Data Workflows

ELIZABETH WICKES, DATA CURATOR

RESEARCH DATA SERVICE

UNIVERSITY OF ILLINOIS URBANA CHAMPAIGN

RESEARCH DATA SERVICE

Workflow Workshop Goals

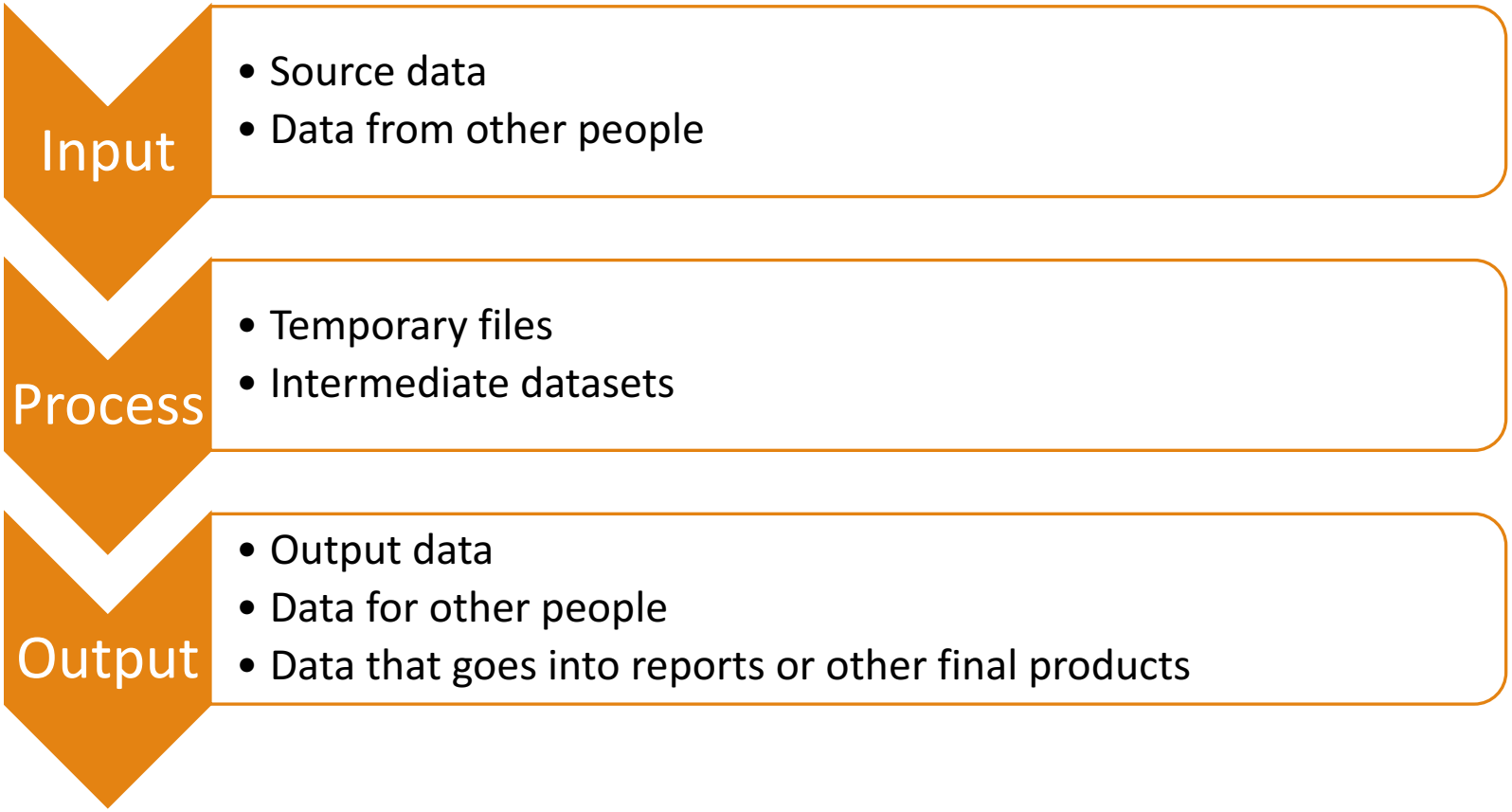
- Know
  - the tools you use
  - the stuff you use
  - where it all lives
  - where it all goes
- Learn
  - How your project workflow works
  - Points where you need clarification
  - How your collaboration with others could be improved
- Practice
  - Mapping out your workflow

RESEARCH DATA SERVICE

Materials

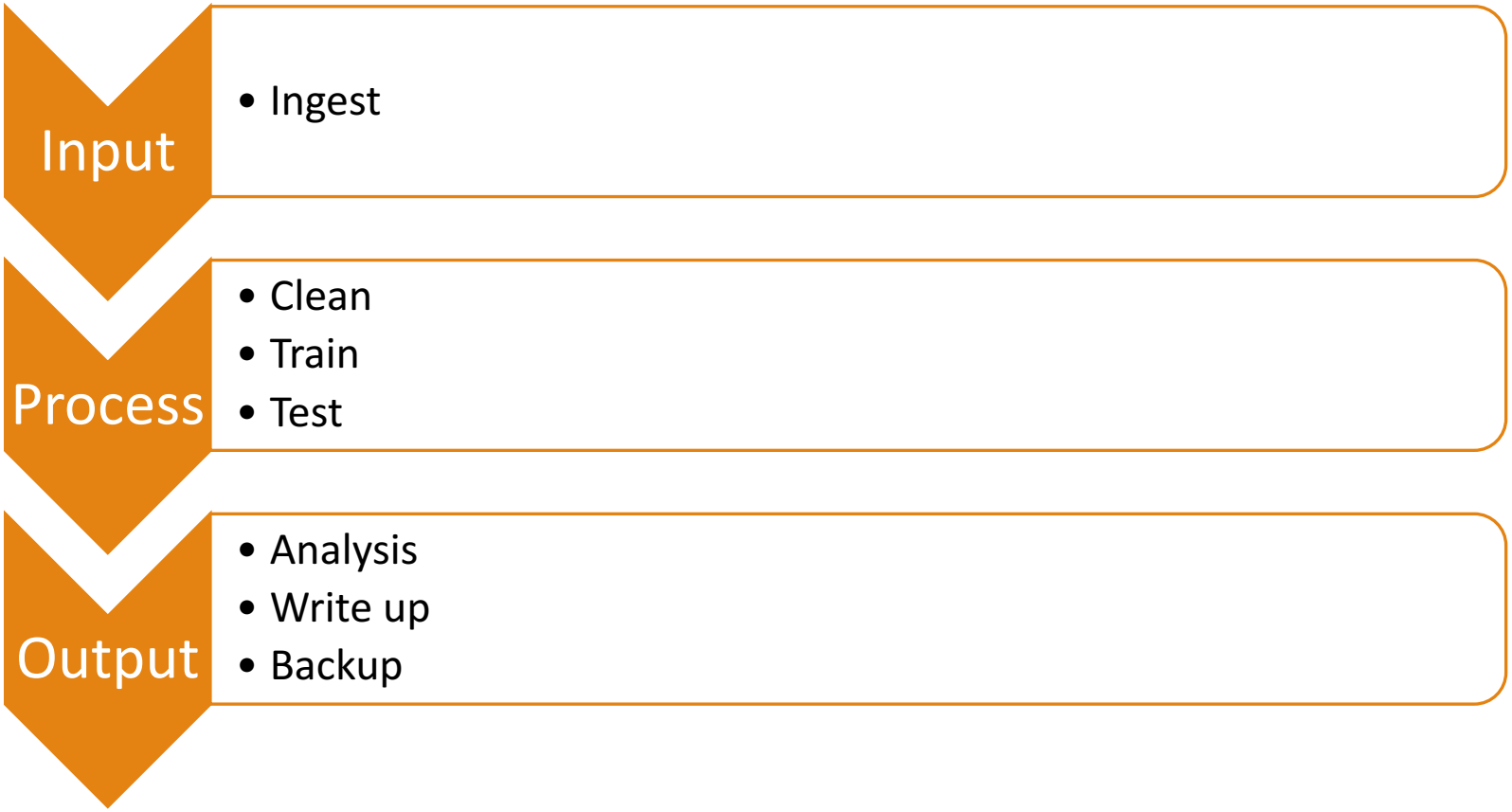
- Preferred:
  - A few pieces of paper
  - A pencil and/or pens in several colors
  - Post it notes in as many colors as you can find
- Minimally:
  - A piece of paper and a writing instrument
- Alternatively:
  - Your imagination

What data do you have?

