Mining Archives: needs for Machine Learning

S. Cavuoti & G. Riccio

INAF - Astronomical Observatory of Capodimonte

Users do what users do best

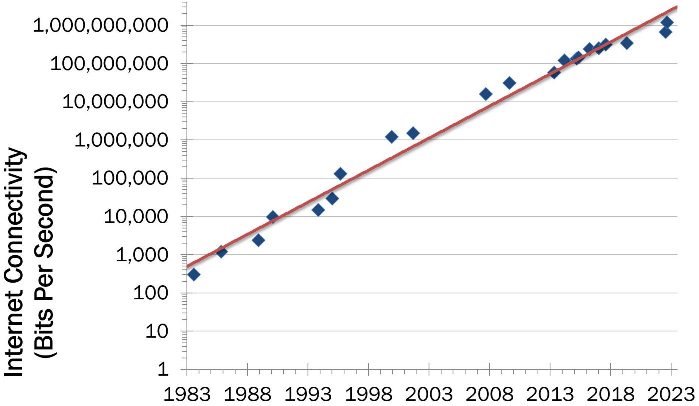
PEANUTS
STUPID ARCHIVE!
ARE YOU COMPLAINING AGAIN? DO YOU REALIZE THAT YOU SPEND ALL YOUR TIME COMPLAINING?
1-7
WHY SHOULDN'T I COMPLAIN?
IT'S THE ONLY THING I'M REALLY GOOD AT!
SCHULZ

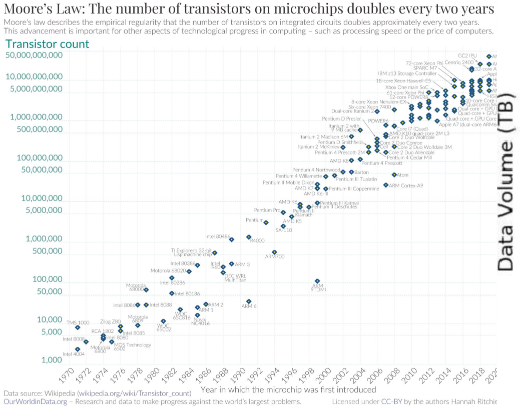
Simple Premise

There are archives already implementing some (or even all) solutions to the problems that I will address, there is no need to jump up and say: “***I’m not guilty***”

Data Tsunami

We are all facing Data Tsunami

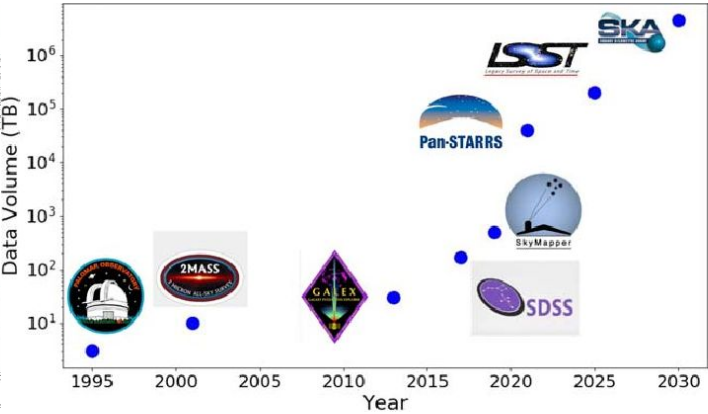



Moore’s Law: The number of transistors on microchips doubles every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

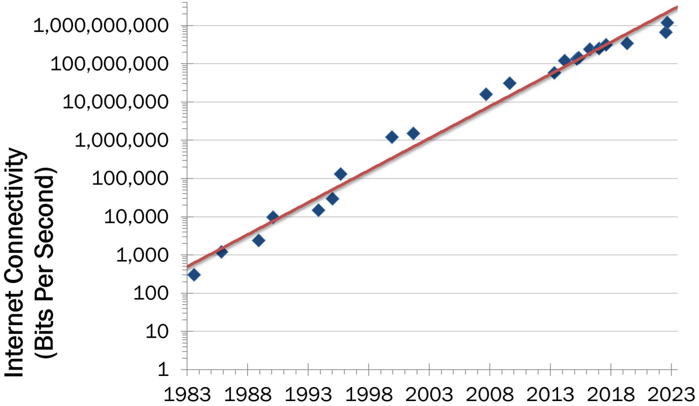
Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

