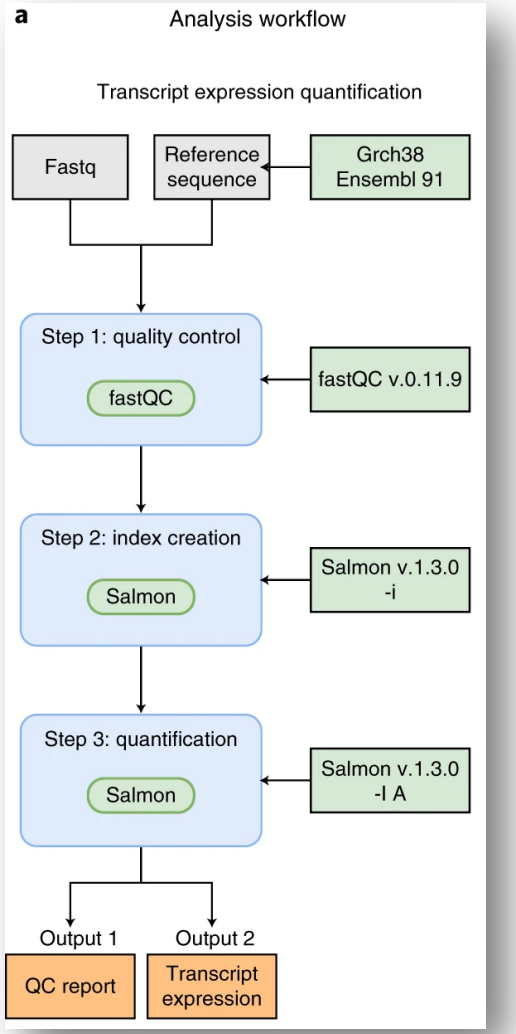
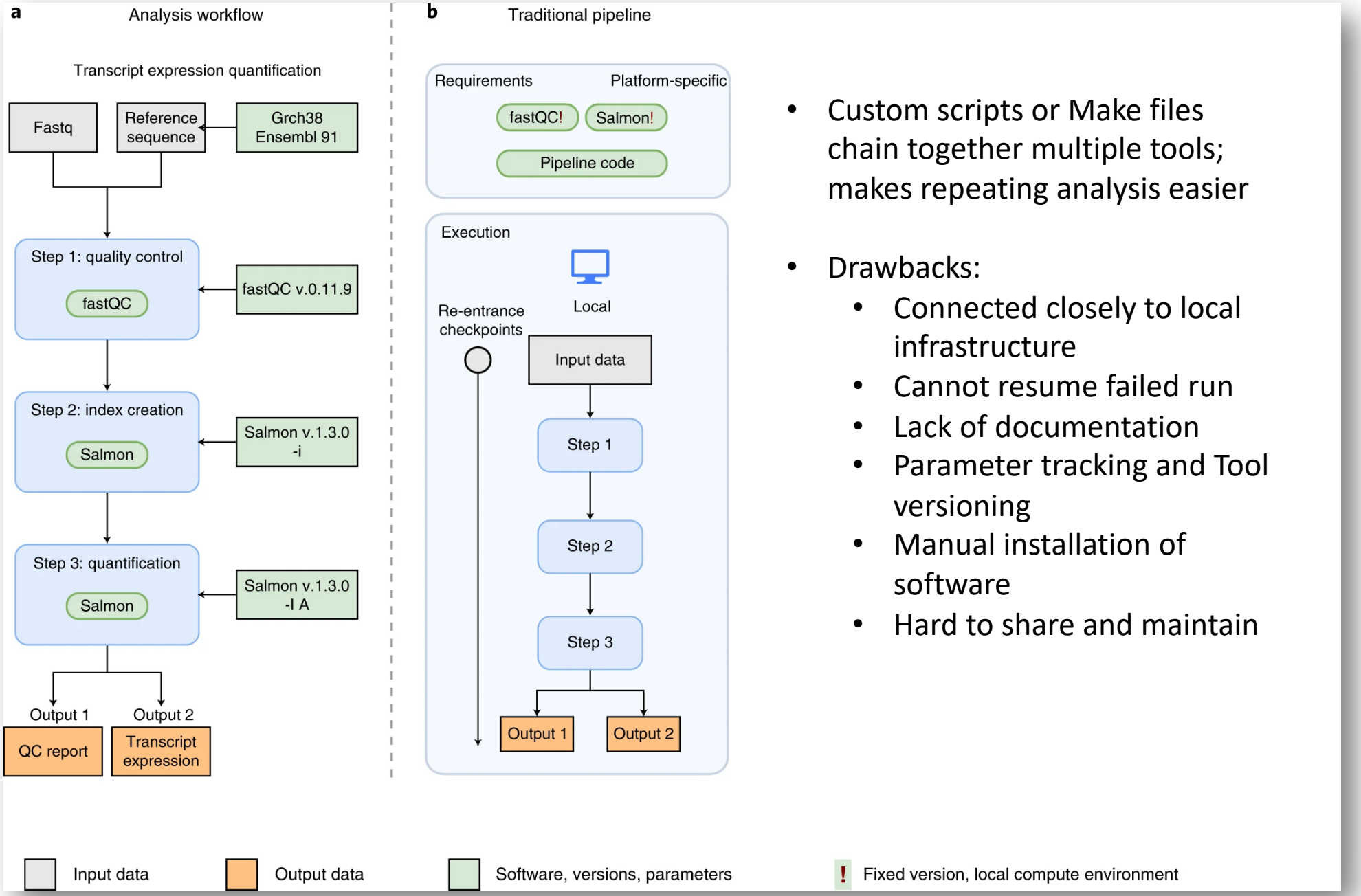
Workflow Management