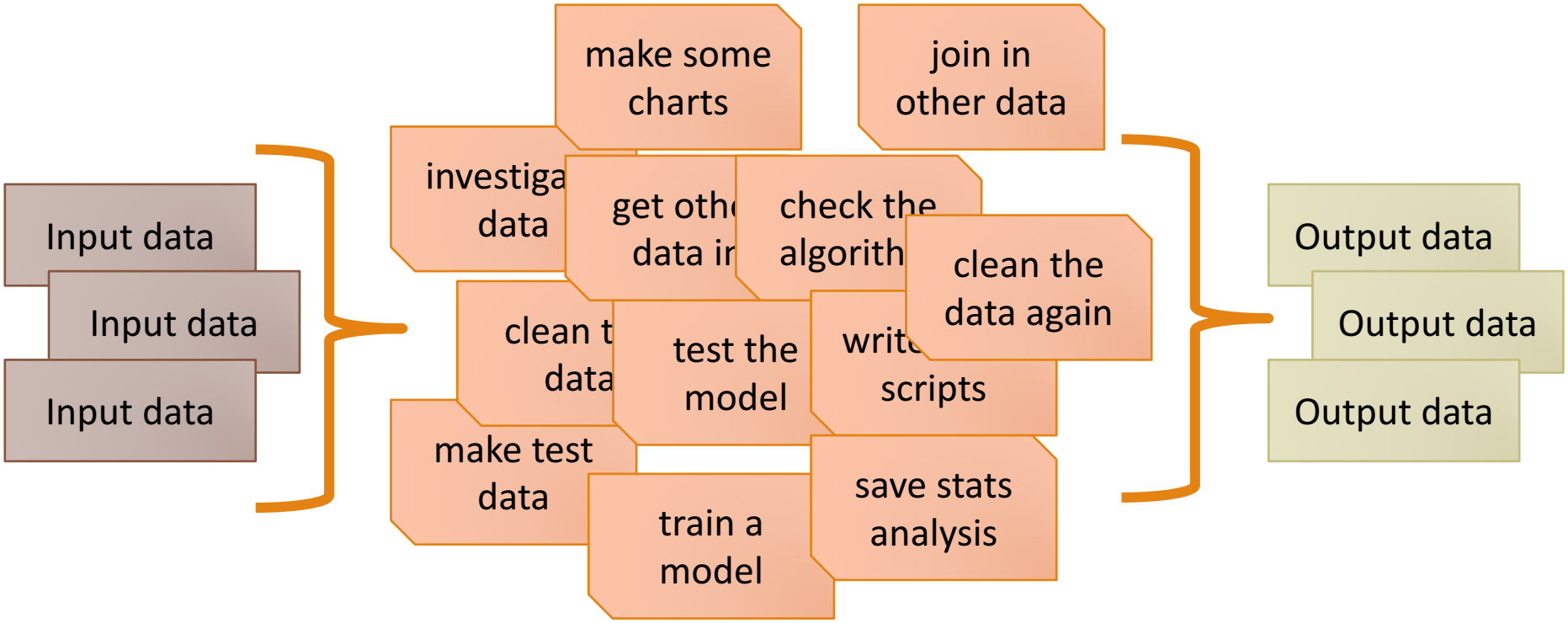


RESEARCH DATA SERVICE

And what do you do to it?

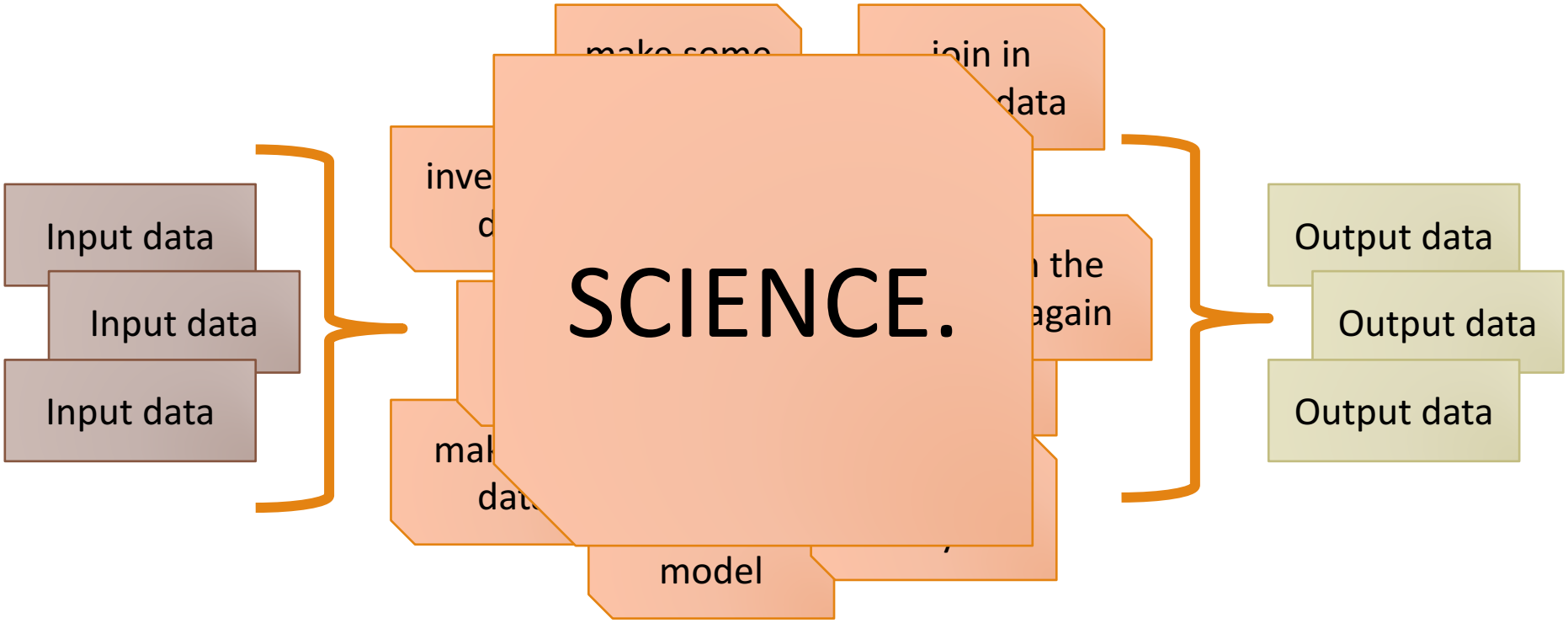


So how do you science?



