Data Engineering 101

Building Data Pipelines

galvanize

Jonathan Dinu
VP of Academic Excellence, Galvanize
jonathan@galvanize.com
@clearspandex

Currently

Zipfian Academy

+

galvanize

Questions? tweet @clearspandex

Outline

- What is Data Engineering?
- Why is Data Engineering?!
- How is Data Engineering!?!
- Data Architectures
- Building a Pipeline (w/ Luigi)
- Q&A

Data Science

Josh Wills
@josh_wills

Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

Reply Retweet Favorite More

Data Engineering

Josh Wills
@josh_wills

Engineer

Data ~~Scientist~~ (n.): Person who is better at statistics than any software engineer and better at software engineering than any ~~statistician.~~ Data Scientist

Reply Retweet Favorite More

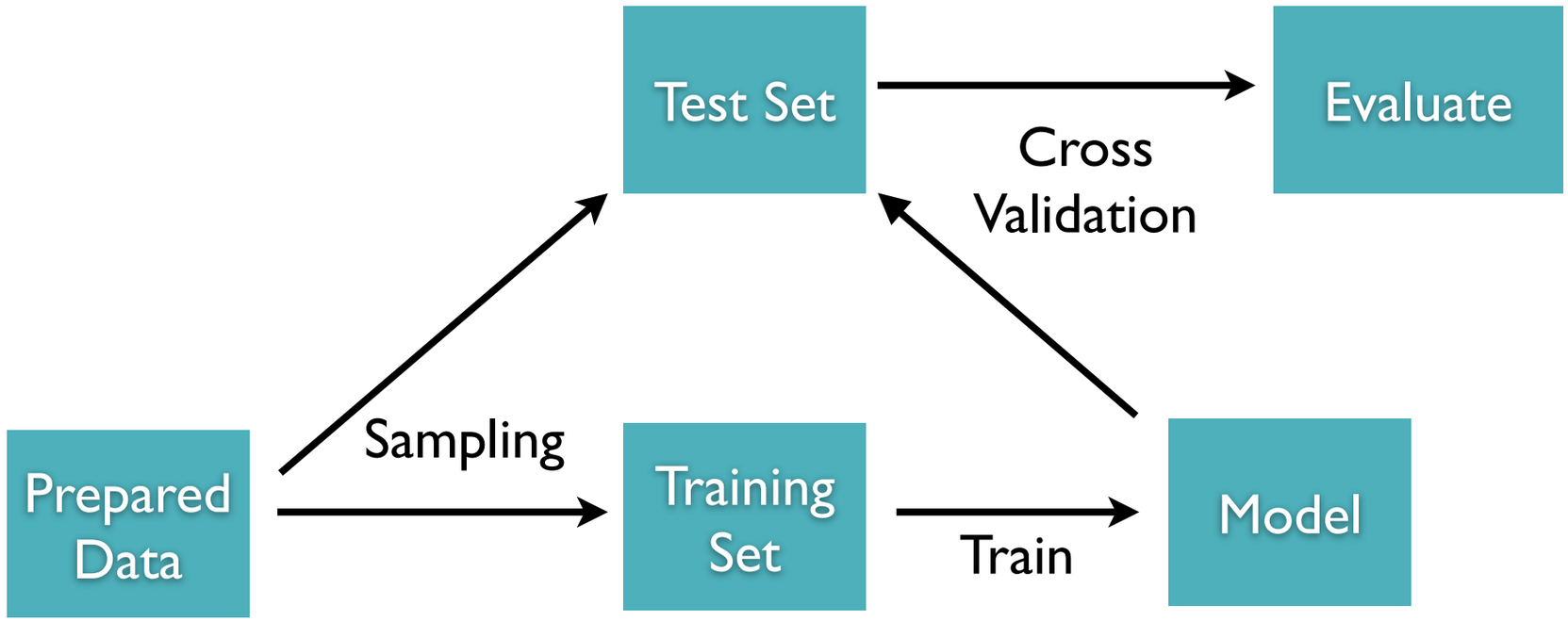
Follow

Questions? tweet @clearspandex

(In Reality)

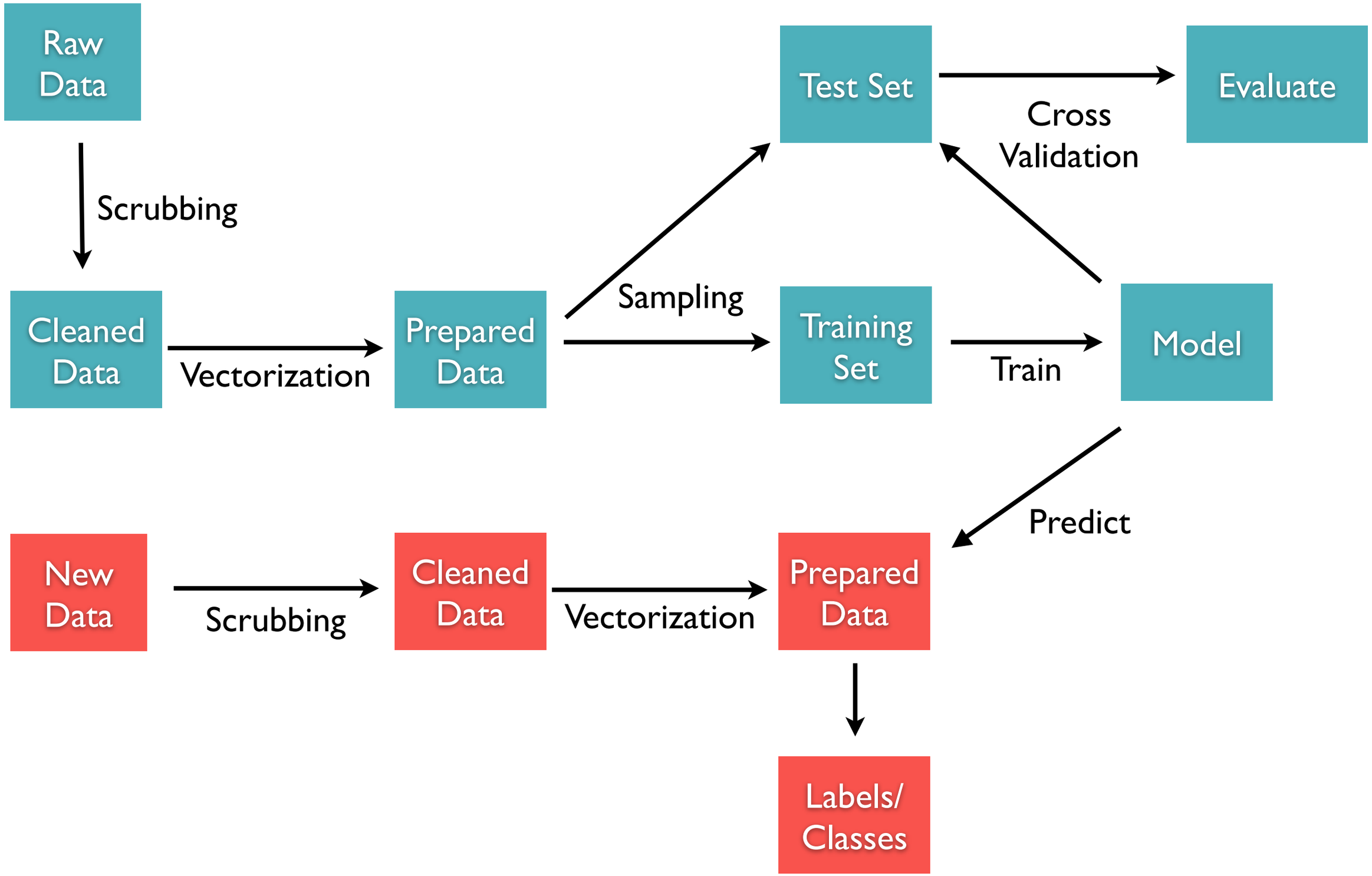
http://resource.mmgn.com/Gallery/full/SBS5NKQM.jpg

