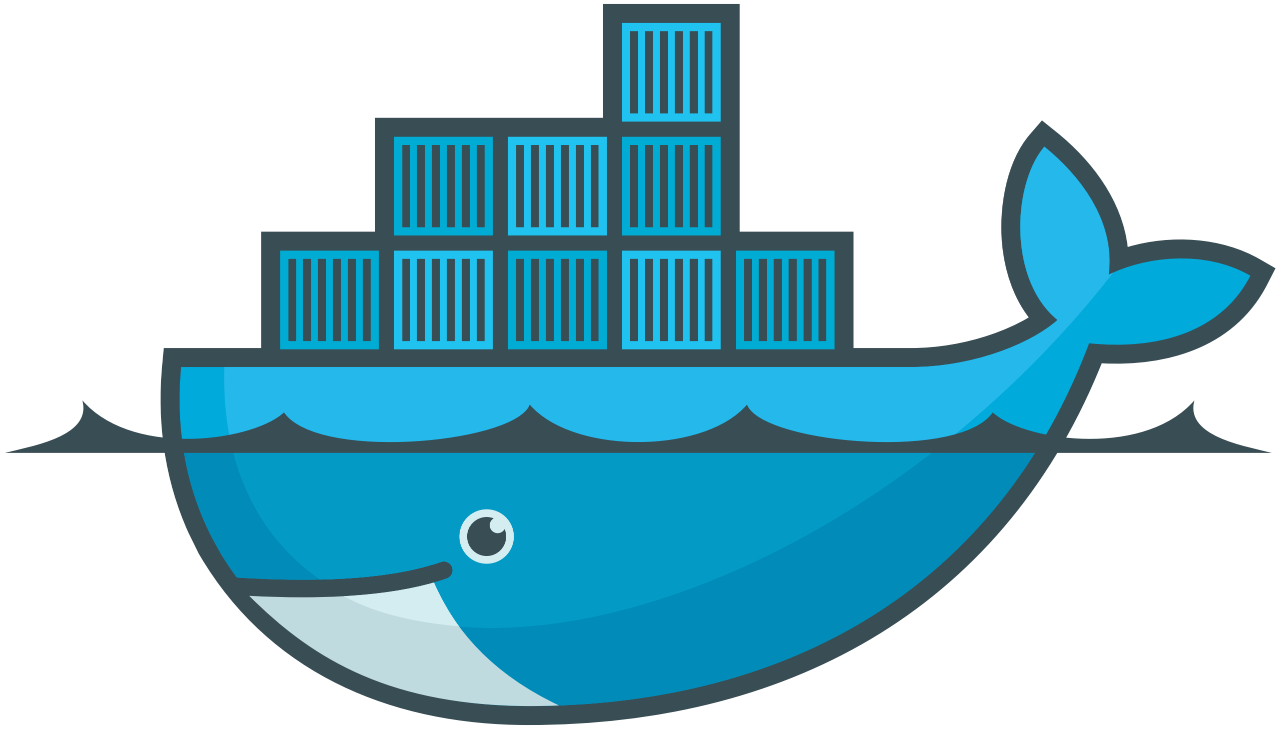
# Docker-encapsulated pipelines

Brief overview of my summer placement

Stefan Dang • Wellcome Trust Sanger Institute • Hinxton, 14.10.14

# Introduction

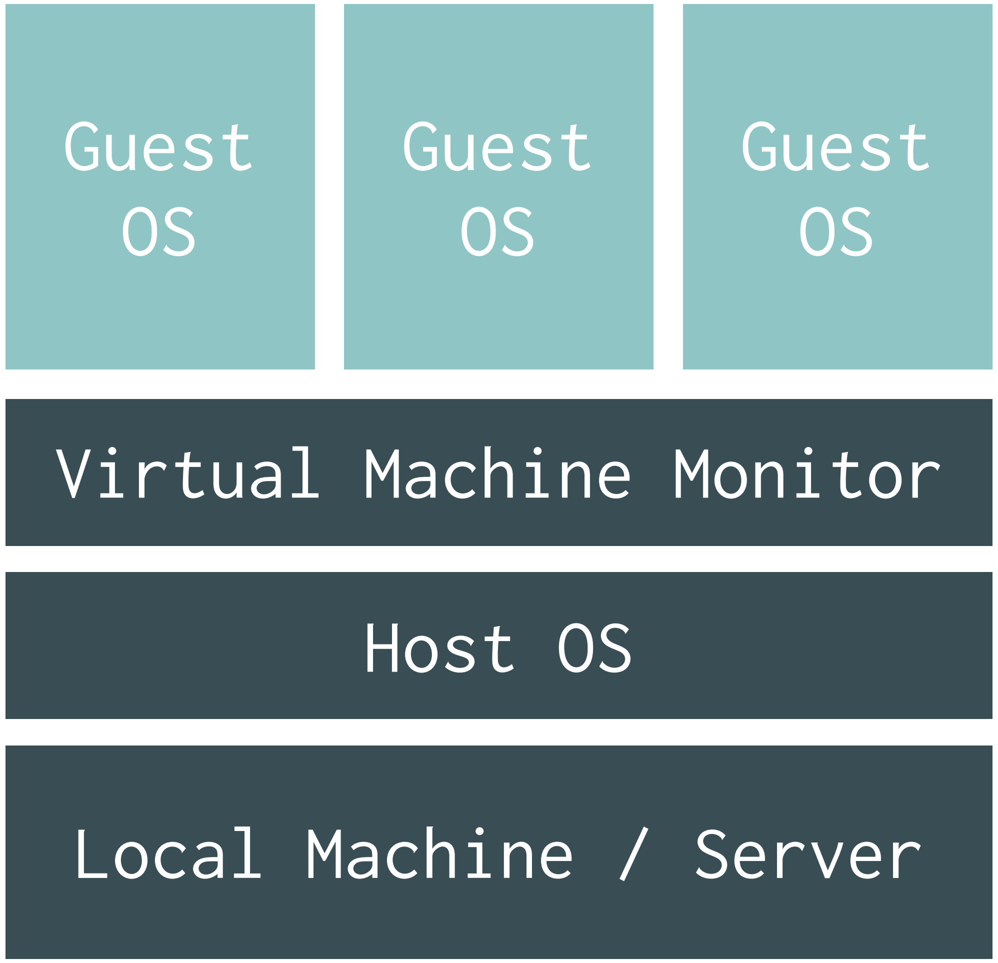


docker

docker.com

# Virtualization

## Virtual Machine



Guest OS

Guest OS

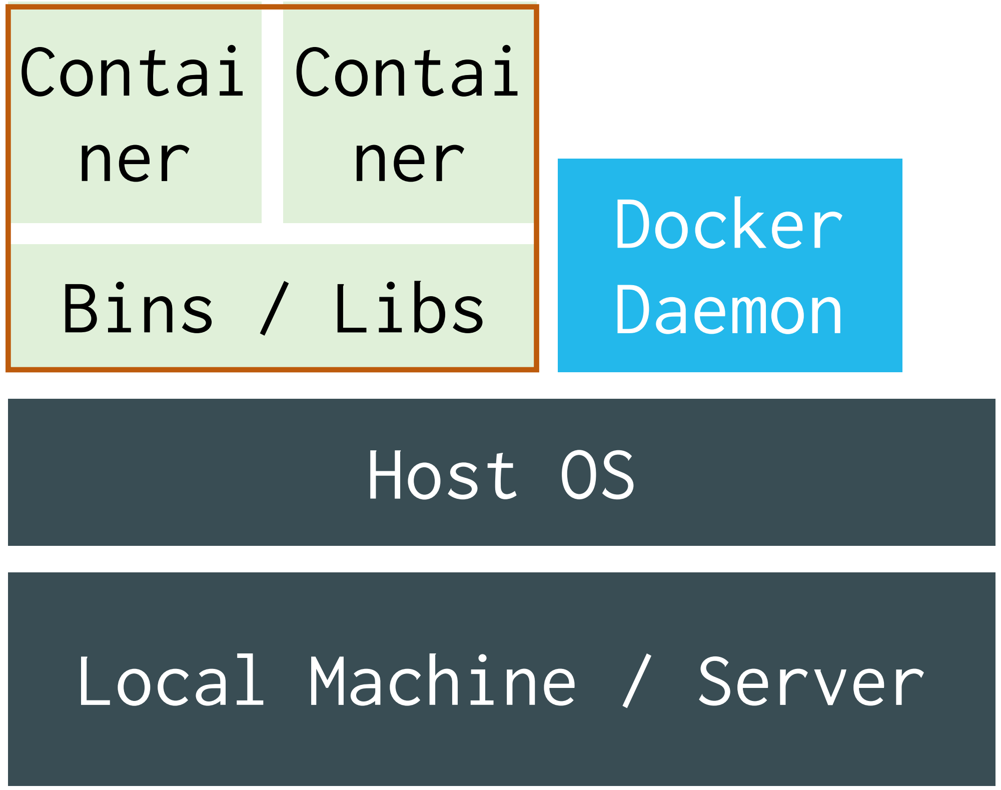
Guest OS

Virtual Machine Monitor

Host OS

Local Machine / Server

## Linux Containers



Container

