Swiss Transport in Real Time: Tribulations in the Big Data Stack

Alexandre Masselot
Soft-shake, Geneva
October 2016

Swiss Transport in Real Time: Tribulations in the Big Data Stack

This is only a POC!!!

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, store, transform

and **visualize "near real time"** data

and achieve ***a posteriori*** **analysis**?

Finding a dataset

- social media
- finance
- sport
- energy
- transport
- log analysis
- meteorology
- bioinformatics
- personalized health
- monitoring
- security
- IOT

Finding a dataset

- social media
- finance
- sport
- energy
- **transport**
- log analysis
- meteorology
- bioinformatics
- personalized health
- monitoring
- security
- IOT

www.voev.ch

www.voev.ch

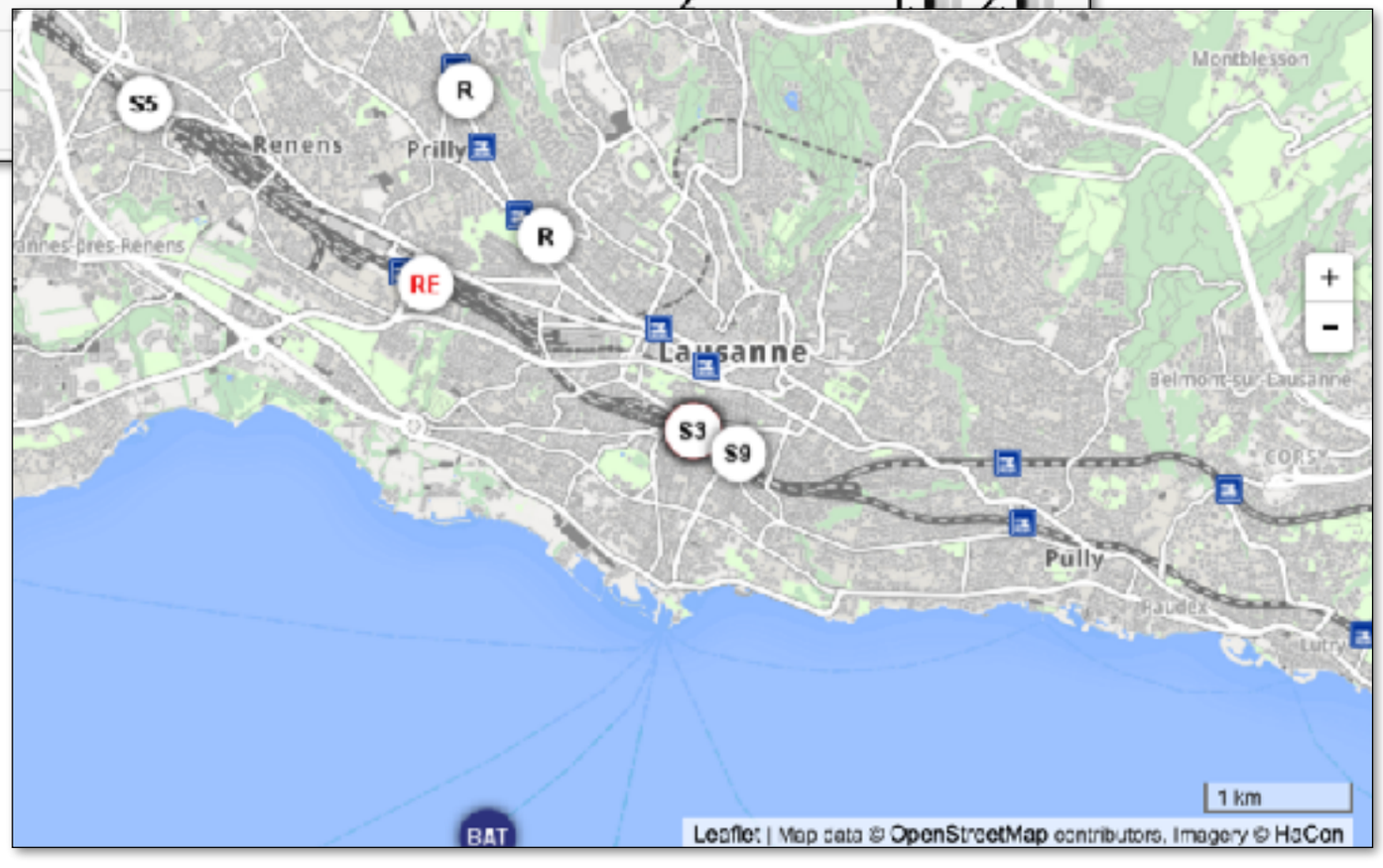
www.voev.ch

AAGL	Autobus AG Liestal
AAGR	Auto AG Rothenburg
AAGS	Auto AG Schwyz
AAGU	AUTO AG URI
AB	Appenzeller Bahnen AG
ABl	Autolinee Bleniesi SA
ABF	Autobusbetrieb Freienbach
AFA	Automobilverkehr Frutigen Adelboden AG
AMSA	Autolinea Mendrisiense SA
AOT	Autokurse Oberthurgau AG
ARAG	Rottal Auto AG
ARBAG	Aletsch Riederalp Bahnen AG
ARL	Autolinee Regionali Luganesi
AS	Autobetrieb Sernftal AG
ASGS	Autotransports Sion-Grône-Sierre
ASm	Aare Seeland mobil AG
AVG	Autoverkehr Grindelwald AG
AVJ	Autotransports de la Vallée de Joux
AWA	Autobetrieb Weesen-Amden
AZZK	Autobus Zürich-Zollikon-Küsnacht
BB	Bürgenstock Bahnen
BBA	Busbetrieb Aarau AAR bus+bahn
BBBW	Bus-Betrieb Binggeli
BDWM	BDWM Transport AG
BGU	BGU Busbetrieb Grenchen und Umgebung AG
BLAG	Busland AG
BLM	Bergbahn Lauterbrunnen-Mürren AG
BLS	BLS AG
BLT	BLT Baselland Transport AG
BLWE	Busbetrieb Lichtensteig-Wattwil-Ebnat-Kappel
BOB	Berner Oberland-Bahnen AG
BOGG	Busbetrieb Olten Gösgen Gäu AG
BOS	BUS Ostschweiz AG
BOS-M	BOS Management AG
BRB	Brienz Rothorn Bahn AG
BRER	Busbetrieb Rapperswil-Eschenbach-Rüti
BRSB	Braunwald-Standseilbahn AG
BSU	Busbetrieb Solothurn und Umgebung AG
BVB	Basler Verkehrs-Betriebe
CGN	CGN SA
CJ	Compagnie des chemins de fer du Jura (C.J.) SA
CROS	Crossrail AG
DBSCH	DB Schenker Rail Schweiz GmbH
DBZ	Dolderbahn Zürich
ETB	Emmentalbahn, Huttwil
FART	Ferrovie Autolinee Regionali Ticinesi
FB	Forchbahn AG
FC	FUNICAR Kursbetriebe AG
FLP	Ferrovie Luganesi SA
FW	Frauenfeld-Wil-Bahn AG
GGB	Gornergrat Bahn AG HBSAG Hafenbahn Schweiz AG
JB	Jungfraubahn AG
LEB	Chemin de fer Lausanne-Echallens-Bercher
LLB	AG für Verkehrsbetriebe Leuk-Leukerbad und Umgebung
LSMS	Schilthornbahn AG
MBC	Transports de la région Morges-Bière-Cossonay SA
MG	Ferrovia Monte Generoso SA
MGB	Matterhorn Gotthard Bahn
MIB	Kraftwerke Oberhasli AG Meiringen-Innertkirchen-Bahn
MOB	Chemin de fer Montreux-Oberland Bernois
MVR	Transports Montreux-Vevey-Riviera SA
NHB	Niederhornbahn
NB	Niesenbahn AG
NStCM	Chemin de fer Nyon-St. Cergue-Morez
OeBB	Oensingen-Balsthal-Bahn
PAG	PostAuto Schweiz AG
PB	PILATUS-BAHNEN AG
RA	RegionAlps SA
RAILG	Railgate AG
RB	RIGI BAHNEN AG
RBL	Regionalbus Lenzburg AG
RBS	Regionalverkehr Bern-Solothurn AG
REGO	Regiobus Gossau AG
RhB	Rhätische Bahn AG
RNCH	DB Schenker Rail Schweiz GmbH
RLC	railCare
RVBW	Regionale Verkehrsbetriebe Baden-Wettingen AG
RVSH	SchaffhausenBus, Regionale Verkehrsbetriebe SH AG
SBB	SBB AG
SBB-D	SBB GmbH
SBC	Stadtbus Chur AG
SBF	Stadtbus Frauenfeld
SBW	Stadtbus Winterthur
SMC	Cie de Chemin de Fer+d'Autobus Sierre-Montana-Crans (SMC) SA
SMGN	Société des Mouettes Genevoises Navigation SA
SMtS	Funiculaire St-Imier - Mont-Soleil SA
SOB	Schweizerische Südostbahn AG
SRTAG	Swiss Rail Traffic AG
SSIF	Società Subalpina di Imprese Ferroviarie S.p.A.
ST	Sursee-Triengen-Bahn
STB	Sensetalbahn AG
STI	Verkehrsbetriebe STI AG
SVB	BERNMOBIL Städt. Verkehrsbetriebe Bern
SWAG	Seilbahn Weissenstein AG
SZU	Sihltal Zürich Uetliberg Bahn SZU AG
THURBO	Thurbo AG
TL	Transports publics de la région lausannoise SA
TMR	TRANSPORTS DE MARTIGNY ET REGIONS SA
TPC	Transports Publics du Chablais SA
TPF	Transports publics fribourgeois SA
TPG	Transports publics genevois
TPL	Trasporti Pubblici Luganesi SA
TPN	Transports Publics de la Région Nyonnaise SA
TRN	Transports Publics Neuchâtelois SA
TRAVYS	TRAVYS SA Transports Vallée de Joux-Yverdon-Sainte-Croix
TSD	Theytaz Excursions Sion
VB	Verkehrsbetriebe Biel
VBD	Verkehrsbetrieb der Landschaft Davos
VBG	VBG Verkehrsbetriebe Glattal AG
VBH	Verkehrsbetriebe Herisau
VBL	Verkehrsbetriebe Luzern AG
VBSG	Verkehrsbetriebe St.Gallen
VBSH	Verkehrsbetriebe Schaffhausen
VBZ	Verkehrsbetriebe Zürich
VMCV	Transports publics Vevey-Montreux-Chillon-Villeneuve
VSSU	Verband Schweizerischer Schifffahrtsunternehmen
VZO	Verkehrsbetriebe Zürichsee und Oberland AG
WAB	Wengernalpbahn AG
WB	Waldenburgerbahn AG
WRS	Widmer Rail Services Personal AG
WSB	Wynental- und Suhrentalbahn AAR bus+bahn
ZB	zb Zentralbahn AG
ZVB	Zugerland Verkehrsbetriebe AG
ZVV	Zürcher Verkehrsverbund ZVV
AES	Ägerisee Schifffahrt AG
BLS	BLS AG Schifffahrt Berner Oberland Thuner- und Brienzersee
BPG	Basler Personenschifffahrt AG
BSG	Bielersee-Schifffahrts-Gesellschaft AG
CGN	CGN SA
FHM	Zürichsee-Fähre Horgen-Meilen AG
LNM	Société de Navigation Lacs de Neuchâtel et Morat SA
NLM	Navigazione Lago Maggiore
SBS	SBS Schifffahrt AG
SGG	Schifffahrts-Genossenschaft Greifensee
SGH	Schifffahrtsgesellschaft Hallwilersee AG
SGV	Schifffahrtsgesellschaft des Vierwaldstättersees
SGZ	Schifffahrtsgesellschaft für den Zugersee AG / Ägerisee
SNL	Società Navigazione del Lago di Lugano SA
SW	Schiffsbetrieb Walensee AG
URh	Schweiz. Schifffahrtsgesellschaft Untersee und Rhein AG
ZSG	Zürichsee-Schifffahrtsgesellschaft AG

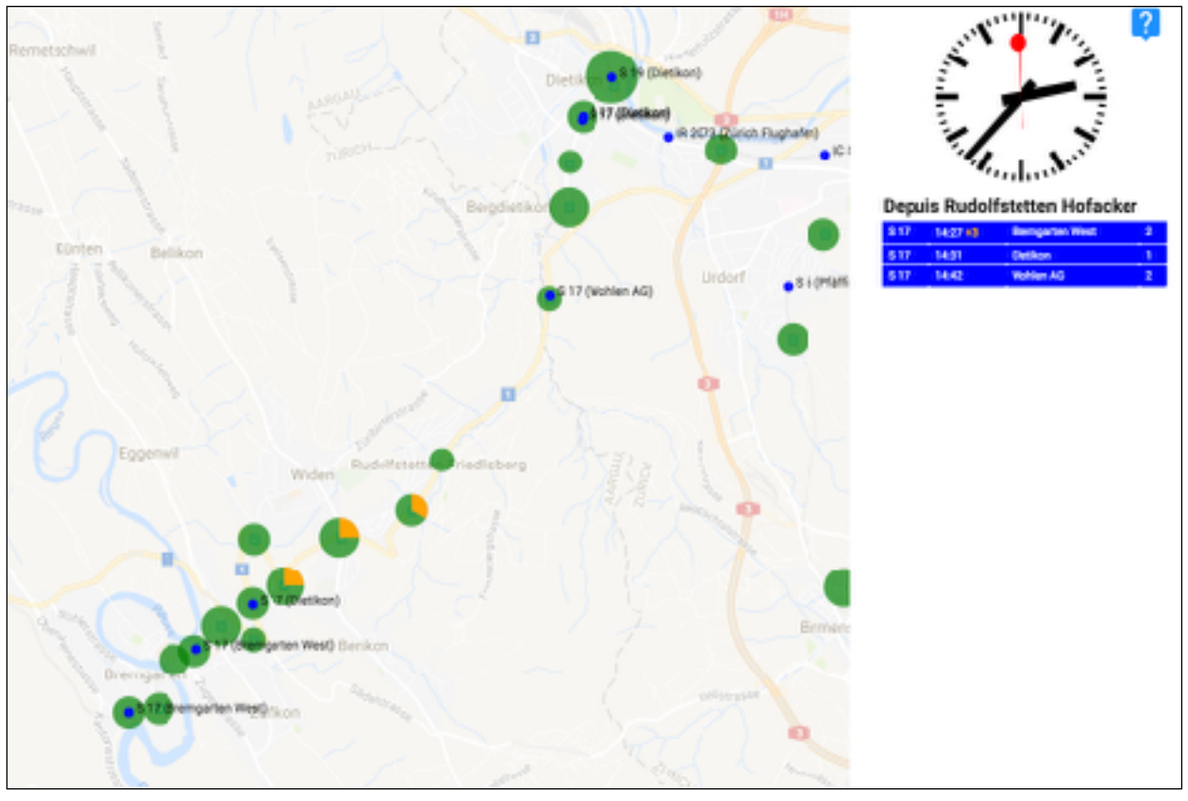
AAGL	Autobus AG Liestal	FC	FUNICAR Kursbetriebe AG	SZU	Sihltal Zürich Uetliberg Bahn SZU AG
AAGR	Auto AG Rothenburg	FLP	Ferrovie Luganesi SA	THURBO	Thurbo AG
AAGS	Auto AG Schwyz	FW	Frauenfeld-Wil-Bahn AG	TL	Transports publics de la région lausannoise SA
AAGU	AUTO AG URI	GGB	Gornergrat Bahn AG HBSAG Hafenbahn Schweiz AG	TMR	TRANSPORTS DE MARTIGNY ET REGIONS SA
AB	Appenzeller Bahnen AG	JB	Jungfraubahn AG	TPC	Transports Publics du Chablais SA
ABl	Autolinee Bleniesi SA	LEB	Chemin de fer Lausanne-Echallens-Bercher	TPF	Transports publics fribourgeois SA
ABF	Autobusbetrieb Freienbach	LLB	AG für Verkehrsbetriebe Leuk-Leukerbad und Umgebung	TPG	Transports publics genevois
AFA	Automobilverkehr Frutigen Adelboden AG	LSMS	Schilthornbahn AG	TPL	Trasporti Pubblici Luganesi SA
AMSA	Autolinea Mendrisiense SA	MBC	Transports de la région Morges-Bière-Cossonay SA	TPN	Transports Publics de la Région Nyonnaise SA
AOT	Autokurse Obe	G	Ferrovia Monte Generoso SA	TRN	Transports Publics Neuchâtelois SA
ARAG	Rottal Auto AG	GB	Matterhorn Gotthard Bahn	TRAVYS	TRAVYS SA Transports Vallée de Joux-Yverdon-Sainte-Croix
ARBAG	Aletsch Rieder	IB	Kraftwerke Oberhasli AG Meiringen-Innertkirchen-Bahn	TSD	Theytaz Excursions Sion
ARL	Autolinee Reg	OB	Chemin de fer Montreux-Oberland Bernois	VB	Verkehrsbetriebe Biel
AS	Autobetrieb Se	VR	Transports Montreux-Vevey-Riviera SA	VBD	Verkehrsbetrieb der Landschaft Davos
ASGS	Autotransport	HB	Niederhornbahn	VBG	VBG Verkehrsbetriebe Glattal AG
ASm	Aare Seeland	B	Niesenbahn AG	VBH	Verkehrsbetriebe Herisau
AVG	Autoverkehr G	StCM	Chemin de fer Nyon-St. Cergue-Morez	VBL	Verkehrsbetriebe Luzern AG
AVJ	Autotransport	eBB	Oensingen-Balsthal-Bahn	VBSG	Verkehrsbetriebe St.Gallen
AWA	Autobetrieb W	AG	PostAuto Schweiz AG	VBSH	Verkehrsbetriebe Schaffhausen
AZZK	Autobus Züric	B	PILATUS-BAHNEN AG		
BB	Bürgenstock B	A	RegionAlps SA		leneuve
BBA	Busbetrieb Aa	AILG	Railgate AG		nen
BBBW	Bus-Betrieb Bi	B	RIGI BAHNEN AG		
BDWM	BDWM Transp	BL	Regionalbus Lenzburg AG		
BGU	BGU Busbetri	BS	Regionalverkehr Bern-Solothurn AG		
BLAG	Busland AG	EGO	Regiobus Gossau AG		
BLM	Bergbahn Lau	nB	Rhätische Bahn AG		
BLS	BLS AG	NCH	DB Schenker Rail Schweiz GmbH		
BLT	BLT Baselland	LC	railCare		
BLWE	Busbetrieb Lic	VBW	Regionale Verkehrsbetriebe Baden-We		
BOB	Berner Oberla	VSH	SchaffhausenBus, Regionale Verkehrsb		
BOGG	Busbetrieb Olt	BB	SBB AG		nd Brienzersee
BOS	BUS Ostschwe	BB-D	SBB GmbH		
BOS-M	BOS Managen	C	Stadtbus Chur AG		
BRB	Brienz Rothor	F	Stadtbus Frauenfeld		
BRER	Busbetrieb Rapperswil-Eschenbach-Rüti	SBW	Stadtbus Winterthur	FHM	Zürichsee-Fähre Horgen-Meilen AG
BRSB	Braunwald-Standseilbahn AG	SMC	Cie de Chemin de Fer+d'Autobus Sierre-Montana-Crans (SMC) SA	LNM	Société de Navigation Lacs de Neuchâtel et Morat SA
BSU	Busbetrieb Solothurn und Umgebung AG	SMGN	Société des Mouettes Genevoises Navigation SA	NLM	Navigazione Lago Maggiore
BVB	Basler Verkehrs-Betriebe	SMtS	Funiculaire St-Imier - Mont-Soleil SA	SBS	SBS Schifffahrt AG
CGN	CGN SA	SOB	Schweizerische Südostbahn AG	SGG	Schifffahrts-Genossenschaft Greifensee
CJ	Compagnie des chemins de fer du Jura (C.J.) SA	SRTAG	Swiss Rail Traffic AG	SGH	Schifffahrtsgesellschaft Hallwilersee AG
CROS	Crossrail AG	SSIF	Società Subalpina di Imprese Ferroviarie S.p.A.	SGV	Schifffahrtsgesellschaft des Vierwaldstättersees
DBSCH	DB Schenker Rail Schweiz GmbH	ST	Sursee-Triengen-Bahn	SGZ	Schifffahrtsgesellschaft für den Zugersee AG / Ägerisee
DBZ	Dolderbahn Zürich	STB	Sensetalbahn AG	SNL	Società Navigazione del Lago di Lugano SA
ETB	Emmentalbahn, Huttwil	STI	Verkehrsbetriebe STI AG	SW	Schiffsbetrieb Walensee AG
FART	Ferrovie Autolinee Regionali Ticinesi	SVB	BERNMOBIL Städt. Verkehrsbetriebe Bern	URh	Schweiz. Schifffahrtsgesellschaft Untersee und Rhein AG
FB	Forchbahn AG	SWAG	Seilbahn Weissenstein AG	ZSG	Zürichsee-Schifffahrtsgesellschaft AG

