Reliable Data Processing

with Minimal Toil

Pieter Coucke, Google
Julia Lee, Slack

SREcon 21
October 14, 2021

Site Reliability Engineering

Pieter Coucke

Technical Program Manager at Google SRE Zürich

Google Workspace
Gmail, Drive, Calendar, Meet and Docs

linkedin.com/in/pcoucke

Photo by Carlos Alberto Gómez Iñiguez on Unsplash

Julia Lee

Senior Software Engineer in Infrastructure at Slack

Leads development of asynchronous compute services


An apple a day
keeps the SRE away
Photo by JK Sloan on Unsplash

Photo by Seb Mooze on Unsplash

Haunted Graveyards

Image: publicdomainpictures.net

Manual Process

Photo by Nicolas Picard on Unsplash

SRE support

Photo by Anita Jankovic on Unsplash

What is a Batch Job?

Photo by Lajos Szabo on Unsplash

Photo by Brett Zeck on Unsplash

Correct and **fresh** results through

- Automated rollouts
- Validations
- Canarying

Photo by Jacek Dylag on Unsplash

Photo by Steve Johnson on Unsplash

Benefits

Photo by Patrick Fore on Unsplash

Challenges

Large blast radius
Data corruption
Downstream delay

Staleness of results
Overloaded backend servers
High cost
Duplicated and divergent logic

Photo by Maria Teneva on Unsplash

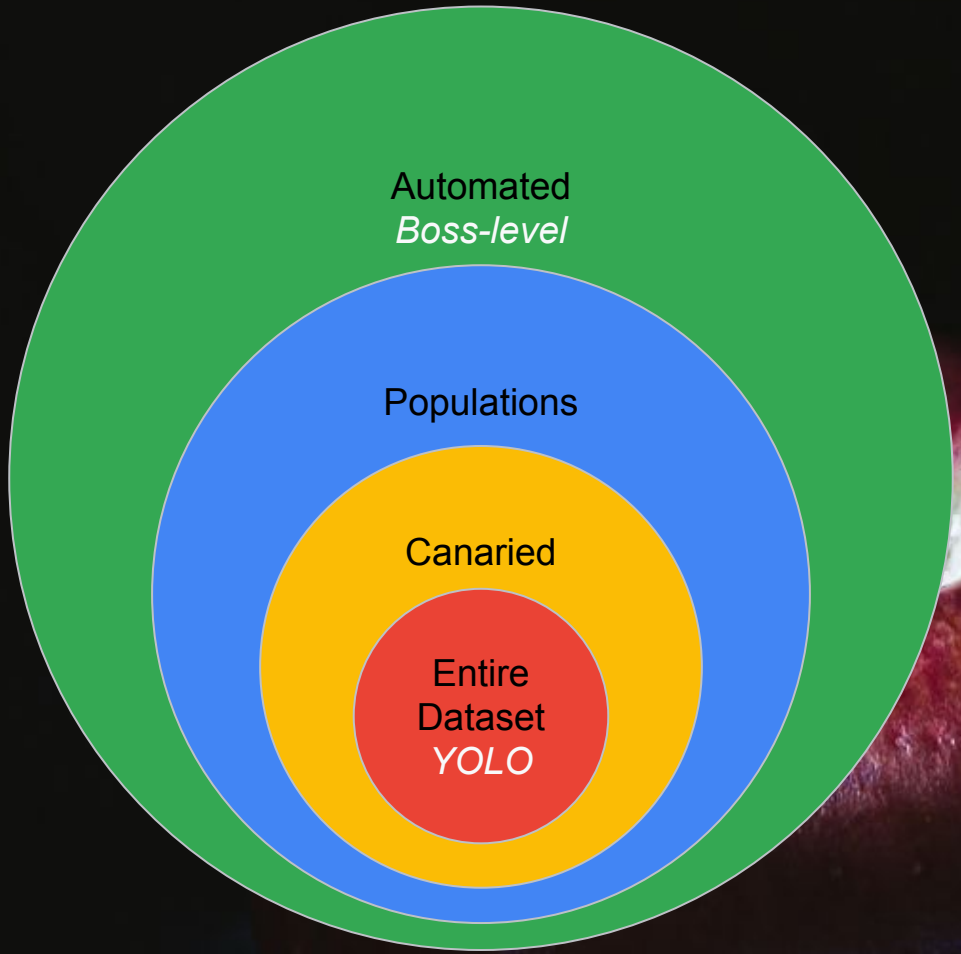
10 PM

Do You Know Where Your Apple Is ?

