DATA MINING

MPAGS Astrophysical Techniques 2023

Kendall Ackley

kendall.ackley@warwick.ac.uk

KNOWLEDGE DISCOVERY IN DATA

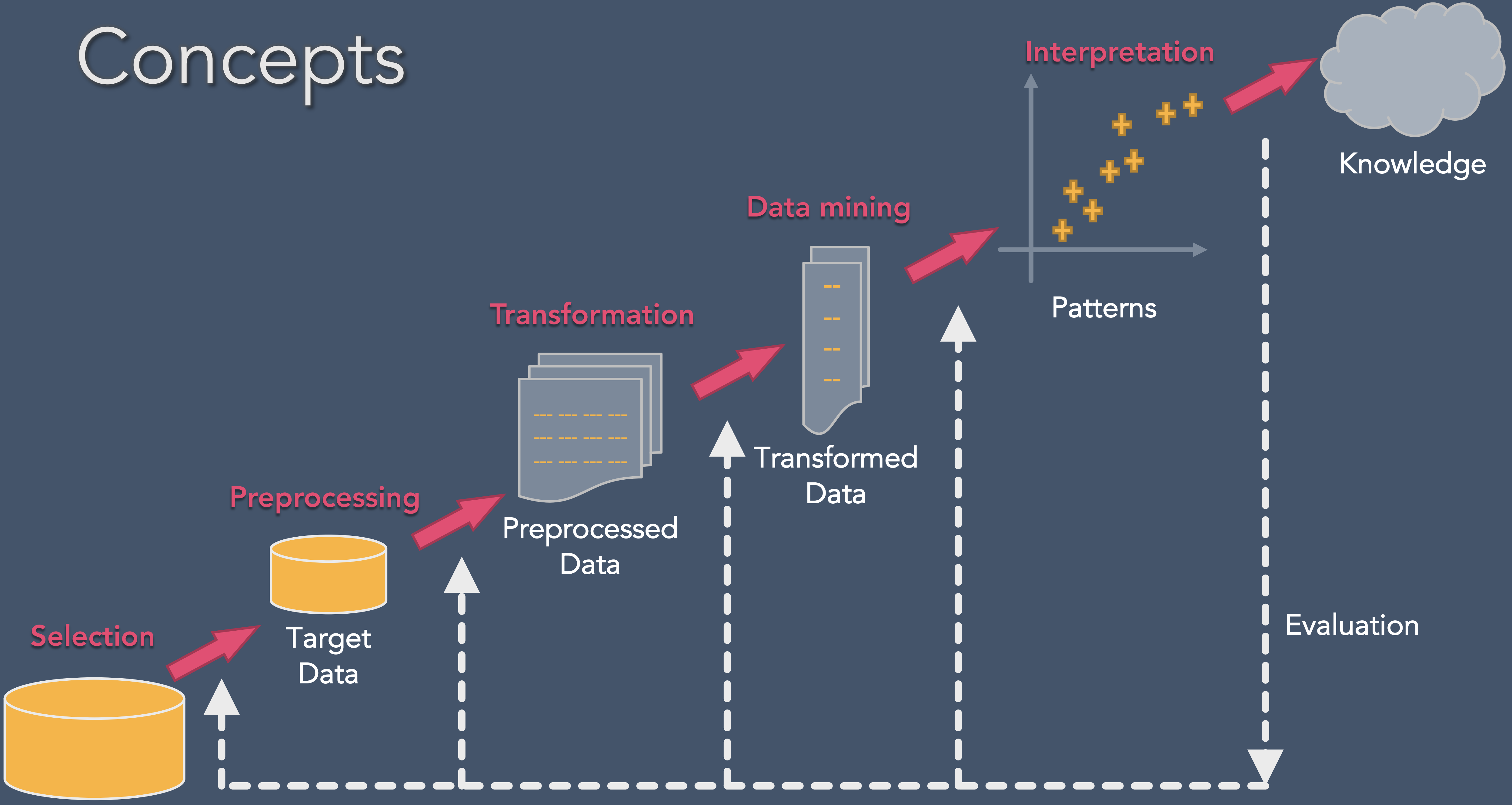
~~DATA MINING~~

MPAGS Astrophysical Techniques 2023

Kendall Ackley

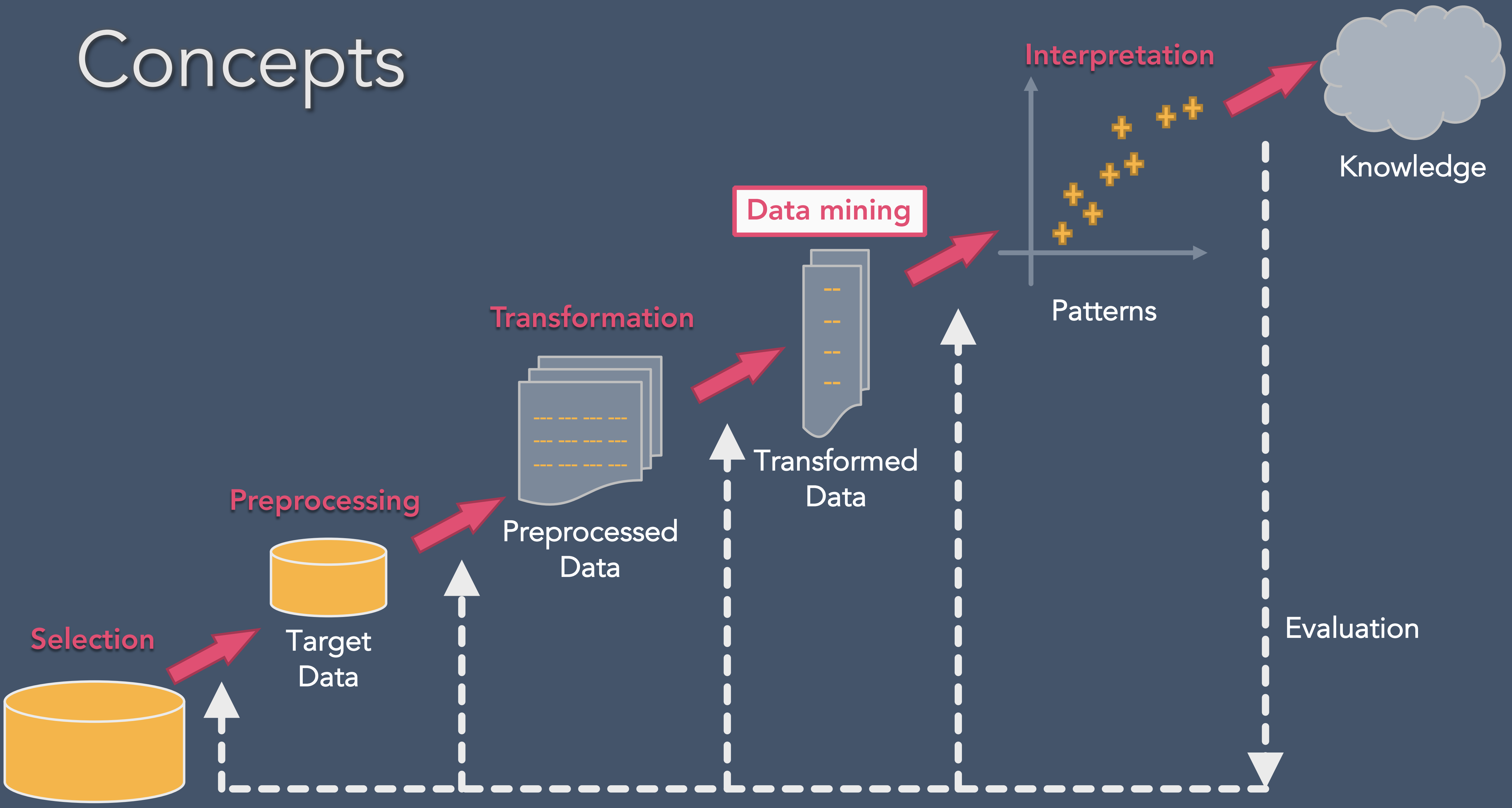
kendall.ackley@warwick.ac.uk

Concepts



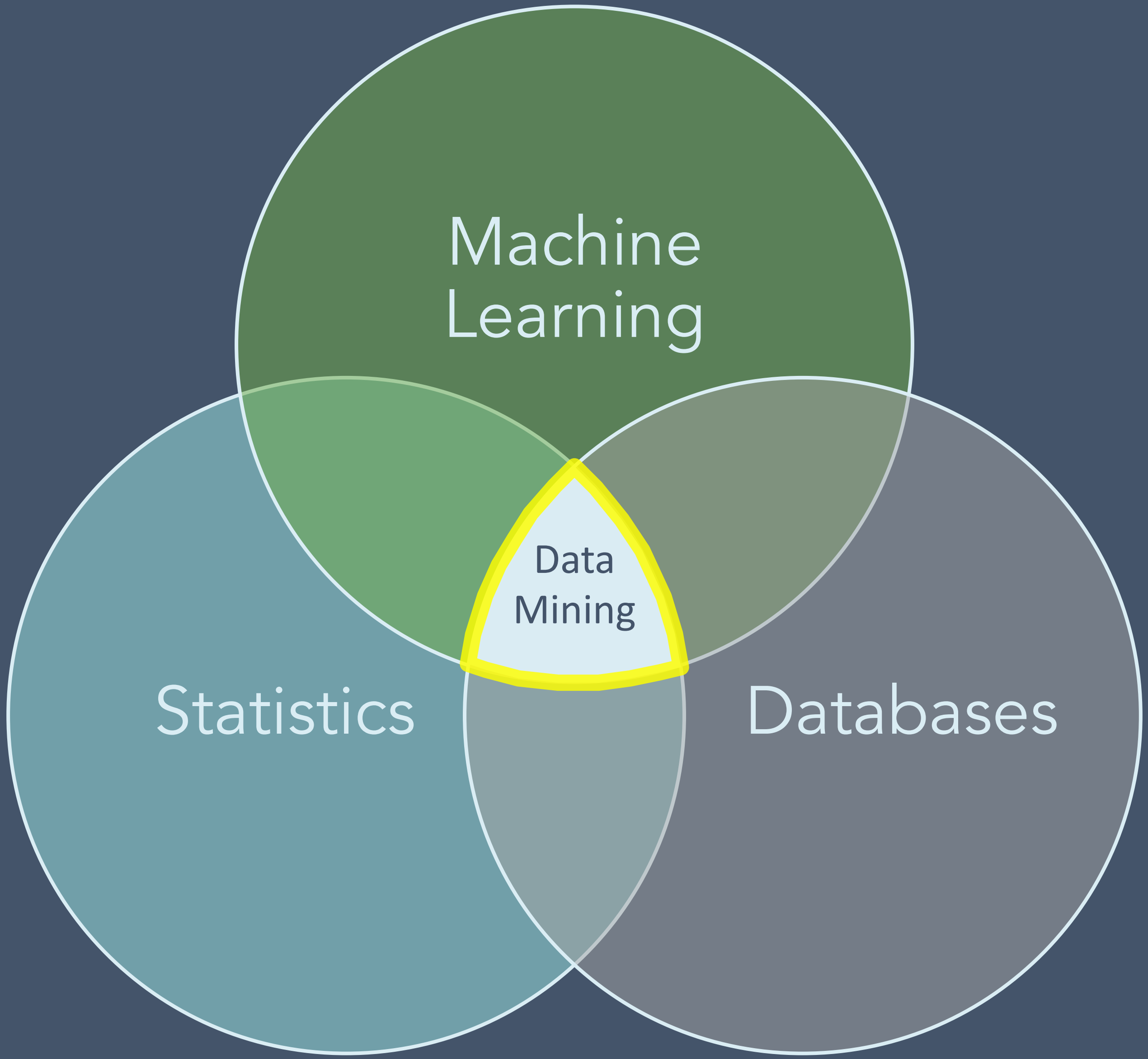
Adapted from Fig 1. Fayyad, Piatetsky-Shapiro & Smyth 1996

Concepts



Adapted from Fig 1. Fayyad, Piatetsky-Shapiro & Smyth 1996

