Breakout Session 3: Track A

Exploration of Cloud Computing for CAZyme Research

Dr. Yanbin Yin (Moderator)
Professor, University of Nebraska Lincoln

Exploration of Cloud Computing for CAZyme Research

Yanbin Yin (UNL)

NIH/ODSS Cloud Supplement Program PI Meeting
1/17/2024

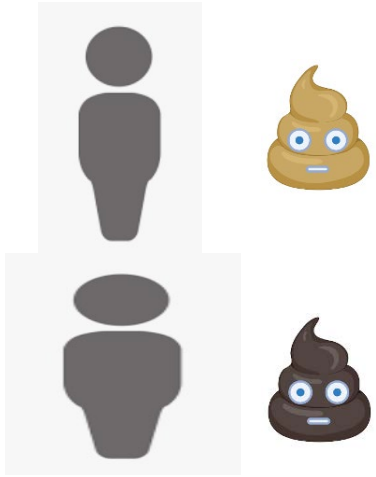
Outline

- Introduction to CAZymes and parent R01 project
- dbCAN tool suite for CAZyme annotation
- Deploy the dbCAN3 web server on AWS

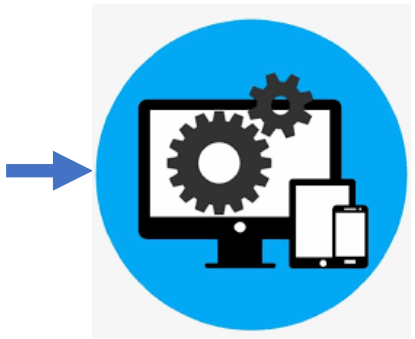
R01 parent grant objective:
Microbiome-based personalized nutrition with bioinformatics tools

Where are CAZymes?

What fibers can you digest?

