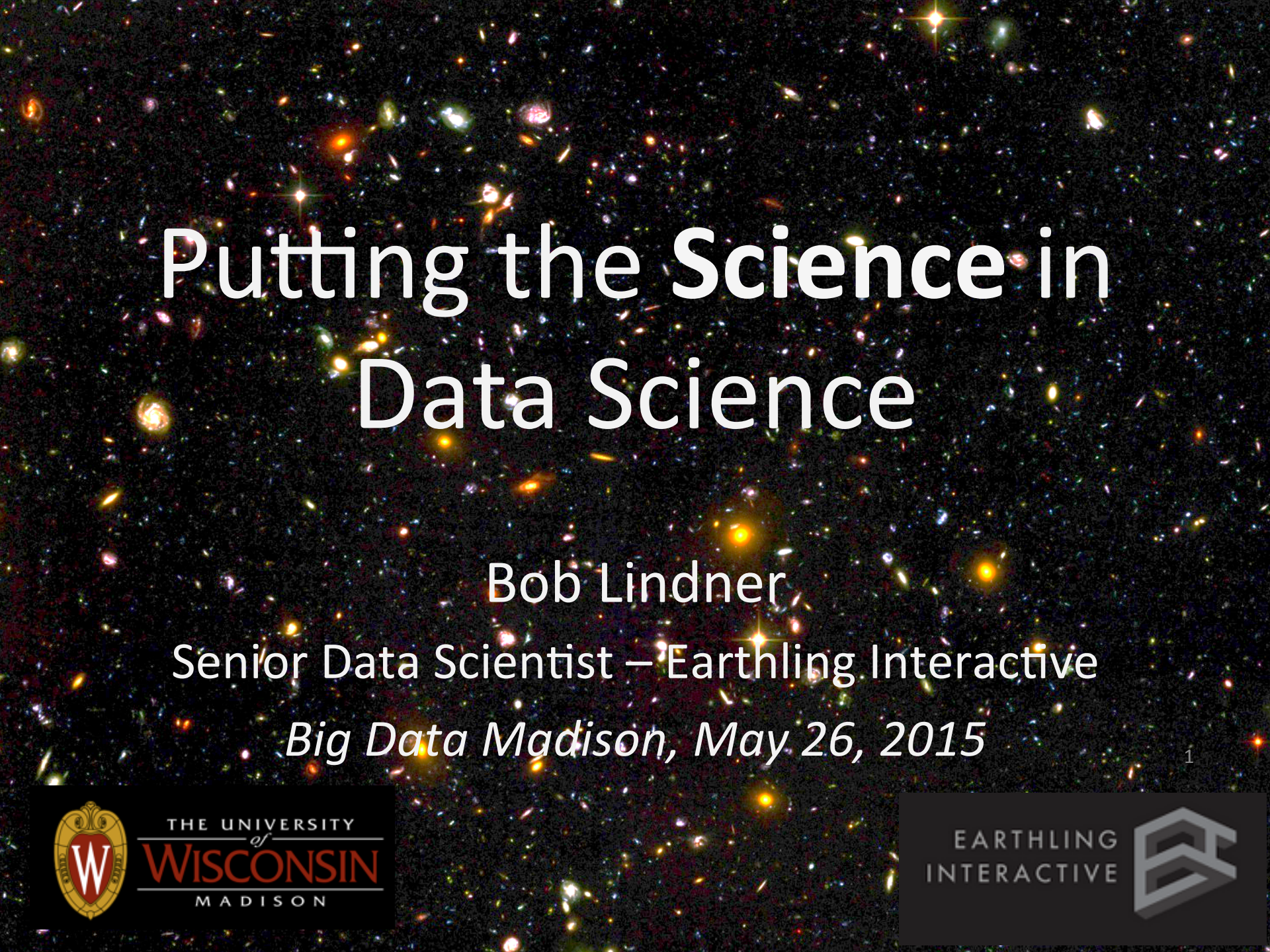


Putting the **Science** in Data Science

Bob Lindner

Senior Data Scientist – Earthling Interactive

Big Data Madison, May 26, 2015

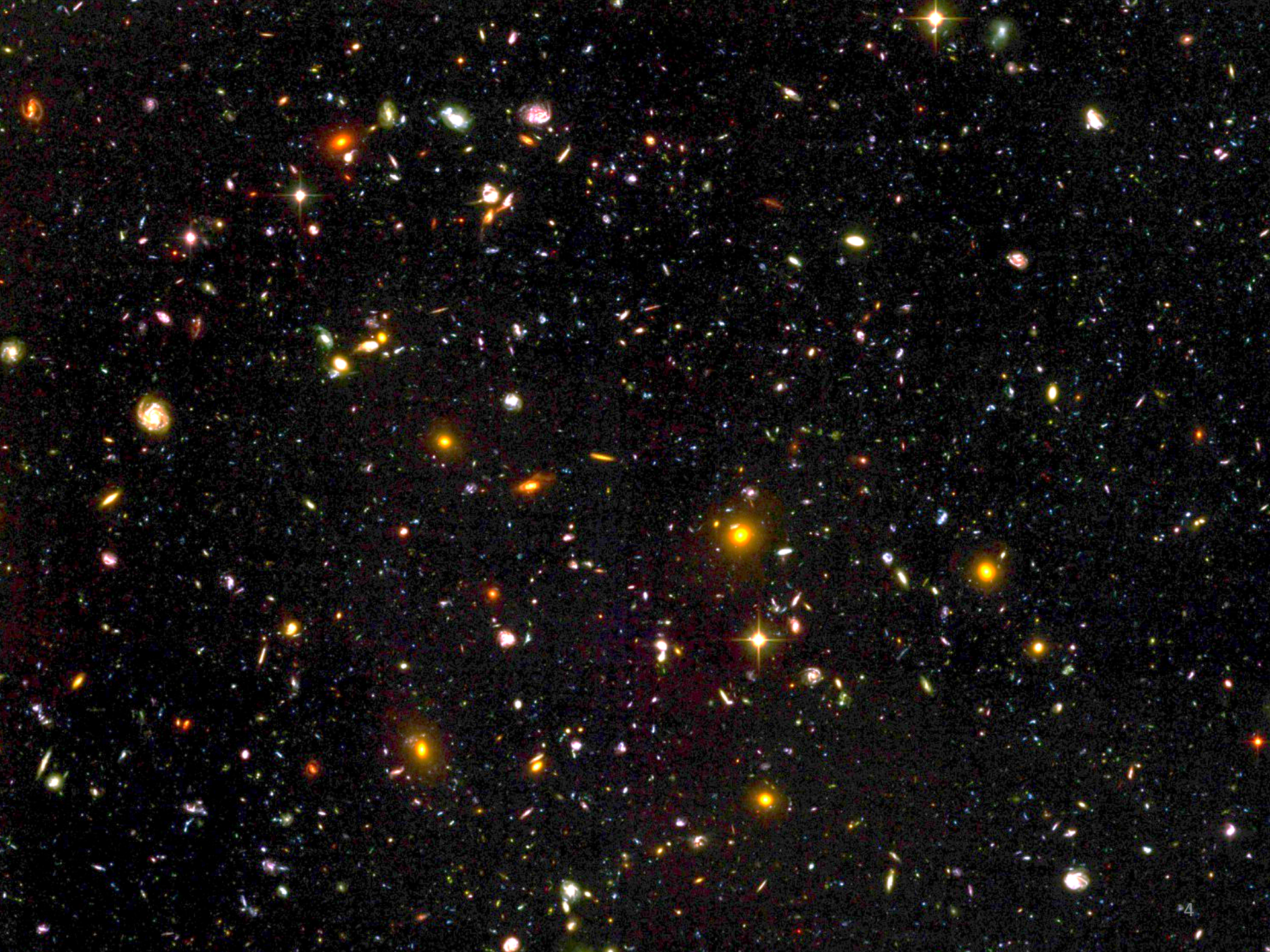
THE UNIVERSITY of WISCONSIN MADISON

EARTHLING INTERACTIVE

Research Collaborators:

Carlos Vera-Ciro (UW-Madison)
Claire Murray (UW-Madison)
Snezana Stanimirovic (UW-Madison)
Patrick Hennebelle (ENS/Paris Observatory)
Miller Goss (NRAO)
Carl Heiles (University of California, Berkeley)
John Dickey (University of Tasmania)
Brian Babler (UW-Madison)
Chang-Goo Kim (Princeton)
Eve Ostriker (Princeton)


1. Motivation




The Universe is full of galaxies

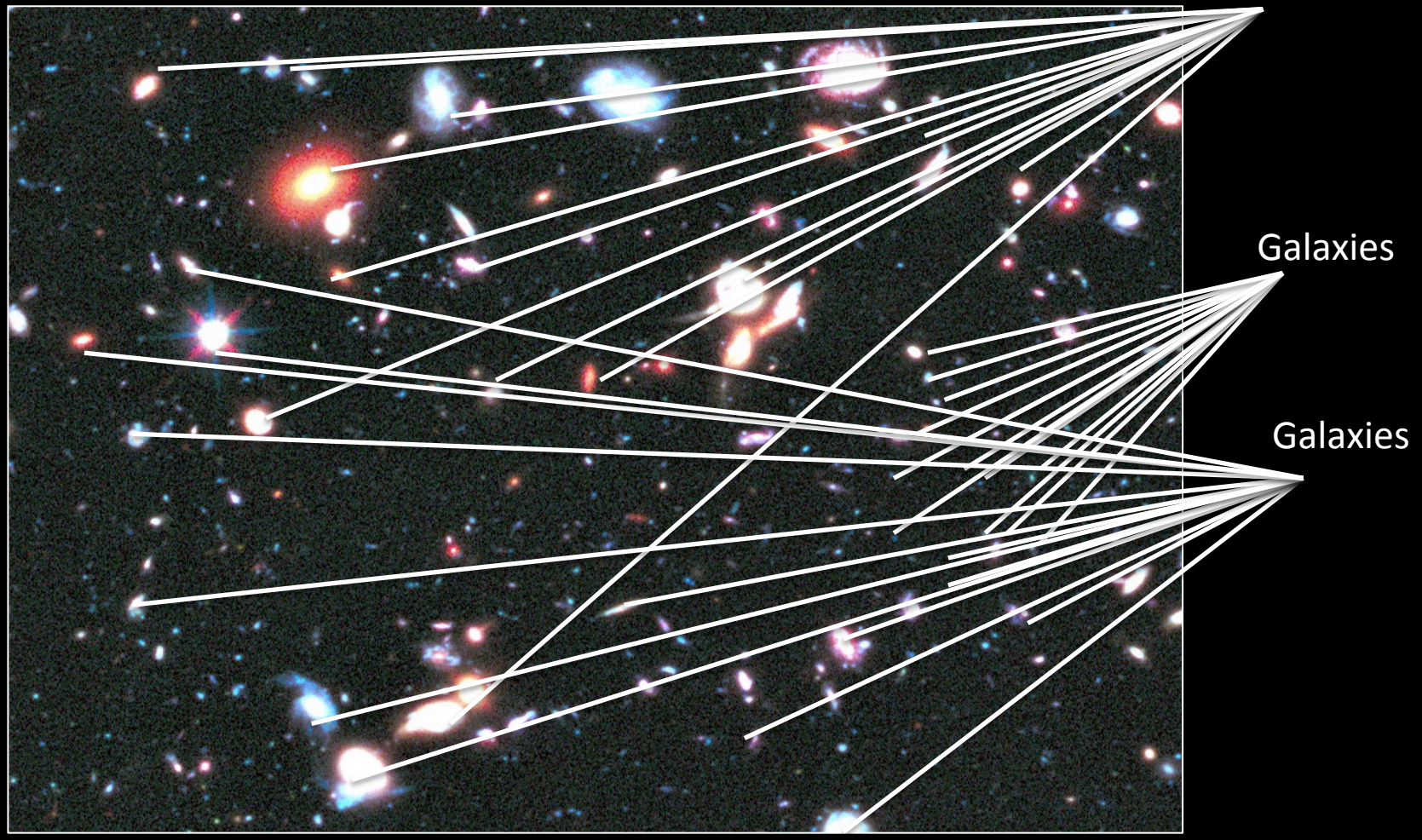

What are they?


Where did they come from?

Hubble Ultra Deep Field

Zoom and Enhance!

Galaxies come in all shapes, sizes, and colors



Examples of Nearby Galaxies

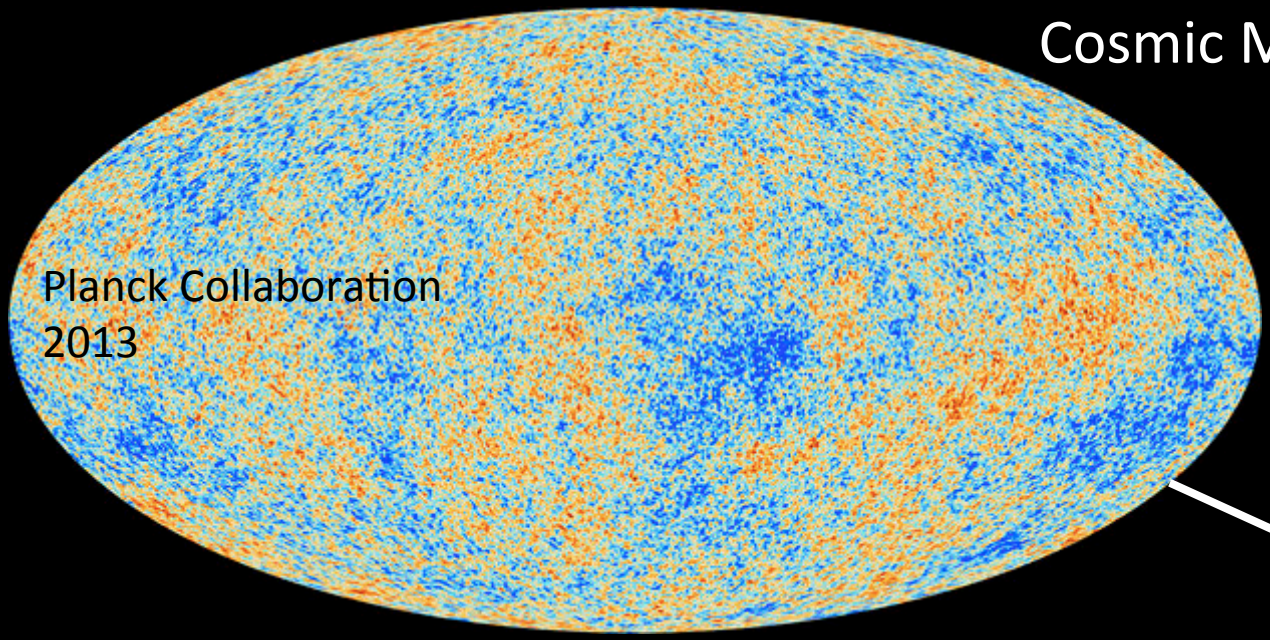
Spiral Galaxy:
Andromeda Galaxy

Elliptical Galaxy
(ESO-325-G004)

