Value in the DETAILS

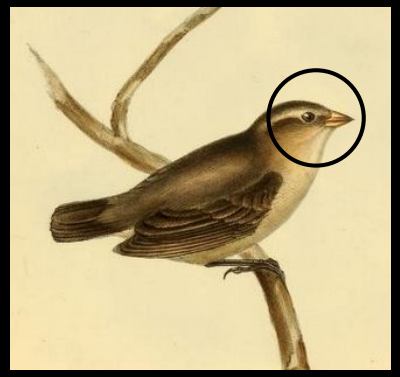
Understanding detailed data through VISUAL EXPLORATION

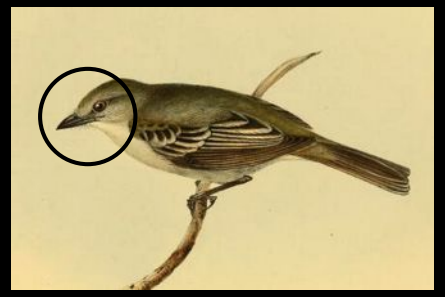
Richard Brath
Rob Harper

uncharted

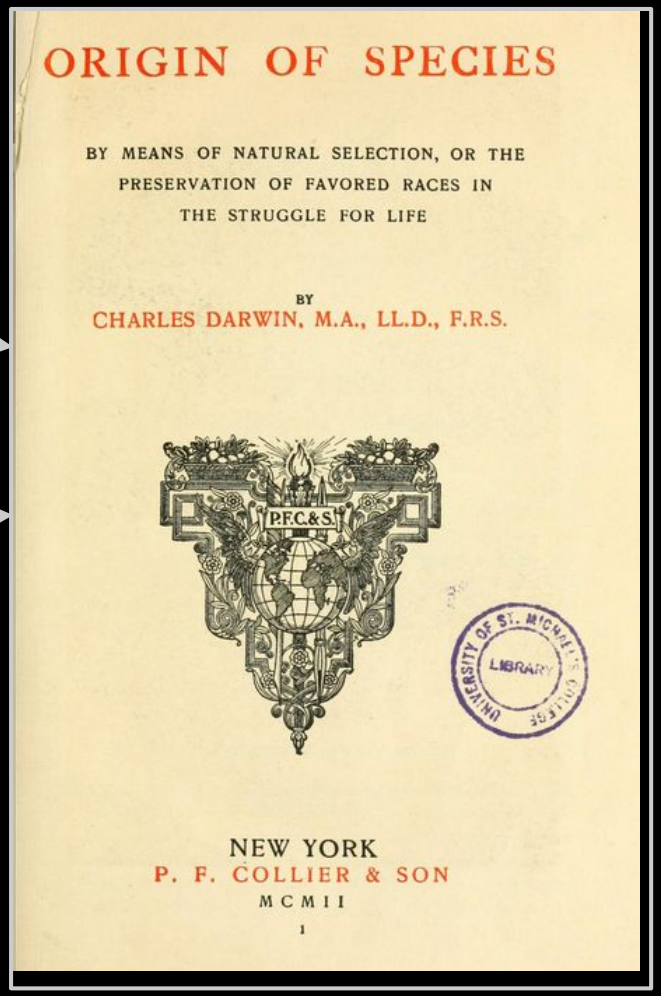

I'll be visually exploring all kinds of interesting patterns in tweets about Trump.
But wait – first what do we mean by visual exploration?

Visual Exploration





Natural Selection?



uncharted

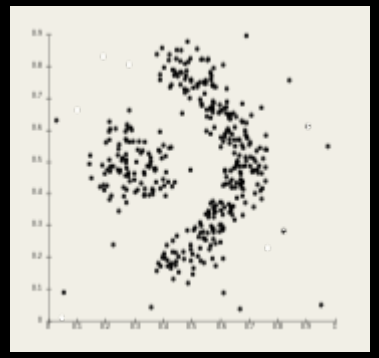
Visual Exploration Process

1. Collect a lot of data
2. Observe some interesting patterns
3. Hypothesize about why they exist
4. Refine and build models

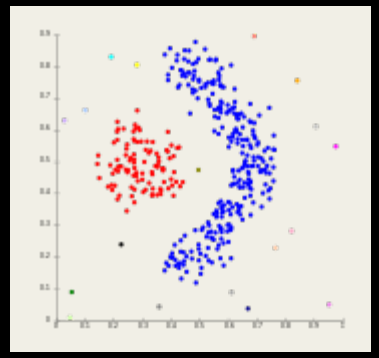
Understanding detail data

Can you see the patterns?

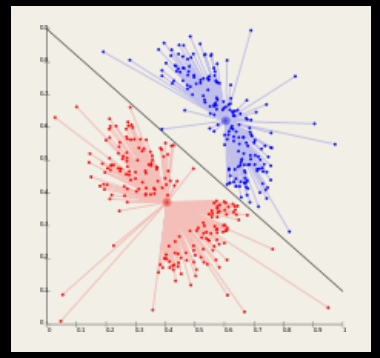
How about these algorithms abilities to find the patterns?



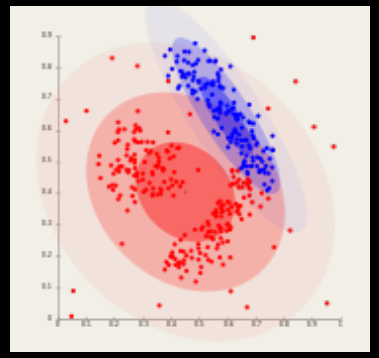
No Clustering



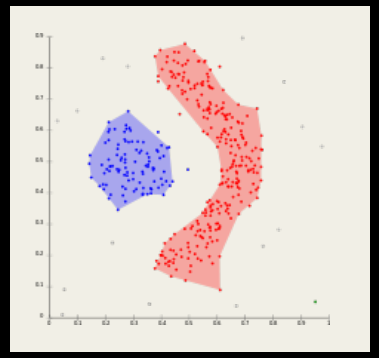
Linkage Clustering



K-Means Clustering



Distribution-based Clustering



Density-based Clustering

https://en.wikipedia.org/wiki/Cluster_analysis

We see patterns all the time, and quite easily

We tend to group things based on visual cues, such as proximity, alignment and containment.

https://www.flickr.com/photos/tobin/111811243/

uncharted

Seeing detailed patterns

Perception can be whole. Once you see it, you won't un-see it.

First publication of the picture probably in Life Magazine:58;7 1965-02-19, p 120.

Also, check out the movie by Wim van de Grind (http://www.michaelbach.de/ot/cog_dalmatian/index.html)

So what?

Powerful human perception system

- Detects **patterns** in complex data
- Can find **patterns** based on different criteria
- Can find **patterns** at different scales

So what?

Exploratory Data Analysis stems from John Tukey's work in the early 1960s. EDA can be characterized by

a. understanding "what is going on here?"
b. graphic representations of data
c. tentative model building and hypothesis generation
d. robust measures, re-expression, and subset analysis
e. skepticism, flexibility.

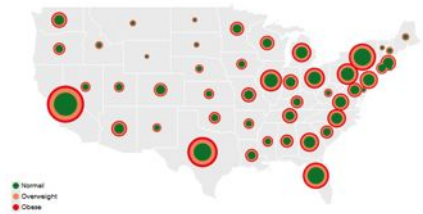
*The goal of Exploratory Data Analysis is to discover **patterns** in data.*

paraphrased from John Behrens, Arizona State University, Principles and Procedures of Exploratory Data Analysis, American Psychological Association, 1997.

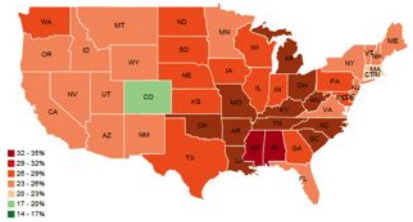
So, why are we summarizing big data into bar charts?

So, why we rolling-up big data into visualizations of 1000 points?

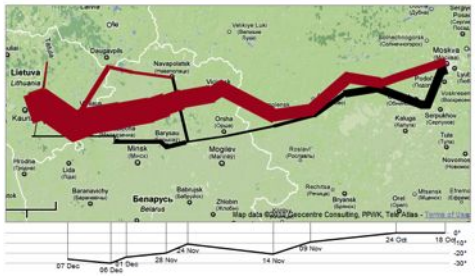
Graduated Symbol Maps



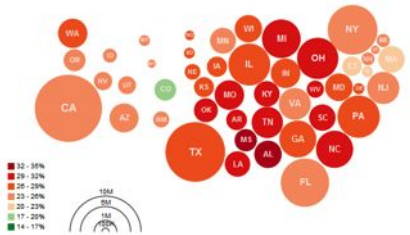
Choropleth Maps



Flow Maps

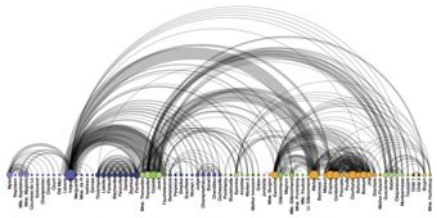


Cartograms



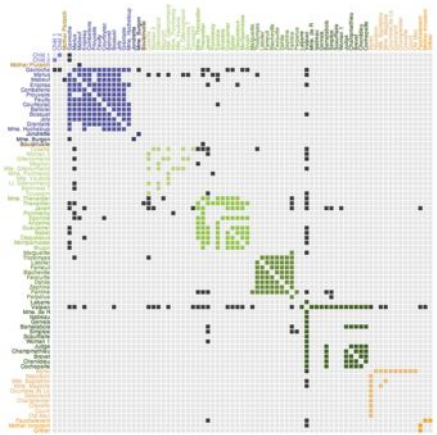
Force-Directed Layout

Arc Diagrams

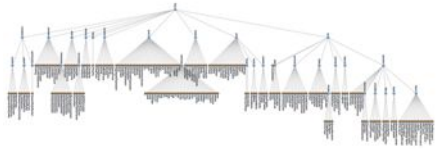


An *arc diagram* uses a one-dimensional layout of nodes, with circular arcs to represent links. While arc diagrams may not convey the overall structure of the graph as effectively as a two-dimensional layout, with a good ordering of nodes it is easy to identify cliques and bridges. And, as with the indented tree layout, multivariate data can easily be displayed alongside nodes. The problem of sorting the nodes in a manner that reveals underlying cluster structure is formally called *seriation*, and has diverse applications in visualization, statistics, and even archaeology!

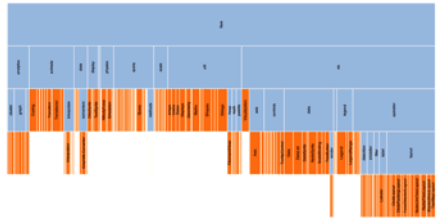
Matrix Views



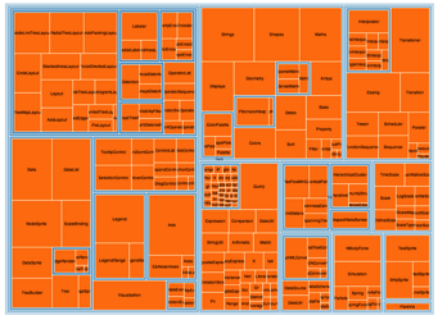
Node-Link Diagrams



Adjacency Diagrams

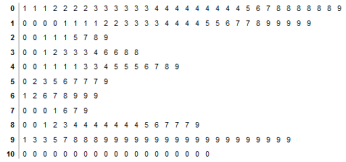


Enclosure Diagrams

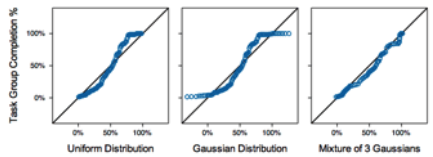


A Tour through the Visualization Zoo
Heer, Bostock, Ogievetsky
http://homes.cs.washington.edu/~jheer/files/zoo/

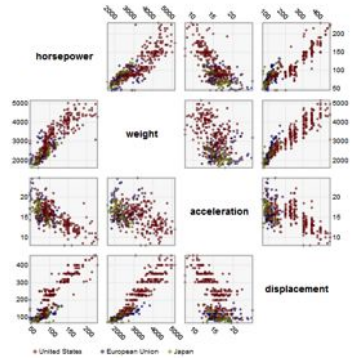
Stem-and-Leaf Plots



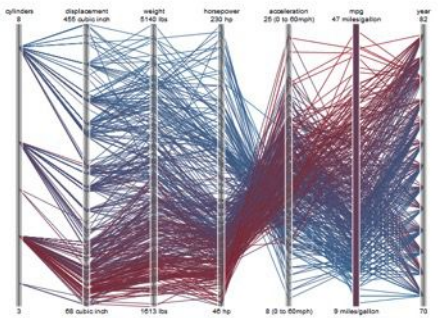
Q-Q Plots



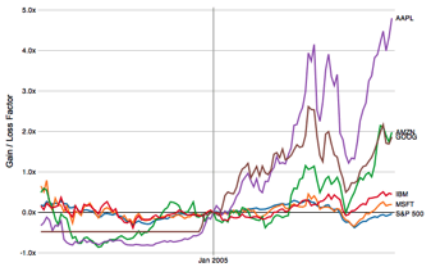
Scatter Plot Matrix (SPLOM)



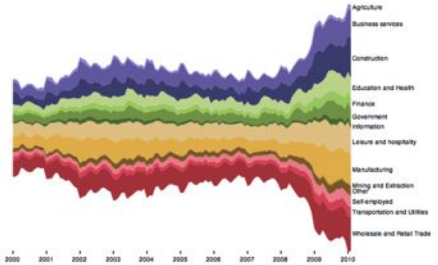
Parallel Coordinates



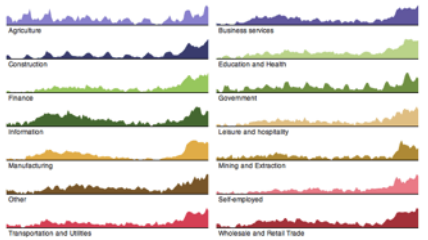
Index Charts



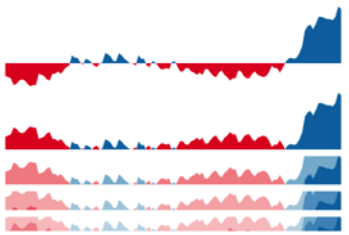
Stacked Graphs



Small Multiples



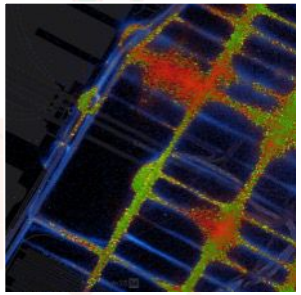
Horizon Graphs



uncharted

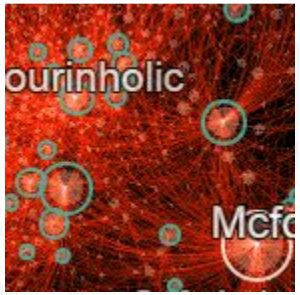
So, how do we visualize 100m data points?

