Accessibility and Jupyter Notebooks

they worked with 10s of disabled people to get the data

Accesibility in this talk they mean disabled people

Possible disabilities:
- visual
- auditory
- motor
- cognitive

Can be constant or tmeporary

Percievable - people need to be able to access all ocntnet in a notebook
 - you can use alt text
 - or audio transcription
 - text is very reliable and flexible
    - reliable - most assistive text can work with it
    - flexible - it can zoom and reflow
    - color - check the contrast; should not be the only source of information;
    - make sure your visualizations are labeled

Operable - people need to be able ot use everything in the notebook
- headings - use markdown headings
- no flashing content (no more than 3 times per second)
- understandable - explicit context
- use plain language - limit jargon

Adopt a checklist!

Try at least one recc from the talk.

---------------

Building metpy for the Long Term

it's a python toolkit for meteorology (since 2008)

used in education and operations

money is not enough to have it be sustainable. You can get lots of money towards the project and still not have enough folks to work it.

So sustainable should be - "the capacity of software to endure"

How to be sustainable?
- automate everything
- bugs inevitable
- painful task? automate!
- tests!
- keep infra in git
- use org accounts instead of individual accounts (where possible)

Use Github Actions Pipelines

CI testing - can end up with combinatorial issues if oyu need to check across lots of versions

Dependabot - multiple upstream deps change and can break you. Best to pin versions in CI to give you control over when versions update.

One bad thing about Dependabot - it doesn't understand Conda

These things make it more sustainable, but workflows don't maintain themselves. They do require work to keep up to date.

Work required to turn contributors to maintainers

------------------

keynote: Open Problems in Single-Cell Analysis is the problem used to illustrate the solutions for the keynote.

ML for drug discovery

-----------------

Diversity Luncheon Keynote

how can we protect vulnerable groups while measuring representation in our communities?

beause they are usually underrepresented - how do you avoid exposing what may the the only member of a group

Who is accountable for what data? Eg company, project, regulator

You can be direct (eg survey), indirect (scrape data), or guess (speaker does not recommend because it can lead to use of stereotypes)

Make sure oyu understand how consent is given or withdrawn? We usually don't consider how to withdraw (esp if anonymous survey)

With sampling - you have to be careful about defining the smallest sample you will share. (eg - don't share if you have 30-50 or less responses from a group)

During surveys - ask the minimal amount of questions. Also, 10-15 min total is best. If ask for too much - the dataset becomes more risky if exposed.

-----------------------

Geophysical Data

central valley of CA has very productive farmland, but severe droughts for 4 years.

Important to know whether the subsurface has clay or not because that affects whether the water drains through

they use EM imaging to try and figure this out

-----------------------

Sci Py 2 Plenary Session

Hypothesis - testing science tools is hard and so we need better tests. So hypothesis provides properties-based testing.

A section about how Python 3.11 is faster (10-60%, depending on code) Also better tracebacks. For Python 3.12 - f-string grammar being added. More parallelism (per-interpreter GIL).Tracing will get faster. Buffer protocol no longer needs a C extension.

A section about Dask.Memory management has been improved (especially if using Pandas) Lots of commercial tools for deploying dask (including K8s) Newest version compatible with Pandas 2.x.

cibuildwheel - JSON output added. Support for Pyton 3.12 beta. Can cross-compile for Windows ARM. PyPy Apple Si support.

conda ecosystem - CUDA12, new solver.

awkward array - allows for manipulating JSON data with numpy idioms. Added dask-awkward optimization for disk access.

pangeo - big data geoscience. They have regular community meetings where people give talks relative to the community. (sign up on website) There's a discouse forum as well.Pangeo Forge - a repo for ocean, weather, and climate science.

cython - Cython 3.0 RC is out now. Final release should be out in the next few weeks. Works better with C++ and is more pythonic.Type annotations have arrived.

-----------------------

Scientific and Technical Publishing

Sharing analysis is a different skillset than doing the research

Jupyter made it easier to combine data and code and make it accessible to people

Quatro is supposed to help solve the last mile of scientific publishing

It builds on pandoc markdown

It renders a Jupyter notebook which  into pdf or word (or both!)

You can publish to Confluence, Github, other places

You can create a website using Quarto and VS Code

also can  make interactive documents

----------

Subpoenas Less Scary

After a company recieves a subpoena, the company leadership has to decide how to fulfill it - completely? Partially? Fight it?

Consider a subpoena as a request for customer data (same as any other customer)

Firefox collects data on a continuum from less to more sensitive.

They have many ways to protect sensitive data. One of the things they do includes PII filtering.

Some of their requirements for PII filtering:
- Favor simplicity
- err on the side of caution
- mainatin control - do it inhouse
- low barrier to entry

An important thing is to monitor your results ot make sure your assumptions continue to hold.

----------------

Taming Black Swans

long-tail distributions are common, but violate our intuition

With his example - the first important thing was to move from linear to a log scale.

Is we see the s curve it's a better first

That above was with the x axis put on the log scale.

Then he wants to put the y-axis on a log scale. This can help expand the "tail" section, but squashes the early part of the data.

You need to have your data visualized well so that you can understand the data correctly.