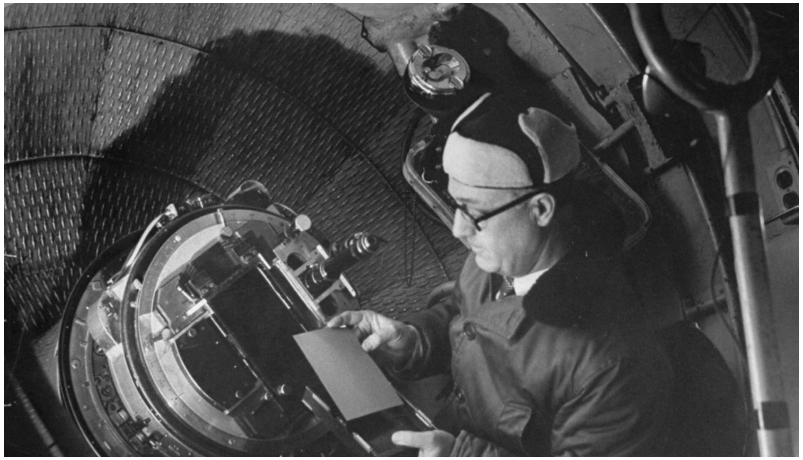
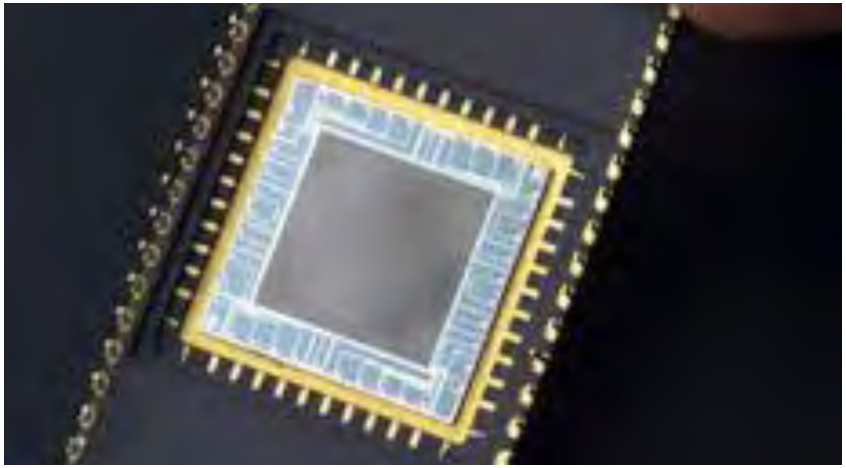
# KNOWLEDGE GAIN IN THE AGE OF HPC AND BIG DATA

SUSANNE PFALZNER

JÜLICH
Forschungszentrum

# SECOND BIG DATA CHALLENGE



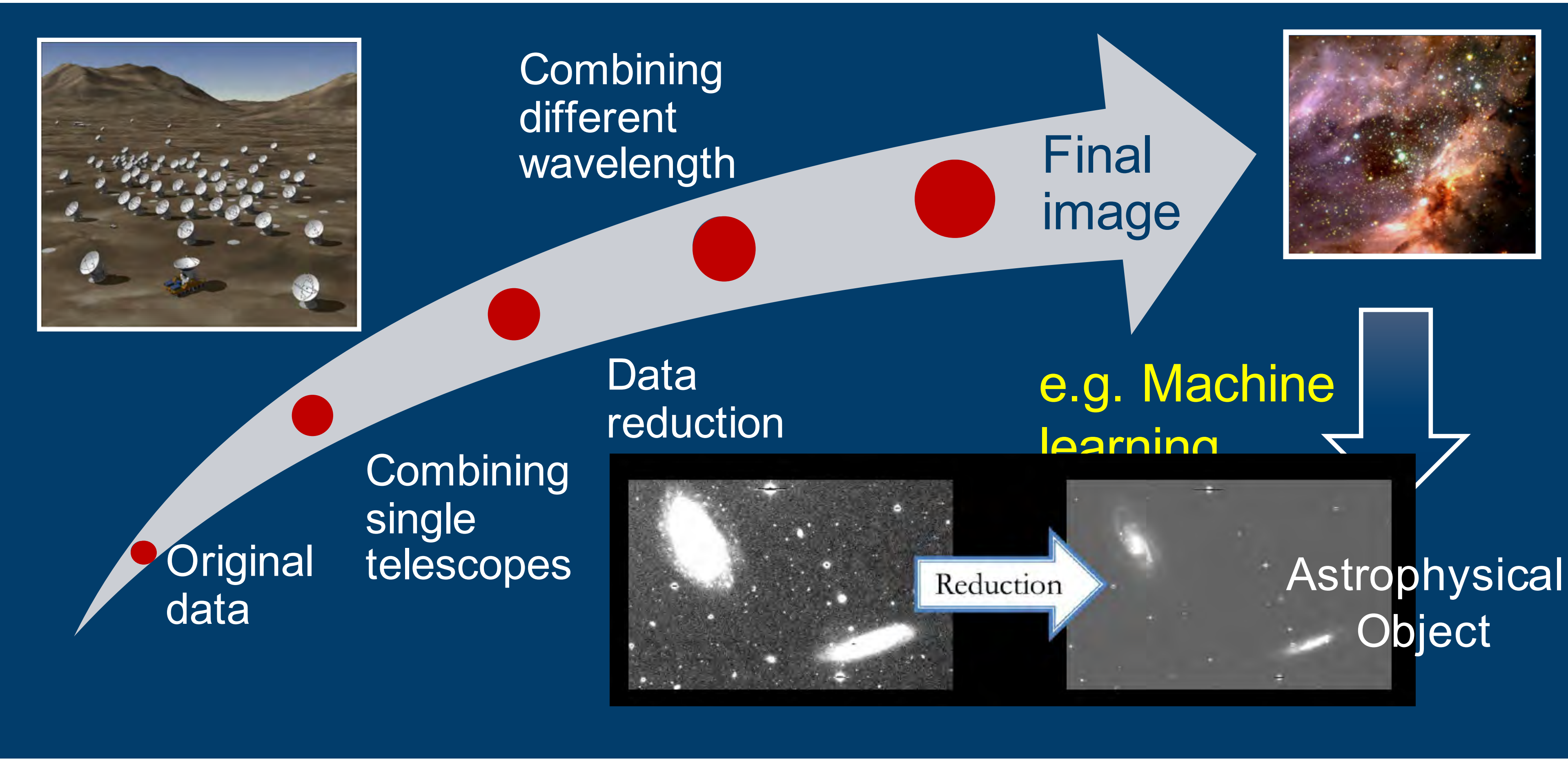


Efficiency around 1%
most of the photons lost!

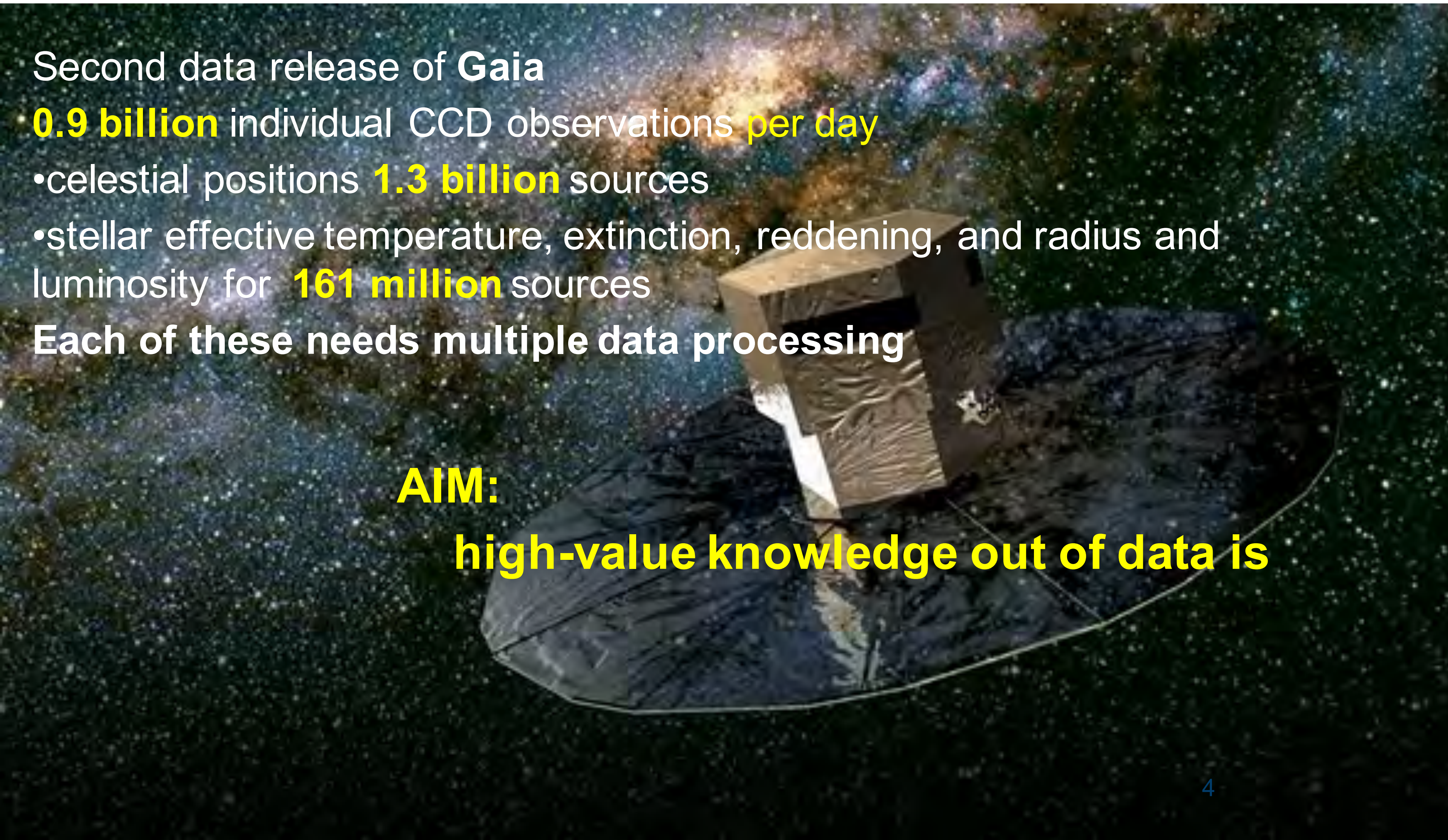
- Extremely high **efficiency** at redder wavebands -- almost 100%!
- Limiting magnitudes increased by four to five magnitudes!

