Harnessing Machine Learning to Study Star Formation

利用機器學習研究恆星形成

Stella Offner
The University of Texas at Austin

Collaborators:
Duo Xu (UVA), Rob Gutermuth (UMass), Colin van Oort (UVT)

NCU-Delta Lecture
March 18

Unlocking the Mystery of our Sun

Why does the Sun have the mass it does?

太陽的物質怎麼來的？

How long does it take to form stars?

製造恆星要多久？

Why is there *only one* star in our Solar System?

Tomorrow!

為何太陽系只有一顆恆星？（明天的講題）

Why do the planets have the properties they do?

如何解釋行星各自的性質？

太陽10年

NASA:
10yr time lapse
of the Sun

從恆星誕生地「分子雲」找答案

Answers lie in the birth places of stars: "molecular clouds"

Outline

1. Life cycle of stars
2. Why use machine learning?
3. Hunting for signatures of stellar feedback: outflows and winds

   利用機器學習，尋找恆星形成過程當中回饋雲氣的「噴流」與「恆星風」

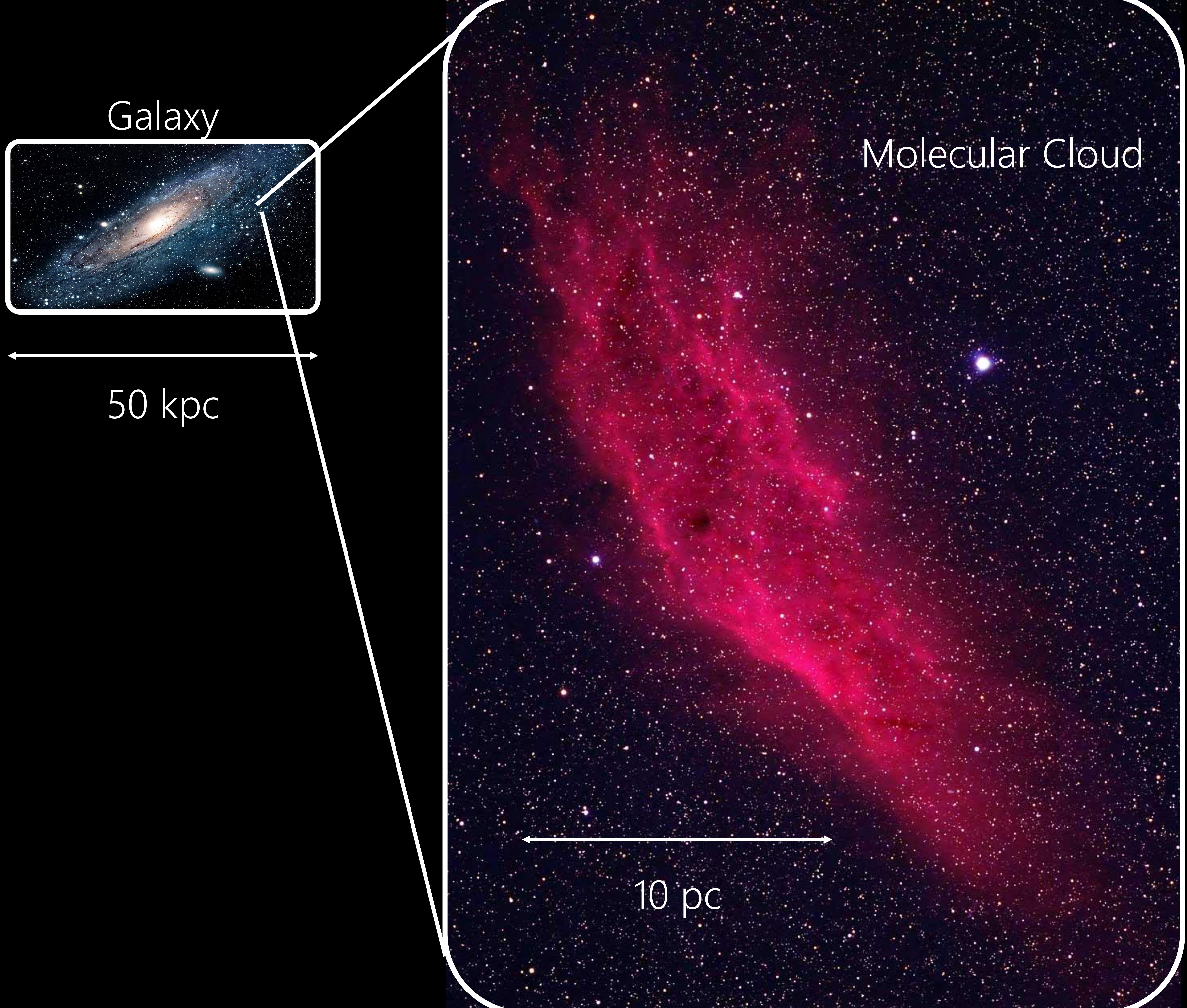
4. Future work & conclusions

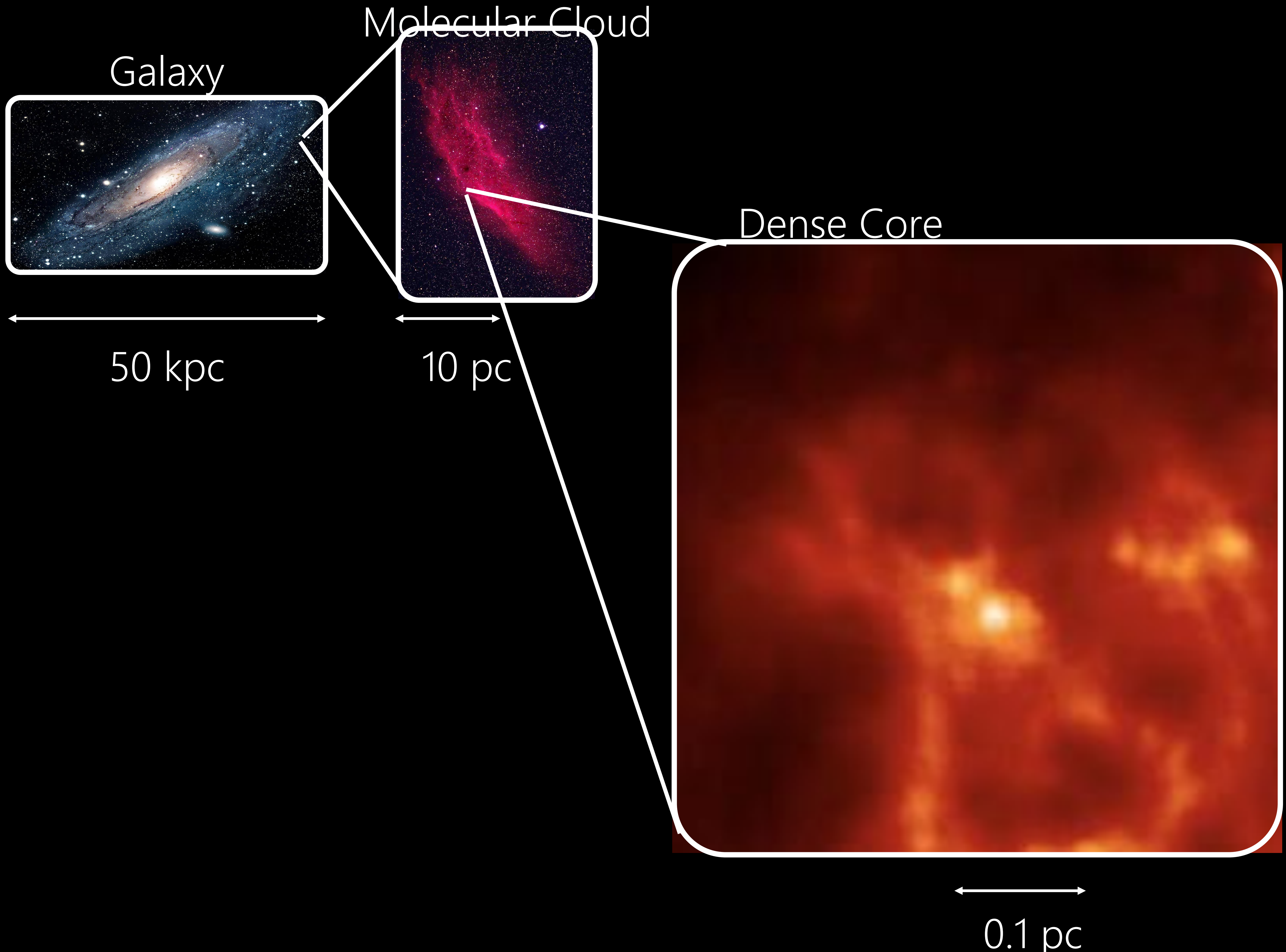
30 Doradus
Credit: NASA, ESA

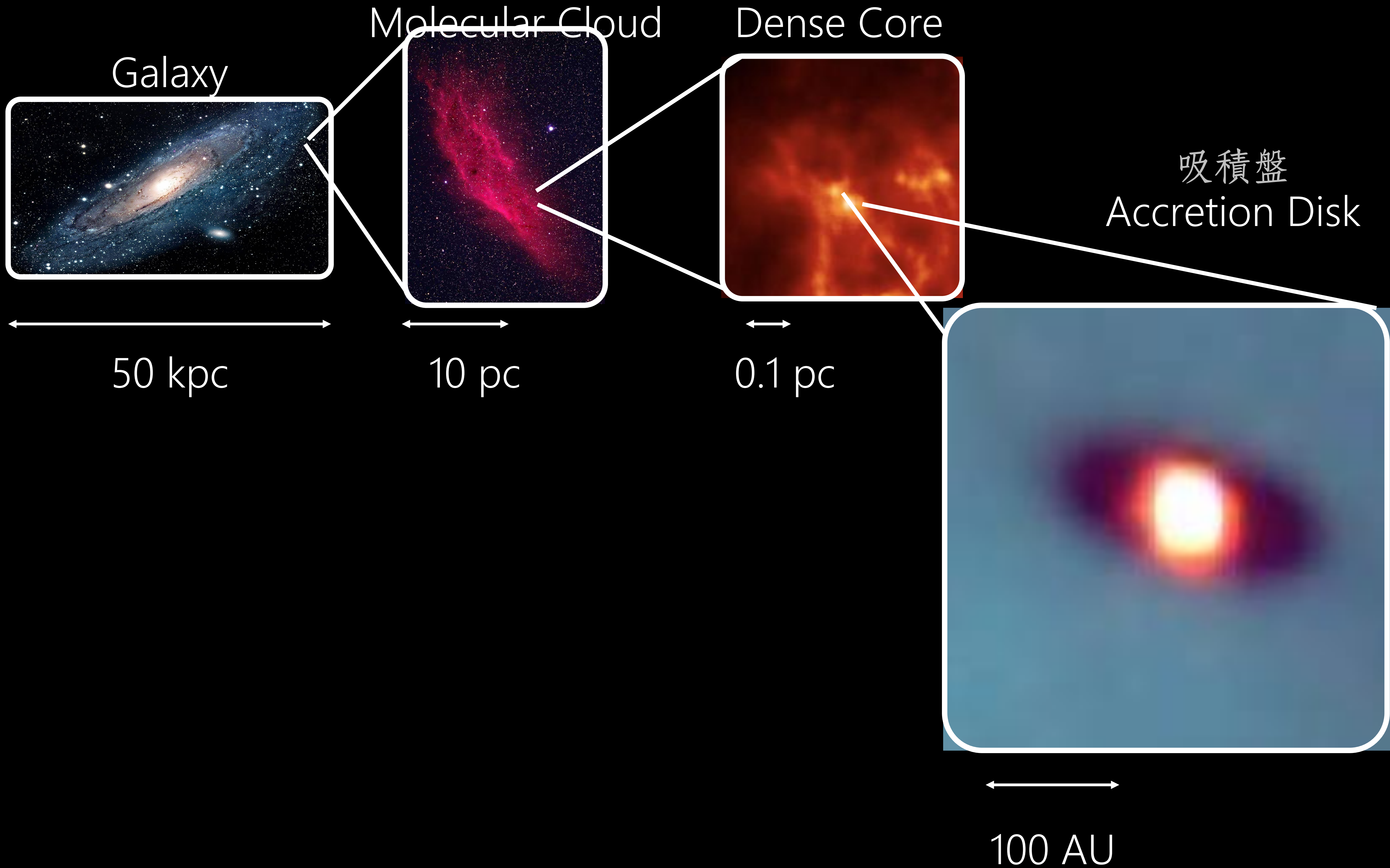
Life Cycle of Stars

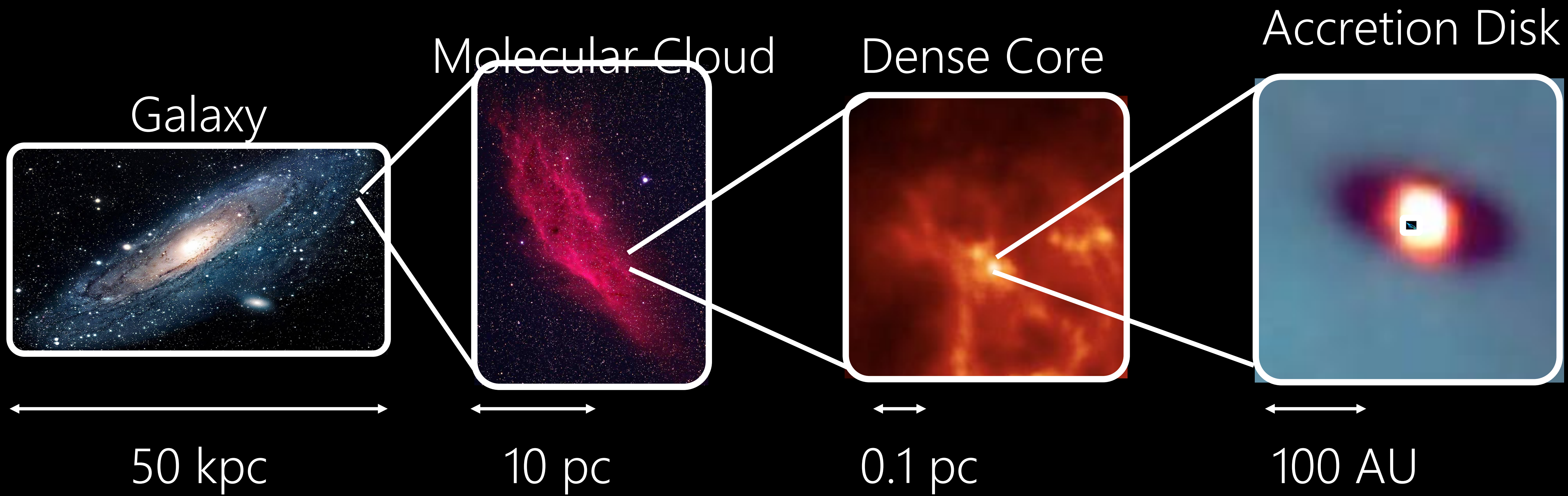
Galaxy

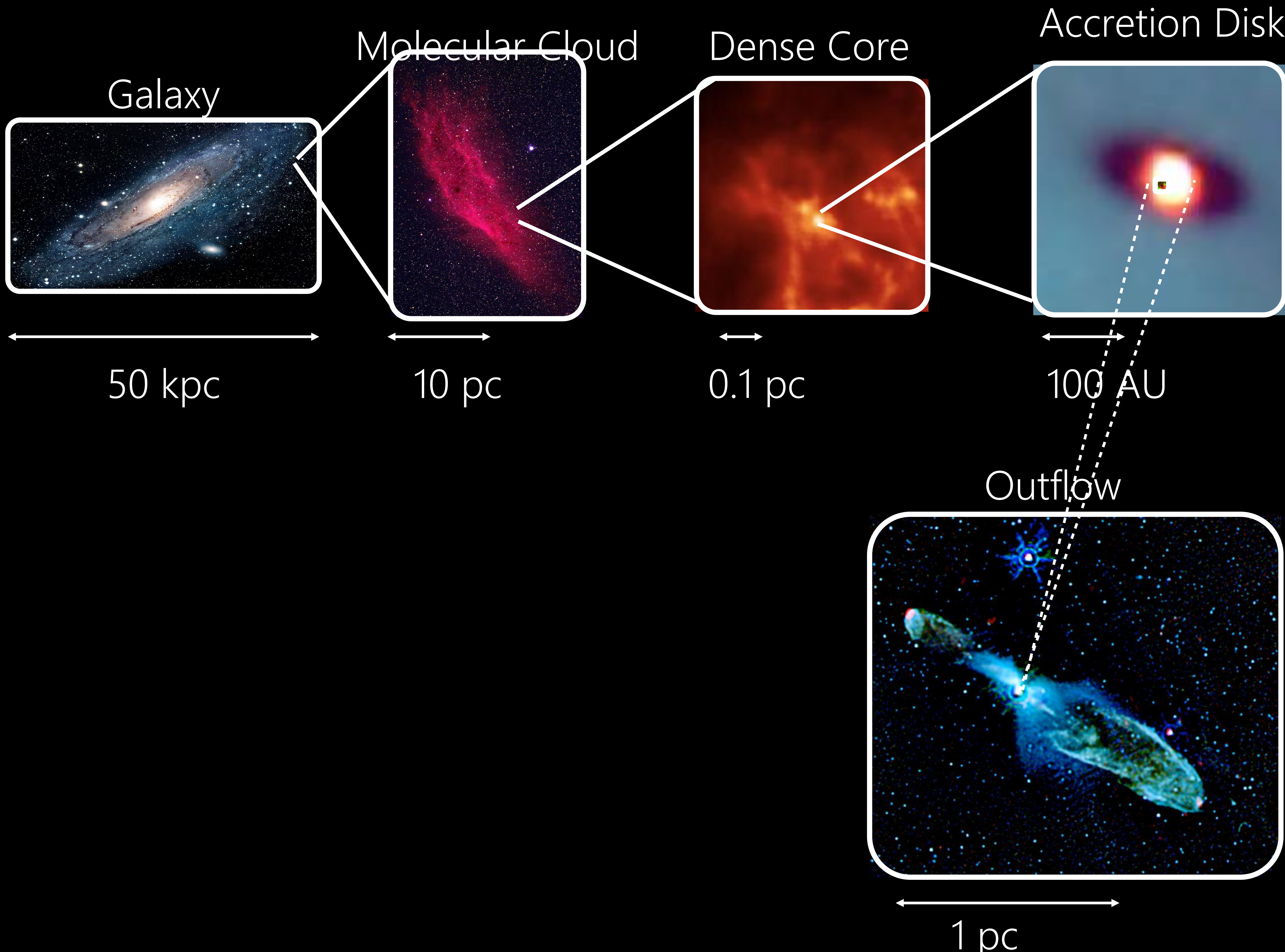
50 kpc

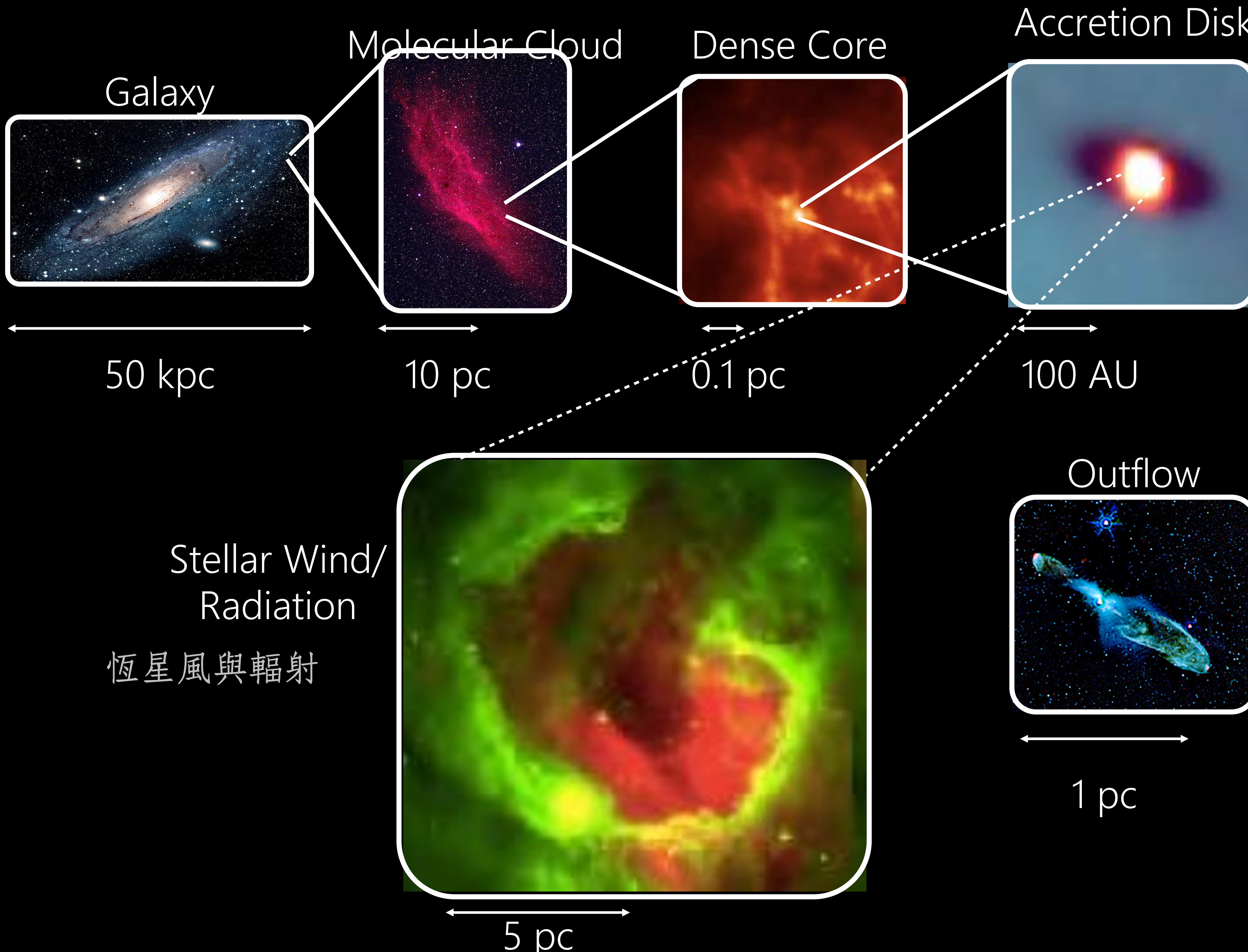

Galaxy
50 kpc
Molecular Cloud
10 pc

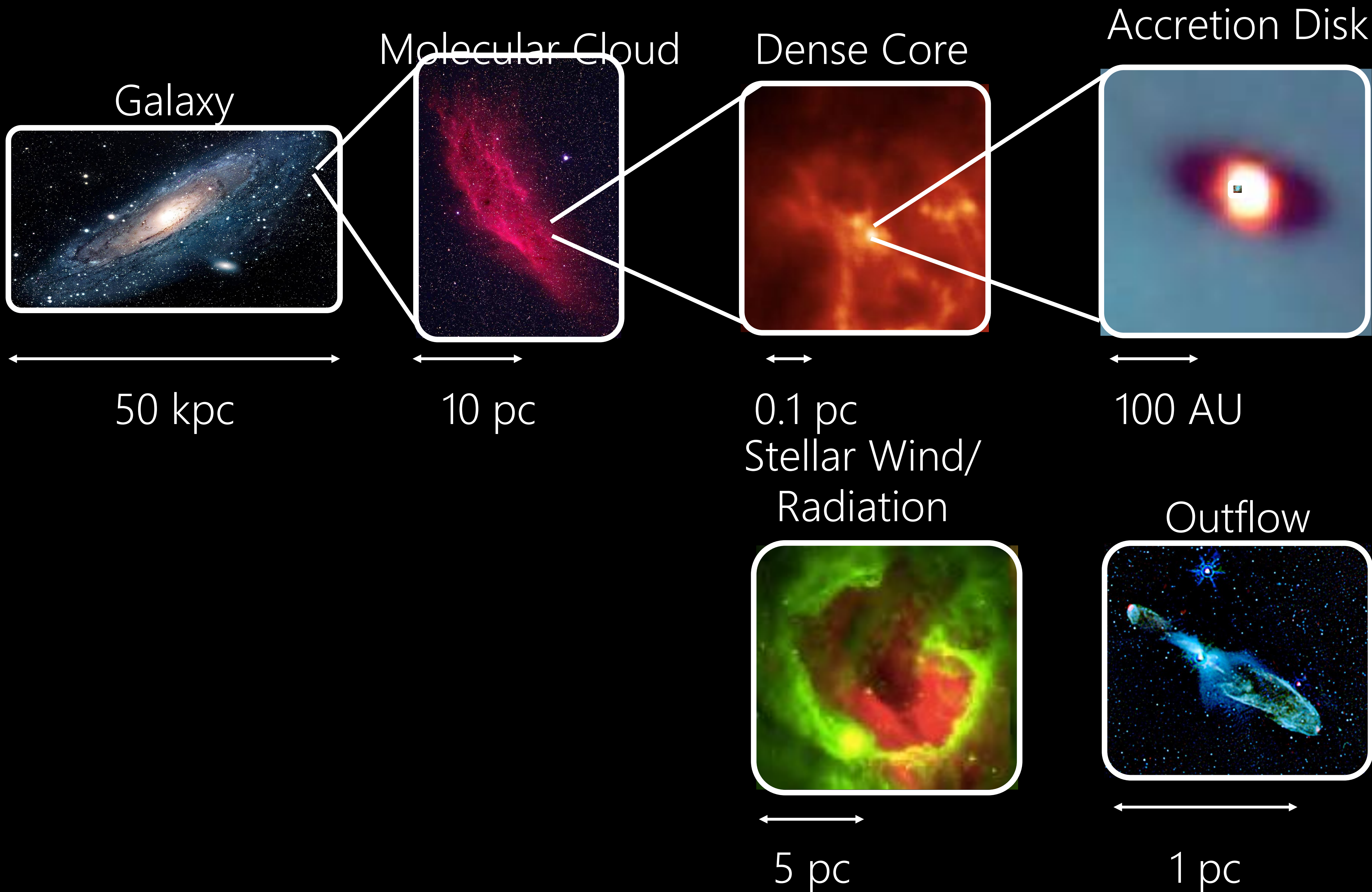

Molecular Cloud
Galaxy
Dense Core
50 kpc
10 pc
0.1 pc

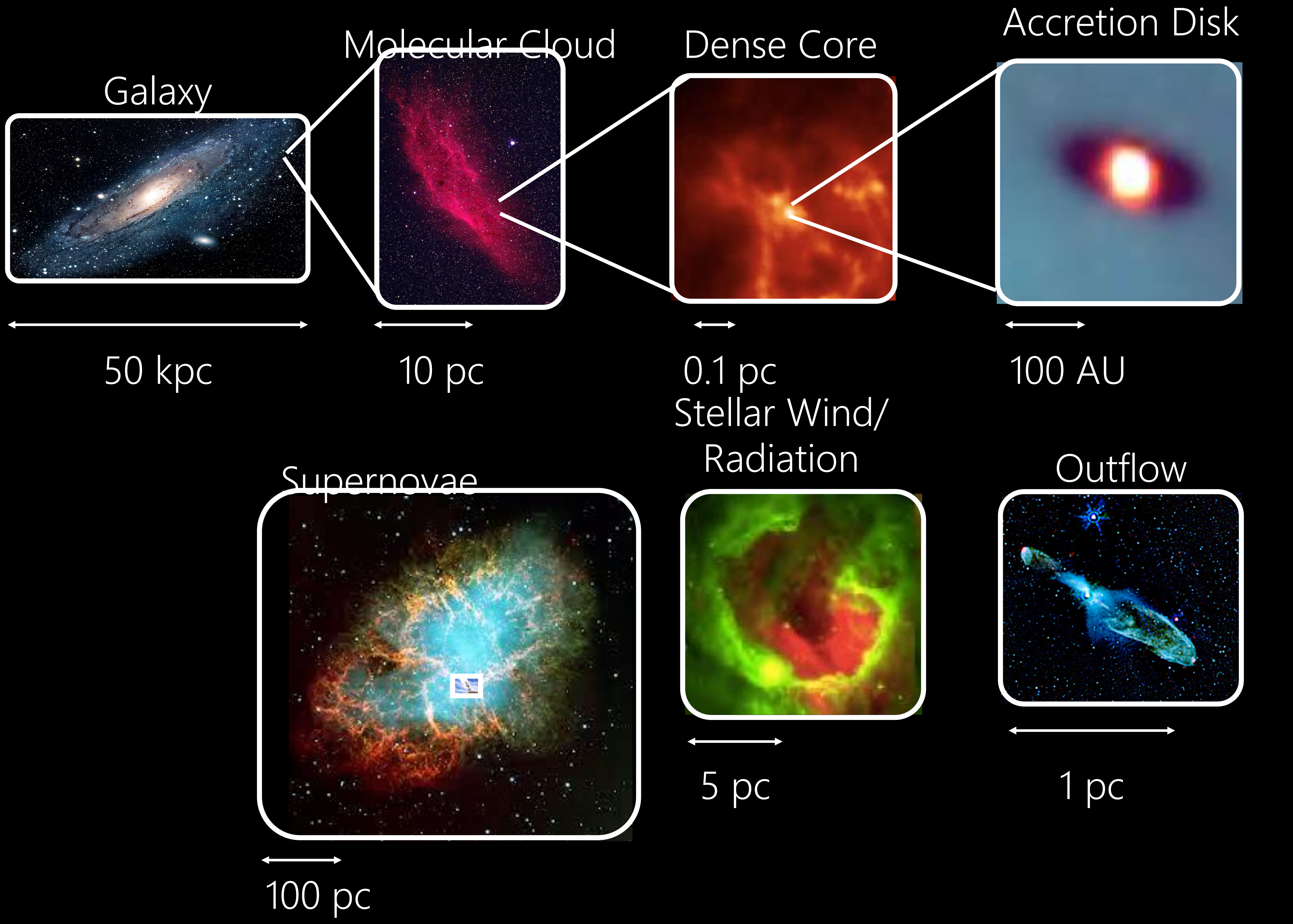

Galaxy
Molecular Cloud
Dense Core
吸積盤
Accretion Disk
50 kpc
10 pc
0.1 pc
100 AU

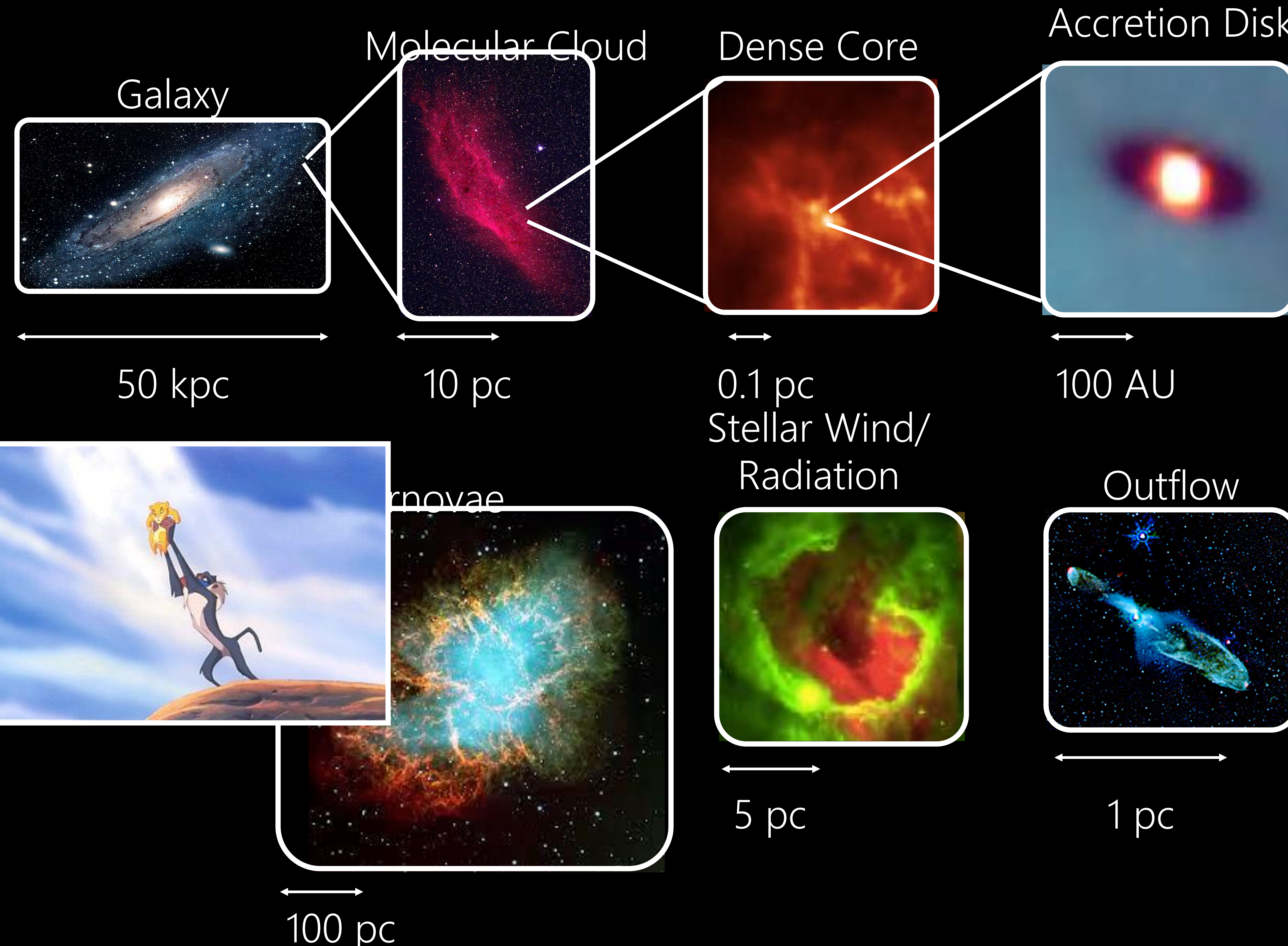

Galaxy
Molecular Cloud
Dense Core
Accretion Disk
50 kpc
10 pc
0.1 pc
100 AU


Galaxy
Molecular Cloud
Dense Core
Accretion Disk
50 kpc
10 pc
0.1 pc
100 AU
Outflow
1 pc


Galaxy
Molecular Cloud
Dense Core
Accretion Disk
50 kpc
10 pc
0.1 pc
100 AU
Stellar Wind/
Radiation
恆星風與輻射
Outflow
1 pc
5 pc

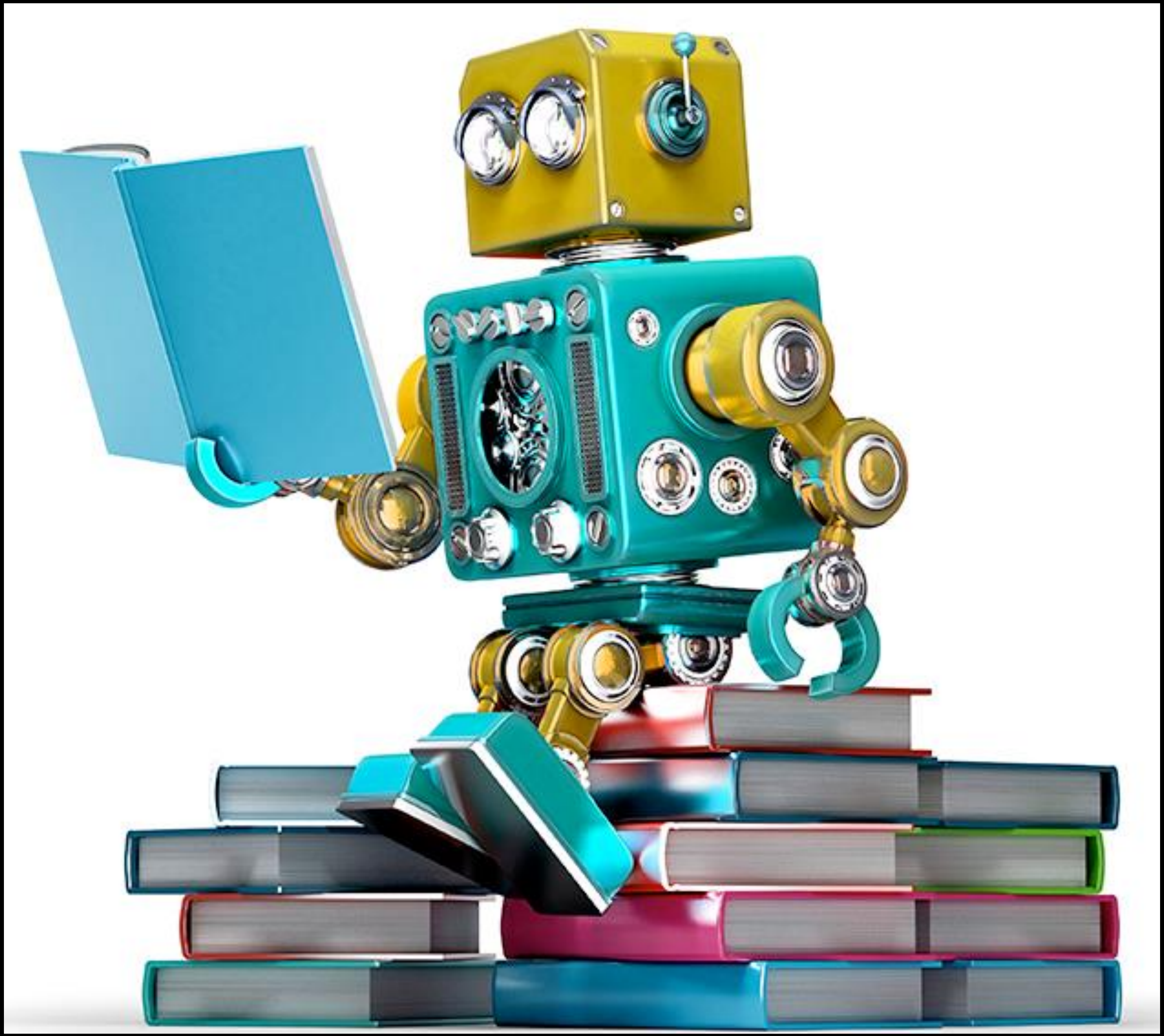