Concepts



> *Data mining is an interdisciplinary field at the intersection of artificial intelligence, machine learning, statistics, and database systems*

https://www.kdd.org/

Data mining applications

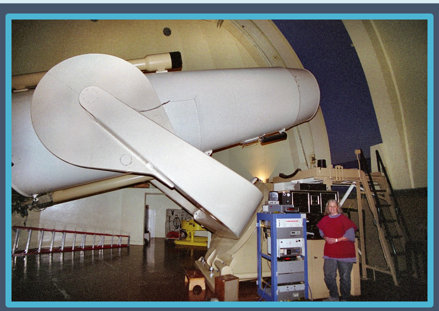
- Financial
  - predicting whether you’ll pay a loan back
- Retail
  - showing you what you want to buy next
- Government
  - determining if you are a security threat
- Healthcare
  - estimating your risk of various diseases
- Sciences
  - knowledge discovery

Astronomical optical sky survey data sets

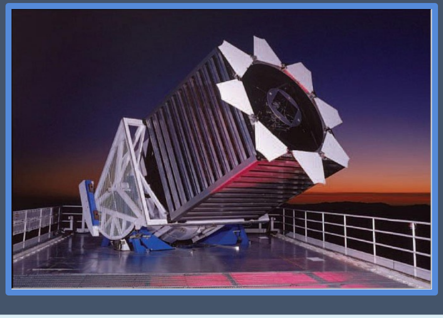
1990s

Last major photographic plate survey complete

- POSS-II (~3TB whole survey digitized)



2000s



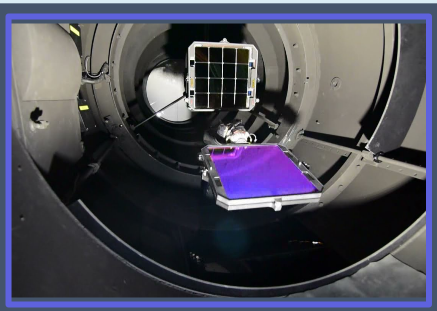
First major digital surveys begin

- SDSS (~10TB per year)

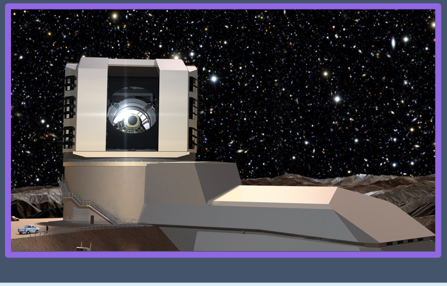
2010s

Wide-field surveys

- ZTF (~1TB **per night**, plus 0.5-1 billion photometry measurements)