**Output is digital**

# DATAWORKFLOWS



# DATA FLOOD



Second data release of **Gaia**

**0.9 billion** individual CCD observations per day

•celestial positions **1.3 billion** sources

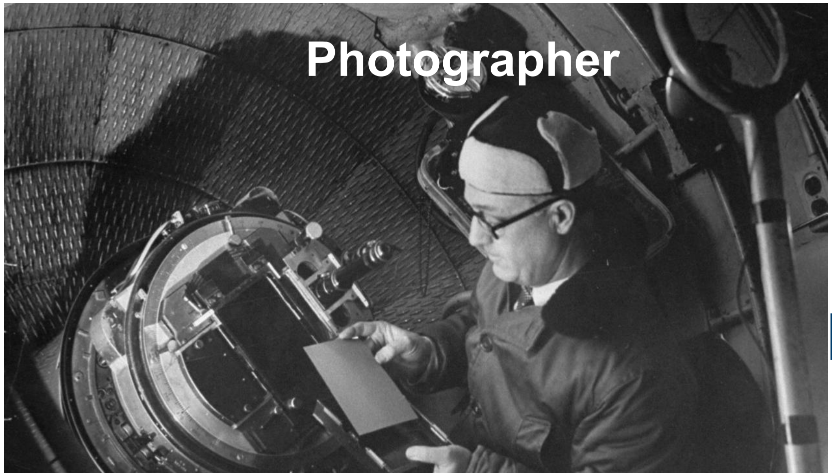
•stellar effective temperature, extinction, reddening, and radius and luminosity for **161 million** sources

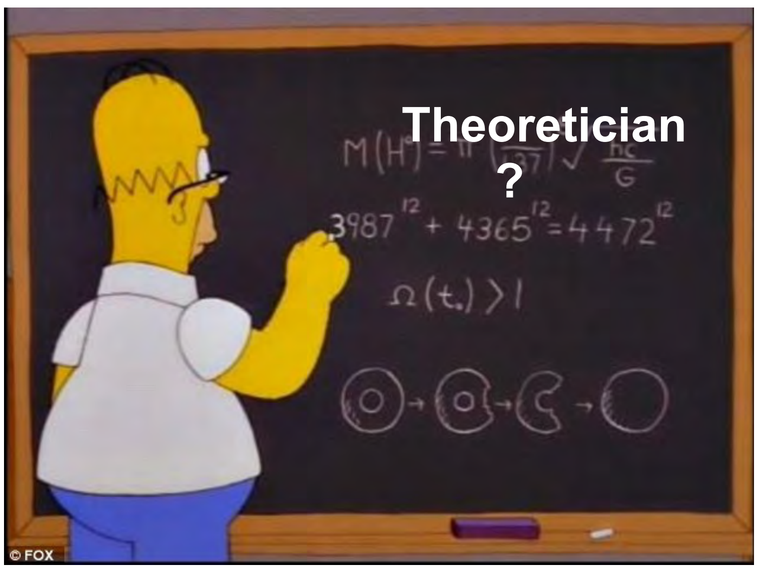
**Each of these needs multiple data processing**

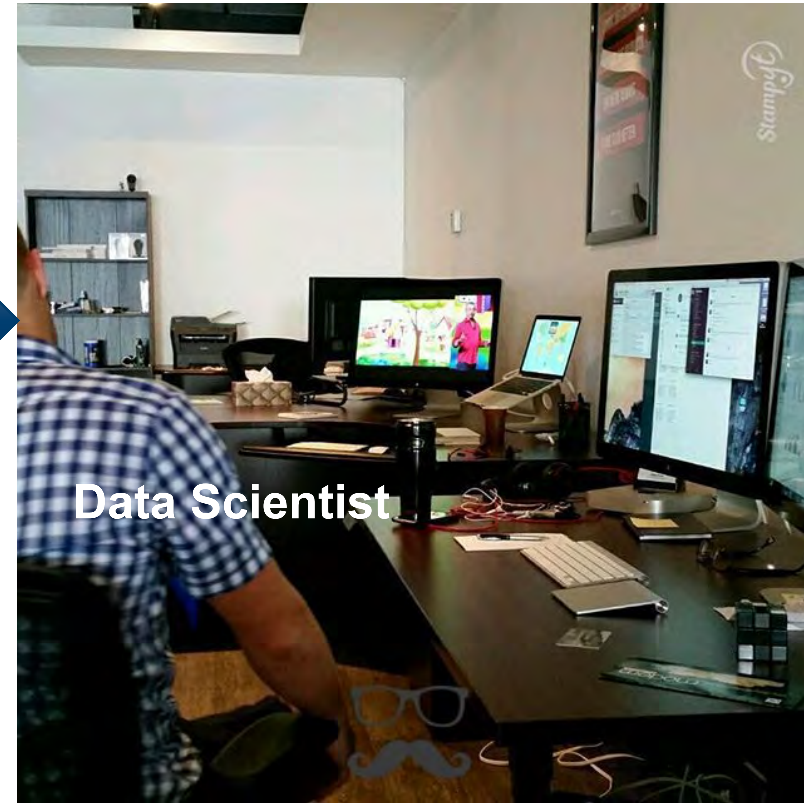
**AIM:**

**high-value knowledge out of data is**

# OBSERVER


Photographer


Theoretician
?
3987^12 + 4365^12 = 4472^12
Ω(t₀) > 1
© FOX


Data Scientist

# ASTROPHYSICAL SIMULATIONS

**Observations** provide only **snapshots** at a certain moment in time

**Theory** (simulation) must create **time sequence**

Several theories –
**Which is the right one?**

**Predictions** from theory **tested by observations**

# ASTROPHYSICAL SIMULATIONS

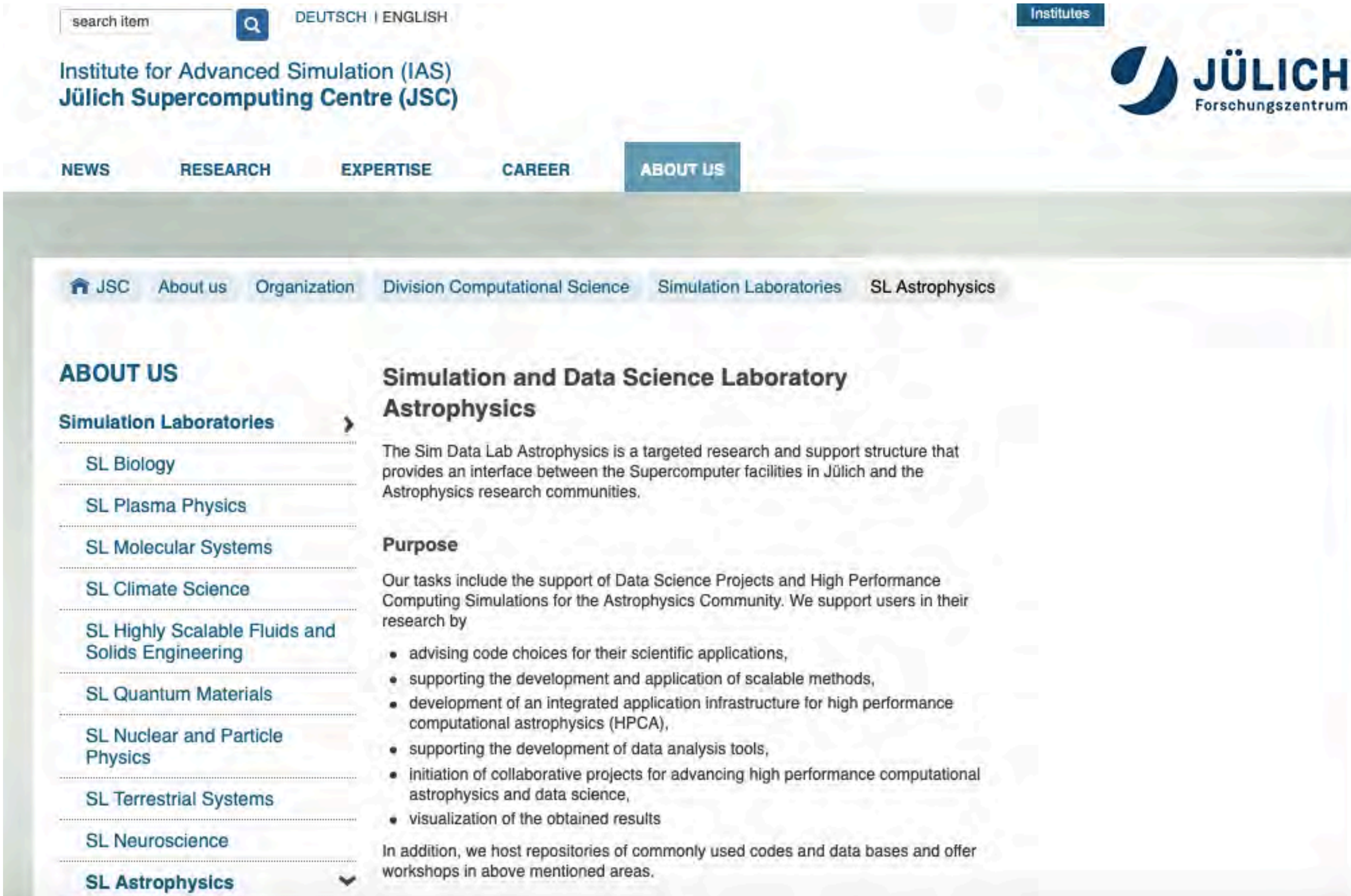
Challenges

- steep spatial gradients, complex geometries, etc.
  (1 AU – 20 000 000 AU)

But also,

- often very different time scales
  years – several million years (Myr)
- No direct comparison with experiment

# SIM AND DATA LAB ASTRO



search item

DEUTSCH | ENGLISH

Institutes

Institute for Advanced Simulation (IAS)
Jülich Supercomputing Centre (JSC)

JÜLICH
Forschungszentrum

NEWS RESEARCH EXPERTISE CAREER ABOUT US

JSC About us Organization Division Computational Science Simulation Laboratories SL Astrophysics

## ABOUT US

**Simulation Laboratories**

- SL Biology
- SL Plasma Physics
- SL Molecular Systems
- SL Climate Science
- SL Highly Scalable Fluids and Solids Engineering
- SL Quantum Materials
- SL Nuclear and Particle Physics
- SL Terrestrial Systems
- SL Neuroscience
- **SL Astrophysics**

## Simulation and Data Science Laboratory Astrophysics

The Sim Data Lab Astrophysics is a targeted research and support structure that provides an interface between the Supercomputer facilities in Jülich and the Astrophysics research communities.