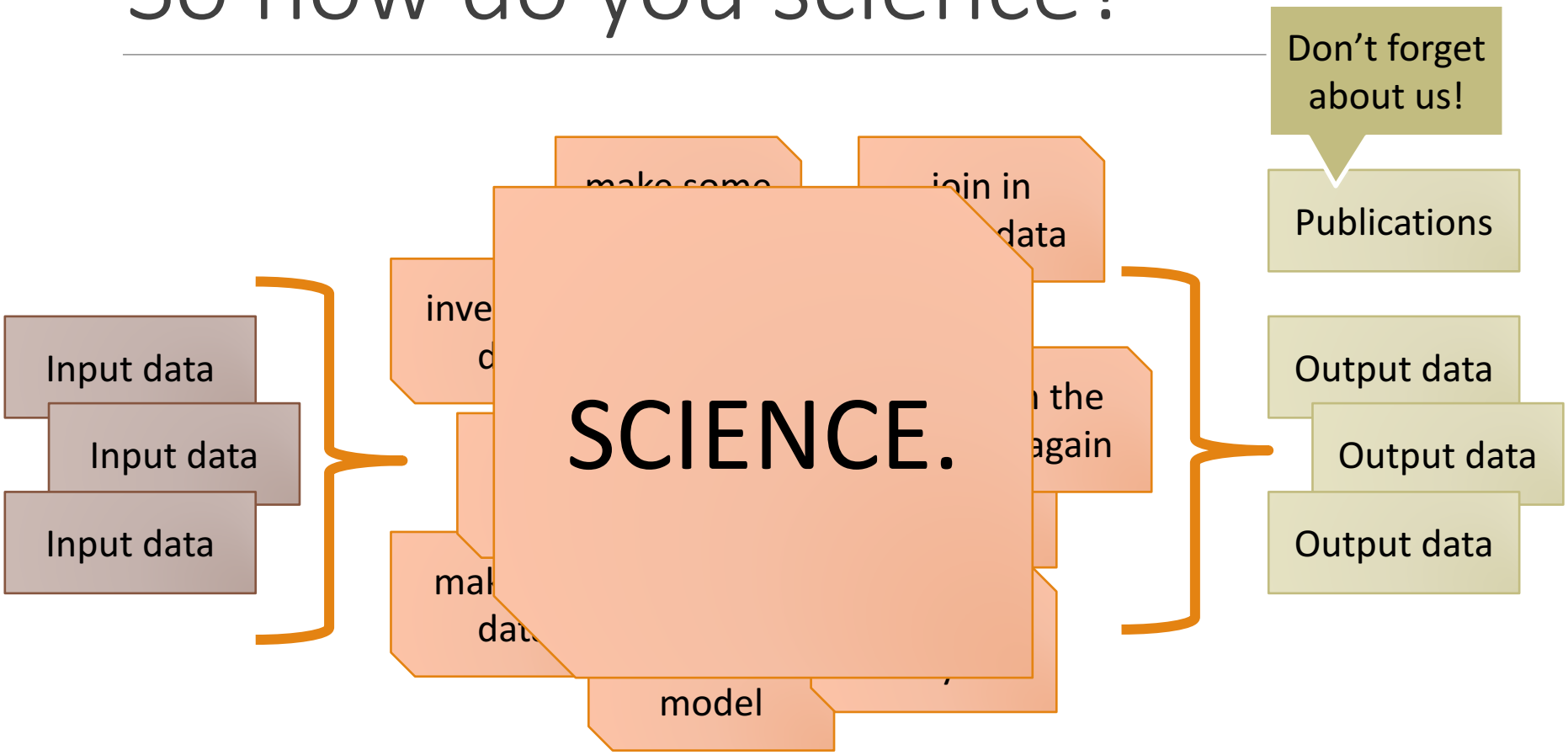
RESEARCH DATA SERVICE

So how do you science?



RESEARCH DATA SERVICE

So how do you science?



RESEARCH DATA SERVICE

But what do I do?

- We’re going to cover an activity to help you think about your projects
- Can be used **prospectively**
  - to help plan
- Or **retrospectively**
  - to pick up the pieces

Choose a project

- Something you're just wrapping up?
- Something you're in the middle of?
- Something you're planning for next year?

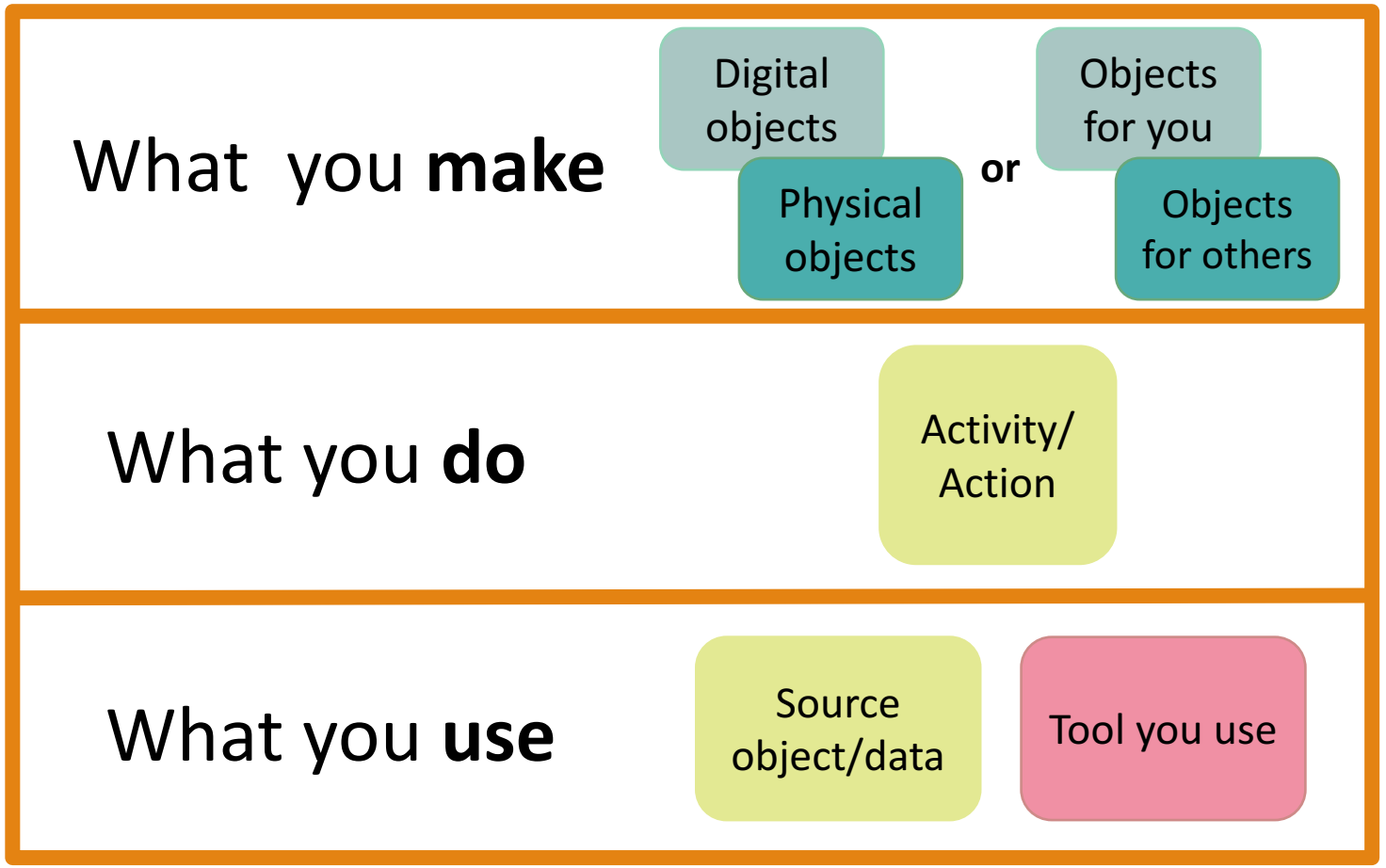
Activity: Workflow Map

- The intention is not to capture every detail of your workflow, but to help you get a feel for the big picture and points where you may need clarification or other help.
- Default to thinking very high level and generalized
- Remember to use specific, short, and meaningful names you'll understand 6 months from now

Approaching an initial workflow

- Think about these 3 questions:
  1. What kind of evidence will help answer your research question?
     - Be as specific as possible, but don't be afraid to generalize at this stage.
  2. What will you do?
     - Use verbs: read, write, script, compute, process, document, etc.
  3. What will you make?
     - Use nouns or named entities: numbers, words, data, graphics, articles, metadata, databases, etc.

