Data Oriented Architectures for Deploying Machine Learning

Andrei Paleyes

Data Science and Artificial Intelligence Research Day

DeKUT, June 2021

About myself

- A software engineer for a decade...
- ... including few years deploying ML in Amazon
- Now PhD student with Neil Lawrence, ML@CL group, University of Cambridge
  - People think we do ML, but we really just do software systems research

Deploying machine learning is hard.

Why do 87% of data science projects never make it into production?

VB Staff July 19, 2019 4:10 AM

Why is it hard?

There many practical reasons.

Why is it hard?

There many practical reasons.

One reason is that data management is a mess.

Why is it hard?

There many practical reasons.

One reason is that data management is a mess.

Modern software systems are API oriented, not data oriented!

What is a service?

A service is

What is a service?

A service is
a piece of software,

What is a service?

A service is
a piece of software,
that provides a function, or many functions, known as "interface"

What is a service?

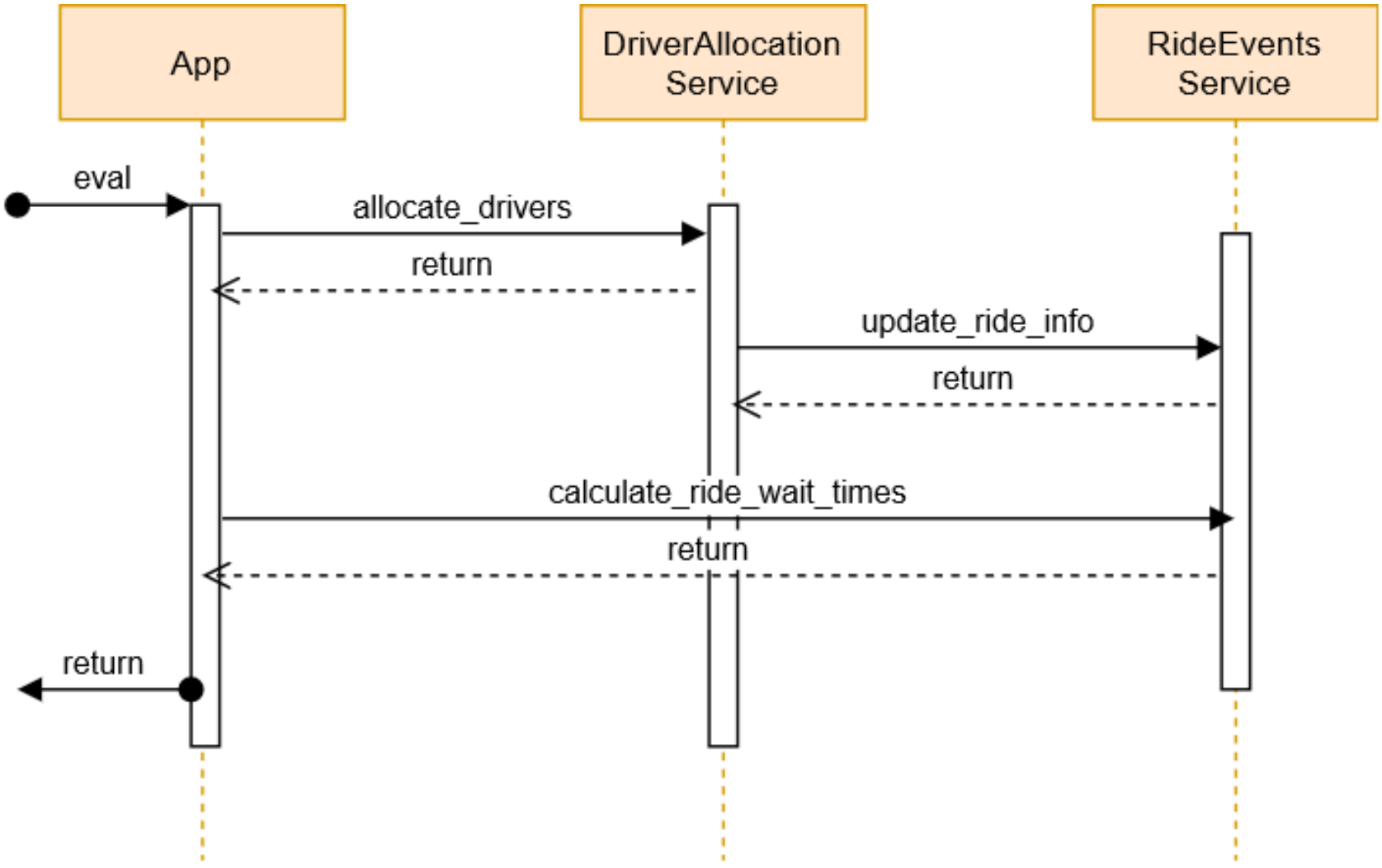
A service is
a piece of software,
that provides a function, or many functions, known as "interface"
that clients* can reuse,

