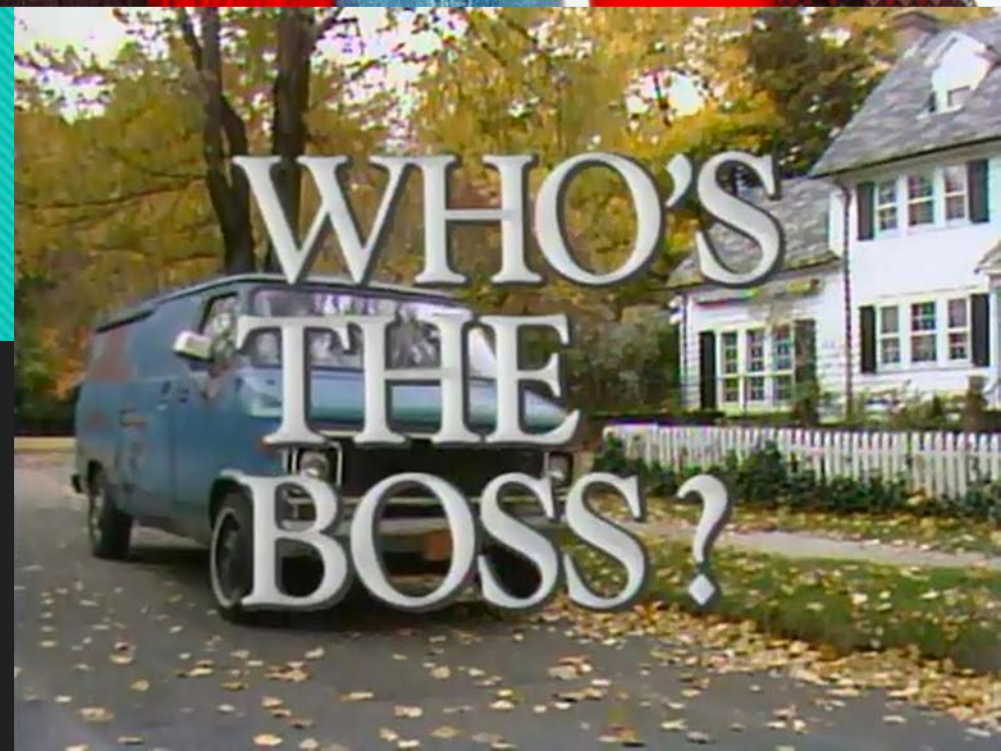
Jamie Rosner
Research Specialist, UBC ARC
CC Bioinformatics National Team, Chair

Bioinformatics vs. IT

Who's the Boss?

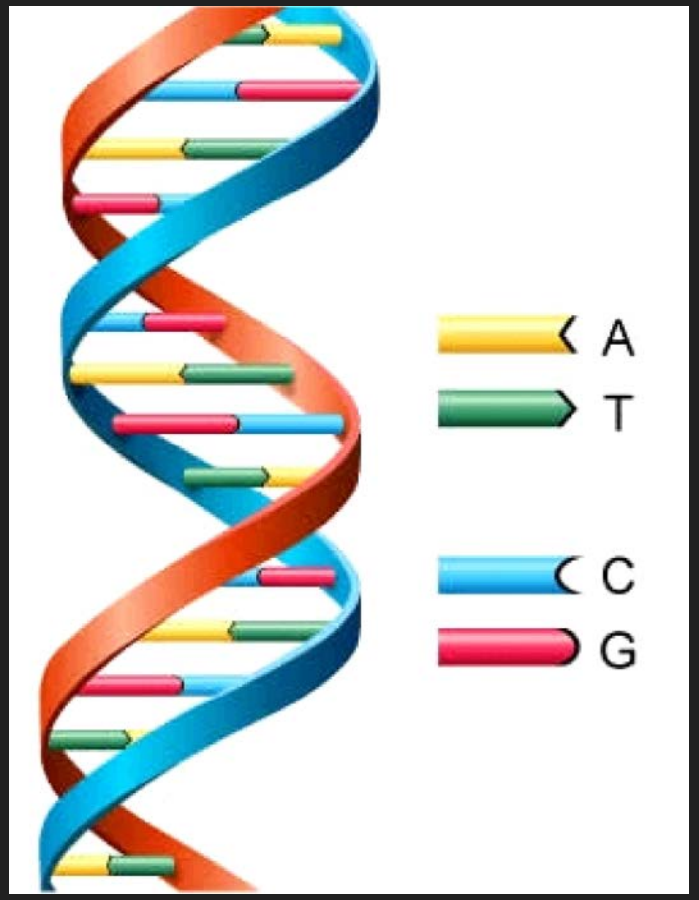
Bioinformatics vs. IT



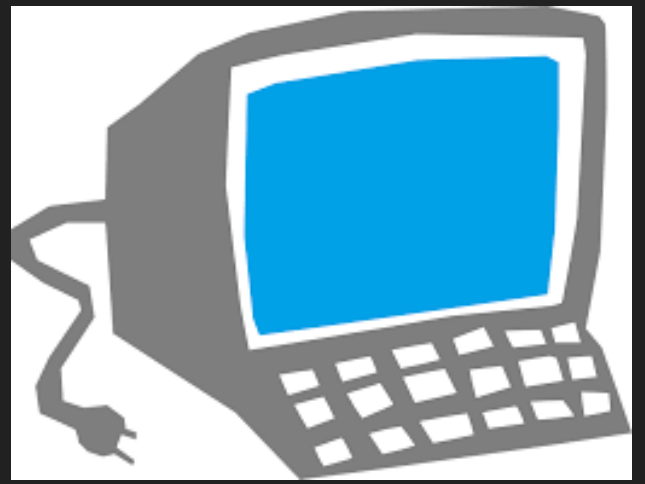
Outline

- What is Bioinformatics?
- What are they trying to do?
- So what's the problem?
- Some solutions

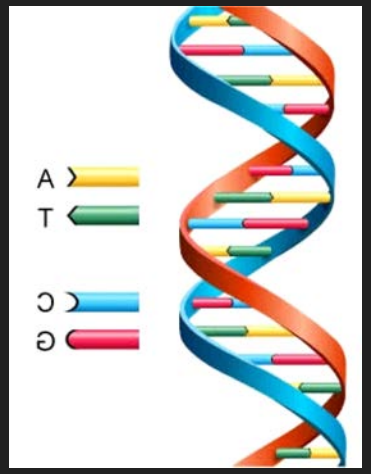
What is Bioinformatics?



+

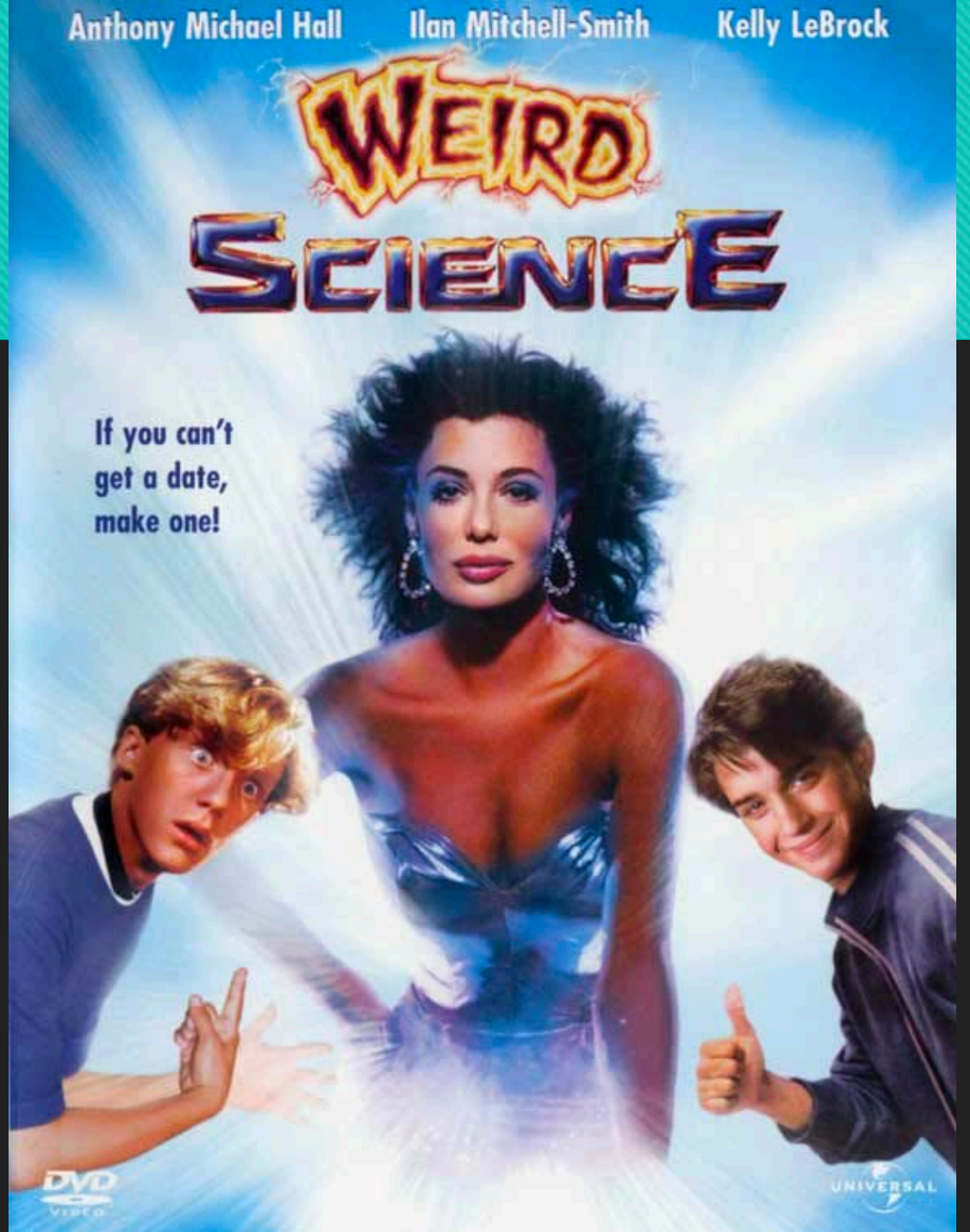


What is Bioinformatics?

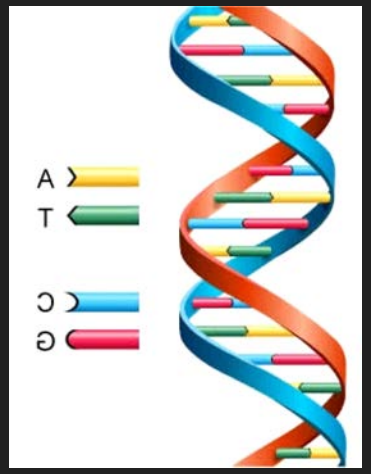


+

=



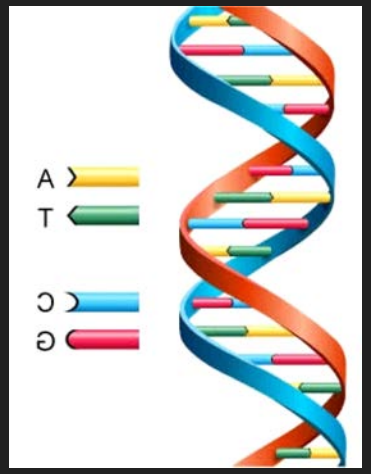
What is Bioinformatics?