MAP



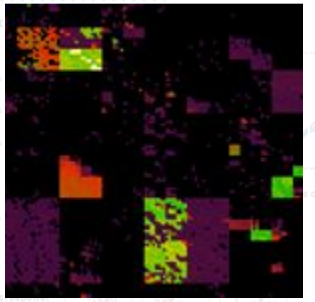
Strata NYC 2014

CONNECT

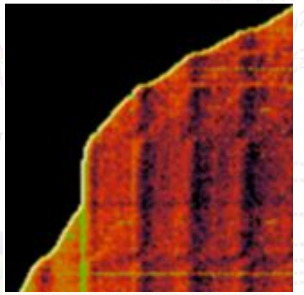


Strata London 2015

ORDER

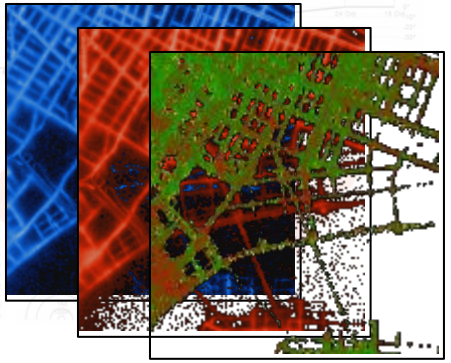


TIME

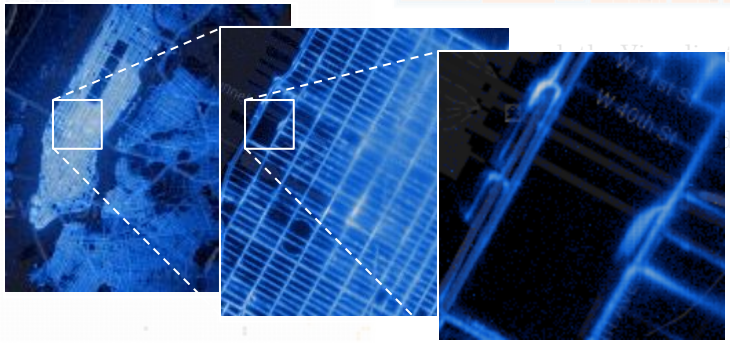


Today

LAYERS

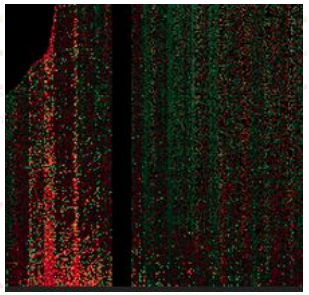


ZOOM



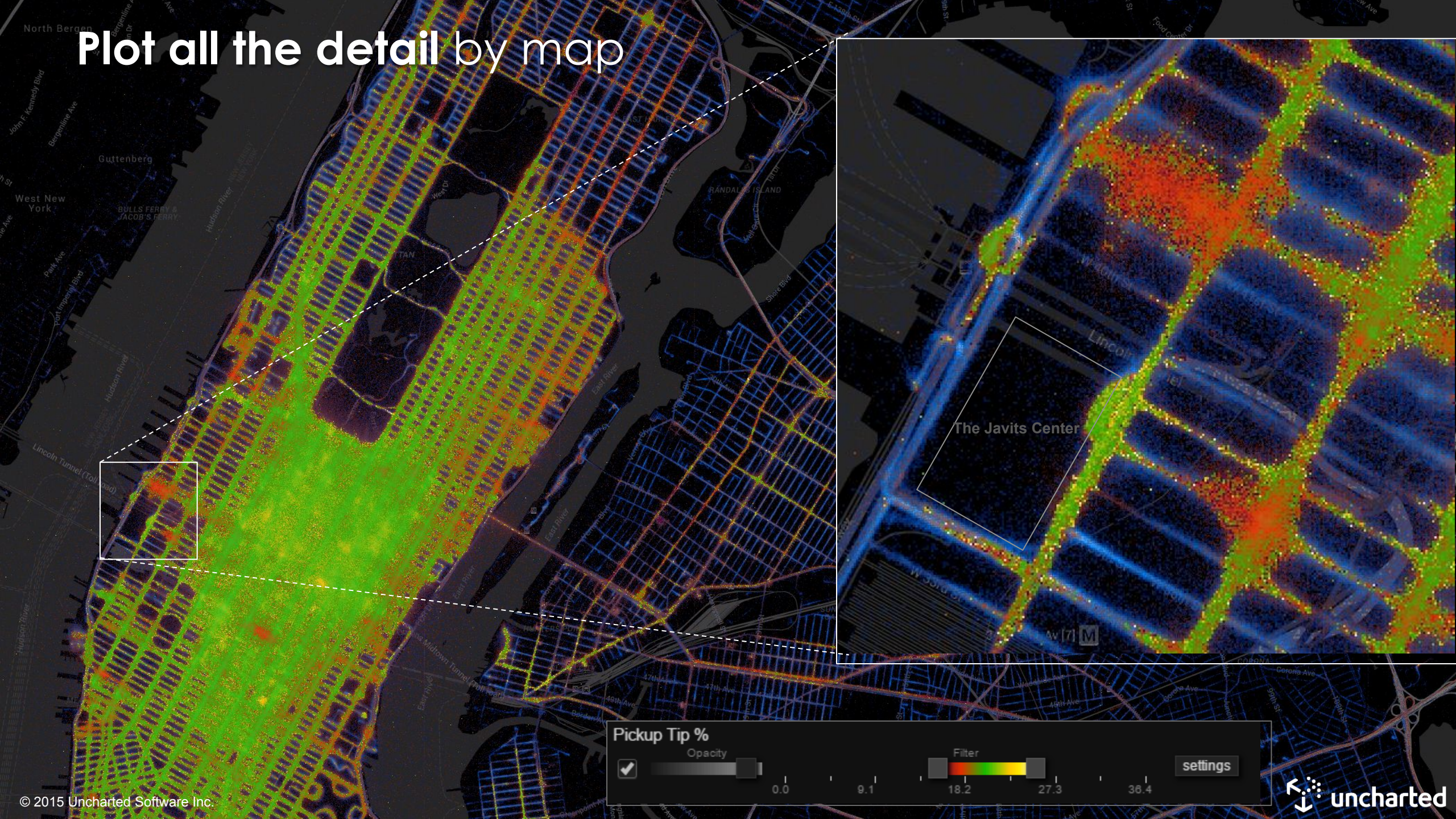
MINE

FILTER

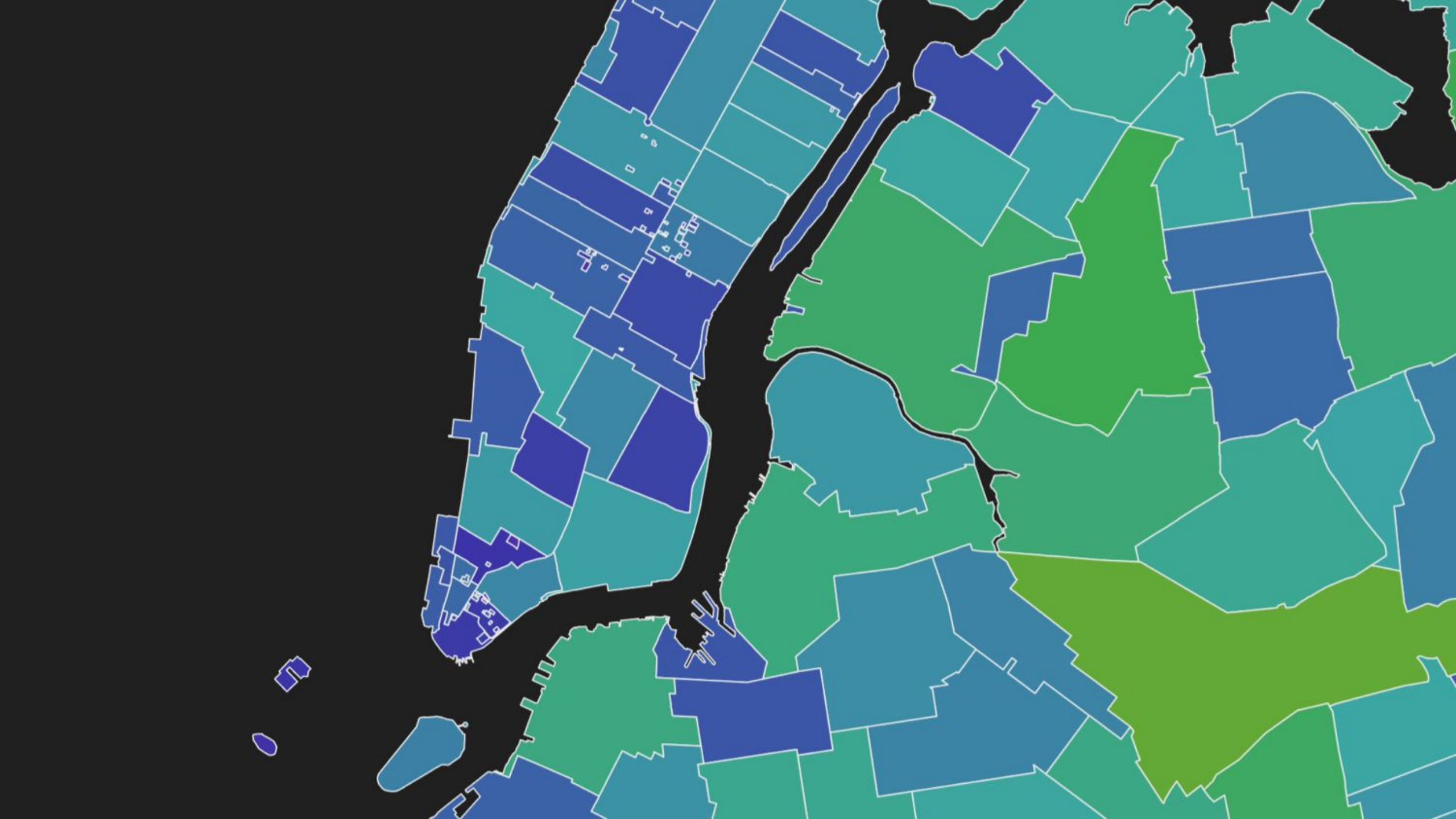


uncharted

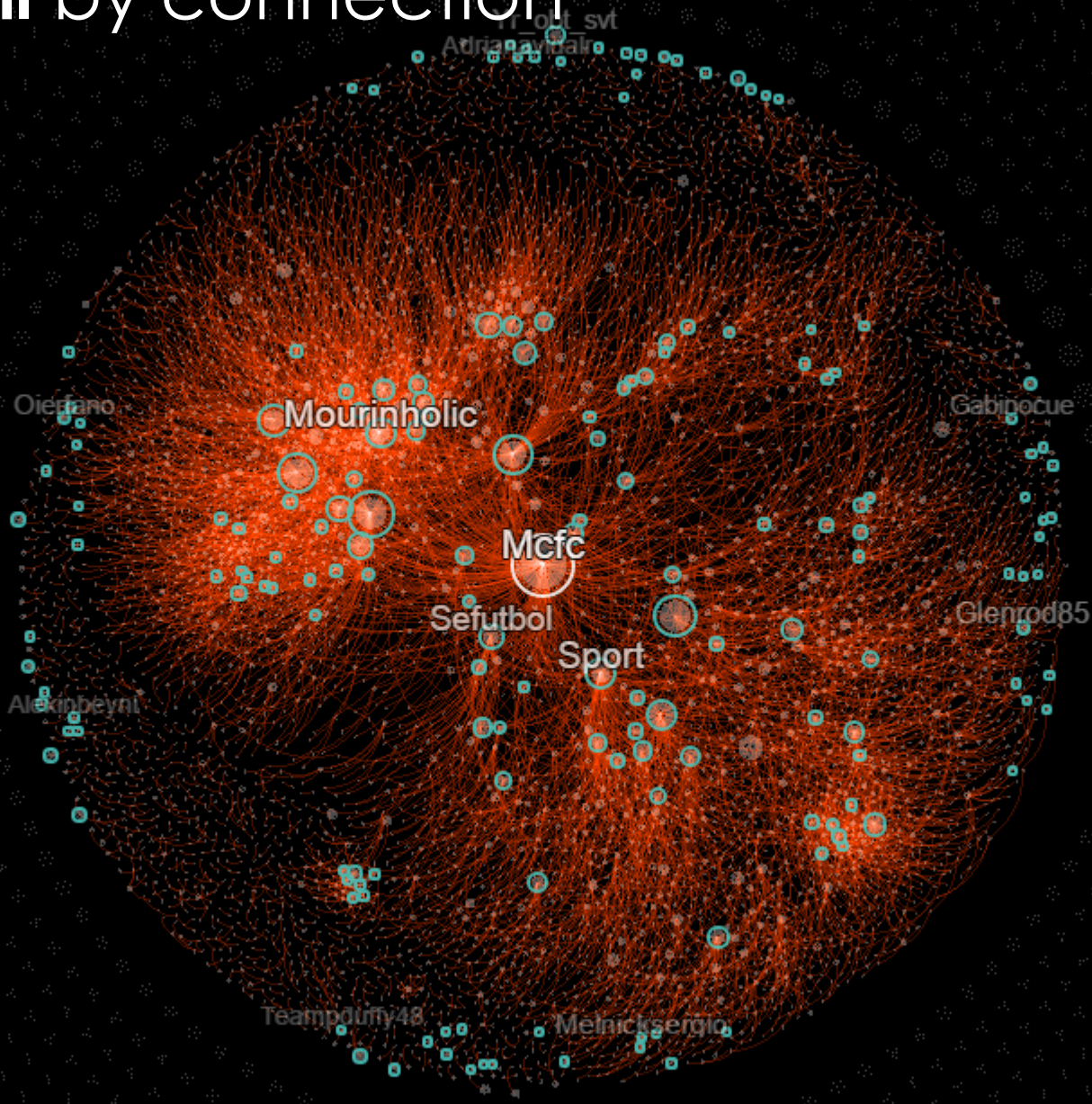
Plot all the detail by map



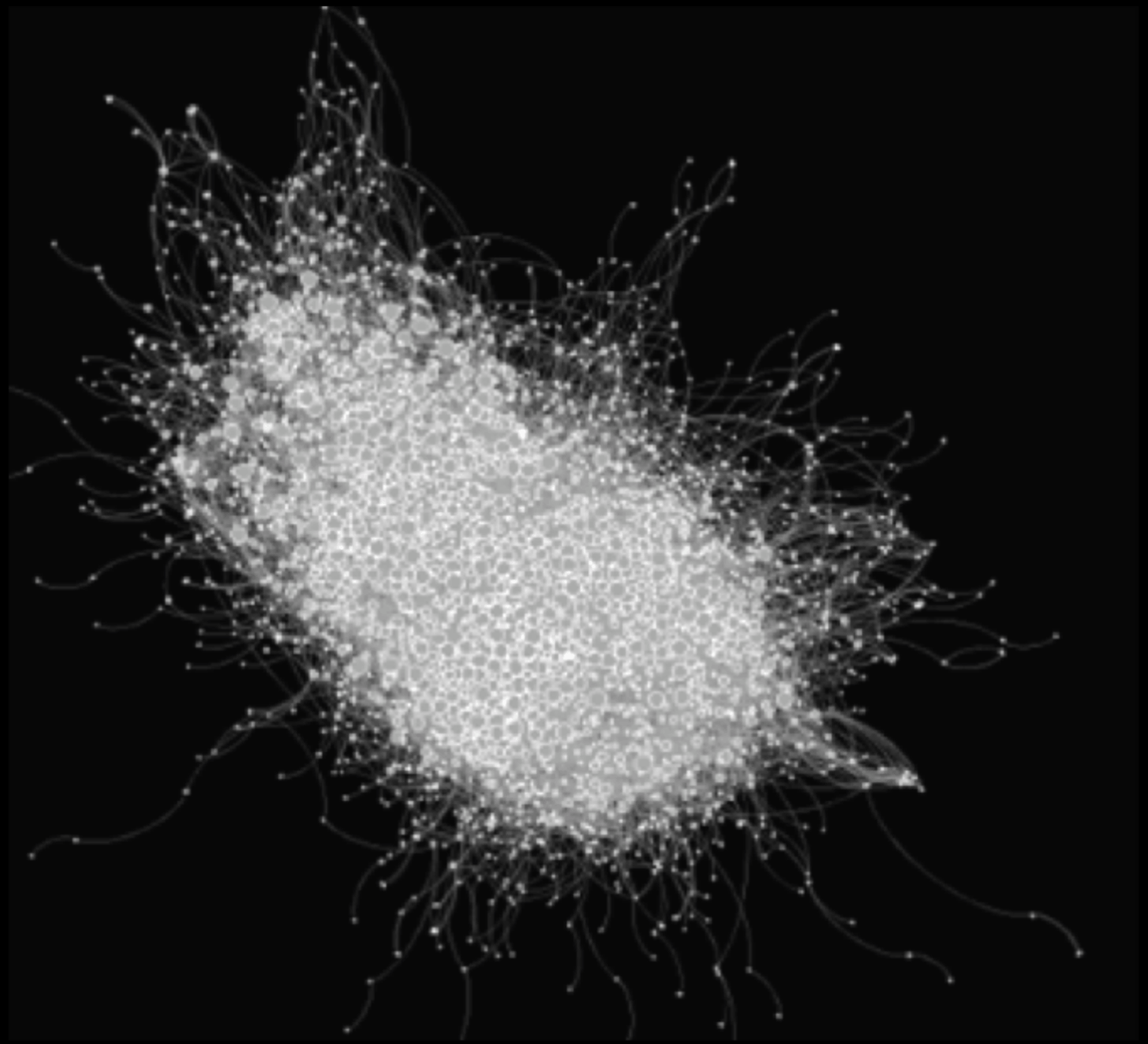