OurWorldinData.org – Research and data to make progress against the world’s largest problems. Licensed under CC-BY by the authors Hannah Ritchie



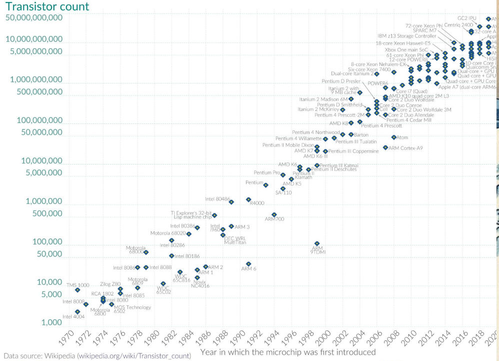
Data Tsunami

We are all facing Data Tsunami



Moore's Law: The number of transistors on microchips doubles every two years

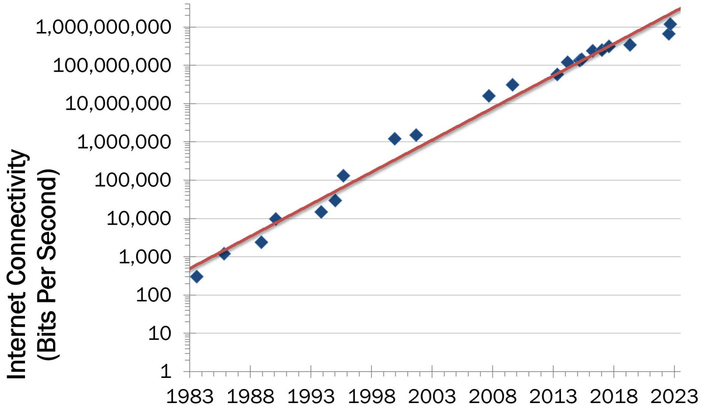
Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

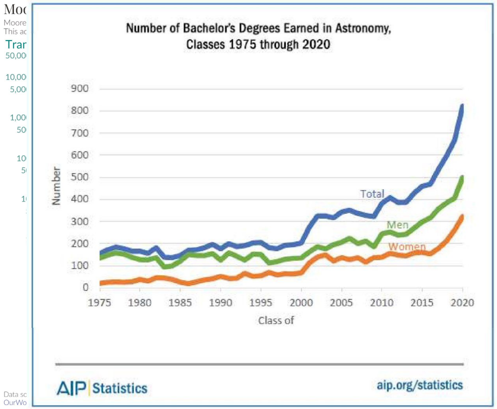


Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)
OurWorldinData.org – Research and data to make progress against the world's largest problems. Licensed under CC-BY by the authors Hannah Ritchie

Data Tsunami

Astronomers are increasing too, even though not at the same speed





More users requiring data, that are by themself increasing faster than the connection.

Data Tsunami

Transferring large datasets from astronomical archives to local machines is inefficient, consuming excessive bandwidth and storage.

One possible solution could be instead of moving data to the code, move the code to the data by providing computational resources directly within the archive infrastructure.

- This Reduced data transfer → Minimizes bandwidth usage and speeds up analysis.

but

- Increased complexity → Requires sophisticated infrastructure and maintenance.
- Security restrictions → Limited user access and constrained software environments.
- Reduced flexibility → Users may not have full control over the computational setup.

Some facilities already offer some computational environments to enable in-place data analysis.

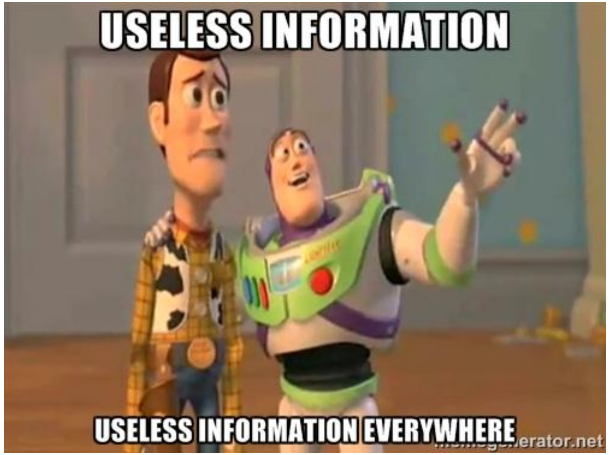
This could be a viable solution but people may want to deal with more than one survey at the time, so the problem is still out there

