Computer Science Every Physicist Should Know

Tim Mattson

Human Learning Group

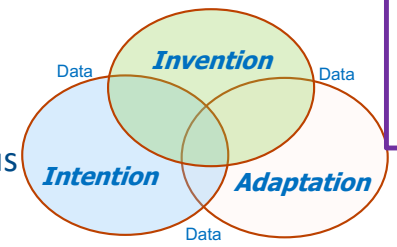
My 40 year Career

[tile]DB
2015
Sparse Array storage engine

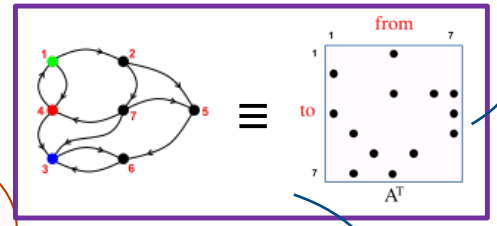

2015
BigDAWG Polystore system

2018-Aug'2023

oneAPI
Future of heterogenous computing



Machine Programming

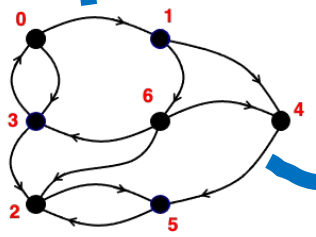


Unified Theory of data and computing

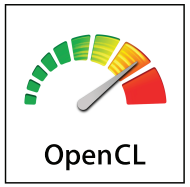
2023-2038
Retirement

Tim Mattson
RIP
1958-2038
Here lies the Christmas bat

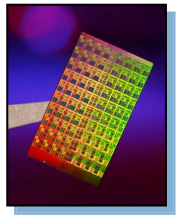
Write, teach, kayak, and think



2013
GraphBLAS

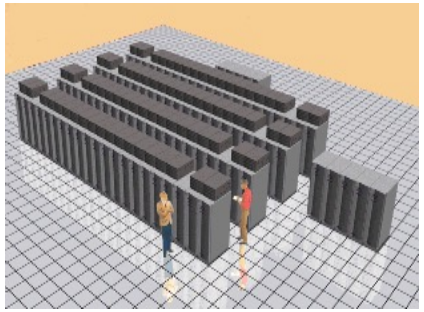


2008
Portable GPU programming


2007
World's First TFLOP chip

2002-2005
Director of life sciences

OpenMP
1997
Portable multithreading


1996
ASCI Red:
World's first TFLOP

PhD chemistry
1979-1985

PostDoc 1986
Caltech Concurrent Computation Project

Jobs with startups …
numerical analysis, signal processing,
scientific computing, and parallel computing
1987-1990

1990-1993
Linda and pre-MPI message passing

intel
1993
Supercomputing Systems Division

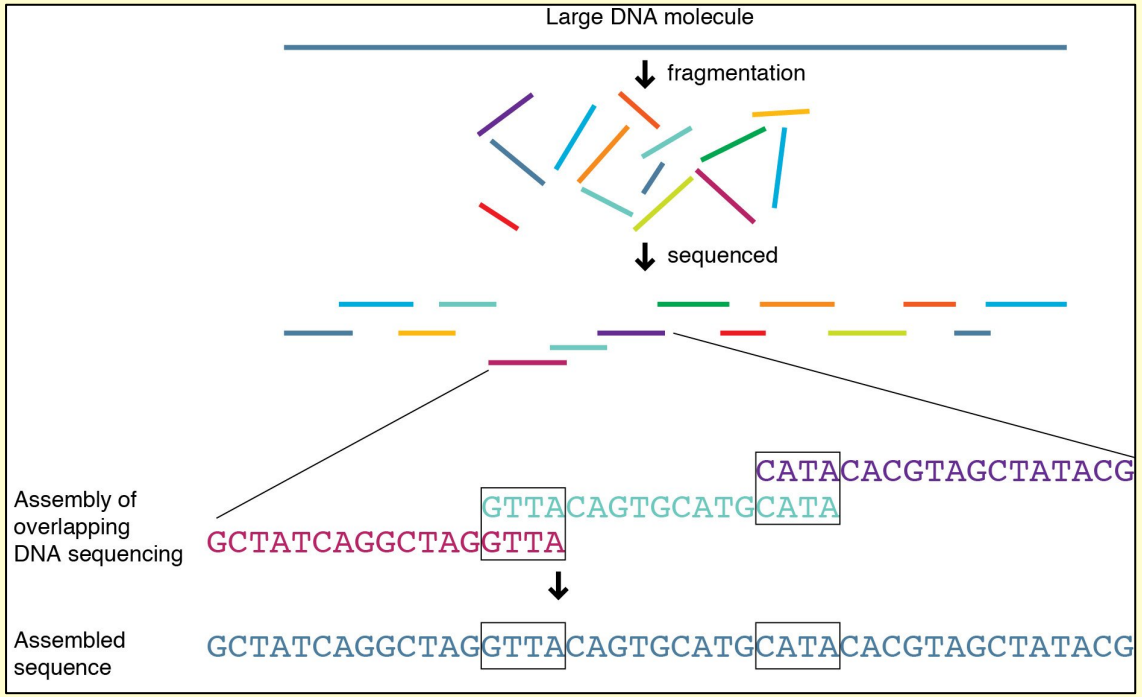
MPI
1994
Message Passing Interface

1994
FDIV Bug

Consider three different experimental systems

A “photograph” of Dark Matter

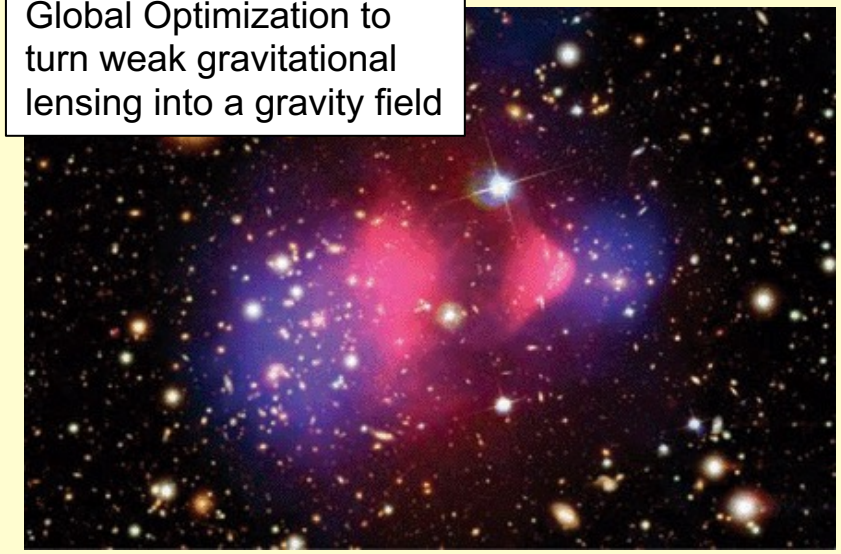
Imaging the event horizon of a black hole



Genome assembly

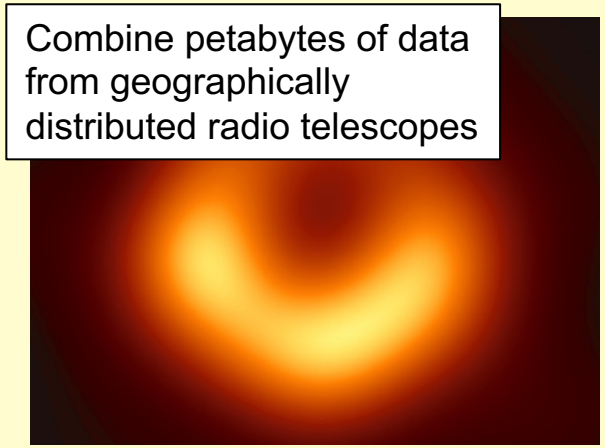
Consider three different experimental systems