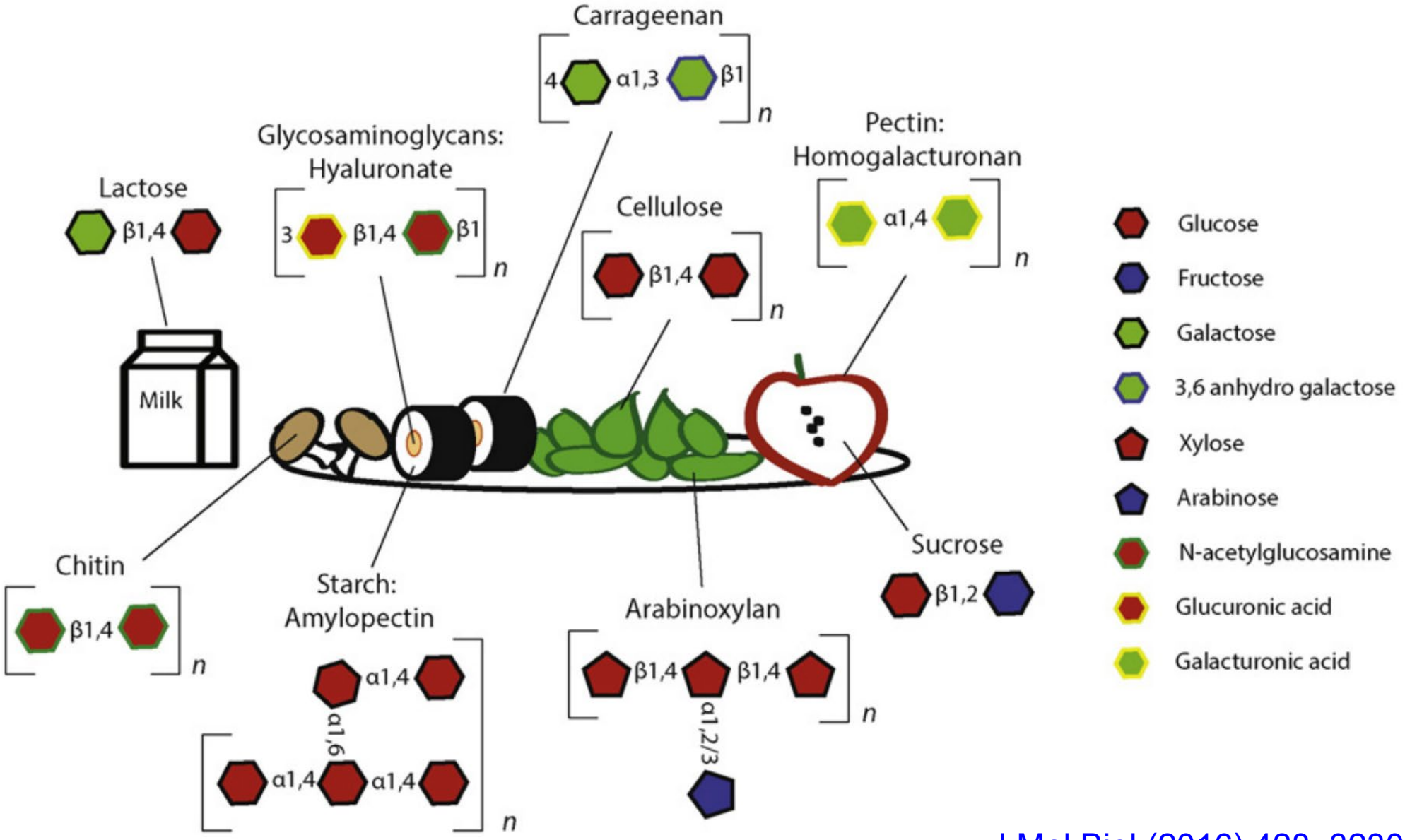


dbCAN



Personalized diet

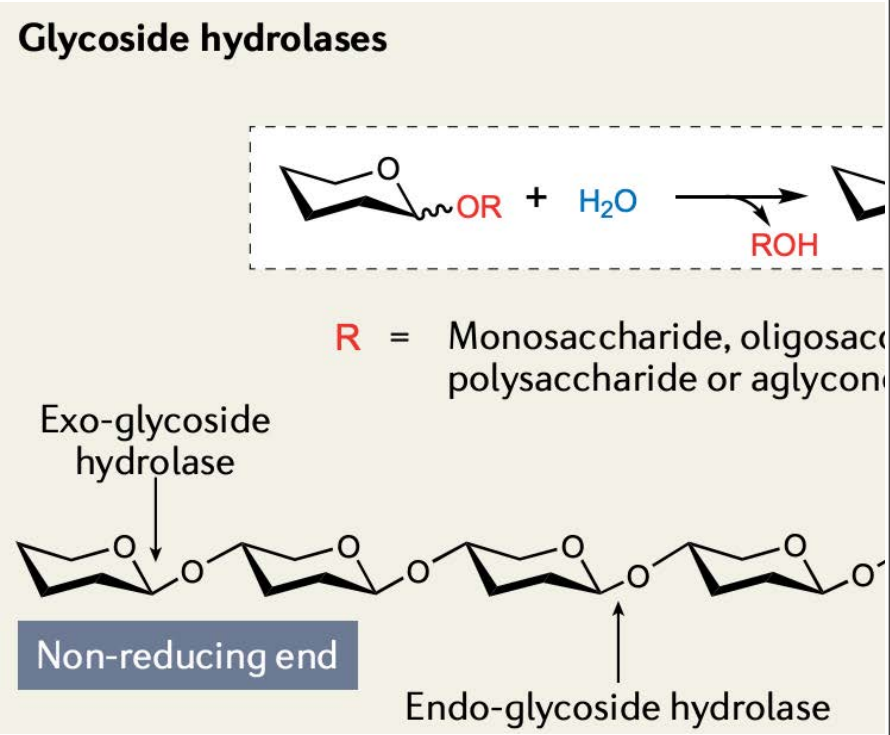
a high diversity of dietary fibers/glycans/carbohydrates

