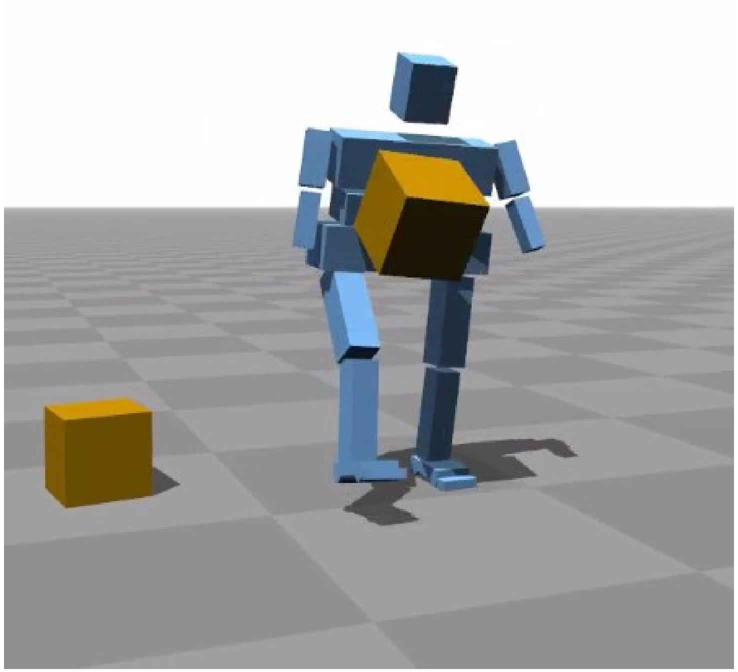
Emergence of Robustness

Nicholas Guttenberg
Earth-Life Science Institute
Tokyo Institute of Technology

Robust Systems



Thomas Geijtenbeek, Michiel van de Panne, A. Frank van der Stappen,
“Flexible Muscle-Based Locomotion for Bipedal Creatures”

Roughly speaking – when you push it, it springs back

Robust Theories

What is “pushing” a theory?

Make some error

Do you still get the same kind of result?

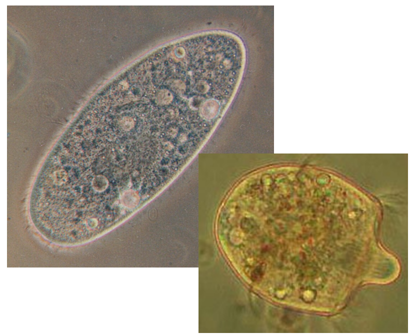
(What does 'the same' mean?)

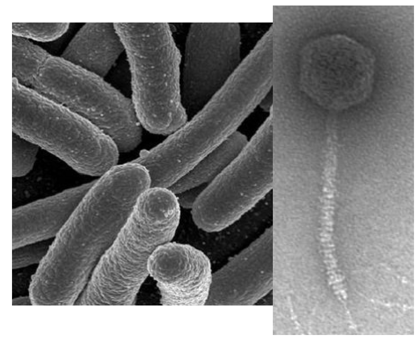
The Unreasonable Effectiveness of ...

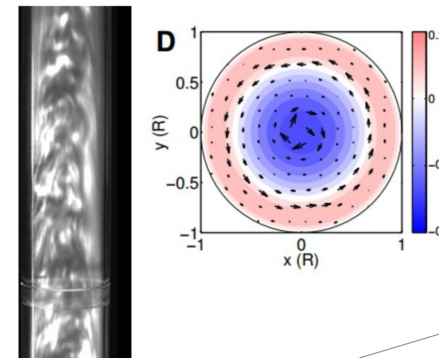
- Unreasonable Effectiveness of Mathematics (Wigner)
- Unreasonable Effectiveness of Science (Robert Haussman)
- Unreasonable Effectiveness of Deep Learning (Yann LeCun)
- The Unreasonable Effectiveness of Recurrent Neural Networks (Andrej Karpathy)

...

Example:
Predatory-prey population dynamics





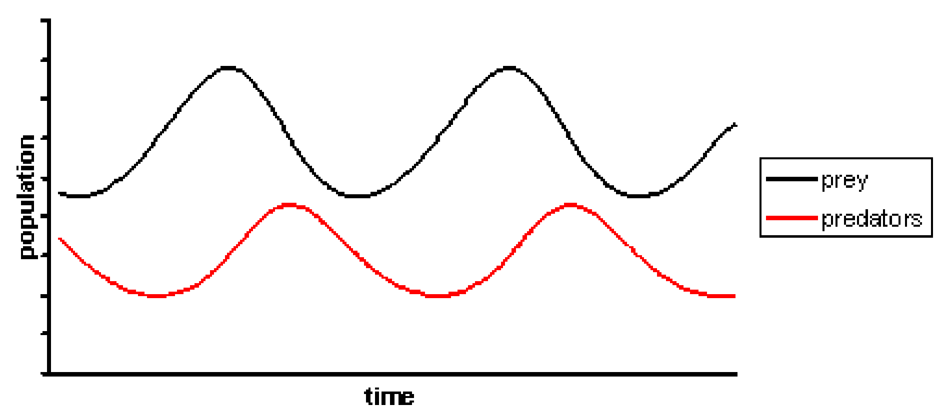


...

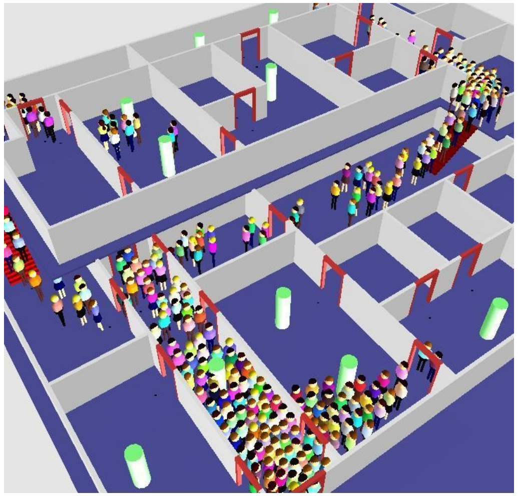
Lotka-Volterra Model

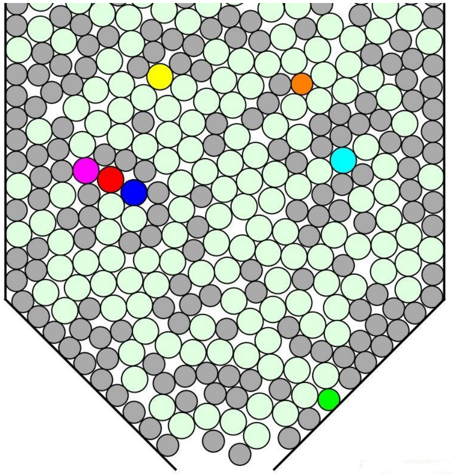
$$\frac{dx}{dt} = \alpha x - \beta xy$$

$$\frac{dy}{dt} = \delta xy - \gamma y$$



Example:
Crowd Models





Ability to Generalize

- Simple models which seem to be able generalize across details of the component 'objects'

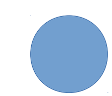
- Influence of details summarized in a few parameters

- Only works when you have large numbers of the components – the individuals are still different

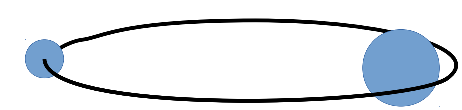
Only the collective behavior is robust.

Infinity is Easier than 3

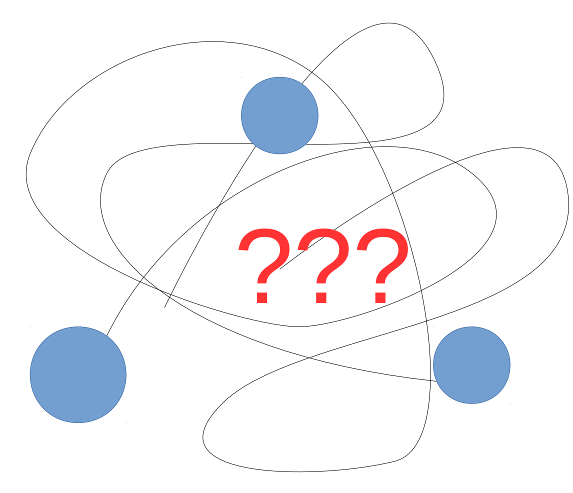
Famous problem in classical mechanics: predict the motion of N bodies



- 1-body: Trivial (just sits there)



- 2-body: Undergrad homework problem (periodic elliptical orbits)



- 3-body: Major problem in physics over the last 350 years. No general algebraic solution is possible.

Infinity is Easier than 3

How about the infinite-body problem?

Statistical mechanics: predict ~ 10^{23} bodies

(but not the details of microscopic motion)

The emergent properties are simpler than the component properties!

Emergence as independence

Emergence: at some scale, the 'appropriate' degrees of freedom change

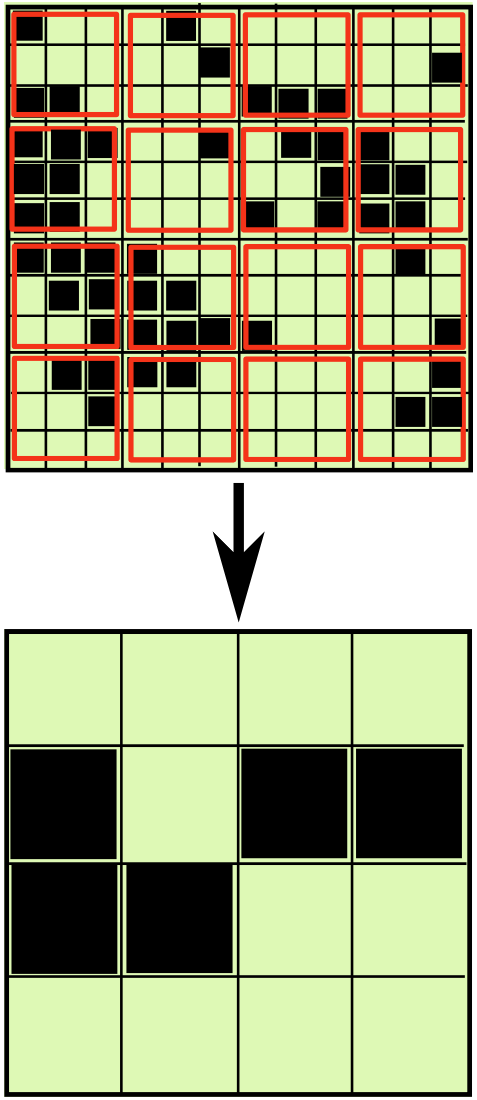
Molecules → Theory of Fluids is hard

Fluid-level description → Theory of Fluids is (relatively) easier

→ Independence from microscopic detail

Coarse Graining

Idea: Look at something on multiple scales, look for the dynamics which are natural to that scale

