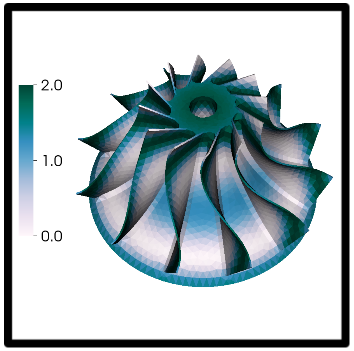
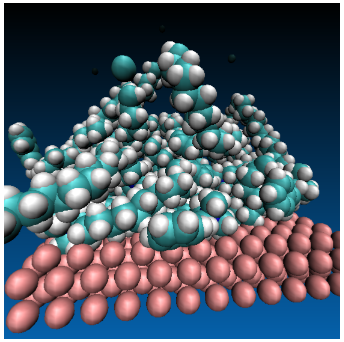
Computational Taxonomy

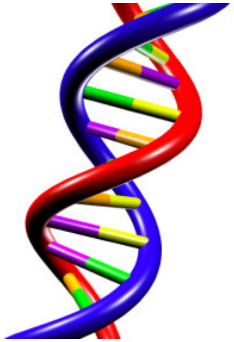
The right solutions for the right research problems

In the field

- In 2015, CHTC's research computing facilitators:
  - Met with 371 researchers from UW Madison*
  - Representing ~57 departments
  - 258 of these researchers were new users of CHTC resources
- Representing a wide variety of:
  - disciplines, research questions, backgrounds
  - computational problems and needs

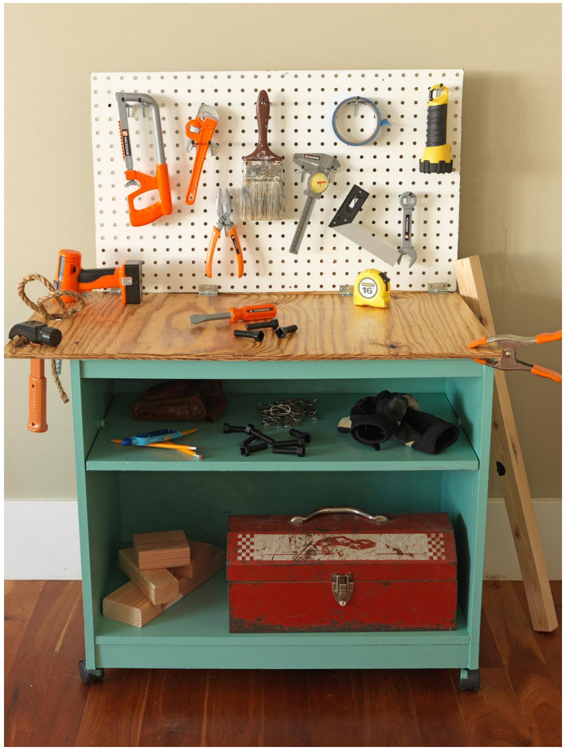







Research matters

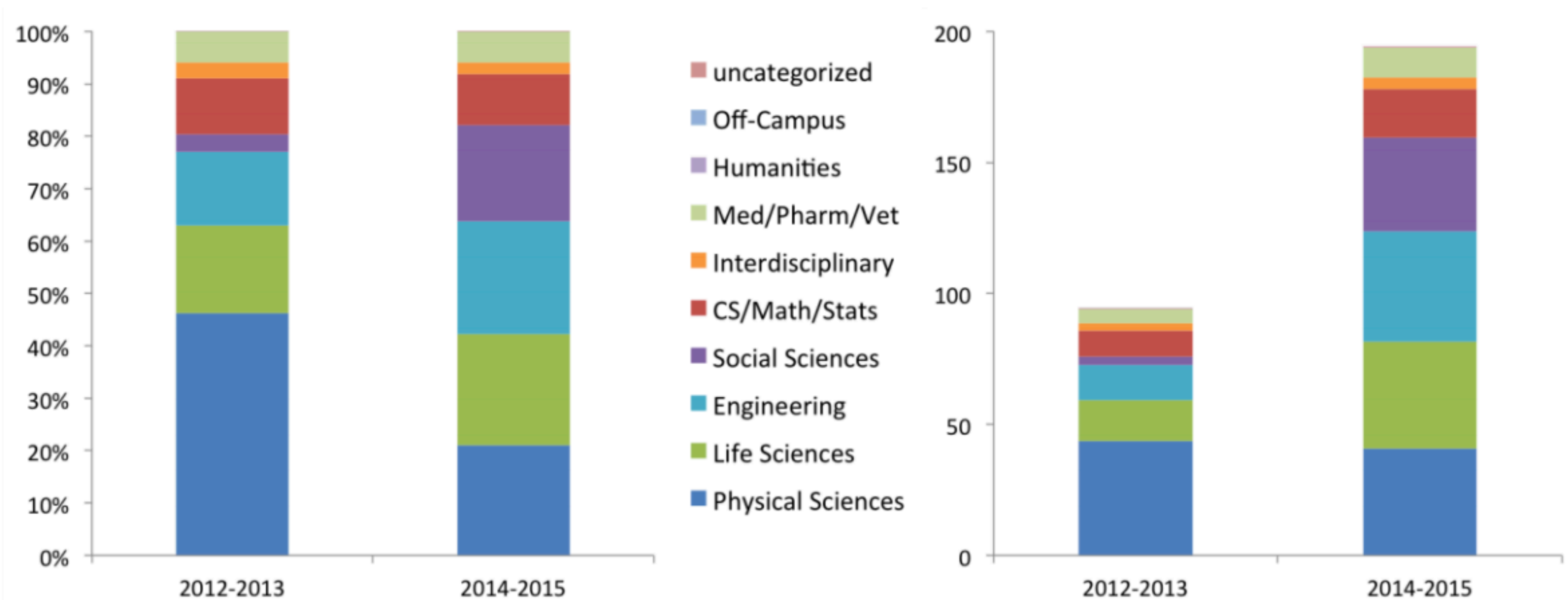
- Which tools to use?
- Understanding computational research problems is a first step to providing appropriate solutions.
- Benefits include:
  - Better resource utilization
  - A broader range of computing-enabled researchers