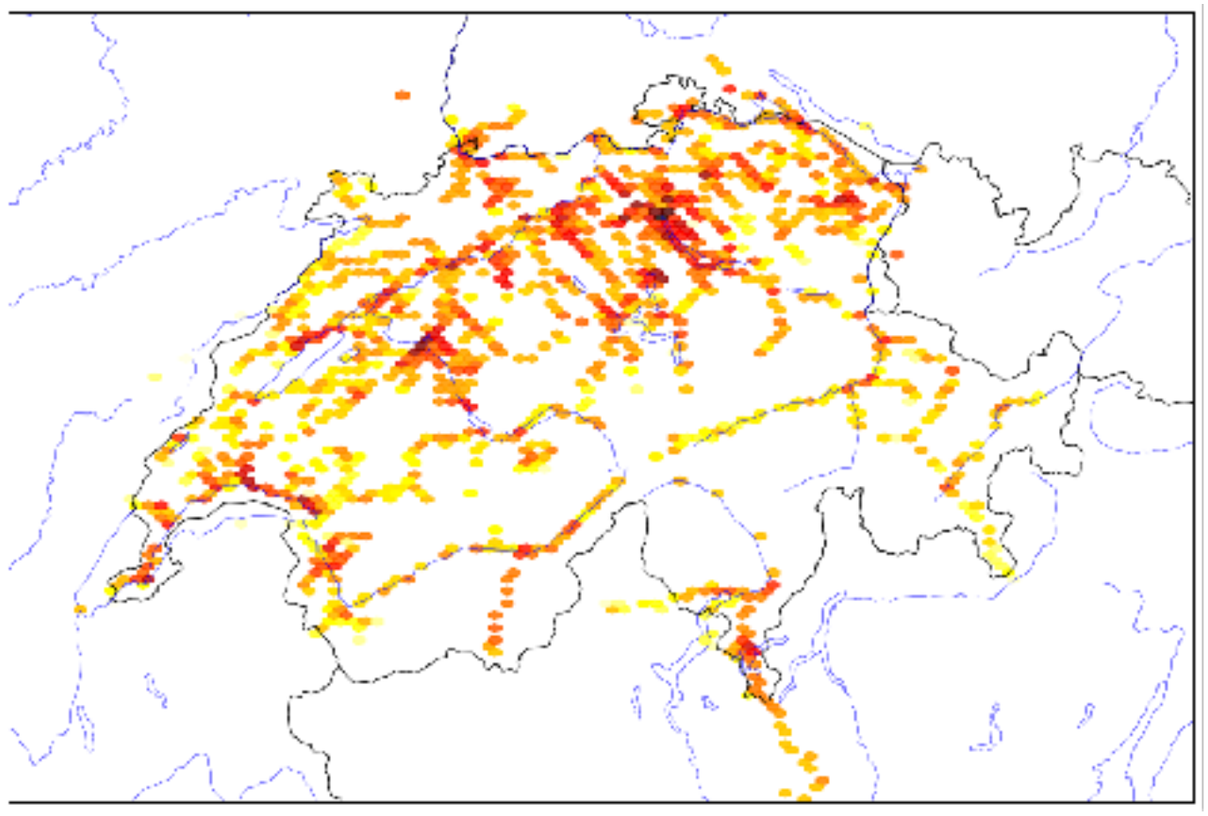
Voyage avec	Dép.	Prévision	À	Voie/Quai/Arrêt
S 3 12375	19:36		**Villeneuve** Lausanne 19:36 - Pully 19:38 - Lutry 19:41 - Cully 19:45 - Vevey 19:52 - La Tour-de-Peilz 19:54 - Burier 19:56 - Clarens 19:58 - Montreux 20:02 - Villeneuve 20:06	6
IC 728	19:42		**Genève-Aéroport** Lausanne 19:42 - Genève 20:18 - Genève-Aéroport 20:27	5
RE 3133	19:42		**Romont** Lausanne 19:42 - Palézieux 19:57 - Romont 20:13	1
IR 1730	19:48		**Genève-Aéroport** Lausanne 19:48 - Morges 19:58 - Nyon 20:14 - Genève 20:30 - Genève-Aéroport 20:39	8
IR 1633	19:50		**Brig** Lausanne 19:50 - Vevey 20:03 - Montreux 20:09 - Aigle 20:20 • St-Maurice 20:31 - Martigny 20:42 - Sion 20:56 - Sierre/Siders 21:07 - Visp 21:24 - Brig 21:32	6
IR 2535	19:50	ca. +4 min	**Luzern** Lausanne 19:50 - Fribourg/Freiburg 20:33 - Bern 20:56 - Zofingen 21:27 - Sursee 21:40 - Luzern 22:00	1
RE 3230				

Gare/Arrêt	Arrivée	Pronostic	Départ	Pronostic	Catégorie et numéro	Voie/Quai	Occupation
Genève-Aéroport			19:03		IR 2535	1	1. 2.
Genève	19:09		19:12			6	1. 2.
Lausanne	19:48	ca. +4 min	19:50	ca. +4 min		1	1. 2.
Fribourg/Freiburg	20:33		20:34			3	1. 2.
Bern	20:56		21:00			8	1. 2.
Neubaustrecke							
Zofingen	21:27		21:28			2	1. 2.
Sursee	21:40		21:41				
Luzern	22:00						

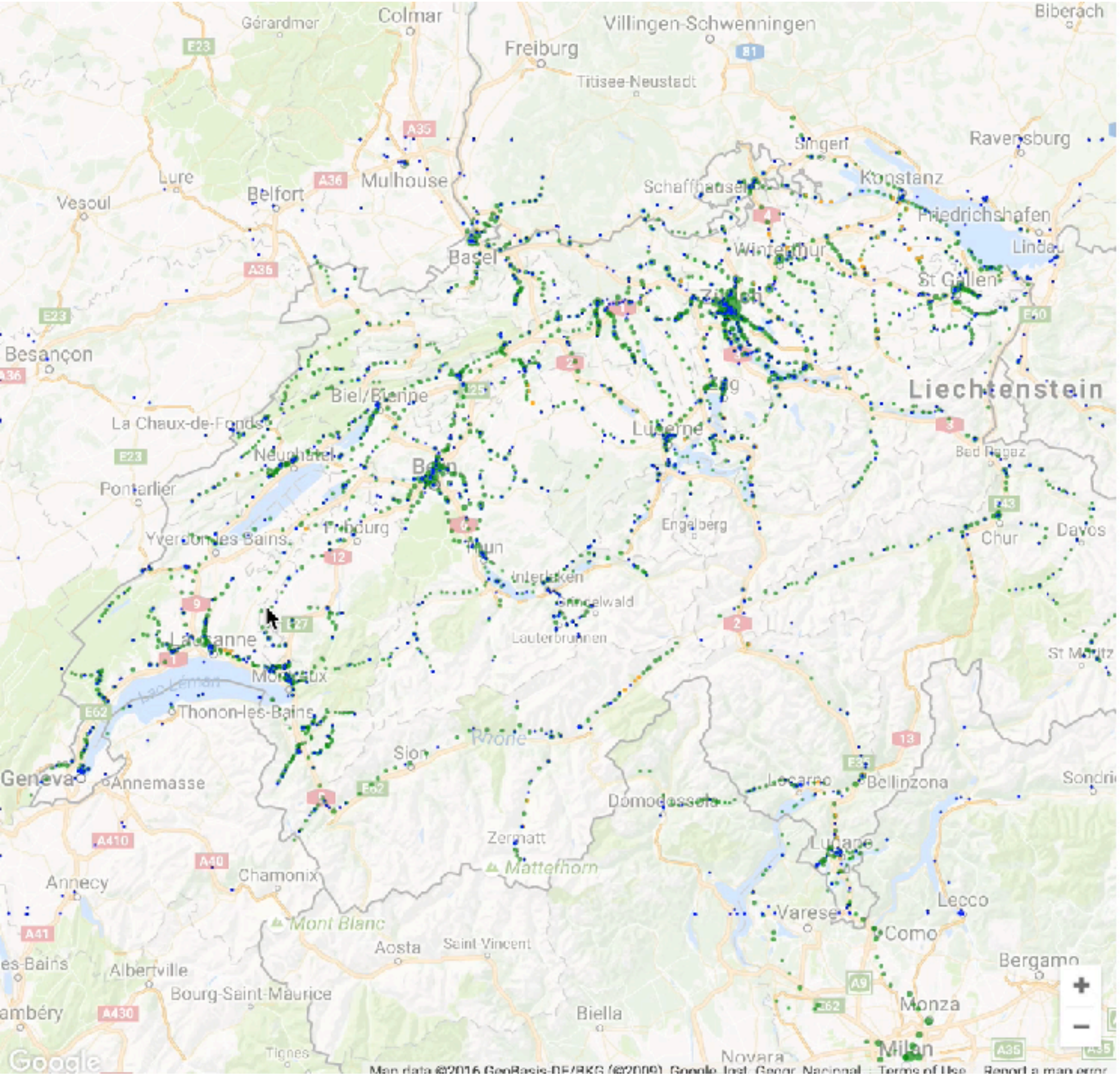


What do we propose?

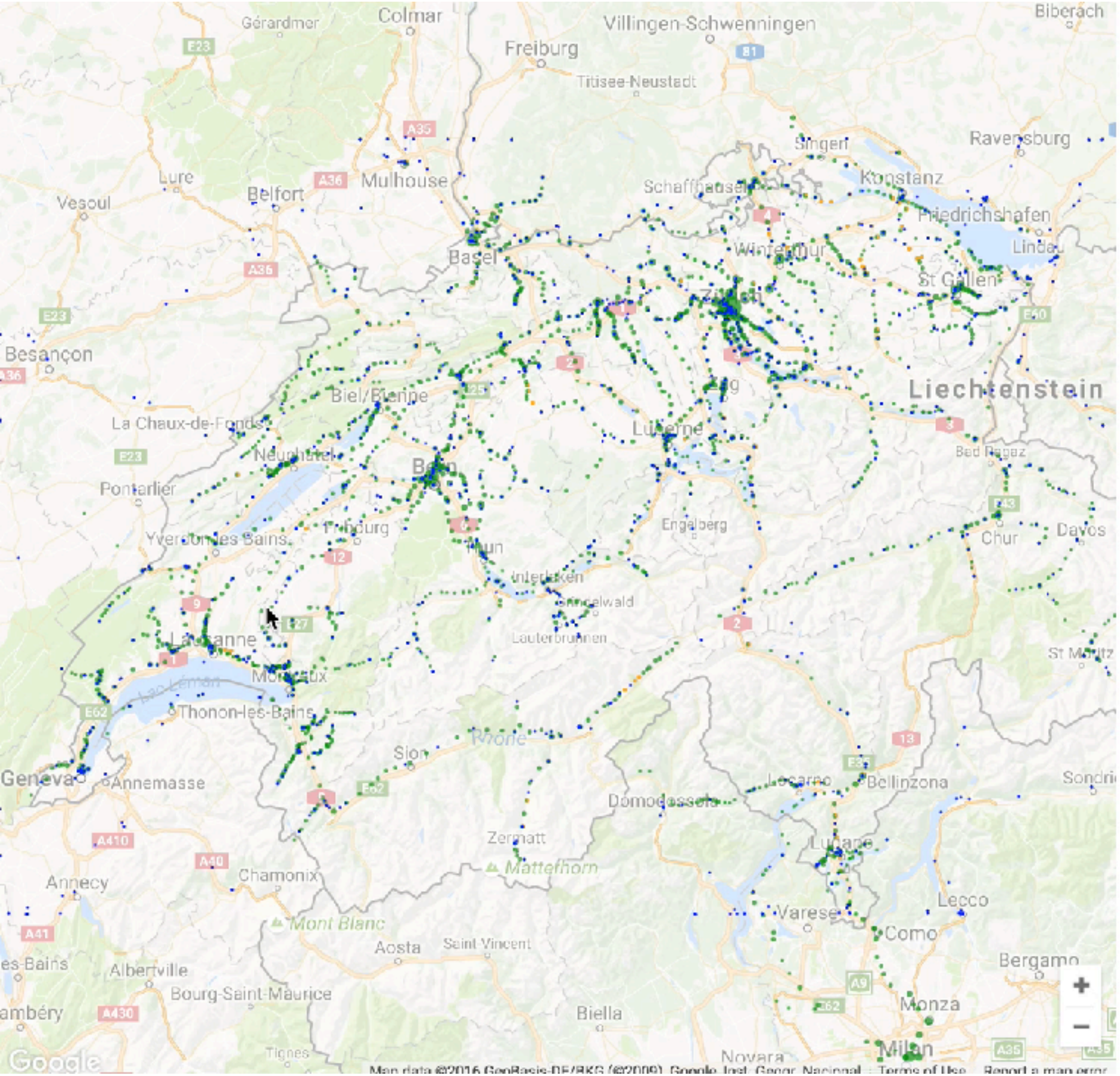




https://github.com/alexmasselot/swiss-transport-realtime



Depuis Fideris



Depuis Fideris

Is it possible to build

a **simple scalable** infrastructure, to

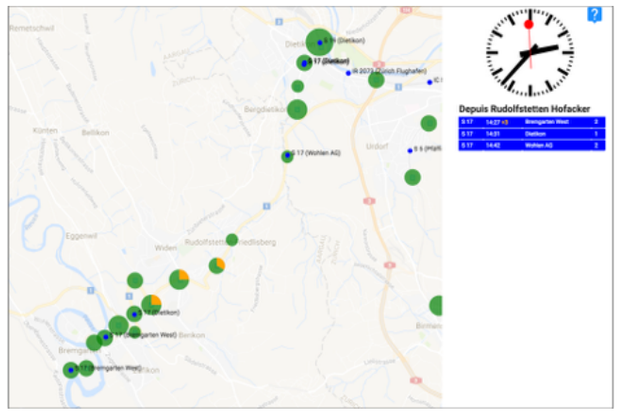
dispatch, transform and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?

vehicles positions

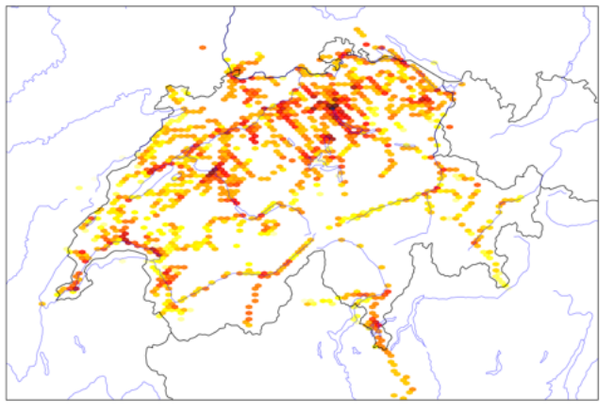
real time



users

station boards

offline



data analysts

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?

