BEYOND DATA STANDARDS

Standardizing & Automating Your Curation Workflows
...Through the Lens of Bioinformatics Workflows & Biomedical Data Curation

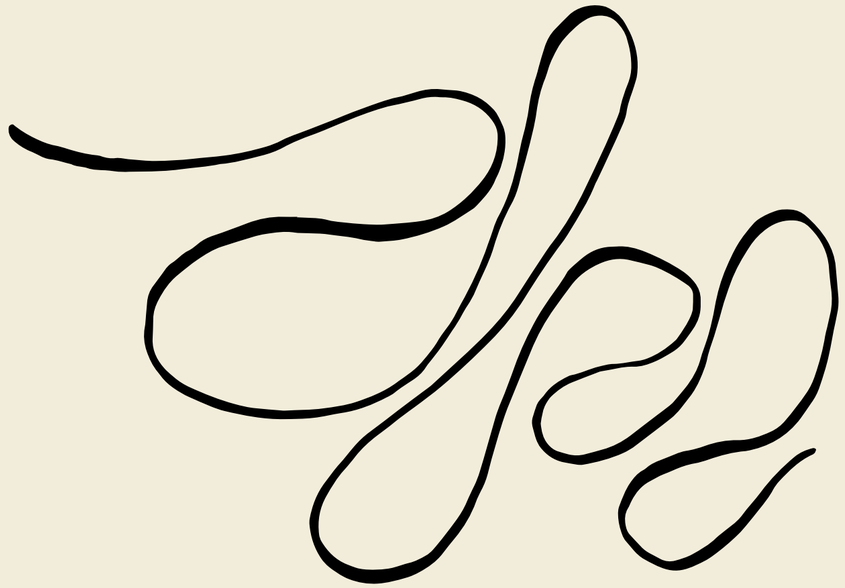
Guest Lecture for LIS 546 (Spring 2021)

Presented by: Kaitlin Throgmorton, MLIS

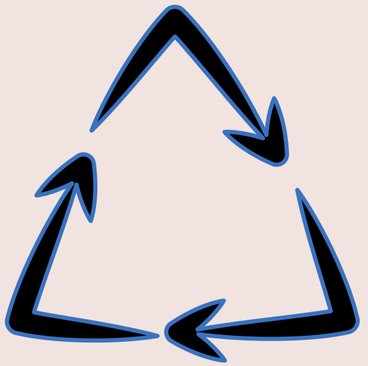
Outcomes:

- **Understand** workflows as both a *concept* and *language*
- **Articulate** the *importance of workflows* in bioinformatics, and *explain how they affect metadata*
- **Spark** an interest in *workflows*, *automation*, *and standardization* beyond (meta)data standards

What **is** a workflow?



A routine?

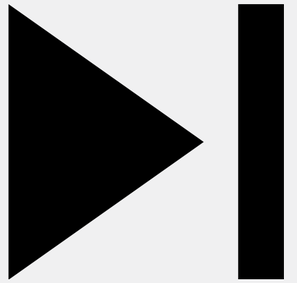


A lifecycle?

Just how **work** ... **flows**? Or gets done?

Pause for a moment.

Consider: did ***you*** execute a workflow today?



Maybe you:
1) made food or a drink?
2) went through a morning routine?
3) completed a series of school-related tasks?

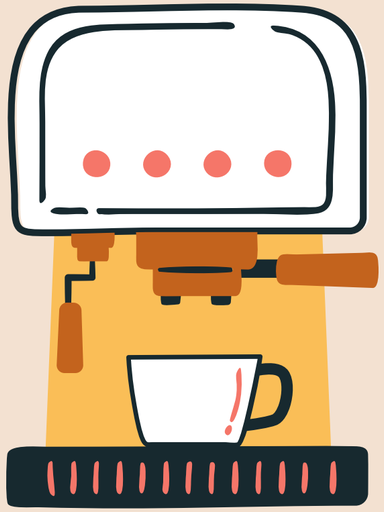
Let's consider **coffee** as an example.

COFFEE- MAKING WORK- **FLOW**

- **input**: currency
  output: bag of coffee

- **input**: whole coffee beans
  output: ground coffee

- **input**: ground coffee
  output: beautiful black liquid (aka, coffee)



What **is** a workflow?