What is a service?

A service is
a piece of software,
that provides a function, or many functions, known as "interface"
that clients* can reuse,
together with policies to control its usage.

*A client can be anything: another software, a person, a hardware.

Service Oriented Architectures



Question

Can we embed proper data care into the software system design?

Desiderata

Explicit data dependencies

Desiderata

Explicit data dependencies

No hidden state

Desiderata

Explicit data dependencies

No hidden state

Trace data from outputs to inputs

Desiderata

Explicit data dependencies

No hidden state

Trace data from outputs to inputs

Ability to reproduce and experiment

What we explore

Data Oriented Architectures (DOA)

Builds upon flow-based programming, dataflow, data streaming.

What we explore

Data Oriented Architectures (DOA)

Builds upon flow-based programming, dataflow, data streaming.

All interactions in the system happen via data interfaces instead of APIs.

What we explore

Data Oriented Architectures (DOA)

Builds upon flow-based programming, dataflow, data streaming.

All interactions in the system happen via data interfaces instead of APIs.

Stateless processing nodes connected by data streams.

What we explore

Data Oriented Architectures (DOA)

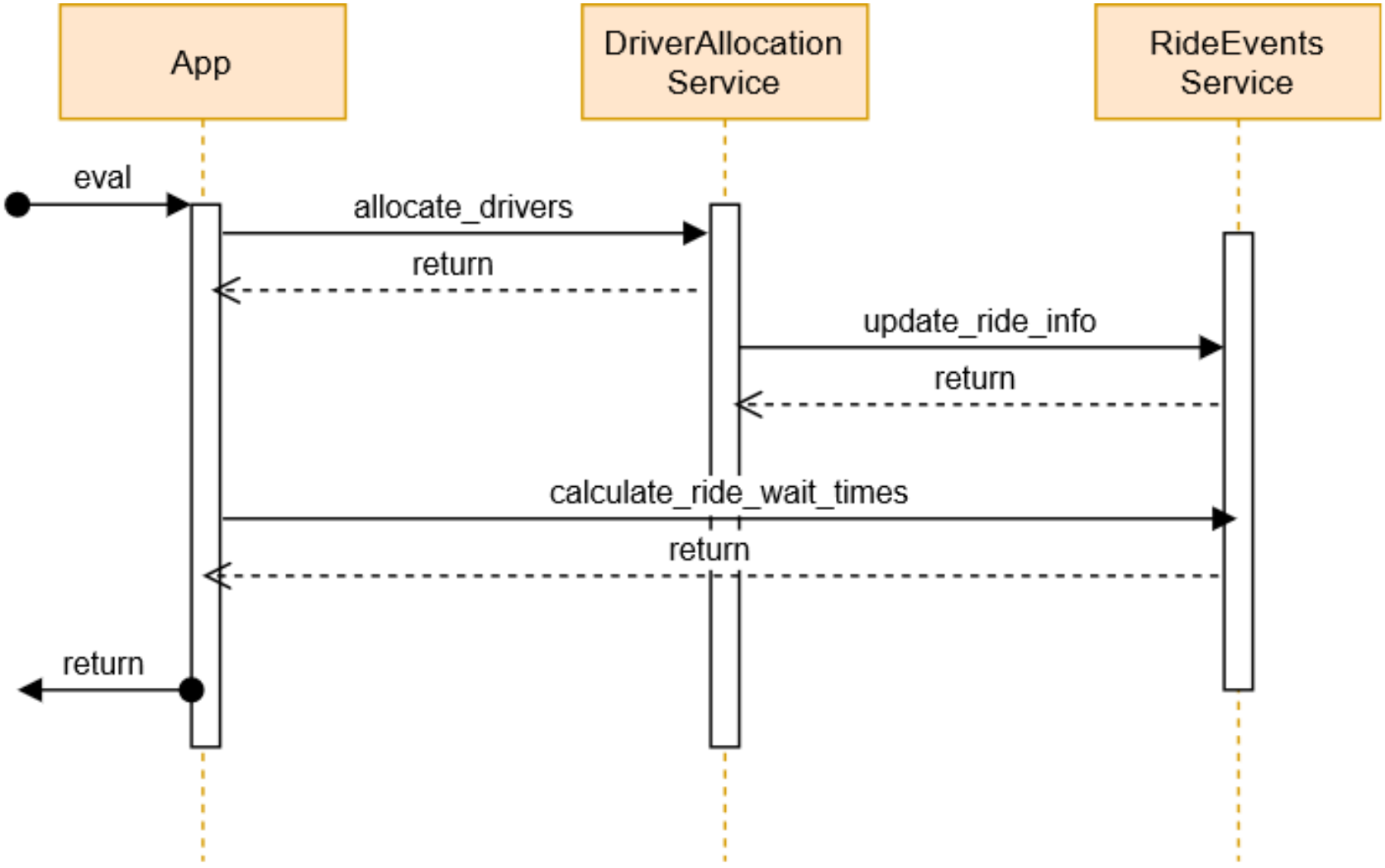
Builds upon flow-based programming, dataflow, data streaming.