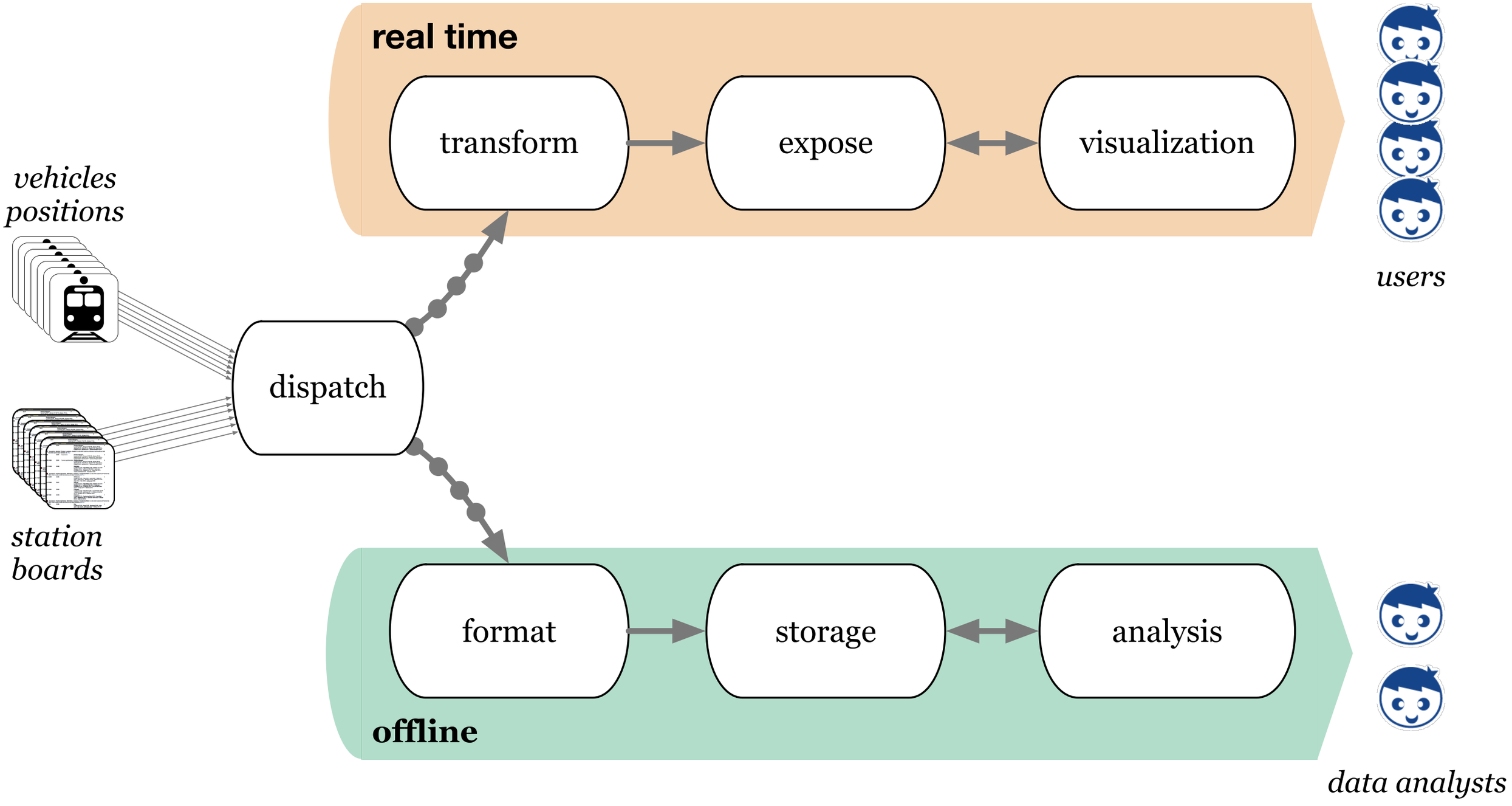
Is it possible to build

a **simple** **scalable** infrastructure, to

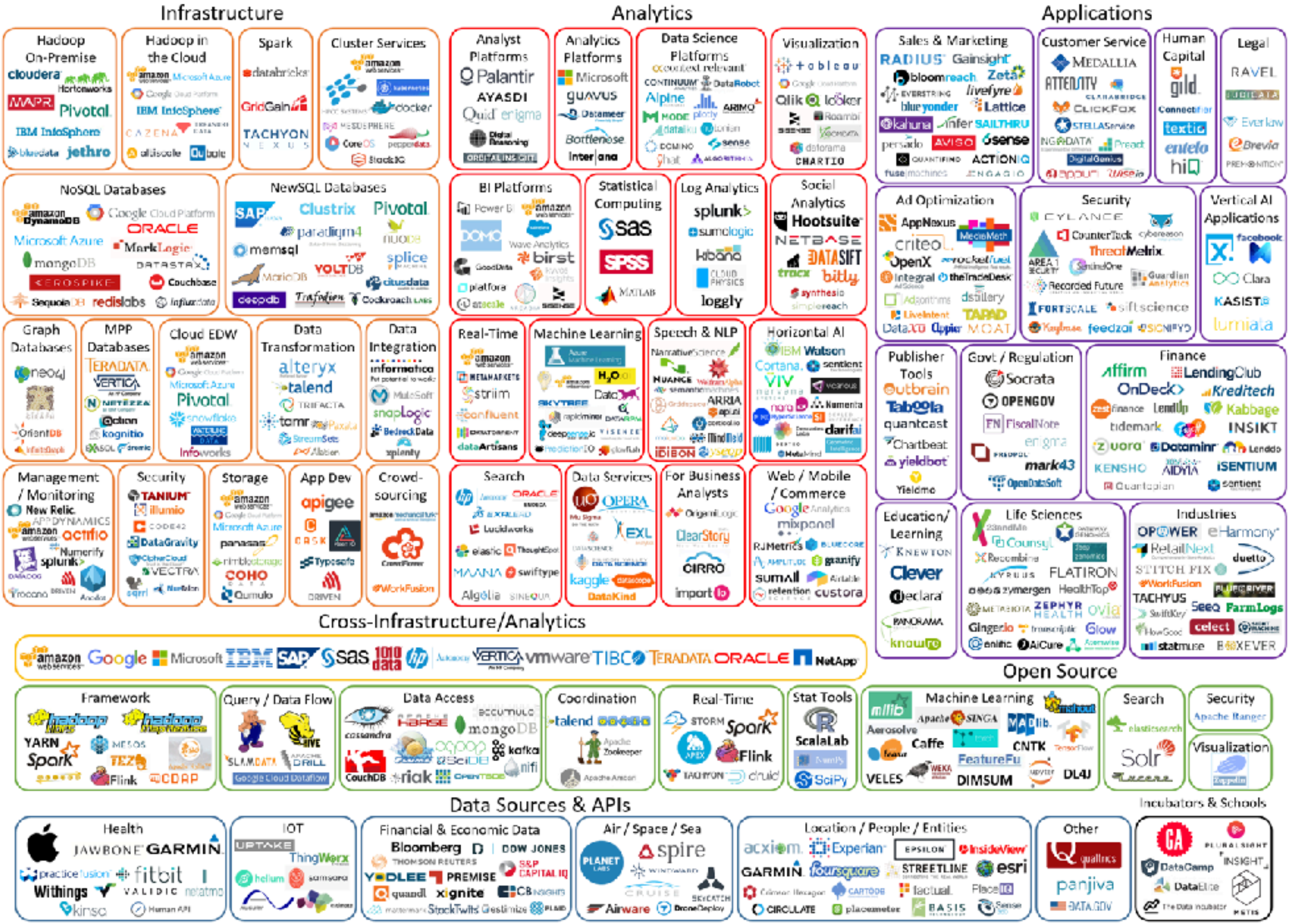
dispatch, transform and **visualize**

“**near real time**” massive data

and achieve ***a posteriori*** **analysis**?


real time
transform
expose
visualization
vehicles positions
dispatch
station boards
format
storage
analysis
offline
users
data analysts

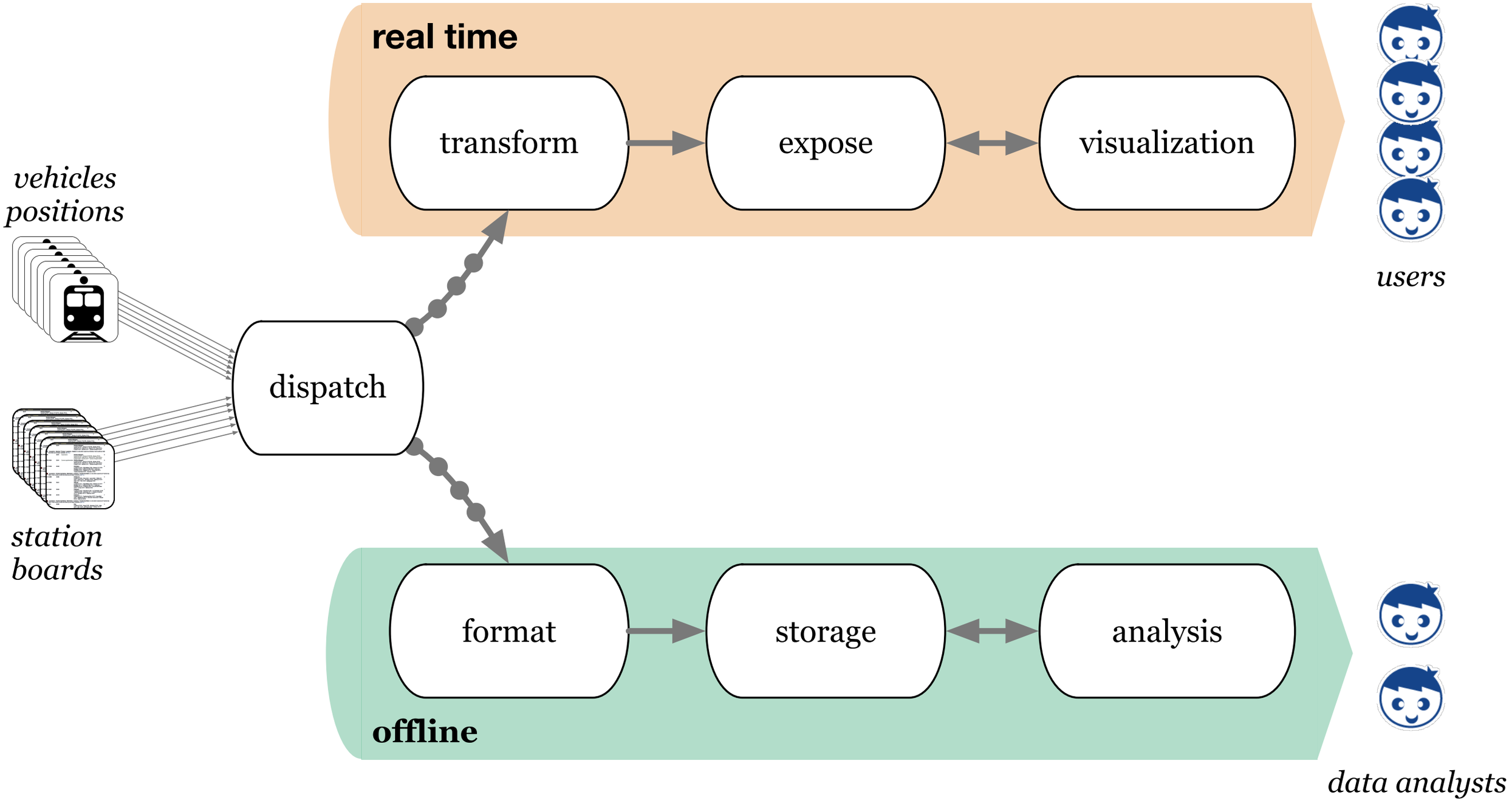
Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)

FIRSTMARK

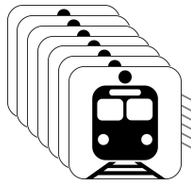

real time
transform
expose
visualization
vehicles positions
dispatch
station boards
format
storage
analysis
offline
users
data analysts

This is only a POC!!!

real time

transform → expose ↔ visualization

vehicles positions



station boards

dispatch

users

format → storage ↔ analysis

offline

data analysts

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?

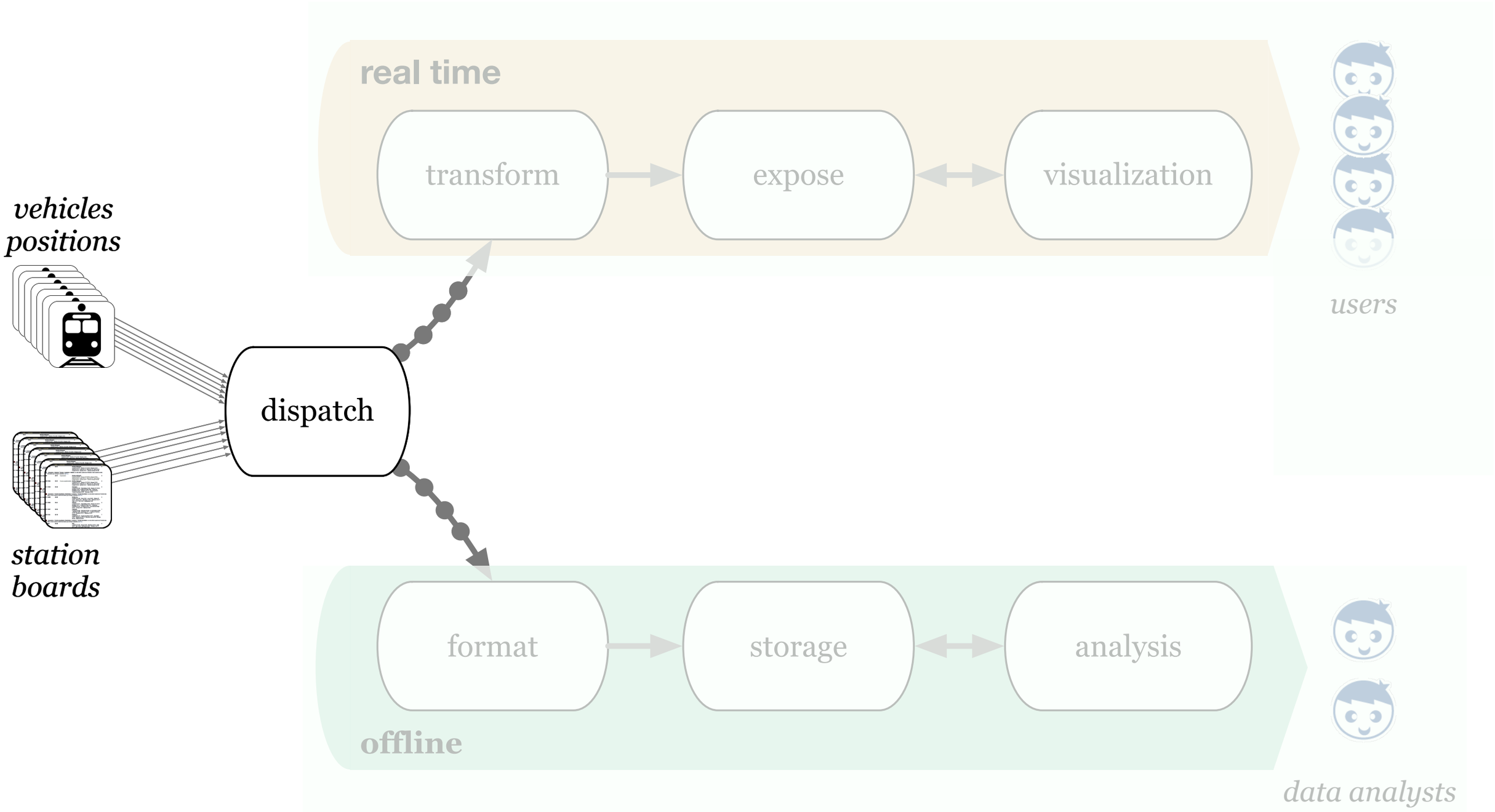
Is it possible to build

a **simple scalable** infrastructure, to

dispatch, **transform** and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?


real time
transform
expose
visualization
users
vehicles positions
dispatch
station boards
format
storage
analysis
offline
data analysts