Global Optimization to turn weak gravitational lensing into a gravity field



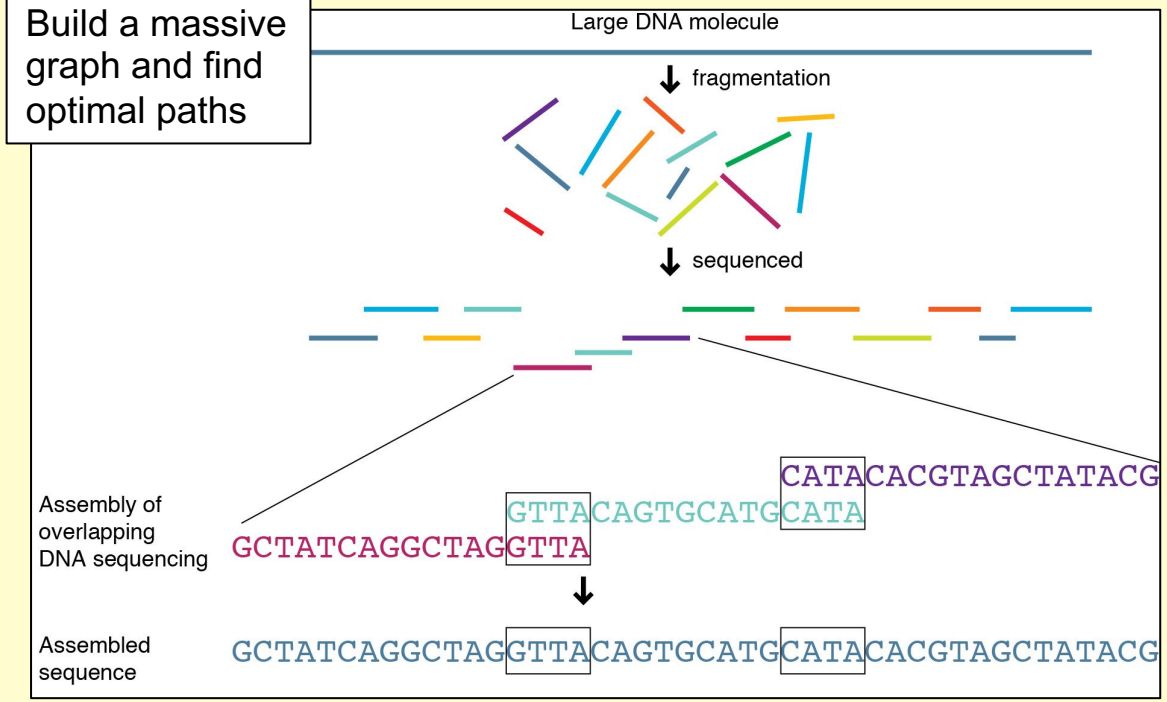
A “photograph” of Dark Matter

Combine petabytes of data from geographically distributed radio telescopes



Imaging the event horizon of a black hole

Build a massive graph and find optimal paths



Genome assembly

Computing doesn’t just connect data to theory … It is an integral part of our experiments.

Computing is fundamental to almost everything we do in physics.

Our Goal for this series of lectures

- There is a body of knowledge from computer science practicing physicists are expected to know.

- Our goal is to survey that body of knowledge … so you have a high level understanding of the key concepts you need from computer science.

- We will cover the following topics
  - What is a computer? How does our software run on a computer?
  - Computer architecture and how to get the most from the computers you use
  - Supercomputing and the need for (and use of) of parallelism
  - Fundamental design patterns and programming models for parallel programming
  - Computer arithmetic: how it works and (more critically) how it fails
  - The use and abuse of random numbers
  - An overview of Data Management technologies

Our Goal for this series of lectures

- There is a body of knowledge from computer science practicing physicists are expected to know.

- Our goal is to survey that body of knowledge … so you have a high level understanding of the core concepts you need from computer science.

- We will cover the following topics
  - What is a computer? How does our software run on a computer?
  - Computer architecture and how to get the most from the computers you use
  - Supercomputing and the need for (and use of) of parallelism
  - Fundamental design patterns and programming models for parallel programming
  - Computer arithmetic: how it works and (more critically) how it fails
  - The use and abuse of random numbers
  - An overview of Data Management technologies

We will meet in room P1C in the Paolotti Building for the next four Fridays (29/11, 6/12, 13/12, 20/12) from 16:30-18:30.

I want these to be student driven lectures. Ask questions. Discuss the material in class. Slow me down when you need more information. Push me faster when you are bored. I'd rather cover a subset of this material well than race through all my PowerPoint slides to cover every topic

How about Artificial Intelligence?

Are we going to spend much time with AI?

AI and Machine learning

- This is a depressingly accurate (though sarcastic) summary of AI and Machine learning.
- We do not understand as much about the world as we like to think.
- Hence, systems that try to be smart often miss important aspects of a problem and don't work very well.
- It is often better to NOT pretend you understand anything and use a brute force approach such as AI

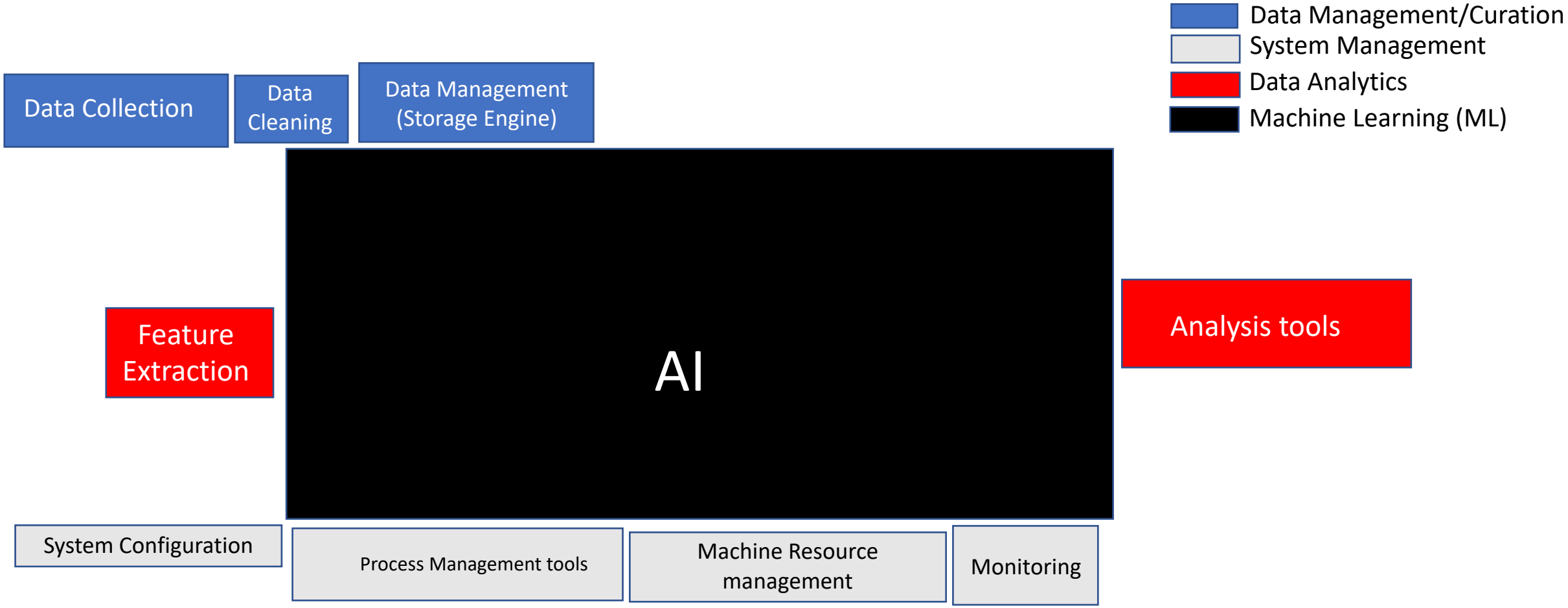
XKCD, Randall Munroe: https://xkcd.com/1838/

Be careful about the data you use to train your models

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.