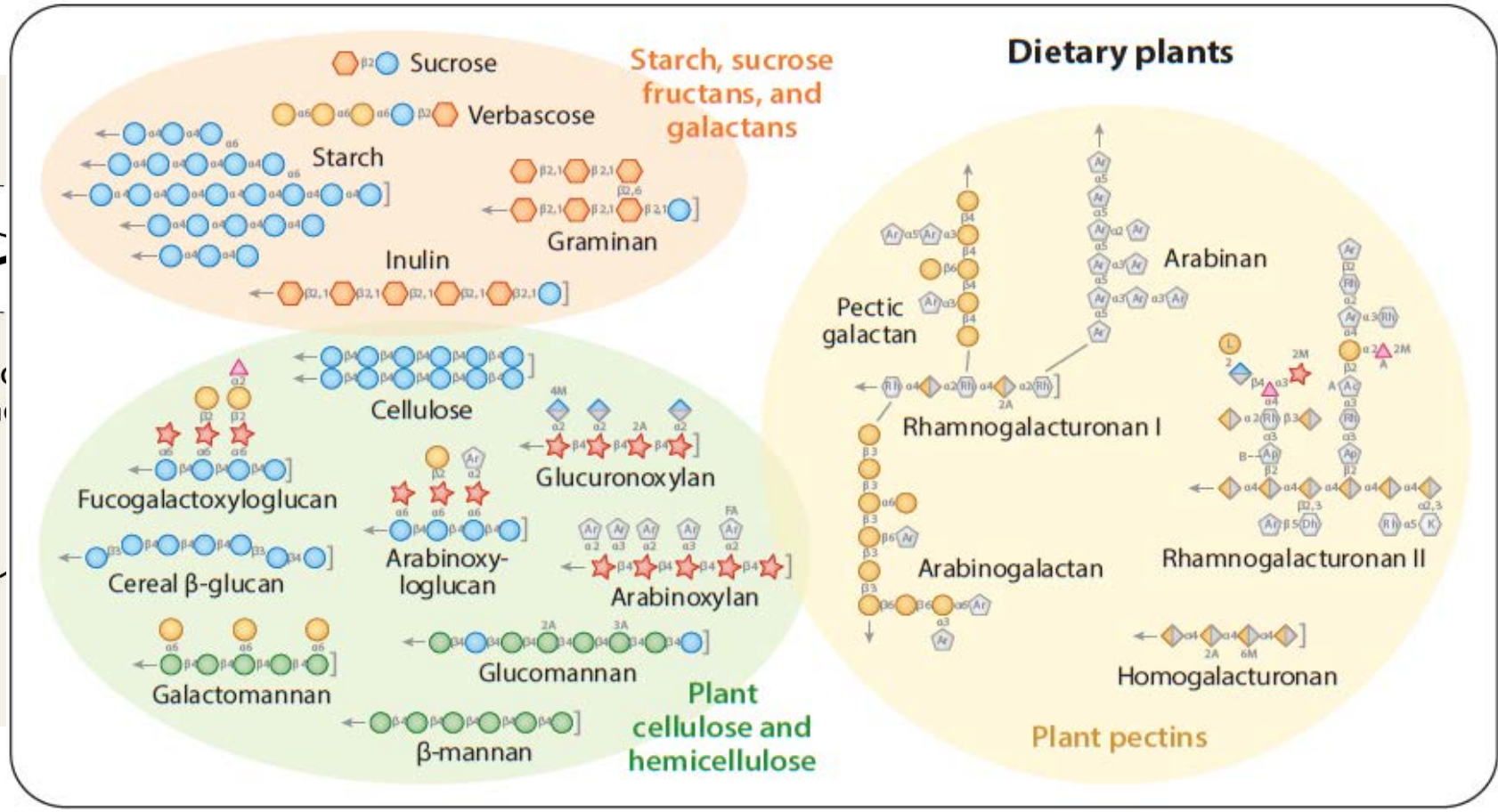


J Mol Biol (2016) 428, 3230–3252

diverse glycosidic linkages exist in the dietary carbs

Nature Reviews Microbiology (2022)