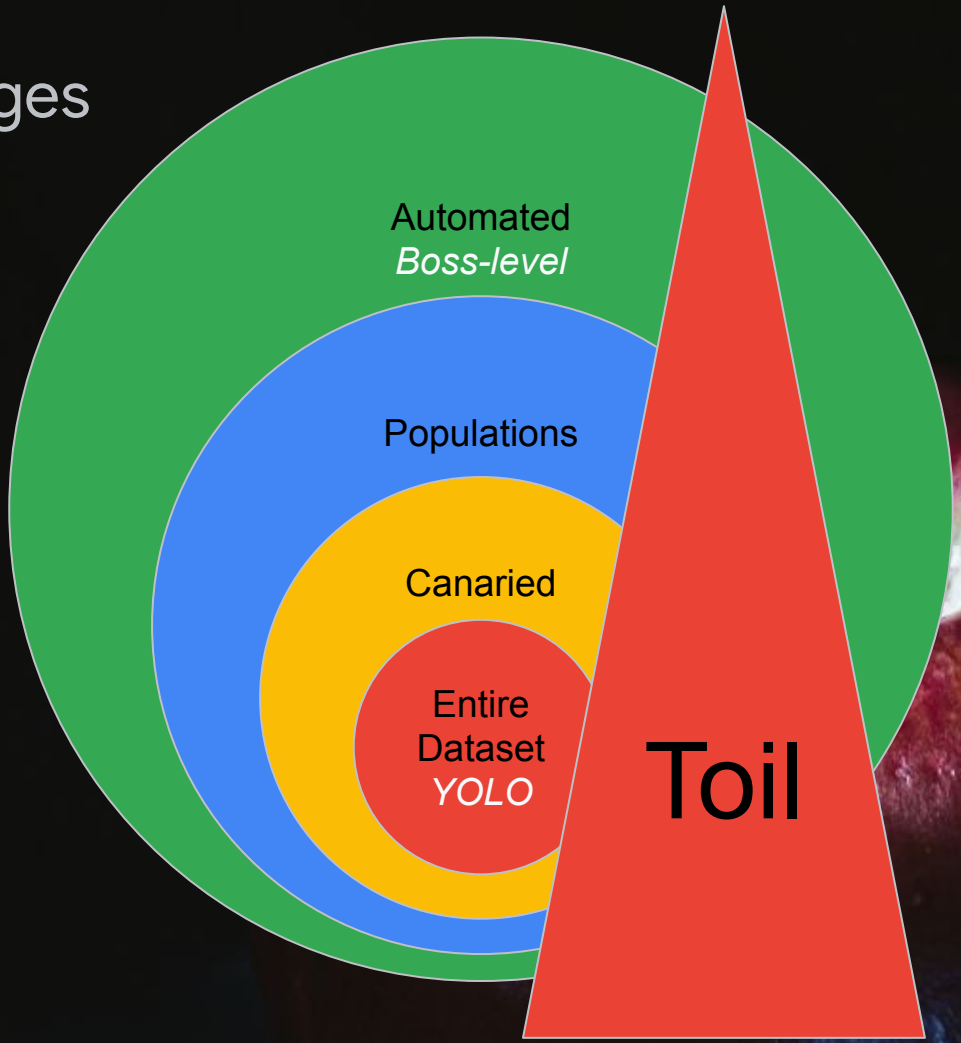
Plan

1. **Identification**
2. **Stop bleeding**
3. **Prioritize**
4. **Scale up**

Safety Levels



Making Changes



Benefits of Standardization

Expertise
Productivity
Incident response
Structured data

Photo by Ana Essentiels on Unsplash

Dry-run

Skip writing

Temp location

Permissions

Photo from pixabay.com

Code Deployment Schedule …

… is not the job run schedule

Release Stages

Photo from maxpixel.net
Photo by IamCristian on Unsplash
Photo by Amit Lahav on Unsplash