The Board & the Pieces



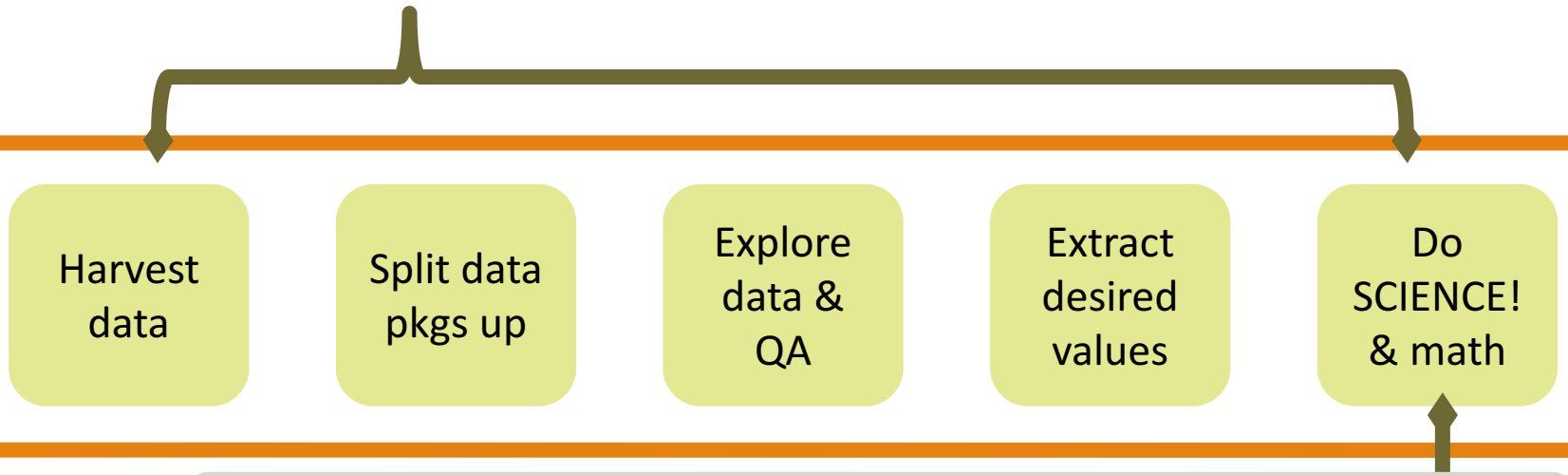
RESEARCH DATA SERVICE

Make this your own

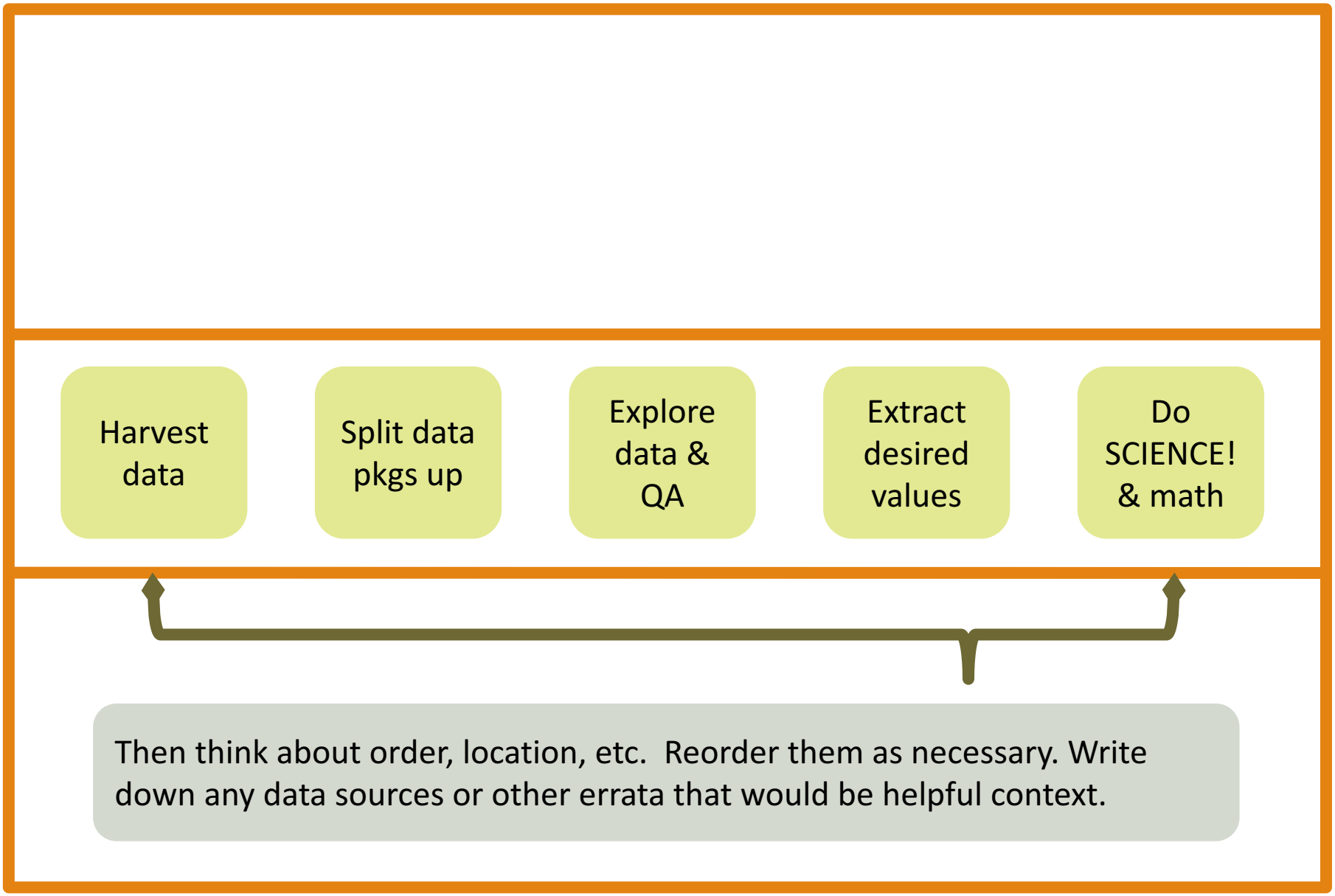
- You know what you do best
- Use your own voice and words
- Just be sure you'll be able to understand them later
- So document your changes, maybe?