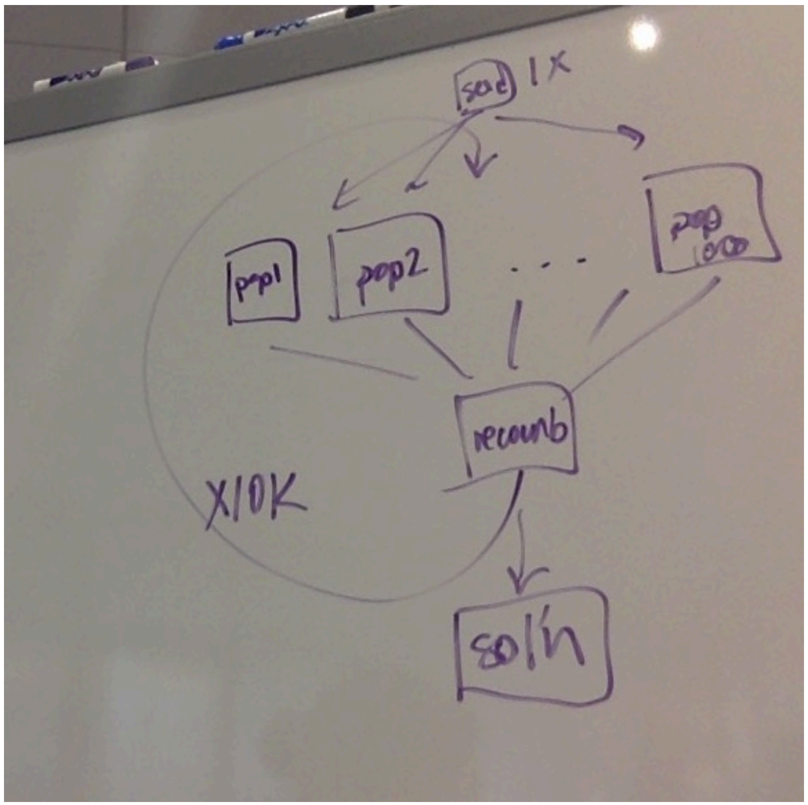
Impact

Figure 2. CHTC-delivered Compute Hours by Research Domain

As a Percentage of Total Usage | Absolute Usage (millions of hours)



A research taxonomy



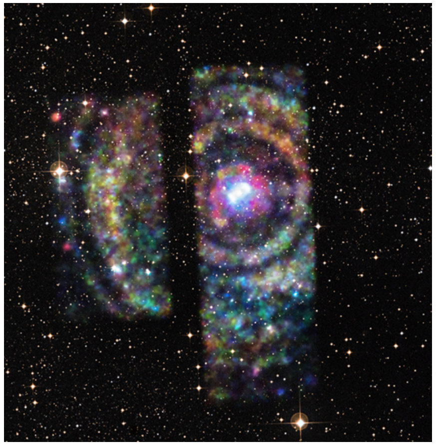
- How to categorize common computational problems in research?
- Think about the "shape" of a research problem:
  - How many "pieces"?
  - Dependent or independent processes?
  - What kind of input/output?