Data Science



Questions? tweet @clearspandex

Data Engineering



Questions? tweet @clearspandex

The Challenge

All the Things we Need to Do

At Scale

scrapy

Hadoop Streaming
(w/ BeautifulSoup4)

Snakebite (HDFS)

mrjob or luigi

Spark ML (pySpark)

Flask

Locally

requests

BeautifulSoup4

pymongo

pandas

scikit-learn/NLTK

Flask

Acquisition

Parse

Storage

Transform/Explore

Vectorization

Train

Model

Expose

Presentation

Questions? tweet @clearspandex

Code

Questions? tweet @clearspandex

Code

(it's the pipes!)

Code

Why Pipelines?

- Always keep raw data
- Data Lineage
- Apply a series of transforms to data.
- Flexible, Modular, Extensible (and testable!)

Code

Why Luigi

- Idempotence
- Checkpointing
- Native Hadoop Support
- But Works for arbitrary scripts (like make!)

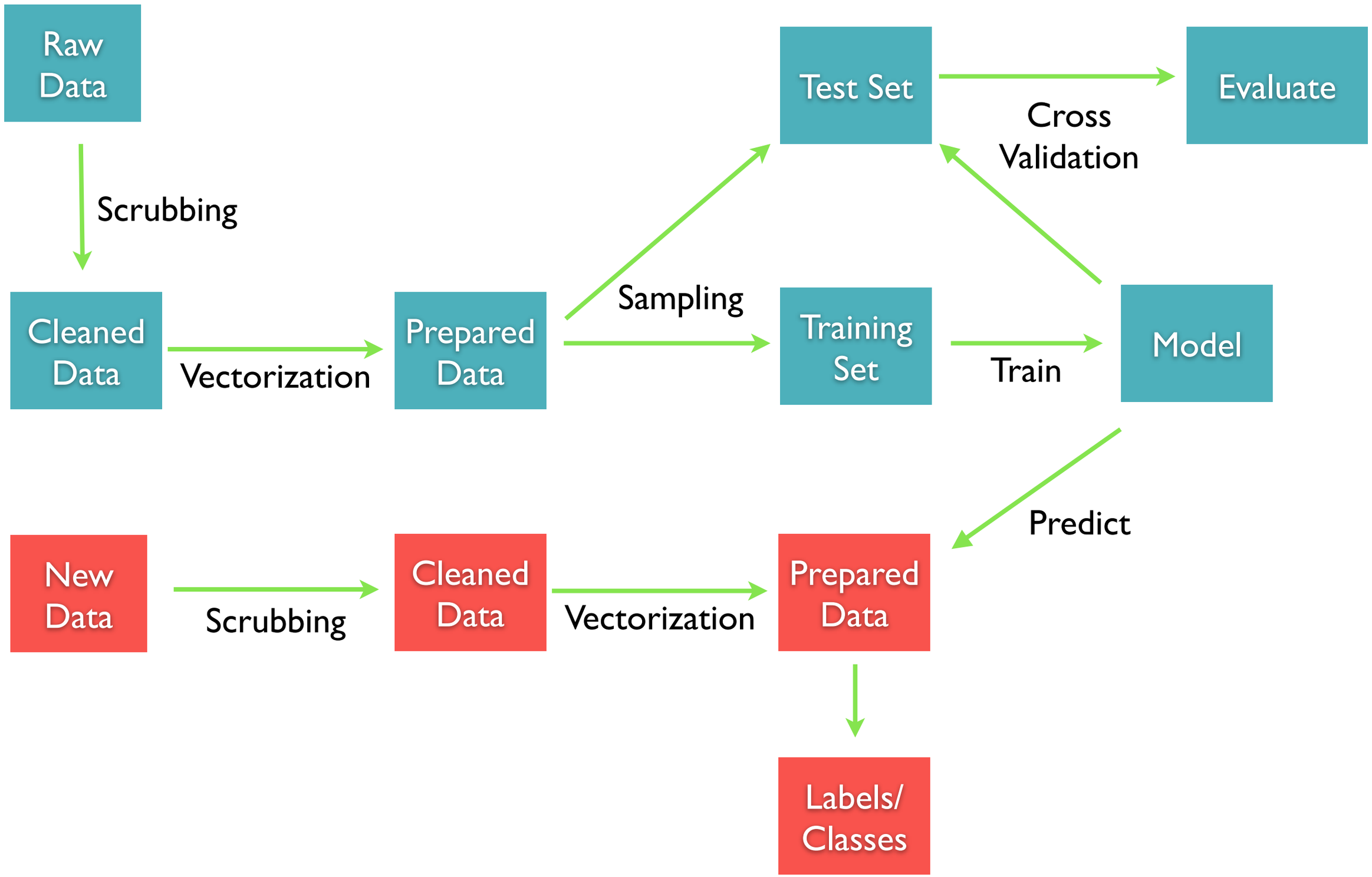
Code

Why Luigi

- Idempotence
- Checkpointing
- Native Hadoop Support
- But Works for arbitrary scripts (like make!)