Spiral Galaxy
Sombrero Galaxy

Galaxies contain:
Stars, gas, and dark matter

Cosmic Microwave Background

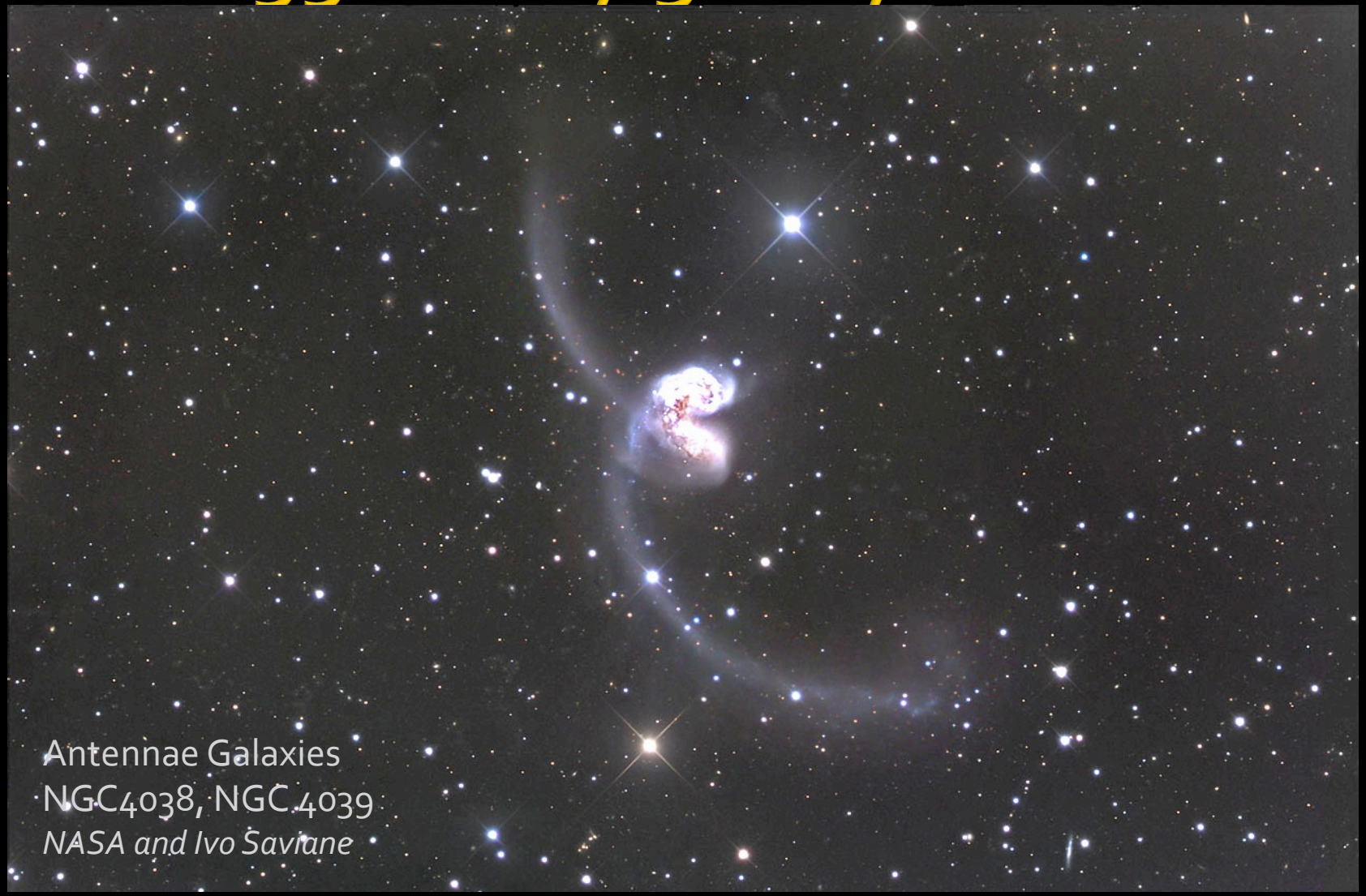


Unexplained question in astrophysics: How does **this** turn into **these?**

+ time, gravity, **star formation**, feedback

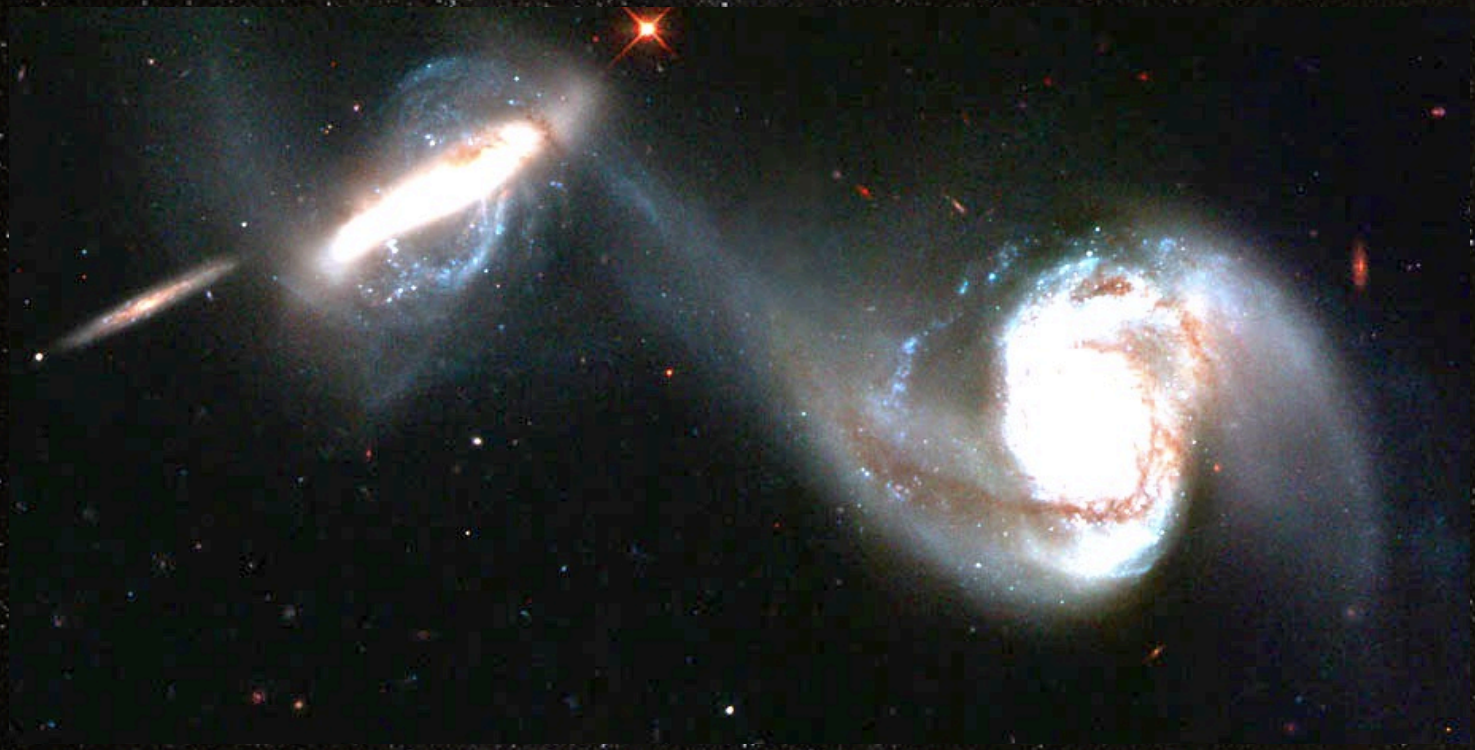
Galactic star forming region, *Hubble Heritage*

Huge bursts of star formation can be triggered by galaxy collisions



Antennae Galaxies
NGC4038, NGC 4039
NASA and Ivo Saviane

We call them "*mergers*"



Arp 87
Hubble Heritage

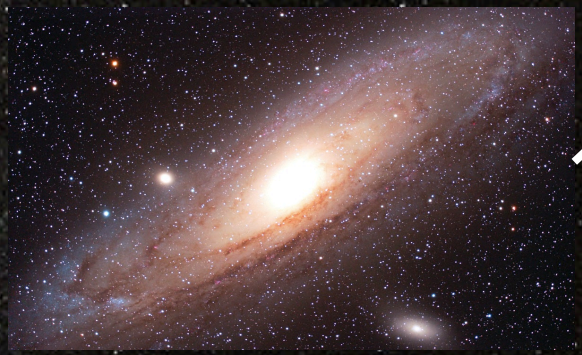
Major Mergers involve two similarly-sized galaxies

NGC 6050
Hubble Heritage

Here is another Major Merger

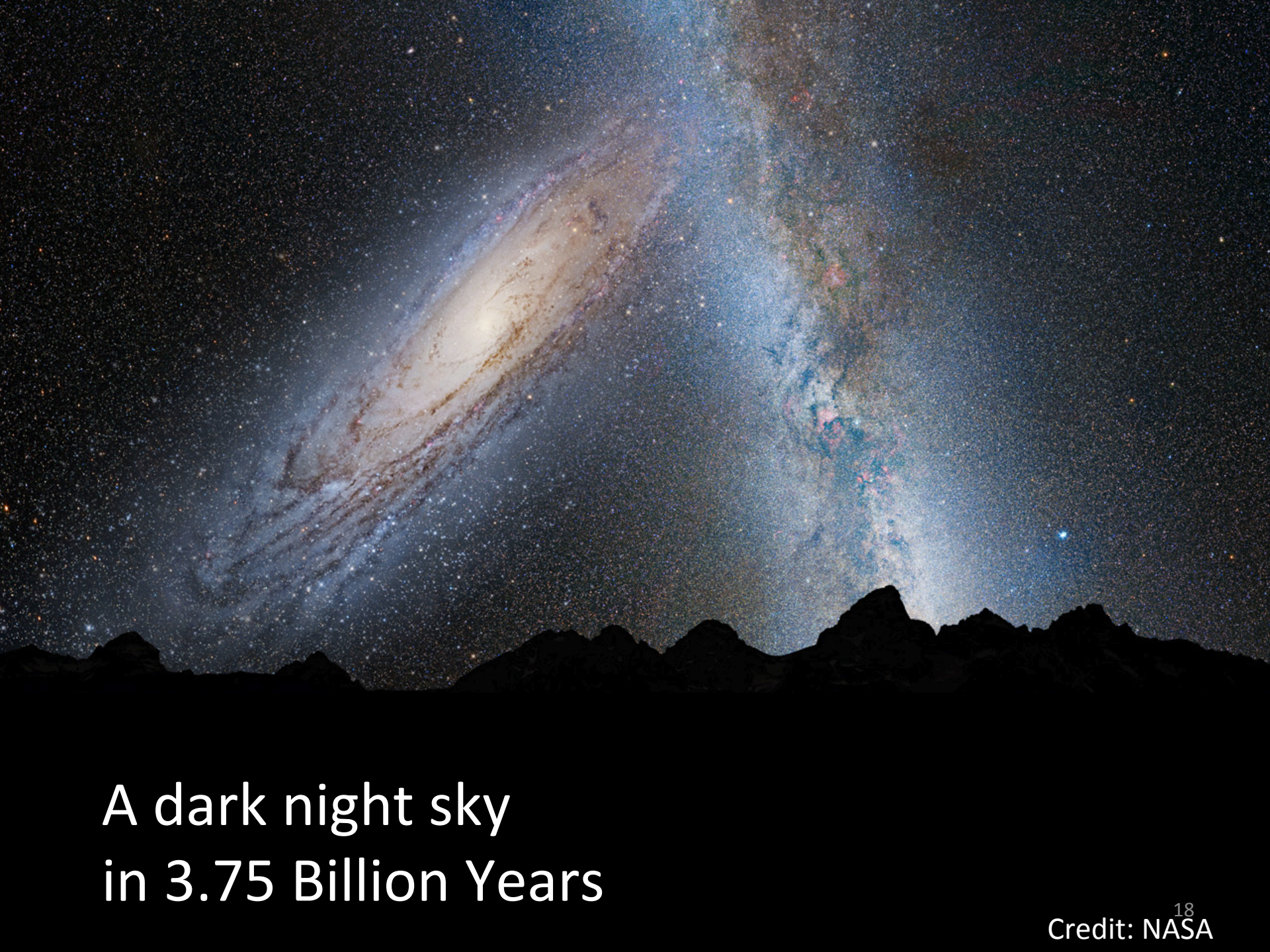
Arp 148
Hubble Heritage

The Milky Way is on a collision course with The Andromeda Galaxy



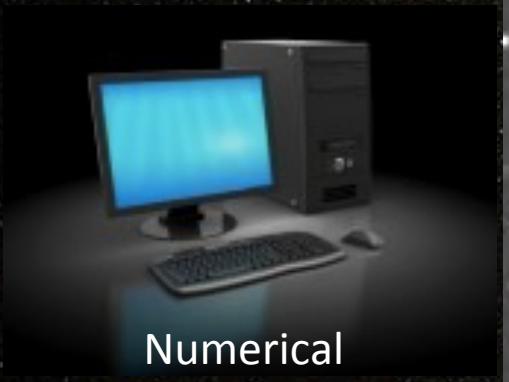
Andromeda Galaxy

Moon



A dark night sky
in 3.75 Billion Years

Credit: NASA



Numerical Simulation

Mihos & Hernquist 1996

Mergers destroy galaxies and **create stars**


Galaxy Evolution
Is controlled by
Star Formation

Galaxy Evolution

Is controlled by

Star Formation

Which is controlled by

Fuel Supply

Fuel Supply

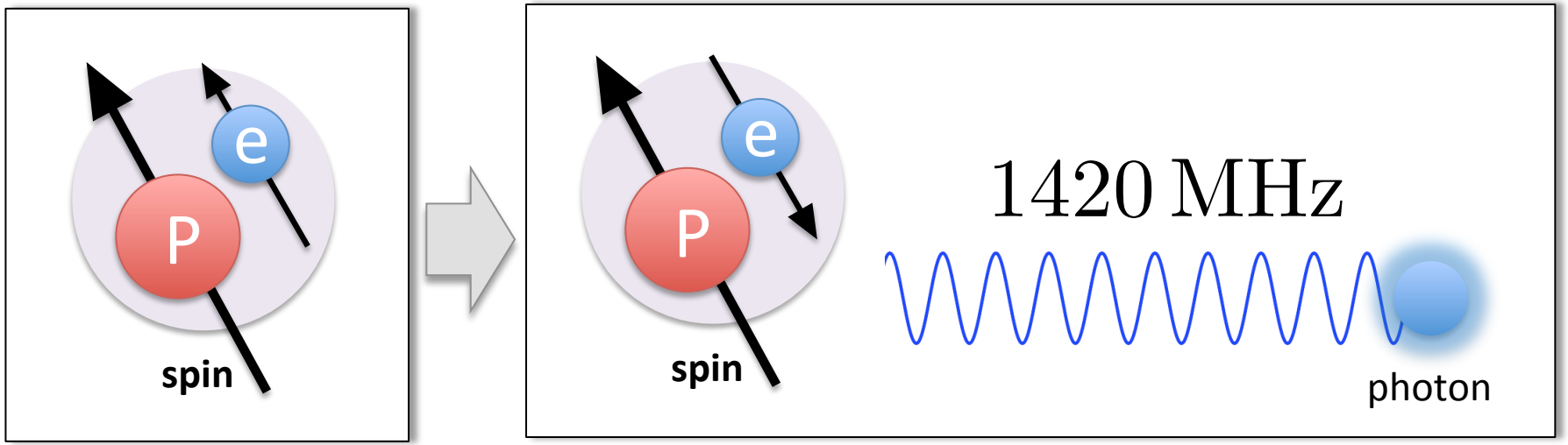
- Neutral Hydrogen gas
- Poorly understood spatial, temp distribution
  Non-linear physics requires numerical simulations

To make progress, we need better observational census of hydrogen gas temperature in galaxies

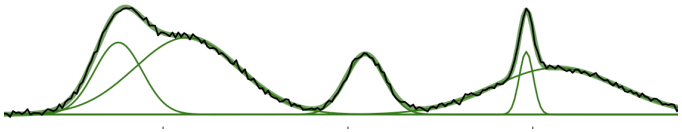

2. The data challenge

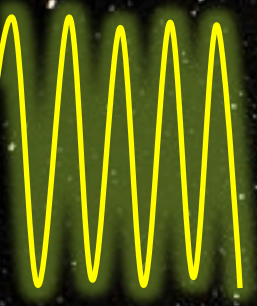
Neutral hydrogen (HI) is important, so how to we detect it?