Container

Bins / Libs

Docker Daemon

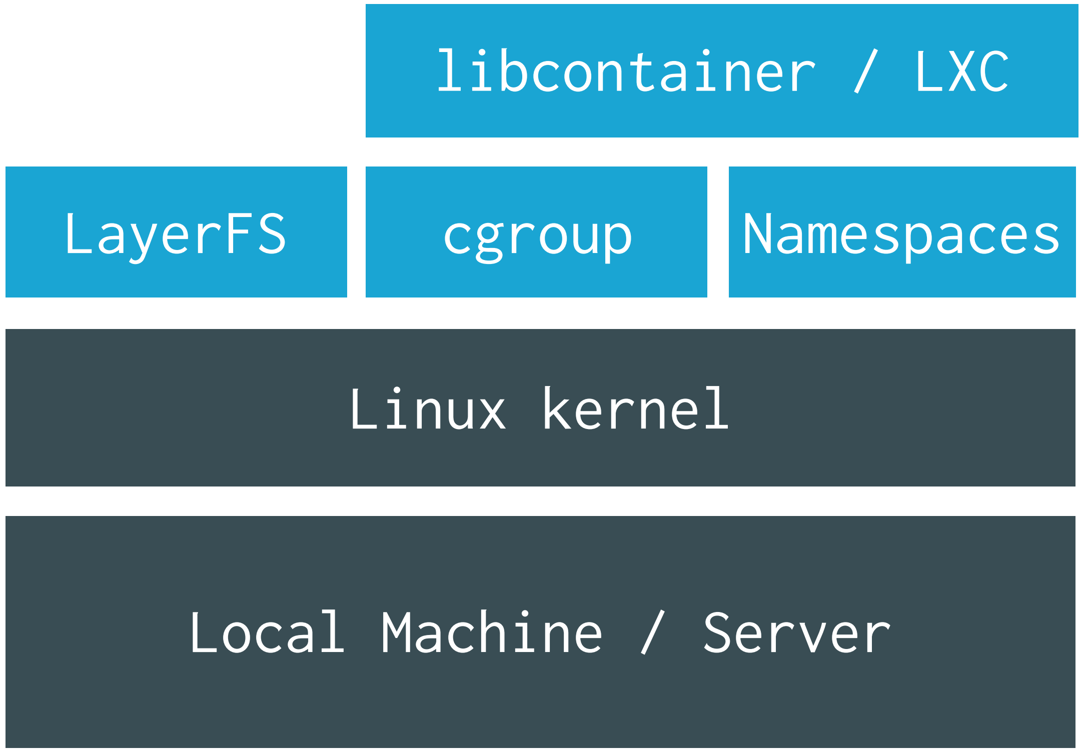
Host OS

Local Machine / Server

docker.com/whatisdocker/

# Docker in Detail

## Underlying Infrastructure



libcontainer / LXC

LayerFS

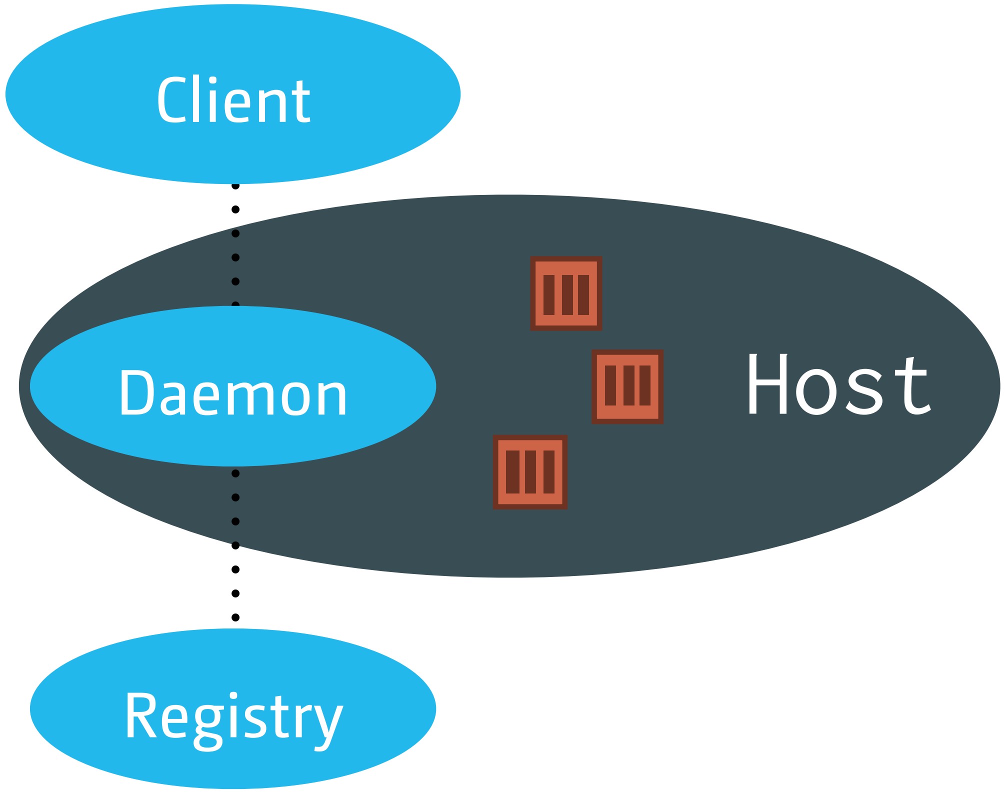
cgroup

Namespaces

Linux kernel

Local Machine / Server

## Client-Server-Application



Client

Daemon

Host

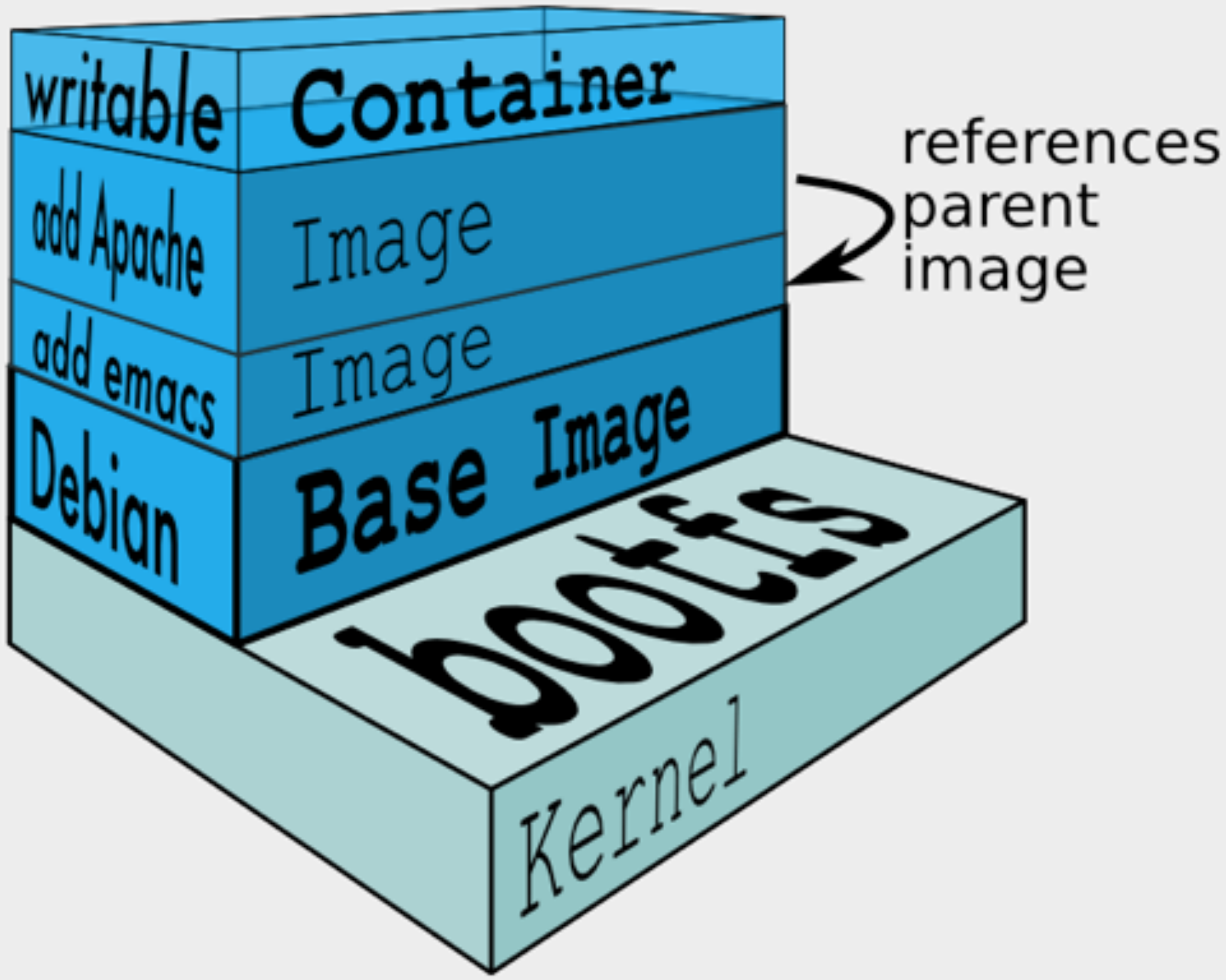
Registry

dikono.vsco.co

