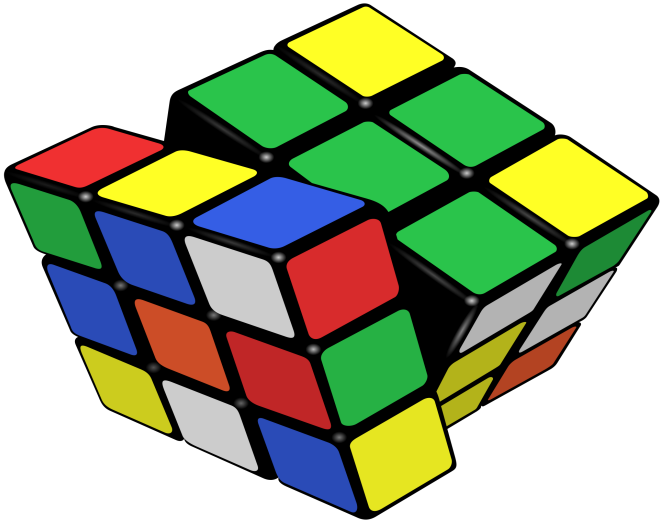
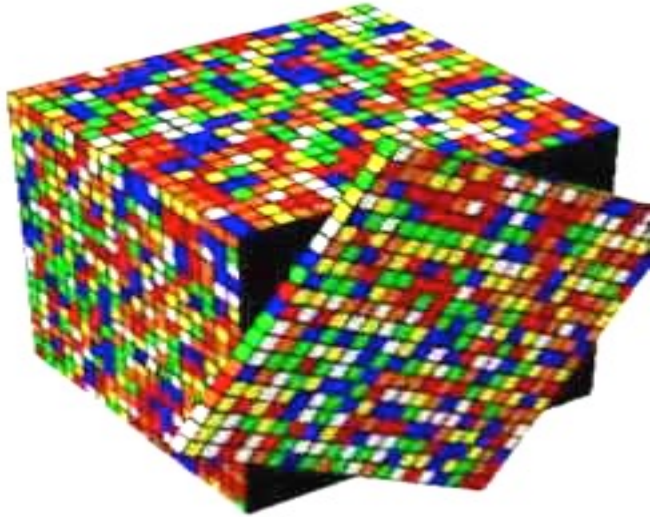
Data Mining in Astronomy

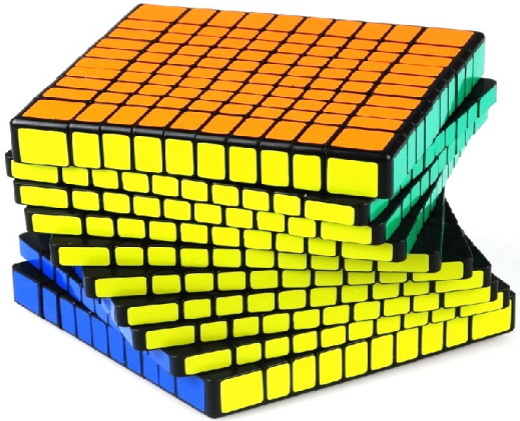
Nadeem Oozeer, PhD

Data Scientist

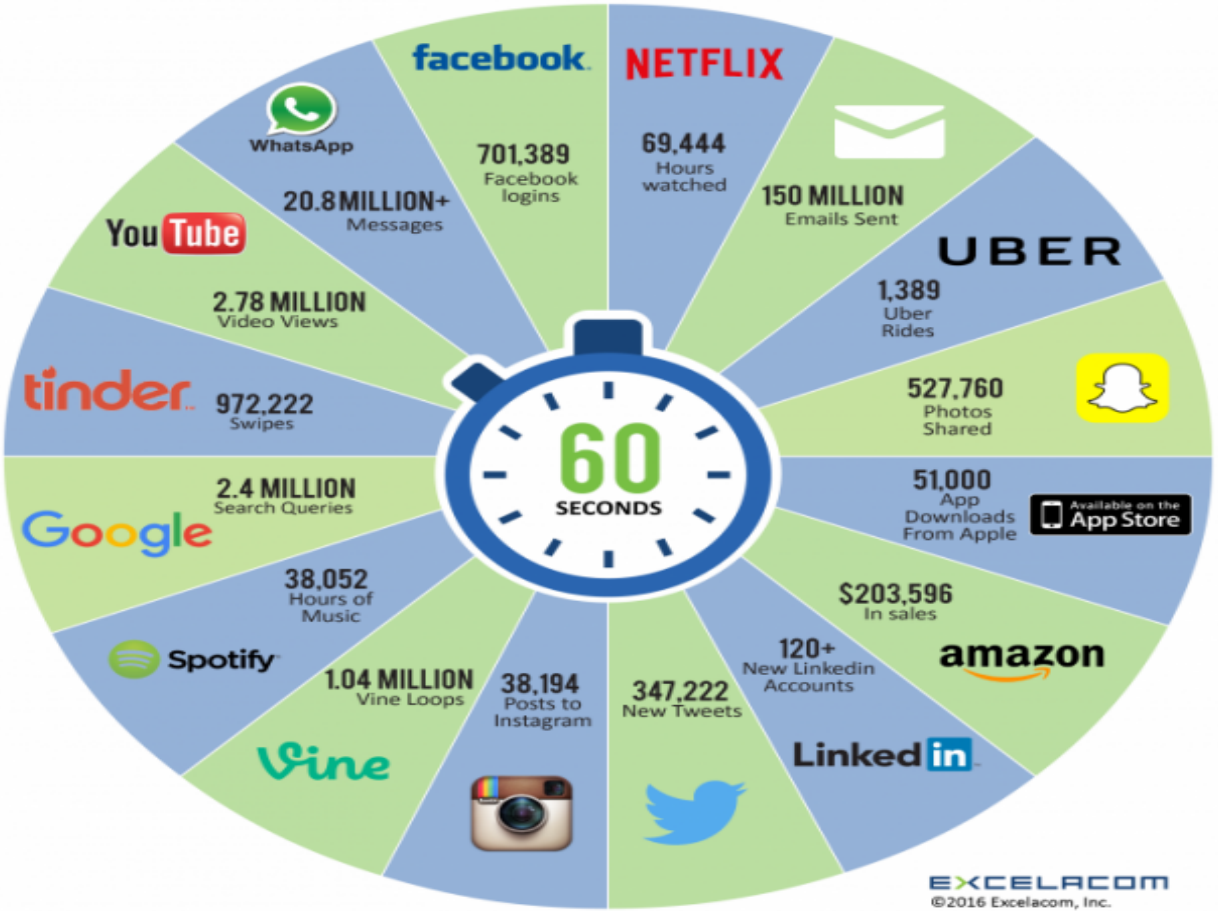
SKA SA / AIMS







2016 What happens in an INTERNET MINUTE?



EXCELACOM

Astronomical Data Deluge

Megadata

In excess of 1 Exabyte of raw data in a single day - more than the entire daily internet traffic

