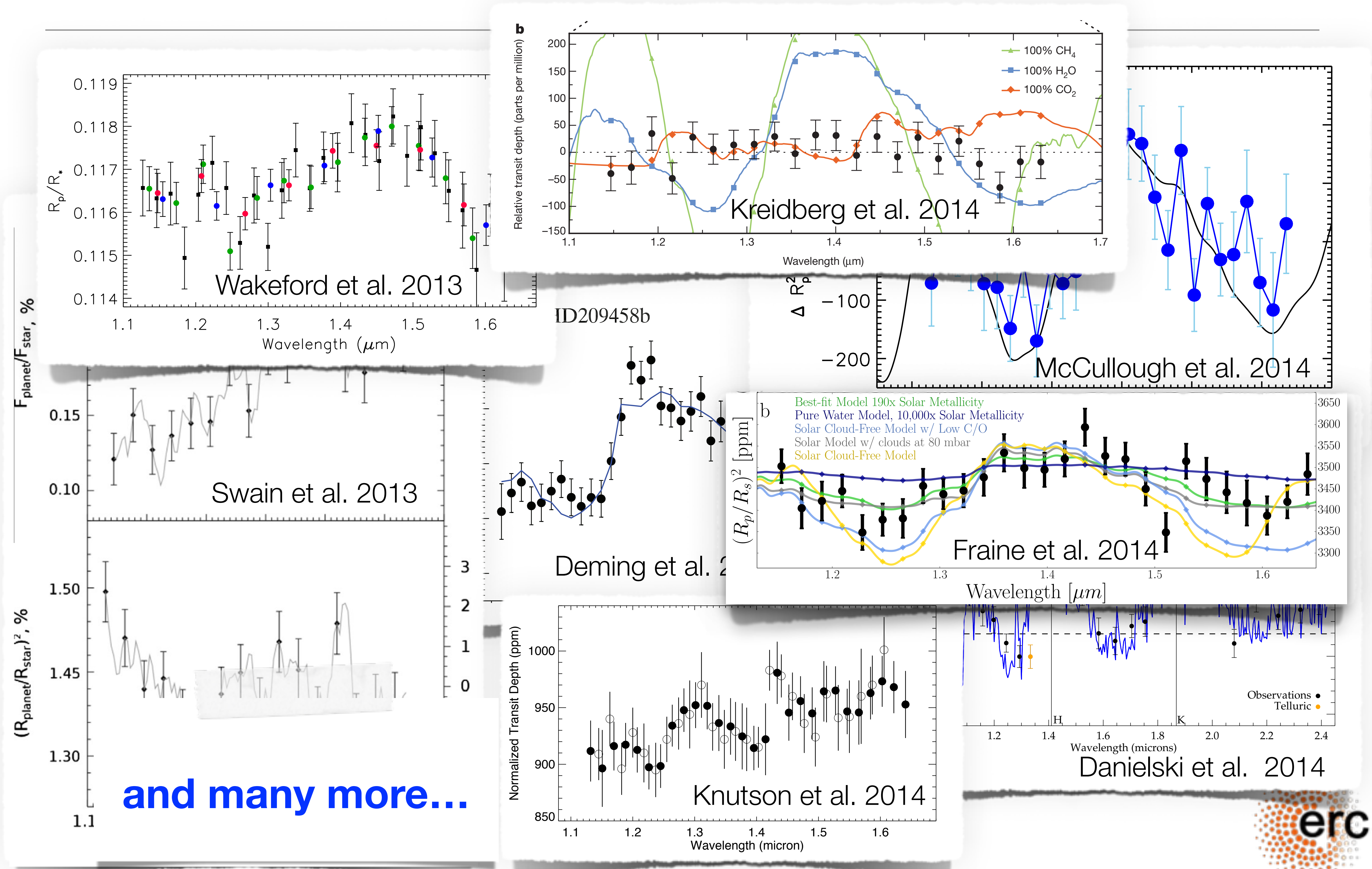
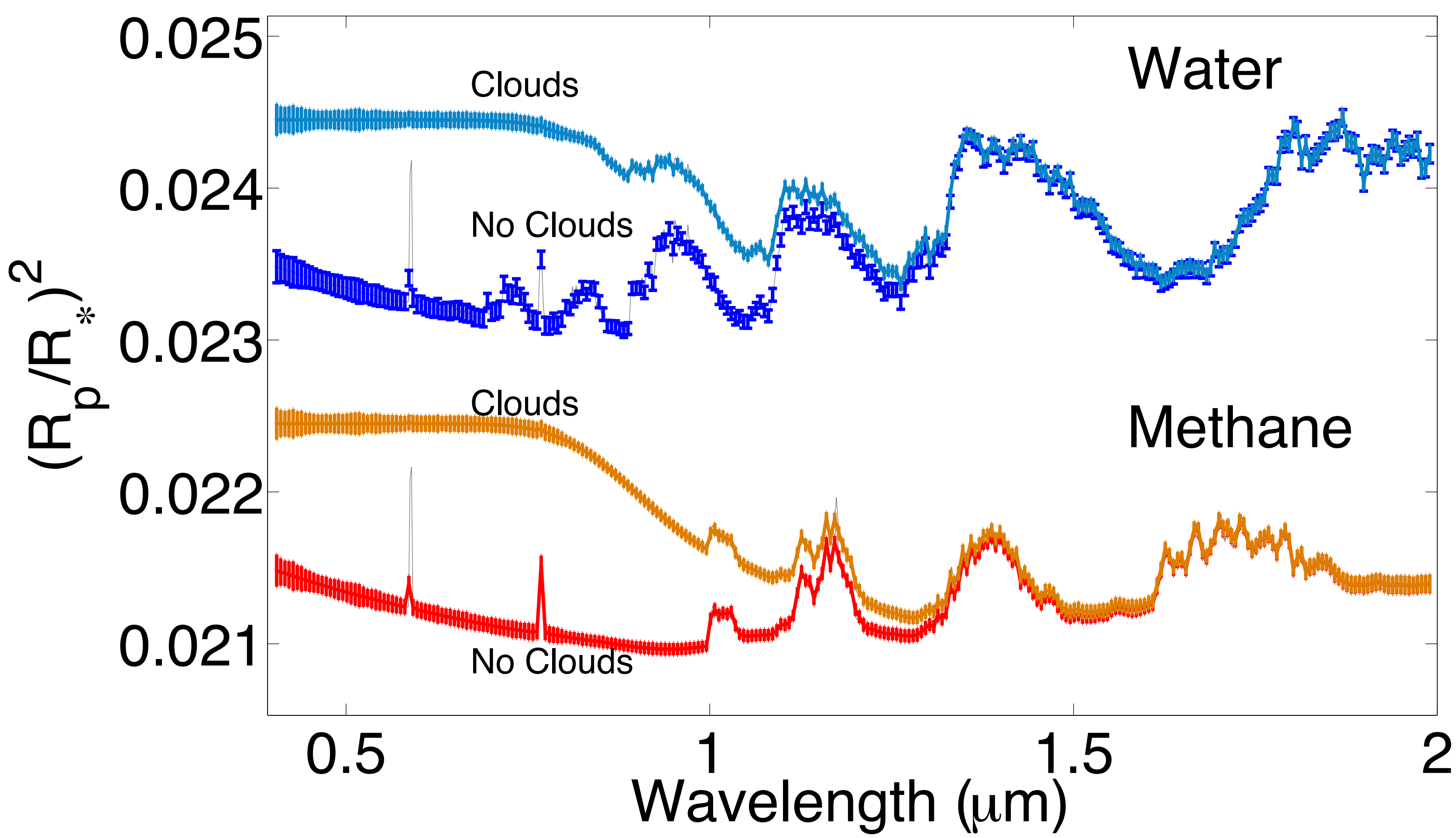

# Developing an integrated approach to exoplanetary spectroscopy

NAM 2015: Ingo P. Waldmann

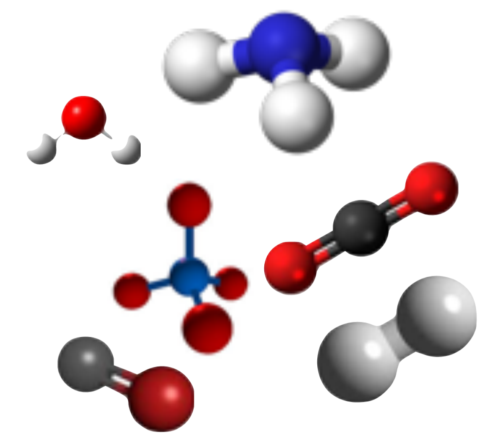
# Exoplanetary spectroscopy is thriving …



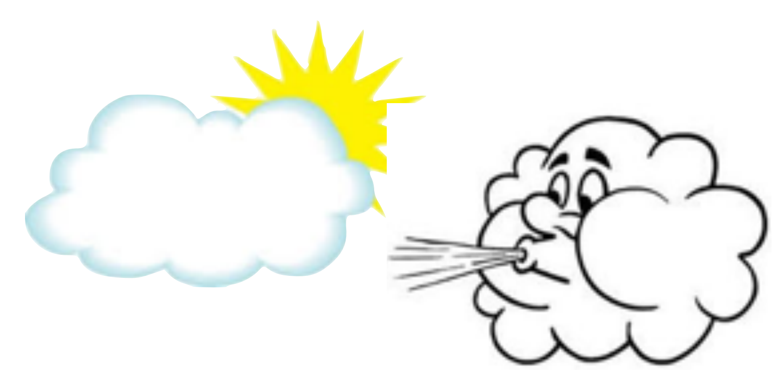
# What we can learn: Interpreting the atmosphere




