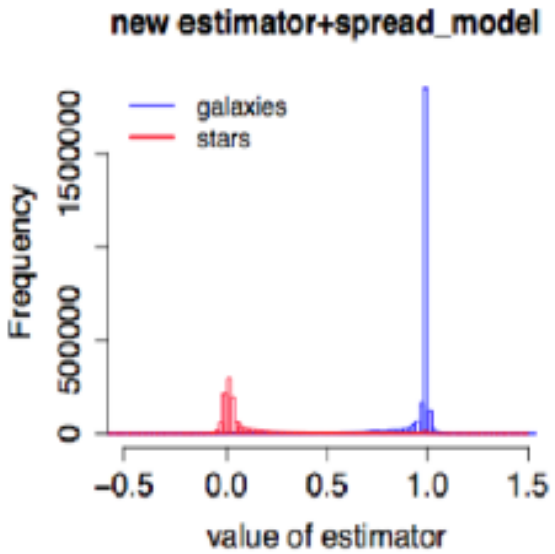
Soares-Santos et al. (2016)
Annis et al. (2016)
Abbott et al. (2016)

Cowperthwaite et al. (2016)

Star/galaxy separation in DES







Soumagnac et al (1306.5236)

GALAXY ZOO 2

- One Million galaxies classified by 100,000 people!

Is the galaxy simply smooth and rounded, with no sign of a disk?

Smooth

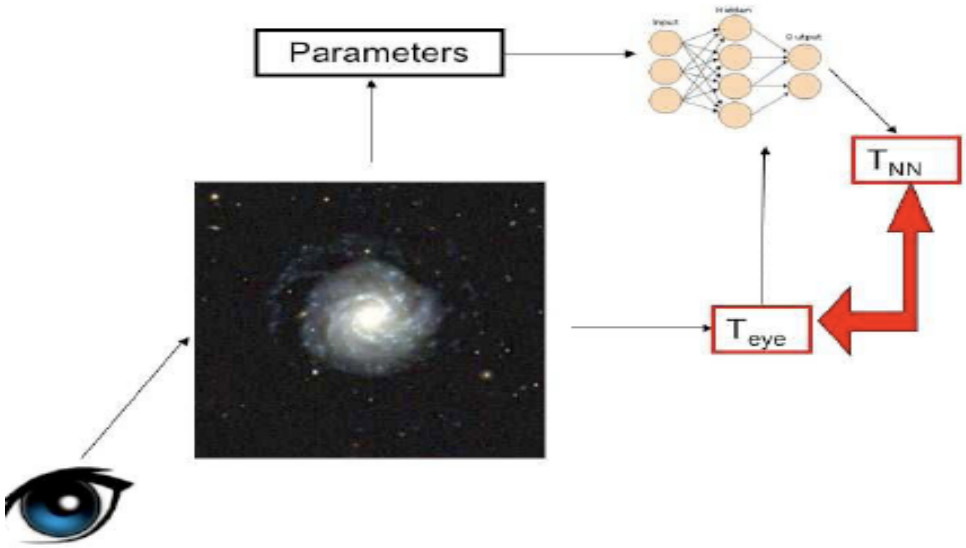
Features or disk

Star or artifact

Need help?

Lintott et al.