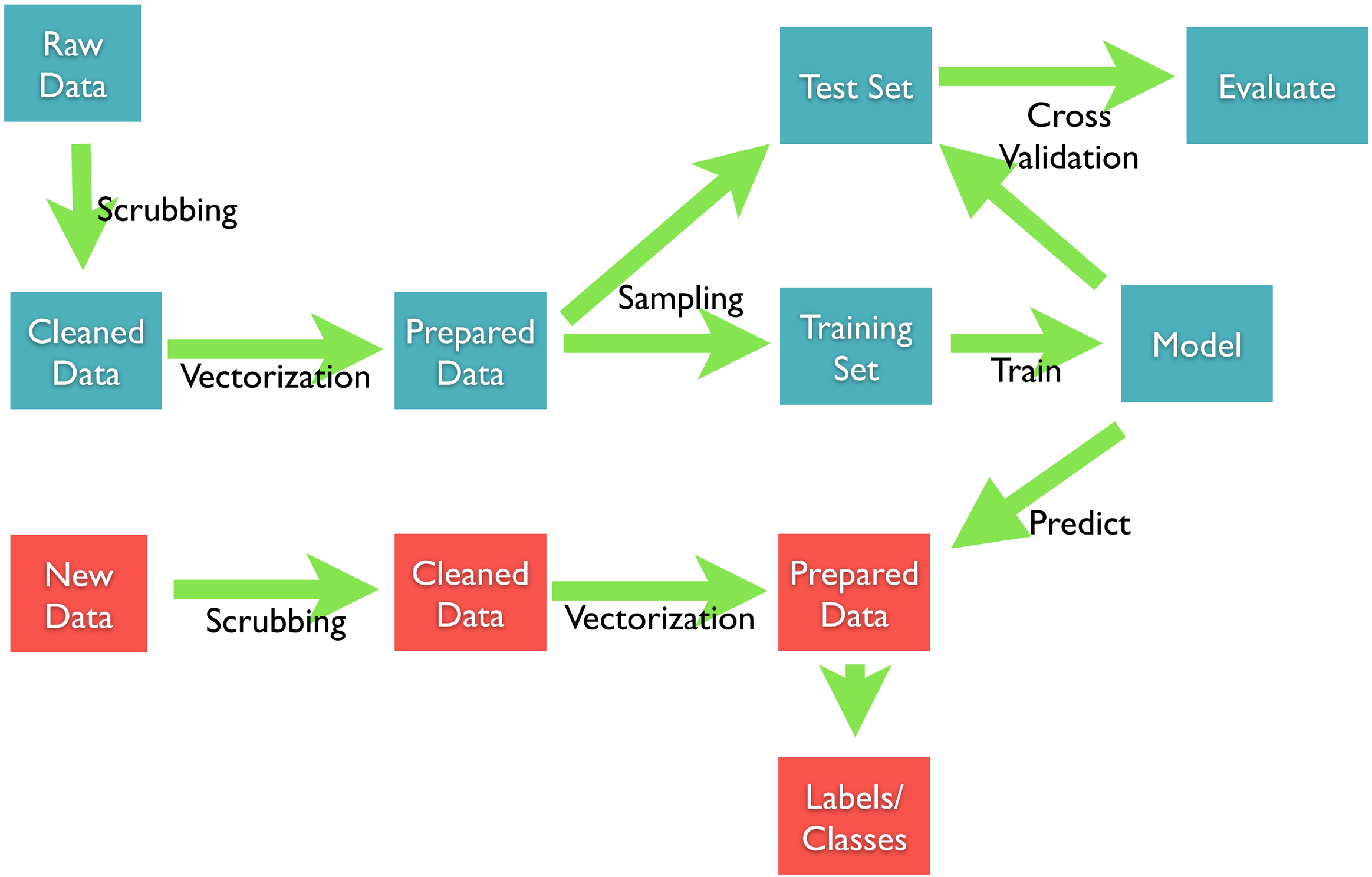
(also has a nice UI and sends emails :)