+

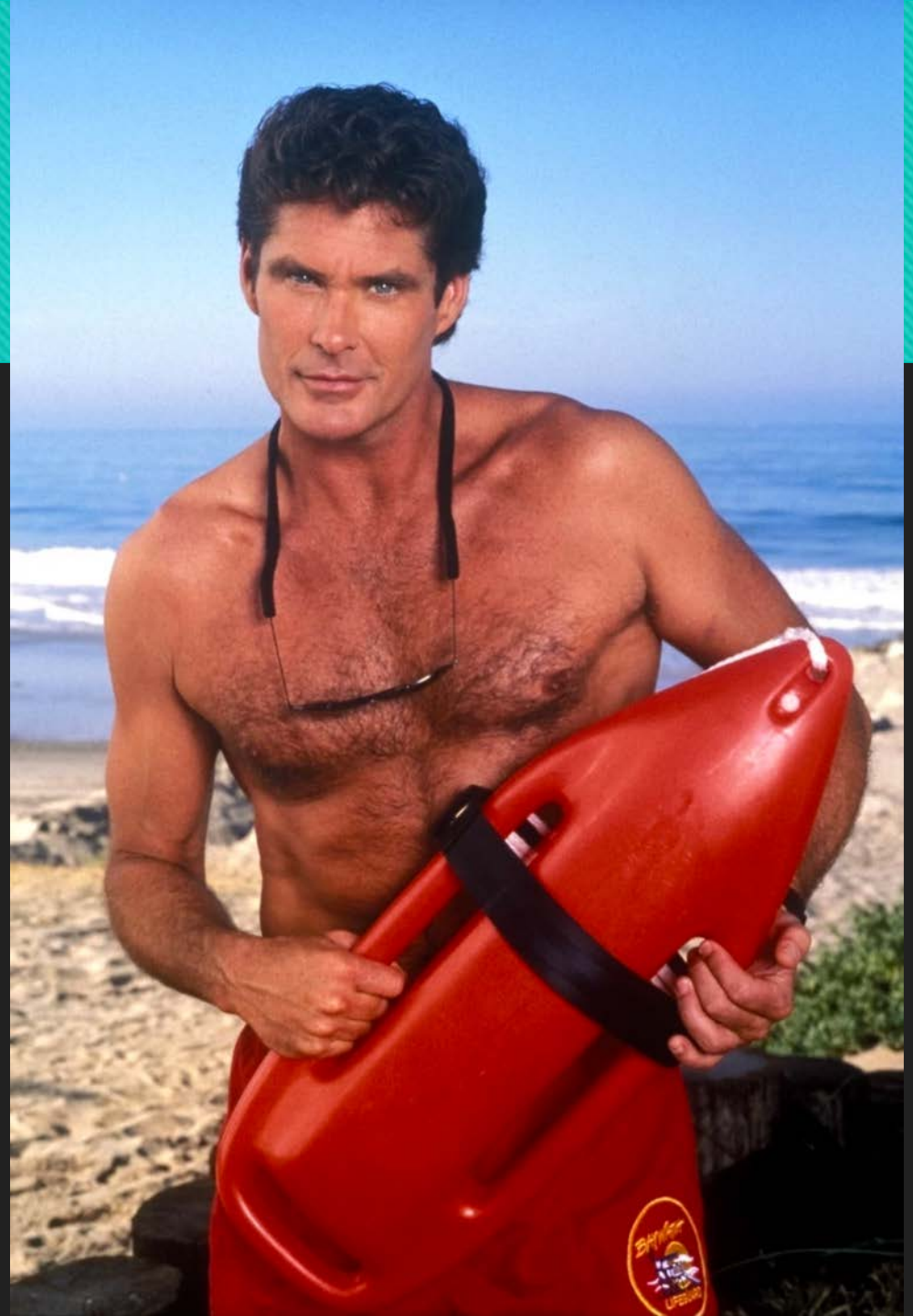
=

What is Bioinformatics?

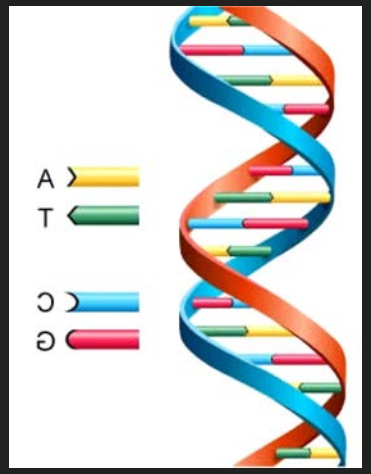


+

=



What is Bioinformatics?

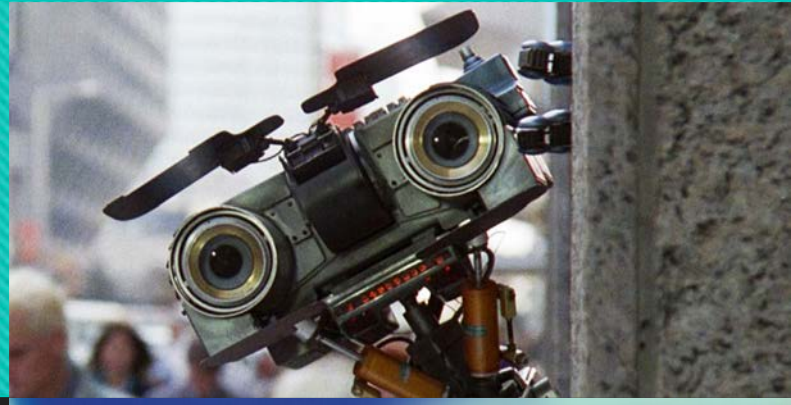


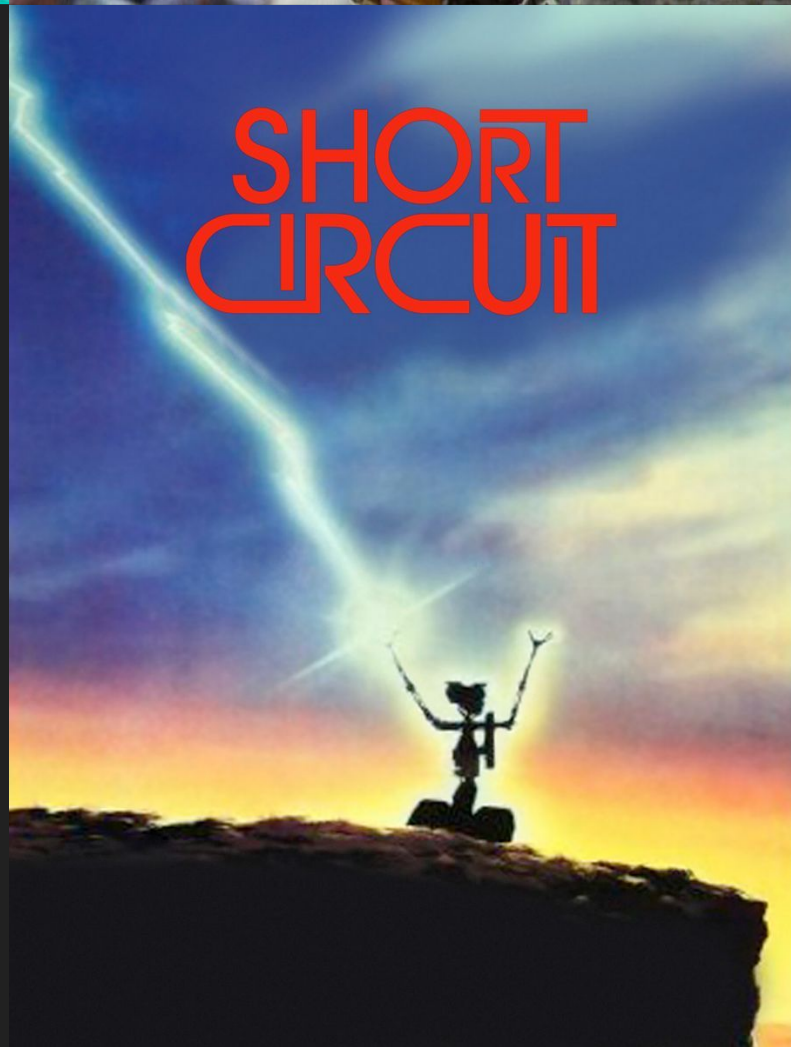
+

=

The Birth of Bioinformatics

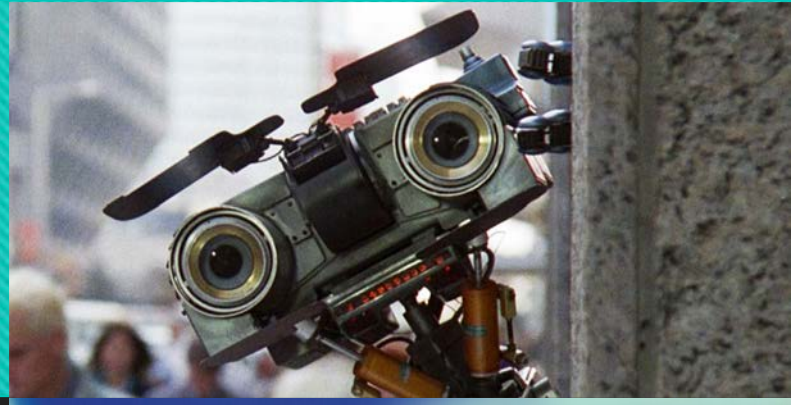
- Human Genome Project (1990-2003)

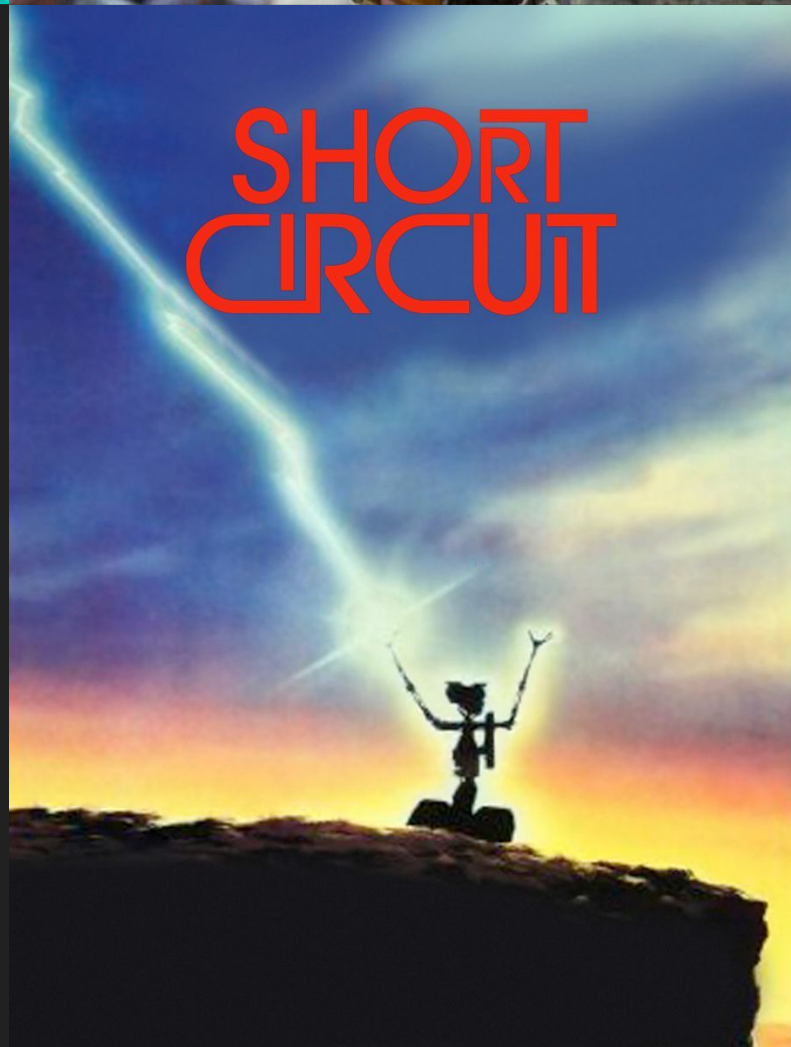




The Birth of Bioinformatics

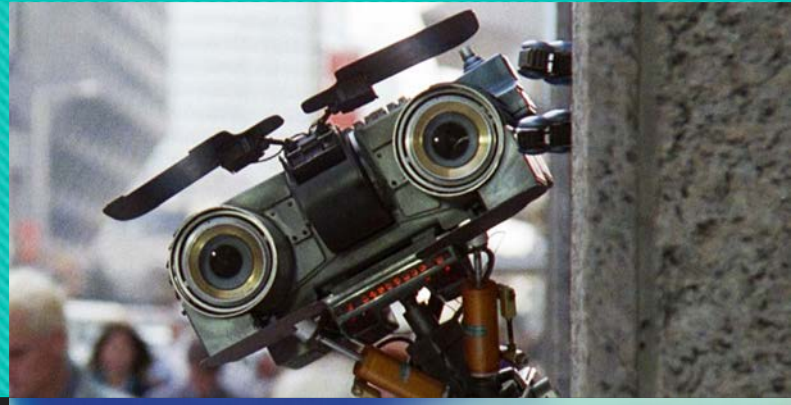
- Human Genome Project (1990-2003)
  - 1st billion bases sequenced – 4 years