Square Kilometre Array

€1.5b

+ A €1.5 billion global science project

+ Astronomers and engineers from more than 70 institutes in 20 countries

3000

+ 3000 dishes, each 15m wide

+ Using enough optical fibre to wrap twice around the Earth

1,000,000 m²

+ A combined collecting area of about one square kilometre

Enough raw data to fill over 15 million 64GB iPods every day

+ Automated data classification = faster with fewer errors
+ Guided search = easier access for scientists and non-scientists alike
+ Frees researchers to be more productive and creative

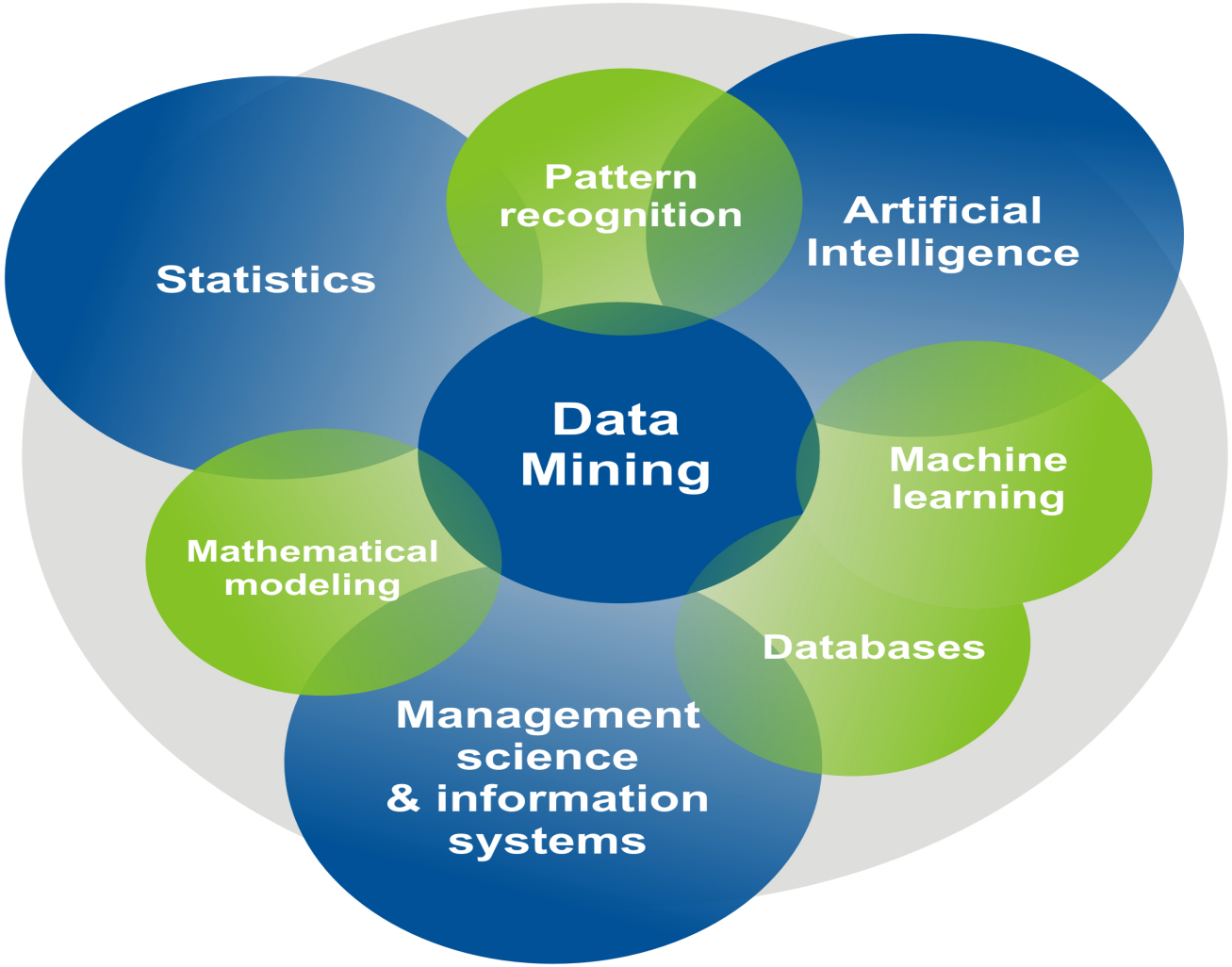
IBM Information Intensive Framework

A prototype software architecture to manage the megadata generated by SKA

IBM

Top image: SPDO/Swinburne Astronomy Productions

Big Data Examples


Statistics
Pattern recognition
Artificial Intelligence
Data Mining
Machine learning
Mathematical modeling
Databases
Management science & information systems

Data Mining & Privacy

- Privacy
  - I want information to be used only for my benefit
- Confidentiality:
  - I want information to go only to those authorized
- Cryptography community understands confidentiality
  - Solid, vetted definitions – Proof techniques
- Not sure if anybody really understands privacy
  - *But confidentiality often sufficient*

Solution

- Data Obfuscation
  - Nobody see the real data
- Summarization
  - Only the needed facts are exposed
- Data Separation
  - Data remains with trusted parties

Can we afford this?

nature

www.nature.com/nature Vol 442 | Issue no. 7098 | 6 July 2006

Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.

That might seem obvious. But despite the efforts of our editors and referees, papers in the biological sciences are still being submitted — and occasionally published — that do not adequately describe the reagents used. Unless efforts are redoubled to eliminate this

established didn't want the author to reveal the sequences, as this would jeopardize its *raison d'être*. This kind of stalemate matters, because it prevents the replication of experiments and inhibits the selection of appropriate controls in subsequent work.