Data Sets

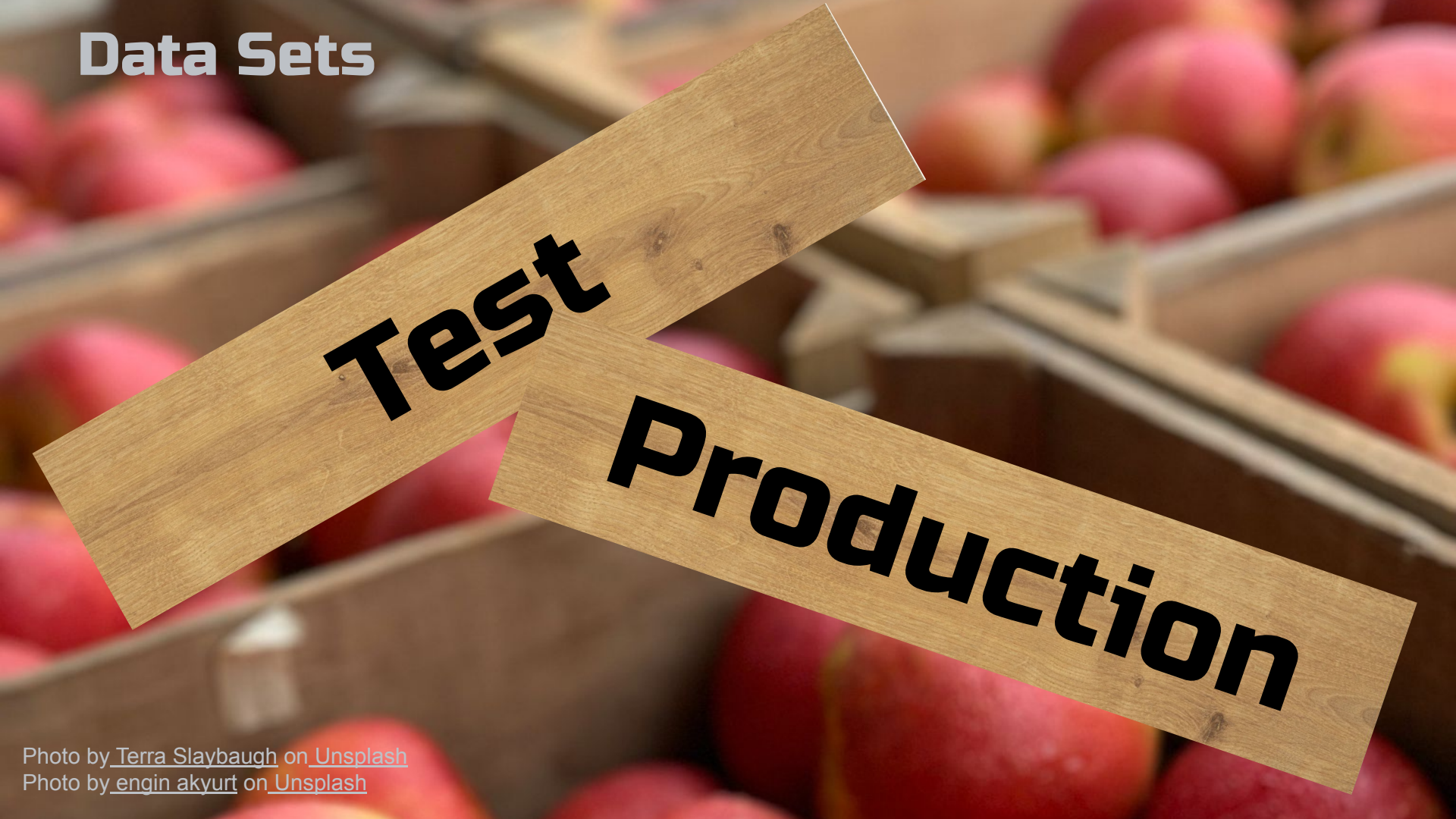


Photo by Terra Slaybaugh on Unsplash
Photo by engin akyurt on Unsplash

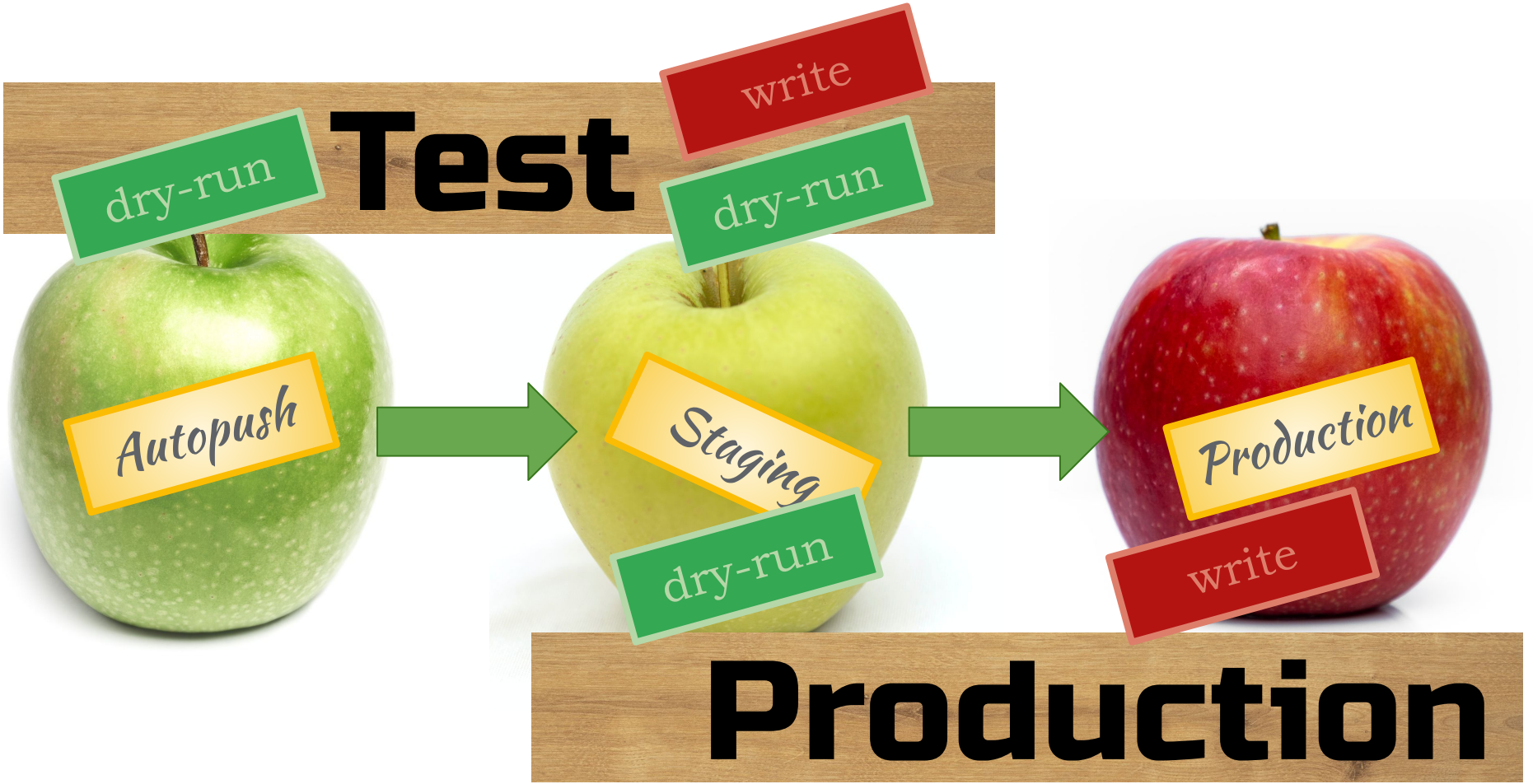

Test
dry-run
write
dry-run
Autopush
Staging
dry-run
Production
write
Production

Photo by Alexandr Dzyuba on Unsplash

False Positives