Epigenomics Data Analysis 2021

Scientific Data Analysis in practice



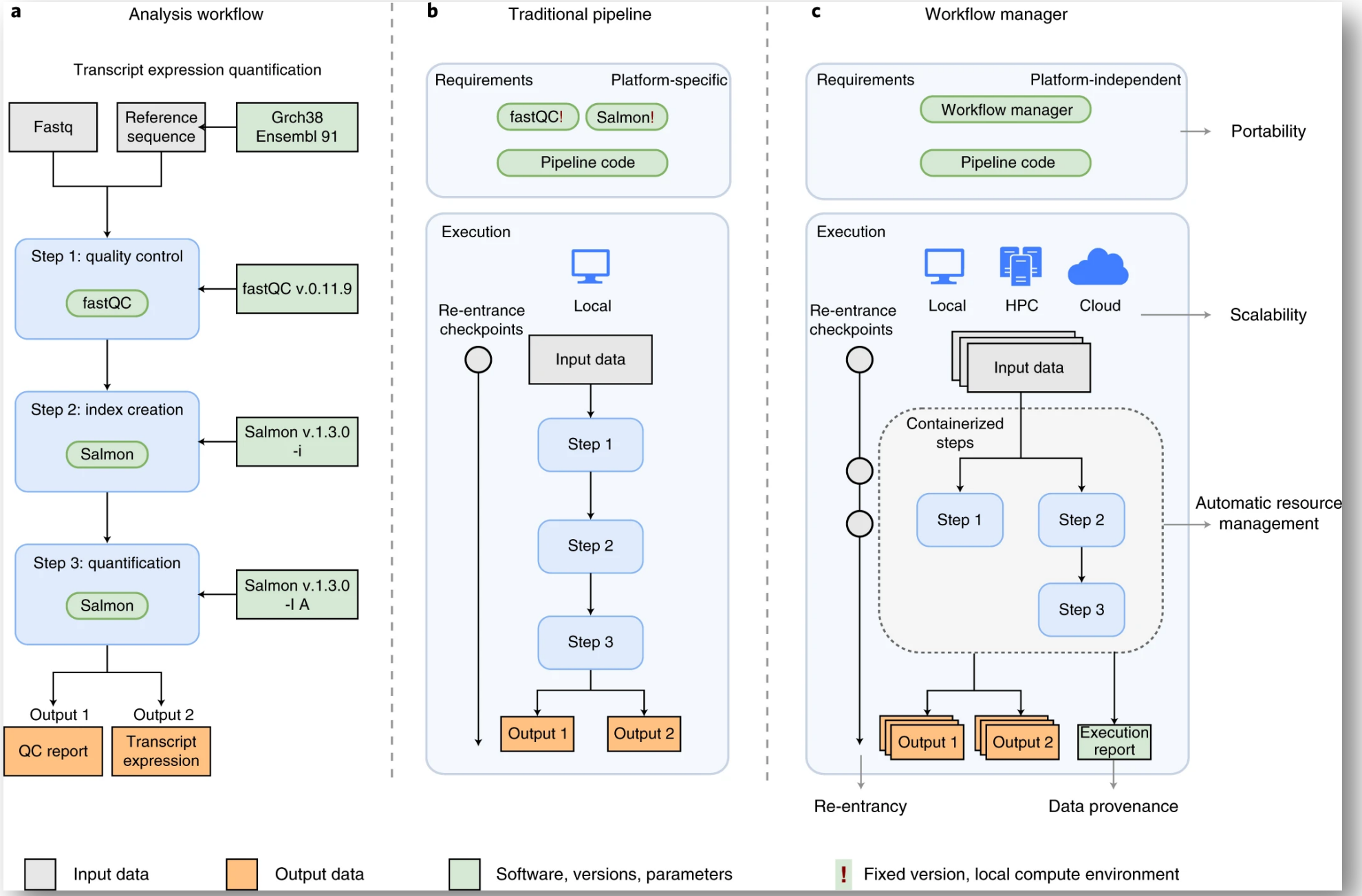
- Often no single tool; many tools need to be combined
- References, databases and software versioned
- Every tool comes with their own set of parameters
- Experimentation is crucial; reruns necessary
- Reproducibility becomes an issue
- Rerun part of the analysis
- Resource management?
- …



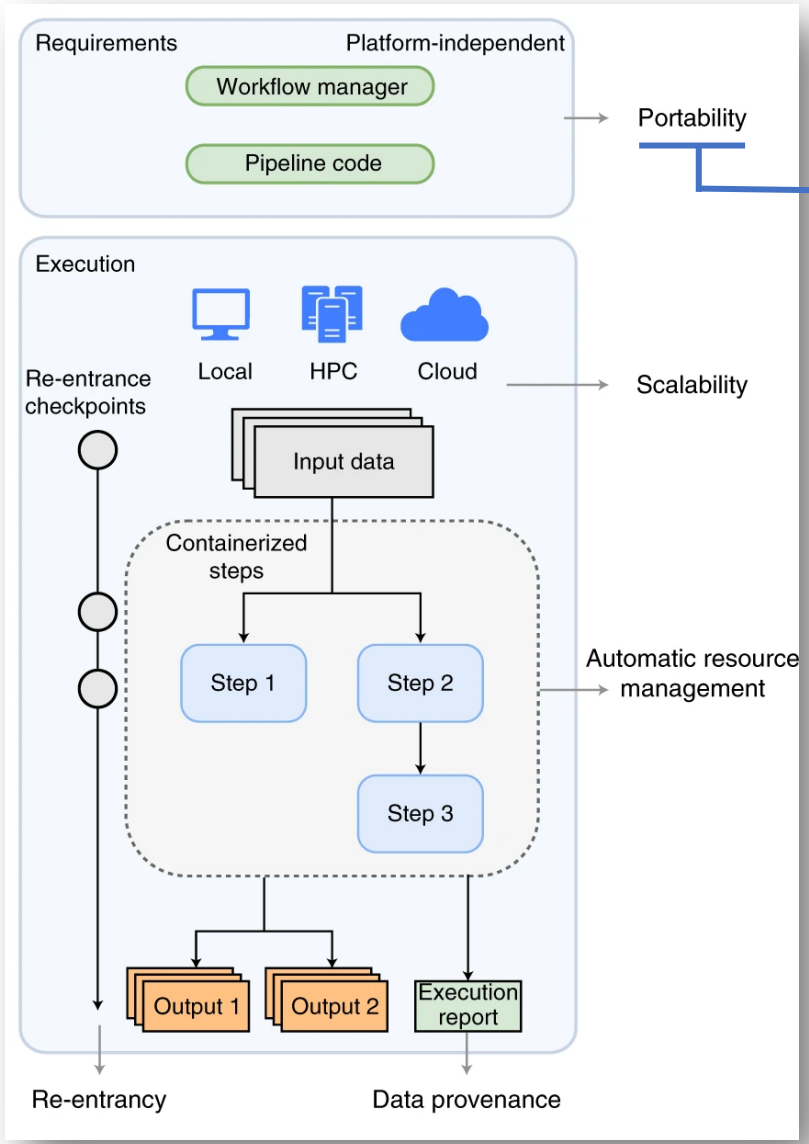
- Custom scripts or Make files chain together multiple tools; makes repeating analysis easier