Some authors claim replication is possible without full sequence information or the details of novel compounds. They say that the materials in question are for sale, enabling anyone to duplicate the paper. This misses the point. Scientific progress revolves around pro-

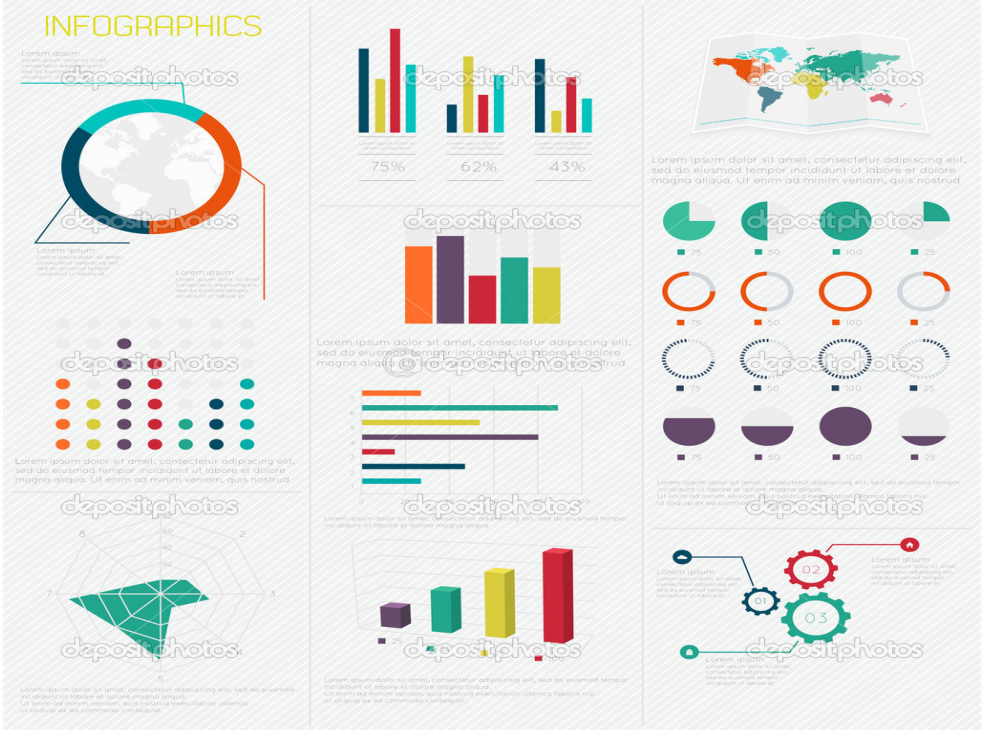
Reproducible Science

means context, quality, trust

means easy access to the sources

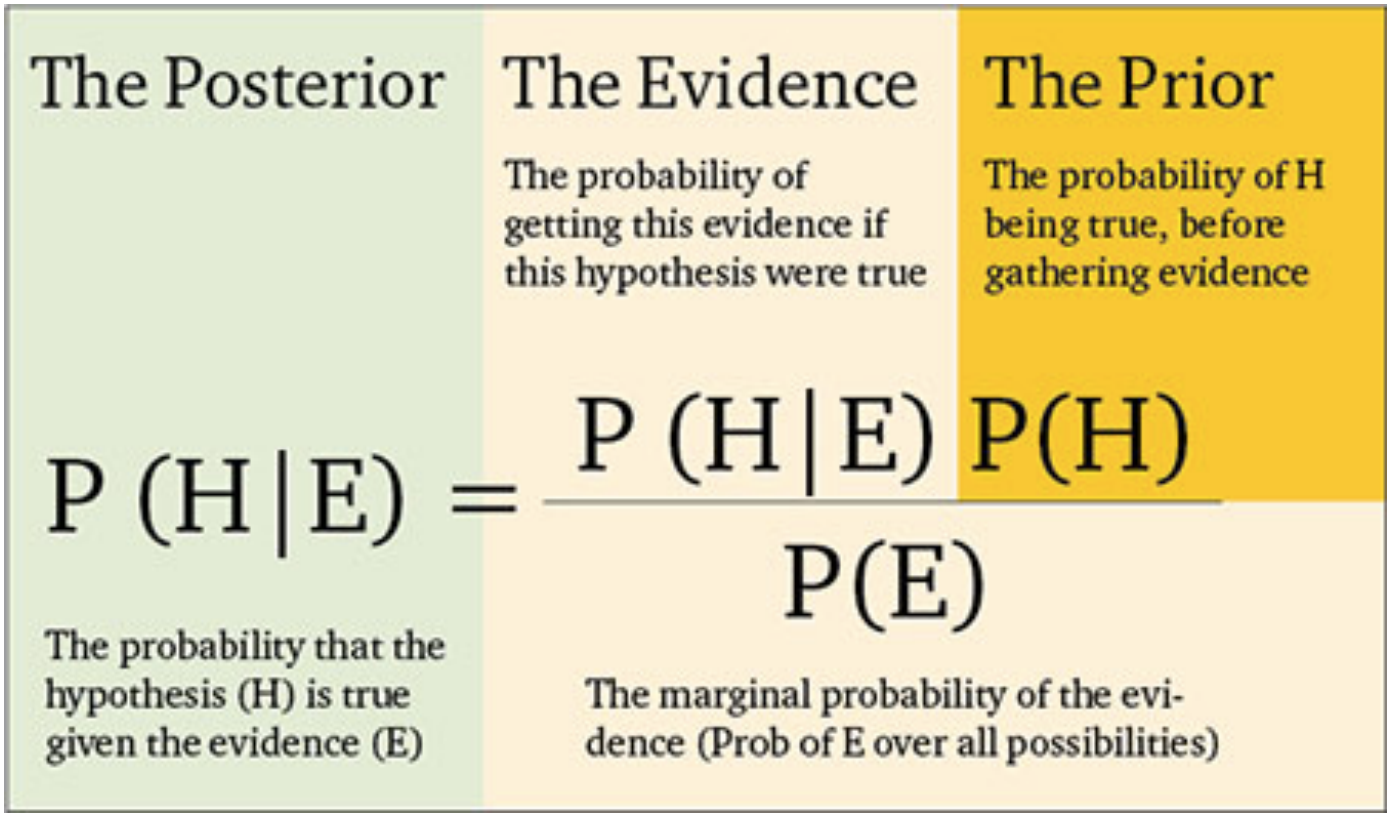
Knowledge Discovery

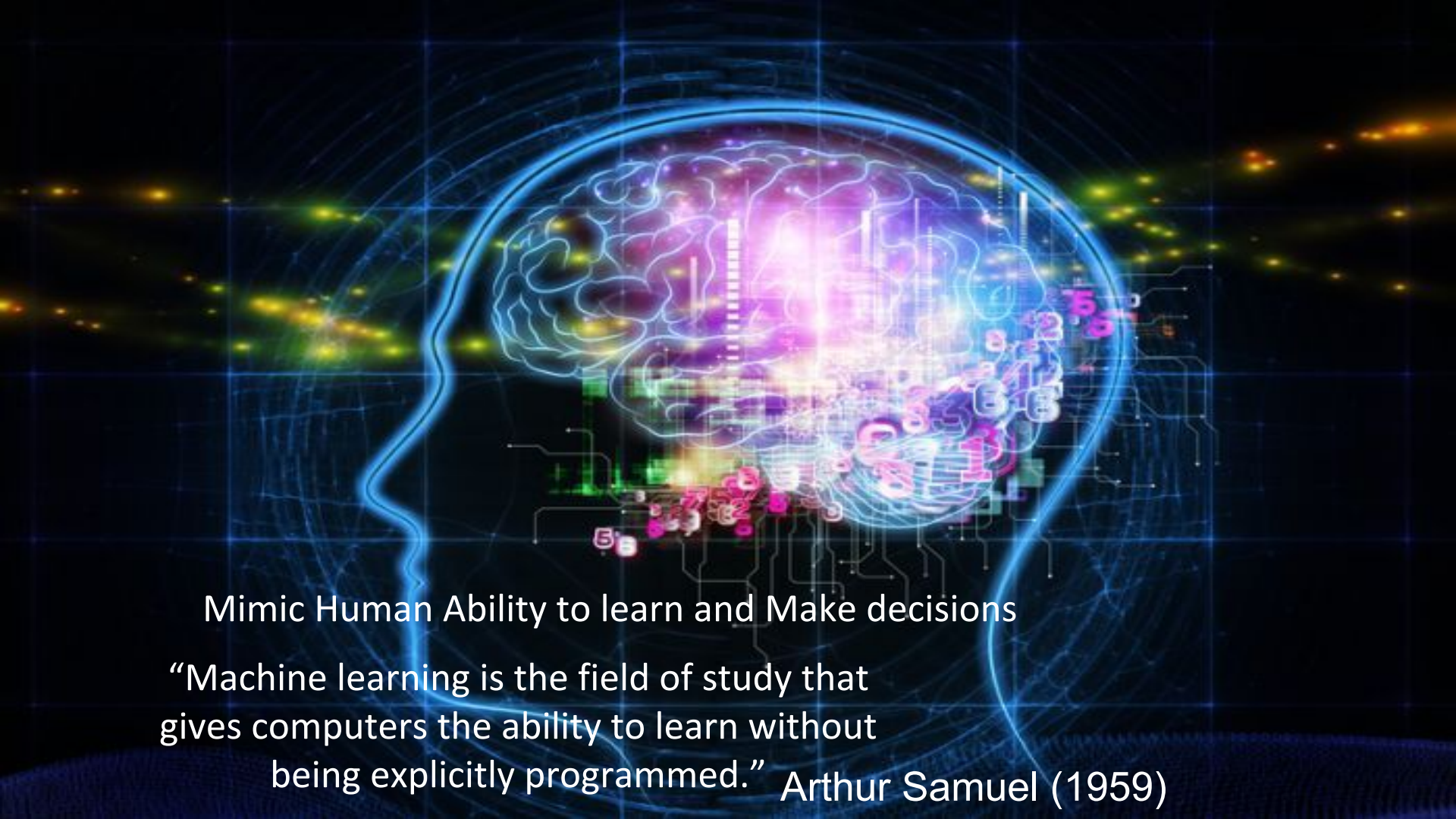
Statistics, Data Mining & Machine Learning



Statistics

Statistic	***Formula***	***Used For***
Sample mean (average)	$\bar{x} = \frac{\sum x}{n}$	Measure of center; affected by outliers
Median	*n* odd: middle value of ordered data *n* even: average of the two middle values	Measure of center; not affected by outliers
Sample standard deviation	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$	Measure of variation; "average" distance from the mean
Correlation coefficient	$r = \frac{1}{n-1} \sum \frac{(x - \bar{x})(y - \bar{y})}{s_x s_y}$	Strength and direction of linear relationship between *X* and *Y*


The Posterior
The Evidence
The Prior
The probability of getting this evidence if this hypothesis were true
The probability of H being true, before gathering evidence
P (H|E) = P (H|E) P(H) / P(E)
The probability that the hypothesis (H) is true given the evidence (E)
The marginal probability of the evidence (Prob of E over all possibilities)



Mimic Human Ability to learn and Make decisions

"Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed." Arthur Samuel (1959)