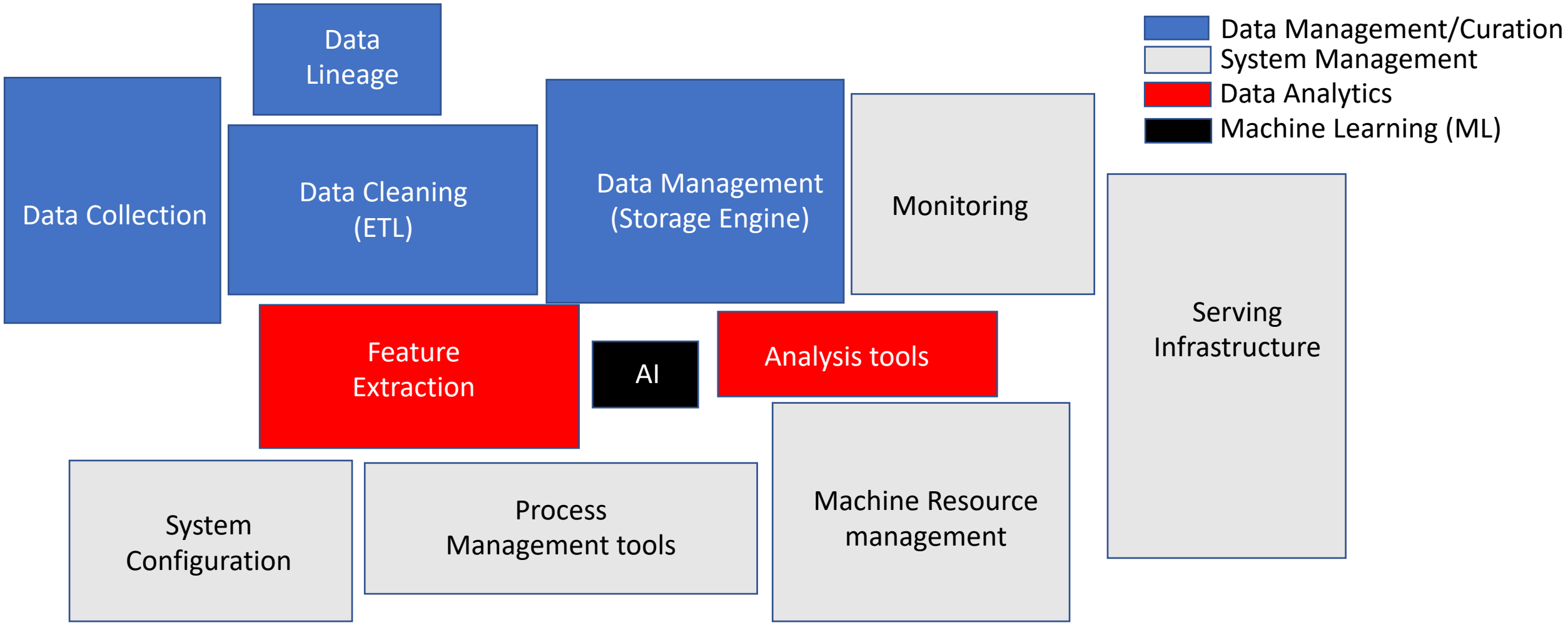
https://www.smbc-comics.com/comic/rise-of-the-machines

The general buzz about AI workflows



Area of Boxes $\propto$ effort expended within a workflow

The reality of priorities of a working data scientist using AI

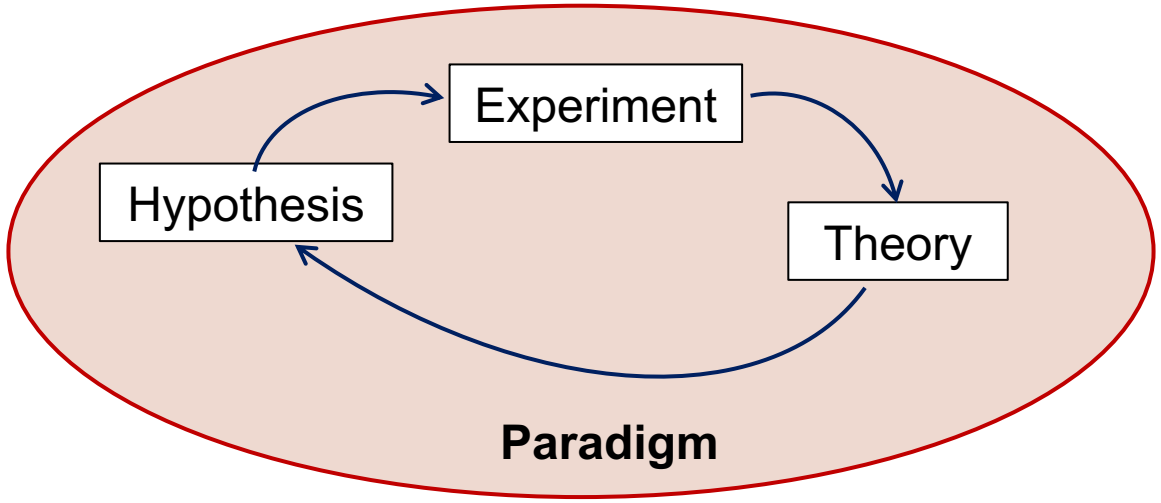


ML code is 2 to 5 percent of the code written by a data scientist*. Most of the code is "glue code" to manage data and system components

*based on figure 1 from "Hidden Technical Debt in Machine Learning Systems", by D. Sculley et. al. from Google.

Artificial Intelligence (AI) and the sciences

- The purpose of science is to gain insight into the physical world.
- AI today mostly just finds patterns and fills in “blanks”.



- It does not explain “its magic” and hence is useless for developing insight.

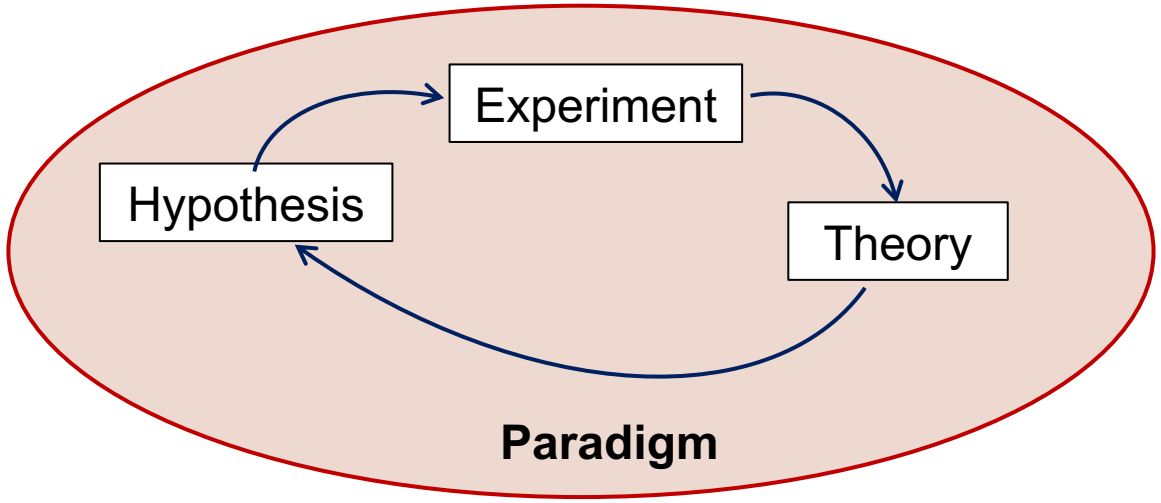
I was actively involved with AI in the old days (the mid 80’s). This was the golden age of symbolic AI. We were going to build “expert systems” and change the world. We failed.

