Interactive Data Exploration in the Browser

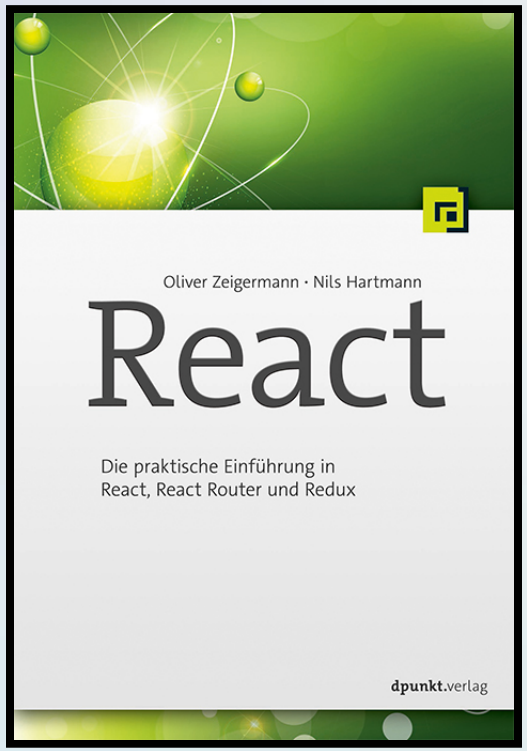
data2day 2016

Oliver Zeigermann / @DJCordhose

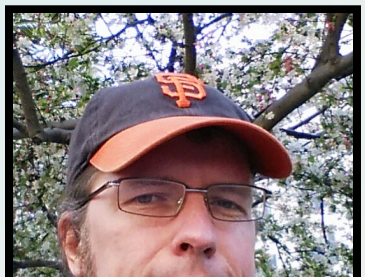
http://bit.ly/data2day-explore


entwickler.press
3.
aktualisierte und erweiterte Auflage
JavaScript für Java-Entwickler
Oliver Zeigermann
schnell + kompakt
Mit ECMAScript 2015 und TypeScript


entwickler.press
JavaScript und TypeScript für C#-Entwickler
Oliver Zeigermann
schnell + kompakt


Oliver Zeigermann · Nils Hartmann
React
Die praktische Einführung in React, React Router und Redux
dpunkt.verlag


embarc



Patrick Tehubijuluw
@Creatuluw

Following

A #dataviz many times is like a report. A static version of #data. To bring out an optimal insight, you'll need interaction and curiosity.

RETWEETS	LIKES
3	6

9:12 AM - 30 Aug 2016

3 6

Try showing your #data from another perspective with #dataviz

@Creatuluw

But why in the Browser

1: Zero-Installation for the user

2: Interactivity coming naturally

Example Project:
Exploring *all* domestic US flights for 2001

http://stat-computing.org/dataexpo/2009/

The raw data

approx. 6 million data sets

```
> wc -l 2001.csv
5967781 2001.csv
```

approx. 600 MB of data

```
> ls -hl 2001.csv
573M Jan 10  2016 2001.csv
```

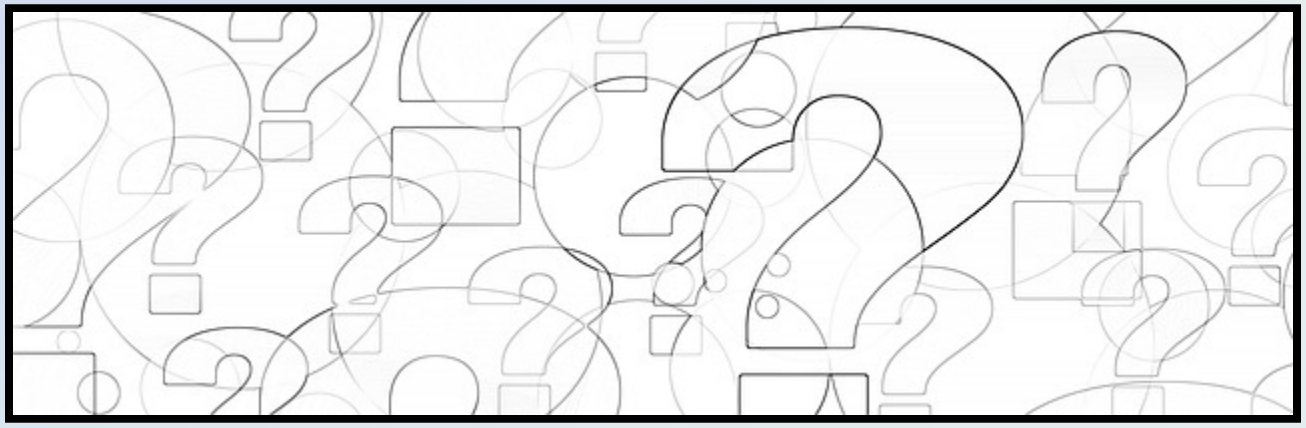
29 colums, data has gaps, but looks consistent

```
> head 2001.csv
Year,Month,DayofMonth,DayOfWeek,DepTime,CRSDepTime,ArrTime,CRSArrTime,Uni
2001,1,17,3,1806,1810,1931,1934,US,375,N700??,85,84,60,-3,-4,BWI,CLT,36
2001,1,18,4,1805,1810,1938,1934,US,375,N713??,93,84,64,4,-5,BWI,CLT,361
2001,1,19,5,1821,1810,1957,1934,US,375,N702??,96,84,80,23,11,BWI,CLT,36
2001,1,20,6,1807,1810,1944,1934,US,375,N701??,97,84,66,10,-3,BWI,CLT,36
```

No specific task or question

Exploring what just might be interesting

Finding the unknown unknowns



Options

1. Google Sheets
2. Elasticsearch / Kibana (ELK)
3. D3 with crossfilter and DC
4. D3 loading Segments from ELK

I: Google Sheets

Import

Up to 2 million cells

We have 9 * 400,000: too many for Google Sheets

```
> cut -f2,3,4,5,6,7,8,9 -d, 09.csv >09_no_month.csv
> awk -F, '$1 >= 10 && $1 <= 15' 09_no_month.csv > 09_very_small.csv
```

```
> ls -lh 09_very_small.csv
-rw-r--r--  1 olli  staff   1.0M Jul 21 21:46 09_very_small.csv
```

```
> wc -l 09_very_small.csv
38272 09_very_small.csv
```

