

DATA MINING

MPAGS ASTROPHYSICAL TECHNIQUES 2020-2021

JOE LYMAN

J.D.LYMAN@WARWICK.AC.UK



KNOWLEDGE DISCOVERY IN DATA ~~DATA MINING~~

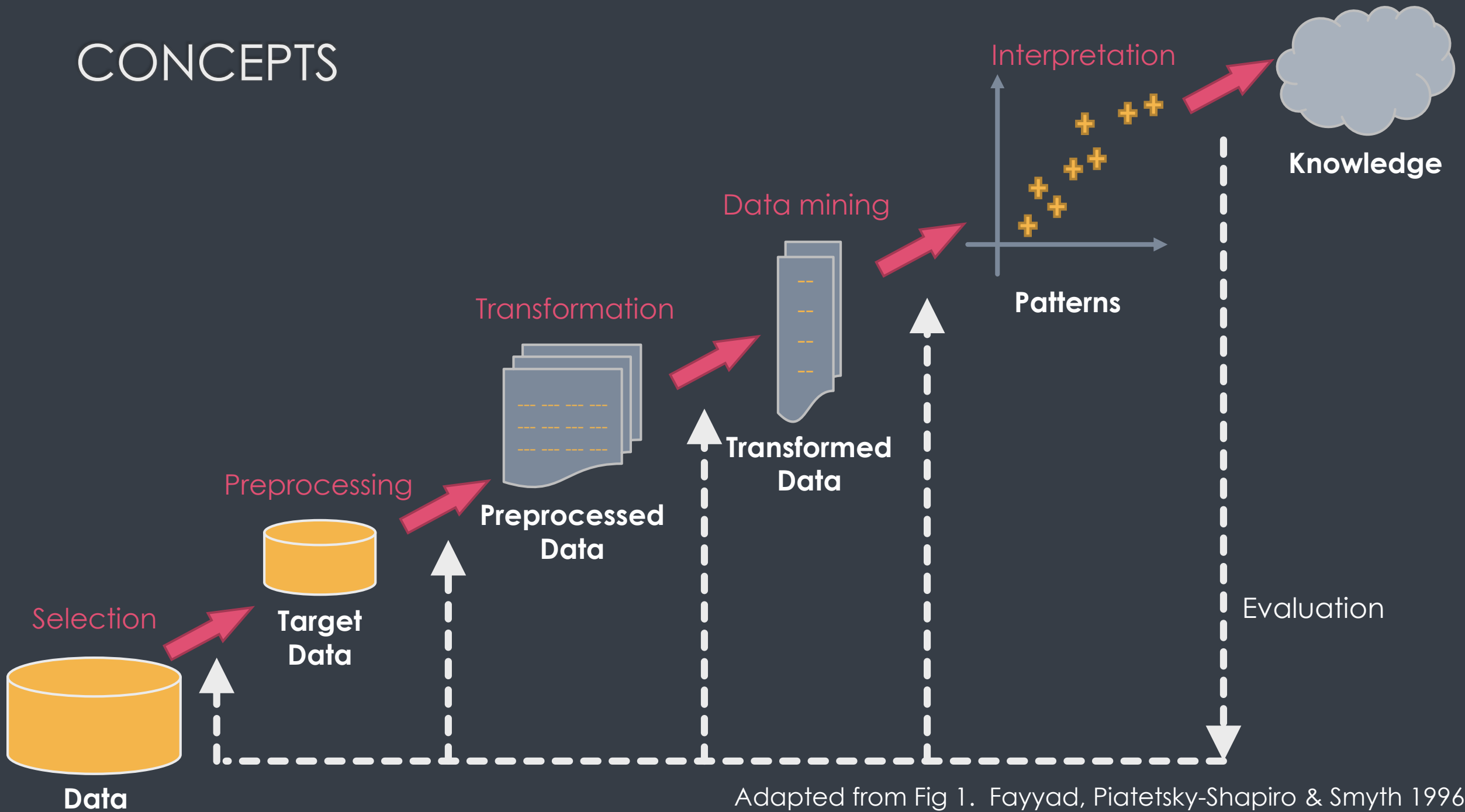
MPAGS ASTROPHYSICAL TECHNIQUES 2020-2021

