Stop writing papers and start creating a knowlege base

Paul Billing-Ross

December 13, 2022

VA Big Data Genomics Group

A Markdown presentation made using Marp

How do you make research repeatable?

- Write down the instructions.

Why is it so hard to write research instructions?

- Research is complex and there are a lot of...
  - Steps
  - Tools
  - Environments
  - Variables
- It's time consuming and tedious
- We wait until we have to write the Methods section
  - Knowledge has been forgotten
- It's hard to describe everything we do because once we learn something, we do it without thinking about it
  - Tell 'em the centrifuge story
- Research is dynamic, it's always changing

How do we write better (bioinformatics) instructions?

- Use computational notebooks to organize you analyses
  - e.g. Jupyter notebooks
  - Code + results + description all in one place
- Don't scimp on the description.
- Don't forget to document the computing environment:
  - How many CPUs/memory/disk did you allocate?
  - Which operating system did you use?

How can we use technology to do repeatable bioinformatics?

- Use **computational notebooks** (e.g. Jupyter notebooks) to organize your code, results, and descriptions and interpretations all in one place.
- Use **version control** (e.g. Git/GitHub) to track changes. Research moves fast and methods change quickly. Use version control to create checkpoints of your research progress.
- Use **project management software** (e.g. GitHub Projects) to track research progress, changes, and issues.
    - Generate a record of how things went sideways
    - Provide transparency to all research members

Hire specialists

A general problem with academic research is that it is done by small research labs that don't have the resources or will to hire specialists. Your research program should have someone who specializes in...

- Project management
- Collaborations
- Information technology
- Bioinformatics
- Statistics
- Data management
- Graphic design
- Web development
- Writing/editing

Don't do any analyses manually

- Program a computer to do all your analyses
  - Excel is great, but don't ever use it, it's not repeatable
  - But I just want to analyze some reaults real quick
    - *No.*
- Use Python
- If you are going to be a baby about it, use R
- If you are going to use Rust or Scala or something you better have a good reason because nobody else knows how to program in Rust or Scala
- Don't ever use Perl.
- Don't mix languages.

Repeatability is a process not a feature

- A feature can be tacked on at the end when you are ready to publish your paper
- A process is something you adhere to every step of the way

Show 'em GitHub

- Version control
- Project management

The most valuable documentation is examples

- "Show, don't tell (but also, do tell)"
- Show all the examples, not just the pretty ones.
  - Nulls
  - Failures
- Science is mostly failures and nulls, but we most discussion is about the positives. That's a lot of knowledge lost!

