Baysian Stats with Python

speaker does not prefer black box methods

an opposite of bayesian is frequentist; ignores the past

when using prior distributions you may have more confidence in some params than others.

you also have to do regularization

Slide contains a formula that shows the relationship between the prior, likelihood, and posterior

bayes_mapvar package takes a bayesian model spec and outputs the posterir mode eistmate and posterior variance estimate

It works with a TensorFlow Probability dictionary

example uses Google's Bard to get data.

Can there be too much parallelism

Immediately says yes, there can be too much parallelism.

To be more subtle - if using more than one scientific python package, they may not play well together.

vendoring your libraries is one potential solution

Disciplined Saddle Programming

domain specific programming for saddle programming

this is for convex optimization functions

CVXPY can solve these functions in a natural way

speaker shows many ways that you can take a 2 constraint problem into a 1 constraint problem

the conic standard form functions as an API

Fast Exploration of the Milky ways

source for the analysis is a CSV file

Blosc2 can be used to compress blocks of data - we're dealing with gigabytes here

Btune plugin lets you choose between speed and compression ratio

In the datasets here - most of the bottleneck is memory bandwidth, not CPU. So single-threading of Python doesn't slow things down.

Open Force Field

they make sure to be agnostic so that reserachers can work together without revealing trade

secrets

Pandera: Beyond Data Validation

speaker wanted to add types to pandas

speaker creates a pydantic-like workflow for pandas

needs more work to be fully compatible with Pandas 2.x

Thar Be Dragons: Ethical Legal Policy Challenges when Measuring Open Source

ethical challenges:

- no one signed up to be your test subject
    - when can we assume consent?
- quantitative/qualitative OSS data is usually not subject ot IRB review
- people don't readily sit in a single dimensional cluster - we could end up erasing people or reducing folks to harmful vectors
- people from vulnerable pops may separate their identities acros smultiple online communites and spaces
    - aggregating the data may "unmask" them
- does anti-aliasing datasets peotentially create opportunities for harm for members of OSS communites?

Legal challenges
- this data is "open" - can I use it?
- this data is "public" - can I use it?
- is this fair use? (this is always changing)
- which license for what?

policy challenges:
- when does a foundatoin speak for a project? A maintainer? A community?
- can foundations "opt-in" communities an dprojects into ecosystem scale research?
-