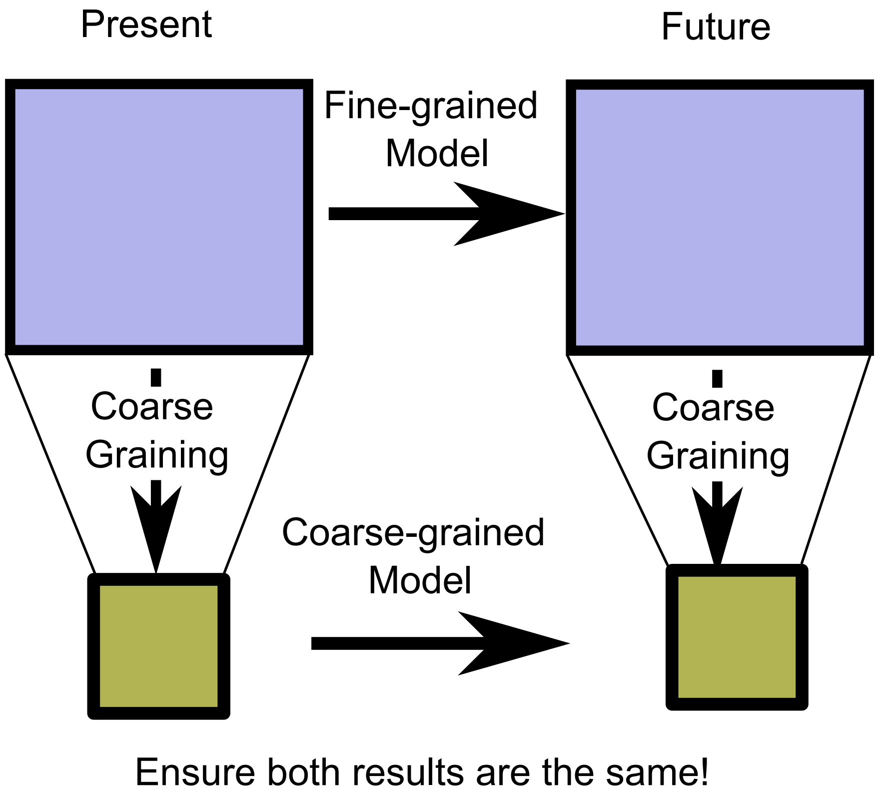


Invented by Leo Kadanoff (block spins) for magnets in 1966, extended by Kenneth Wilson, then spread to particle physics, ...

Coarse Graining

Start with a fine-grained model that can predict the fine-grained future state.

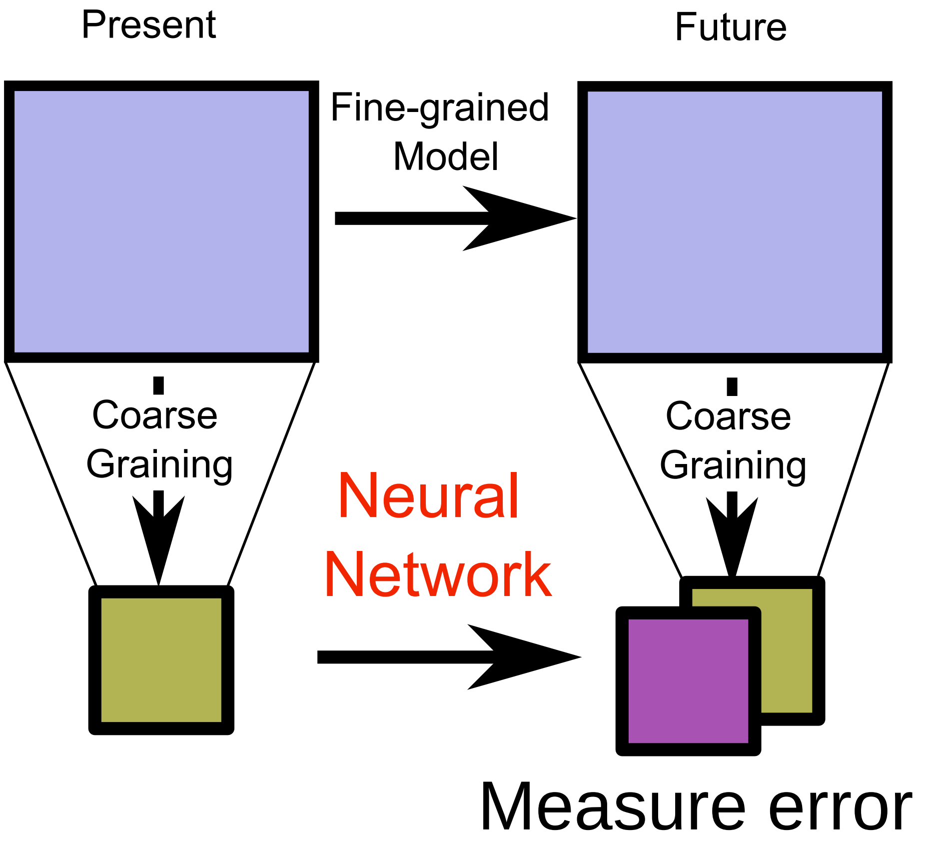
Then, derive an exact predictive model entirely in terms of the coarse-grained scale **(hard!)**



Applications: Detecting scale transitions

Instead of deriving exact coarse-grained theory, use general purpose machine learning framework to build a predictor.

Error is a function of scale – detect natural scale separations.



Toy example

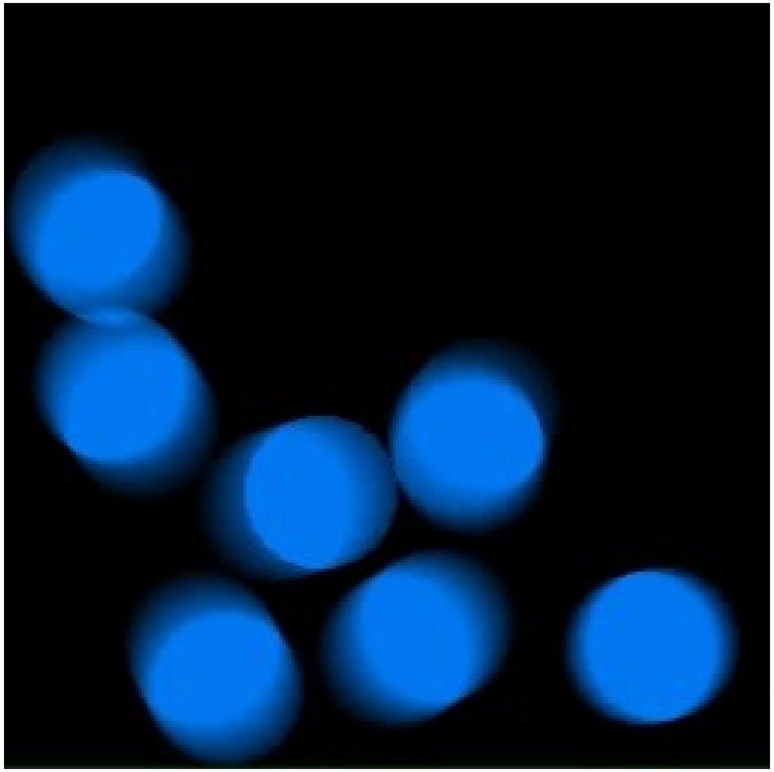
- Simulations with two very different types of particles

