Session 152: Data Science in the Cloud in Under an Hour

SOA Antitrust Compliance Guidelines
SOA Presentation Disclaimer

Data Science In The Cloud In Under an Hour

Sarah Moore, ASA, MAAA
Shea Parkes, FSA, MAAA

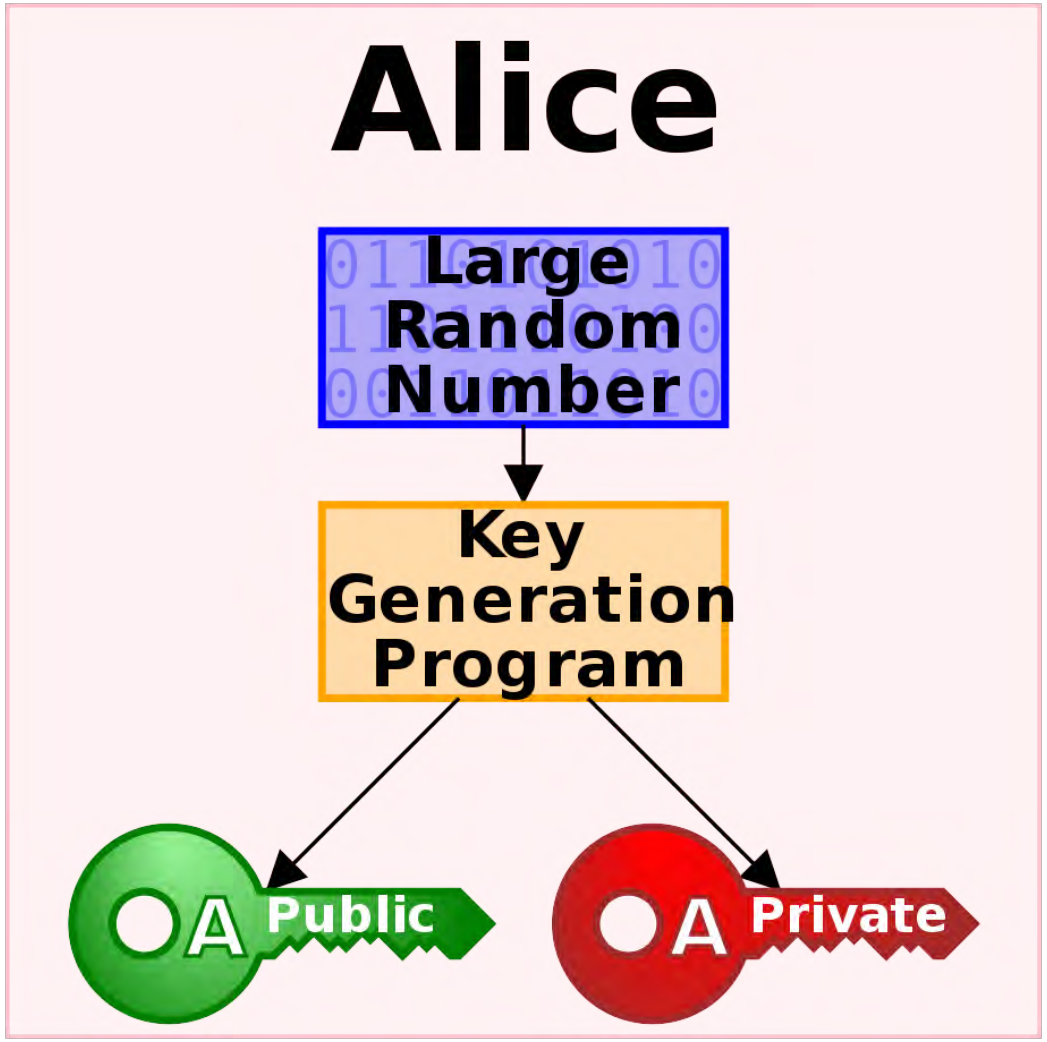
Milliman

Limitations

- The views expressed in this presentation are those of the presenter, and not those of Milliman or the Society of Actuaries. Nothing in this presentation is intended to represent a professional opinion or be an interpretation of actuarial standards of practice.