Data Tsunami Mitigation

Not all the users are interested to the same data

Just to make the easiest example:

if you want to study galaxies you don’t need to retrieve data of Stars



I don't want the universe all at once

For a typical Machine Learning experiment (actually for any kind of science) what you need is of course a **subset of data points** and a **subset of the information** available for that point.

I don't need (at least usually) to download everything is available in the archive.

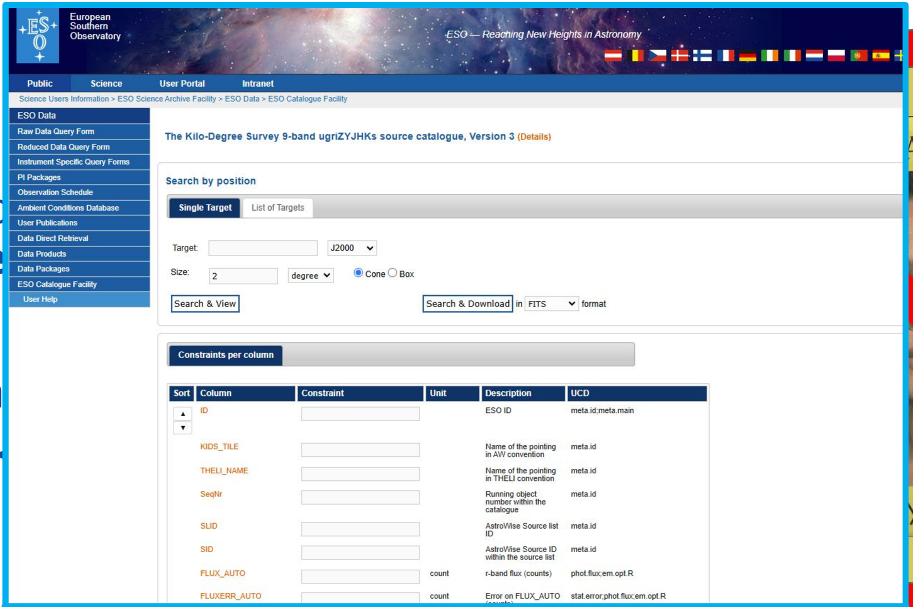
I need services that allow me to select only the data I need.

I don’t want the universe all at once

For a typical Machine Learn
any kind of science) what yo
of data points and a **subse**
available for that point.

**I don’t need (at least usua
is available in the archive.**

**I need services that allow
need.**



European Southern Observatory

ESO — Reaching New Heights in Astronomy

Public | Science | User Portal | Intranet

Science Users Information > ESO Science Archive Facility > ESO Data > ESO Catalogue Facility

- ESO Data
- Raw Data Query Form
- Reduced Data Query Form
- Instrument Specific Query Forms
- PI Packages
- Observation Schedule
- Ambient Conditions Database
- User Publications
- Data Direct Retrieval
- Data Products
- Data Packages
- ESO Catalogue Facility
- User Help

The Kilo-Degree Survey 9-band ugriZYJHKs source catalogue, Version 3 (Details)

Search by position

Single Target | List of Targets

Target: J2000

Size: 2 degree ● Cone ○ Box

Search & View

Search & Download in FITS format

Constraints per column

Sort	Column	Constraint	Unit	Description	UCD
▲ ▼	ID			ESO ID	meta.id;meta.main
	KIDS_TILE			Name of the pointing in AW convention	meta.id
	THELI_NAME			Name of the pointing in THELI convention	meta.id
	SeqNr			Running object number within the catalogue	meta.id
	SLID			AstroWise Source list ID	meta.id
	SID			AstroWise Source ID within the source list	meta.id
	FLUX_AUTO		count	r-band flux (counts)	phot.flux;em.opt.R
	FLUXERR_AUTO		count	Error on FLUX_AUTO	stat.error;phot.flux;em.opt.R