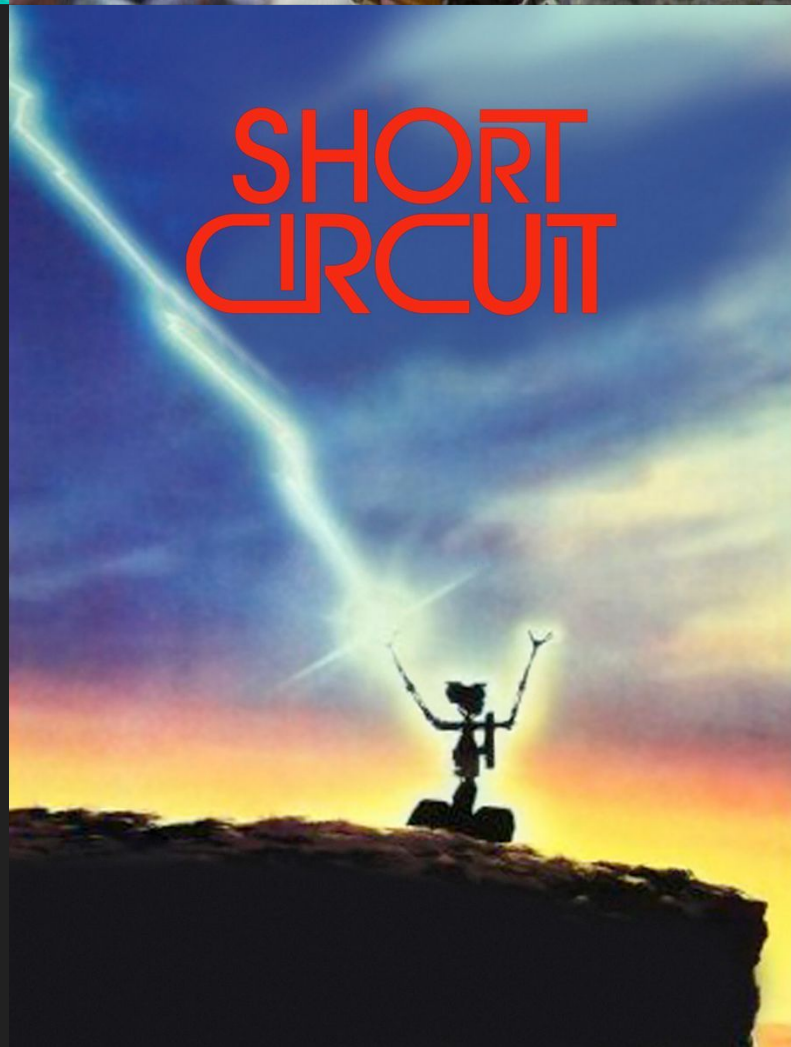




The Birth of Bioinformatics

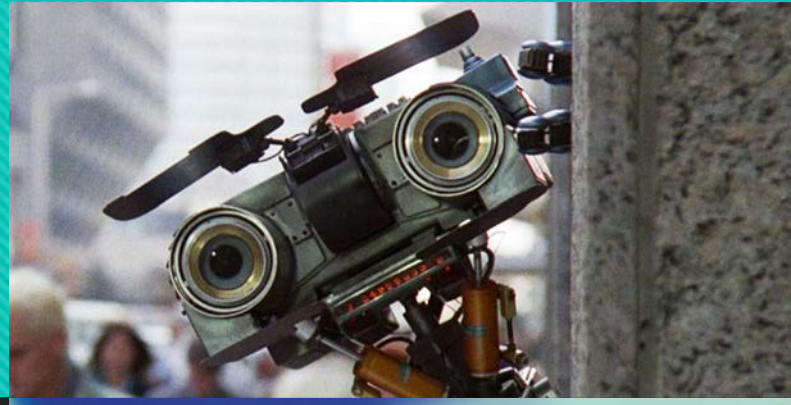
- Human Genome Project (1990-2003)
  - 1st billion bases sequenced – 4 years
  - 2nd billion bases sequenced - 4 months