- Drawbacks:
  - Connected closely to local infrastructure
  - Cannot resume failed run
  - Lack of documentation
  - Parameter tracking and Tool versioning
  - Manual installation of software
  - Hard to share and maintain

Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* (2021).

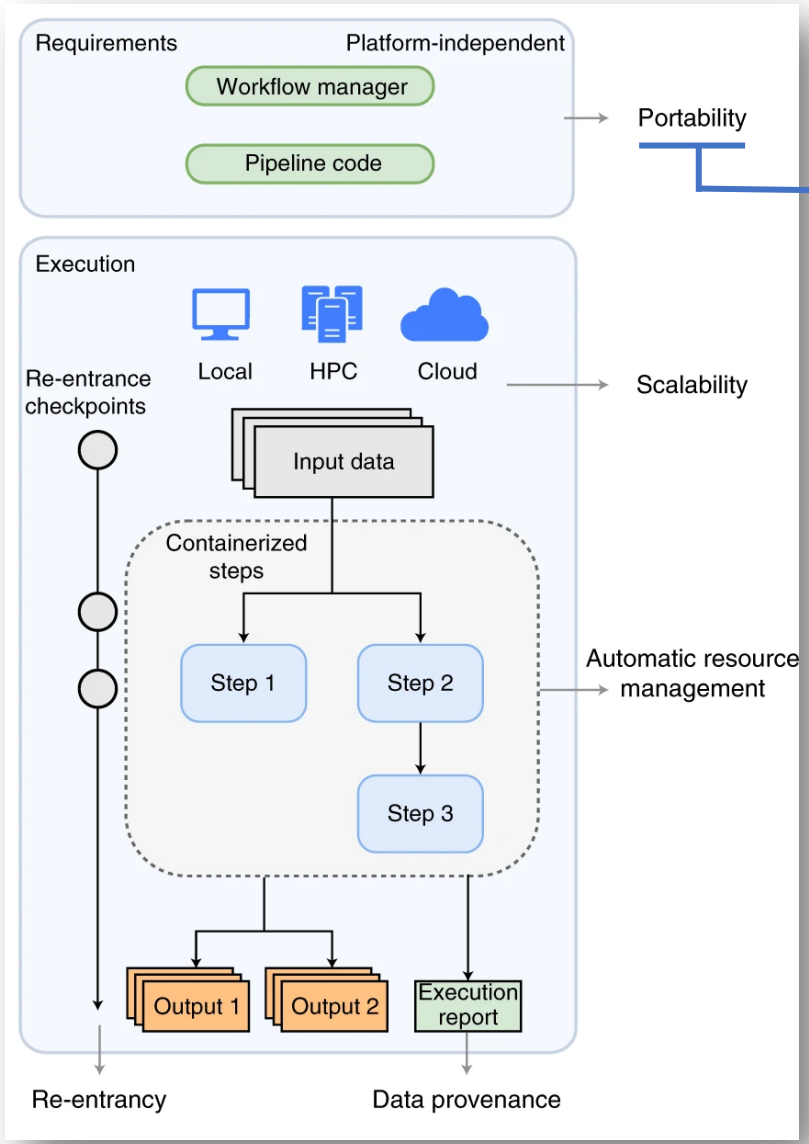


Wratten, L., Wilm, A. & Göke, J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* (2021).



