Return of the Kraken: Bioinformatics at Scale

Cab Maddux, Day Zero Diagnostics

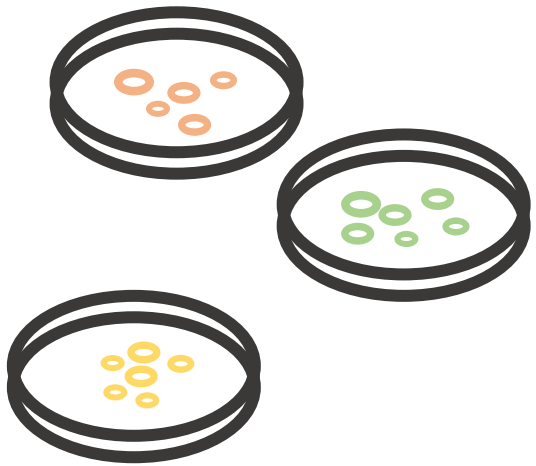
Our Agenda

- 🦠 Bugs!
- 🧬 Making Sense of Sequencing Data
- 🦑 Return of the Kraken
- ☸️ Bioinformatics at Scale
- 🎓 Research + Engineering 👩🏽‍💻

The Take Home

Really hard/important problems (🦠) requiring specialized scientific expertise (🧬, 🦑) is going to necessitate researchers (🎓) and engineers (👩🏽‍💻) working together on solutions that scale (☸).

Bugs!



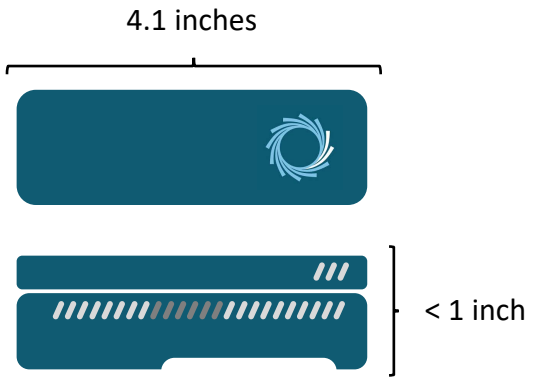
To treat patients with suspected bacterial infections, isolated bacteria are grown in culture, then tests are performed to identify pathogenic bacteria and determine appropriate treatment.

Growing bacteria in culture for species identification and antibiotic susceptibility testing generally takes at least two days.

Bugs!

DAYZERO DIAGNOSTICS

It's become technically and economically feasible to utilize whole genome sequencing, bioinformatics software and machine learning to identify pathogenic bacterial species and profile antibiotic resistance for patients with suspected bacterial infections in hours.



