

# Fully reproducible data analysis with Snakemake and Bioconda

Johannes Köster

Centrum Wiskunde & Informatica, Amsterdam  
Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston

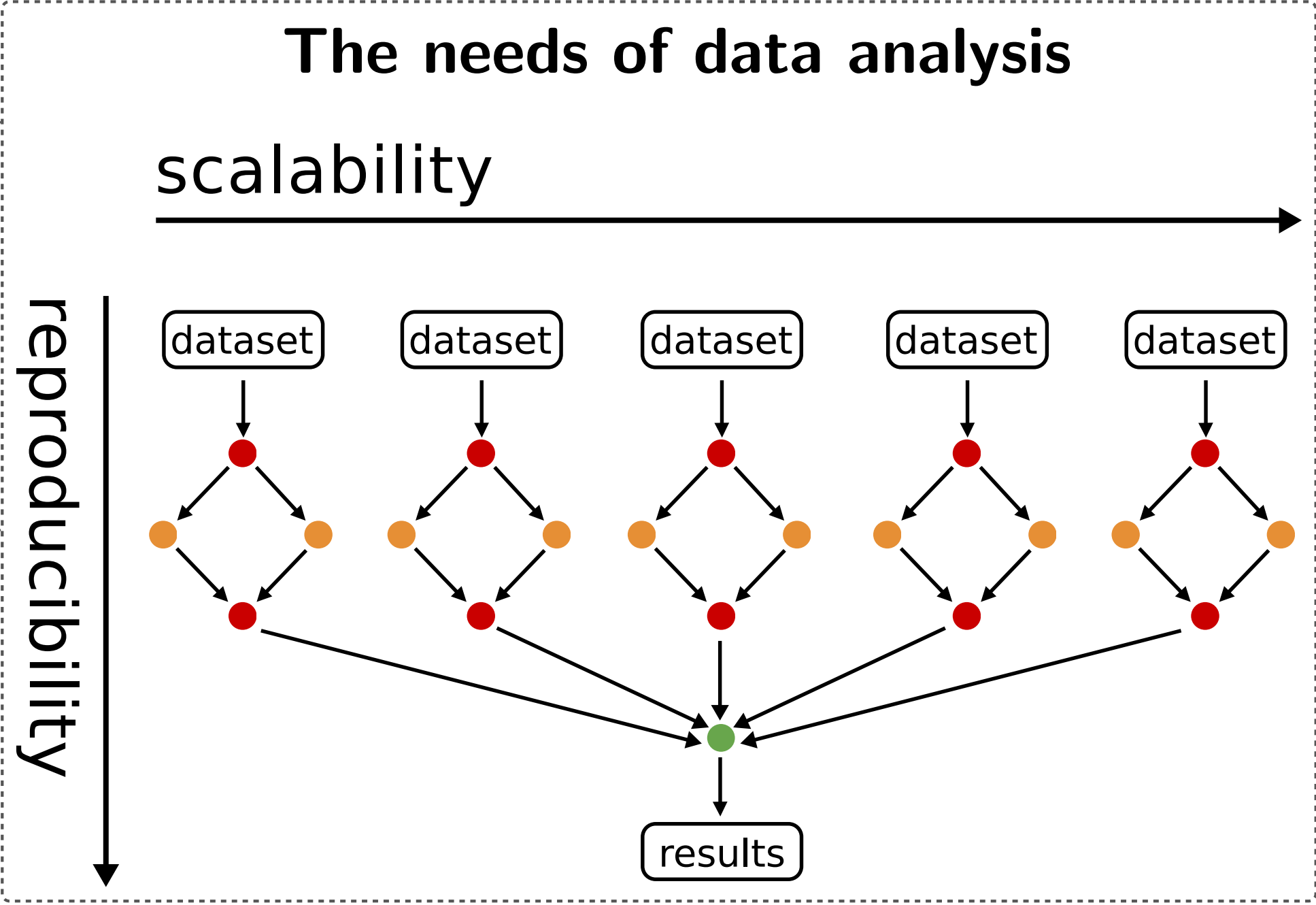
Data analysis

Snakemake

Bioconda

- Handle tens to thousands of samples via **parallelization**.
- **Avoid redundant computations** when changing datasets or parameters.

- **Document** parameters, tools, versions.
- **Execute and deploy** without manual intervention.



- Isolation allows **conflicting versions** on the same system.
- Exact versions ensure **full reproducibility**.



automatic installation  
via the versatile **Conda**  
**package manager**

**Define isolated software environments per rule**

```
channels:  
- bioconda  
- r  
dependencies:  
- bwa ==0.7.4  
- samtools ==1.1
```

**Define workflows via generic rules**

```
rule mytask:  
input:  
    "reference.fasta",  
    "reads/{dataset}.fastq"  
output:  
    "mapped/{dataset}.bam"  
environment:  
    "software.yaml"  
resources:  
    mem_gb=4  
shell:  
    "bwa mem {input} | "  
    "samtools view -b > {output}"
```

use **shell commands**,  
**scripts** (R, Python),  
and **tool wrappers**

**Dependencies**  
between rules are deter-  
mined **automatically**.

**implicit paralleliza-  
tion** to compute servers  
and clusters

**Snakemake formalizes,  
documents and executes  
data analyses**

Used by various high-impact studies:

Genome of the Netherlands:  
GoNL consortium. *Nature Genetics* 2014.  
Cancer:  
Townsend et al. *Cancer Cell* 2016.  
Schramm et al. *Nature Genetics* 2015.  
Martin et al. *Nature Genetics* 2013.  
Ebola:  
Park et al. *Cell* 2015  
IPSC:  
Burrows et al. *PLOS Genetics* 2016.  
Computational methods:  
Ziller et al. *Nature Methods* 2015.  
Schmied et al. *Bioinformatics* 2015.  
Brinda et al. *Bioinformatics* 2015  
Chang et al. *Molecular Cell* 2014.  
Marshall et al. *Bioinformatics* 2012.



**Bioinformatics software  
installation is  
heterogeneous**

```
source("https://bioconductor.org/biocLite.R")  
biocLite("BSSeq")  
cp lib/and44/111/*_so lib  
cp lib/and44/*_so lib  
cp * $PREFIX  
install.packages("matrixplot")  
cnae ../../my_project/  
-CHORE_MODULE_PATH=/dev/seqan/util/cnae  
-SEQAN_INCLUDE_PATH=/dev/seqan/include  
make install  
easy_install snakemake  
make install  
./configure --prefix=/usr/local  
make  
make install  
apt-get install bwa  
yum install python-hbpy  
cpan -i bioperl
```

**Bioconda normalizes software  
installation via easy to create  
package recipes**

```
package:  
name: seqtk  
version: 1.2  
source:  
fn: v1.2.tar.gz  
url: https://github.com/lh3/seqtk/archive/v1.2.tar.gz  
requirements:  
build:  
- gcc  
- zlib  
run:  
- zlib  
about:  
home: https://github.com/lh3/seqtk  
license: MIT license  
summary: Seqtk is a fast and lightweight tool for processing seqe:  
test:  
commands:  
- seqtk seq
```



Works for **any lan-  
guage** (R, Python,  
C/C++, Rust, Perl, ...).

- over **1500 packages**
- over **100 maintain-  
ers**

**By combining Snakemake and  
Bioconda, data analyses become  
reproducible with minimal effort**

```
# clone workflow repository  
$ git clone https://github.com/user/workflow  
  
# install Snakemake  
$ conda install snakemake  
  
# execute workflow  
# (software dependencies are handled automatically)  
$ snakemake -s Snakefile
```