Hard Spheres

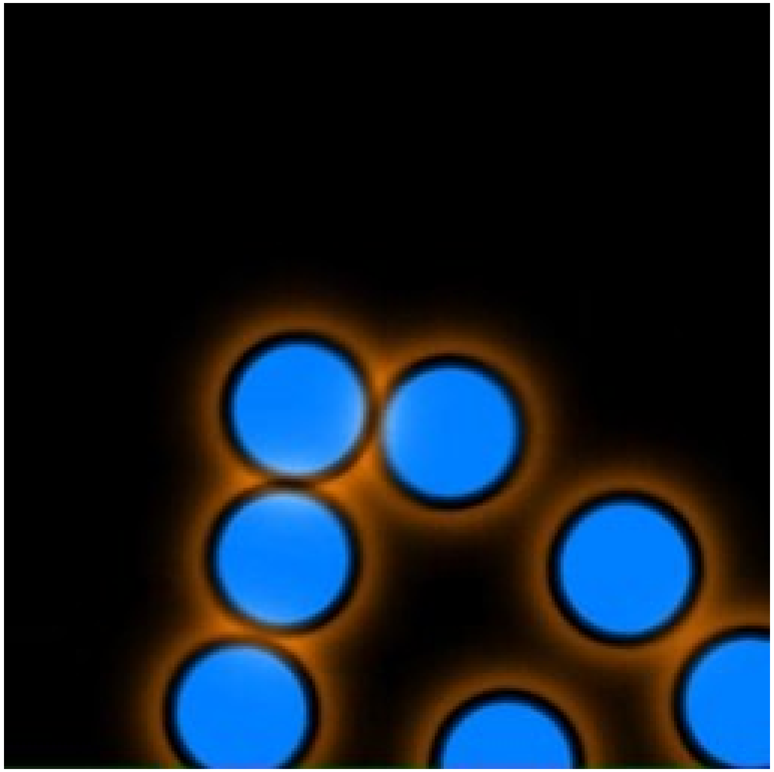
Lennard-Jones

Toy example

Hard Spheres

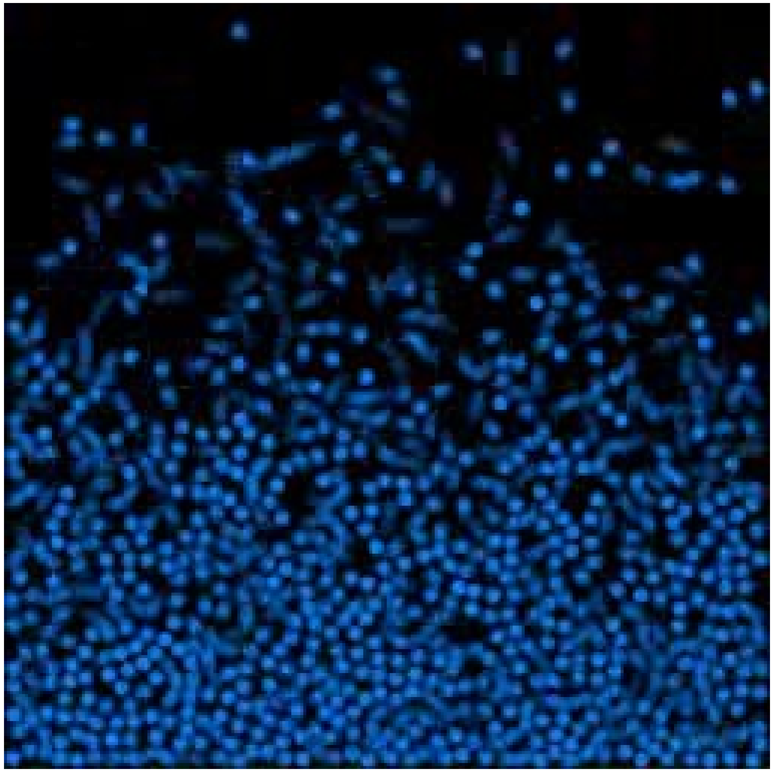


Lennard-Jones

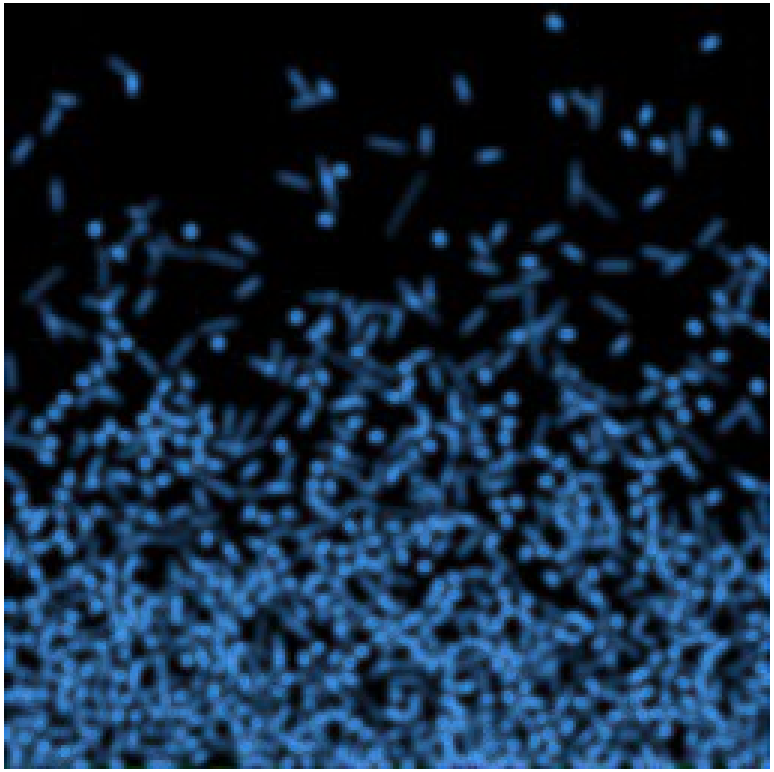


Toy example

Hard Spheres

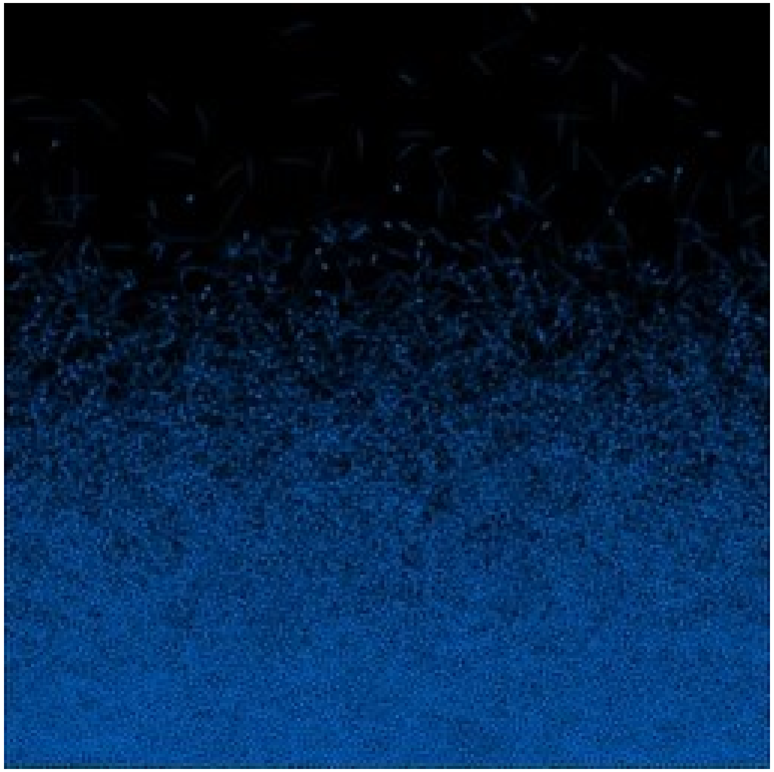


Lennard-Jones

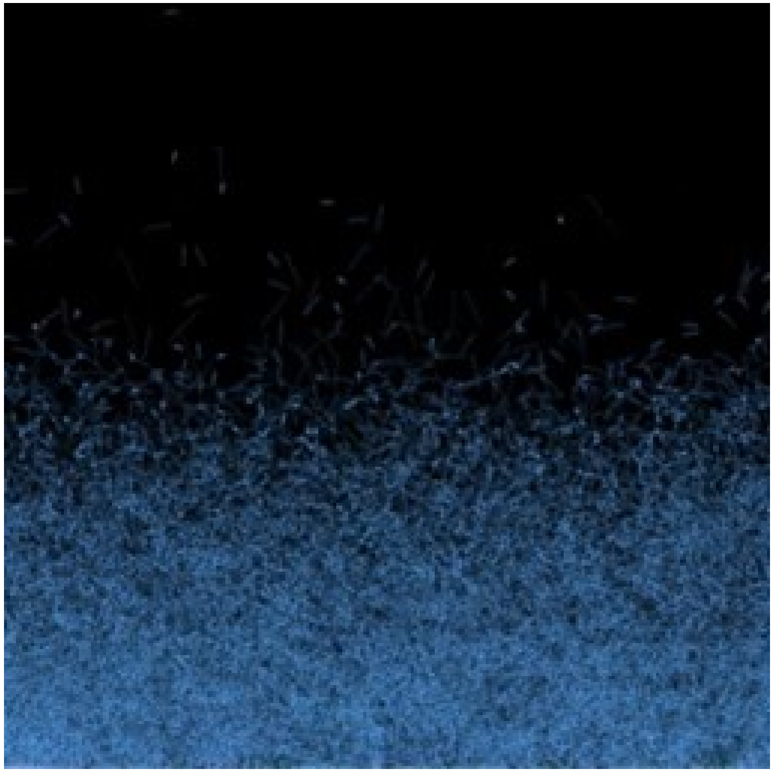


Toy example

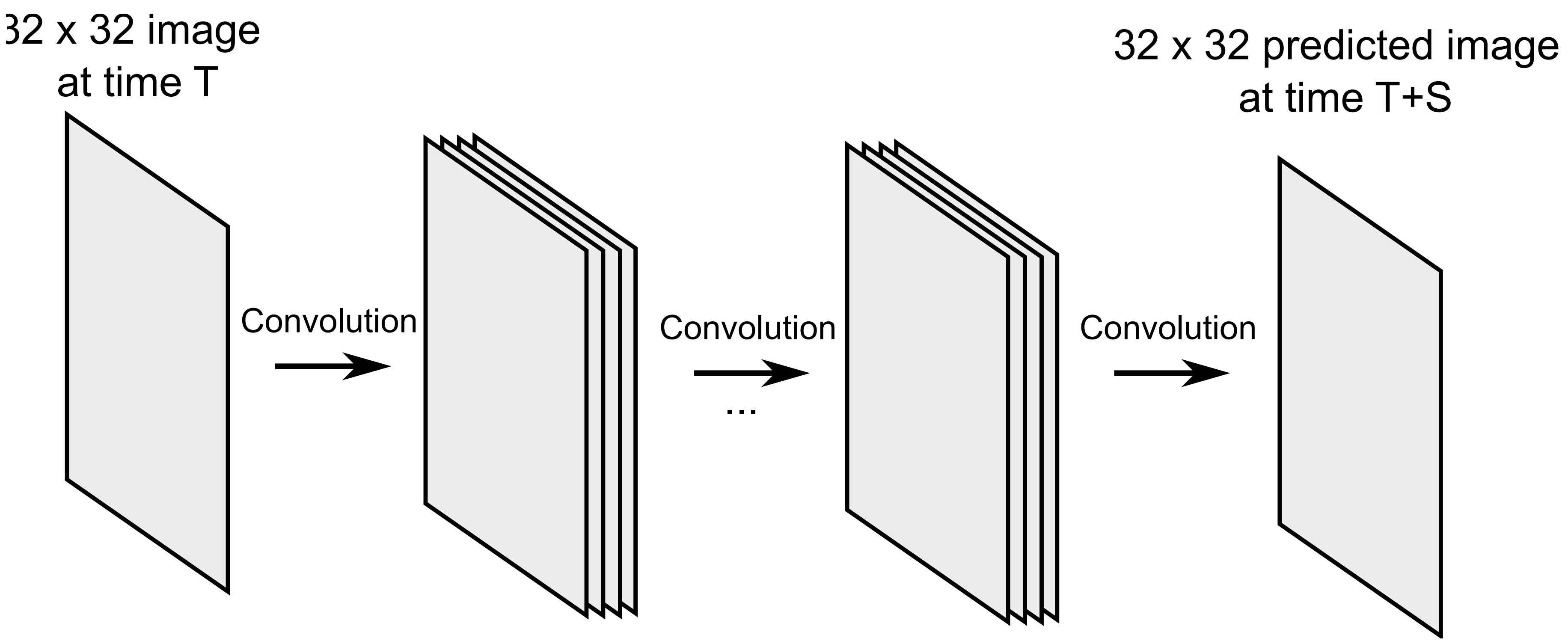
Hard Spheres



Lennard-Jones



Applications: Detecting scale transitions

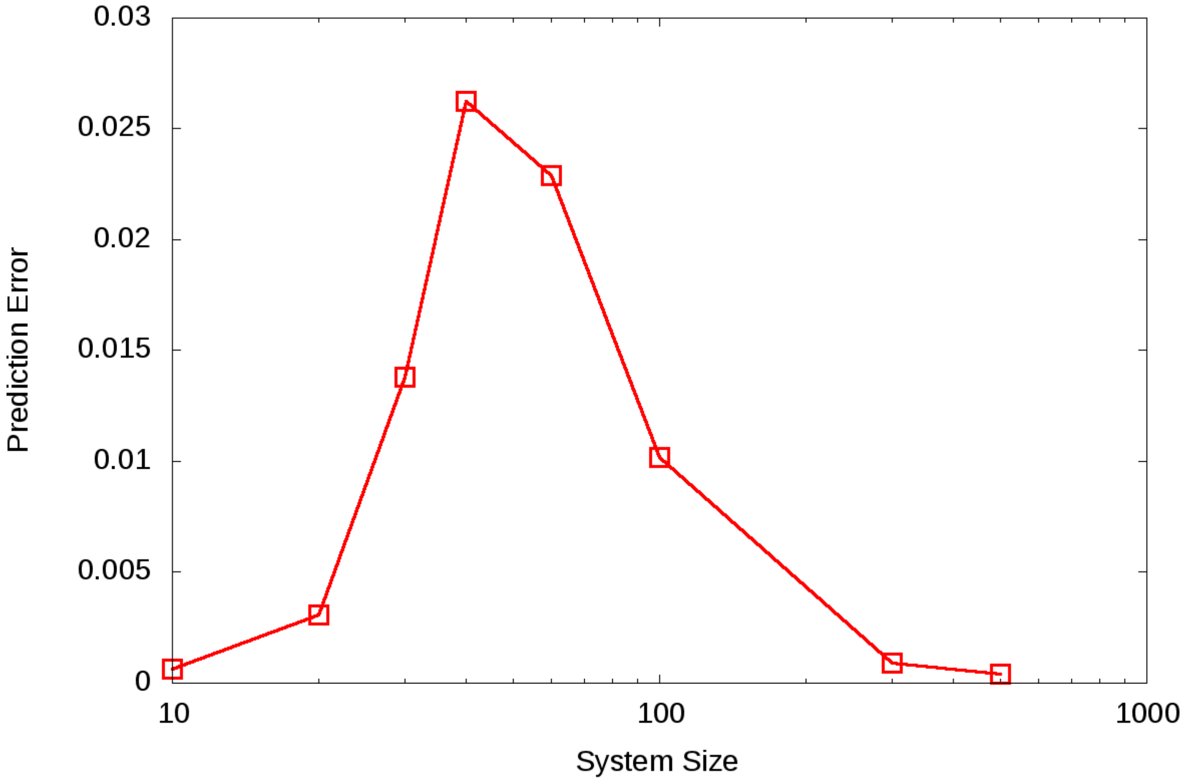


Time offset S is chosen such that an object with constant velocity moves a fixed number of pixels at each scale

