Live Data Products!

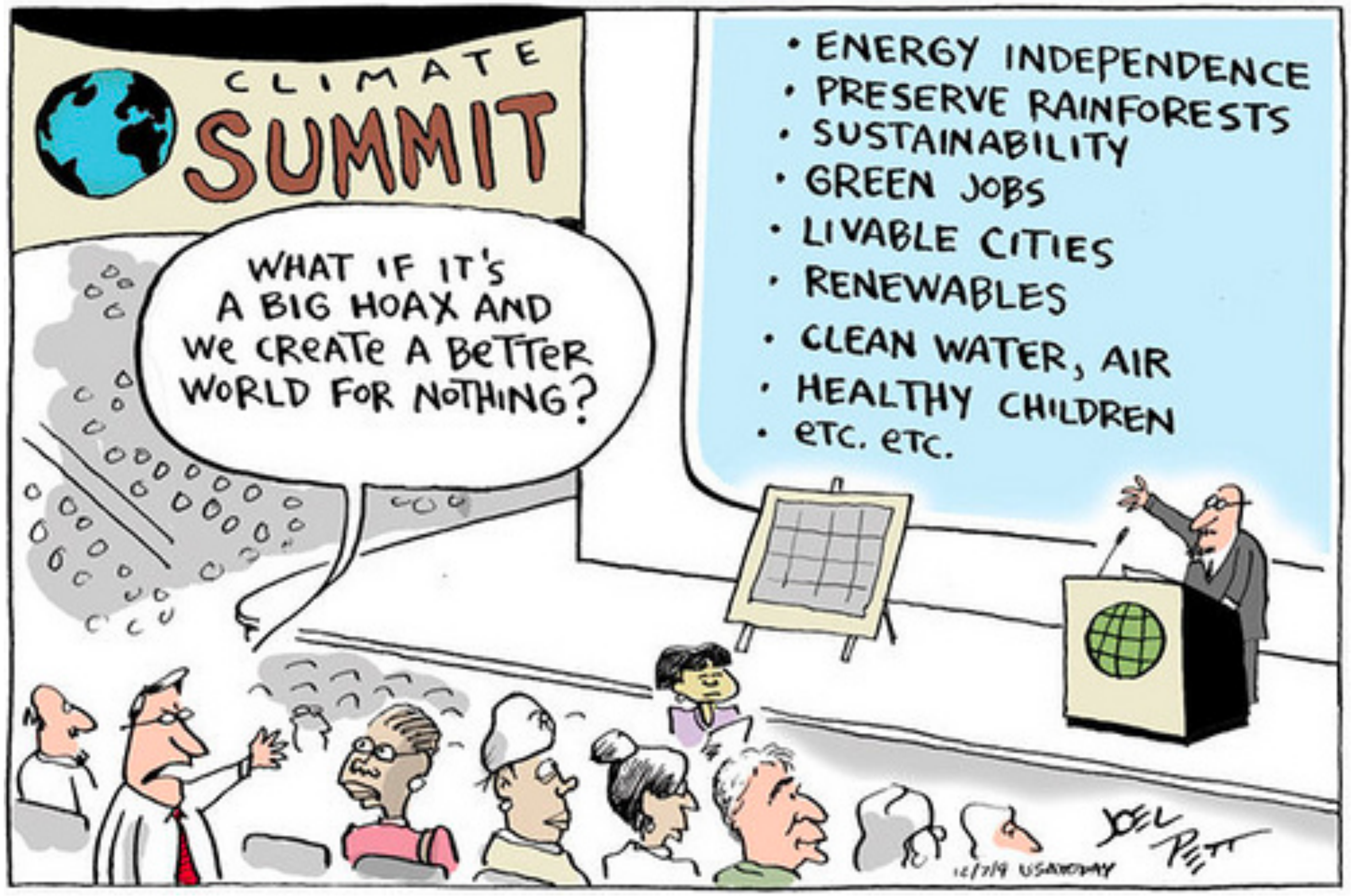
~~Live Data Products!~~

~~Whatever Titus Wants~~

How to sucker scientists into doing better software development by creating a community of practice around open science

(by producing live data products)

Maybe it’s worthwhile even if the community building & #openscience thing fails :)



