Overview of the 2019 Open-Source IR Replicability Challenge (OSIRRC 2019)

Ryan Clancy, Nicola Ferro, Claudia Hauff,
Jimmy Lin, Tetsuya Sakai, Ze Zhong Wu

Vision

Source: saveur.com

Vision

The ultimate candy store for information retrieval researchers!

Source: Wikipedia (Candy)

Vision

The ultimate candy store for information retrieval researchers!

See a result you like?
Click a button to recreate those results!

Really, *any* result?
(not quite… let's start with batch *ad hoc* retrieval experiments on standard test collections)

What is this, really?

Repeatability: you can recreate your own results again

We get this "for free"

Replicability: others can recreate your results (with your code)

Our focus

Reproducibility: others can recreate your results (with code they rewrite)

Stepping stone…

ACM Artifact Review and Badging Guidelines

Why is this important?

Good science

Sustained cumulative progress

Armstrong et al. (CIKM 2009): Little empirical progress made from 1998 to 2009

Why? researchers compare against weak baselines

Yang et al. (SIGIR 2019): Researchers *still* compare against weak baselines

How do we get there?

Open-Source Code!

A good start, but far from enough…

TREC 2015 "Open Runs"
79 submitted runs…

Voorhees et al. Promoting Repeatability Through Open Runs. *EVIA 2016*.

Number of runs successfully replicated

Voorhees et al. Promoting Repeatability Through Open Runs. *EVIA 2016*.

How do we get there?

Open-Source Code!

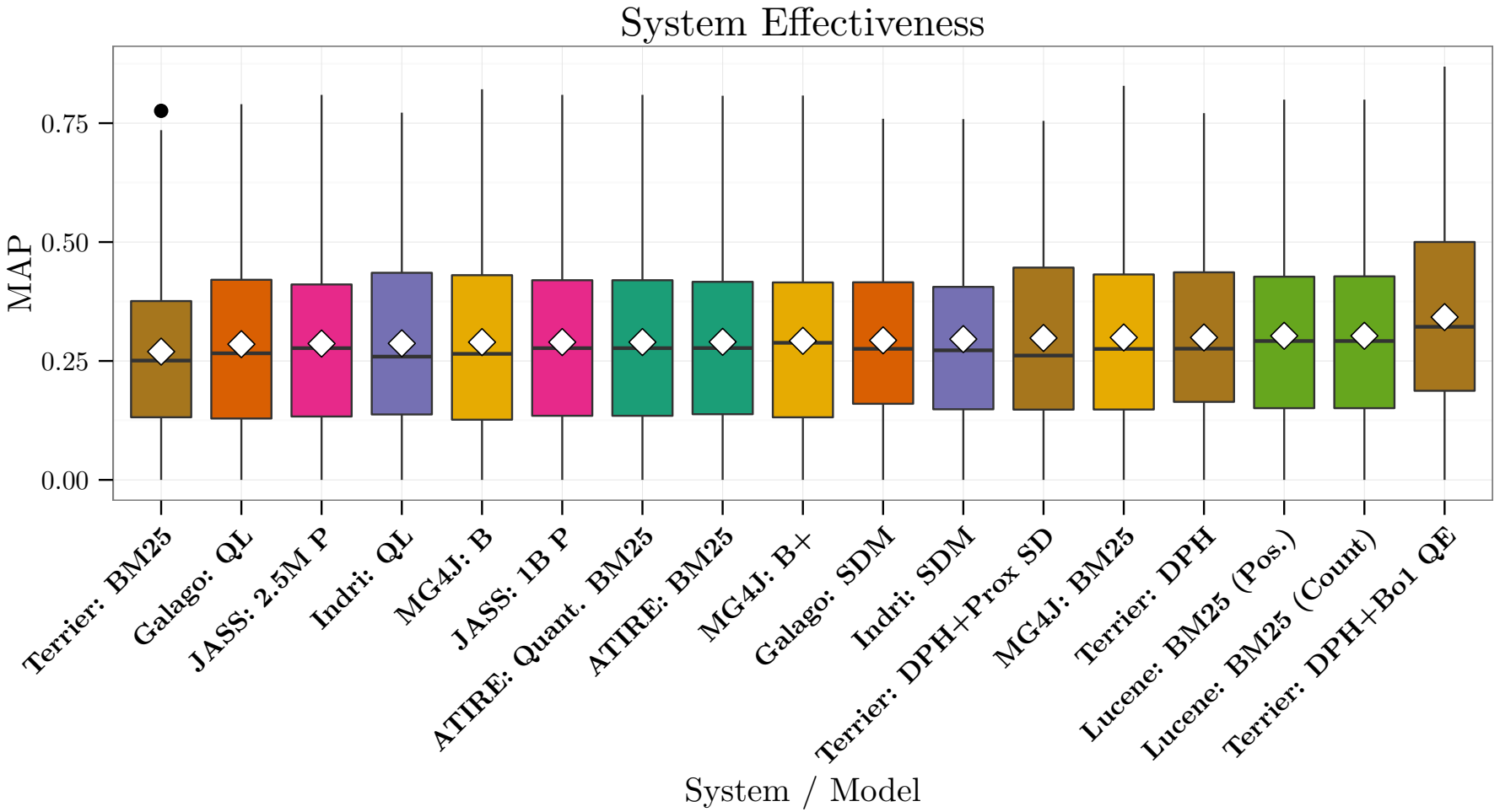
A good start, but far from enough…

Ask developers to show us how!

