Objective discovery of dominant dynamical regimes

Bryan Kaiser

KITP Machine Learning for Climate

November 2, 2021

LA-UR-21-31428

Los Alamos
NATIONAL LABORATORY

Image: NASA

Outline & Outlook

- ***Looking backward*** at how scientific discoveries have occurred can give us insights into how to build machines for scientific discovery.

- ***Looking forward,*** how might we build AI architectures specifically for science?

Image: NASA

Los Alamos
NATIONAL LABORATORY

Dominant Dynamical Regimes:

empirical & non-asymptotic equation reduction approximations with a long and fruitful history in geophysical fluid dynamics

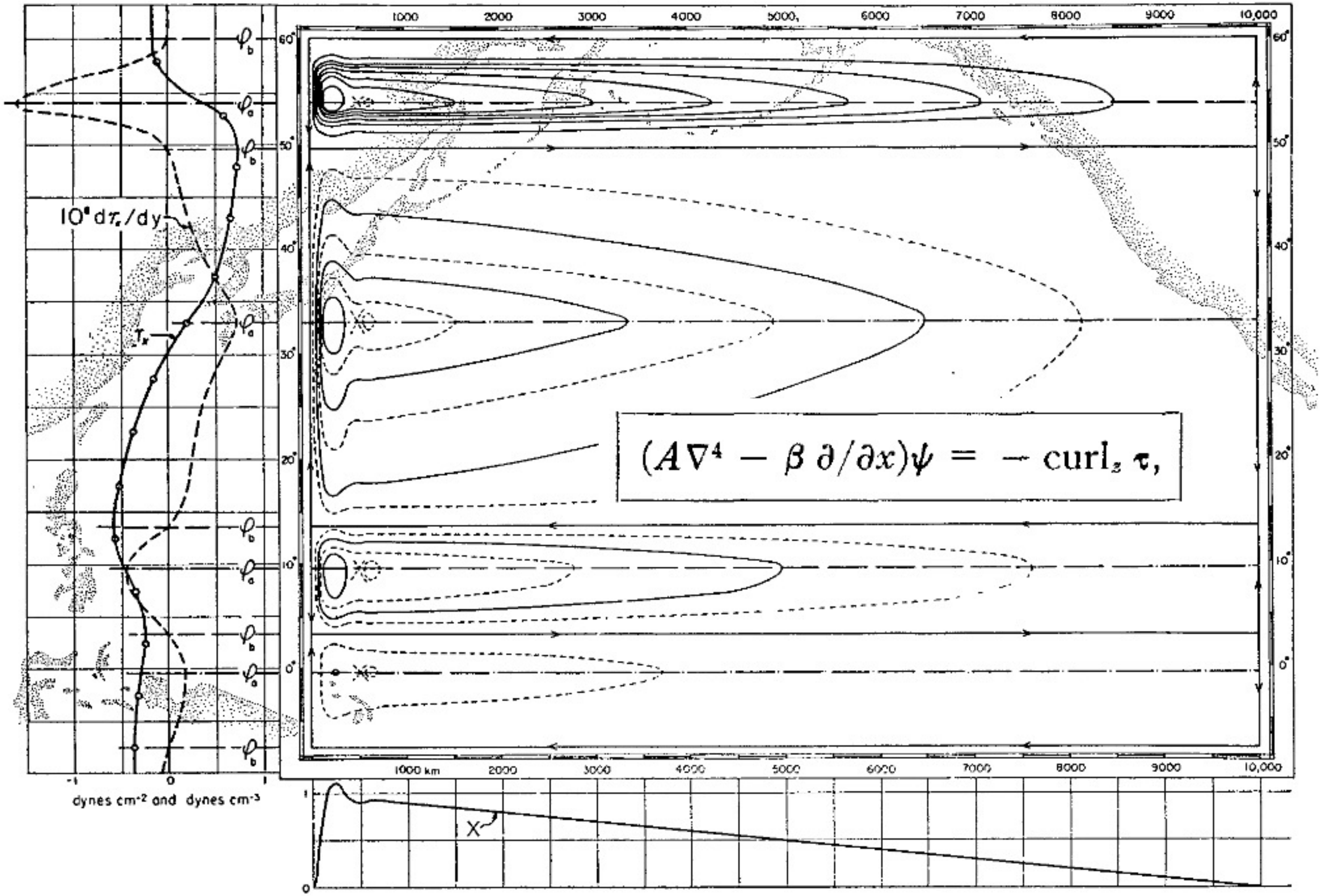
Juan Saenz (LANL), Maike Sonnewald (Princeton & U. Washington), & Daniel Livescu (LANL)

Machine Learning for Turbulence (MeLT)

Los Alamos
NATIONAL LABORATORY

Munk's ocean circulation model utilizes ad hoc scaling arguments

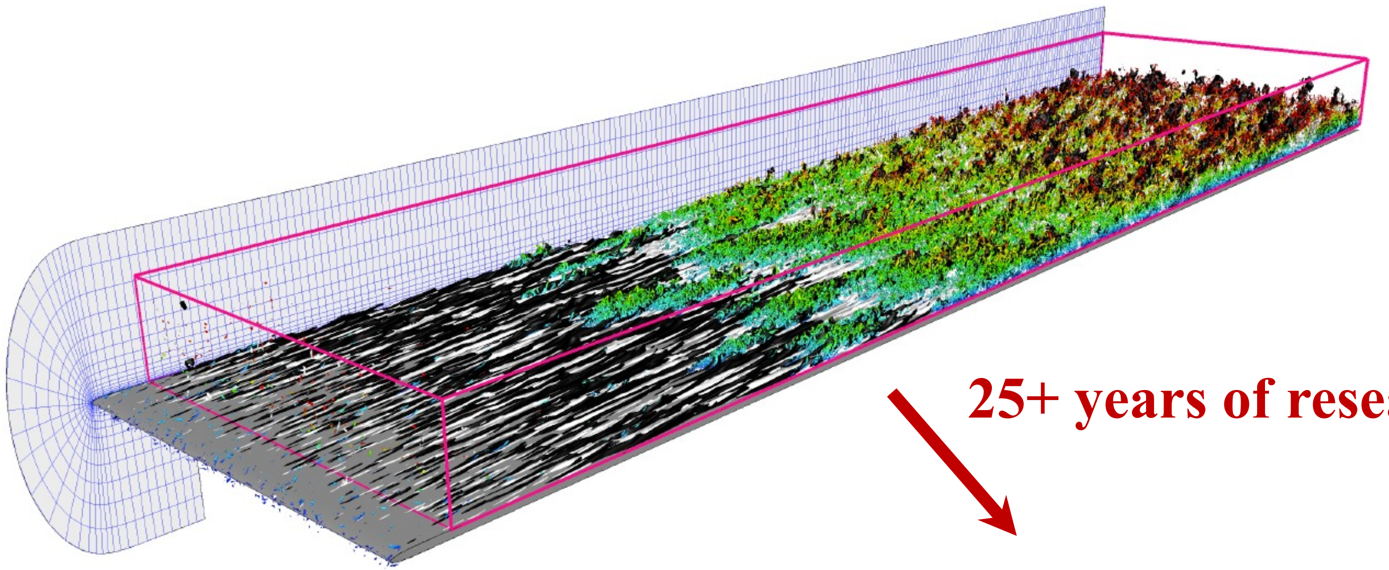
Munk's barotropic vorticity model is derived from the Navier-Stokes equations by neglecting equation terms according to geometric and ***empirical scaling arguments***



Munk, "On the wind-driven ocean circulation" (1950)