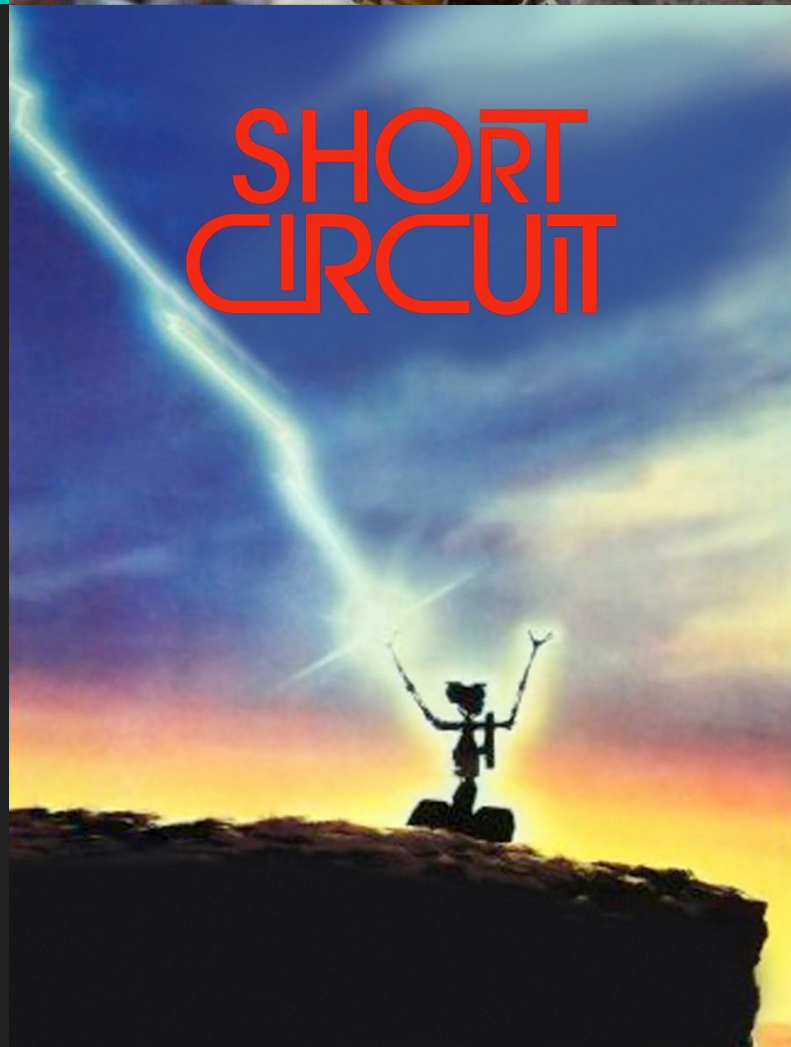




The Birth of Bioinformatics

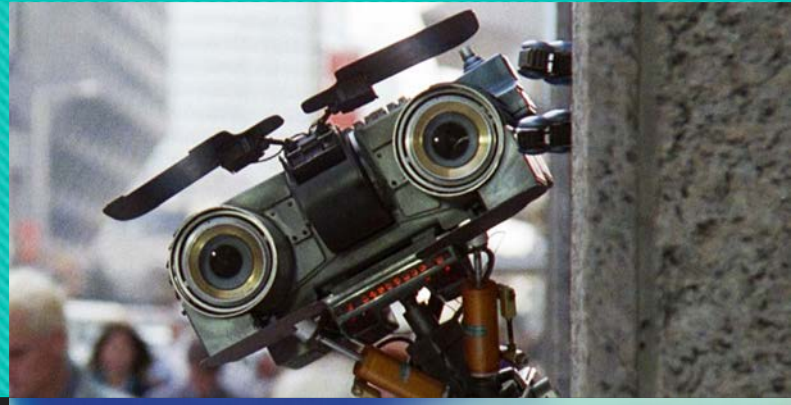
- Raw genome files → 200 – 300GB

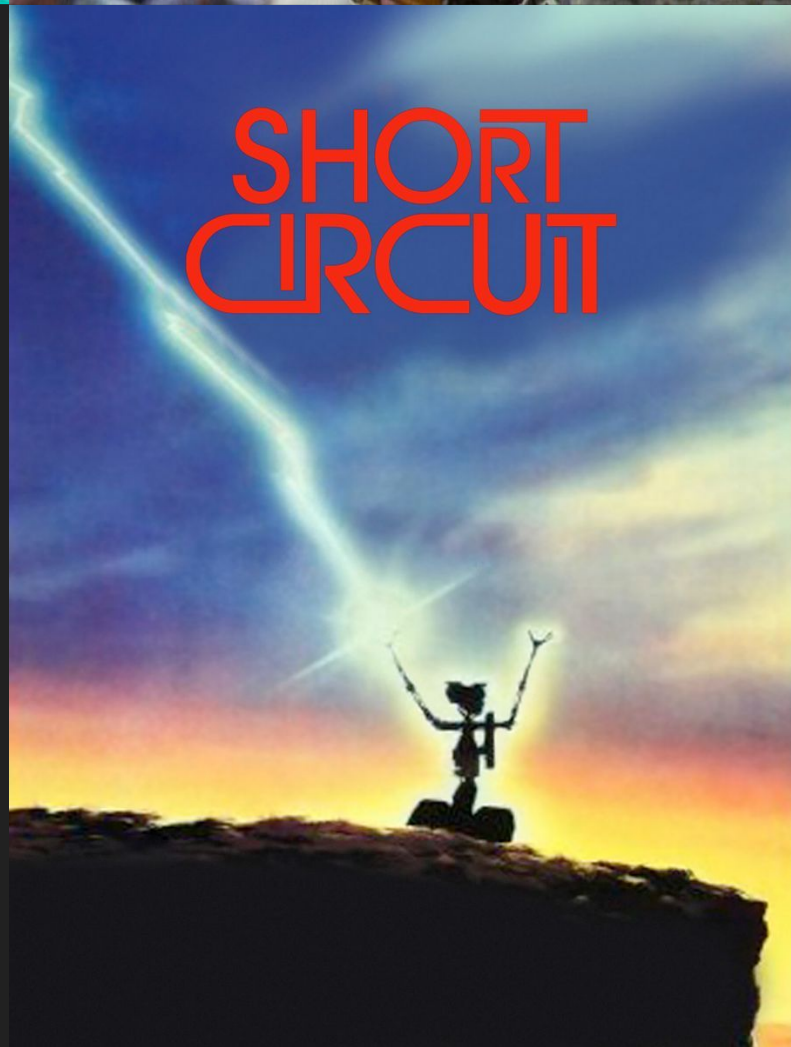




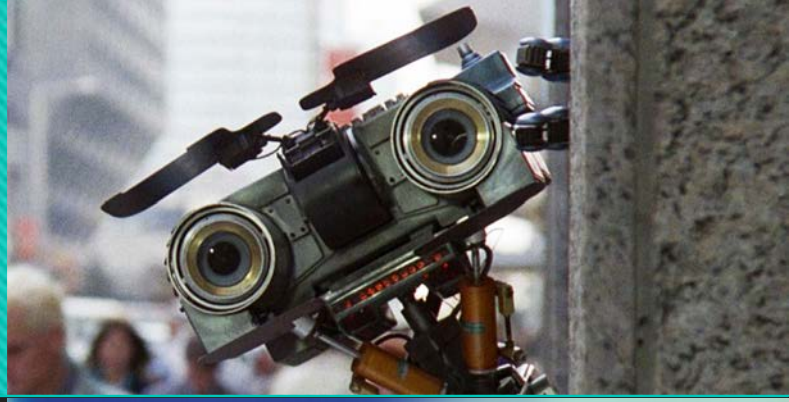
The Birth of Bioinformatics

- Raw genome files → 200 – 300GB
- Microsoft Excel <= 2GB

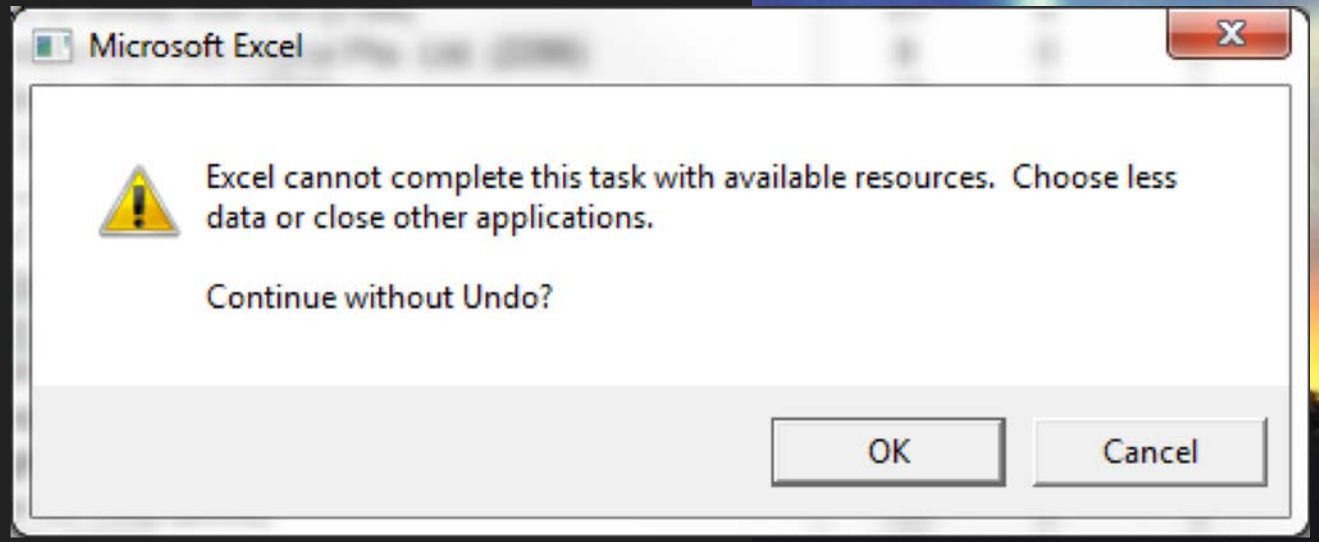




The Birth of Bioinformatics



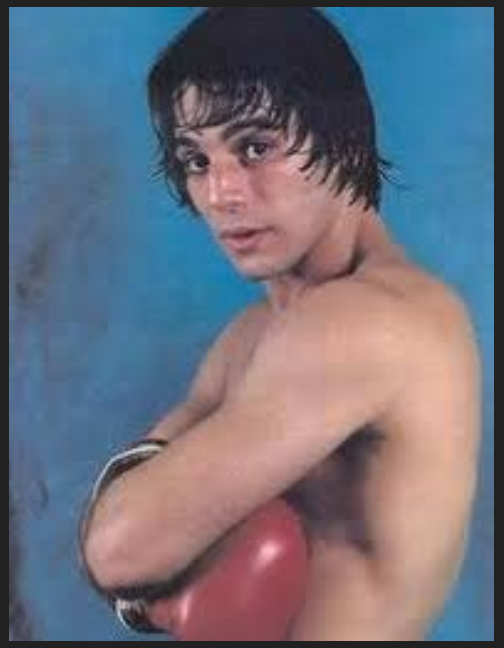
- Raw genome files → 200 – 300GB
- Microsoft Excel <= 2GB



Facts of Life

- Big data
- Advanced algorithms
- Computing power / HPC
- Command-line, coding, etc.

