

Temporality in Data Science Education

Elliott Hauser University of Texas, Austin

Will Sutherland University of Washington, Seattle

Site

- Two week NSF-funded workshop at an information school
- Educate practicing researchers on data science / data management tools
 - Participants from ecology, genomics, chemistry, seismology, computer science, and others

Containers



Notebooks



Data Management / Storage



Workflows

Snakemake

Version Control



(Github)

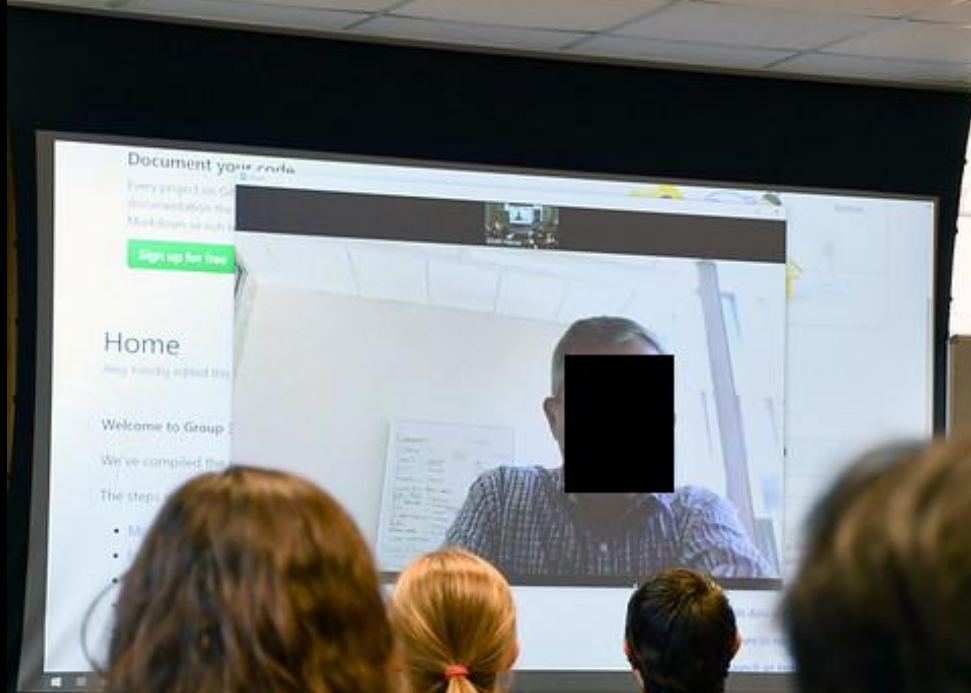
Cloud Computing



Machine Learning



Elliott




Data


- 21 participant interviews, 9 instructor interviews (30 total)
- Two weeks participant observation
- Group Slack chats
- Collaborative notetaking via HackMD
- GitHub code repositories and documentation created by participants
- Instructor presentations
- Participant final presentations



Data


- Important exchanges between participants


 [redacted] Jul 21st, 2019 at 2:36 PM
@ [redacted] yups I got the same error (edited)



20 replies


 [redacted] 8 months ago
the workflow_simple.cwl file works when I give the input 'cwl-runner workflow_simple.cwl workflow_simple-job.yaml'. Not sure if that counts as getting it working 🤔




 1 

 [redacted] 8 months ago
Yeh same workflow_simple working for me but not the workflow