"a sequence of operations to complete a process"

(Mallawaarachchi, 2018)

"using the outputs of one process, as the input for another, and chaining these steps together to form one large process that can be executed as one flow — ideally through one computational command"

Why do we need workflows?

Short answer:

NGS.

Long answer:

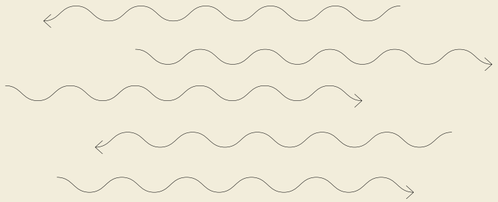
Many reasons.

We'll get there.

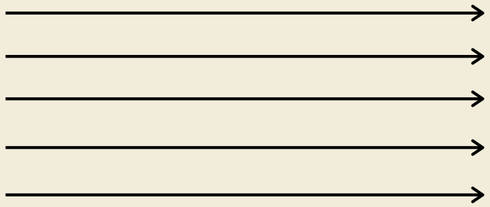
Back up.

What's **NGS**?

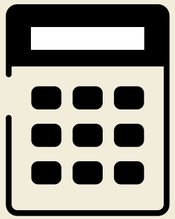
(CDC, n.d.; Gertner, 2021)



unaligned



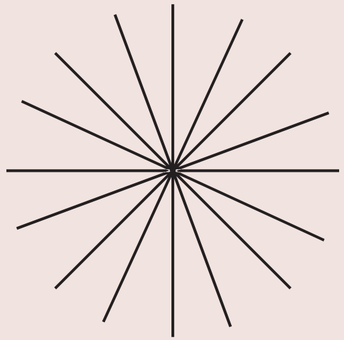
ALIGNED



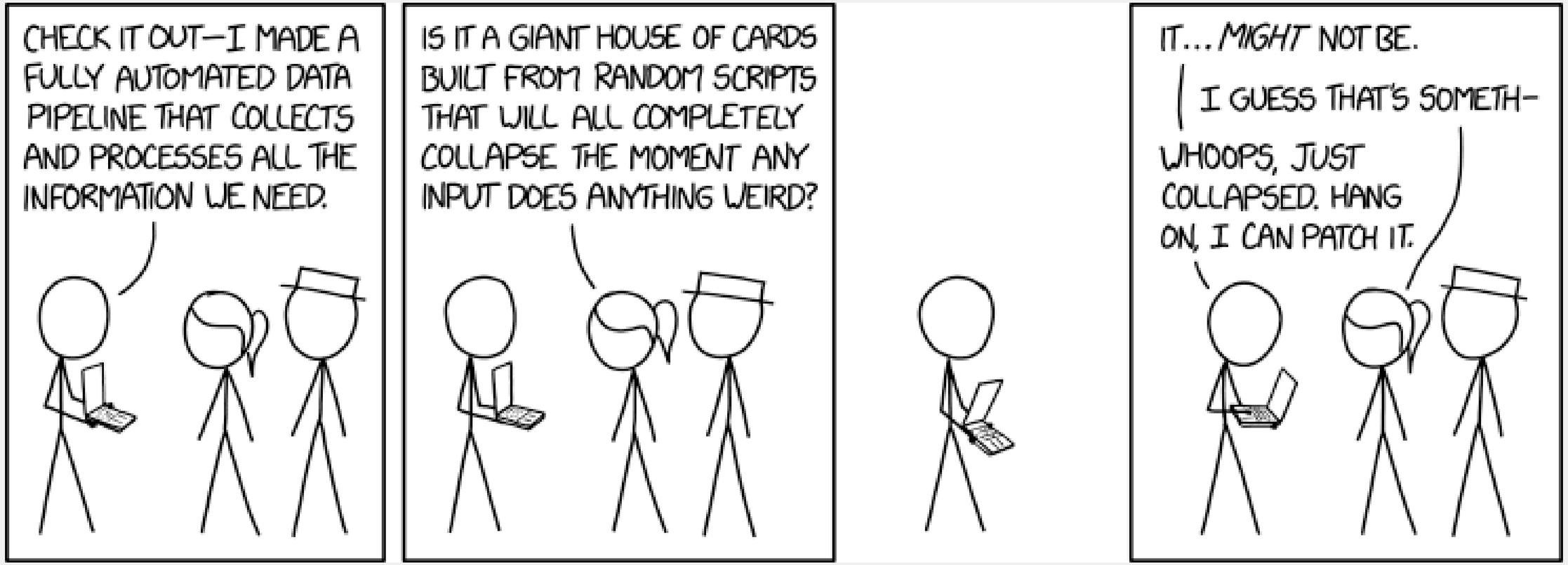
COUNTS

(Behjati & Tarpey, 2013; GDC, n.d.; Starmer, 2017)

These steps get managed with **workflows**.