uncharted



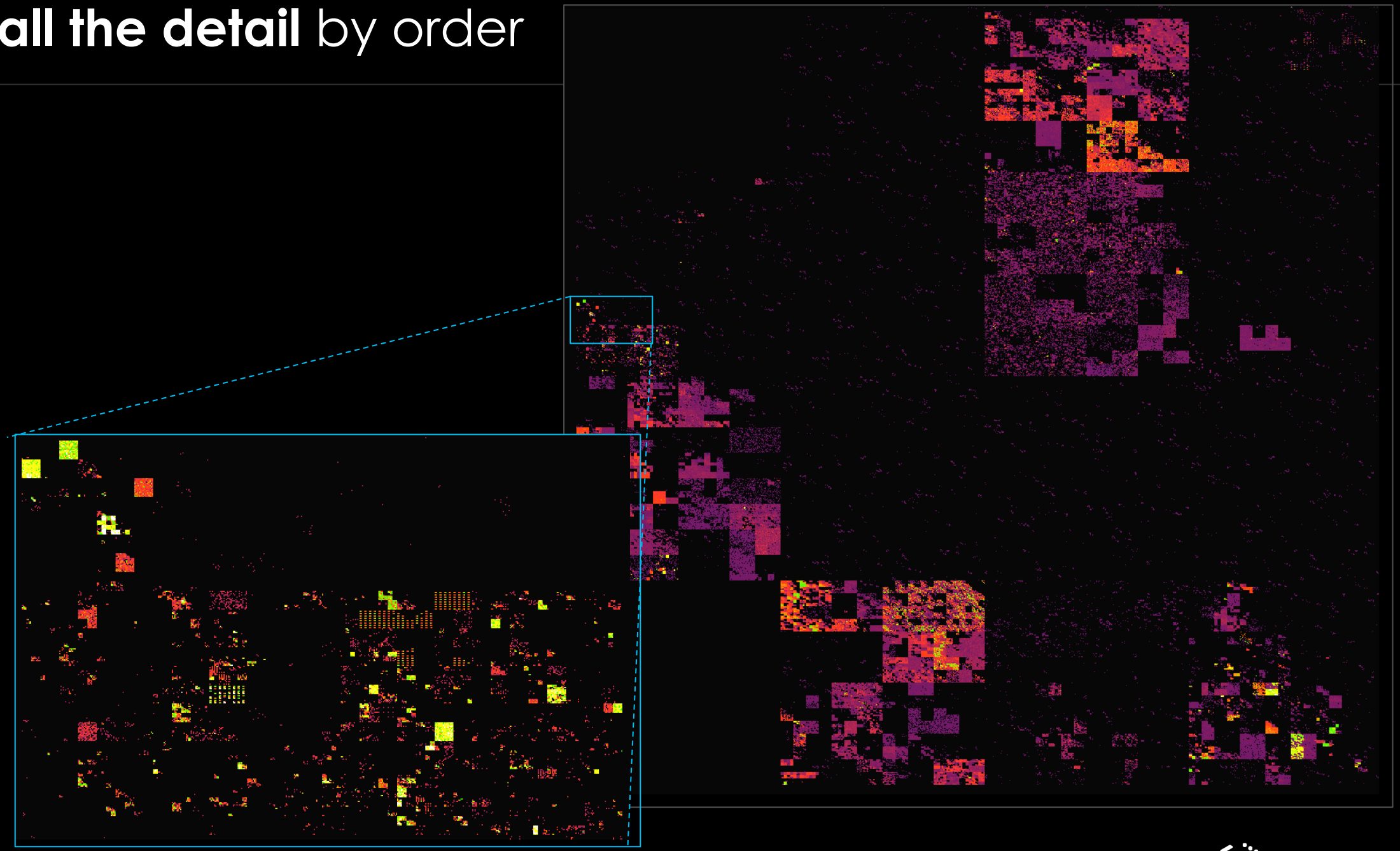
Plot all the detail by connection



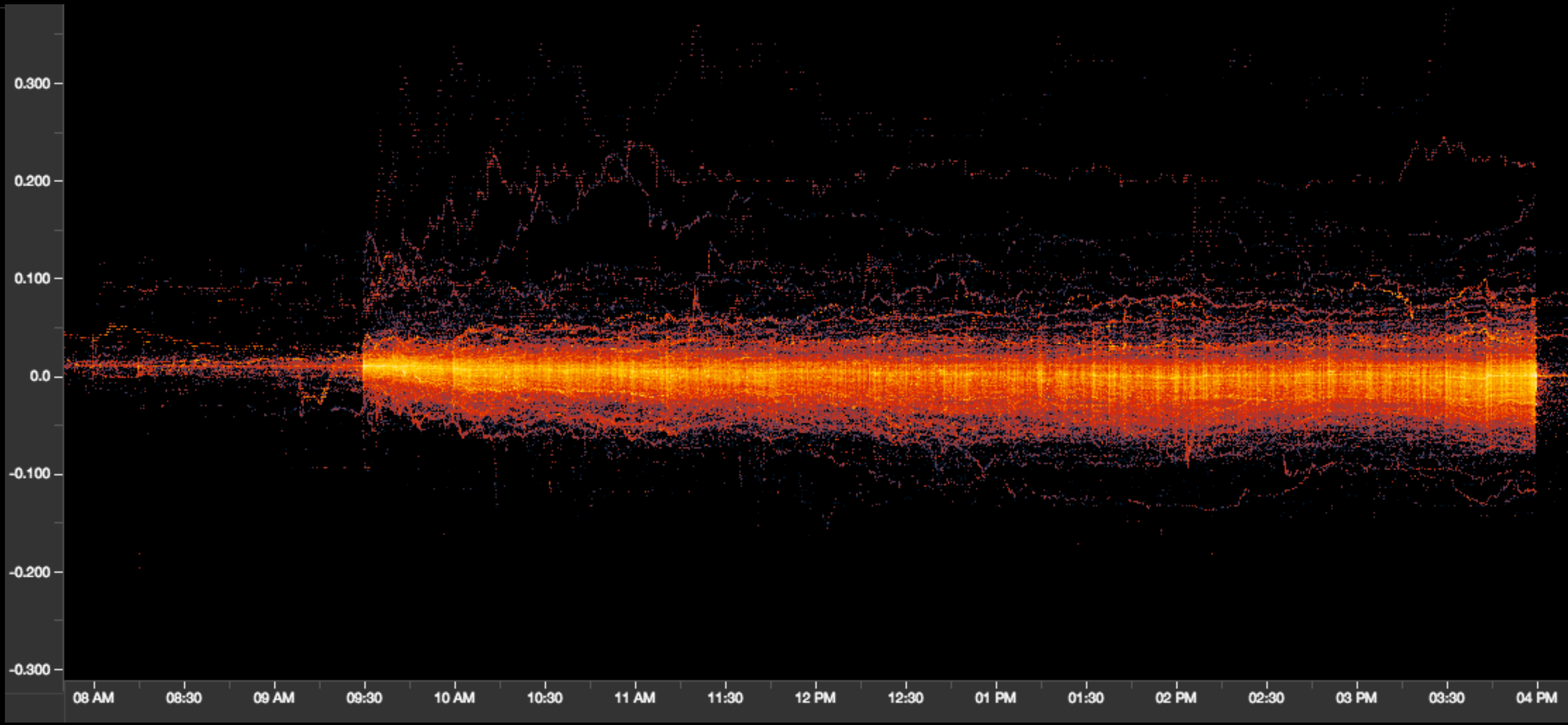
uncharted



Plot all the detail by order

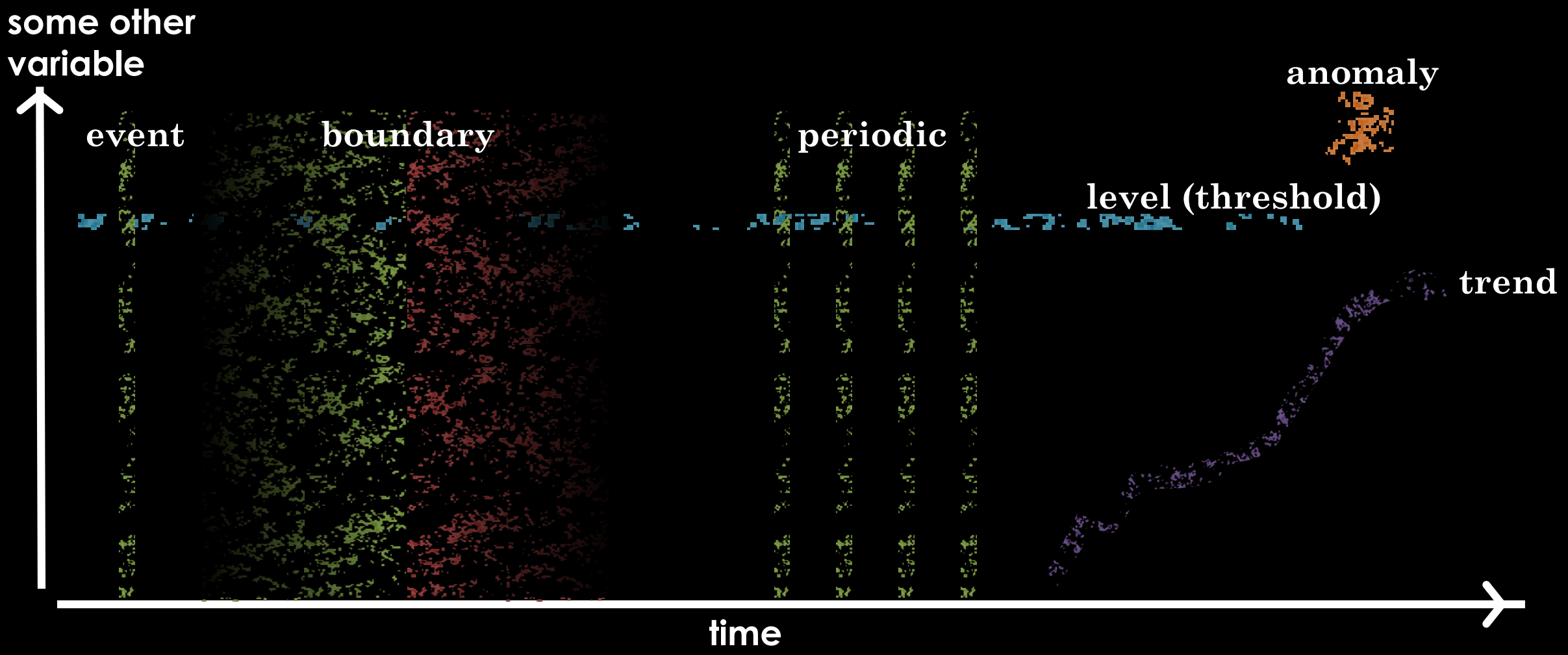


Plot all the detail by time



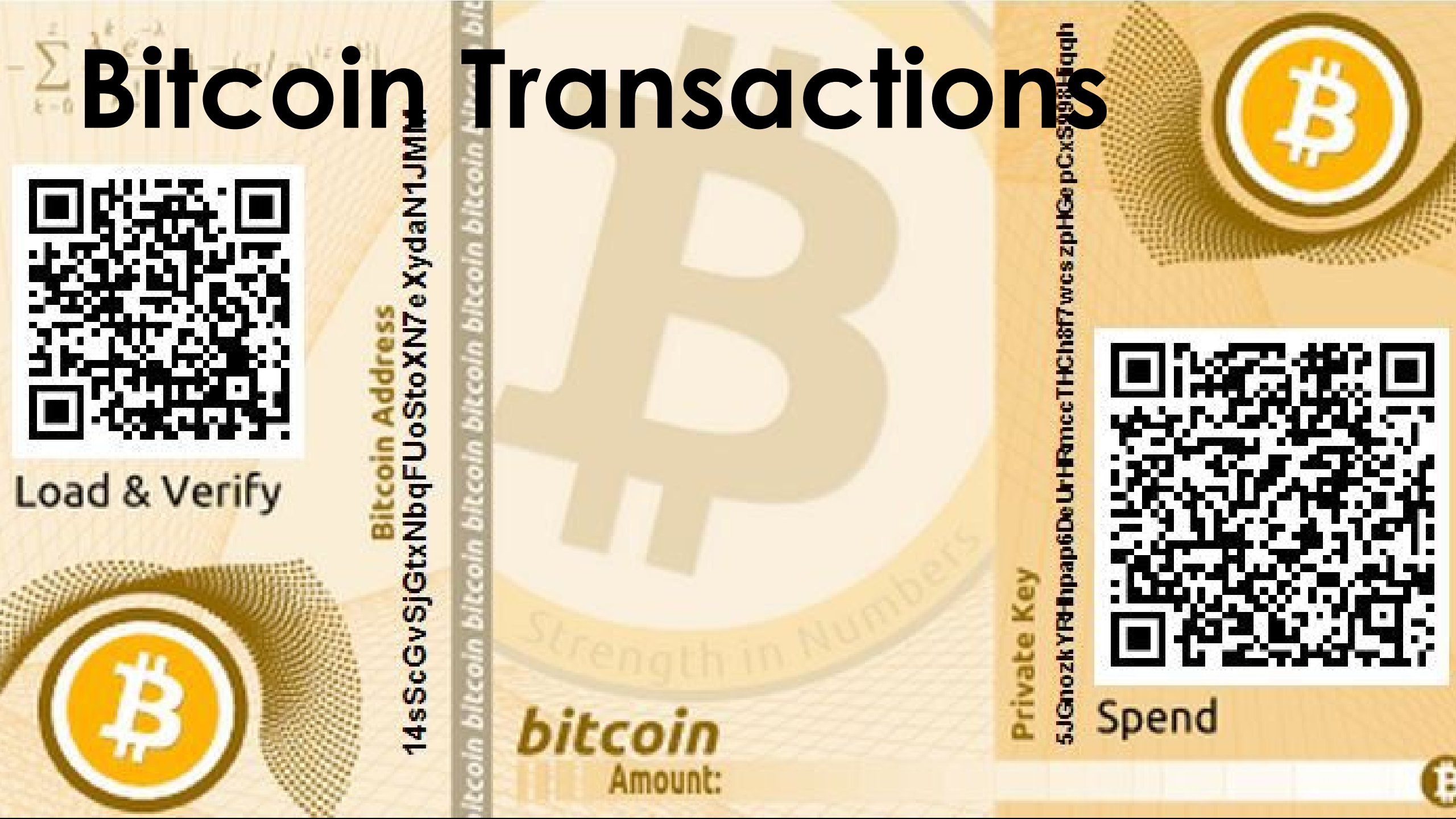
uncharted

Visual Patterns in Time Plots

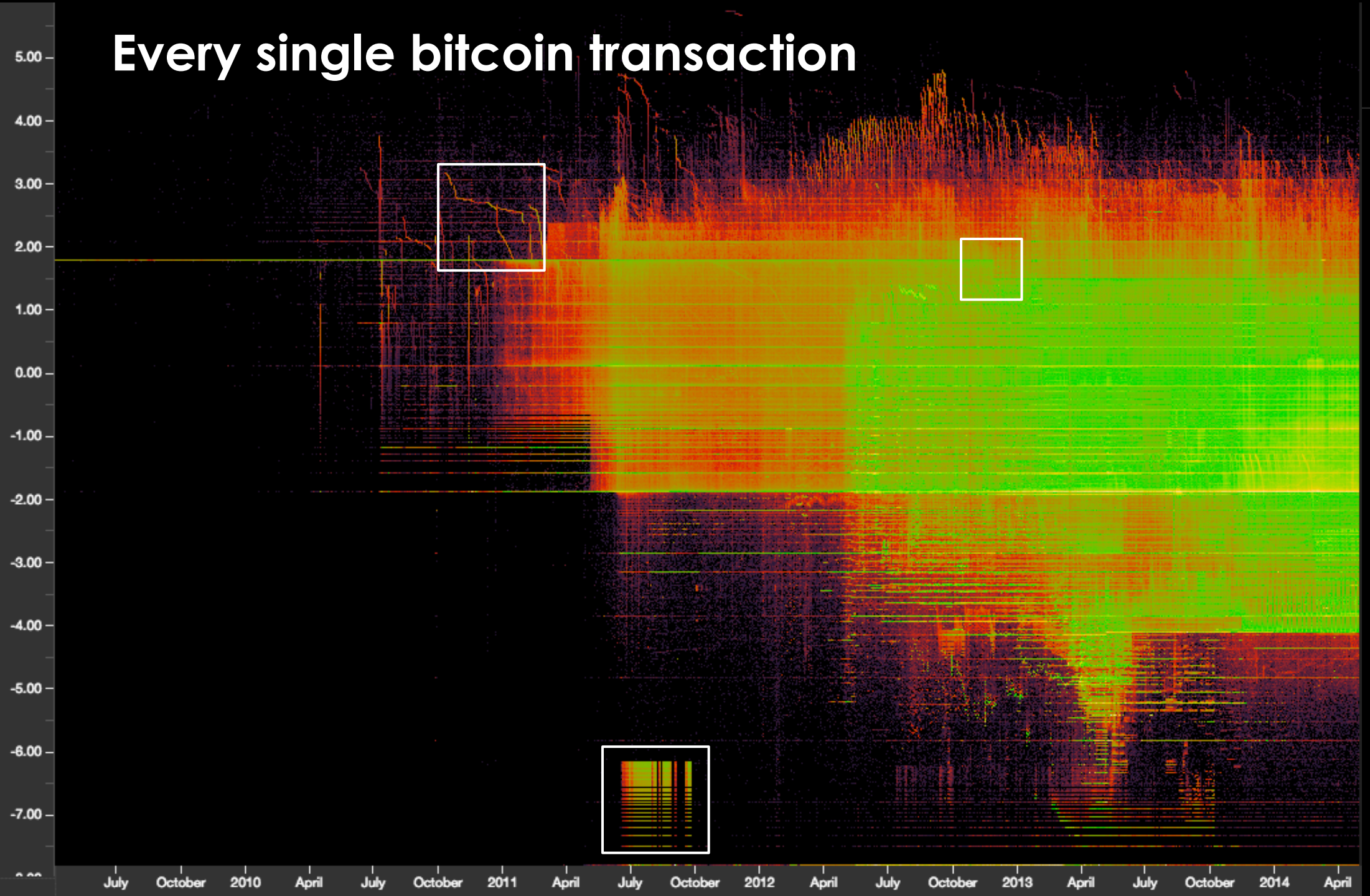


uncharted

Bitcoin Transactions



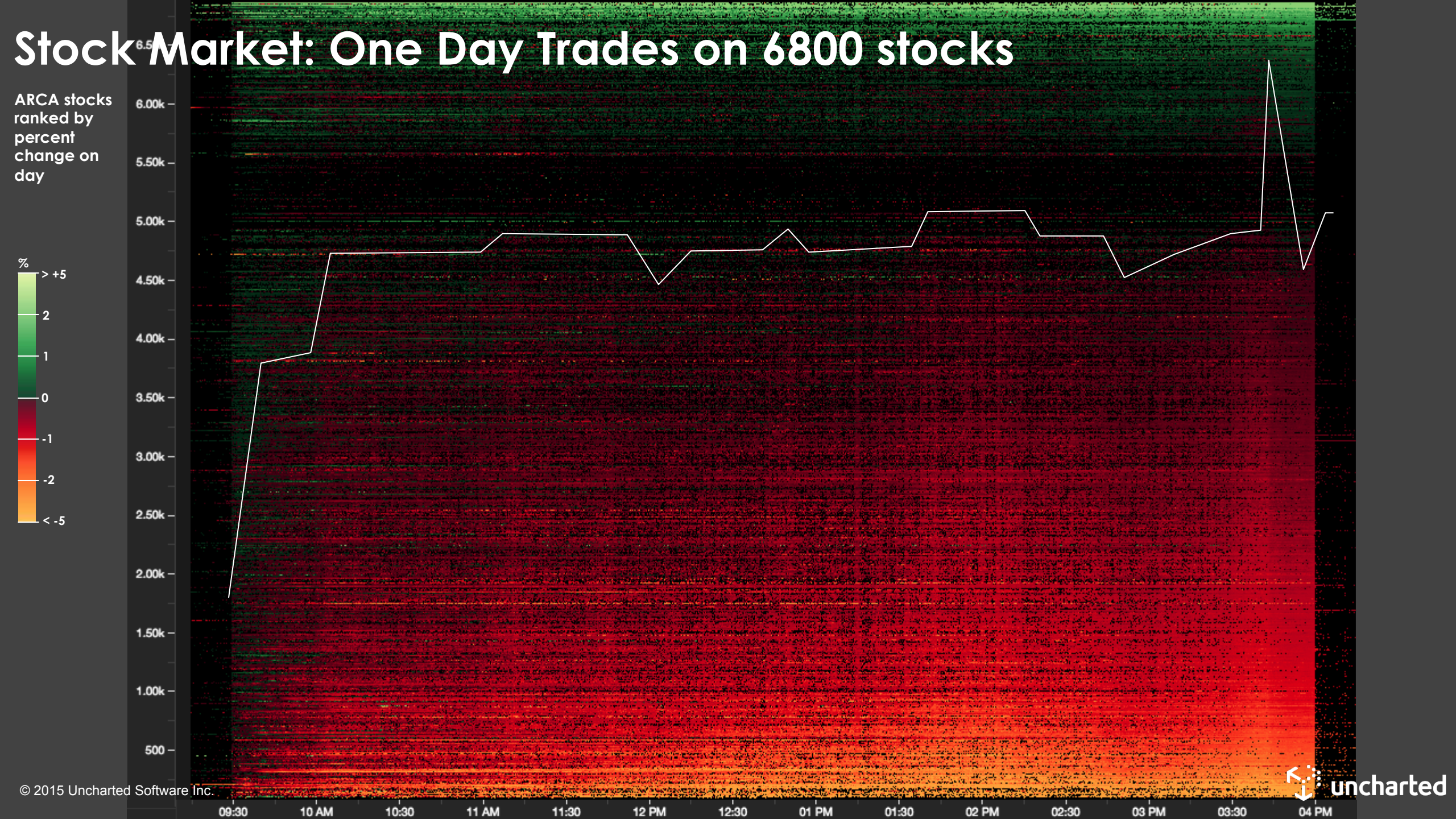
Every single bitcoin transaction