Public Key Cryptography

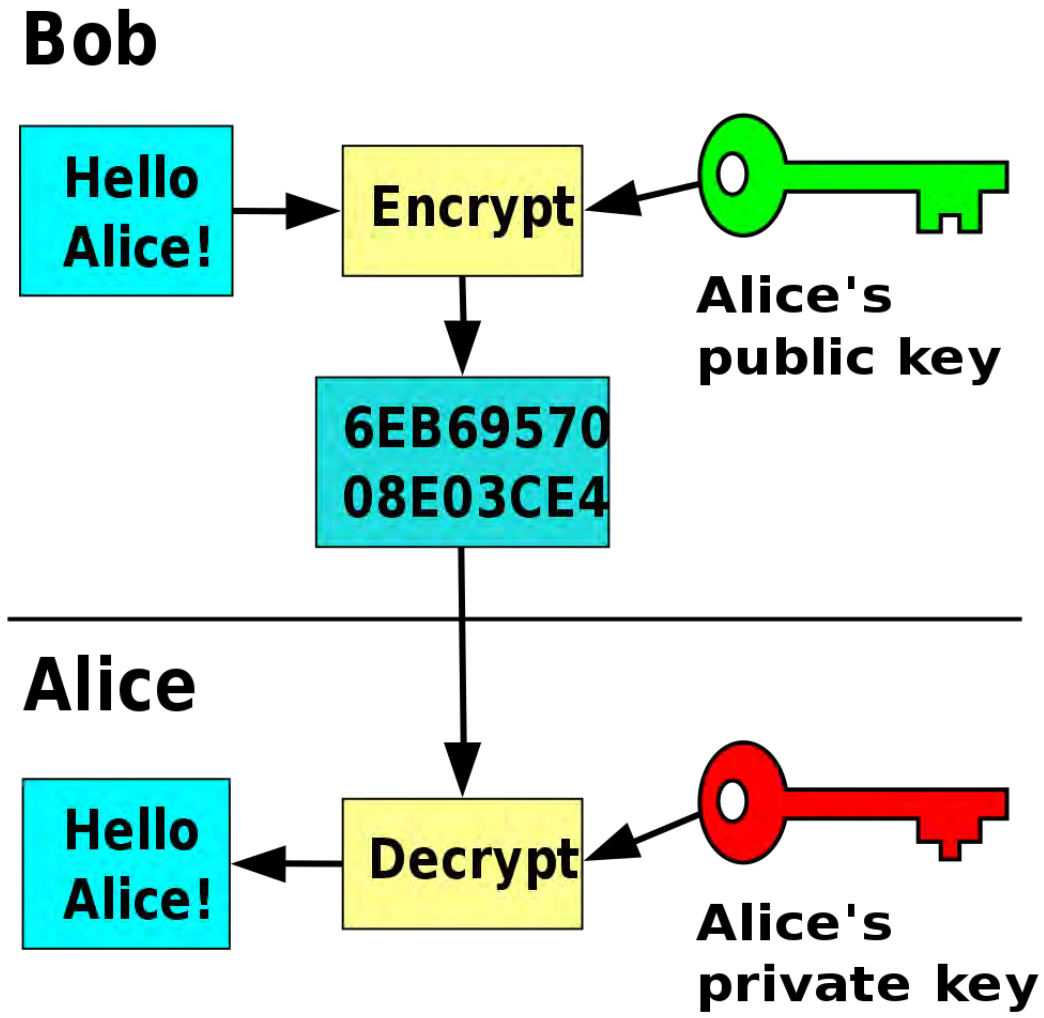
Generate your keys



https://en.wikipedia.org/wiki/Public-key_cryptography

Public Key Cryptography

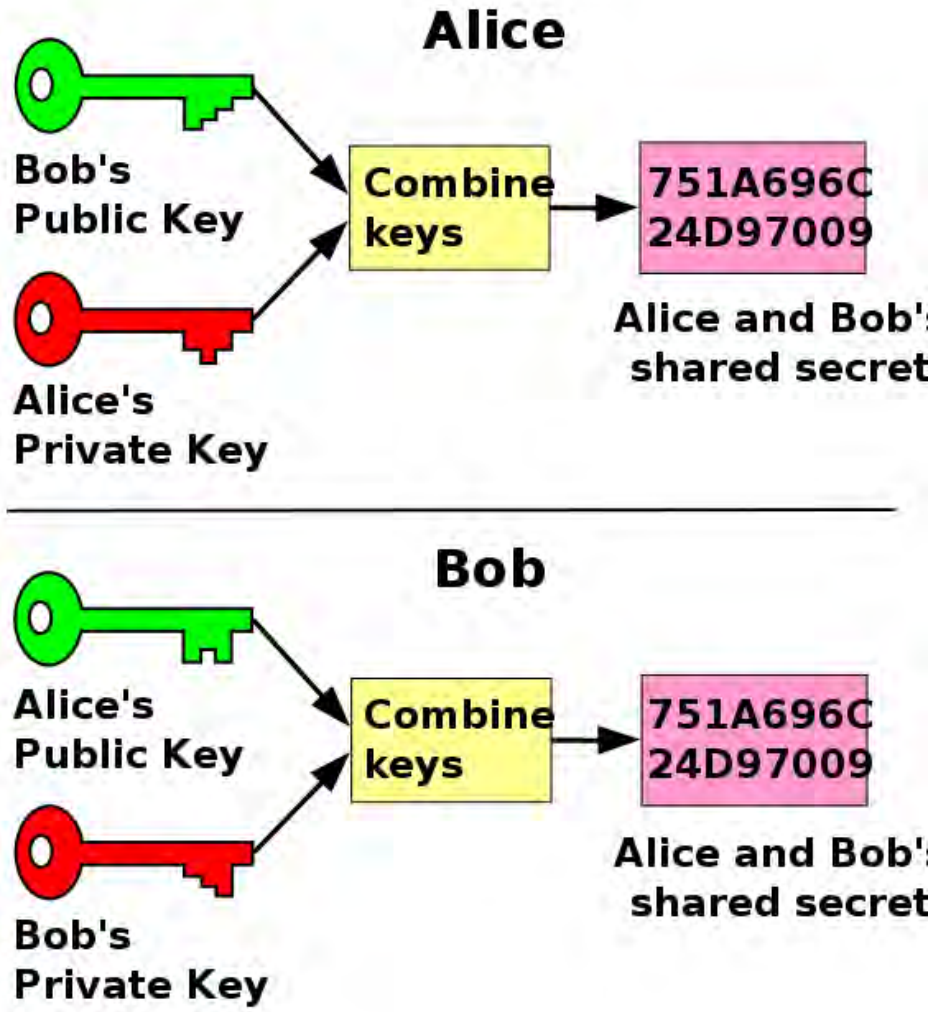
One-way encryption