So today, I’m a bit of an AI curmudgeon*.

*Curmudgeon: An ill-tempered (and usually old) person full of resentment and stubborn notions.

Artificial Intelligence (AI) and the sciences

- The purpose of science is to gain insight into the physical world.



- AI today mostly just finds patterns and fills in "blanks"
- It does

Recent developments have forced me to stop being so hard on AI-as-Science

There are problems that even when we know the physics, we can't produce the insights we seek. Brute force computing just won't make it

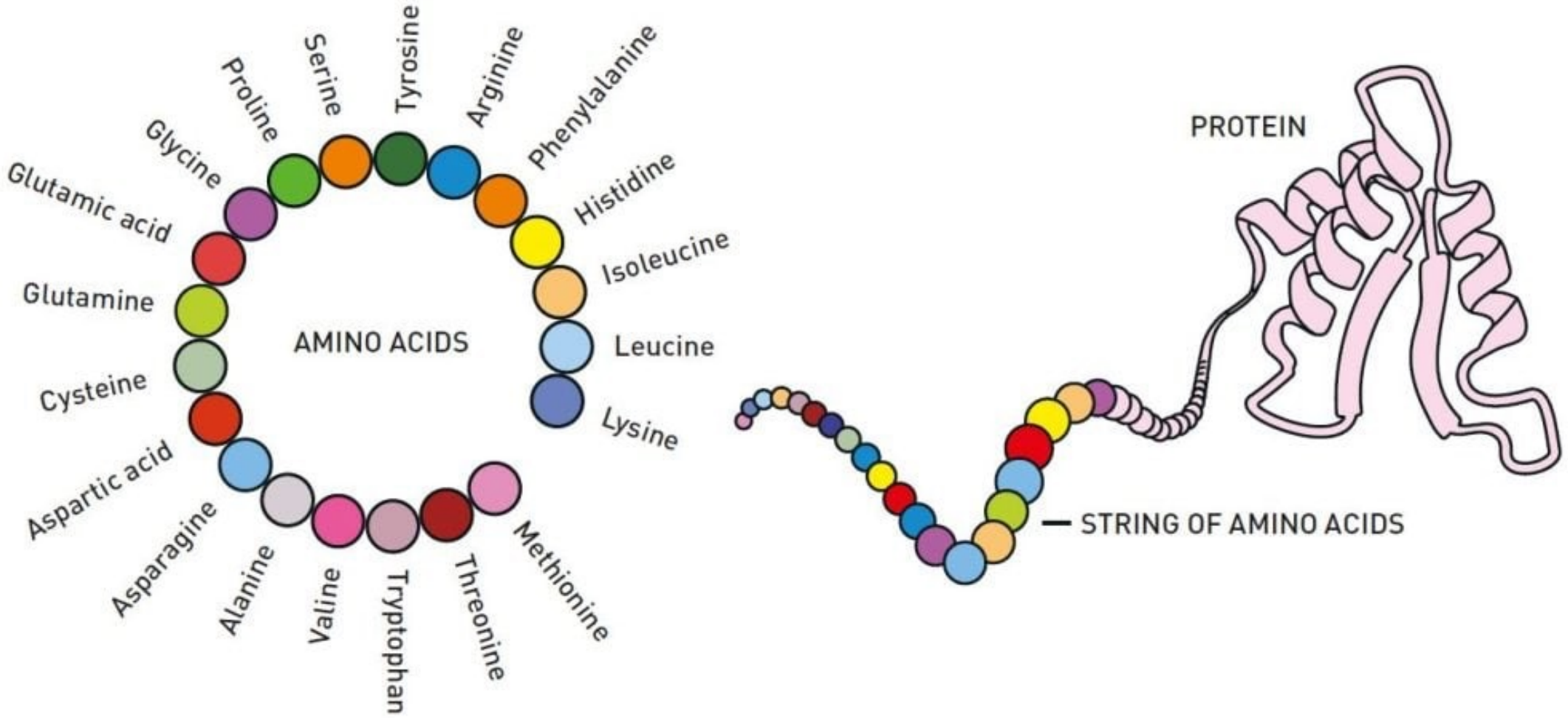
I was actively involved with AI in the old days (the mid 80's). This was the golden age of symbolic AI. We were going to build "expert systems" and change the world. We failed.

So today, I'm a bit of an AI curmudgeon*.

*Curmudgeon: An ill-tempered (and usually old) person full of resentment and stubborn notions.

Protein Folding

Proteins are the fundamental molecular building blocks of life. DNA codes mRNA which is translated into a sequence of amino acids which fold into the final, functional protein. Protein function depends on its 3D shape.



Nobel Prize in 1972 (to Christian Anfinsen) for work proving that all the information needed to determine the 3D shape is contained in the linear sequence … hence, read an mRNA and you know the 3D ahape and function.

Protein Folding Physics has been known for decades

- Electrons move much faster than nuclei, so for motions of nuclei, electron motions smear-out into a force field.

- There are two terms in the forces … bonded and nonbonded (electrostatic).

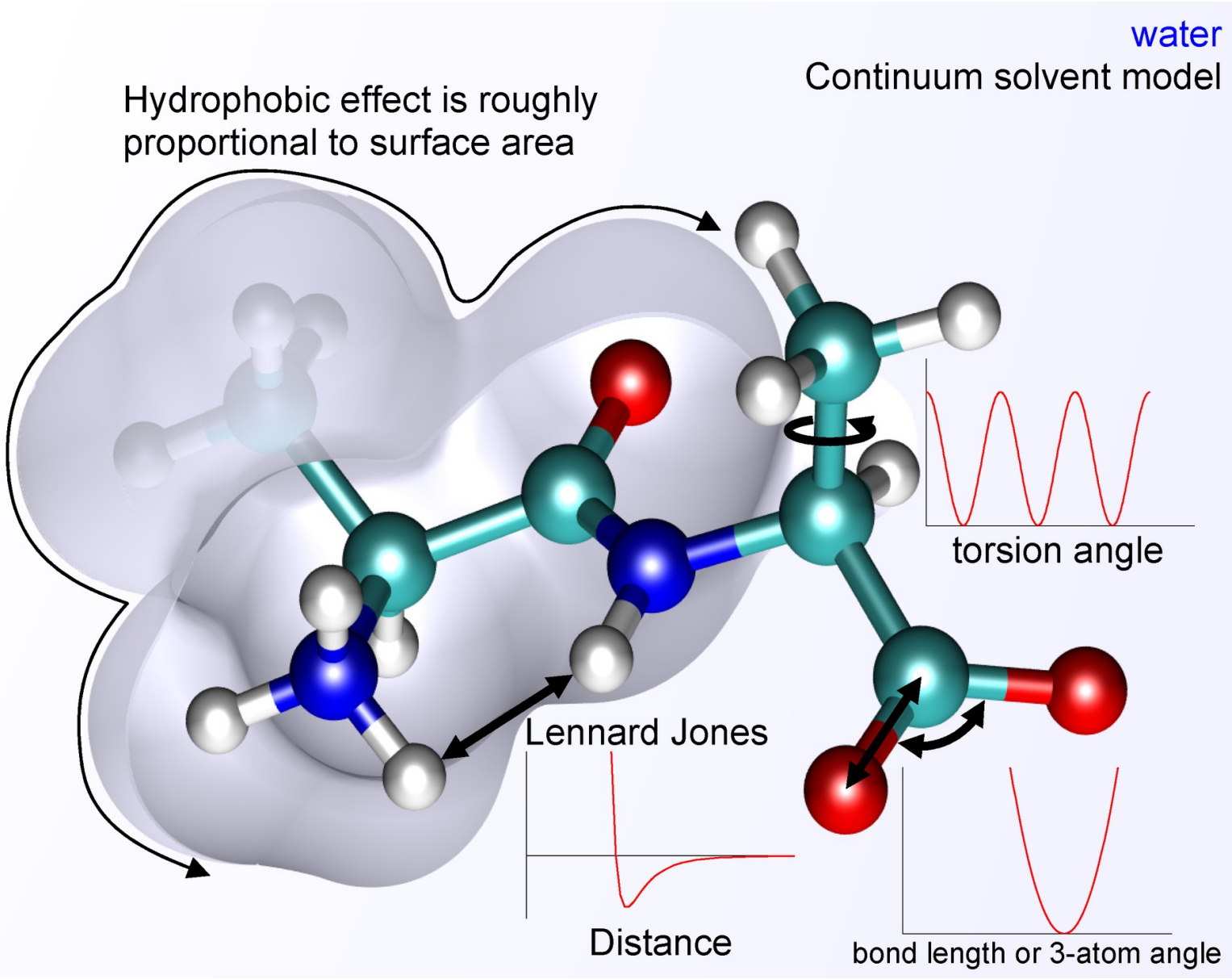
- For example, stretching a bond between atoms i and j relative to their equilibrium length, l_{0ij}, is just Hooke's law (with spring constant, k_{ij})