Filters

DayofMonth	UniqueCarrier	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance
10	US			4	MCO	CLT	468
10	US			-2	PIT	DCA	205
10	US			-4	CLT	ORD	599
10	US			-6	MYR	CLT	156
10	US			-5	CLT	CHS	168
10	US			-7	DFW	CLT	936
10	US			-2	CLT	IAH	913
10	US			0	RDU	CLT	130
10	US			90	LGA	GSO	461
10	US			-5	DCA	TPA	814
10	US			-3	SYR	DCA	298
10	US			132	CLE	PHL	363
10	US			-4	DCA	TPA	814
10	US			-4	SYR	DCA	298
10	US			-1	CLT	AVL	92
10	US			-7	FLL	CLT	631
10	US			1	CLT	DFW	936
10	US			-3	CMH	CLT	346
10	US	158	20	-6	FLL	PHL	992
10	US	62	-5	-6	PIT	CLT	366
10	US	20	-11	-6	GSO	CLT	83
10	US	121	-22	-5	PHL	TPA	920
10	US	66	-1	-6	ORF	PIT	330

Sort A → Z

Sort Z → A

Filter by condition...

Filter by values...

Select all - Clear

AS

CO

DL

HP

MQ

OK

Cancel

Filters: All flights between JFK and STL

Physically change the data

	A	B	C	D	E	F	G	H
1	DayofMonth	UniqueCarrier	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance
5211	10	TW	119	-12	-10	JFK	STL	892
5337	10	TW	120	188	-1	JFK	STL	892
5529	10	TW	145	162	0	JFK	STL	892
5602	10	TW	125	113	61	JFK	STL	892
17748	11	TW	NA	NA	-5	JFK	STL	892
31944	15	TW	124	23	41	JFK	STL	892
32054	15	TW	131	9	32	JFK	STL	892
32214	15	TW	128	-2	11	JFK	STL	892
32268	15	TW	121	201	230	JFK	STL	892

Filters Views

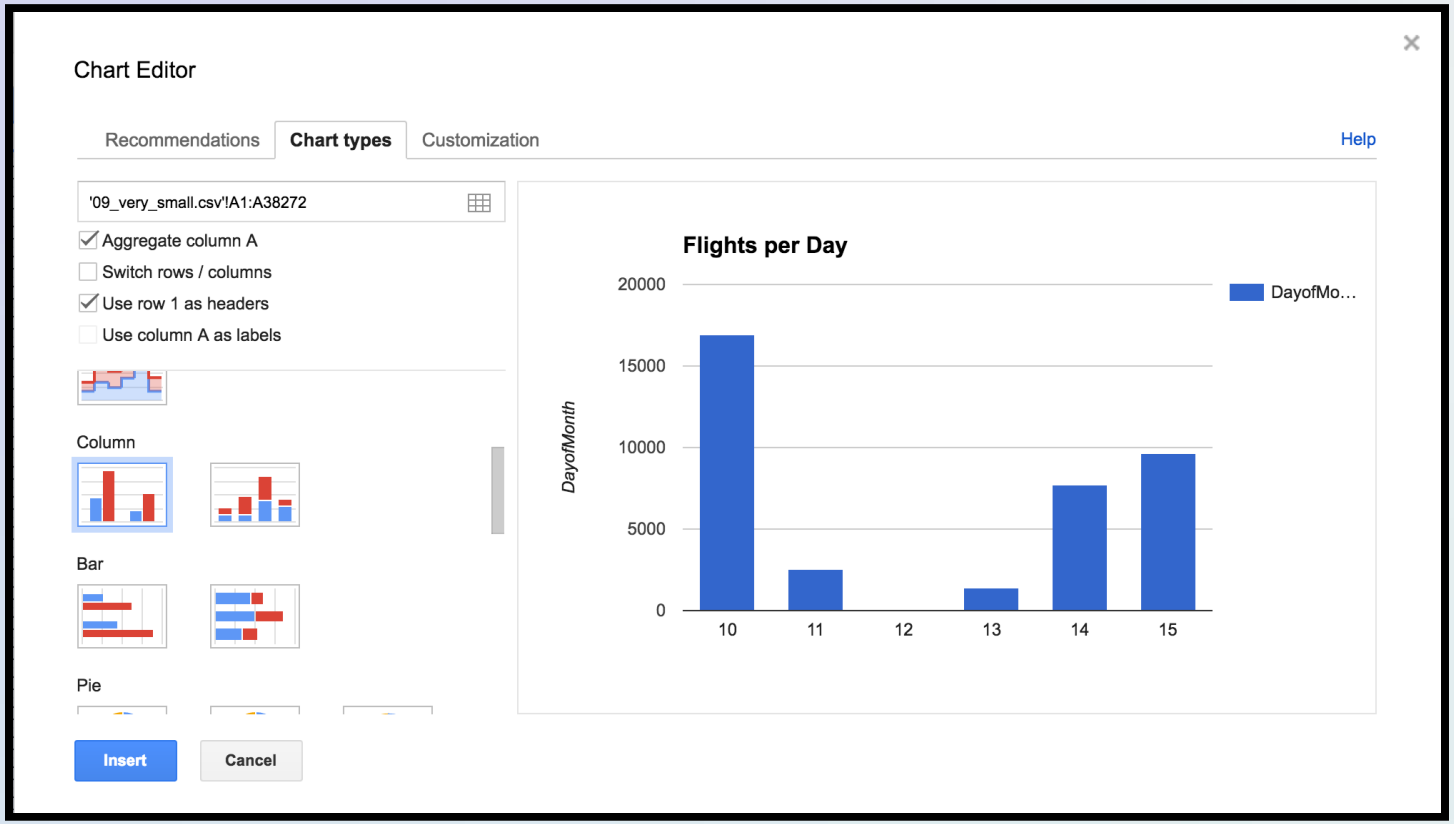
Just a view on the data, more than one possible