2020s



Next generation surveys

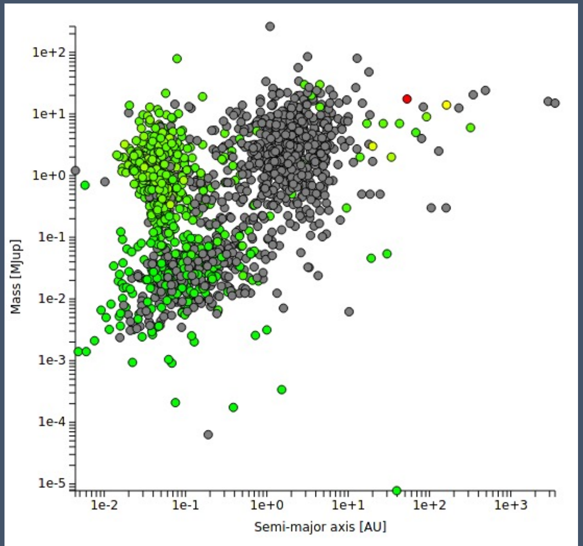
- LSST (~20TB **per night**, ~200PB whole survey)

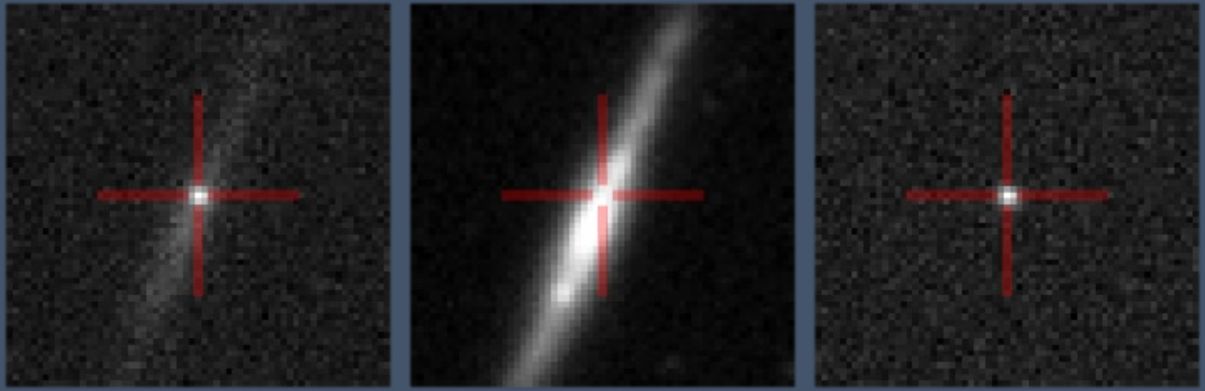
Data resources – Optical sky (wide)

Survey	Area (fraction of sky)	Filters	Depth (mag)	URL
SDSS	~1/3 (northern)	ugriz	~22 (ugr) ~21 (iz)	https://www.sdss.org/
PTF/ZTF	~3/4 (northern)	gr	~20-21	https://irsa.ipac.caltech.edu/Missions/ptf.html https://irsa.ipac.caltech.edu/Missions/ztf.html
Legacy Survey	~1/3 (high Galactic latitudes)	grz	~25 (g) ~24 (r) ~23 (z)	https://www.legacysurvey.org
Pan-STARRS	~3/4 (northern)	grizy	~23 (gri) ~22 (z) ~21 (y)	https://panstarrs.stsci.edu/
SkyMapper	~1/2 (southern)	uvgriz	~20 (uz) ~22 (g,r) ~21 (i)	https://skymapper.anu.edu.au
ATLAS	~1/10 (southern)	ugriz	~22 (ui) ~23 (gr) ~21 (z)	https://astro.dur.ac.uk/Cosmology/vstatlas/
DES	~1/8 (southern)	grizY	~25 (gr) ~24 (iz) ~22 (Y)	https://www.darkenergysurvey.org/
EGaPS (= VPHAS+, IPHAS, UVEX)	~1/15 (Galactic plane)	UgriHα	~21 (gr)	https://www.vphasplus.org/ http://www.iphas.org/ https://www.astro.ru.nl/uvex/

Data resources – Optical sky (new)