(Schematic – actual architecture has more layers, noise layers to prevent overfitting, etc)

Applications: Detecting scale transitions

Prediction Error on Lennard-Jones video

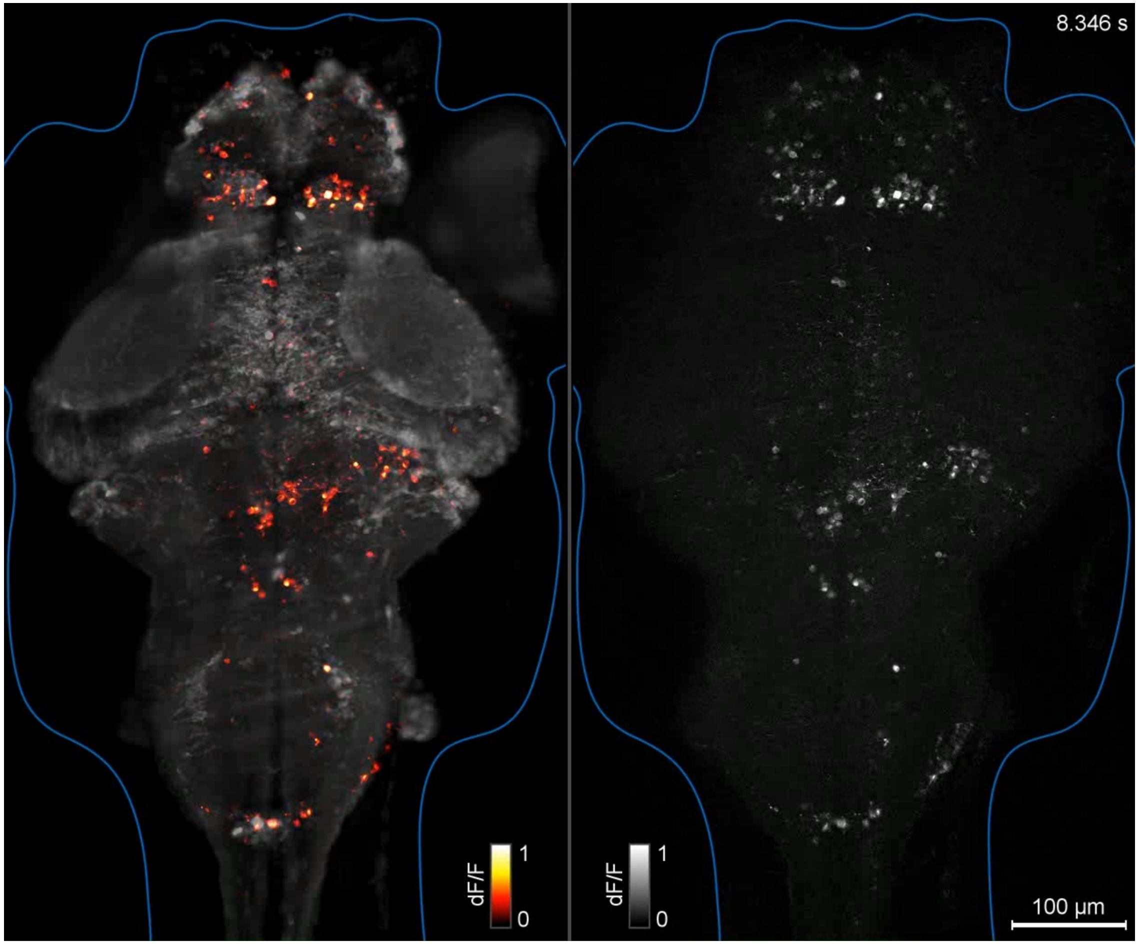


Peak corresponds to roughly 1 particle per pixel

Applications: Detecting scale transitions

Prediction Error on Zebrafish brain (with Dror, Ray)

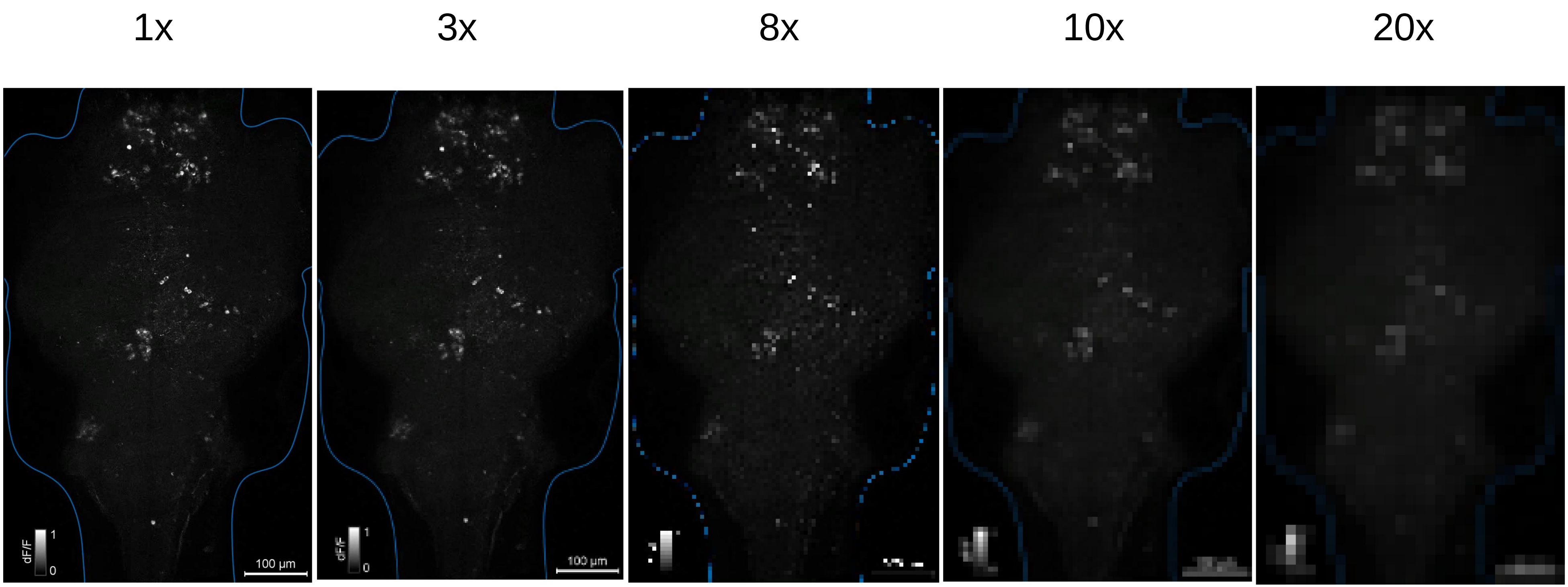
Data from Ahrens, M.; Orger, M.; Robson, D; Li, J. M.; Keller, P.; "Whole-brain functional imaging at cellular resolution using light-sheet microscopy", Nature Methods **10** (2013)



Applications: Detecting scale transitions

Prediction Error on Zebrafish brain (with Dror, Ray)

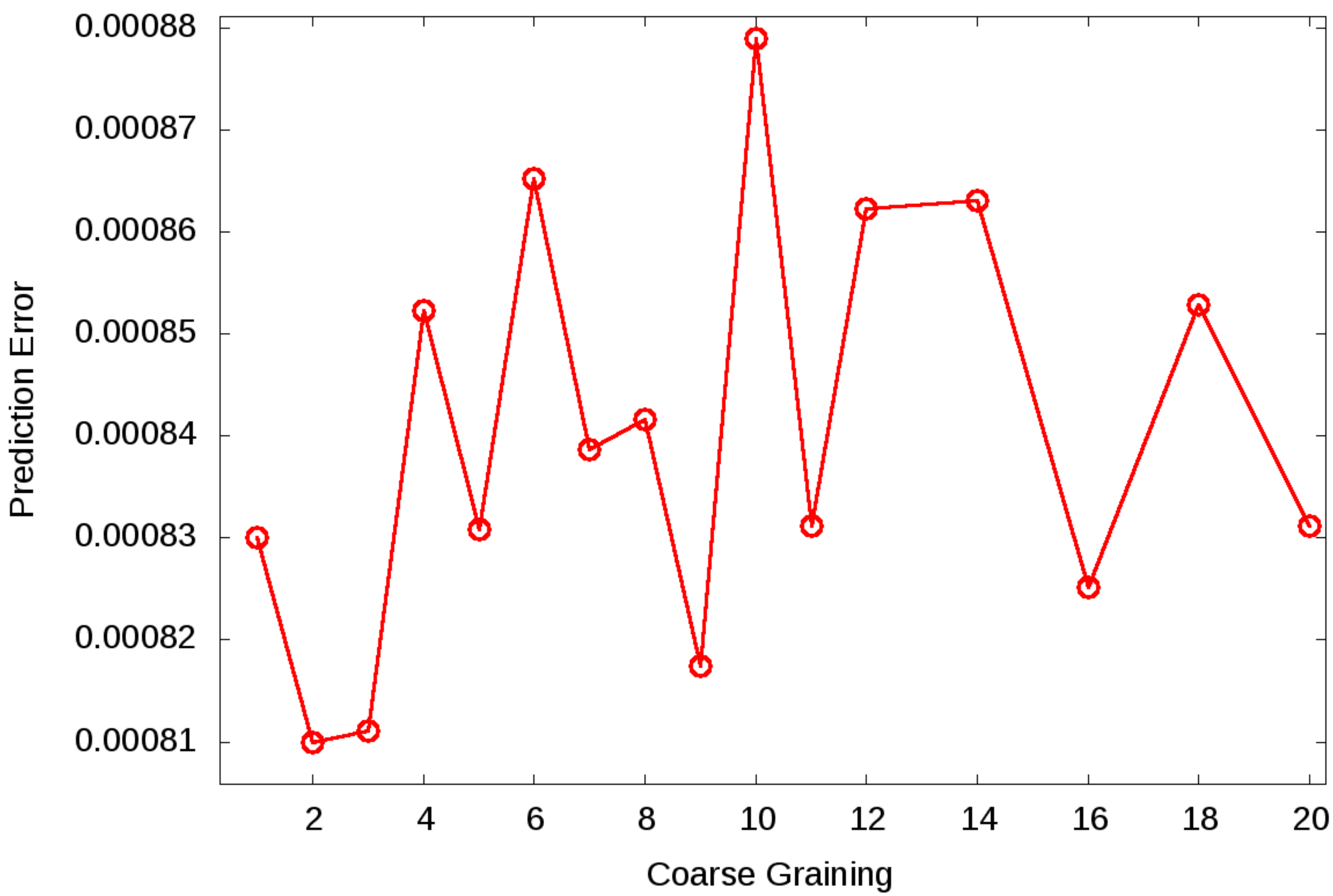
Data from Ahrens, M.; Orger, M.; Robson, D; Li, J. M.; Keller, P.; “Whole-brain functional imaging at cellular resolution using light-sheet microscopy”, Nature Methods **10** (2013)