uncharted

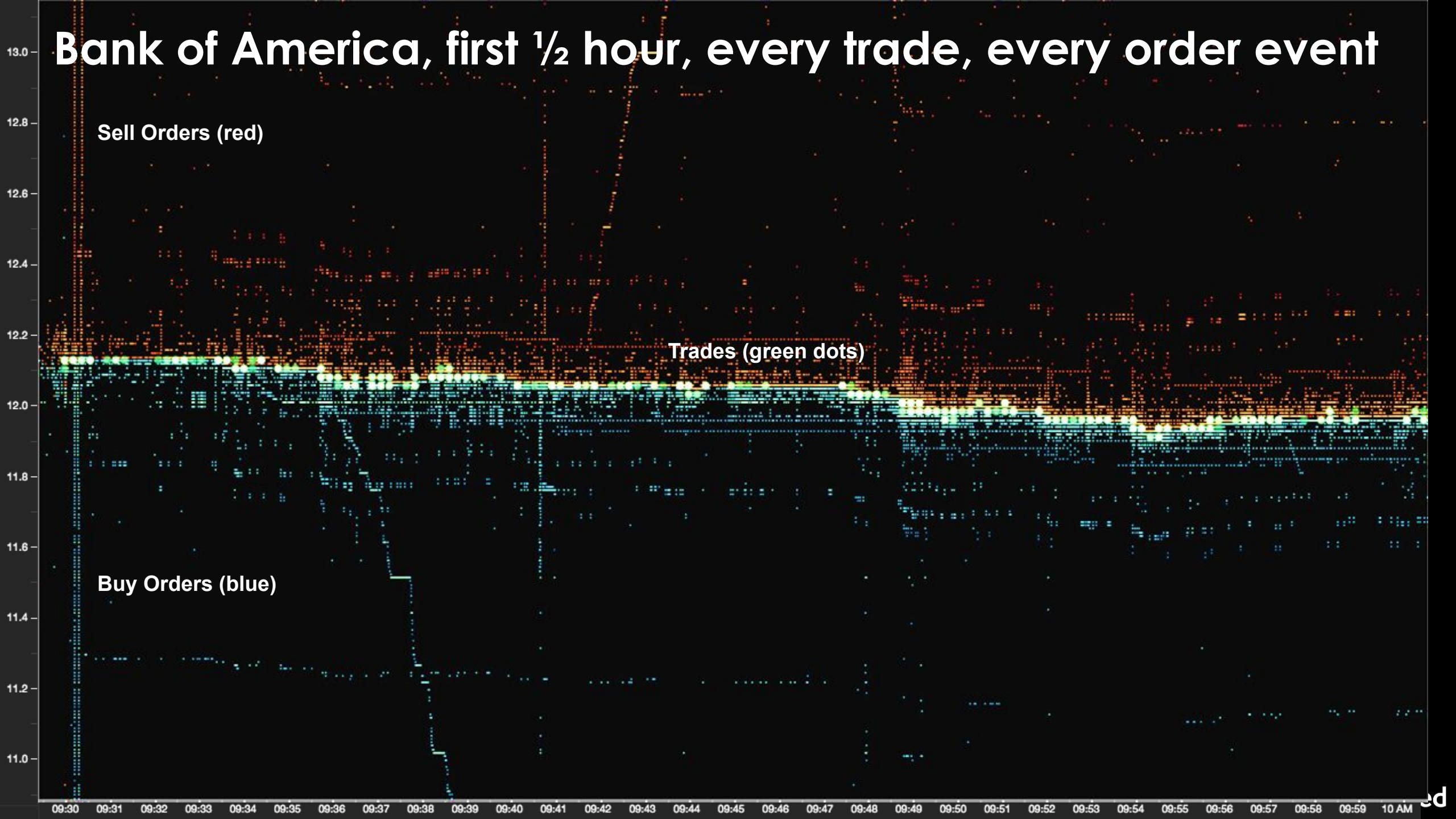

Financial Markets

Stock Market: One Day Trades on 6800 stocks

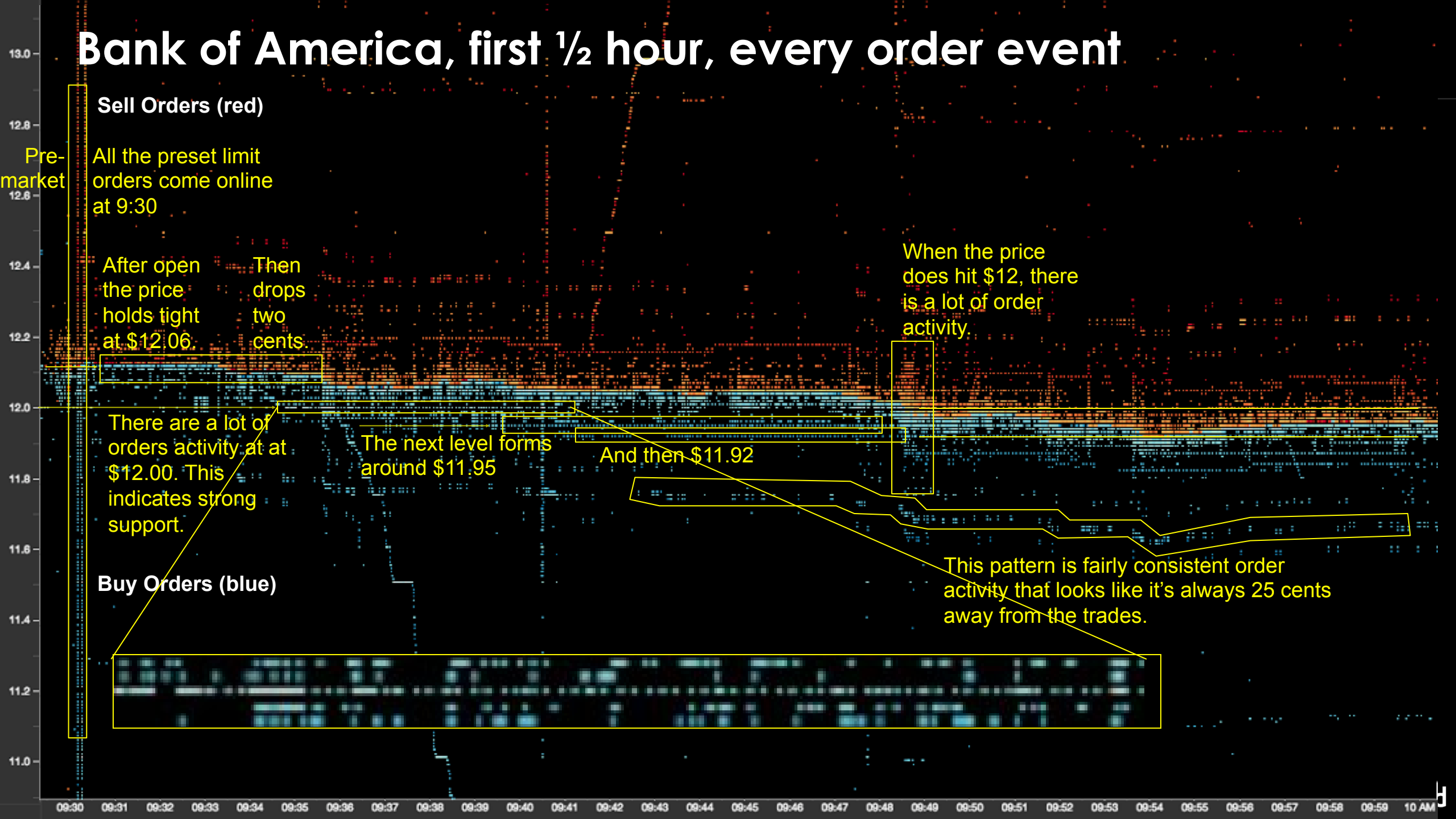


uncharted

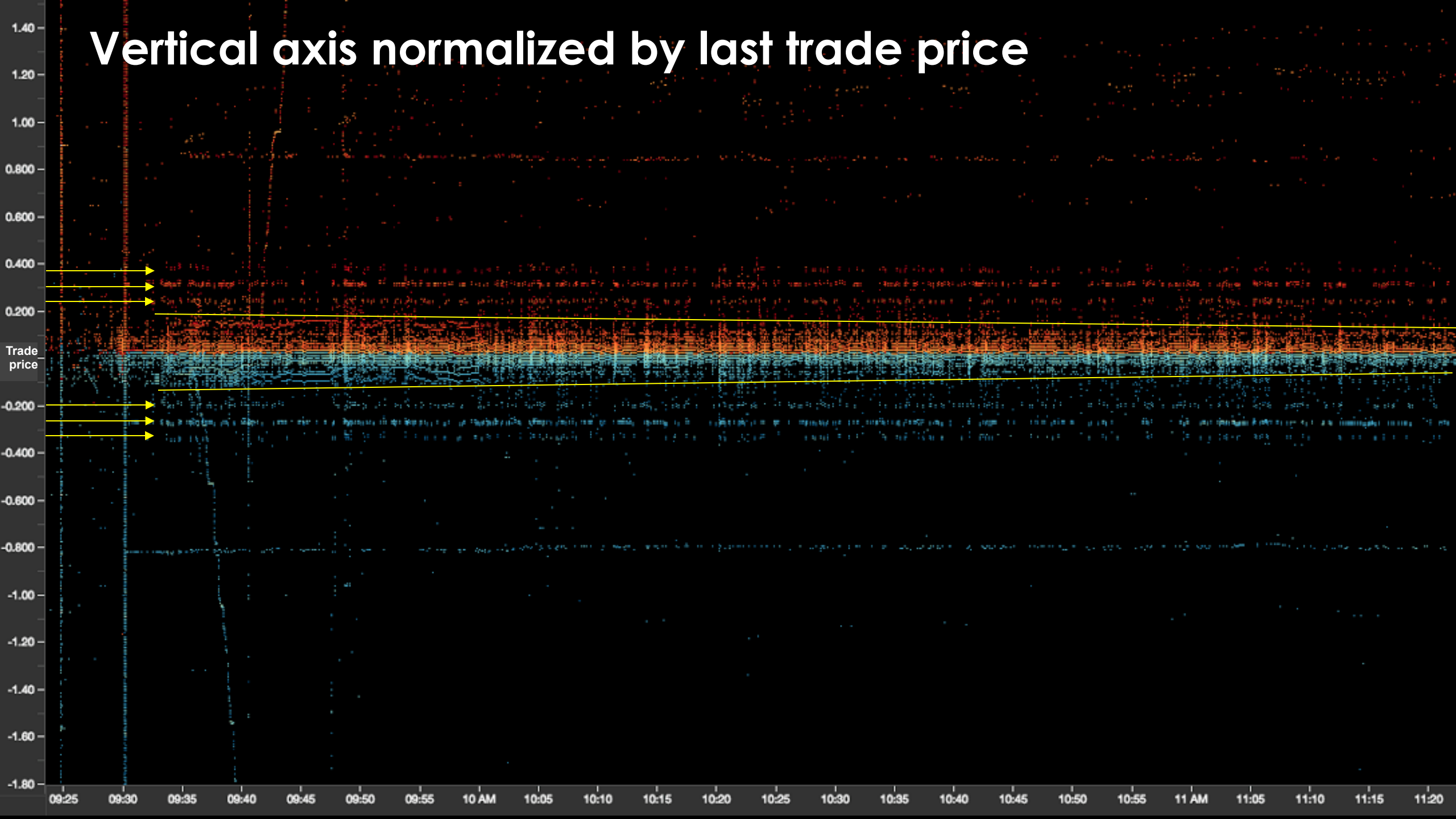
Bank of America, first ½ hour, every trade, every order event



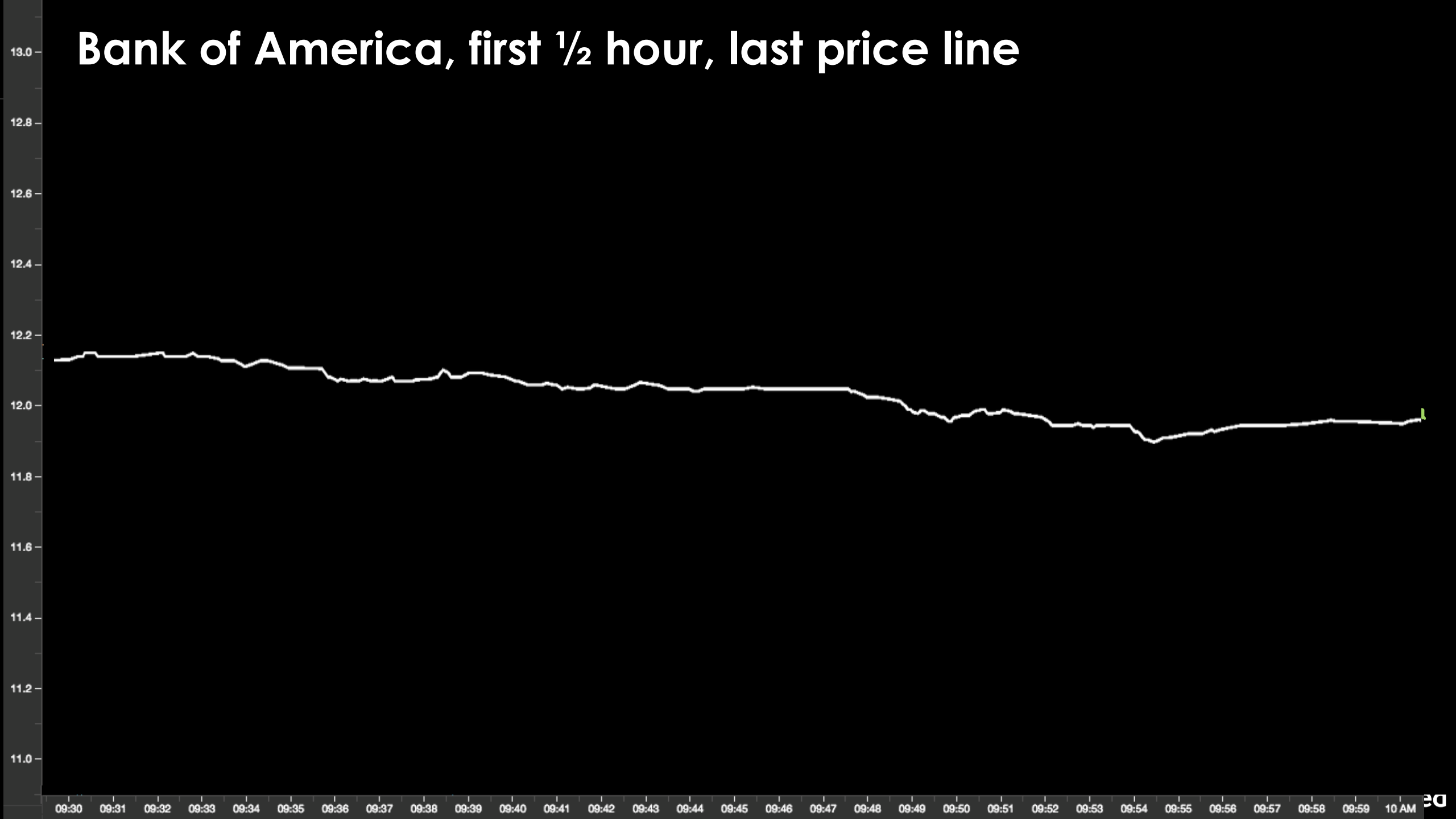
Bank of America, first ½ hour, every order event



Vertical axis normalized by last trade price



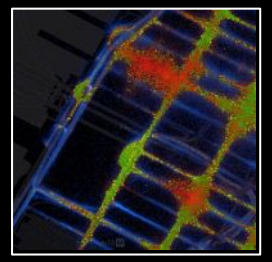
Bank of America, first ½ hour, last price line