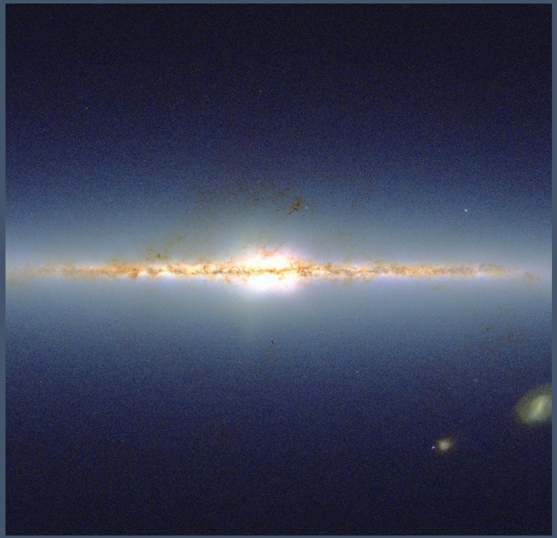
Resource	Description	Depth (mag)	URL
ZTF brokers	User-friendly interfaces to large data streams of transient alerts (supernovae, novae, outbursts etc.). Will also host LSST alerts	gr~21	e.g. Lasair: https://lasair.roe.ac.uk/ ALeRCE: https://alerce.online/
Gaia Alerts	Alerts are triggered by any Gaia source changing in brightness above some threshold	G~20	http://gsaweb.ast.cam.ac.uk/alerts/home
Transient Name Server (TNS)	IAU-designated repository for all discovery and classification reports of new transients		https://www.wis-tns.org/
Exoplanet Catalogues	Databases of exoplanet discoveries		http://www.openexoplanetcatalogue.com/ http://exoplanet.eu/catalog/

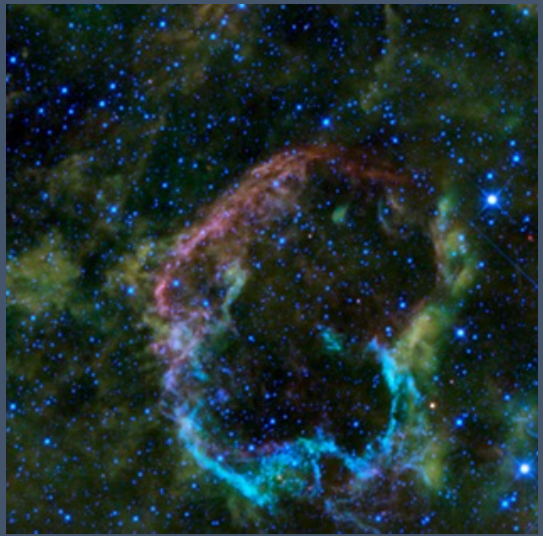




Data resources – Infrared sky (wide)

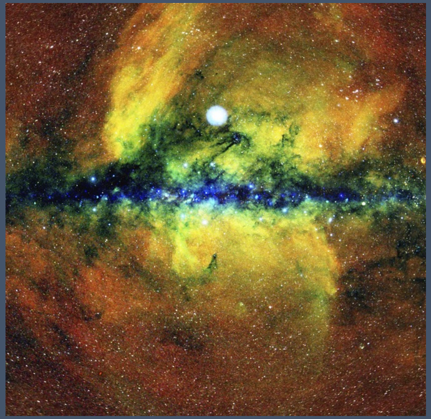
Survey	Area (fraction of sky)	Filters	Depth (mag)	URL
2MASS	~1	JHK	~16	https://irsa.ipac.caltech.edu/Missions/2mass.html
UKIDSS	~1/5	JHK	~18	http://wsa.roe.ac.uk/
VHS	~1/2	YJHK	~20	https://www.vista-vhs.org/
WISE	~1	3-22 micron	~17-8	https://wise2.ipac.caltech.edu/docs/release/allsky/

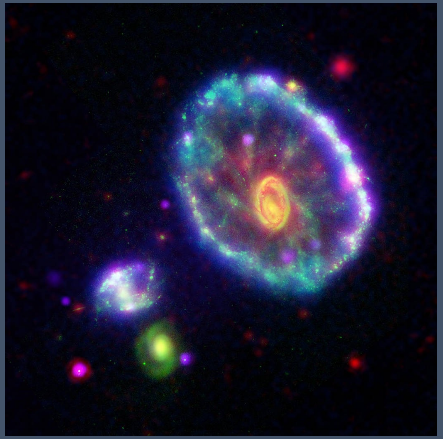




Data resources – Radio/UV/Xray sky (wide)