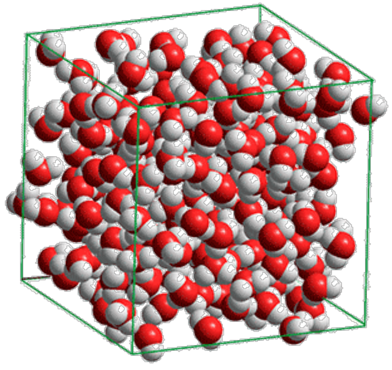
Overview

- Problems
  - Particle Simulation
  - Aggregation
  - Optimization
  - Data I: Analysis
  - Data II: Generation
- Solutions
  - high performance computing
  - high throughput computing
  - large memory

Particle Simulation

- **Problem**:
  - Model behavior of many particles in a system over time.
- **Examples**:
  - astronomy, engineering (materials, civil, electrical, industrial, nuclear), chemistry, geosciences, physics
- **Typical solution**:
  - multi-core (multi-server) software; a typical HPC cluster



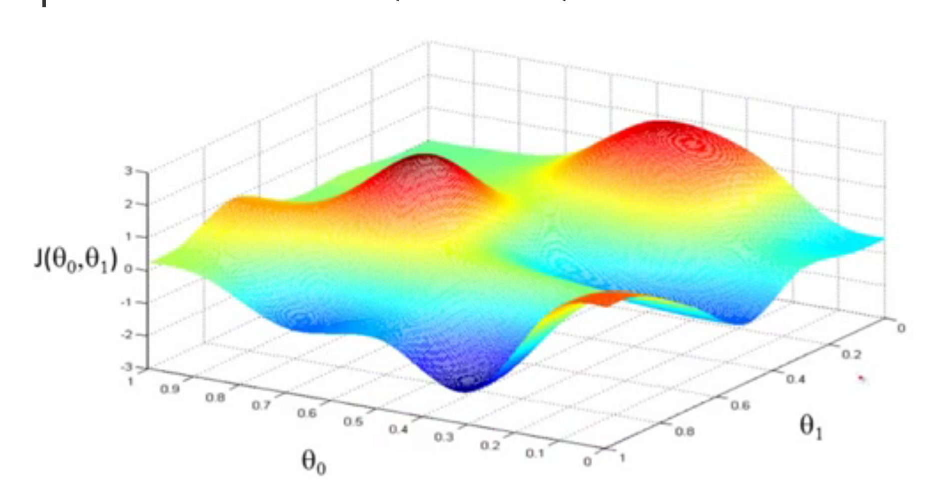


• Optimization

- **Problem:**
  - Find the best solution given a starting state and lots of unknown variables.
- **Examples**:
  - optimization, genetic algorithms, Monte Carlo. machine learning
  - economics, psychiatry, computer science, math, stats
- **Solution**:
  - varies: multi-core software,