21cm Hyperfine "spin flip" transition

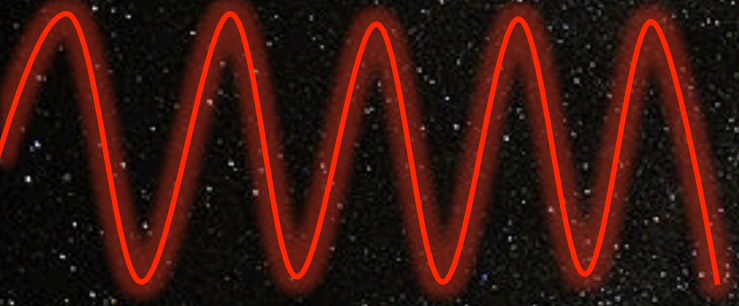


- Unaffected by Earth's atmosphere
- 21cm signals can help measure Hydrogen gas supply in galaxies
- A "spectrum" of 21cm signal shows the amount of gas at each line-of-sight velocity

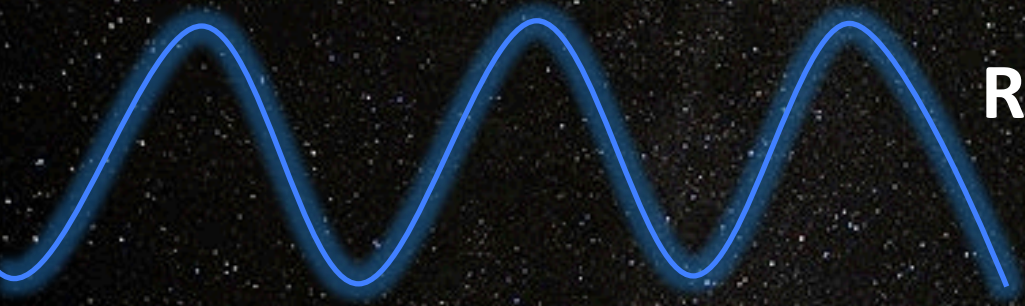




Optical light:
Wavelength $\simeq 500\,\mathrm{nm}$

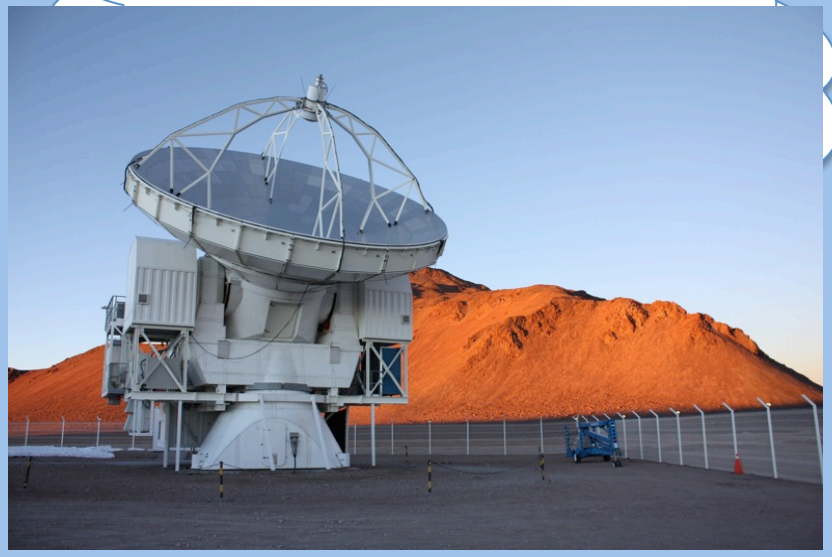


"Submillimeter" light:
Wavelength $\lesssim 1\,\mathrm{mm}$



Radio-wave light:
Wavelength
$\sim 1\,\mathrm{m}$

IRAM 30m Telescope



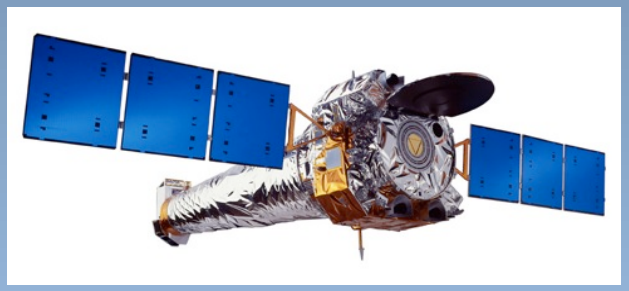
APEX Telescope

Green Bank Telescope

Very Large Array

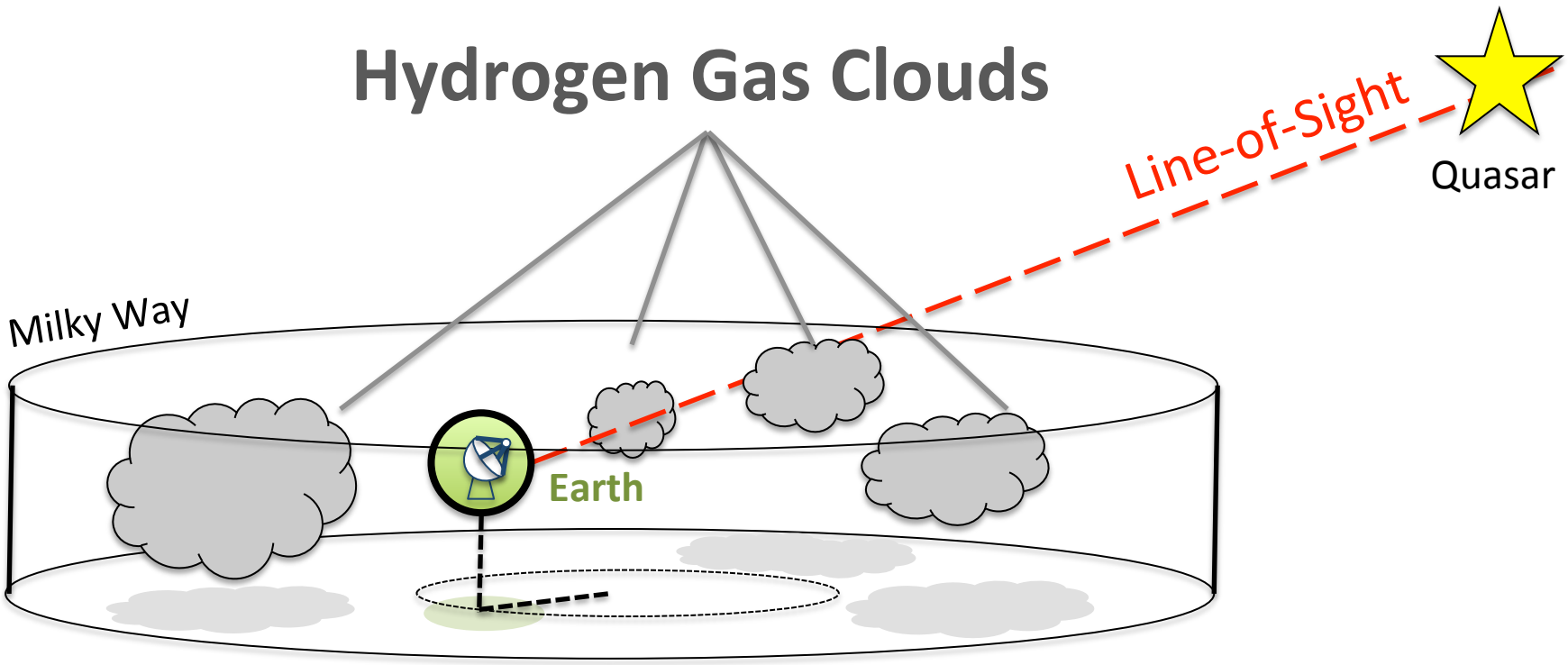
Hubble

Chandra



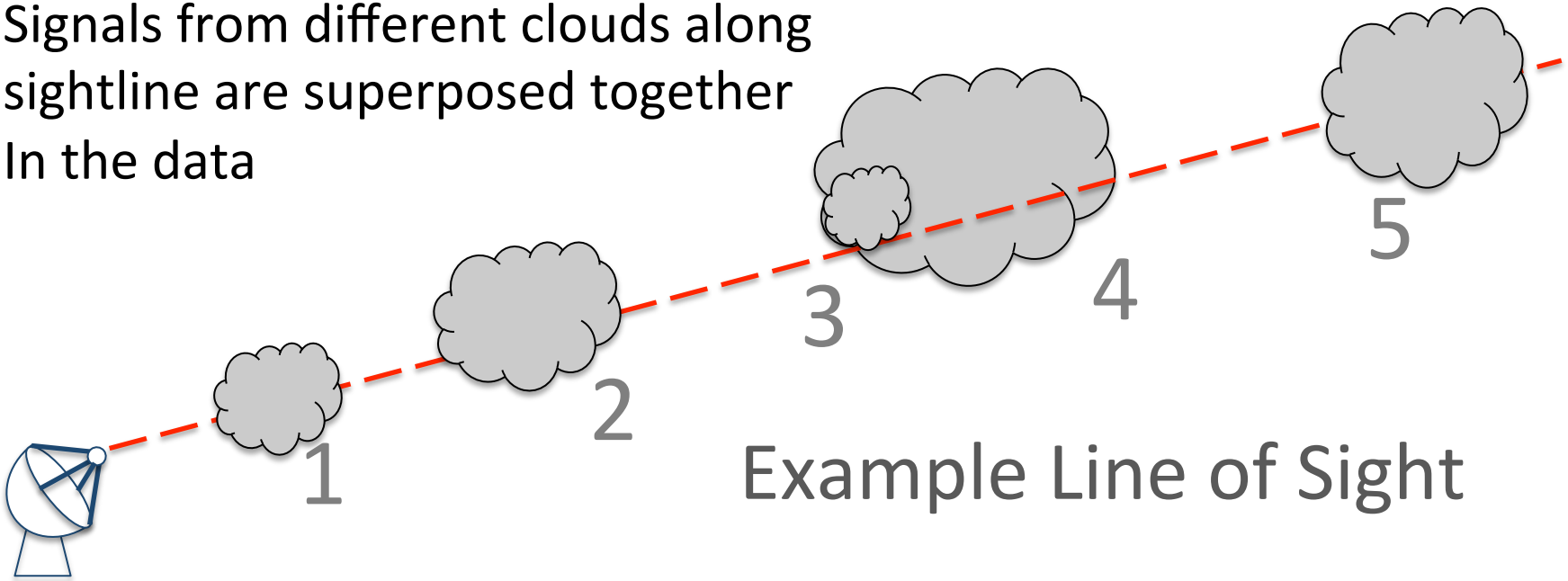
Spitzer

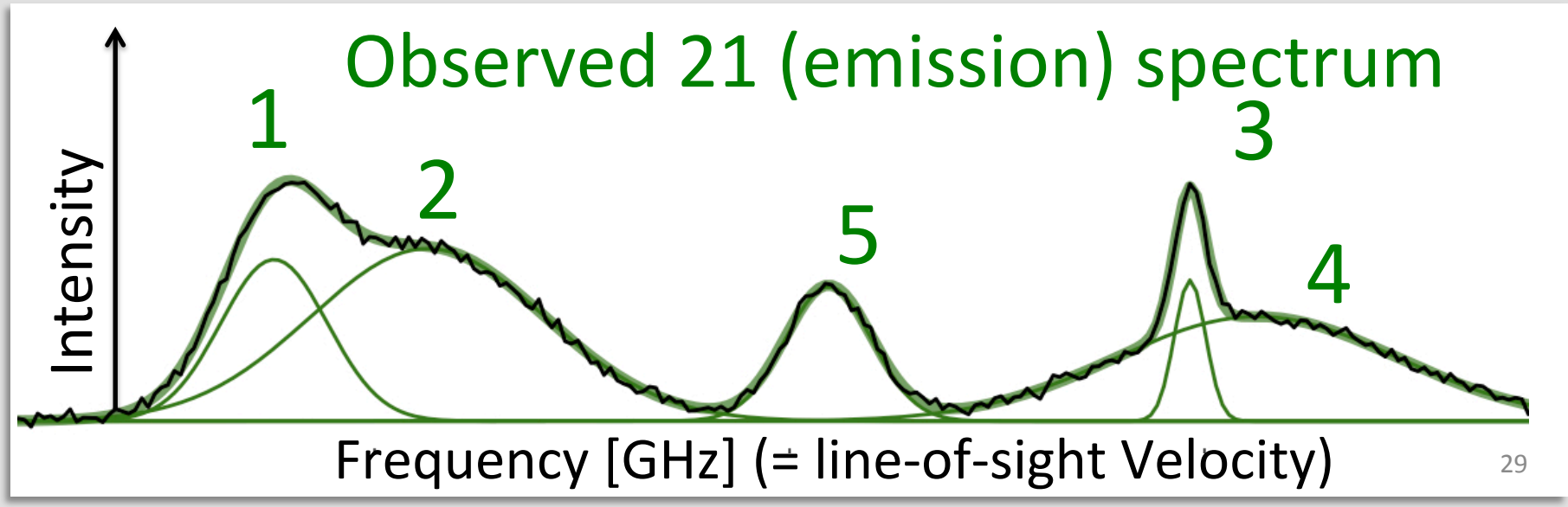
Hydrogen Gas Clouds



Bright background radio source measures intervening Hydrogen

Signals from different clouds along sightline are superposed together In the data