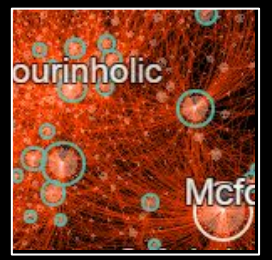
Technical Approach

Exploratory Big Data Analysis

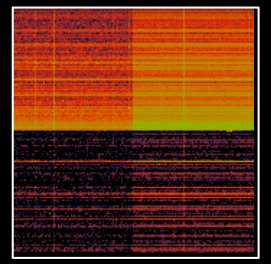
MAP



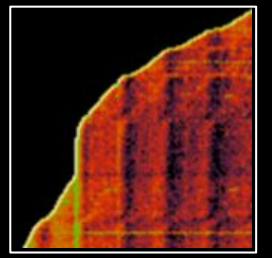
CONNECT



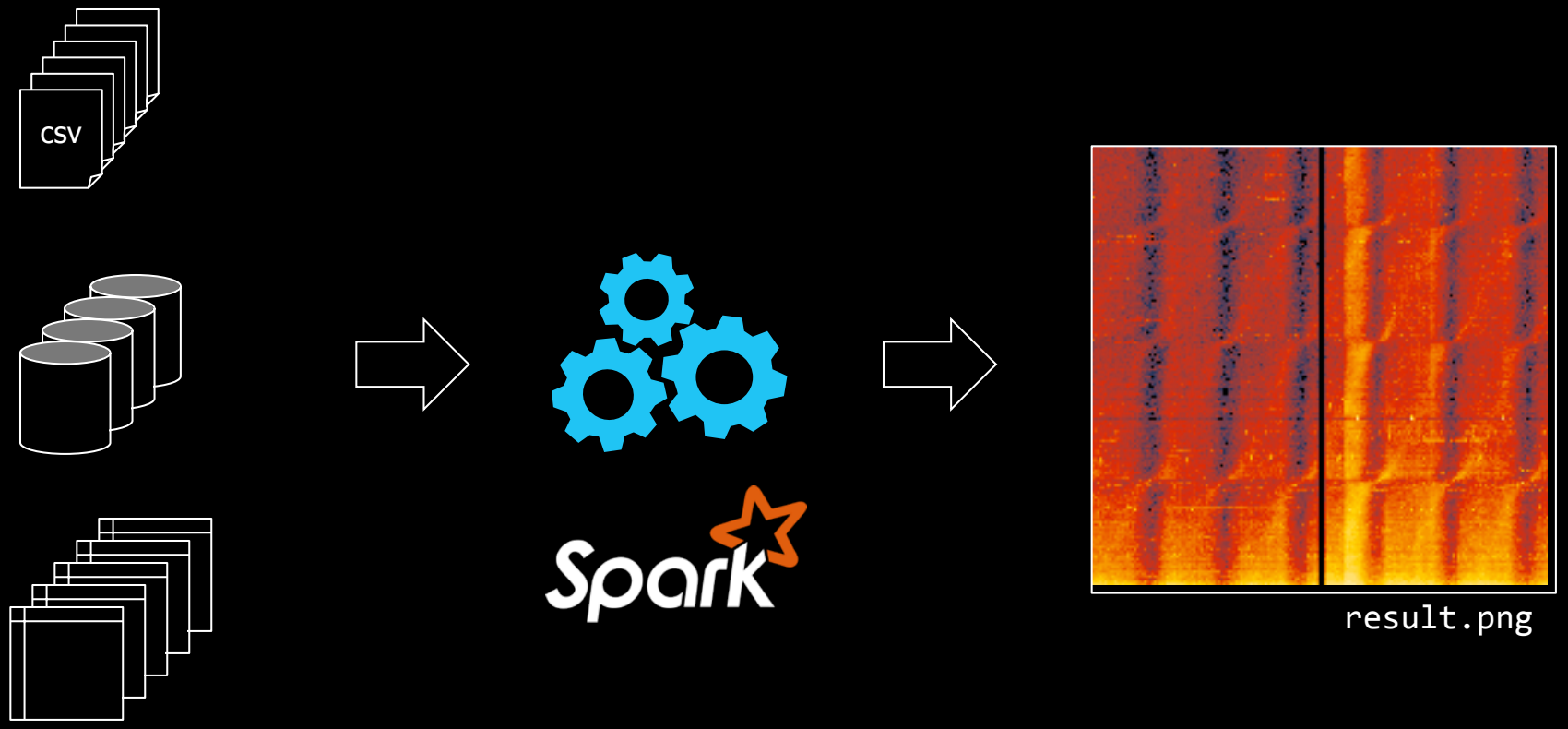
ORDER



TIME

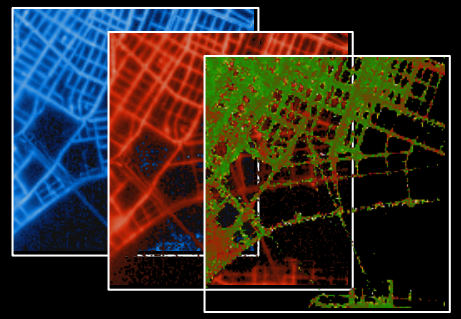


uncharted

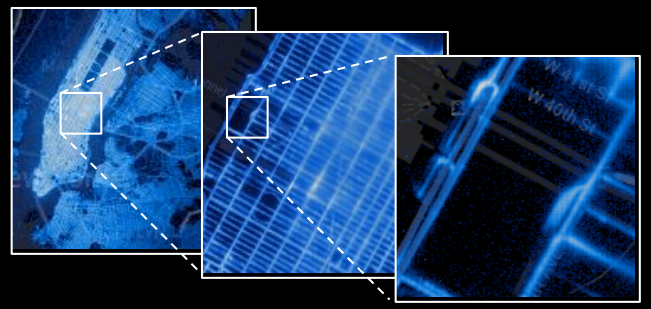


Exploratory Big Data Analysis – Rich Interactions

LAYERS

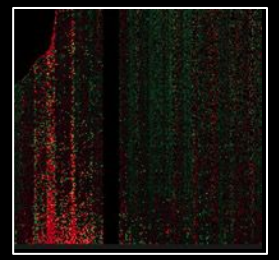


ZOOM



MINE

FILTER



uncharted

as a Service

Expressive API

Technical Approach

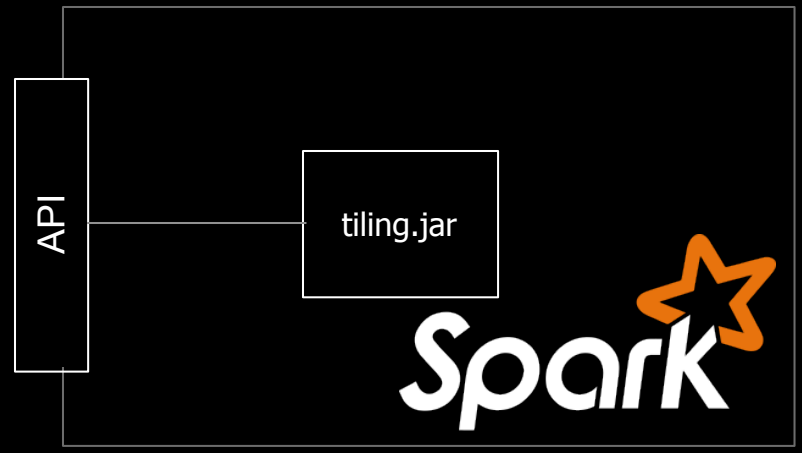
Data Pipeline

- Loading
- Filtering
- Transformation
- Sentiment
- Serialization

Results Generation

- Input mapping
- Projection
- "Pixel"-level analytics
- Area-level analytics
- Dataset-level analytics

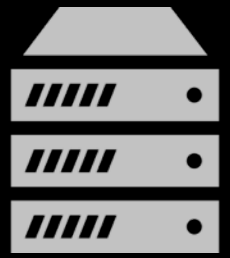
Spark Cluster as a Service



uncharted

Technical Approach

configuration
sentiment
density
distributions
transformations
term frequencies
summary statistics
data drilldown
filters