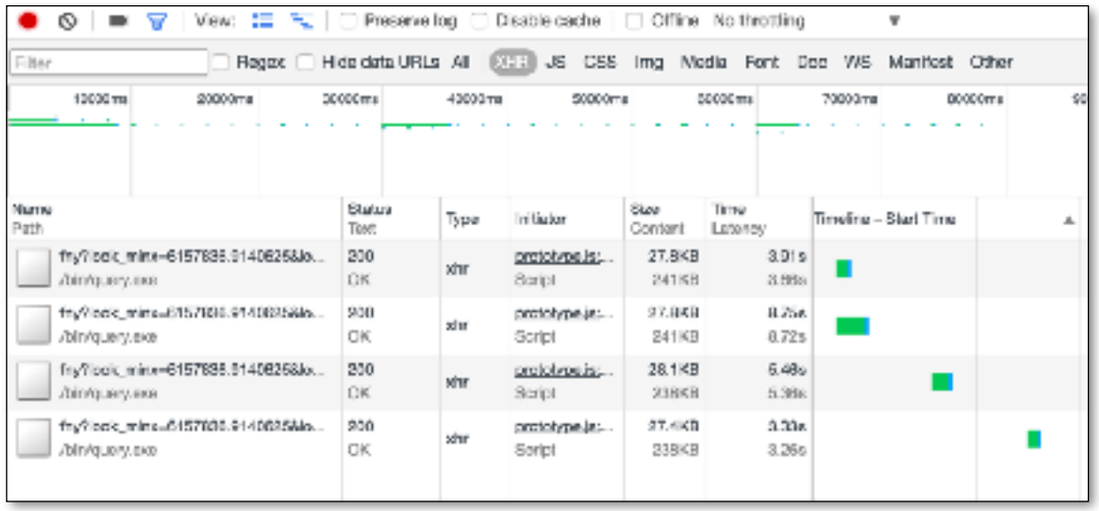
Acquire

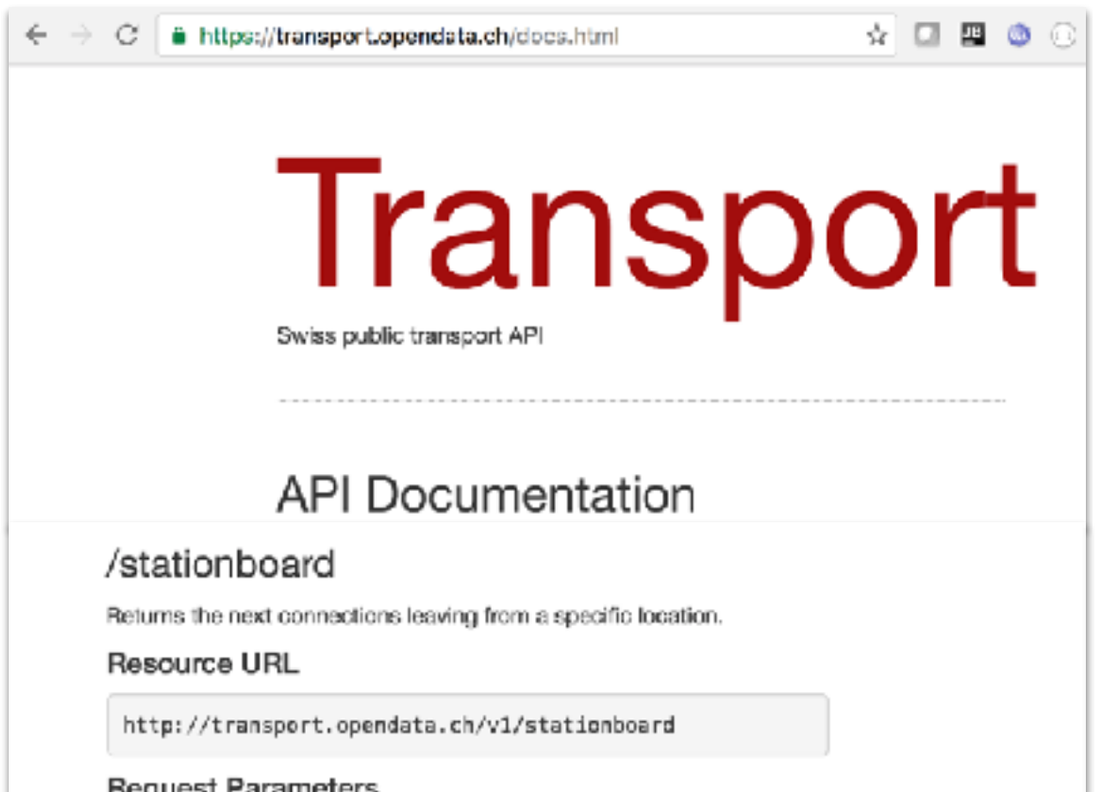
vehicles positions

SBB rest api

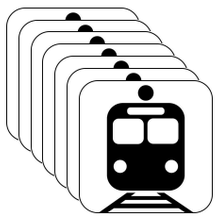
station boards

OpenData transport api





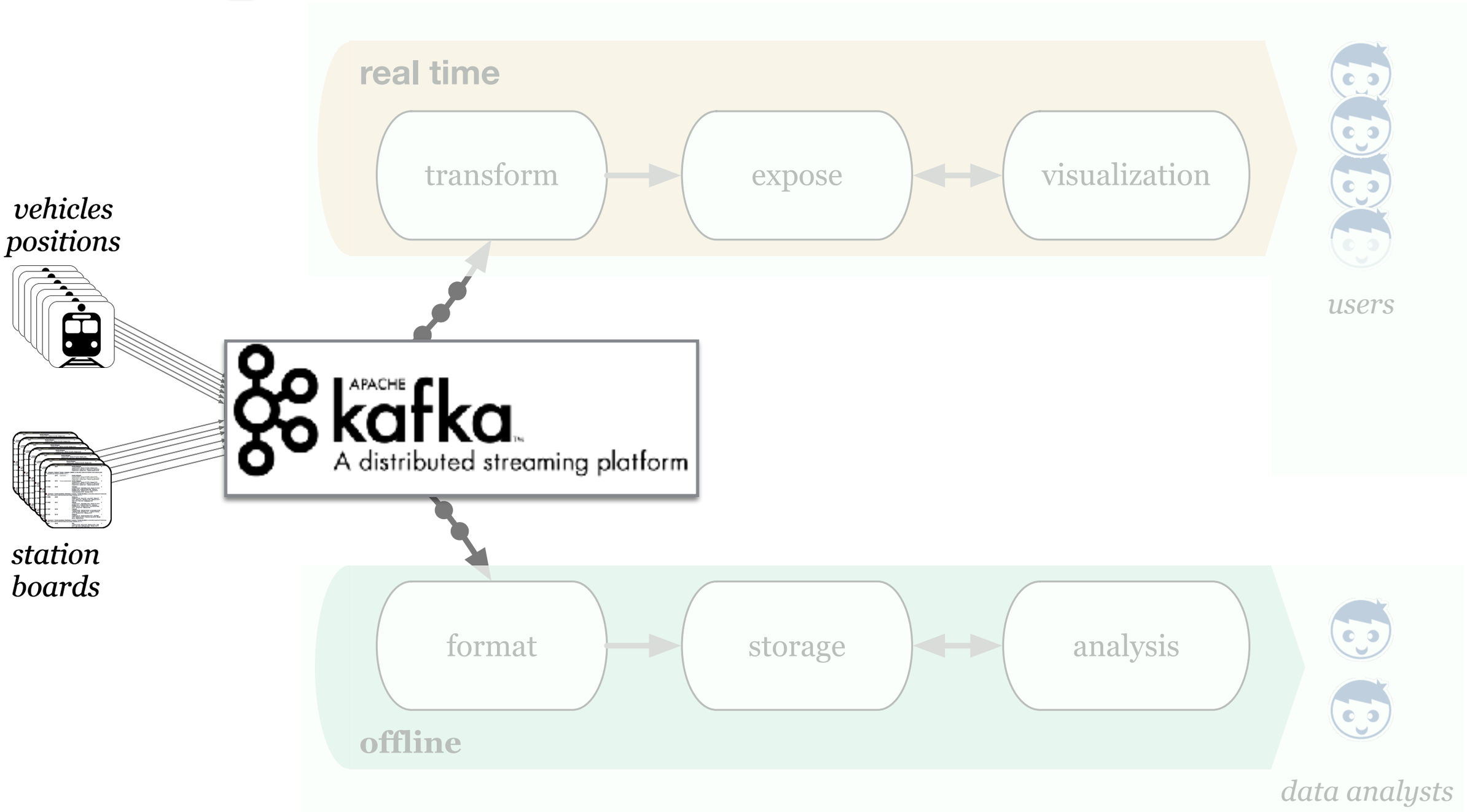
```
{
   id: 12345xyz,
   category: IR,
   name: IR 72928,
   destination: Alpnach,
   position: {
      lat: 46.940582,
      lon: 8.275442
   }
}
```



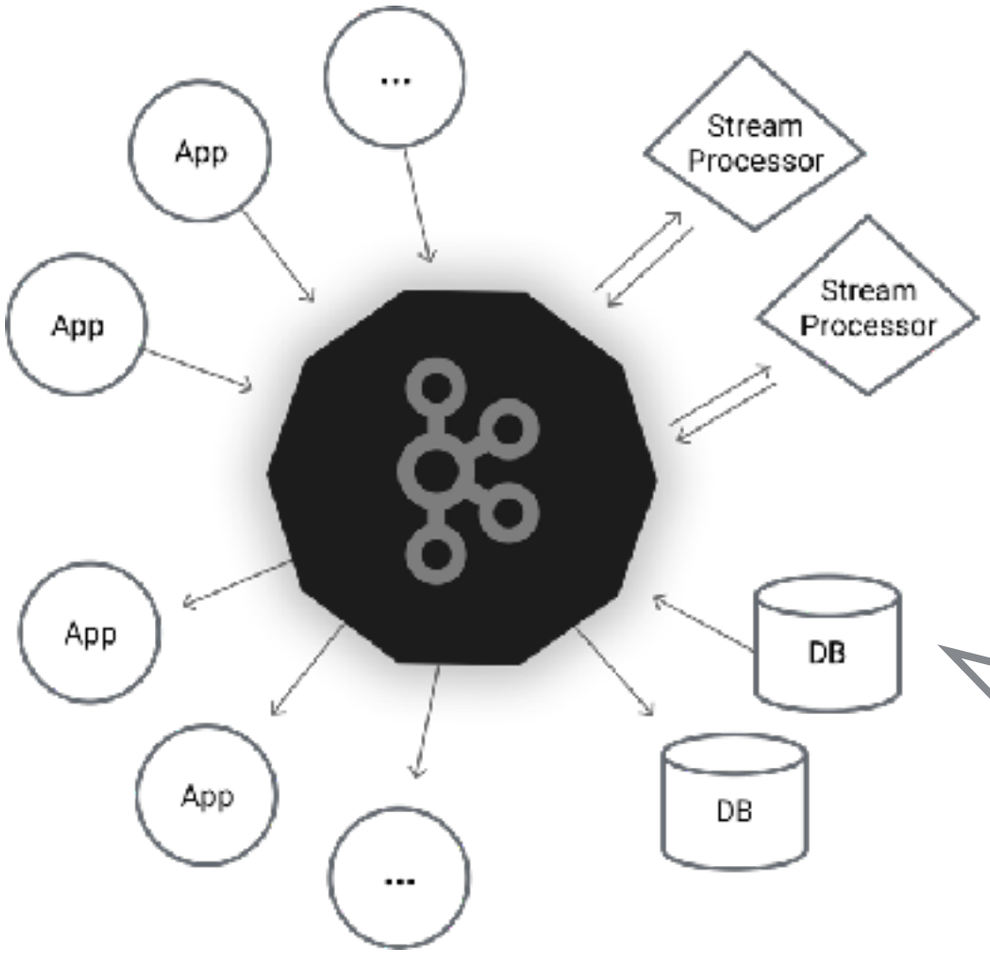
```
{
   id: 12345xyz,
   category: IR,
   name: IR 72928,
   destination: Alpnach,
   position: {
      lat: 46.940582,
      lon: 8.275442
   }
}
```

```
{
   station: {
     name: Lausanne,
     location: {lat, long}
   },
   departures: [
     {
       to:Domodossola,
       time: 20:13,
       delayed: 4,
       prognosis: {
        capacity2nd: 3,
        capacity1st: 1
       }
   },
   {…}
```

Dispatch


real time
transform
expose
visualization
users
vehicles positions
APACHE kafka
A distributed streaming platform
station boards
format
storage
analysis
offline
data analysts

Events are streamed to

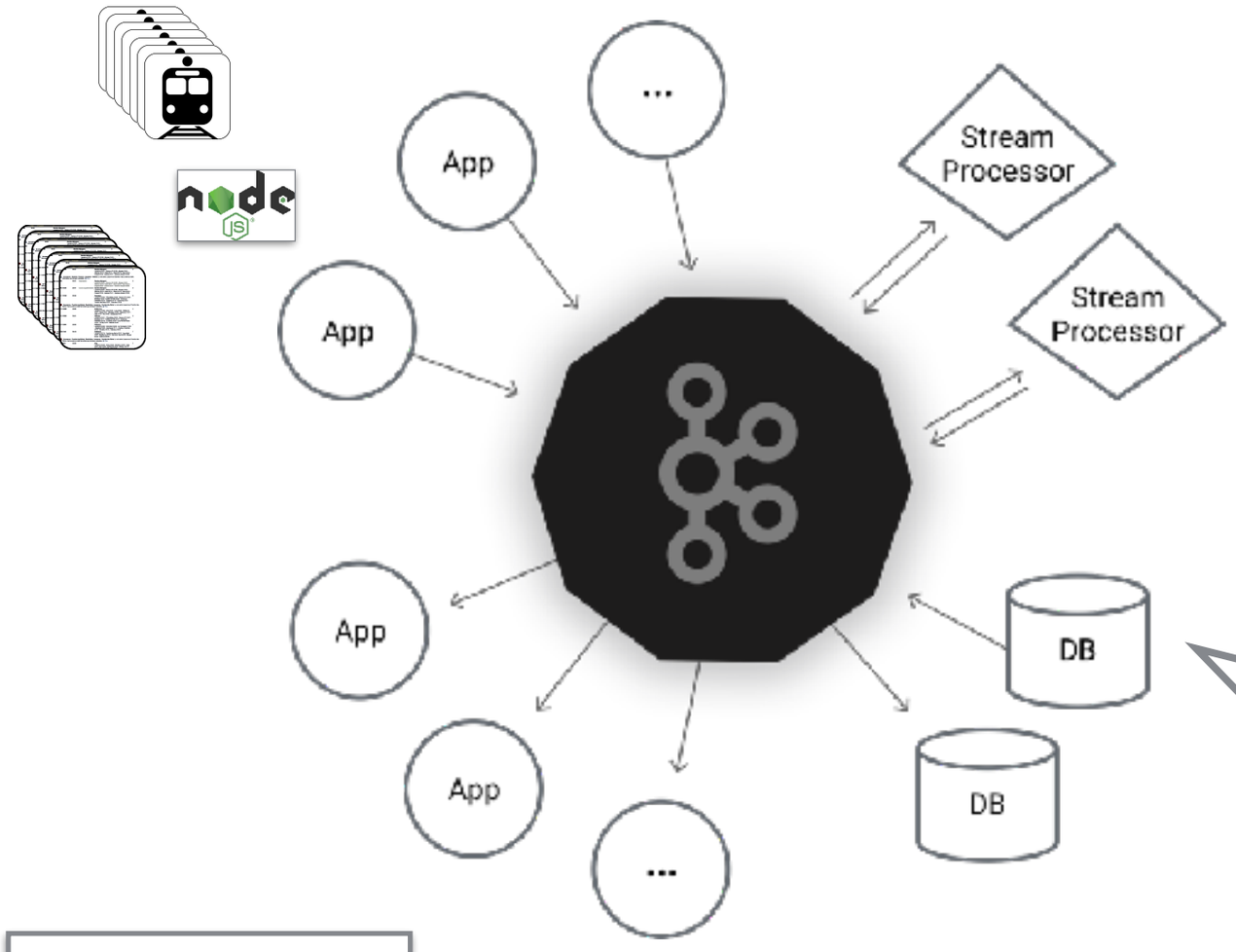


*"Kafka is used for building **real-time data pipelines** and streaming apps. It is horizontally **scalable**, **fault-tolerant**, wicked fast, and runs in production in thousands of companies."*

kafka.apache.org

Events are streamed to

APACHE kafka™
A distributed streaming platform



real time

offline

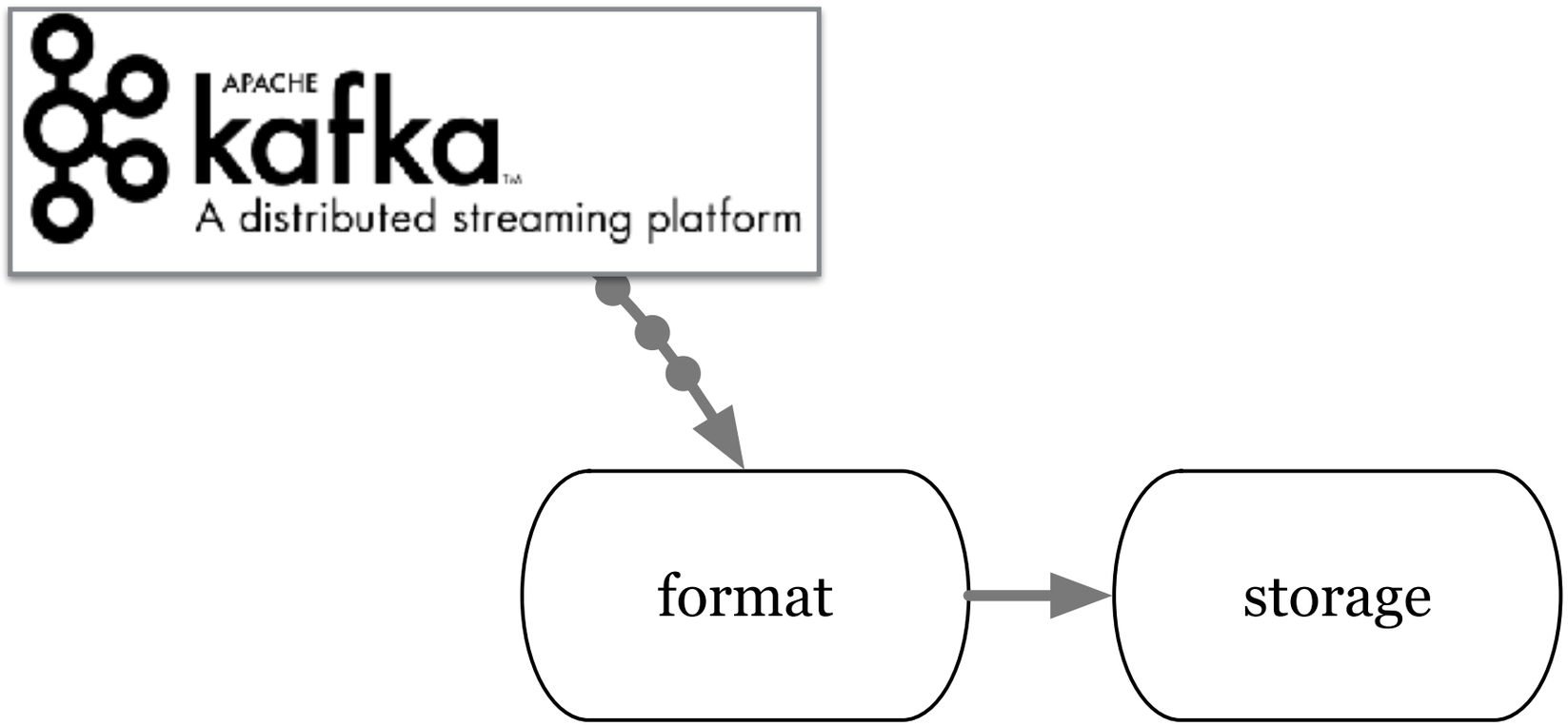
*"Kafka is used for building **real-time data pipelines** and streaming apps. It is horizontally **scalable**, **fault-tolerant**, wicked fast, and runs in production in thousands of companies."*