$$E_{Bond} = \frac{k_{ij}}{2}(l_{ij} - l_{0ij})^2$$

- Electrostatic terms are Coulomb's and Lennard Jones potentials:

$$E_{Coulomb} = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}}$$

$$E_{LJ} = 4\varepsilon\left[(\sigma/r_{ij})^{12} - (\sigma/r_{ij})^{6}\right]$$



Protein Folding Physics has been known for decades

- Electrons move much faster than nuclei, so for motions of nuclei, el
fo
- Th
bo
- Fo
be
ec
la

water

Continuum solvent model

Hydrophobic effect is roughly

Knowing the physics doesn't really help us. A protein with just 100 Amino Acids has at least 10^{47} different 3D structures. It would take supercomputers running full bore for the age of the universe to explore all those structures by brute force.

So how should we fold proteins? Humans use homologous reasoning (the sequence is similar to this known protein so it should fold the same way) and heuristics

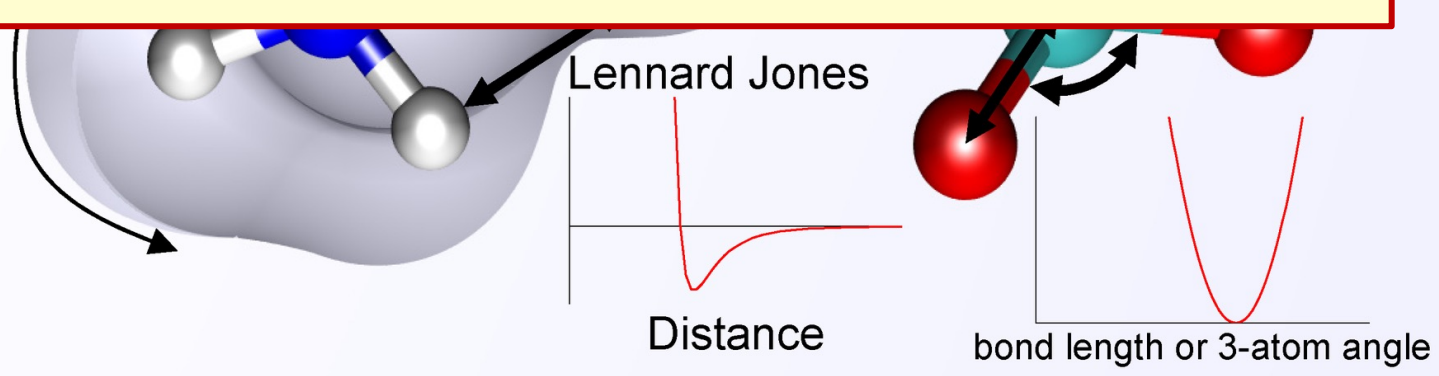
But this sort of reasoning (analogy and heuristics) is exactly what AI does well.

So much science rides on connecting a proteins linear sequence to its 3D shape … it just makes sense to use AI.

- Electrostatic terms are Coulomb's and Lennard Jones potentials:

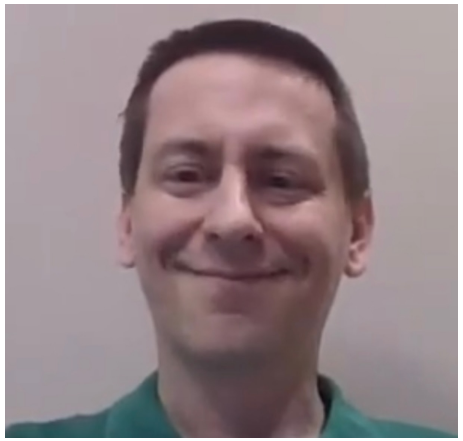
$$E_{Coulomb} = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}}$$

$$E_{LJ} = 4\varepsilon\left[(\sigma/r_{ij})^{12} - (\sigma/r_{ij})^{6}\right]$$



By Edboas - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=4194424

2024 Nobel prize in Chemistry: Rosetta and Alpha fold



Demis Hassabis **John Jumper**

Google Deep Mind

David Baker

University of Washington, Seattle, WA

- Using AI technology first developed for playing the game, Go, a team developed an AI system for deducing protein structure … Alpha fold.
- Uses a combination of techniques but at the heart of the system is the same transformer technology at the heart of famous Large Language Model (LLM) systems such as ChatGPT.

- Protein design software (Rosetta), later adapted to use AlphaFold. Resulted in ab initio protein design solutions to design new proteins not found in nature.
- This could have a revolutionary impact on humanity with new drugs, new materials, and more

Image credit: Google Deep Mind

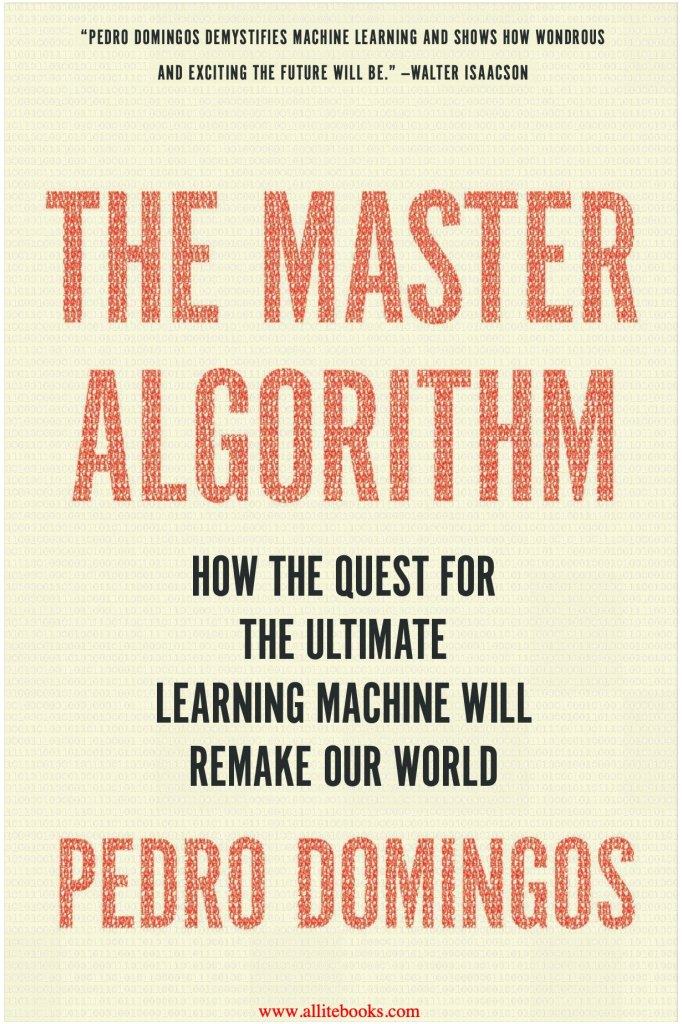
https://www.youtube.com/watch?v=1FD9Br7ybqk
National Academies - Earth and Life Studies

AI is not foundational to physics … many of you will not need anything beyond a casual knowledge of AI.