Bioinformaticians are like Tony



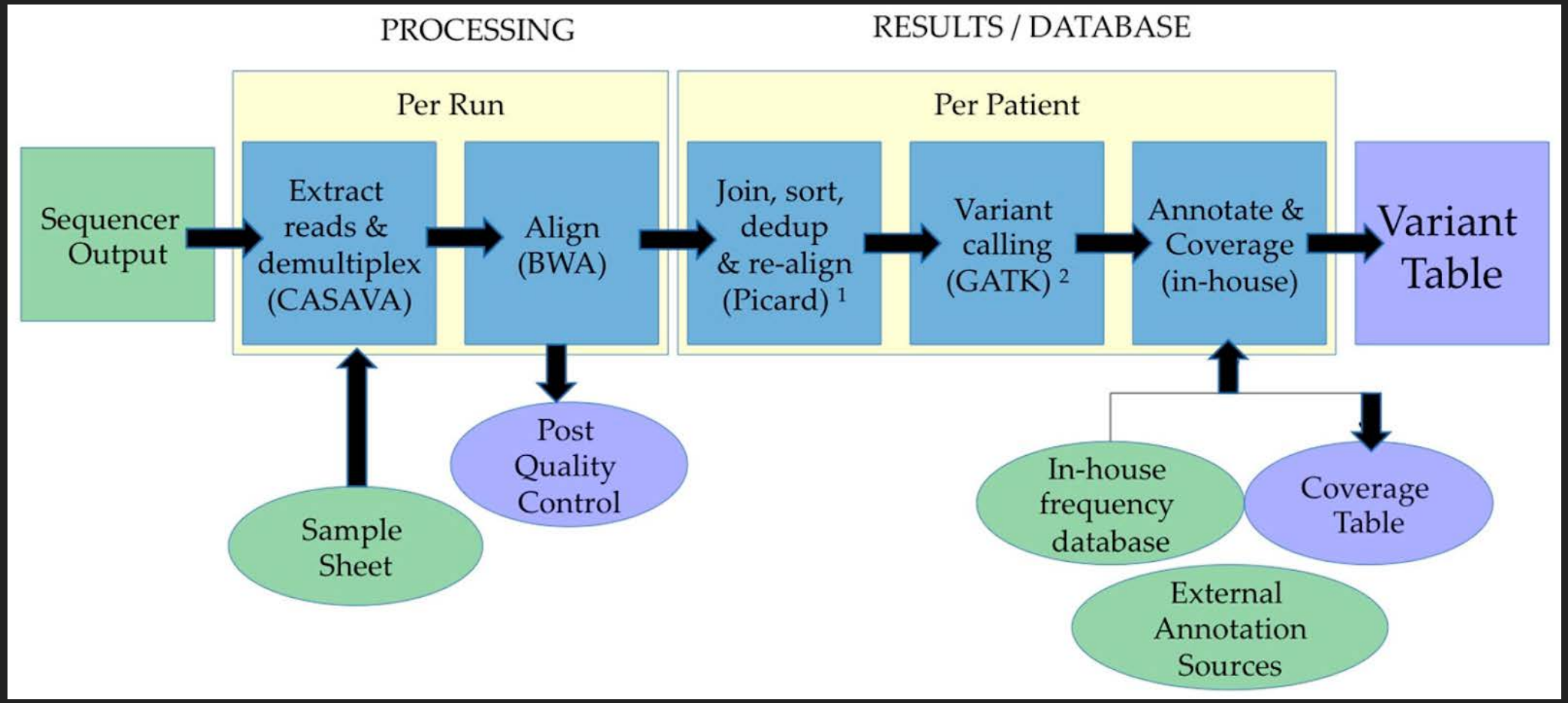
Bioinformaticians are like Tony

Bioinformaticians are like Tony

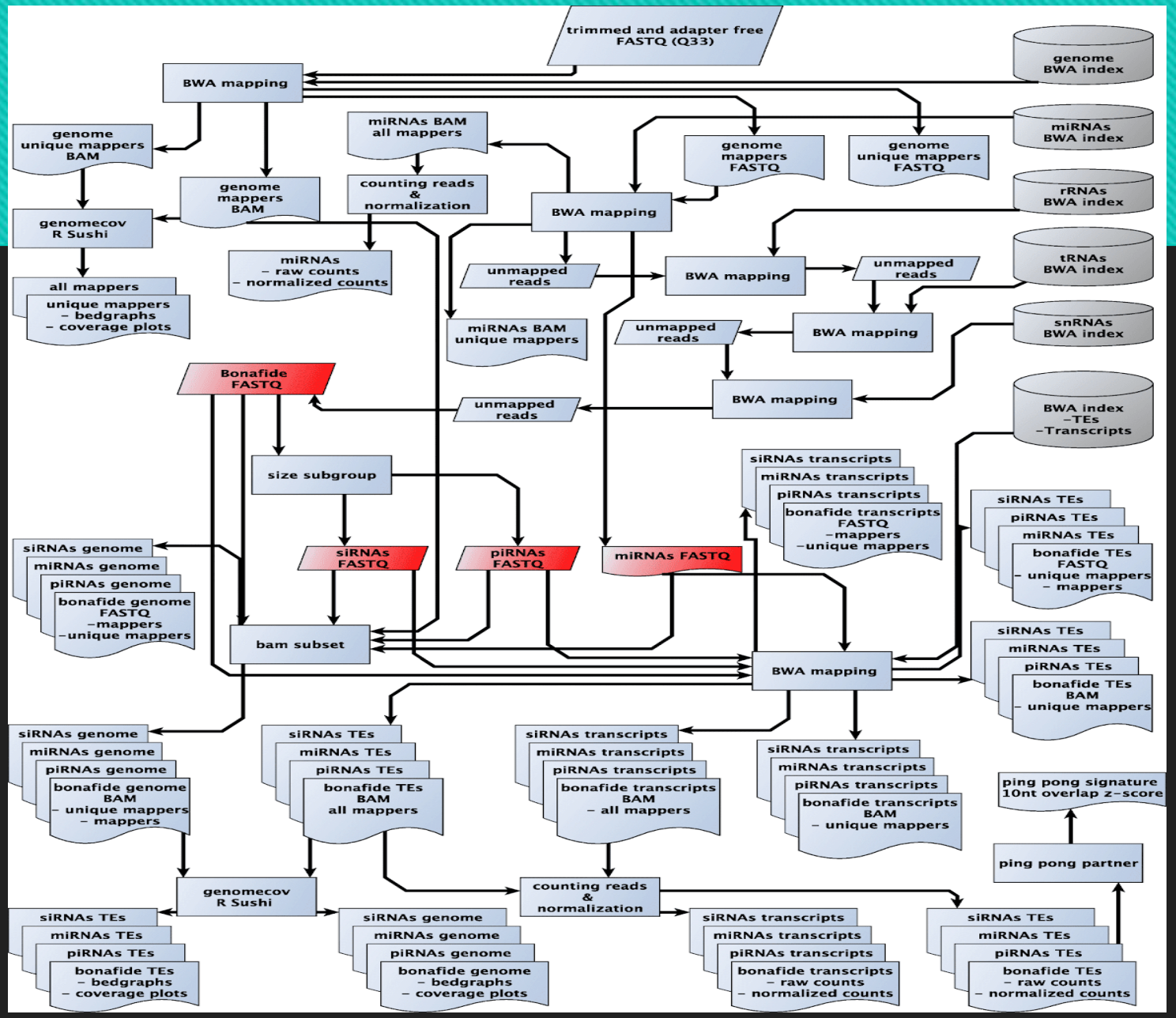
Bioinformaticians are like Tony

IT folk are like Angela

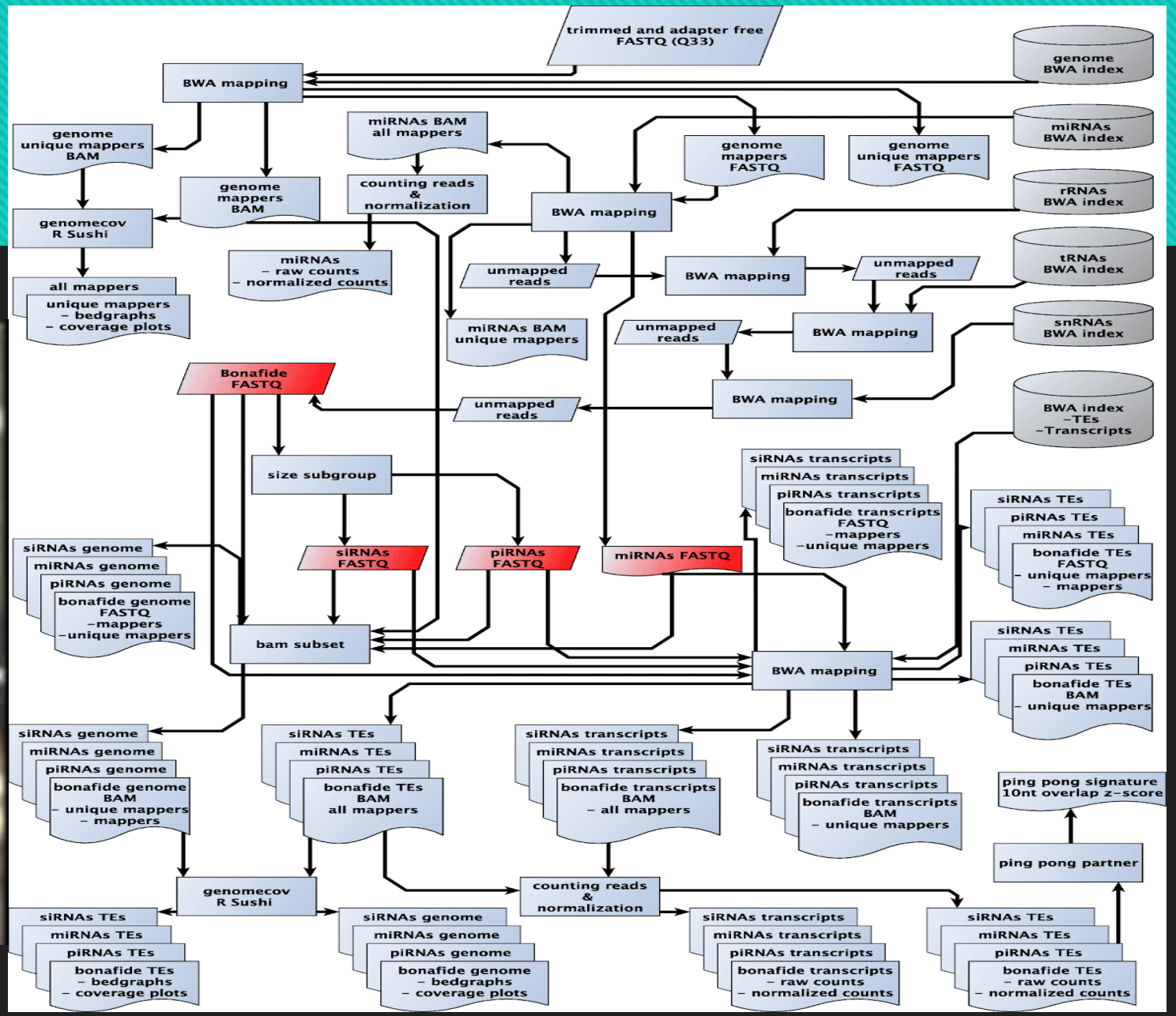
Typical Workflow



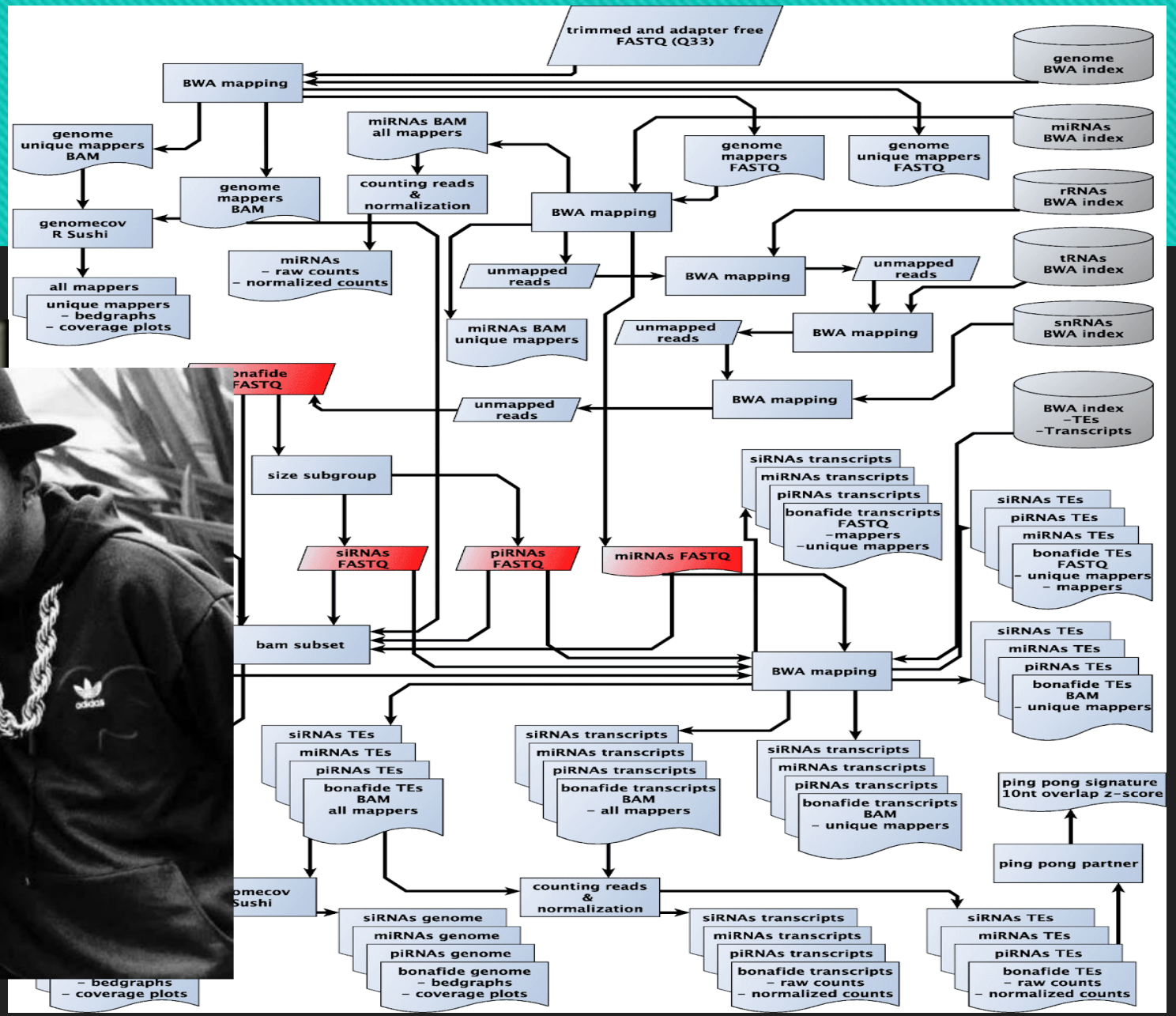
Typical workflow



Typical workflow



Typical workflow



Typical workflow

What are the problems?

- Let's have a look…

Data is Big

- 300 Gb / sample
- Projects 10s or 100s of terabytes
- Reference databases
- Public datasets
- Data transfer
- It's getting bigger

Data is Big

- 300 Gb / sample
- Projects 10s or 100s of terabytes
- Reference databases
- Public datasets
- Data transfer
- It's getting bigger

…Solutions?

Novice users

- Students

Novice users

- Students
- Poorly trained

Novice users

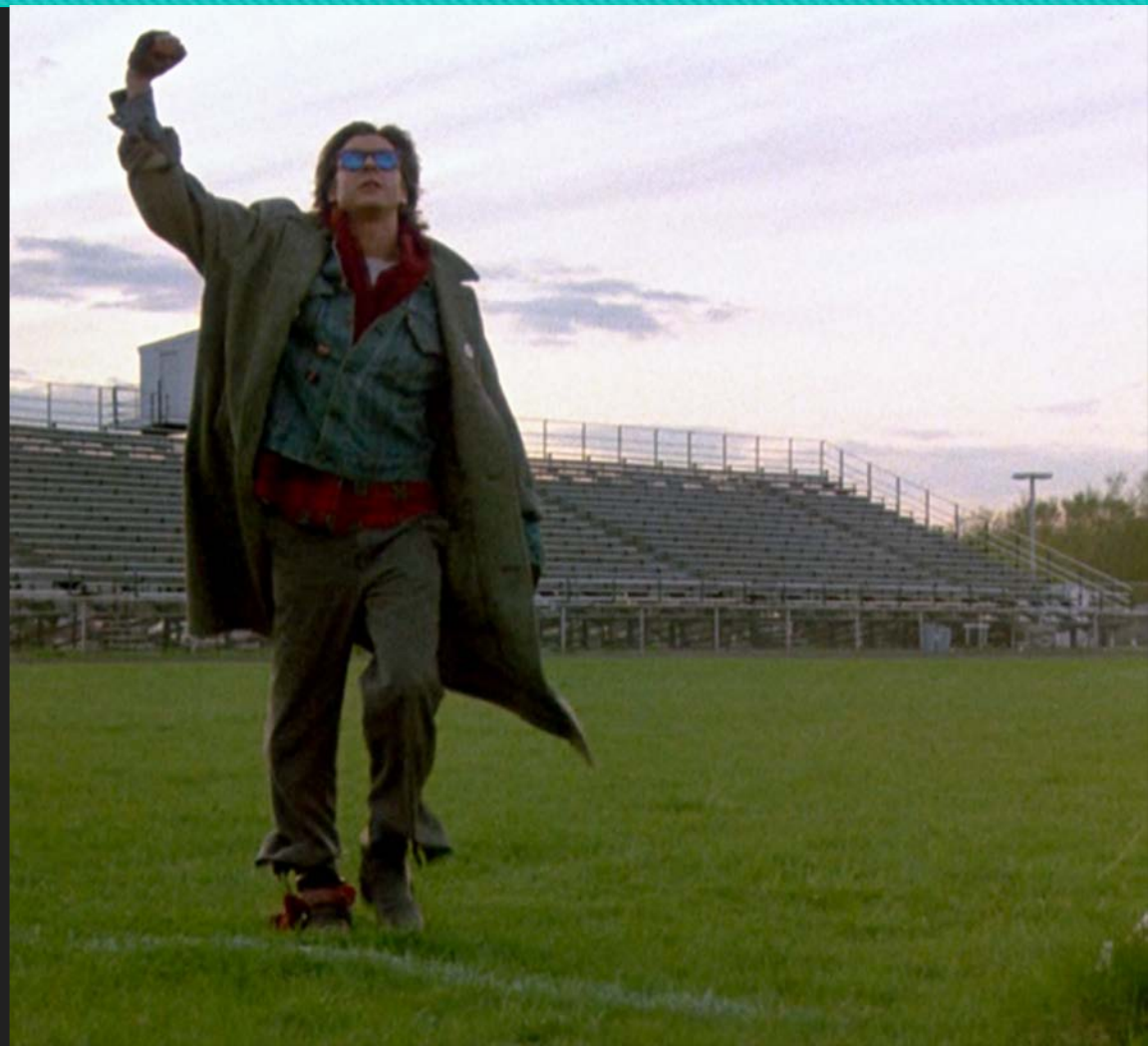
- Students
- Poorly trained
- Often little guidance

Inexperienced users + bad code

What's the worst
that could happen?