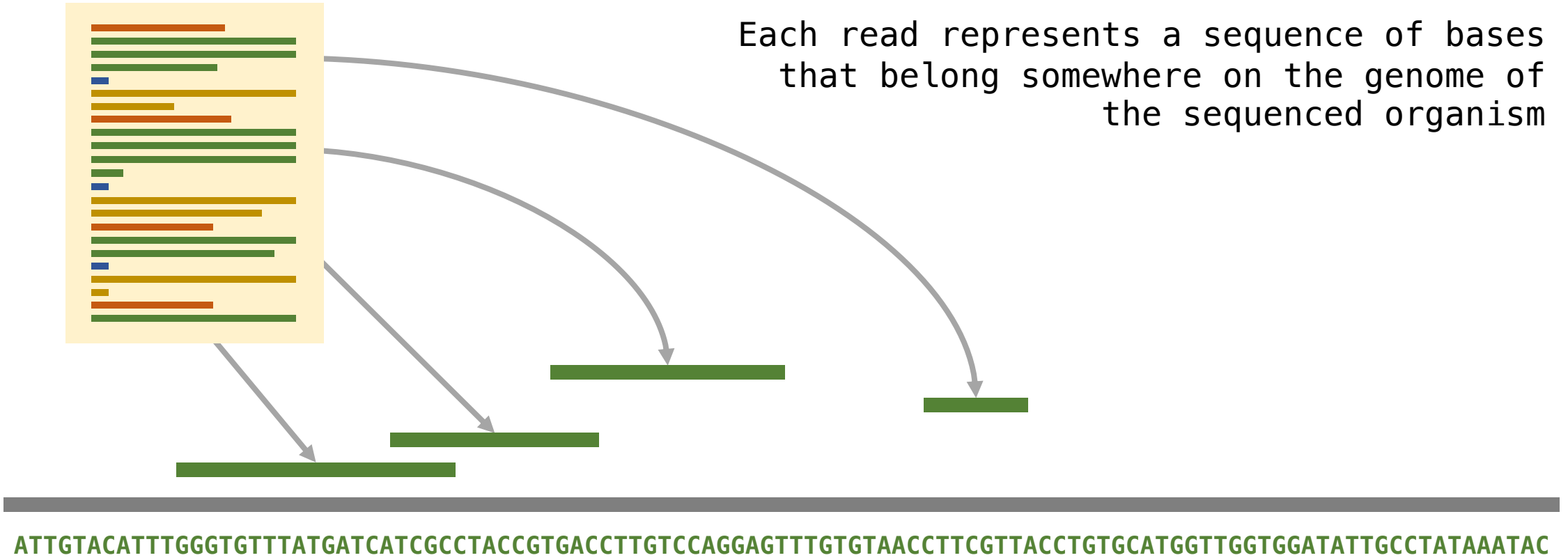
Oxford **N**anopore **T**echnologies

MinION sequencer

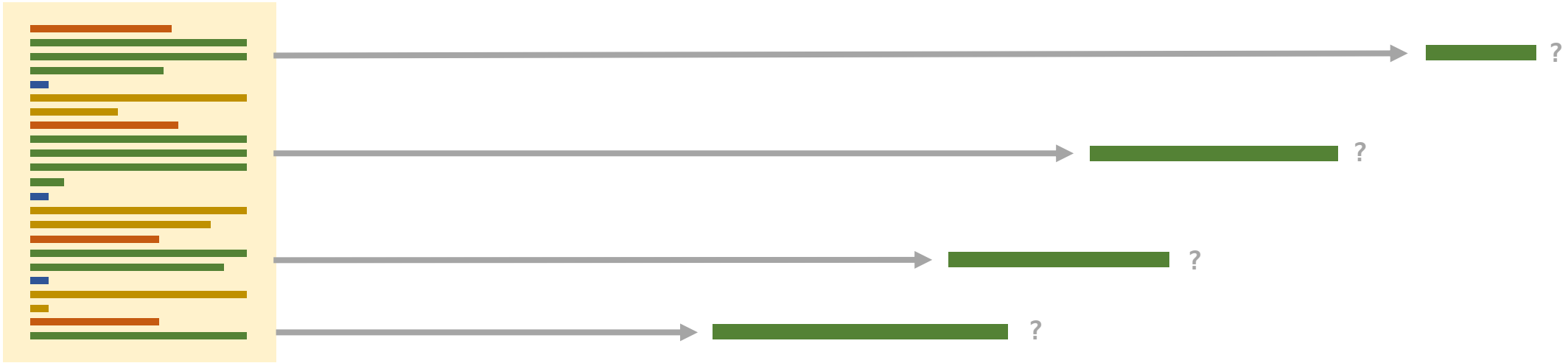
Starts at $1,000

Making Sense of Sequencing Data

Making Sense of Sequencing Data



Making Sense of Sequencing Data



If we don’t know what organism we’re sequencing, this becomes an even more difficult problem

Return of the Kraken

DerrickWood / **kraken**

Watch 28 | Star 183 | Fork 104

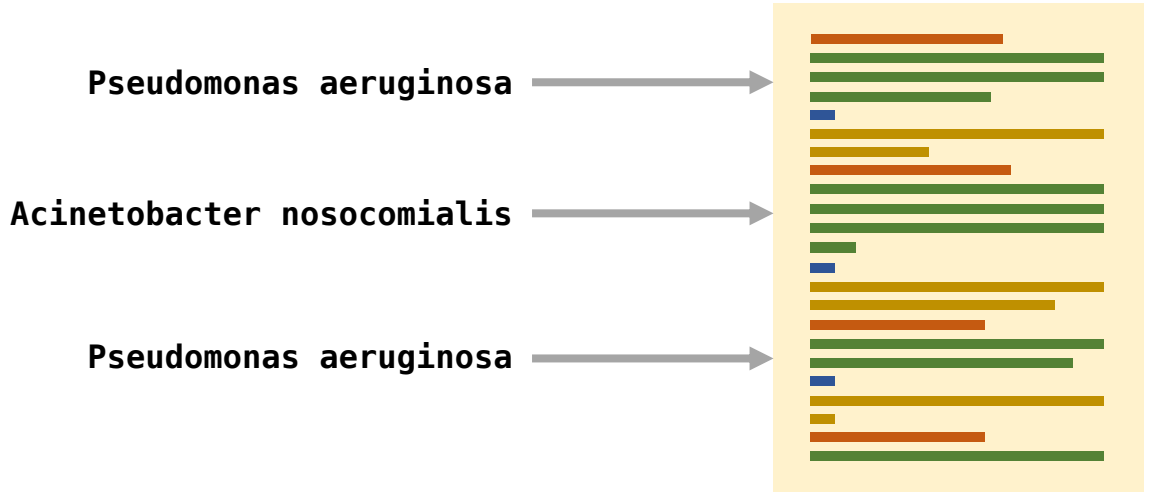
Code | Issues 31 | Pull requests 12 | Actions | Projects | Wiki | Security | Insights