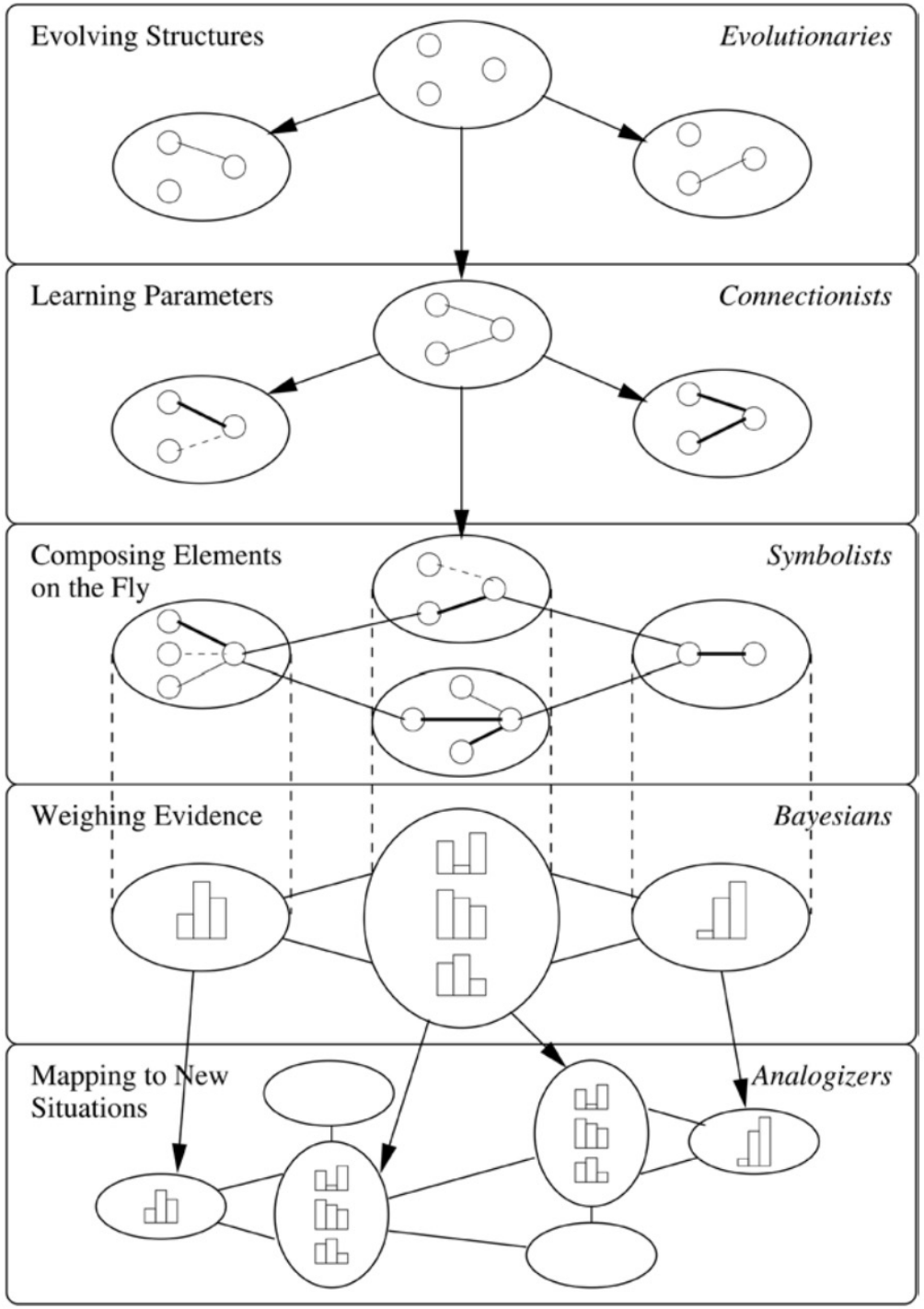
It is also a huge topic that deserves its own series of focused lectures … therefore we will not cover it.

Well … other than a few comments about the future of AI

The five tribes of AI

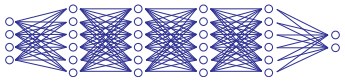


This book is old (2015), but its still one of the best AI-overview books I've found.

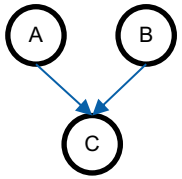


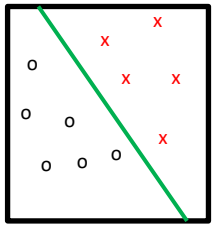
The five tribes of AI

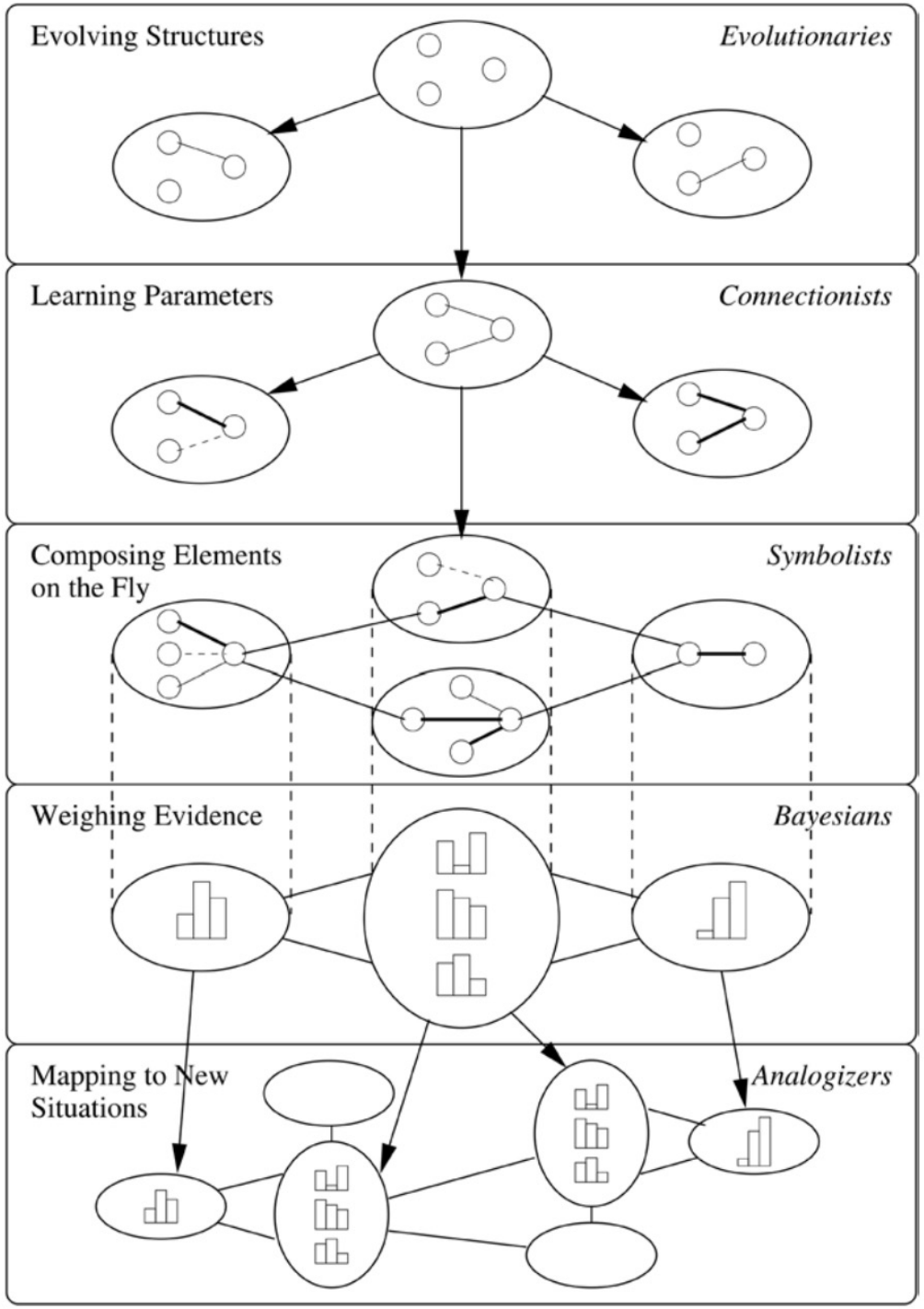
- **Evolutionaries**
  - Genetic algorithms → Genetic programs: for learning as evolution
- **Connectionists**
  - Brain inspired, Neural networks with Lots of layers to get to Deep Learning
- **Symbolists**:
  - Rules (e.g. first order predicate logic) connecting input to a training set.
- **Bayesians**
  - Graphical models trained by Bayesian statistics
- **Analogizers**
  - Find similarity groups … clusters, support vectors, subgraphs