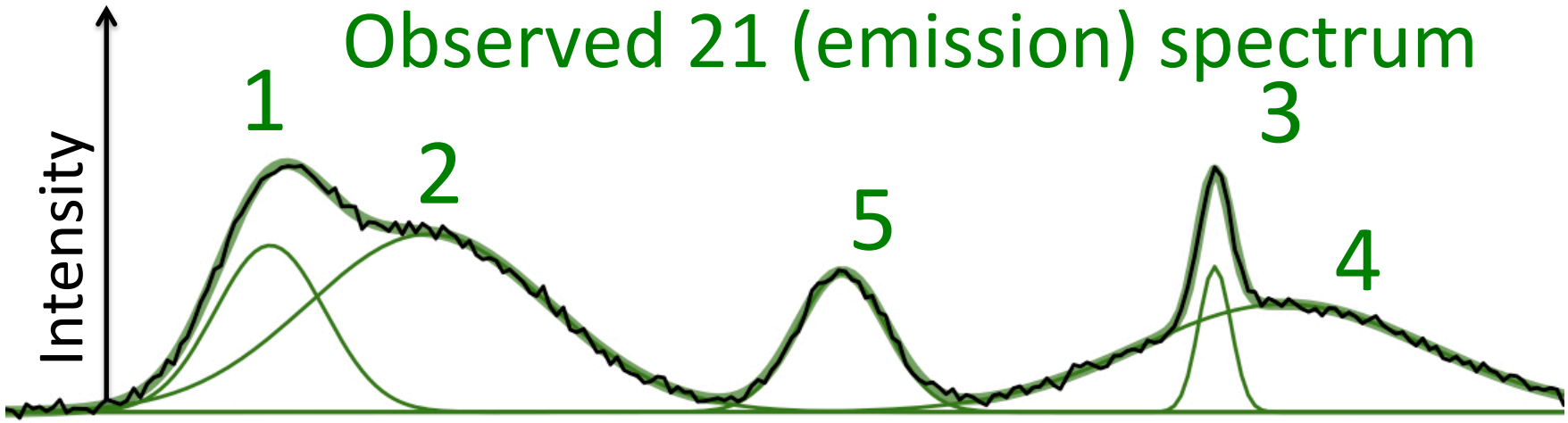




Model:

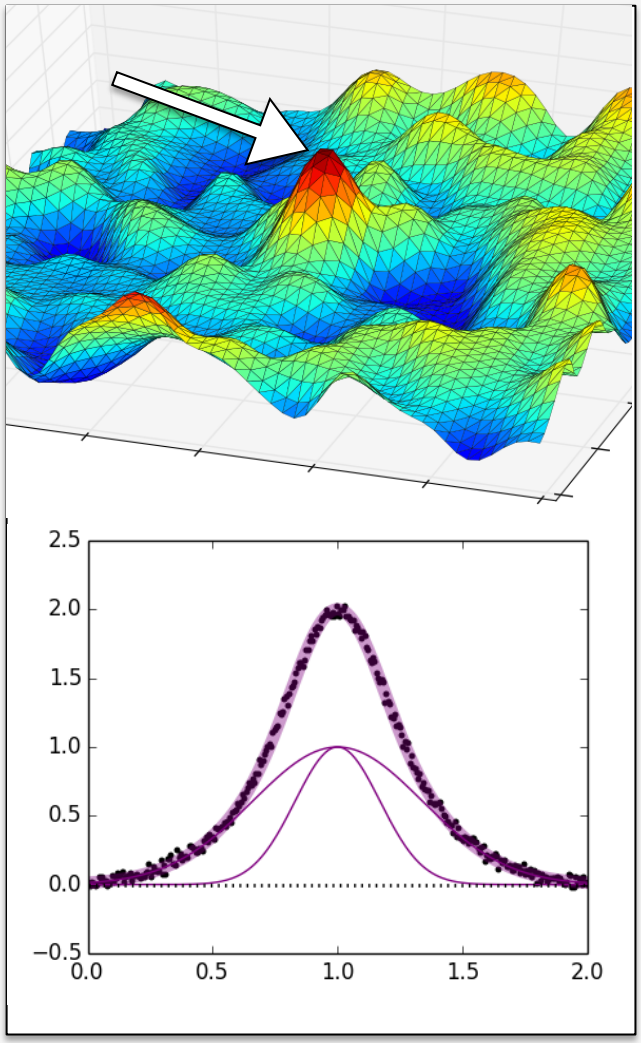
$$\tau(v) = \Sigma_{i=1}^{N} \tau_i e^{-(v-v_i)^2/2\sigma_i^2}$$

- Each component has a *physical* interpretation. **Must get them right.**



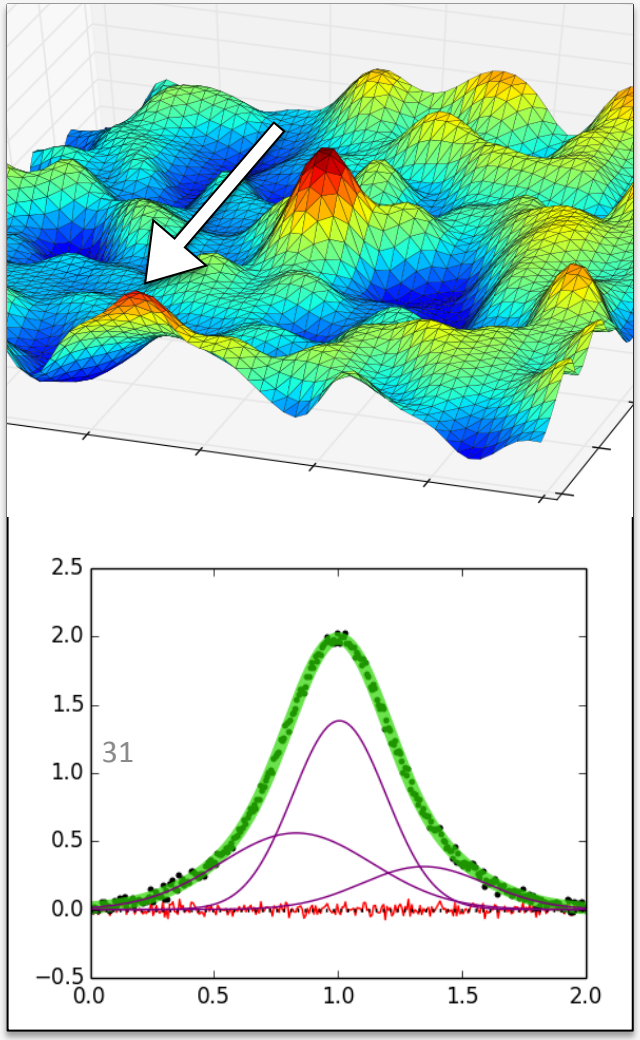
Parameter space is non-convex: local optima everywhere!

Desired solution



Correct model fits the data best

Stuck in Local optimum

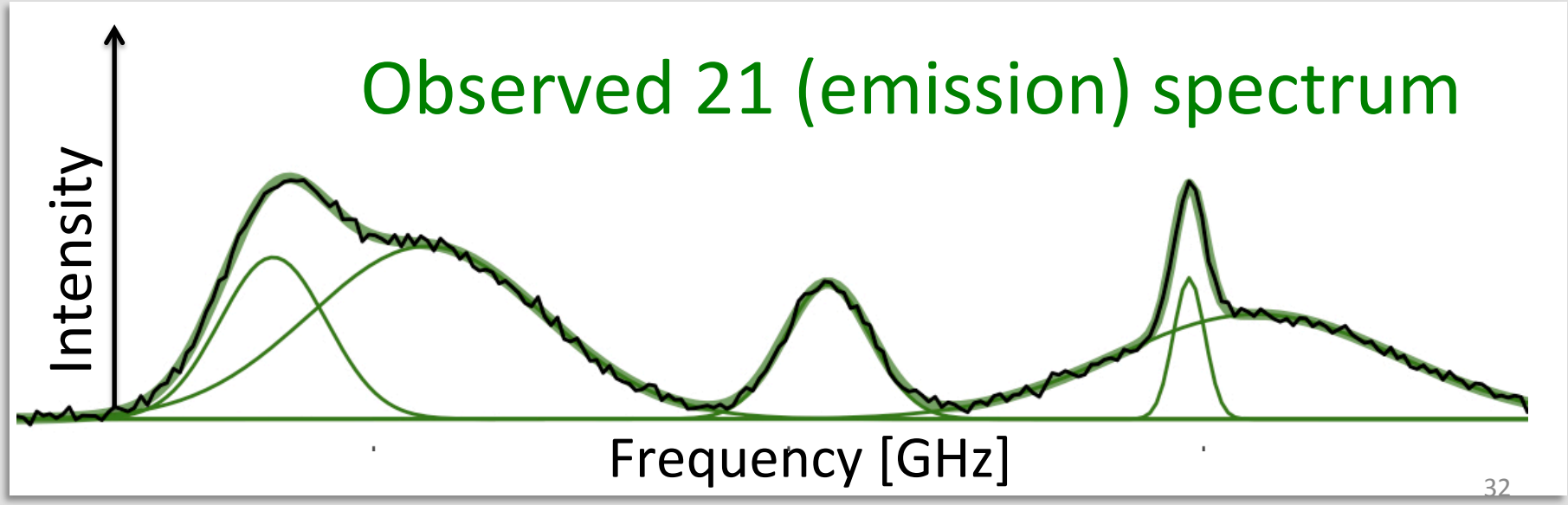


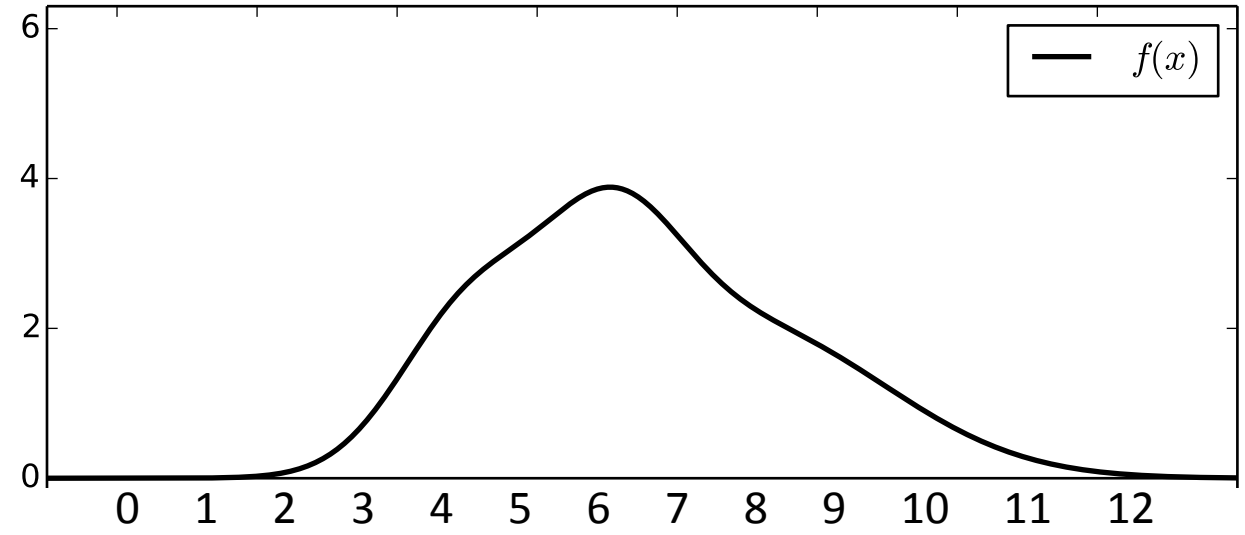
Incorrect model, still fits data well

The Challenge

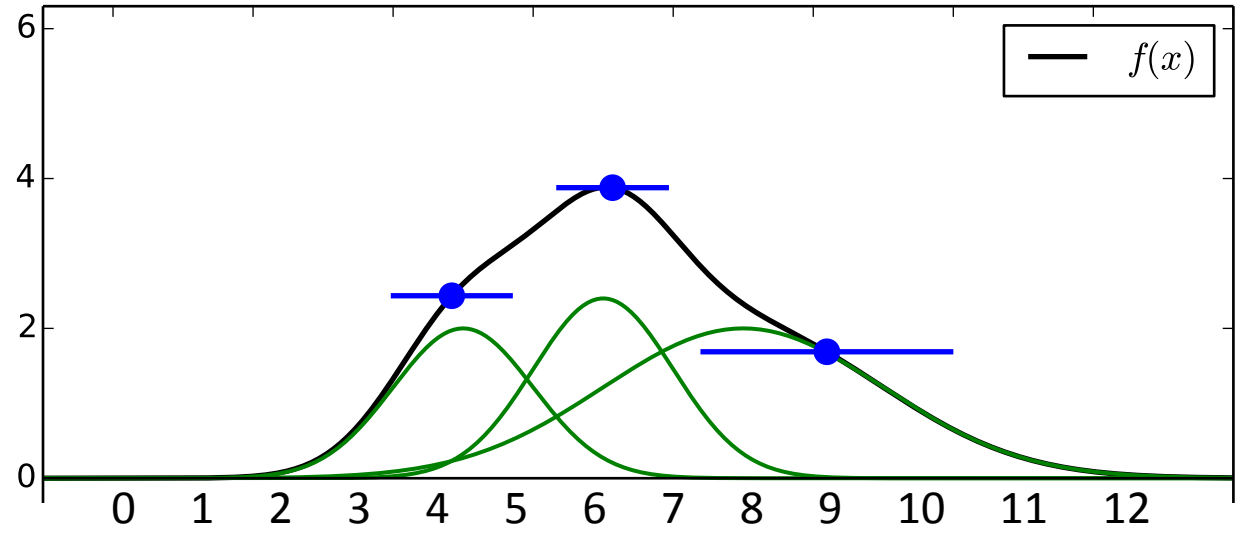
- The **numbers** and shapes of individual components are not known
- Optimization-based fitting algorithms like Levenberg-Marquardt or Gaussian Mixture Models **require initial Guesses** for all parameters.

- Need a way to quickly guess locations of components **before** fitting.