# Dockerfile

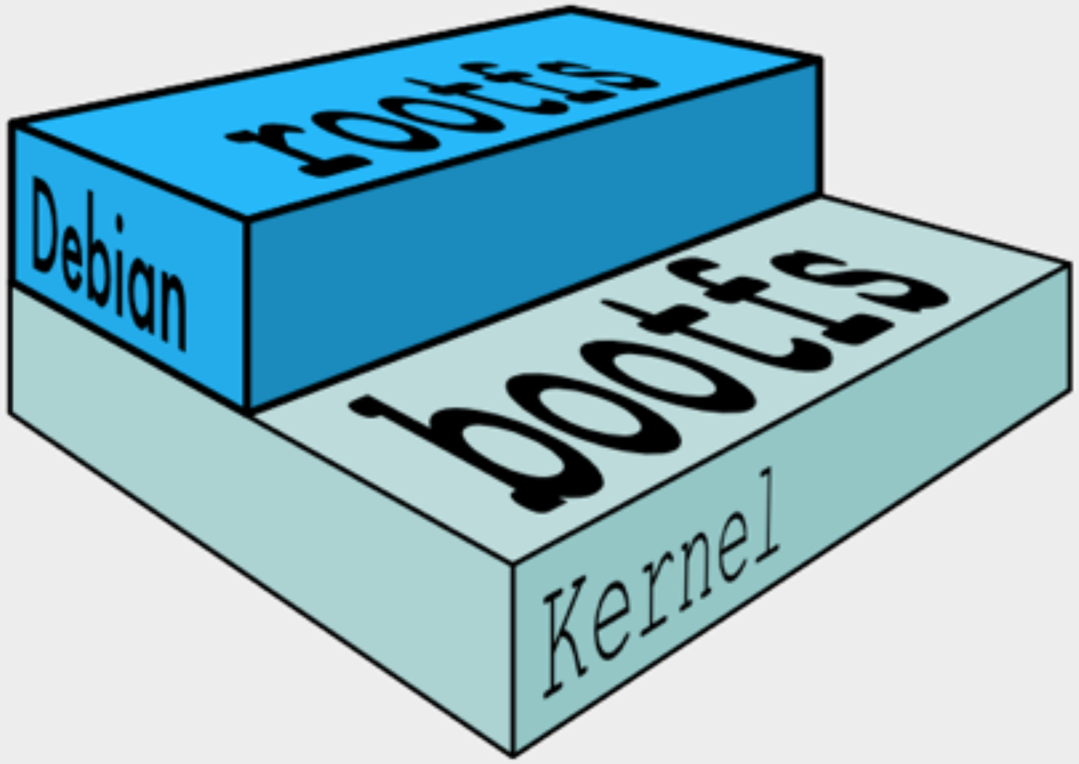
```
FROM debian:jessie
MAINTAINER Stefan Dang <sd15@sanger.ac.uk>

ENV FOO_VERSION 0.6.7

# Download dependencies & sources
RUN apt-get update && sudo apt-get install -yqq [Dependencies]

# Build foo
RUN git clone -b $FOO_VERSION https://github.com/foo.git
WORKDIR ./foo
RUN ./configure && make && make install

COPY ./entrypoint.sh /entrypoint.sh
ENTRYPOINT [”/entrypoint.sh”]
```

```
$ docker build -t ImageName ./DockefileDir/
```