composition

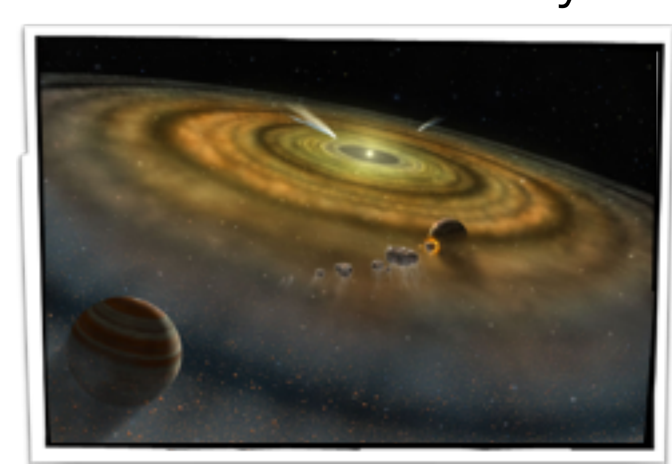


clouds + dynamics

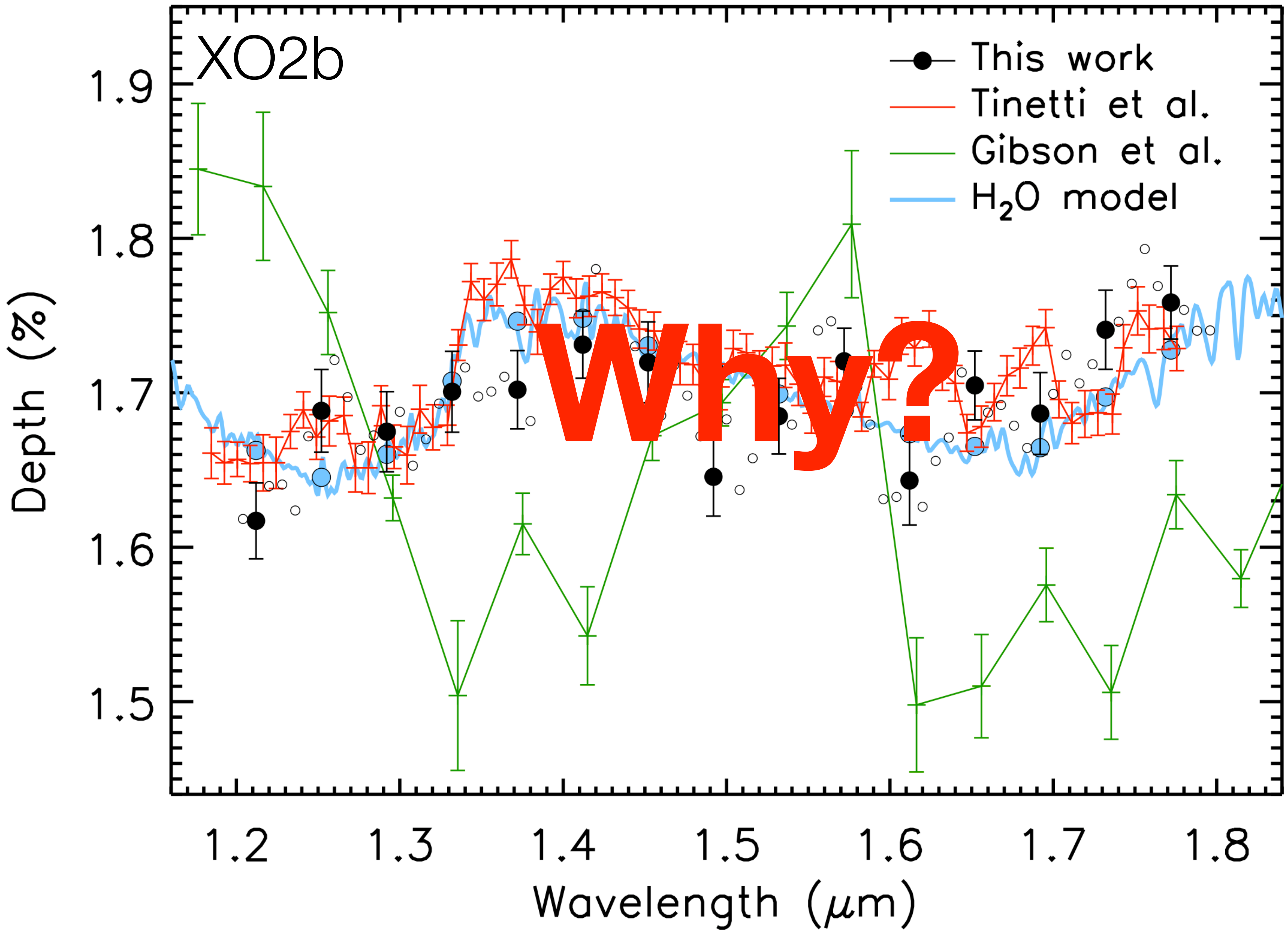


other stuff

formation history



# …, but we not always agree…




Crouzet et al. (2012)

# The data and the models

## Data

- Low Signal-to-Noise data
- Analysis often depending on parametric solutions
- Inherent biases and error covariances associated to each analysis framework
- Coherent analysis of multiple data sets
- Non-parametric approach

## Models

- Highly correlated parameter spaces
- User defined molecule selections/inputs
- Molecular opacity Line Lists
- Self-consistent vs data quality
- Full mapping of correlated likelihoods
- Non-parametric approach

# The data and the models

## Data

- Low Signal-to-Noise data
- Analysis still often depending on parametric solutions
- Inherent biases and error covariances associated to each analysis framework
- Coherent analysis of multiple data sets
- Non-parametric approach

## Models

- Highly correlated parameter spaces
- User defined molecule selections/inputs
- Molecular opacity Line Lists
- Ab initio models vs data driven
- Full mapping of correlated likelihoods
- Non-parametric approach

# Non-parametric data analysis

What if we will never know the instrument response of most generic instruments and don't know how to calibrate it?

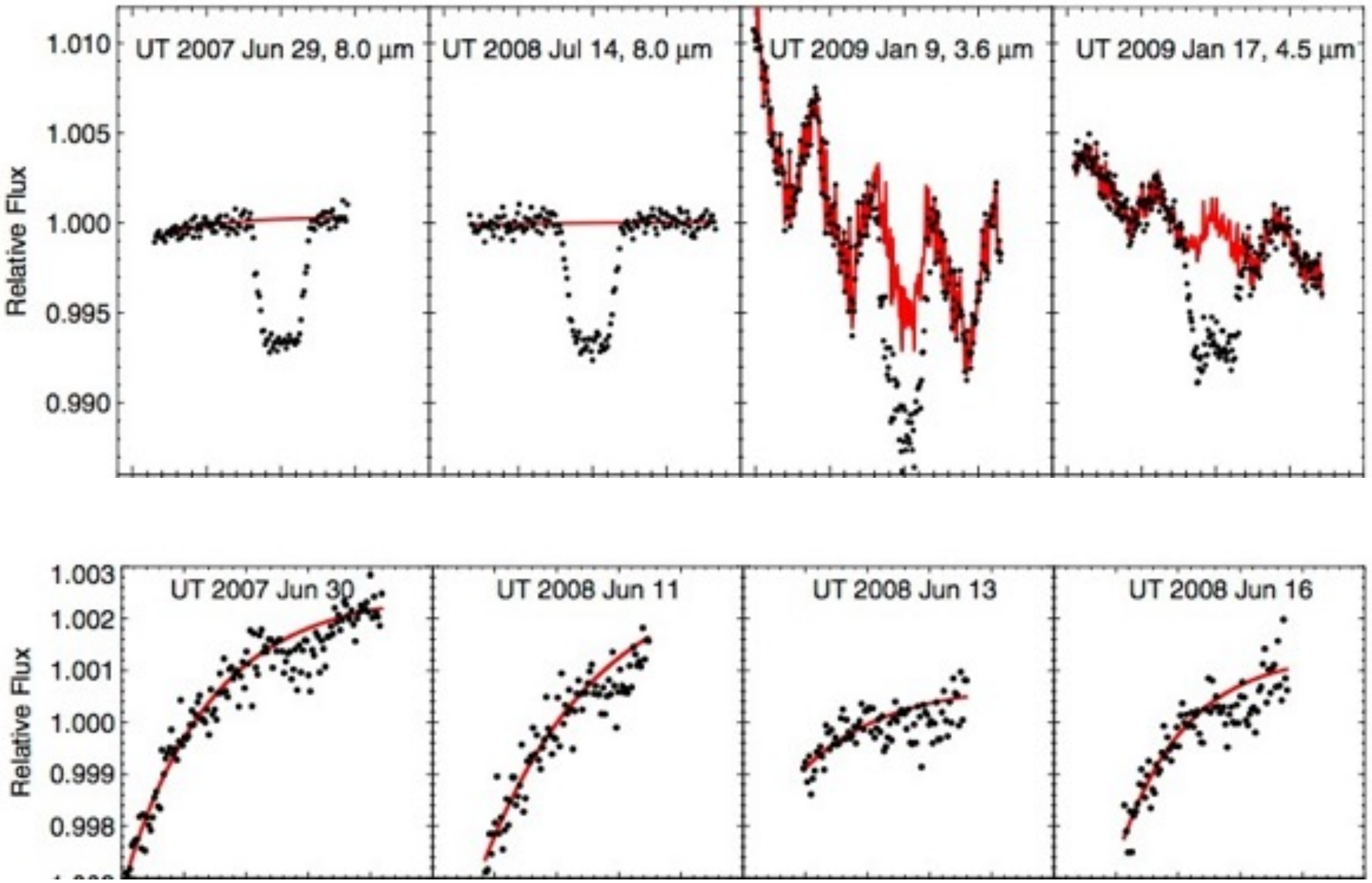
Can we still do something with the data?

- Cleaning the data e.g. wavelet decomposition
- De-trending the data using statistics
    - Supervised machine learning
    - Unsupervised machine learning