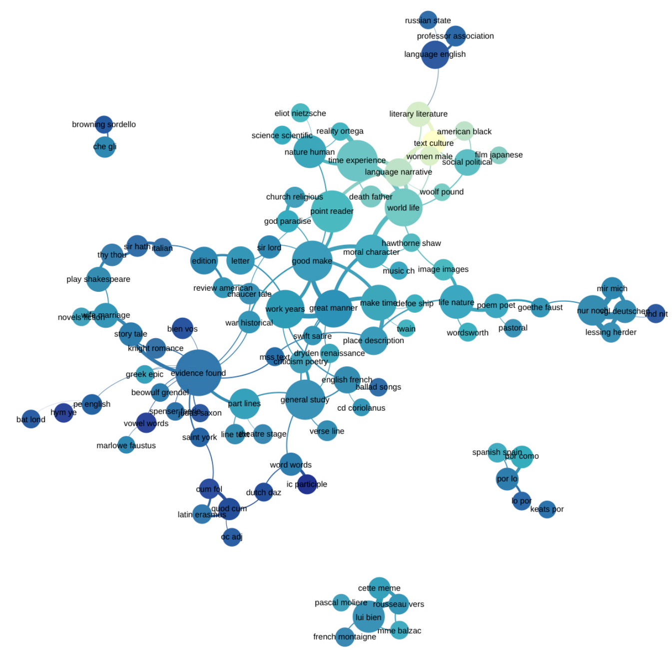
  high throughput workflow,

  GPUs



• Aggregation

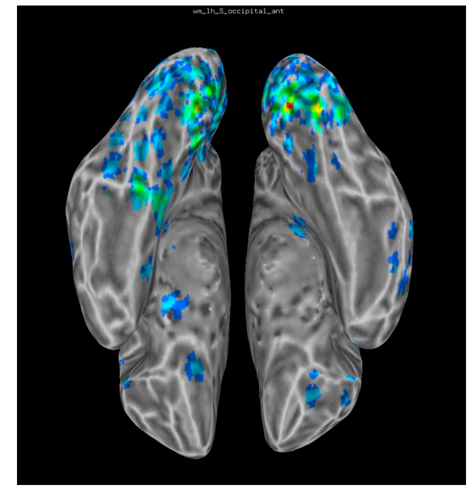
- **Problem**:
  - Take a large amount of raw data and form some sort of summary: either aggregated or differentiated.
- **Examples**:
  - genome assembly, phylogenetic trees, topic modeling
  - genetics, biostatistics, statistics, education policy, geosciences, pharmacy

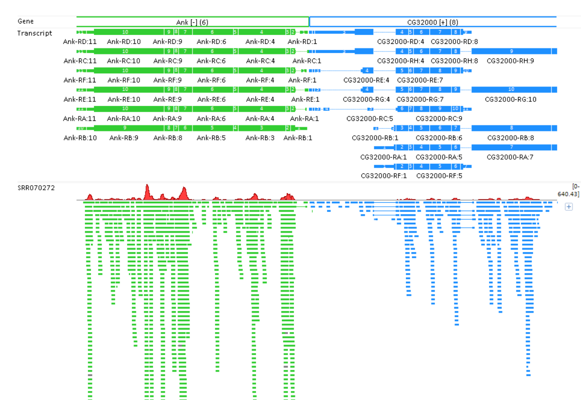


- **Solution**:
  - varies, often requires a high amount of memory

• Data I: Analysis

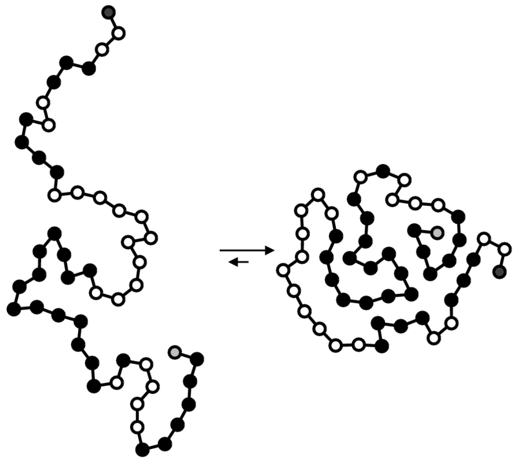
- **Problem**:
  - Analyze many independent pieces of data.
- **Examples**:
  - image analysis (e.g. fMRI), genetic data, climate/hydrological models
  - pyschology/psychiatry, genetics, forestry, engineering, zoology, animal sciences, biochemistry, botany
- **Solution**:
  - running many independent jobs - high throughput computing





Data II: Generation

- **Problem**:
  - Generate data at a large scale for further analysis.
- **Examples**:
  - parameter sweeps, Monte Carlo methods, protein folding/docking
  - economics, statistics, engineering, drug discovery, biochemistry
- **Solution**:
  - running many independent jobs - high throughput computing

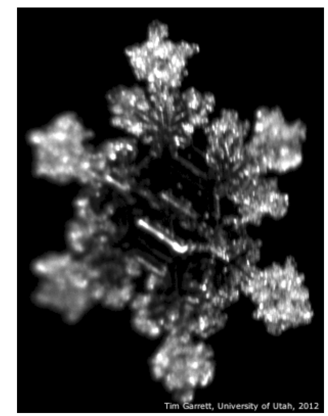


Overview

- Problems
  - Particle Simulation
  - Aggregation
  - Optimization
  - Data I: Analysis
  - Data II: Generation
- Solutions
  - high performance computing
  - high throughput computing
  - large memory

Caveats

- It can be useful to think of research problems in these broad categories.
  - Facilitate immediate understanding of problems
  - Recognize need for diverse solutions
- However, each research problem is unique.
  - Treat each problem individually
  - Fully understand problem first, then seek solution
  - Try new things
- Complement technical solutions with human assistance.