### Purpose

Our tasks include the support of Data Science Projects and High Performance Computing Simulations for the Astrophysics Community. We support users in their research by

- advising code choices for their scientific applications,
- supporting the development and application of scalable methods,
- development of an integrated application infrastructure for high performance computational astrophysics (HPCA),
- supporting the development of data analysis tools,
- initiation of collaborative projects for advancing high performance computational astrophysics and data science,
- visualization of the obtained results

In addition, we host repositories of commonly used codes and data bases and offer workshops in above mentioned areas.

# OBSERVATIONS VS. SIMULATIONS

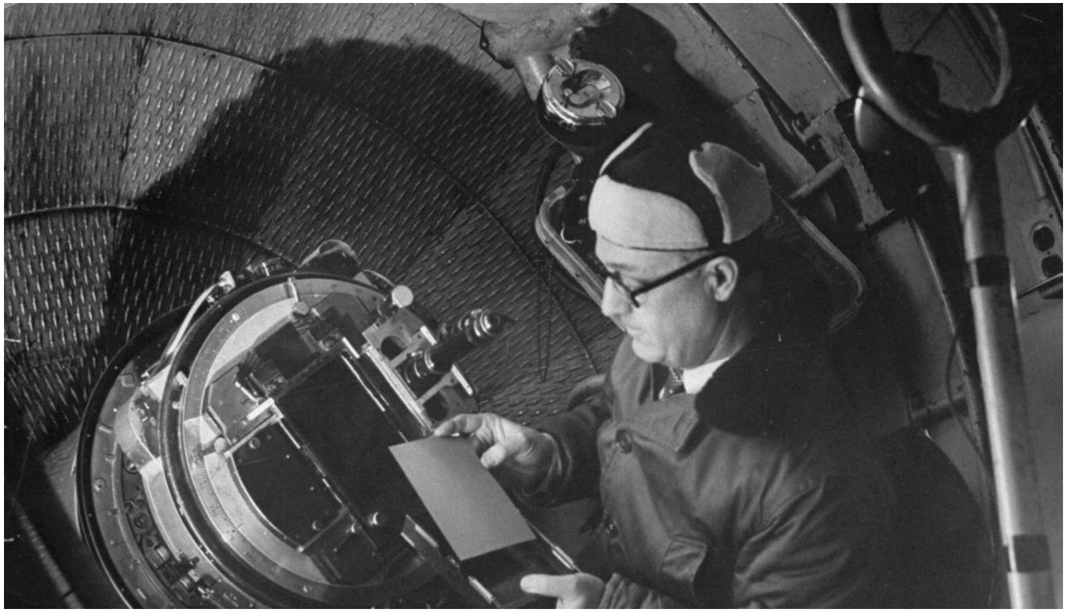
**Comparison between TWO SIMULATIONS**

observational data

theoretical models

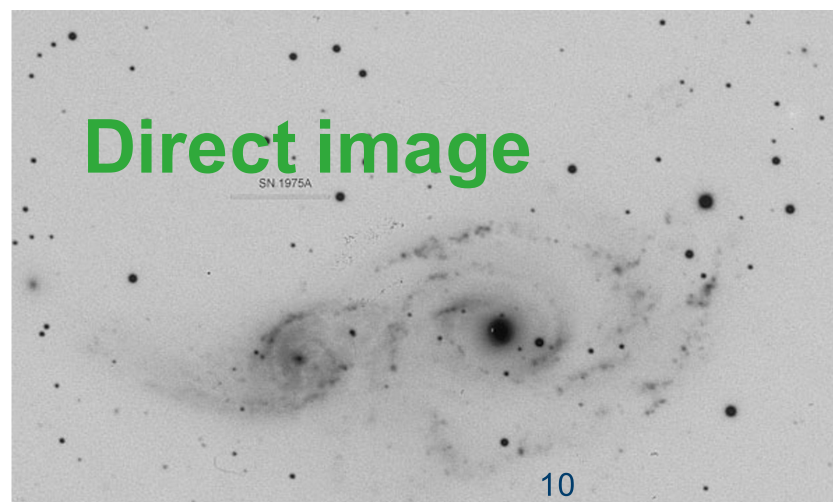
**„It looks the same", is not enough!**

# FIRST BIG DATA CHALLENGE



**Photographic plates**

First objective, permanent record of astronomical phenomena



# FIRST COMPUTERS

...the amount of astronomical data was surpassing the capacity of the Observatories to process it

Pickering and his Computers standing in front of Building C at the Harvard College Observatory, 13 May 1913

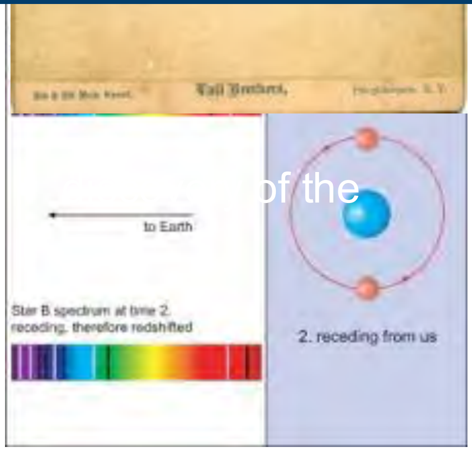
# BEYOND PROCESSING DATA ...

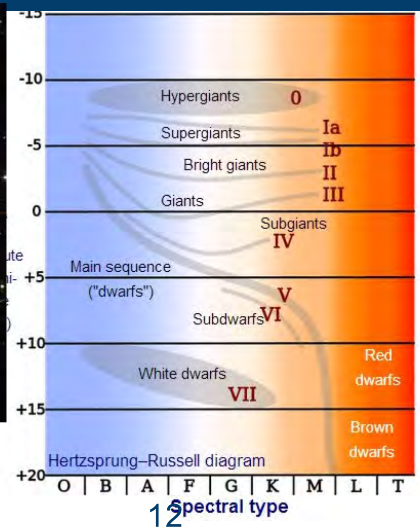
**Antonia Maury**
**First spetroscopic binary star**

**Williamina Fleming**
**Horseshoe Nebula**

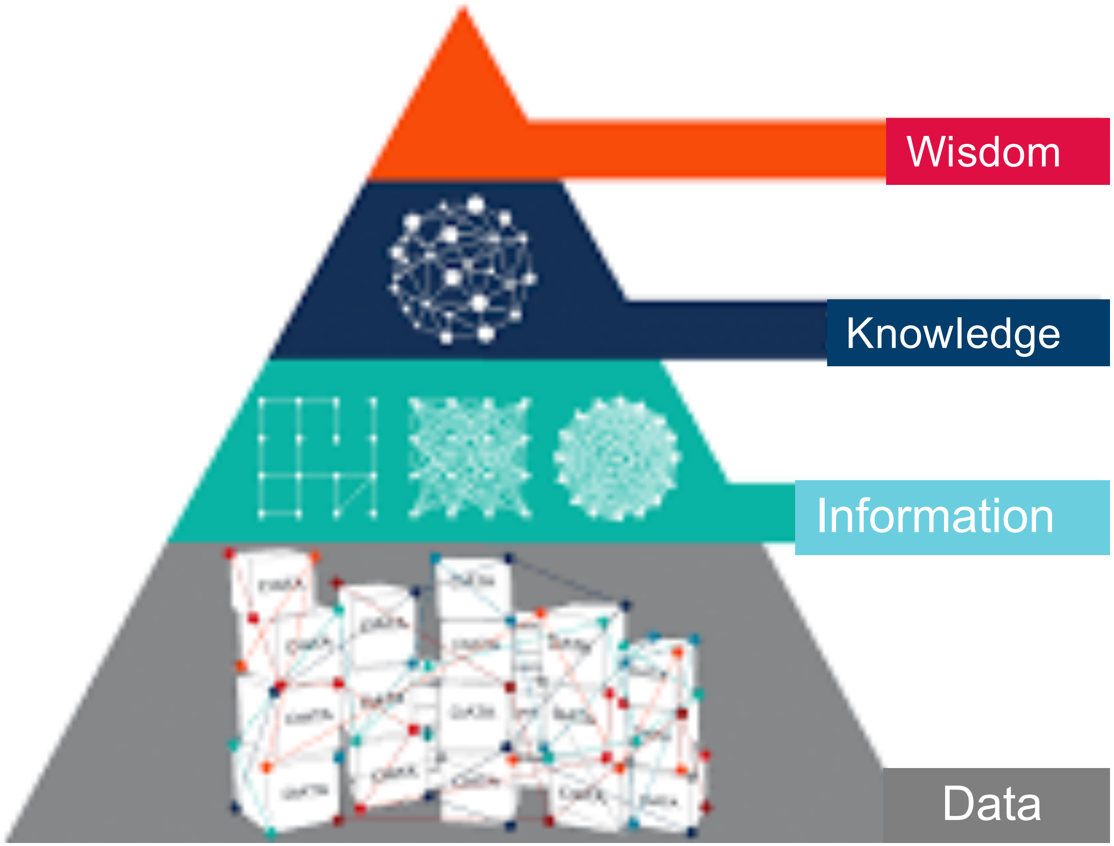
**Henrietta Swan Leavitt**
**Luminosity in Chephides**