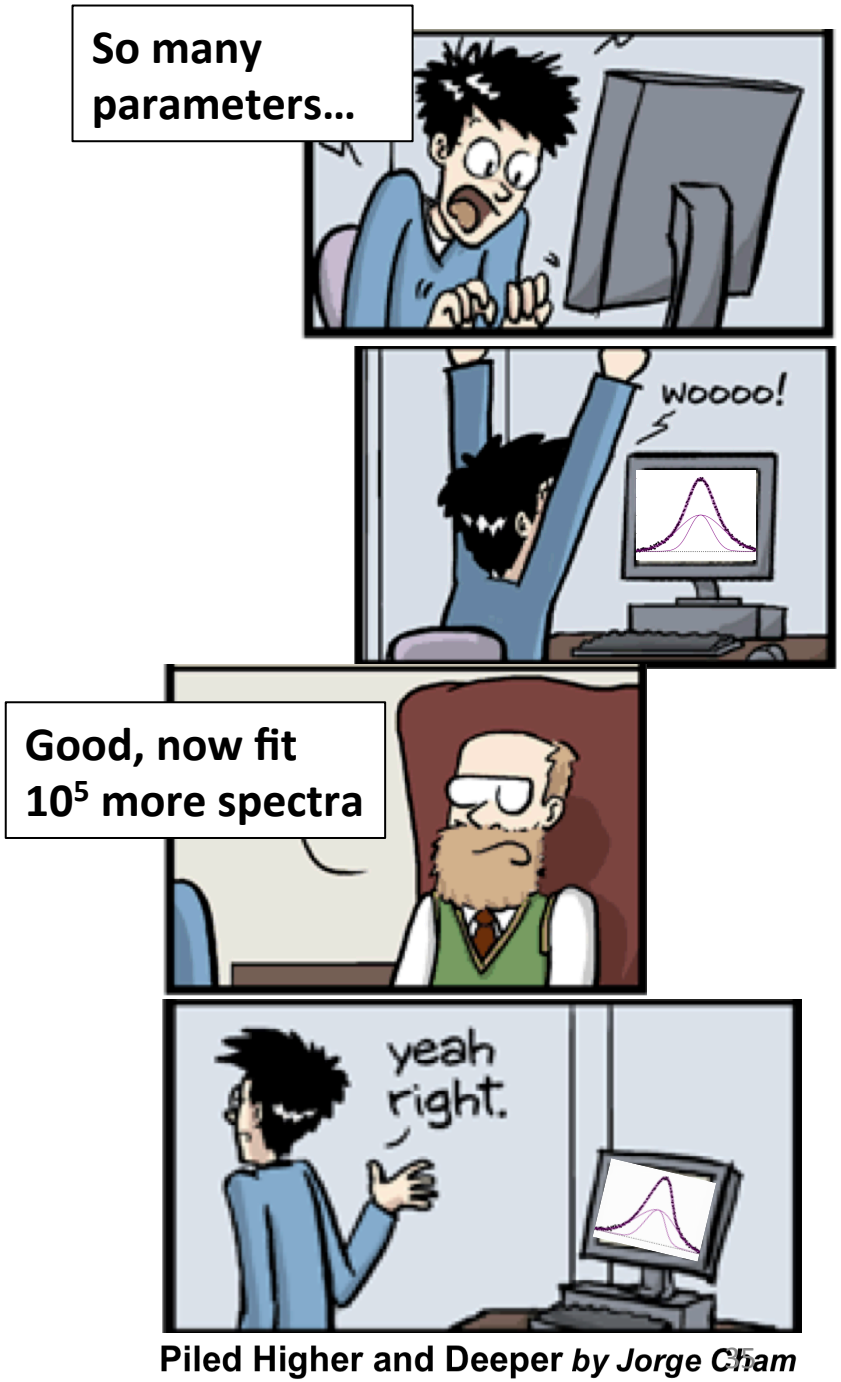


- How many components do you see?
- Where?



- How did you analyze the thousands of individual data points so fast? **Human vision.**

- Current best practice: use **“human”** to find initial parameter values.
  - Usually very good fits
  - Humans vary
    - Non-reproducible results
  - Cannot scale to fit large datasets
    - 30m per spectrum!



Piled Higher and Deeper *by Jorge Cham*

The Future: Big Spectral Data

- Human-assisted Gaussian decompositions have worked until now because of the relatively sparse data available
- Square Kilometer Array (below), MeerKAT, and ASKAP will provide orders-of-magnitude larger data volumes than previous surveys.
- Without autonomous algorithms, the data will not be understood

Square Kilometer Array
www.skatelescope.org

The Future: Big Spectral Data

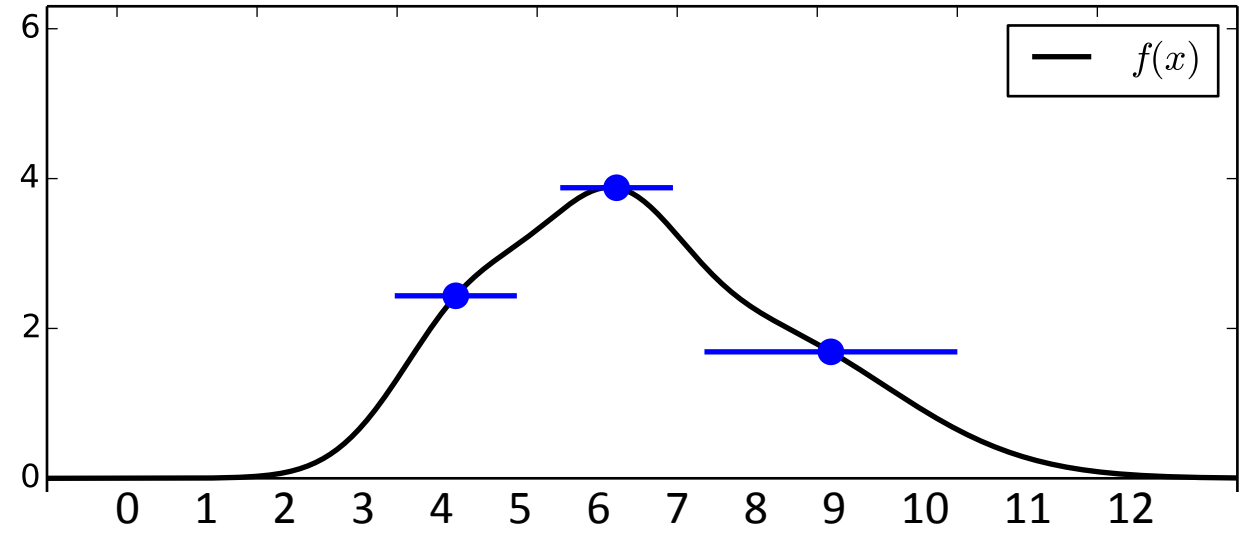
- 160 GB/s per antenna
- Several Petabits/s for full telescope
- Millions of 16k-channel 21cm spectra.

(www.skatelescope.org)

3. A Data Science-based solution:

Autonomous Gaussian Decomposition (AGD)

Lindner et al. 2015, AJ, 149, 138



Recast the problem in computer vision terms:

Finding "bumps" in spectra is equivalent to finding **local minima of negative curvature:**

(aka differential Spectroscopy)

$$\begin{cases} f > 0 \\ d^2 f / df^2 < 0 \\ d^3 f / df^3 = 0 \\ d^4 f / df^4 > 0 \end{cases}$$

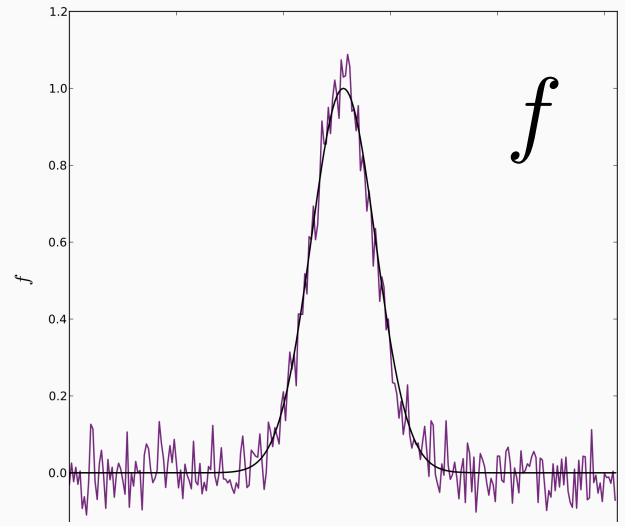
The *real* challenge: data with **noise**

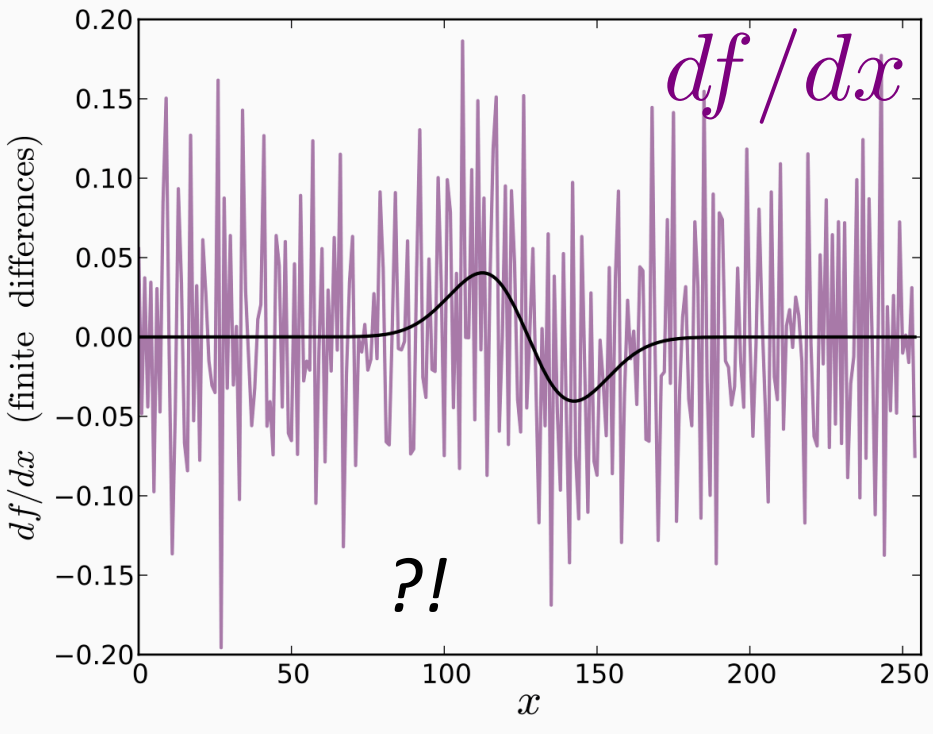
- Using finite-difference numerical derivative:

$$\frac{df}{d\nu}(\nu_i) = \frac{f(i) - f(i-1)}{\nu_i - \nu_{i-1}}$$

-

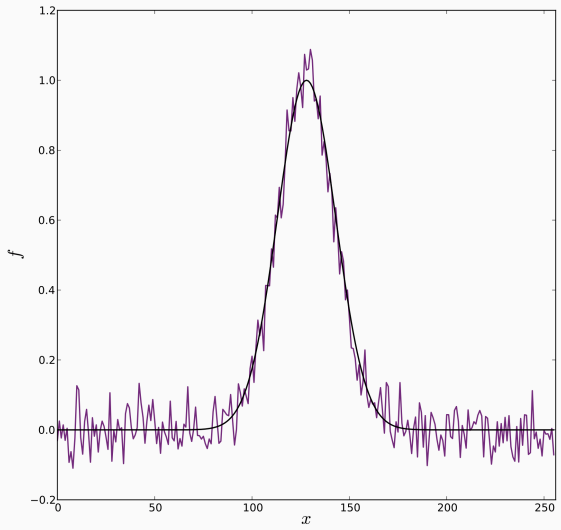
- Finite-difference differentiation **amplifies noise**, making it impossible to locate real local maxima or minima





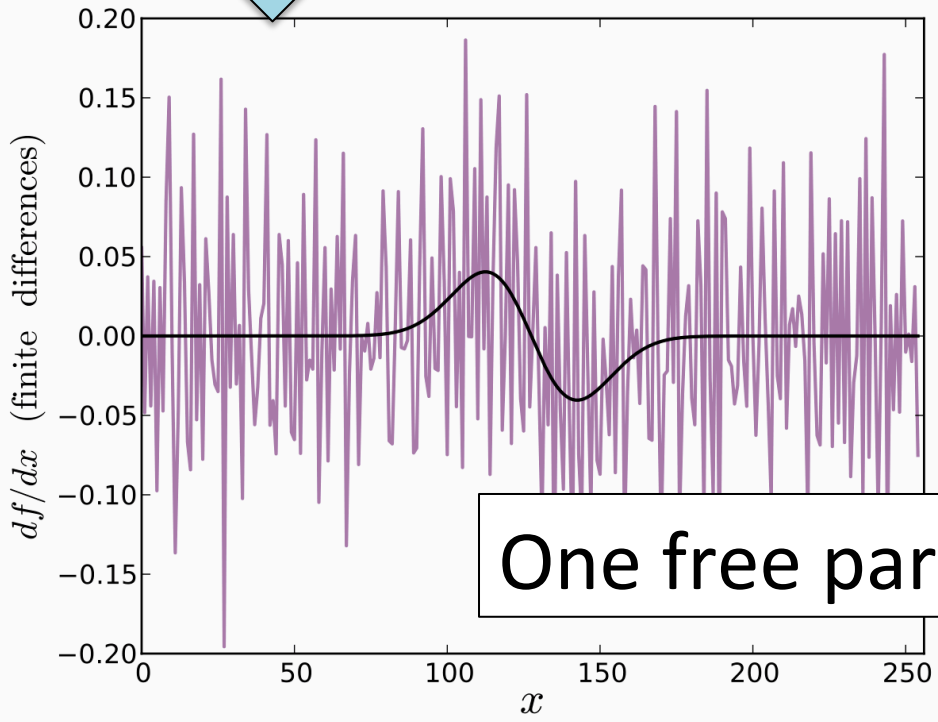
Finite-difference Differentiation

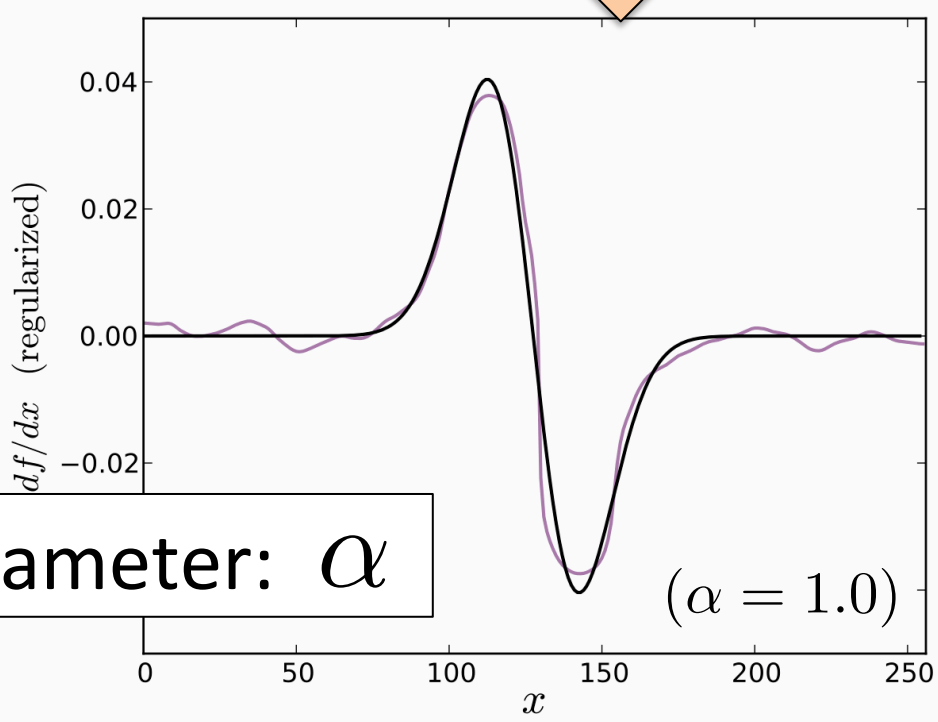
$$\frac{d}{dx}$$



Regularized Differentiation

$$\frac{d}{dx}$$

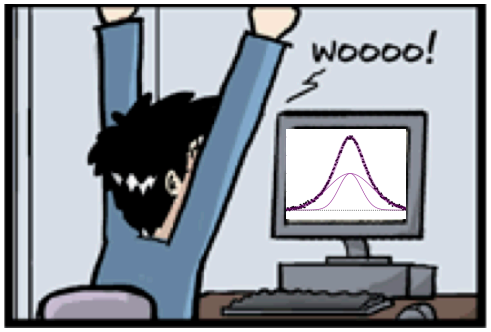




One free parameter: α

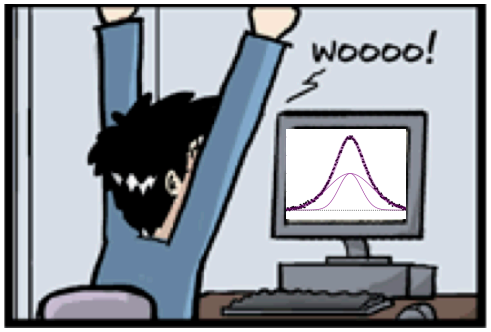
Training the algorithm with supervised machine learning

- Supervised machine learning: Feed the computer a collection of input/output pairs so it can learn a mapping from input to output.
- In our case, input = example spectra, output = correct Gaussian decompositions.