Portability

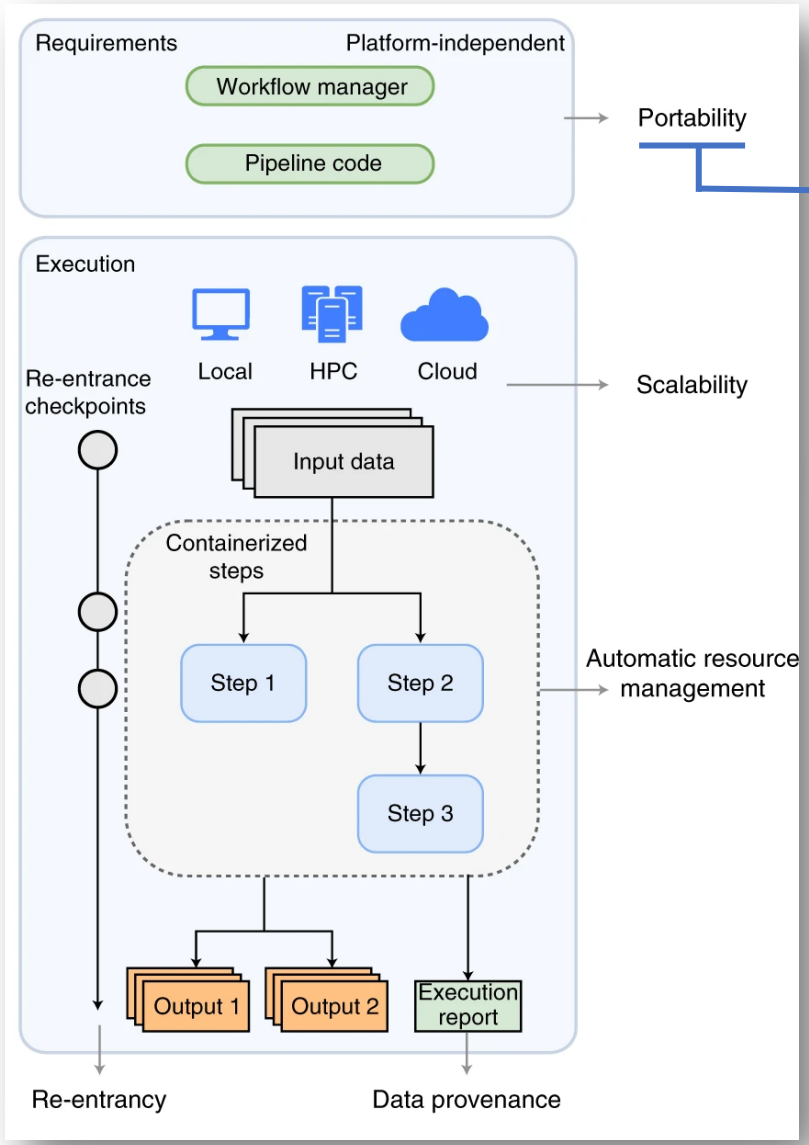
Execution with the same functionality across different platforms and over time.



Portability

Execution with the same functionality across different platforms and over time:

- Package Managers
  - Automate the process of installing and configuring software
  - Obtain all tools and dependencies with a single command
  - Homebrew, conda, …

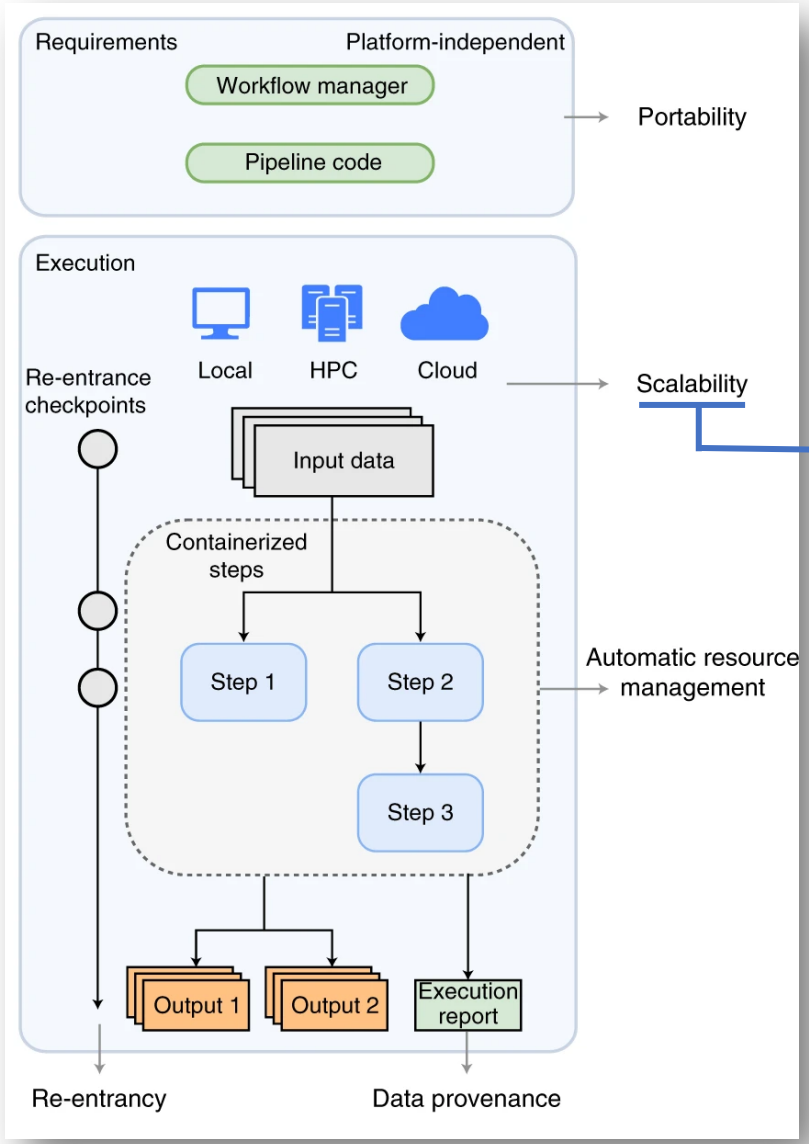


Portability

Execution with the same functionality across different platforms and over time:

- Package Managers
  - Automate the process of installing and configuring software
  - Obtain all tools and dependencies with a single command
  - Homebrew, conda, …
- Containerization Software
  - “Lightweight” technology
  - Combines software, dependencies and operating system in a self-contained “package”
  - Docker, Singularity
  - BioContainers: prebuilt containers 1000s of tools

OS & software version can influence results; use these techniques for increased reproducibility!