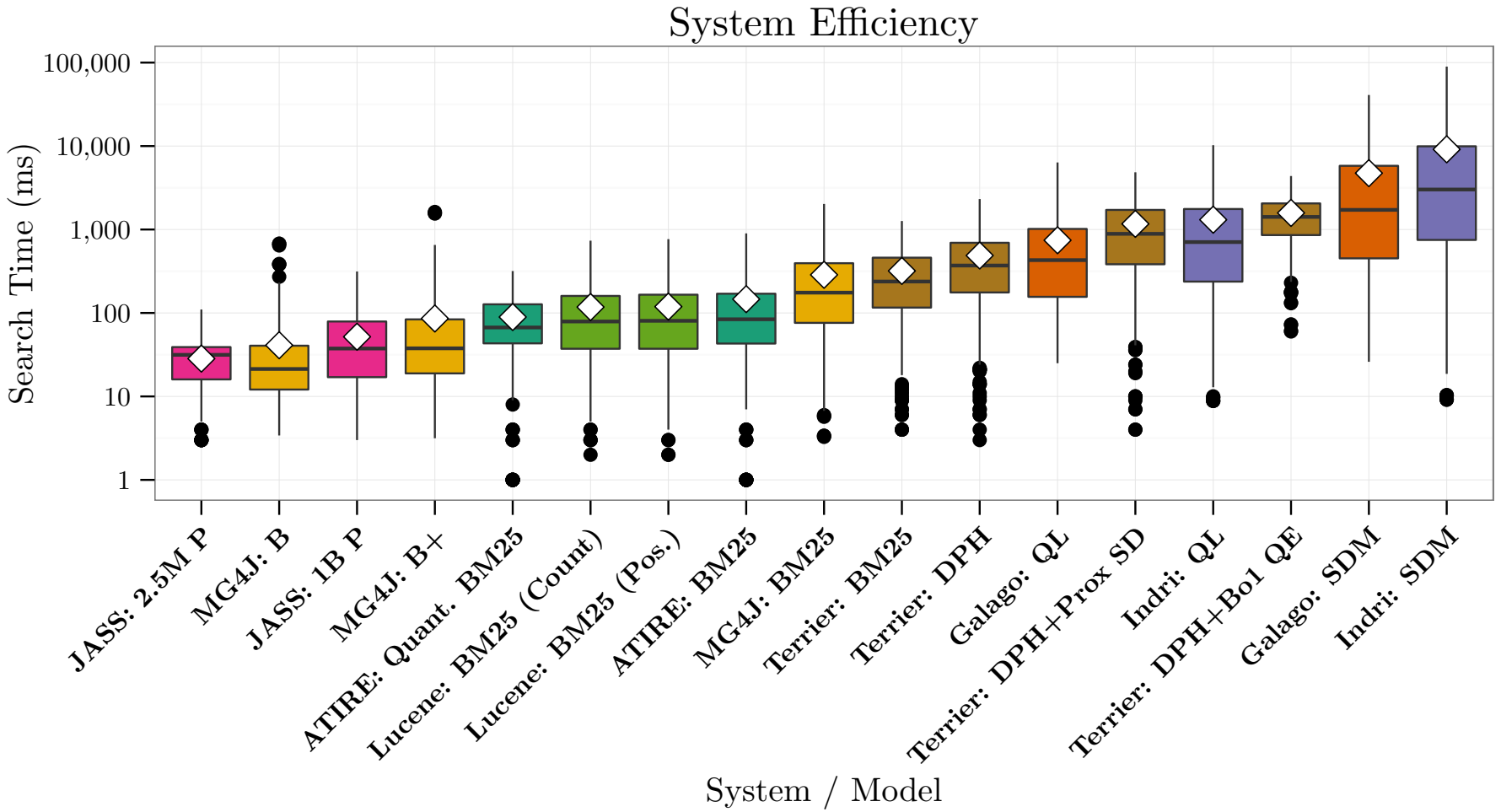
Open-Source IR Reproducibility Challenge (OSIRRC), SIGIR 2015 Workshop on Reproducibility, Inexplicability, and Generalizability of Results (RIGOR)

Participants contributed end-to-end scripts for replicating *ad hoc* retrieval experiments