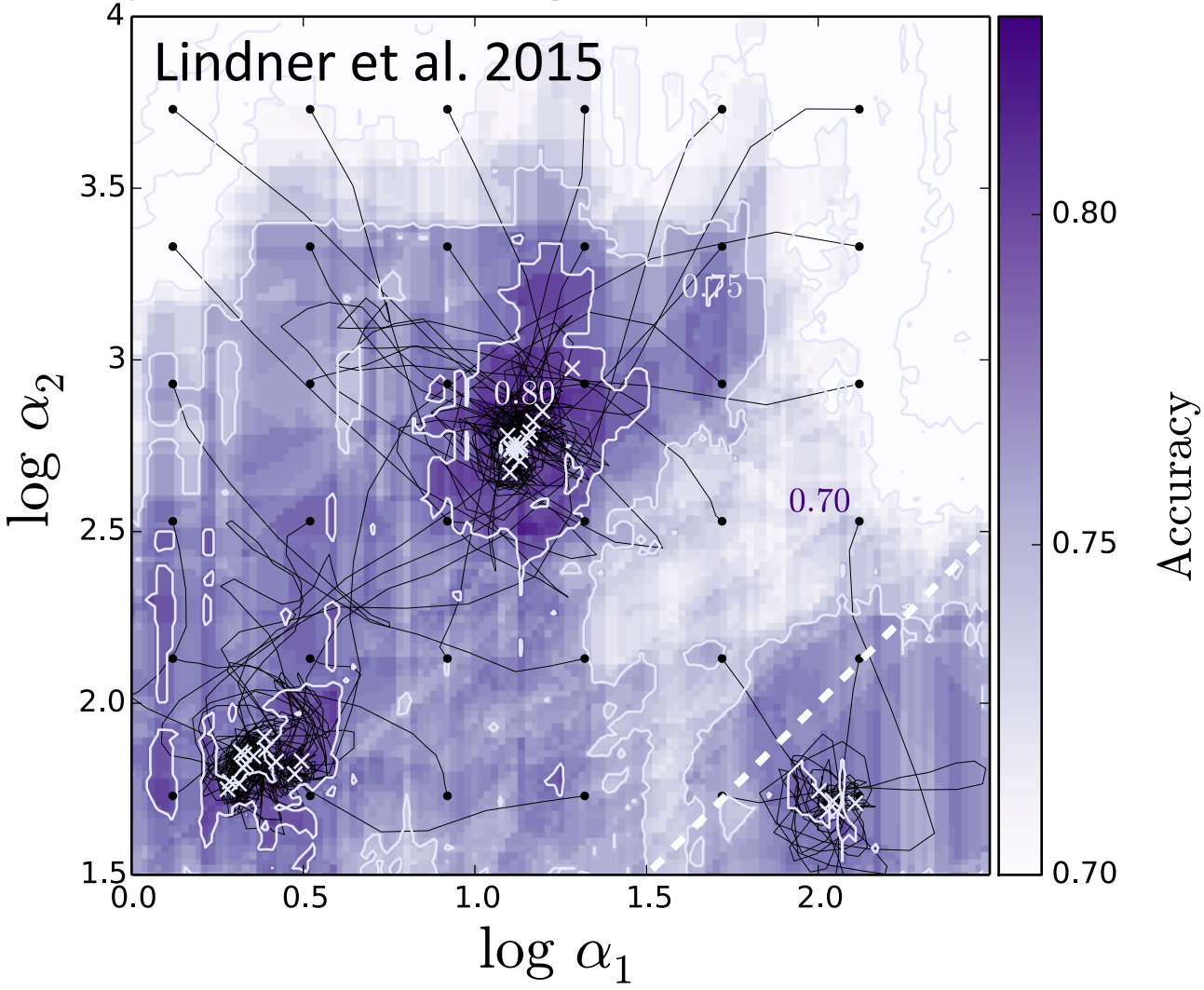


Training the algorithm with supervised machine learning

- Maximize Balanced F1 score between true and guessed components
- Method: Gradient descent with momentum
- Training set: subset of manual decompositions or manually constructed examples



Example training result on real data



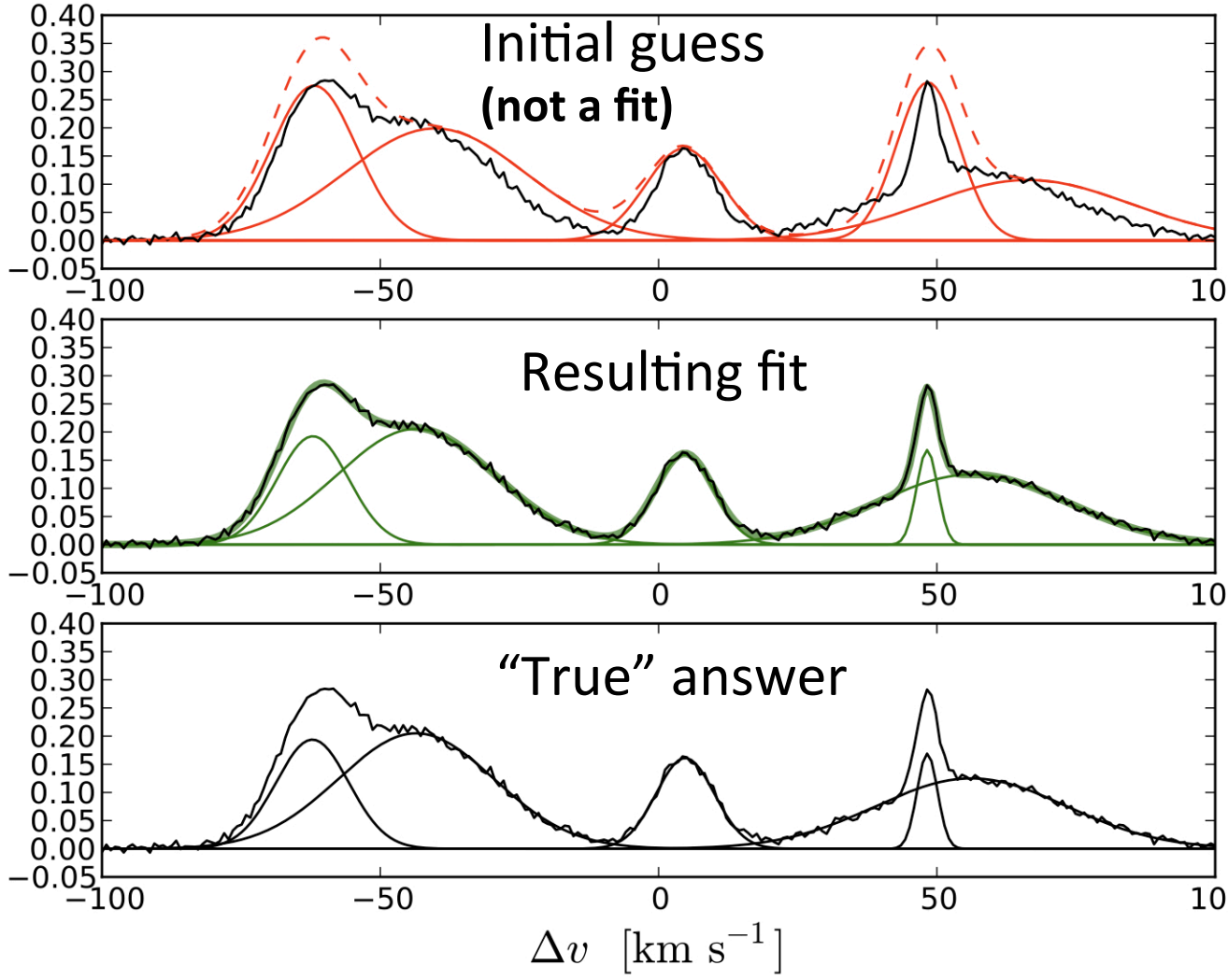
Densely sampled parameter space image produced using the UW-Madison HTCondor cluster

Example process of "Guess" → "Best Fit" after training.

Red: Output of the code is initial guesses of positions, widths and peaks of components

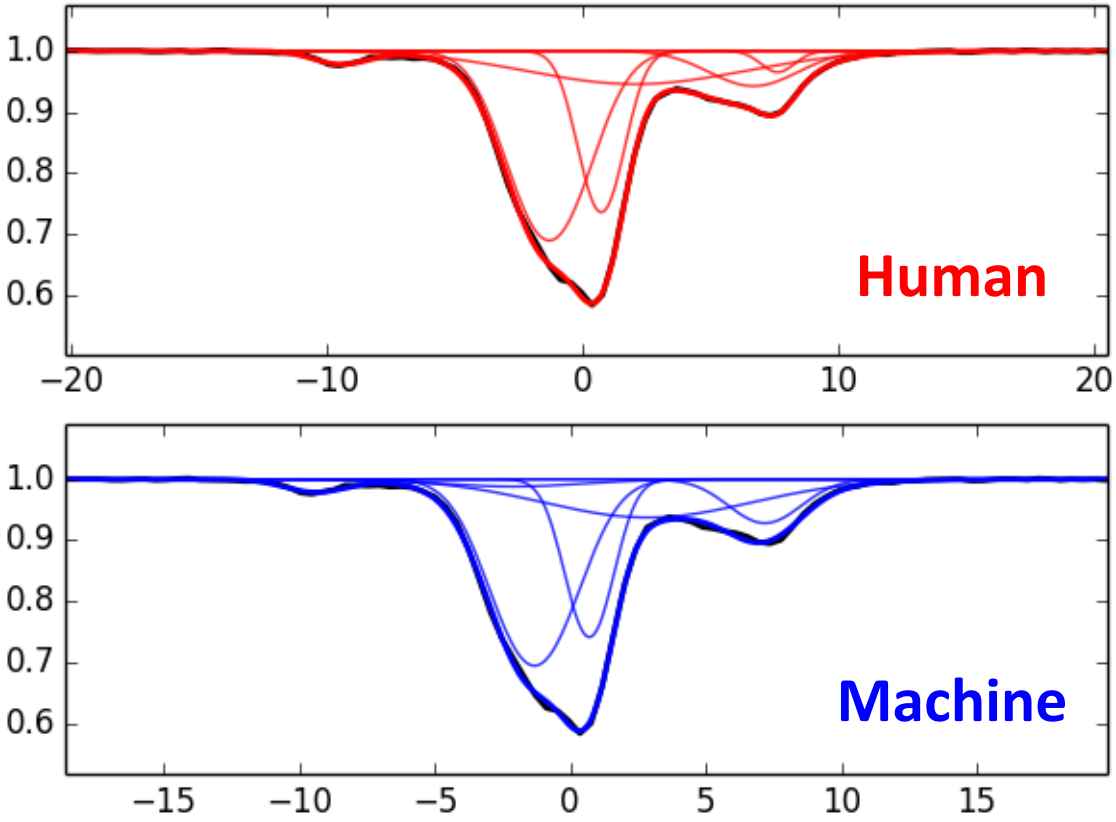
Green: The result after using the initial guesses for a traditional least-squares fit.