Data Engineering



Questions? tweet @clearspandex

Data Engineering



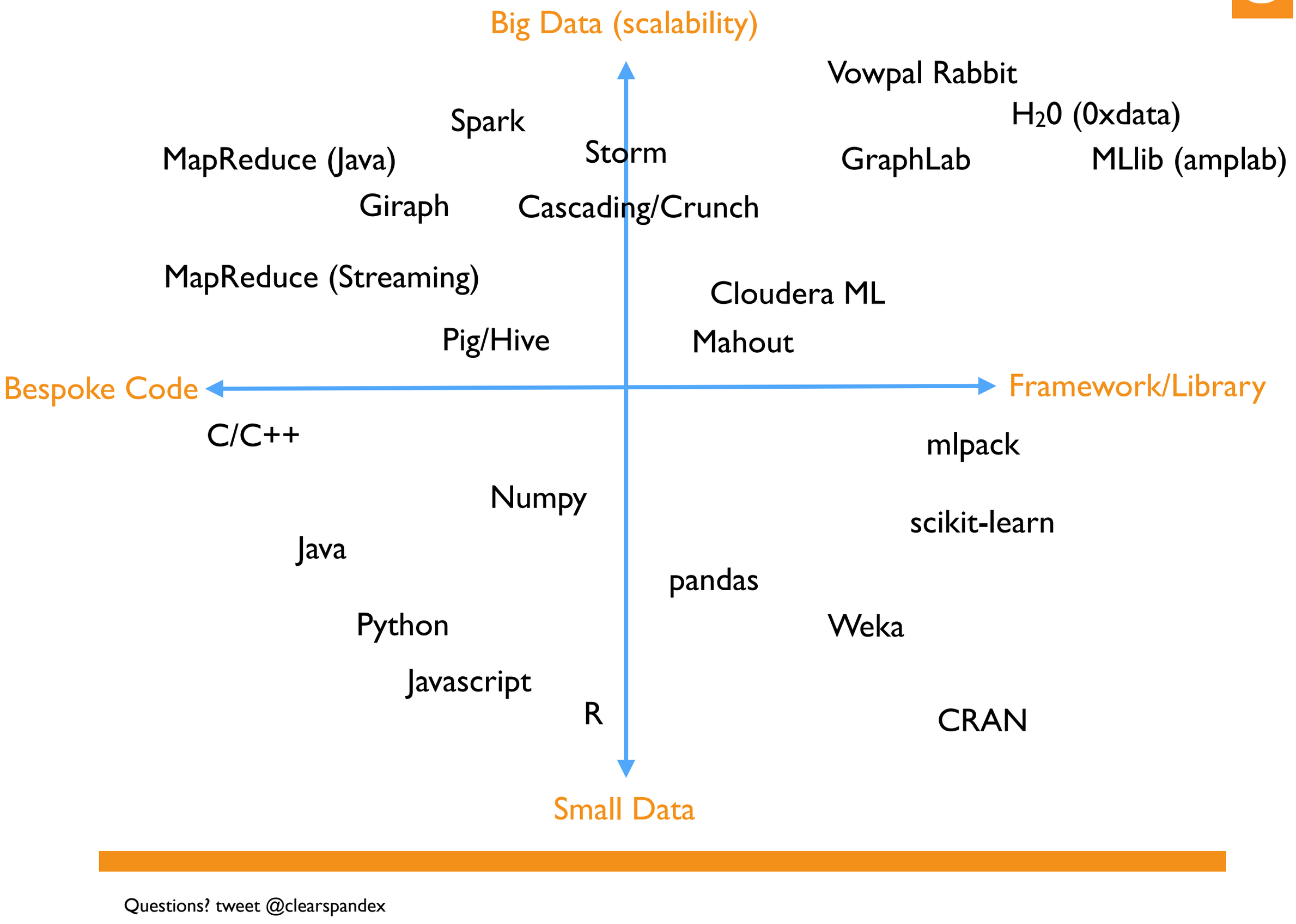
Questions? tweet @clearspandex

Code

LIVE CODE

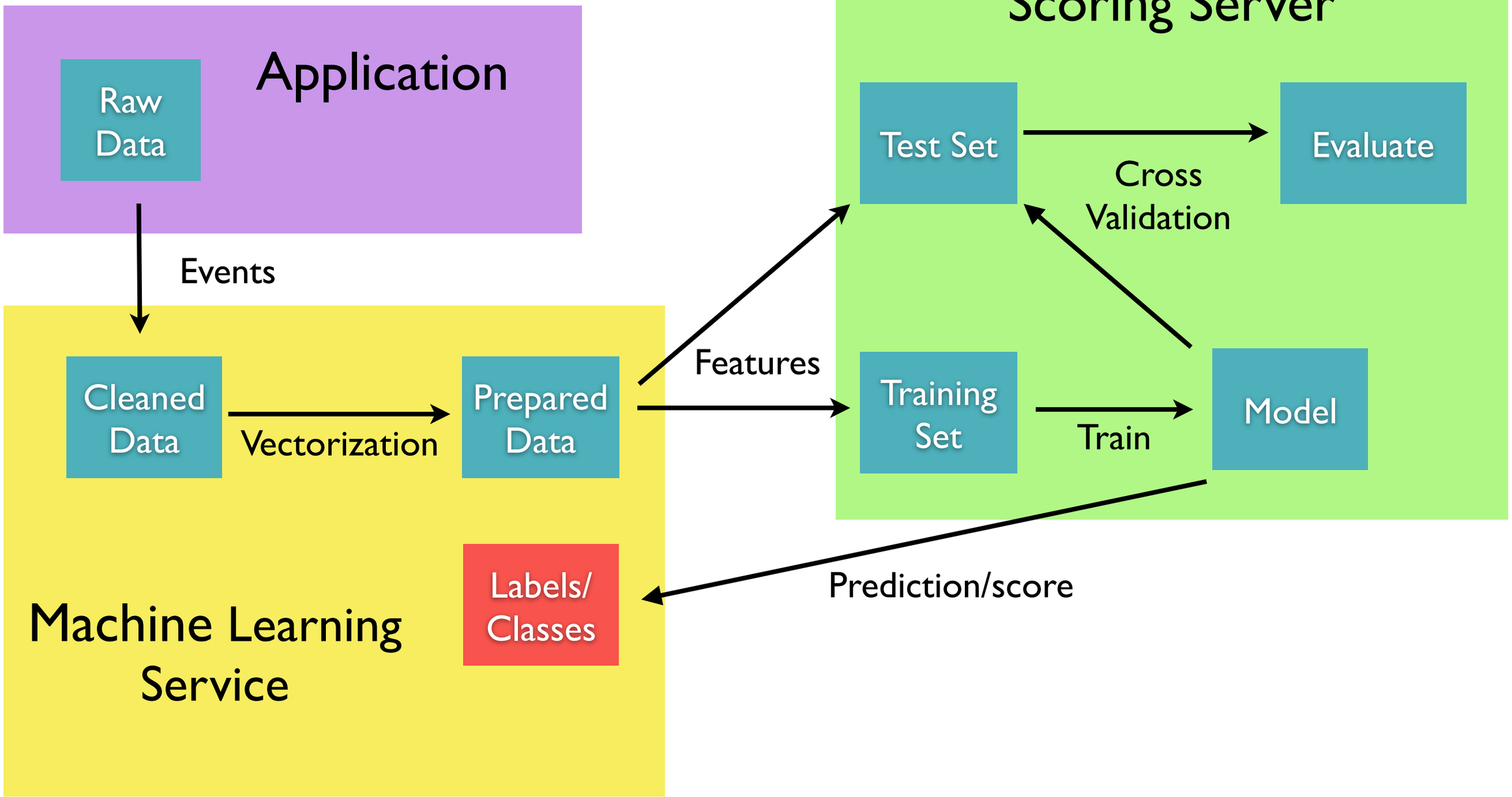
Questions? tweet @clearspandex

All the Things we Have to Do it!