 [redacted] 8 months ago
Reproducible science is a myth 🤔

 1 

 [redacted] 8 months ago
.yml file says this

Data

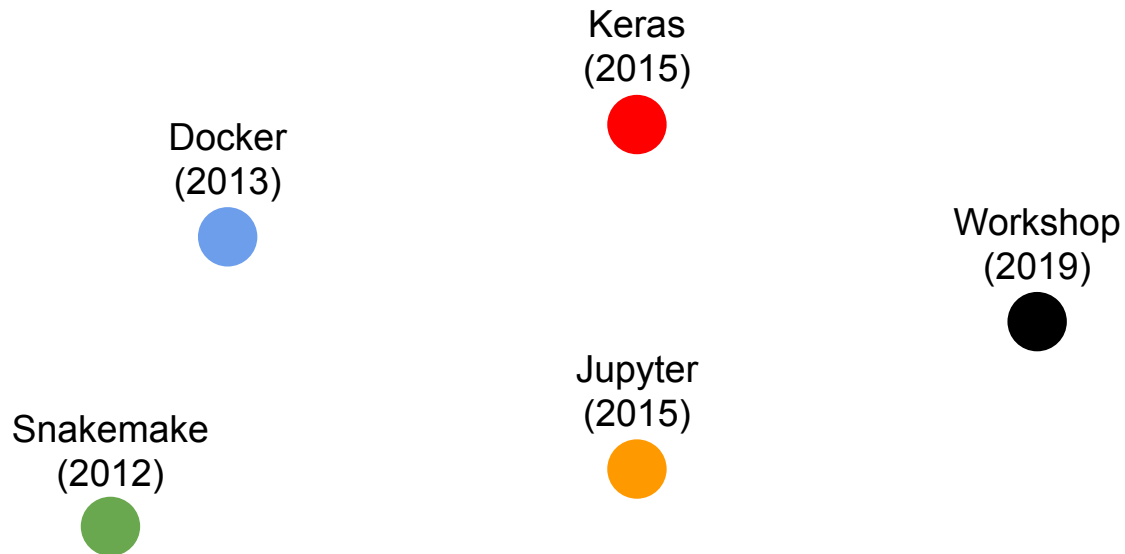
- 21 participant interviews, 9 instructor interviews (30 total)
- Two weeks participant observation
- Group Slack chats
- Collaborative notetaking via HackMD
- GitHub code repositories and documentation created by participants
- Instructor presentations
- Participant final presentations

“machine learning with Keras”

“reproducibility with Docker”

“workflows with Snakemake”

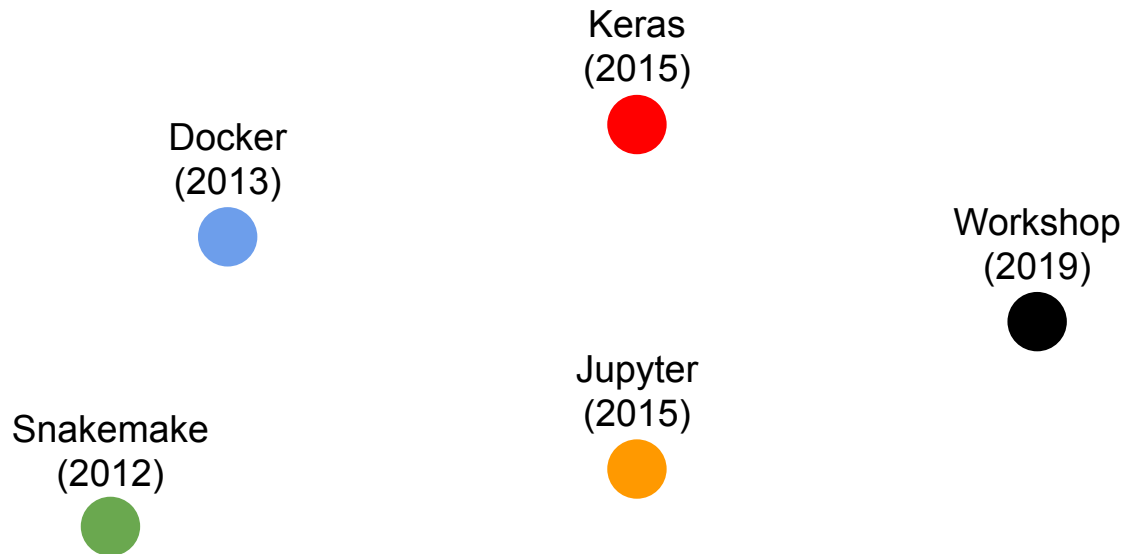
“Five Years Ago”