Departure Delay > 200 minutes

Name: DepDeplay gt 200 **Range:** A1:H38272

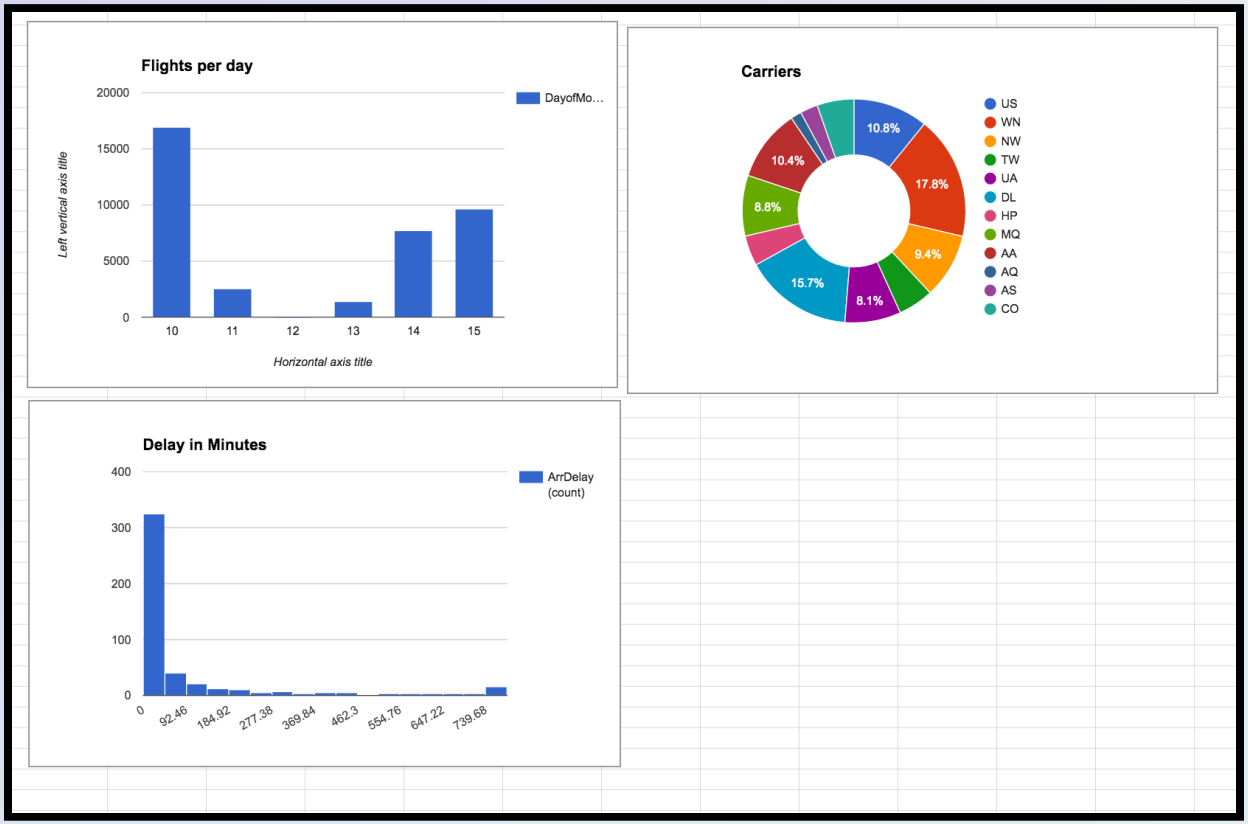
	A	B	C	D	E	F	G	H
1	DayofMonth	UniqueCarrier	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance
2	14	AA	203	1109	1114	SJU	PHL	1576
3	15	AA	158	1018	1032	BOS	MIA	1258
4	15	AA	212	991	991	SJU	BWI	1565
5	14	NW	22	857	871	CLE	DTW	95
6	14	NW	126	832	847	DEN	DTW	1123
7	10	AA	124	662	667	MCO	DFW	984
8	13	DL	211	621	640	LAX	ATL	1946
9	10	AQ	29	606	603	KOA	HNL	163
10	13	NW	113	602	597	SLC	MSP	991
11	13	NW	32	594	587	CMH	DTW	155
12	15	AA	124	564	579	SJU	MIA	1045

Diagram



How many flights per day?