Black: The "true" answer



Example performance on real data

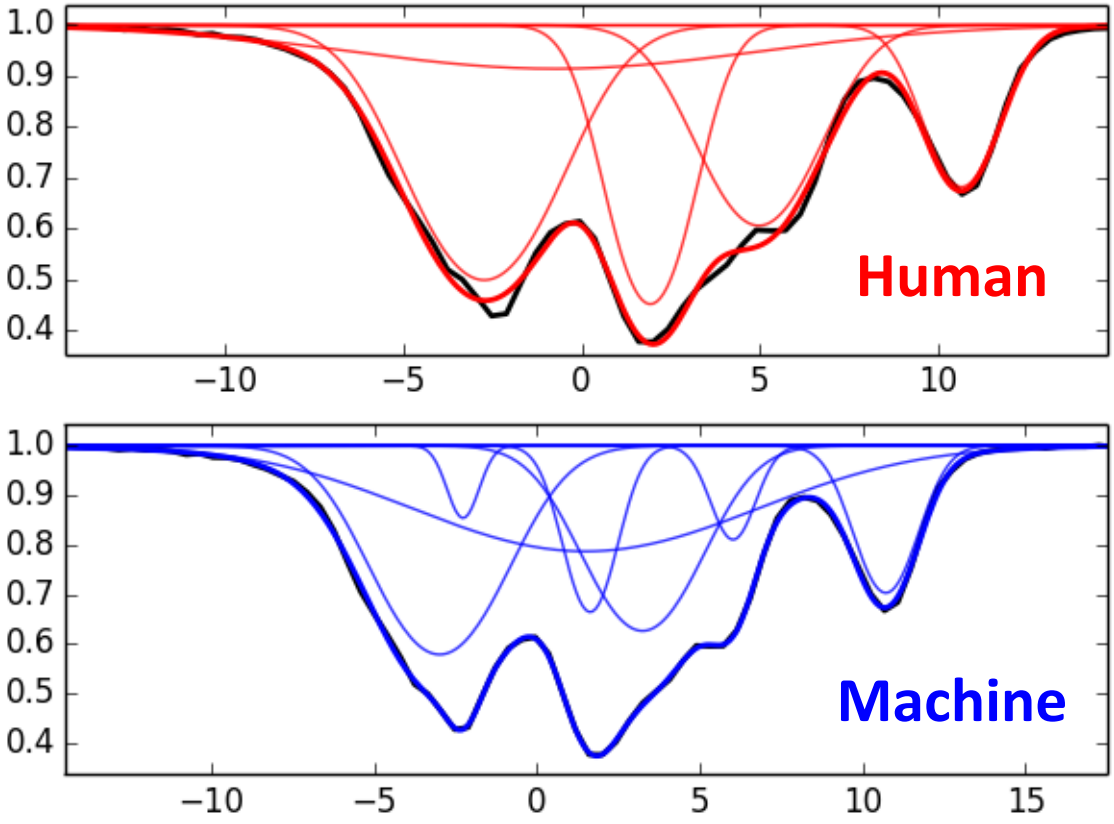
Moderate complexity, high signal-to-noise



4C16.09 spectrum from 21-SPONGE survey (Murray et al. 2015)

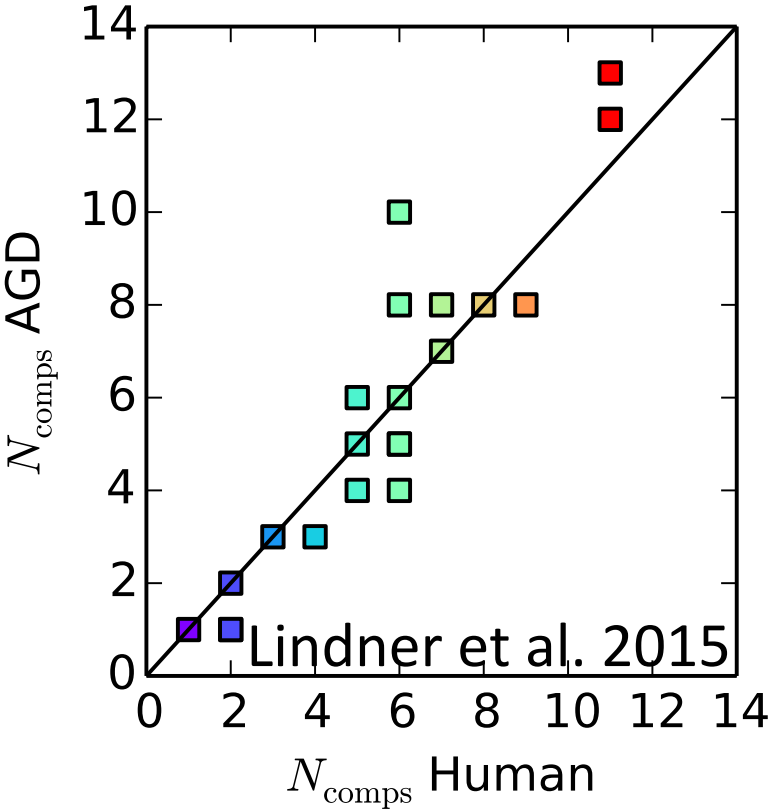
Example performance on real data

Very complex, high signal-to-noise

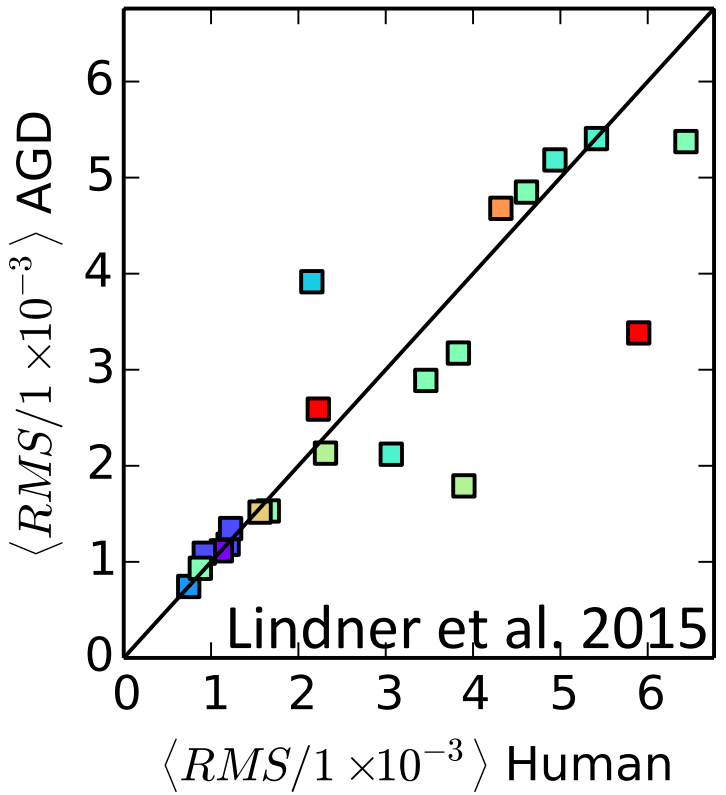


3C154 spectrum from 21-SPONGE survey (Murray et al. 2015)

Aggregate performance on real data



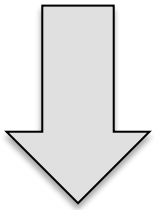
Number of components guessed.

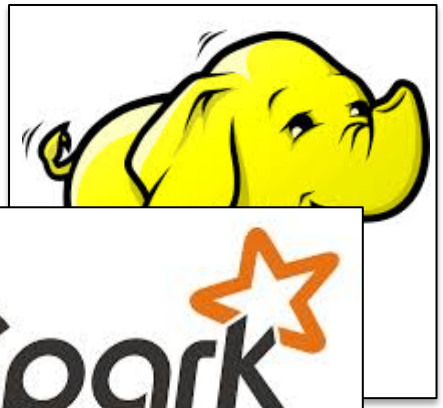


Residuals in resulting best-fit

Scaling AGD to massive spectral datasets:

Many high-throughput solutions available





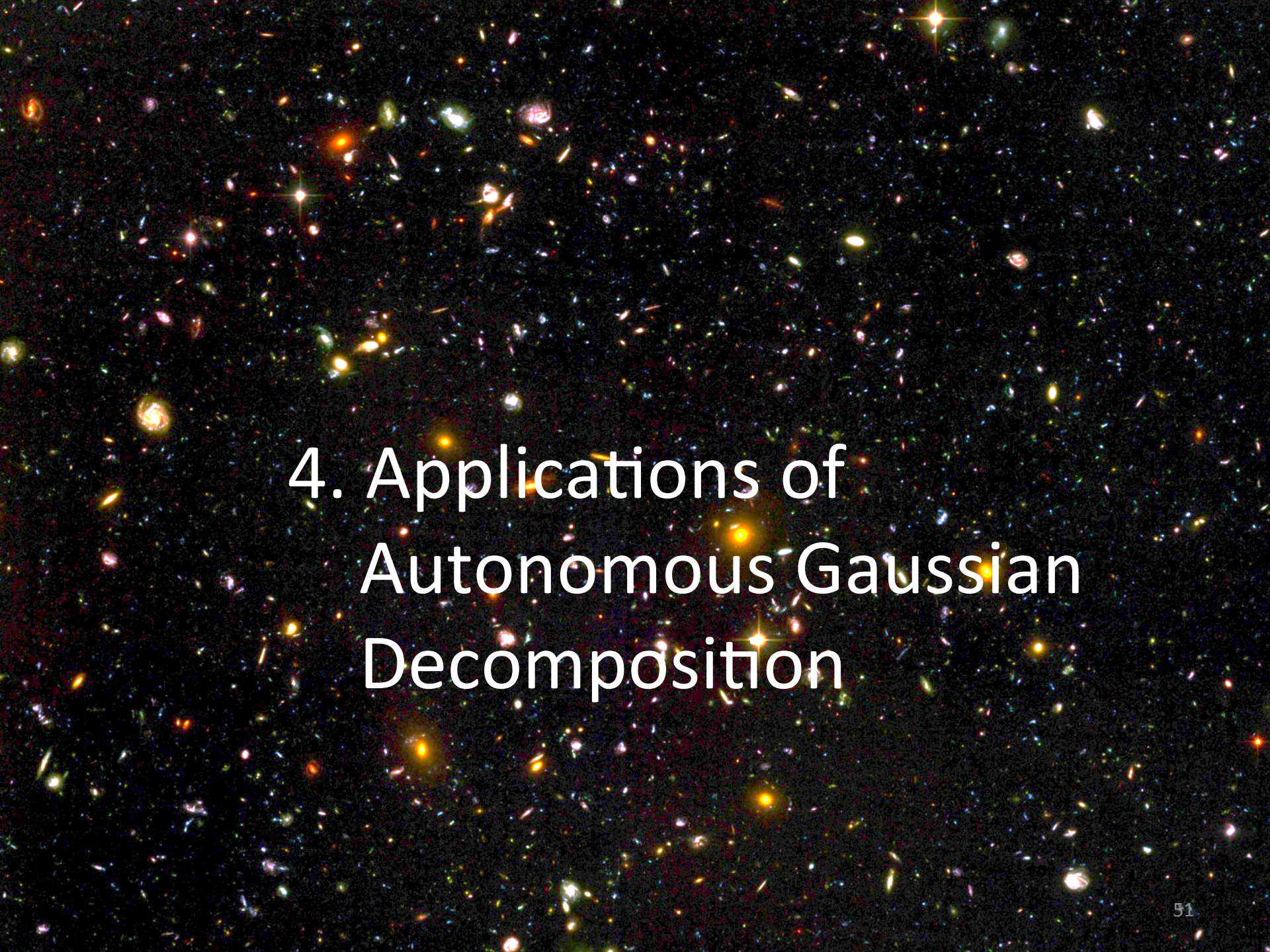
AGD Performance

- About 1 second for guess + fit on single thread standard linux hardware.

- HTCondor can use thousands of cores

- Commercial solution: 20.16 USD = 12hour on AWS c3.8xlarge = 1.4 million spectra decomposed.

- Apache Spark deployment will scale time and price linearly with number of spectra.