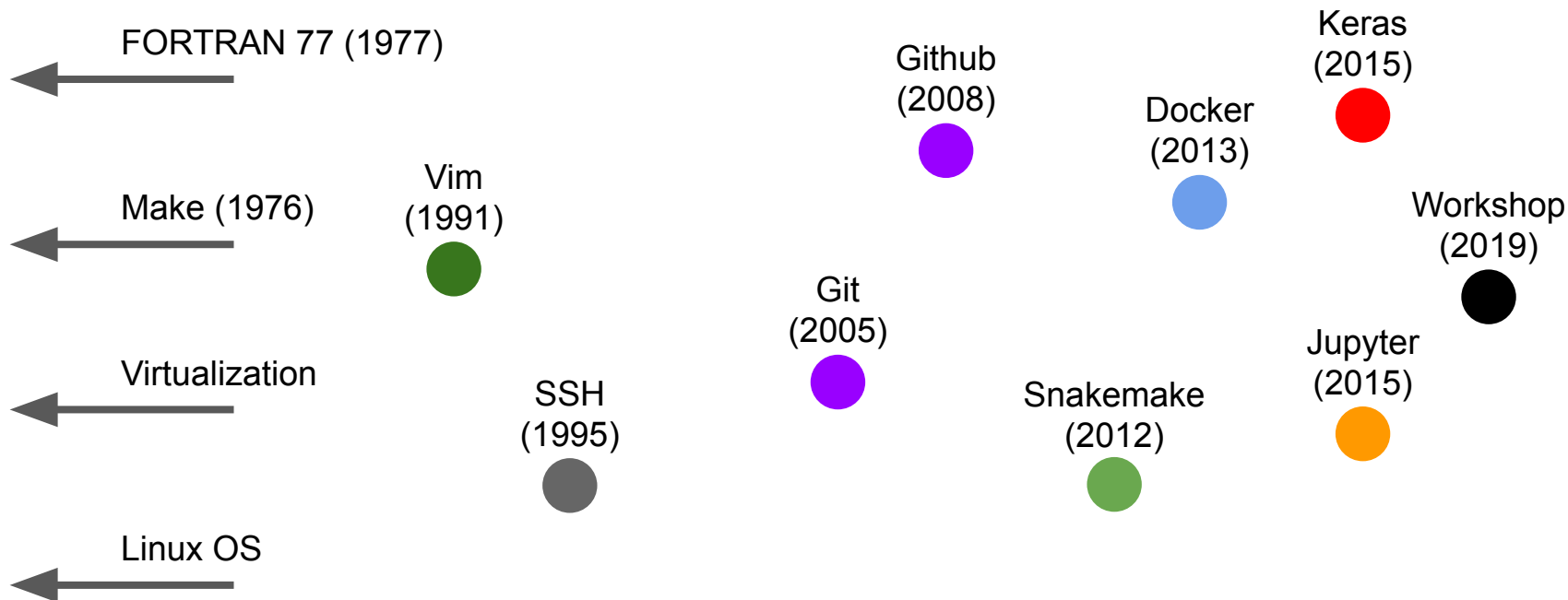
...and five years hence

“...of course, no one wants to think about that the technologies we’re learning today will be obsolete in five years’ time, but that’s another story” (I10)

“Five Years Ago”

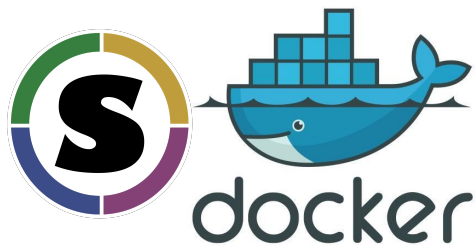


“Five Years Ago”