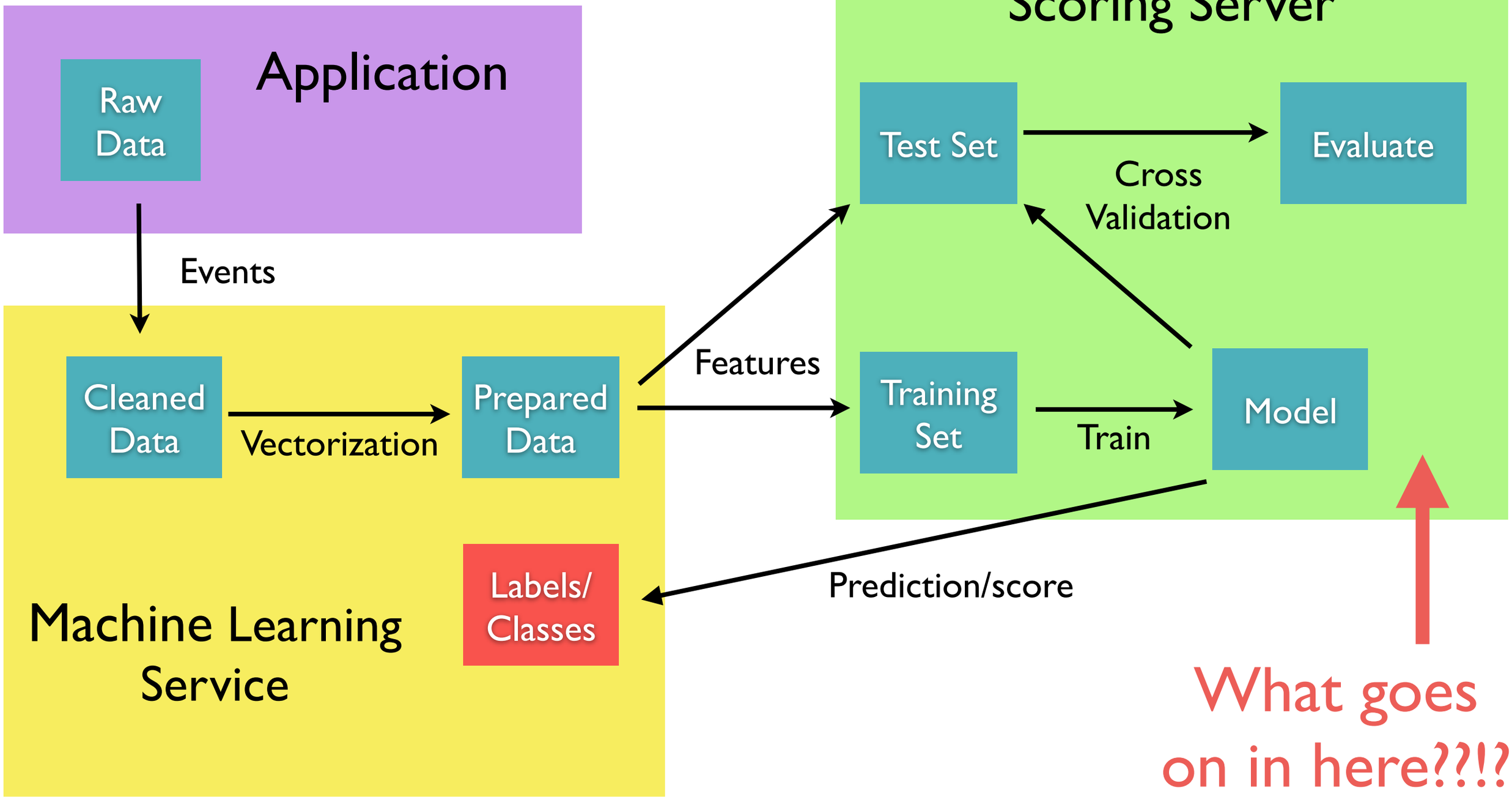


Questions? tweet @clearspandex

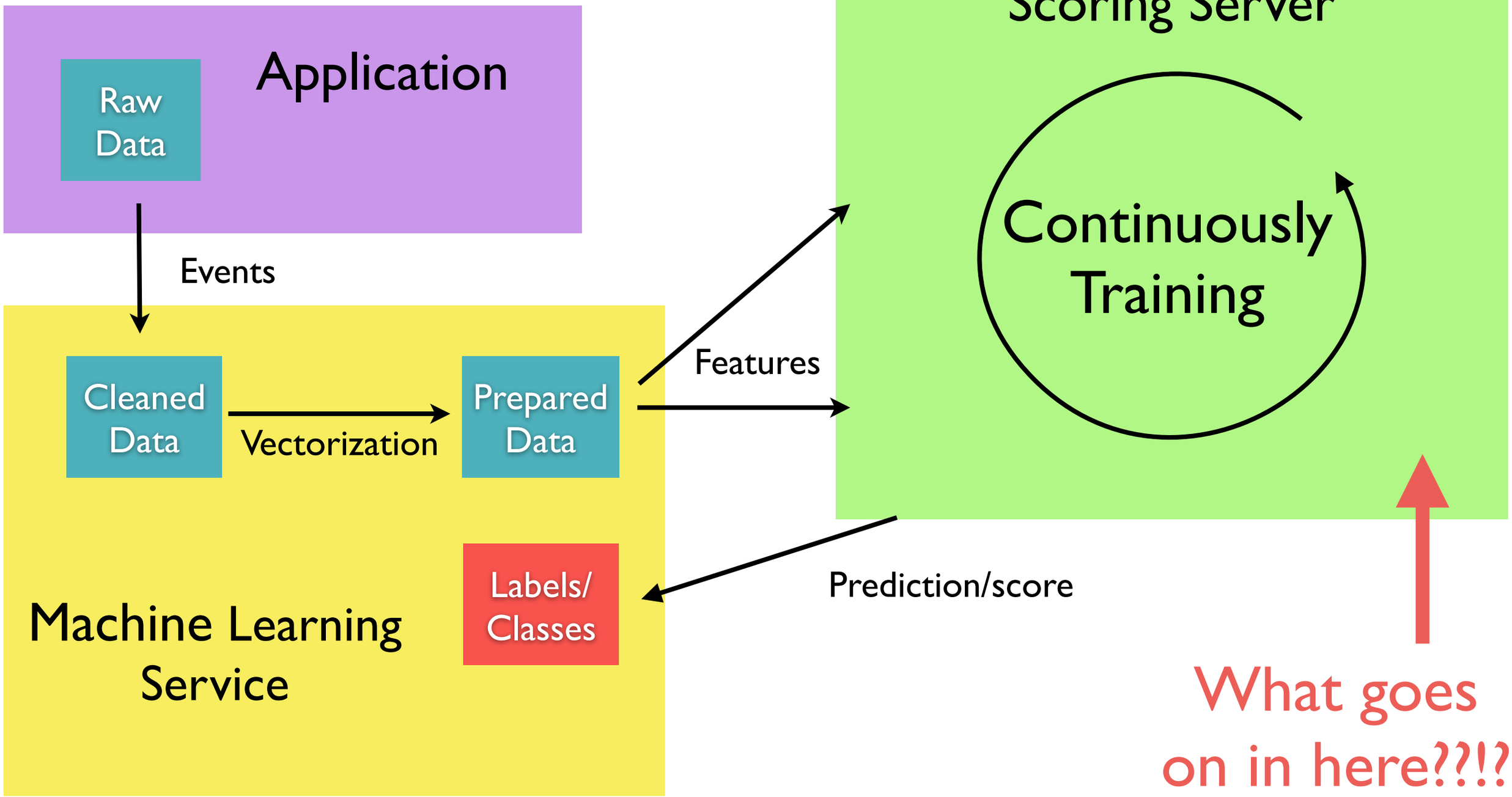
Machine Learning