Message Queue

RabbitMQ

Shared Storage

Redis

Spark Cluster as a Service

API

Executor
Executor
Executor
Executor
Executor

tiling.jar

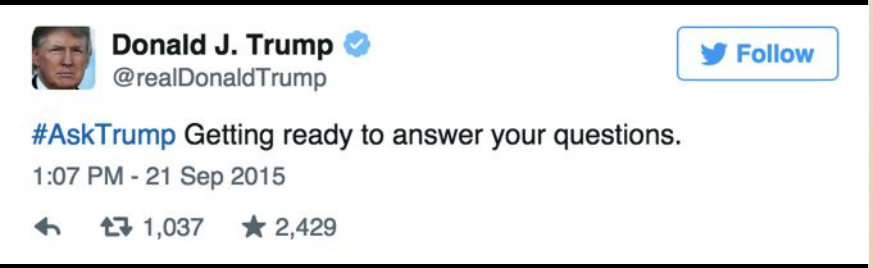
Spark

uncharted

Technical Approach

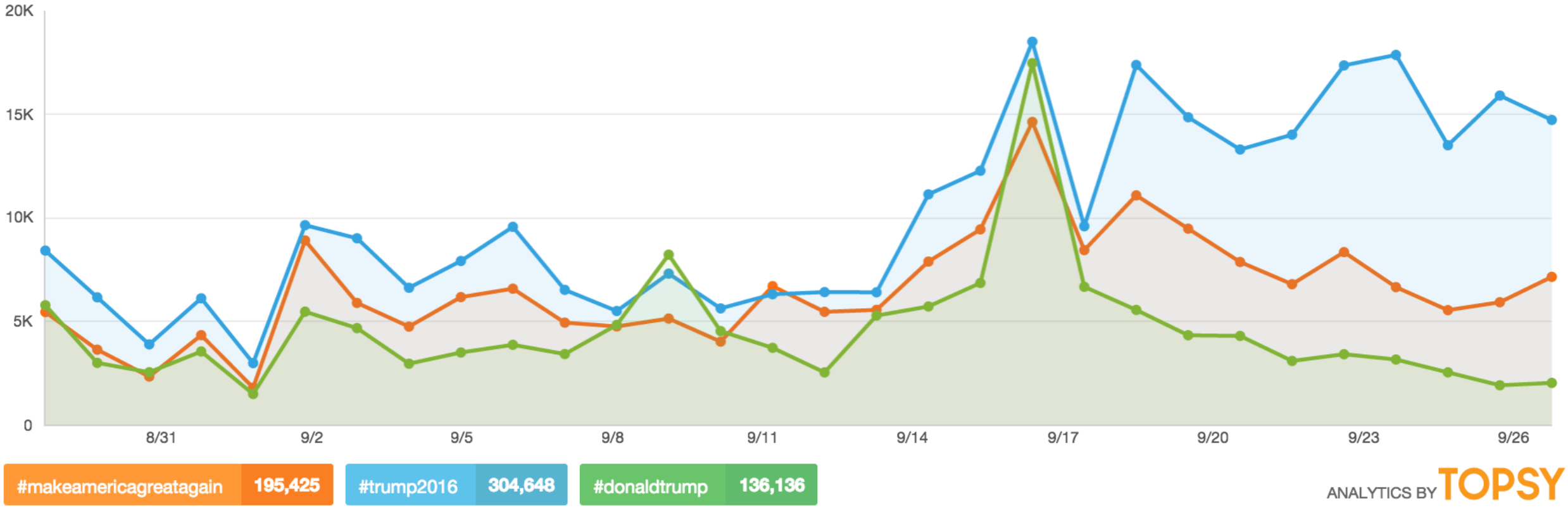
github.com/unchartedsoftware

Trump Tweets



Tweets per day: #makeamericagreatagain, #trump2016, and #donaldtrump

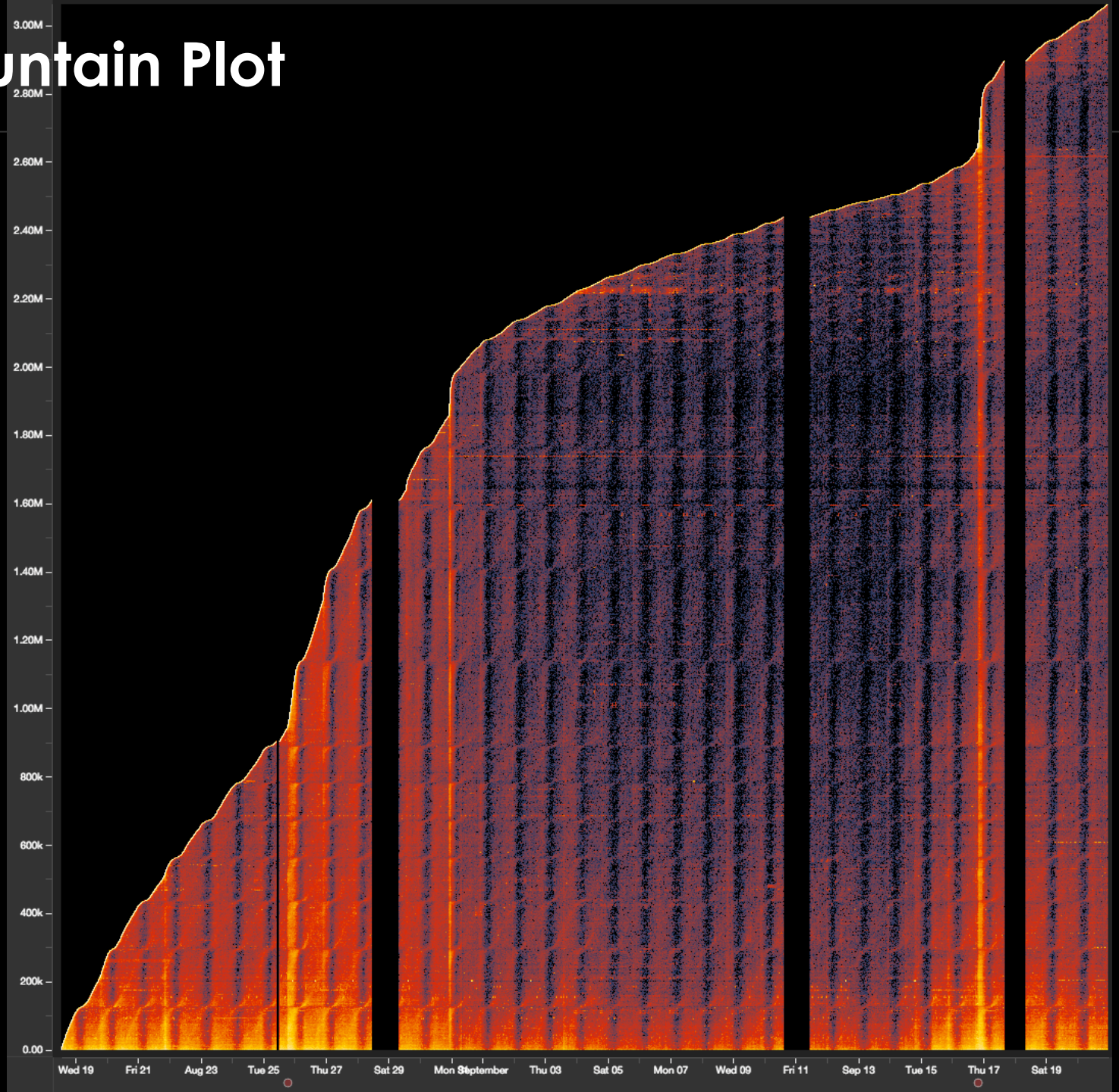
August 28th — September 27th



Continue

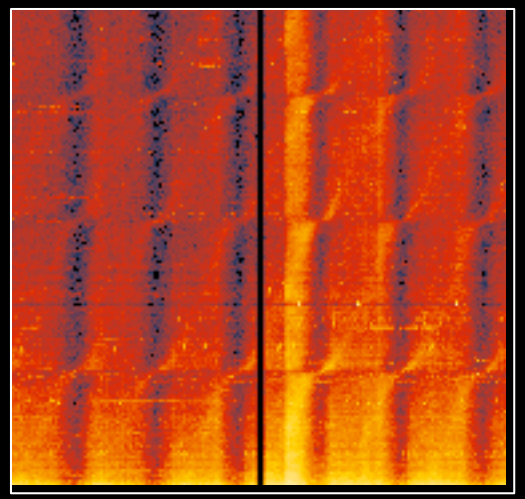
uncharted

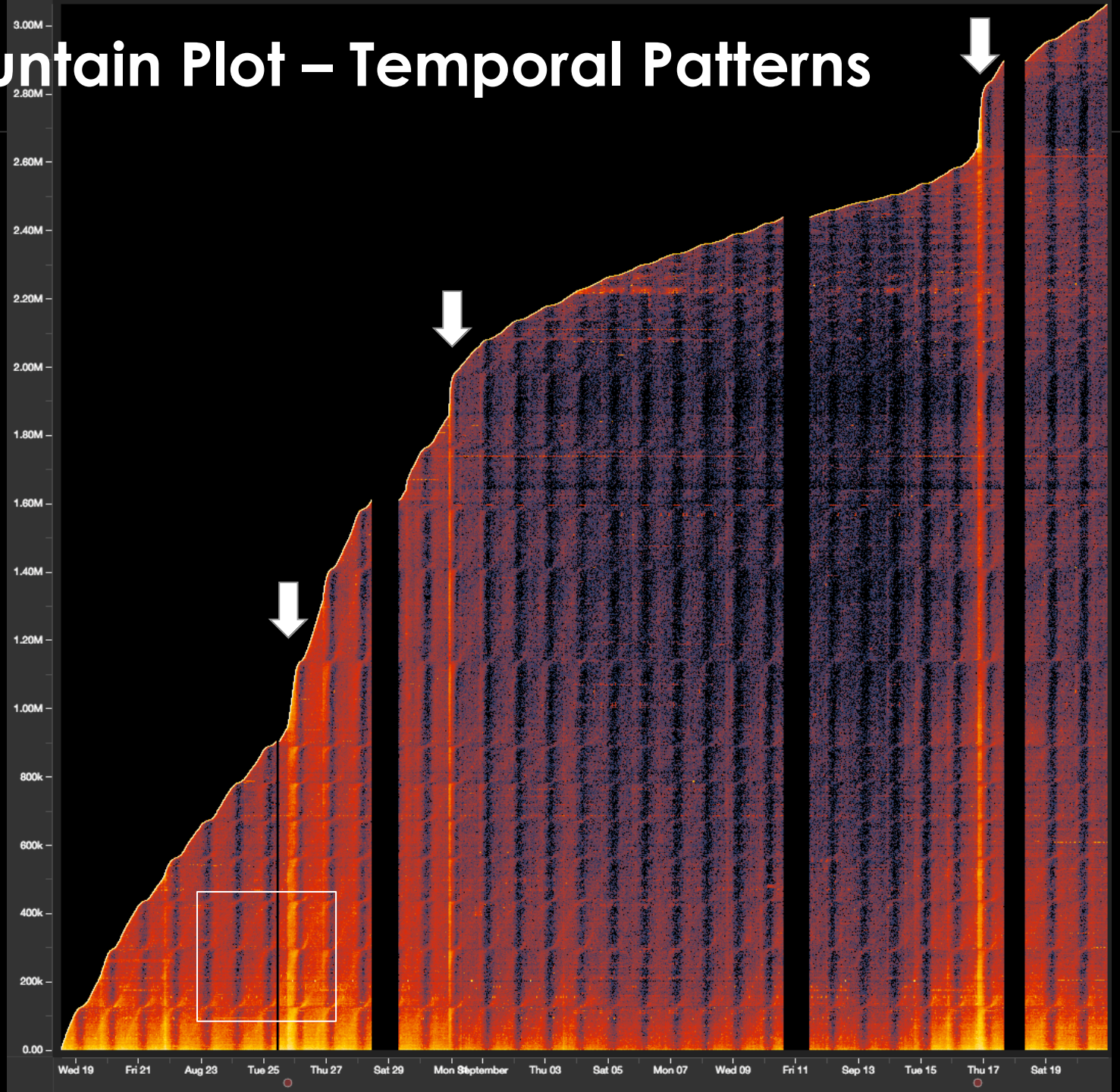
Trump Mountain Plot



uncharted

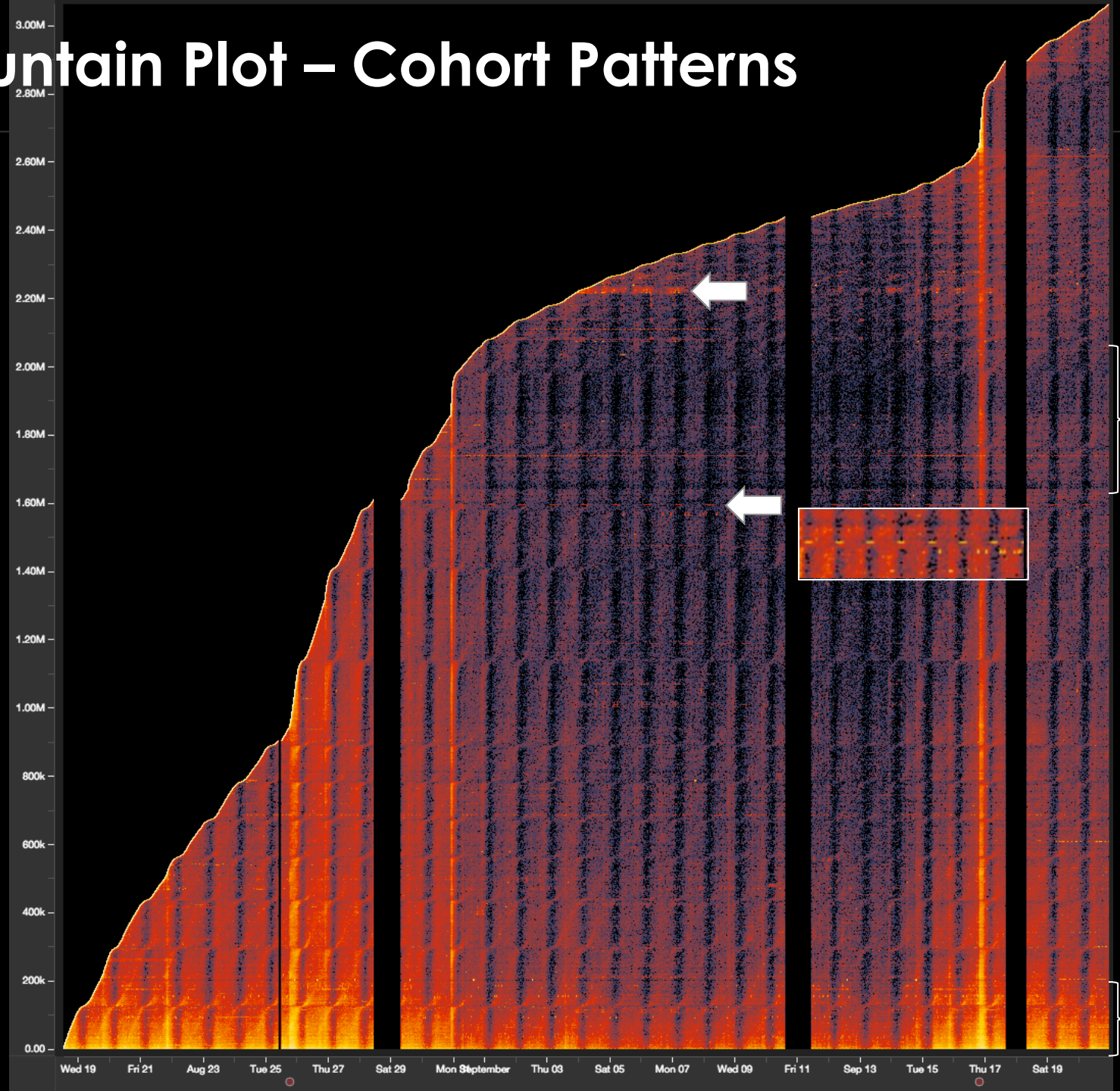
Trump Mountain Plot – Temporal Patterns





uncharted

Trump Mountain Plot – Cohort Patterns

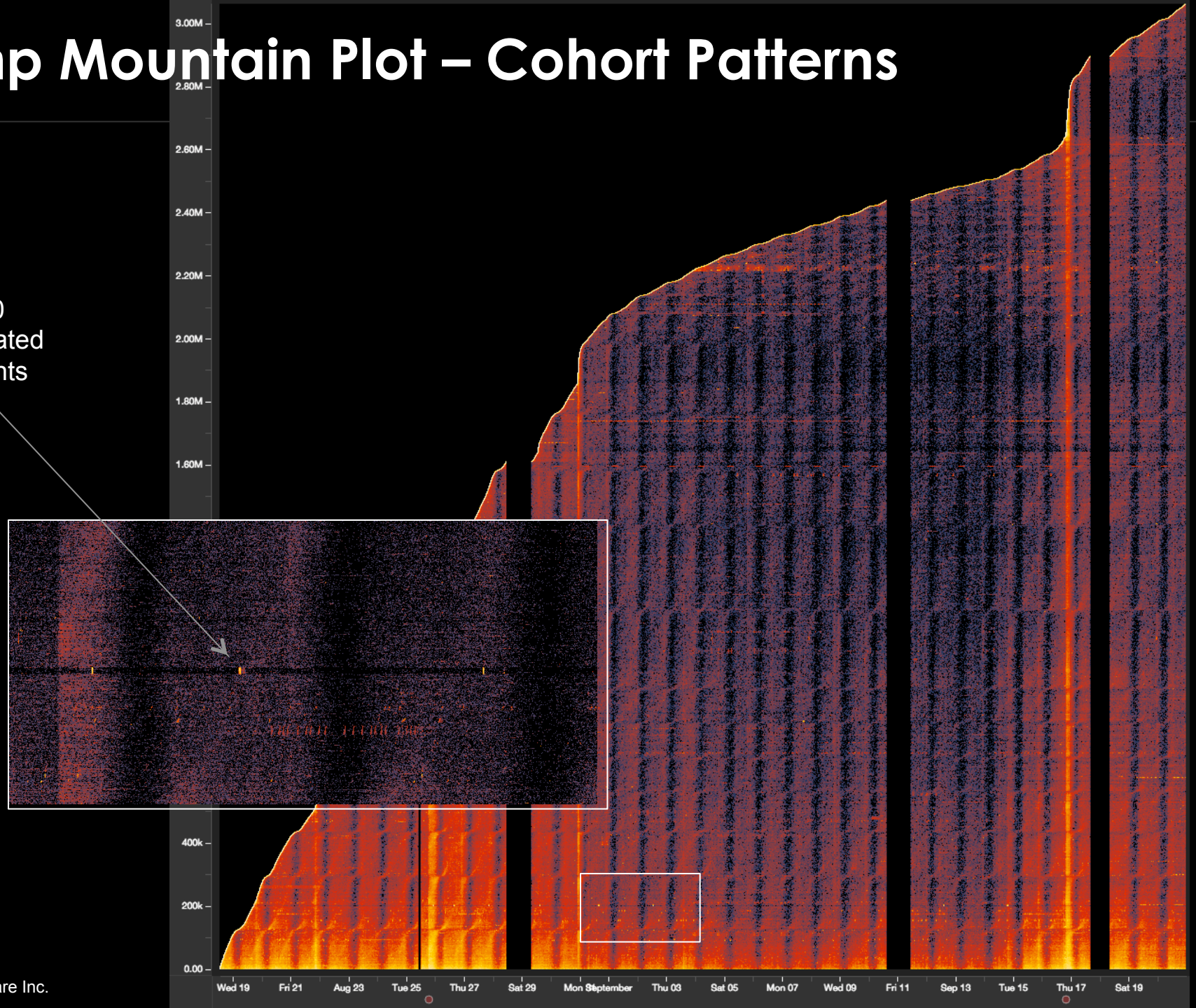