**Annie Jump Cannon**
**Stellar Classification**



# THE STEEP CLIMB TO KNOWLEDGE




**Each step up** the pyramid answers questions about and **adds value** to the initial data

# IT IS ALL IN THE DATA, IS IT REALLY?


**Facts**

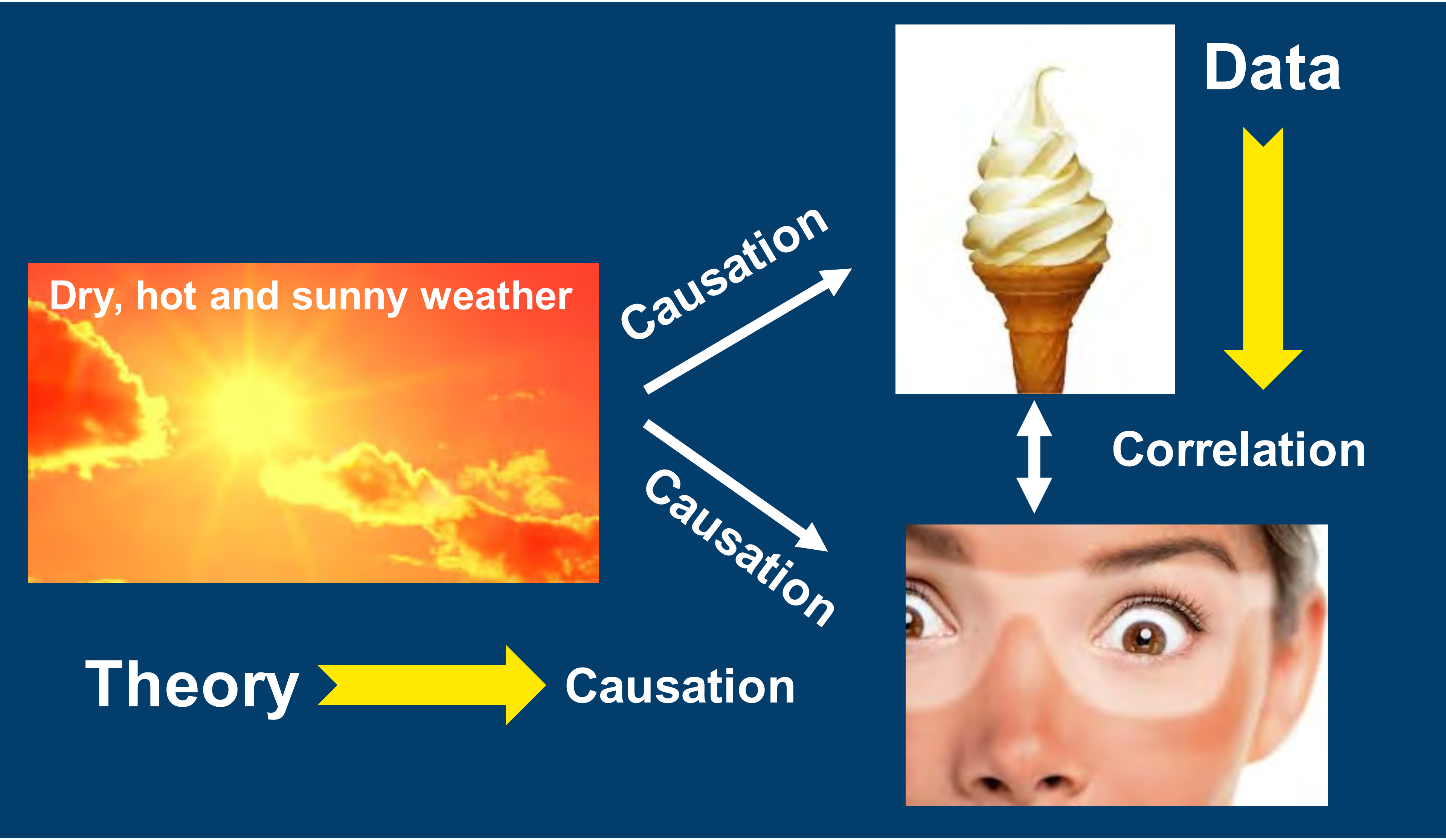
**Correlations**

?