Applications: Detecting scale transitions

Prediction Error on Zebrafish brain (with Dror, Ray)

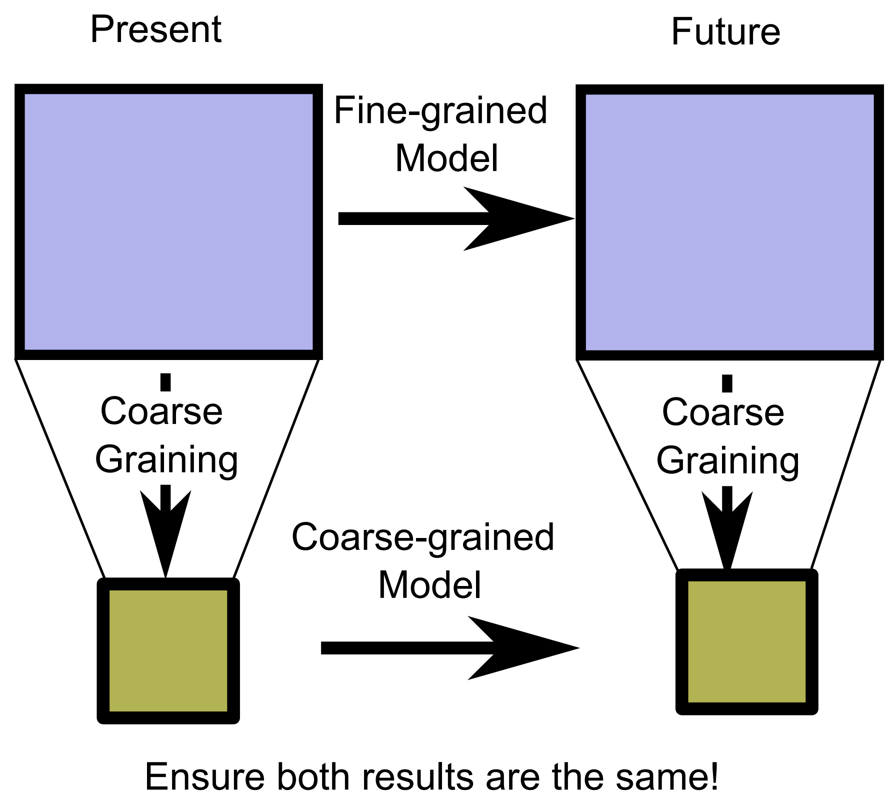
Data from Ahrens, M.; Orger, M.; Robson, D; Li, J. M.; Keller, P.; "Whole-brain functional imaging at cellular resolution using light-sheet microscopy", Nature Methods **10** (2013)



Coarse Graining

Observation from physics: when you do the **exact** coarse-graining, it often has the same mathematical form as the fine-grained theory.

But with different parameter values



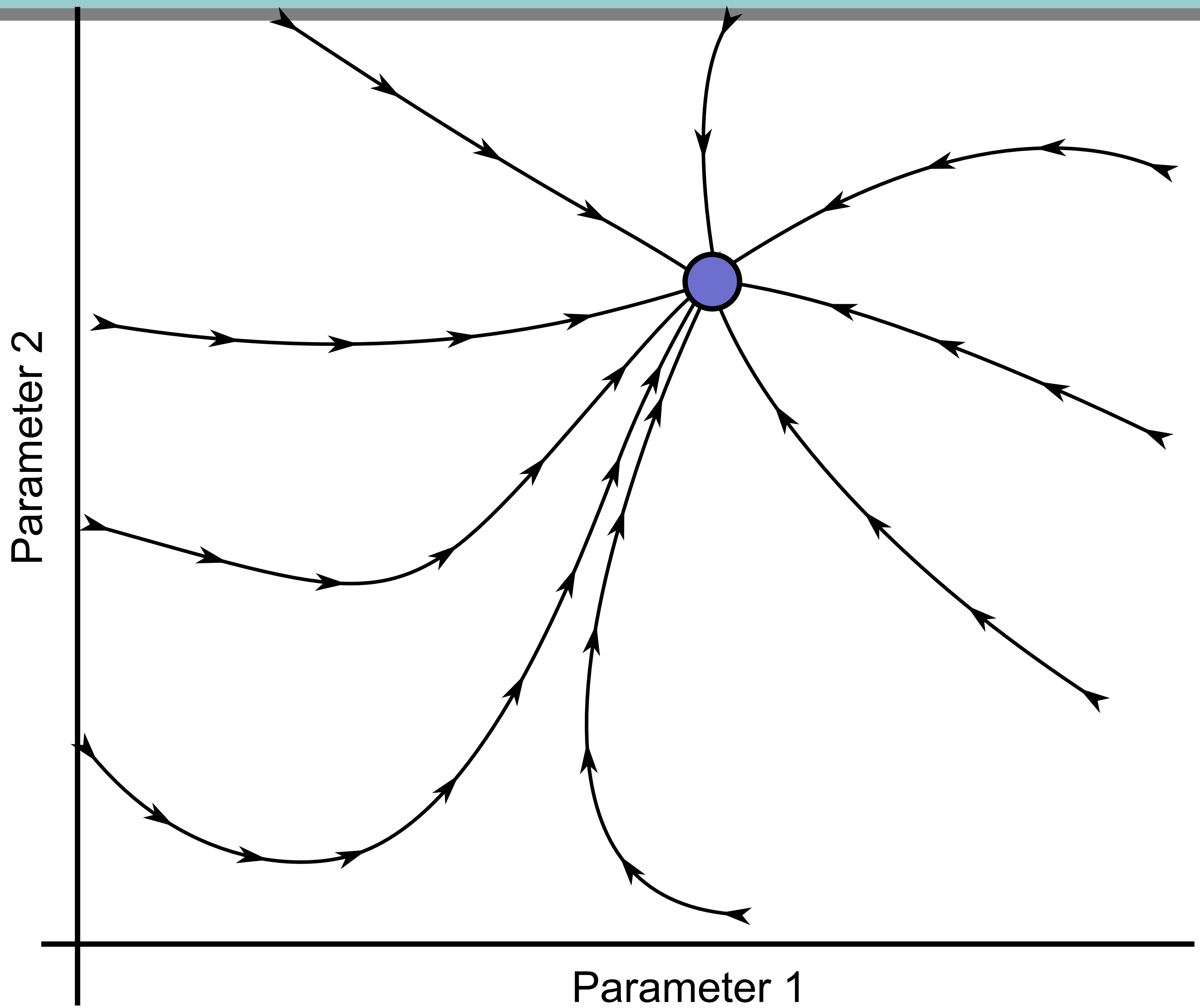
Effective Theories

Both models equally correct at explaining the coarse-grained behavior.

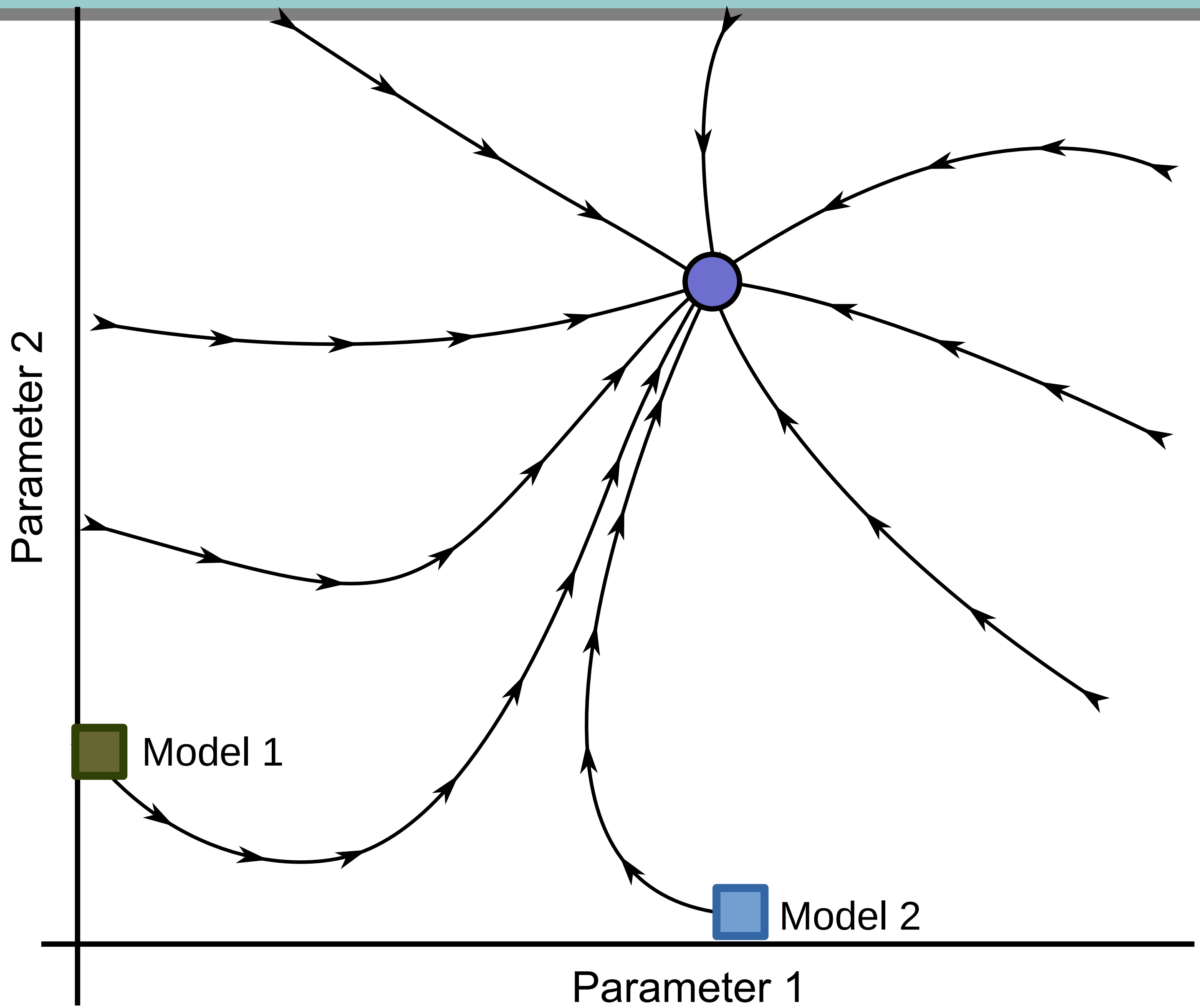
The coarse-grained model is an 'effective theory' for the system, above a certain scale

In general, families of equivalent theories for the same system, associated by changes in scale.