Machine Learning



Machine Learning



Overview

Modern Architectures

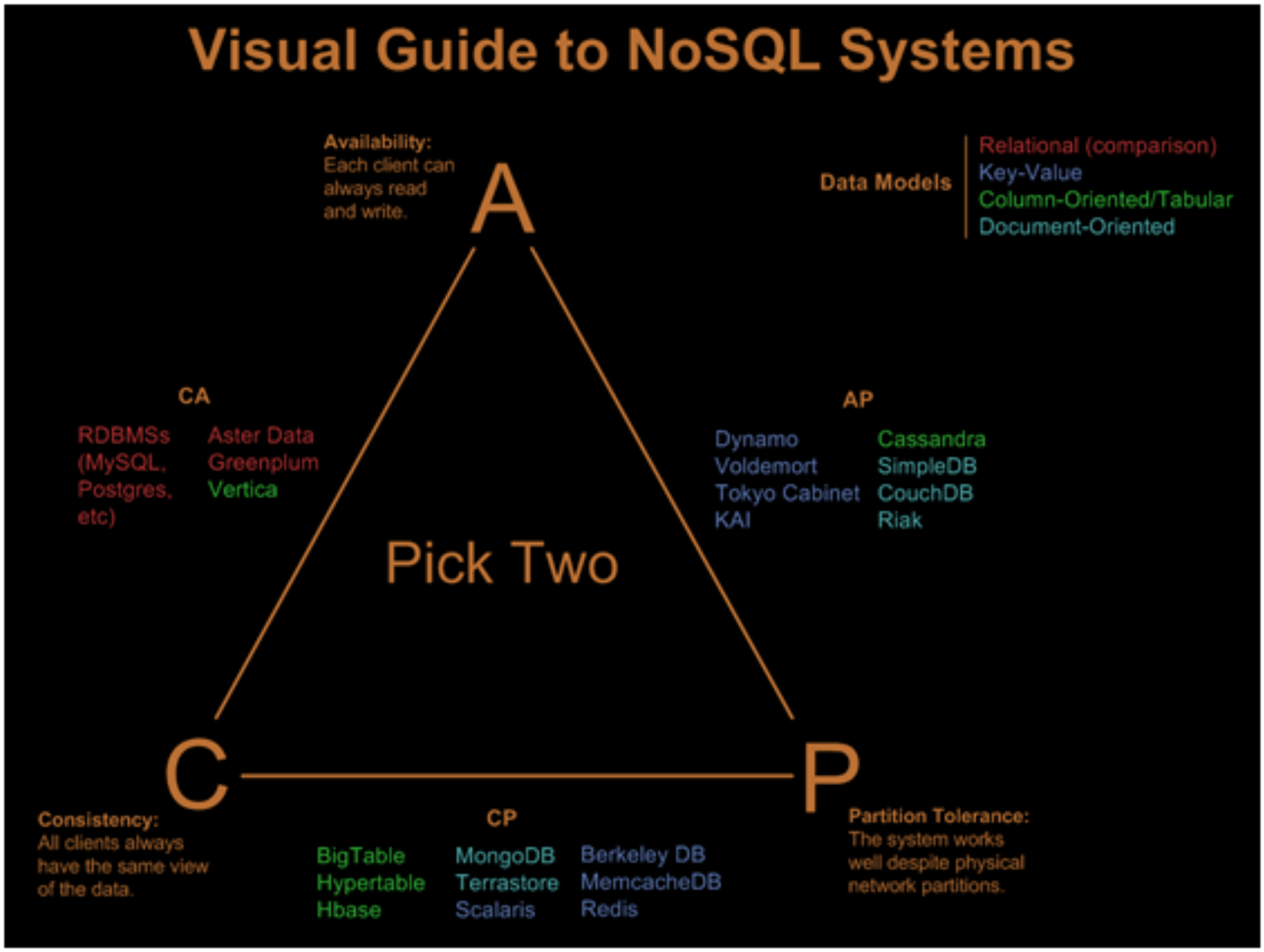
Questions? tweet @clearspandex

Architecture