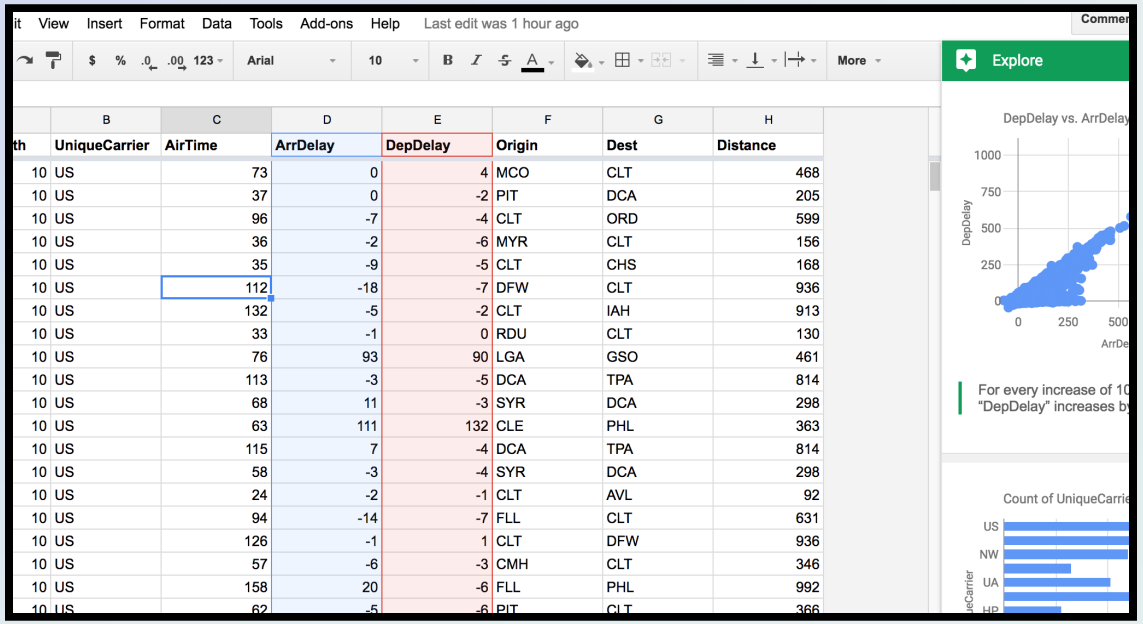
Mini Dashboard



flights per day / carriers / delay

Explore

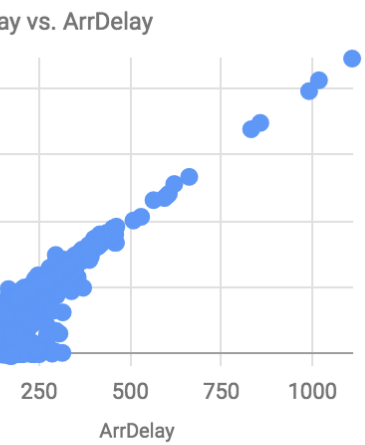
The best is yet to come

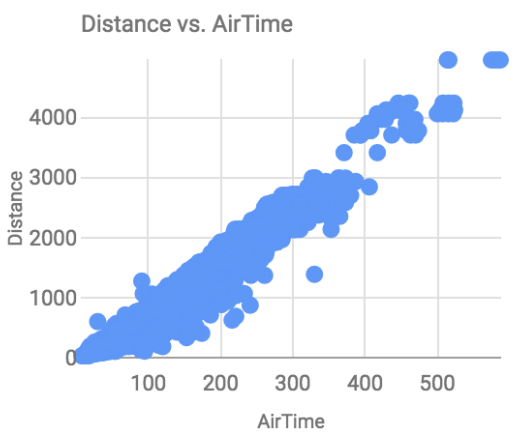


th	UniqueCarrier	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance
10	US	73	0	4	MCO	CLT	468
10	US	37	0	-2	PIT	DCA	205
10	US	96	-7	-4	CLT	ORD	599
10	US	36	-2	-6	MYR	CLT	156
10	US	35	-9	-5	CLT	CHS	168
10	US	112	-18	-7	DFW	CLT	936
10	US	132	-5	-2	CLT	IAH	913
10	US	33	-1	0	RDU	CLT	130
10	US	76	93	90	LGA	GSO	461
10	US	113	-3	-5	DCA	TPA	814
10	US	68	11	-3	SYR	DCA	298
10	US	63	111	132	CLE	PHL	363
10	US	115	7	-4	DCA	TPA	814
10	US	58	-3	-4	SYR	DCA	298
10	US	24	-2	-1	CLT	AVL	92
10	US	94	-14	-7	FLL	CLT	631
10	US	126	-1	1	CLT	DFW	936
10	US	57	-6	-3	CMH	CLT	346
10	US	158	20	-6	FLL	PHL	992
10	US	62	-5	-6	PIT	CLT	366

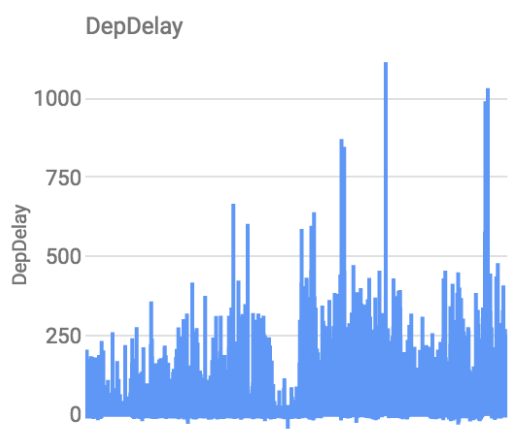
Correlations, Outliers, etc.