“Theory Space”

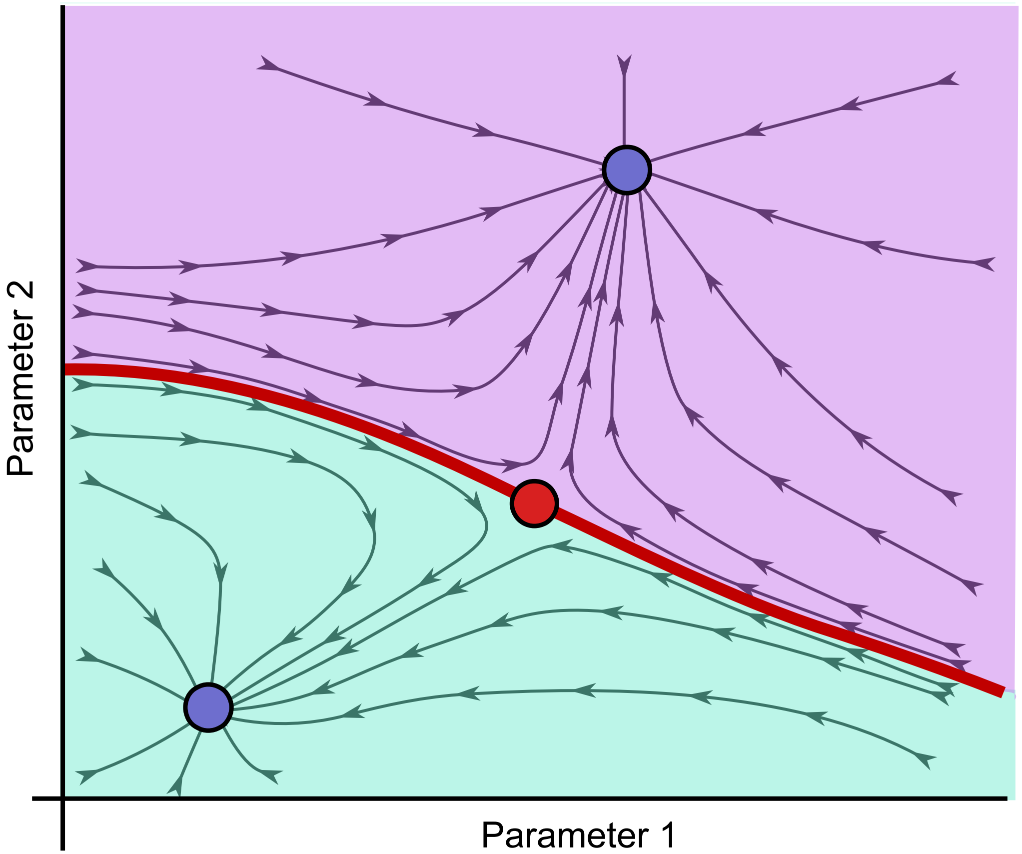


“Theory Space”



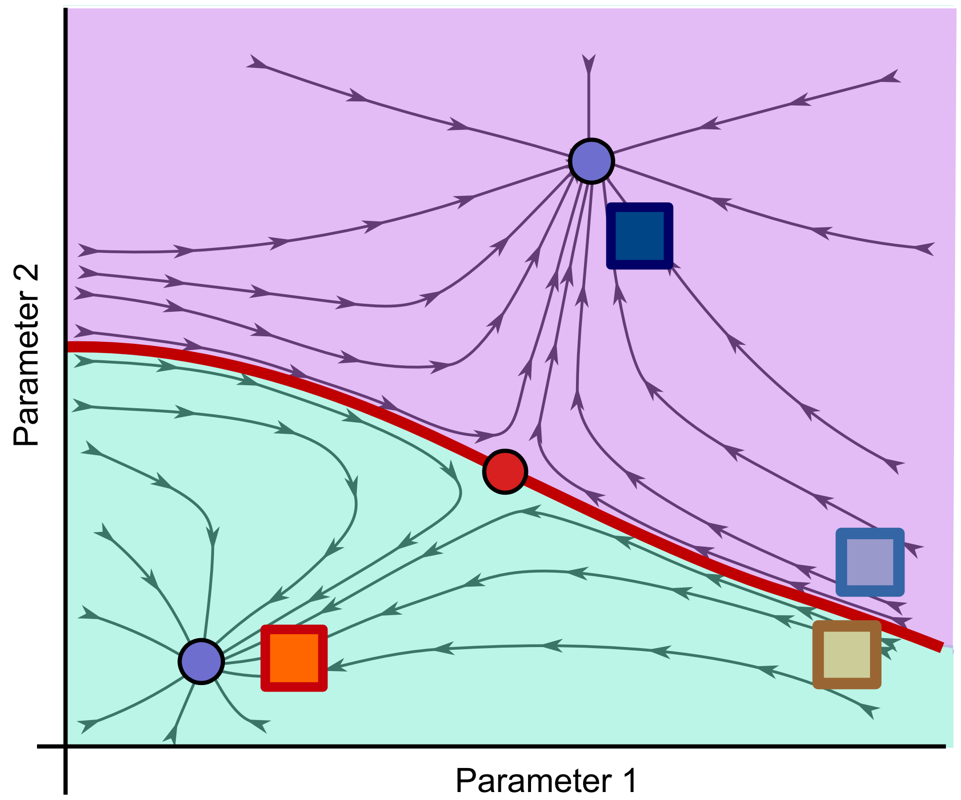
Emergence of Sharp Distinctions

- Multiple attractors: depending where you start, some distinct set of outcomes.



Breakdown of Sharp Distinctions

- In complex cases (biology, neurology), not necessarily at an asymptotic limit.
  - Different cases further along than others
- Even if we see something that seems sharp, there may be blurry examples

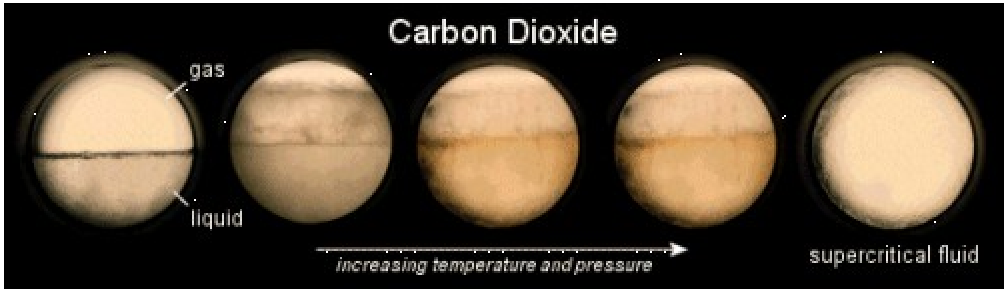


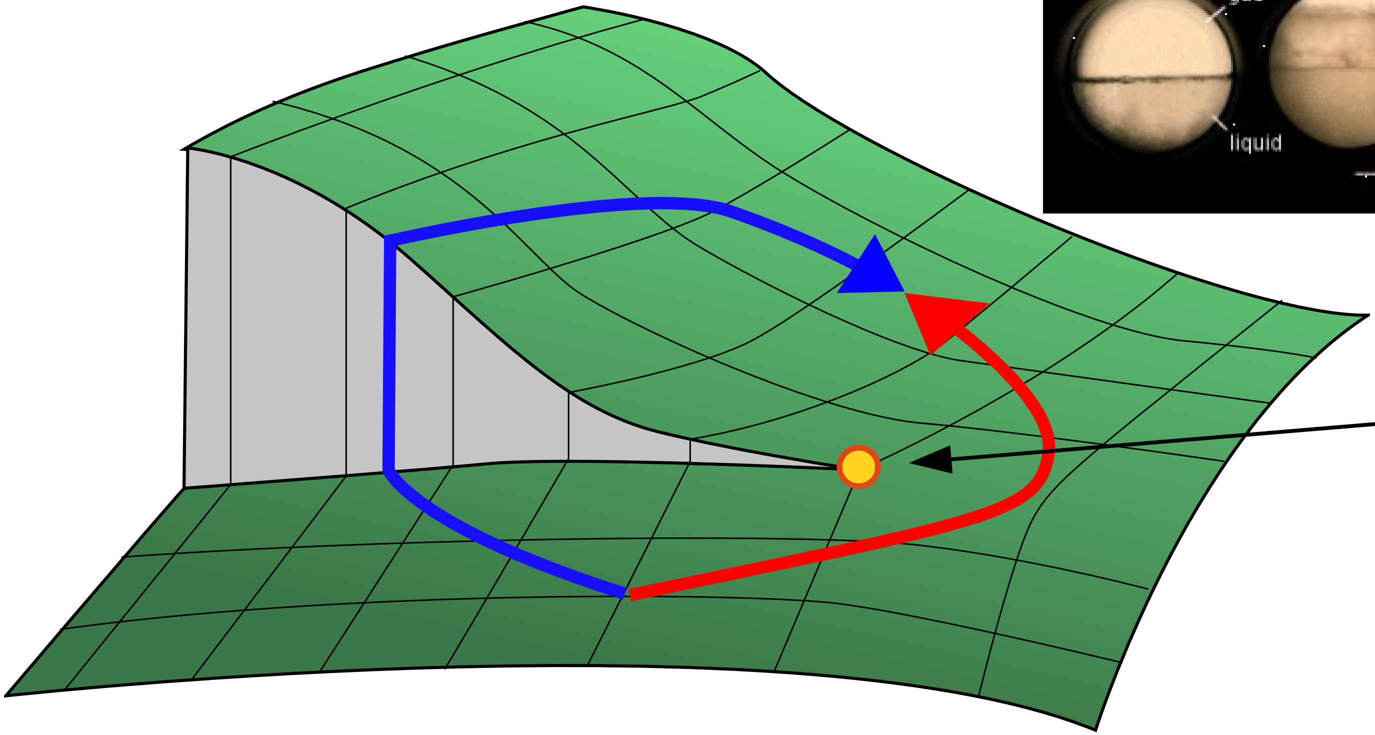
Finding blurry cases doesn't rule out an asymptotic distinction