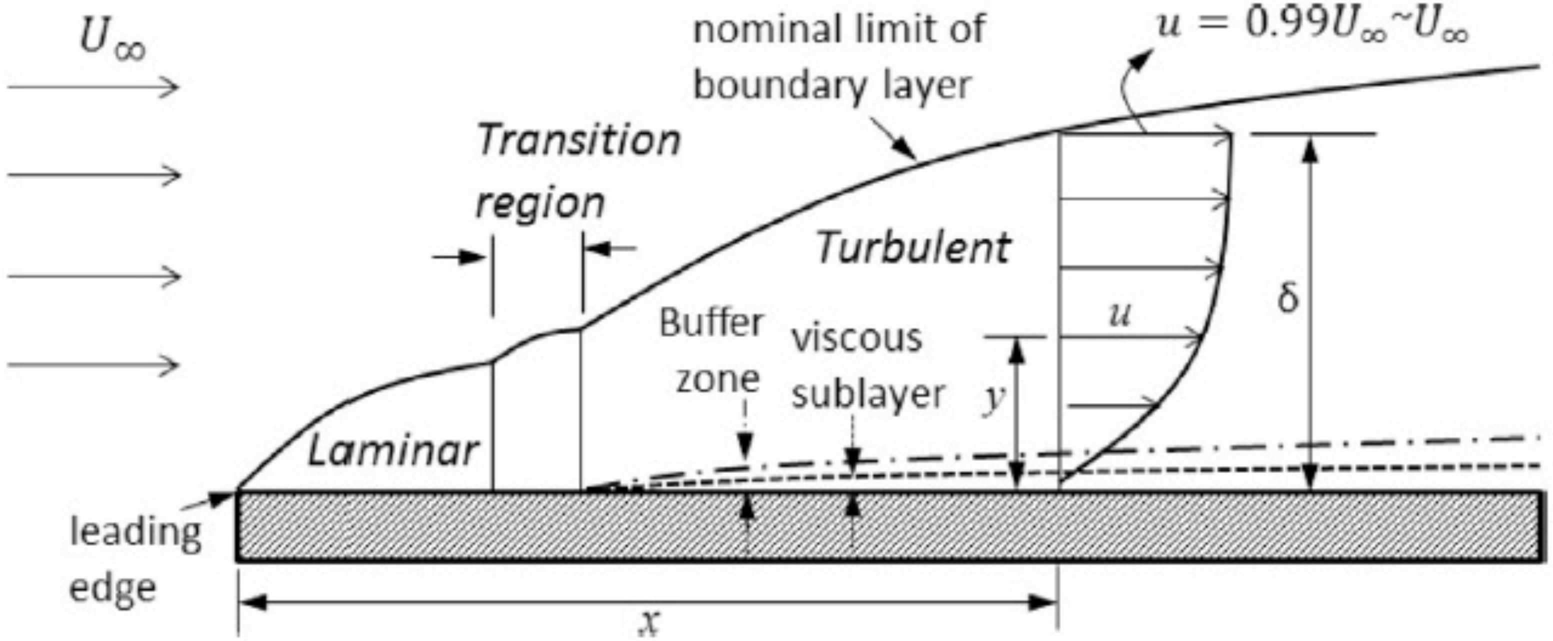
Los Alamos NATIONAL LABORATORY

There is ***no explicit universal verification for scale analysis*** for non-asymptotic dynamics



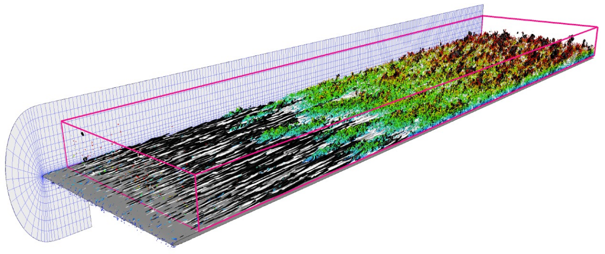
1) High quality empirical data
e.g. DNS by Zaki (2013) (*left*)

25+ years of research: Prantdl (1904) to von Karman (1930)

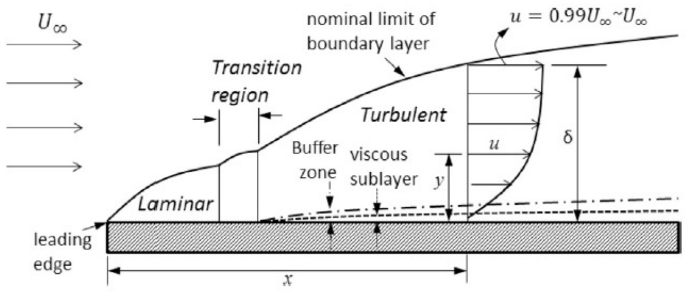


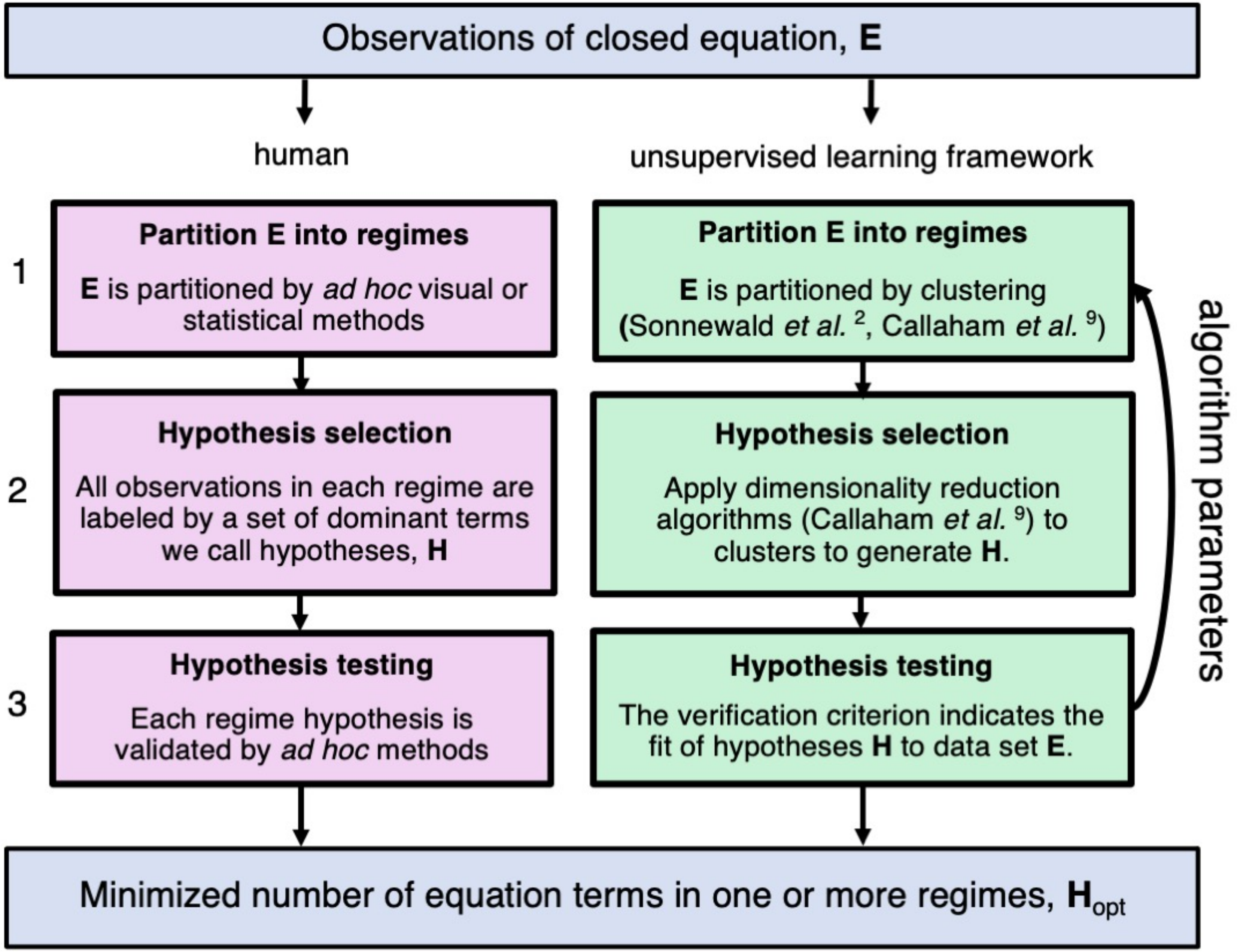
2) Domain knowledge
Image: Shahmohamadi & Rashidi (2017) (*right*)