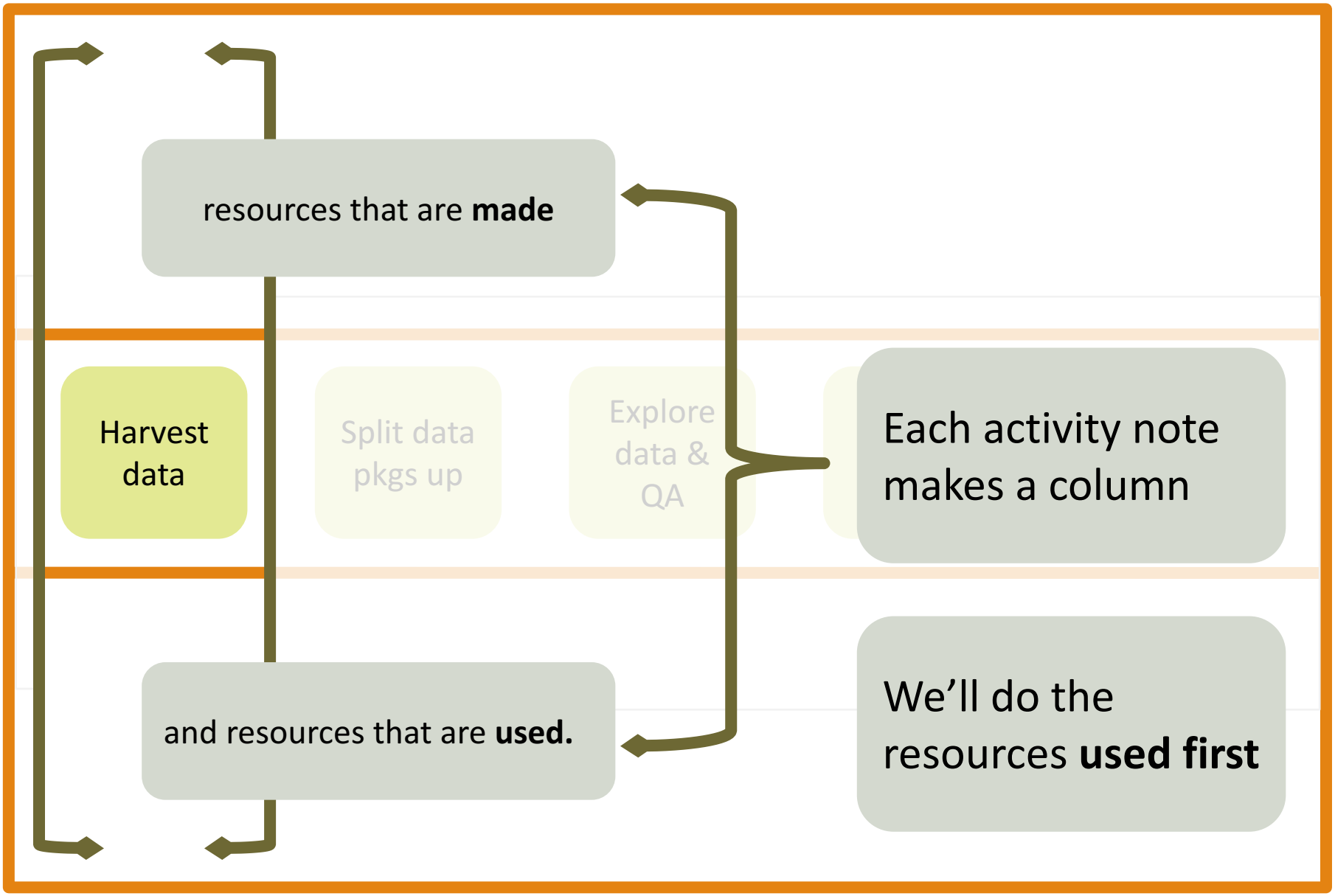
Start with your activities:

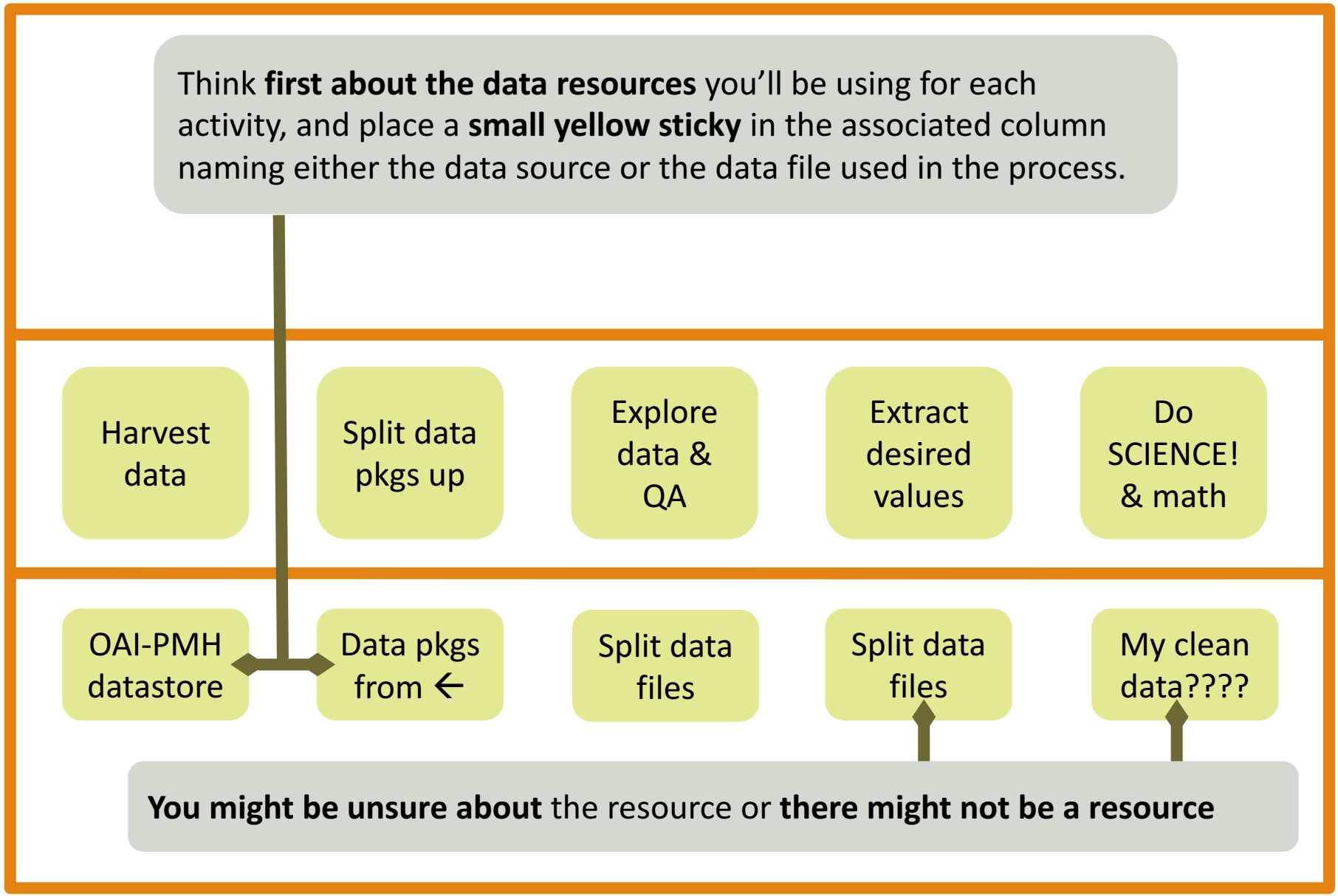
lay out about 5-7 big yellow stickies in a row in the center, and write down **what you will do** – action statements please



