Open Science Grid

HTCondor: BLAST

Monday, 3:30pm

Zach Miller <zmiller@cs.wisc.edu>

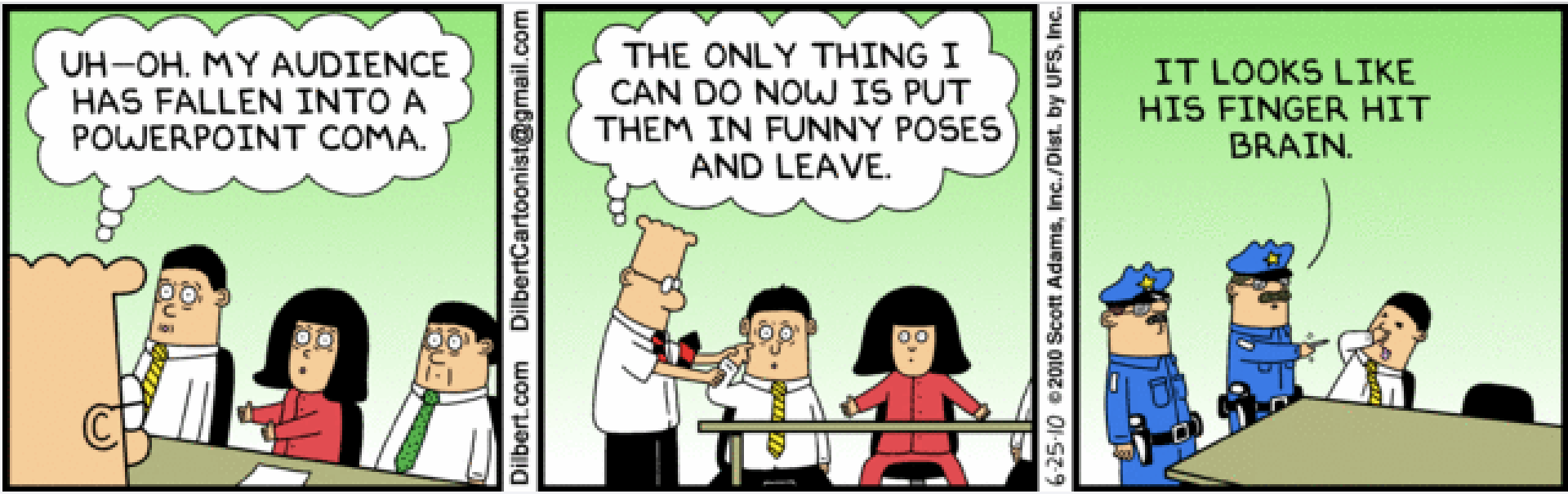
Flightworthy Team

University of Wisconsin-Madison

Before we begin…

- Any questions on the lectures or exercises up to this point?

I hope you're not getting too tired



Open Science Grid

BLAST

- Up to now, you've done toy examples
  - Simple, easy to use
  - Illustrate basics of what you need to know
  - The Mandlebrot set is cool… but a toy
- Let's try out a real application: BLAST
  - More complex, not so easy to use

First, some honesty

- I am a computer scientist
- I am not a biologist
- My knowledge of BLAST is shallow
- But it's way cooler application than what we've done so far!

BLAST Description

From the BLAST web page:

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Blast Description (My understanding)

- Biologists have sequences:
  - Nucleotides in DNA: ACGTTGCA…
  - Amino acids in proteins: GECVASR…
- They also have databases of lots of sequences
  - From lots of organisms, from tiny bacteria to humans
- BLAST helps them answer questions:
    - Which bacterial species have a protein that is related in lineage to another protein?
    - What other genes encode proteins that exhibit structures or motifs such as ones that have just been determined?
    - …
- BLAST is widely used and considered important