Los Alamos
NATIONAL LABORATORY

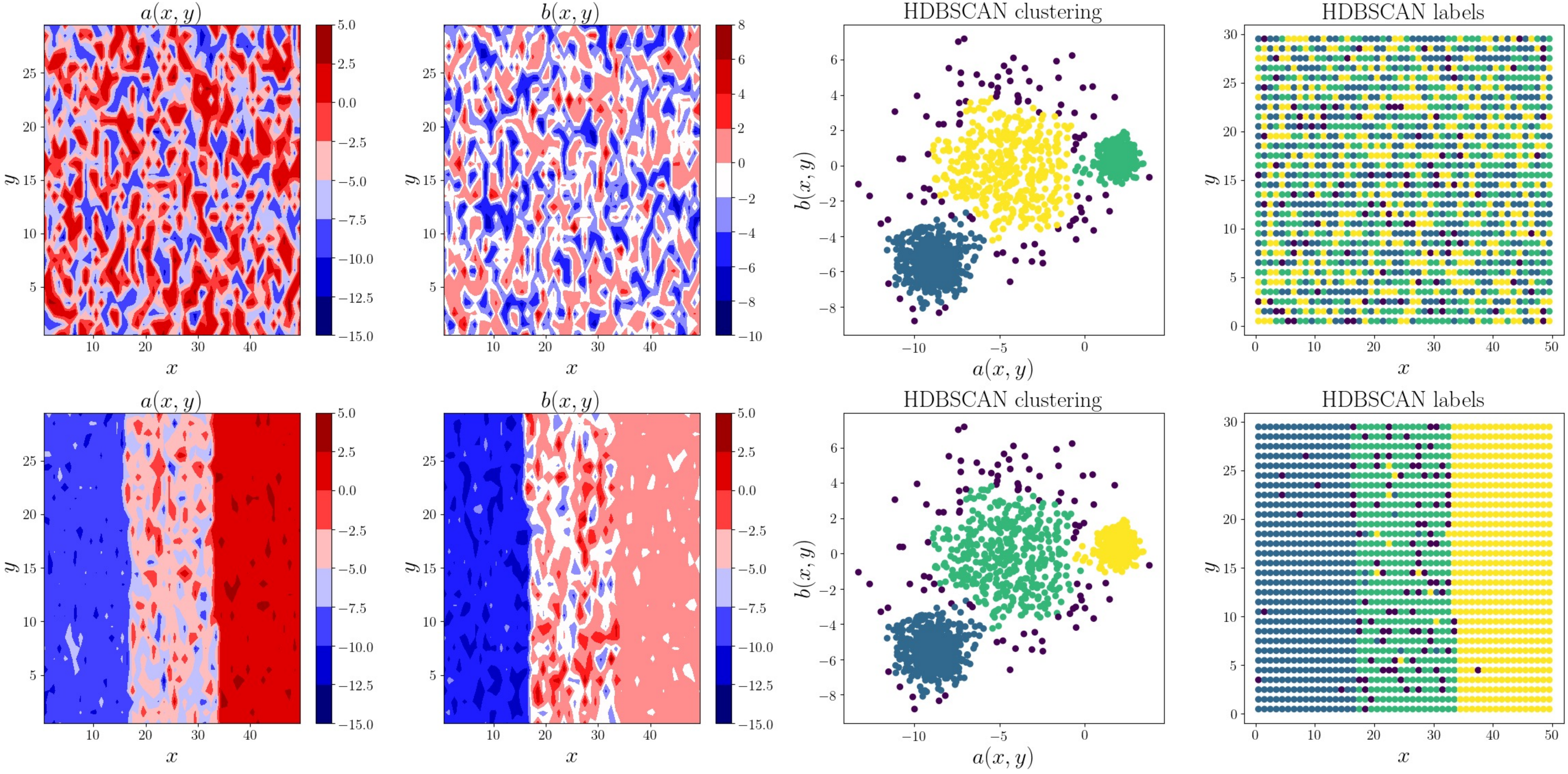


Building on the work of Sonnewald et al. (2019) and Callaham et al. (2021), we show that: ***dynamical regimes can be identified without a priori domain knowledge.***





***Partition* E = [a,b],** equation-term clusters exploit sparsity in the *equation-space*



Verification criteria for dominant dynamical regimes

Given observations of **equation terms**: $$\mathbf{E} = \left[\frac{\partial u}{\partial t}, u\frac{\partial u}{\partial x}, v\frac{\partial u}{\partial y}, w\frac{\partial u}{\partial z}, fv, \frac{1}{\rho}\frac{\partial p}{\partial x}, \nu\frac{\partial^2 u}{\partial x^2}, \nu\frac{\partial^2 u}{\partial y^2}, \nu\frac{\partial^2 u}{\partial z^2}\right],$$

Choose **hypotheses** for which terms are dominant: $$\mathbf{H} = \{0, 0, 0, 0, 1, 1, 0, 0, 0\}$$

To find the **optimal hypotheses…**

$$\mathbf{H}_{\text{opt}} = \begin{cases} \underset{\mathbf{H}}{\text{argmax}}\, \mathcal{V}(\mathbf{E}, \mathbf{H}) & \text{if} \quad \max \mathcal{V}(\mathbf{E}, \mathbf{H}) > \mathcal{V}(\mathbf{E}, \mathbf{1}) \\ \mathbf{1} & \text{if} \quad \max \mathcal{V}(\mathbf{E}, \mathbf{H}) \leq \mathcal{V}(\mathbf{E}, \mathbf{1}) \end{cases}$$

…according to the **verification criterion:**

$$\mathcal{V}(\mathbf{E}, \mathbf{H}) = \frac{\sum_{n=1}^{N} w_n \cdot \mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n)}{\sum_{n=1}^{N} w_n},$$

$$\boxed{\mathcal{M}_n(\mathbf{e}_n, \mathbf{h}_n) = \frac{\Gamma_n}{1 + \Omega_n} = \frac{\text{magnitude gap}}{\text{magnitude spread of dominant terms}}}$$

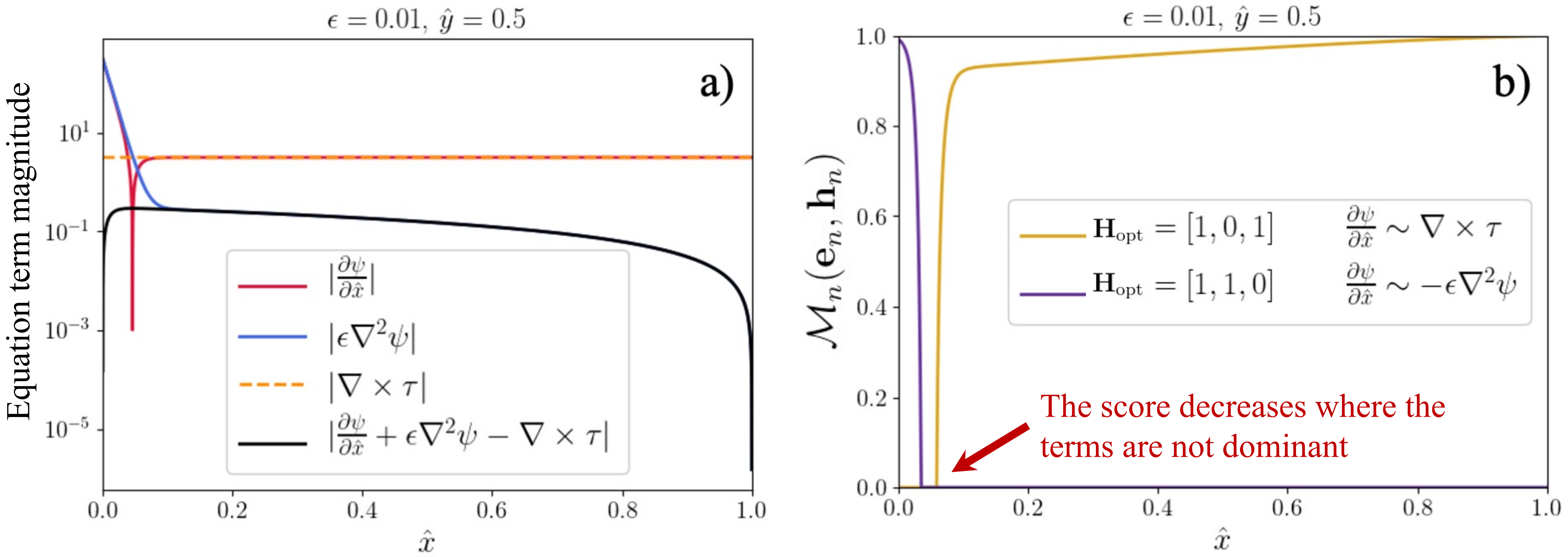
Los Alamos
NATIONAL LABORATORY

Verification criteria: 1D asymptotic example

Munk-like Vallis (2017) barotropic vorticity model:

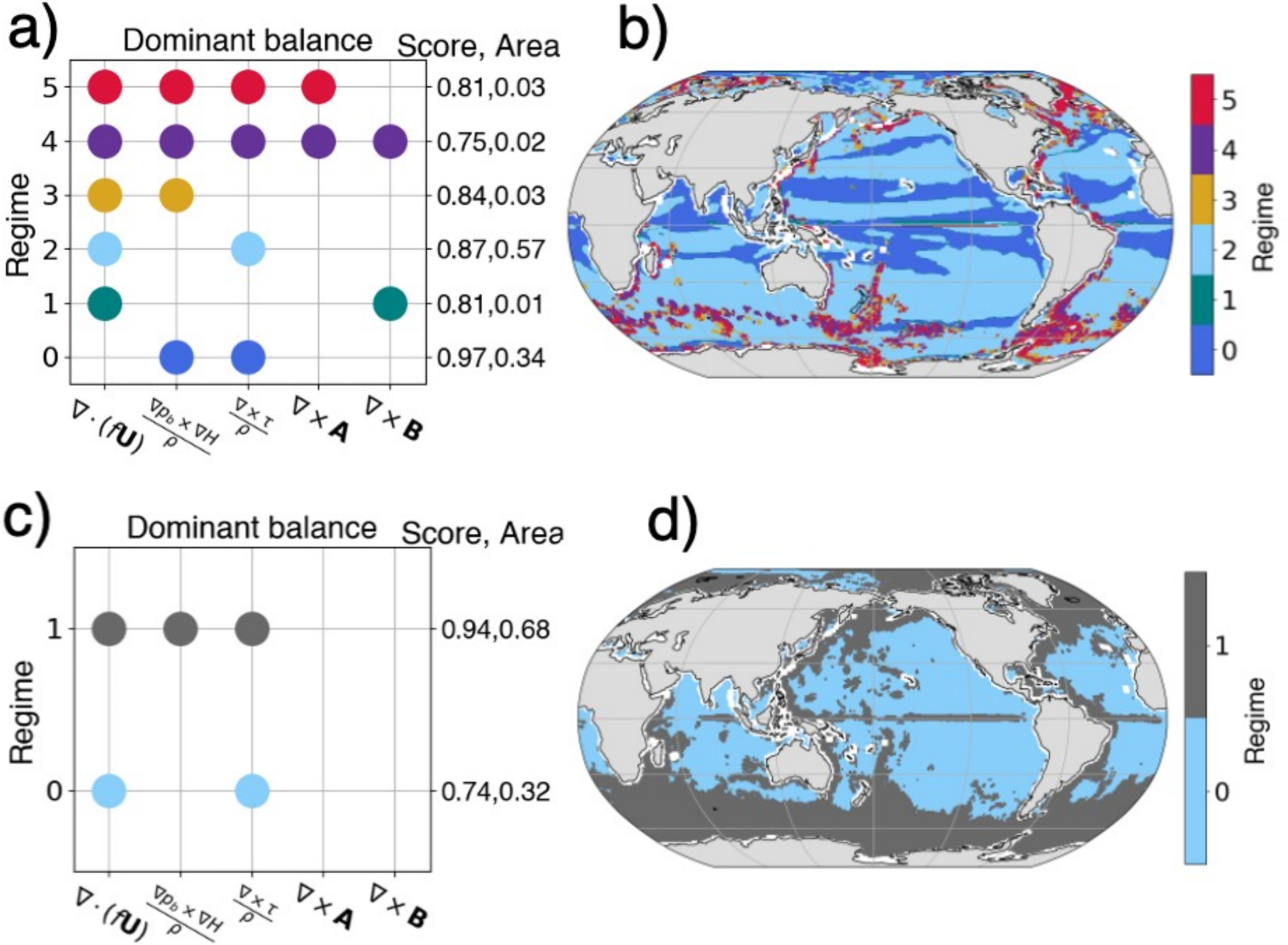
$$\overbrace{\frac{\partial \psi}{\partial \hat{x}}}^{\substack{\text{advection of}\\ \text{planetary vorticity}}} + \overbrace{\epsilon \nabla^2 \psi}^{\substack{\text{diffusive}\\ \text{torque}}} = \overbrace{\nabla \times \boldsymbol{\tau}}^{\substack{\text{wind}\\ \text{stress curl}}},$$

Solution: $\psi(\hat{x}, \hat{y}) \approx (1 - x - \mathrm{e}^{-\hat{x}/\epsilon})\pi \sin(\pi \hat{y}),$