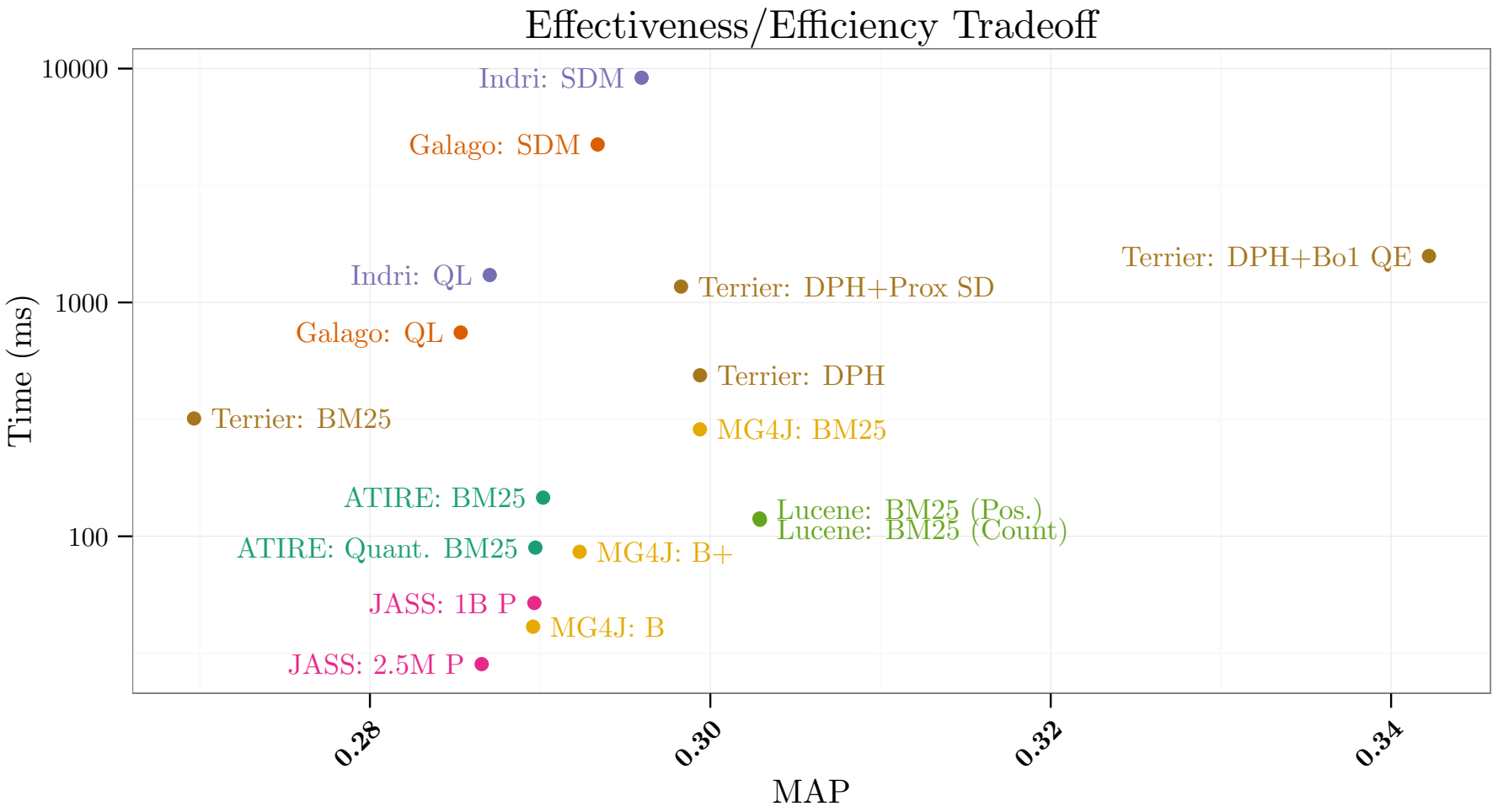
Lin et al. Toward Reproducible Baselines: The Open-Source IR Reproducibility Challenge. *ECIR 2016*.



7 participating systems, GOV2 collection



7 participating systems, GOV2 collection



7 participating systems, GOV2 collection

How do we get there?

Open-Source Code!

A good start, but far from enough…

Ask developers to show us how!