kafka.apache.org

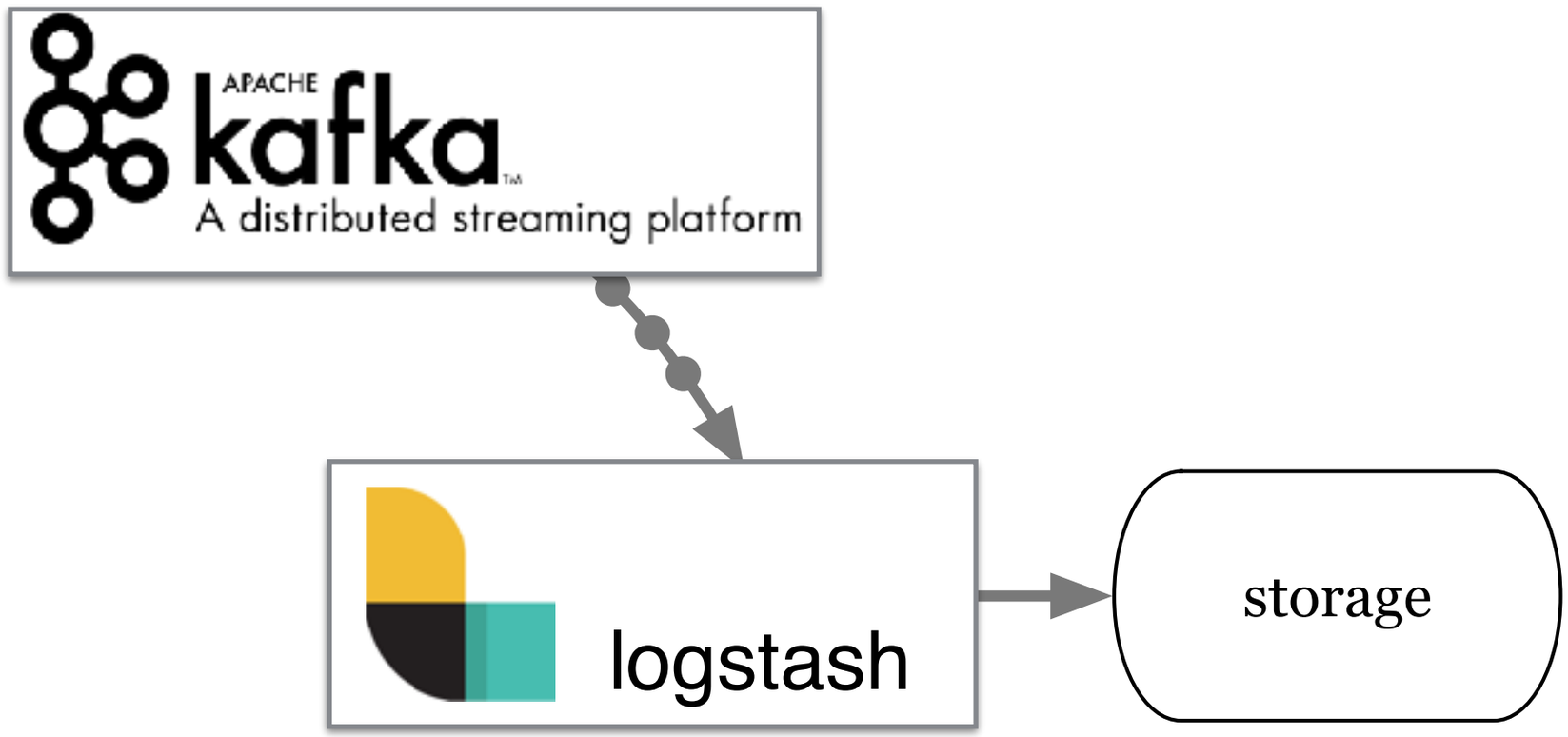
TIMTOWTDI

Kafka, RabbitMQ, ZeroMQ...

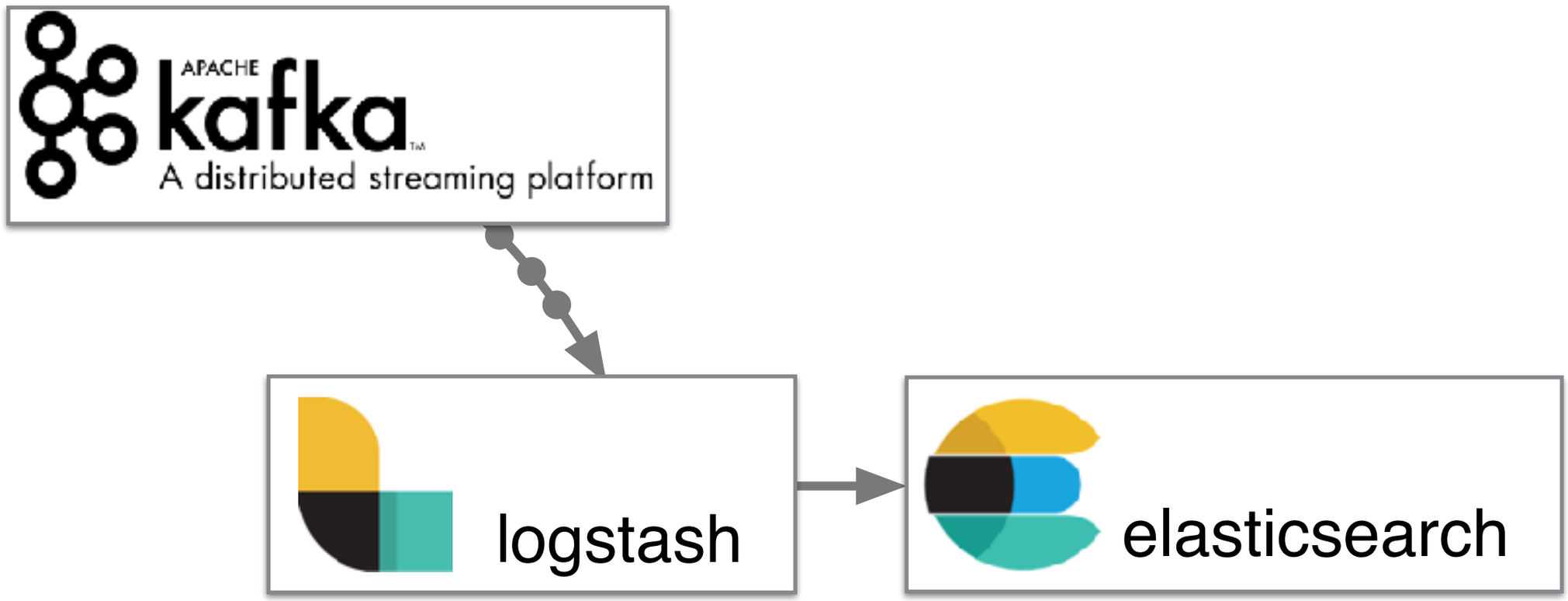
Store


APACHE
kafka
A distributed streaming platform
format
storage

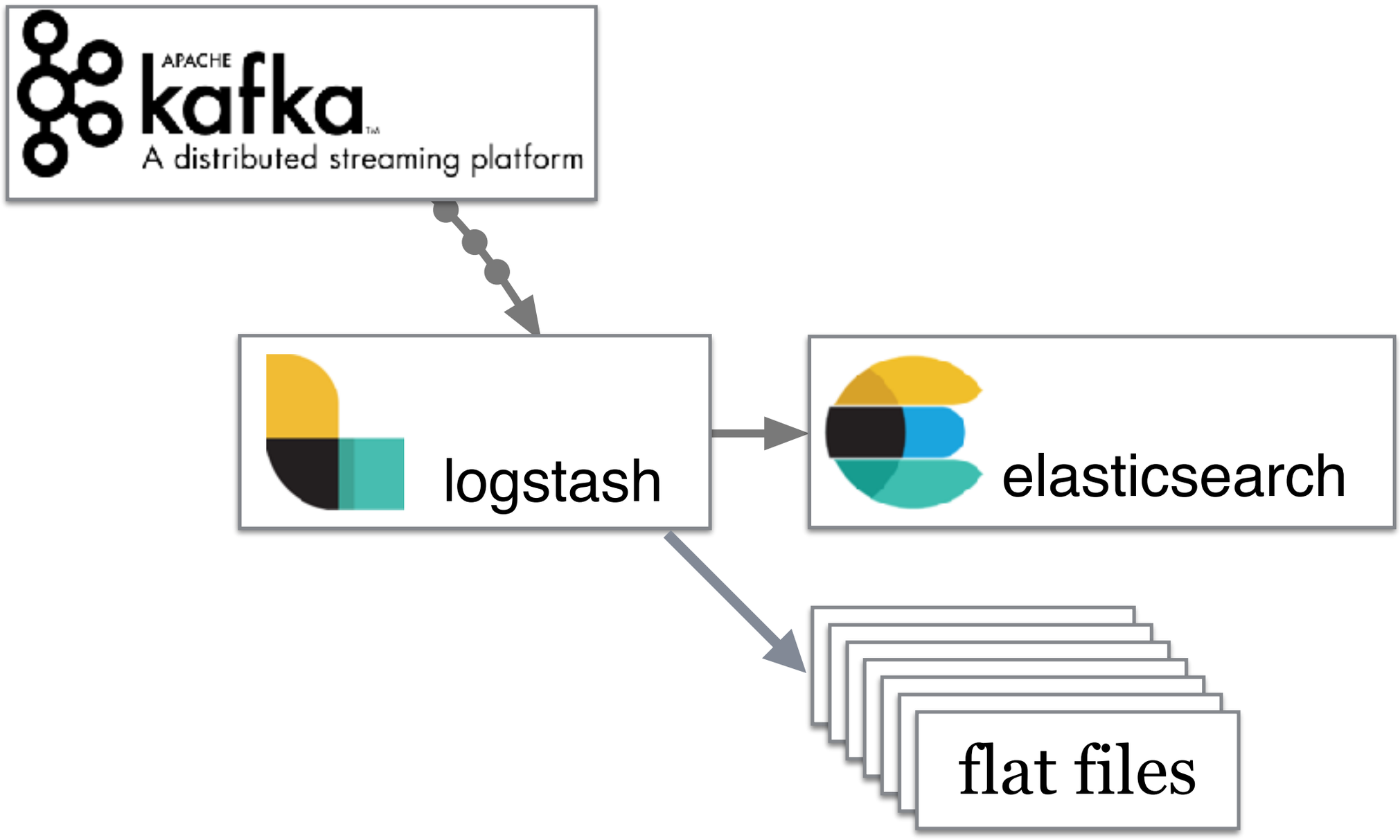
Store


APACHE
kafka
A distributed streaming platform
logstash
storage

Store


APACHE
kafka
A distributed streaming platform
logstash
elasticsearch

Store

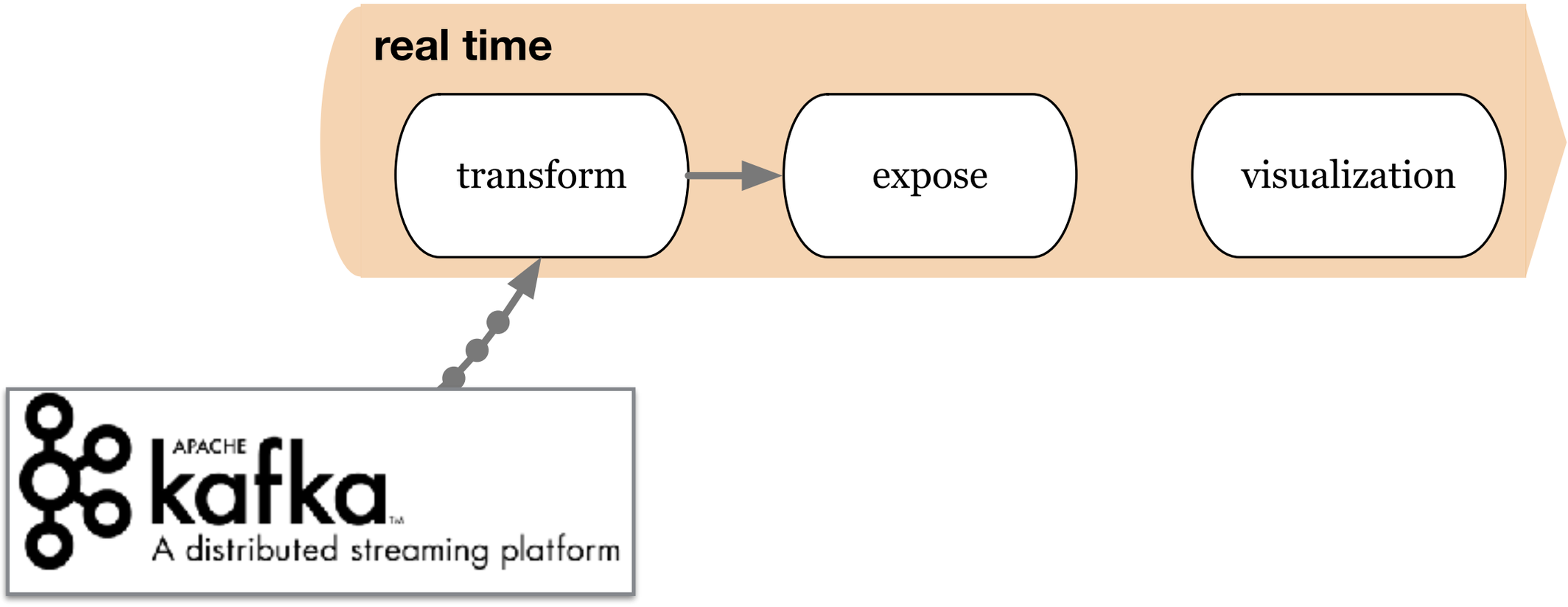

APACHE
kafka
A distributed streaming platform
logstash
elasticsearch
flat files

TIMTOWTDI

Logstash, Flume, Filebeat…

TIMTOWTDI

Elasticsearch, HBase, Cassandra...


real time
transform
expose
visualization
APACHE
kafka
A distributed streaming platform

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

"near real time" massive data

and achieve ***a posteriori*** **analysis**?

Stream transformation

- We have an input flow of events and want to:
  - know if a train is stopped into a station;
  - know if a train as exited the network;
  - expose an aggregated station board.

- We need to:
  - digest the input flow;
  - process with temporary state persistance;
  - be able to expose snapshots.

Stream transformation

- **Scala** is The language for Big Data (functional & OO)
- **Akka (actors):**
  - lightweight entities (one per train, per station);
  - easy asynchronous communications;
  - the perfect use case.
- **Play** framework for REST service, configuration etc.

TIMTOWTDI

Spark Streaming, Storm, Flink...