Causal Connection

Knowledge

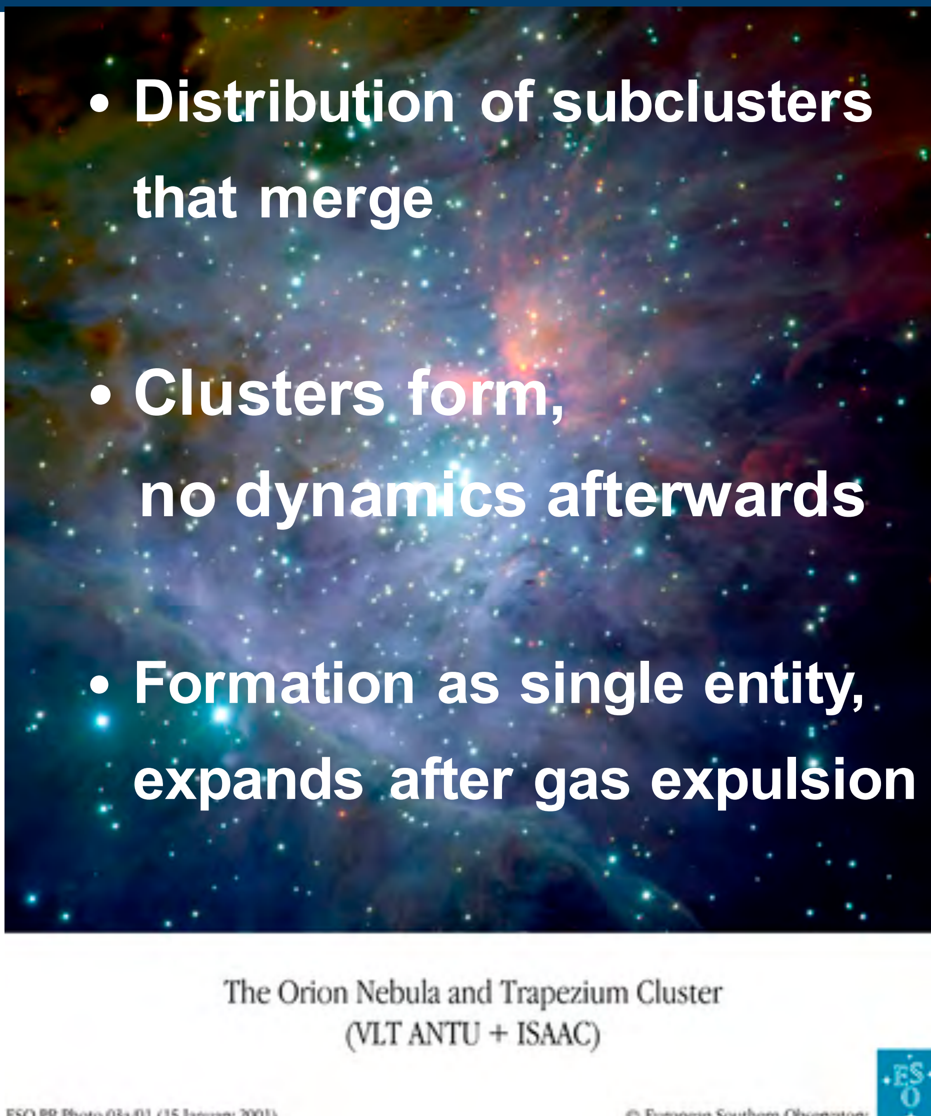
# CORRELATIONS VS CAUSAL CONNECTION



# EXAMPLE: DATA ANALYSIS

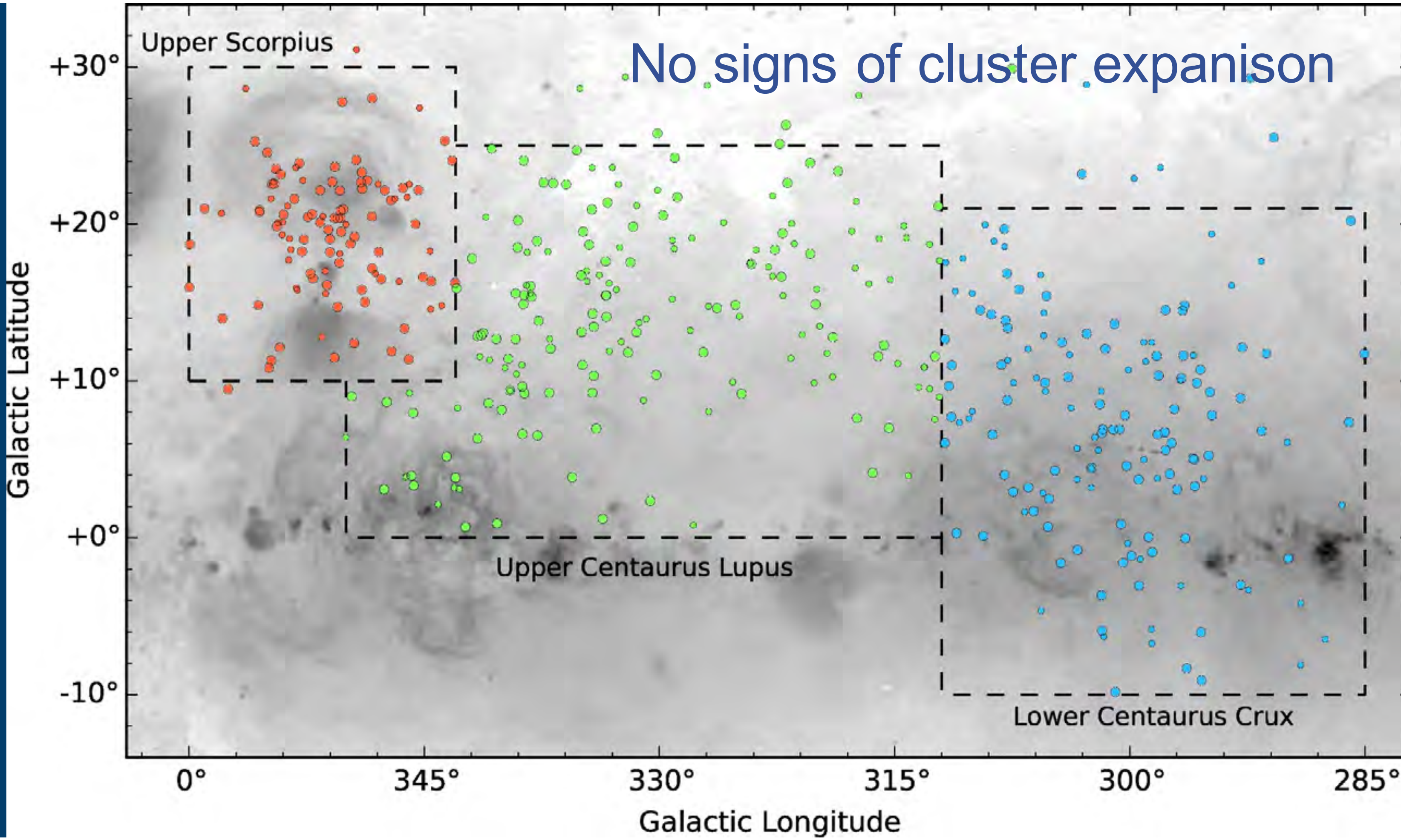
# EXAMPLE:
# STAR CLUSTER FORMATION

- Distribution of subclusters that merge
- Clusters form, no dynamics afterwards
- Formation as single entity, expands after gas expulsion