DerrickWood / **kraken2**

Watch 24 | Star 267 | Fork 91

Code | Issues 97 | Pull requests 14 | Actions | Projects | Wiki | Security | Insights

Kraken is a taxonomic sequence classifier that assigns taxonomic labels to DNA reads



Return of the Kraken

ATTGTACATTTGGGTGTTTATGATCATCGCCTACCGTGACCTTGTCCAGGAGTTTGTGTAACCTTCGTTACCTGTG

ATTGTACATTTGGGTGTTTATGATCATCGCCTACCGTGACCTTGTCCAGGAGTTTGTGTAACCTTCGTTACCTGTG

A**TTGTACATTTGGGTGTTTATGATC**ATCGCCTACCGTGACCTTGTCCAGGAGTTTGTGTAACCTTCGTTACCTGTG

AT**TGTACATTTGGGTGTTTATGATCA**TCGCCTACCGTGACCTTGTCCAGGAGTTTGTGTAACCTTCGTTACCTGTG

ATT**GTACATTTGGGTGTTTATGATCAT**CGCCTACCGTGACCTTGTCCAGGAGTTTGTGTAACCTTCGTTACCTGTG

ATTG**TACATTTGGGTGTTTATGATCATC**GCCTACCGTGACCTTGTCCAGGAGTTTGTGTAACCTTCGTTACCTGTG

85% Pseudomonas aeruginosa

12% Acinetobacter nosocomialis

1% Staphylococcus pseudintermedius

Return of the Kraken

DB

The 350G classification data structure must be loaded into memory at runtime – **recommended use of RAM-backed disk.**

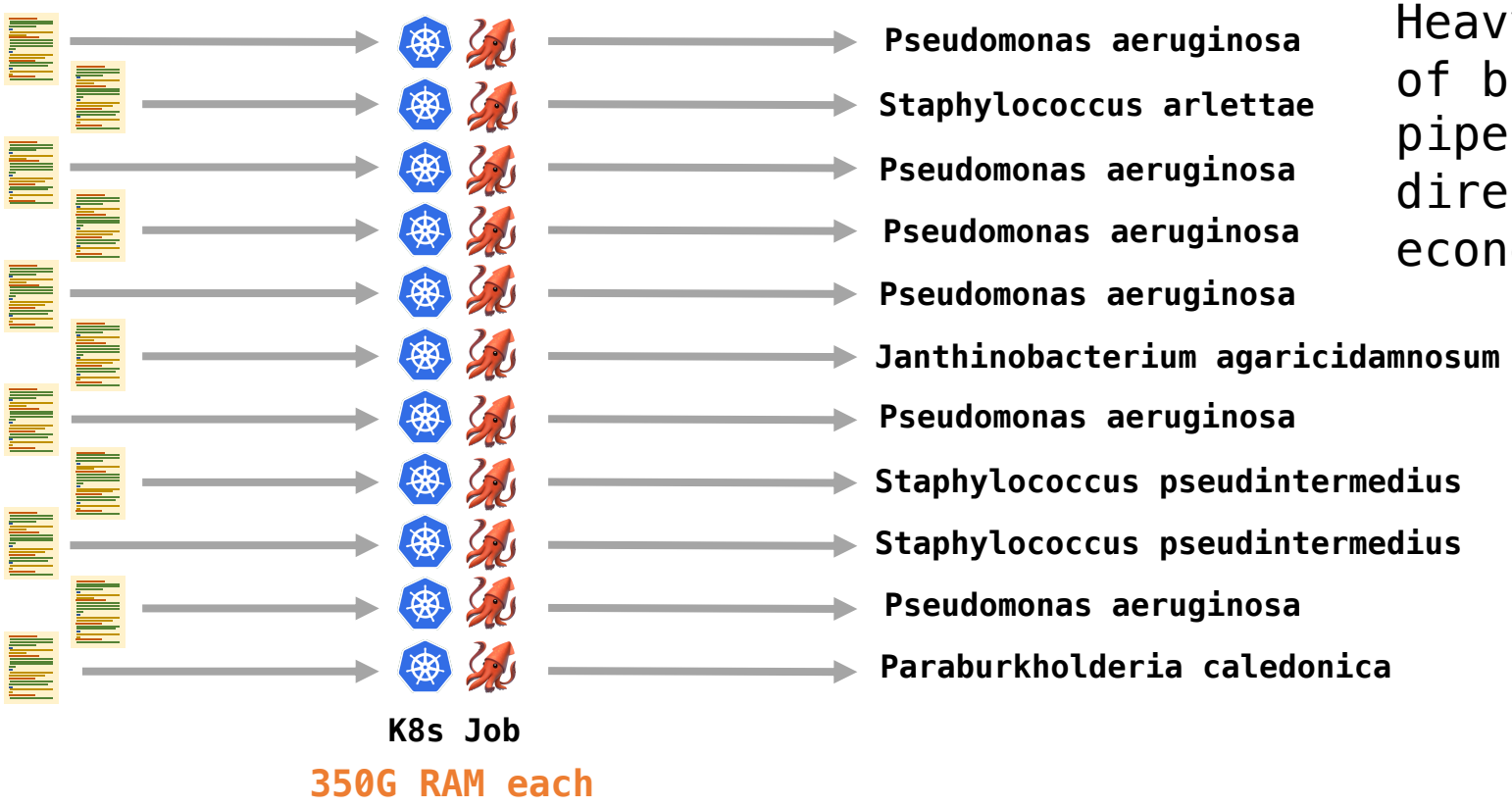
Once loaded into memory, classification of sequencing data containing 500M to 2B bases takes seconds to tens of minutes.