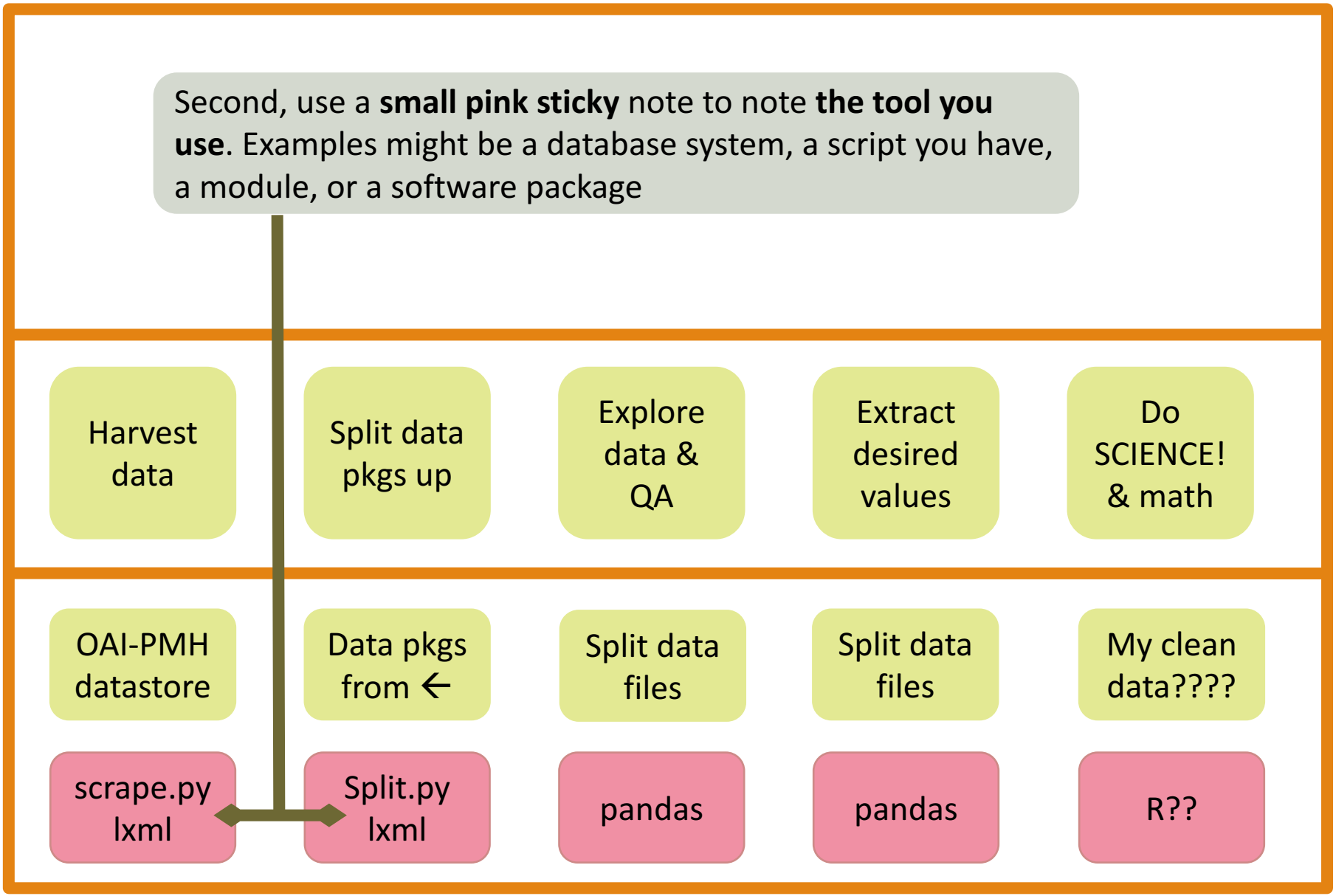
Fine to be very general about activities. The point is to note that you'll do them! Also fine to end your workflow at a meaningful breakpoint.


Harvest data
Split data pkgs up
Explore data & QA
Extract desired values
Do SCIENCE! & math
Then think about order, location, etc. Reorder them as necessary. Write down any data sources or other errata that would be helpful context.


resources that are **made**
Harvest data
Split data pkgs up
Explore data & QA
Each activity note makes a column
and resources that are **used.**
We'll do the resources **used first**


Think **first about the data resources** you'll be using for each activity, and place a **small yellow sticky** in the associated column naming either the data source or the data file used in the process.
Harvest data
Split data pkgs up
Explore data & QA
Extract desired values
Do SCIENCE! & math
OAI-PMH datastore
Data pkgs from ←
Split data files
Split data files
My clean data????
You might be unsure about the resource or **there might not be a resource**

Second, use a **small pink sticky** note to note **the tool you use**. Examples might be a database system, a script you have, a module, or a software package



Use as many as you need.
Okay to repeat!

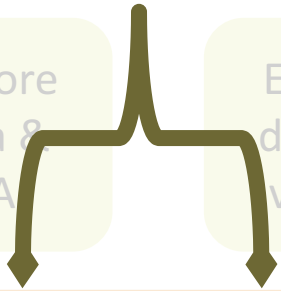
Harvest data

Split data pkgs up

Explore data & QA

Extract desired values

Do SCIENCE! & math



OAI-PMH datastore

Data pkgs from ←

Split data files

Split data files

My clean data????

scrape.py lxml

Split.py lxml

pandas

pandas

R??