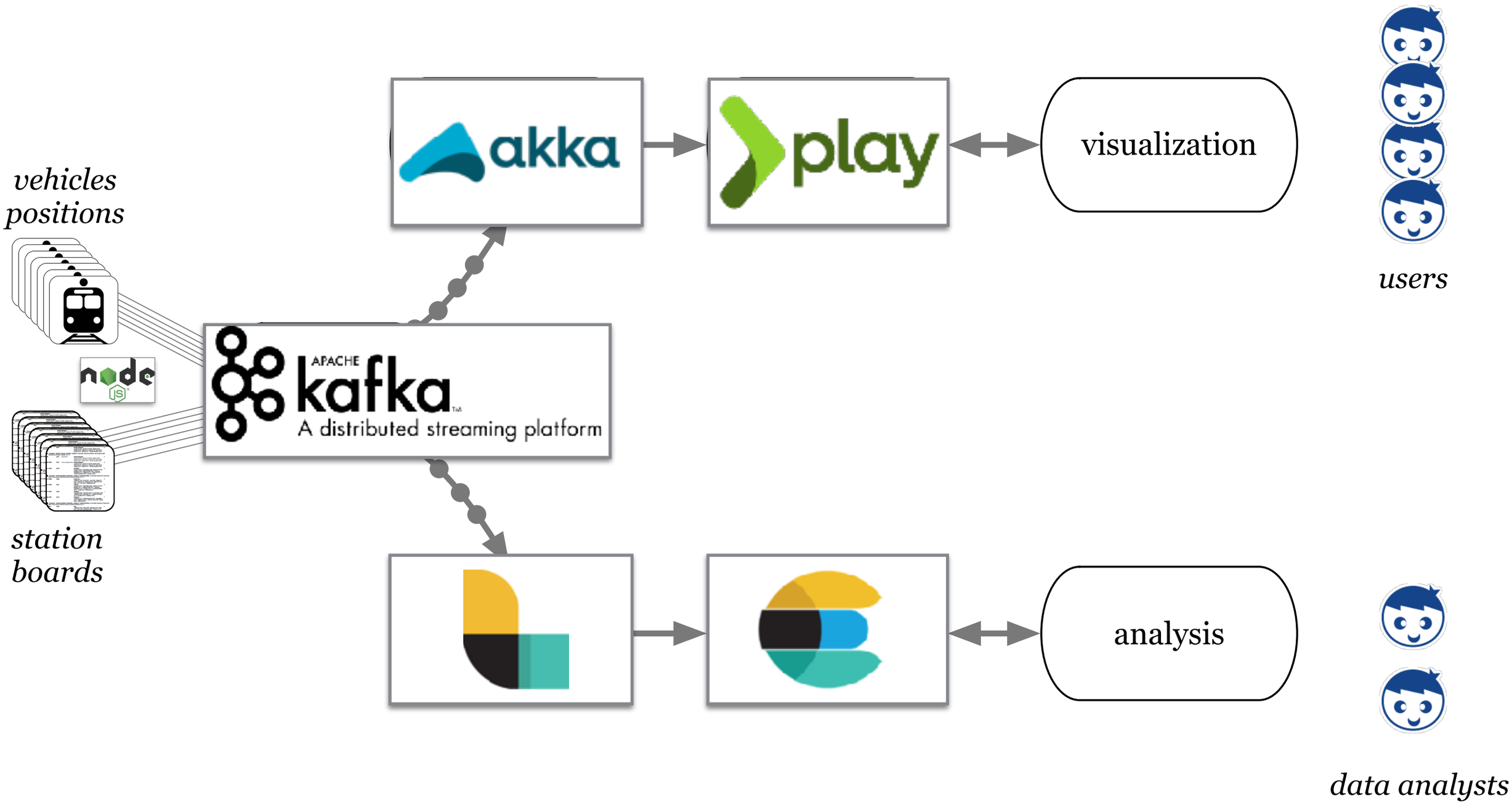
TIMTOWTDI

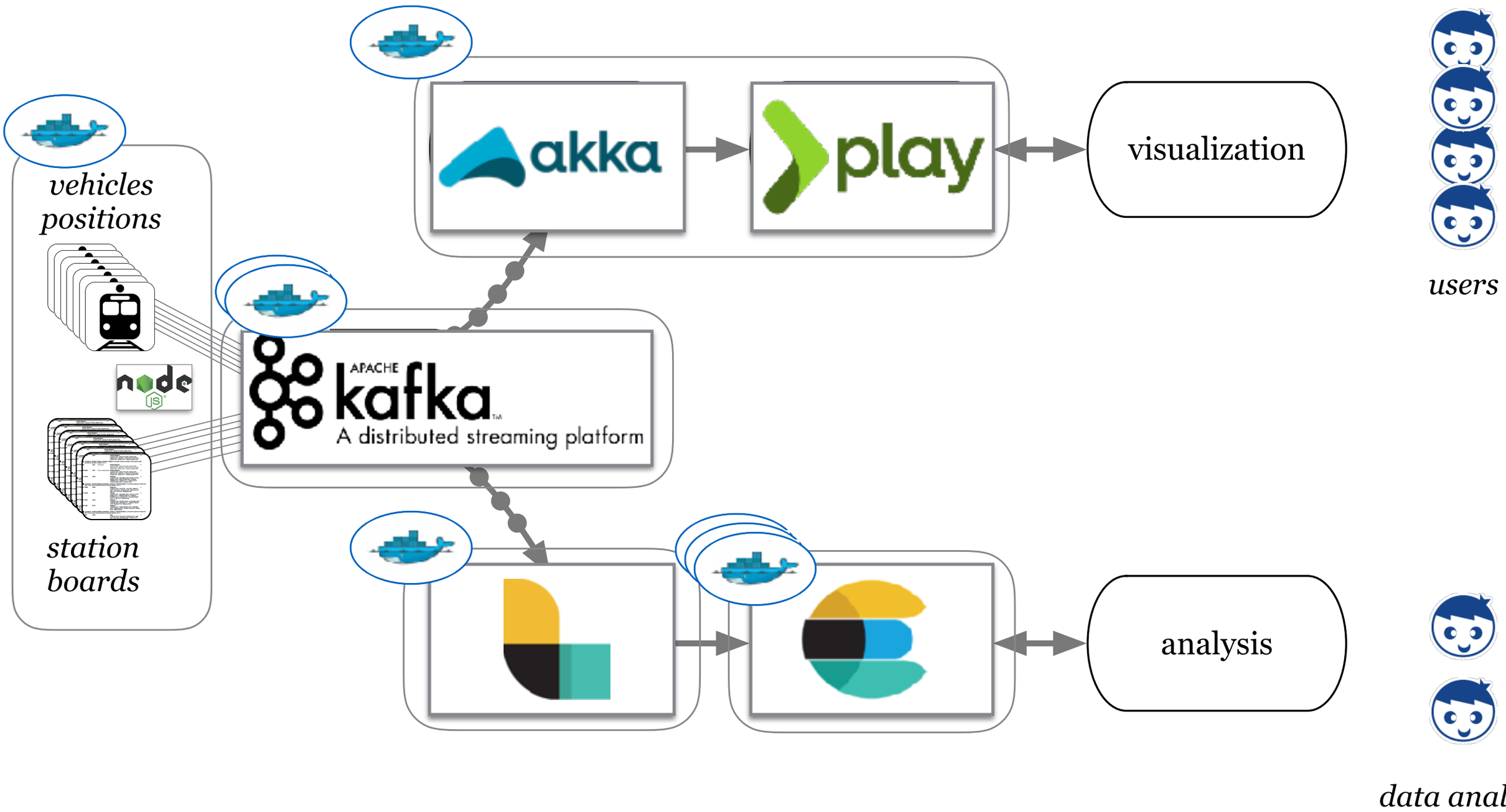
Spark Streaming, Storm, Flink...

DevOps

: putting everything together

- The “simple” infrastructure is not so light;
- A developper should have everything on his/her laptop without polluting the machine;
- Docker comes to the rescue:
  - lightweight containers,
  - pre-existing images,
  - docker-compose to describe the infrastructure
  - deploy directly to AWS or GCE.


akka
play
visualization
vehicles positions
APACHE kafka
A distributed streaming platform
station boards
analysis
users
data analysts


vehicles positions
station boards
akka
play
visualization
users
APACHE kafka
A distributed streaming platform
analysis
data analysts

Performance: 2 numbers

Performance: 2 numbers

15x faster ajax queries (vs SBB rest)
to gather 30 times more trains

Performance: 2 numbers

15x faster ajax queries (vs SBB rest)
to gather 30 times more trains

15% CPU: nodeJS + kafka + akka + play

Name	Zone	Machine type	Recommendation	In use by	Internal IP	External IP	Connect
gce-swiss-transport-rt	europe-west1-b	1 vCPU, 3.75 GB			10.132.0.2	104.199.35.228	SSH

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?

Is it possible to build

a **simple** **scalable** infrastructure, to

dispatch, transform and **visualize**

“**near real time**” massive data

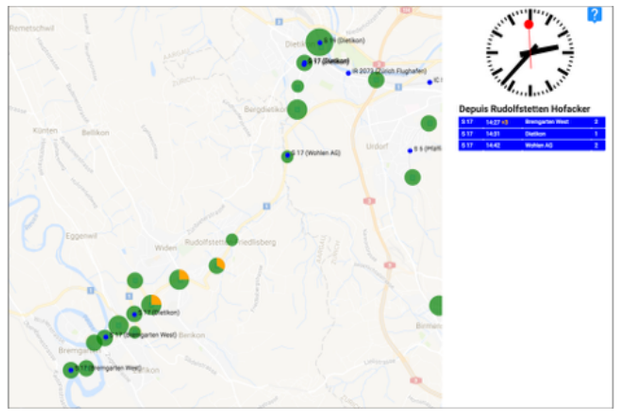
and achieve ***a posteriori*** **analysis**?

A scalable infrastructure

Kafka	partitioning and zookeeper
Logstash	? (but naturally recover on failure)
Elasticsearch	partitioning
Spark streaming	distributed by essence & write ahead logs
Akka 	aka cluster, supervisors & failure strategy
Docker	Kubernetes, AWS, GCE, Exoscale

vehicles positions

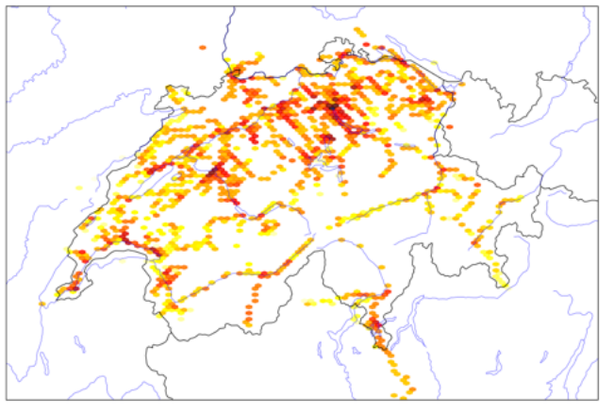
real time



users

station boards

offline



data analysts

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?

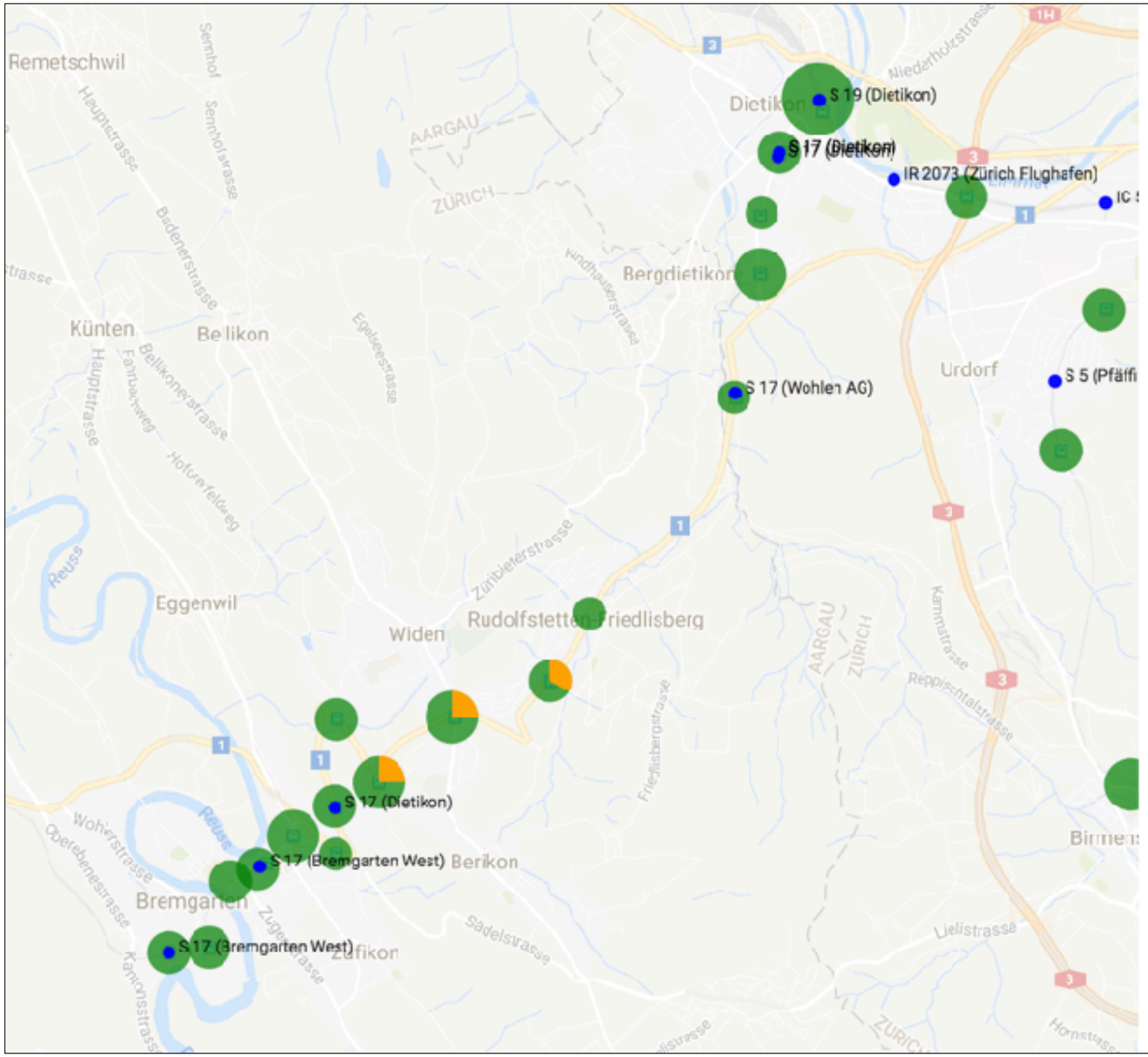
Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

"near real time" massive data

and achieve ***a posteriori*** **analysis**?



Depuis Rudolfstetten Hofacker

S 17	14:27 +3	Bremgarten West	2
S 17	14:31	Dietikon	1
S 17	14:42	Wohlen AG	2

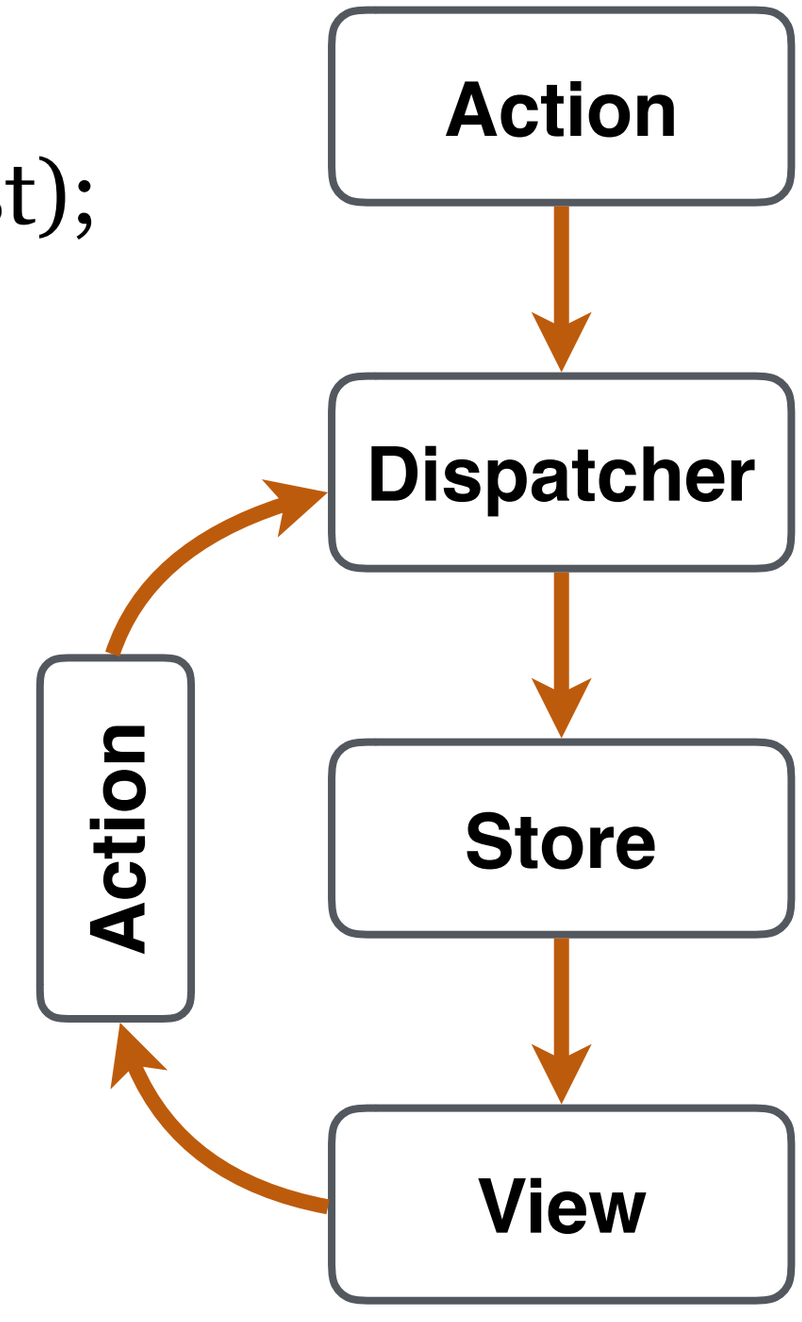
Data-Driven Documents

JS for large data set

- Only a rendering library (but fast);
- Use a flux architecture;
- Built by Facebook.

React JS for large data set

- Only a rendering library (but fast);
- Use a flux architecture;
- Built by Facebook.



JavaScript for big data viz

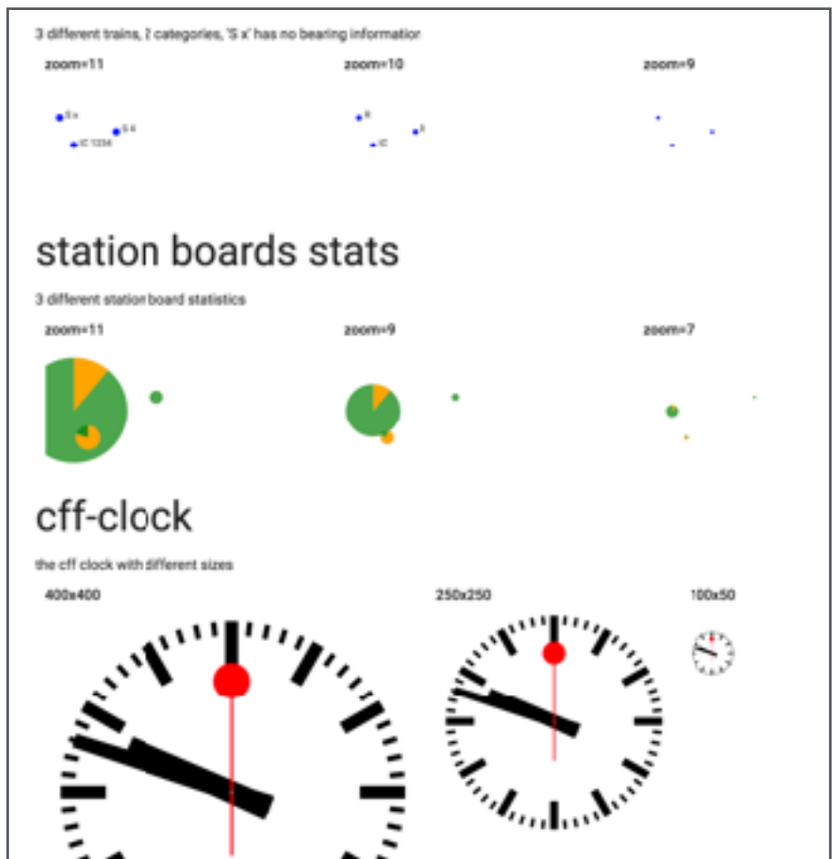
- React can handle viz >100k elements (don’t show them individually!)

JavaScript for big data viz

- React can handle viz >100k elements (don't show them individually!)
- Beware of performance issue;

JavaScript for big data viz

- React can handle viz >100k elements (don't show them individually!)
- Beware of performance issue;
- Testing is not an option.



Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

“near real time” massive data

and achieve ***a posteriori*** **analysis**?

Is it possible to build

a **simple scalable** infrastructure, to

dispatch, transform and **visualize**

"near real time" massive data

and achieve ***a posteriori*** **analysis**?