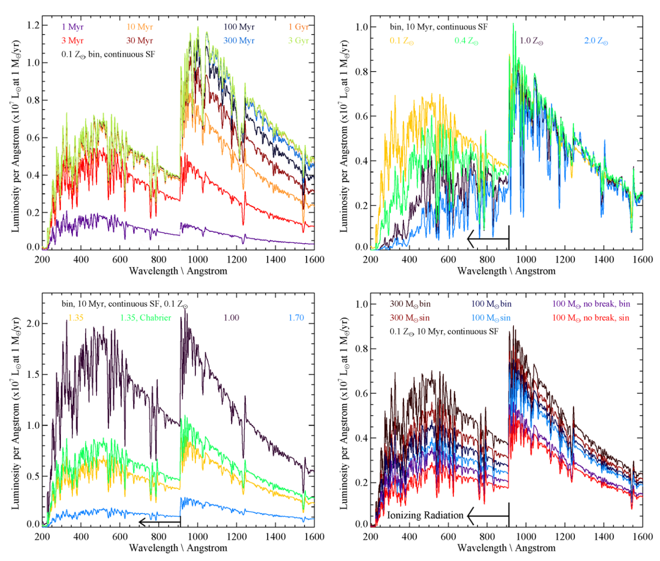
Survey	Area (fraction of sky)	Wavelengths	Depth	URL
FIRST	~1/4	~21cm	~1 mJy	https://sundog.stsci.edu/
GALEX	~1 (but significant gaps)	UV (135-280nm)	~20 mag	https://archive.stsci.edu/missions-and-data/galex
ROSAT	~1	Soft X-ray (~2 keV)	~3x10^{-12} erg/cm^2/s	https://heasarc.gsfc.nasa.gov/docs/rosat/rosat3.html
eROSITA (ongoing)	~1 (but practically 1/2 for open data)	Soft and Hard X-ray (2-30 keV)	~10^{-14} erg/cm^2/s	https://www.mpe.mpg.de/eROSITA





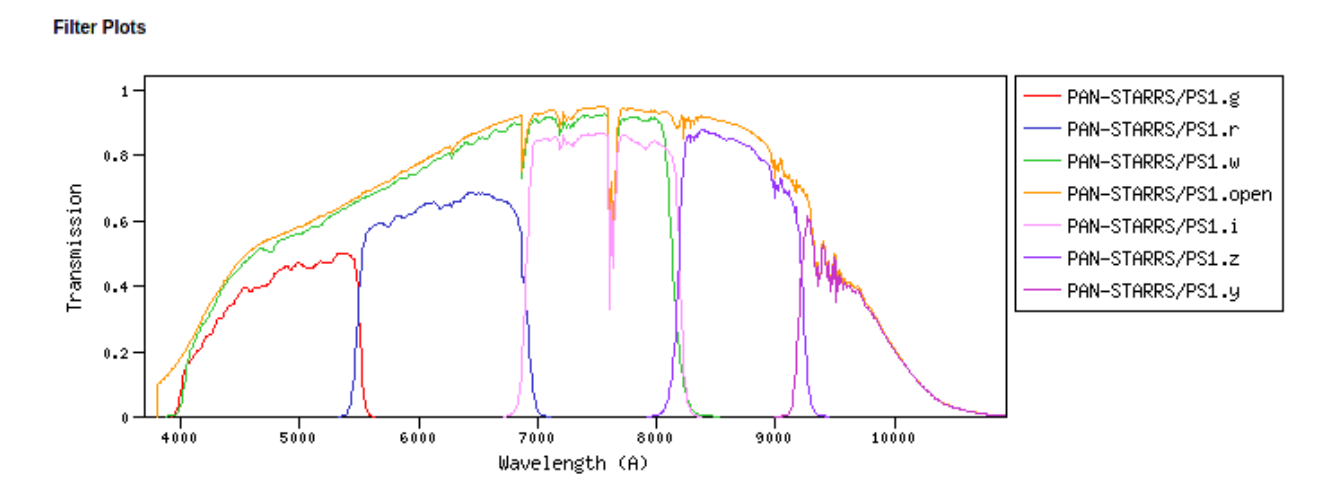
Data resources – Simulations

Resource	Content	URL
IllustrisTNS/ EAGLE/ Horizon-AGN	Hydrodynamical cosmological simulation	https://www.tng-project.org/ http://icc.dur.ac.uk/Eagle/ https://www.horizon-simulation.org/
BPASS	Binary stellar population synthesised SEDs	https://bpass.auckland.ac.nz/
CosmoSim	Semi-analytic cosmological simulations, galaxy catalogues	https://www.cosmosim.org/cms/data/



Data resources – Misc

Resource	Content	URL
Filter Profile Service	Standard format filter profiles for all major surveys to compare photometry, generate SEDs etc.	http://svo2.cab.inta-csic.es/theory/fps/
NIST Atomic Spectra Database	Atomic lines database for spectral line identification.	https://physics.nist.gov/PhysRefData/ASD/lines_form.html
ADS abstracts	Digital library portal for researchers in astronomy	https://ui.adsabs.harvard.edu/
NASA HEARSARC	Primary archive for NASA's (and other space agencies') missions studying electromagnetic radiation	https://heasarc.gsfc.nasa.gov/