Galaxy
Molecular Cloud
Dense Core
Accretion Disk
50 kpc
10 pc
0.1 pc
100 AU
Stellar Wind/
Radiation
Outflow
5 pc
1 pc


Galaxy
Molecular Cloud
Dense Core
Accretion Disk
50 kpc
10 pc
0.1 pc
100 AU
Stellar Wind/
Radiation
Supernovae
Outflow
5 pc
1 pc
100 pc


Galaxy
Molecular Cloud
Dense Core
Accretion Disk
50 kpc
10 pc
0.1 pc
100 AU
rnovae
Stellar Wind/
Radiation
Outflow
100 pc
5 pc
1 pc

Machine Learning

or Artificial Intelligence

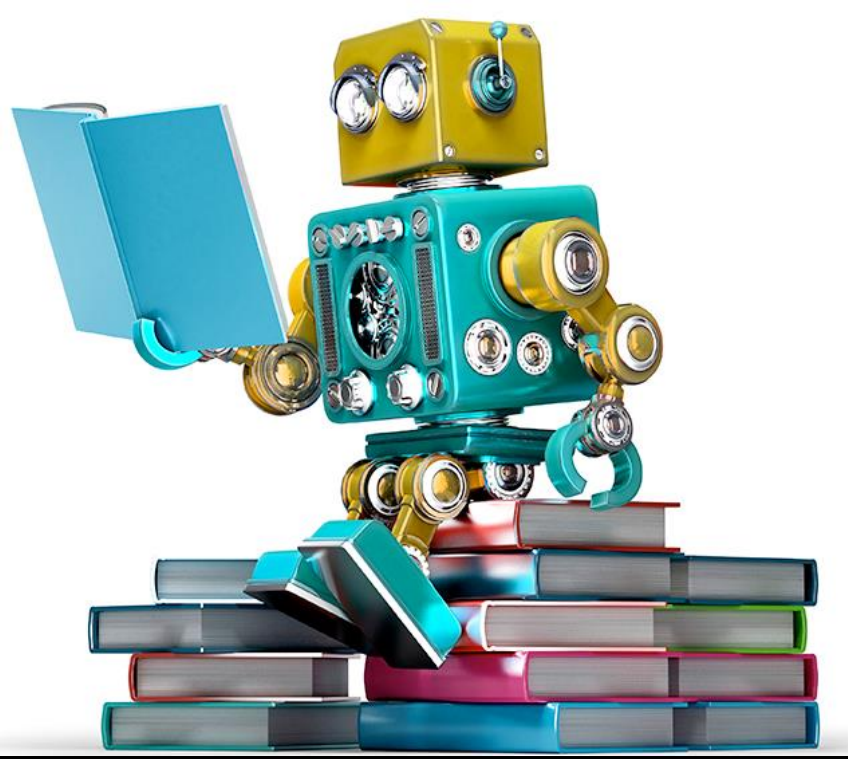


Machine Learning

or Artificial Intelligence

noun :: field of computer science that gives computers the ability to learn without being explicitly programmed. 讓電腦自我學習，而非執行特定程式

— Arthur Samuel



Machine Learning

or Artificial Intelligence

noun :: field of computer science that gives computers the ability to learn without being explicitly programmed.

— Arthur Samuel

1. Look for other things like (but not exactly like) this one: *supervised*

Machine Learning

or Artificial Intelligence

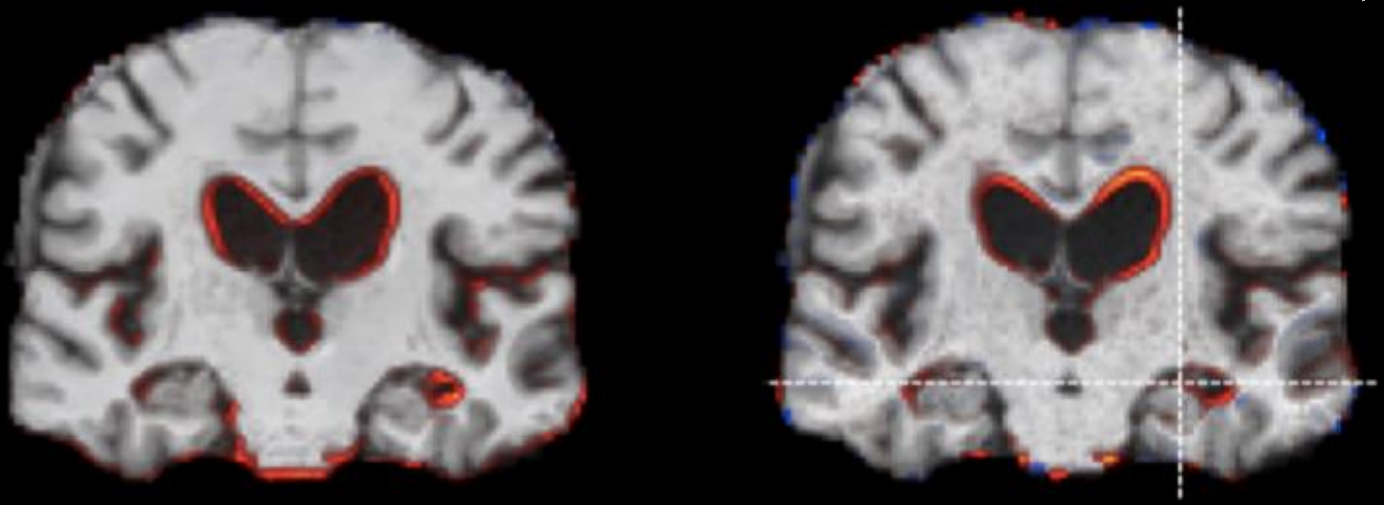
noun :: field of computer science that gives computers the ability to learn without being explicitly programmed.
— Arthur Samuel

1. Look for other things like (but not exactly like) this one: *supervised*

監督學習

2. Find a pattern / something interesting: *unsupervised*

無監督學習



3. Make new images / similar data: *generative*

具衍生特性

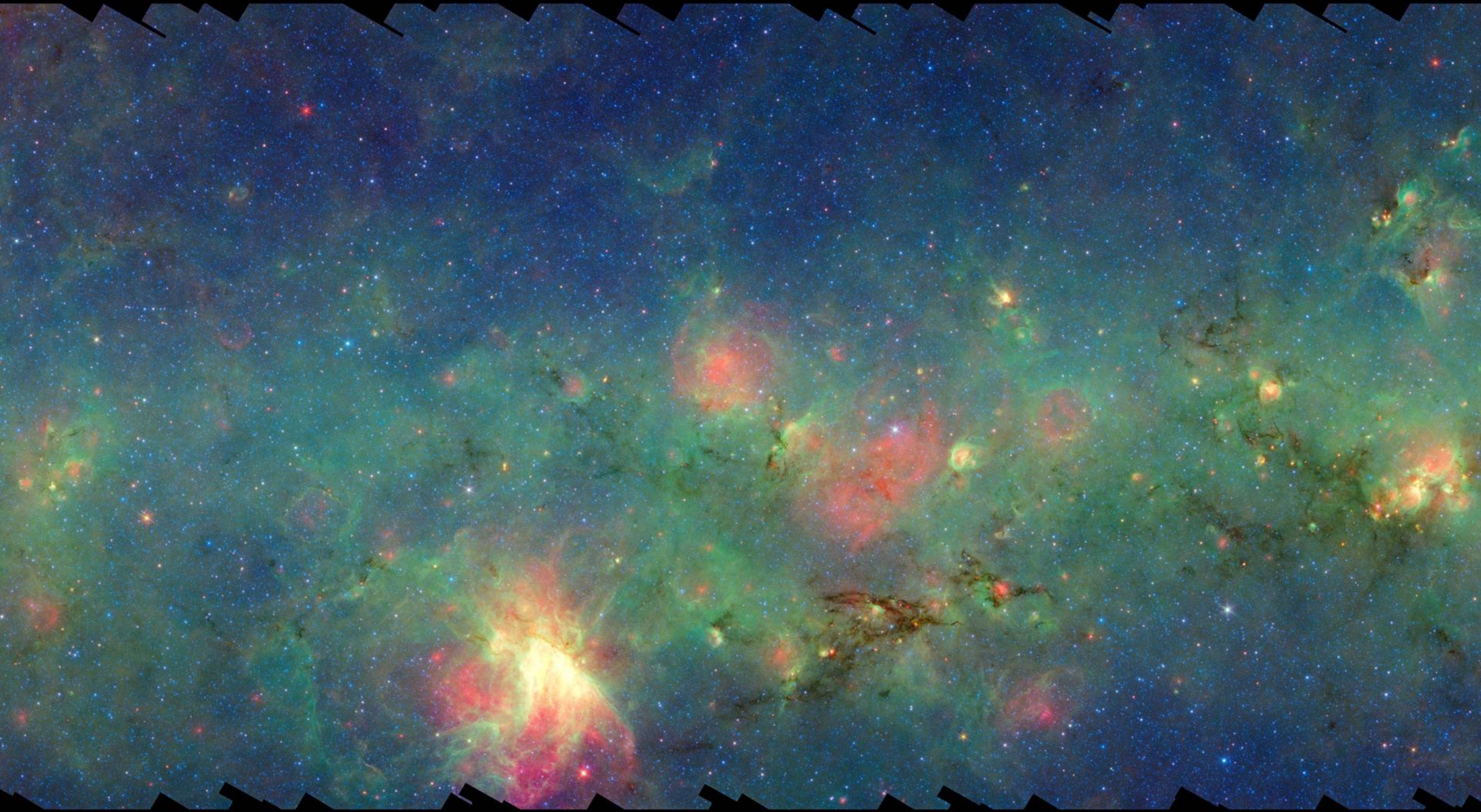
Why machine learning?

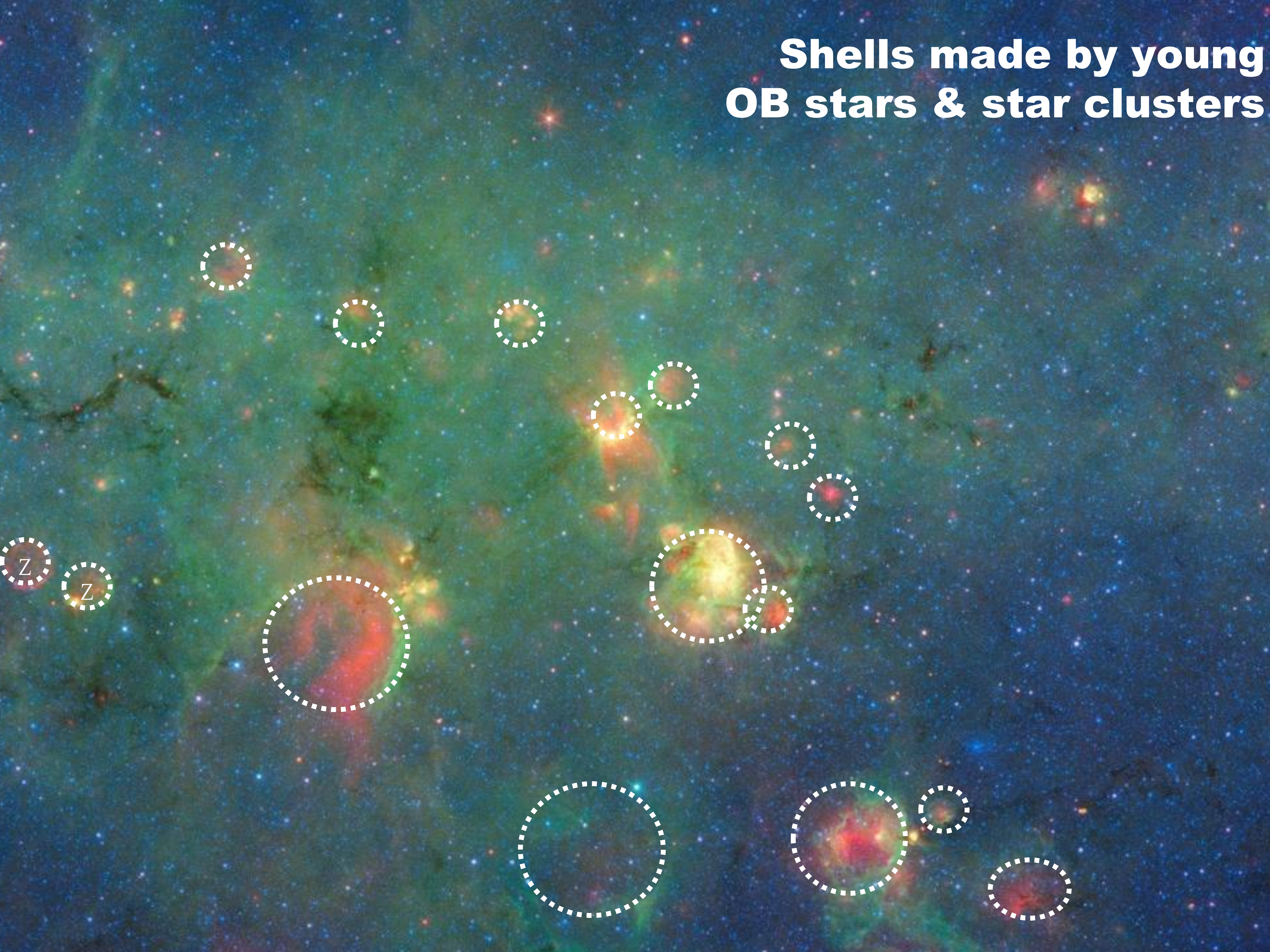
....Star formation is *messy*

恆星誕生的環境複雜

Dust Emission 星際塵埃的輻射

Spitzer Space Telescope Galactic Plane Survey



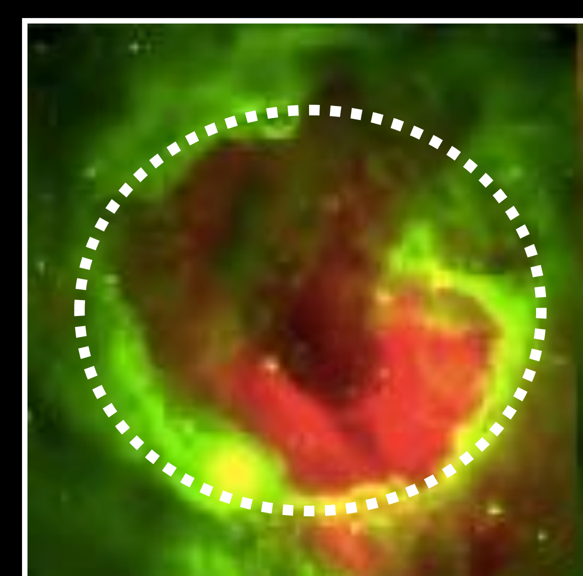

Shells made by young
OB stars & star clusters
Z
Z

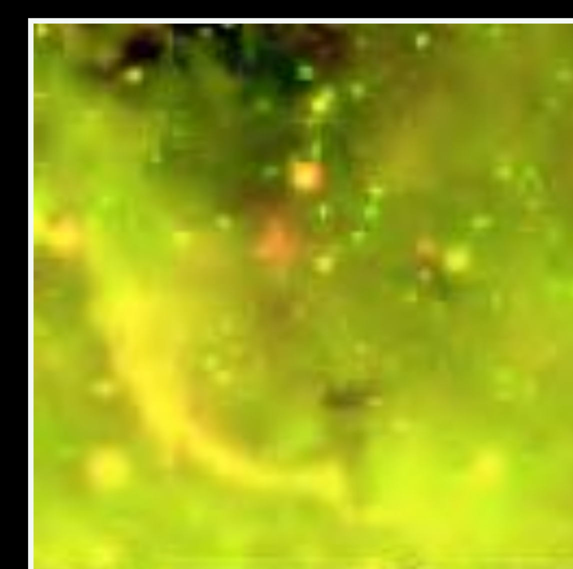
Why machine learning?