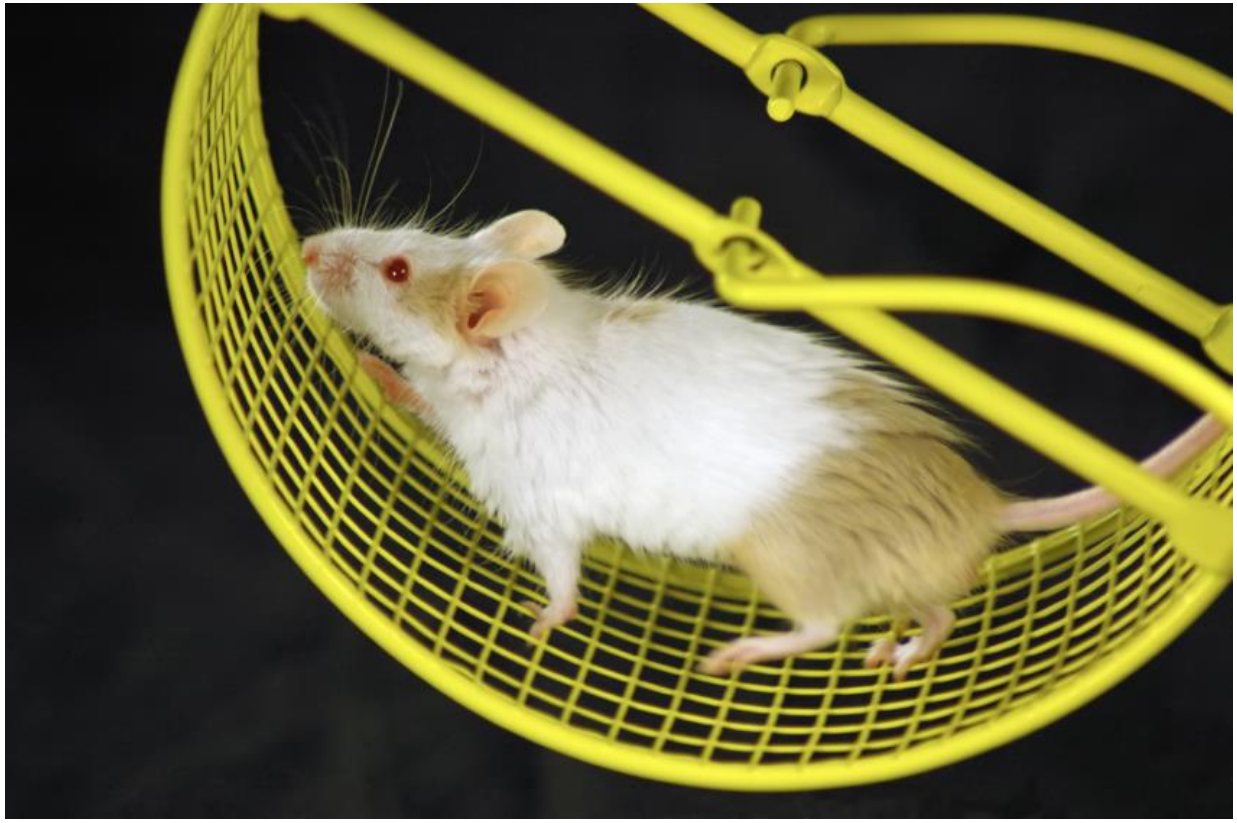
Is this just string comparison?

- It's harder than just comparing two strings: Is "GCTA == GCTA"?
- BLAST can find "similar" sequences, based on metrics that biologists determine.
  - "Similar" means this is more computationally expensive than just string comparison
- BLAST is a very popular program to ask these questions

BLAST exercise

- The final set of exercises have you run queries with BLAST
- They are a bit arbitrary, because I know less about the underlying biology
- But it's a real application with real data!
- Your challenge: run a bunch of BLAST queries and summarize the results. Do it all within a DAG

Time to try it out!



Questions?

- Questions? Comments?
- Feel free to ask me questions later:

  Zach Miller <zmiller@cs.wisc.edu>
- Upcoming sessions
  - Now – 4:30: HTC in action (Live!)
  - 4:30pm – 5:30pm
    - Hands-on exercises
    - Finish up earlier exercises
    - Try out BLAST
  - 5:30 – 7:00: Dinner, on your own
  - 7:00 – 9:00: **Optional** evening work session