Annu. Rev. Microbiol (2017) 71:349–69

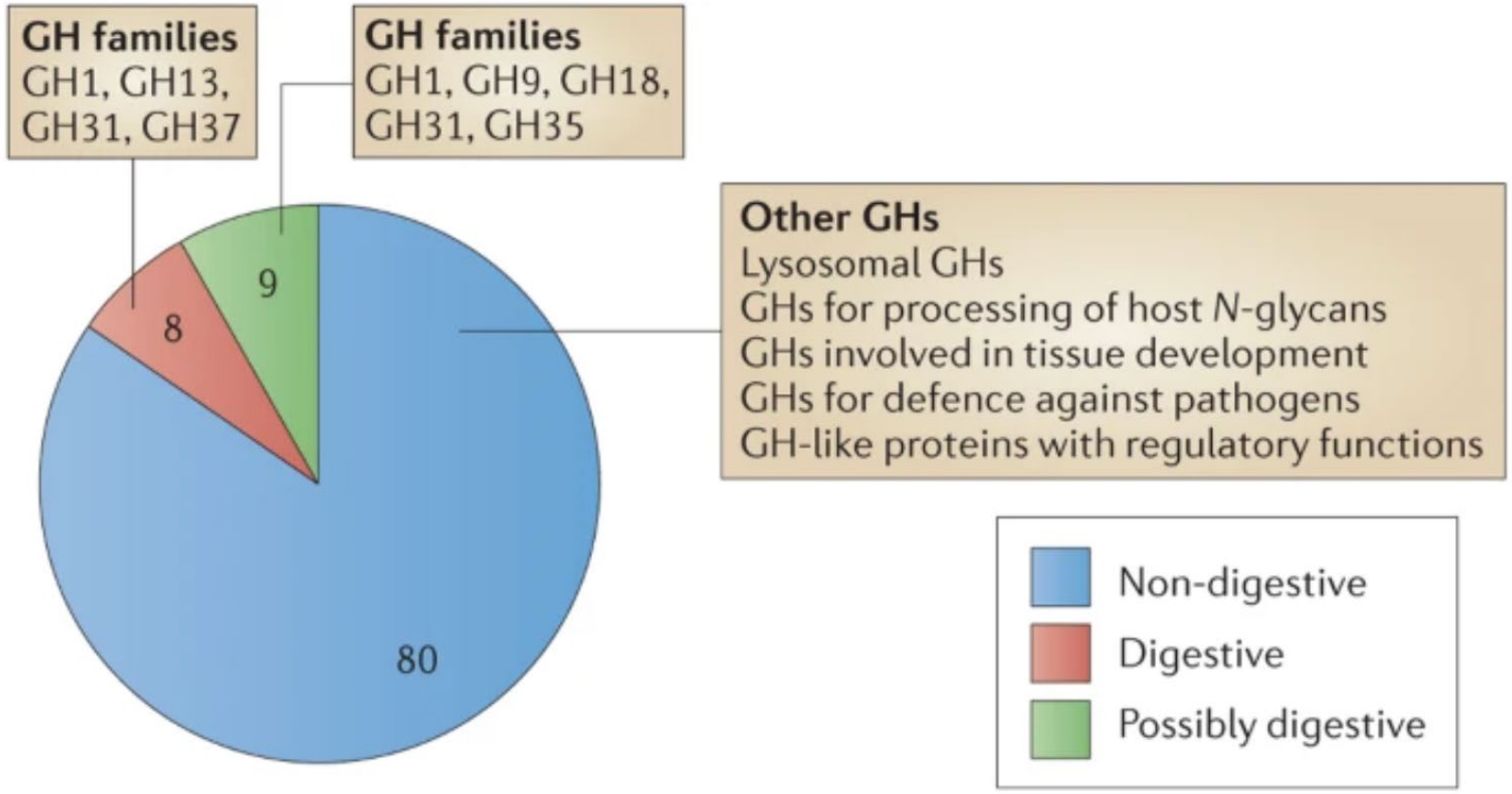
diverse glycosidic linkages need various CAZymes to break

http://www.cazy.org/

- GlycosylTransferases (GTs): **115**
- **Glycoside Hydrolases (GHs): 172**
- Polysaccharide Lyases (PLs): **42**
- Carbohydrate Esterases (CEs): **19**
- Auxiliary Activities (AAs): **17**
- Carbohydrate-Binding Modules (CBMs): **89**