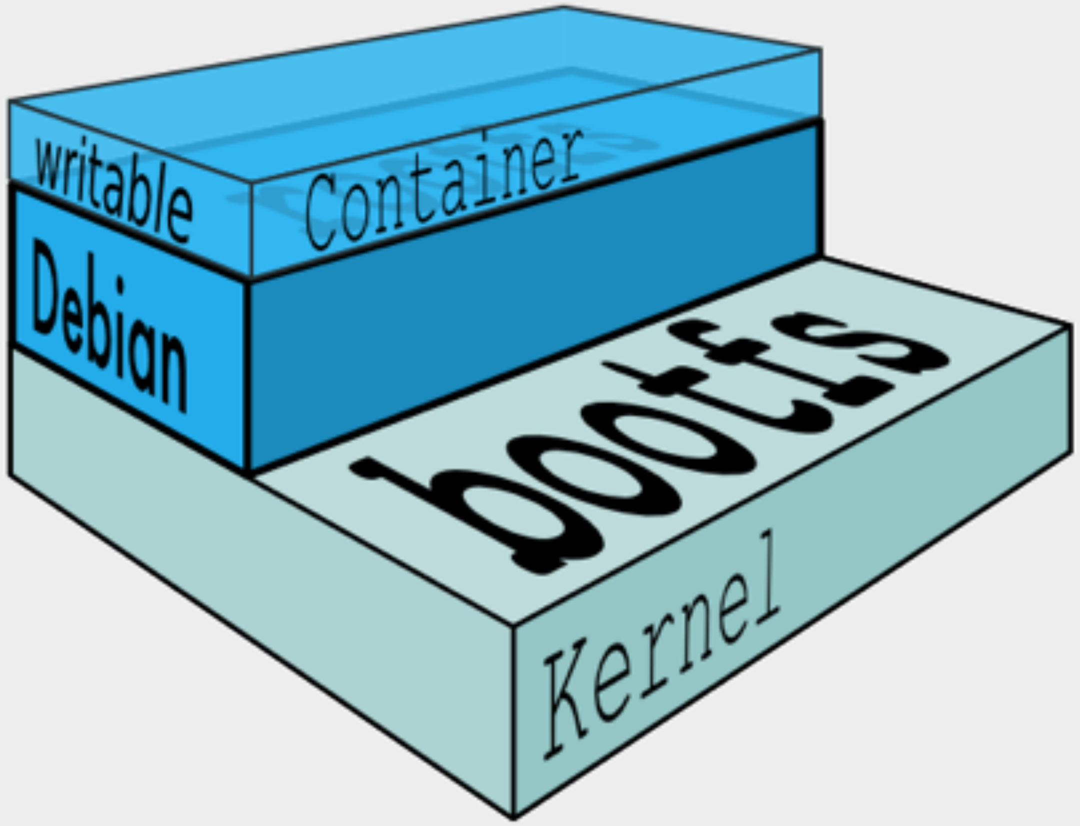


docs.docker.com/terms/image/

```
$ docker run -it ImageName
```



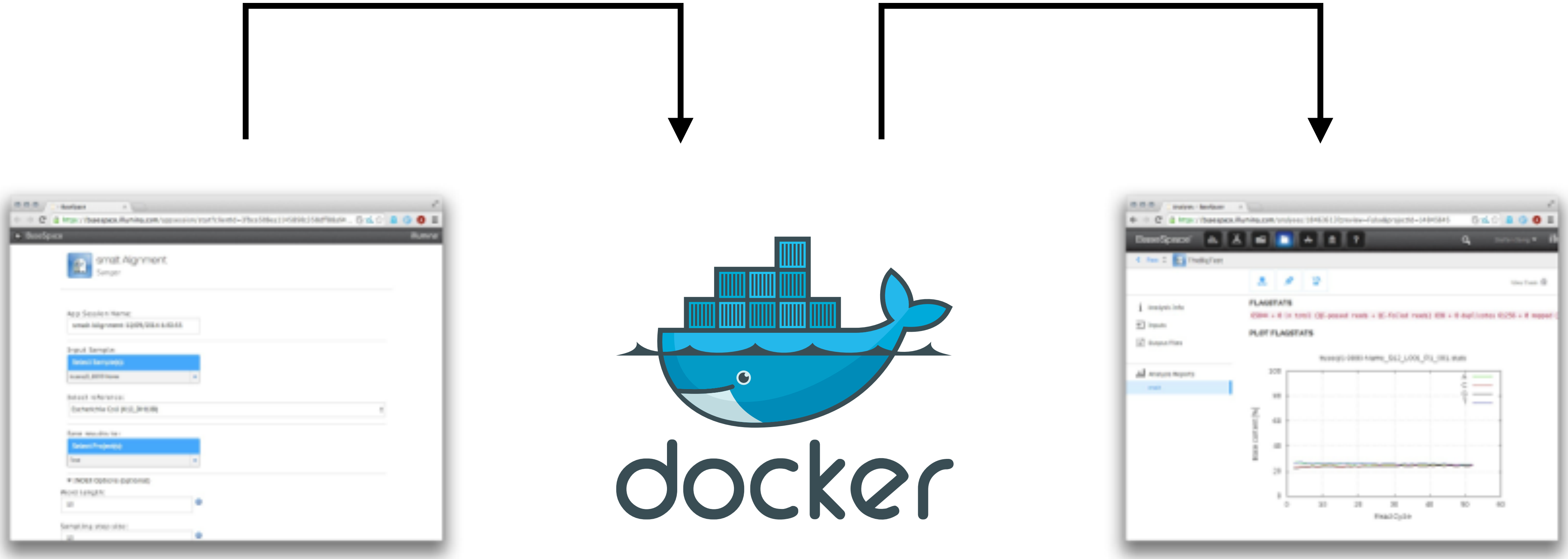
docs.docker.com/terms/image/



docs.docker.com/terms/image/

# Use Cases

# Illumina BaseSpace Native Apps



## Input Form

JSON

Javascript

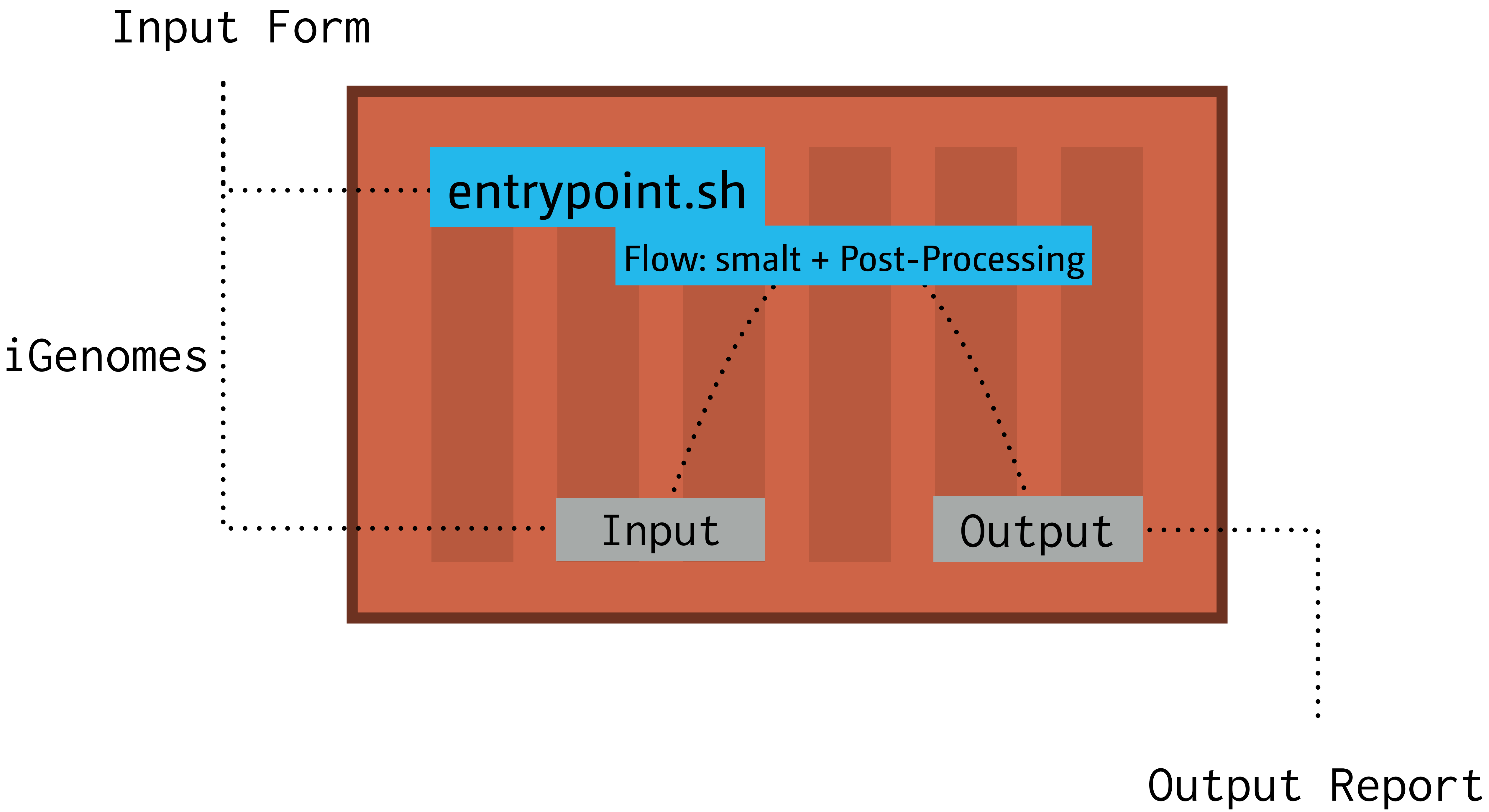
## App / Pipeline

AWS EC2

## Output Form

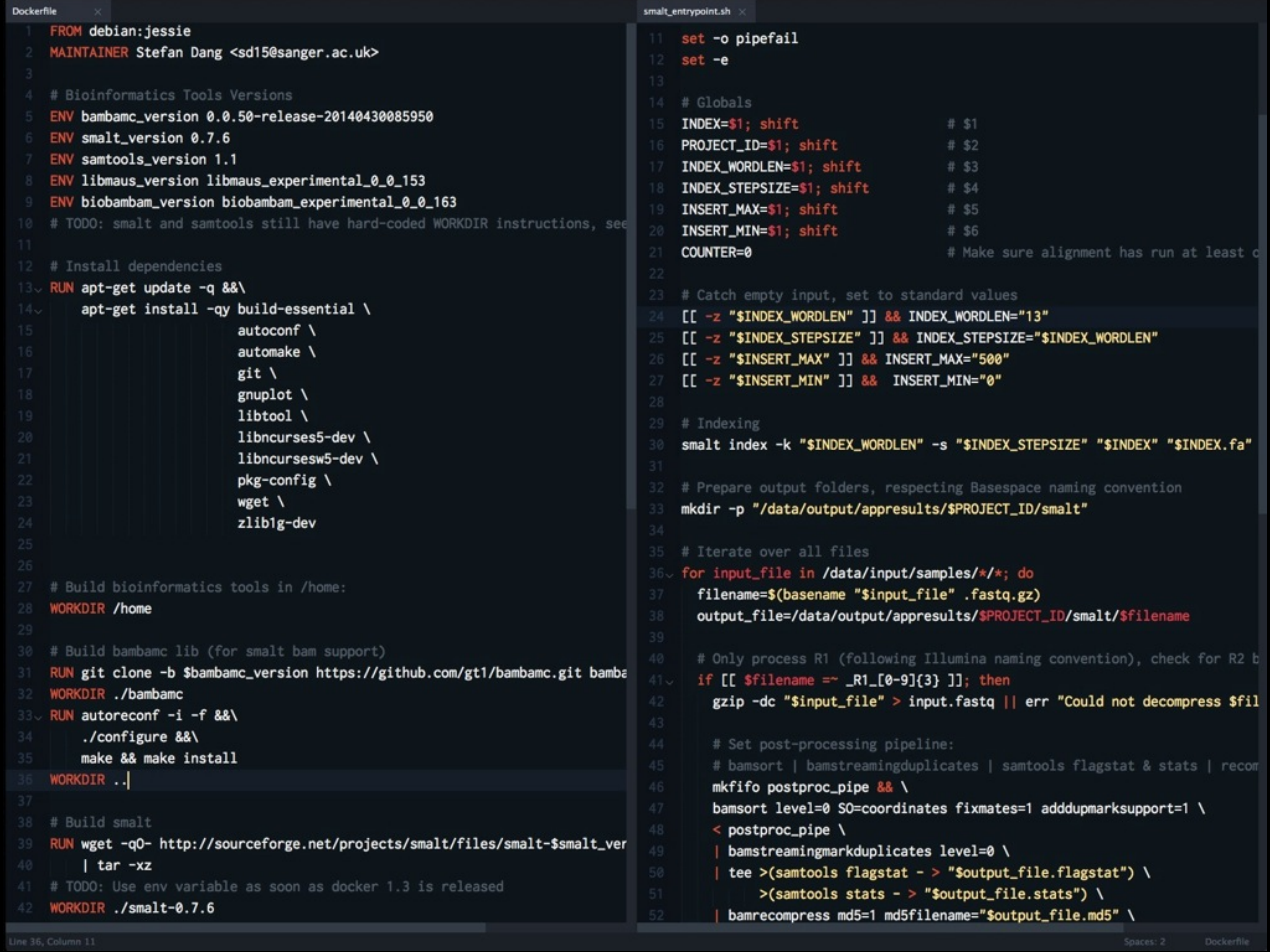
dotfluid

# Smalt Docker Container



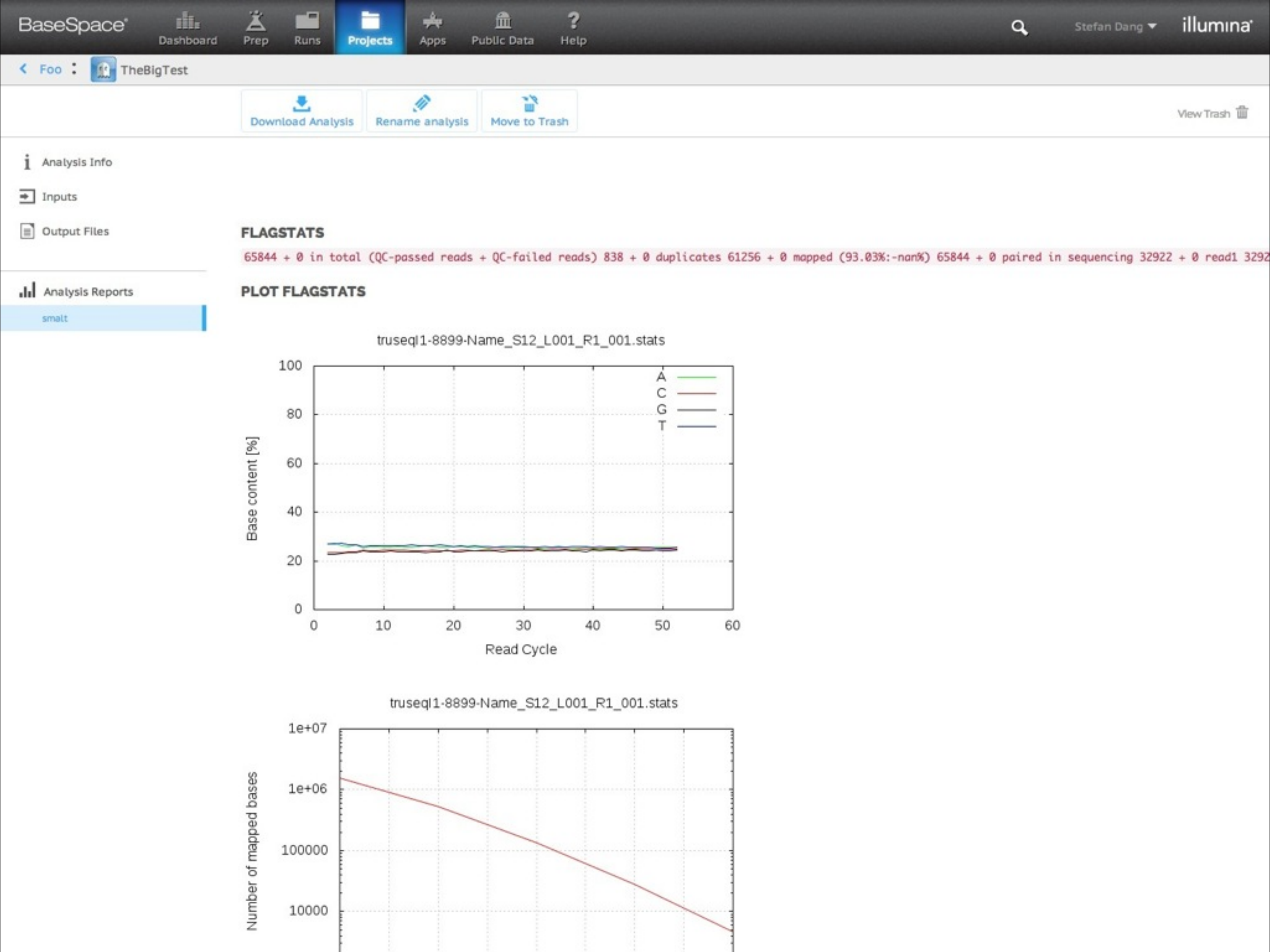
# Smalt Flow (gist)

```
$ smalt index

$ smalt map \
  > bamsort \
  | bamstreamingmarkduplicates \
  | tee >(samtools flagstat) >(samtools stats) \
  | bamrecompress md5=1 index=1 \
  > "out.bam”

$ plot-bamstats
```



```
FROM debian:jessie
MAINTAINER Stefan Dang <sd15@sanger.ac.uk>

# Bioinformatics Tools Versions
ENV bambamc_version 0.0.50-release-20140430085950
ENV smalt_version 0.7.6
ENV samtools_version 1.1
ENV libmaus_version libmaus_experimental_0_0_153
ENV biobambam_version biobambam_experimental_0_0_163
# TODO: smalt and samtools still have hard-coded WORKDIR instructions, see

# Install dependencies
RUN apt-get update -q &&\
    apt-get install -qy build-essential \
                        autoconf \