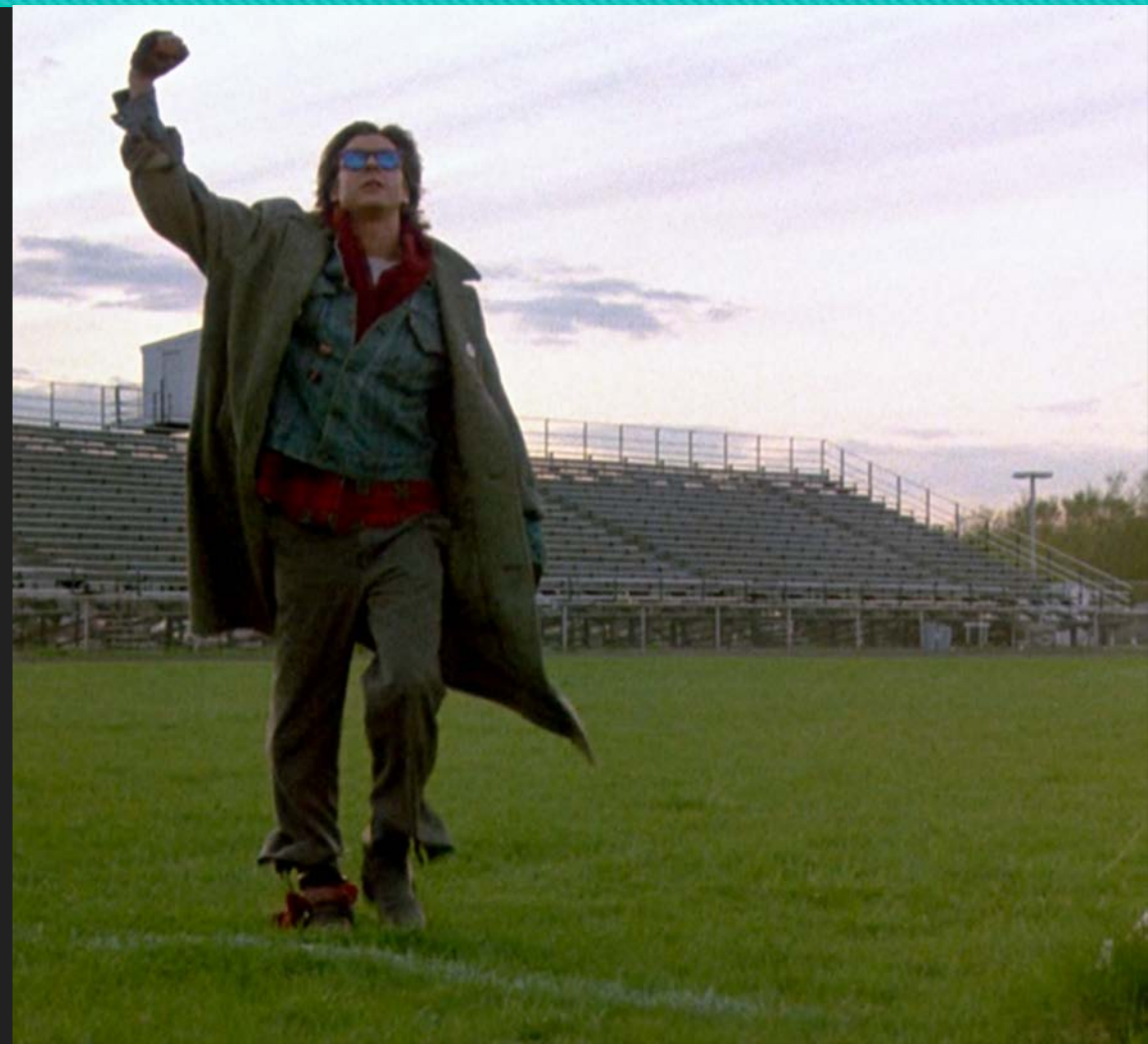
Novice users

Solutions?



Novice users

- Training
  - Software Carpentry
  - Compute Canada Webinars
  - UBC Research Computing Summer School
  - Bash / Python interactive online tutorials
  - Canadian Bioinformatics Helpdesk
    → bioinformatics.computecanada.ca