…There is *a lot* of data…

數據量大

Milky Way Citizen Science Project (MWP)

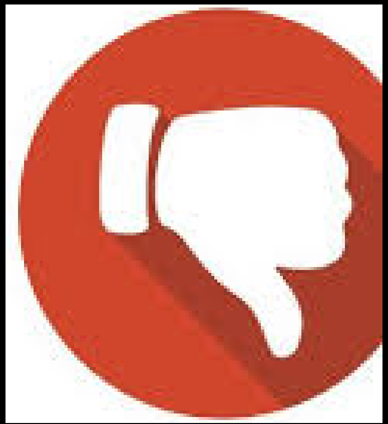




✔ ?

https://www.zooniverse.org/projects/povich/milky-way-project

Human Identifications



1. Engage the public in science!
2. Numerous!
3. Free!
4. Can identify atypical cases!

1. Different opinions.
2. Need simple instructions.
3. Can be "hangry."

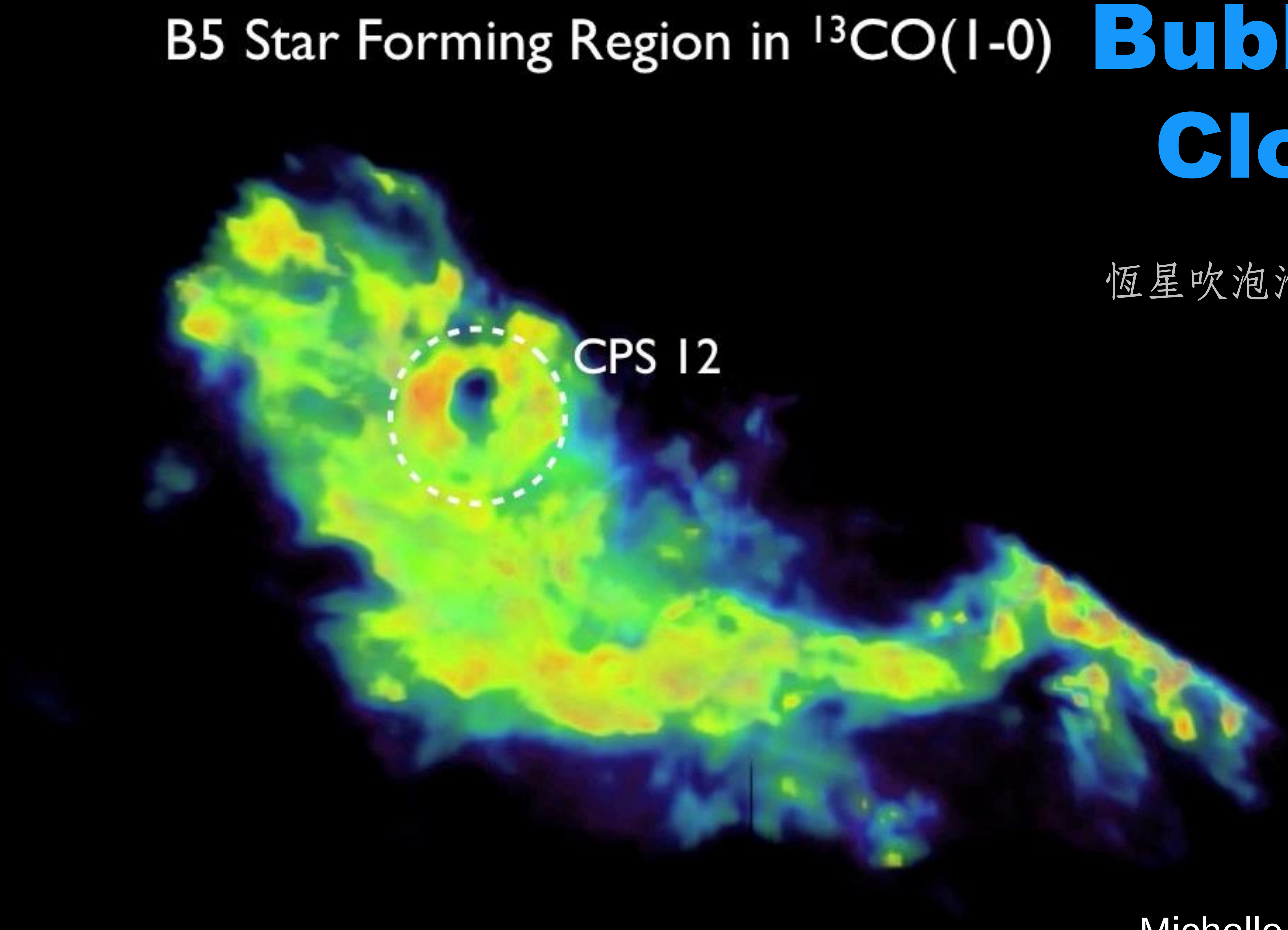
Even experts don't know the "right" answer.

Why machine learning?

….Star formation is messy

…*and* 3+ Dimensional

Bubbling Clouds!



恒星吹泡泡

Michelle Borkin
Human with Applied Physics PhD, topic: data viz

A Bubbling Nearby Molecular Cloud

H. Arce et al. 2011, movie courtesy of A. Goodman

How does feedback shape clouds & star formation?

Part I. Stellar Winds

Part II. Outflows

利用機器學習指認恆星回饋
其所誕生的雲氣

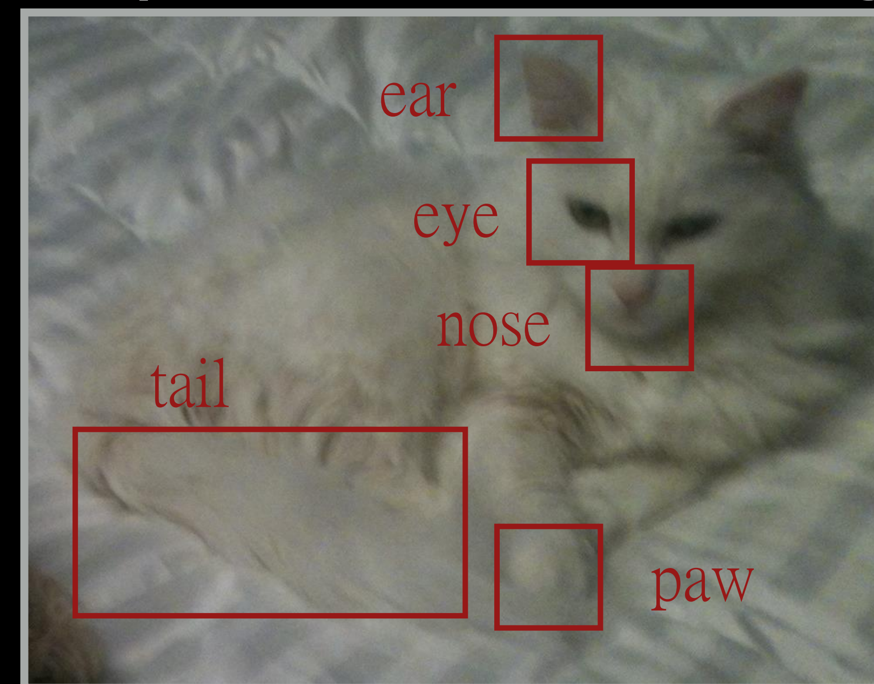
Use machine learning to identify stellar feedback!

Convolutional Neural Network

卷積神經網路

先告訴機器要學什麼；叮著它學

Supervised Learning



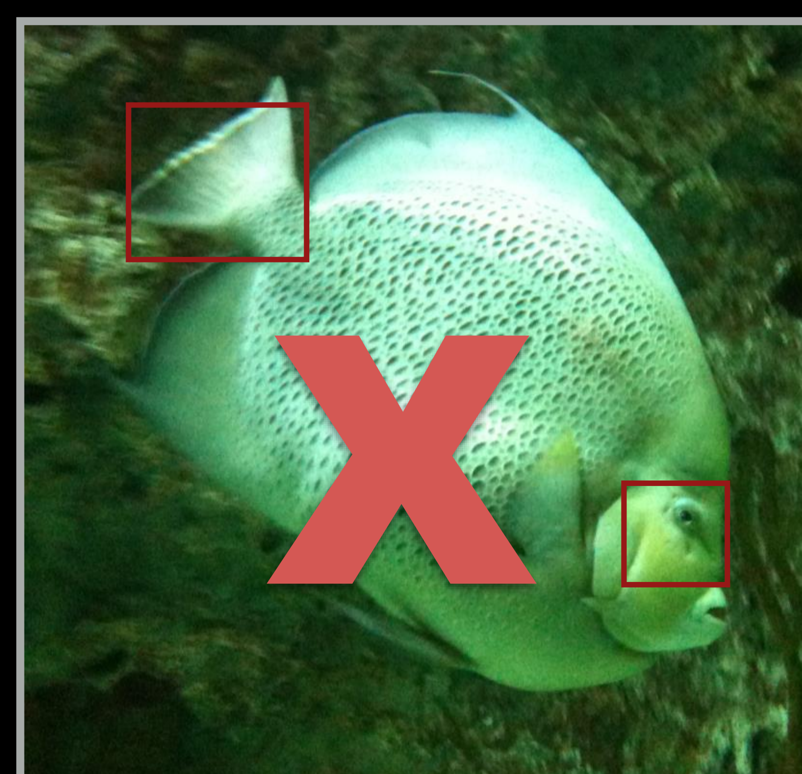
"Cat"

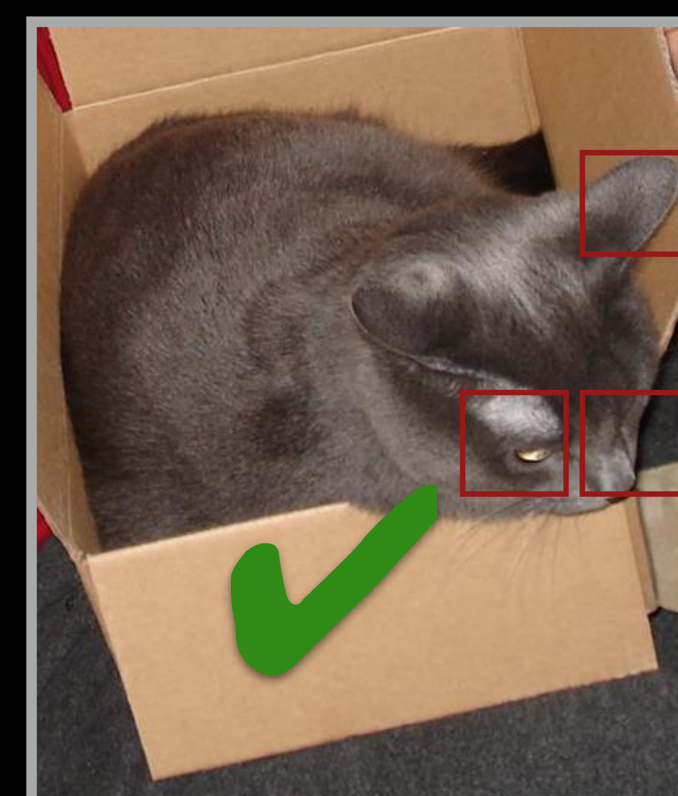
Characteristics

"deep learning"

深度學習

Colin van Oort





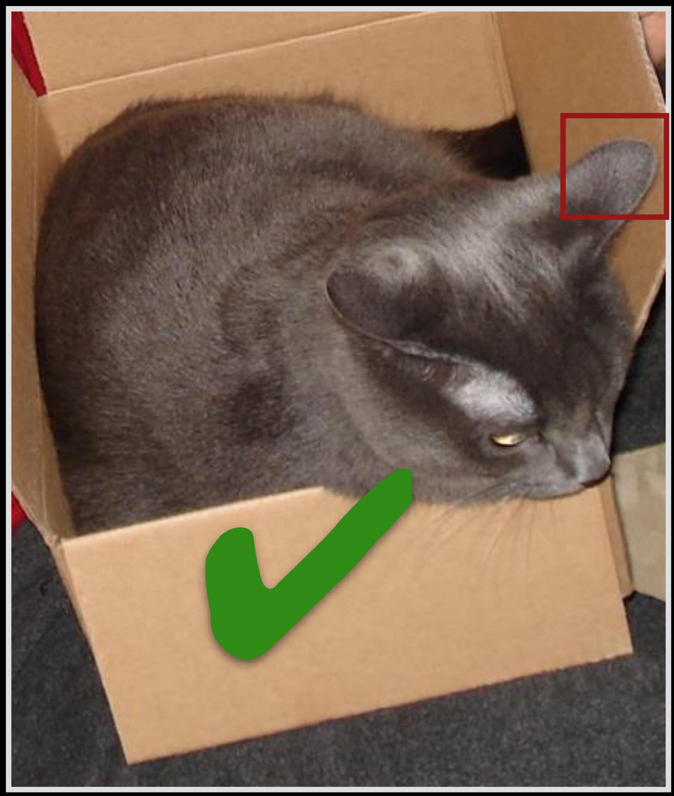
Convolutional Neural Network

"deep learning"

Supervised Learning

Training Set

Not an ear,
but a set of
patterns:
"filters"



Colin van Oort

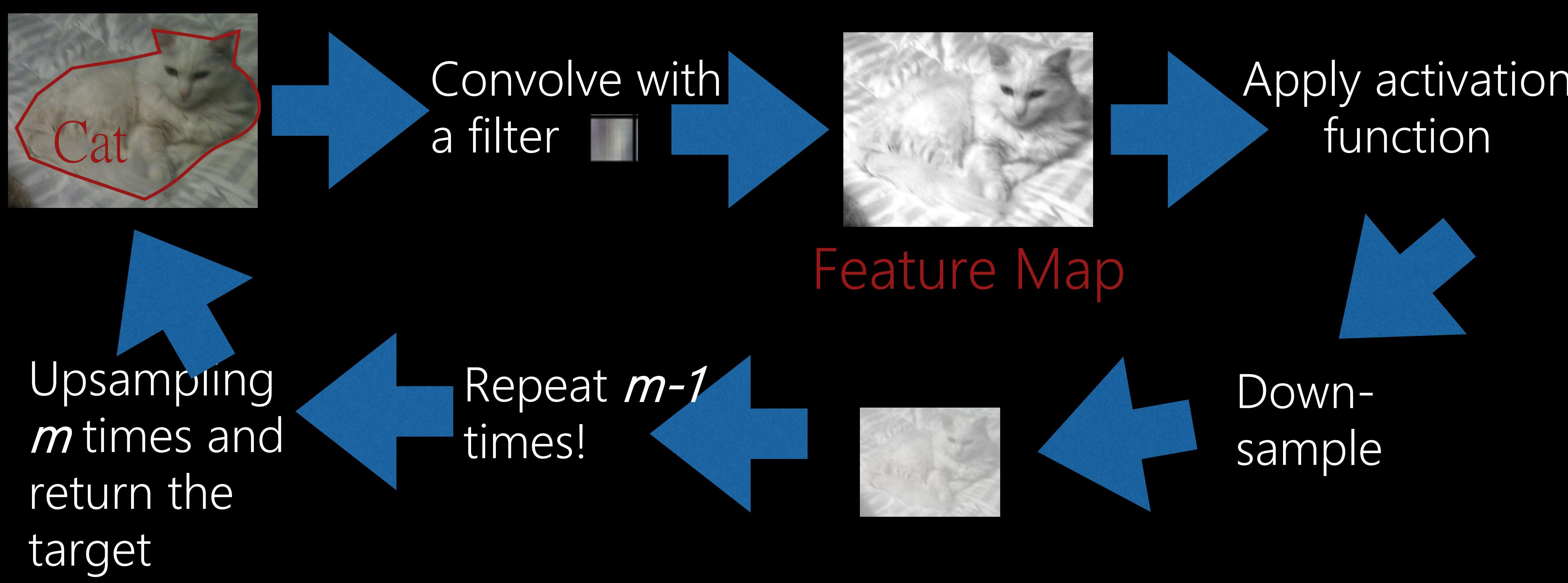
Convolutional Approach to Structure Identification (CASI)

"Denoising Convolutional Autoencoder"

Remove noise / everything not of interest

Be able to reproduce input image

Training Data



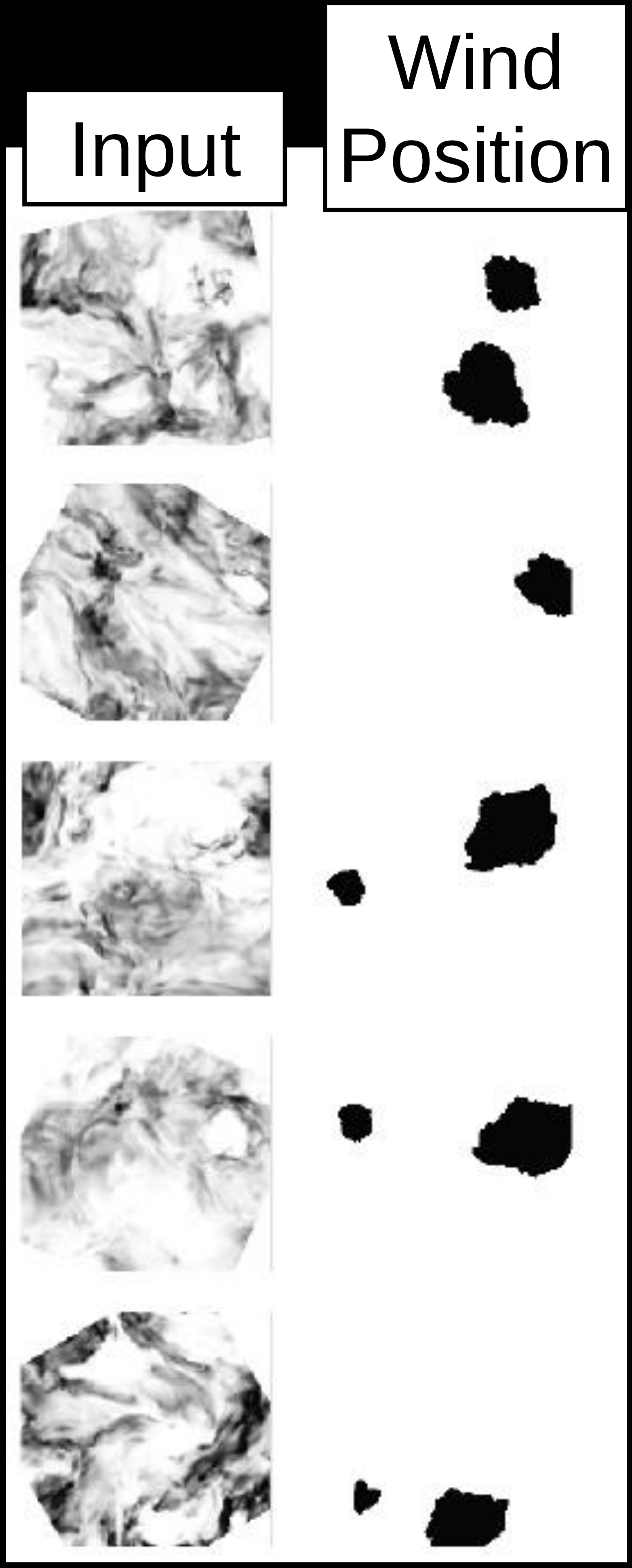
Constructing the Training Set

輸入的學習數據

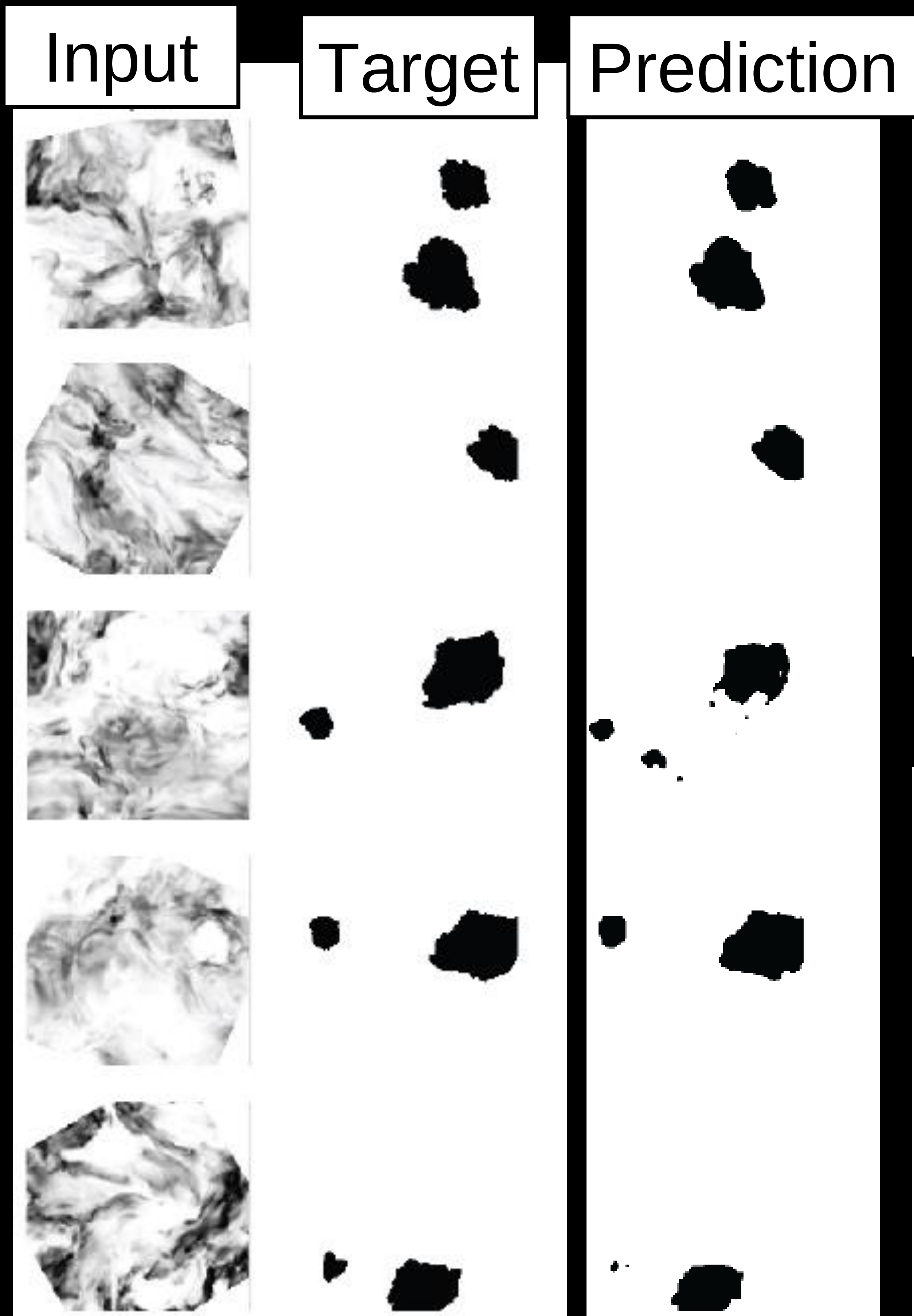
- Gas density　氣體密度
- $^{12}CO/^{13}CO$ emission
- Wind tracer field　風向
- Different times, magnetic field strengths, star properties
- With / without noise