                        automake \
                        git \
                        gnuplot \
                        libtool \
                        libncurses5-dev \
                        libncursesw5-dev \
                        pkg-config \
                        wget \
                        zlib1g-dev


# Build bioinformatics tools in /home:
WORKDIR /home

# Build bambamc lib (for smalt bam support)
RUN git clone -b $bambamc_version https://github.com/gt1/bambamc.git bamba
WORKDIR ./bambamc
RUN autoreconf -i -f &&\
    ./configure &&\
    make && make install
WORKDIR ..

# Build smalt
RUN wget -qO- http://sourceforge.net/projects/smalt/files/smalt-$smalt_ver
    | tar -xz
# TODO: Use env variable as soon as docker 1.3 is released
WORKDIR ./smalt-0.7.6
```

```
set -o pipefail
set -e

# Globals
INDEX=$1; shift                   # $1
PROJECT_ID=$1; shift              # $2
INDEX_WORDLEN=$1; shift           # $3
INDEX_STEPSIZE=$1; shift          # $4
INSERT_MAX=$1; shift              # $5
INSERT_MIN=$1; shift              # $6
COUNTER=0                         # Make sure alignment has run at least o

# Catch empty input, set to standard values
[[ -z "$INDEX_WORDLEN" ]] && INDEX_WORDLEN="13"