JOE LYMAN

J.D.LYMAN@WARWICK.AC.UK

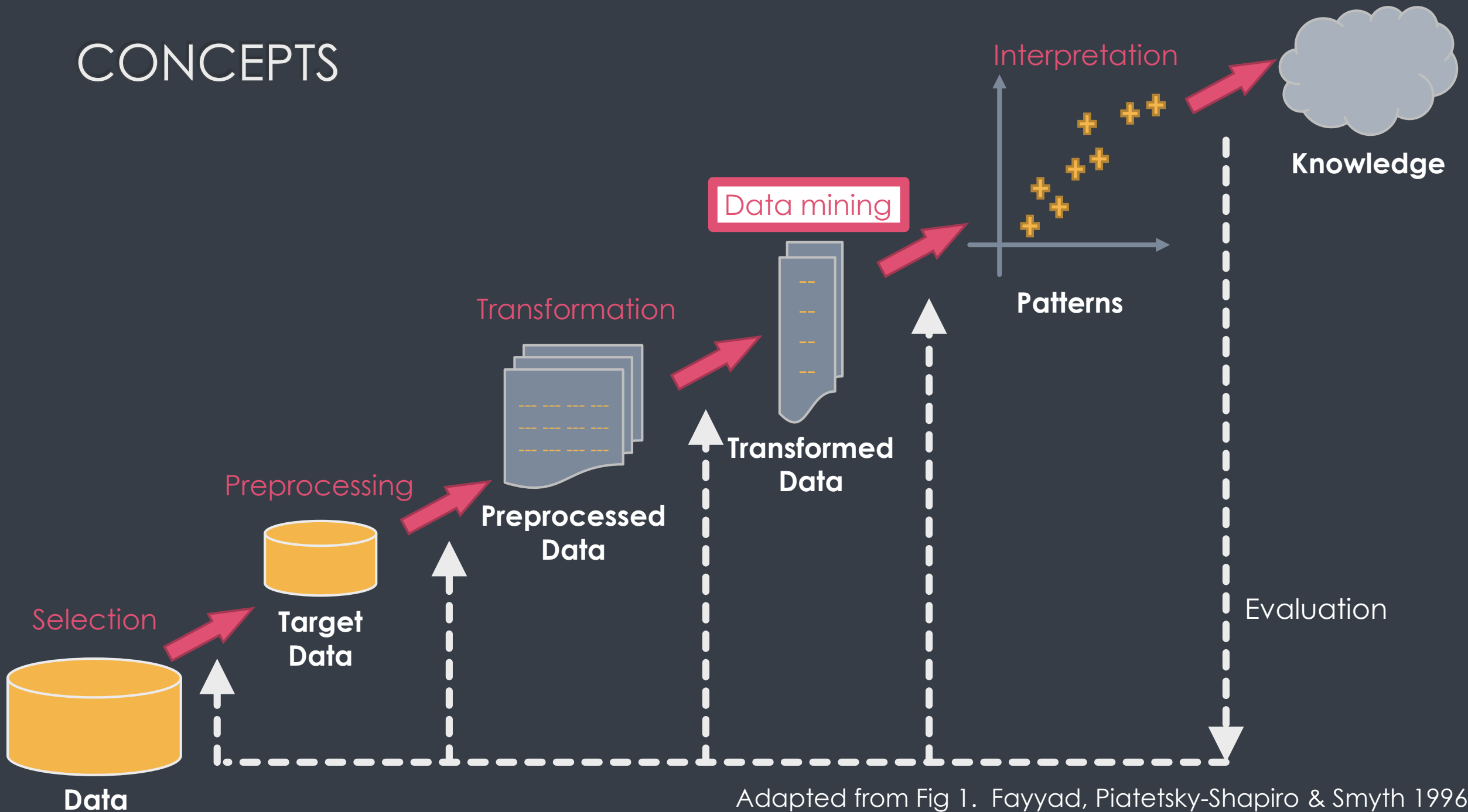


CONCEPTS



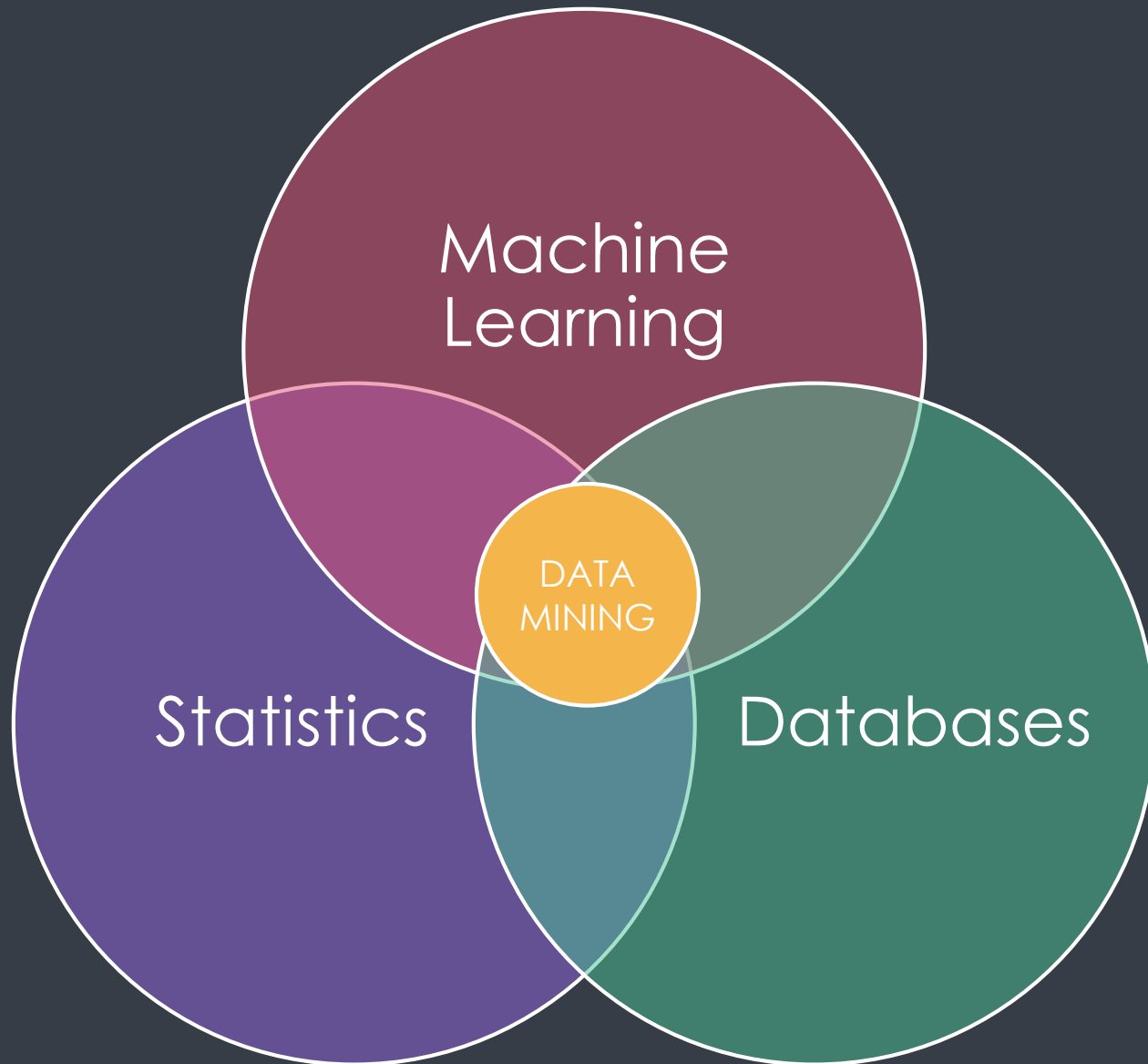
Adapted from Fig 1. Fayyad, Piatetsky-Shapiro & Smyth 1996

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

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

.. and anywhere there are large datasets

ASTRONOMICAL OPTICAL SKY SURVEY DATA SETS

Last major photographic plate survey complete

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

Wide-field surveys

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

1990s

2000s

2010s

2020s

First major digital surveys begin

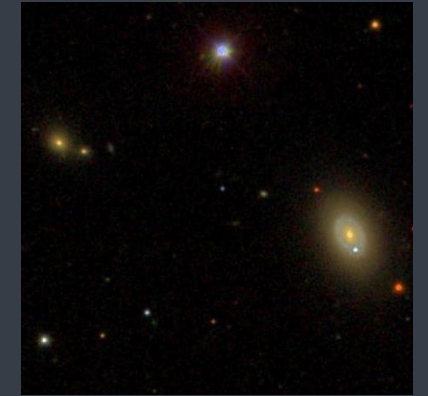
- SDSS (~10TB per year)