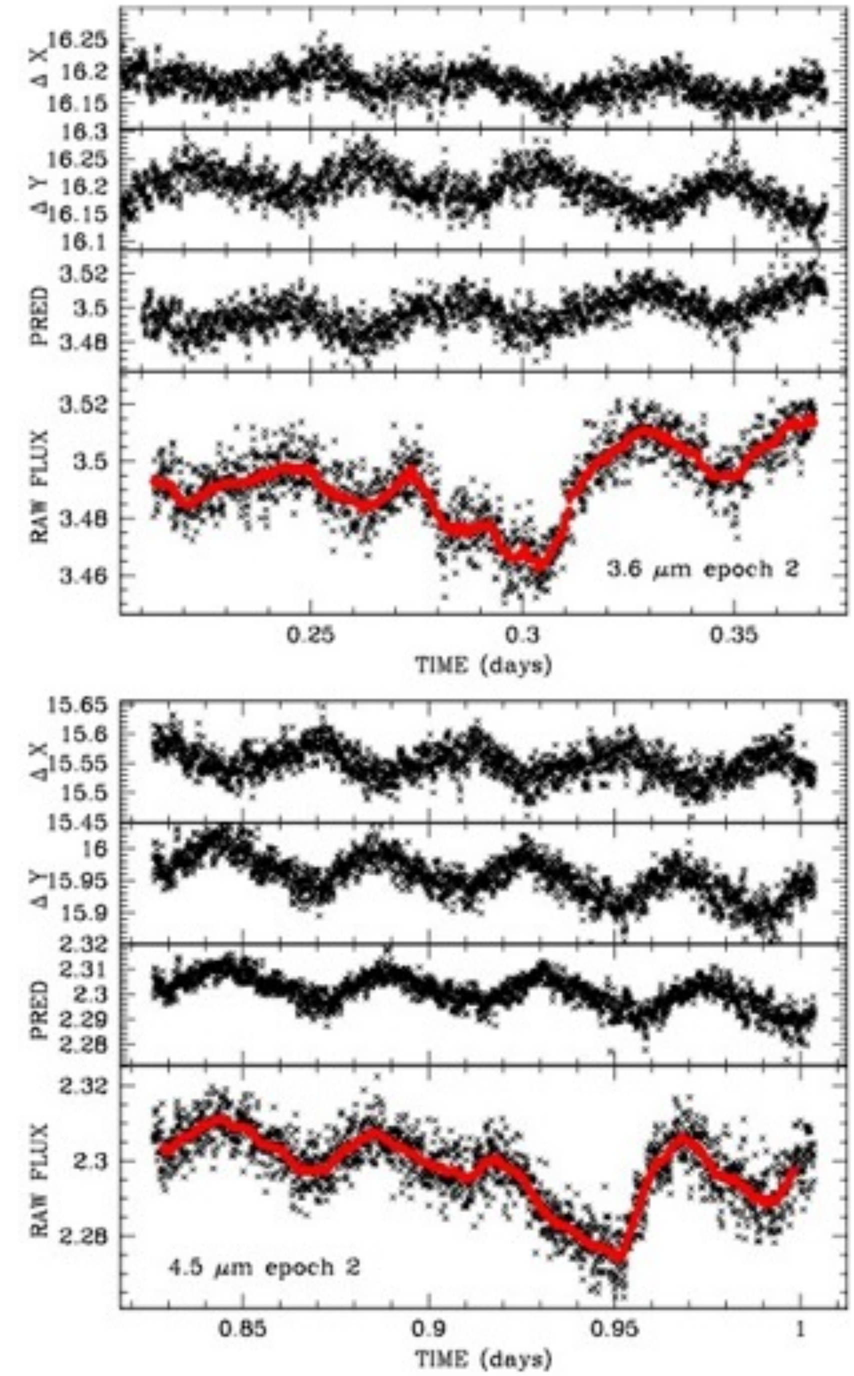
# Example: Spitzer - IRAC

## The issue of persistence and inter and intra pixel variations

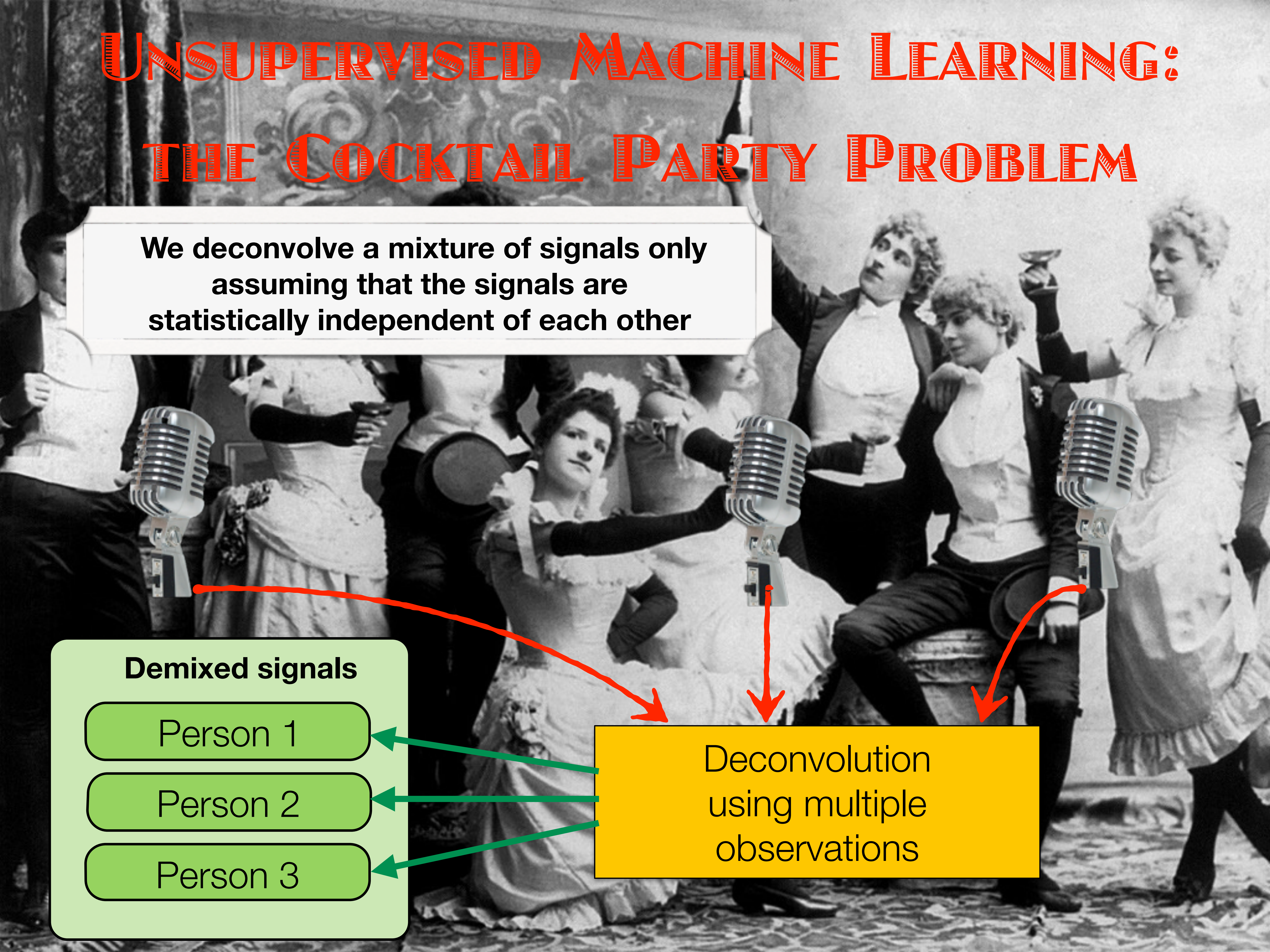
Spitzer/IRAC, GJ436b




Knutson et al. 2011




Beaulieu et al. 2011


Unsupervised Machine Learning:
the Cocktail Party Problem
We deconvolve a mixture of signals only assuming that the signals are statistically independent of each other
Demixed signals
Person 1
Person 2
Person 3
Deconvolution using multiple observations

# Using information entropy to de-correlate data

$$x_k = a_{k,1} s_a + a_{k,2} s_{wn} + \sum_{l=3}^{N_{sn}} a_{k,l} s_{sn}$$

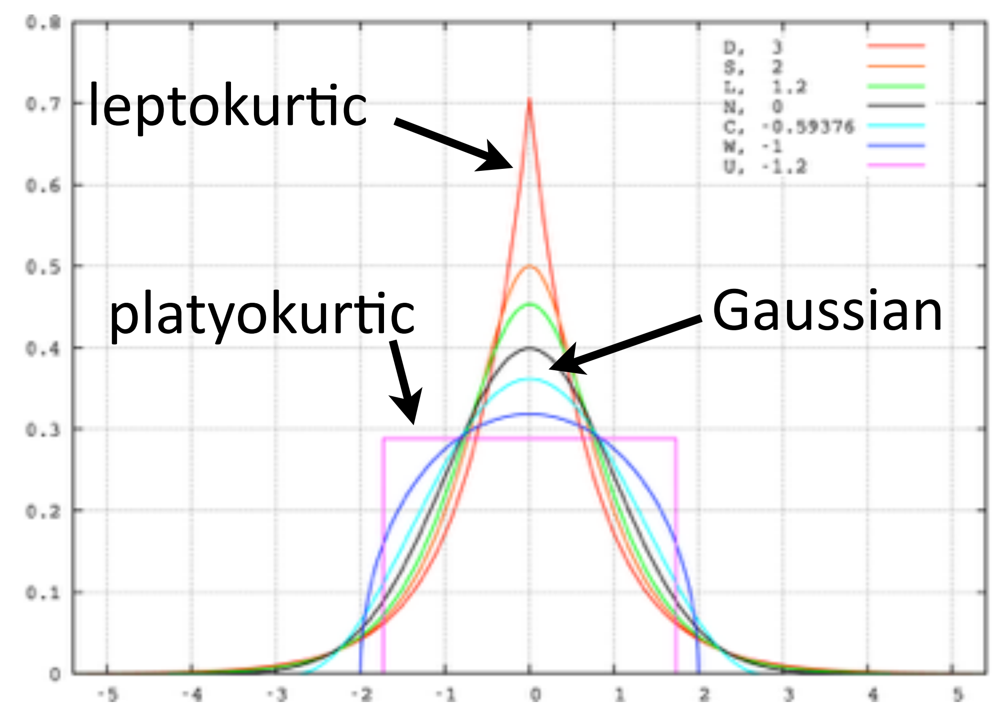
observed time series

astrophysical source

white noise

systematic noise

Using kurtosis as measure of non-Gaussianity




We optimise the statistical independence of the source signals, s, by minimising their respective Shannon entropies.

$$\mathrm{H}(\mathbf{y}) = -\int p(\mathbf{y}) \log_2 p(\mathbf{y}) \mathrm{d}\mathbf{y}$$

observations

mixing matrix

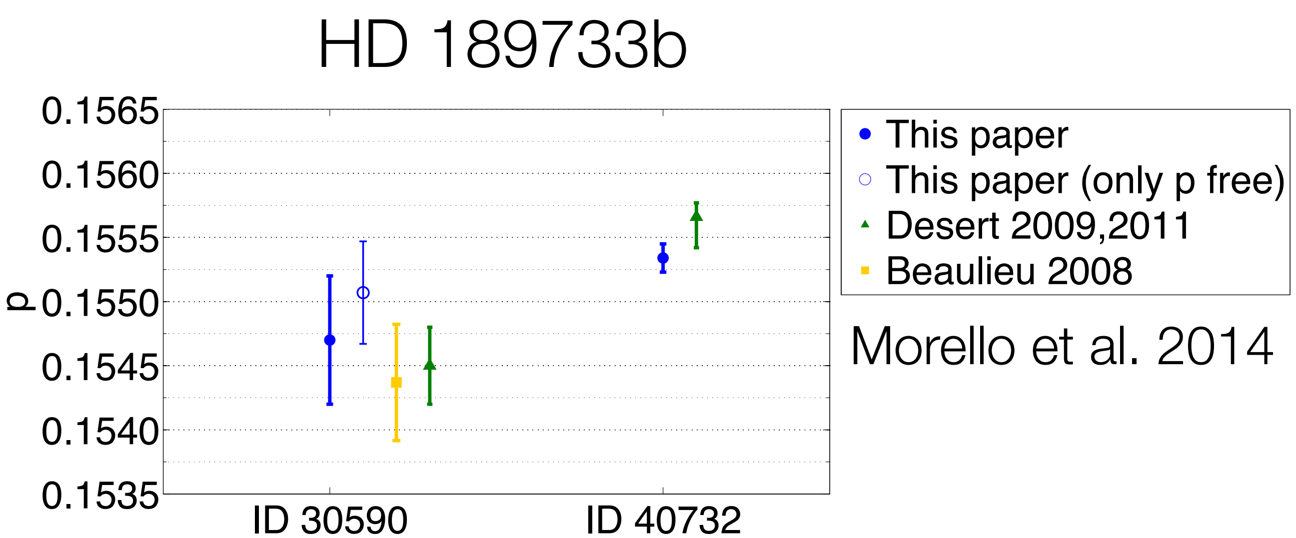
X = AS

signals

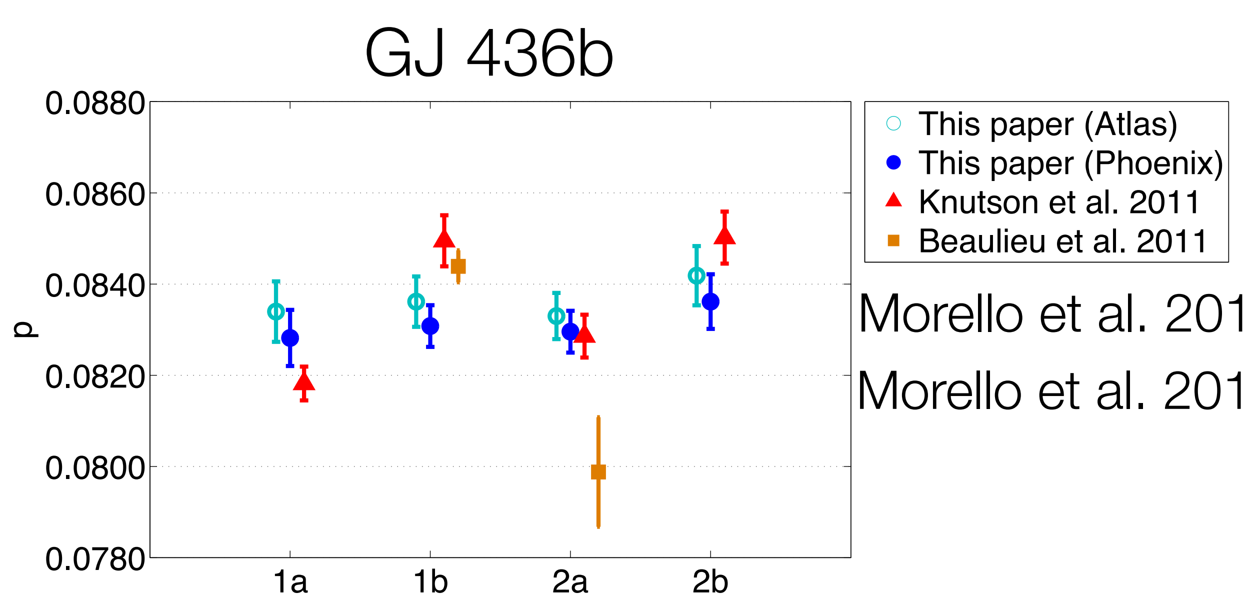
Waldmann 2012

# Unsupervised learning: De-trending Spitzer/IRAC

De-trending Spitzer **photometry** without prior assumptions gives (for the first time) consistent results across data sets