Questions? tweet @clearspandex

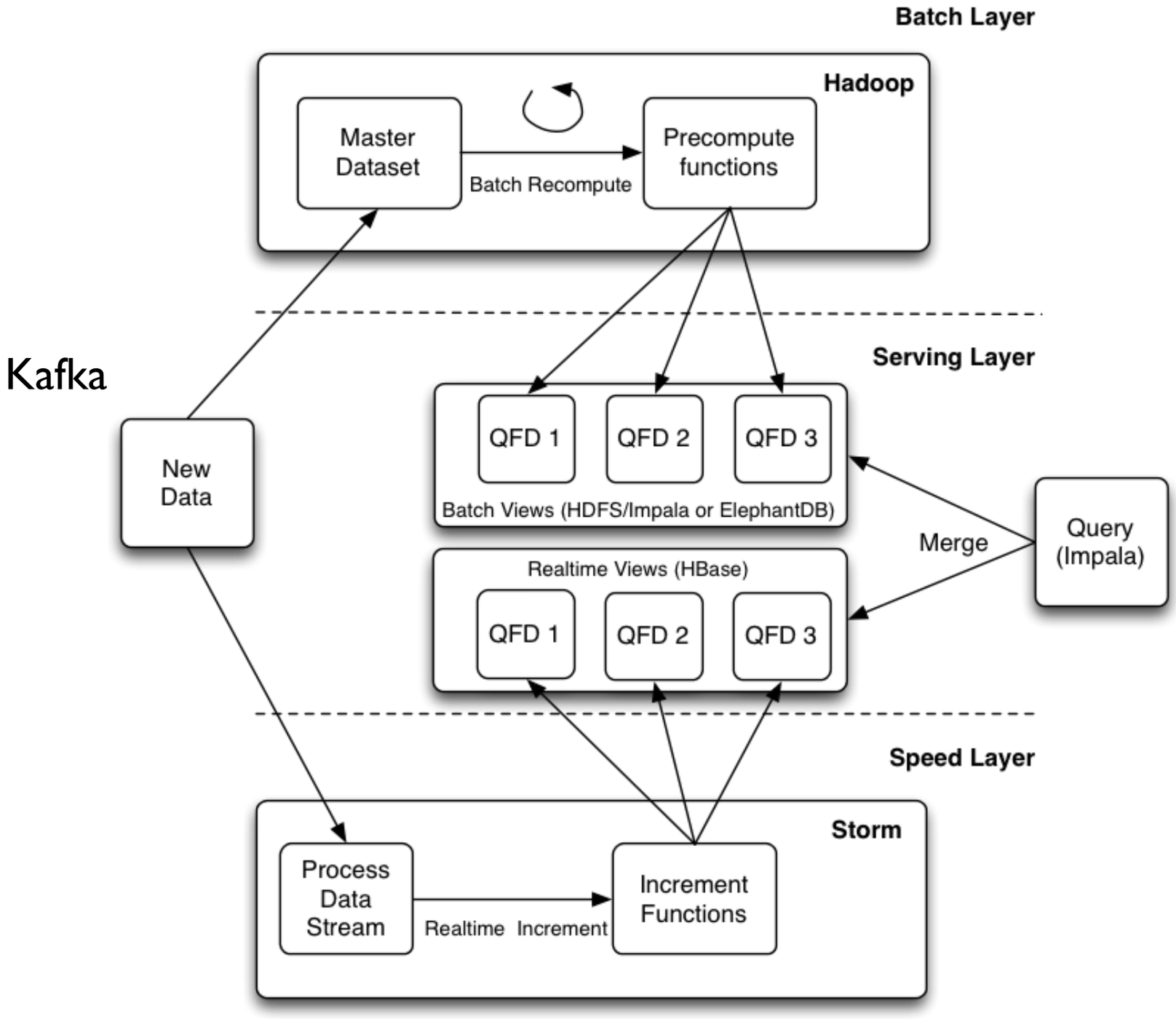
Source: http://www.fwdlife.in/bizarre-buildings-around-the-globe/