Next generation surveys

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

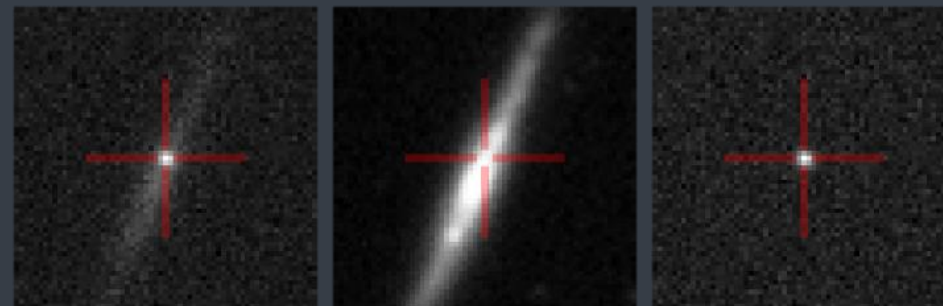
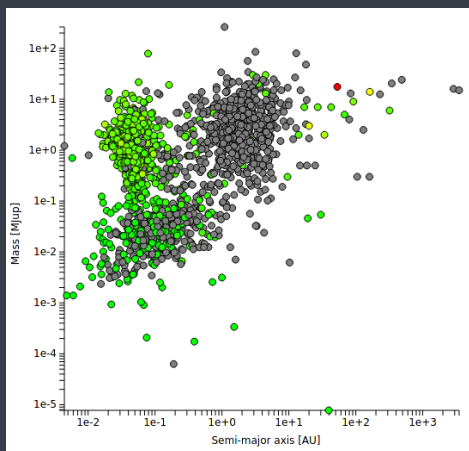
DATA RESOURCES – OPTICAL SKY (WIDE)

Survey	Area (fraction of sky)	Filters	Depth (mag)	URL
SDSS	~1/3 (northern)	ugriz	~22 (ugr) ~21 (iz)	https://www.sdss.org/
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)	UgriHa	~21 (gr)	https://www.vphasplus.org/ http://www.iphas.org/ https://www.astro.ru.nl/uvex/



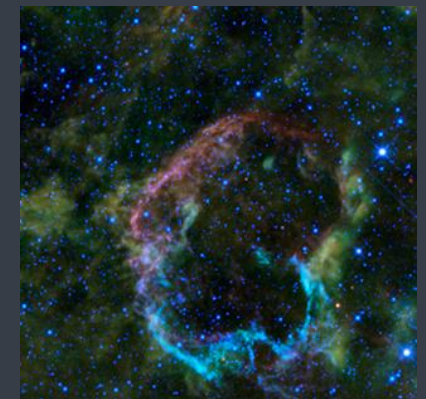
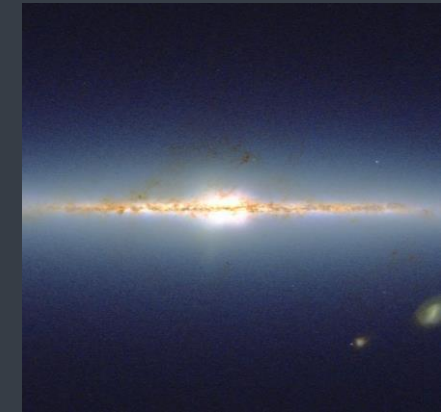
DATA RESOURCES – OPTICAL SKY (NEW)

Resource	Description	Depth (mag)	URL
ZTF brokers	User-friendly interfaces to large data streams of alerts. 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	any	https://www.wis-tns.org/
Exoplanet Catalogues	Databases of exoplanet discoveries	?	http://www.openexoplanetcatalogue.com/ http://exoplanet.eu/catalog/



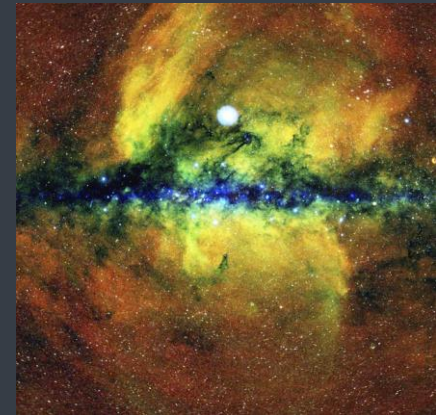
DATA RESOURCES – INFRARED SKY (WIDE)

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://irsa.ipac.caltech.edu/Missions/wise.html