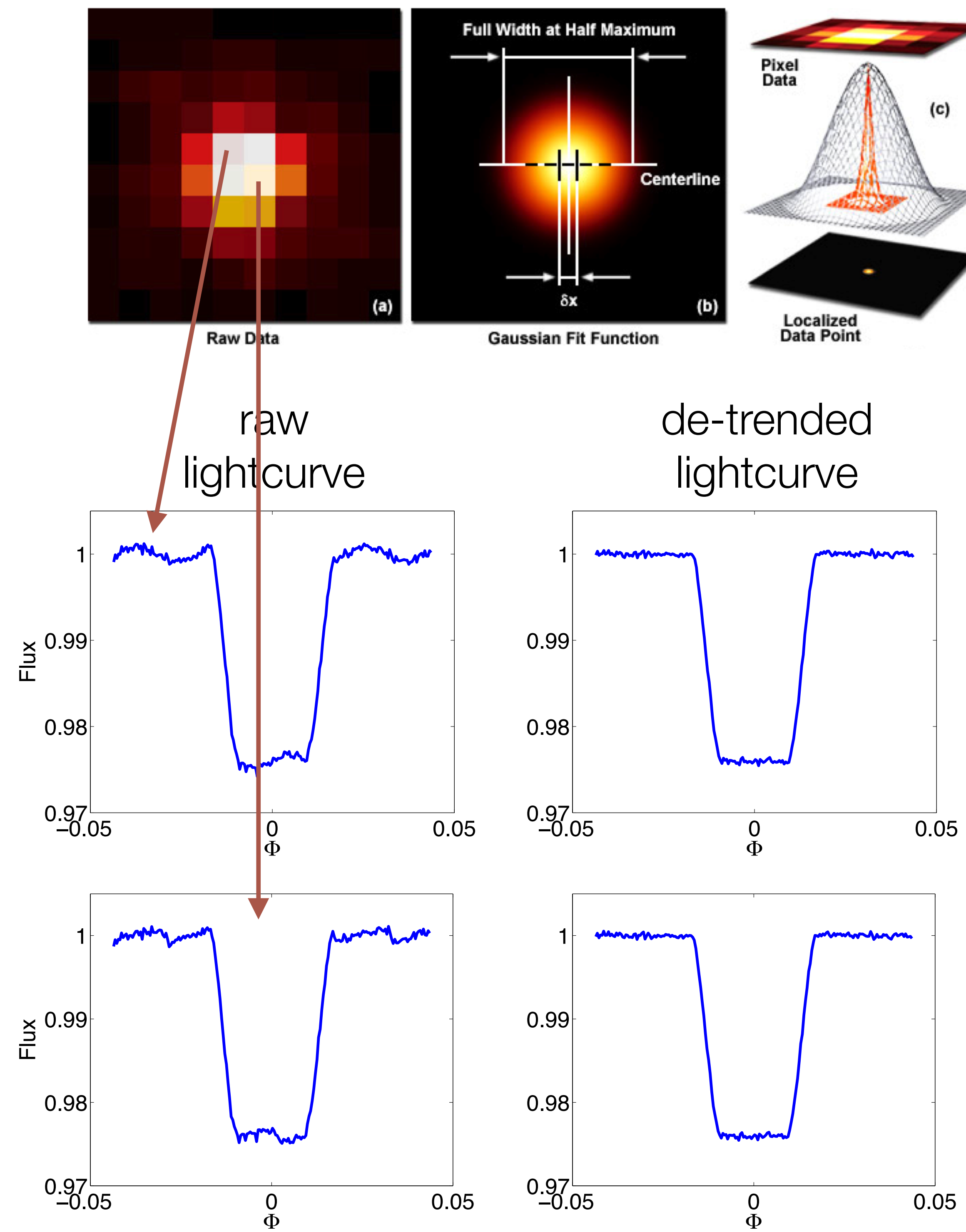



Morello et al. 2014



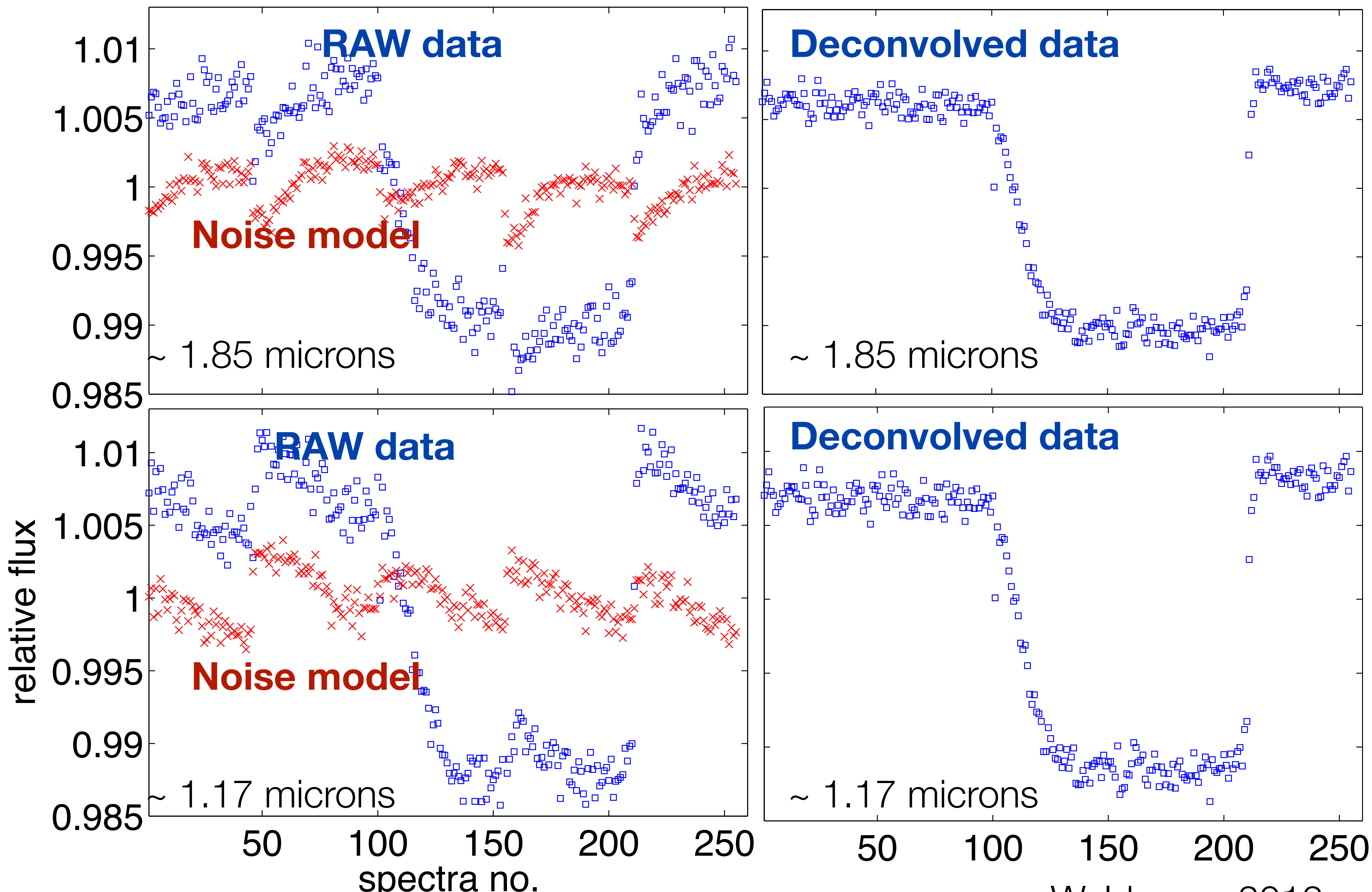

Morello et al. 2015a

Morello et al. 2015b




Morello et al. 2015

# XO1b spectroscopy observed by Hubble/NICMOS




Waldmann 2012

# The data and the models

## Data

- Low Signal-to-Noise data
- Analysis still often depending on parametric solutions
- Inherent biases and error covariances associated to each analysis framework
- Coherent analysis of multiple data sets
- Non-parametric approach

## Models

- Highly correlated parameter spaces
- User defined molecule selections/inputs
- Molecular opacity Line Lists
- Ab initio models vs data driven
- Full mapping of correlated likelihoods
- Non-parametric approach

# The data and the models

## Data

- Low Signal-to-Noise data
- Analysis still often depending on parametric solutions
- Inherent biases and error covariances associated to each analysis framework
- Coherent analysis of multiple data sets
- Non-parametric approach

## Models

- Highly correlated parameter spaces
- User defined molecule selections/inputs
- Molecular opacity Line Lists
- Ab initio models vs data driven
- Full mapping of correlated likelihoods
- Non-parametric approach

# Models: The TauREx retrieval framework

- A signal-processing approach
- New custom built high temp. line-lists
- Advances in Bayesian sampling
- Advances in Pattern recognition
- Advances in large scale automation