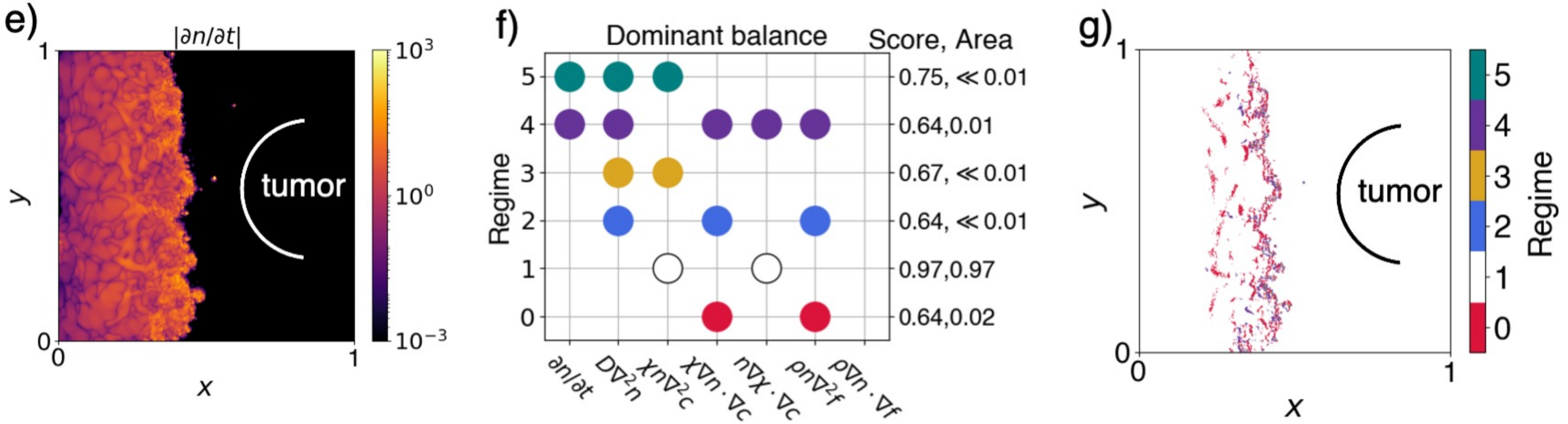
Los Alamos NATIONAL LABORATORY

Oceanic barotropic vorticity equation example



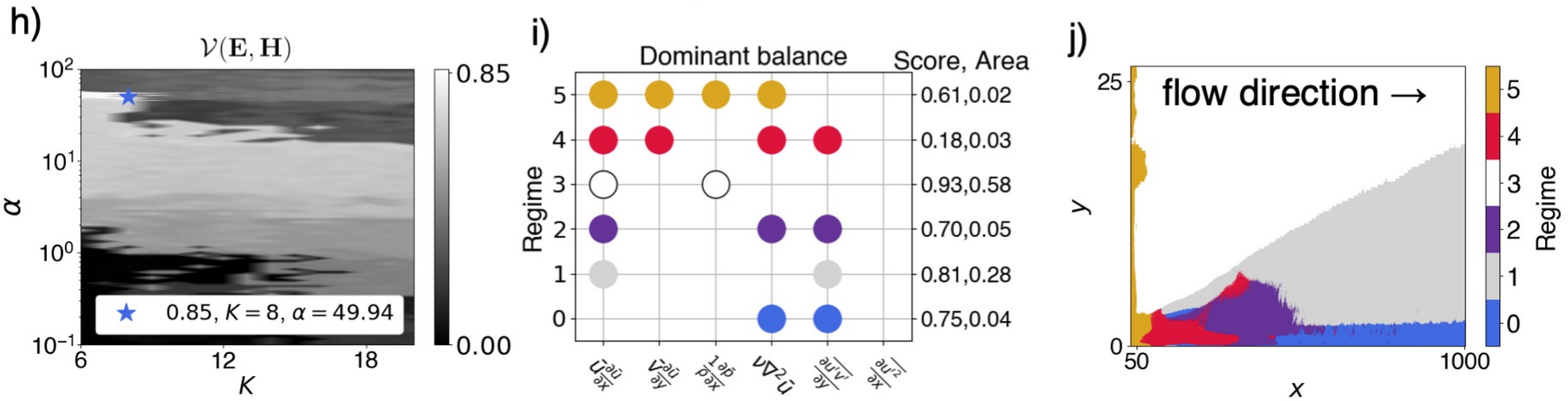
Los Alamos
NATIONAL LABORATORY

Tumor angiogenesis / endothelial cell growth equation example



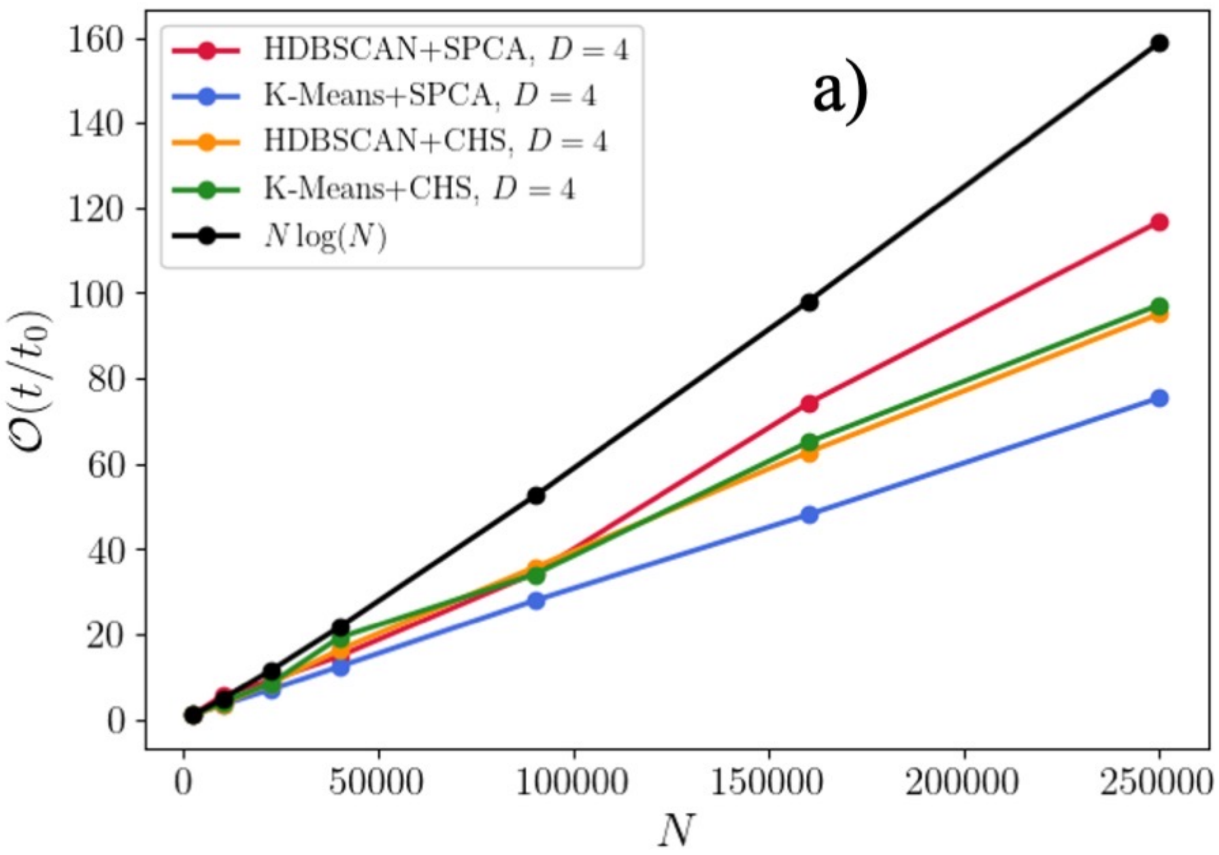
Los Alamos NATIONAL LABORATORY

Reynolds-averaged turbulent boundary layer equation example

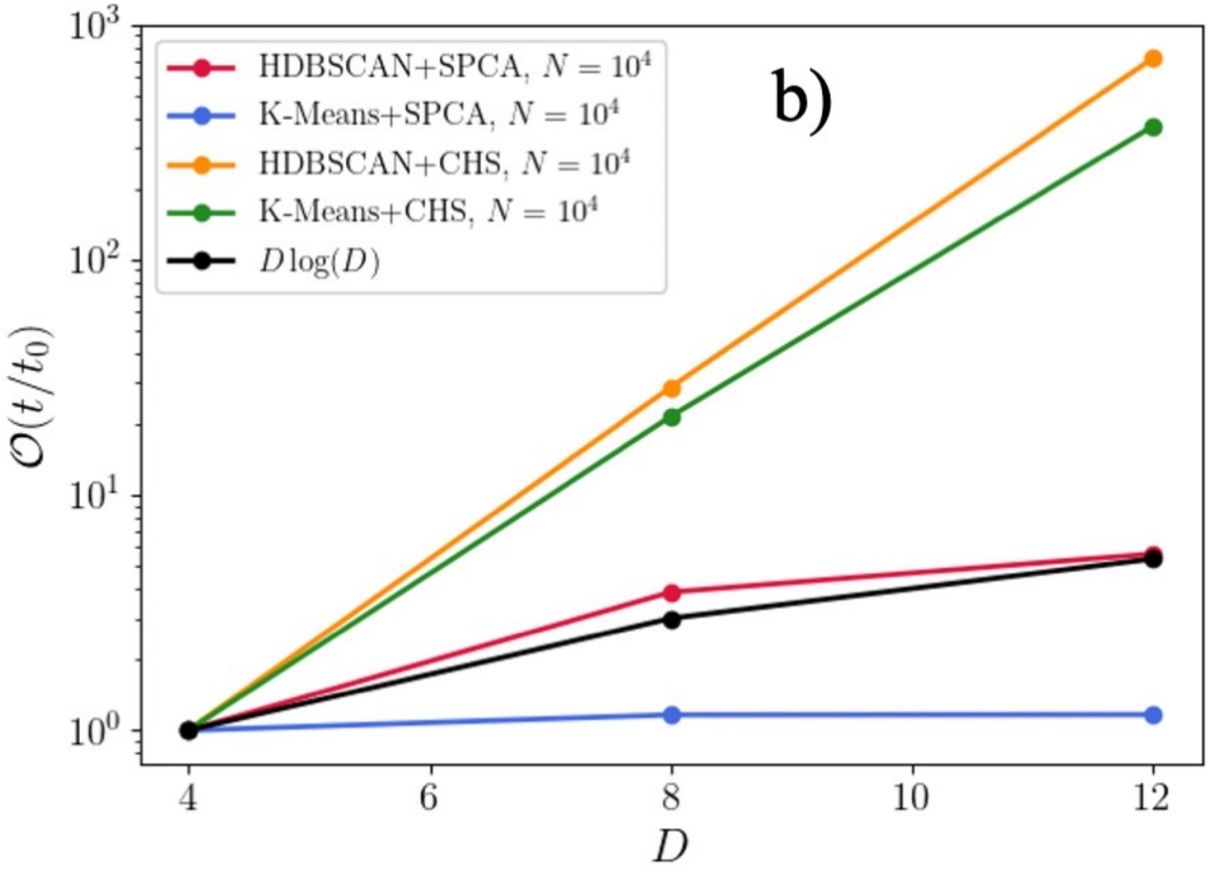


Los Alamos
NATIONAL LABORATORY

Time complexity depends upon chosen clustering & dimensionality reduction algorithms



N = number of samples in data set



D = number of equation terms

Los Alamos NATIONAL LABORATORY

Conclusions

- We have formulated the partitioning & classification of dominant dynamical regimes as an ***optimization problem.***

- We have proposed a ***verification criteria*** that:
    1. Is consistent with domain knowledge
    2. Allows regimes to be identified with no *a priori* domain knowledge

Los Alamos NATIONAL LABORATORY

Image: NASA

Machines that hypothesize

reflections on *generalizability & interpretability*
AI/ML methods for science

Juan Saenz (LANL) & Ismael Boureima (LANL)

Machine Learning for Turbulence (MeLT)

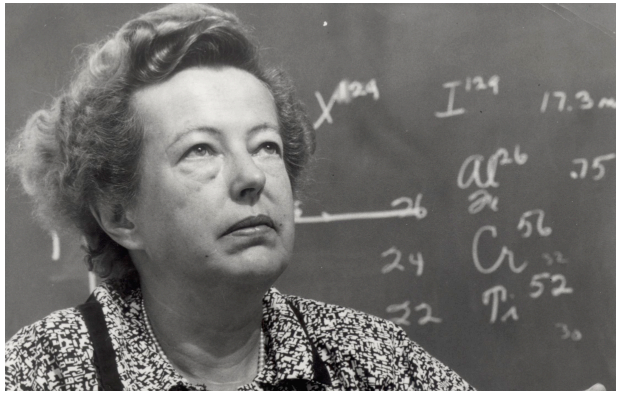
Los Alamos NATIONAL LABORATORY

What is scientific intelligence?