Pseudomonas aeruginosa

Acinetobacter nosocomialis

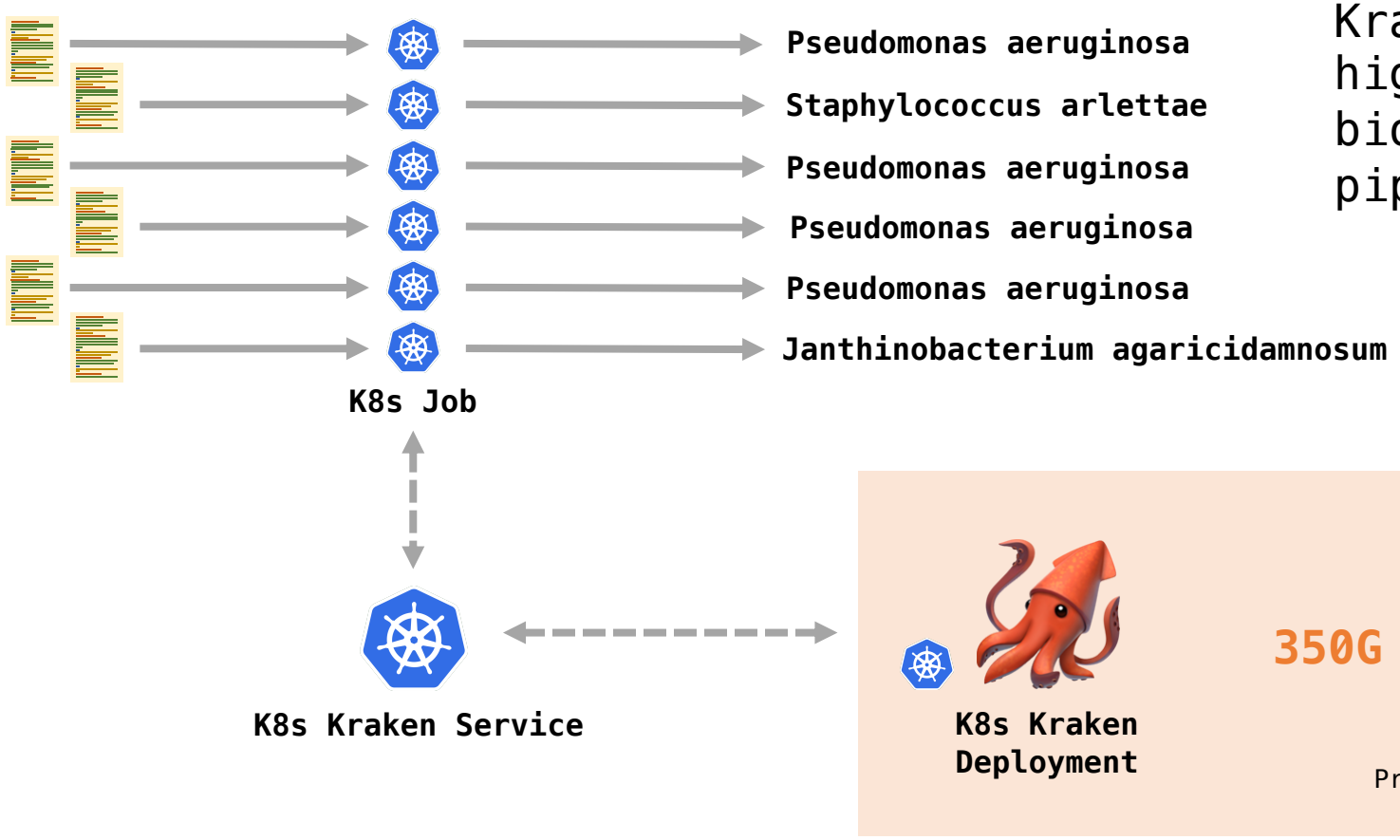
Staphylococcus pseudintermedius

Bioinformatics at Scale



Heavy parallelization of bioinformatics pipelines using Kraken directly is not economically feasible

Bioinformatics at Scale



Centralization of a Kraken service makes highly parallel bioinformatics pipelines possible

Bioinformatics at Scale

Bioinformatics at Scale

HELM

Non-HIPAA compliant

HIPAA compliant

Bioinformatics at Scale

```
volumes:
   - cache-volume
     emptyDir:
         medium: Memory
initContainers:
   - name: prepkrakendb
     image: ubuntu:18.04
     command: [
       “cp”,
       “/krakendb/krakenDB/{database.kdb, database.idx}”,
       “/cache/krakenDB”
     ]
```

There are some complexities associated with this deployment that require engineering expertise

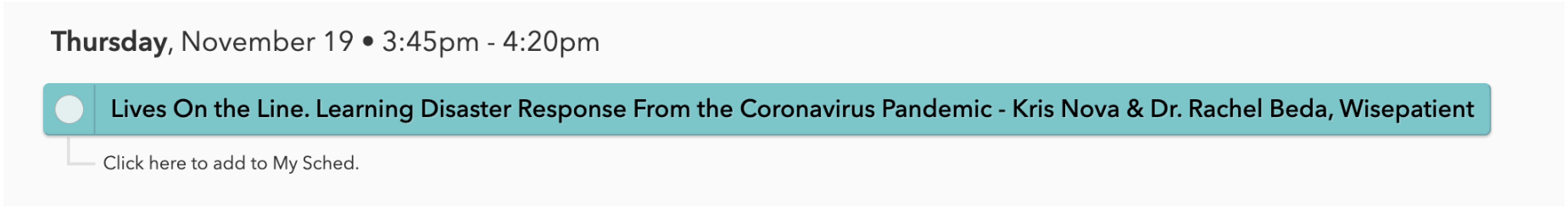
emptyDir with medium: Memory ignores sizeLimit #63126

Open RenaultAI opened this issue on Apr 25, 2018 · 16 comments

The Take Home

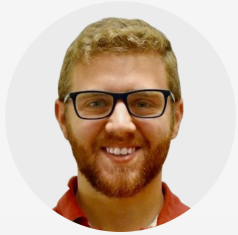
Really hard/important problems (🦠) requiring specialized scientific expertise (🧬, 🦑) is going to necessitate researchers (🎓) and engineers (👩🏽‍💻) working together on solutions that scale (☸).

🎓 Research + Engineering 👩🏽‍💻



Go watch this talk! Support talks, projects, initiatives that apply best of class infrastructure solutions to scientific research and healthcare.

🎓 Research + Engineering 👩🏽‍💻



Zach Munro, Software Engineer @ Day Zero Diagnostics

https://www.linkedin.com/in/zachary-munro/

https://github.com/zmunro

Tim Farrell, Manager, Bioinformatics Data Engineering
@ Day Zero Diagnostics

https://www.linkedin.com/in/tim-farrell-8003bb42/

https://github.com/tmfarrell

Cab Maddux, Principal Software Engineer @ Day Zero Diagnostics

https://www.linkedin.com/in/cabell-maddux/

https://github.com/cmaddux


KEEP CLOUD NATIVE
EVERYWHERE
KubeCon
CloudNativeCon
North America 2020
Virtual