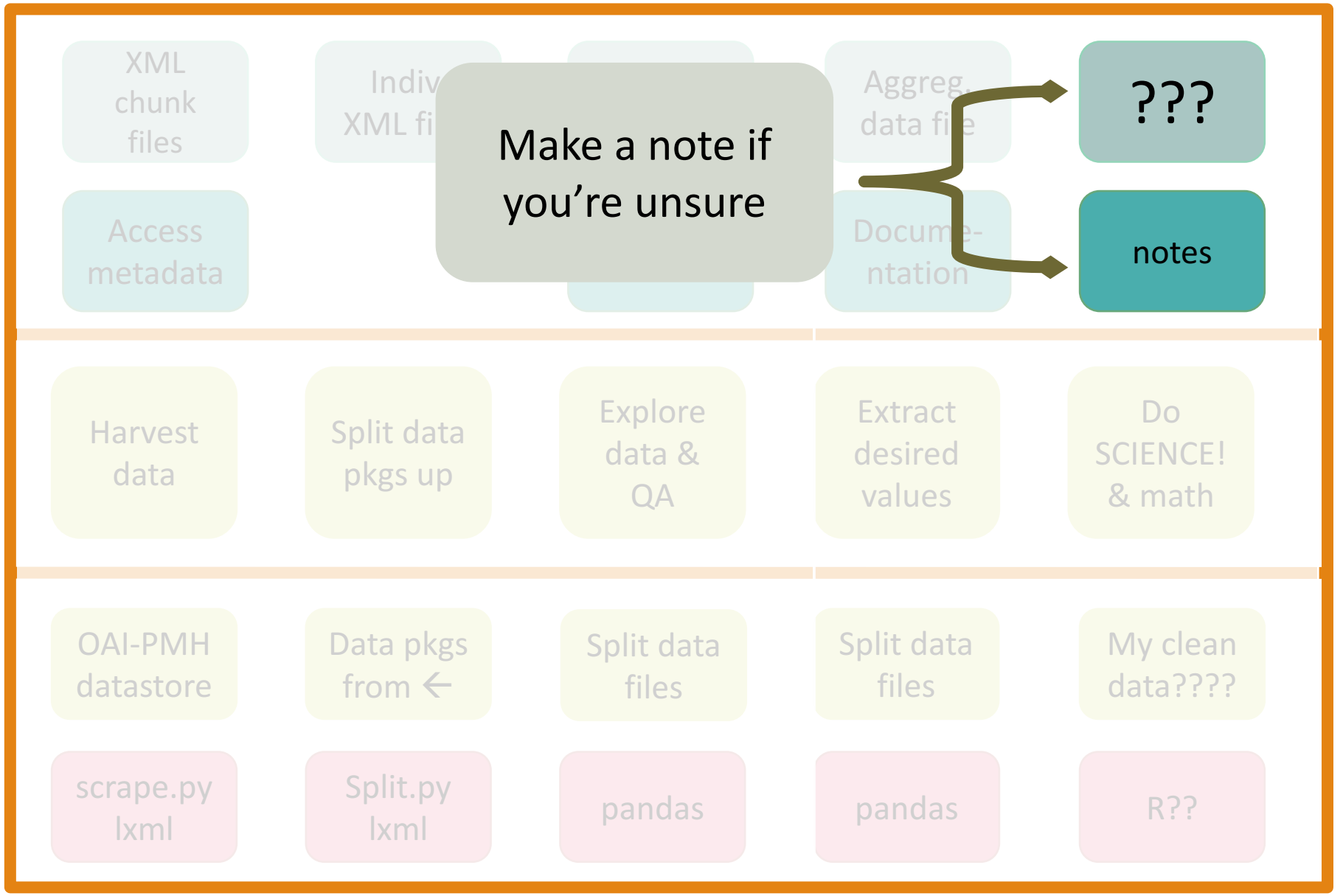
XML chunk files

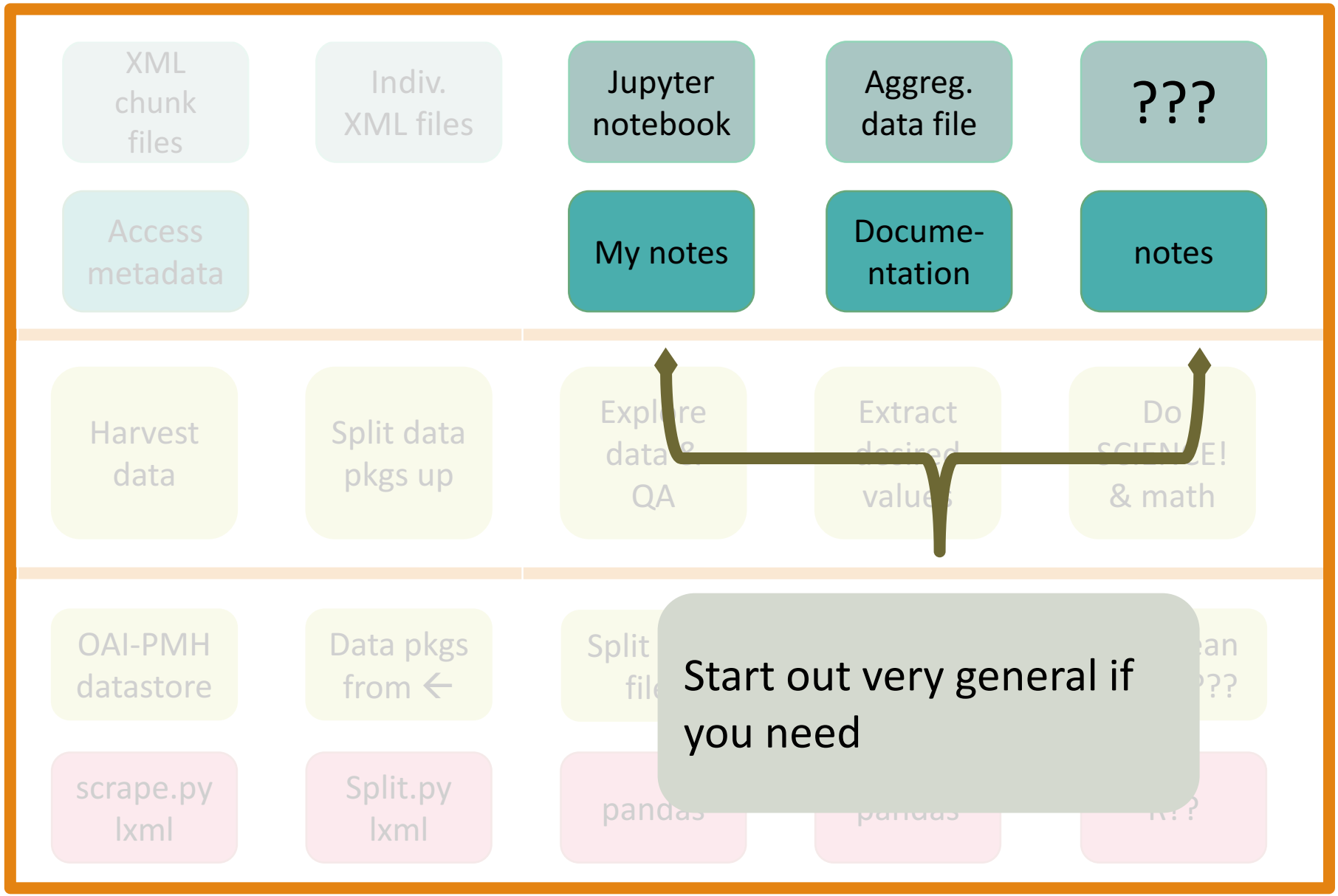
Note the data products that you'll be making