It worked, but…

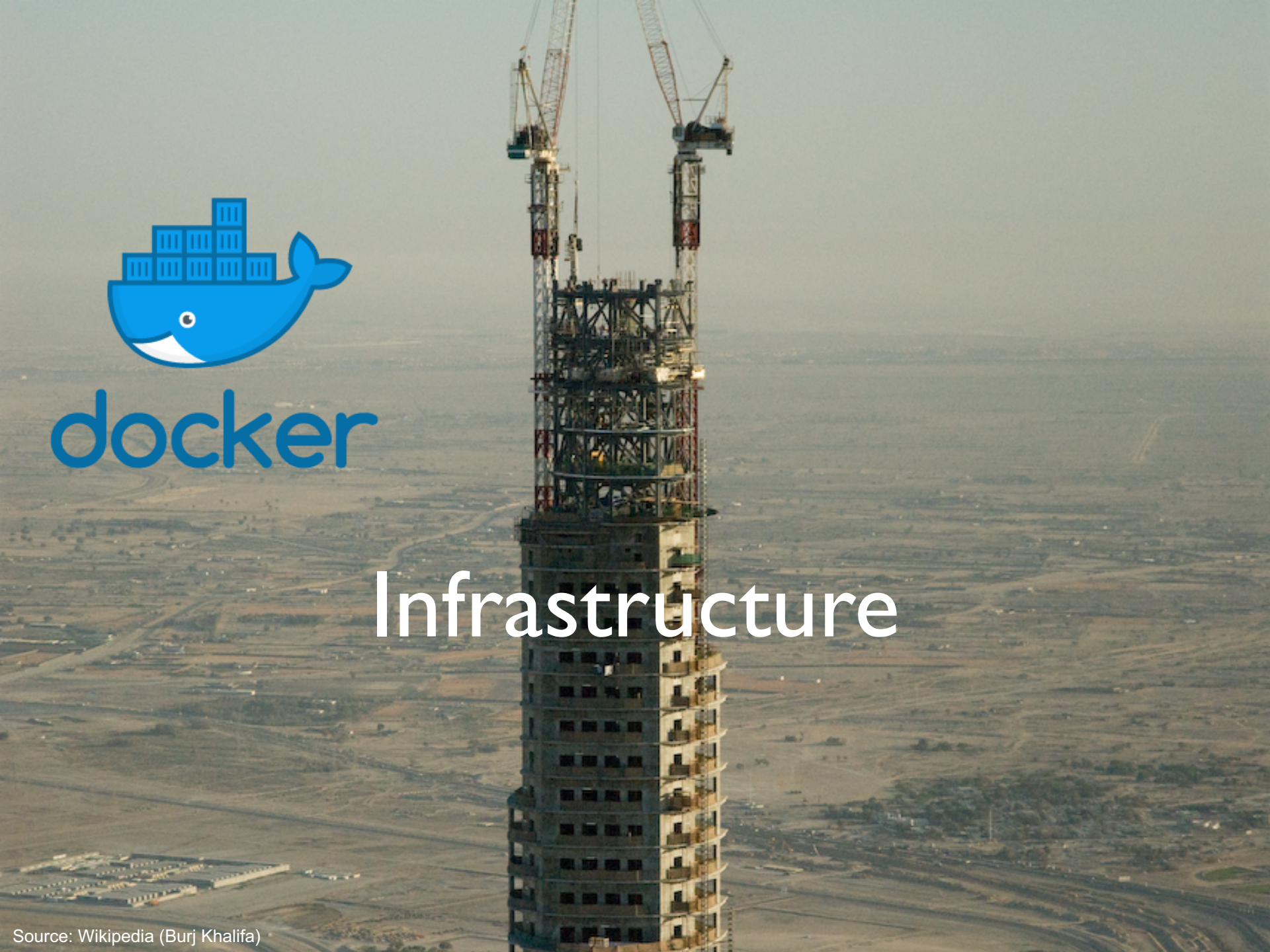
What worked well?

We actually pulled it off!

What didn't work well?

Technical infrastructure was brittle

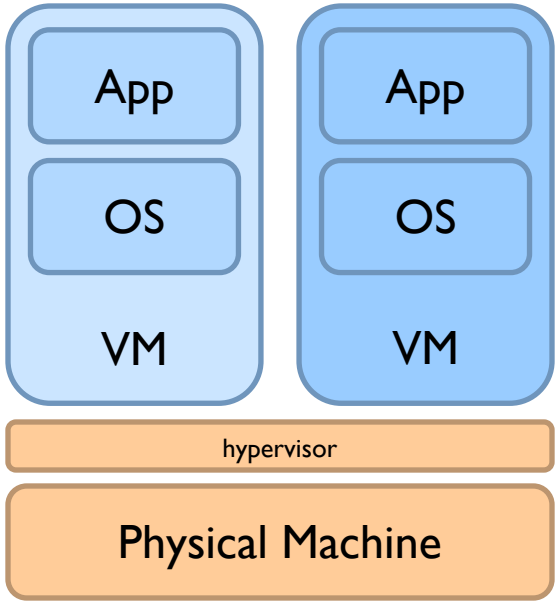
Replication scripts too under-constrained