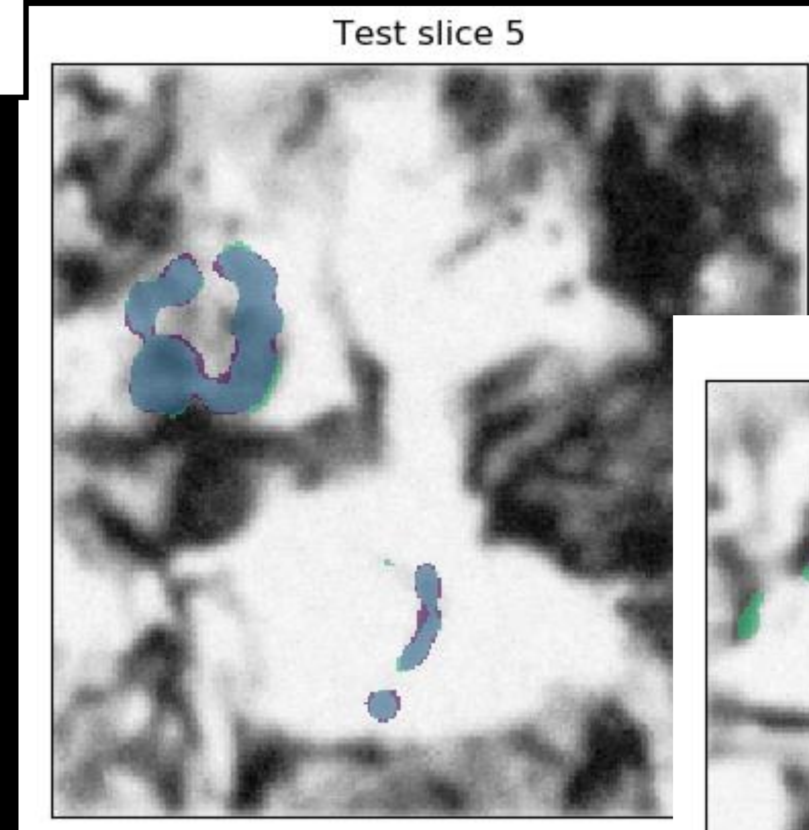
van Oort et al. 2019

^{12}CO emission & binary wind tag slices

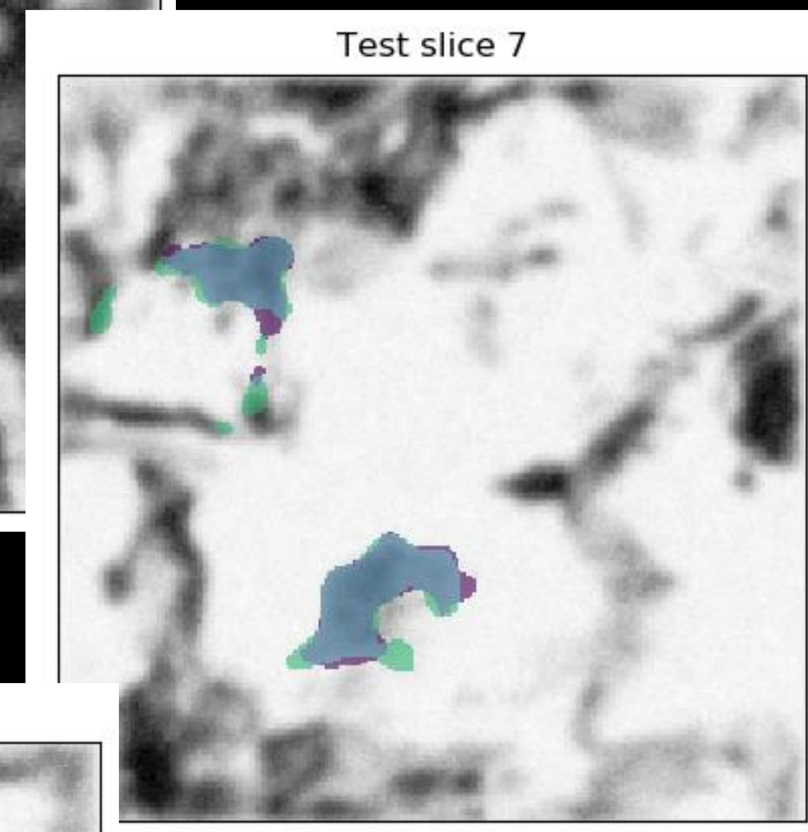


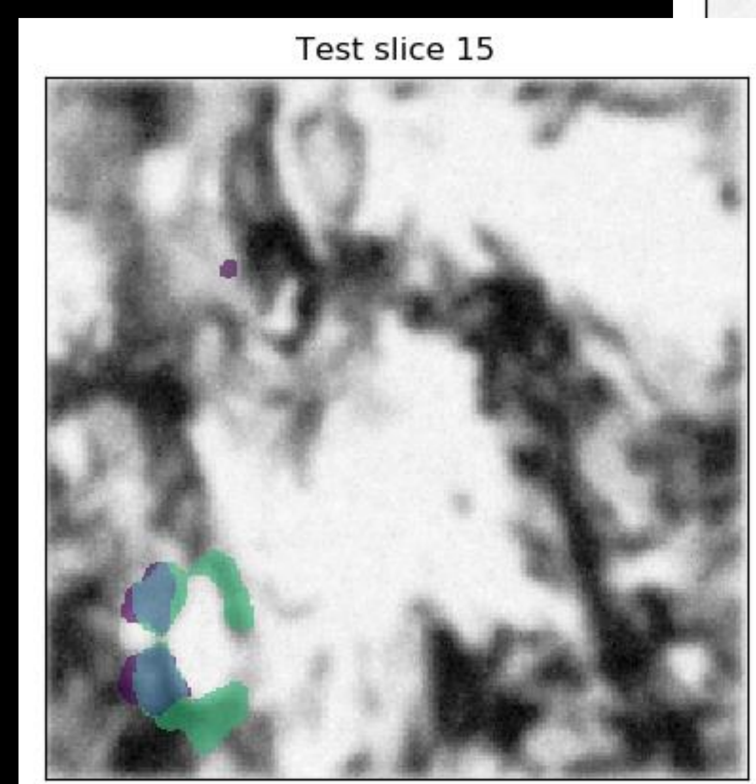
Feedback Recovery (2D slices)





Simulation Testing





True Positive

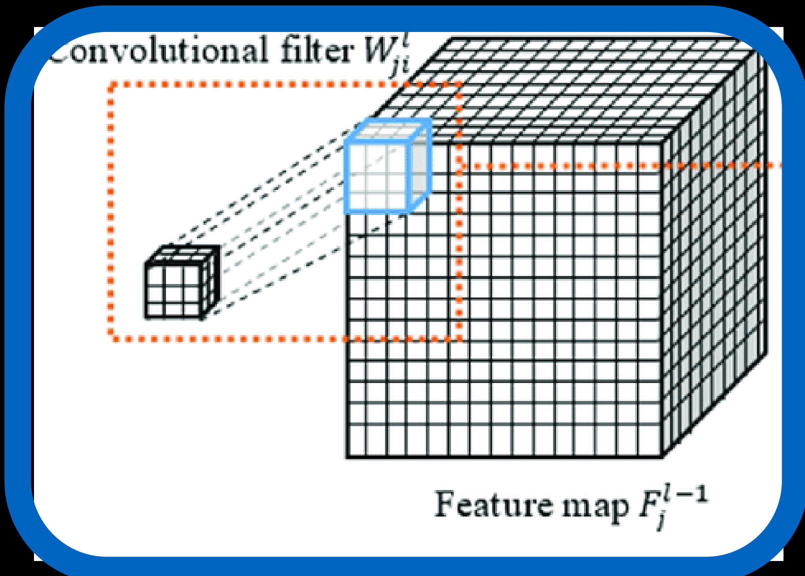
False Positive

False Negative

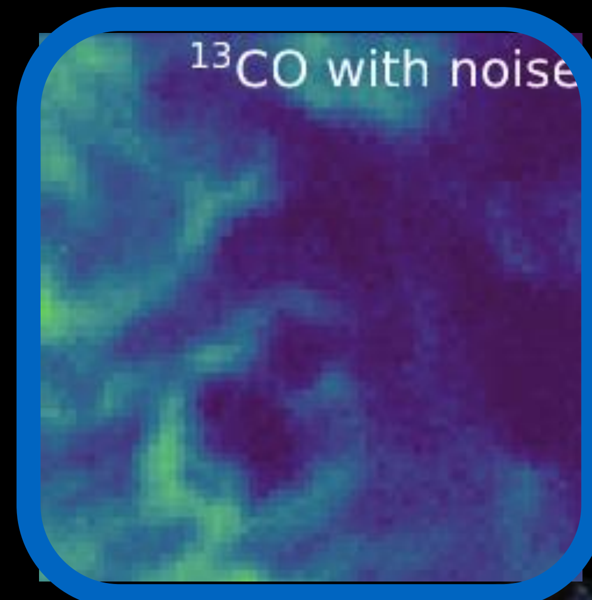
van Oort et al. 2019

Bubbles Expand! ⇒ CASI-3D

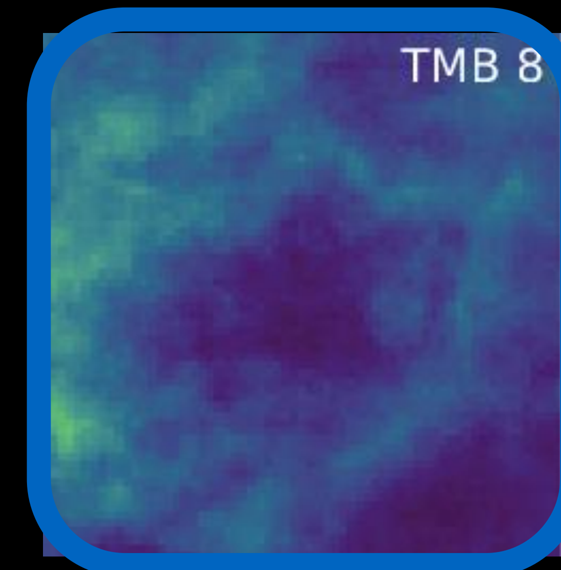
3D Convolution



Train on Simulations



Apply to Taurus cloud

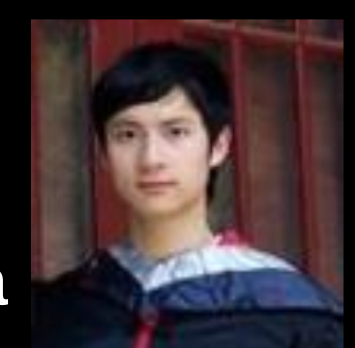


Compare to visual identifications

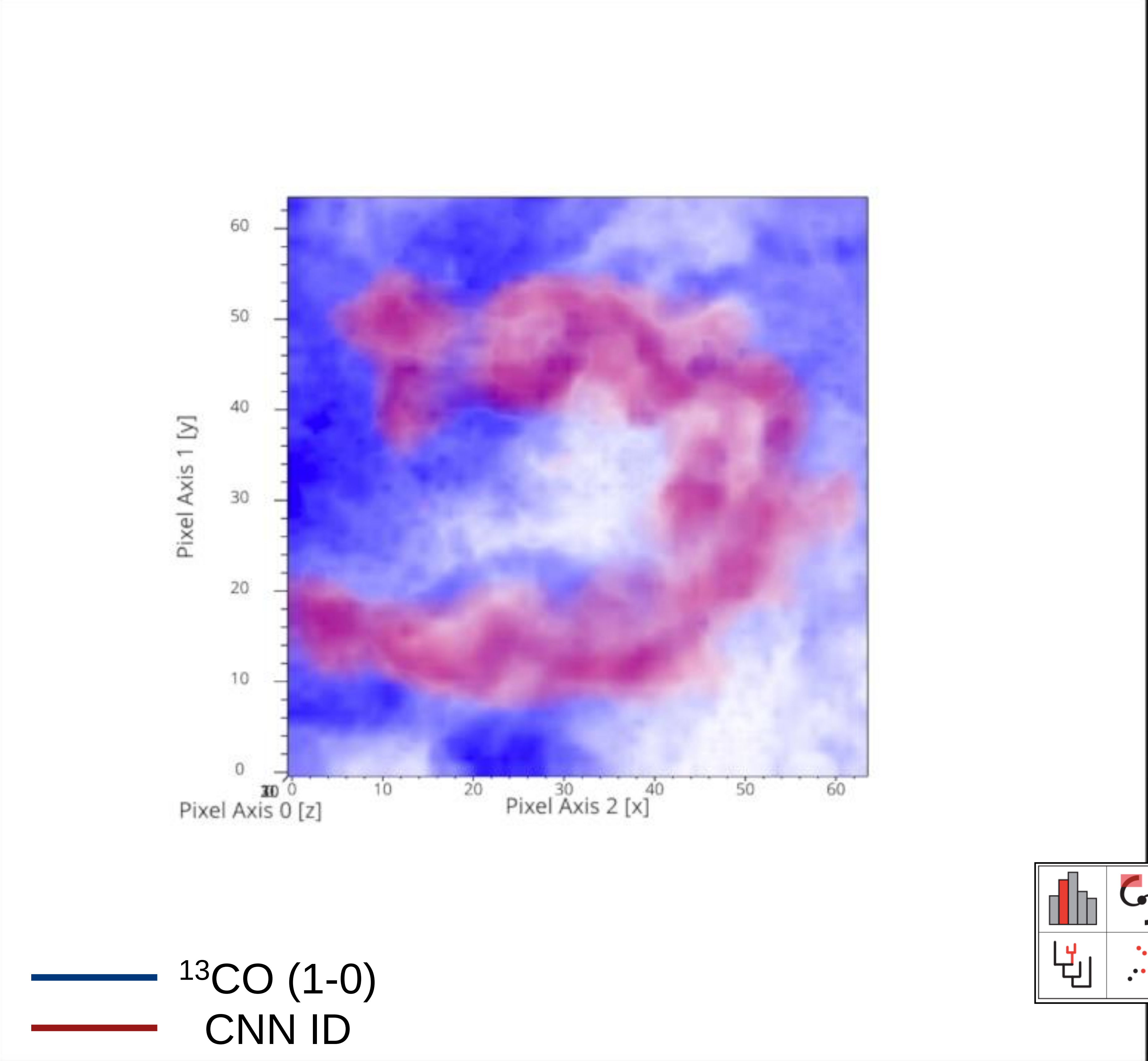
跟「人眼」看到的比較

d=140 pc
Forming low-mass stars (m* < 3 Msun)
Mapped in 12CO, 13CO by FCRAO

Xu et al 2020a



Identification of Shell TMB8



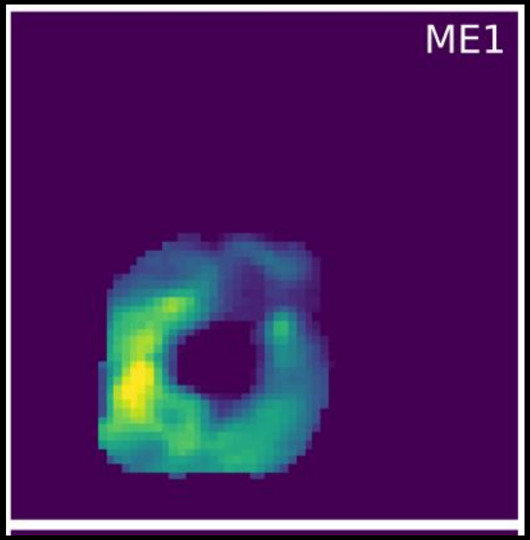
Identifies feedback with pixel level accuracy.

Xu et al 2020a

A Tale of Two Models: ME1 & MF

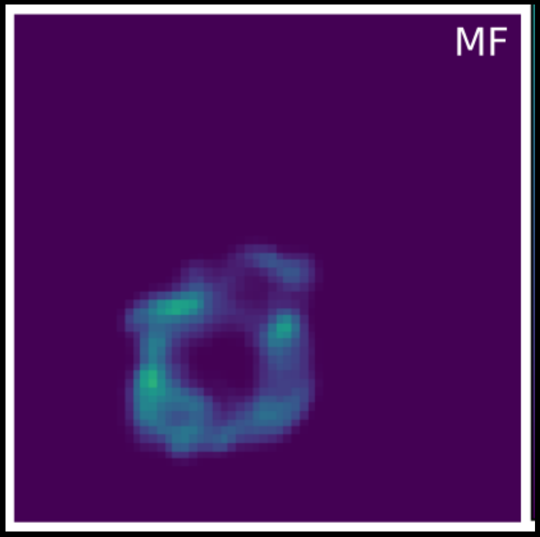
立體像素當中的全部輻射

- ME1 trains on total emission in voxels that contain some feedback gas
  - Similar to how human visual identification works



立體像素當中包含的質量

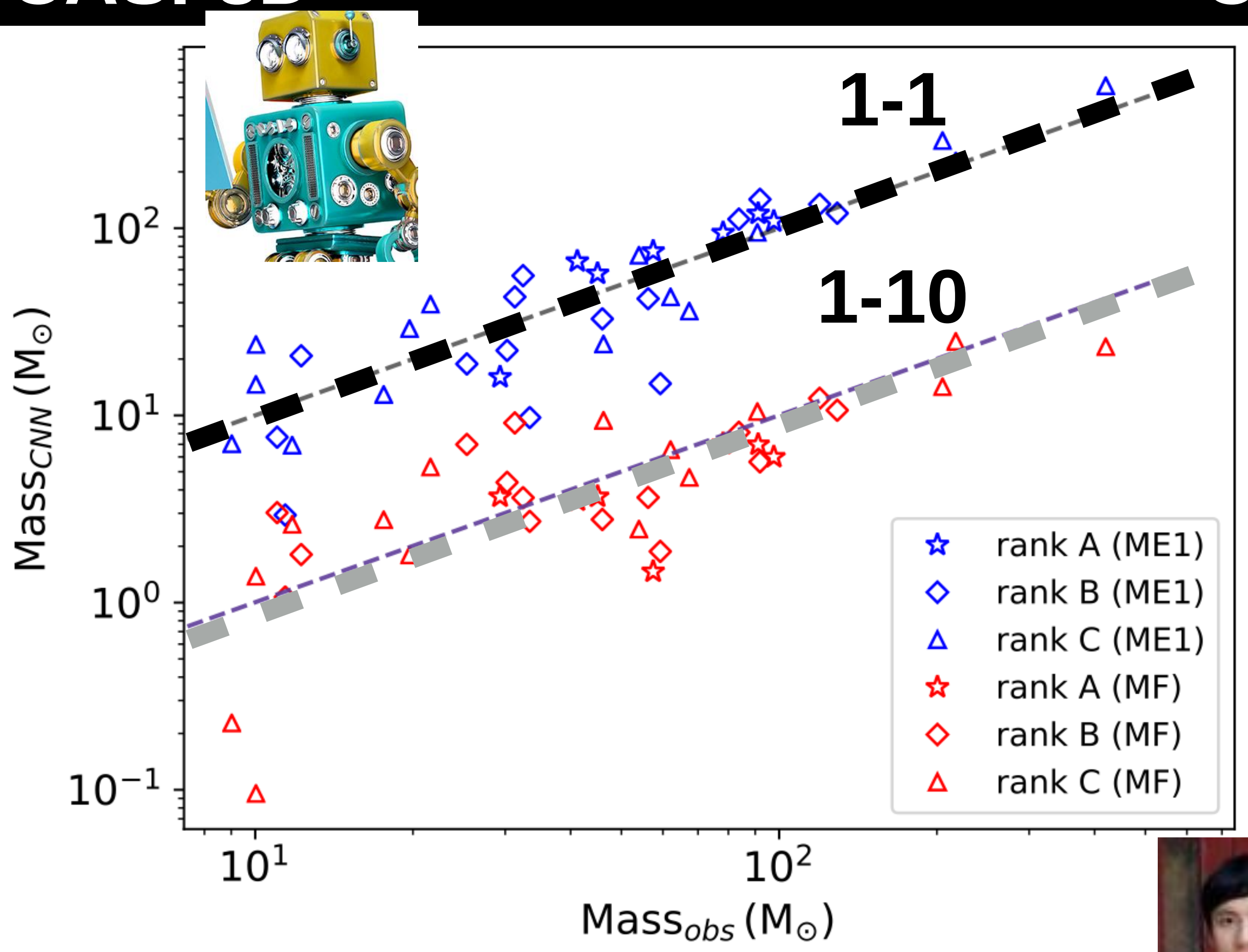
- Model MF trains on the fraction of mass contained in voxels as mapped into spectral space
  - Learns mapping between emission and true feedback mass



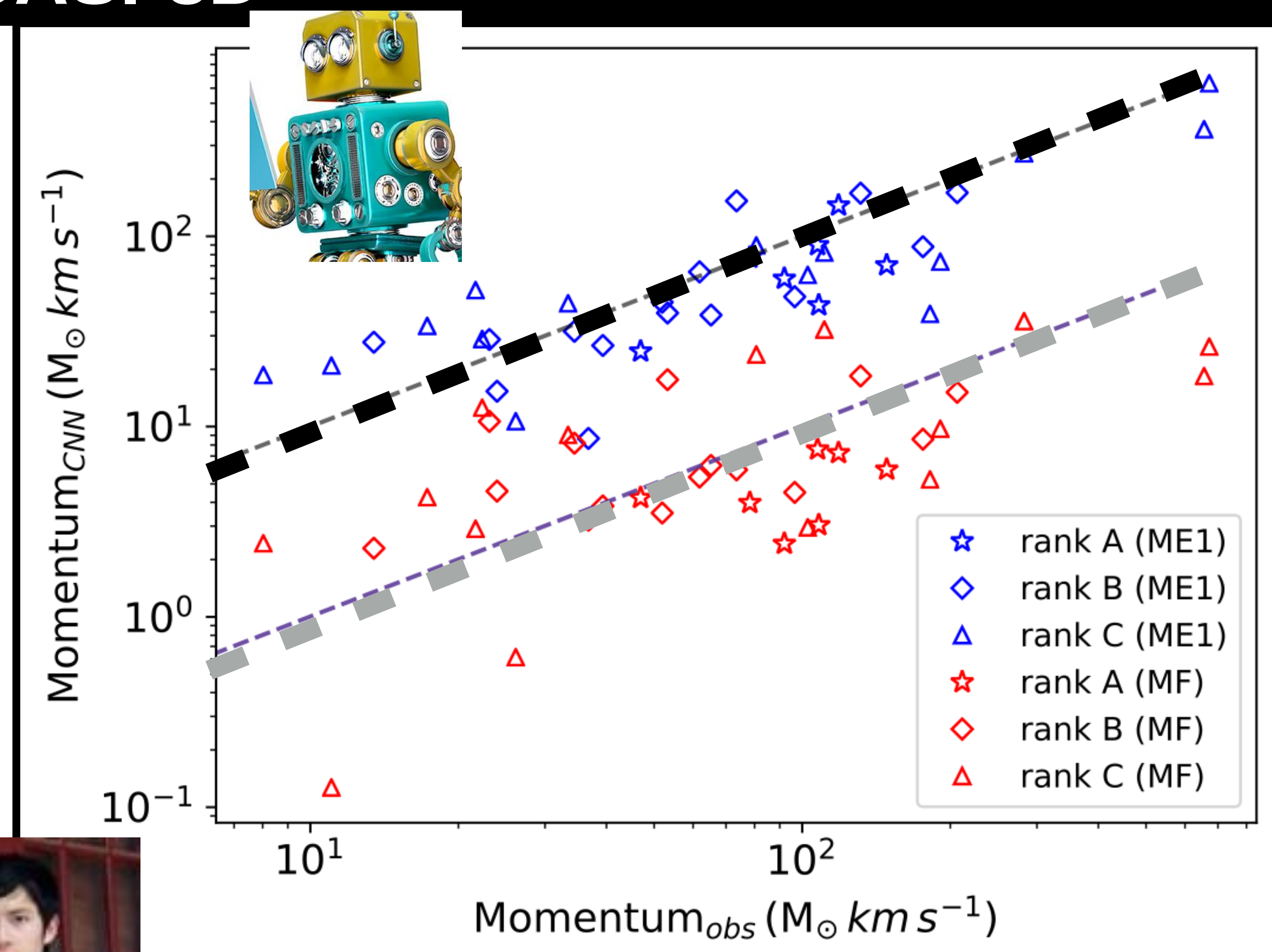
Xu et al 2020a

A Tale of Two Models: ME1 & MF

CASI-3D



CASI-3D



Mass of Human ID'd Shell

Mom. of Human ID'd Shell

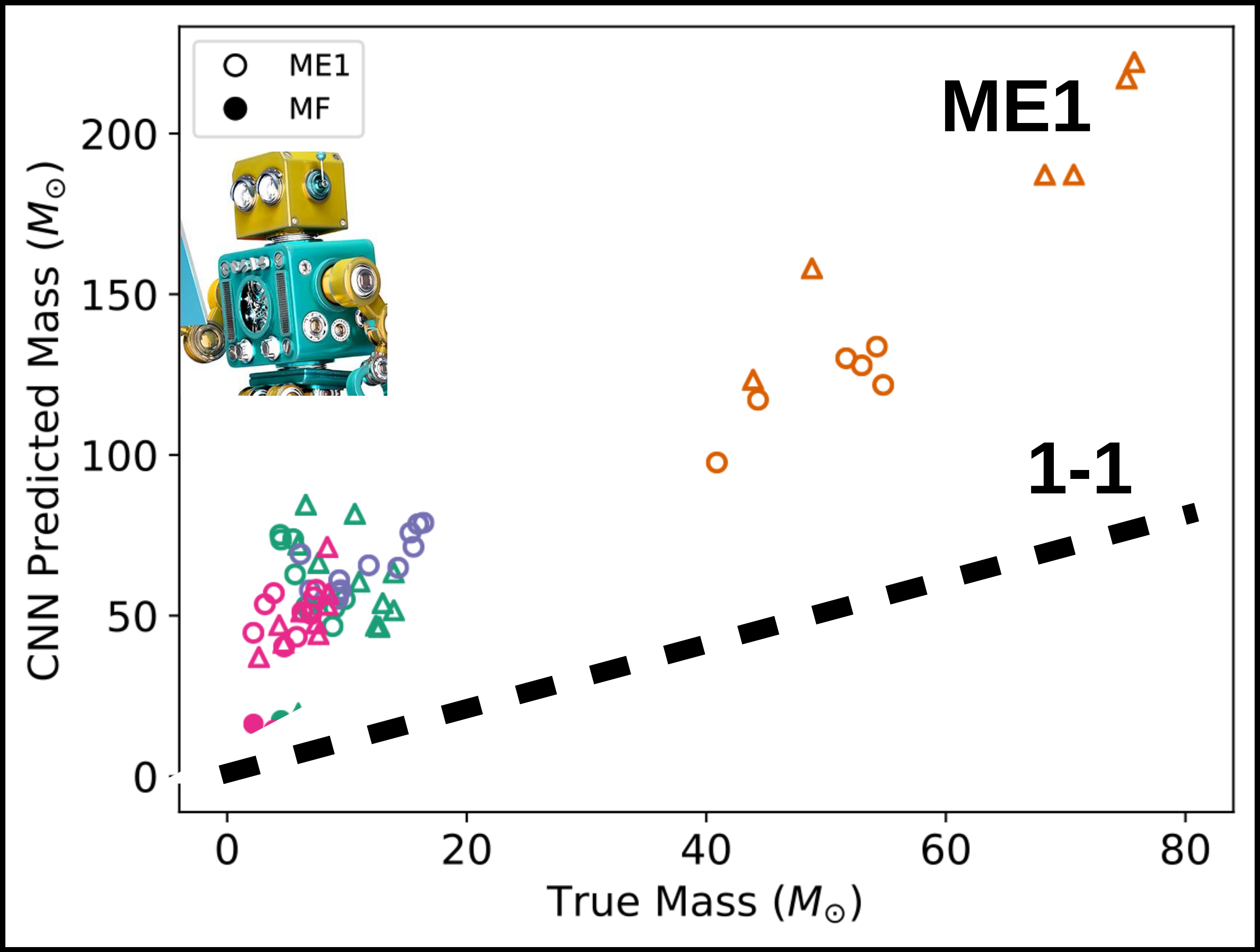
- ME1 agrees very well with previous 'by eye' estimates.
- Model MF predicts 10x lower mass & momentum
- Which to believe??