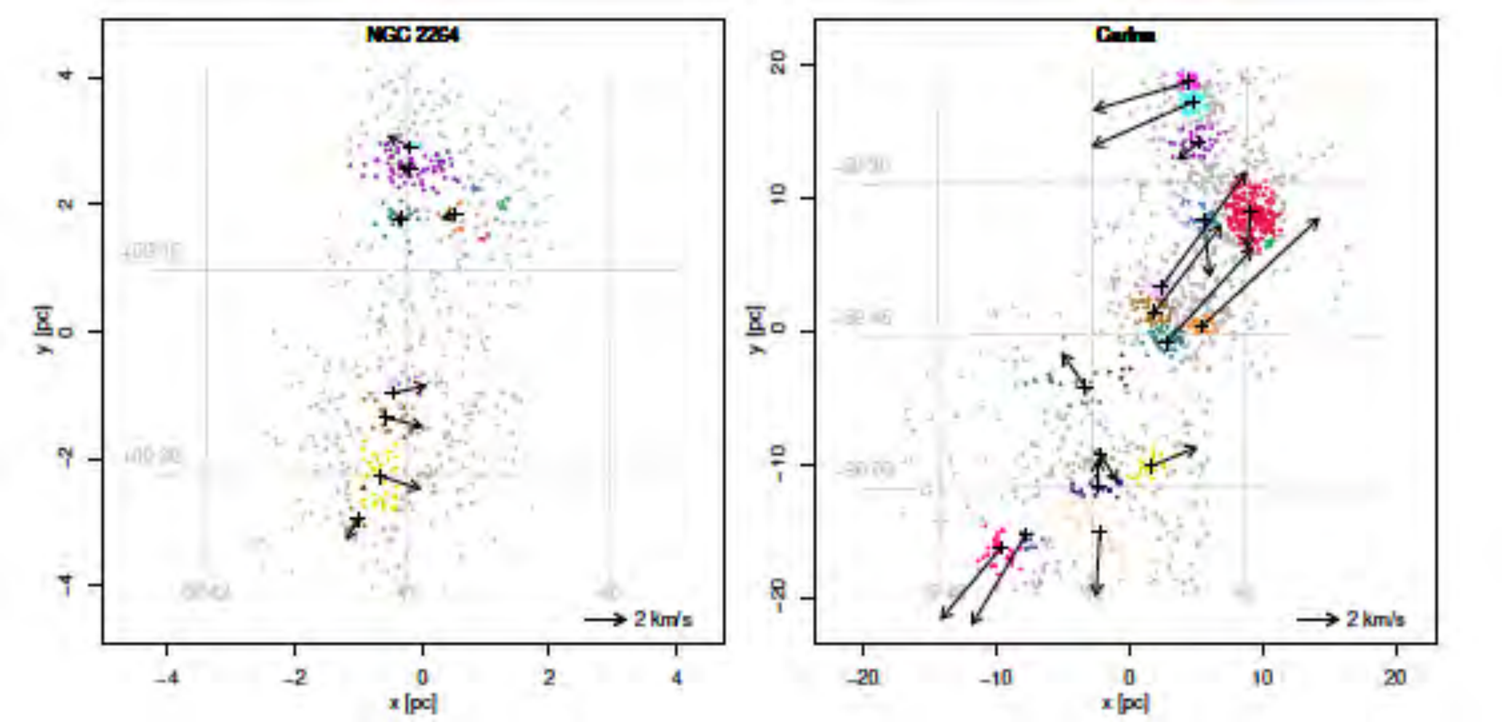



JÜLICH
Forschungszentrum

# ANALYSIS OF GAIA DATA



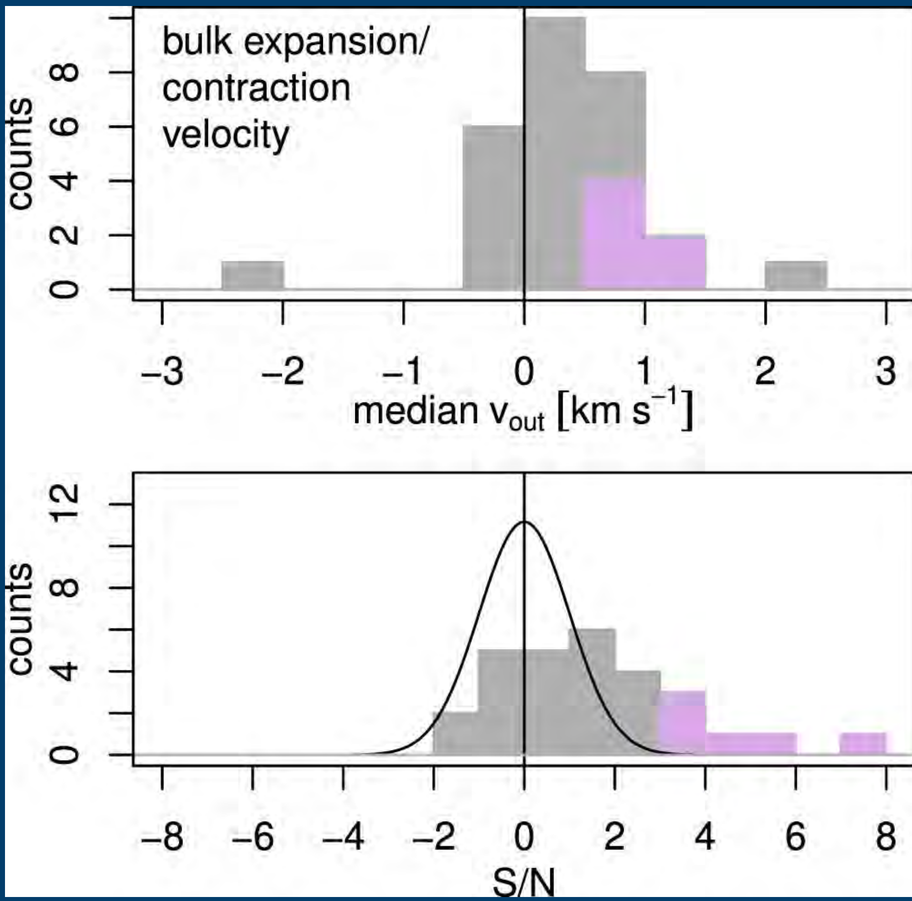
# OVER-COMING ONE'S OWN EXPACTATION




**No signs of subcluster merging**

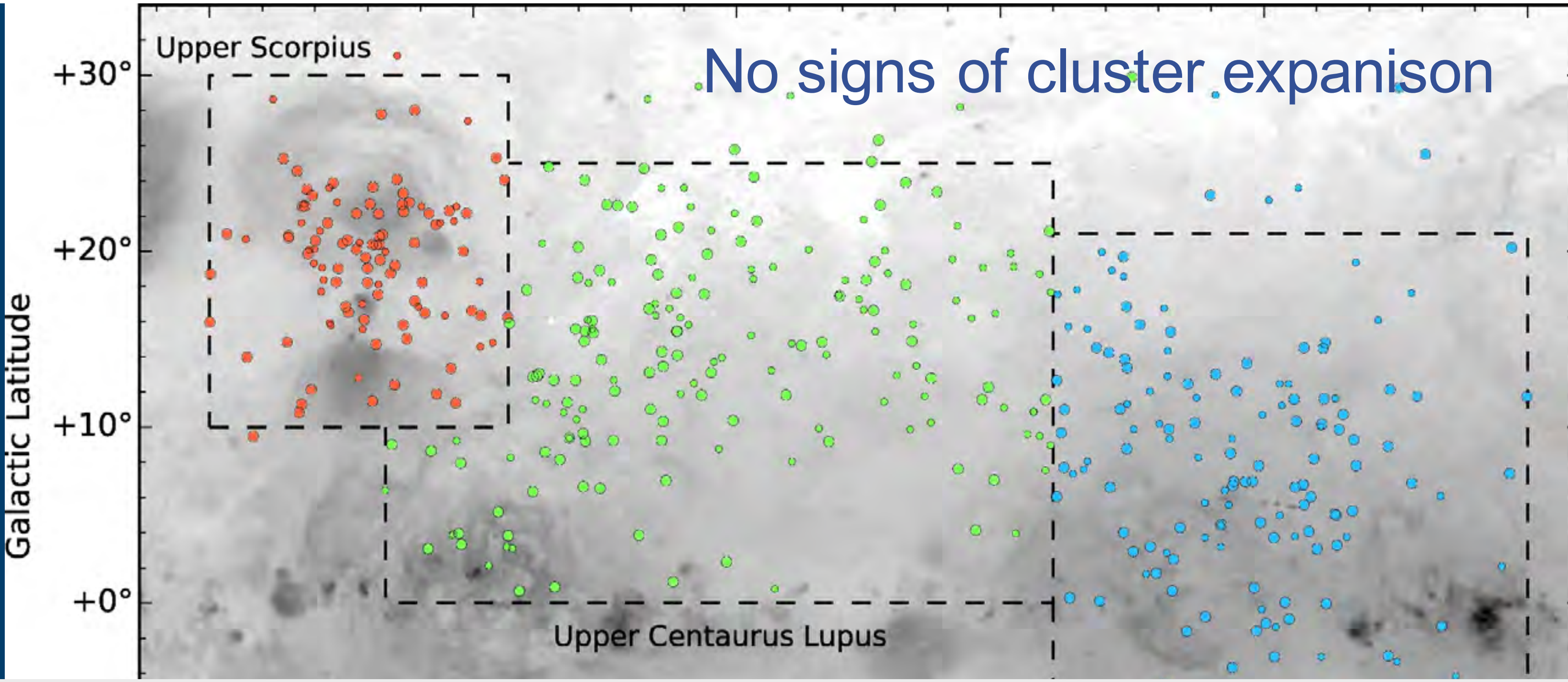
Kuhn et al. arXiv:1807.06085

# OVER-COMING ONE‘S OWN EXPACTATION




velocity dispersion in the individual clusters shows they expand

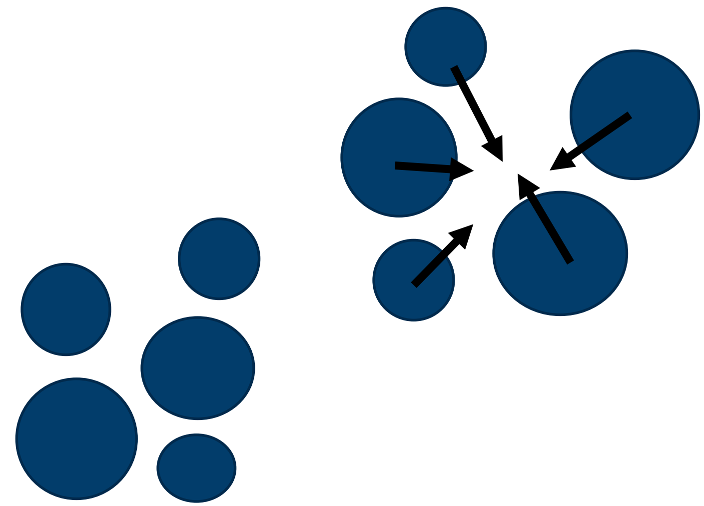
Kuhn, M. et al. 2019




# PROBLEM:
# MODEL ASSUMPTIONS IN DATA ANALYSIS

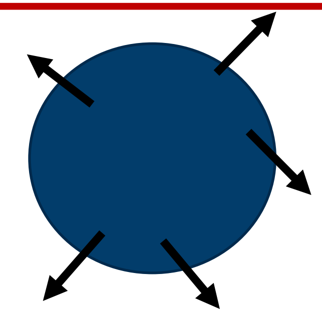
# EXAMPLE:
# STAR CLUSTER FORMATION

- Distribution of subclusters that merge



- Clusters form,
  no dynamics afterwards

- Formation as single entity,
  expands after gas expulsion



# LESSON 1:

MACHINE LEARNING DOES NOT AVOID
HUMAN MISTAKES, BUT CAN HARDWIRES THEM

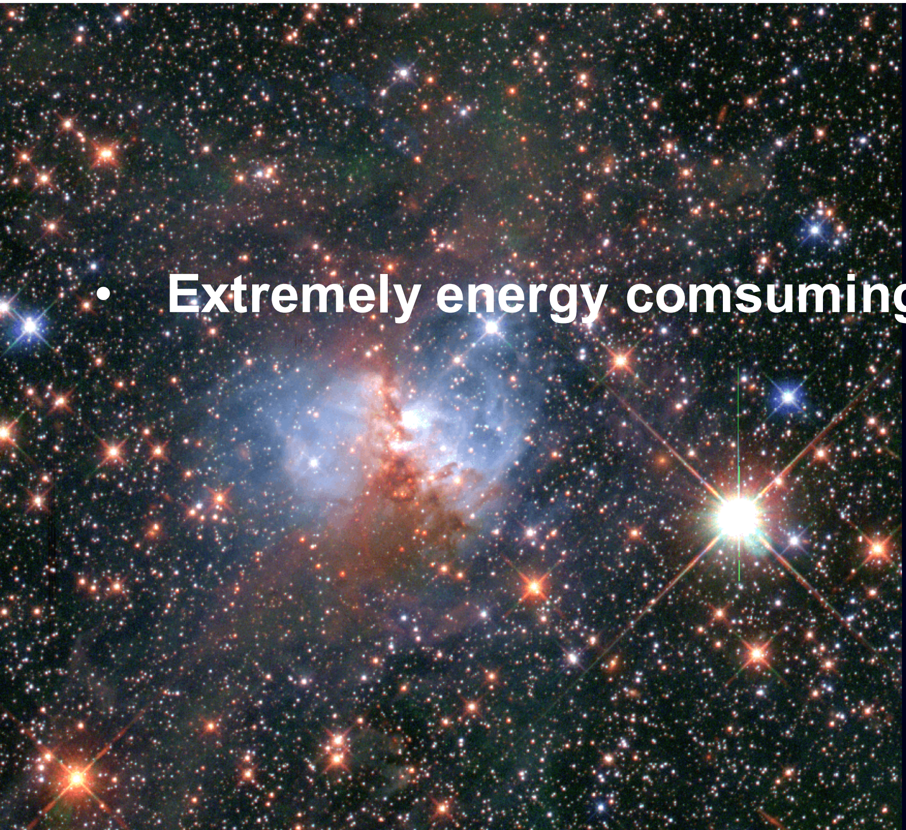
MODEL ASSUMPTIONS
CAN CORRUPT DATA ANALYSIS

DATA CHALLENGE REQUIRES
LEARNING FROM OUR MISTAKES NOW

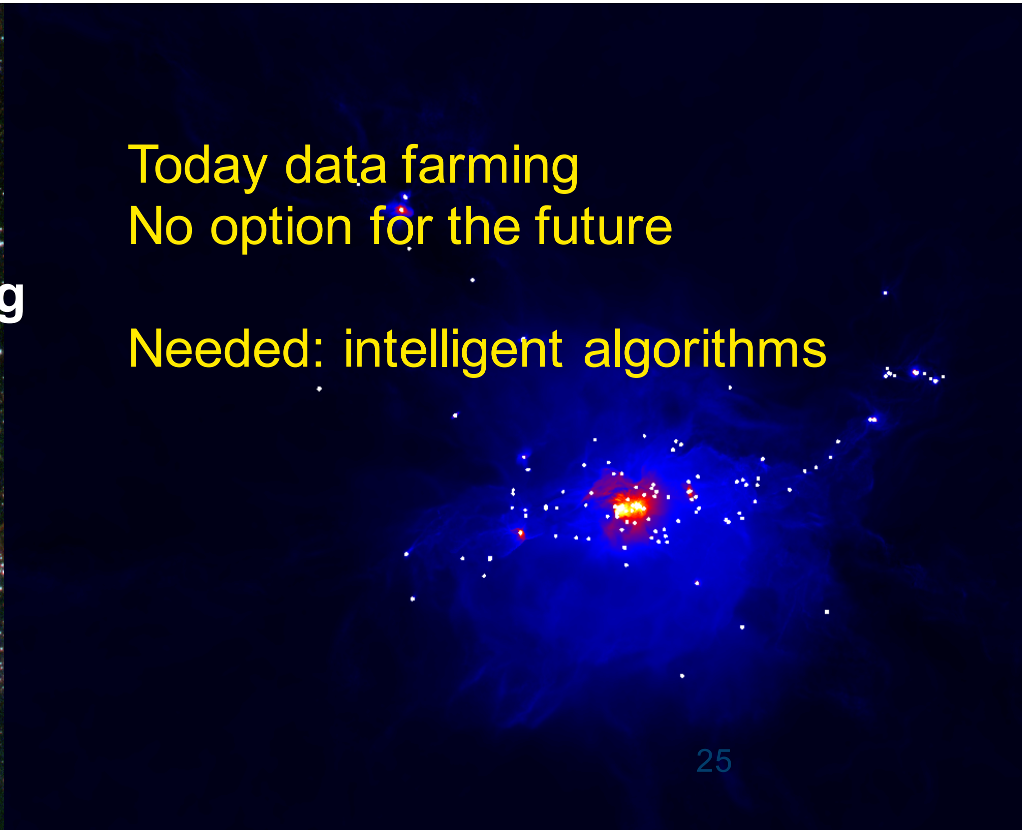
OVERCOMING ONES PREJUDICES PAYS

# POTENTIAL PROBLEMS IN CREATING THE DATA

- Model assumption transferred to data analysis
- Progress limited to more of the same
- Correlation vs causation