A knowledge base lowers the bar for scientific contribution

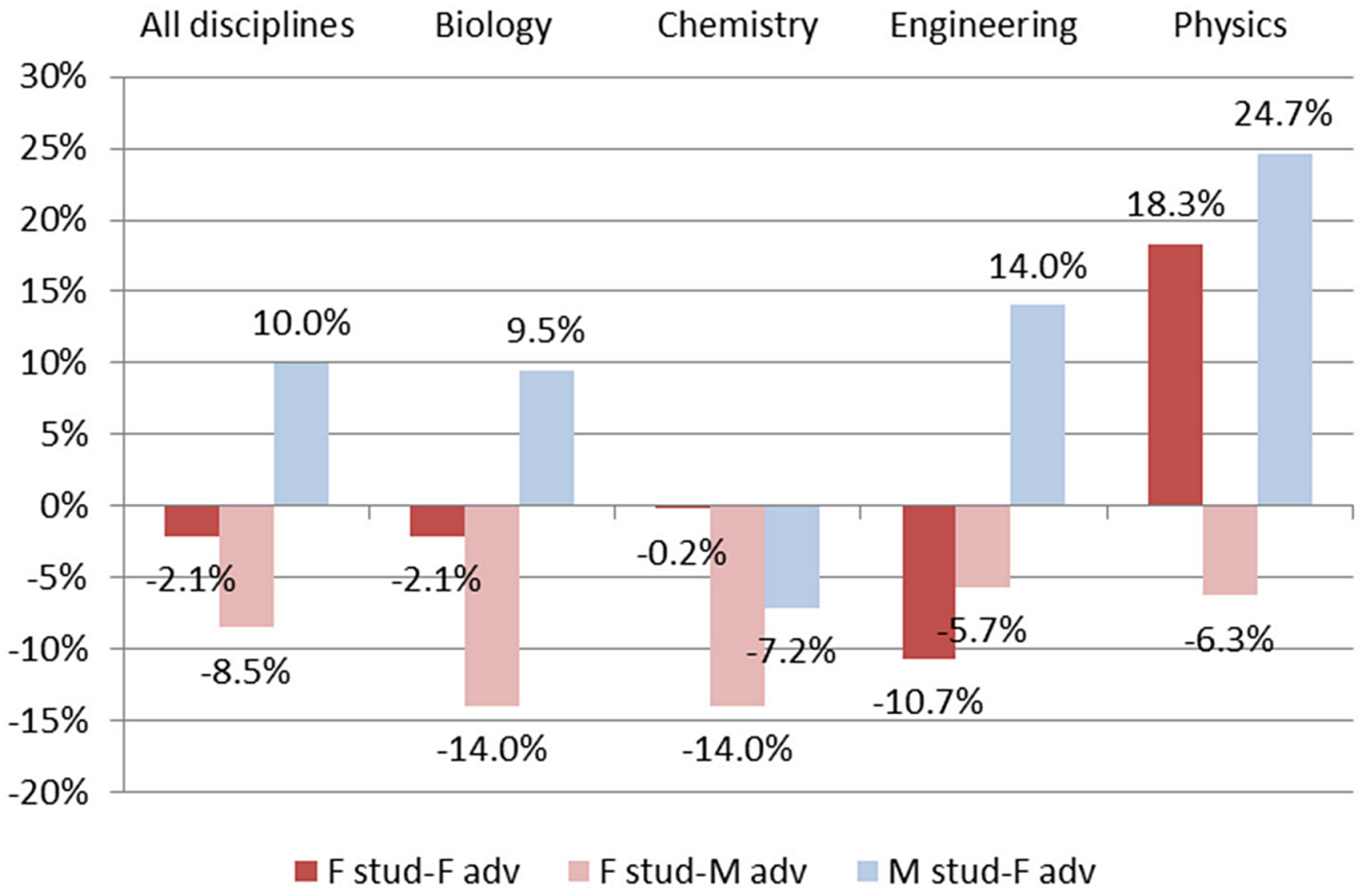
- And that's a good thing.
- Everyone who participates in science wants to contribute
- Who can make first/last author contributions to a journal article?
    - (Senior) graduate students, post-docs, professors
- Who can answer a question and add it to a knowledge base?
    - Rotation students, interns, undergraduates, high school students
- Publishing journal articles is hard. Lots of researchers struggle to publish.

How should we view researchers who don't publish?

- These researchers spent 5-8 years doing research and made no contribution to science.
    - Why were they awarded a doctorate?
- Using journal articles as the sole method of conveying knowledge and measuring researcher contributions is flawed & inequitable.

Inequality of journal publications by gender

Gender and the Publication Output of Graduate Students: A Case Study



What are the advantages of using text as a medium?