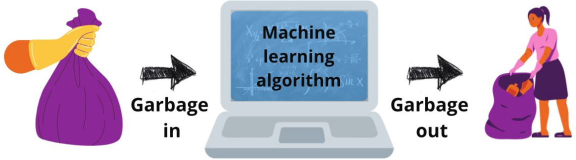
**ESO portal, just as an example:
I have to download every single column and I have very limited possibility to perform a complex query
Even if a lot of survey are on the portal they do not “talk” one with the others**

I don’t want the universe all at once

It should be obvious but:

Data quality flags are really crucial when dealing with machine learning

It is important that they are available and well documented



I don’t want the universe all at once

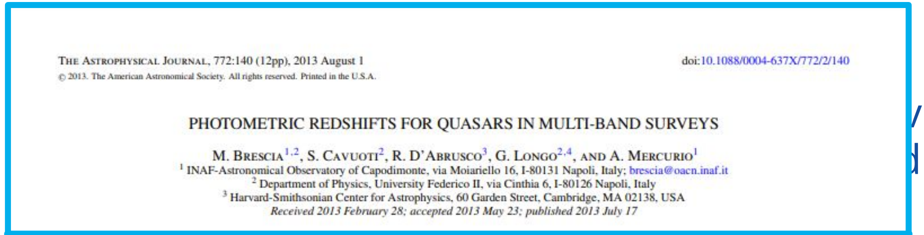
When comparing objects across multiple surveys, users need to cross-match them based on celestial coordinates.

This can happen on different levels:

- I have a catalogue of sources and I want the matching entries (**ideal solution would be to run the cross-match on the server rather than download everything and perform the match on my side**)
- I want sources with a given counterpart from a different survey, since this task could be often usual **add a pre calculated ID referencing to some of the most important surveys** would be useful sparing a lot of computation

CLOUD MONITORING EMERGENCY REPORT ...
WE MAY HAVE SERIOUS PROBLEMS WITH DATA-MERGING!
© CLOUDTWEAKS.COM
CHRISTIAN MIRRA 2016

I don't want the universe all at once



The Astrophysical Journal, 772:140 (12pp), 2013 August 1 doi:10.1088/0004-637X/772/2/140

© 2013. The American Astronomical Society. All rights reserved. Printed in the U.S.A.

PHOTOMETRIC REDSHIFTS FOR QUASARS IN MULTI-BAND SURVEYS

M. Brescia[1,2], S. Cavuoti[2], R. D'Abrusco[3], G. Longo[2,4], and A. Mercurio[1]

[1] INAF-Astronomical Observatory of Capodimonte, via Moiariello 16, I-80131 Napoli, Italy; brescia@oacn.inaf.it

[2] Department of Physics, University Federico II, via Cinthia 6, I-80126 Napoli, Italy

[3] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA

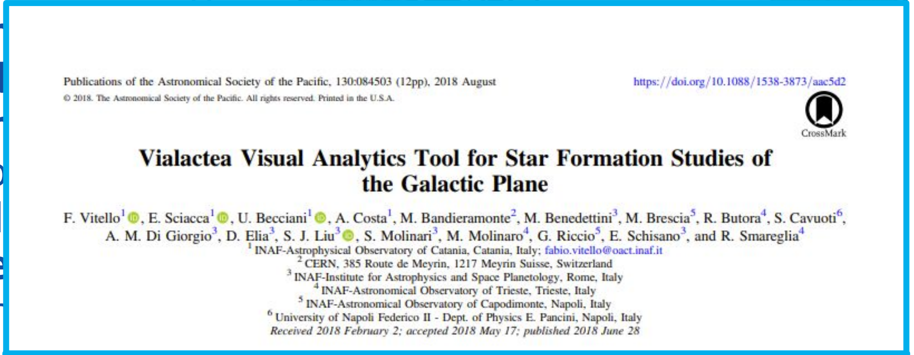
Received 2013 February 28; accepted 2013 May 23; published 2013 July 17

4 surveys involved

veys, users need
dinates.

CLOUD MONITORING EMERGENCY REPORT ...
WE MAY HAVE SERIOUS PROBLEMS WITH DATA-MERGING!