Cantarel B. et al. 2009, Nucleic Acids Res
Lombard V. et al. 2014, Nucleic Acids Res

Human encodes 17 food digesting GHs



NATURE REVIEWS | **MICROBIOLOGY,** doi:10.1038/nrmicro3050, Kaoutari, 2013

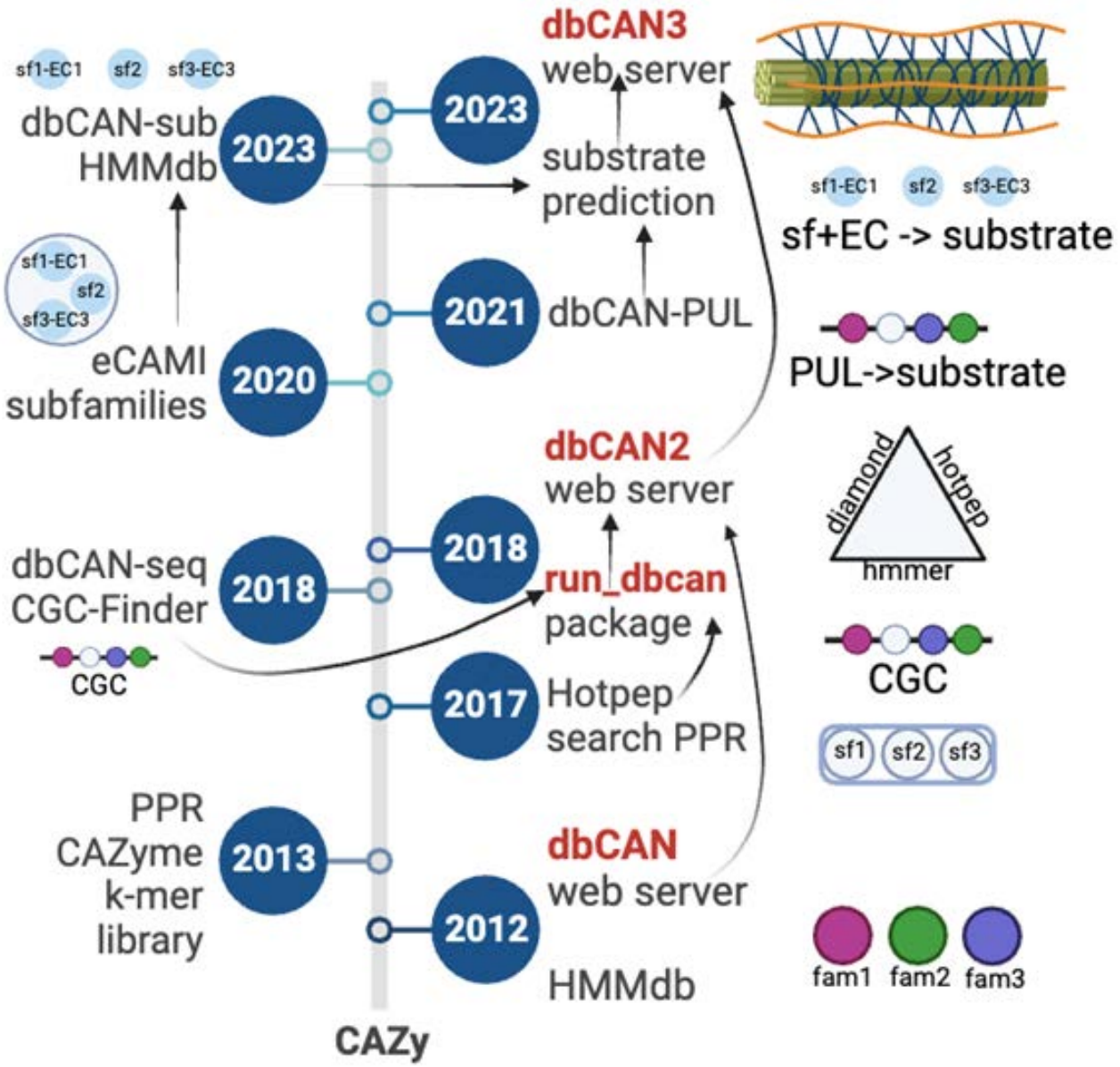
gut bacteria dedicate > 6% of their genes to CAZymes