- “AI is the science of making machines capable of performing tasks that would require intelligence if done by humans” - Marvin Minsky
- “Intelligence measures an agent’s ability to achieve goals in a wide range of environments” - Legg & Hutter
- Intelligence is the ability to efficiently acquire skills, not mastery of a single skill. [Chollet (2020)]
- Psychologists question the concept of intelligence as a single, undifferentiated capacity. [Adams (2012)]

Scientific intelligence is the measure of a scientist's skill at generating *falsifiable* and *causal* models of Nature in the form of *symbolic* hypotheses and theories.

Maria Goeppert Mayer,
Nobel Prize in Physics (1963)

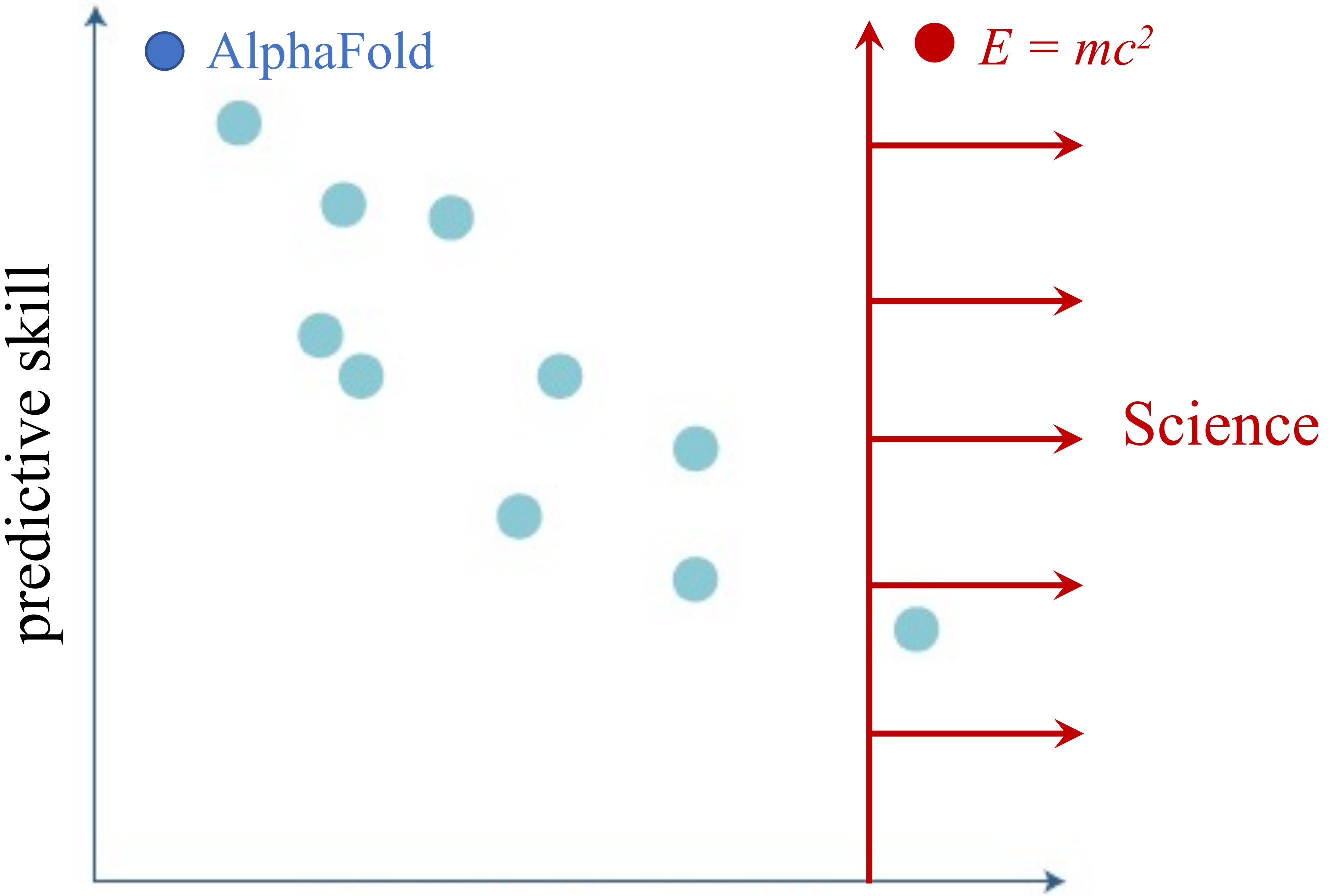


?

Machine scientists

Los Alamos
NATIONAL LABORATORY

Machine learning answers questions of statistical association
Science answers questions of causality with symbolic hypotheses



Post hoc explainability/interpretability/intelligibility of predictions

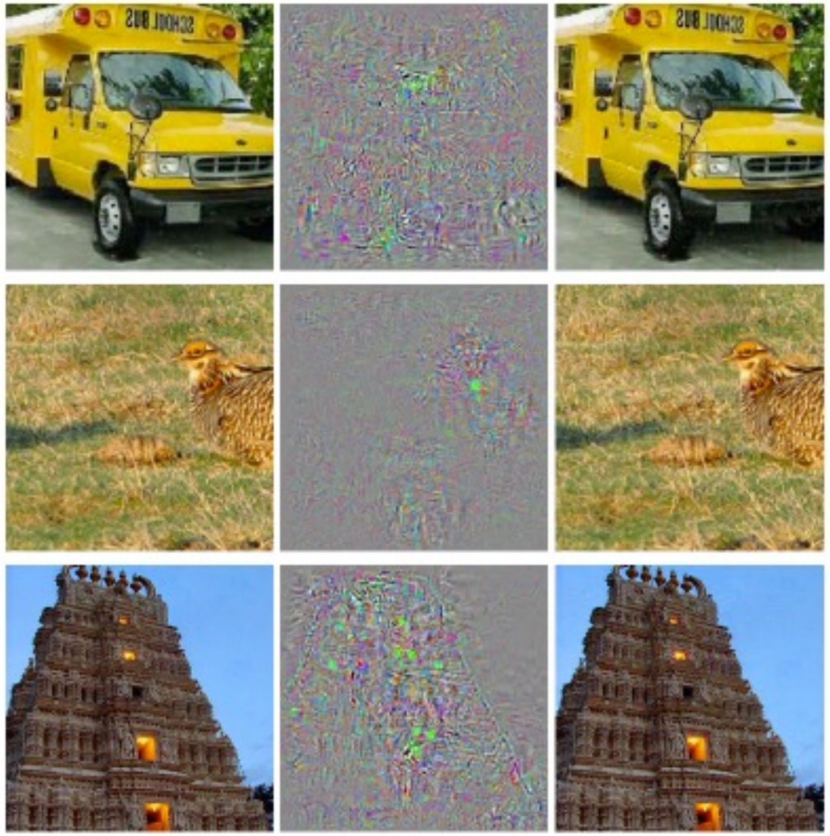
Other salient axes:

- Computational complexity
- Scope of applicability

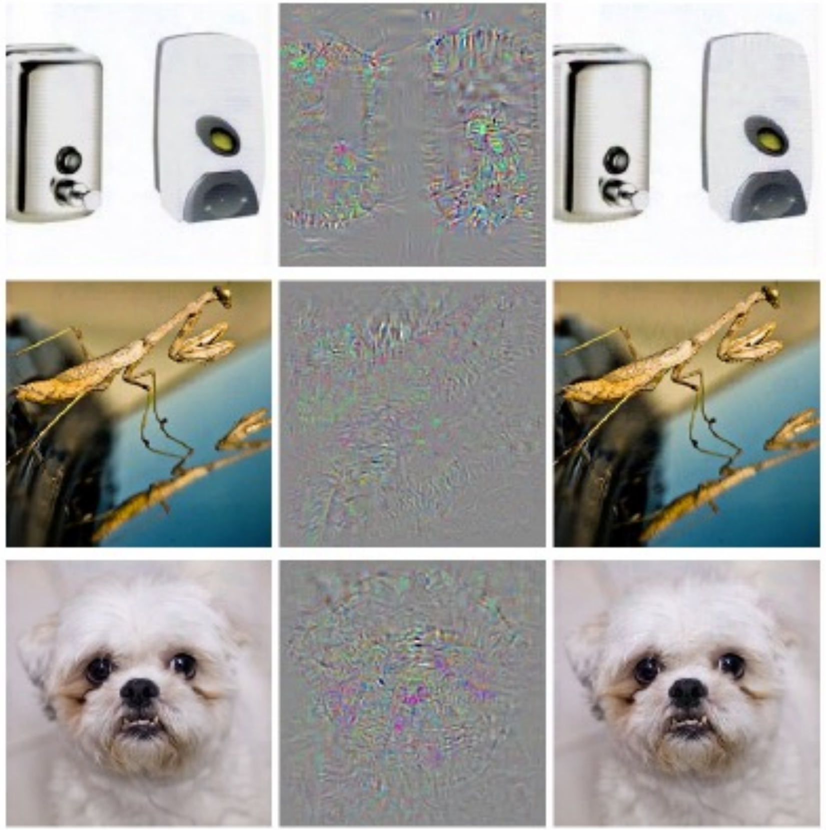
"solving the pathway of protein folding, along with the dynamics of protein processes, is a different type of challenge from predicting protein structure"
Fersht (2020)

Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," (2019)

Machine learning answers questions of statistical association



Right hand column all labeled as “ostrich”



Right hand column all labeled as “ostrich”

Concept learning is different statistical learning

Los Alamos NATIONAL LABORATORY

Szegedy et al. “Intriguing properties of neural networks” (2014)

Science answers questions of causality with symbolic hypotheses

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Judea Pearl, "The Seven Tools of Causal Inference with Reflections on Machine Learning," 2018

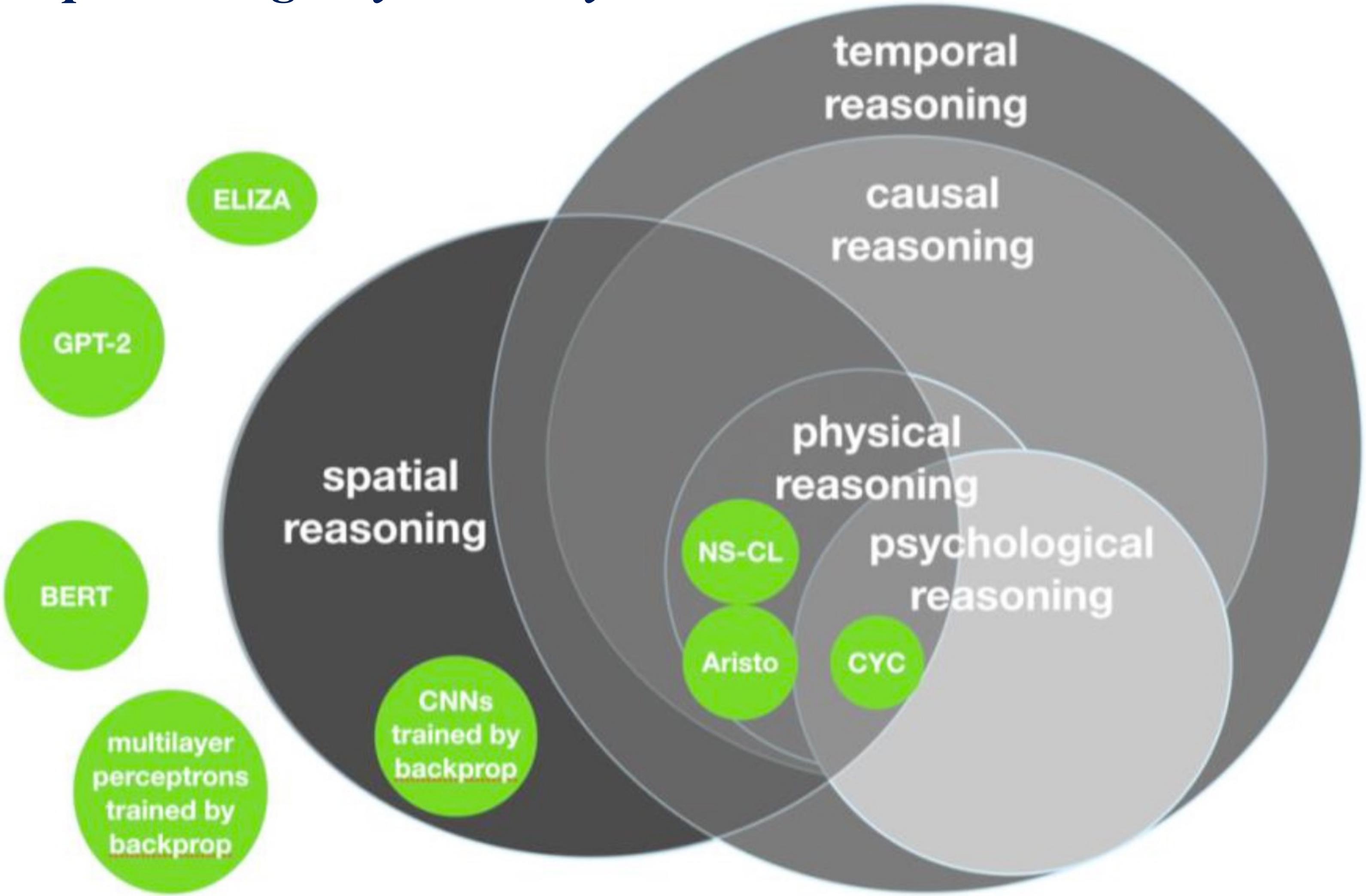
"As long as our system optimizes some property of the observed data, however noble or sophisticated, … we are back to **[the association level]** ***of the hierarchy, with all the limitations this level entails."***

We can't wait for the 50% chance of AGI in 2140, if ever …

…but we can develop hybrid deep learning – symbolic systems now

Will human-like reasoning eventually emerge from a sufficiently large neural network?

(deep learning folks, e.g. R. Sutton, say yes)



Los Alamos NATIONAL LABORATORY

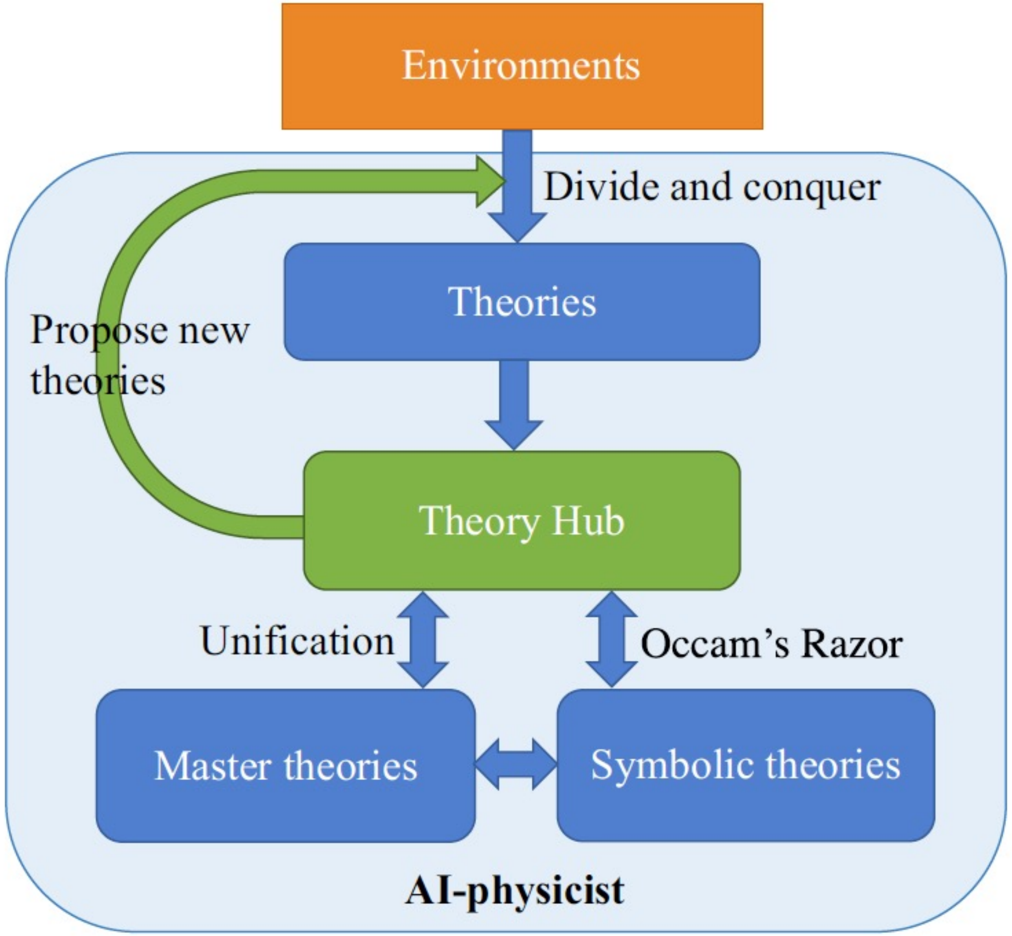
Gary Marcus, "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence," (2020)