Lots of files

- Bioinformatics pipelines generate many many files

Lots of files

Where X is large, # of files = X

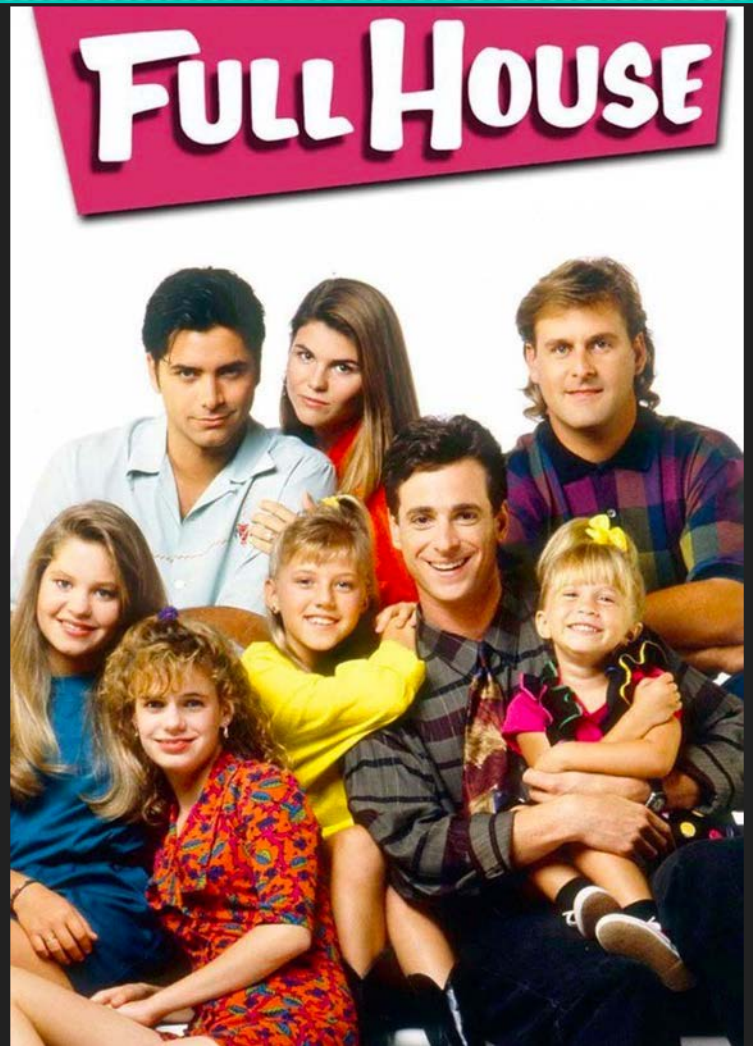
Lots of files

Where X is large, # of files = X

Lots of files

Where X is large, # of files = X

- Full house



Lots of files

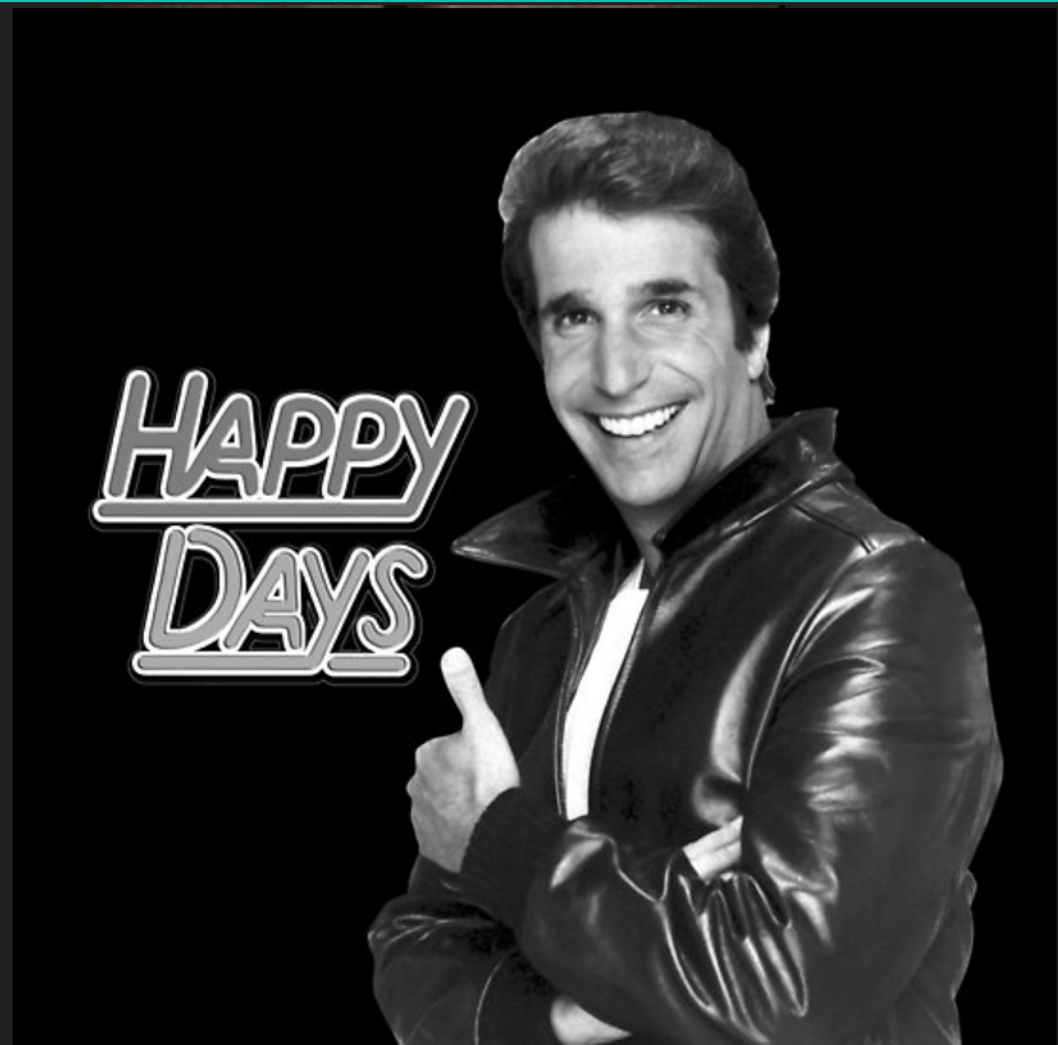
Where X is large, # of files = X

- Full house
- Eight is enough

Lots of files

Where X is large, # of files = X

- Full house
- Eight is enough
- Just delete them all…



Lots of files

- Example – MISO software can generate over 100,000 files per sample

Lots of files

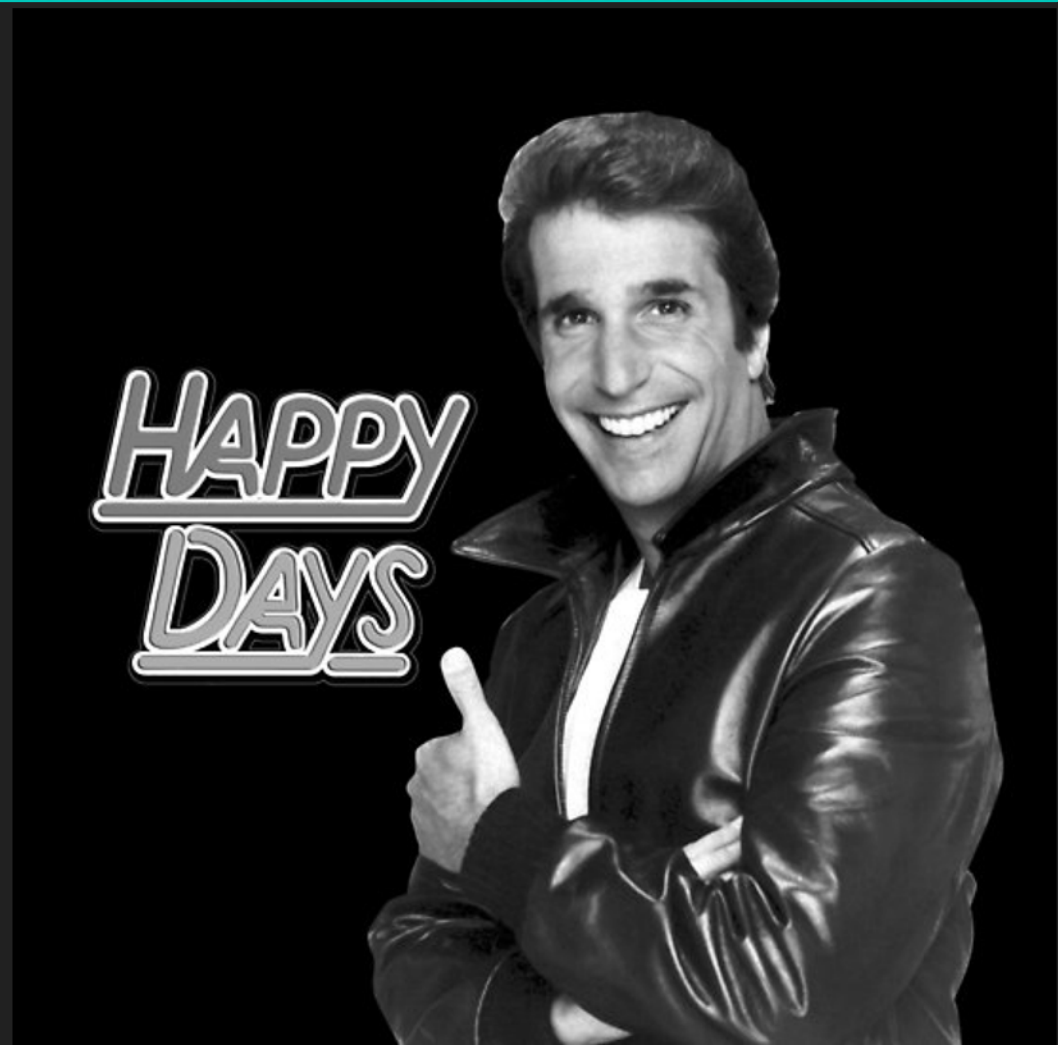
- No more file creation…

Lots of files

- Solution?
  - tar or zip files
  - “Low inode” warning

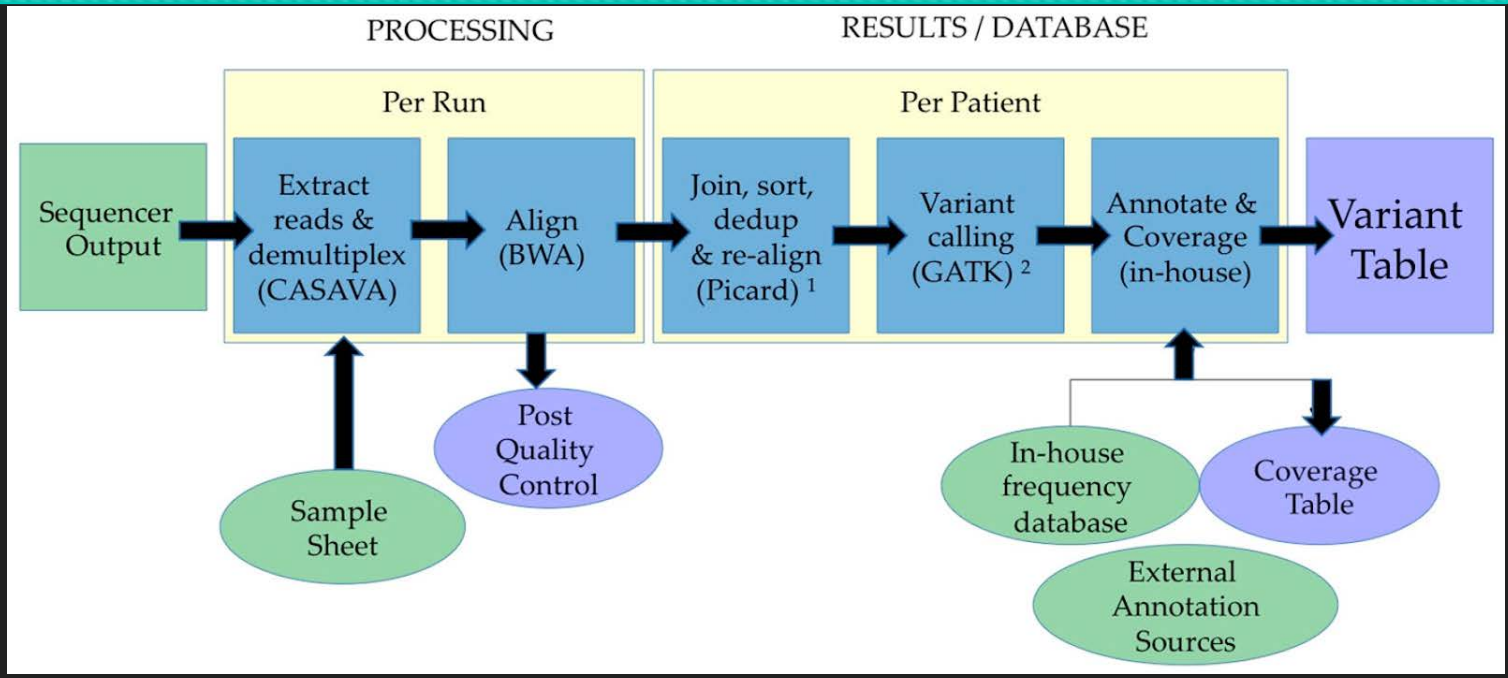
Lots of files

- ;)



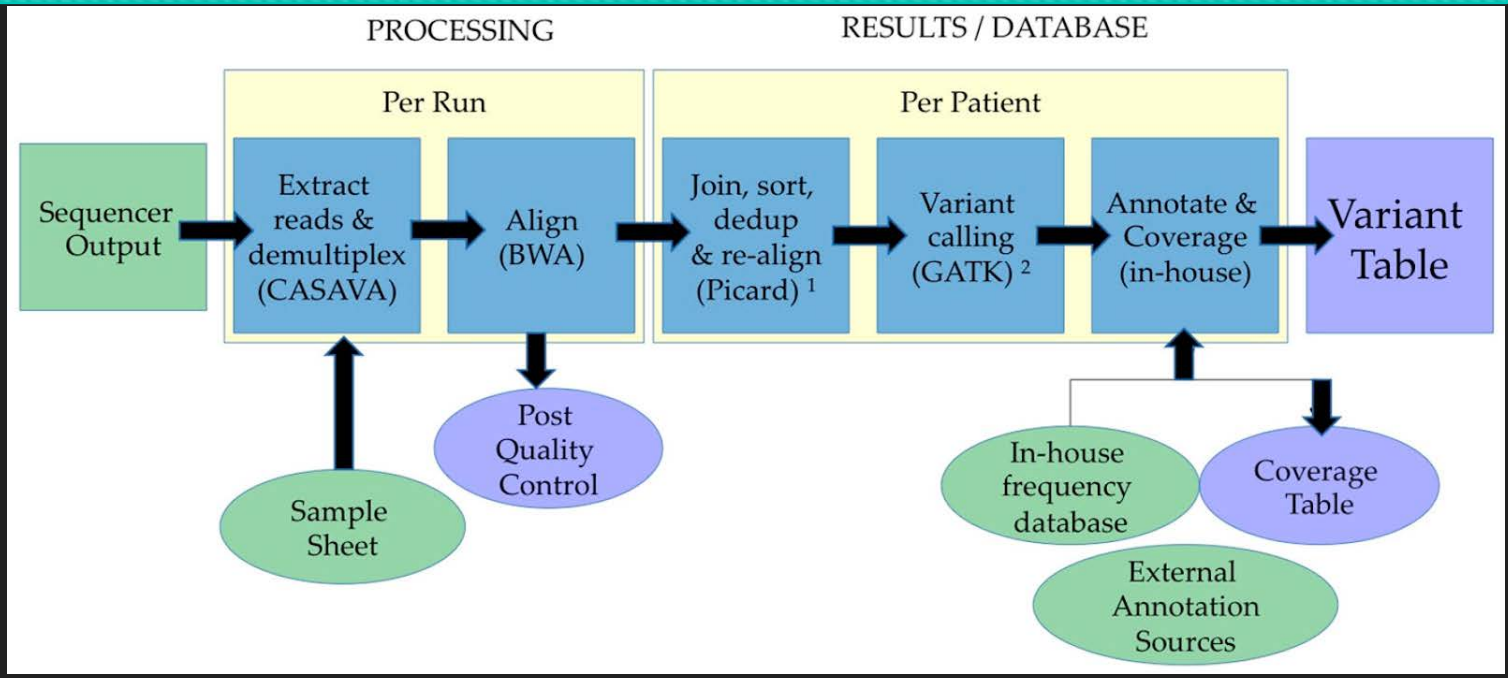
Poorly suited pipelines

- Not well suited for general purpose HPC systems
- I/O bound processes
- Many large memory/CPU jobs
- Complex software environments



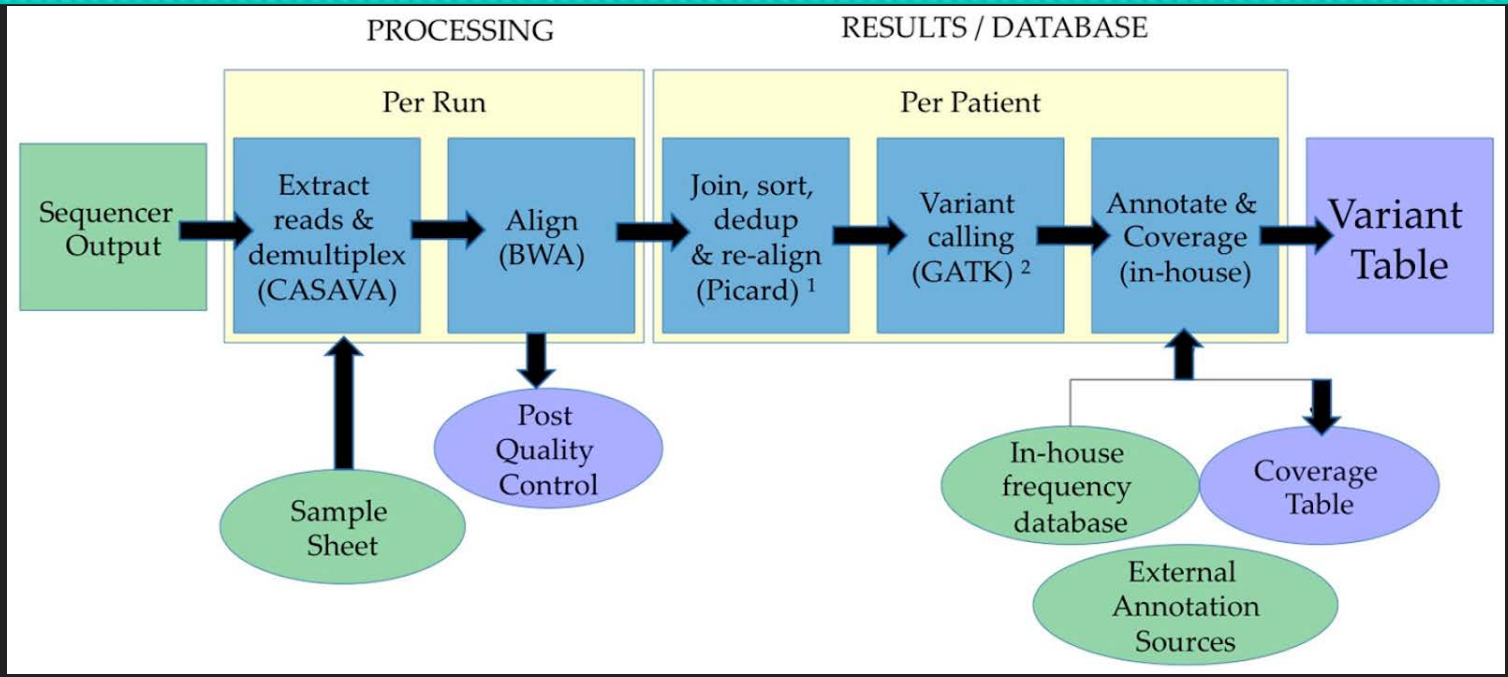
Poorly suited pipelines

- Solutions?