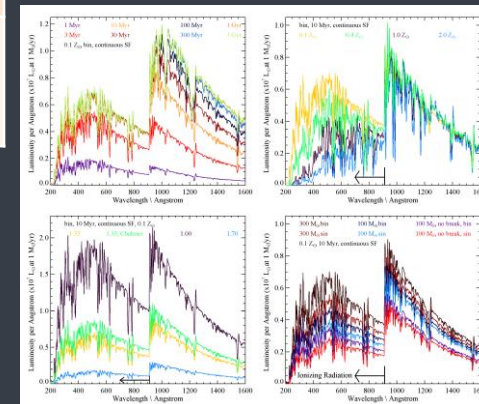
DATA RESOURCES – RADIO/UV/XRAY SKY (WIDE)

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)	~ 3×10^{-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

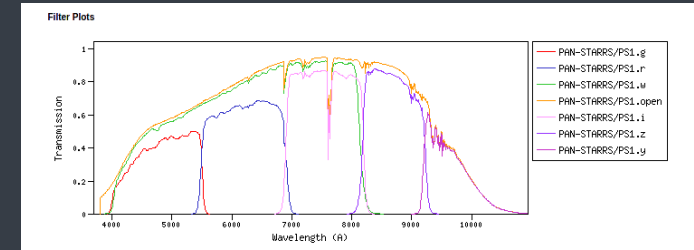


DATA RESOURCES – SIMULATIONS

Resource	Content	URL
IllustrisTNS/ EAGLE	Hydrodynamical cosmological simulation	https://www.tng-project.org/ http://icc.dur.ac.uk/Eagle/
BPASS	Binary stellar population synthesised SEDs	https://bpass.auckland.ac.nz/



DATA RESOURCES – MISC

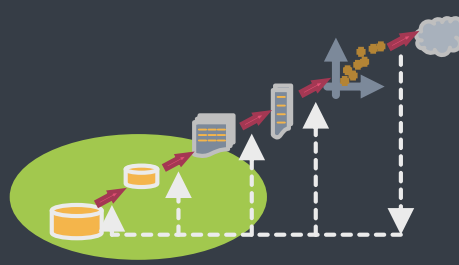


Resource	Content	URL
Filter Profile Service	Standard format filter profiles for all major surveys to compare photometry, generate SEDs etc.	http://svo2.cab.inta-csic.es/theory/fps/
NIST Atomic Spectra Database	Atomic lines database for spectral line identification.	https://physics.nist.gov/PhysRefData/ASD/lines_form.html

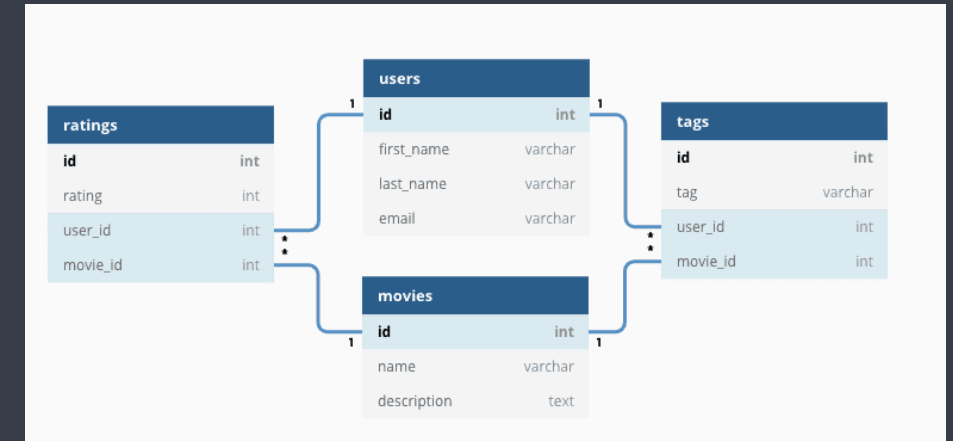
ASTRONOMICAL DATA WAREHOUSES (VIRTUAL OBSERVATORIES)

- VIRTUAL OBSERVATORY [HTTPS://WWW.IVOA.NET/ASTRONOMERS](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](https://cds.u-strasbg.fr)
 - SIMBAD – EXCELLENT “QUICKLOOK” 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](https://irsa.ipac.caltech.edu)
 - FRIENDLY INTERFACE TO MANY LARGE (MAINLY US) INFRARED PROJECTS DATABASES