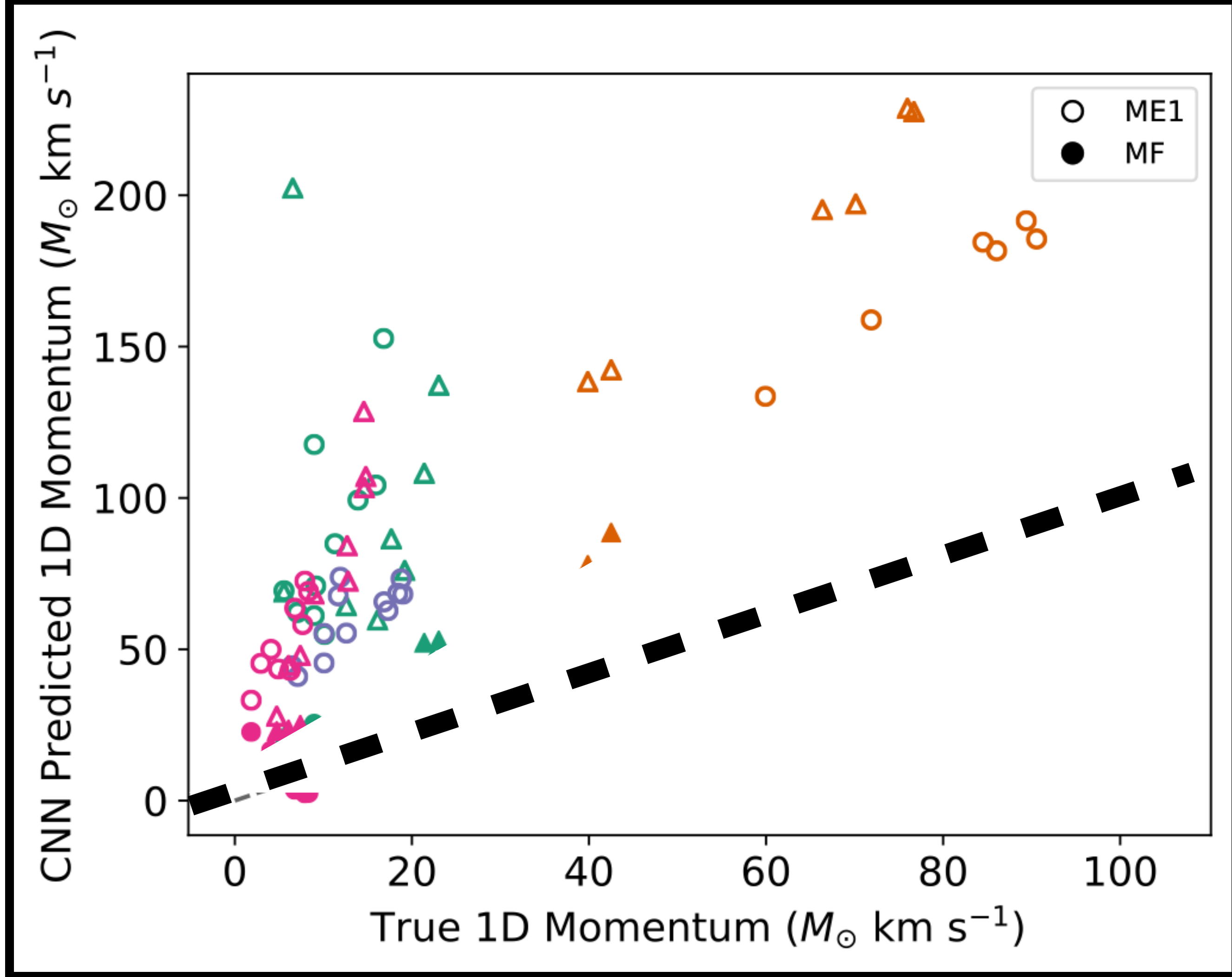
兩種方式估計結果不同

Xu et al 2020a

Human Error

Xu et al 2020a





True Mass (Simulation) **True Momentum (Simulation)**

- **Line-of-sight emission not associated with feedback increases estimated mass and momentum by a factor of 10!**

MF 經過學習，得到正確的質量與動量結果不同

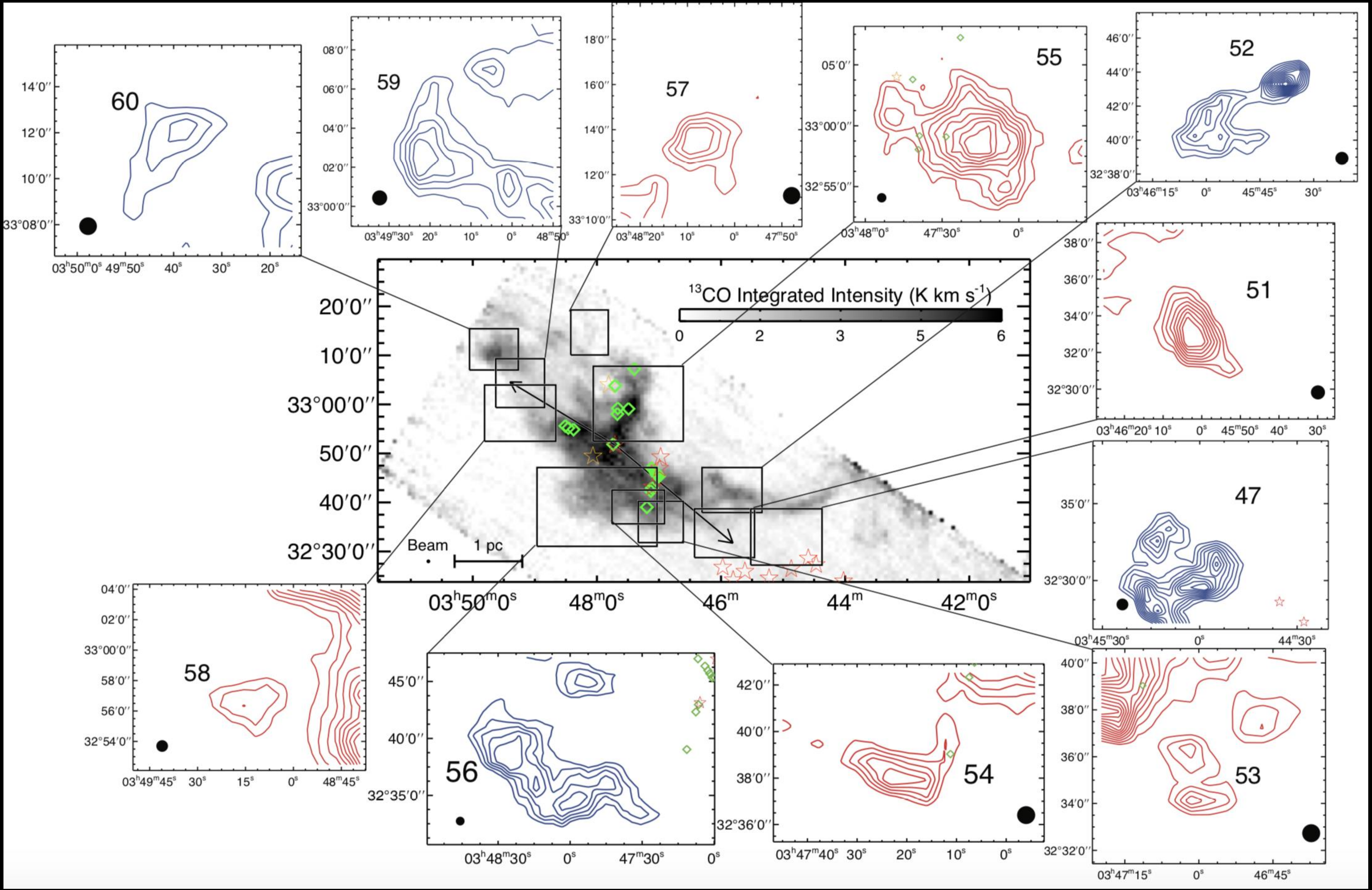
- **CASI-3D ME1 represents how humans find bubbles.**

Protostellar Outflows

應用在指認原恆星的噴流

Human-Identified Outflows

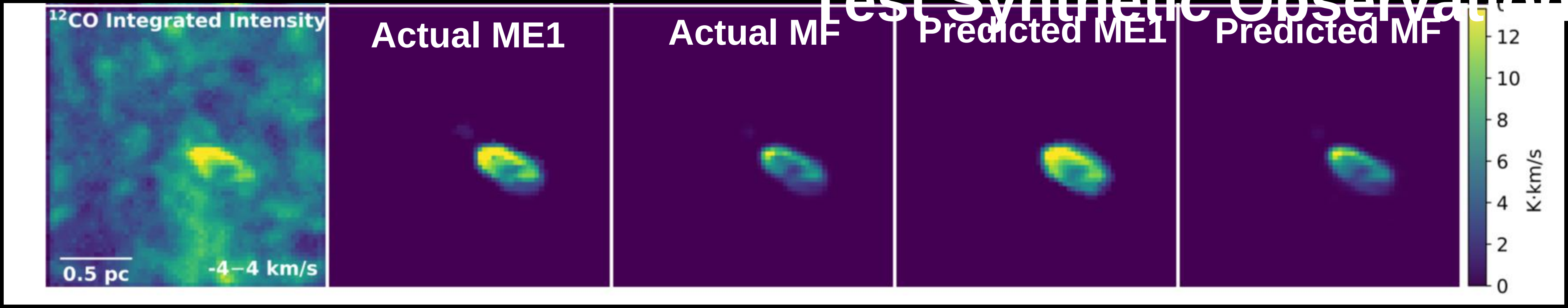
Arce et al. 2010



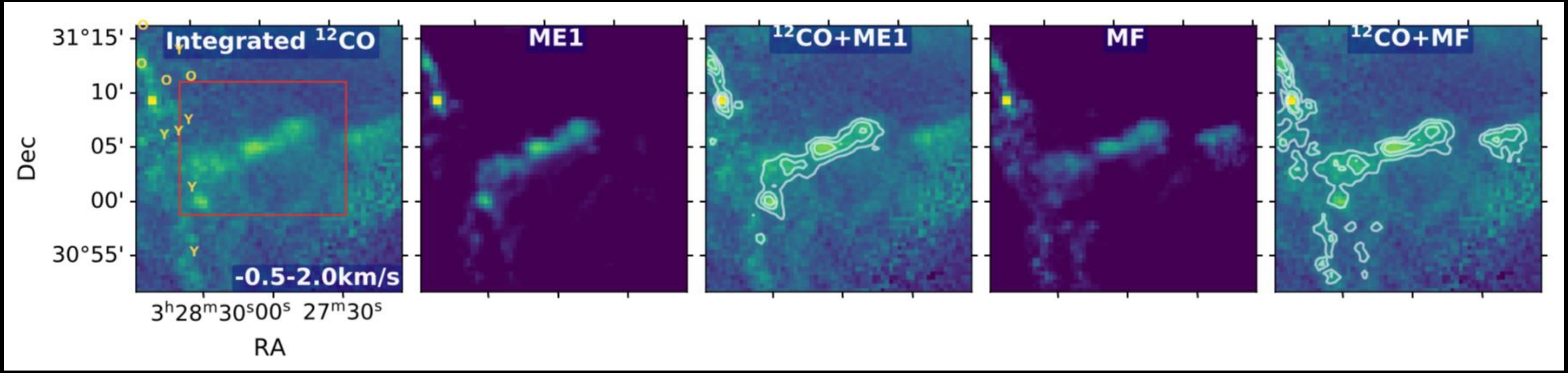
Re-Train CASI-3D for Outflows

Xu et al 2020b

Test Synthetic Observation



Actual Perseus Outflow

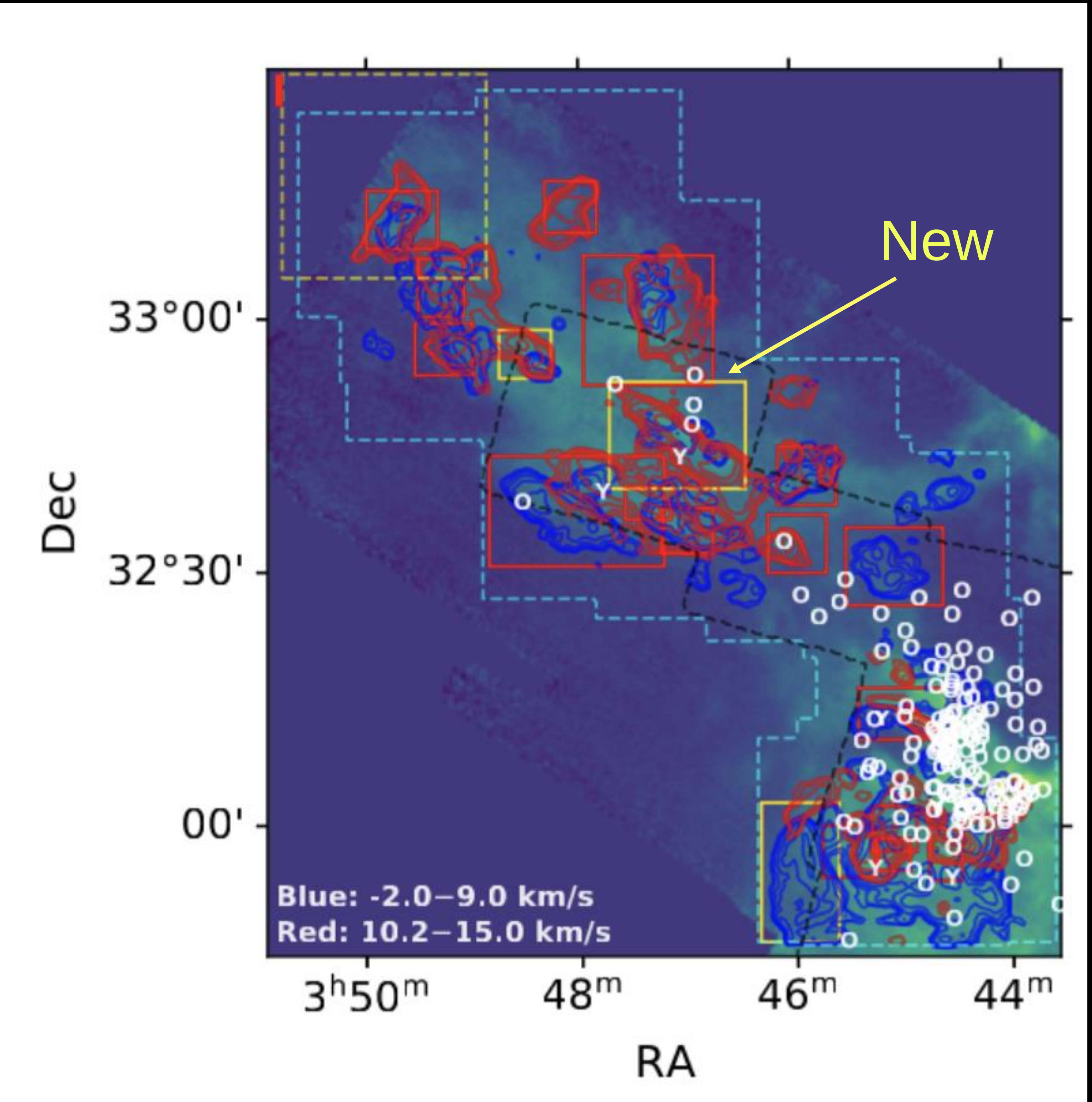


Protostellar Outflows

機器學習找到的噴流

Machine-Identified Outflows

- **Identifies all 60 known visually identified outflows**
  已知60個全找到
- **Identifies 20 new outflows!**
  另發現20個
- **Identifies outflows in confused regions!**
  即使在複雜區域…