Photo by Andre Taissin on Unsplash

Two-phase mutation pattern

Photo by CHUTTERSNAP on Unsplash

Photo by Sıla Ergin on Unsplash

Start-up Tests

Photo by Andrew Johnson on Unsplash

Process Exit Code

Photo by Yaya The Creator on Unsplash

Photo by Jen Theodore on Unsplash

Data Validation

Photo by Vesela Vladimirova on Unsplash

A/B testing

Photo by James Yarema on Unsplash

Resource Overloading

Photo by Ashlee Brown on Unsplash

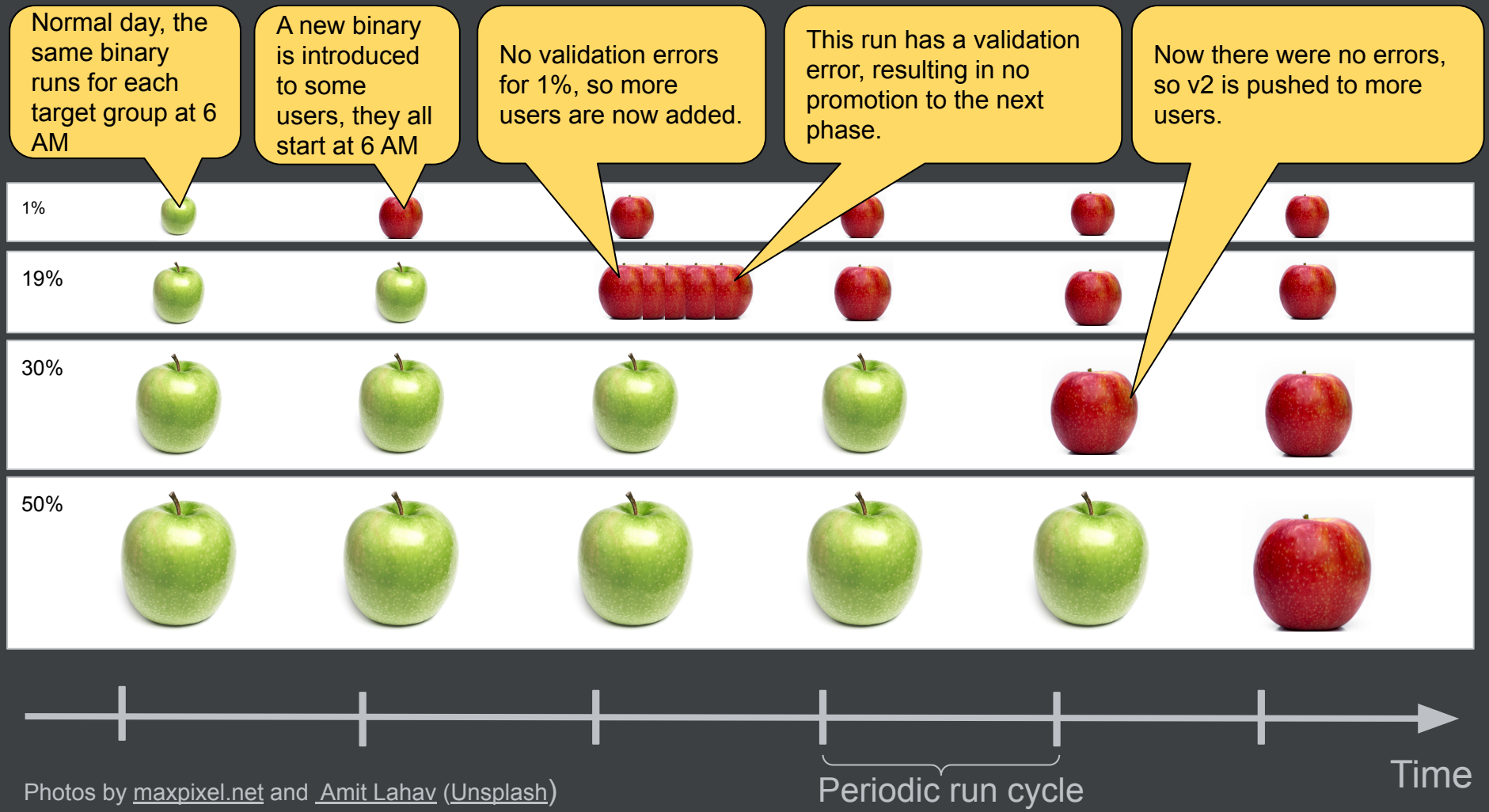
Photo by Timotheus Fröbel on Unsplash

Cascading

Photo by Andra Ion on Unsplash

Canarying

Photo by Ale Maciel on Unsplash



Photos by maxpixel.net and Amit Lahav (Unsplash)