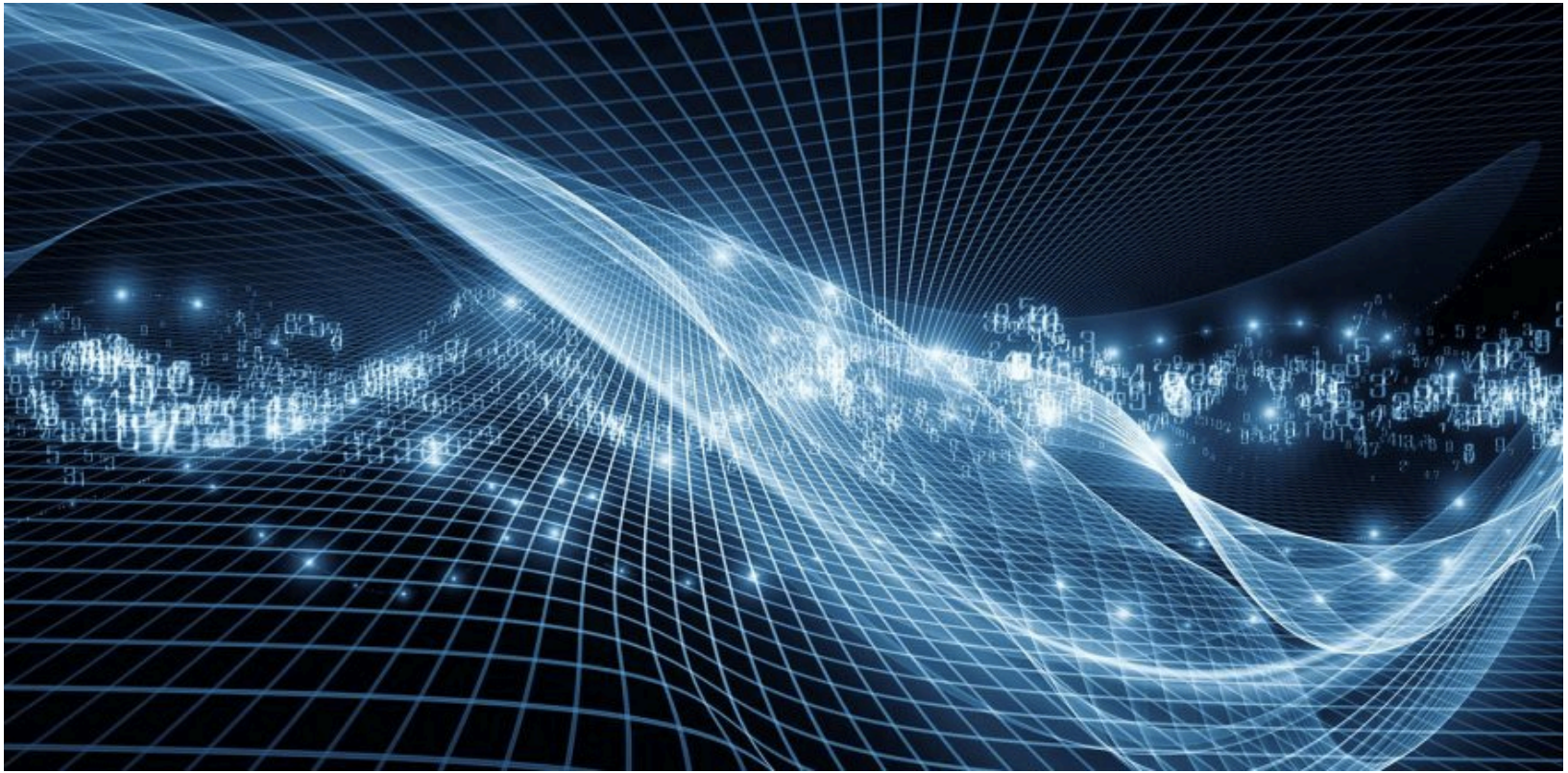
Machine learning:

- **Supervised vs Unsupervised.**
  - ***Supervised learning*** - the presence of the outcome variable is available to guide the learning process.
    - there **must** be a training data set in which the solution is already known.
  - ***Unsupervised learning*** - the outcomes are unknown.
    - cluster the data to reveal meaningful partitions and hierarchies

Why Data Mining?

- Classification
- Clustering
- Associations
- Visualization
- Summarization
- Serendipity …

When is Data enough?



The Spitzer Infrared Nearby Galaxies Survey (SINGS) Hubble Tuning-Fork

The Spitzer Space Telescope observed 75 galaxies as part of its SINGS (Spitzer Infrared Nearby Galaxies Survey) Legacy Program. The galaxies are presented here in a Hubble Tuning-Fork diagram, which groups galaxies according to the morphology of their nuclei and spiral arms. The designation of these galaxies and their placement in the diagram is based on their visible-light appearance. The main goal of the SINGS program is to characterize the infrared properties of a wide range of galaxy types. The images of the galaxies are composites created from data taken by IRAC (the Infrared Array Camera) at 3.6 and 8.0 μm, and MIPS (the Multiband Imaging Photometer for Spitzer) at 24 μm.

The infrared range probed by these and other observations taken for the SINGS project allows for the detailed study of star formation, dust emission, and the distribution of stars in each galaxy. Light from old stars appears as blue in the images, while the lumpy knots of green and red light are produced by dust clouds surrounding newly born stars. The elliptical galaxies on the left are almost entirely made of old stars, while spiral galaxies like our own Milky Way are rich in young stars and the raw materials for future star formation.

More information can be found at:
http://sings.stsci.edu/

Unbarred Spirals

Weak Bulge

Strong Bulge

Intermediate Spirals

Ellipticals

Strong Bulge

Irregulars

Weak Bulge

Barred Spirals

Poster and composite images created from SINGS observations by Karl D. Gordon (Oct 2007)
Blue=IRAC 3.6μm (stars)
Green=IRAC 8μm
(aromatic features from dust grains/molecules)
Red=MIPS 24μm (warm dust)

SPITZER SPACE TELESCOPE

SINGS Team

Robert Kennicutt, Jr. (Principle Investigator), Daniela Calzetti (Deputy Principle Investigator), Charles Engelbracht (Technical Contact), Lee Armus, George Bendo, Caroline Bot, Brent Buckalew, John Cannon, Daniel Dale, Bruce Draine, Karl Gordon, Albert Grauer, David Hollenbach, Tom Jarrett, Lisa Kewley, Claus Leitherer, Aigen Li, Sangeeta Malhotra, Martin Meyer, John Moustakas, Eric Murphy, Michael Regan, George Rieke, Marcia Rieke, Helene Roussel, Kartik Sheth, J.D. Smith, Michele Thornley, Fabian Walter & George Helou