Bacterium	Total CAZymes	GH	GT	PL	CE	Total CBMs
Bacteroides thetaiotaomicron VPI-5482	386	263	87	16	20	31
B. xylanisolvens XB1A*	349	224	81	22	22	26
B. vulgatus ATCC-8482	279	177	78	7	17	18
B. fragilis 638R	223	138	78	1	6	26
Roseburia intestinalis XB6B4*	175	115	46	0	14	11
Butyrivibrio fibrisolvens 16/4*	115	75	37	0	3	31
Ruminococcus champanellensis 18P13*	87	54	12	9	12	34
Bifidobacterium adolescentis ATCC15703	94	54	37	0	3	6

Gut Microbes 3:4, 289-306; 2012

1000 (species) x 100 (genes) = 100,000 CAZymes

Brief history of dbCAN development



Web server:
https://bcb.unl.edu/dbCAN2

300,000+ jobs in 10 years
8,000+ email addresses

Python package:
https://github.com/linnabrown/run_dbcan

dbCAN3 automated carbohydrate-active enzyme & substrate annotation

Home | Annotate | Download | Example result | Help | About us | AWS mirror site

You are here: Home > Annotate

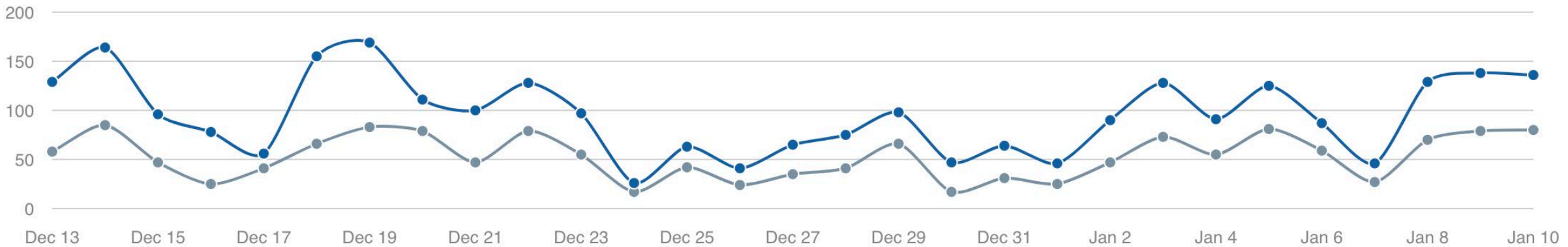