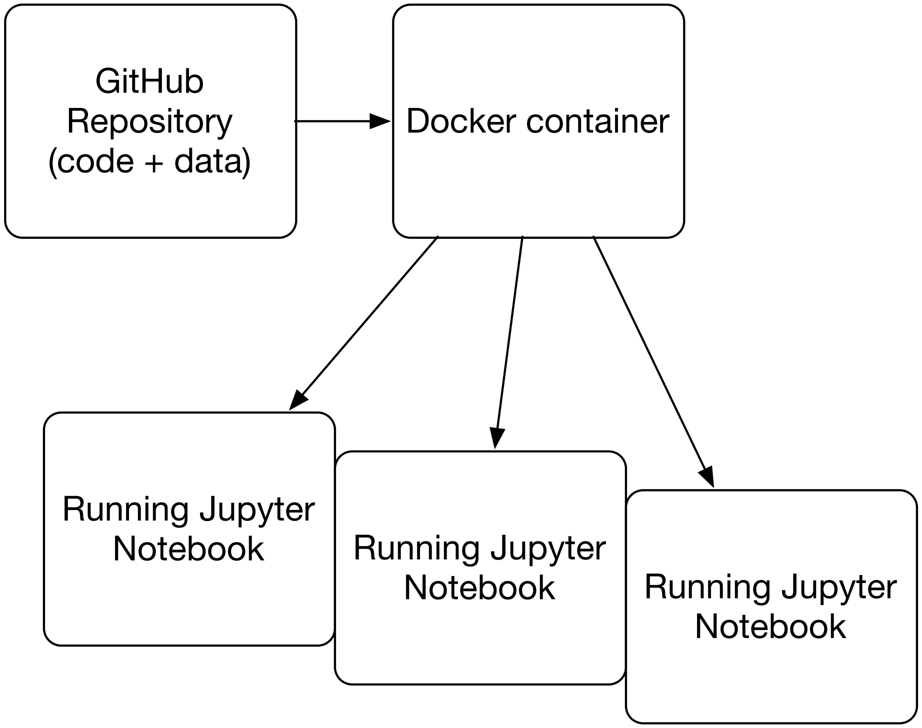
Packaging code and data -> cloud

mybinder.org demo:

- https://losc.ligo.org/tutorials/
- https://github.com/ngs-docs/2018-ggg201b/tree/master/lab0-monty-hall

Note, this all trivially supports forking, editing, updating etc.

mybinder.org: single-click deployment



Our technology for this is **rapidly improving**:

- Binderhub, JupyterHub, cloud functionality, etc are all moving at light speed*

**not literally*

- e.g. JupyterHub is being developed by Berkeley for data8.org, their intro data class; running ~1000 simultaneous Jupyter Notebooks in the cloud for undergrads!

- One missing link is **data handling**.
  - Often not practical to copy ALL the data into the image.
  - Dynamic retrieval
  - Large volumes

Some ideas

- Hook Binder / Jupyter Hub directly into data archives (e.g. Dataverse is thinking about this) – single click to code and data import, data viz, & data exploration.
- For ongoing data gathering, automatically publish or update notebooks as new data arrives, cut new releases, place on Zenodo, and generate a DOI; then reverse link from data catalogs to relevant time periods.
  - Can also use this to construct higher-granularity "versioned" catalogs; n.b. ref genomics.
- Include ”control” checklists (aux channels, known instrument events, etc.) in a single report so non-Laura/Neils can interpret.
- Publish "live"papers tying data + code + analysis (c.f. "Scientific Paper of the Future")

Many more ideas came out at binder workshop! http://ivory.idyll.org/blog/2017-binder-workshop.html

Live data products – why!?

Data's not the new oil. Data becomes *more* valuable as it's used and re-used.