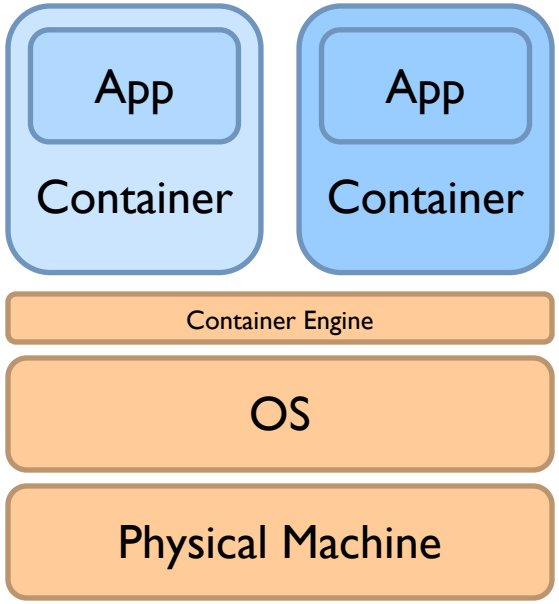
Infrastructure

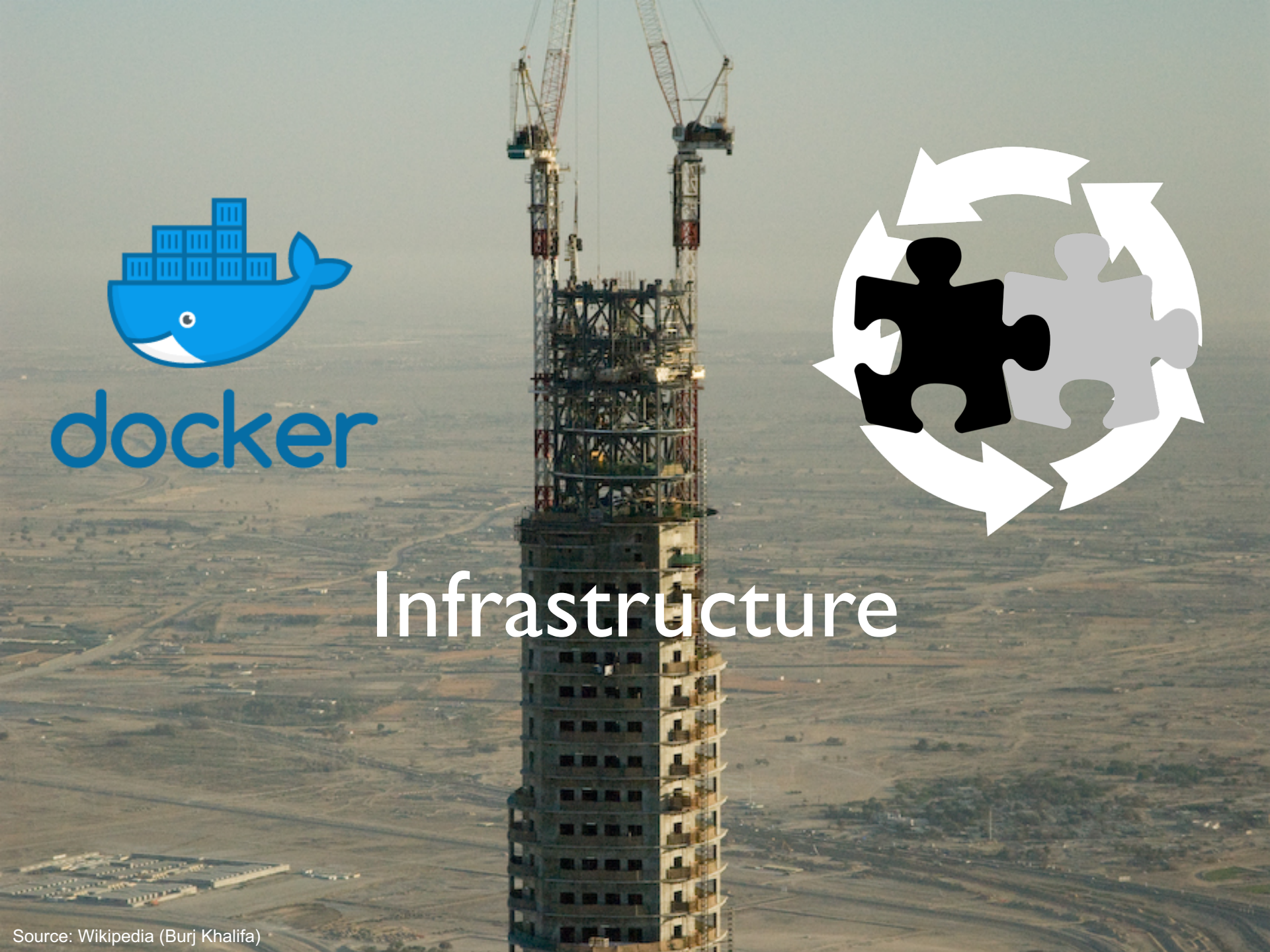
Source: Wikipedia (Burj Khalifa)