Human solutions

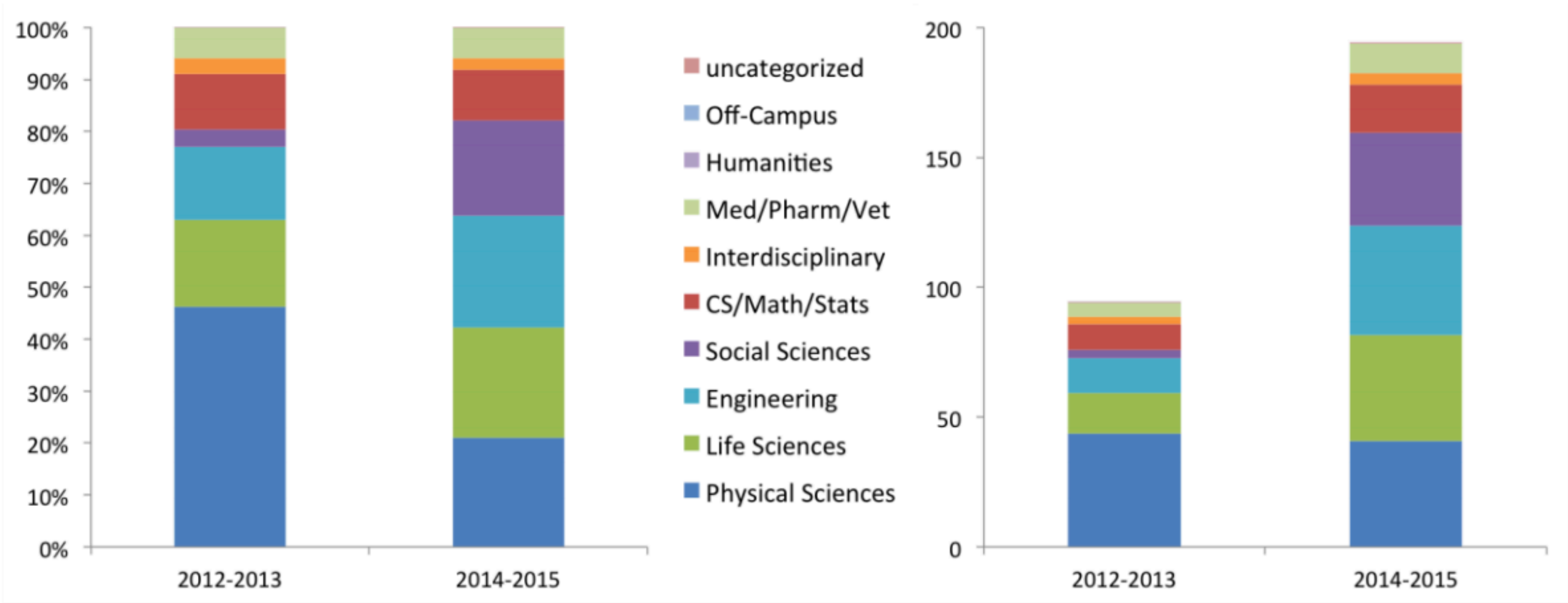
- **Matchmaking**
  - Identify researcher problems and match them to solutions [including people]
  - Bring together people with the same problem
- **Training and support**
  - Help researchers implement appropriate solutions
- **Advocacy**
  - Communicate common problems to computational experts who can provide solutions

Impact

Figure 2. CHTC-delivered Compute Hours by Research Domain

As a Percentage of Total Usage

Absolute Usage (millions of hours)

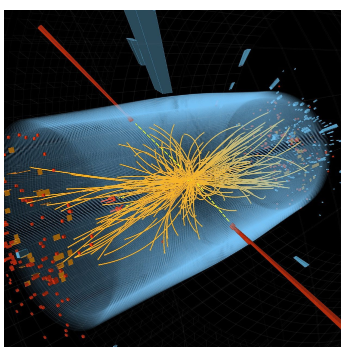


Impact

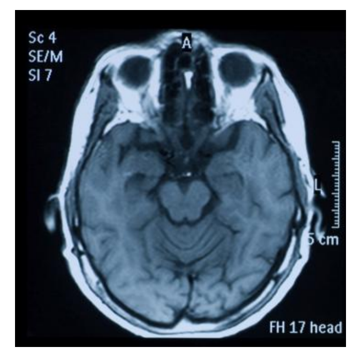
- Significant increases over two years in usage from researchers in:
  - Life Sciences (from 17% to 21% of total usage)
  - Social Sciences (from 3% to 18%).
- Roughly 95% of CHTC-delivered usage to these groups (including Open Science Grid hours) has been on high throughput compute systems.
- Another 4% has been on large-memory machines.

Summary

- Learn something about your users
  - Identify common, basic problems
  - Their problems → appropriate solutions
- Appreciate particularity
- Include non-technical, human solutions
- Watch your compute hours increase and diversify!



$$
\begin{aligned}
\textbf{max.} \quad & c^T x \\
\textbf{s.t.} \quad & Ax \leq b \\
& x \geq 0
\end{aligned}
$$



Questions?