https://en.wikipedia.org/wiki/Public-key_cryptography

Public Key Cryptography

Two-way encryption

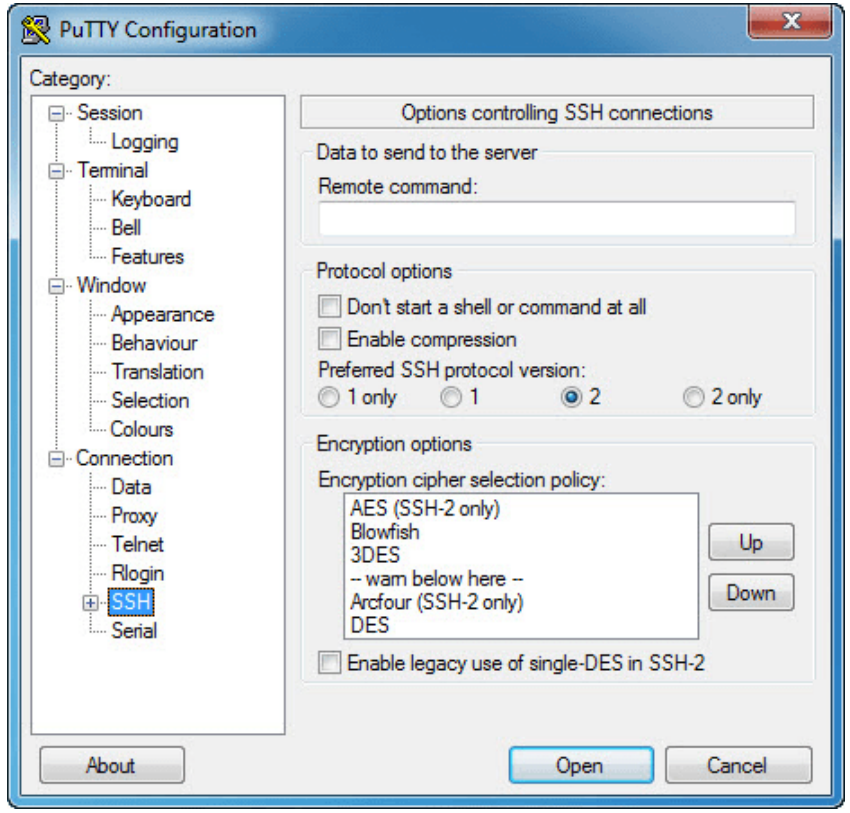


https://en.wikipedia.org/wiki/Public-key_cryptography

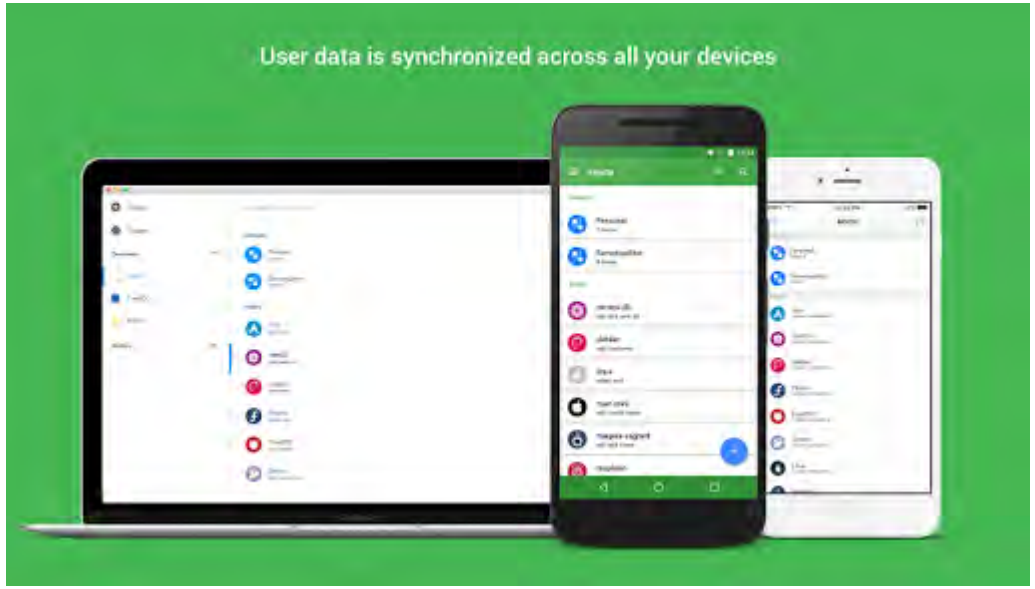
SSH Clients

Built-in to Mac OSX and Linux,