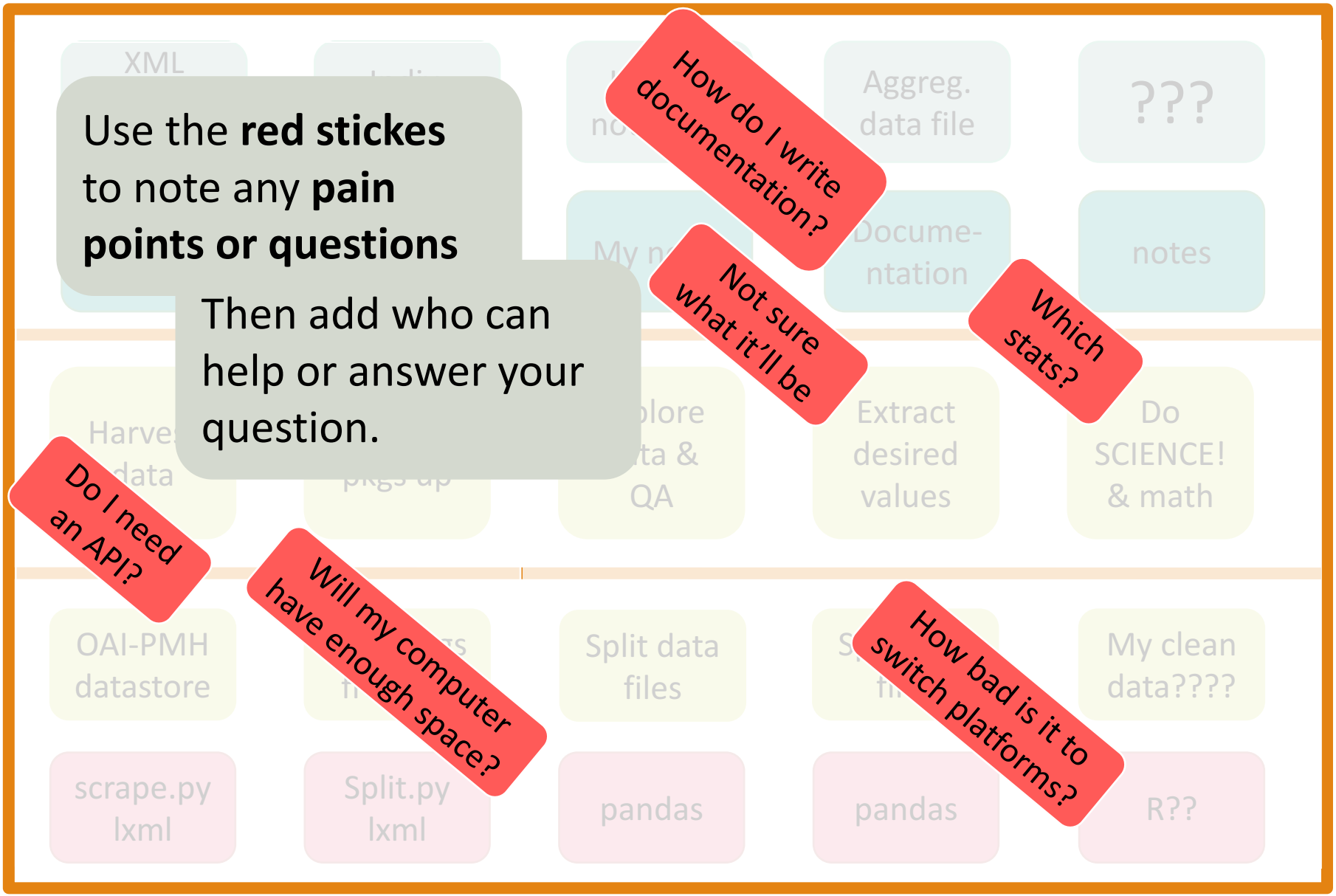
Access metadata

Use another color to distinguish another kind of data type or purpose (e.g.if that data will go to another human)

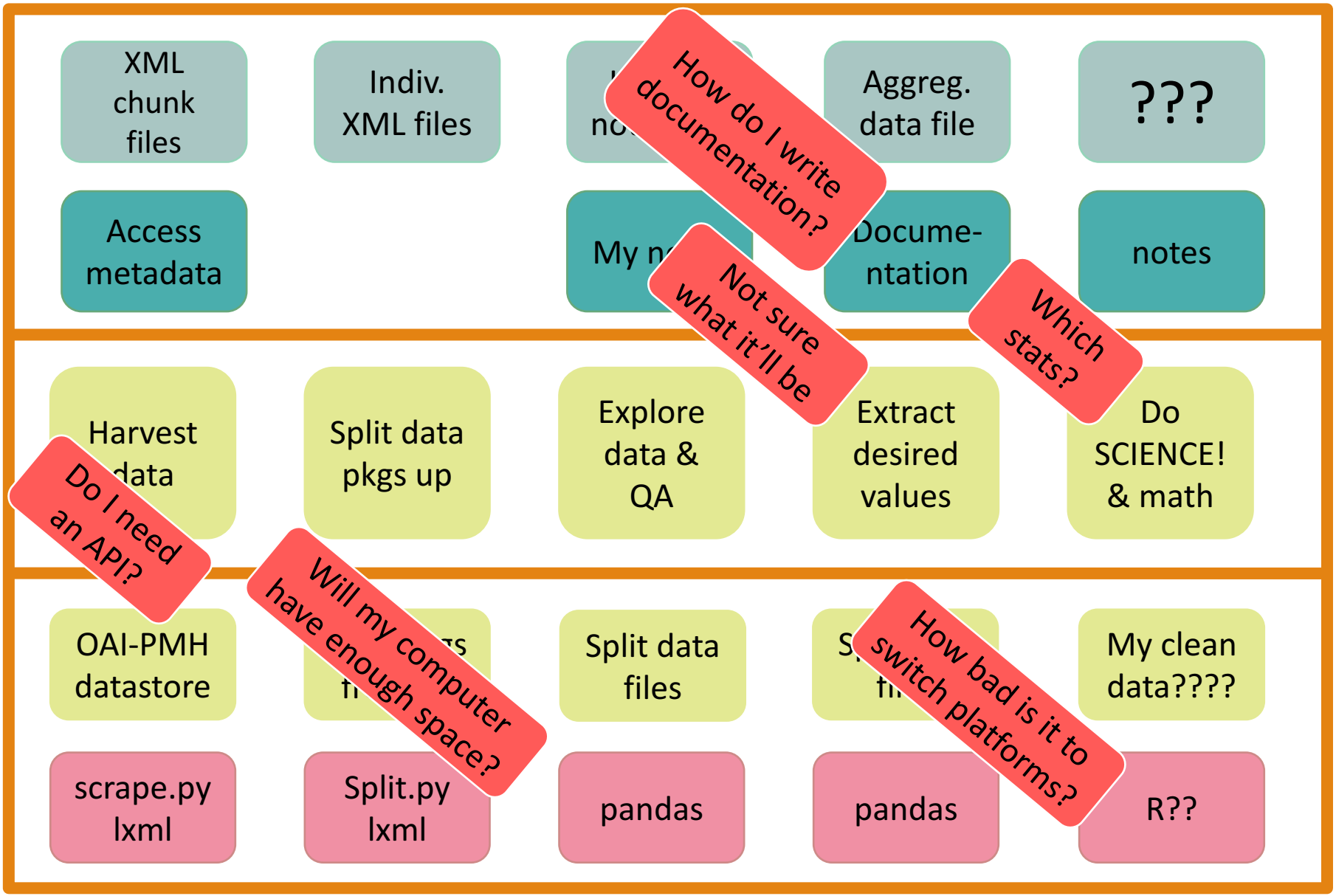
Harvest data	Split data pkgs up	Explore data & QA	Extract desired values	Do SCIENCE! & math
OAI-PMH datastore	Data pkgs from ←	Split data files	Split data files	My clean data????
scrape.py lxml	Split.py lxml	pandas	pandas	R??


XML chunk files
Indiv XML fi
Aggreg. data file
Access metadata
Docume-ntation
Make a note if you're unsure
???
notes
Harvest data
Split data pkgs up
Explore data & QA
Extract desired values
Do SCIENCE! & math
OAI-PMH datastore
Data pkgs from ←
Split data files
Split data files
My clean data????
scrape.py lxml
Split.py lxml
pandas
pandas
R??


XML chunk files
Indiv. XML files
Jupyter notebook
Aggreg. data file
???
Access metadata
My notes
Docume-ntation
notes
Harvest data
Split data pkgs up
Explore data & QA
Extract desired values
Do SCIENCE! & math
OAI-PMH datastore
Data pkgs from ←
Split file
Clean ???
scrape.py lxml
Split.py lxml
pandas
pandas
R??
Start out very general if you need



RESEARCH DATA SERVICE


XML chunk files
Indiv. XML files
Aggreg. data file
???
Access metadata
Docume-ntation
notes
How do I write documentation?
Not sure what it'll be
Which stats?
Harvest data
Split data pkgs up
Explore data & QA
Extract desired values
Do SCIENCE! & math
Do I need an API?
Will my computer have enough space?
How bad is it to switch platforms?
OAI-PMH datastore
Split data files
My clean data????
scrape.py lxml
Split.py lxml
pandas
pandas
R??

Now take another look

- Are there deadlines you can trace back and add?
- Looking at the stuff that you are making:
  - What folders do you need?
  - Where should those folders be?
  - What should your file names be?
- Looking at the tools you use:
  - What documentation do you need about them to understand your project in a few years or for another person to take it up?
  - Do you need to save/backup the software or scripts to include as a reference in a future project?
- Add annotations to your board to indicate this. Use the back of your worksheet to document the folder structure.