Cite us: dbCAN3 | dbCAN2 | dbCAN

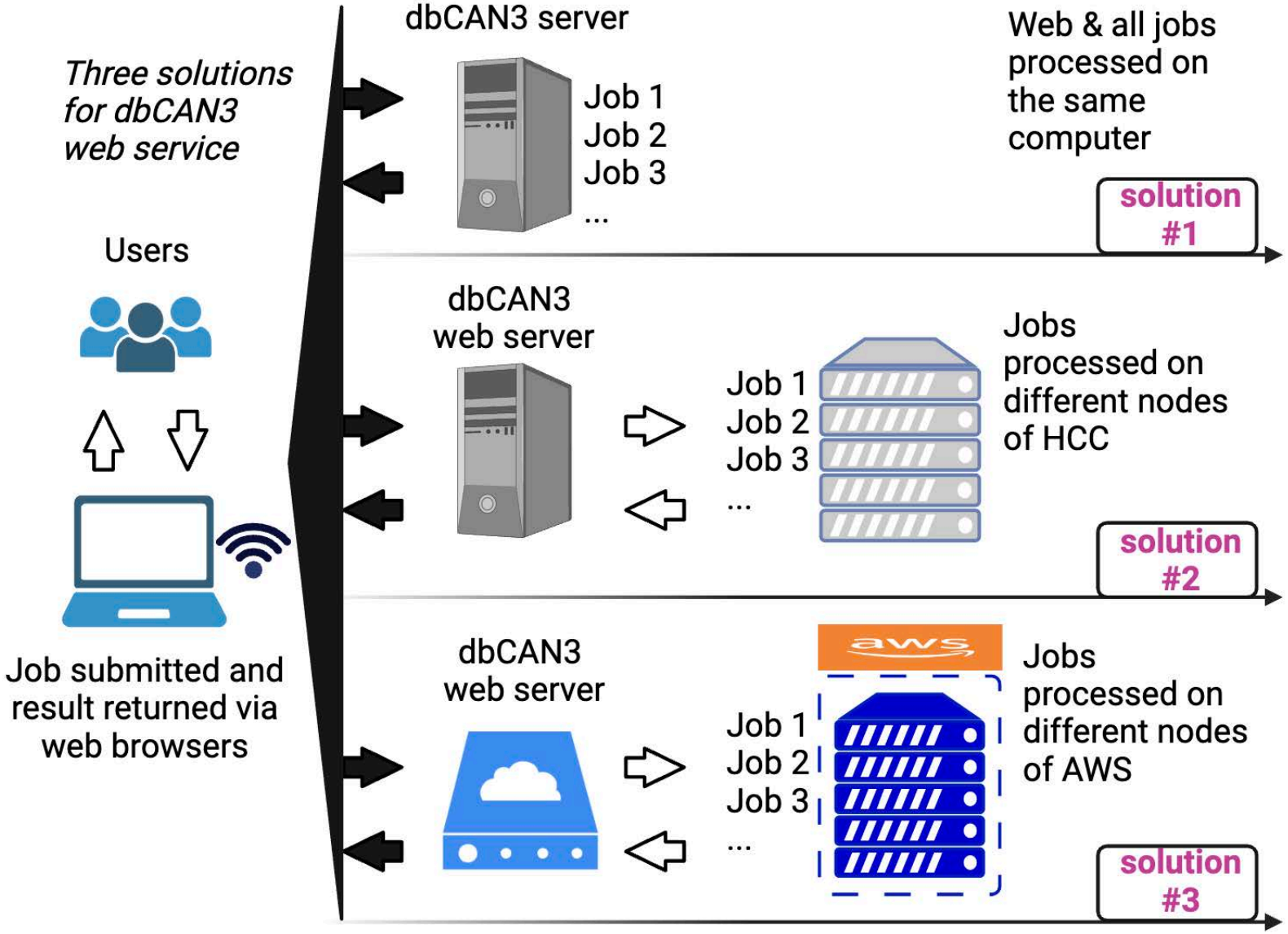
Annotate proteins using DIAMOND, HMMER via CAZy, dbCAN, dbCAN-sub respectively

Server Info: Working
Running Jobs: 1
Pending Jobs: 1
Completed Jobs (2023): 42007

Note: We encourage users to leave your email address if submitting an entire genome or proteome; the result page will be emailed to you when the job is done.
8/2/2023: dbCAN HMMdb v12.0 is released; see readme.txt for details. The DIAMOND db is also updated (7/26/2023).
2/11/2023: You can now predict substrates for CAZymes and CGCs! Please do not use ":" in your FASTA sequence names. This will cause problem in substrate prediciton.
5/25/2022: If your gff file is from NCBI, please check the last column, replace 'Name' tag with 'ID', and 'ID' with 'Name' (only affects CGC predictions).

For future announcement, please follow us on Twitter.