splitting the work

Photo by Lewis Fagg on Unsplash

extending to target population

Targeting Library

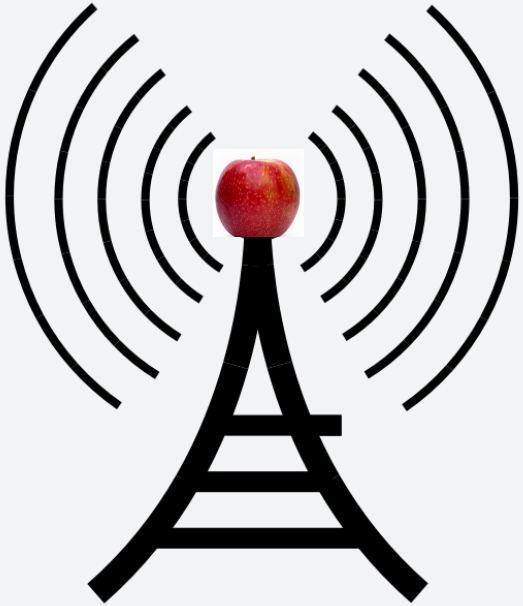
Photo by Alexander Schimmeck on Unsplash


keep it
fresh
Photo from rawpixel.net

long running jobs

Photo by Leonardo Yip on Unsplash

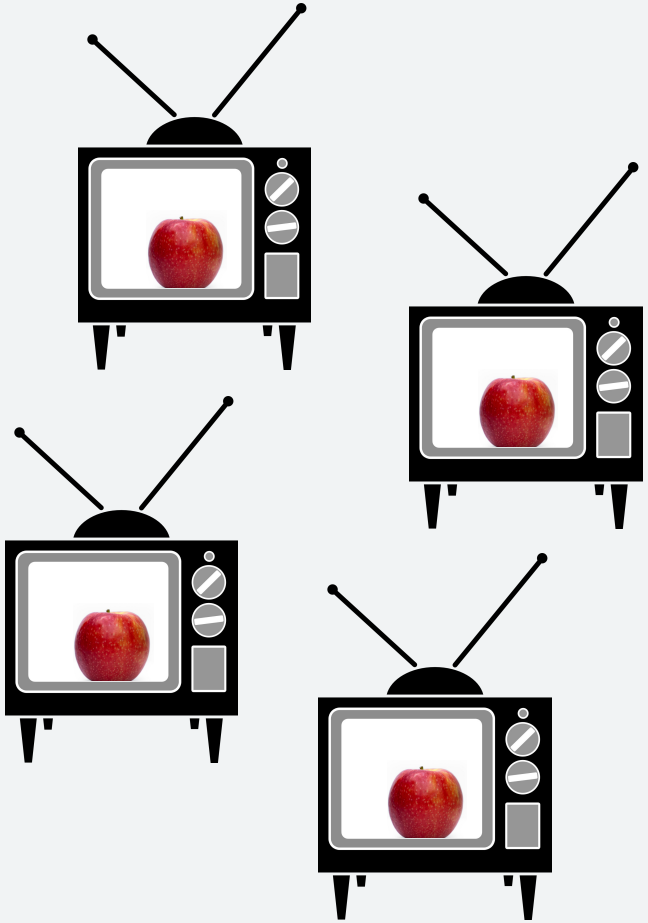
batch jobs



event-based processing

Crate photo by Jen Theodore on Unsplash
Radio Tower image from freesvg.org
Apple photo by Amit Lahav on Unsplash

event-based processing



Radio Tower image from freesvg.org
TV image from maxpixel.net
Apple photo by Amit Lahav on Unsplash

Julia Lee

Senior Software Engineer in Infrastructure at Slack

Leads development of asynchronous compute service

Photo by David Orlando Us De Paz on Unsplash

Data processing comes with a unique set of challenges and risks, but we can mitigate these with the best-practices we employ for servers.

Photo by JESUS ECA on Unsplash

Thanks!

Site Reliability Engineering

Special thanks to

Rita Sodt, *Julia Lee*, *Athena Vawda*, *Rich Feit*, *Alex Cebrian*, *Glen Sanford*, *John Lunney*, *Salim Virji*, *Jennifer Petoff*, *Todd Underwood*, *Yuval Greenfield*, *Joe Kearney* and *Jason Lee*

Images: unsplash.com and others (see speaker notes)