- Anyone can contribute
- Easy to search (command-F, Google, AI)
- Convert to other formats
  - **slide deck**, PDF, html, website, visualization, speech
- Track changes & contributors in version control
- Transformed to fit different use cases, like finetuning a **AI language model**

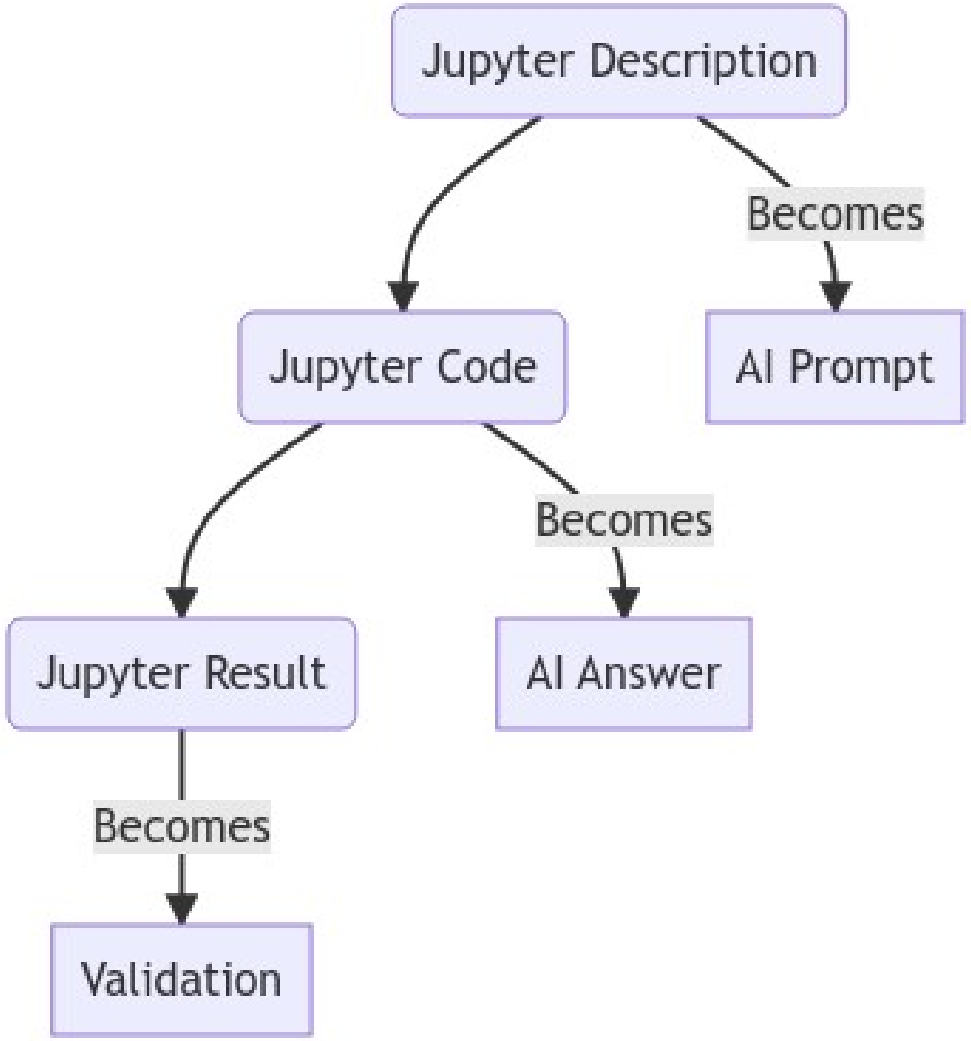
Communicating with humans is over, only talk to computers

- You aren't going to be compelled by any of my reason. Incentives aren't there.
- AI is going to make science repeatable
- Ways of doing science:
  - Small science: Do it yourself, with your hands
  - Big science: Program a computer to do it
  - A lot of big science: Teach a computer to program it

Why is that going to make science repeatable?