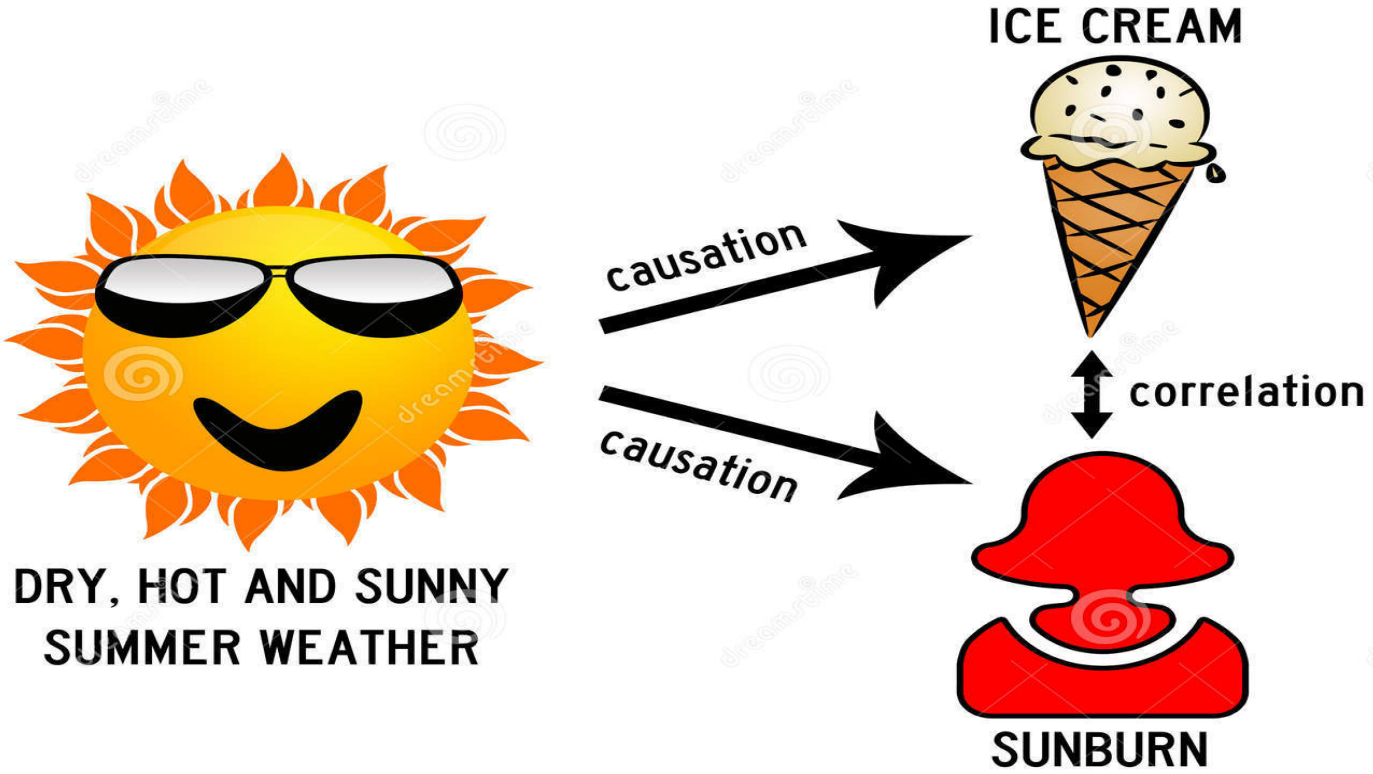
Clustering:

- Clustering is the task of gathering samples into groups of similar samples according to some predefined similarity or dissimilarity measure

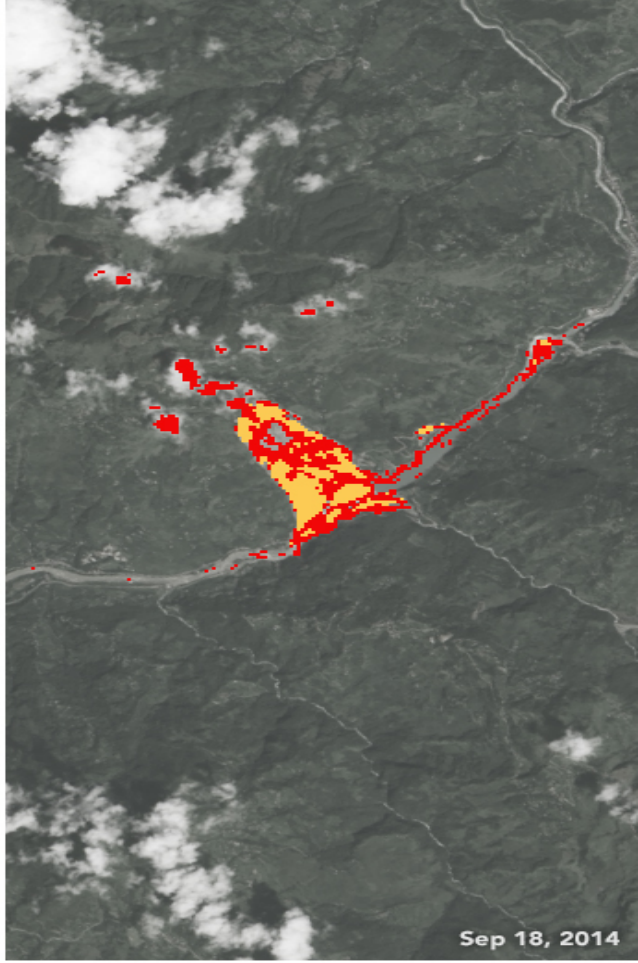
sample

Cluster/group

Associations, A & B & C occur frequently



Summarization



Landslide Detection

Possible

Probable

Project

Where do I get data?

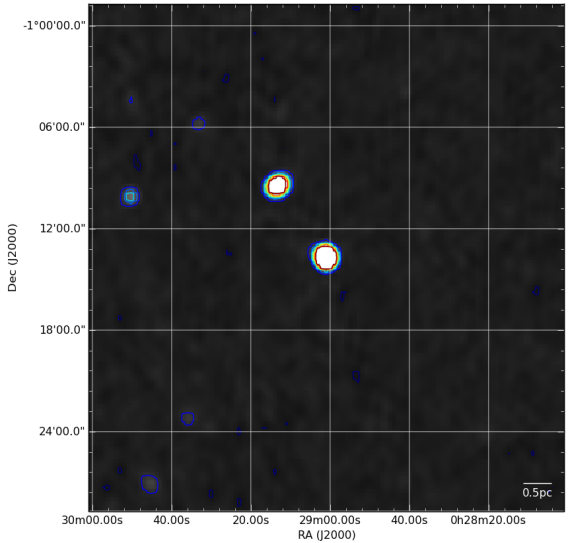
- IVOA
  - VO
- VLA, **Vizier**, NED, …
  - Extract all AGN from Veron paper within 0< dec < -90
  - Plot the redshift distribution of the sources you have obtained