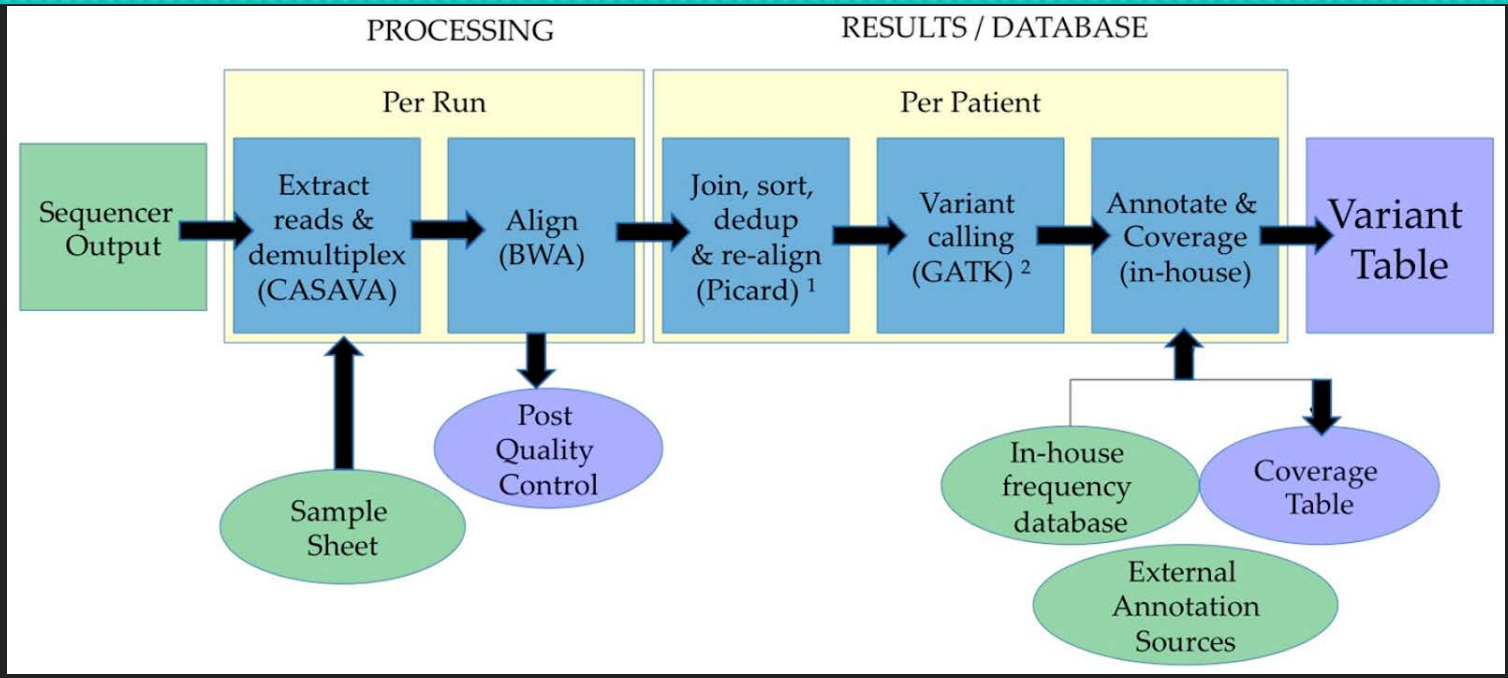
Poorly suited pipelines

- I/O Bound
  - Isilon over Lustre
  - Hadoop?
  - Local /tmp on a thin node



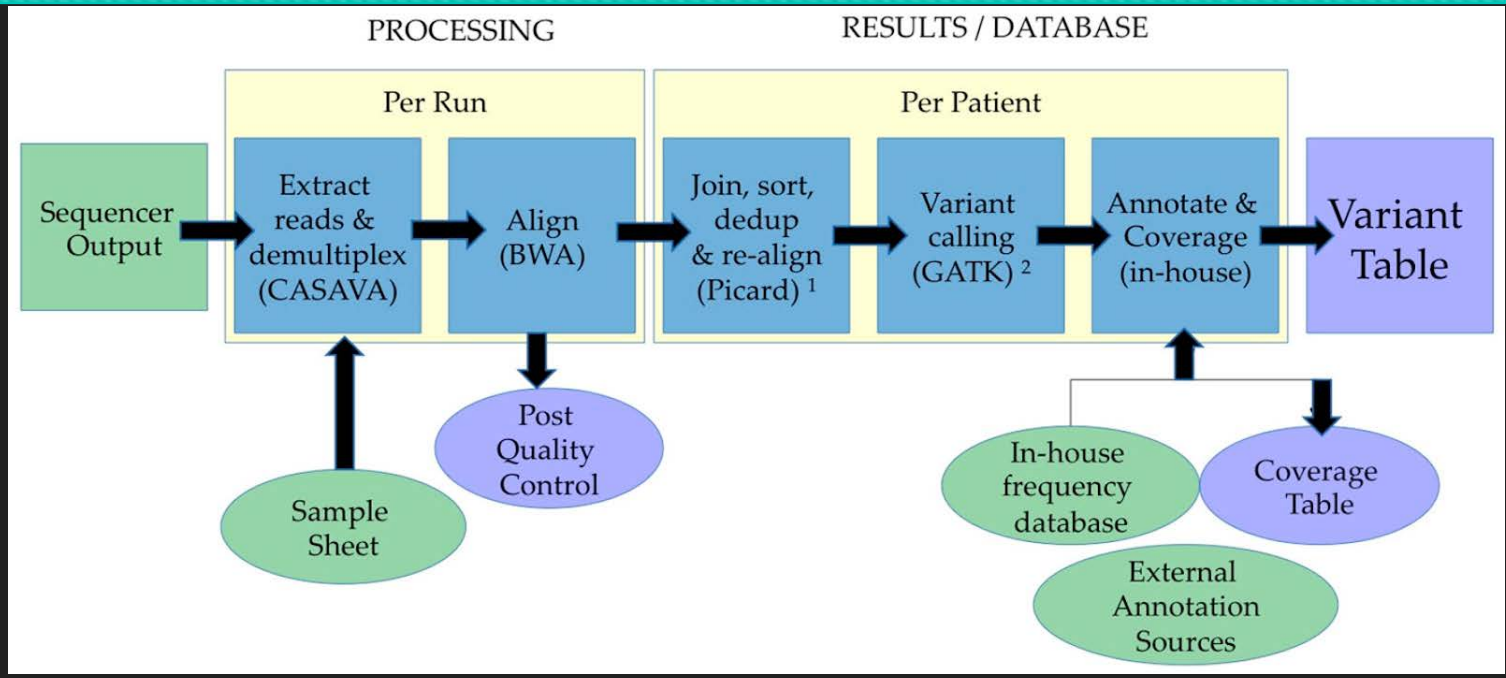
Poorly suited pipelines

- Memory/CPU bound
  - Better code
  - Fat/fast nodes



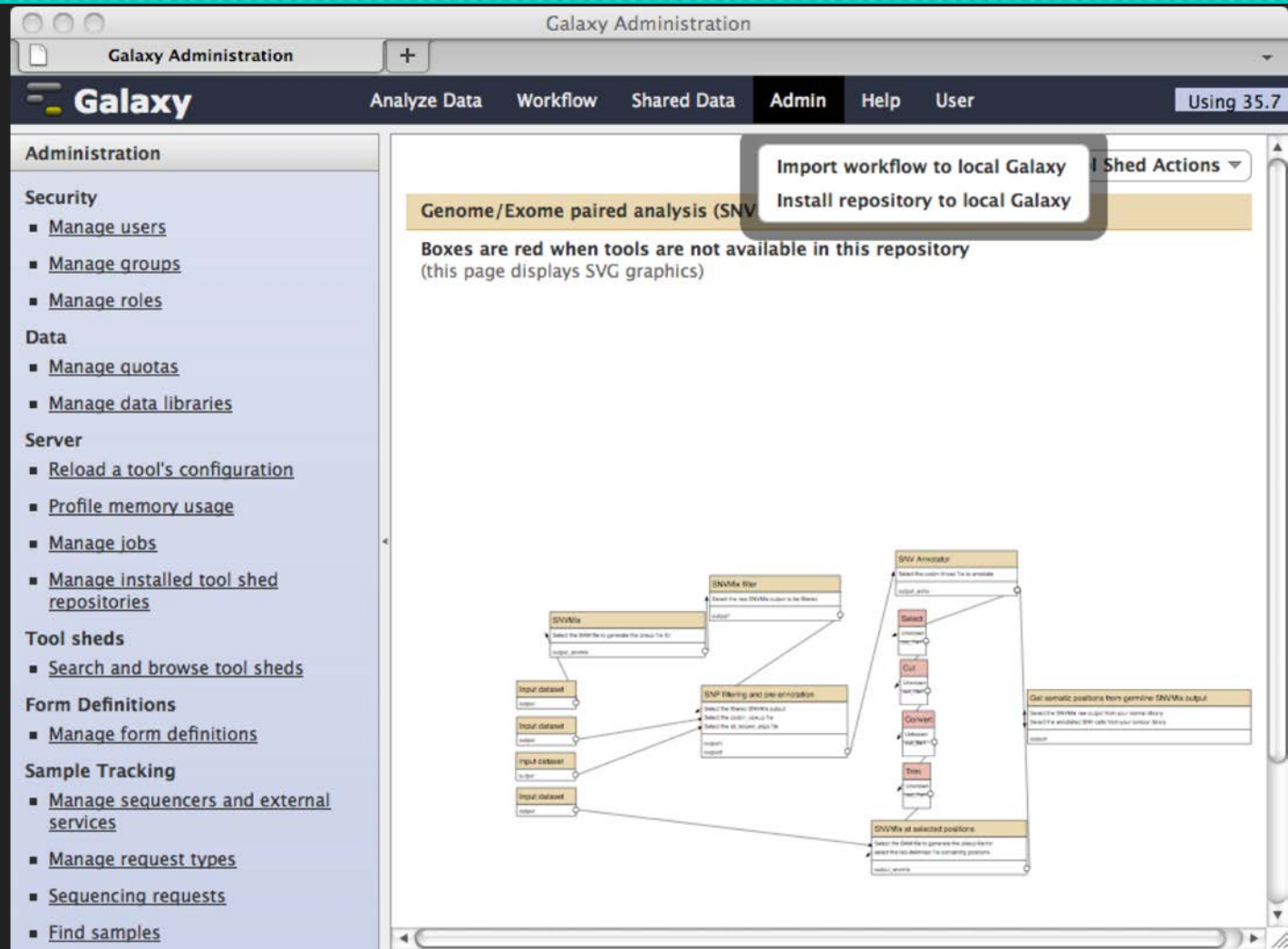
Poorly suited pipelines

- Complex software environments
  - Modules or containers
  - RLIBS and virtualenv
  - Platforms



Galaxy → GenAP

- Dedicating servers to run Galaxy
- GenAP → uses Galaxy and Compute Canada infrastructure on the backend.



Sensitive data

- Large general purpose systems are not accessible
- Need local, or at least regional solutions
  - GSC
  - UBC ARC DRI

Conclusion

- Training
- Move to specialized systems
- Appropriate use on existing systems
- … so “Who’s The Boss?”