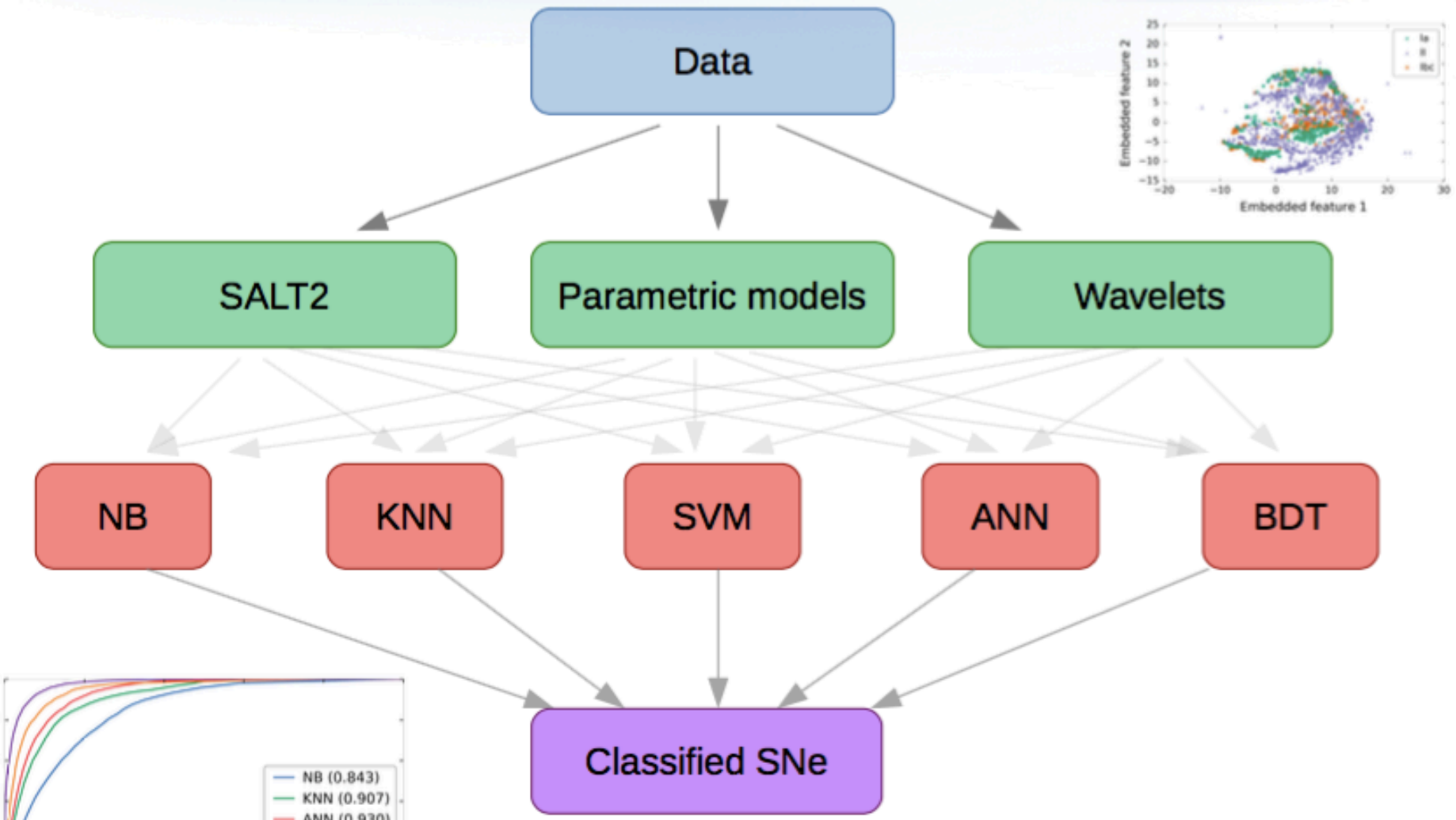
Galaxy zoo and machine learning



		GALAXY ZOO		
		Elliptical	Spiral	Star/Other
A	ELLIPTICAL	91%	0.08%	0.5%
N	SPIRAL	0.1%	93%	0.2%
N	STAR/OTHER	0.3%	0.3%	96%