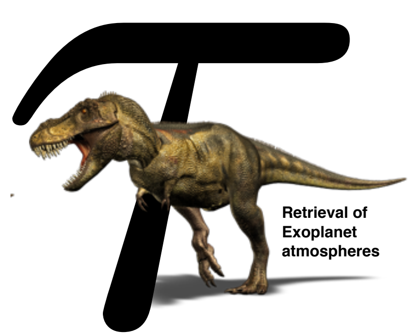
## current retrievals

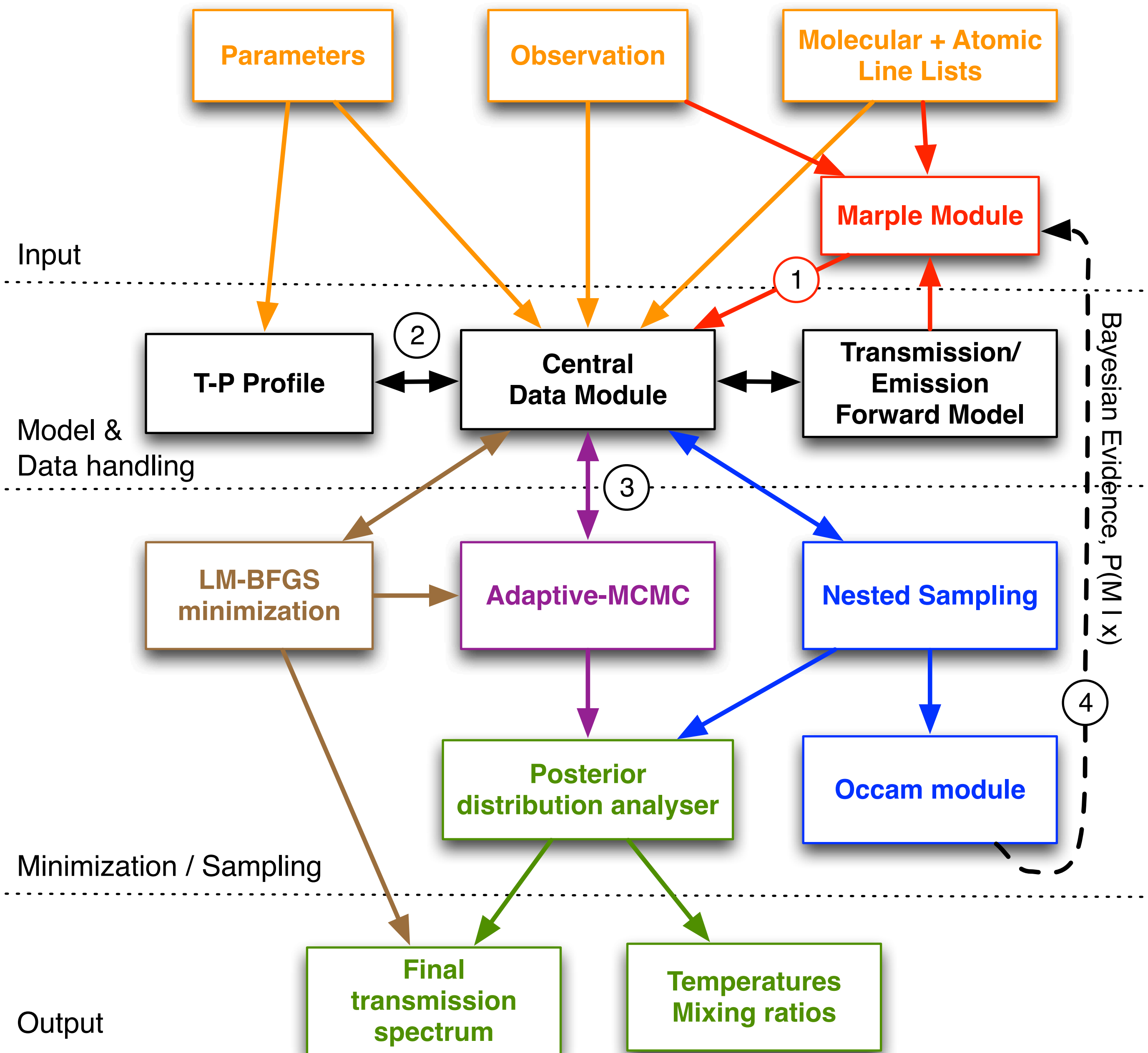
Fletcher et al. 2007
Terrile et al. 2008
Madhusudhan & Seager 2009
Lee et al. 2011
Line et al. 2012
Benneke & Seager 2012
Griffith 2014

Waldmann et al. 2015a,b; Rocchetto et al. in prep

# Tau-REx - Next Gen atmospheric retrieval

- **Fully Bayesian Retrieval**
  - MCMC
  - Nested Sampling
- **Custom made opacity line-lists** from the ExoMol project
- Prior composition selection through **pattern recognition software**
- **Full parallelisation for cluster computing**




Waldmann et al. 2015a,b; Rocchetto et al. in prep

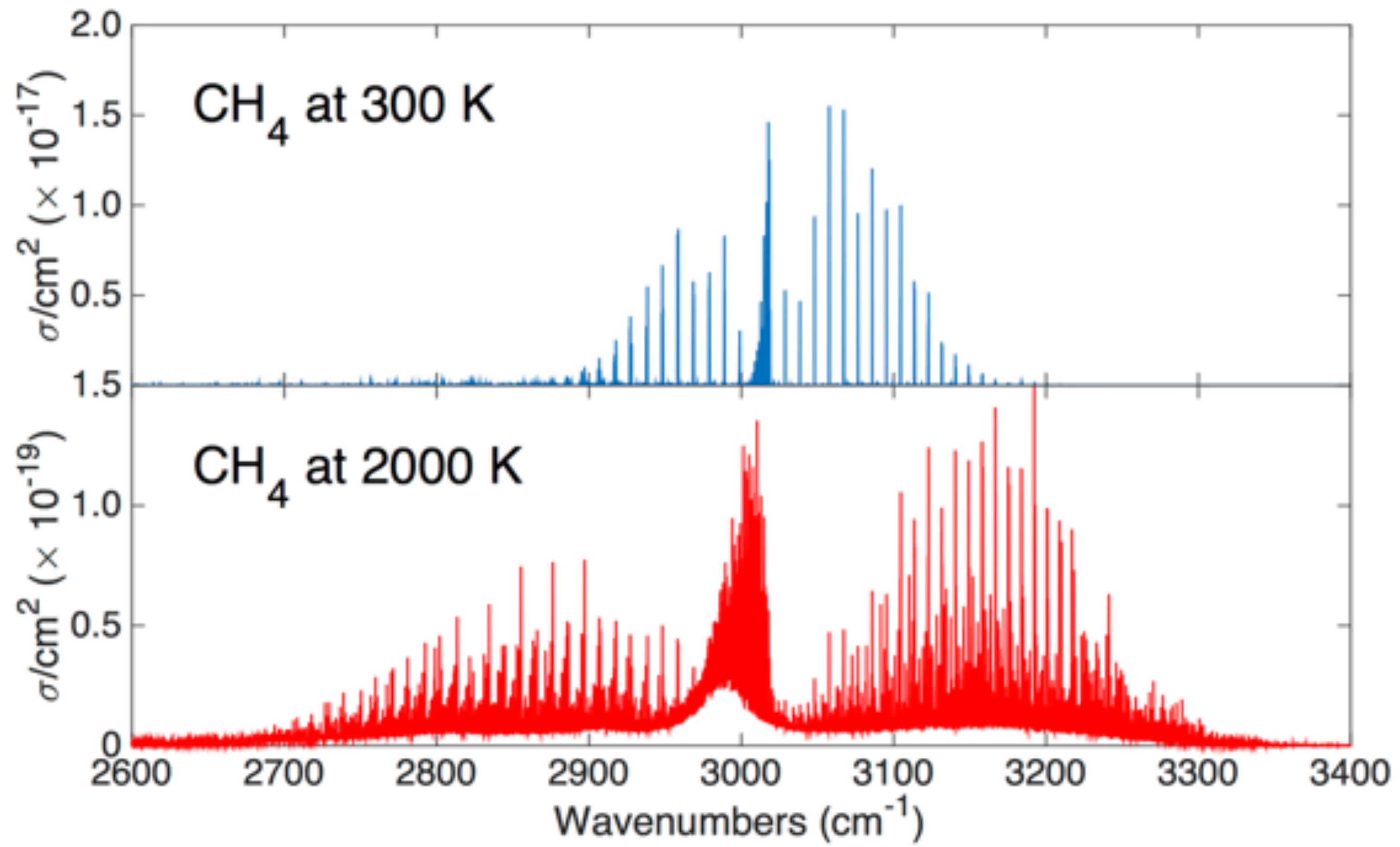
# Custom built line lists

High temperature ExoMol line-lists

Line-by-line forward model

Non-linearly sampled for optimal computation

Exact line broadening




Waldmann et al. 2015a,b; Rocchetto et al. in prep

# The Marple Module

- Constrain prior space by finding most likely absorbers.
- Custom built pattern recognition
- Based on ‘eigenface’ facial recognition

Waldmann et al. 2015a,b; Rocchetto et al. in prep

# Emission Spectroscopy: The issue with TP profiles…

- In emission spectroscopy we must solve for the atmospheric opacities as well as the atmospheric temperature-pressure profile.

- The Temperature-Pressure profiles are degenerate and notoriously hard to constrain.

**Parametric TP-profiles**

+ Easy to implement
+ Few (<10) free parameters
+ Good convergence in low S/N regimes

- Can only fit TP-profile within predetermined functional form
- Potentially unrealistic assumptions