Either a series of scripts and programs connected with a workflow language, or using a full system.



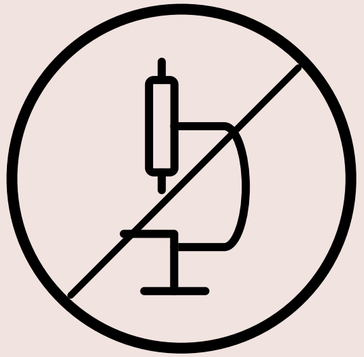
https://xkcd.com/2054/

BIO-
INFOR-
MATICS
WORK-
FLOWS

- Fully open standard languages: **CWL** & **WDL** (widdle)
- Languages built on open source: **Nextflow** & **Snakemake**
- Workflow management systems: **Taverna**, **Galaxy**, **Unipro UGENE**

[see 'workflow tools' slide at end for links]

Cool.
But... I'm into
library science, not
science science.



I get it.
But this has
applications for
data curation
at large.

5

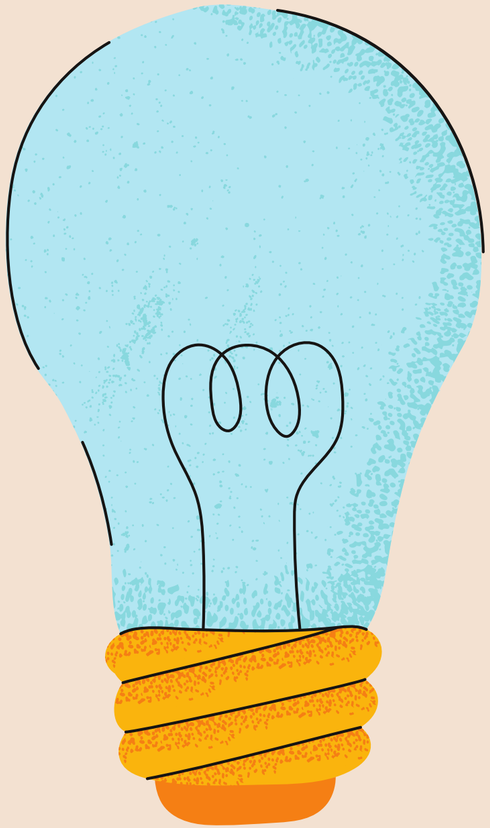
Workflow Implications for Data Curators

1:

Data quality analysis metrics (i.e., **metadata attributes**) and data quality rubrics (i.e., **intake protocols**) are increasingly in the data curator's domain.

2:

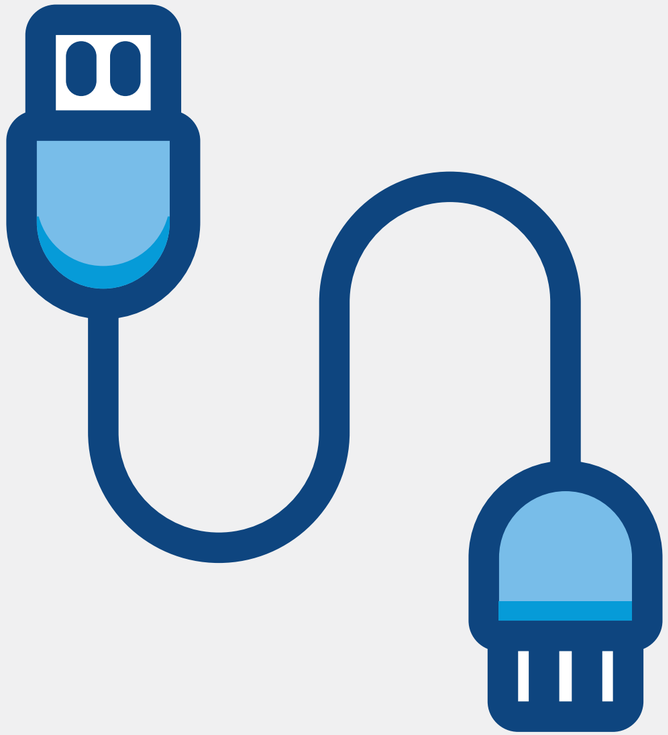
Standardization, and automation, of **more than just data** — but also the entire process and workflow used to create and process it — is needed.