Sheets can automatically find this

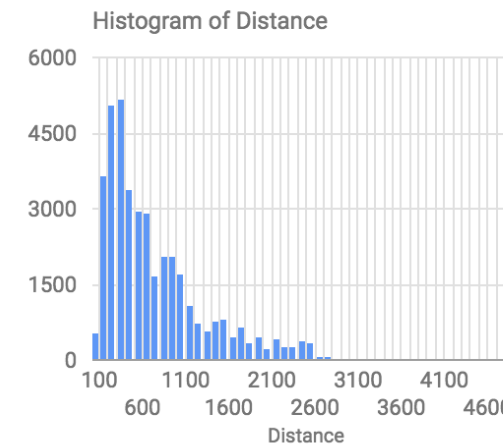




For every increase of 100 in “AirTime”,



Outlying values for “DepDelay”: peaks at



Ranges from 31 to 4962, with most value

What I have learned

Without asking

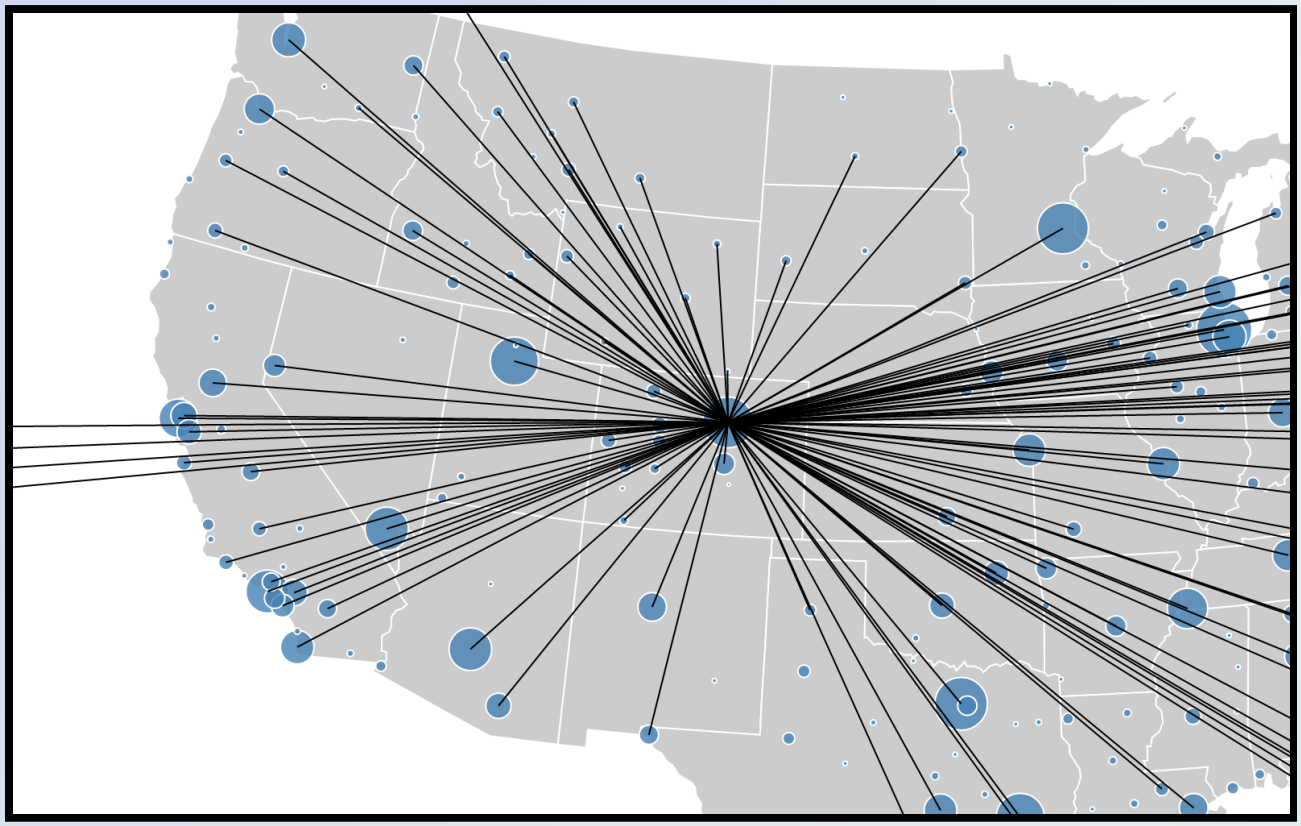
- long arrival delays are caused by departure delays
- per 100 minutes you can fly around 857 miles
- some flights where delayed more than 1000 minutes (more than 16 hours)
- there seems to be a flight distance of 4962 miles in the US (around 10 hours of flight)

II: D3 with crossfilter and DC

Things get more interactive and connected

Using full data set of September (10x the data)

D3.js: Dynamic graphics in the browser



Data based vector graphics (SVG)

http://d3js.org/

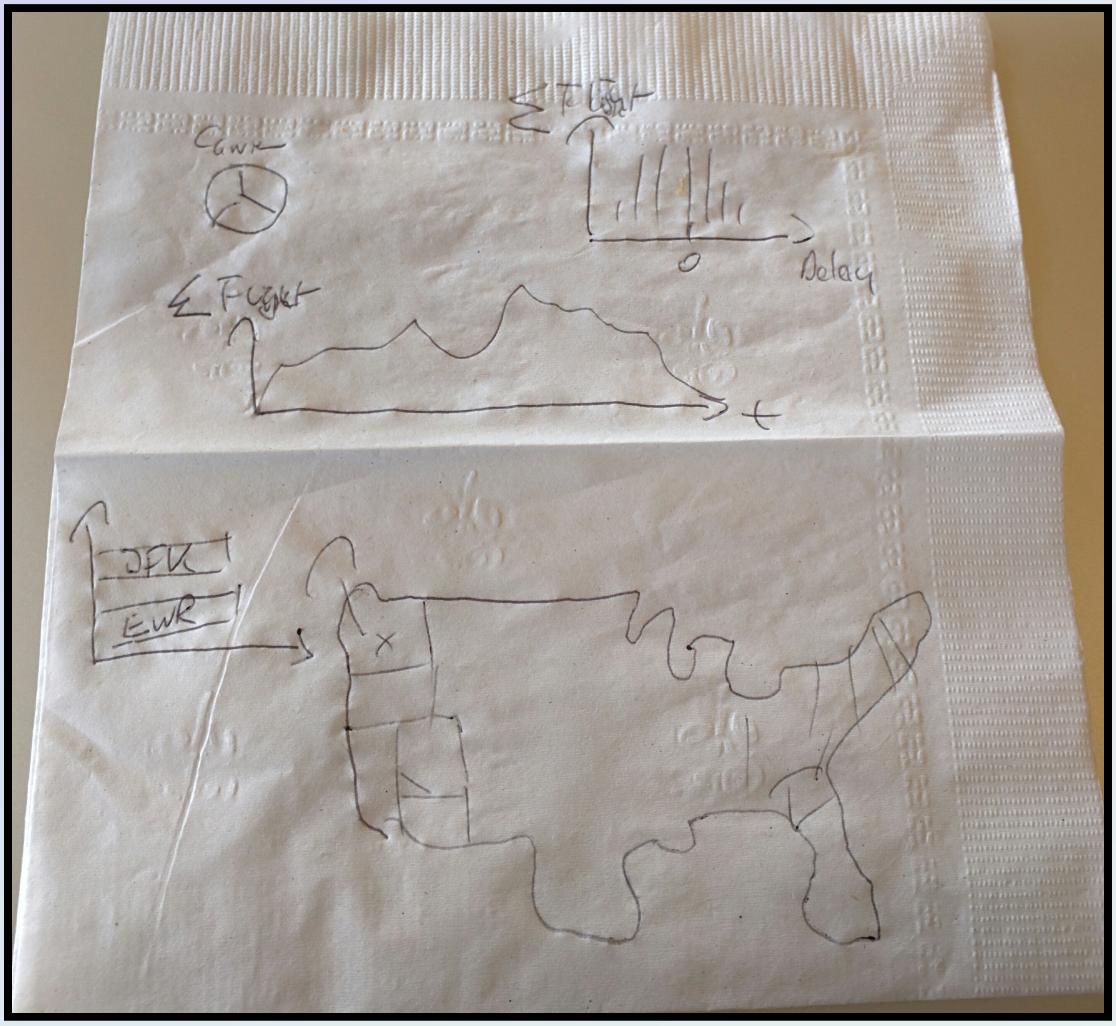
Crossfilter: Filtering millions of data sets in real time