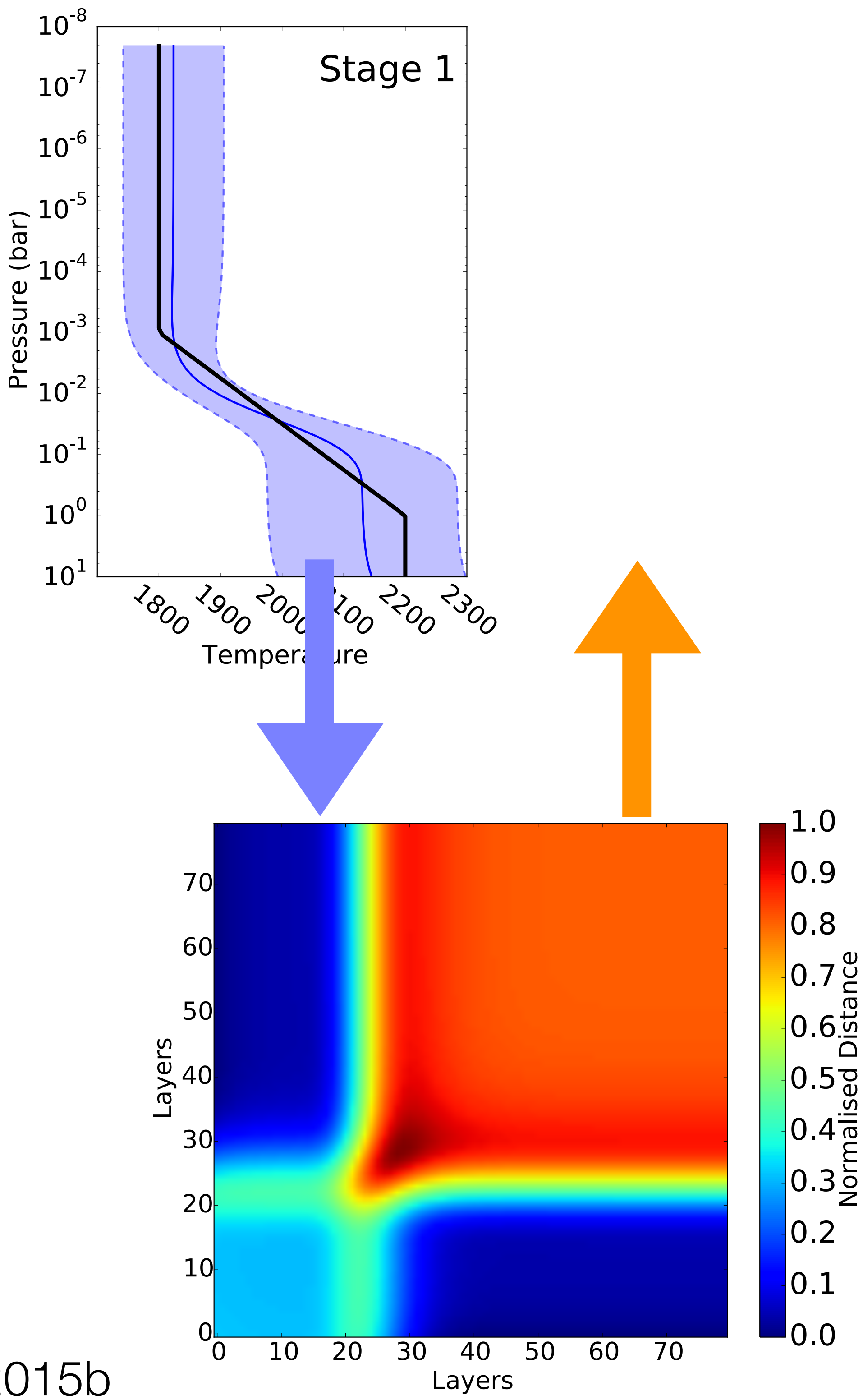
**Layer-by-layer TP-profiles**

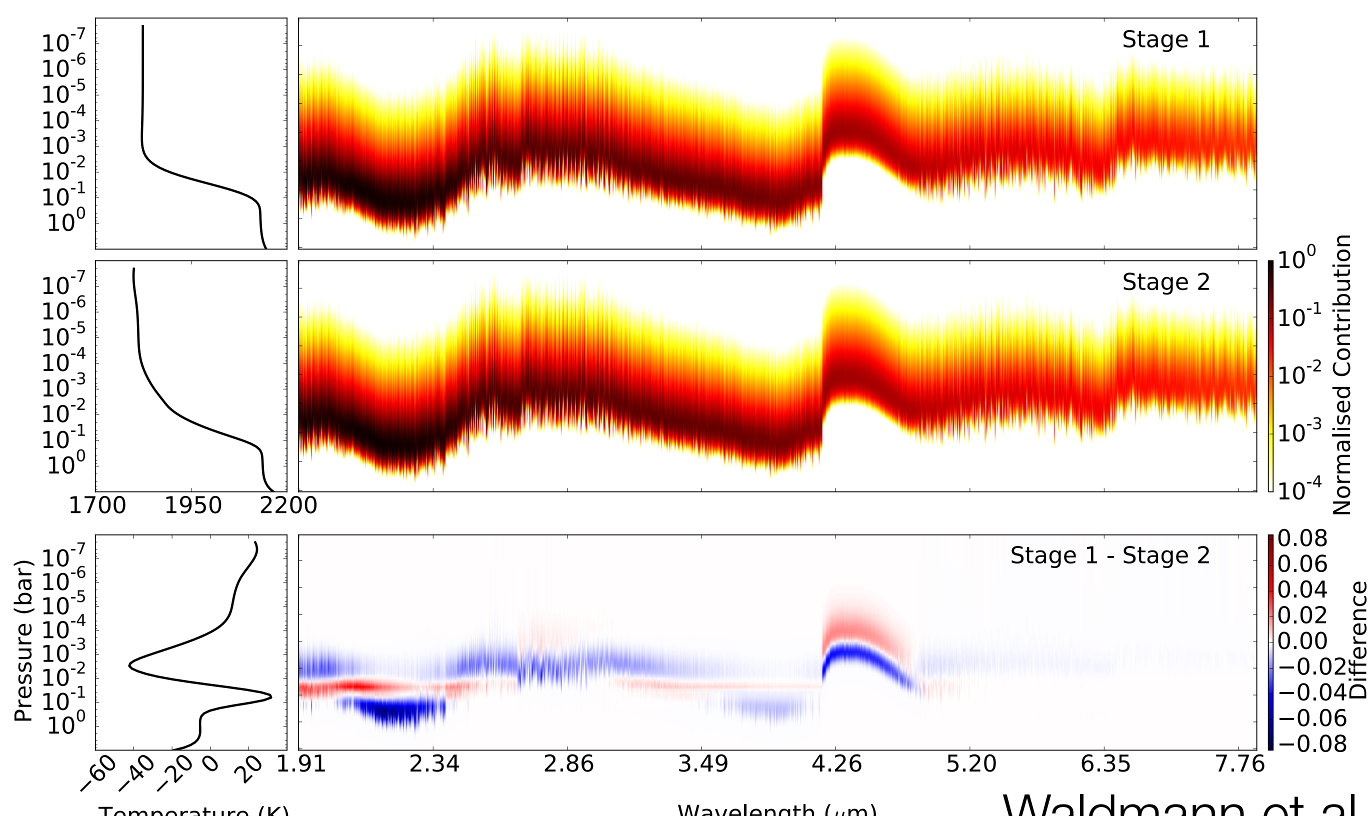
+ Very objective -> No assumptions on atmosphere.

- Many (>30) free parameters
- Poor convergence properties in low S/N regimes

# Emission Spectroscopy: The issue with TP profiles…

- Compute parametric solution
- Obtain temperature-pressure covariance
- Relax parametric solution to layer-by-layer TP-profile, using the covariance as convergence aid
- Achieves a 'fine-tuning' of the parametric solution.




Waldmann et al. 2015b

# The data and the models

## Data

- Low Signal-to-Noise data
- Analysis still often depending on parametric solutions
- Inherent biases and error covariances associated to each analysis framework
- Coherent analysis of multiple data sets
- Non-parametric approach

## Models

- Highly correlated parameter spaces
- User defined molecule selections/inputs
- Molecular opacity Line Lists
- Ab initio models vs data driven
- Full mapping of correlated likelihoods
- Non-parametric approach

# The data and the models

## Data

- Low Signal-to-Noise data
- Analysis still often depending on parametric solutions
- Inherent biases and error covariances associated to each analysis framework
- Coherent analysis of multiple data sets
- Non-parametric approach

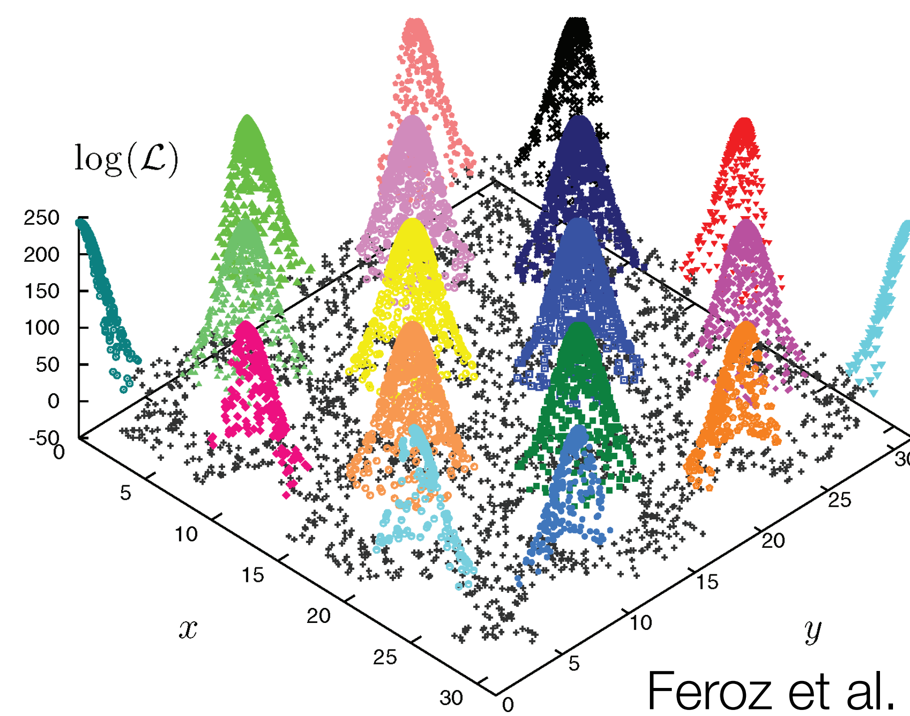
## Models

- Highly correlated parameter spaces
- User defined molecule selections/inputs
- Molecular opacity Line Lists
- Ab initio models vs data driven
- Full mapping of correlated likelihoods
- Non-parametric approach

# Coherently mapping the correlated parameter spaces

- Working towards an integrated approach
- From **raw data -> final molecular abundances** in one coherent retrieval.
- Unified likelihood space correctly propagates all errors and correlations

- Using MCMC and Nested Sampling
- Fully map the likelihood space
- Global model selection
  (Bayesian Evidence)
- Nested model selection
  (Savage Dickey Ratio)




Feroz et al. 2009

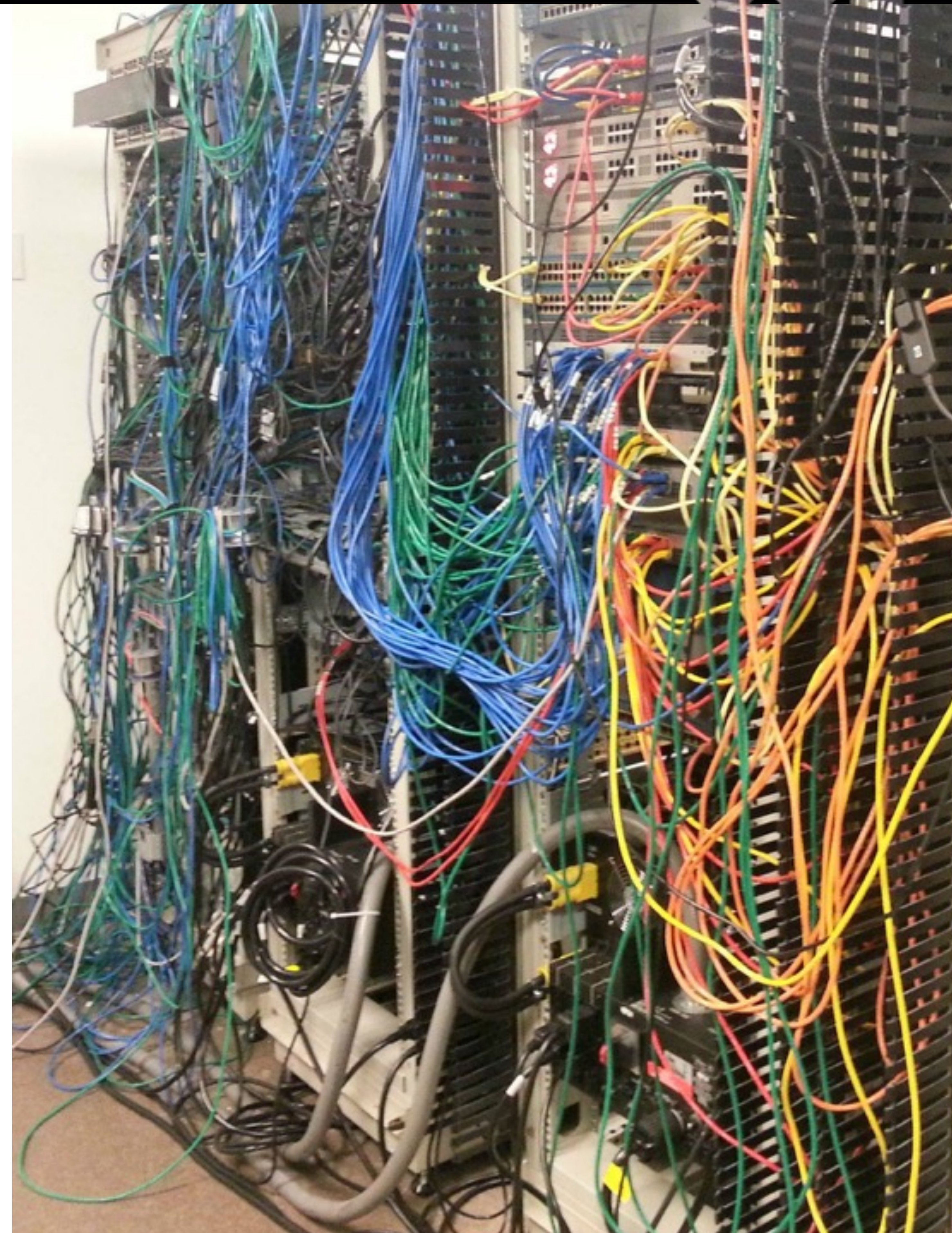
# Building the infrastructure

Data de-trending + retrieval models unified in objective Python & C++

Full MPI and OpenMP support for cluster computing

Intelligent de-bugging algorithms and failsafes

Providing the infra-structure for future missions, e.g.: JWST, ARIEL, etc



# Conclusion

- Both data analysis and models need to move towards a less heuristic footing
- In data analysis, Independent Component Analysis, Gaussian Processes and Wavelet approaches become established now
- In atmospheric retrieval models, the synergy between machine learning and fully Bayesian approaches are key
- Both data analysis and models need to work together to fully constrain error bars and map potential biases
- Important infrastructure for JWST and ARIEL