Breakdown of Sharp Distinctions

Super-critical phenomena

- A distinction may only be **locally** sharp.
  - Usually corresponds to a critical point – super-critical





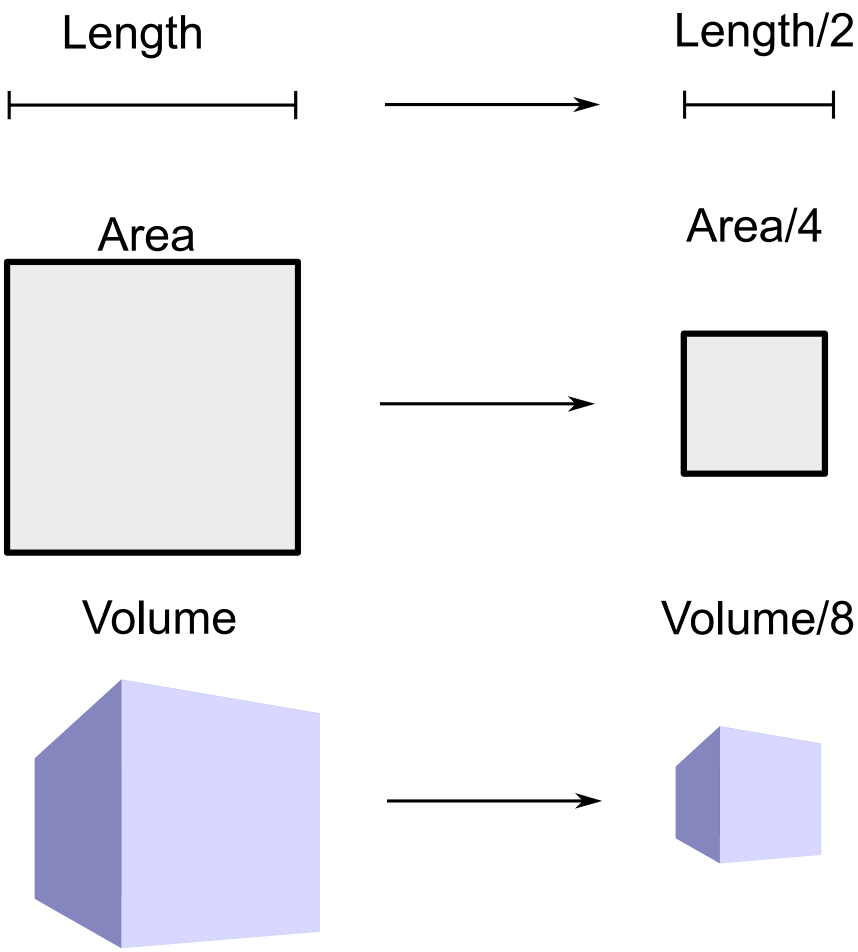
Critical Point

Scaling

The dependence of each parameter on scale usually has an asymptotic **scaling** form:

The way that it changes remains constant at large scales.

- At constant density, things that measure mass scale like the cube of the length, etc.



Example: Allometric Scaling

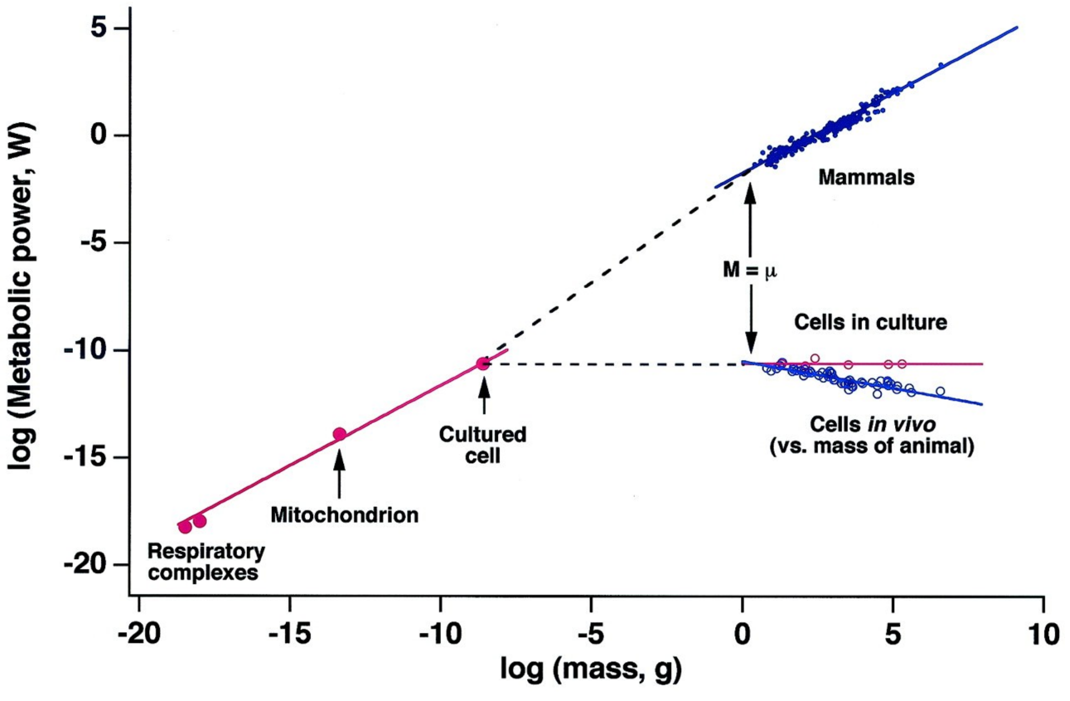
Muscle strength ~ cross-sectional area

Body weight ~ volume

Beyond a certain size, animals would not be able to lift their own body weight.

Example: Allometric Scaling

Many such inter-related scaling laws in biology



Geoffrey West:

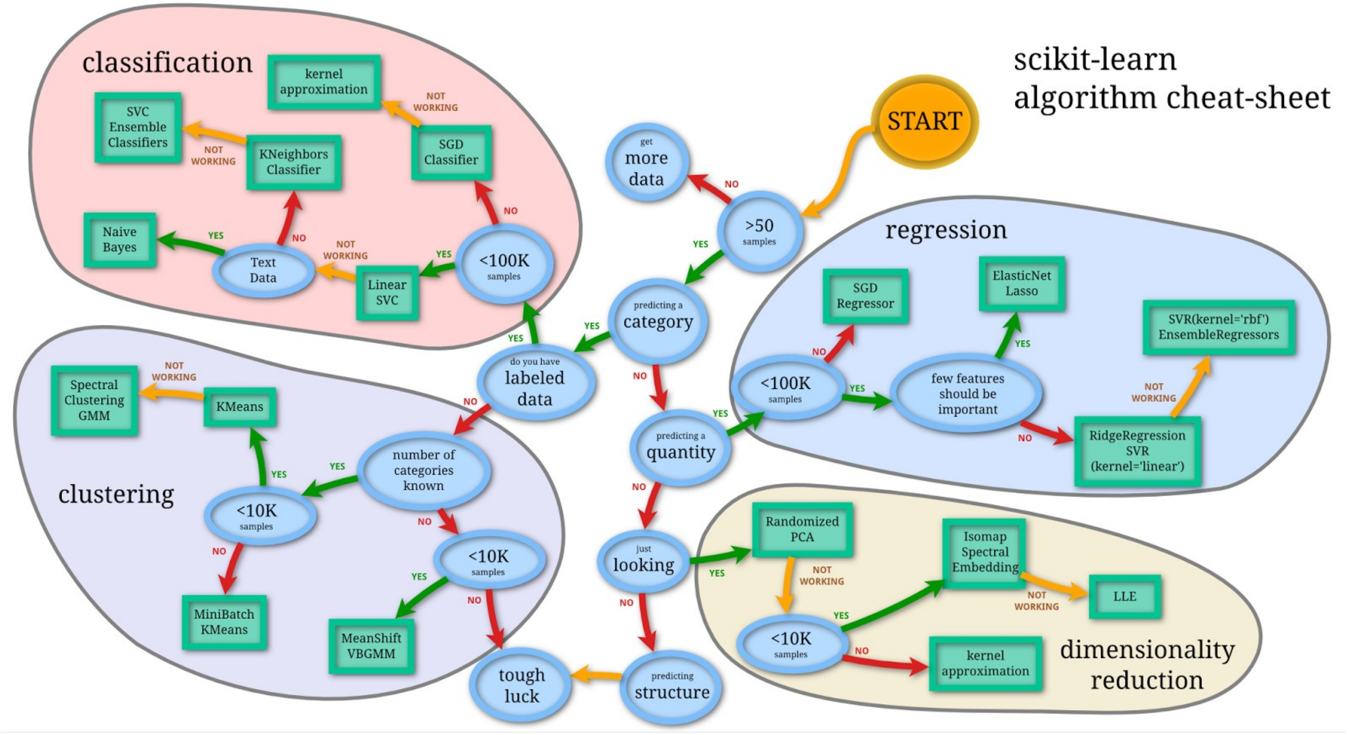
Scaling of pressure in branching veins → heart rate, metabolic constraints, etc.

Approach is now being used for cities

West, G; “Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals” PNAS **99** (2002)

Example: Scaling in Machine Learning

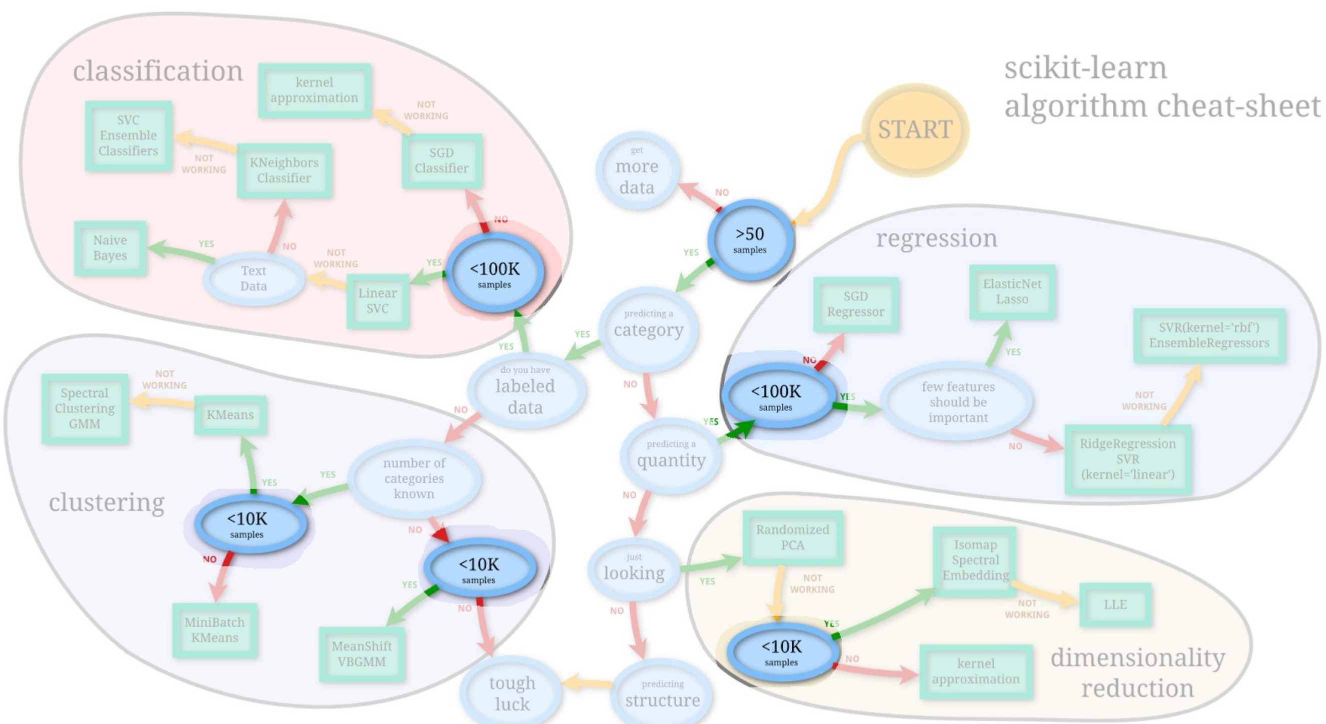
Machine learning has a diverse set of algorithms – why?