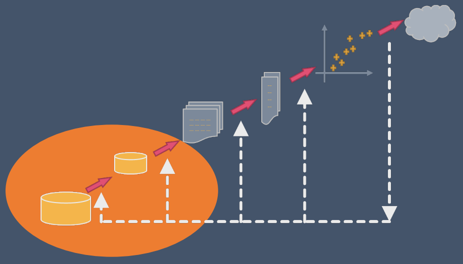


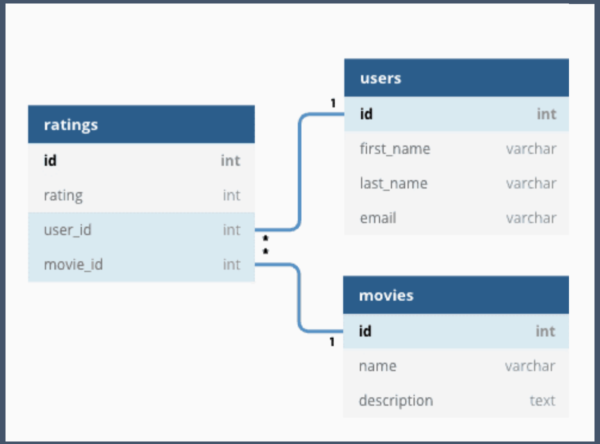
Astronomical Data Warehouses

(Virtual Observatories)

- Virtual Observatory https://www.ivoa.net/astronomers
  - Sets standards for astronomical data to enable easier data warehousing
  - Links to various VO-compliant software
- Strasbourg https://cds.u-strasbg.fr
  - SIMBAD – excellent "quick-look" tool for finding a wealth of information on objects
  - VIZIER – very large collection of diverse astronomical catalogues
    - Often data associated with publications are hosted here
  - ALADIN – Nice interactive sky atlas with plenty of integration to visualise SIMBAD/VIZIER data
- IRSA https://irsa.ipac.caltech.edu
  - Friendly interface to many large (mainly US) projects' databases
- MAST https://archive.stsci.edu/
  - Access to multiple space-based mission data archives (mostly NASA)
  - Gaia, JWST, XMM, SwiftUVOT, Hubble, TESS, Kepler, etc.

Databases 101





ratings	
id	int
rating	int
user_id	int
movie_id	int

users	
id	int
first_name	varchar
last_name	varchar
email	varchar

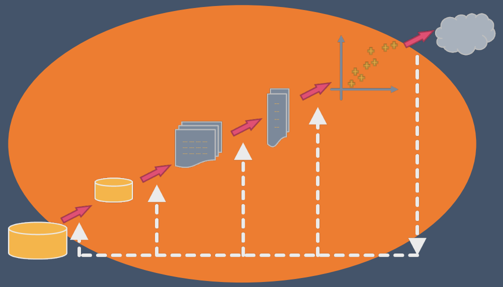
movies	
id	int
name	varchar
description	text

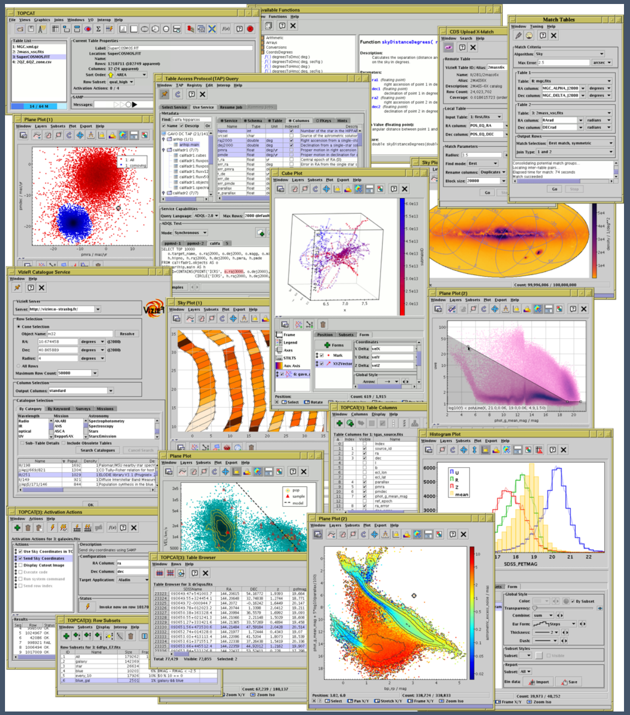
- Most data resources build on a "relational" database
  - A schema defines tables
  - tables define columns
  - columns can be linked between tables
- SQL is the language used to interrogate relational databases
  - Many variants!
  - Reasonably quick to learn enough for most use cases – LOTS of resources online
  - e.g.