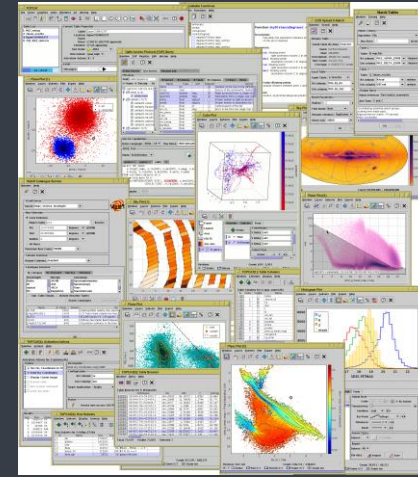
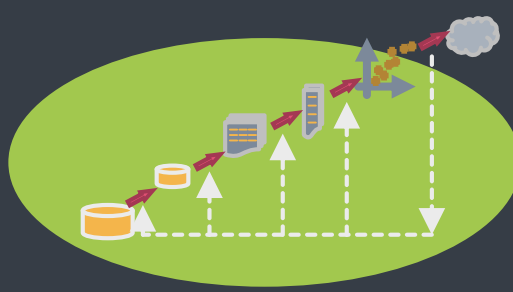
DATABASES 101



- 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](http://skyserver.sdss.org/DR16/en/tools/search/searchhome.aspx)

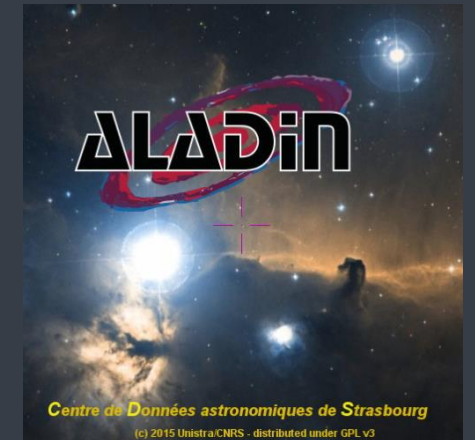


TOOLS



- TOPCAT [HTTP://WWW.STAR.BRIS.AC.UK/~MBT/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/](https://aladin.u-strasbg.fr/)
 - EXCELLENT QUICK VISUALISATION OF SURVEY IMAGING
 - GOOD CATALOGUE QUERYING TOOLS



- ASTROQUERY [HTTPS://ASTROQUERY.READTHEDOCS.IO/EN/LATEST/](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()

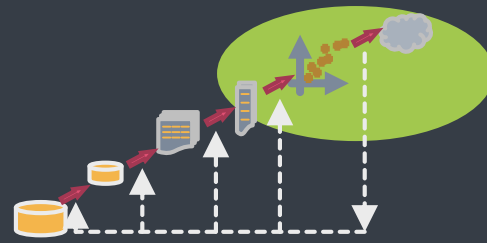
Out[2]: [u'INTORGAL',
u'XMM-EPIC',
u'SUZAKU',
u'XMM-OM-OPTICAL',
u'XMM-OM-UV',
u'HERA',
u'Herschel',
u'ISO']

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 4 row(s)
  4: HERA 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](https://pandas.pydata.org/pandas-docs/stable/user_guide/dsintro.html#dataframe)



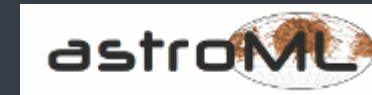
- ASTROPY TABLE

- HUMAN-FRIENDLY INTERFACES TO DATA TABLES – BETTER TO WORK WITH THAN RAW NUMPY ARRAYS
- [HTTPS://DOCS.ASTROPY.ORG/EN/STABLE/TABLE/](https://docs.astropy.org/en/stable/table/)



- ASTROML

- ASTRO-SPECIFIC MACHINE LEARNING AND DATA-MINING TOOLS
- [HTTP://WWW.ASTROML.ORG/](http://www.astroml.org/)



- SCIKIT-LEARN

- ACCESSIBLE MACHINE LEARNING TOOLKIT – VERY EASY TO DIVE INTO
- [HTTPS://SCIKIT-LEARN.ORG/STABLE/](https://scikit-learn.org/stable/)



- TENSORFLOW AND PYTORCH

- DEEP-LEARNING TOOLKITS – SIGNIFICANT LEARNING CURVES BUT EXTREMELY POWERFUL
- [HTTPS://WWW.TENSORFLOW.ORG/](https://www.tensorflow.org/)
- [HTTPS://PYTORCH.ORG/](https://pytorch.org/)