$$A \vee B = \neg A \wedge \neg B$$

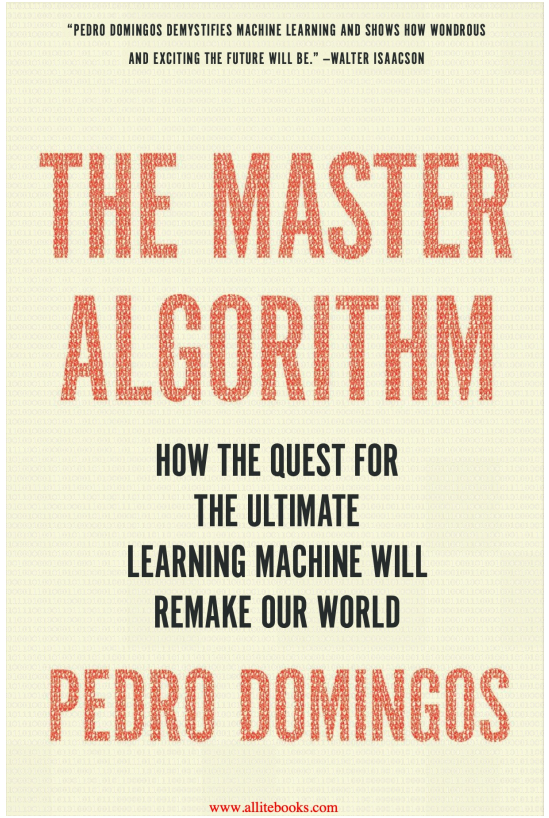


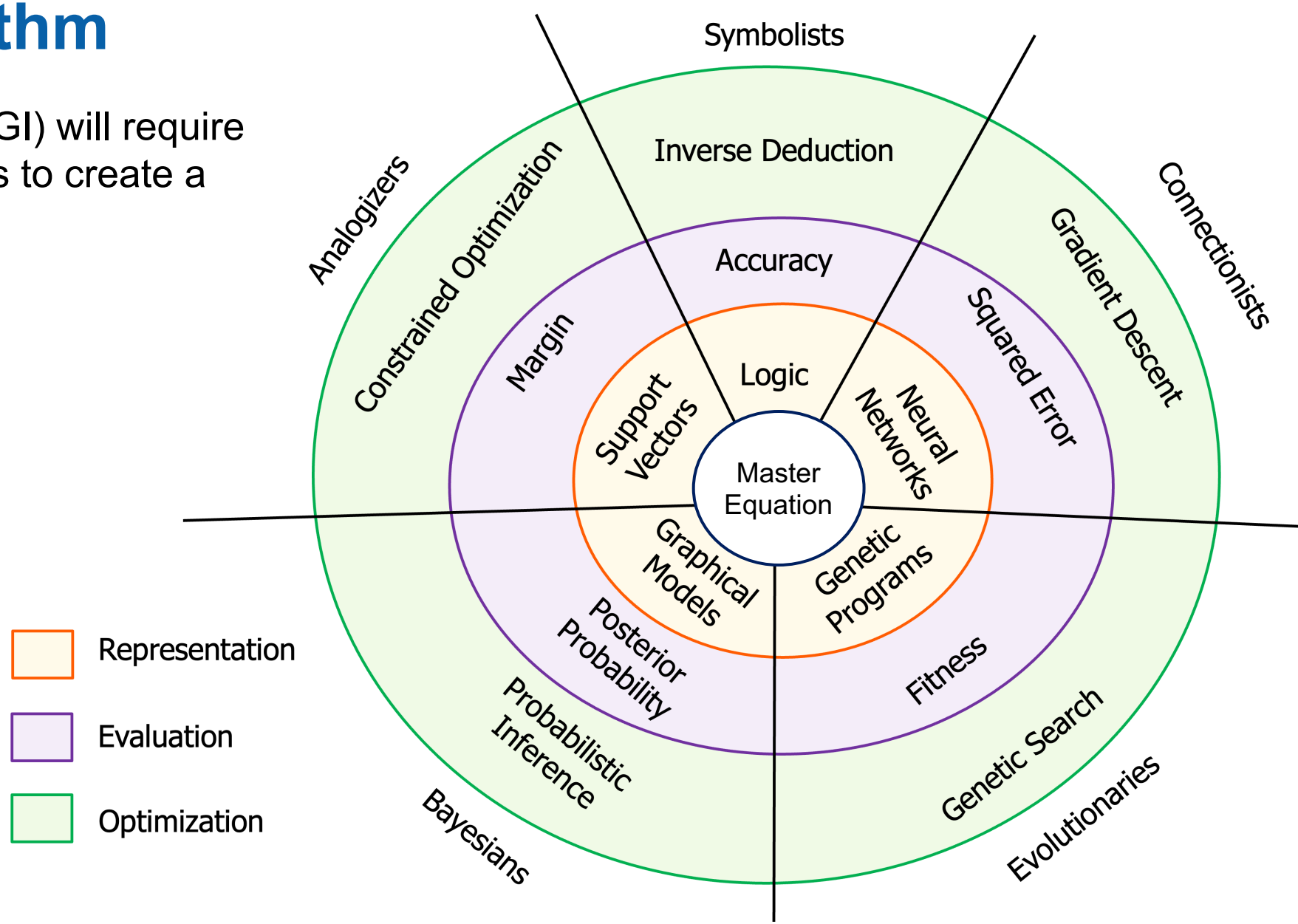




The Master Algorithm

Artificial General Intelligence (AGI) will require a synthesis across the five tribes to create a single, master algorithm