[[ -z "$INDEX_STEPSIZE" ]] && INDEX_STEPSIZE="$INDEX_WORDLEN"
[[ -z "$INSERT_MAX" ]] && INSERT_MAX="500"
[[ -z "$INSERT_MIN" ]] &&  INSERT_MIN="0"

# Indexing
smalt index -k "$INDEX_WORDLEN" -s "$INDEX_STEPSIZE" "$INDEX" "$INDEX.fa"

# Prepare output folders, respecting Basespace naming convention
mkdir -p "/data/output/appresults/$PROJECT_ID/smalt"

# Iterate over all files
for input_file in /data/input/samples/*/*; do
  filename=$(basename "$input_file" .fastq.gz)
  output_file=/data/output/appresults/$PROJECT_ID/smalt/$filename

  # Only process R1 (following Illumina naming convention), check for R2 b
  if [[ $filename =~ _R1_[0-9]{3} ]]; then
    gzip -dc "$input_file" > input.fastq || err "Could not decompress $fil

    # Set post-processing pipeline:
    # bamsort | bamstreamingduplicates | samtools flagstat & stats | recom
    mkfifo postproc_pipe && \
    bamsort level=0 SO=coordinates fixmates=1 adddupmarksupport=1 \
    < postproc_pipe \
    | bamstreamingmarkduplicates level=0 \
    | tee >(samtools flagstat - > "$output_file.flagstat") \
          >(samtools stats - > "$output_file.stats") \
    | bamrecompress md5=1 md5filename="$output_file.md5" \
```