- I have a catalogue of sources and I want the matching entries (**ideal solution would be to run th cross-match on the server rather than d everything and perform the match on m**
- I want sources with a given counterpart fro survey, since this task could be often usual **calculated ID referencing to some of the important surveys** would be useful sparin computation



Publications of the Astronomical Society of the Pacific, 130:084503 (12pp), 2018 August https://doi.org/10.1088/1538-3873/aac5d2

© 2018. The Astronomical Society of the Pacific. All rights reserved. Printed in the U.S.A.

CrossMark

Vialactea Visual Analytics Tool for Star Formation Studies of the Galactic Plane

F. Vitello[1], E. Sciacca[1], U. Becciani[1], A. Costa[1], M. Bandieramonte[2], M. Benedettini[3], M. Brescia[5], R. Butora[4], S. Cavuoti[6], A. M. Di Giorgio[3], D. Elia[3], S. J. Liu[3], S. Molinari[3], M. Molinaro[4], G. Riccio[5], E. Schisano[3], and R. Smareglia[4]

[1] INAF-Astrophysical Observatory of Catania, Catania, Italy; fabio.vitello@oact.inaf.it

[2] CERN, 385 Route de Meyrin, 1217 Meyrin Suisse, Switzerland

[3] INAF-Institute for Astrophysics and Space Planetology, Rome, Italy

[4] INAF-Astronomical Observatory of Trieste, Trieste, Italy

[5] INAF-Astronomical Observatory of Capodimonte, Napoli, Italy

[6] University of Napoli Federico II - Dept. of Physics E. Pancini, Napoli, Italy

Received 2018 February 2; accepted 2018 May 17; published 2018 June 28

7 surveys involved

I don’t want the universe all at once

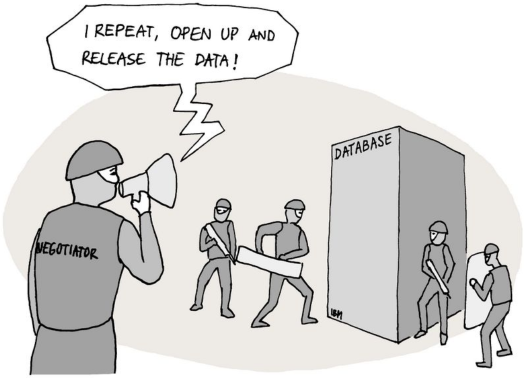
In order to train a deep learning model **we need the images.**

Not all surveys provide images and even in that case for most of them it is not straightforward to retrieve them

We are often interested in small cutout surrounding the centroid of the object but we had to download entire plates and crop them locally, leading to significant inefficiencies. Users have access to raw, full-frame data, which can be useful for context and custom processing.

Bandwidth-heavy: Downloading entire plates is costly in terms of storage and data transfer. Although it would be computationally expensive for the server this effort in terms of computing power would be balanced from the limited usage of bandwidth.

Implementing a server-side cutout service that allows users to request only the relevant portion of an image directly from the archive. Some archives like SDSS and Pan-STARRS offer this, but it is not universal and they are implemented in a very different way and they could not be exactly what a user need (going back to the initial problem).



Dataedo /cartoon

I don’t want the universe all at once

most of them it is not

ntro

, lea

inefficiencies: Users have access to raw, full-frame data, which

context and custom processing.