Use Case

Morphology
of
Active Galactic Nuclei

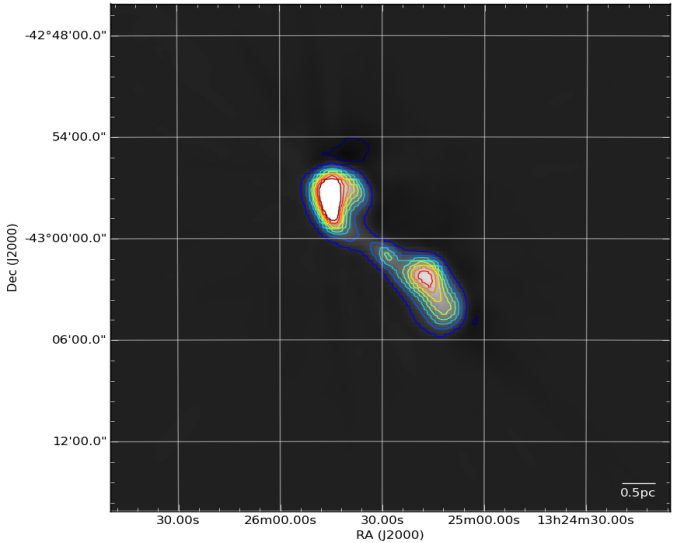
Radio Sources

Point Sources



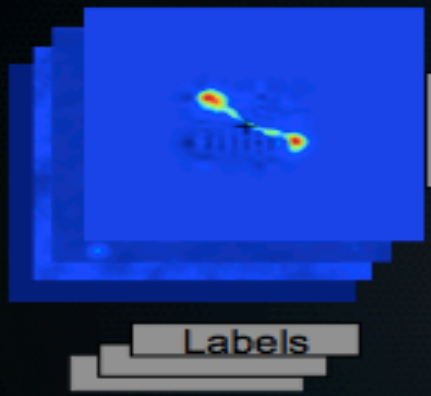
PKS 0026-014

Extended Sources



Centaurus A

ML Classification

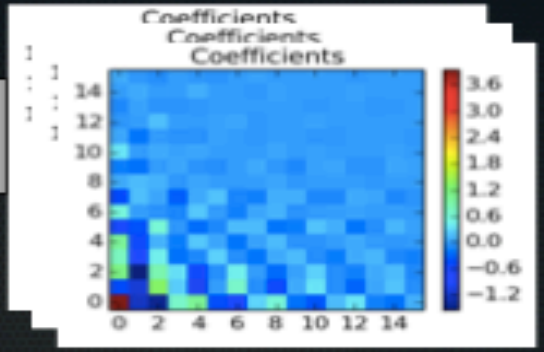


Labels

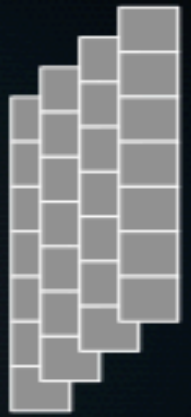
Shapelets Decomposition

Feature Extraction

Shapelet Matrices

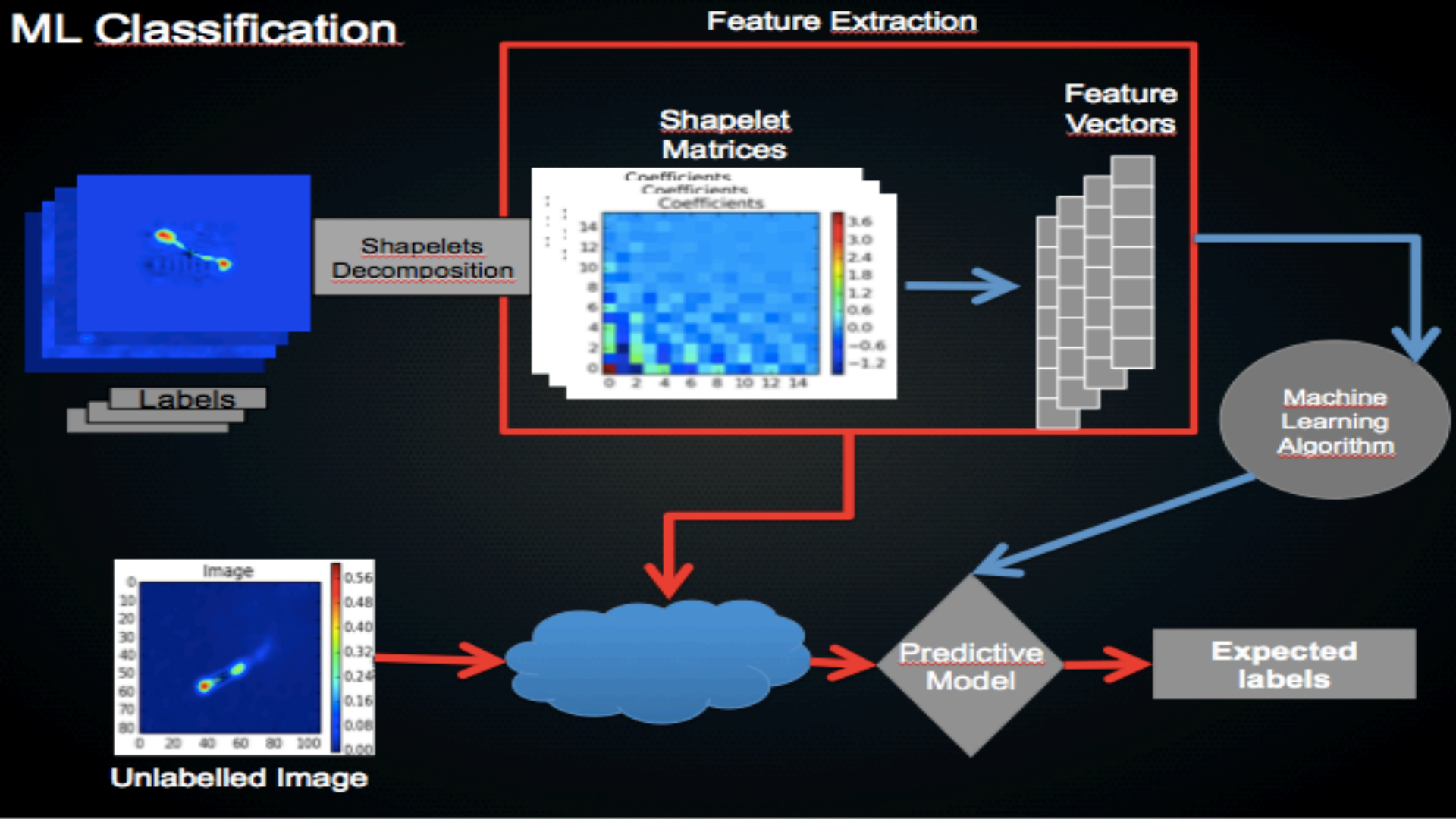


Feature Vectors



$$\mathbf{S}_{(i,\theta)} = \begin{bmatrix} f_{(0,0)} & f_{(0,1)} & f_{(0,2)} & \cdots & f_{(0,15)} \\ f_{(1,0)} & f_{(1,1)} & f_{(1,2)} & \cdots & f_{(1,15)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{(15,0)} & f_{(15,1)} & f_{(15,2)} & \cdots & f_{(15,15)} \end{bmatrix}$$