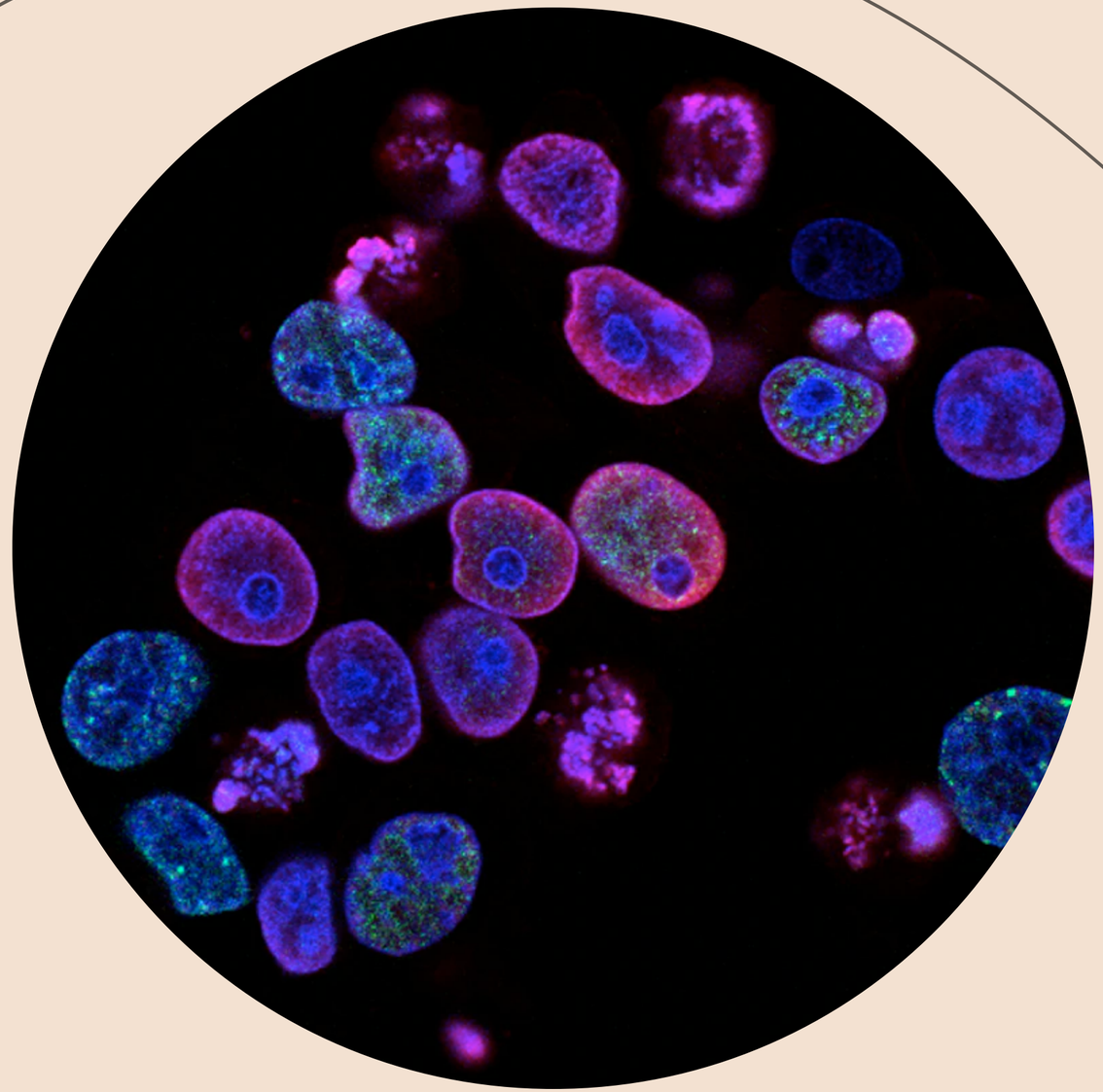


This can also **increase transparency**, **reveal bias**, and **promote more equitable practices**.