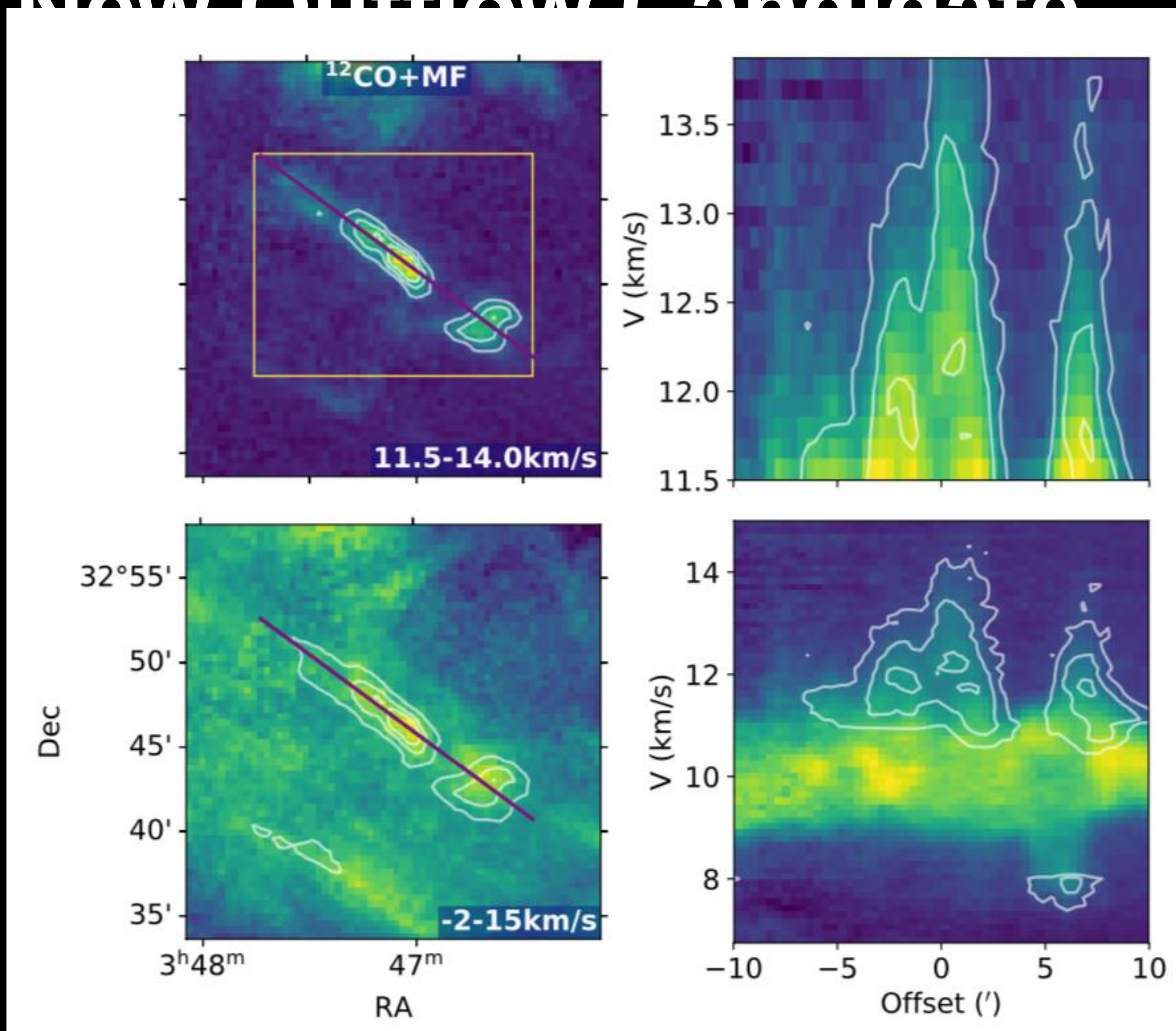
Y= young star
O = older young star

Cluster With ~100 young stars

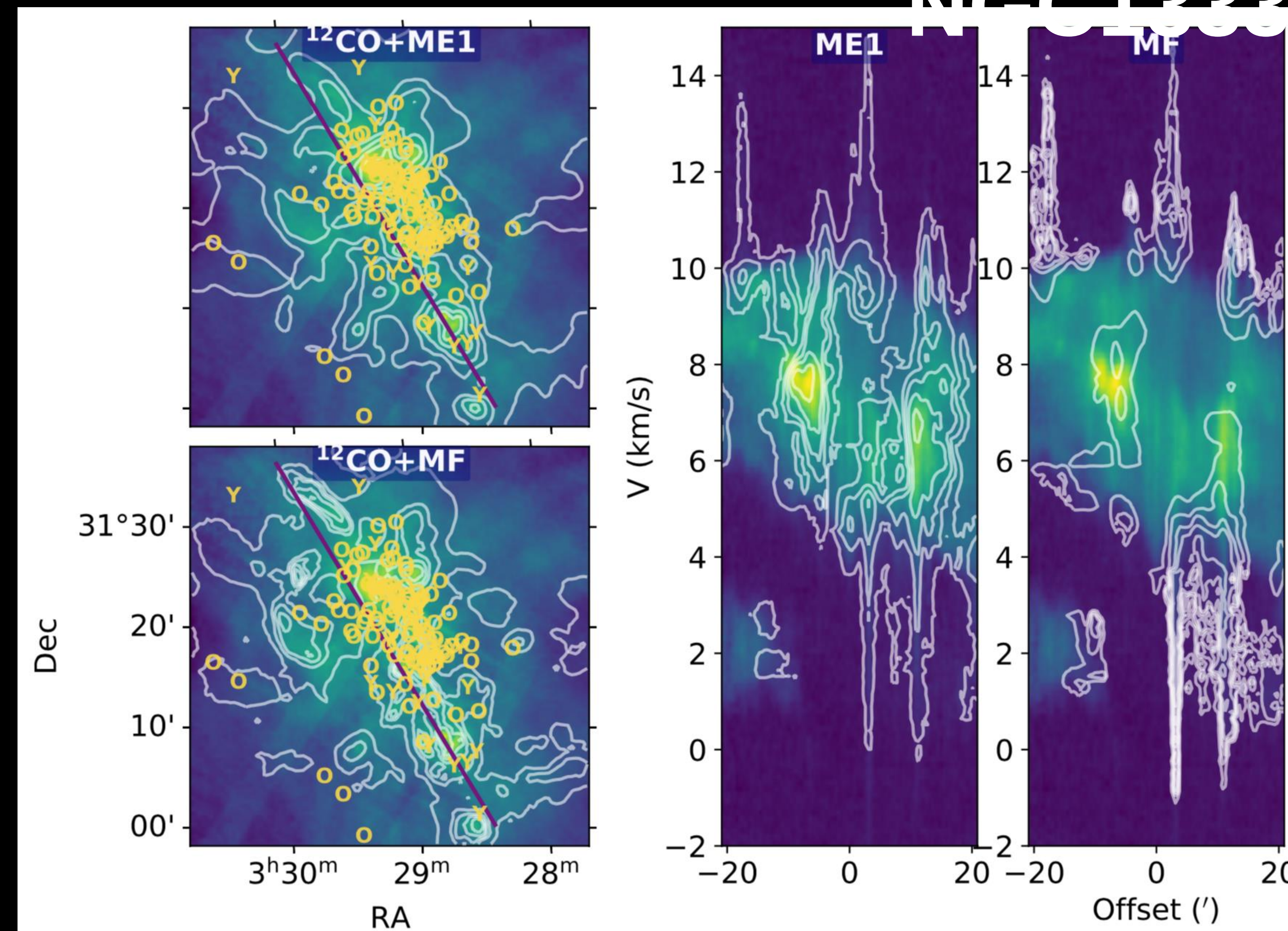
Xu et al 2020b

Protostellar Outflows

New Outflow Candidate



NGC1333

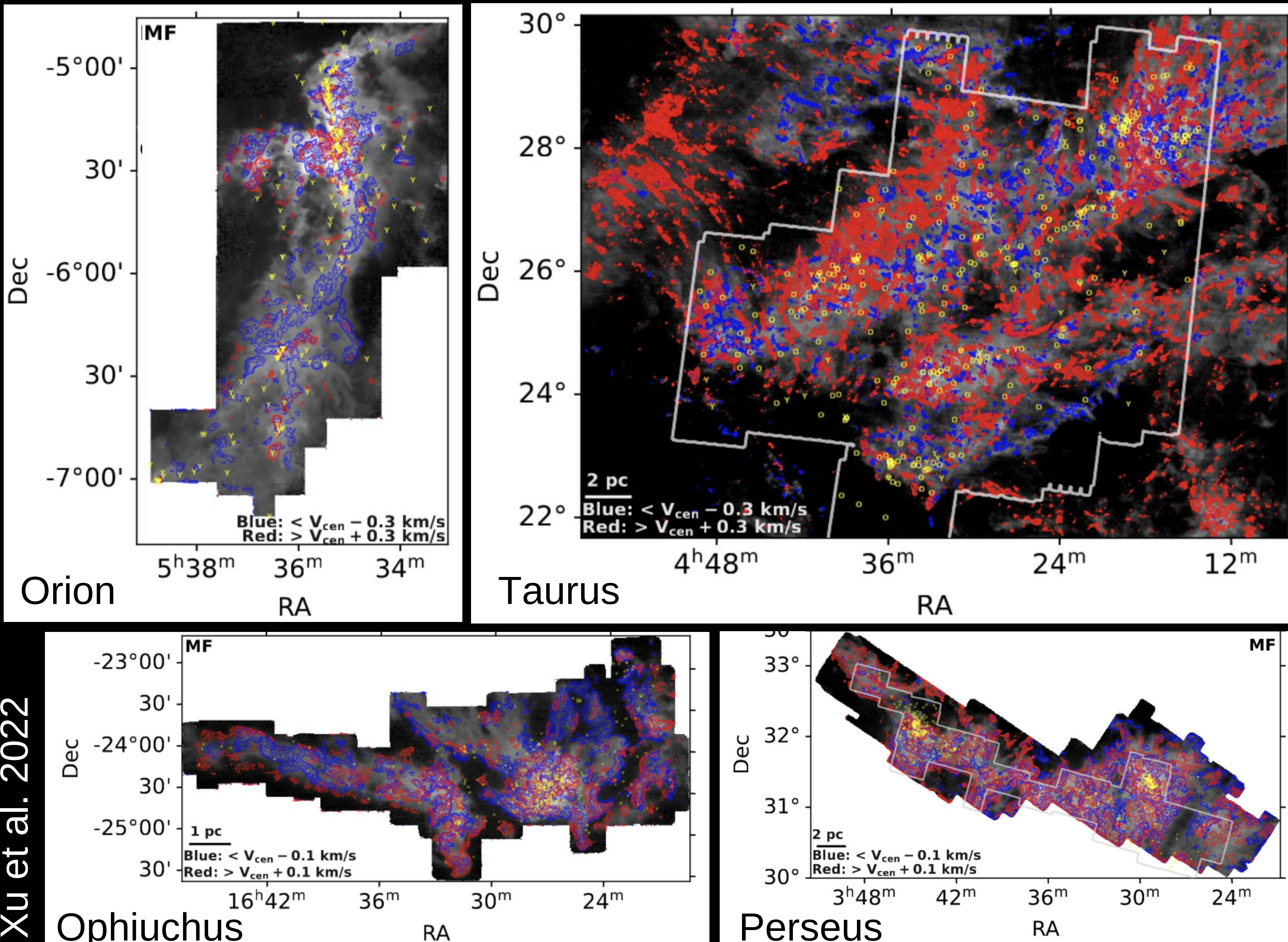


- **Excludes most cloud emission; allows more exact mass, momentum and energy estimates.**

- **Properties of individual outflows are similar due to cancelation of errors but total outflow impact is underestimated by at least 30% (missing outflows).**

Xu et al 2020b

Cloud Outflow Survey



Xu et al. 2022

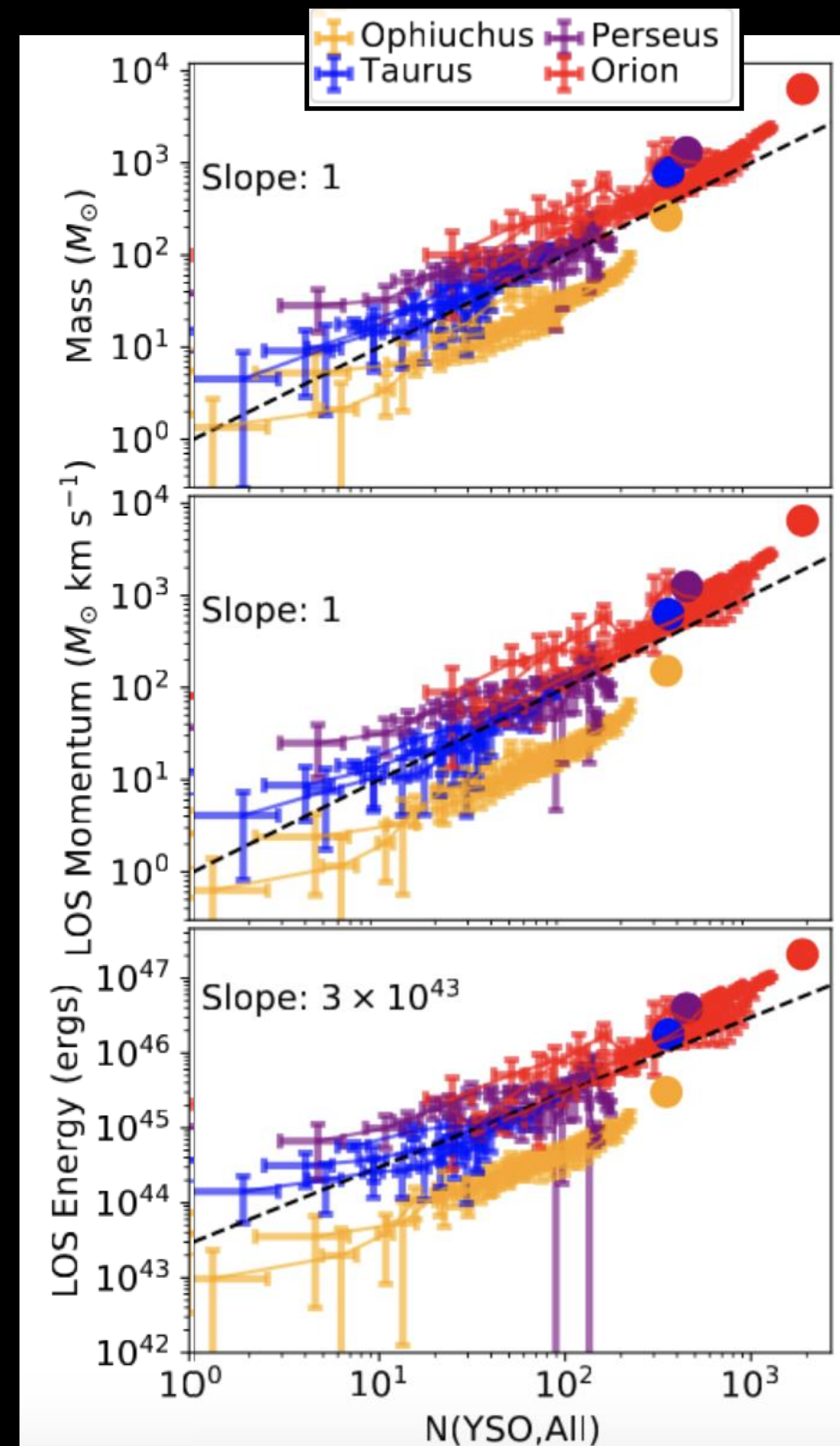
Feedback Scales with NYSO

- All follow same feedback-young stellar object (YSO) relationship

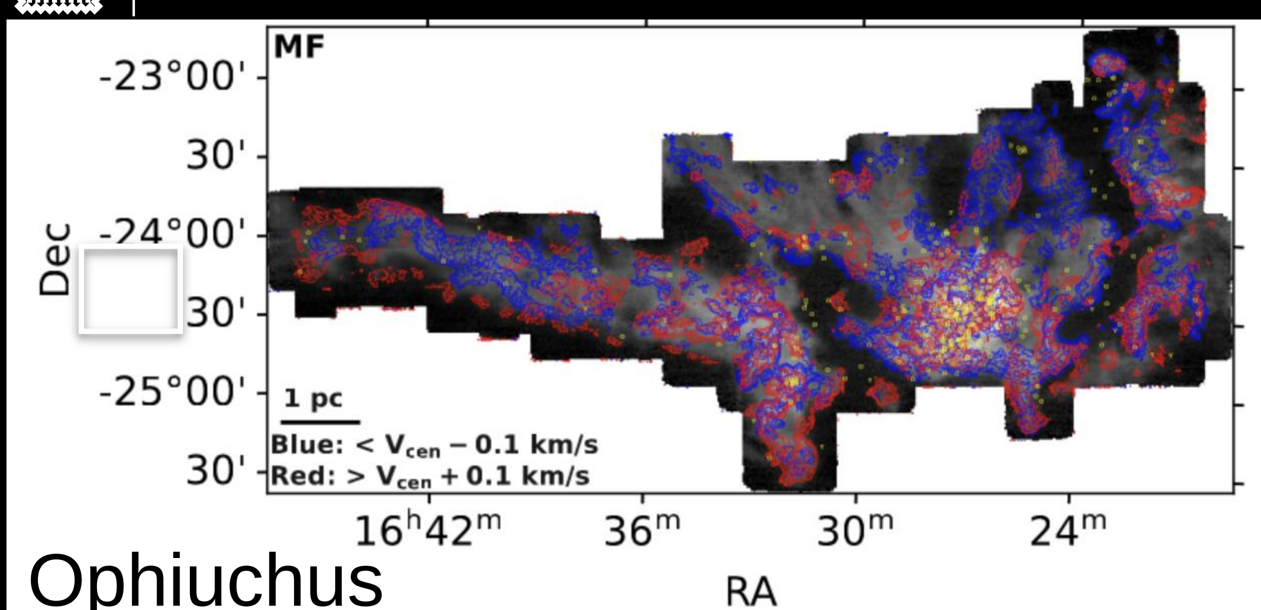
- ~ 1 Msun outflow mass per YSO
- ~ 1 Msun km/s outflow momentum per YSO 每顆年輕恆星噴出1個太陽質量的物質以及1 Msun km/s 的動量

- Energy injection offsets turbulent dissipation in all 4 regions
提供的能量足夠湍流消耗

- Ophiuchus is under-performing
□ ▦♀□ 但蛇夫座恆星形成區稍有不足

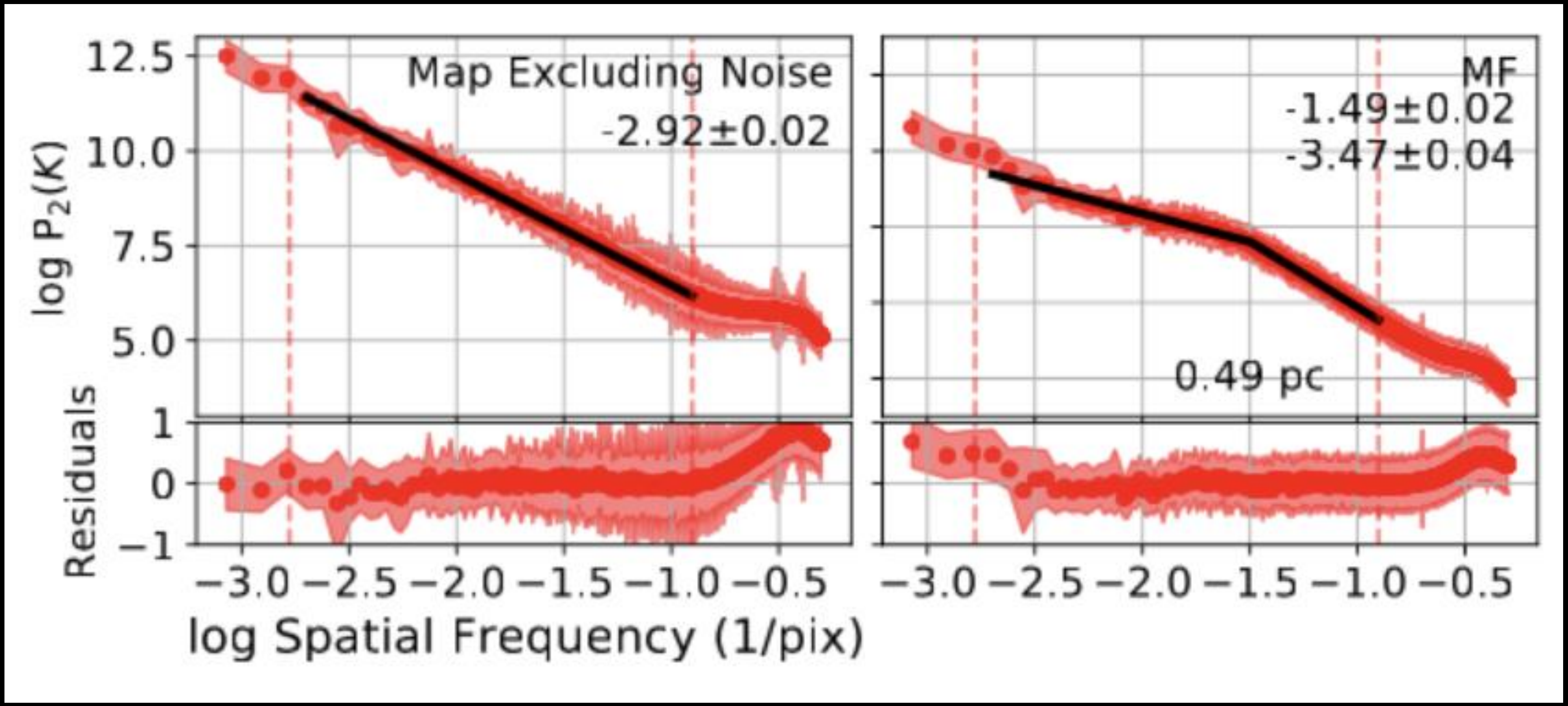


Xu et al. 2022



Characteristic Input Scale

Spatial Power Spectrum (SPS)



Xu et al 2022

- **SPS of the outflow emission shows a break in all 4 clouds**
- **Interpretation: outflow size and momentum injection scale**
- **Break point ranges from 0.27 pc (Orion) - 0.65 pc (Perseus)**

Star Formation is Messy

- **Protostellar outflows are important in clouds from Taurus to Orion**

- **Feedback can offset turbulent dissipation: star-forming regions can self-regulate**
  - **Young stars influence their forming neighbors**
  - **Star formation lasts longer…**

以致於延長了恆星形成的
時間尺度

- **Astronomers can move beyond catalogs (and visual inspection!) to 3D images**

30 Doradus
Credit: NASA, ESA

What’s next?

STAR FORmation in **G**aseous **E**nvironments (**STARFORGE**)

Grudic et al. 2021, Guszejnov et al. 2021, Grudic ea 2022

Conclusions

- CASI-3D is a general CNN method:
  - Can be used to identify structure in spectral cubes
  - Can be used to estimate observational biases like radiative transfer effects / opacity & projection effects
- Feedback is **everywhere** in star-forming clouds
- Previous bubble/wind feedback impact over-estimated (**mass and momentum lower by a factor of 10!**)

  之前高估了恆星風回饋的影響
- Previous outflow feedback impact under-estimated due to **missing at least 30% of outflow activity**

  而低估了恆星噴流的影響
- Impact is **significant** compared to cloud kinetic energy

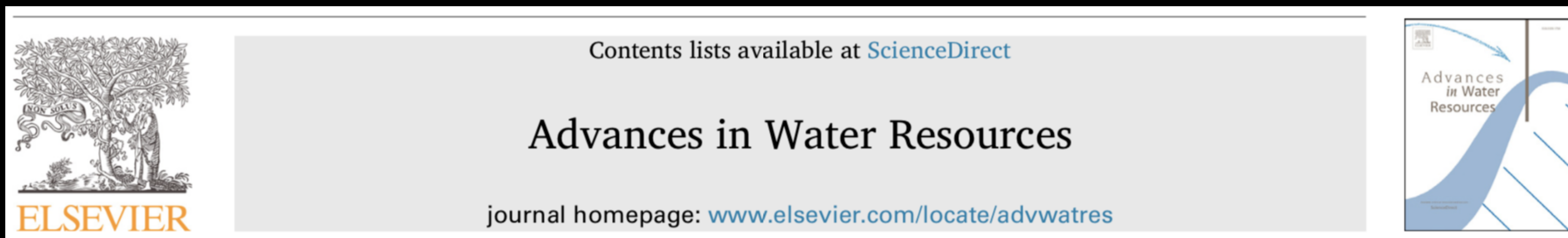
Thank you!

https://gitlab.com/casi-project/casi-2d

https://gitlab.com/casi-project/casi-3d

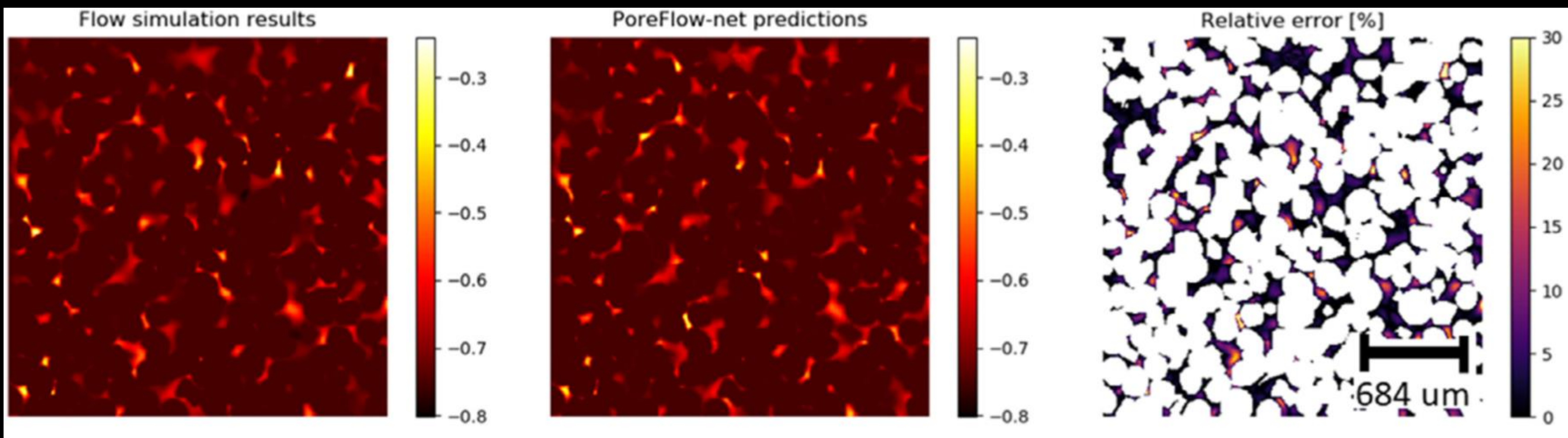
Predicting Fluid Flow Through Porous Media

在孔狀介質中預測流向

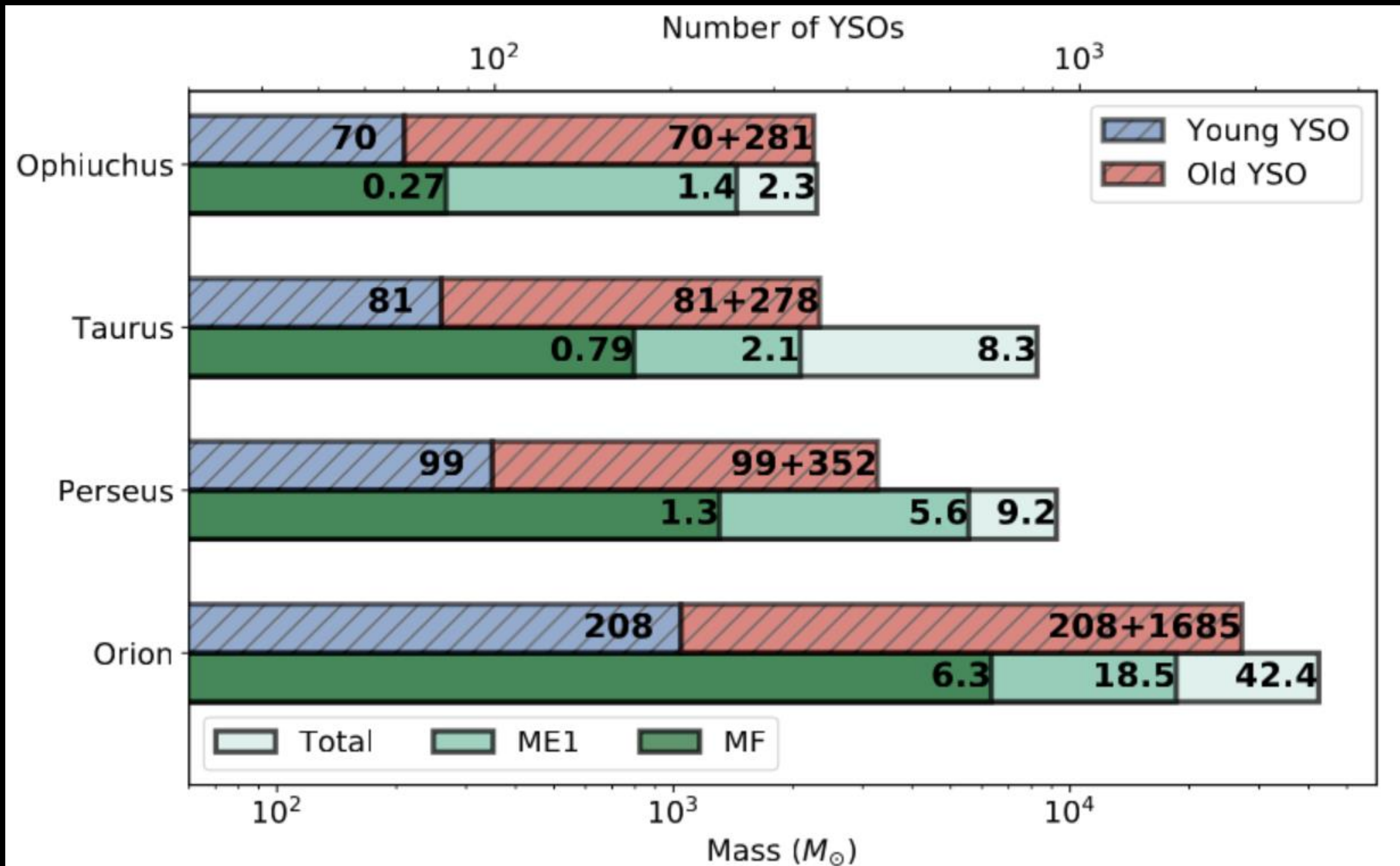


PoreFlow-Net

Santos, Xu et al 2020



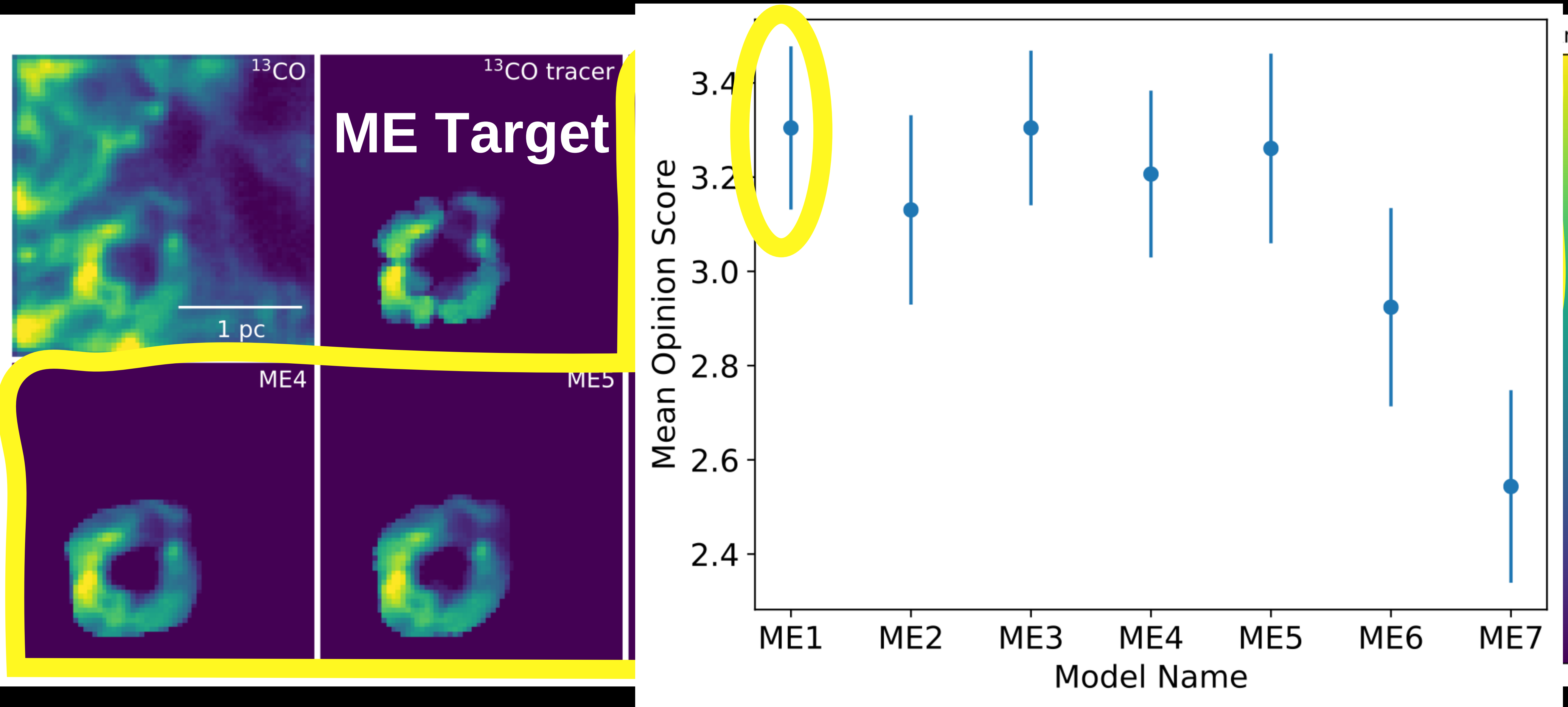
- **Accurate flow predictions in less than 1sec**
- **Results for granular rocks, carbonates and "consolidated media"**
- **"shows the successful application of a disruptive technology (physics-based training of machine learning models) to the digital rock physics community"**