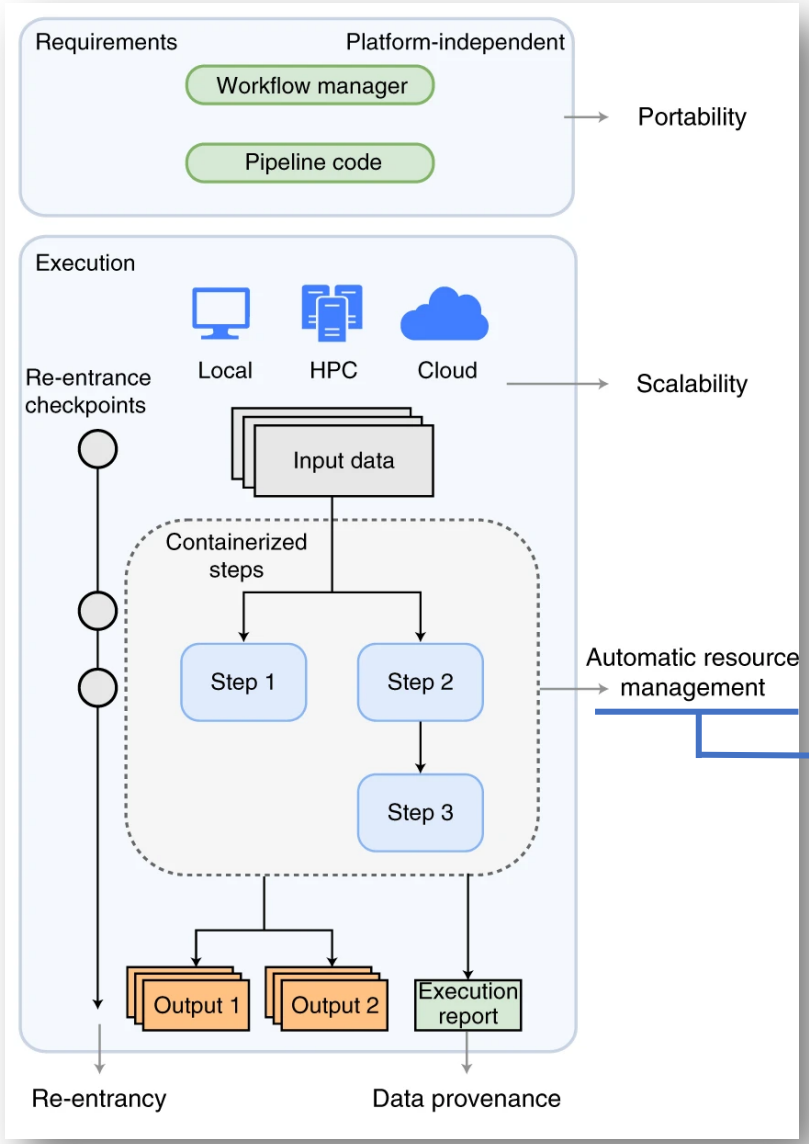


Scalability

Ability to handle any size and quantity of input data:

- Built-in support for local computer, high-performance computing and cloud computing.

Choose the right execution infrastructure to enable efficient analysis of small and large datasets!



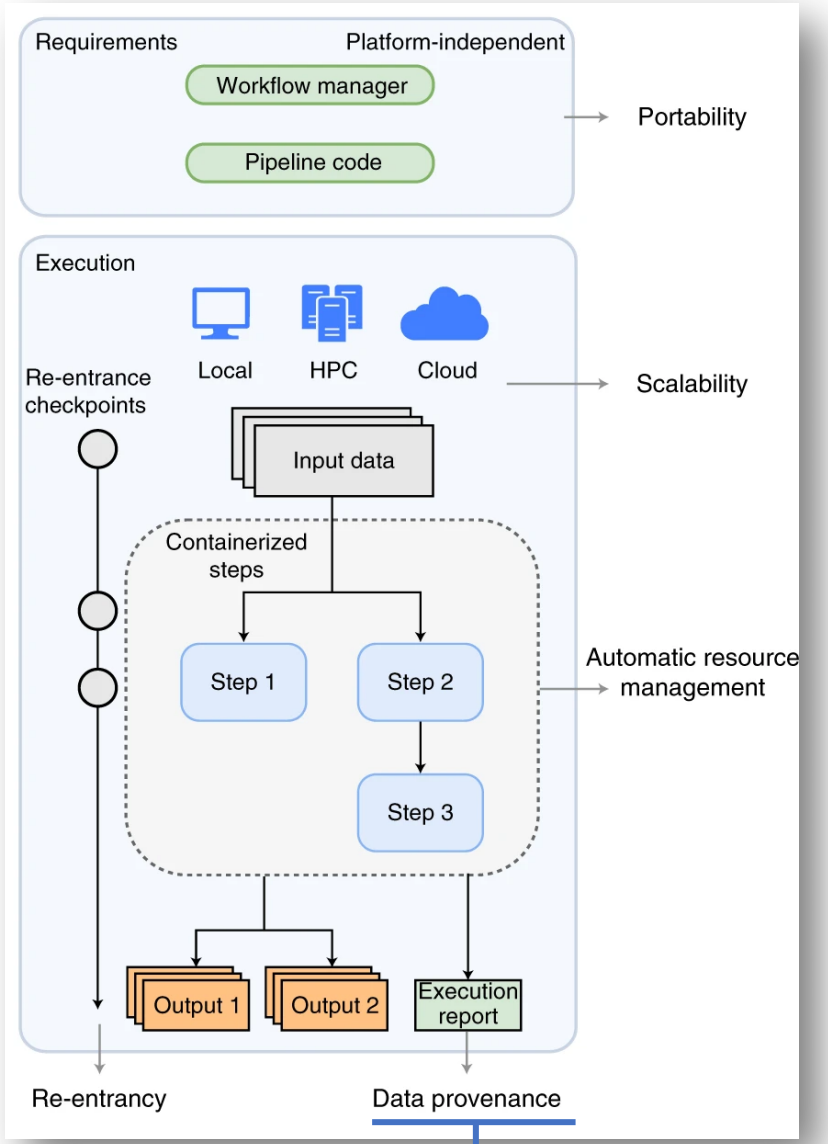
Resource Management

Execution each step of a pipeline with sufficient (but not excessive) resources:

- Resources: Time, CPU, Memory
  - Multiple steps have different requirements
  - WFM high degree of flexibility
- Efficient parallelization!
  - Dynamic scheduling, reduce unused resources
- Many pipeline managers have implementations of most common task managers (Slurm, PBS, …)

Effective handling of large datasets and minimize bottlenecks that increase running time!

slurm
workload manager

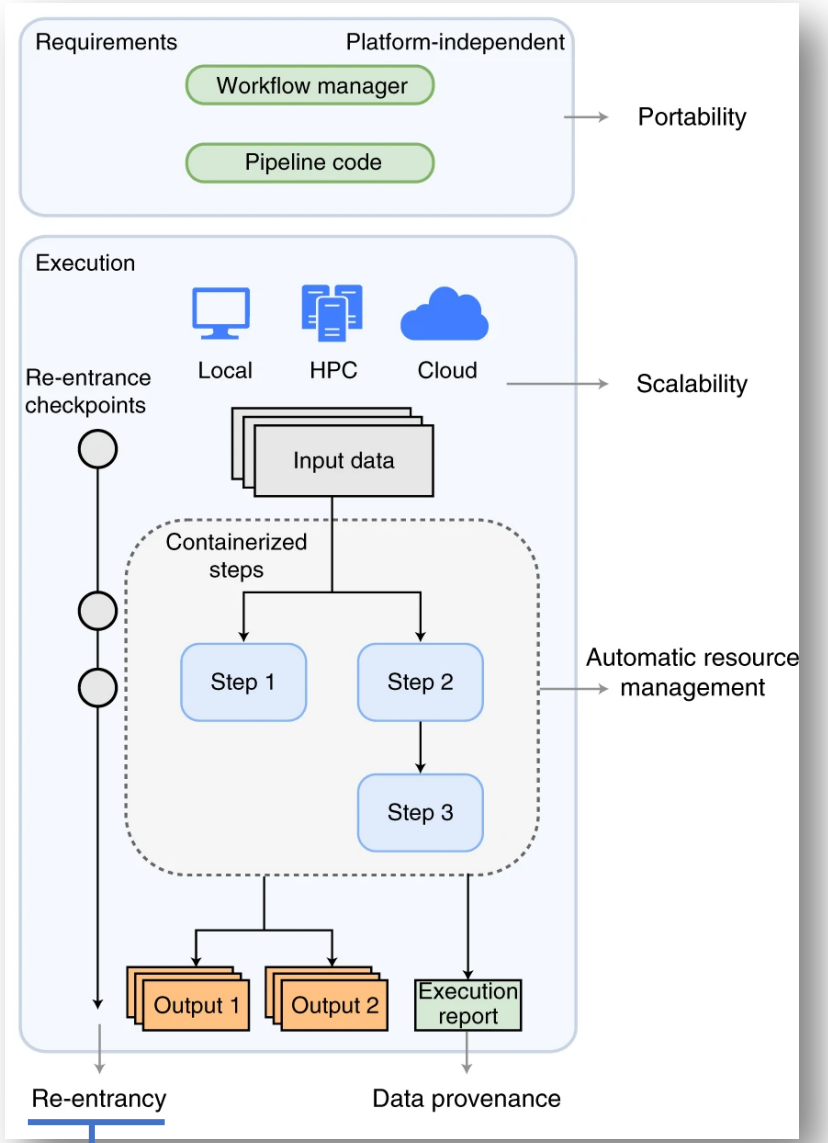


Data provenance

Any change in software version, parameters or reference annotation version can alter results:

- WFM automates tracking of input parameters and tool versions
- Detailed information in execution report
- Workflow itself can be archived and versioned

High level documentation enables transparency, code sharing and long-term reproducibility!



Re-entrancy

Run a pipeline from its last succesfully executed step, rather than from the beginning:

- Error in pipeline; only repeat necessary steps
- Avoid recalculation of very long steps
- Cache intermediate results and data files
- Save significant time and resources

Re-entrancy saves significant time and compute resources and is **a key advantage** of workflow managers!

Rapid rise in manager software…

The three technologies bioinformaticians need to be using right now

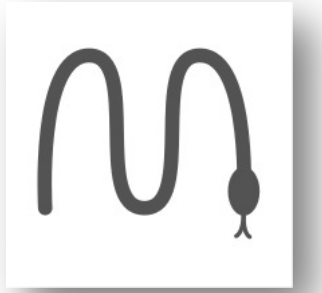
12TH AUGUST 2019 / BIOMICKWATSON / COMMENTS OFF

2. Workflow management systems

It's a law of the universe that every bioinformatician will say they have a "pipeline", but often I find this is actually a bash, perl or python script. They were great for their time, but it is time to move on.