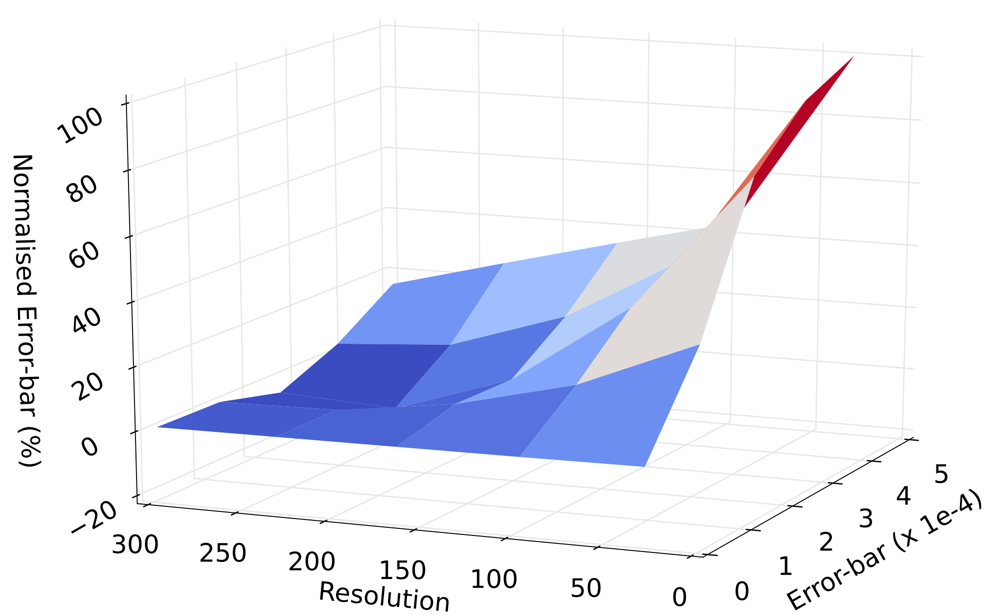
Thank you

# Additional stuff

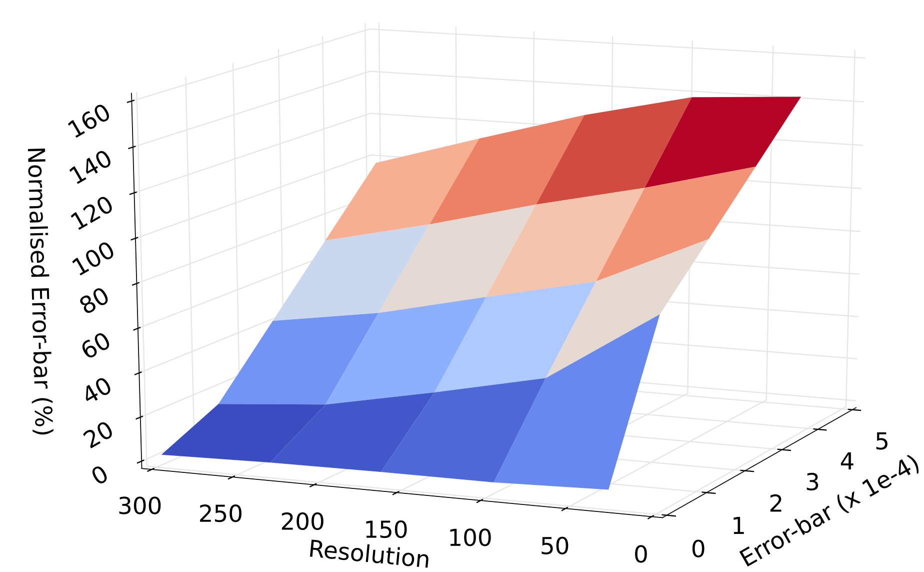
# Not all parameters behave the same way

- Ability to retrieve molecular abundance depends on the nature of the molecule
- Absorbers across a broad wavelength range are less prone to low resolution data
- Temperature retrieval is almost solely dependent on Signal-to-Noise