- Computers are dumb, they don't understand things (yet)
- But they can recognize patterns
- Before you can get them to run your analyses, you need to train them
- Train them on different combinations and permutations of an analysis
- Validate your model on a test set
- Then the computer performs the analysis x1000
    - And doesn't make mistakes*
    - And doesn't forget*

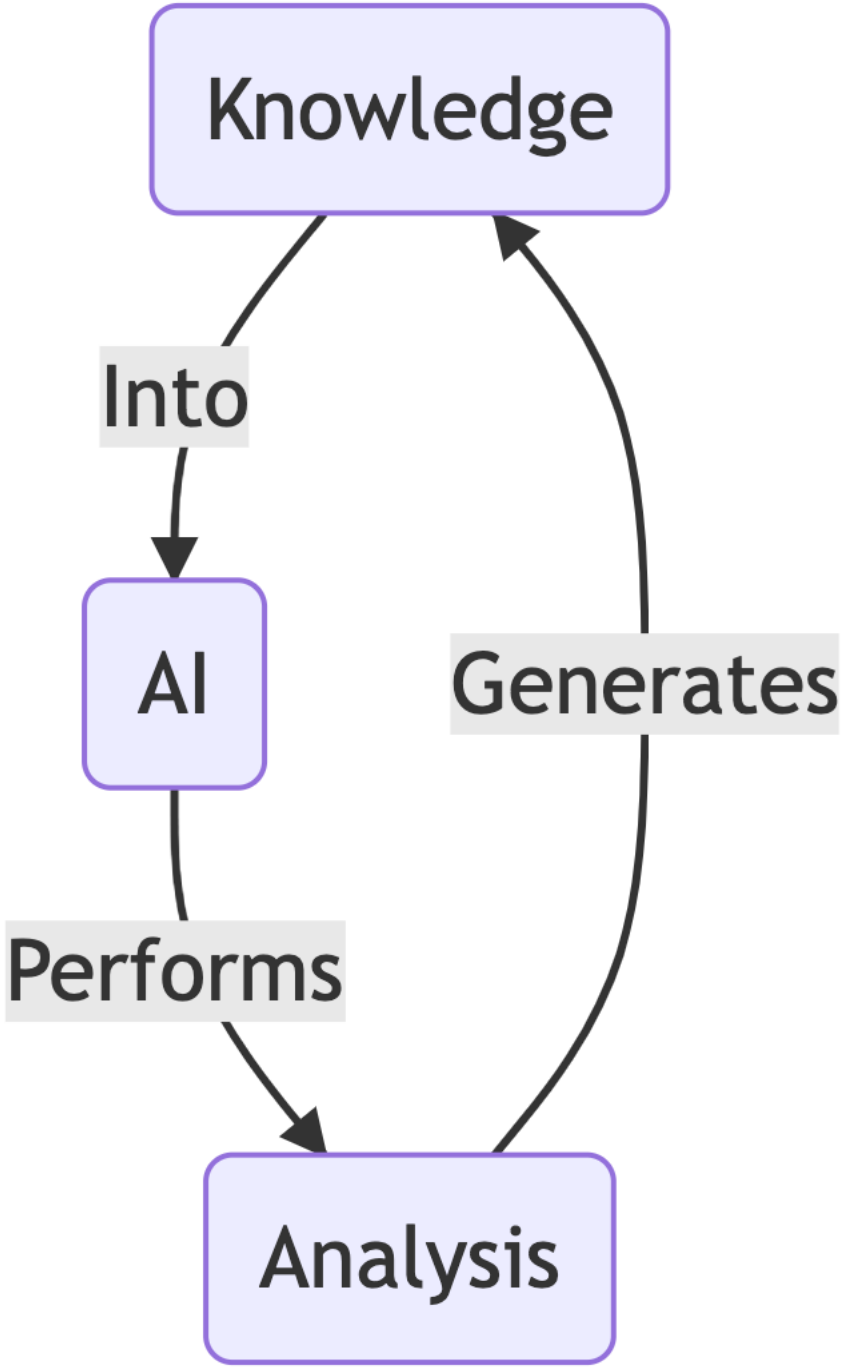
Jupyter as a structure for training AI



Get in early

The Goal: Create a feedback loop of knowledge and analysis

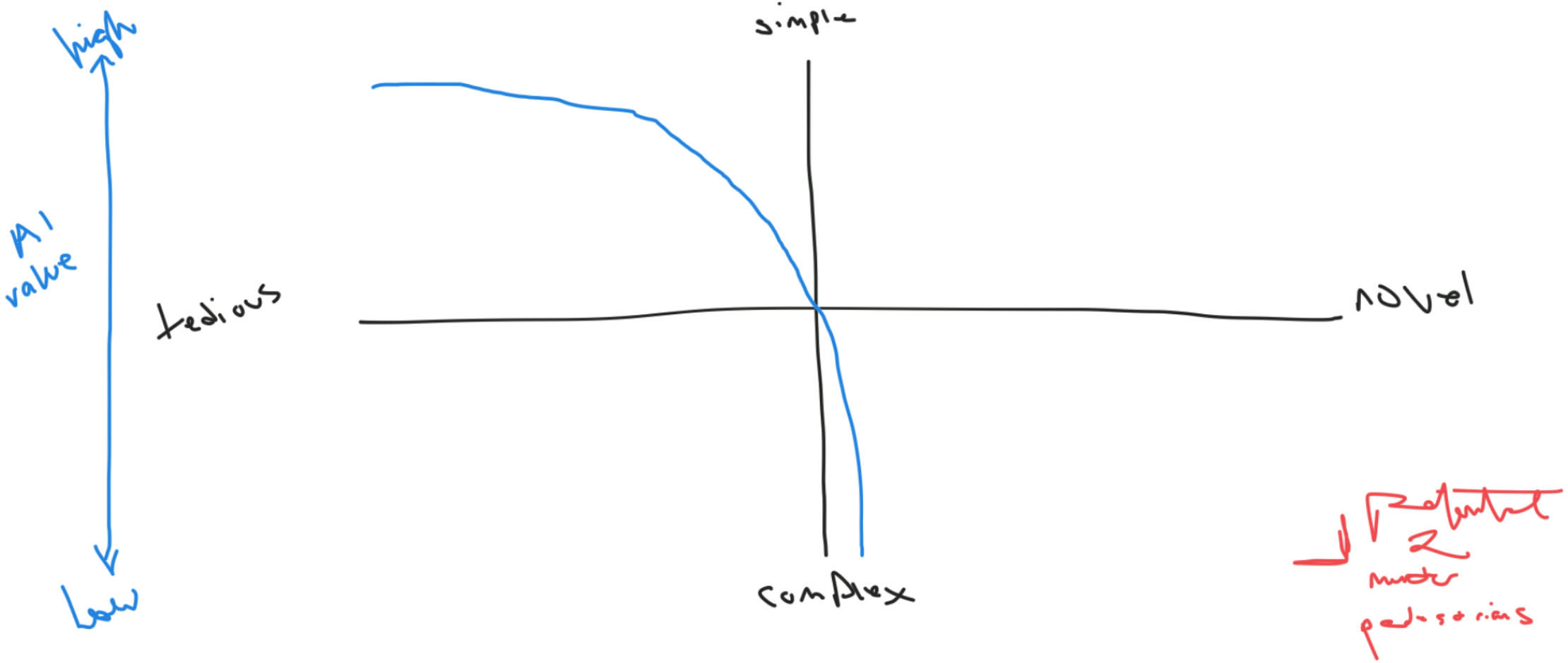
1. Add knowledge to your knowledge base/text file
2. Train an AI to run analyses described in the knowledge base
3. Cajole the AI to perform more analyses
4. Generate additional knowledge
5. Repeate step 1



How can I generate value from an AI?

Show 'em. [DRAWING TIME]

How can I generate value from an AI?



What am I going to use an AI language model for?

- Answering questions
- Describing protocols
- Troubleshooting errors
- Essential lab minutia
    - How should samples be organized in the freezer?
    - Where are buffer reagants stored?
    - Where is the document for scheduling lab meetings?

How should I go about creating a knowledge base?

- Create a text file and check it into version control.

Show 'em. [DEMO TIME]

Model for building a knowledge base