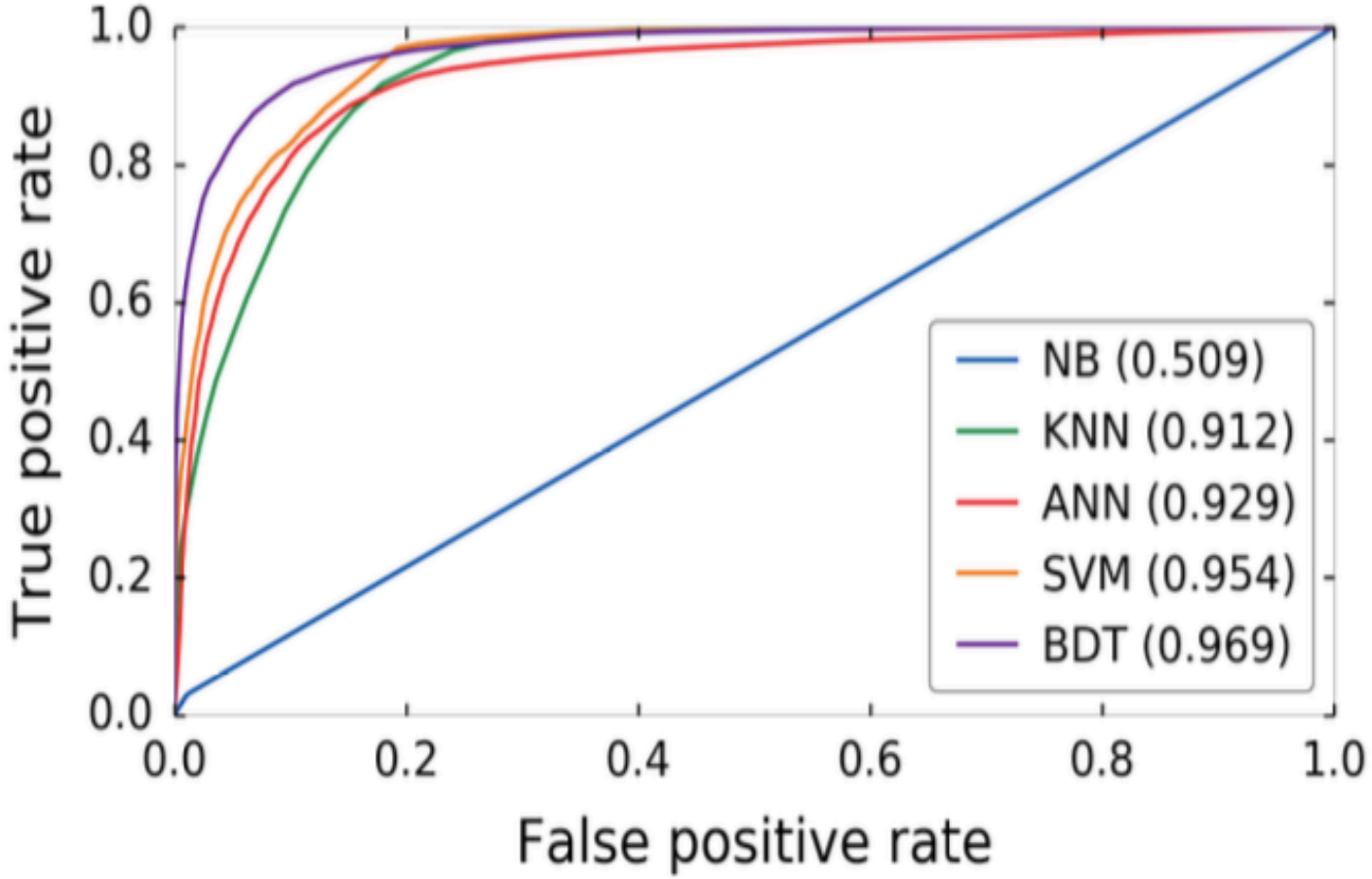
Banerji, OL et al. (0908.2033)