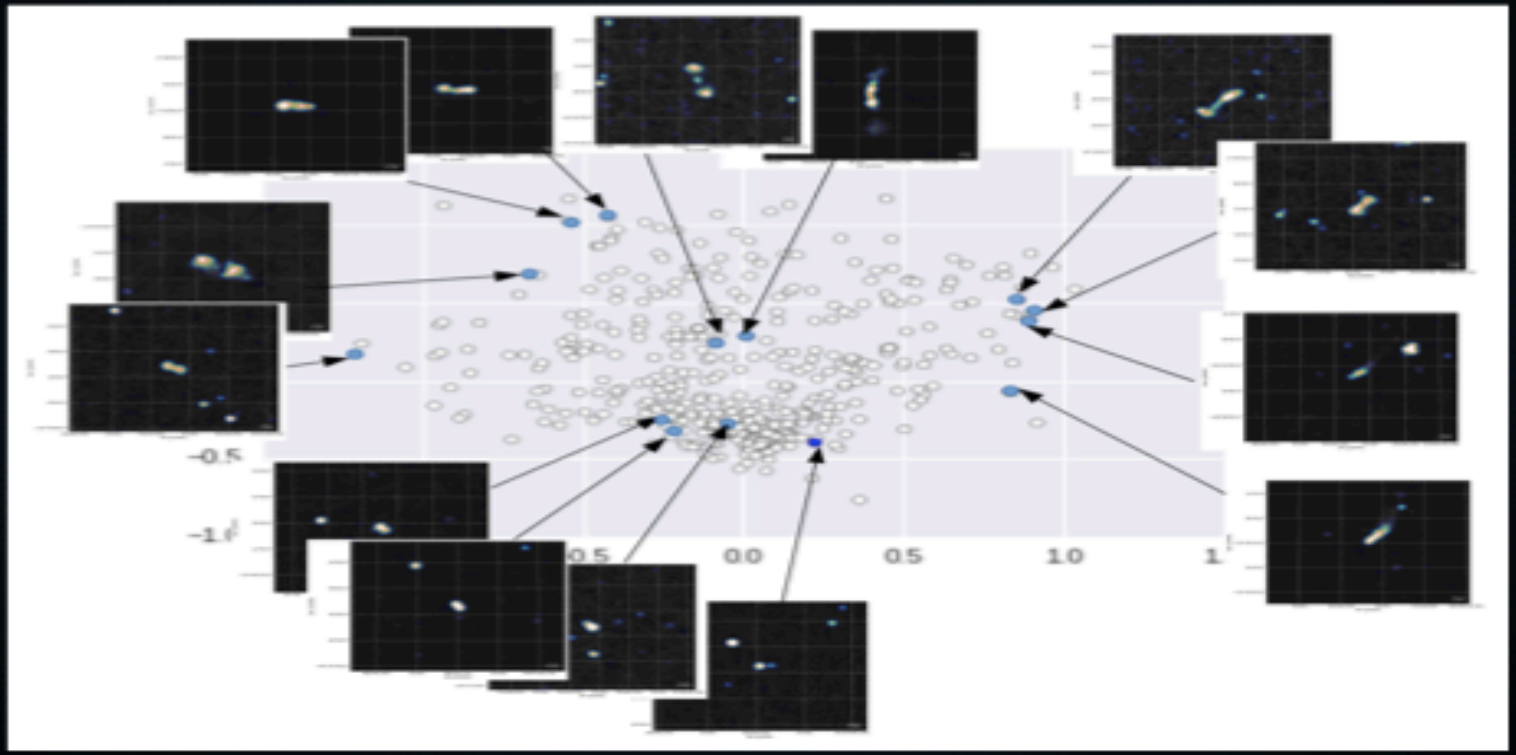
$$\hat{\mathbf{x}_{(i,\theta)}} = \frac{\mathbf{x}_{(i,\theta)}}{|\mathbf{x}_{(i,\theta)}|} = \begin{bmatrix} g_{(0,0)} \\ \vdots \\ g_{(0,15)} \\ \vdots \\ \vdots \\ g_{(15,0)} \\ \vdots \\ g_{(15,15)} \end{bmatrix}$$


ML Classification
Feature Extraction
Shapelet Matrices
Feature Vectors
Shapelets Decomposition
Labels
Machine Learning Algorithm
Predictive Model
Expected labels
Image
Unlabelled Image

Feature Visualisation: isomap

Angular Variation

Triple Gaussian



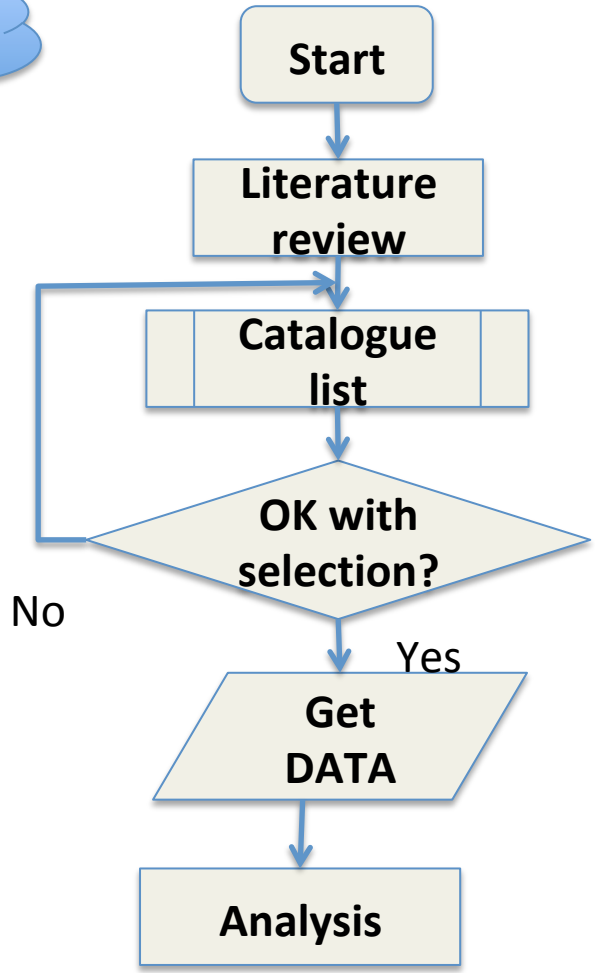
Single Gaussian

Elliptical

Use Case

- Morphology of AGN from FIRST
- Mine archive for list of AGN
- Images
  - Radio
  - Optical
  - Overlay
- Summary

Morphology of AGN



Data Mining

Virtual Observatory

Toolbox, e.g. Python

Science