All interactions in the system happen via data interfaces instead of APIs.

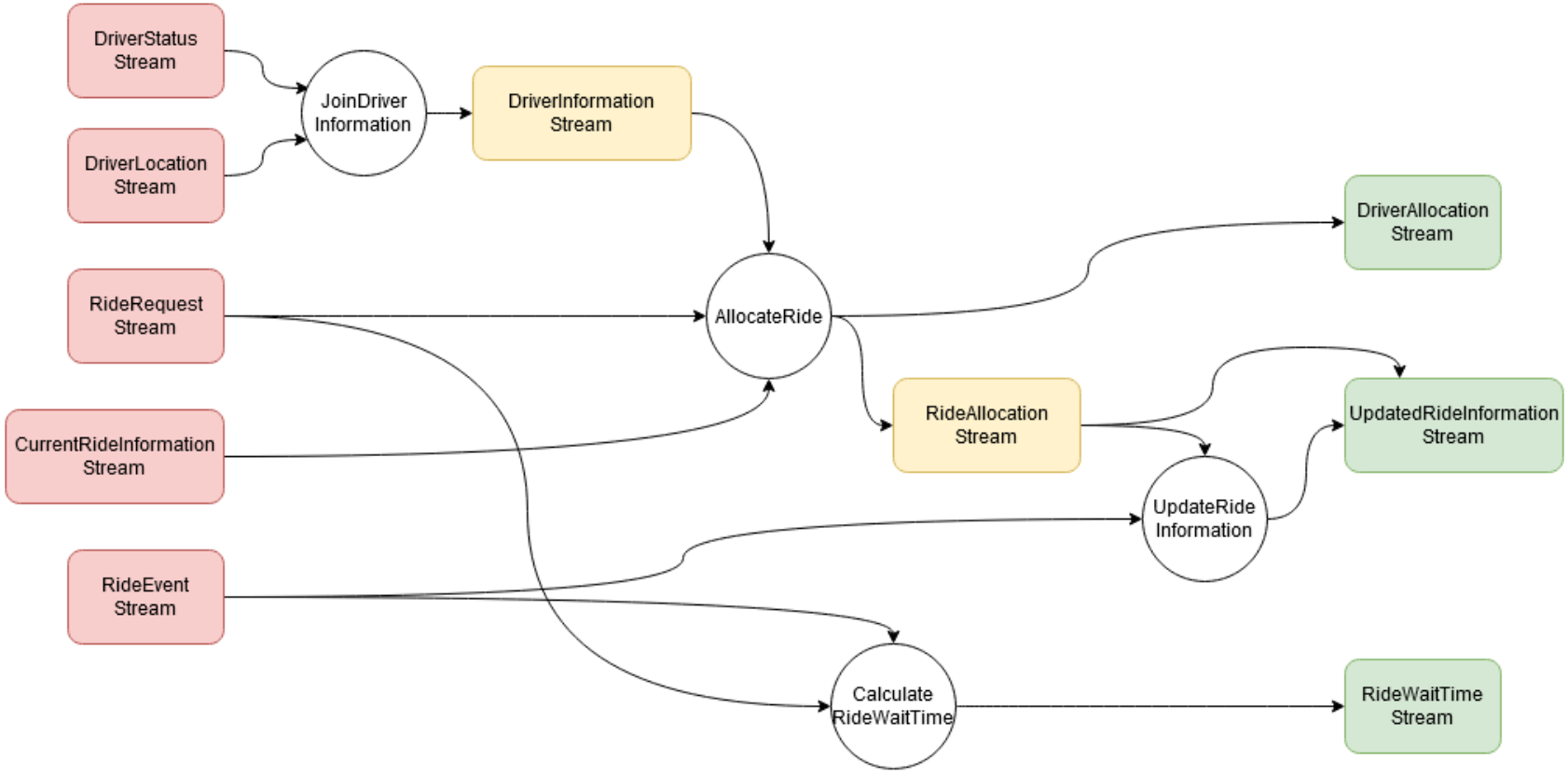
Stateless processing nodes connected by data streams.