- http://square.github.io/crossfilter/
- Can filter up to millions of data sets in real time in the browser
- Data sets are indexed when loaded
- You specify what you want to filter using dimensions

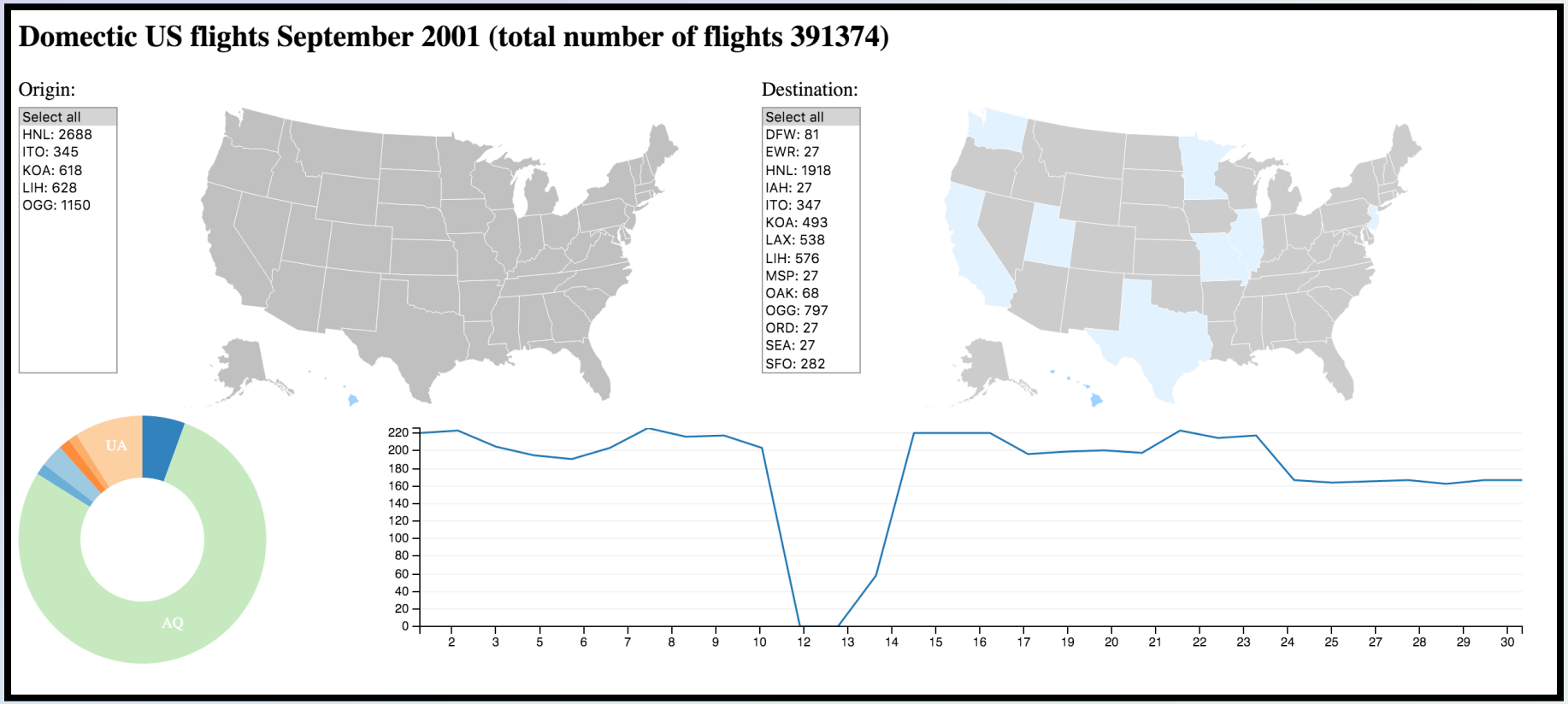
dc.js: Charting with D3 and Crossfilter

- http://dc-js.github.io/dc.js/
- Prebuilt integration of D3 and Crossfilter
- offers a couple of diagram types that are useful for interactive big data
  - bar
  - pie
  - many more

Initial Design



Implementation using a static dump of September



Run (have a little bit patience)

III: ELK

Things scale

Using all of 2001 (once more 10x the data)

Data needs to be (physically) close to the interaction to make it fast and thus most useful