Theories



Questions? tweet @clearspandex

Architecture

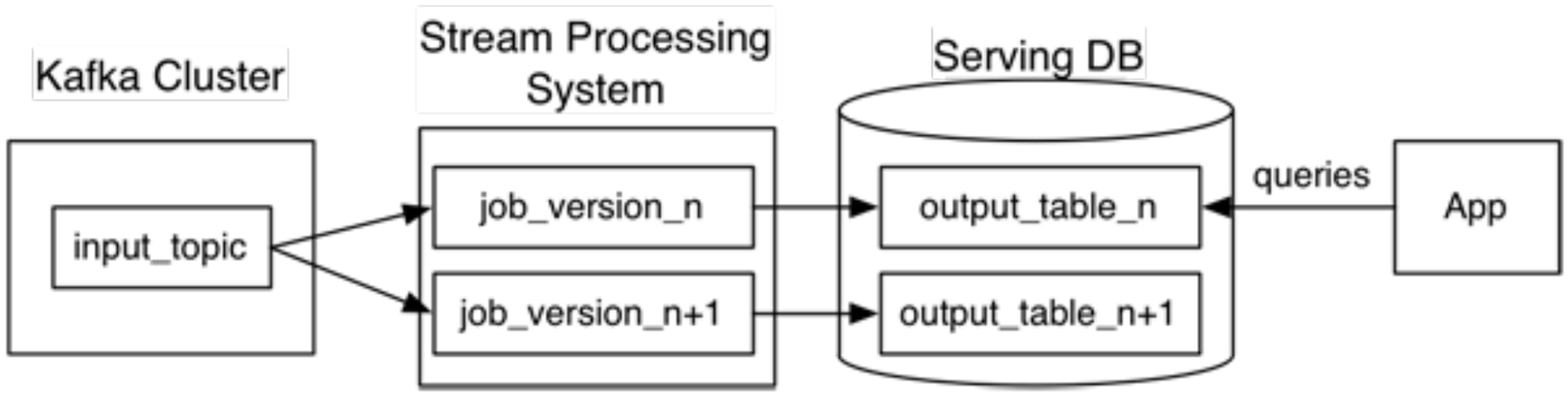


LAMBDA ARCHITECTURE

Questions? tweet @clearspandex

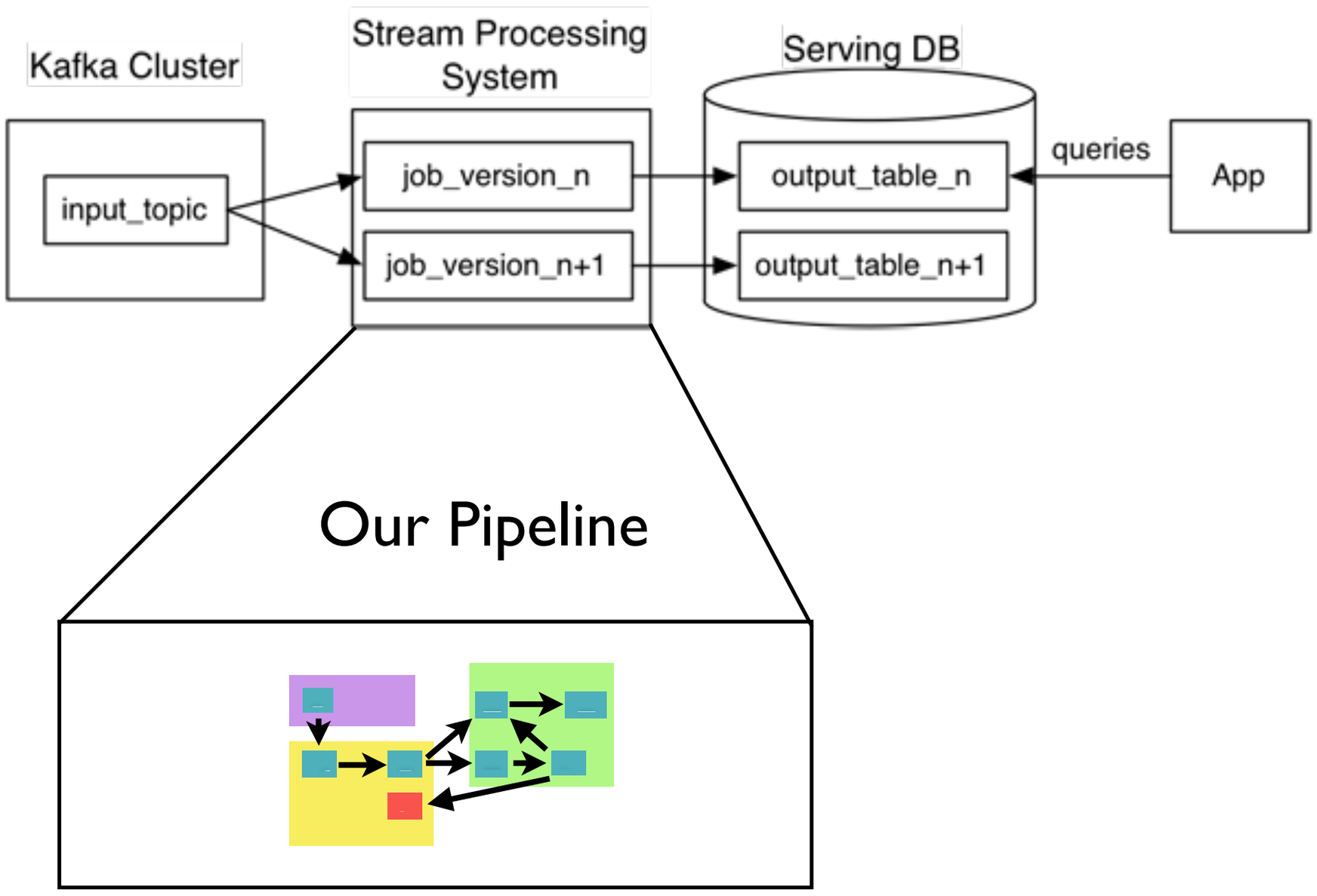
Source: http://slides.com/gchomatas/delivering-a-big-data-ready-minimum-viable-product

Architecture



Source: http://radar.oreilly.com/2014/07/questioning-the-lambda-architecture.html

Architecture



Questions? tweet @clearspandex

Source: http://radar.oreilly.com/2014/07/questioning-the-lambda-architecture.html

Architecture

> So, why the excitement about the Lambda Architecture? I think the reason is because people increasingly need to build complex, low-latency processing systems. What they have at their disposal are two things that don't quite solve their problem: a scalable high-latency batch system that can process historical data and a low-latency stream processing system that can't reprocess results. By duct taping these two things together, they can actually build a working solution.
>
> - Jay Kreps

Source: http://radar.oreilly.com/2014/07/questioning-the-lambda-architecture.html

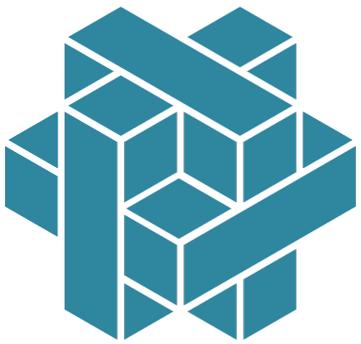
METRICS

Coda Hale

@coda

github.com/codahale

METRICS EVERYWHERE



Zipfian Academy

+

galvanize

Data Science Immersive

Masters in Data Science

Data Engineering Immersive

Weekend Workshops

Questions? tweet @clearspandex

Goals

We're Hiring!

- Full-time Instructors
- TAs
- Mentor (volunteer)

Questions? tweet @clearspandex

Questions?

Thank You!

Jonathan Dinu
VP of Academic Excellence, Galvanize
jonathan@galvanize.com
@clearspandex