Discovery, traceability, experimentation come by design.

Example: data flow graph



Example: data flow graph



Related work

Paradigms

- Actor model
- Dataflow
- Flow-based programming
- Data meshes

Related work

Data processing

- Apache Spark
- Google Dataflow
- Node-RED

Stream processing

- Apache Kafka
- AWS Kinesis

Related work

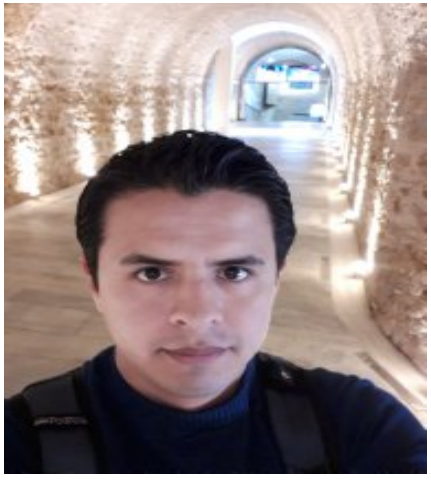
Lot more

- Akka
- Zio
- Ray
- Luigi
- Apache NiFi
- Apache Flink
- Faust

What we are up to

- Survey on challenges in deploying ML
- Comparison of FBP and SOA in ML deployment context
- Investigation of ML systems fairness
- Machine learning on data streams
- Real time reinforcement learning

ML@CL



Christian Cabrera

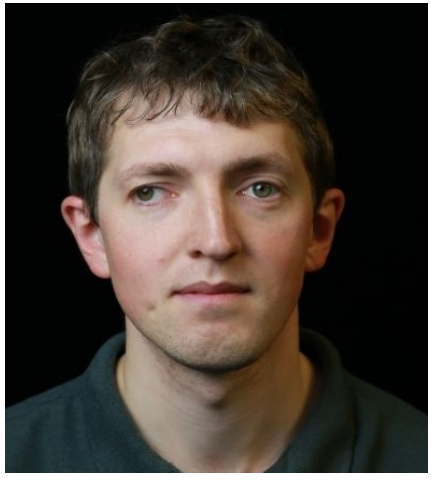
Eric Meissner

Neil Lawrence

Jess Montgomery

Pierre Thodoroff

Markus Kaiser



Andrei Paleyes

Carl Henrik Ek

Summary

- Deploying ML is hard
- Can we embed better data care in our software designs?
- Proposal: Data Oriented Architectures
- Building on data streaming, flow-based programming

https://paleyes.info
https://mlatcl.github.io/
Get in touch: ap2169@cam.ac.uk