Rapid rise in manager software…

- Many workflow management systems have been developed



… and hundreds more

Three types of Workflow Managers

- Graphical Interface
  - No programming experience needed; drag and drop
- Domain-Specific Language (DSL)
  - Dedicated programming language
- Programming-library-based
  - Developed within and for a specific language

Galaxy PROJECT

- Web-based platform for bioinformatic workflows
  - Open source; more than 7,500 citations
- Designed for biologists to work with their own and public data
- “App store” for bioinformatic tools
  - 8140 tools (https://toolshed.g2.bx.psu.edu)
- Abundance of learning resources

Galaxy PROJECT

Genomics, HTS

Quality control: FastQC, MultiQC, Trim Galore!, etc
Alignment: BLAST, Diamond, etc
Mapping: Bowtie2, STAR, HISAT2, BWA, segemehl, etc
Assembly: Unicycler, SPades, Quast, etc
Transcriptomics: FeatureCounts, DESeq2, Trinity, Salmon, etc
RNA: LocARNA, RNAfold, RNAz, RNAplot, etc
Variant Calling: FreeBayes, Gemini, VCFTools, SnpEff, etc
Peak Calling: MACS2, Piranha, PEAKachu, etc
Epigenetics: Bismark, metilene, bwameth, MethylDackel, etc
deepTools, SAM Tools, HicExplorer, Picard, EMBOSS, etc

File and metatools

Get data: USCS, Uniprot
Send data: GenomeSpace Exporter
Convert

Text tools

Text manipulation
Filter and Sort
Join, Substract and Group
Statistics

Metagenomics

MetaPhlAn2, HUMAnN2, VSearch, ...
QIIME, Mothur
MEGAHIT, MetaSpades, ...

Annotation, ontologies

SortMeRNA, Aragorn, Roary, Prokka, Augustus, KOBAS, Glimmer, antiSMASH, etc

Proteomics, Metabolomics, Chemistry

OpenMS, PeptideShaker, SearchGUI, MADLIquant, etc
JMol Editor, Docking, etc
OpenBabel, ChemFP, OMG, QED, etc

Domain-specific language (DSL) workflow managers

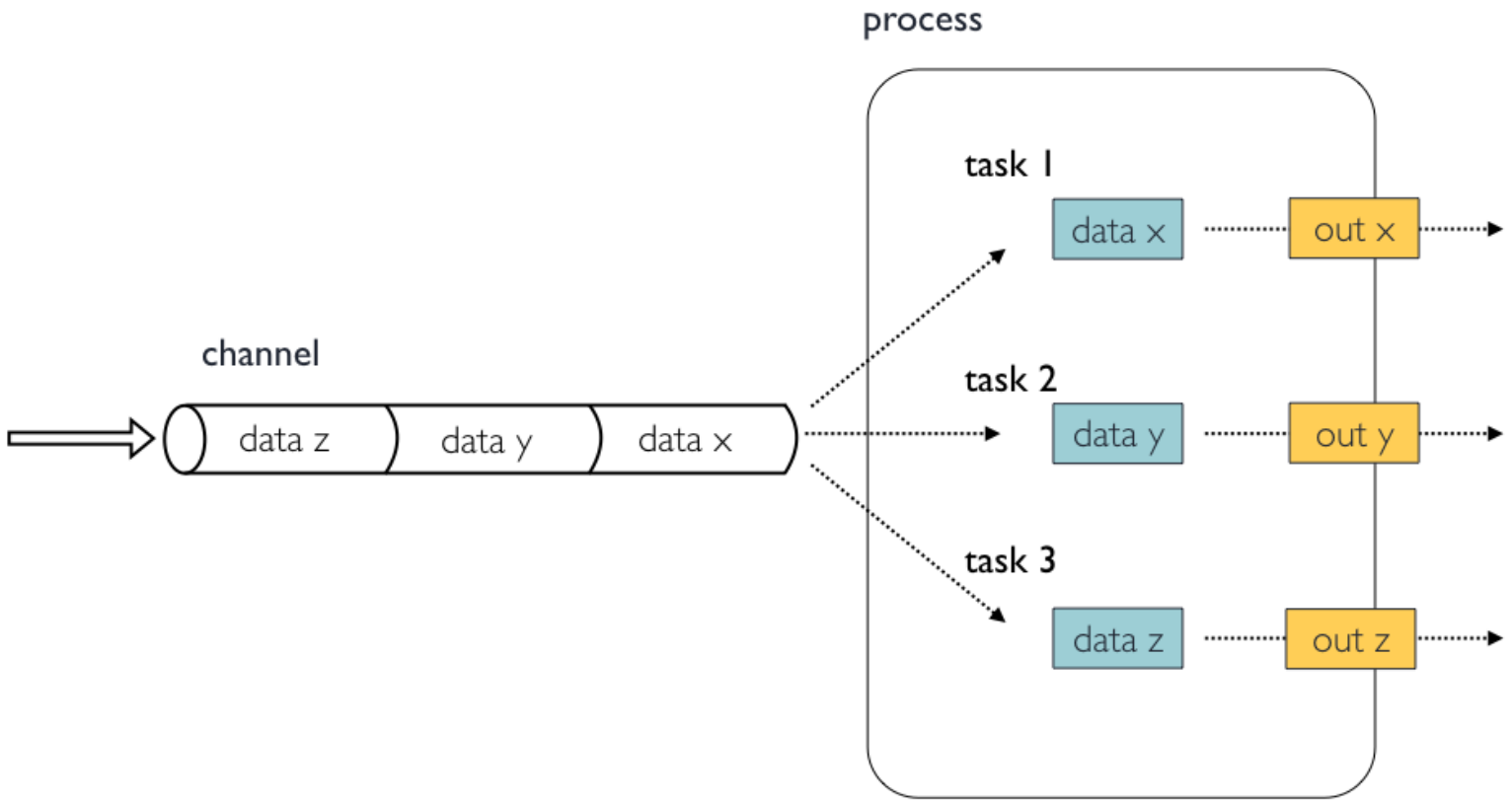
- DSL is a programming language developed to meet a specific need in a particular domain
- WFM implemented in DSL
  - Reproducible, robust and portable
  - Incorporate existing tools
  - Reusable modules
- Maximal flexibility, but can have a bit of an initial learning curve

(GUIs being developed to improve accesibility)