4.5 months of data

A. What is the train occupancy during weekdays, between Lausanne and Geneva?

B. When are the train the most delayed?

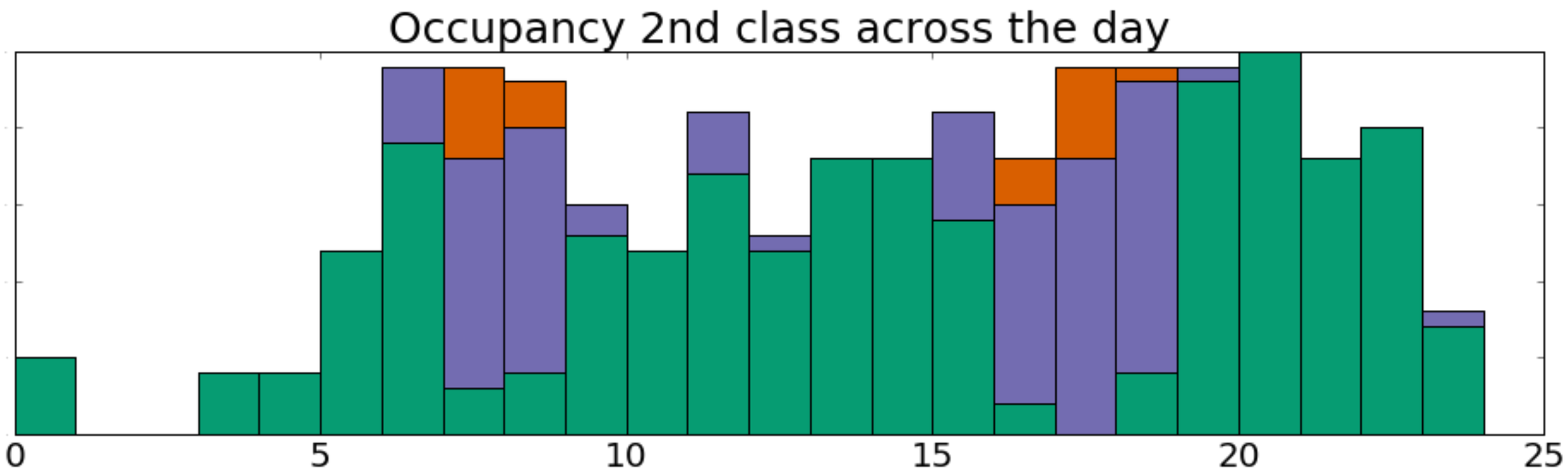
C. Where are the train the most delayed?

A. Lausanne-Genève: when to have a seat?

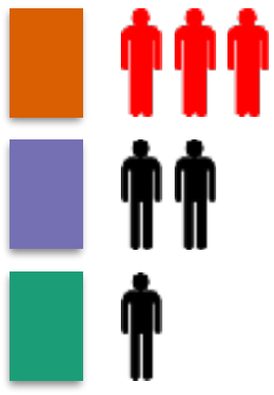
Lausanne-Genève: when to have a seat?

	Gare/Arrêt		Heure	Durée	Change	Voyage avec	Information
	Correspondances du Ma, 18.10.16						
1		Lausanne	dép. 16:51	0:50	0	RE	1. 2.
		Genève	arr. 17:41				
2		Lausanne	dép. 16:58	0:47	1	S 4, ICN	1. 2.
		Genève	arr. 17:45				
3		Lausanne	dép. 17:12	0:36	0	IR	1. 2.
		Genève	arr. 17:48				
4		Lausanne	dép. 17:18	0:42	0	IR	1. 2.
		Genève	arr. 18:00				

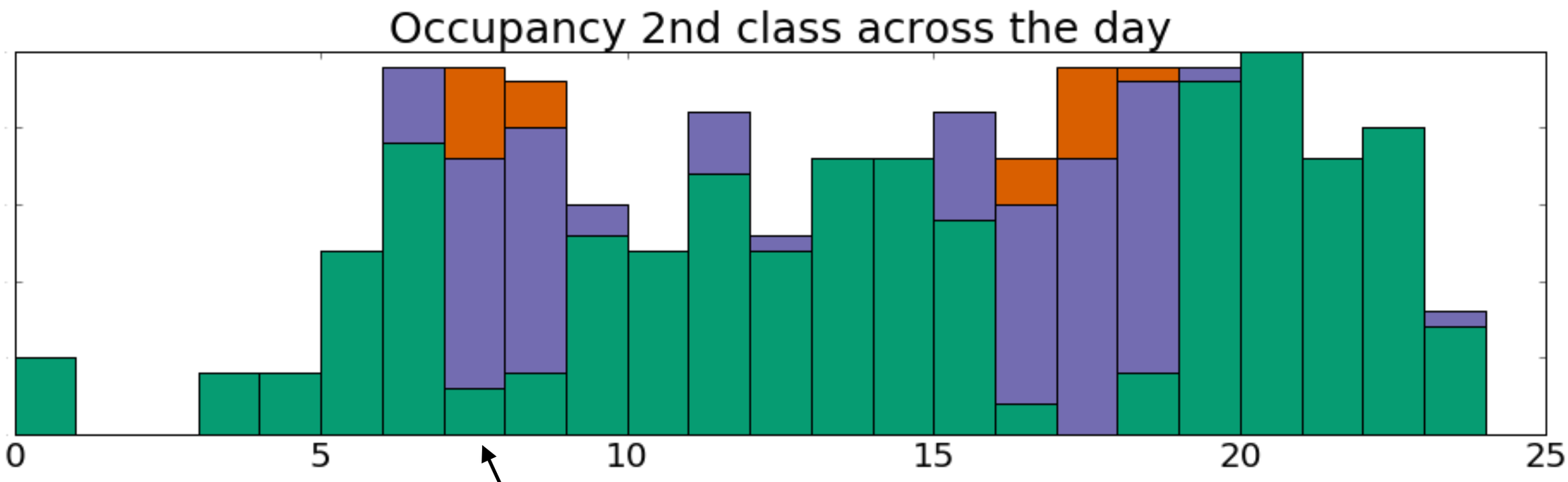
Lausanne-Genève: when to have a seat?



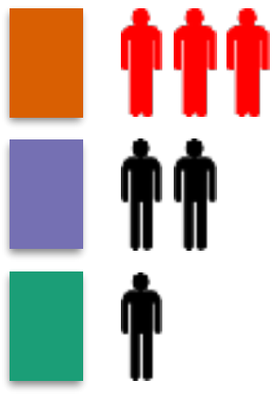
	Gare/Arrêt		Heure	Durée	Change	Voyage avec	Information
Correspondances du Ma, 18.10.16							
1		Lausanne	dép. 16:51	0:50	0	RE	1. 2.
		Genève	arr. 17:41				
2		Lausanne	dép. 16:58	0:47	1	S 4, ICN	1. 2.
		Genève	arr. 17:45				
3		Lausanne	dép. 17:12	0:36	0	IR	1. 2.
		Genève	arr. 17:48				
4		Lausanne	dép. 17:18	0:42	0	IR	1. 2.
		Genève	arr. 18:00				



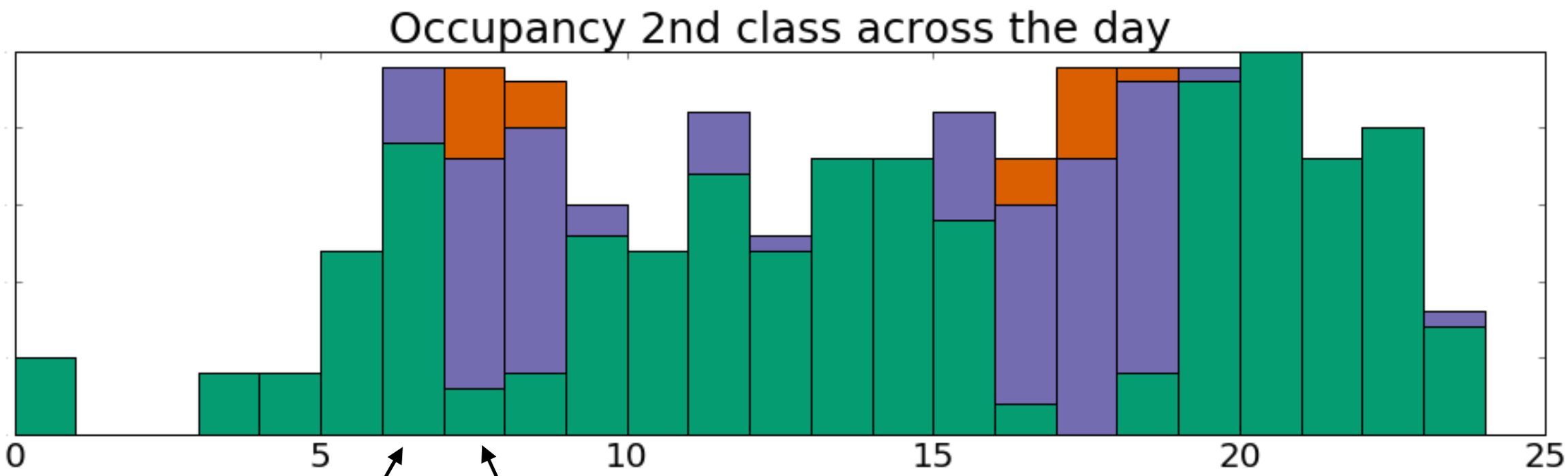
Lausanne-Genève: when to have a seat?



Good luck
in finding a spot!

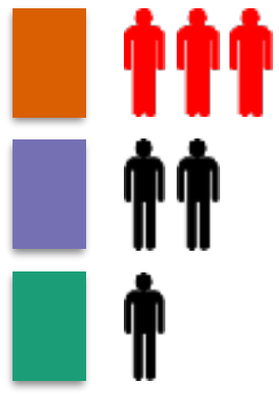


Lausanne-Genève: when to have a seat?

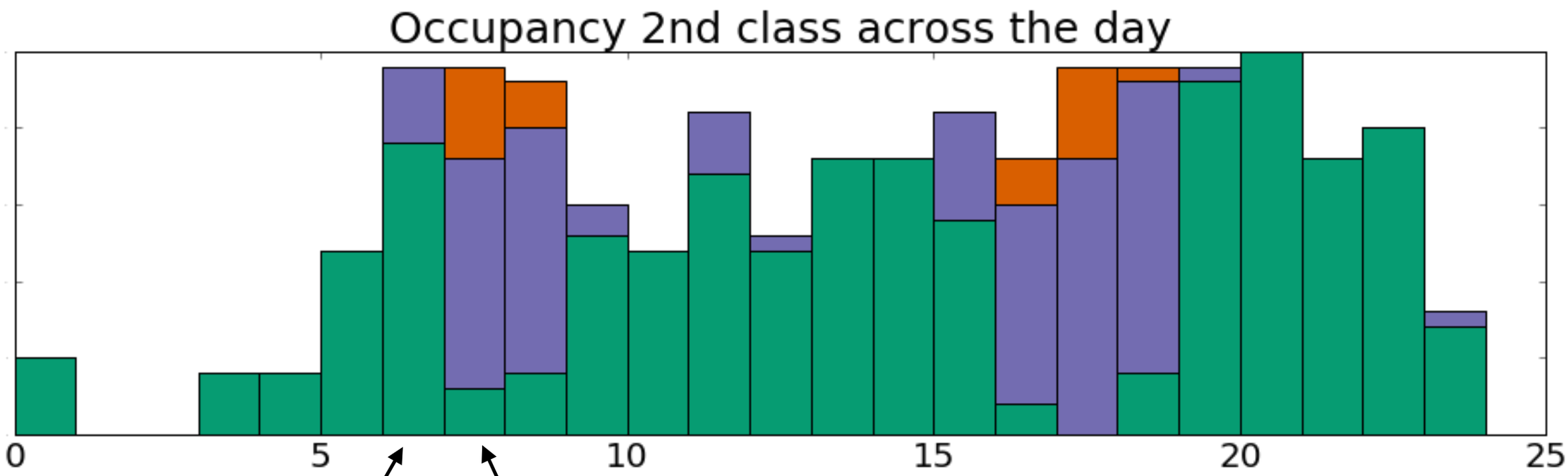


Wake up earlier!

Good luck
in finding a spot!

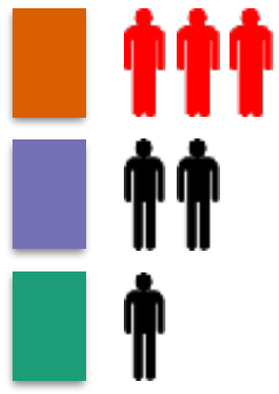


Lausanne-Genève: when to have a seat?



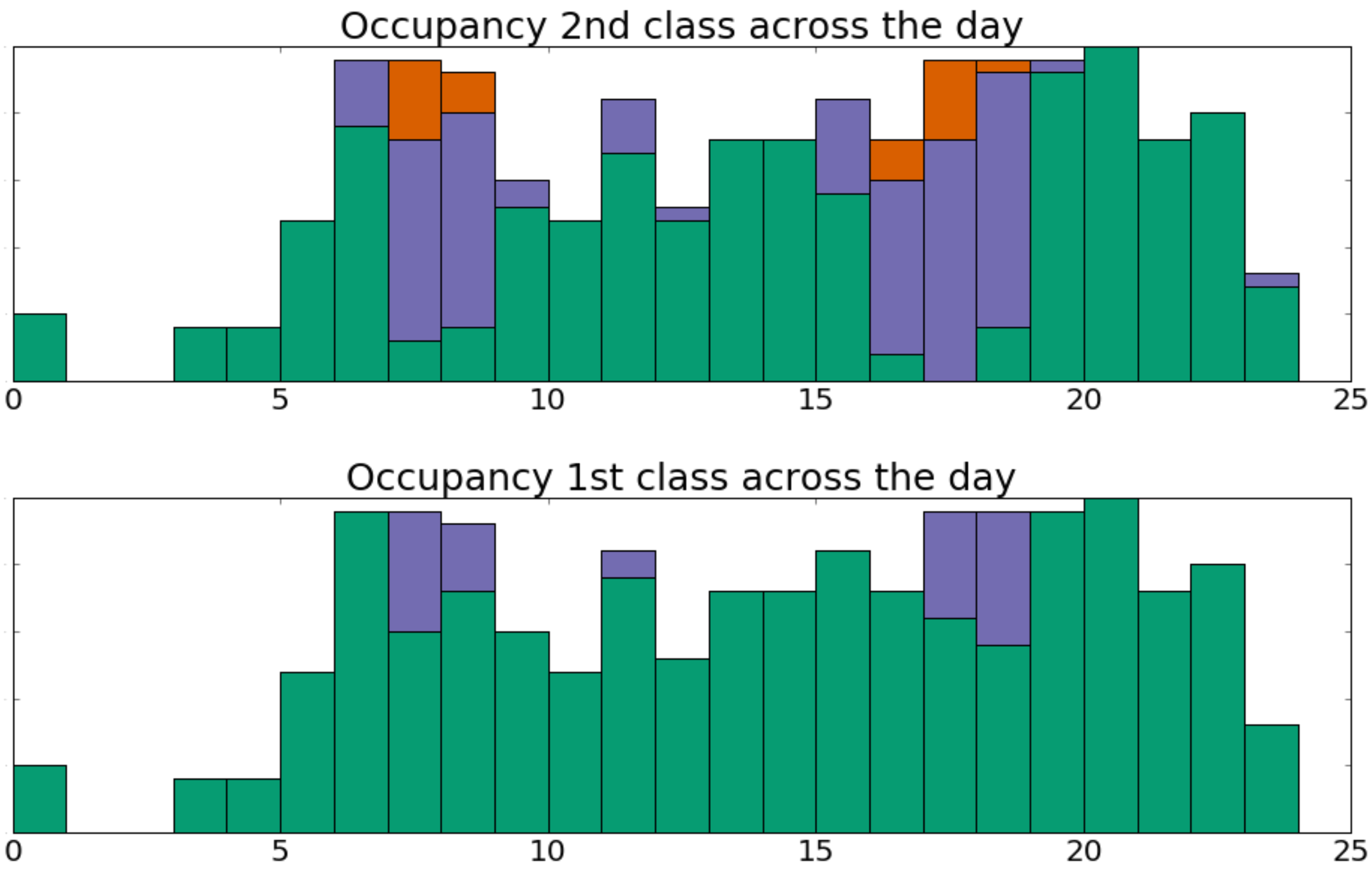
Wake up earlier!

Good luck
in finding a spot!

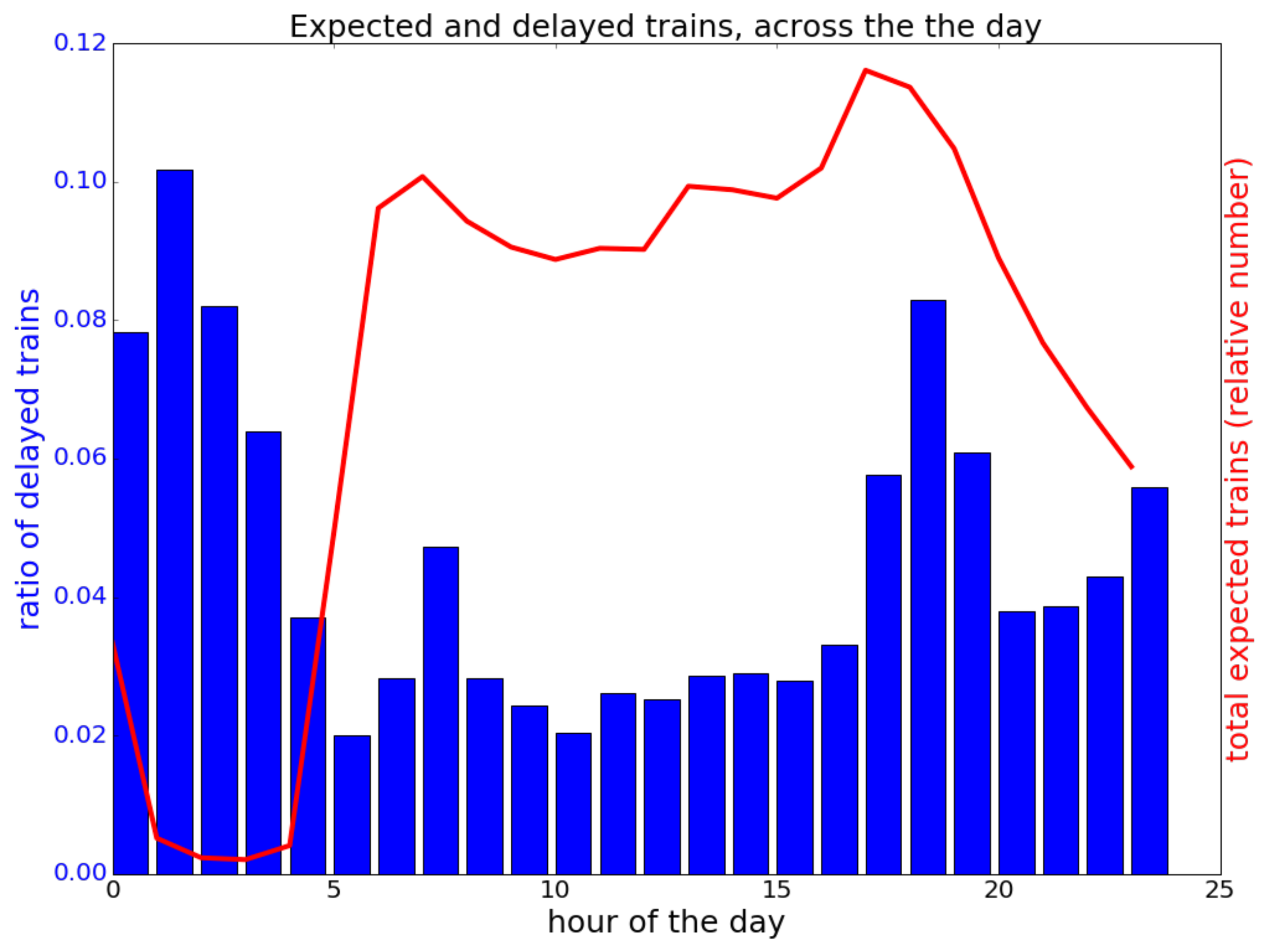


or pay…

Lausanne-Genève: when to have a seat?