3:

Workflows add **portability** and create **interoperability** — both of which align with FAIR principles.





 # 4:

Workflows are also used to verify published results in bioinformatics — highlighting the importance of **data validation in curation**.

 # 5:

Data curators can assist with **good documentation** and SOPs, as well as small scripts, and chains of scripts — **bash** is especially useful for this.

= standard operating procedure

FOR REFLECTION

What workflow could you start to standardize today?

How can you apply the concept of portable, repeatable, standardized workflows to your data curation work?

How can published workflows reduce data bias?

What would a data curation workflow language look like? What kinds of commands would you want to execute?

REFERENCES

Citations and acknowledgements.

Behjati, S. & Tarpey, P.S. (2013). What is next generation sequencing? *Archives of Disease in Childhood: Education & Practice, 98*(6):236-8. doi: 10.1136/archdischild-2013-304340.

Centers for Disease Control (CDC). (n.d.). COVID Data Tracker: Genomic Data Surveillance. https://covid.cdc.gov/covid-data-tracker/?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fcases-updates%2Fvariant-surveillance%2Fgenomic-surveillance-dashboard.html#datatracker-home

Genomics Data Commons (GDC). (n.d.). RNA-Seq Alignment Workflow. National Cancer Institute. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#rna-seq-alignment-workflow

Gertner, J. (2021). Unlocking the COVID Code. New York Times. https://www.nytimes.com/interactive/2021/03/25/magazine/genome-sequencing-covid-variants.html

Mallawaarachchi, V. (2018). Bioinformatics Workflow Management Systems. *Towards Data Science.* https://towardsdatascience.com/bioinformatics-workflow-management-systems-cc3edd97be79

Starmer, J. (2017). A gentle introdution to RNA-seq. StatQuest. https://www.youtube.com/watch?v=tlf6wYJrwKY

Thanks to: Bruno Grande, who answered some of my questions about bioinformatics workflows.

Note: All images free from either Canva or Unsplash.

WORKFLOW TOOLS

All the workflow things.

Languages & Tools:

- Common Workflow Language (CWL)
  - CWL User Guide
- Workflow Description Language (WDL)
- NextFlow
- Snakemake

Command Line / Bash:

- PLOS Comp Bio | Ten Simple Rules for Getting Started with Command Line Bioinformatics
- Jeroen Janssens | Data Science at the Command Line
- Kade Killary, Medium | Command Line Tricks for Data Scientists

BIOINFORMATICS RESOURCES

For those interested in knowing more about bioinformatics, biomedical data curation, and/or genetics/genomics.

To better understand parts of this lecture:

- A Gentle Introduction to RNA-Seq | StatQuest | YouTube
- *The Gene: An Intimate History* | Siddhartha Mukherjee (ISBN: 978-1432837815)

In general:

- Bioinformatics for Beginners | Coursera
- Biostars | A bioinformatics forum
- OMGenomics | YouTube channel about bioinformatics
- *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome* | Alondra Nelson (ISBN: 978-0807033029)
- *A Brief History of Everyone Who Ever Lived: The Human Story Retold Through Our Genes* | Adam Rutherford (ISBN: 978-1615194186)

LEARNING LAB

Want to try some of this out? This is a bit of a shameless plug*, but you might want to tinker around in Synapse, where you can access open biomedical data, and use our various programming clients for command line, Python, and R.

*__Full disclosure:__ I work at Sage Bionetworks, a health and biomedical data nonprofit which built Synapse.

Access open biomedical data and try out programmatic clients with Synapse:

- Synapse
  - Docs
  - Clients
- Data portals built on Synapse:
  - Alzheimer's Disease Data Portal
  - Neurofibromatosis Data Portal

Or: document your own data curation workflow for your class project:

- Identify a part of your project that has a series of steps that could be repeated by someone else and document them
- Work with a partner to ensure its repeatable for and sensible to someone other than yourself!

ALL MATERIALS FROM THIS LECTURE ARE AVAILABLE IN THIS GITHUB REPO:

HTTPS://GITHUB.COM/KTHROG/LIS-546-GUEST-LECTURE

Feel free to reach out to me with questions at kaitlin@kaitlinthrogmorton.com.

P.S. We're hiring for **data curators** and **data analytics interns** at Sage!

https://sagebionetworks.org/open-positions/

SageBionetworks