VMs



Containers


App
Container
App
Container
Container Engine
OS
Physical Machine

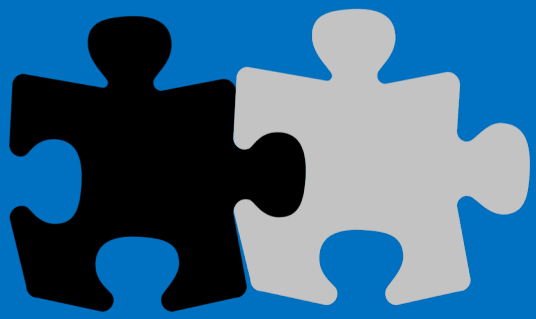


Infrastructure

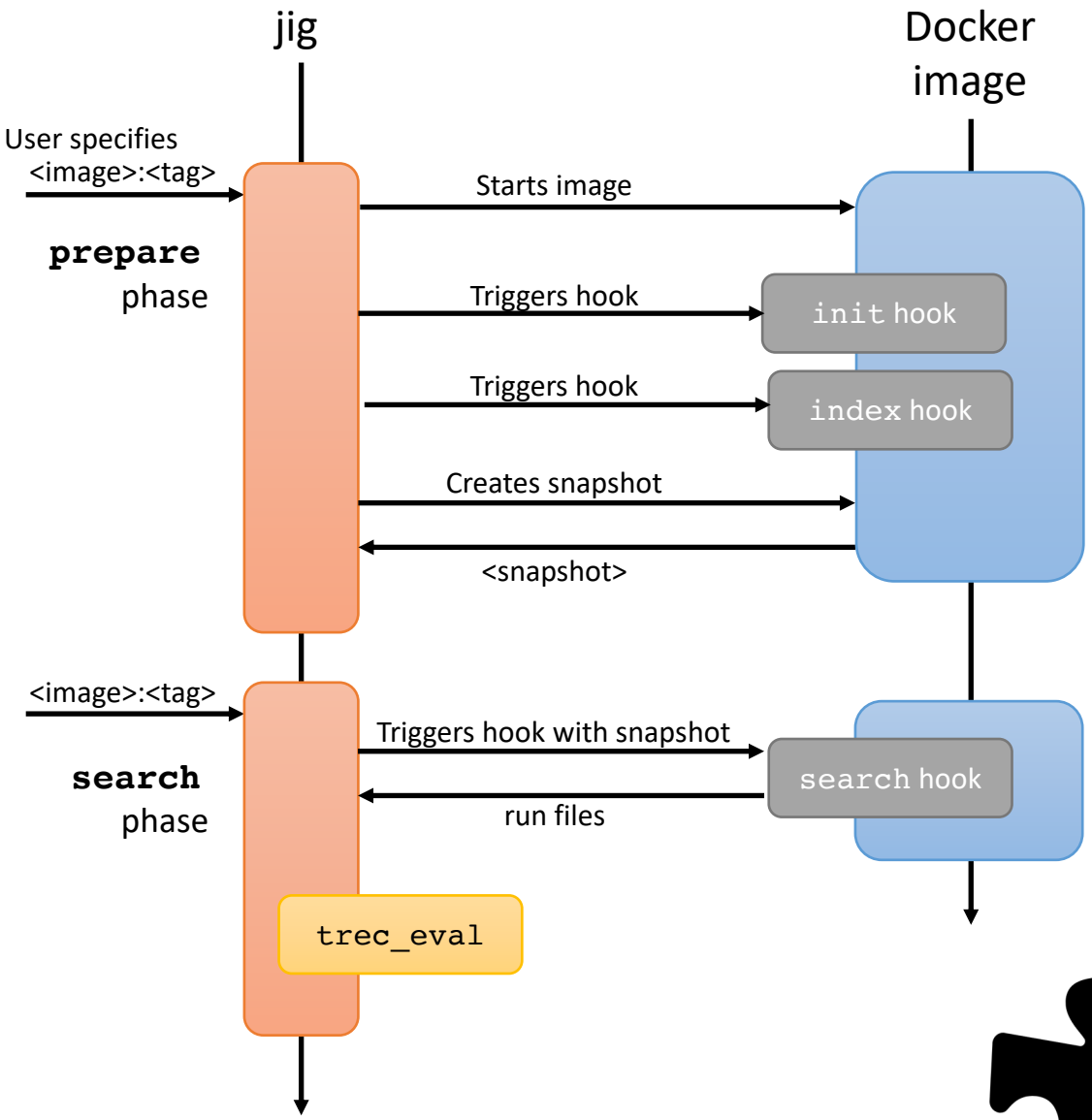
Source: Wikipedia (Burj Khalifa)

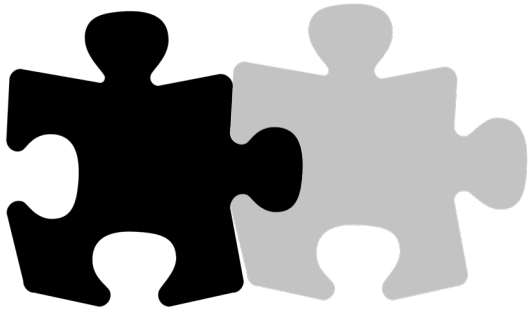
Workshop Goals

1. Develop common Docker specification for capturing *ad hoc* retrieval experiments – the "jig".
2. Build a library of curate images that work with the jig.