Photometric Classification of Supernovae



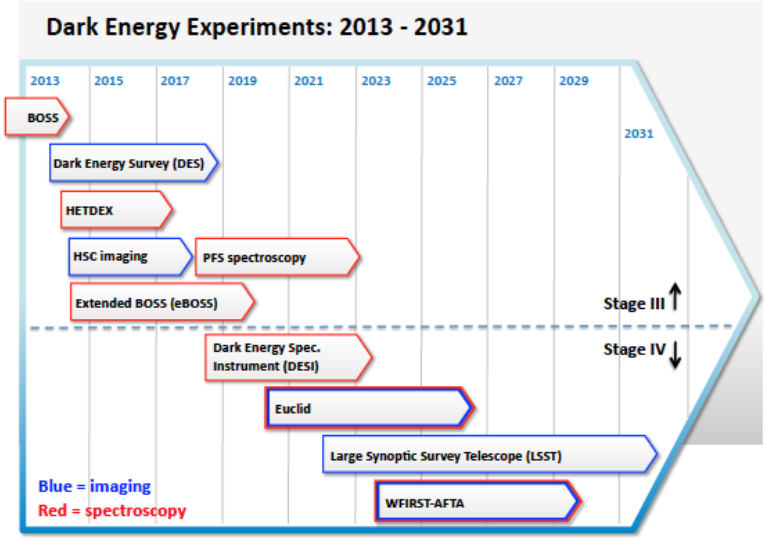
Lochner et al.
arXiv: 1603.00882

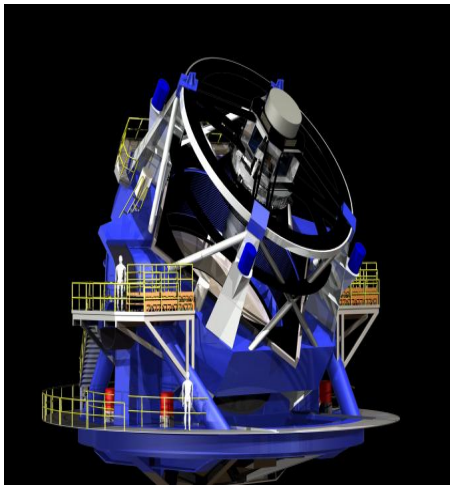
Feature extraction with Wavelet

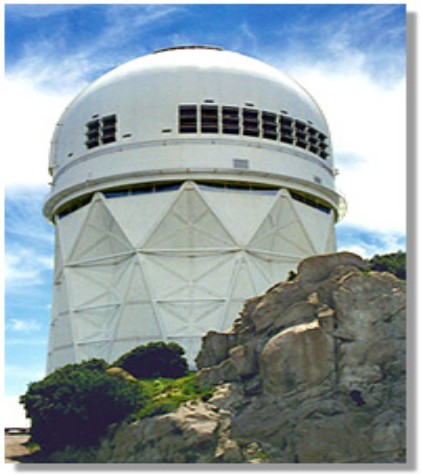


Lochner et al. (2016)

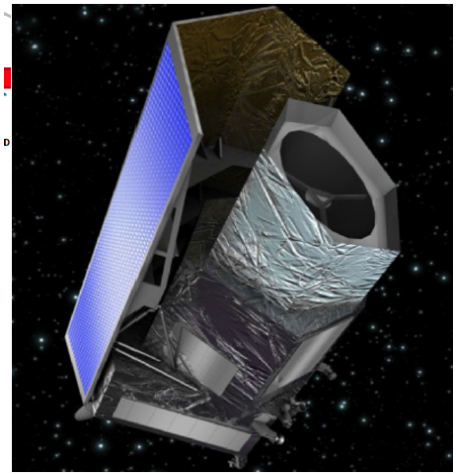
The era of DESI, Euclid, LSST,…



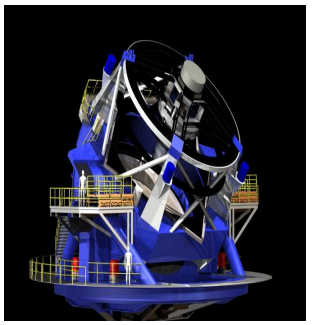




Mayall 4-Meter Telescope



DES vs. LSST



	DES	LSST
Telescope	4 meters	8 meters
Field-of-View	3 sq-deg	9.6 sq-deg
Survey Area	5,000 sq-deg	18,000 sq-deg
Camera	570 megapixels	3,200 megapixels
Cadence	2 / yr / band	~100 / yr / band
Raw Data	1 TB / night	15 TB/ night
Reduced	2.5 PB	Few 100 PB
Catalog	6×10^8 objects	2×10^{10} objects

Euclid Forecast

What	Euclid	Before Euclid
Galaxies at 1<z<3 with good mass estimates	~2×10^8	~5×10^6
Massive galaxies (1<z<3) w/ spectra	~few x 10^3	~few tens
Hα emitters/metal abundance in z~2-3	~$4\times10^7/10^4$	~10^4/~10^2?
Galaxies in massive clusters at z>1	~2×10^4	~10^3?
Type 2 AGN (0.7<z<2)	~10^4	<10^3
Dwarf galaxies	~10^5	
T_{eff} ~400K Y dwarfs	~few 10^2	<10
Strongly lensed galaxy-scale lenses	~300,000	~10-100
z > 8 QSOs	~30	None

Big Data, Big collaborations:
How many collaborators can one have?

2dF: 30
DES: 500
Planck:400
Euclid: 1200
LHC: 4000

Within the primates there is a general relationship between the size of the brain and the size of the social group. Scaling it to humans gives 150 people. This is the cognitive limit to the number of people with whom one can have stable interaction (Dunbar 1992).

Some points for discussion (based on panel discussion of ATI-PS meeting in Jan 2016)

- What can we learn from other fields?
- What can we learn from other Big Data initiatives elsewhere in the world?
- How to connect applications to computer science foundations (networks etc.)
- How to move from ‘deterministic’ algorithms to computation in the presence of uncertainties.
- Would Big Data lead to more (or less) ‘deep thinking' in Physics problems?

Summary

- Science is going ‘industrial revolution’
- Both spatial and time domains
- Great training of PhDs, beyond academia.
- Will Big Data produce bigger knowledge?

  (it depends it part on Nature…)