Andreas Mueller, “Peekabo: Andy's Computer Vision and Machine Learning Blog”

Example: Scaling in Machine Learning

Machine learning has a diverse set of algorithms – why?



Andreas Mueller, “Peekabo: Andy's Computer Vision and Machine Learning Blog”

Example: Scaling in Machine Learning

- Different algorithms **scale differently**
  - Amount of data ('big data' revolution)
  - Dimensionality of data (high dimension → linear, low dimension → non-linear)
  - Computational time
  - Computational memory
  - ...

In different regions of the problem-space, different methods are (asymptotically) best

Example: Scaling in Machine Learning

- Big pushes for modern neural network approach: understand the performance scaling
- Find and change unfavorable scalings
- As a result, discovered new methodologies
  - Slow convergence of convolutional networks → RBMs, Glorot Initialization
  - 'Vanishing gradient' problem in recurrent neural networks → LSTM units
  - Momentum methods, etc

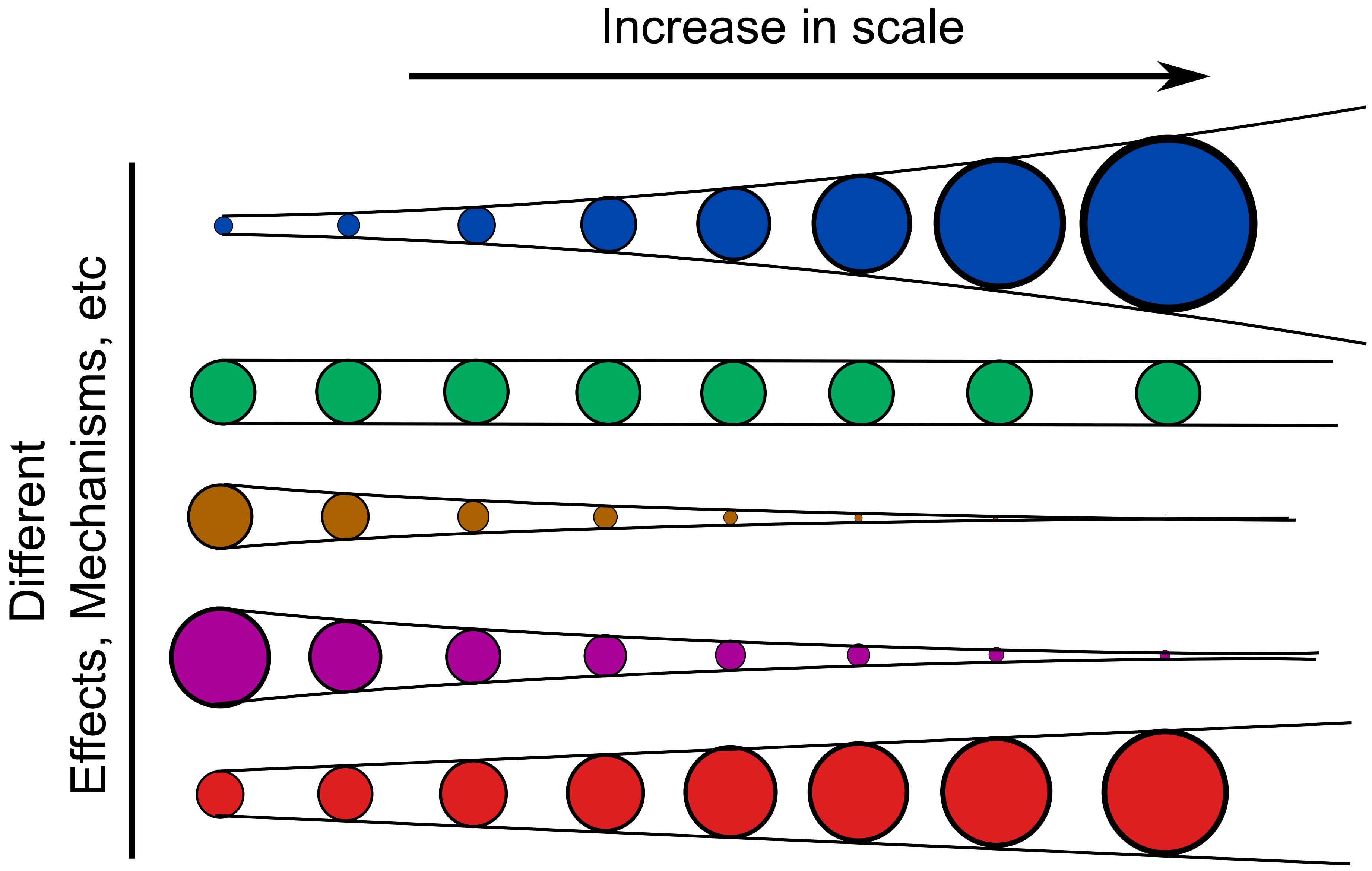
How Scaling Leads to Simplicity

At the start, I promised to talk about why simple models work at all

Decompose theory into distinct effects that can influence an outcome. Each effect can scale differently.

At large scales, differences between importance become magnified.

How Scaling Leads to Simplicity



How Scaling Leads to Simplicity

So at large enough scales, only a few things still matter.

Can think of 'errors' in a theory as terms relative to reality. Each error is associated with a parameter. If zero, the theory is correct.

Only errors which scale as quickly as the dominant effects remain relevant at large scale

Applications: Deep Learning

- Connection between multi-layer neural networks and 'Renormalization Group' (Pankaj Mehta, David J. Schwab, 2014)

- Multi-layered networks learn to implement coarse-graining, preserve only scale-crossing features

- Even if there are many important effects in the data, only a few will be dominant at each scale

What about Life?

Does this work for life in general? Allometric scaling seems to suggest it might. However...

We are large compared to our DNA, but our DNA still has a dominant effect at our scale

(Kunihiko Kaneko: idea of 'Minority Control')

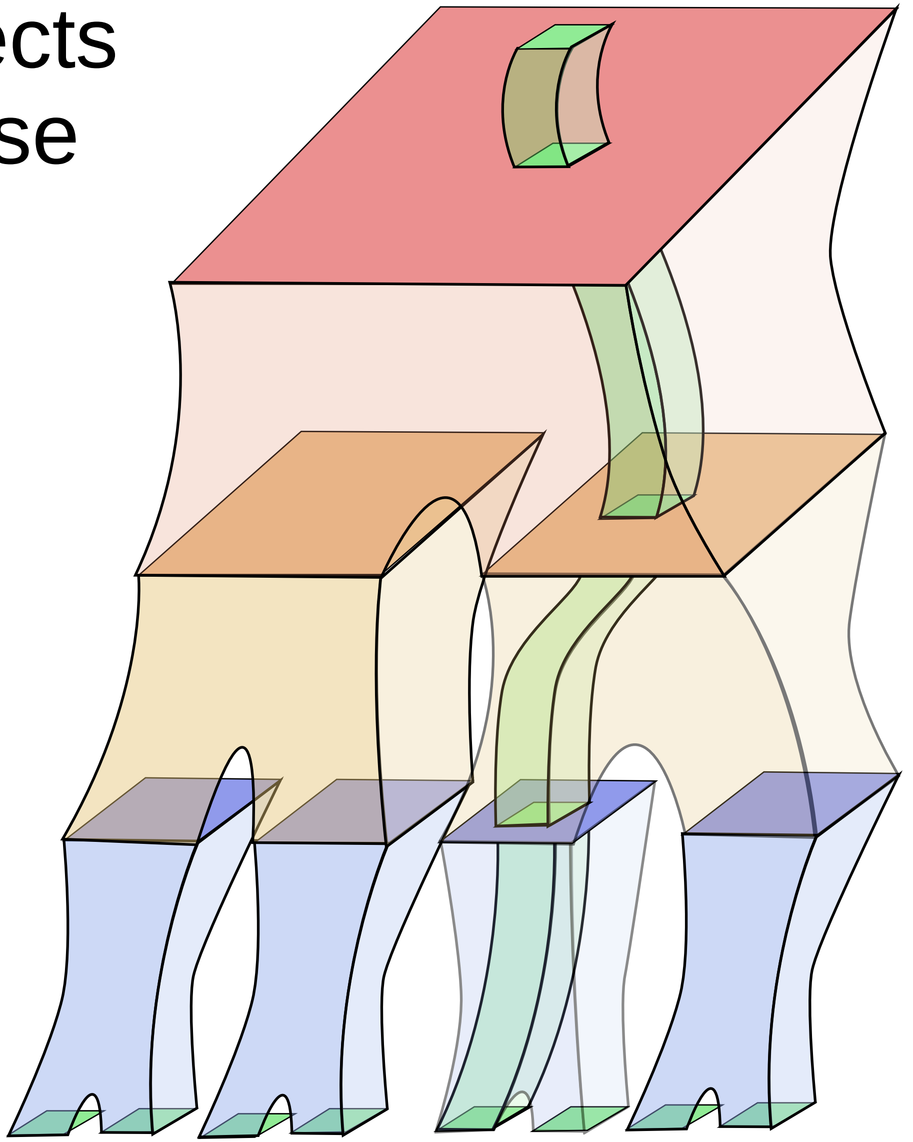
Viruses, other small things can still kill us

What about Life?

These scale-crossing objects seem to be relatively sparse

Perhaps this is functional:

Provides access to small scales without being dominated by them



Conclusions

- Large systems tend to become simpler with scale due to divergence of the relative importance of their component processes
- Because of this, many microscopic theories are equivalently good for describing the same emergent phenomenon – the phenomenon itself is robust to details.
- Some systems appear to build in exceptions
  - What does this do, and why does it happen?