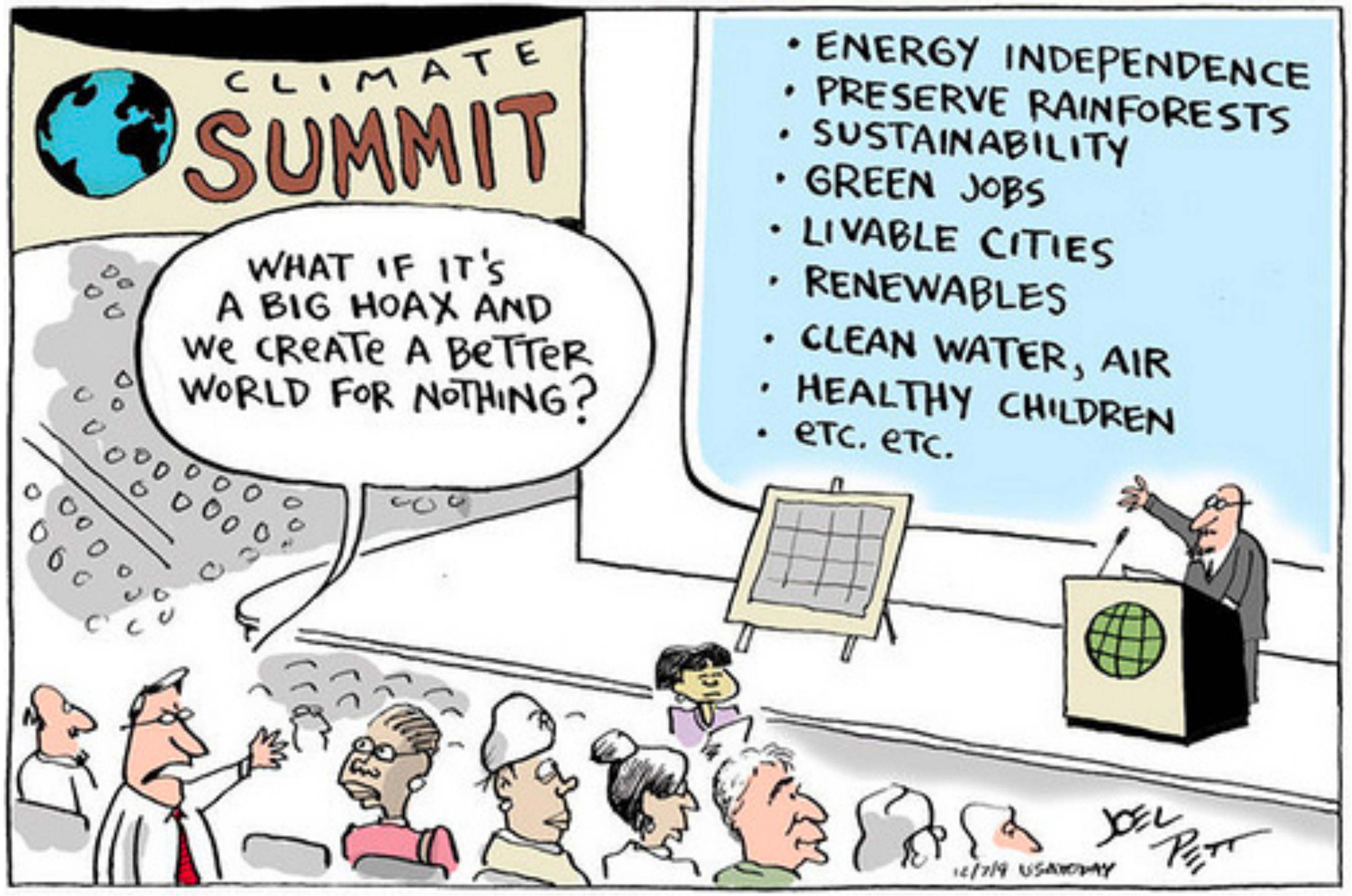
- Simplifies on-boarding of new people
- Simplifies exploratory analysis even for experienced community members
- Support ad hoc integration with other communities (e.g. EM sources)
- Supports greater participation by broader community & can be used as evidence of impact (see: NSF)
- Unleash the serendipity!
  - (Try out crazy ideas faster so they can be rejected faster)
  - The problem is that it is hard to measure missed opportunities - the "unknown unknowns."

(Incidentally, if you release all your code and make it easy to run, and run it a lot, you’ll end up with better code and better coding processes.)

Other things worth mentioning --

- The Journal of Open Source Software
  - Periodically release and review software, receive peer reviewed DOI (citation)
  - Rewards software development (maybe)

- Permissionless annotation mechanisms – see hypothes.is
  - Allow arbitrary annotation of signal, catalog by community members;
  - Just indicate it’s untrusted (and/or have spam filtering)
  - Can use to build ad hoc micro-ontologies, link between databases, etc.


CLIMATE SUMMIT
WHAT IF IT'S A BIG HOAX AND WE CREATE A BETTER WORLD FOR NOTHING?
• ENERGY INDEPENDENCE
• PRESERVE RAINFORESTS
• SUSTAINABILITY
• GREEN JOBS
• LIVABLE CITIES
• RENEWABLES
• CLEAN WATER, AIR
• HEALTHY CHILDREN
• etc. etc.
JOEL PETT
12/7/9 USATODAY