Event-driven

Early-engagers

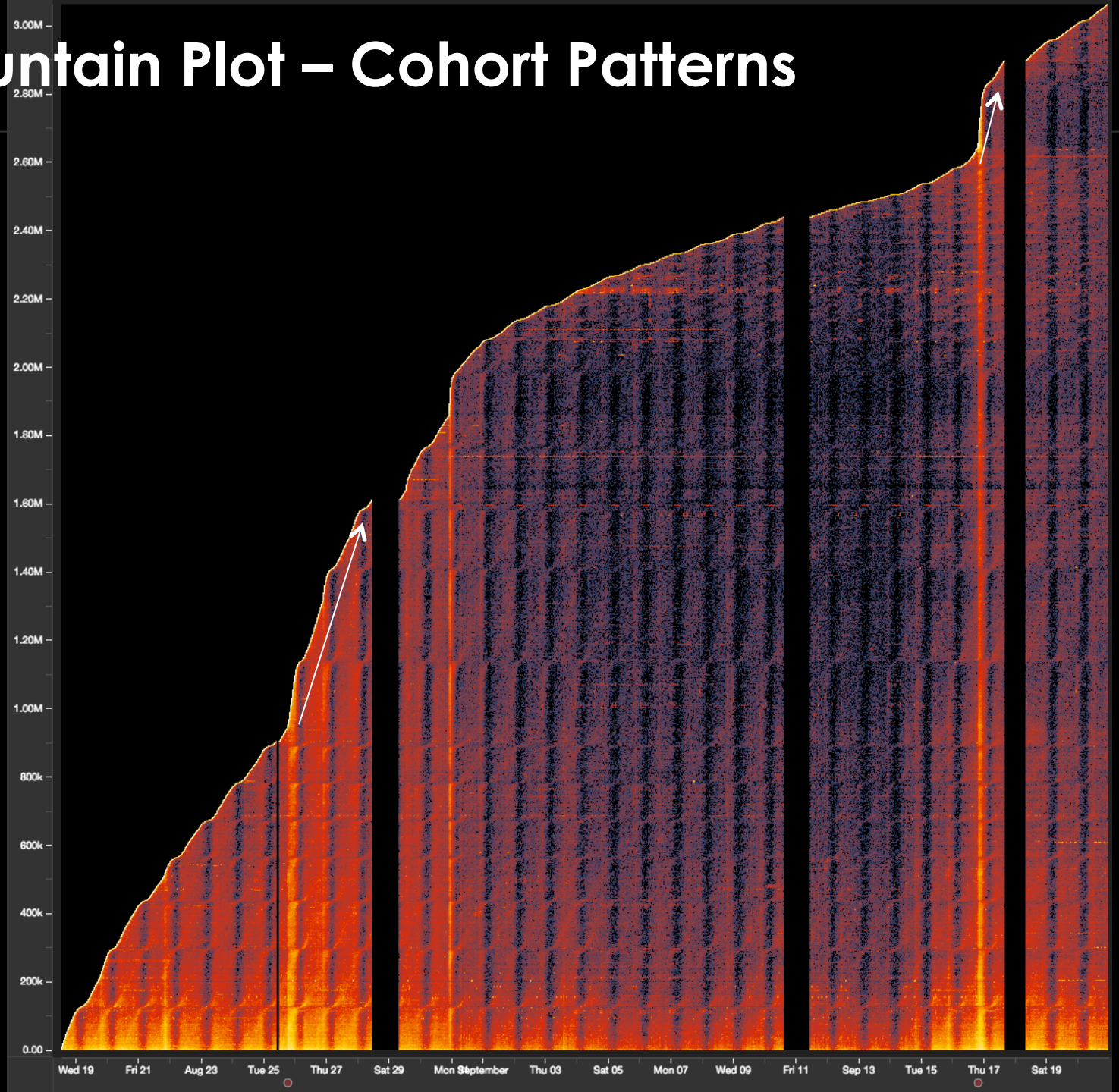
uncharted

Trump Mountain Plot – Cohort Patterns



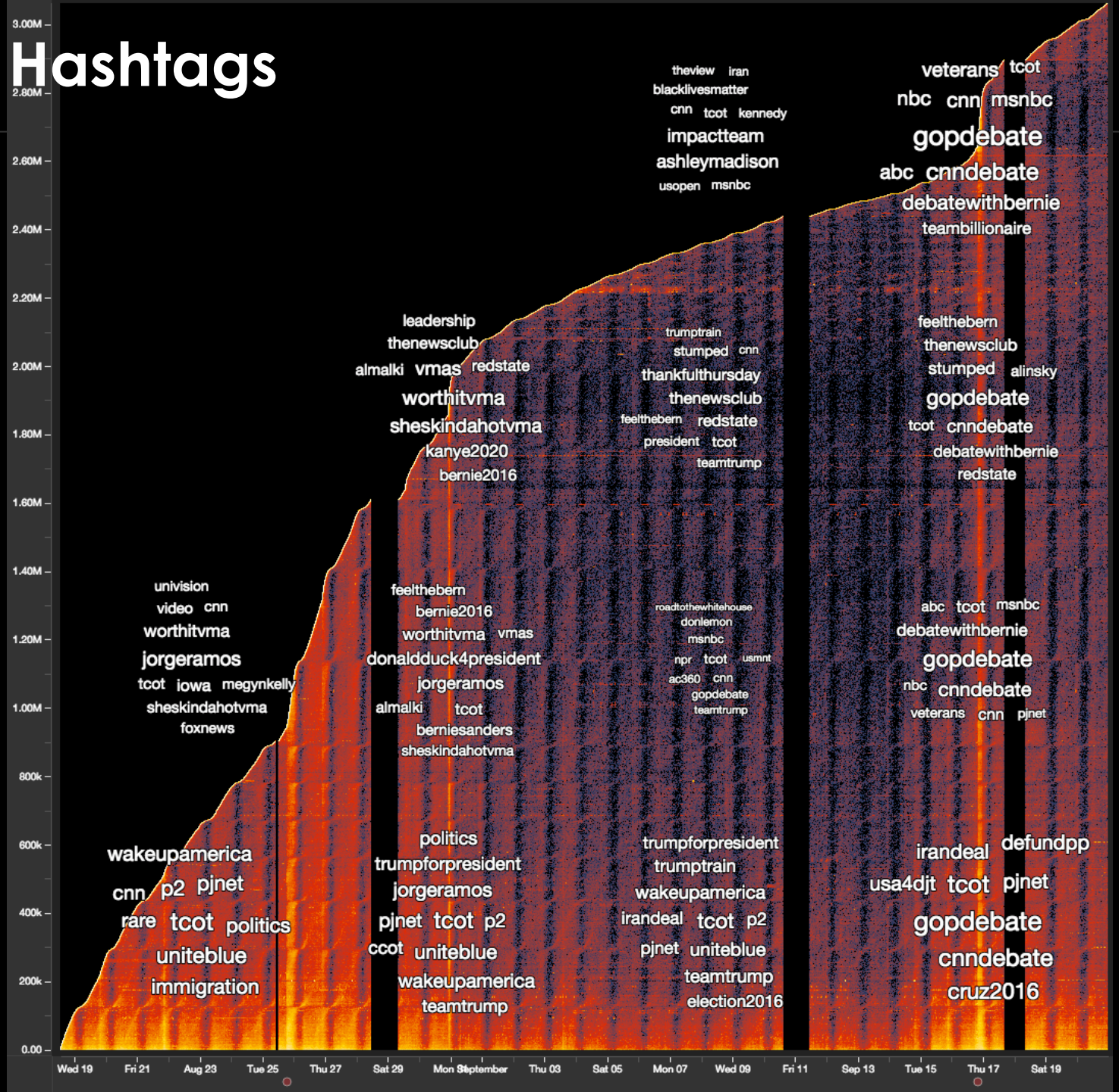
uncharted

Trump Mountain Plot – Cohort Patterns



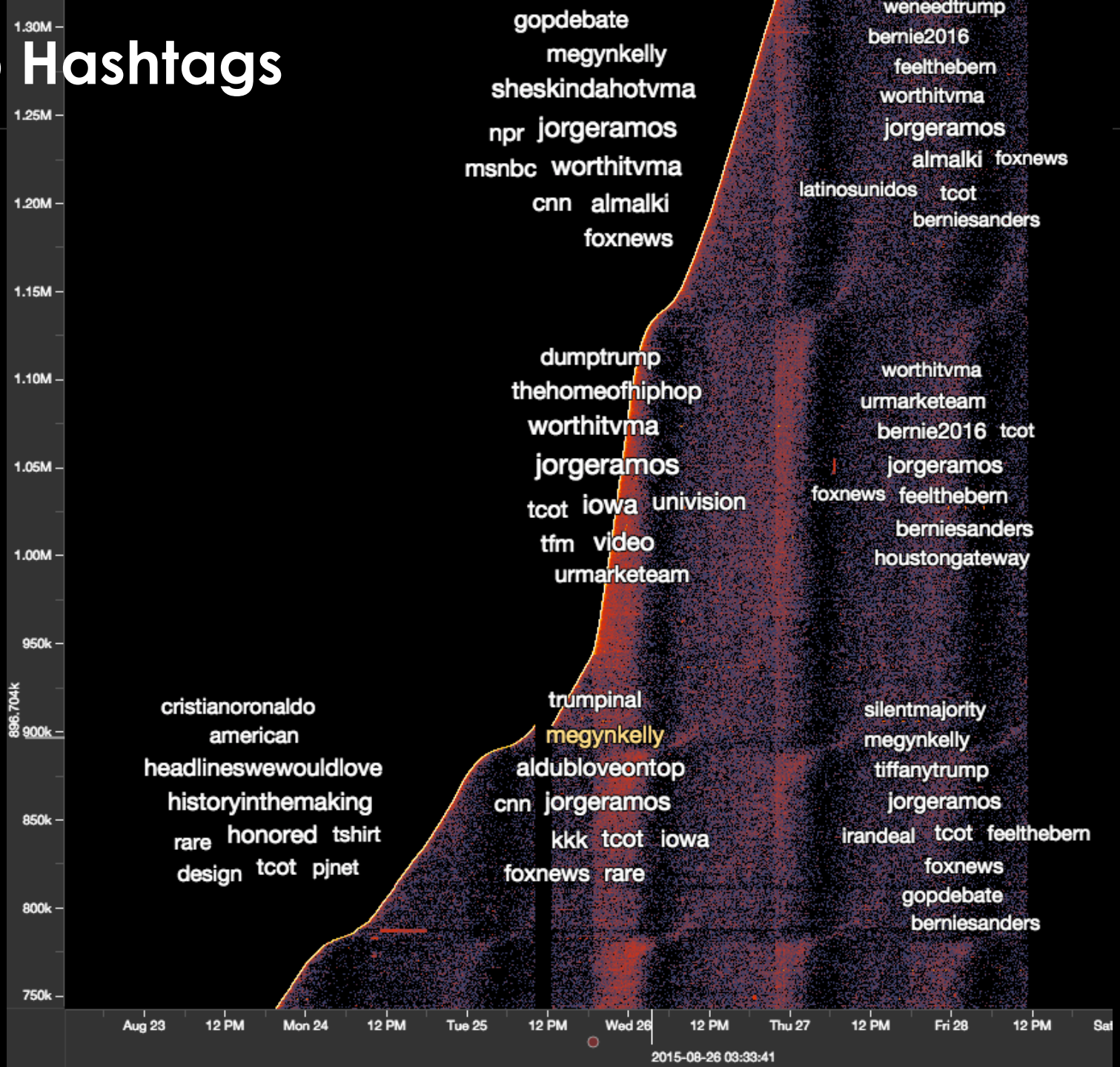
uncharted

Trump Top Hashtags



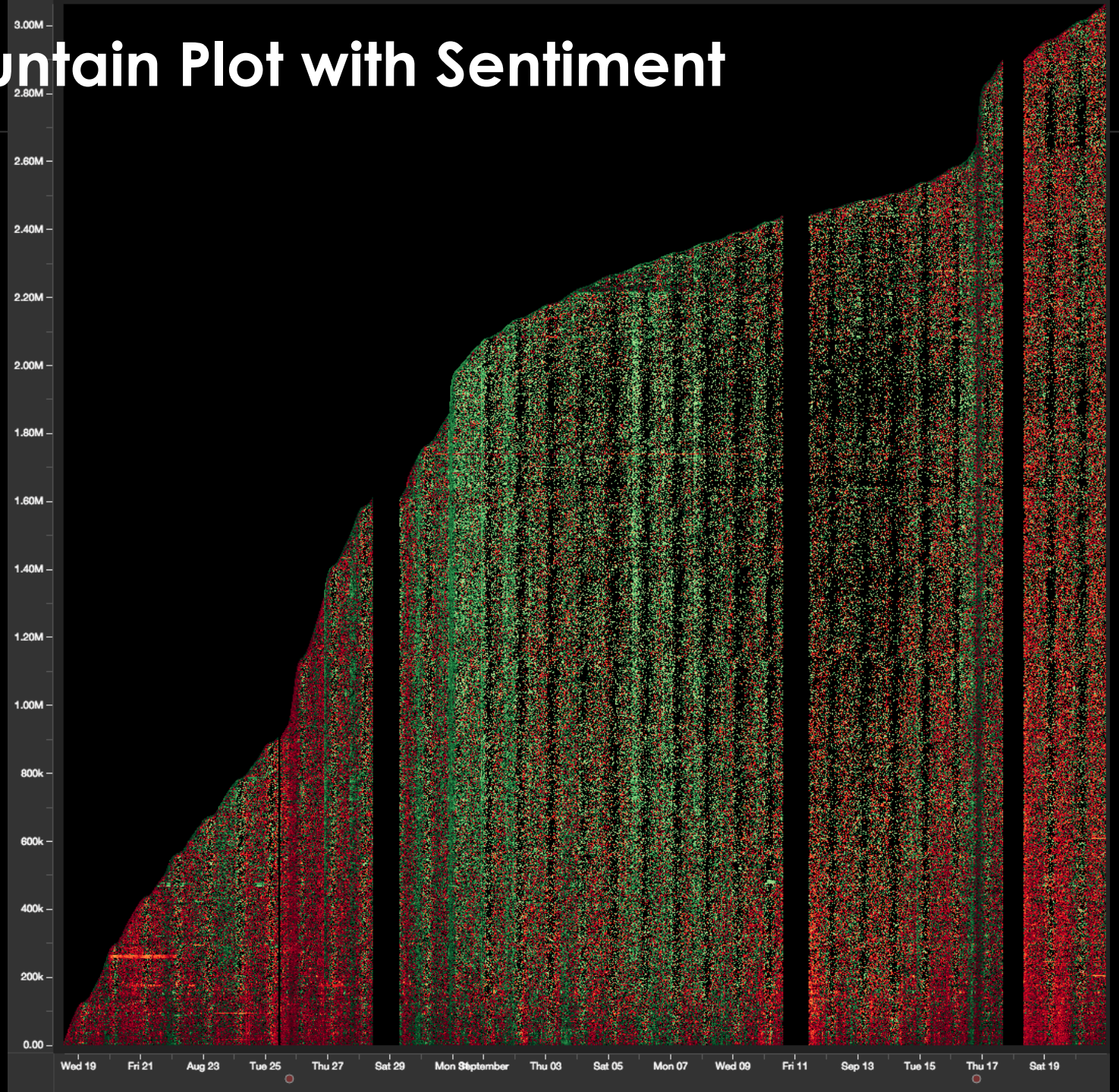
uncharted

Trump Top Hashtags



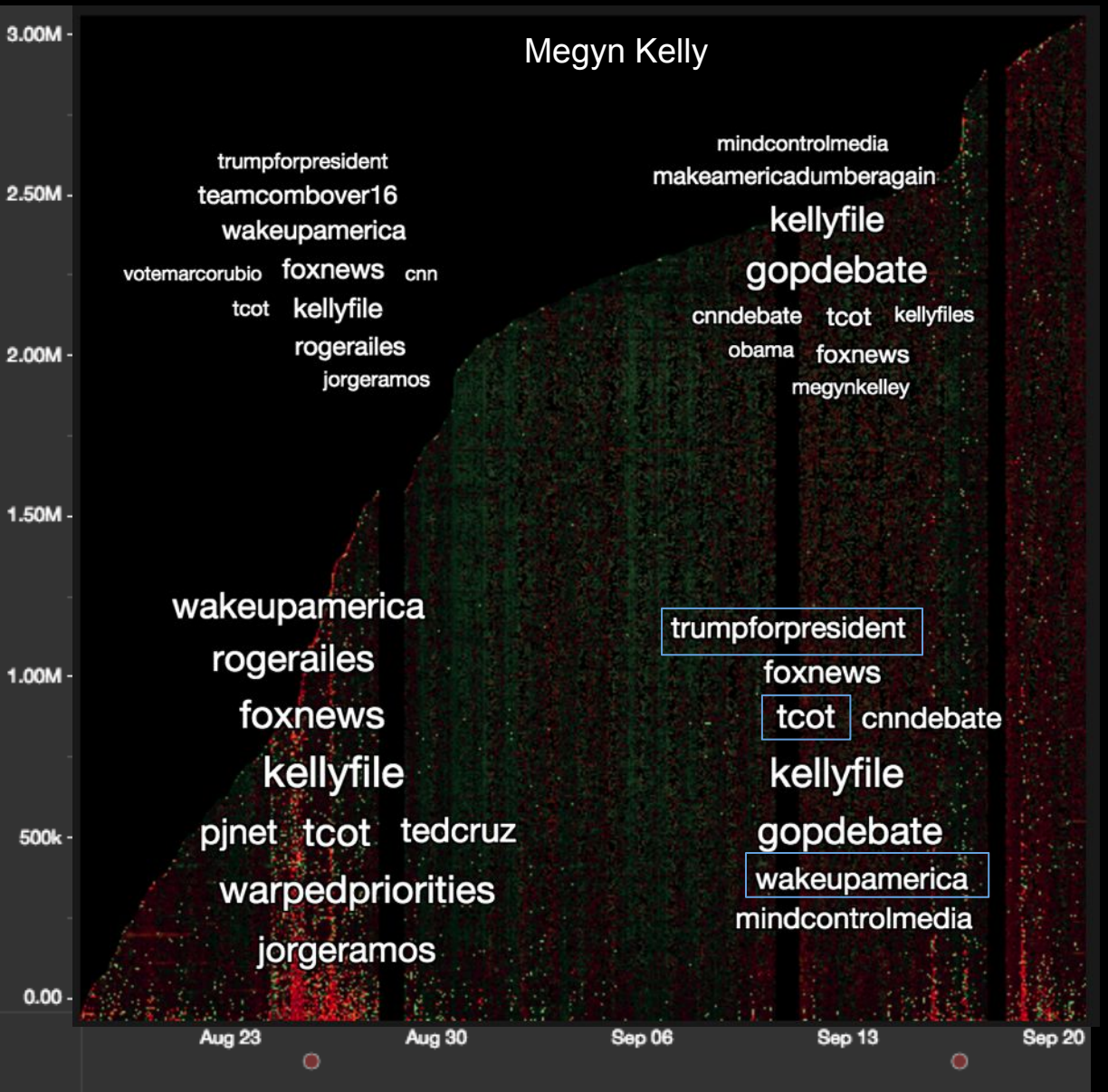
uncharted

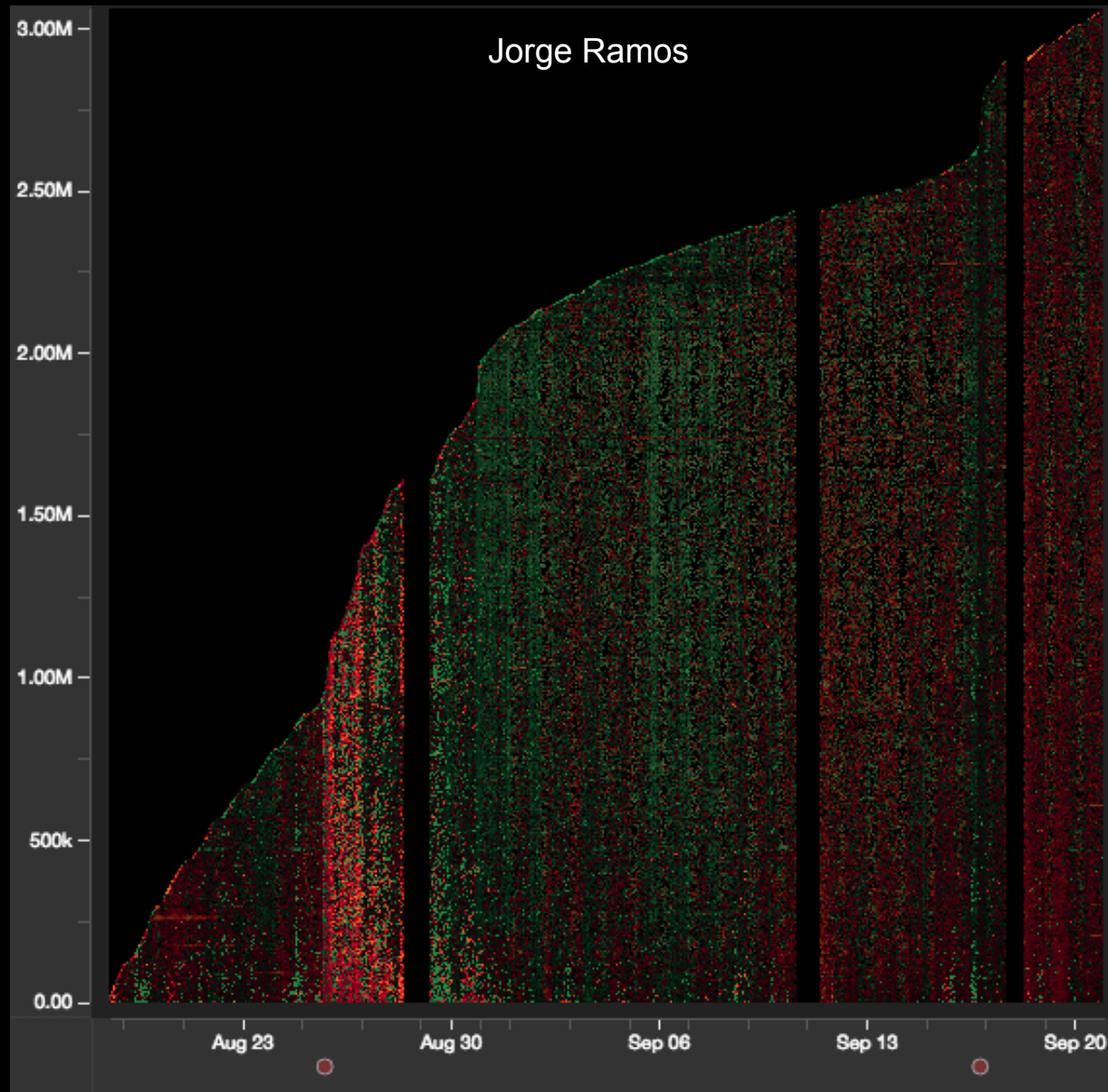
Trump Mountain Plot with Sentiment



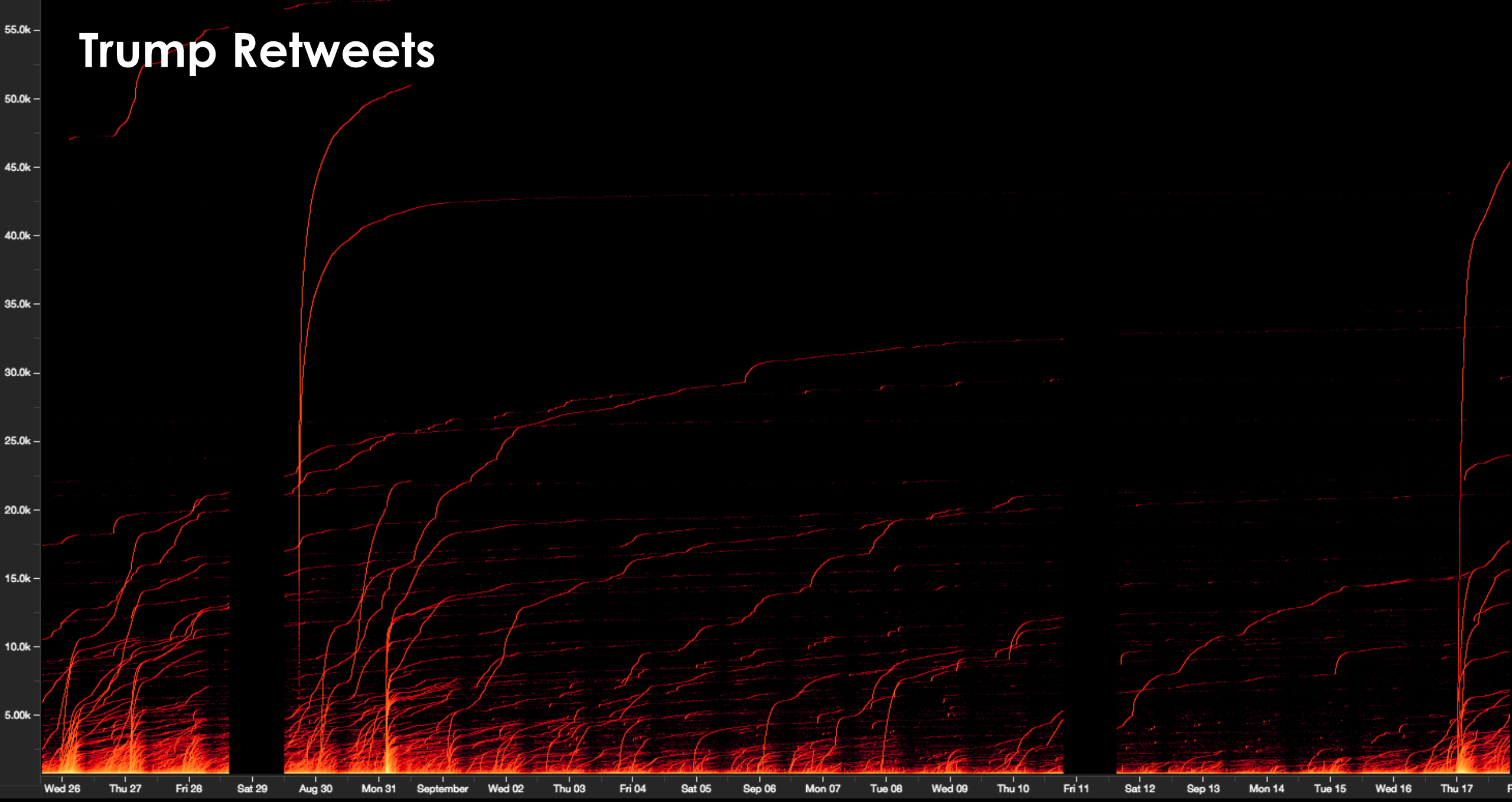
uncharted

Trump Mountain Plot with Sentiment – Megyn / Jorge



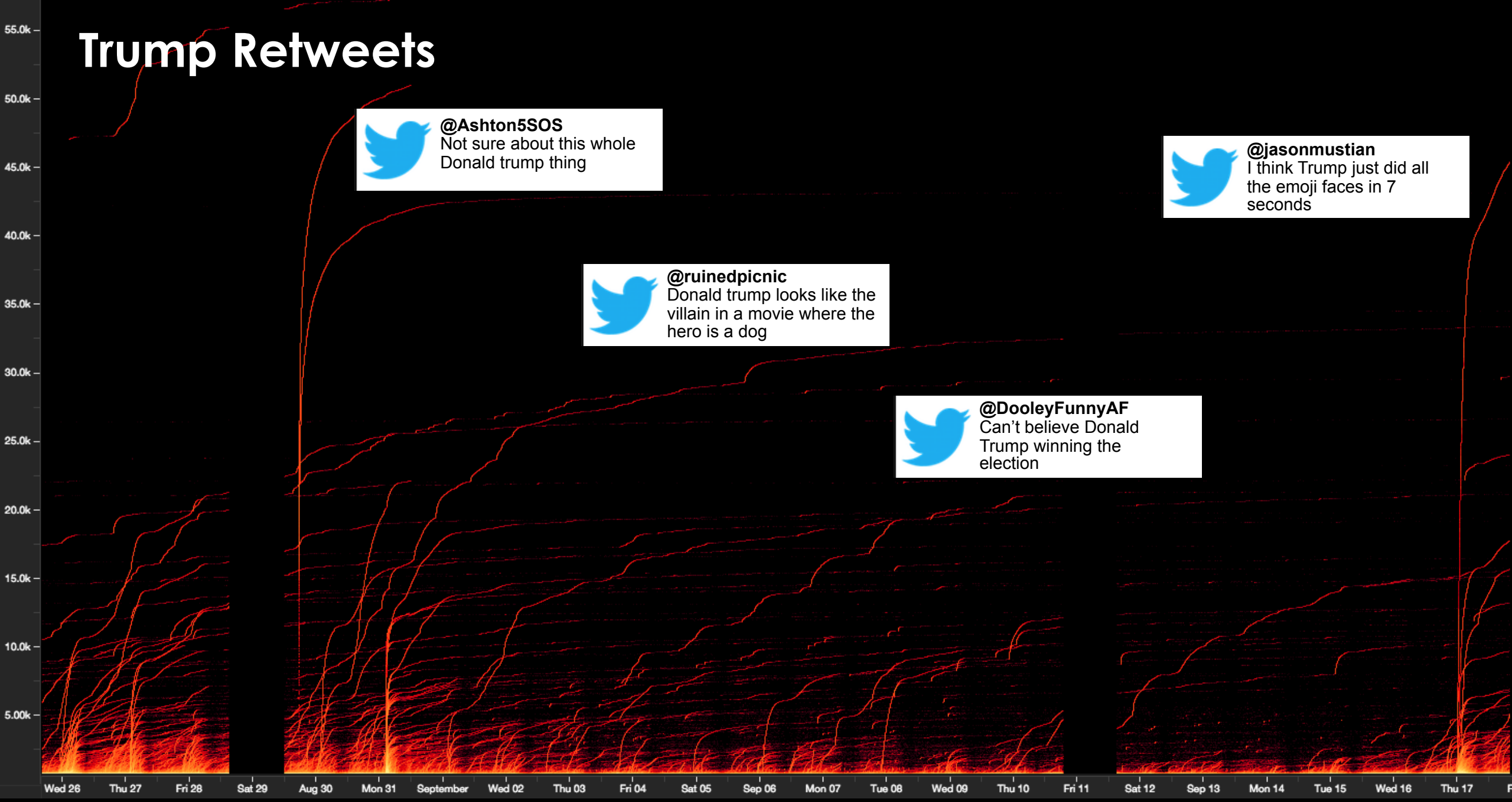


Trump Retweets



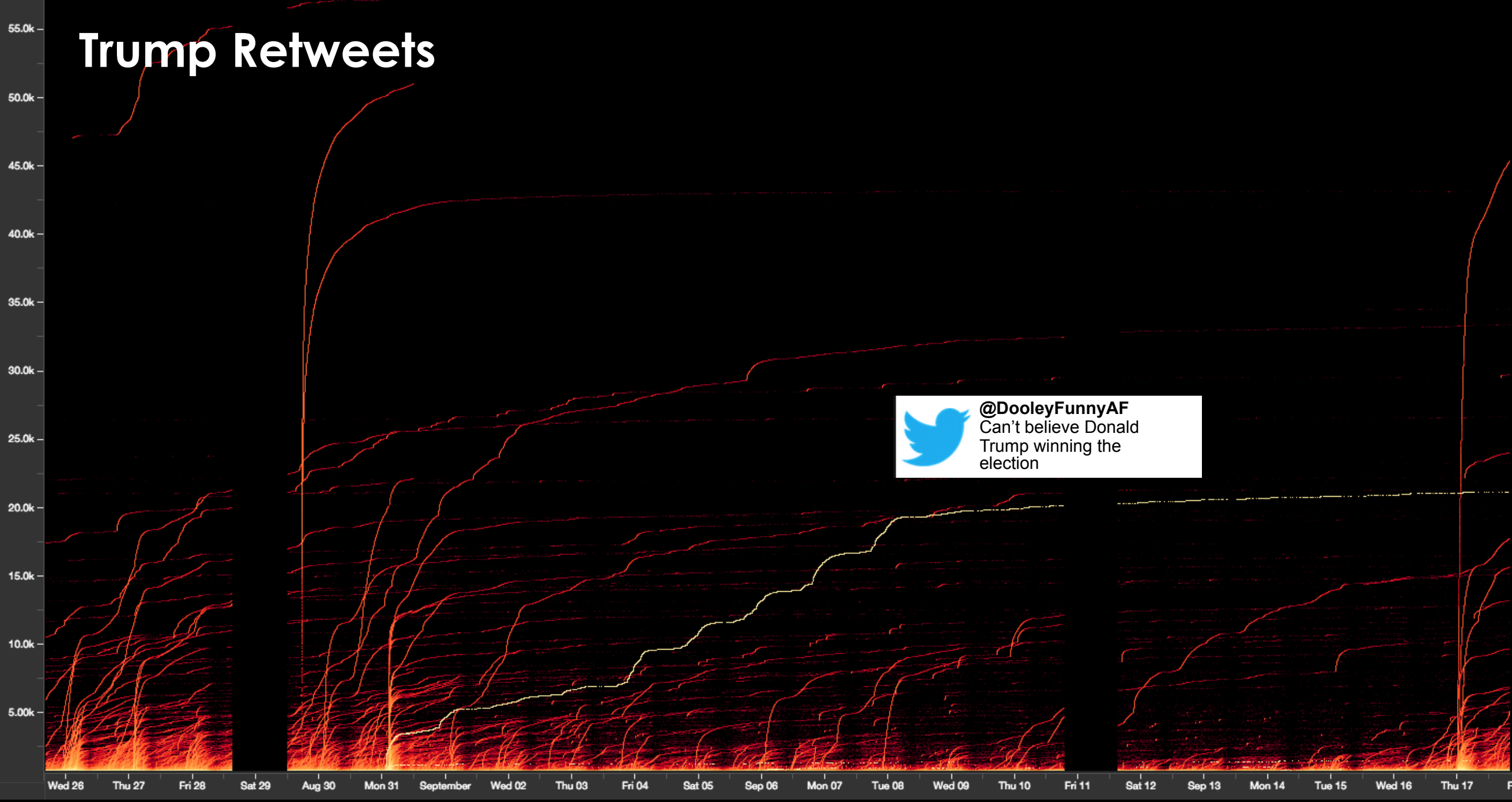
uncharted

Trump Retweets