- Extremely energy comsuming



Today data farming
No option for the future

Needed: intelligent algorithms

# BEYOND PROCESSING DATA ...

Antonia Maury
First spetroscopic binary star

Williamina Fleming
Horseshoe Nebula

Henrietta Swan Leavitt
Luminosity in Chephides

Annie Jump Cannon
Stellar Classification

Unlike human computers, digital computers are (still) unable to

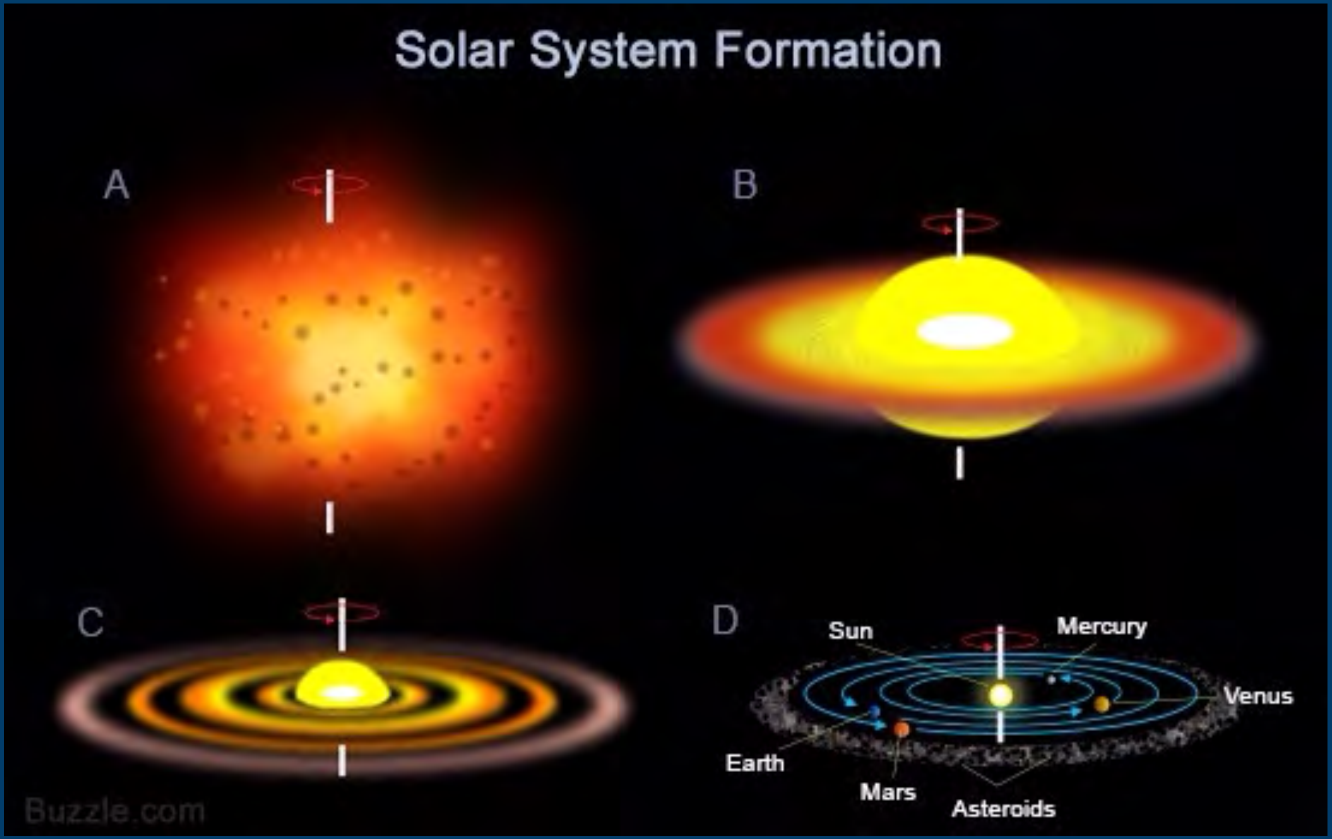
- ask for causation → No model development
- see the unexpected → Risk: more of the same

# EXAMPLE: SIMULATIONS

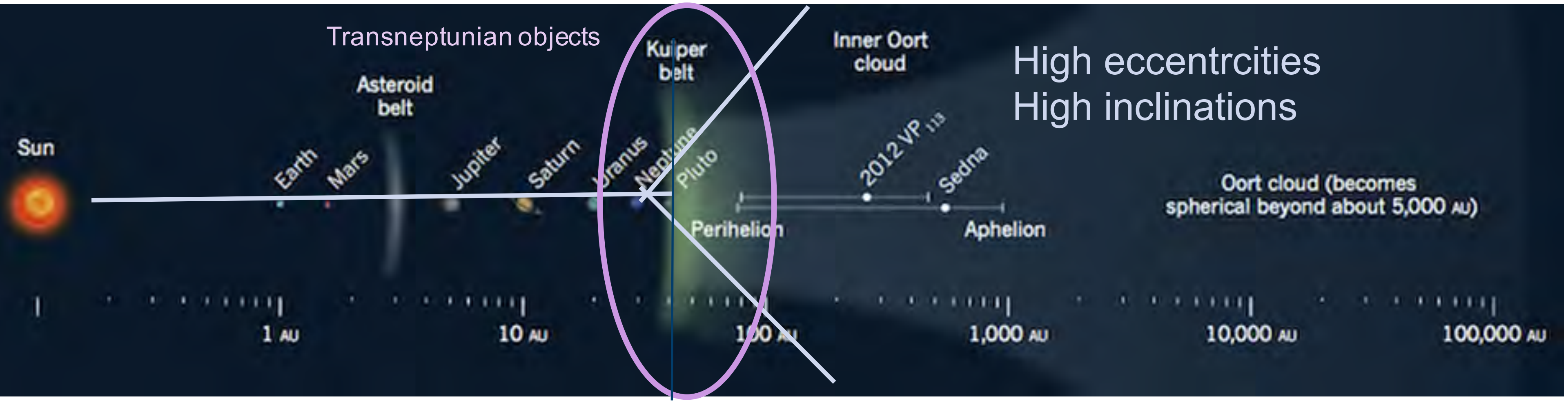
# CHARACTERISTICS OF A GOOD ANALYTICAL MODEL

- Matches observation
- Reproducible
- Simple
- Fit into the general theory
- Falsifiable
- Makes prediction

# FORMATION OF THE SOLAR SYSTEM



# EXAMPLE:
# SOLAR SYSTEM FORMATION

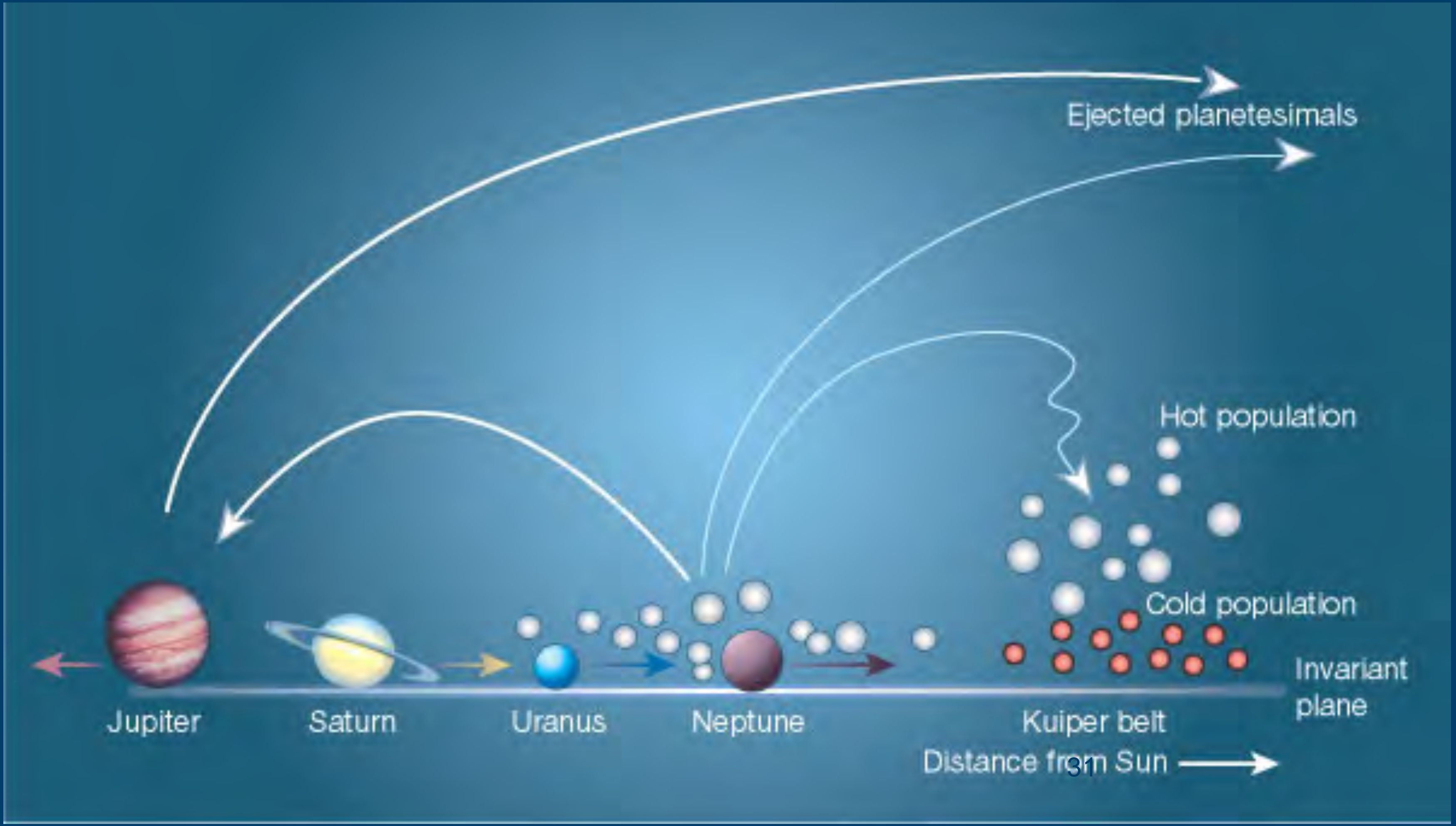



## …, BUT BEYOND NEPTUNE THINGS ARE DIFFERENT

- Cut-off in mass beyond Neptune: 1000 fewer objects than expected
- Most objects have high inclinations, eccentric orbits

# POPULAR EXPLANATION: NICE MODEL

- TNOs were originally between Saturn and Neptune
- Scattered outwards due to movement of planets