How can we build ***scientifically intelligent machines?***

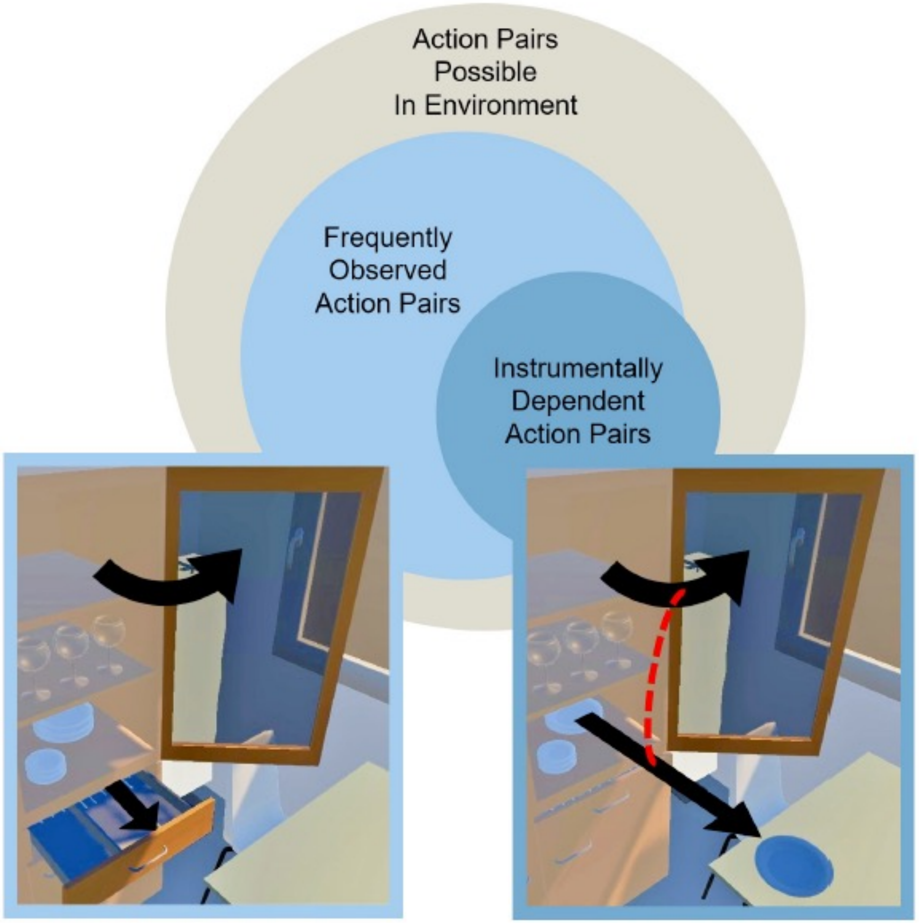
General properties of machine scientists
- Inputs: prior knowledge, data
- Outputs: symbolic hypotheses and theories
- Iterates over the scientific method
- Manipulates hypotheses with symbolic logic

Technical challenges:
- Computational complexity
- The symbol grounding problem [Harnad 1990]
- Machine reasoning [Sparkes 2010, Bottou 2014]
- Mathematics [Davis 2020]

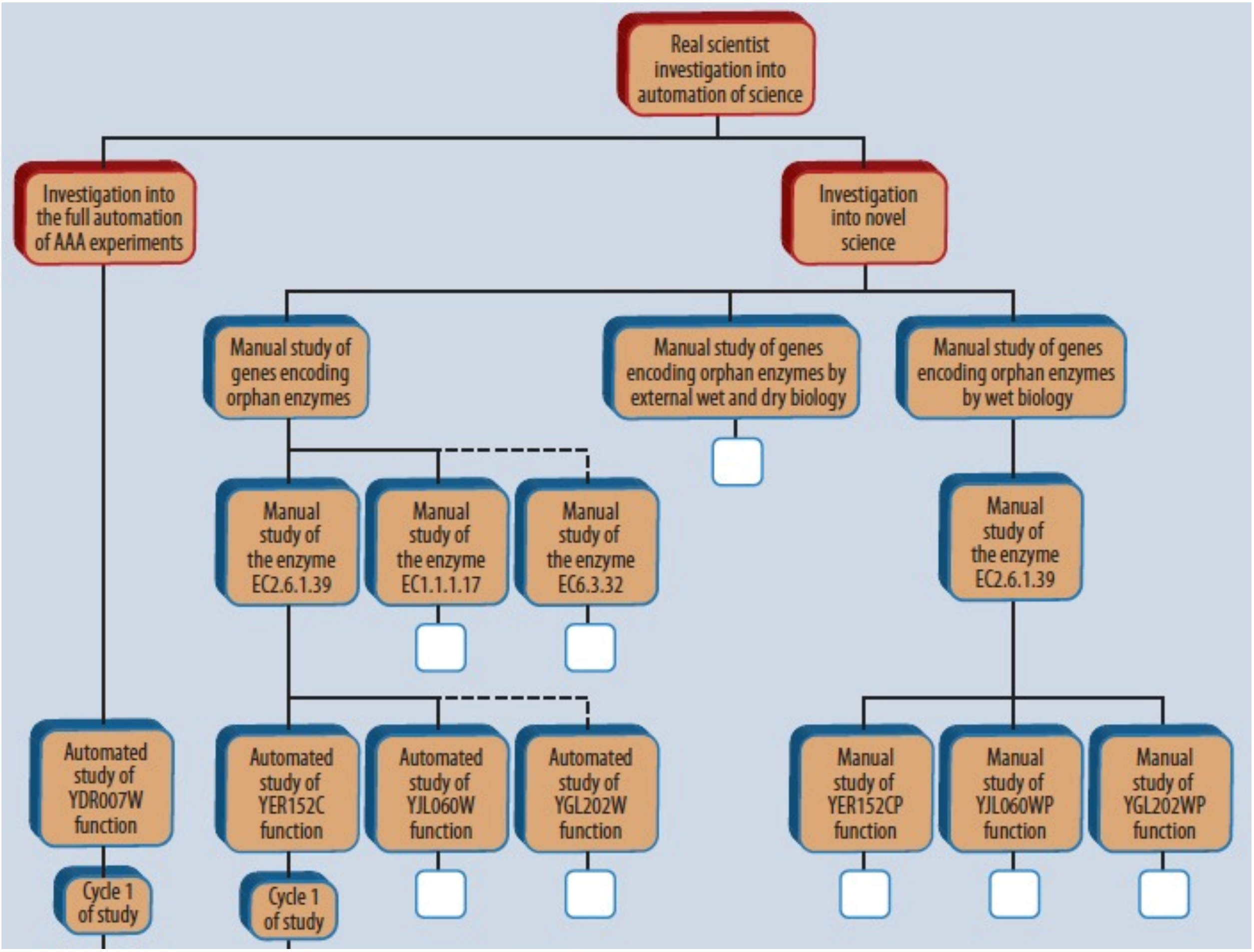
Philosophical challenges:
- Scope of applicability: breadth of inquiry
- Value: ranking the relative importance of theories



Example: Wu & Tegmark (2019) "AI physicist" uses symbolic regression and the above graph to find the most accurate and broadly applicable symbolic expressions.



Example: Uhde *et al.* (2020) "Robot as Scientist" uses virtual reality simulation, causal graphs, and experiments to reduce the search space required to predict the effects of robot motion.



Example: King *et al.* (2009) robot functional genomicist uses abductive logic, prior knowledge, and experiments to identify gene encodings that cause protein functions in yeast.

Los Alamos
NATIONAL LABORATORY

Thank you for your time and attention