ELK Stack

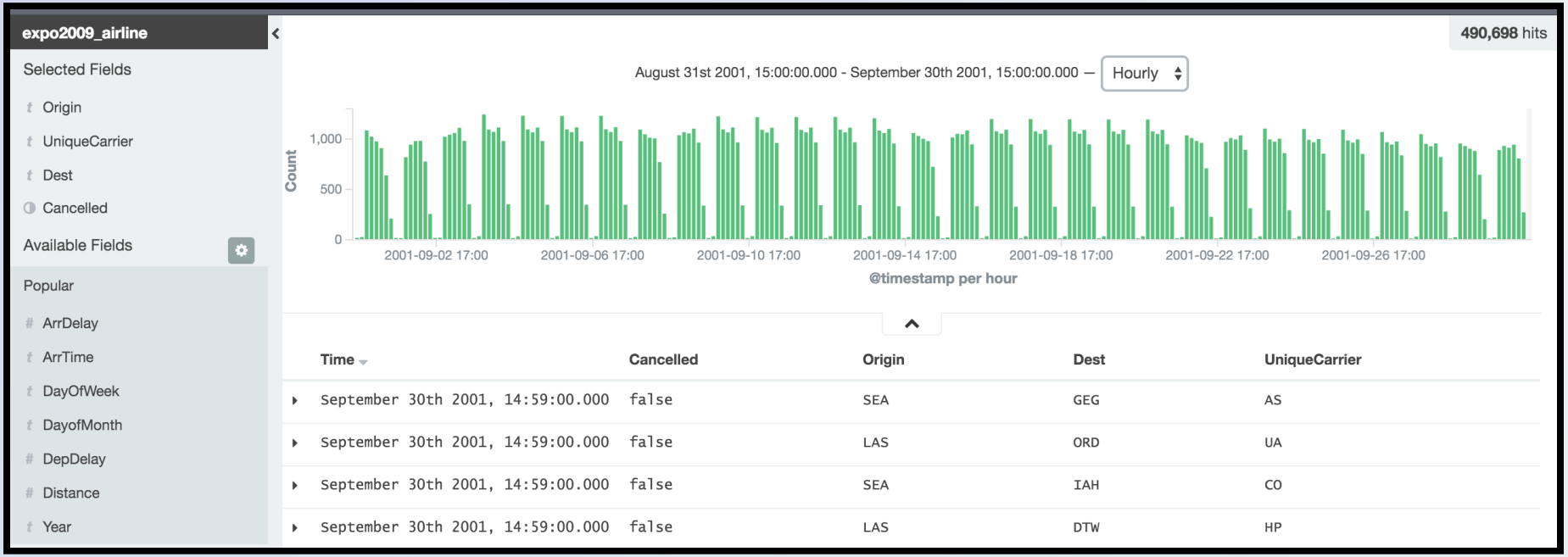
- Elasticsearch
  - Search and Analytics Engine
  - Indexes and stores structured or unstructured data
  - Offers query language to search or aggregate data
- Logstash
  - process data and store into Elasticsearch
  - Ruby based import description
- Kibana
  - interactive querying
  - visualization (in dashboards)

Kibana

- generic frontend for Elasticsearch
- browser based
- allows for dashboards
- also allows to make arbitrary adhoc queries

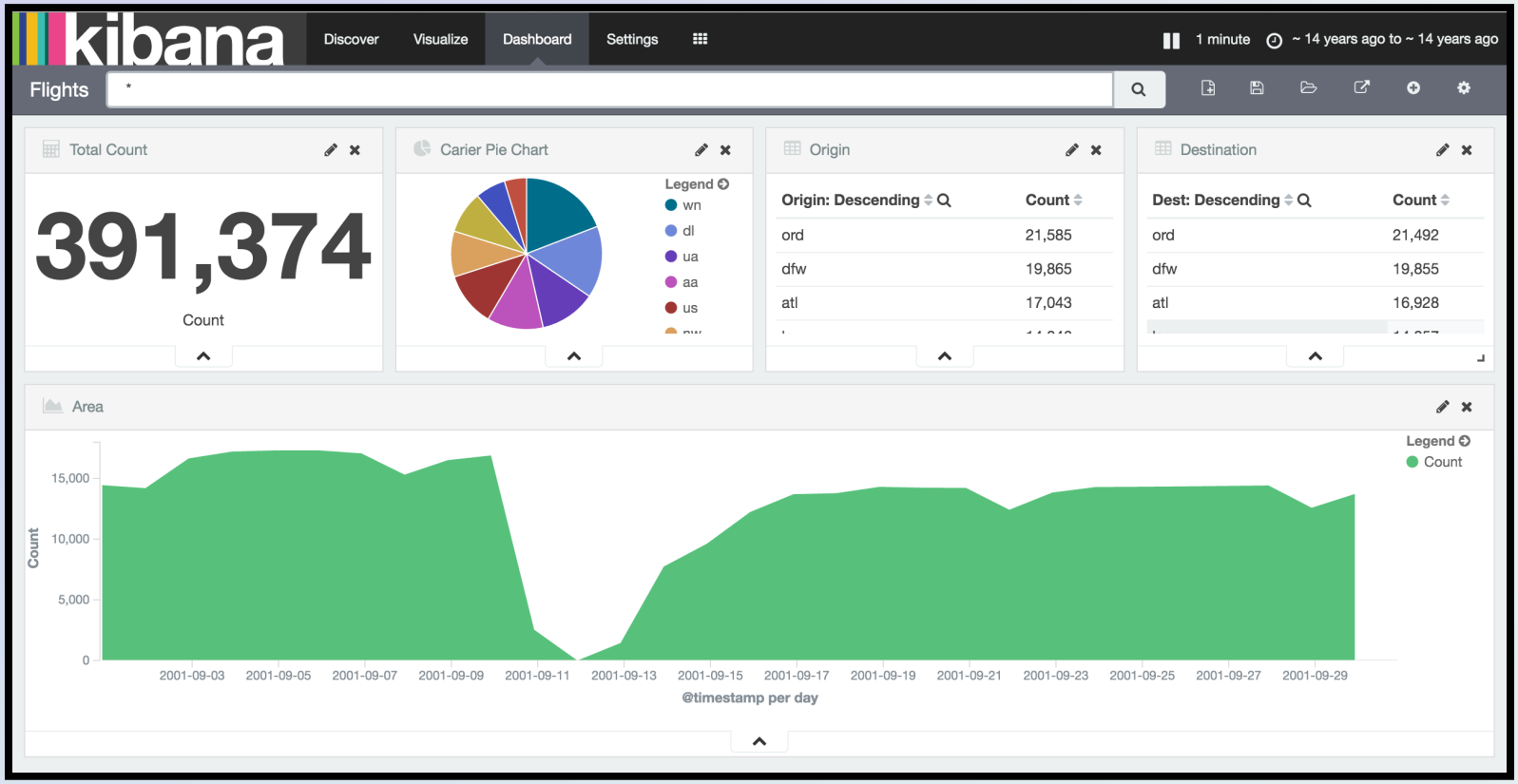
Discover Data

using Kibana adhoc queries



Demo: filter for '*Cancelled:false*'