Number of YSOs
10^2
10^3
Young YSO
Old YSO
Ophiuchus
70
70+281
0.27
1.4
2.3
Taurus
81
81+278
0.79
2.1
8.3
Perseus
99
99+352
1.3
5.6
9.2
Orion
208
208+1685
6.3
18.5
42.4
Total
ME1
MF
10^2
10^3
10^4
Mass ($M_\odot$)

Xu et al 2020a

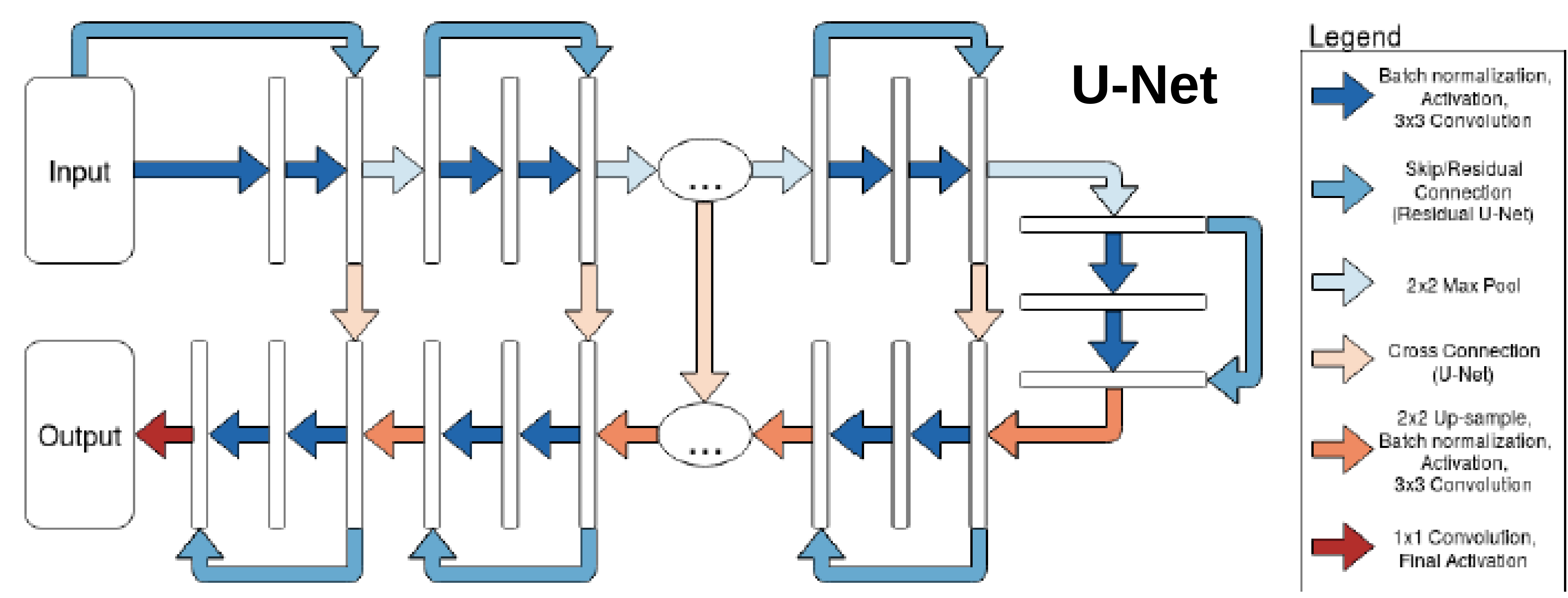
CASI-3D Models



- **Evaluate hyper-parameter choices & training sets**
- **Models ME1-ME7 train on emission associated with feedback**
- **Use Mean Opinion Score to test performance on observed data.**

5: Excellent
4: Good
3: Average
2: Fair
1: Poor

Convolutional Approach to Shell Identification (CASI)



- U-Net architecture
- 4 residual blocks
- Mean squared error loss function
- Input image is 256^2
- 200 epochs, <~5% convergence
- Data: 1/3 training, 1/3 validating, 1/3 testing

https://gitlab.com/casi-project/casi-2d

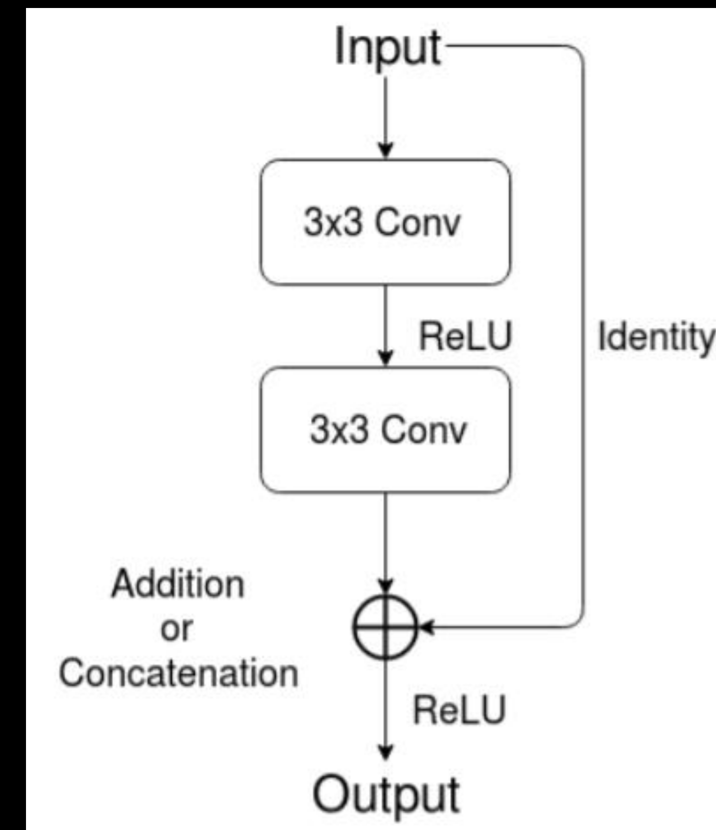
van Oort et al. 2019

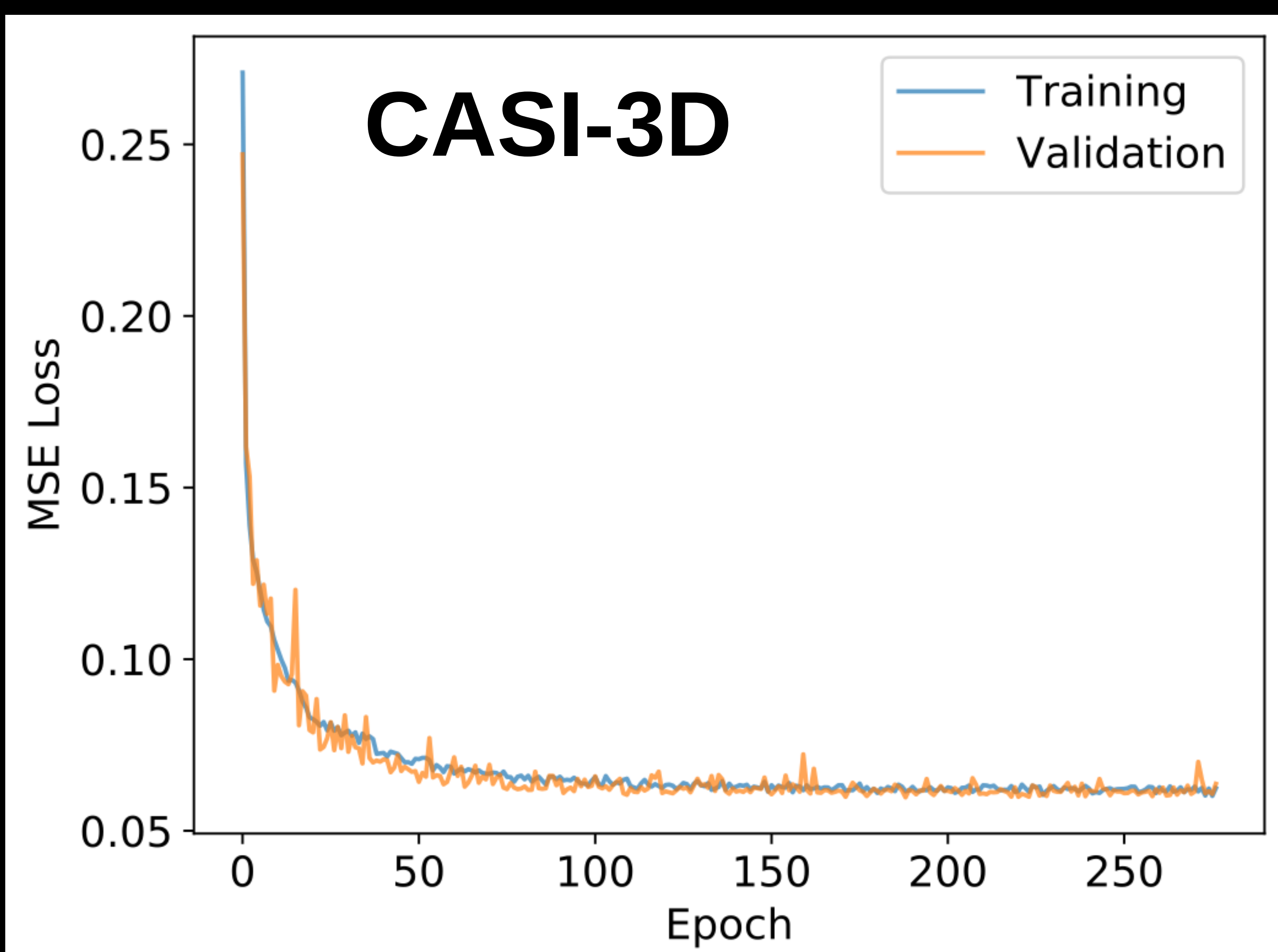
CNN Details

Table 1. Network Hyper-Parameters[a]

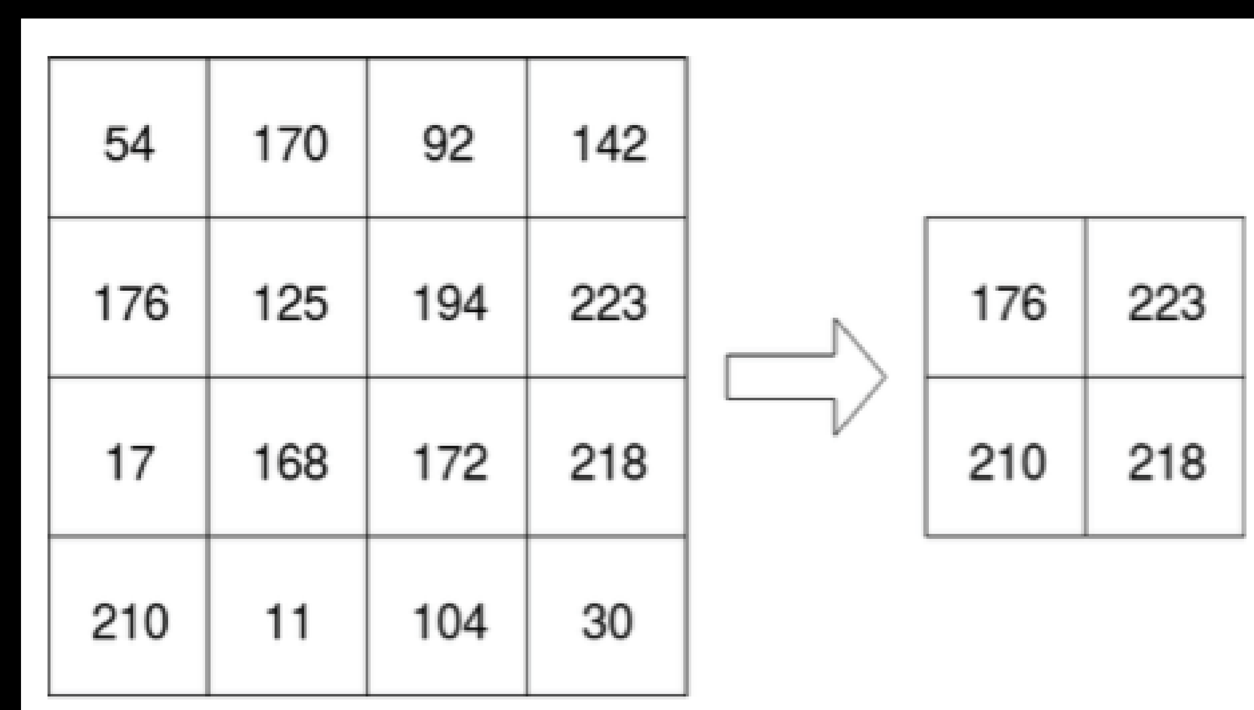
Name	Definition	Symbol
Depth	Blocks in the compressive/decompressive legs.	D
Number of Filters	Filter alloted for each convolution operation.	F
Batch Size	Samples provided at each training iteration.	B
Noise Strength	Standard deviation of the white Gaussian noise applied to network inputs.	σ

Basic Residual Block

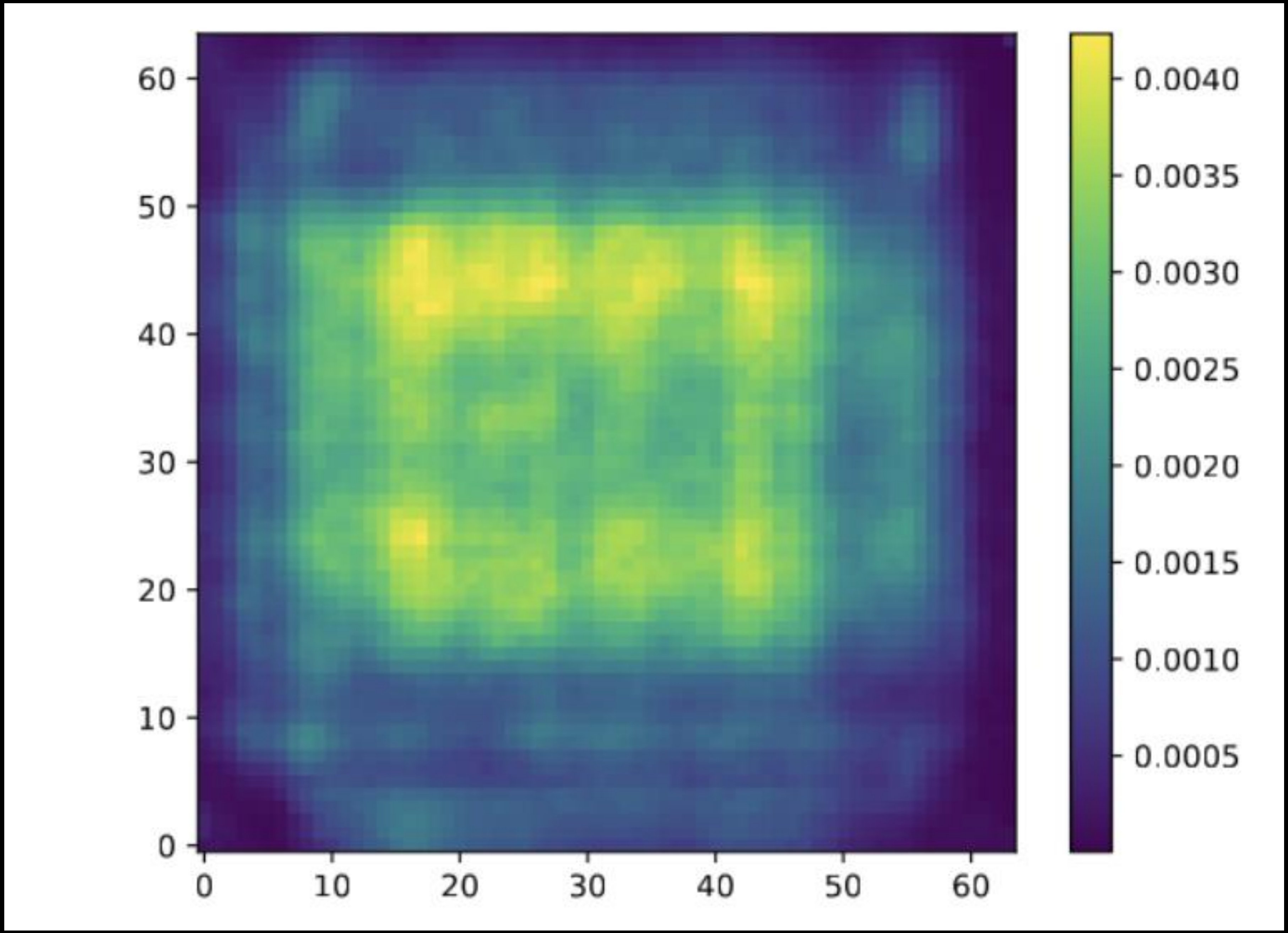




Max pooling



CASI-3D Window Sensitivity

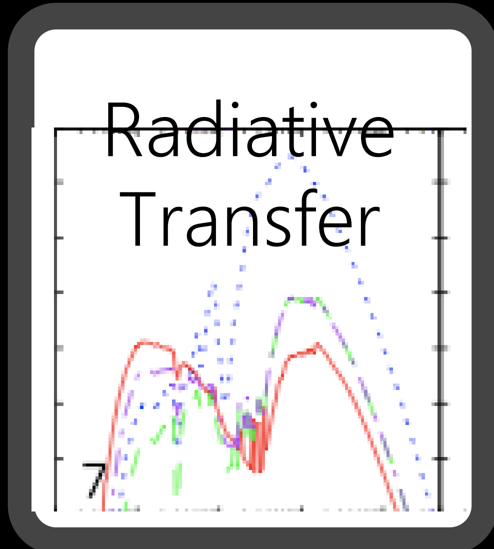


The integrated response of a voxel in the cube: the fraction of voxels predicted to be associated with feedback in the same position in the stack of cubes summed over all velocity channels.