3. Take over the world!
   (encourage adoption, broaden to other tasks, etc.)


jig
Docker image
User specifies <image>:<tag>
prepare phase
Starts image
Triggers hook
init hook
Triggers hook
index hook
Creates snapshot
<snapshot>
<image>:<tag>
search phase
Triggers hook with snapshot
search hook
run files
trec_eval



Source: Flickr (https://www.flickr.com/photos/m00k/15789986125/)

17 images
13 different teams

Focus on newswire collections: Robust04, Core17, Core18
Official runs on Microsoft Azure

Thanks Microsoft
for free credits!

Anserini (University of Waterloo)
Anserini-bm25prf (Waseda University)
ATIRE (University of Otago)
Birch (University of Waterloo)
Elastirini (University of Waterloo)
EntityRetrieval (Ryerson University)
Galago (University of Massachusetts)
ielab (University of Queensland)
Indri (TU Delft)
IRC-CENTRE2019 (Technische Hochschule Köln)
JASS (University of Otago)
JASSv2 (University of Otago)
NVSM (University of Padua)
OldDog (Radboud University)
PISA (New York University and RMIT University)
Solrini (University of Waterloo)
Terrier (TU Delft and University of Glasgow)

Robust04

49 runs from 13 images

Images captured diverse models:

query expansion and relevance feedback

conjunctive and efficiency-oriented query processing

neural ranking models

Core I7

12 runs from 6 images

Core18

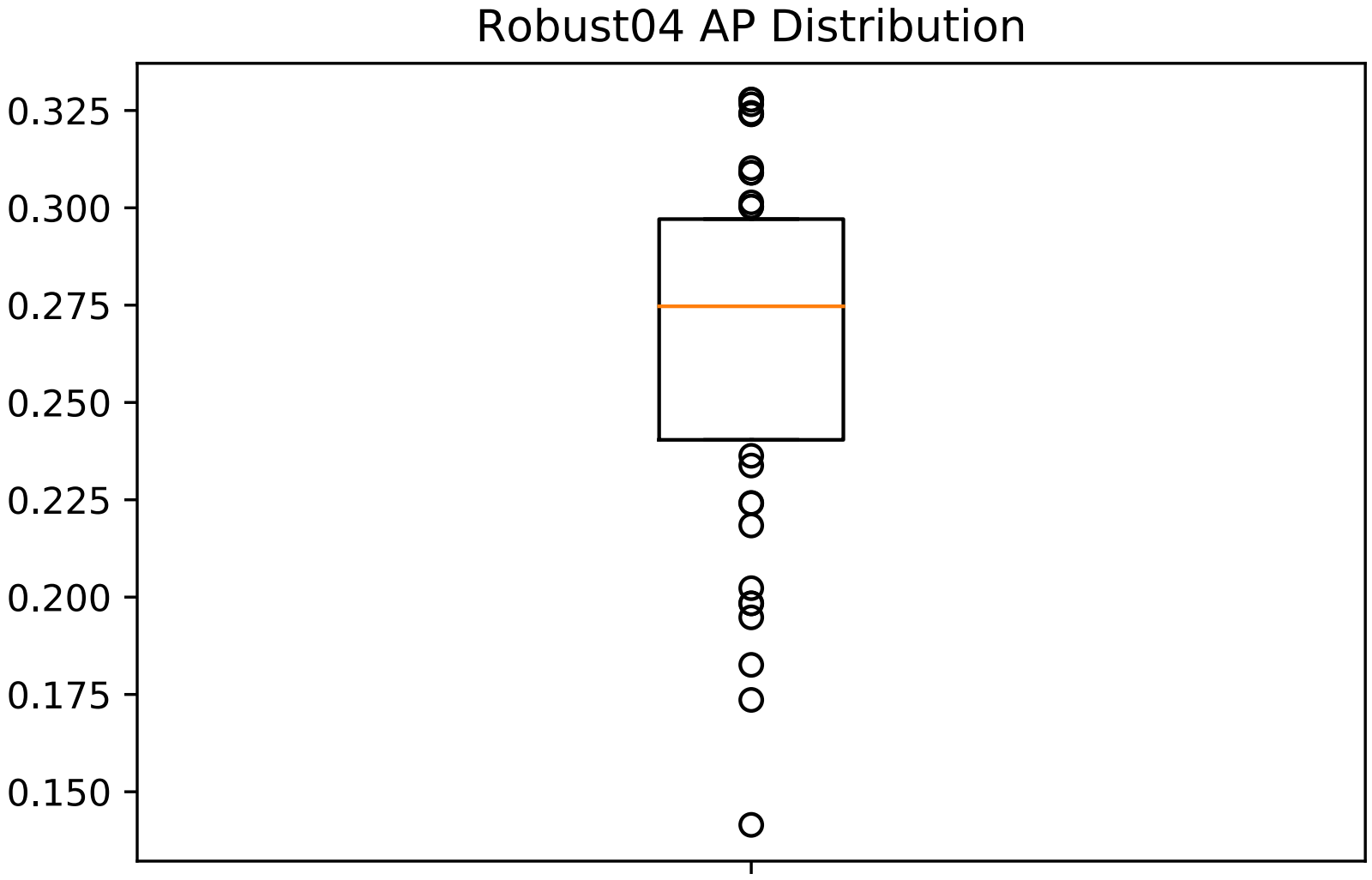
19 runs from 4 images

Robust04

49 runs from 13 images

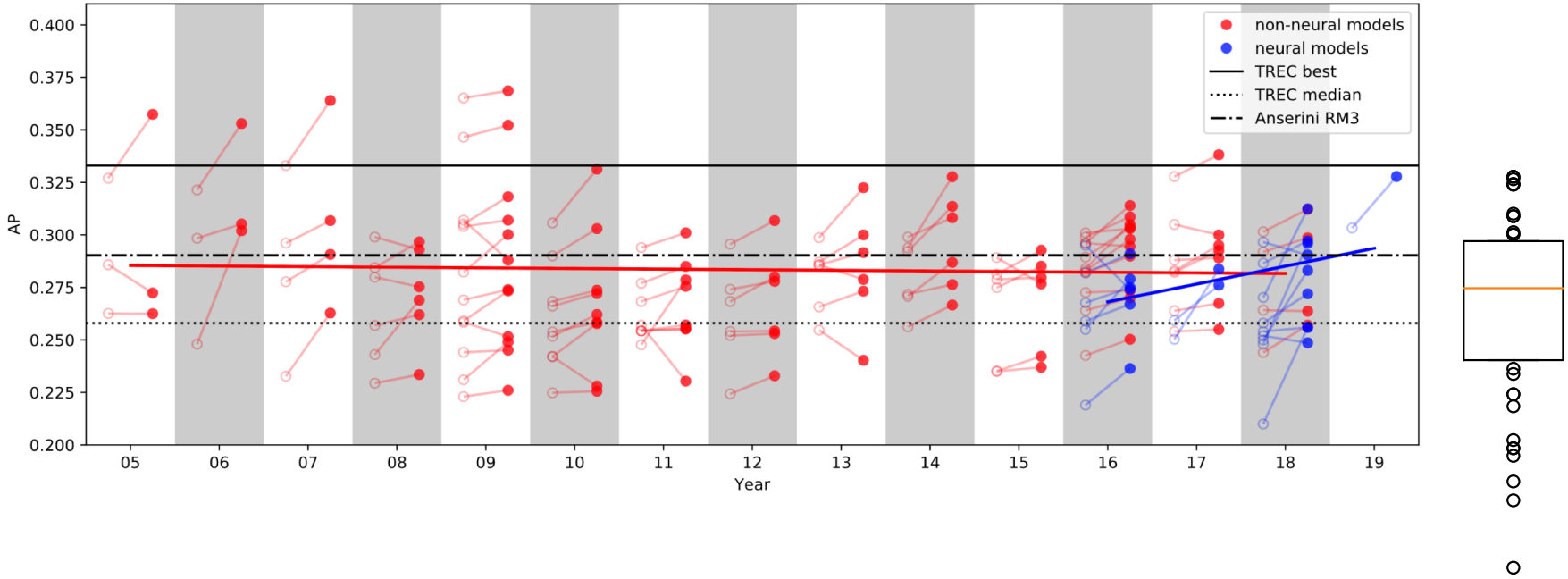
Source: Time Magazine

Source: Washington Post



TREC best – 0.333

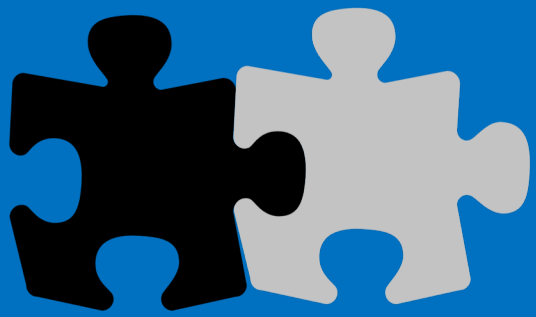
TREC median (title) – 0.258


non-neural models
neural models
TREC best
TREC median
Anserini RM3
AP
0.400
0.375
0.350
0.325
0.300
0.275
0.250
0.225
0.200
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
Year

Workshop Goals

✓ 1. Develop common Docker specification for capturing *ad hoc* retrieval experiments – the "jig".

✓ 2. Build a library of curate images that work with the jig.



? 3. Take over the world!
(encourage adoption, broaden to other tasks, etc.)

Source: flickr (https://www.flickr.com/photos/39414578@N03/16042029002)