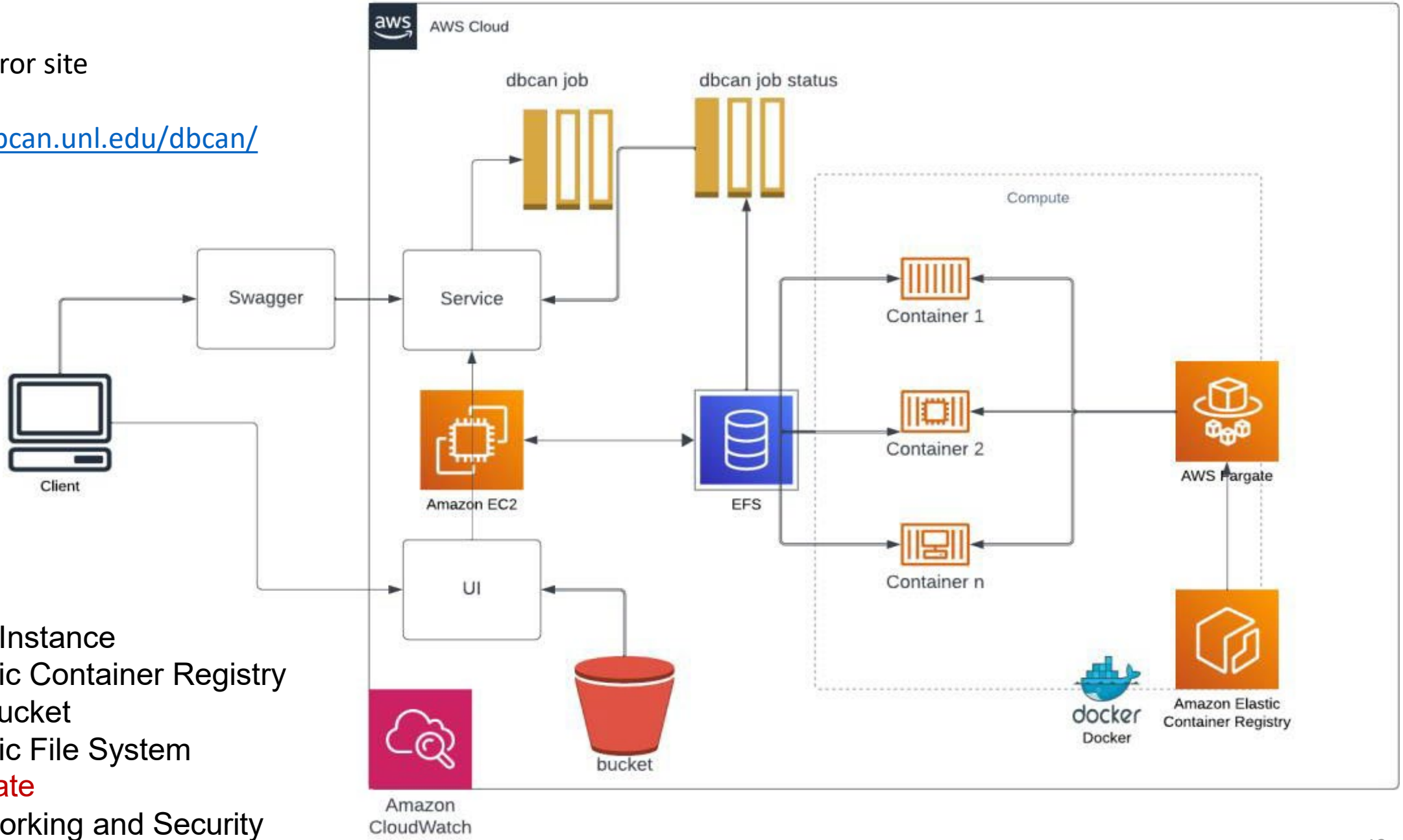
desktop solution

HCC solution

AWS solution

AWS mirror site

http://dbcan.unl.edu/dbcan/



- EC2 Instance
- Elastic Container Registry
- S3 Bucket
- Elastic File System
- Fargate
- Networking and Security

Comparison of computational efficiency: on-prem vs AWS fargate

# of Parallel jobs	Start Time	End Time	Duration
1 On-Prem job	17:37	17:39	2 min
1 Cloud job	17:37	17:44	7 min
10 On-Prem jobs	18:32	18:39	7 min
10 Cloud jobs	18:33	18:39	7 min
50 On-Prem jobs	8:52	9:17	25 min
50 Cloud jobs	8:59	9:05	6 min

On-prem solution is more efficient for individual or fewer jobs
AWS solution offer competitive performance, especially when scaling up to handle more jobs

Acknowledgements

Students:

Qiwei Ge
Yuchen Yan
Jinfang Zheng
Jerry Akresi
Xinpeng Zhang
Nishanth Vangara