Have to install client on Windows…

Putty (https://www.putty.org/)
- Free, tried and true



Termius (http://www.termius.com/)
- Paid, cross-platform, cloud sync

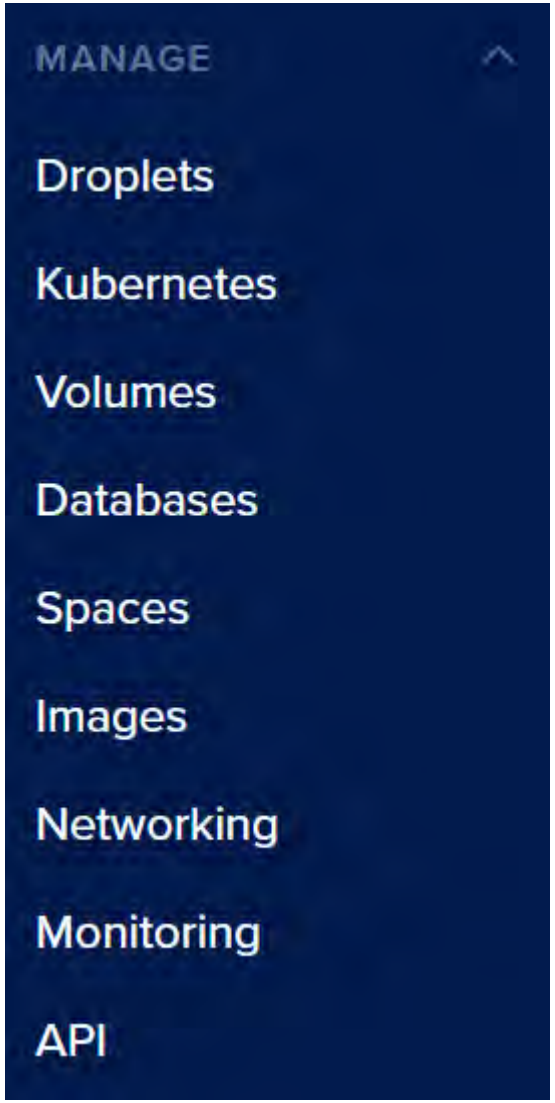


… and countless more options available.