Bandwidth-heavy: Downloading entire plates is costly in terms c

the

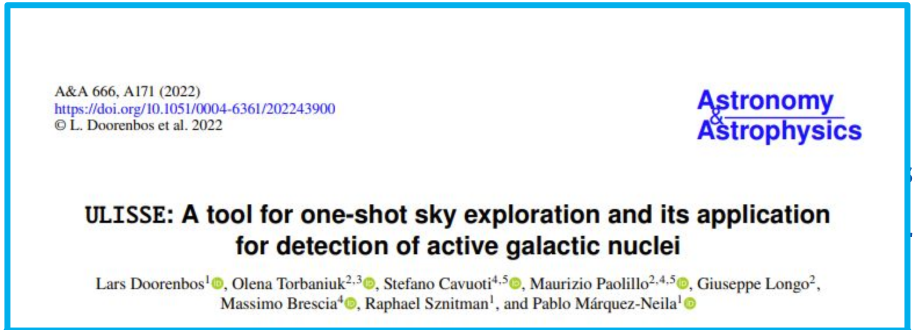
ed us

ers to request only the

e archives like SDSS and

implemented in a very

need (going back to the



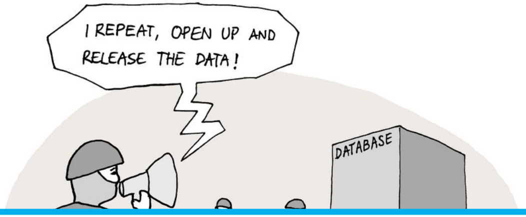
A&A 666, A171 (2022)
https://doi.org/10.1051/0004-6361/202243900
© L. Doorenbos et al. 2022

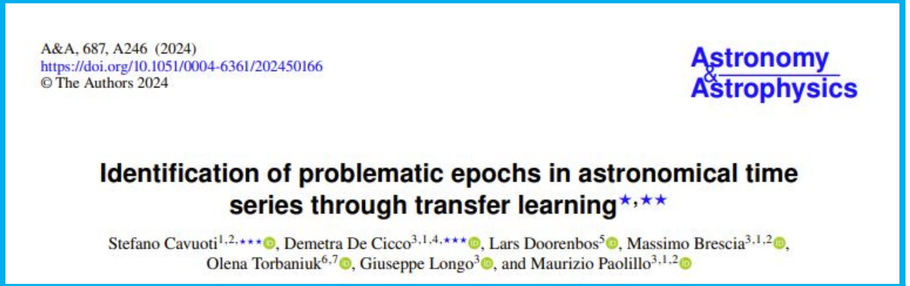
Astronomy & Astrophysics

ULISSE: A tool for one-shot sky exploration and its application for detection of active galactic nuclei

Lars Doorenbos[1], Olena Torbaniuk[2,3], Stefano Cavuoti[4,5], Maurizio Paolillo[2,4,5], Giuseppe Longo[2], Massimo Brescia[4], Raphael Sznitman[1], and Pablo Márquez-Neila[1]

~100k 73x73px 3-band thumbnails





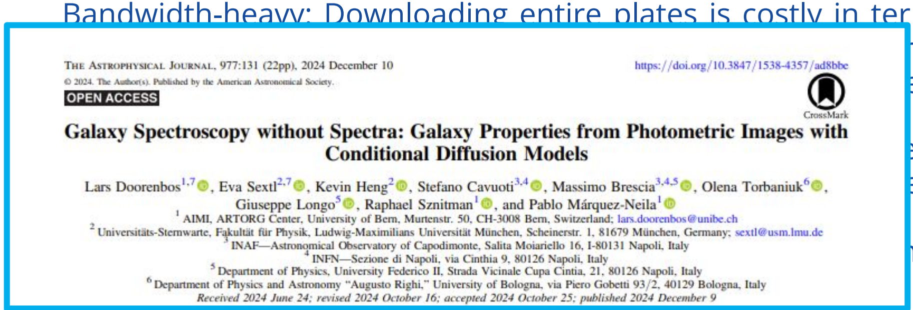
A&A, 687, A246 (2024)
https://doi.org/10.1051/0004-6361/202450166
© The Authors 2024

Astronomy & Astrophysics

Identification of problematic epochs in astronomical time series through transfer learning[★,★★]

Stefano Cavuoti[1,2,★★★], Demetra De Cicco[3,1,4,★★★], Lars Doorenbos[5], Massimo Brescia[3,1,2], Olena Torbaniuk[6,7], Giuseppe Longo[3], and Maurizio Paolillo[3,1,2]

~1M 51x51px single band thumbnails



The Astrophysical Journal, 977:131 (22pp), 2024 December 10
https://doi.org/10.3847/1538-4357/ad8bbe
© 2024. The Author(s). Published by the American Astronomical Society.
OPEN ACCESS
CrossMark

Galaxy Spectroscopy without Spectra: Galaxy Properties from Photometric Images with Conditional Diffusion Models

Lars Doorenbos[1,7], Eva Sextl[2,7], Kevin Heng[2], Stefano Cavuoti[3,4], Massimo Brescia[3,4,5], Olena Torbaniuk[6], Giuseppe Longo[5], Raphael Sznitman[1], and Pablo Márquez-Neila[1]