Error-bar on $H_2O$ abundance




Error-bar on CO abundance

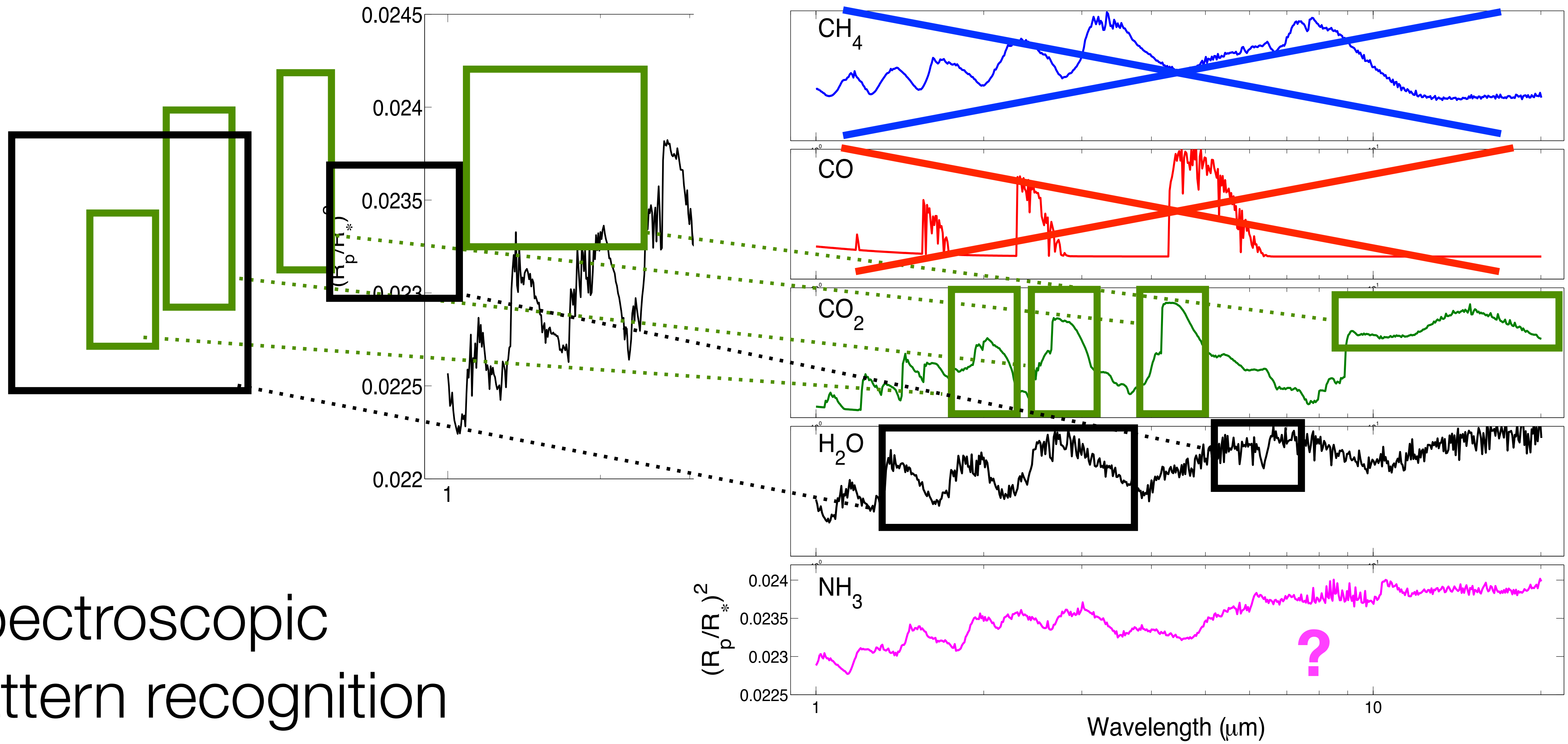



arXiv:1409.2312

# The Marple Module - Constraining the prior space

Observed Spectrum

Individual molecules

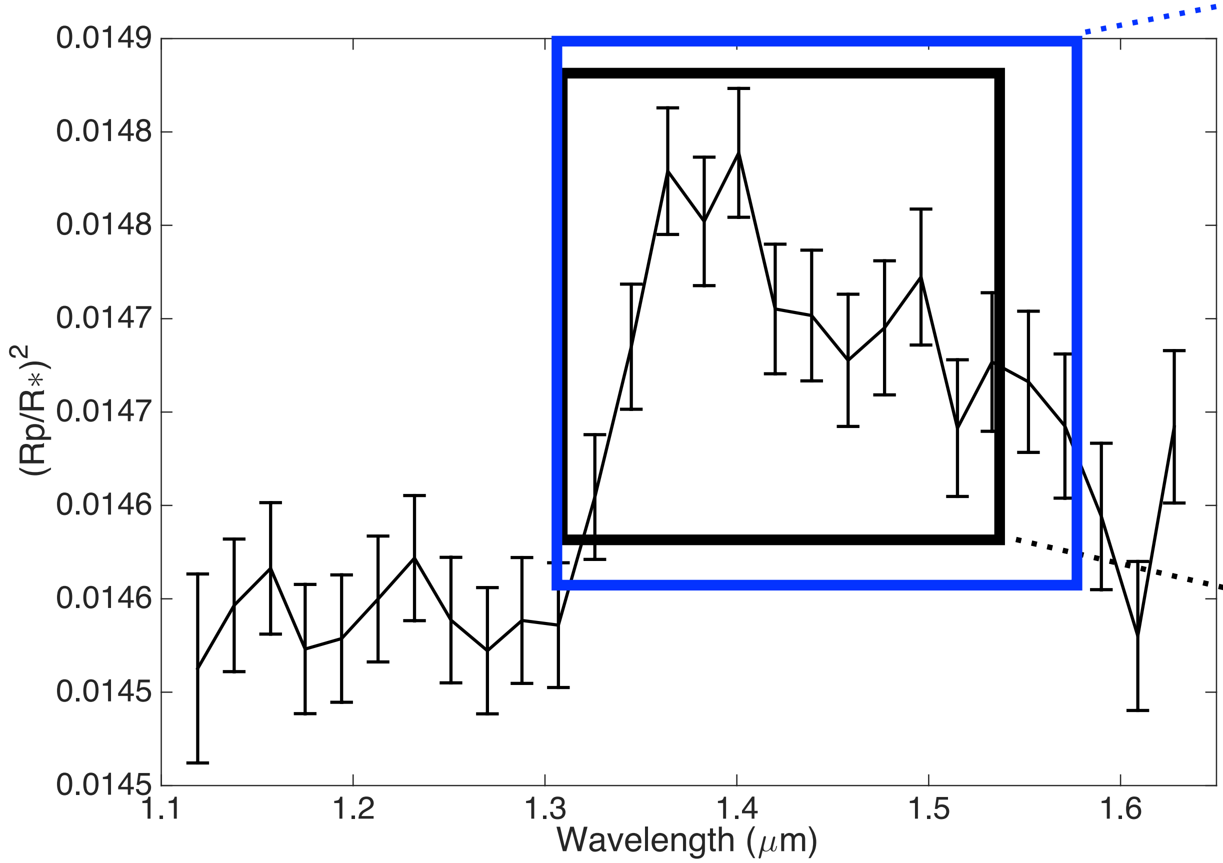



Spectroscopic pattern recognition

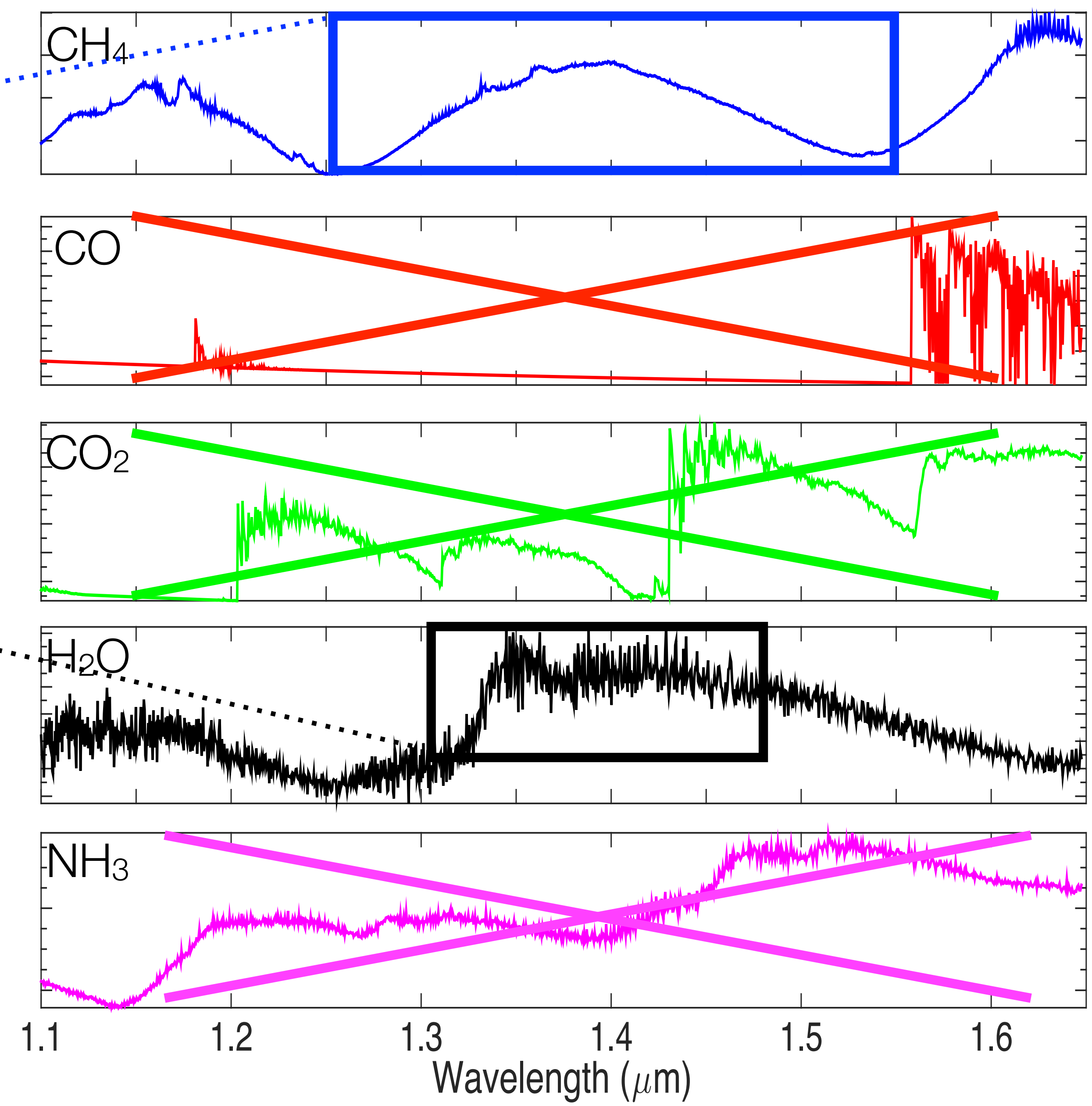
arXiv:1409.2312

# Constraining the prior space - the Marple module

WFC3 HD209458b
(Deming et al. 2013)




Individual molecules




<u>The Marple module:</u>

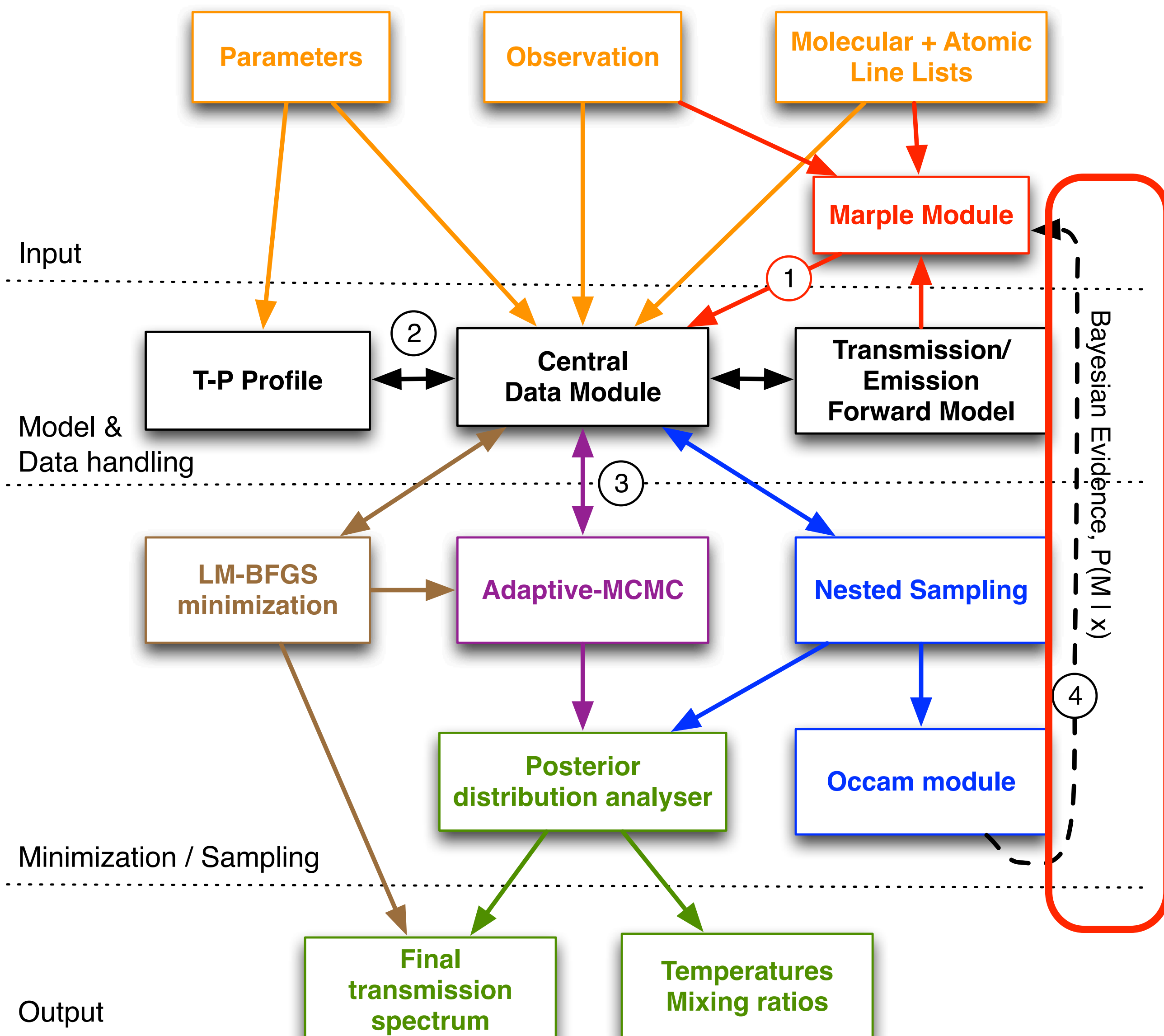
Spectroscopic pattern recognition

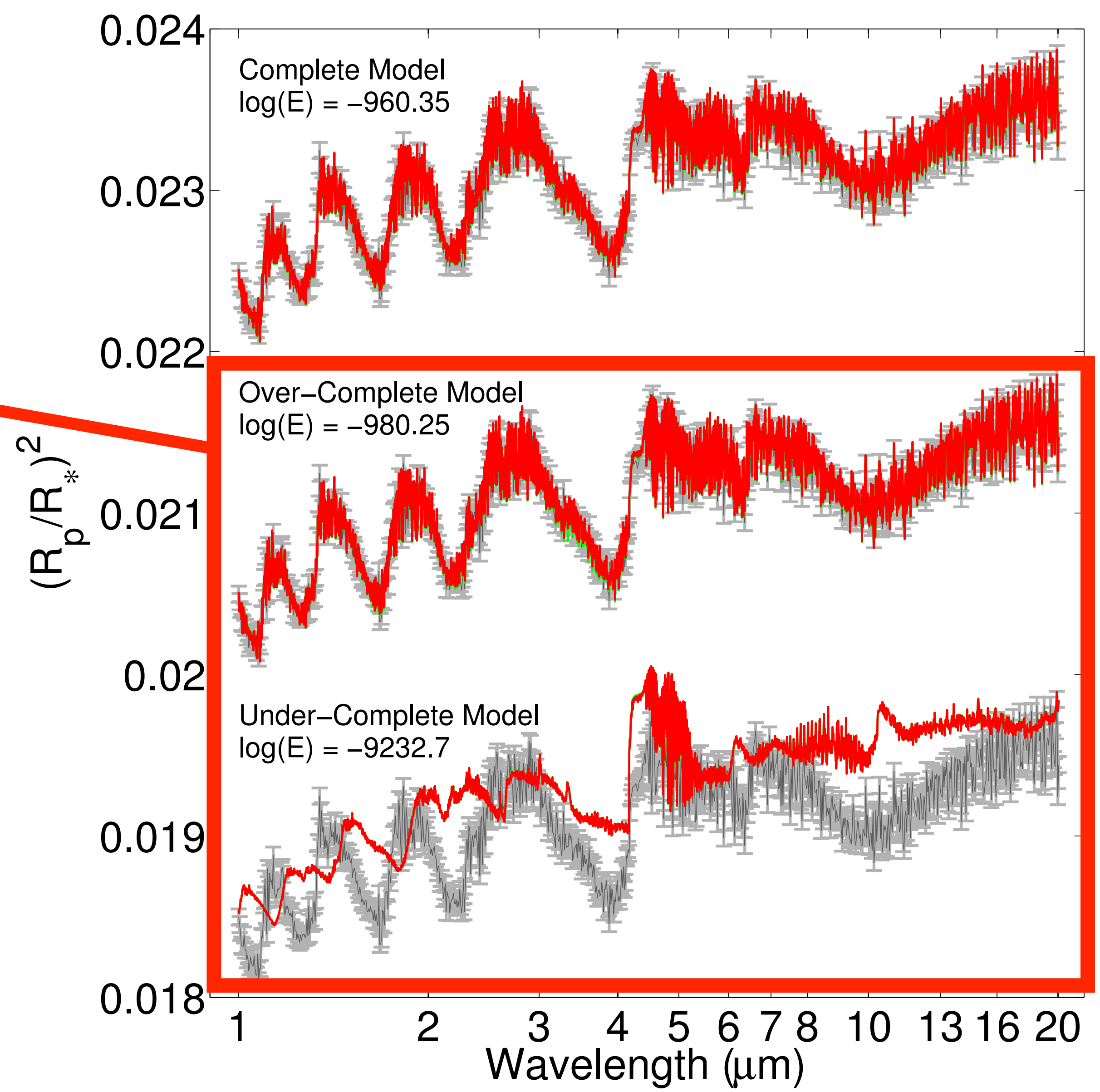
arXiv:1409.2312

# Bayesian Model Selection

Testing for over-complete and under-complete models

$$E = \int P(\boldsymbol{\theta}|\mathcal{M})P(\mathbf{x}|\boldsymbol{\theta},\mathcal{M})\mathrm{d}\boldsymbol{\theta}$$




arXiv:1409.2312