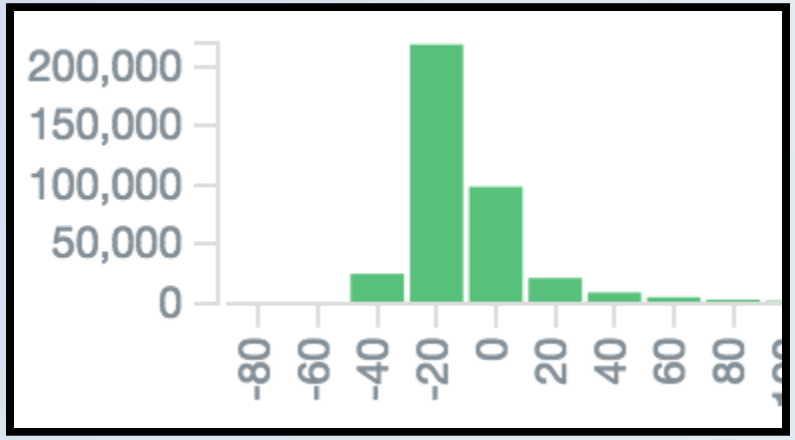
Flights Dashboard #1



Clicks trigger requests, responses update graphics

Demo

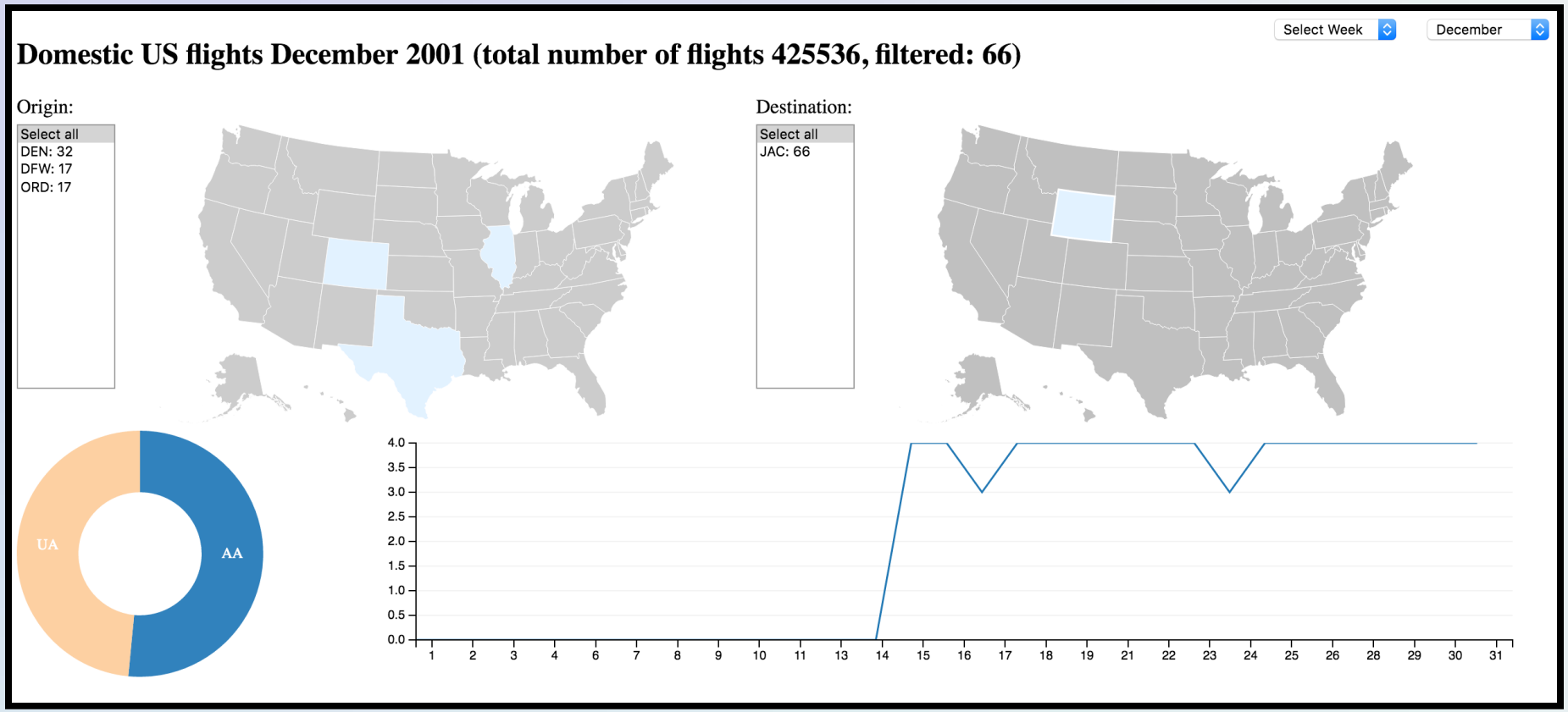
Adding a Departure Delay Barchart



IV: D3 loading Segments from ELK

- Elasticsearch delivers data over HTTP/JSON using query
- D3/Crossfiler/DC display partion of data
- Partition could be month
- Manually selected by user

DC/D3/Crossfilter Dashboard loading segments from Elasticsearch



Final Matrix

	Google Sheets	D3	ELK	Segmented D3	iPython
One tool for designer / user					
Required effort					
Easy to create new dashboard					
Interactivity					
Unlimited data size					
Offline					
Unrestricted widgets					
Auto Refresh					
Repl		Console	Query	(Console)	
Turn Query into Dashboard					
Add your category here					

Thank you

Questions / Discussion

Code for all examples: https://github.com/DJCordhose/big-data-visualization/code

Slides: http://bit.ly/data2day-explore

Oliver Zeigermann / @DJCordhose