[1] AIMI, ARTORG Center, University of Bern, Murtenstr. 50, CH-3008 Bern, Switzerland; lars.doorenbos@unibe.ch
[2] Universitäts-Sternwarte, Fakultät für Physik, Ludwig-Maximilians Universität München, Scheinerstr. 1, 81679 München, Germany; sextl@usm.lmu.de
[3] INAF—Astronomical Observatory of Capodimonte, Salita Moiariello 16, I-80131 Napoli, Italy
[4] INFN—Sezione di Napoli, via Cinthia 9, 80126 Napoli, Italy
[5] Department of Physics, University Federico II, Strada Vicinale Cupa Cintia, 21, 80126 Napoli, Italy
[6] Department of Physics and Astronomy “Augusto Righi,” University of Bologna, via Piero Gobetti 93/2, 40129 Bologna, Italy
Received 2024 June 24; revised 2024 October 16; accepted 2024 October 25; published 2024 December 9

~300k 64x64px 5-band thumbnails

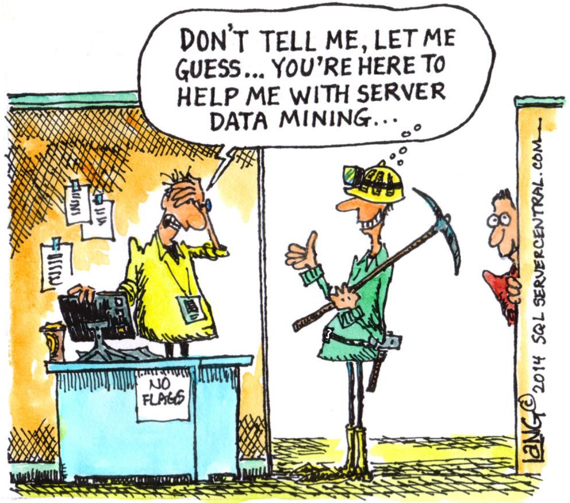
I Don't Want to Miss a Thing

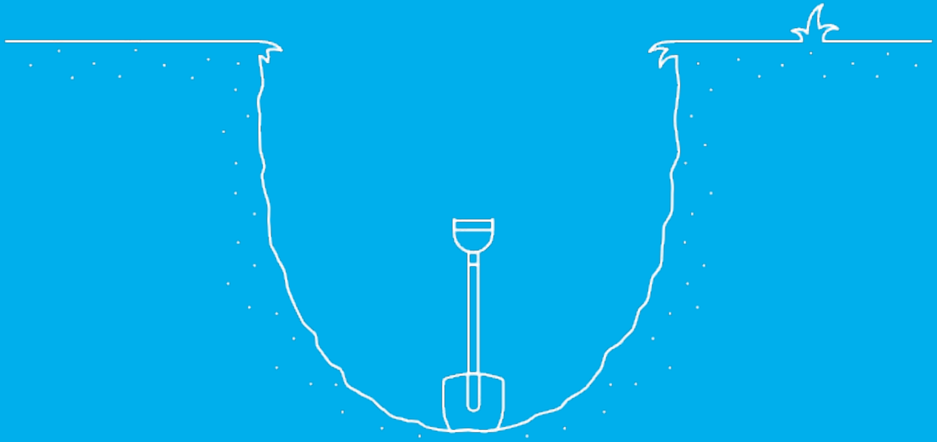
On the other hand I need to have control on missing data. Usually in astronomical DB a value could be missing for mainly three reason:

A. Observed but not detected
B. Observed but the object has been masked for a problem
C. Not observed

It is really important to distinguish at least between A. & B. from C. since I would deal with them in a very different way.

Thank you for your attention!


DON'T TELL ME, LET ME GUESS... YOU'RE HERE TO HELP ME WITH SERVER DATA MINING...
NO FLAGS
LANG© 2014 SQL SERVERCENTRAL.COM



Possible Discussion Points

- computational environments to enable in-place data analysis
- complex query infrastructure
- importance of data quality flags
- in-place cross-match
- precomputed cross-match with surveys
- on-the-fly cutout service
- missing data information