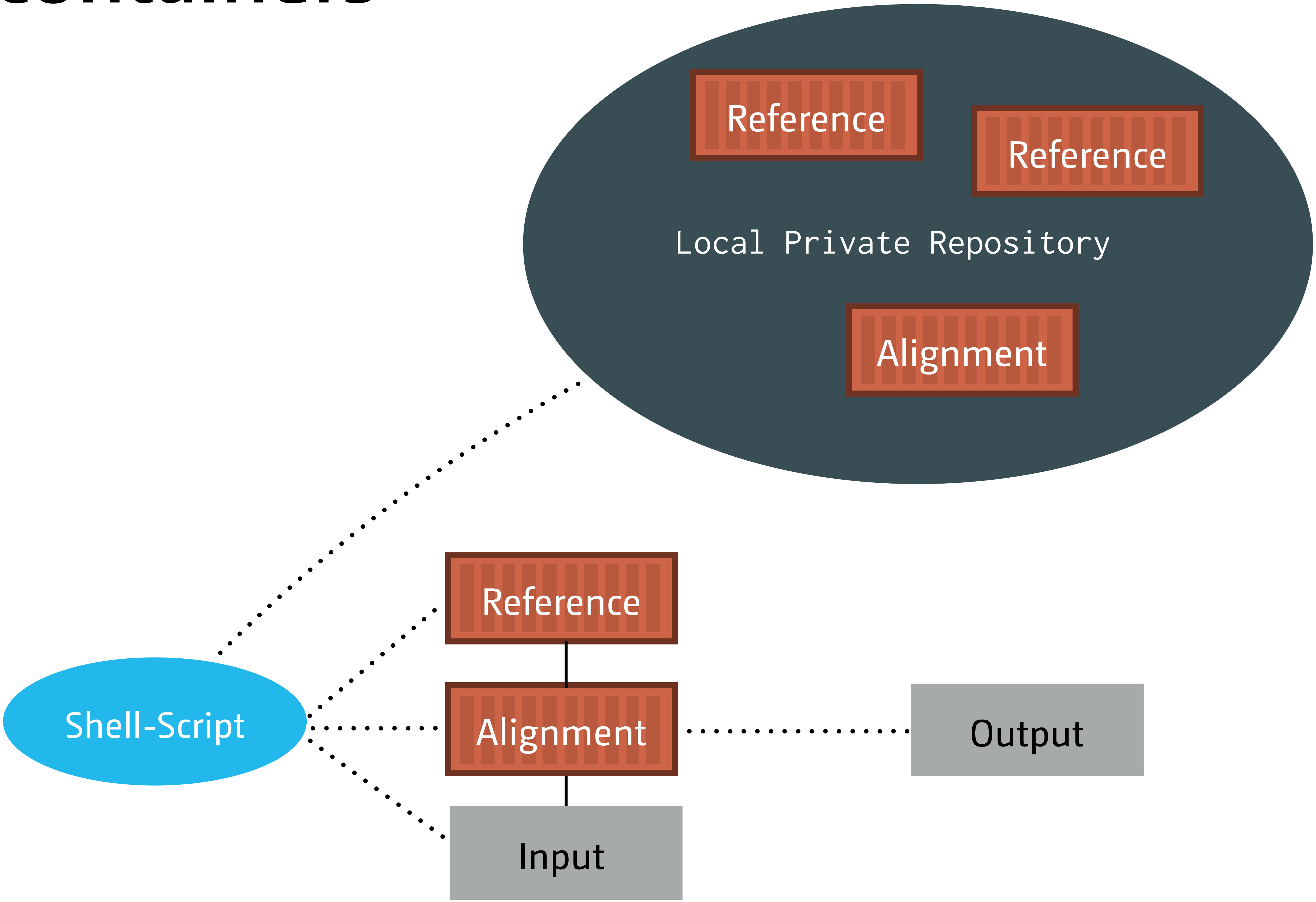


### FLAGSTATS

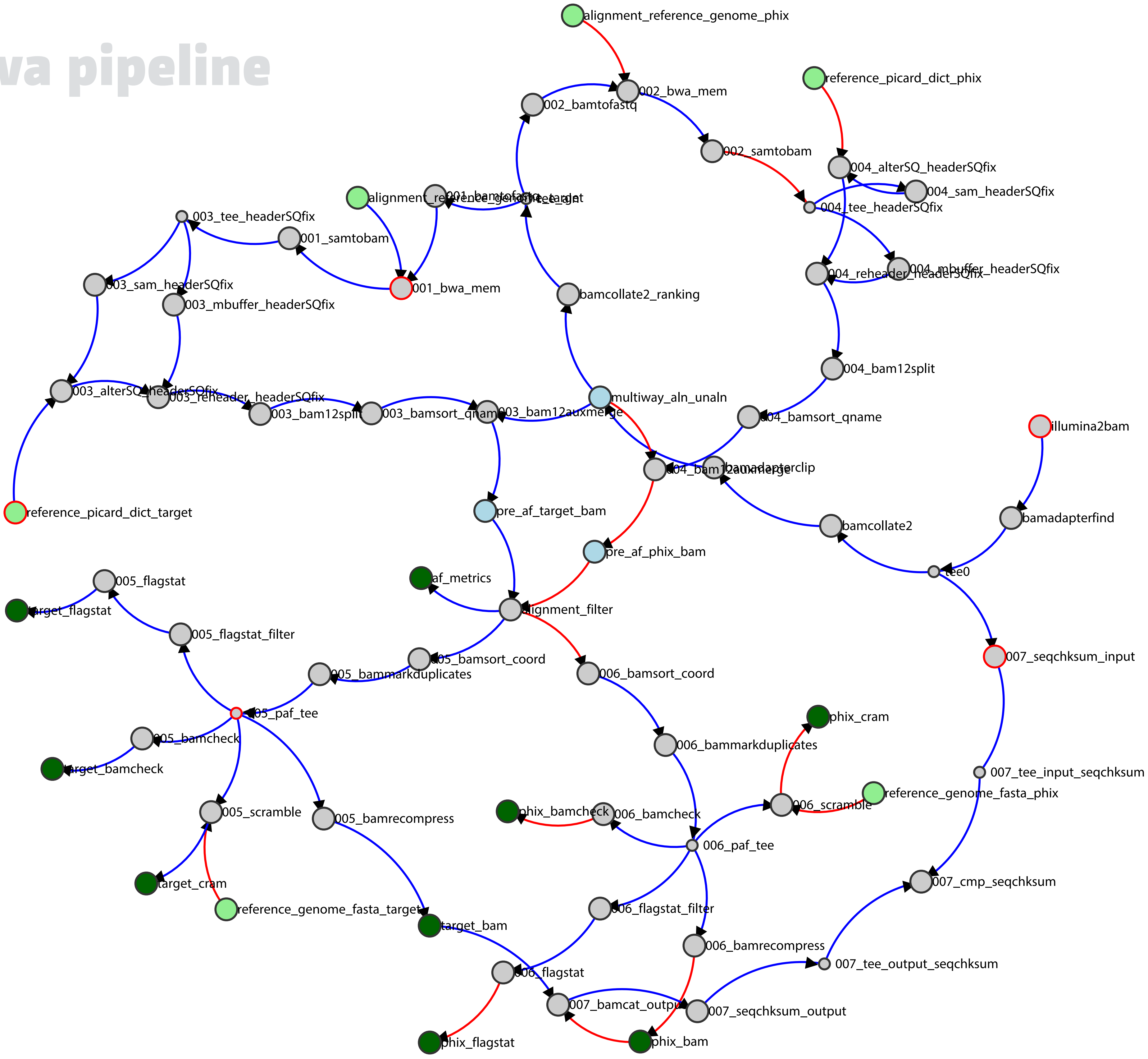
```
65844 + 0 in total (QC-passed reads + QC-failed reads) 838 + 0 duplicates 61256 + 0 mapped (93.03%:-nan%) 65844 + 0 paired in sequencing 32922 + 0 read1 3292
```

### PLOT FLAGSTATS

# Pipelines in orchestrated Docker containers



# p4 bwa pipeline



libmaus<br>biobambam

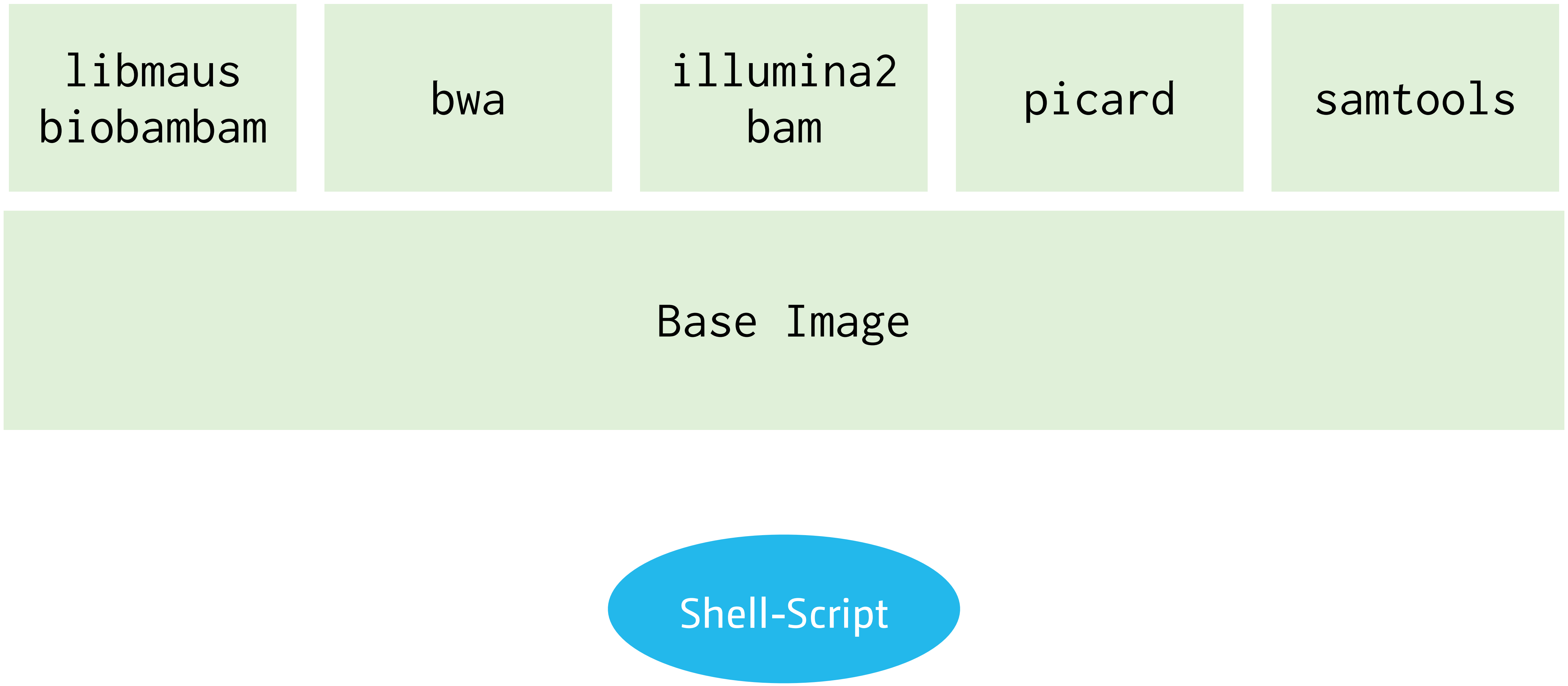
bwa

illumina2<br>bam

picard

samtools

# Docker Build System



libmaus
biobambam

bwa

illumina2
bam

picard

samtools

Base Image

Shell-Script

# Overhead

```
$ time docker run ubuntu:14.04
  0.01s user 0.02s system 2% cpu 1.007 total

$ docker images
  REPOSITORY                    VIRTUAL SIZE
  ubuntu:14.04                   194.9 MB
  debian:jessie                  120.0 MB
  p4                            1259.0 MB
  p4_flattened                   484.1 MB
  p4_build-system                294.2 MB
  PhiX_iGenome                    16.9 MB
```

# Summary

# Drawbacks

## System requirements

- 64bit, Linux 3.8 or later kernel, Storage Driver Support
- Boot2Docker

## Security

- docker daemon requires root privileges
- less mature / tested compared to VMs
- less isolation compared to VMs

# Advantages

Open Source

- github, Go, Apache 2.0 license

Speed & Maintenance

- predictable
- repeatable
- managed

stepf

# Thanks

Marina Gourtovaia

Kevin Lewis

David Jackson

# Thank you!

# Questions?