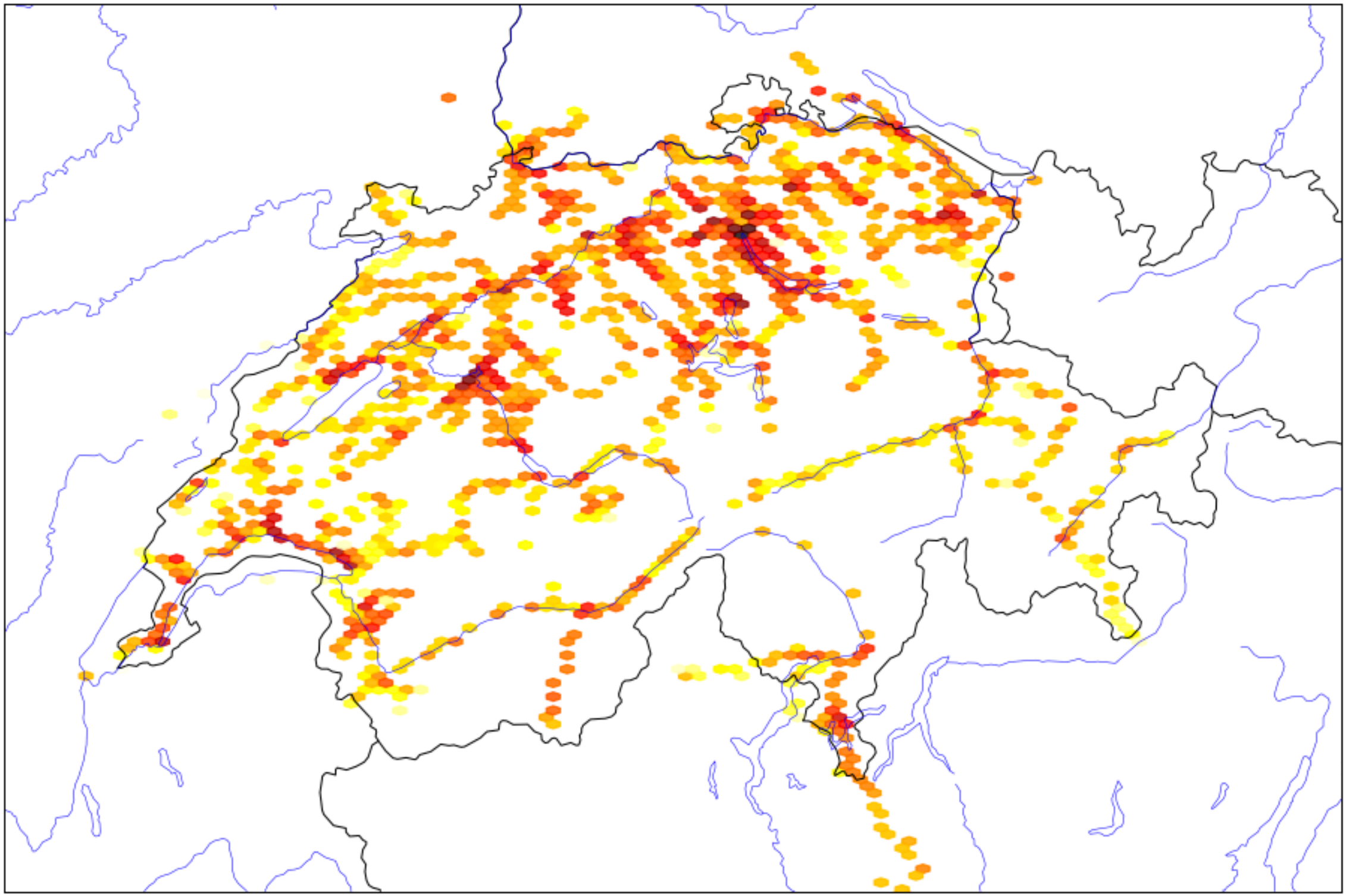


B. When are the trains most delayed?

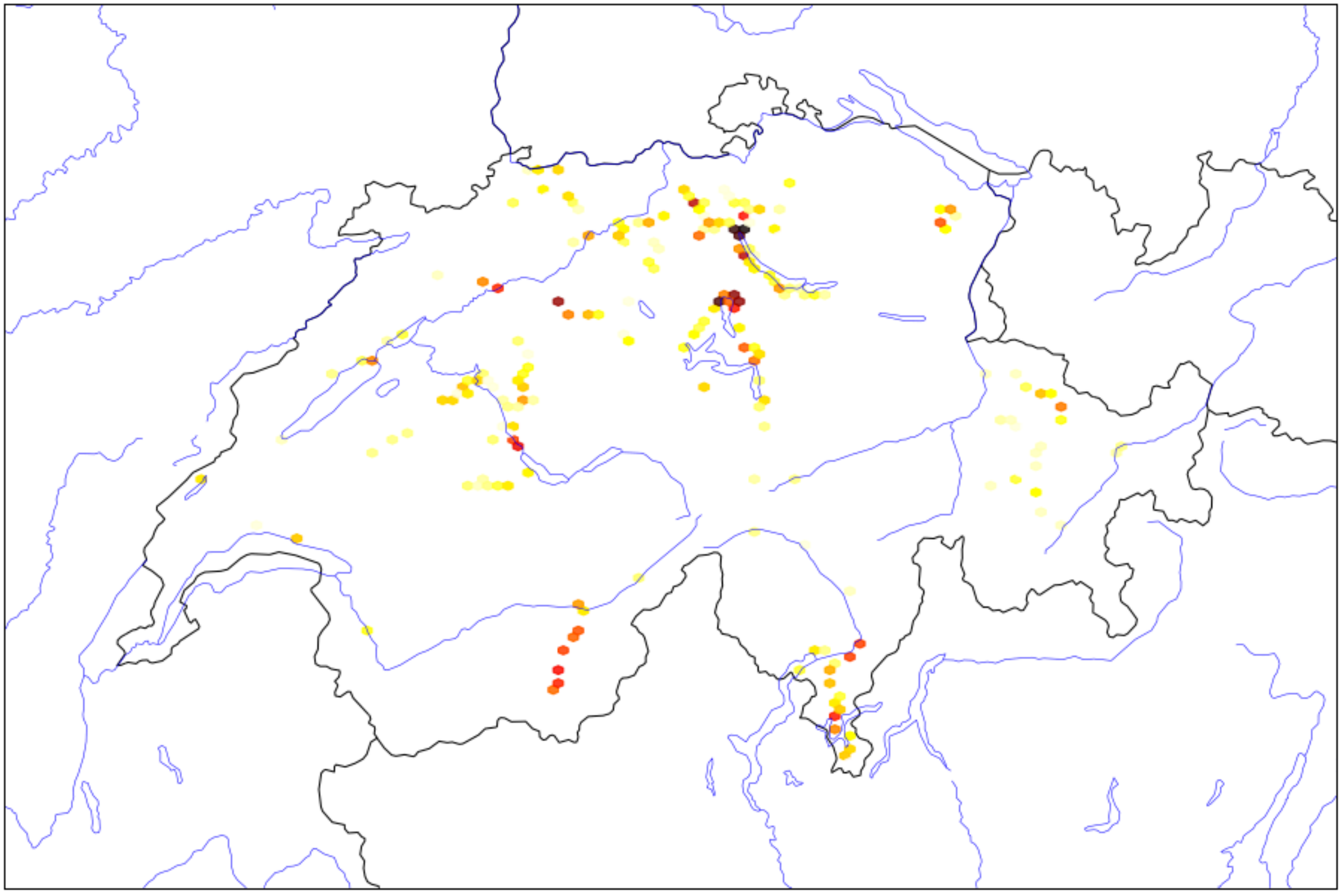

Expected and delayed trains, across the the day
ratio of delayed trains
total expected trains (relative number)
hour of the day
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0
5
10
15
20
25

C. Where are the trains most delayed?

Trains Expected



Trains Delayed



Data analysis tooling…

…or “reproducible science”

An invitation to reproducible computational research

David L. Donoho

– Author Affiliations

Department of Statistics, Stanford University, Stanford, CA 94305, USA, donoho@stanford.edu

a data science notebook

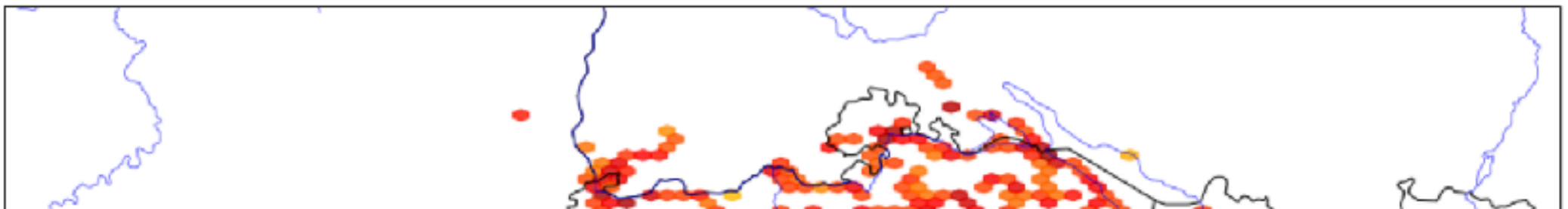
In [2]:
```
parser = lambda date: pd.datetime.strptime(date, '%a %b %d %Y %H:%M:%S GMT+0100 (CET)')
```

In [69]:
```
df_station_boards = pd.read_csv(path_station_boards,
                                sep='\t',
                                nrows=1000,
                                parse_dates=[1,4],
                                date_parser=parser
                               ).fillna(0)
```

In [71]:
```
df_station_boards['delta_mesure'] = df_station_boards.apply(lambda row:(row.timestamp-row.stop_departure).seconds if (r
df_station_boards['stop_departure_hour'] = df_station_boards.stop_departure.apply(lambda t:t.hour)
df_station_boards['is_delayed'] = df_station_boards.stop_delay.apply(lambda d:d>0)
```

In [11]:
```
m=setup_map()
x = list(df_station_boards_last.stop_lon.dropna())
y = list(df_station_boards_last.stop_lat.dropna())
x,y=m(x,y)
hb = m.hexbin(np.array(x),np.array(y),gridsize=130, mincnt=10, cmap='hot_r', alpha=0.8, bins='log')

plt.show()
```

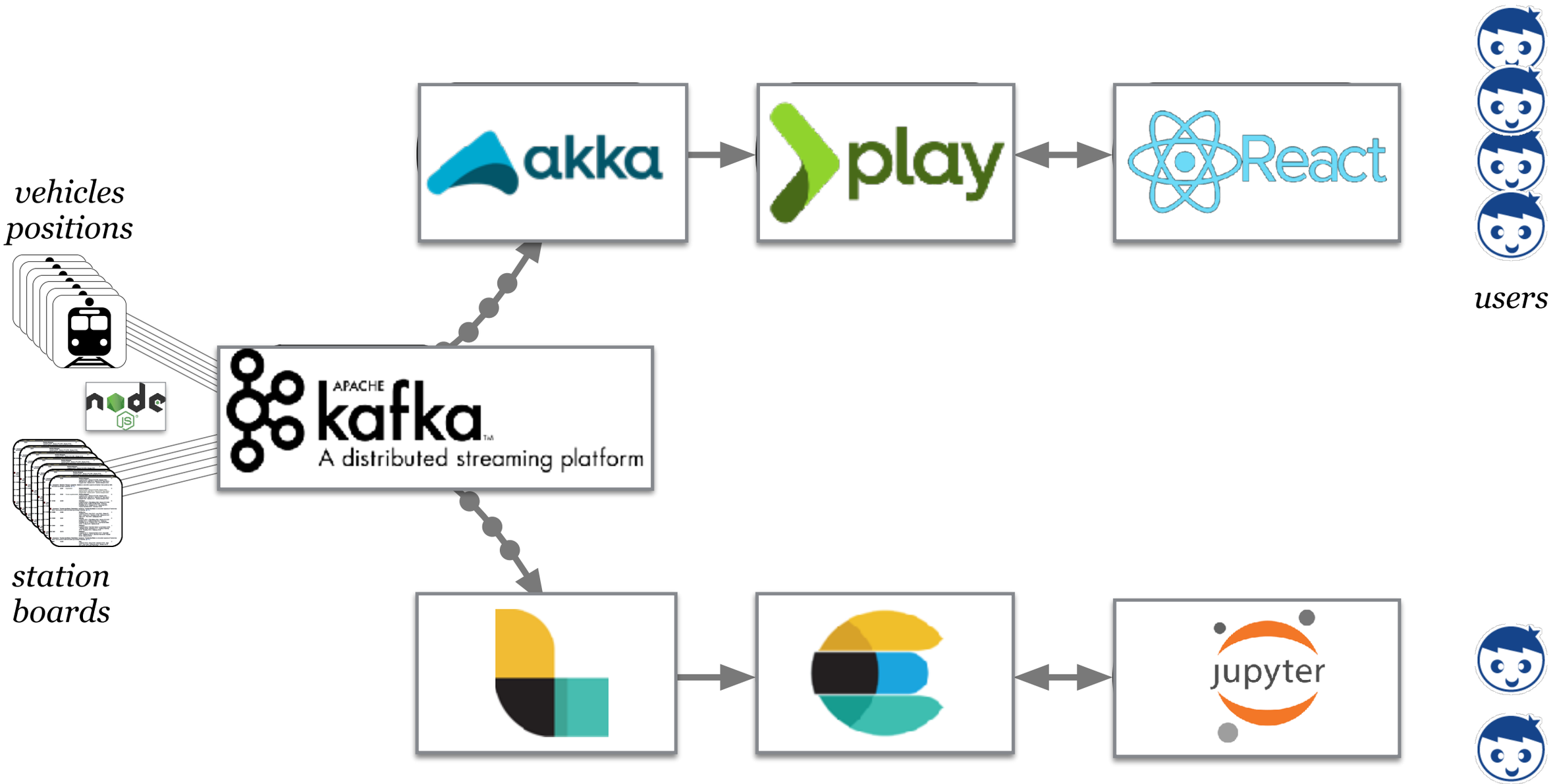


a data science notebook

- Web application
- Interactively edit and run pieces of code (analysis steps)
- Inclined towards Python (although other languages are available)
- Beware of performance with large dataset (sample data or use Spark mode)

TIMTOWTDI

Jupyter, Zeppelin, RStudio…



https://github.com/alexmasselot/swiss-transport-realtime

This is only a POC!!!

vehicles positions

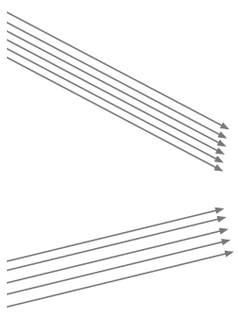
station boards

akka → play ↔ React → *users*

kafka (A distributed streaming platform)

jupyter → *data analysts*

https://github.com/alexmasselot/swiss-transport-realtime



users

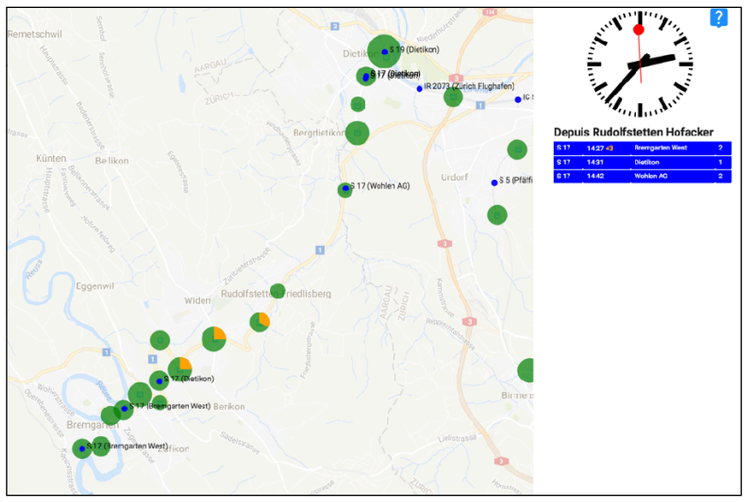
data analysts

Meetup

Big Data Romandie

Nov 8th 7 pm, Genève

"Banknote Recognition System"
(Machine Learning)



Demo on OCTO stand

Nov 10th 6 pm, Genève

"Data Science & Machine Learning:
Explorer, Comprendre Et Prédire"

amasselot@octo.com

@alex_mass