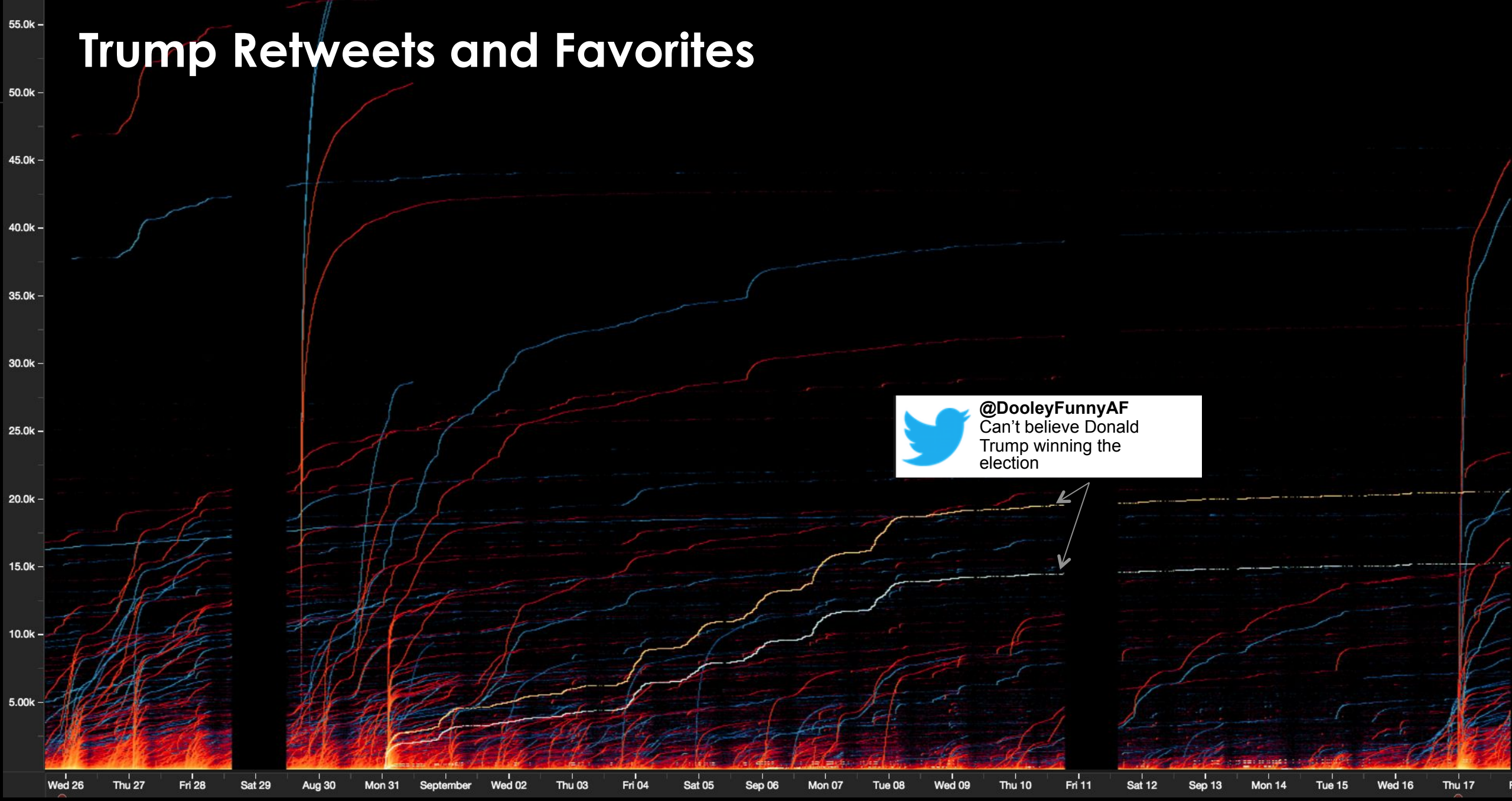
uncharted

Trump Retweets



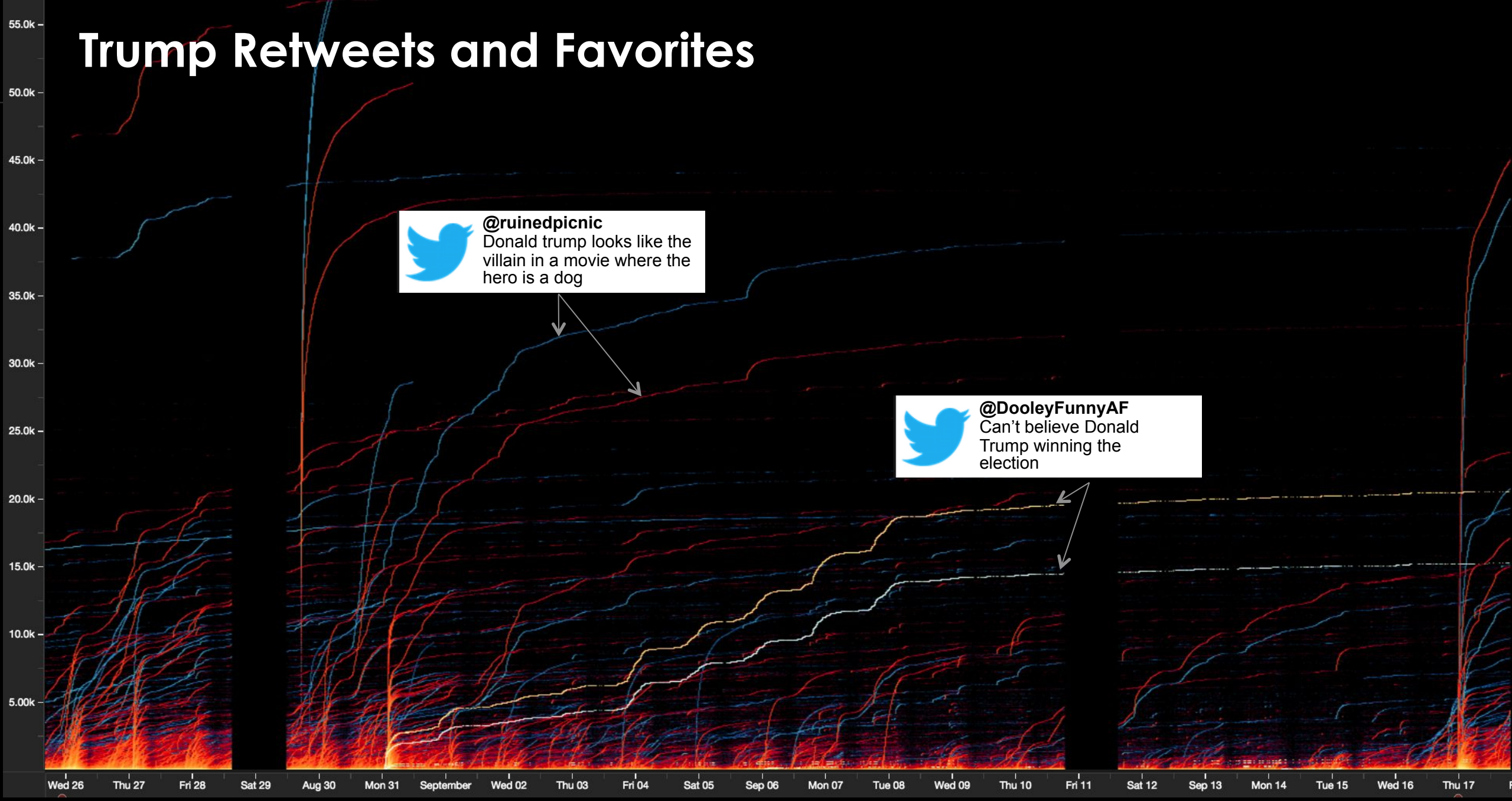
uncharted

Trump Retweets and Favorites



uncharted

Trump Retweets and Favorites

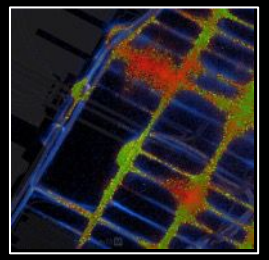


uncharted

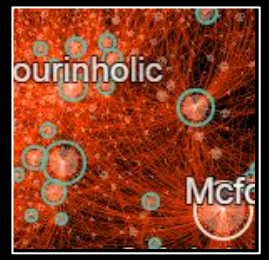
KEY TAKE AWAYS

1. Plot all the data

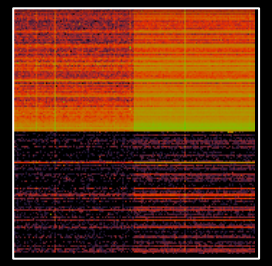
MAP



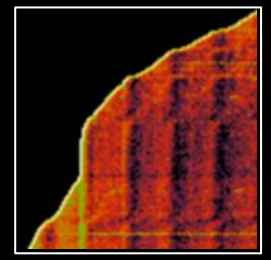
CONNECT



ORDER

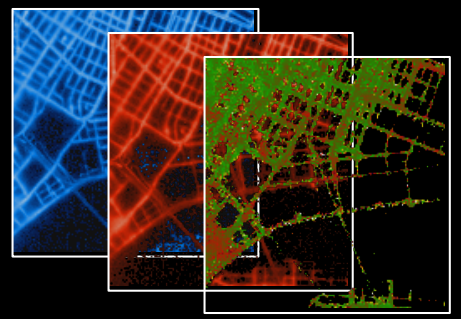


TIME

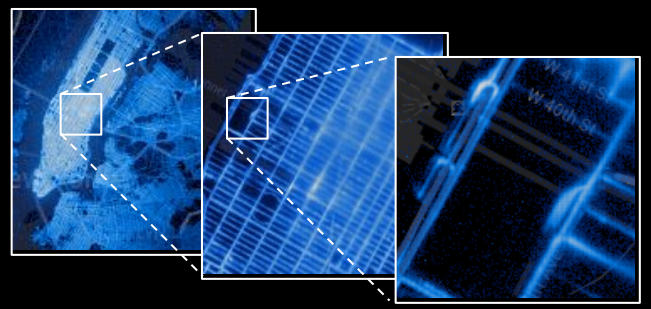


2. Explore it

LAYERS

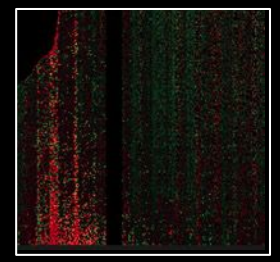


ZOOM



MINE

FILTER



uncharted

More Info

Richard Brath
rbrath@uncharted.software
416-203-3003 x 242

Robert Harper
rharper@uncharted.software
@rdharper