Big Cloud Providers

Bite-Sized Cloud Provider

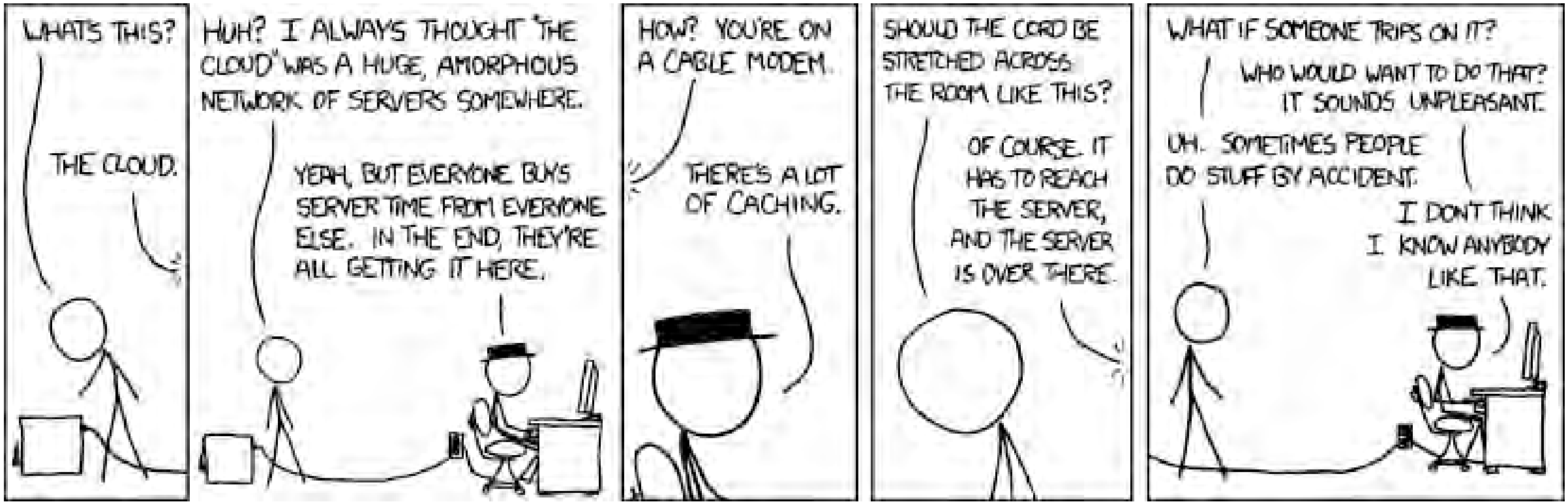
https://www.digitalocean.com/



The steps

- Make an SSH key with Termius
- Upload public key to Digital Ocean account
- Launch private Virtual Machine (VM) on Digital Ocean
  - Use a container-centric OS (e.g. CoreOS)
  - Digital Ocean injects our public key into this VM when it is created
- Add firewall rule to only allow SSH
- Use Termius to connect to our VM
  - If using CoreOS, username is “core”
  - VM will send us its own public ~key at this time.
- Launch data science container: https://github.com/jupyter/docker-stacks
  - `docker run -d -p 8888:8888 jupyter/scipy-notebook`
- Get URL w/ access token for Jupyter notebook (`docker logs -f …`)
- Use Termius to securely forward port 8888 from VM to local PC
- Use local web browser to access Jupyter notebook with provided URL
- Use Jupyter notebook interface to:
  - Upload data, Run analysis, Download results
- Destroy VM

Obligatory xkcd: https://xkcd.com/908/



<img src="//imgs.xkcd.com/comics/the_cloud.png" title="There's planned downtime every night when we turn on the Roomba and it runs over the cord." alt="The Cloud">

Cautions and Caveats

- You should always involve your local IT department.
- Cloud vendors impose limits until they trust you.
- You really don't want to run the same VM for weeks at a time.
  - You become responsible for security updates and a whole mess of other things.
  - You also would get charged for a lot of idle time…
- You should learn about cloud-native data storage soon after trying this.
  - And you need to be mindful of data transfer charges.
- You likely want to get away from managing VMs at all.
  - Which is a large reason the big cloud vendors are so complex.
  - You should also look at custom Data Science cloud "wrappers" (e.g. DataBricks, Data Robot)
- Refine cloud firewall to only allow specific IPs.
- Don't store ePHI / PII without entering proper agreements.
  - Digital Ocean won't enter BAAs yet even…
- Be careful when choosing which cloud data center to ensure compliance with data sovereignty

Milliman

Thank you

Shea Parkes
shea.parkes@milliman.com

Other container examples

- docker run -d -p 8888:8889 -e JUPYTER_ENABLE_LAB=yes jupyter/scipy-notebook
  - Still need to open logs to find URL with access token.
  - The above maps to 8889 just to avoid conflict with basic Jupyter
- docker run -d -p 8787:8787 -e PASSWORD=soa rocker/rstudio
  - No token URL needed, just browse to https://127.0.0.1:8787 (after forwarding port)
  - Username is “rstudio”, password is whatever you make it above.