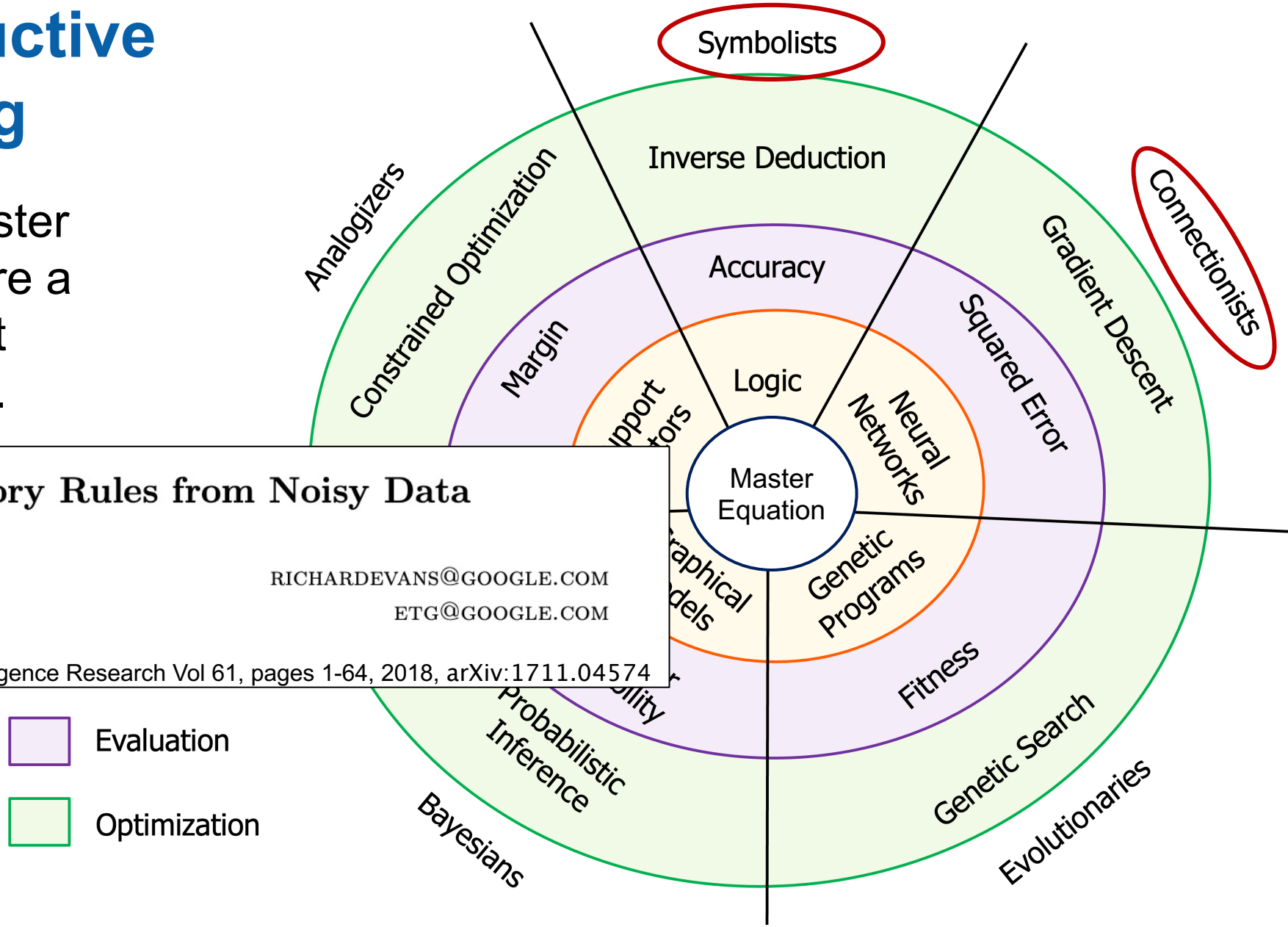




Artificial General Intelligence (AGI) is AI that matches or surpasses humans across a wide range of cognitive tasks (as opposed to current AI focused on single tasks).

Differentiable Inductive logic programming

- We do not have the master algorithm yet, and we are a long ways from AGI, but progress is being made.



Learning Explanatory Rules from Noisy Data

Richard Evans RICHARDEVANS@GOOGLE.COM
Edward Grefenstette ETG@GOOGLE.COM
DeepMind, London, UK

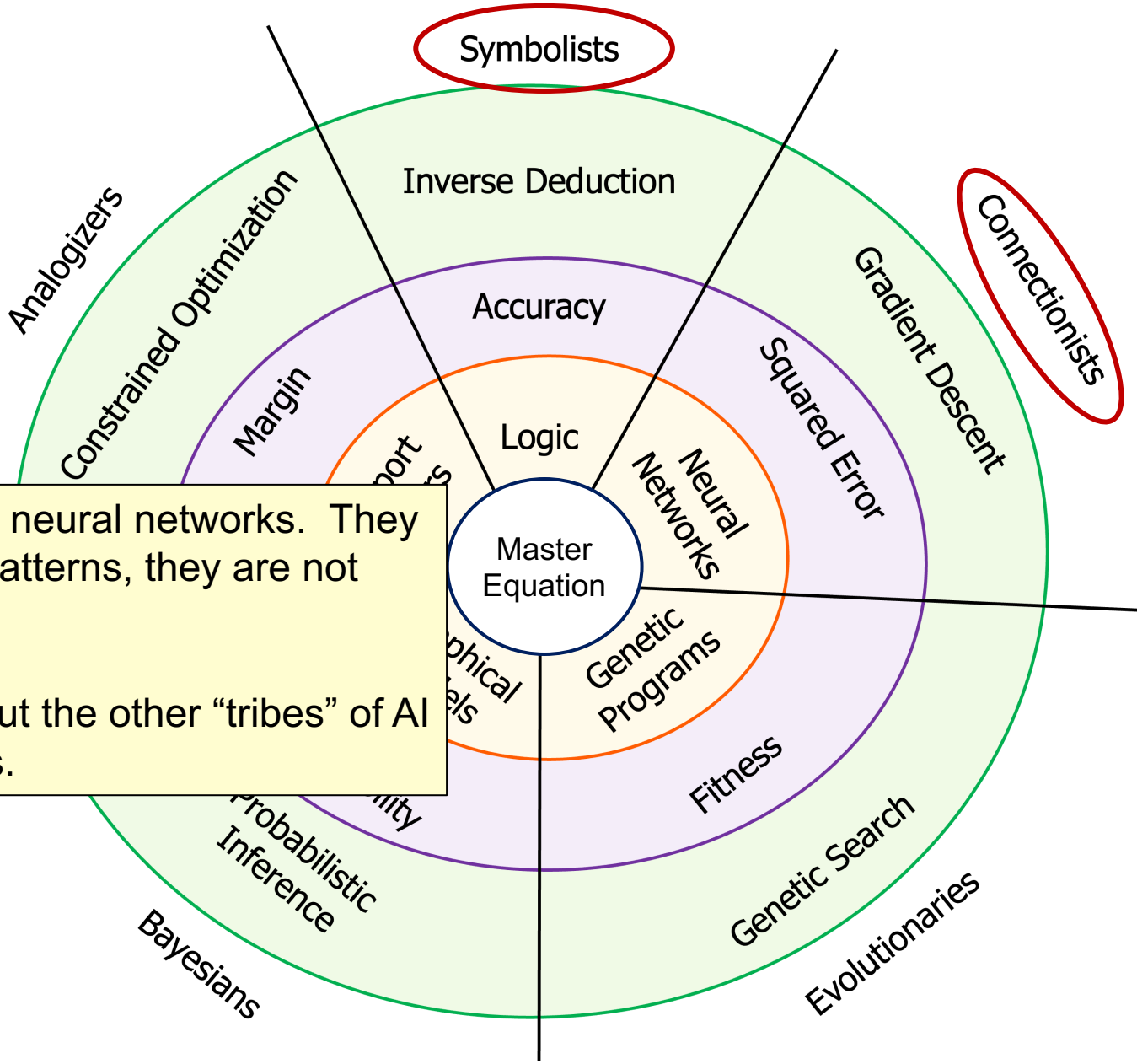
Journal of Artificial Intelligence Research Vol 61, pages 1-64, 2018, arXiv:1711.04574

Differentiable Inductive logic programming

- We do not have the master algorithm yet, and we are a long ways from AGI, but progress is being made.

… so don't fall prey to the irrational exuberance over neural networks. They are powerful but if the goal is insight not finding patterns, they are not enough.

It may not be fashionable, but I urge you to learn about the other "tribes" of AI in addition to Neural Networks.



And that’s it for AI … let’s move to our core material.

What is a computer and how do we use them?