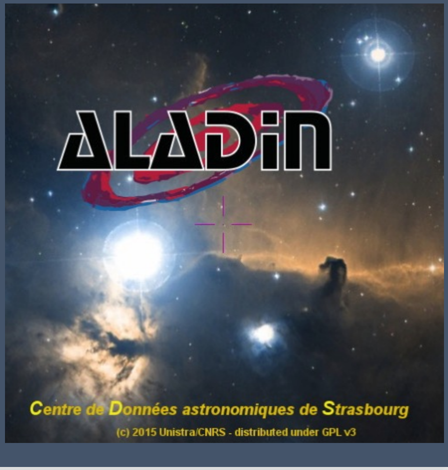
```
select mjd, mag, mag_error, filter from photometry where
name = "delta_scuti";
```

- Always refer to the schema and usage documentation for the data resource
  - Descriptions of tables and columns
  - Non-SQL (e.g. GUI) interfaces to searching
  - Example queries
  - e.g. http://skyserver.sdss.org/dr16/en/tools/search/searchhome.aspx

Tools

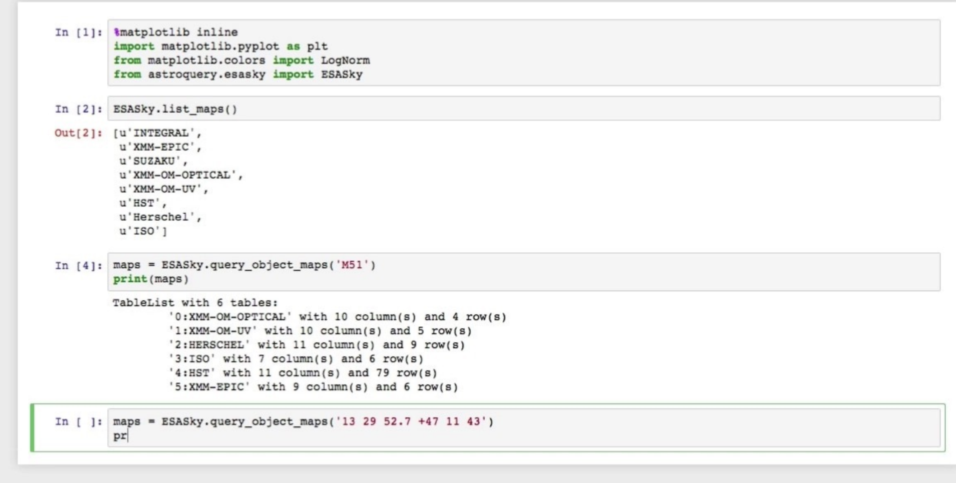






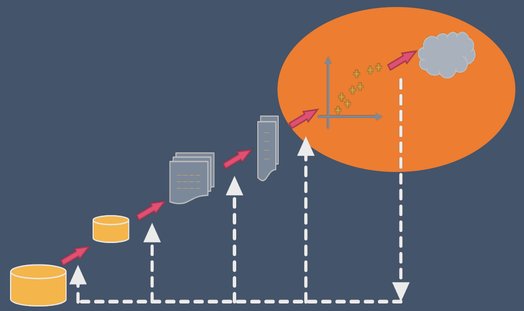
ALADiN
Centre de Données astronomiques de Strasbourg
(c) 2015 Unistra/CNRS - distributed under GPL v3

- TOPCAT http://www.star.bris.ac.uk/~mbt/topcat/
  - Lots of features for querying a whole range of resources
  - Built in analysis such as plotting, statistics
- ALADIN https://aladin.u-strasbg.fr/
  - Excellent quick visualisation of survey imaging
  - Good catalogue querying tools
- Astroquery https://astroquery.readthedocs.io/en/latest/
  - Programmatic access to databases in python
  - close relation to astropy – vastly streamlines retrieval to analysis



```
In [1]: %matplotlib inline
        import matplotlib.pyplot as plt
        from matplotlib.colors import LogNorm
        from astroquery.esasky import ESASky

In [2]: ESASky.list_maps()

Out[2]: [u'INTEGRAL',
         u'XMM-EPIC',
         u'SUZAKU',
         u'XMM-OM-OPTICAL',
         u'XMM-OM-UV',
         u'HST',
         u'Herschel',
         u'ISO']

In [4]: maps = ESASky.query_object_maps('M51')
        print(maps)

        TableList with 6 tables:
                '0:XMM-OM-OPTICAL' with 10 column(s) and 4 row(s)
                '1:XMM-OM-UV' with 10 column(s) and 5 row(s)
                '2:HERSCHEL' with 11 column(s) and 9 row(s)
                '3:ISO' with 7 column(s) and 6 row(s)
                '4:HST' with 11 column(s) and 79 row(s)
                '5:XMM-EPIC' with 9 column(s) and 6 row(s)

In [ ]: maps = ESASky.query_object_maps('13 29 52.7 +47 11 43')
        pr
```

Analysis (in python)



- Pandas DataFrames
  - Close representation of a database table in python – widely used across data science
  - https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html#dataframe
- Astropy Table
  - Human-friendly interfaces to data tables – better to work with than raw numpy arrays
  - https://docs.astropy.org/en/stable/table/
- AstroML
  - Astro-specific Machine learning and data-mining tools
  - http://www.astroml.org/
- Scikit-learn
  - Accessible machine learning toolkit – very easy to dive into
  - https://scikit-learn.org/stable/
- Tensorflow and Pytorch
  - Deep-learning toolkits – significant learning curves but extremely powerful
  - https://www.tensorflow.org/
  - https://pytorch.org/