- Nextflow, Snakemake, …

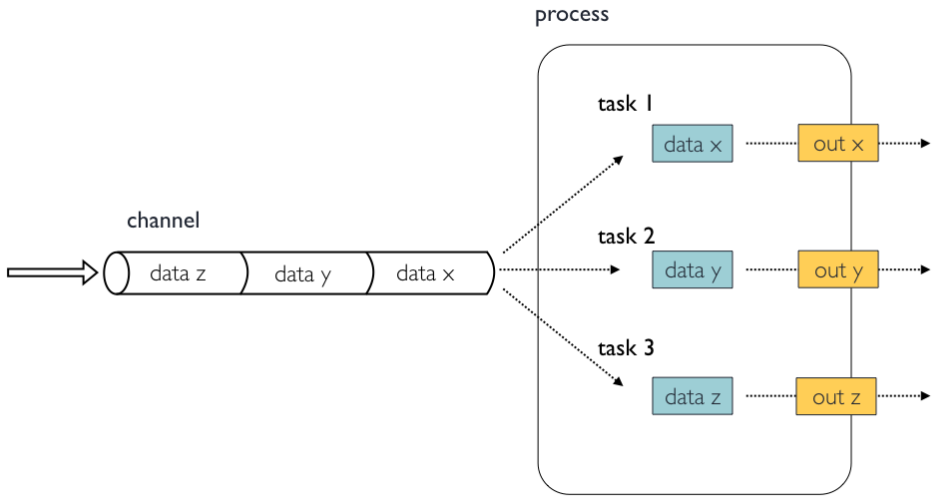
nextflow

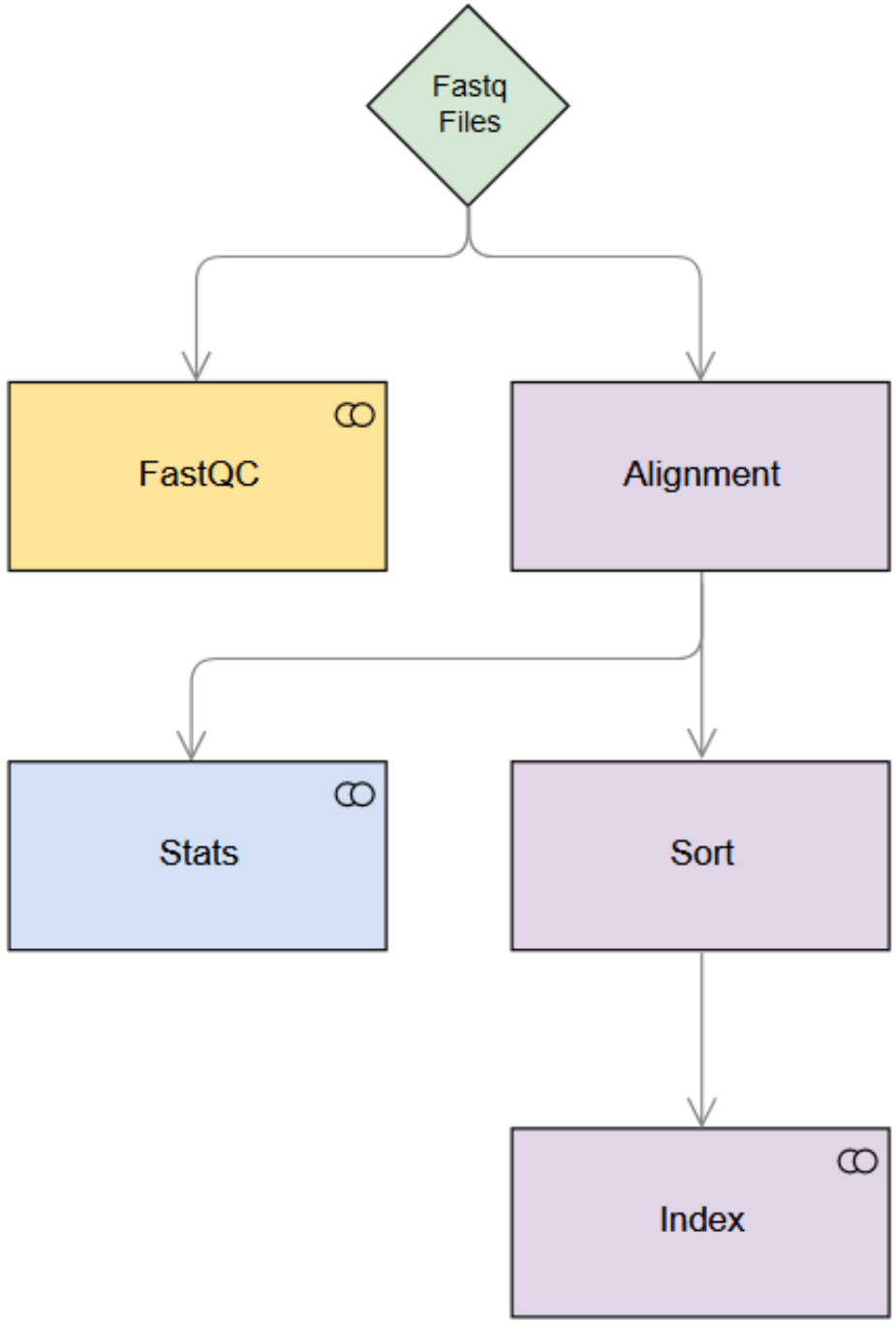
- Free, open source software tool
- Written in Groovy



nextflow

- Free, open source software tool
- Written in Groovy





No need to reinvent the wheel!

- Often good idea to look for existing pipelines
- Best-practice pipeline collections

- 41 released pipelines, 24 under development: https://nf-co.re/pipelines
- Methylseq, cutandrun, chipseq, atacseq, hiC, …

Labs

- Run nf-core on uppmax

- To understand the basic components in a Nextflow run
- To be able to find relevant nf-core pipelines

Use a workflow manager!

Whenever you:

- Use more than 1 tool
- Your analysis contains long calculation steps
- Expect to repeat your analysis
- Expect to publish or share your analysis results
- Want to avoid tricky software installation