Versions, Labels,
Names

Containers



(Singularity)

docker

Workflows

Snakemake

Machine Learning



Notebooks



Data Management /
Storage



Cloud Computing

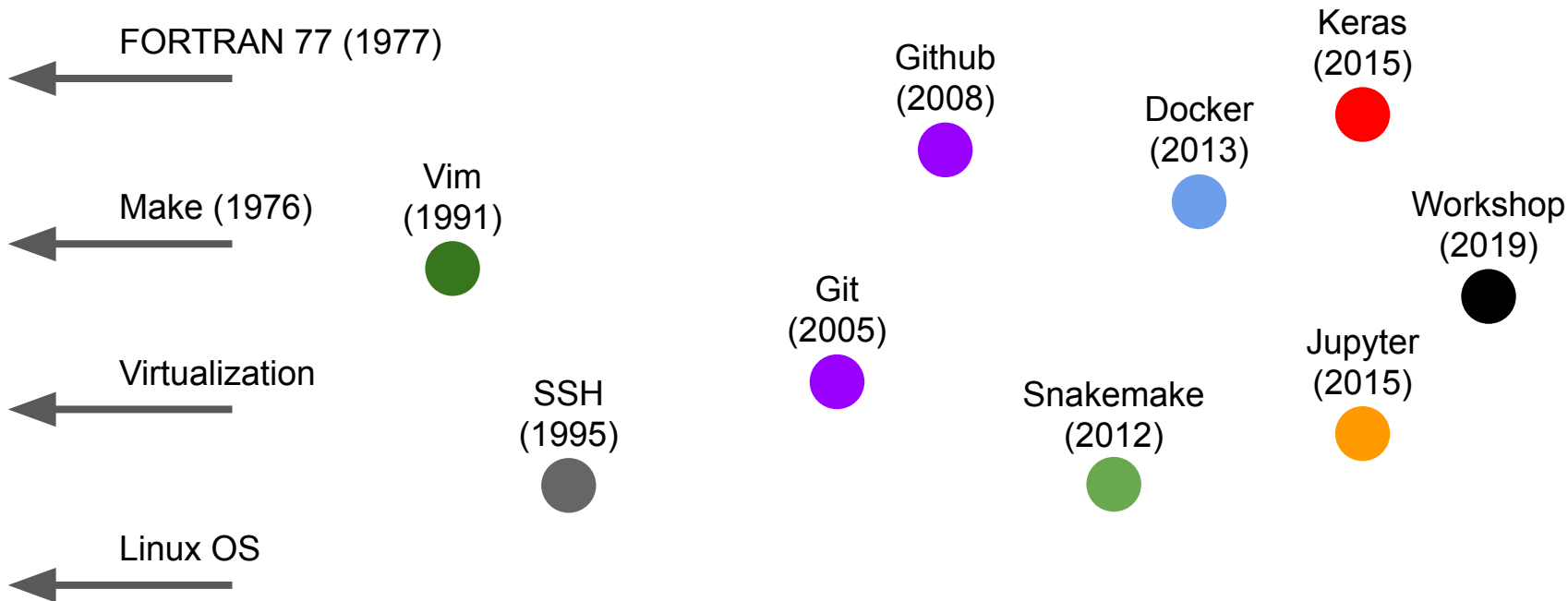


Version
Control



(Github)

“Five Years Ago”



Takeaways

- Problematize what precisely we are teaching
 - There is a difference between teaching tools and teaching cyberinfrastructure
 - We need to think about precisely what a data science curriculum includes
- Dual commitments to the reliable, and to the cutting edge
 - The historical mantra of *in medias res*
- Data science as a nascent problematic

Extra Slides

“Yeah mostly I struggled with all of these, like everything we talked about at the beginning was so abstract. Like I’ve never heard about virtual machines before which I guess is a pretty basic concept in computer science but I had never, like I didn’t have a mental structure to hang all this on. Like what’s happening in JetStream and now we’re going to CyVerse and atmosphere something and trying to figure where, like we’re SSH’ing into this and we’re building a container and like trying to model that in my brain, it was not happening in the first few days...” (P9)