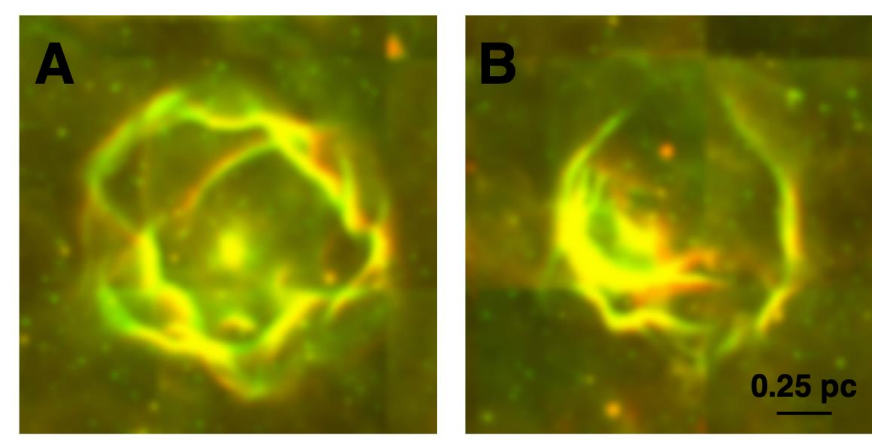
simulation data is a good repre

Synthetic Observations

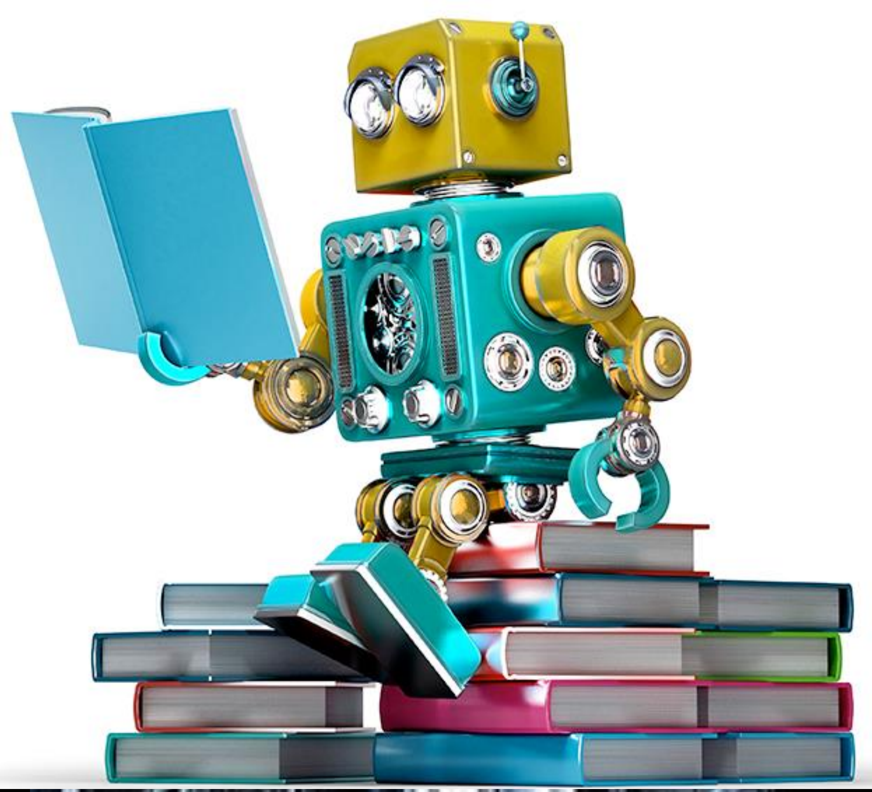
Duo Xu



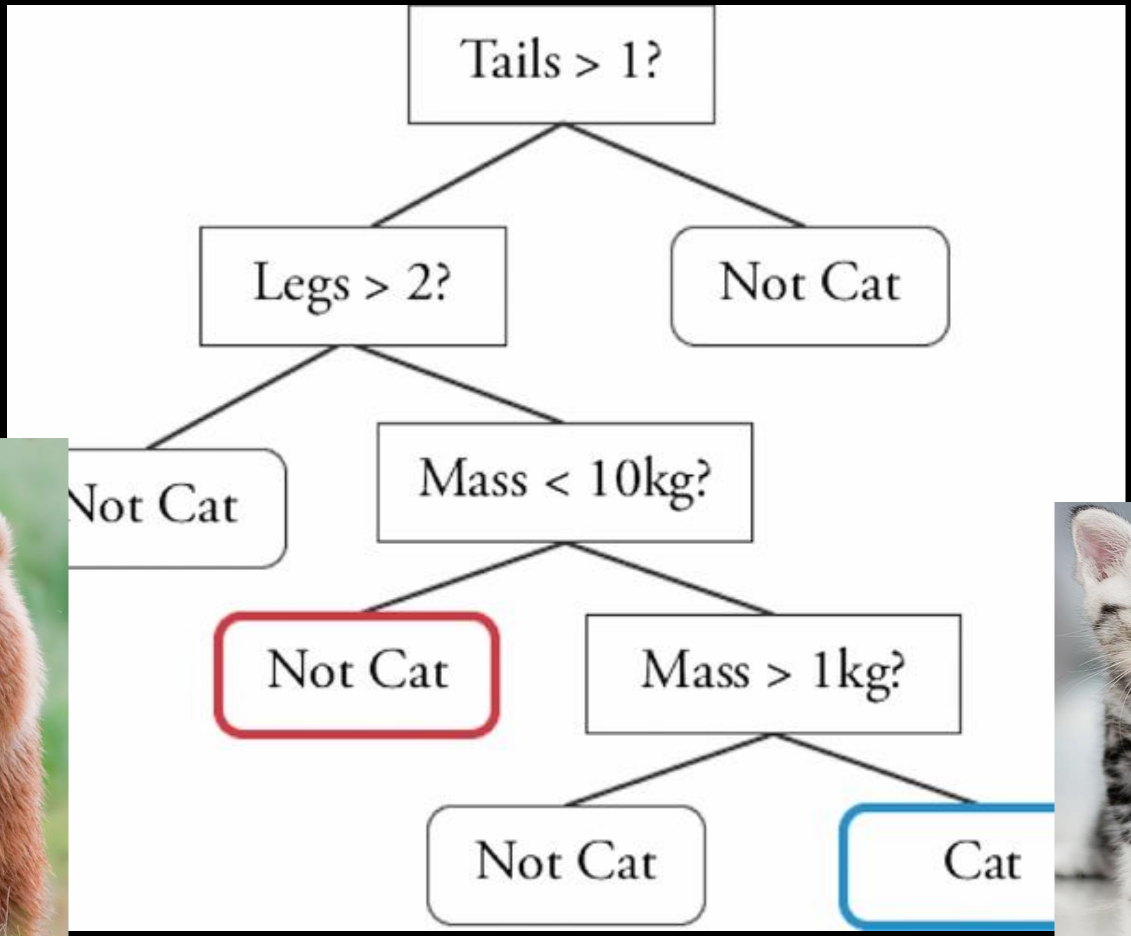
Training Set

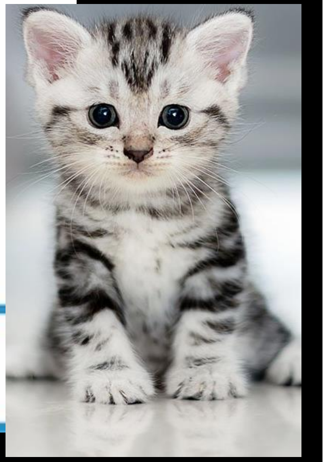


Machine Learning



"Random Forest" Approach

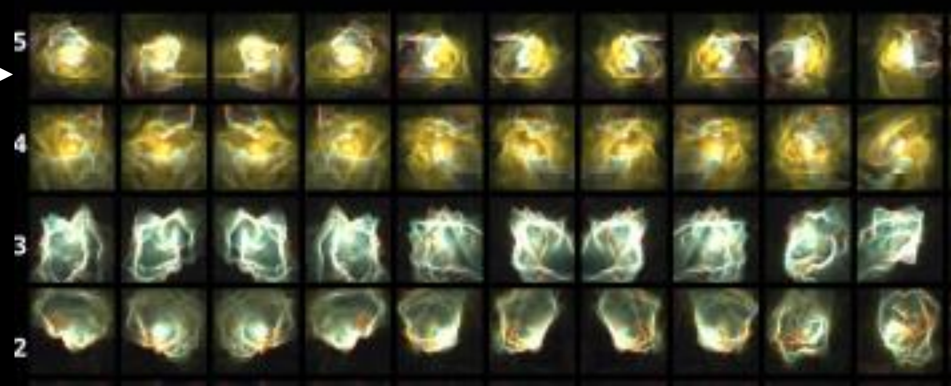




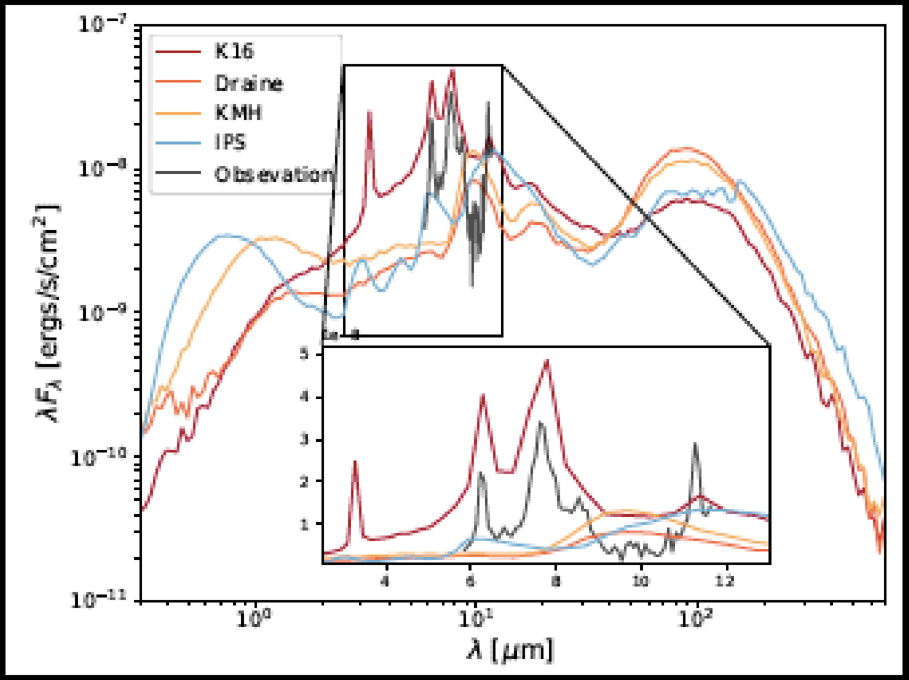
BRUT, Beaumont ea 2014

Training Data

- Synthetic dust emission (3 color)



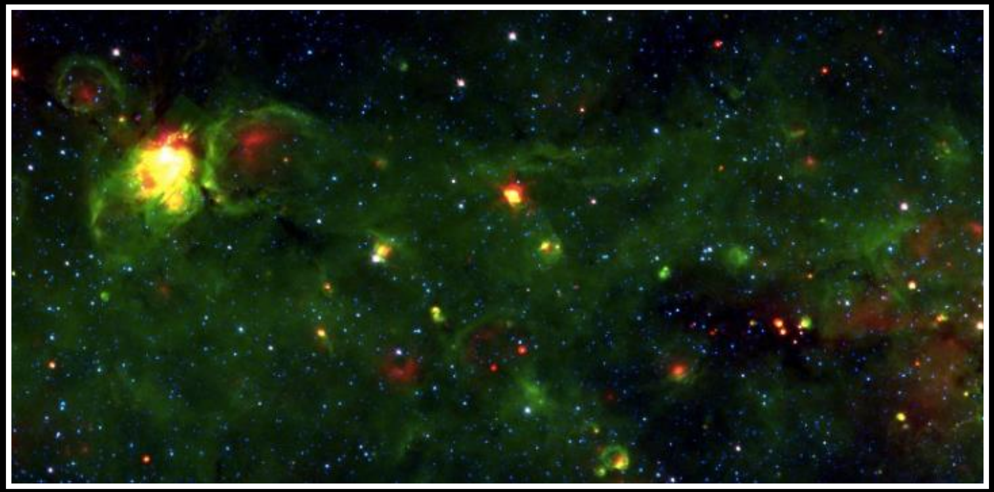
- Dust model with PAHs



- Different times, magnetic field strengths, star properties, viewing angles
- With / without noise
- Centered on shell

+

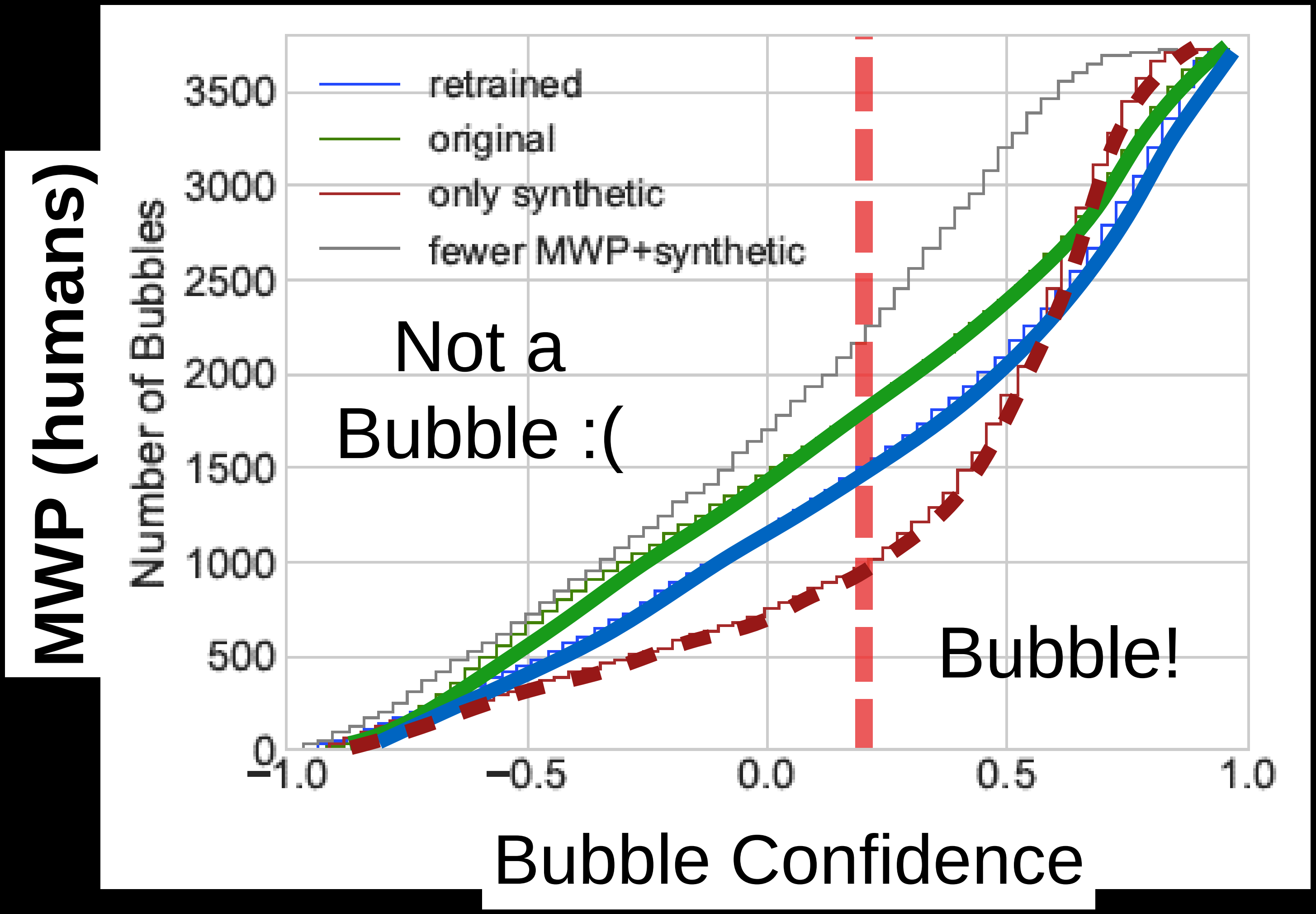
- With / without observational data identifications

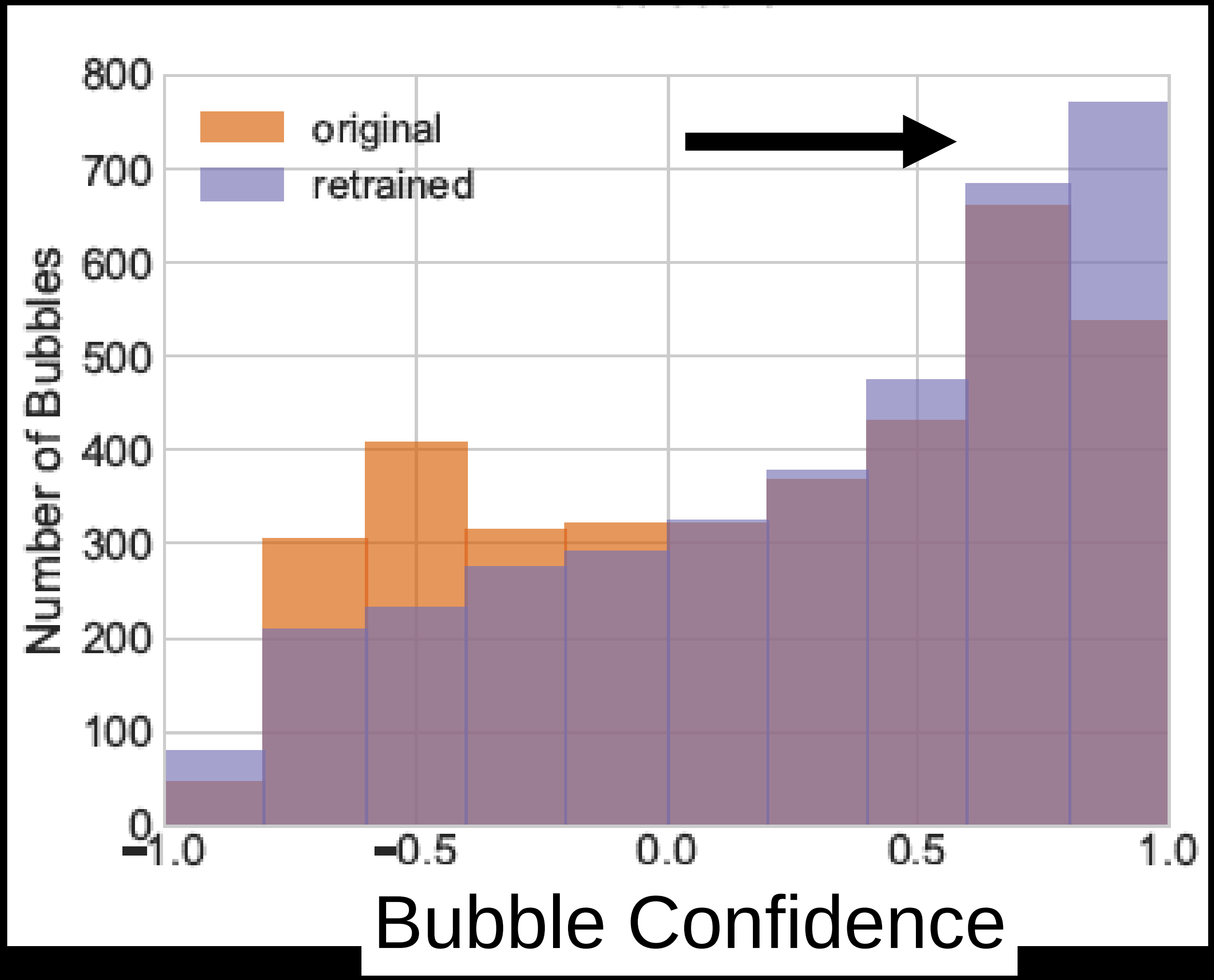


Xu & Offner 2017

Train with Simulations, Apply to Real Data

Xu & Offner 2017

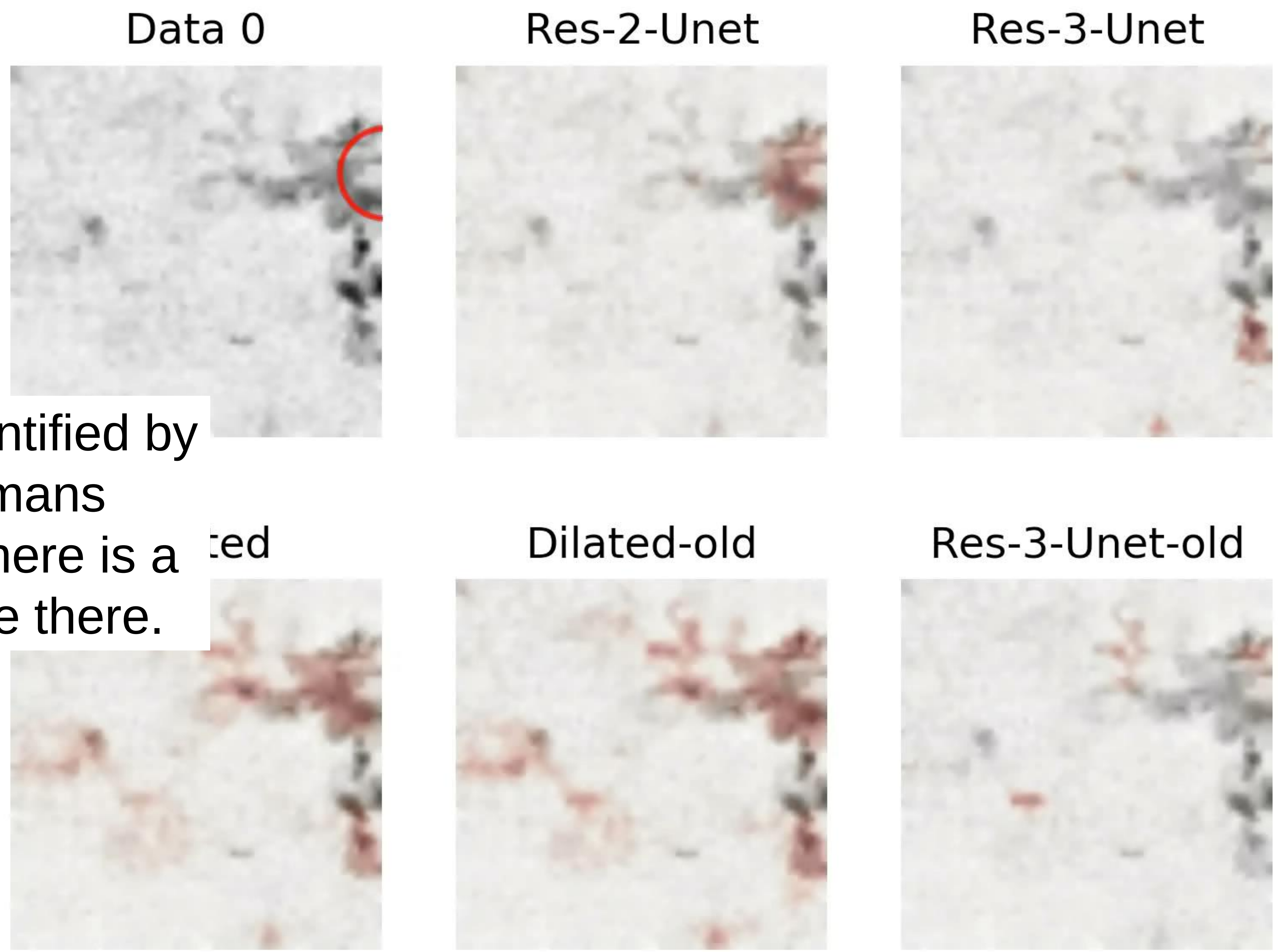




- **Confidence increases for visually identified bubbles.**
- **For simulation-only training, some bubbles (as expected) have lower scores… know your training set!**
- **Best results: test set samples all types of bubbles**

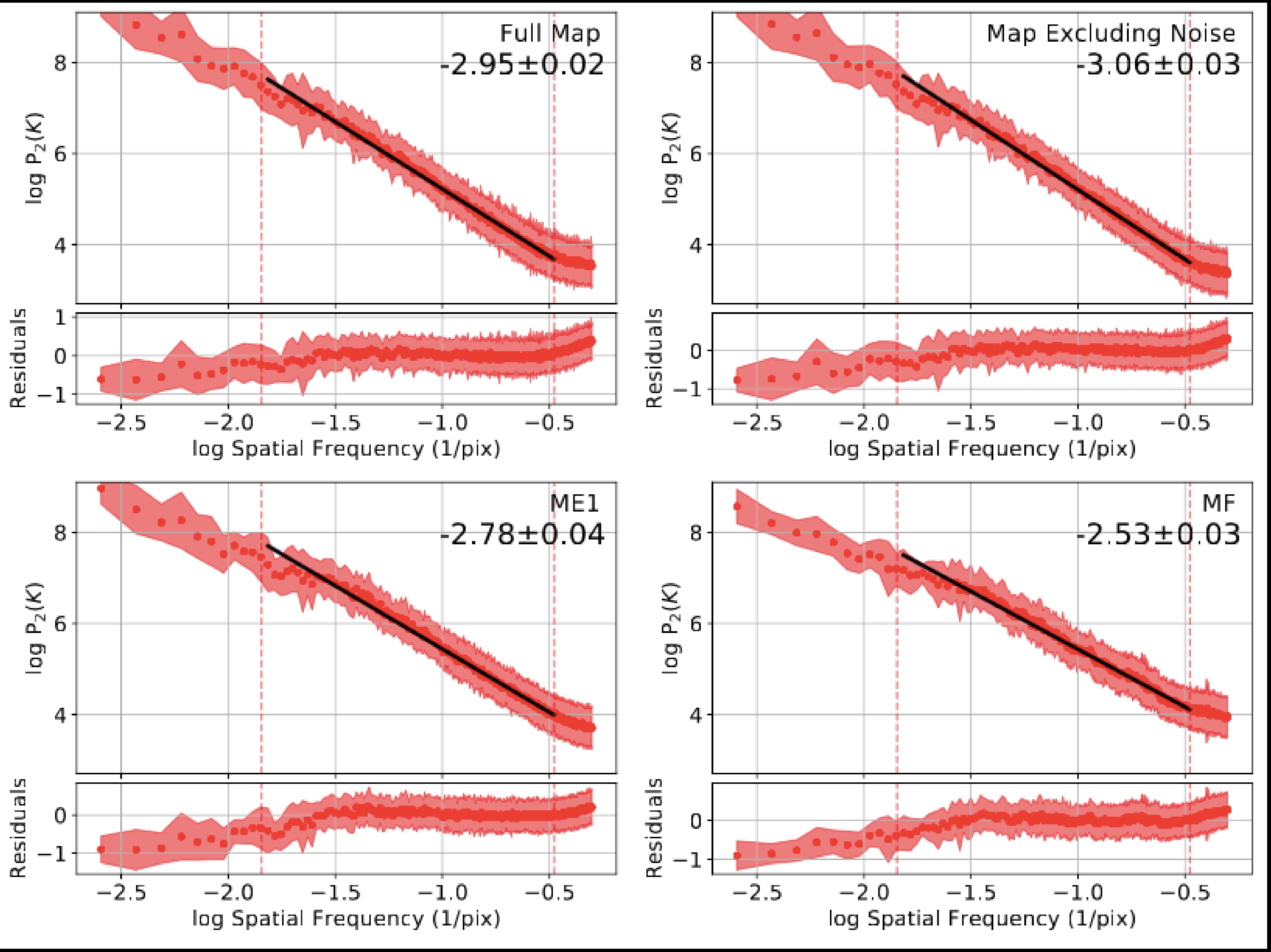
Feedback Recovery (3D)

Taurus
13CO



Not identified by humans
BUT there is a source there.

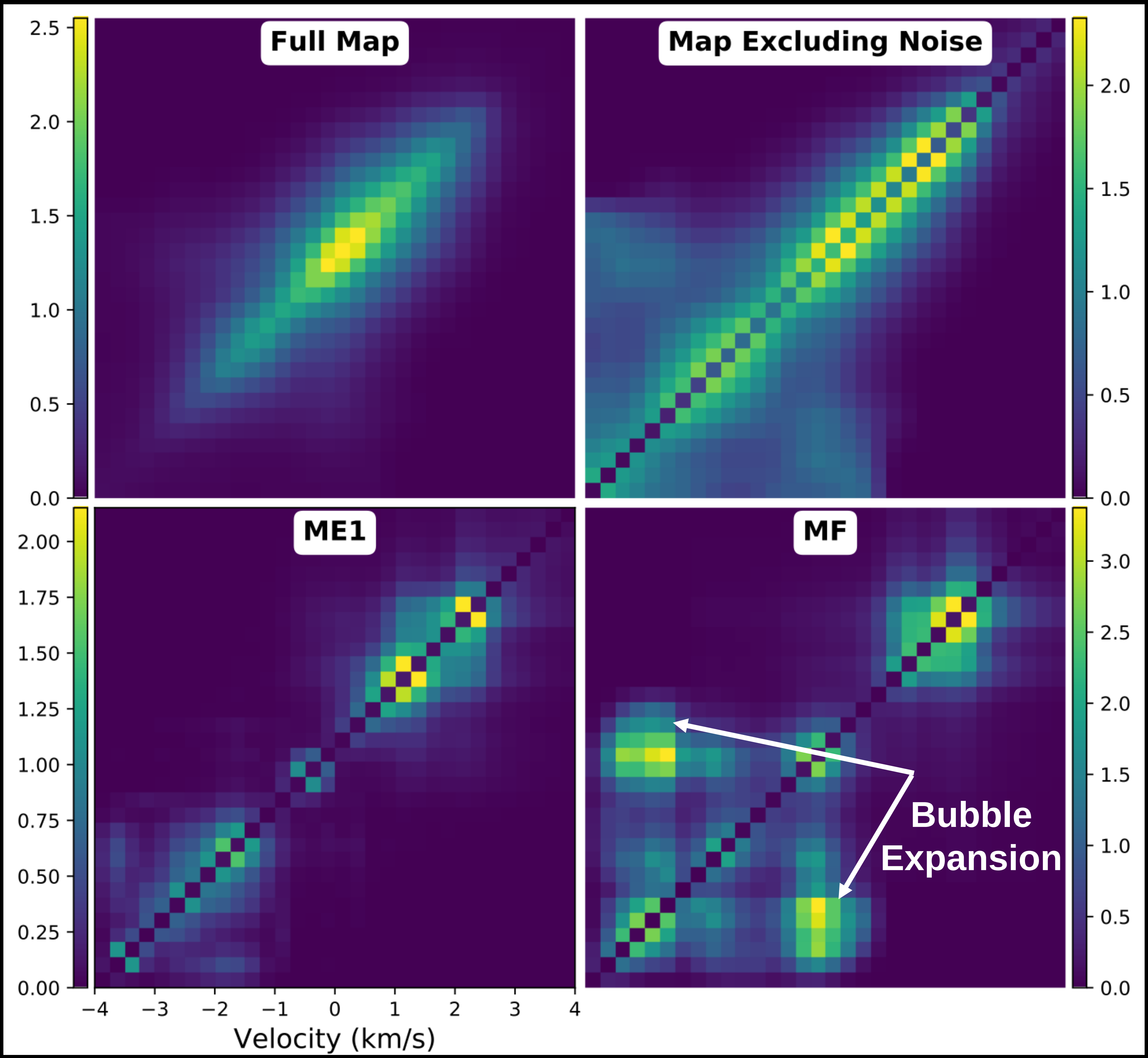
Spatial Power Spectrum (SPS)



- SPS of Taurus intensity flattens with feedback
- Predicted by Offner & Liu 2018 using models.

Xu et al 2020

PCA is Sensitive to Feedback



- **Taurus PCA covariance matrix shows evidence of feedback.**
- **Predicted by Boyden et al. 2016 using models.**

$$C_{jk} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}X_{ik},$$

$$X_{ij} = T(r_i, v_j) - [\sum_{k=1}^{n} T(r_k, v_j)]/n,$$

C = covariance

T = spectral cube

Xu et al 2020