# NICE MODEL CAN EXPLAIN

1. Late-Heavy bombardment
2. Hot Kuiper belt (90%) very well,
3. Families of Asteroids
4. Low mass of Mars
5. etc.

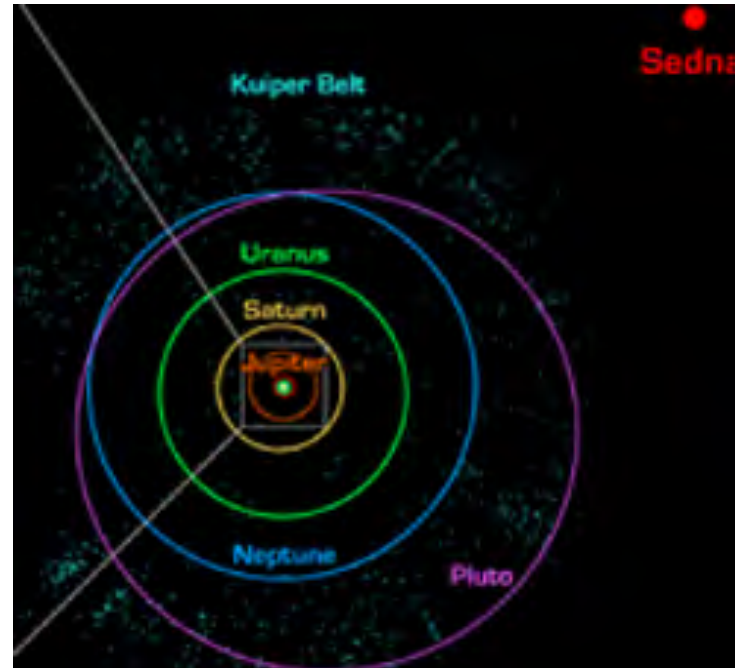
# ..., SOME ARE EXTREME

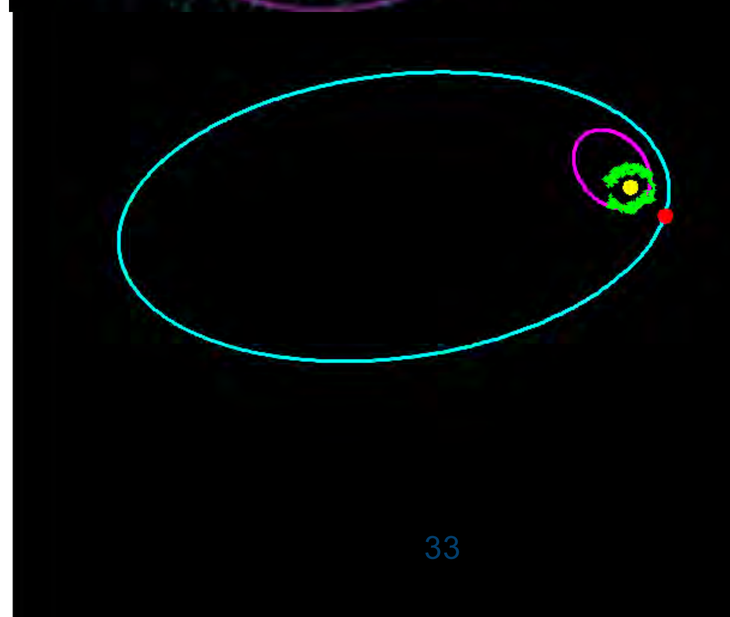
| | Sedna | 2012 $VP_{113}$ |
|---|---|---|
| Perhelion: | 76 AU | 80 AU |
| Apelion: | 937 AU | 446 AU |
| Period: | 11400 yr | 4274 yr |
| Eccentricity: | 0.8527 | 0.694 |

**High eccentricity NOT caused by planets**
(Gaidos et al. 2005)

**Not predicted by Nice model**

**Planet Nine?**



# THE PROBLEM WITH THE NICE MODEL

1. ~~Late-Heavy bombardment~~
2. Hot Kuiper belt (90%) very well,
3. Families of Asteroids
4. Low mass of Mars
5. etc.

Match to observations not self-consistent

Each “match“ for different initial condition and subset of simulations

No predictions

Not falsifiable

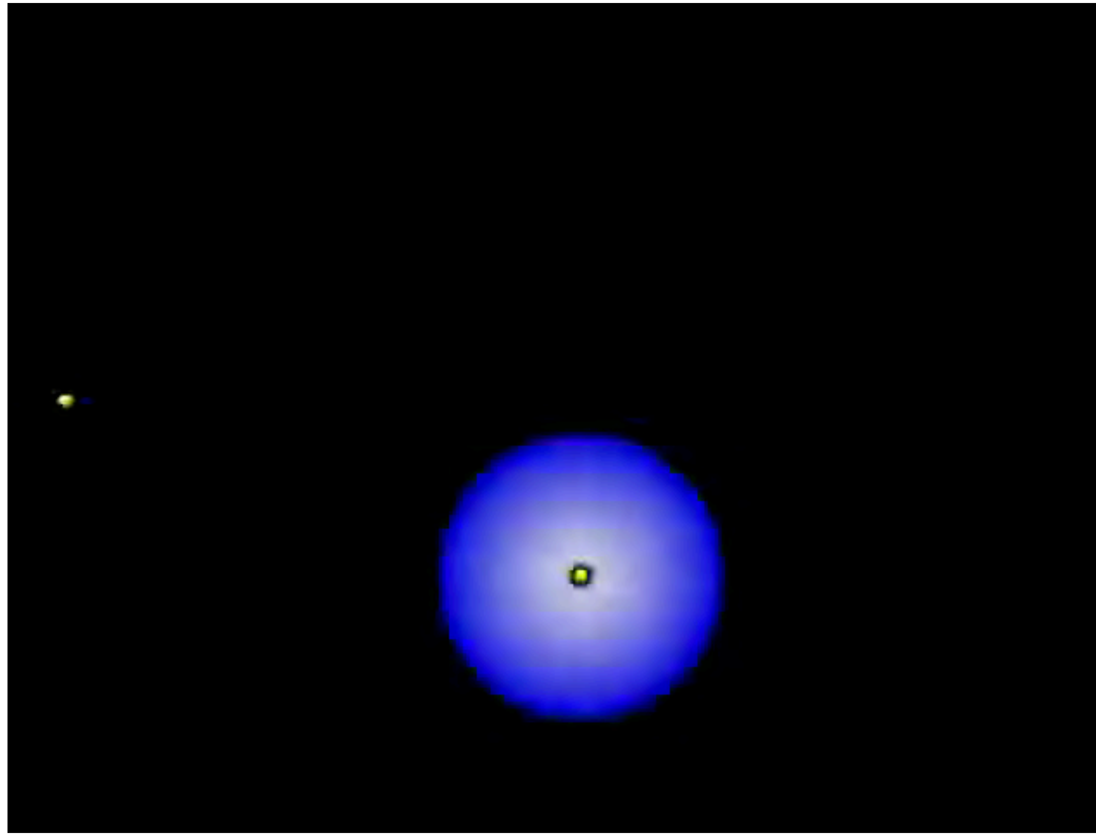
# ALTERNATIVE: CLOSE STELLAR FLYBY

First suggested by
Kobayashi & Ida (2001)
Kenyon & Bromley (2004)

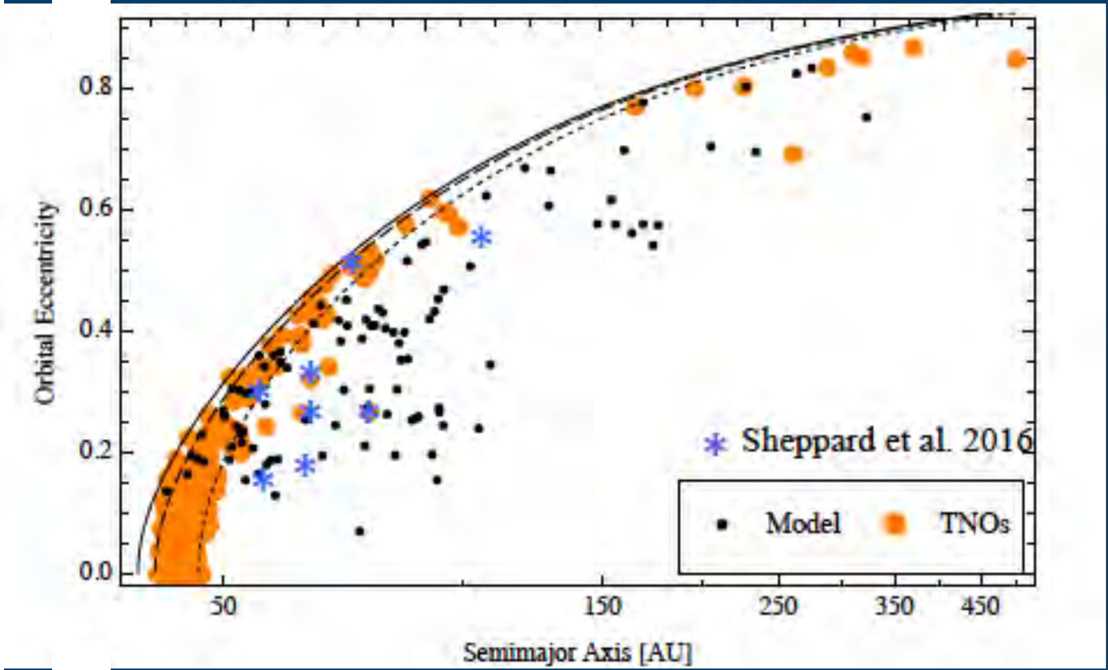
Simulated some specific cases and found that **inclined, eccentric orbits can be obtained by flyby**

But did not get
- Sednoids
- cold Kuiper belt

- Resolution was too low
- Too small initial disc



# FLY-BY REPRODUCES TRANSPENTUNIAN OBJECTS




Fly-by of star with

Mass: 0.5 $M_{sun}$

Perihelion distance: 100 AU

Inclination: $60^0$

Solar disc: > 100 AU

reproduces

- 30 AU drop
- Kuiper belt
- Sednoids

Pfalzner et al. ApJ 2018

# LESSON 2:

LOOKING AT PHENOMENA IN ISOLATION IS DANGEROUS

MODELS HAVE TO BE FALSIFIABLE

MODEL PREDICTIONS ARE ESSENTIAL

# CHARACTERISTICS OF A GOOD ANALYTICAL MODEL

- Matches observation
- Reproducible
- Simple
- Fit into the general theory
- Falsifiable
- Makes prediction

**Criteria should be fulfilled by any model**

# FIRST DATA CHALLENGE