4. Applications of
Autonomous Gaussian
Decomposition

Simulations

Physical quantities

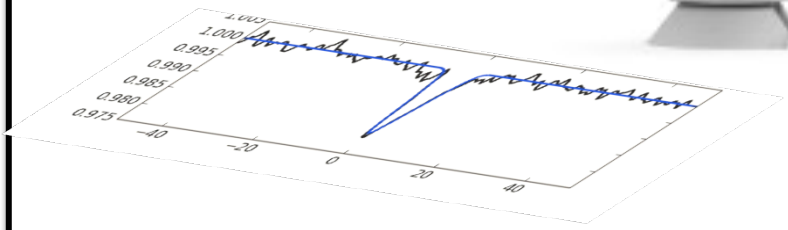
Synthetic spectra

$AGD()$

Gaussian components

$\tau_i, \sigma_i^v, \Delta v_i,$

Observations



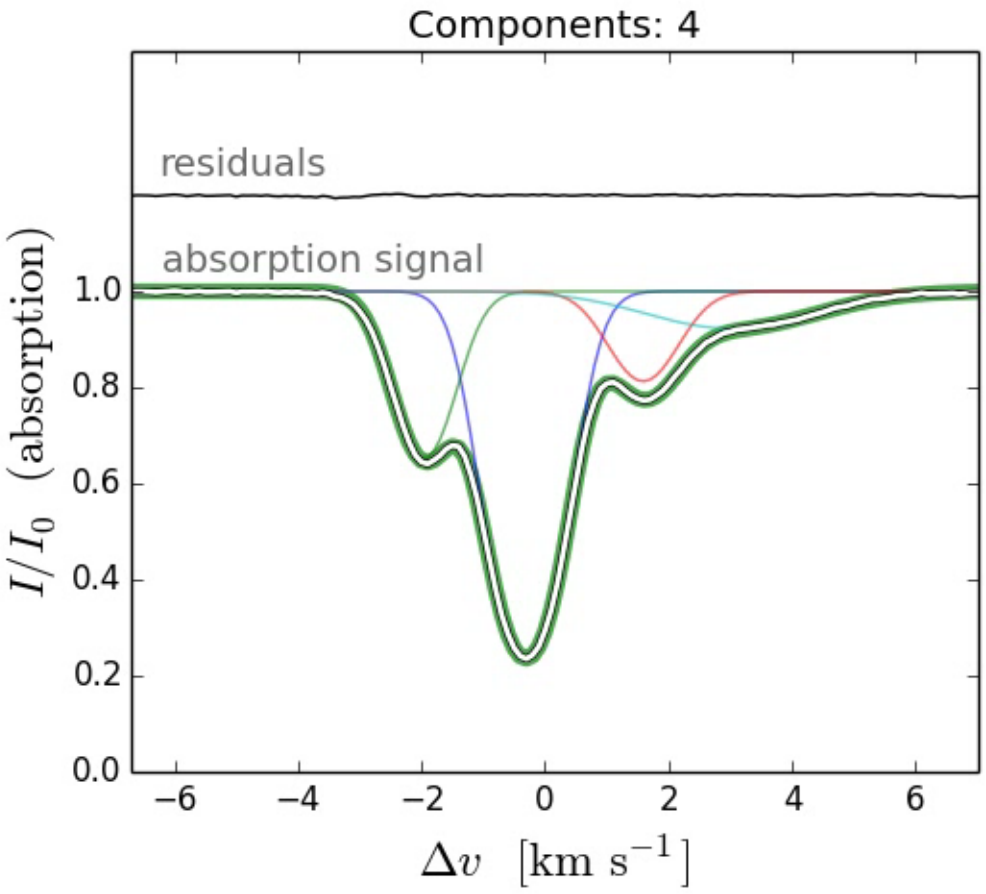
Observed spectra

$AGD()$

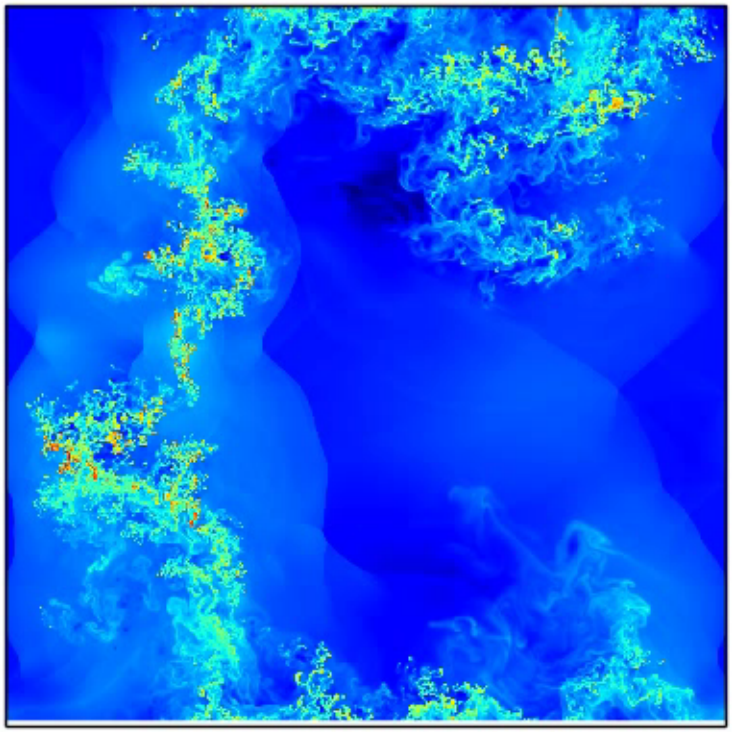
Gaussian components

$\tau_i, \sigma_i^v, \Delta v_i,$

Example decompositions of synthetic spectra

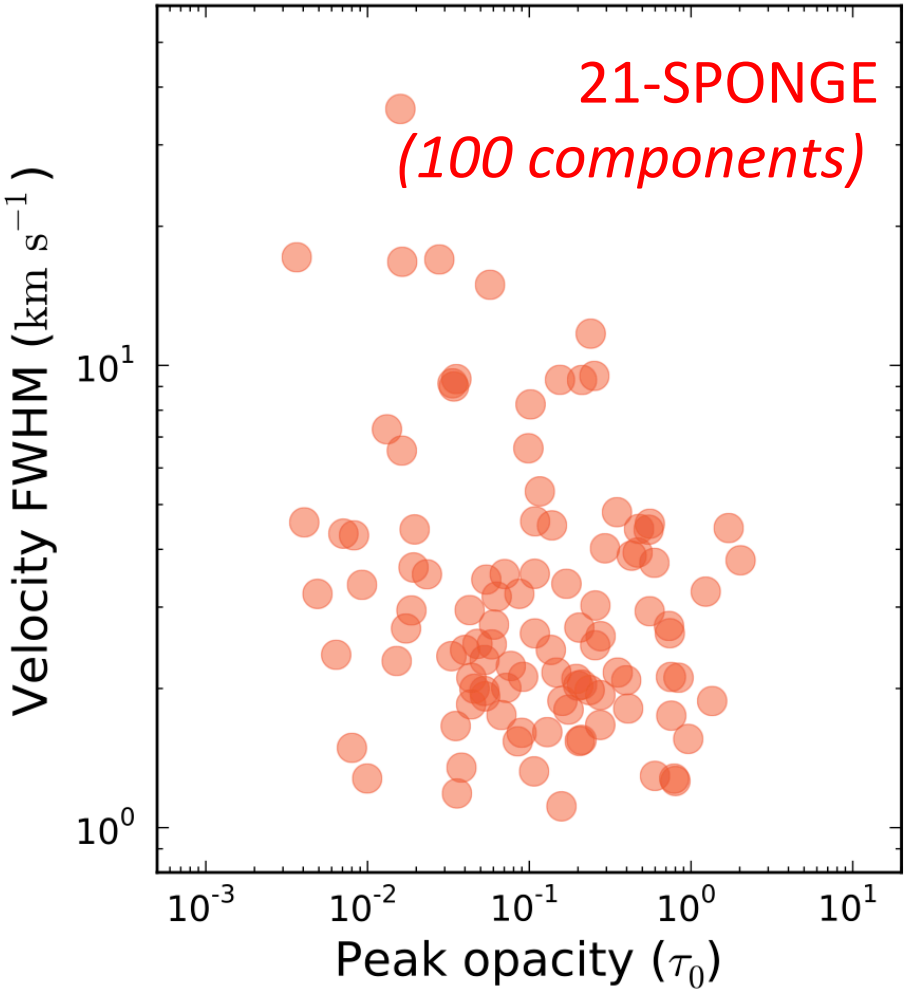


Results of AGD

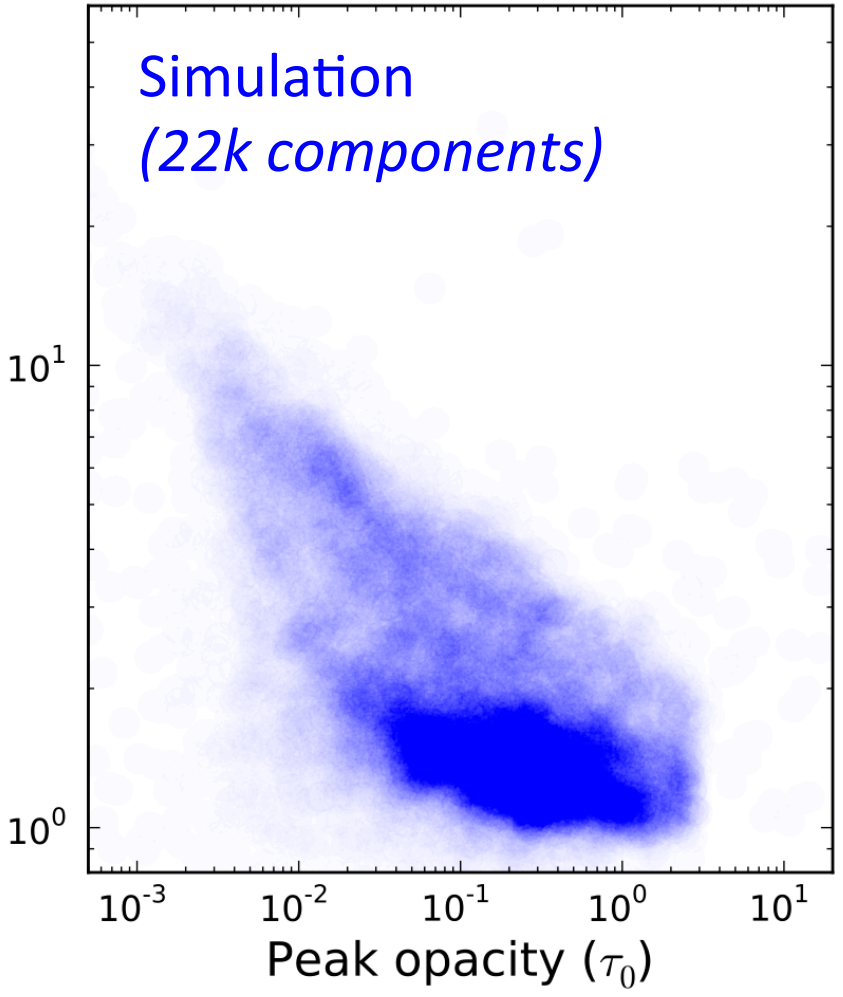


Simulation:
Hennebelle & Audit (2007)

Physical structures in data finally revealed with large data volumes



21cm spectra from Murray et al. (2015)

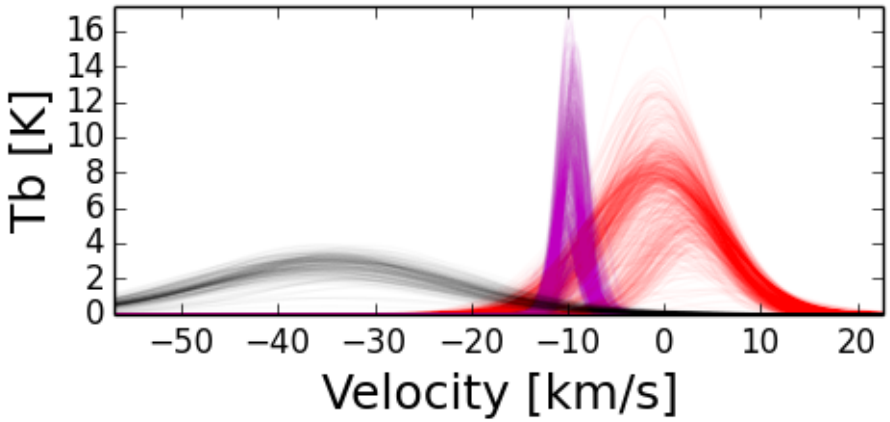


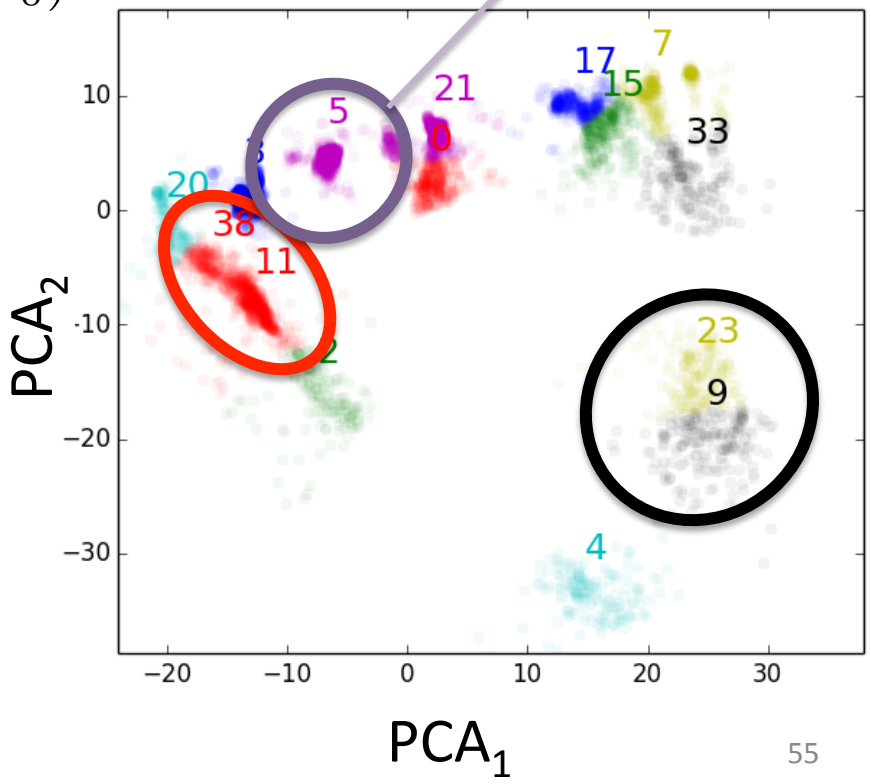
Synthetic spectra from:
Hennebelle & Audit (2007)

Gaussian decomposition

- 32x32 pixels, 1024 spectra
- 8400 components
- Manual decomposition finds 6700 (C. Heiles)
- K-means clustering on $(\tau_0, \Delta v, v_0)$ to find related "clusters" of components

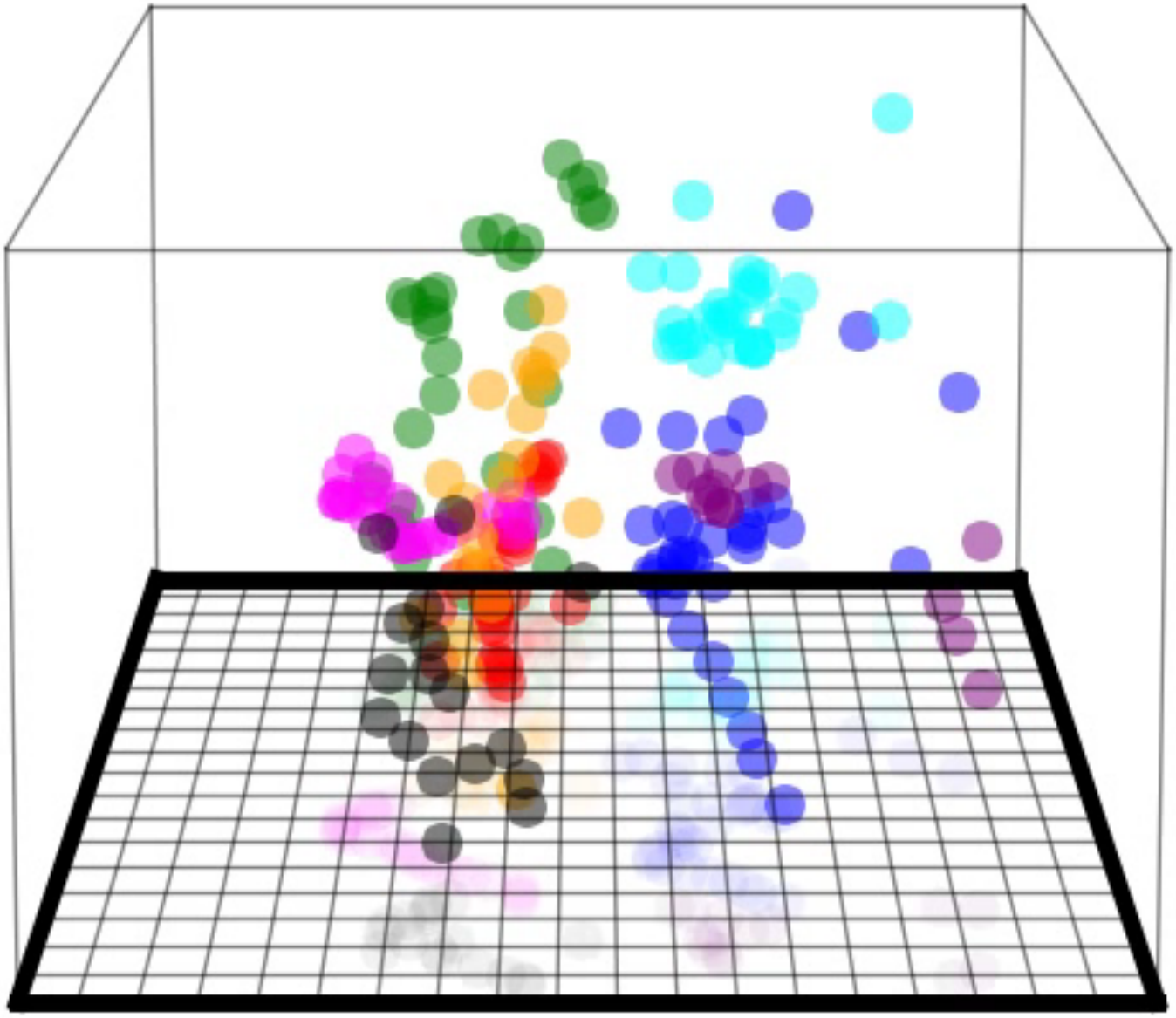
"Cluster 5" has narrowest mean velocity width





Distinct clusters (clouds) from Audit & Hennebelle (2007) simulation

Component Features $(\tau_0, \Delta v, \langle v \rangle, x)_i$ reduced from $4\mathrm{D} \rightarrow 3\mathrm{D}$ using PCA.



Ward minimum-variance Hierarchical clustering

The End. Thank you.

- The "Autonomous Gaussian Decomposition" (AGD) algorithm was published in The Astronomical Journal in April 2015.
- AGD is implemented as a Python/C module named GaussPy . It will be open sourced, but is currently in testiing in a scientific closed beta. (gausspy.software@gmail.com)
- The Square Kilometer Array will be fully operational around the year 2020.