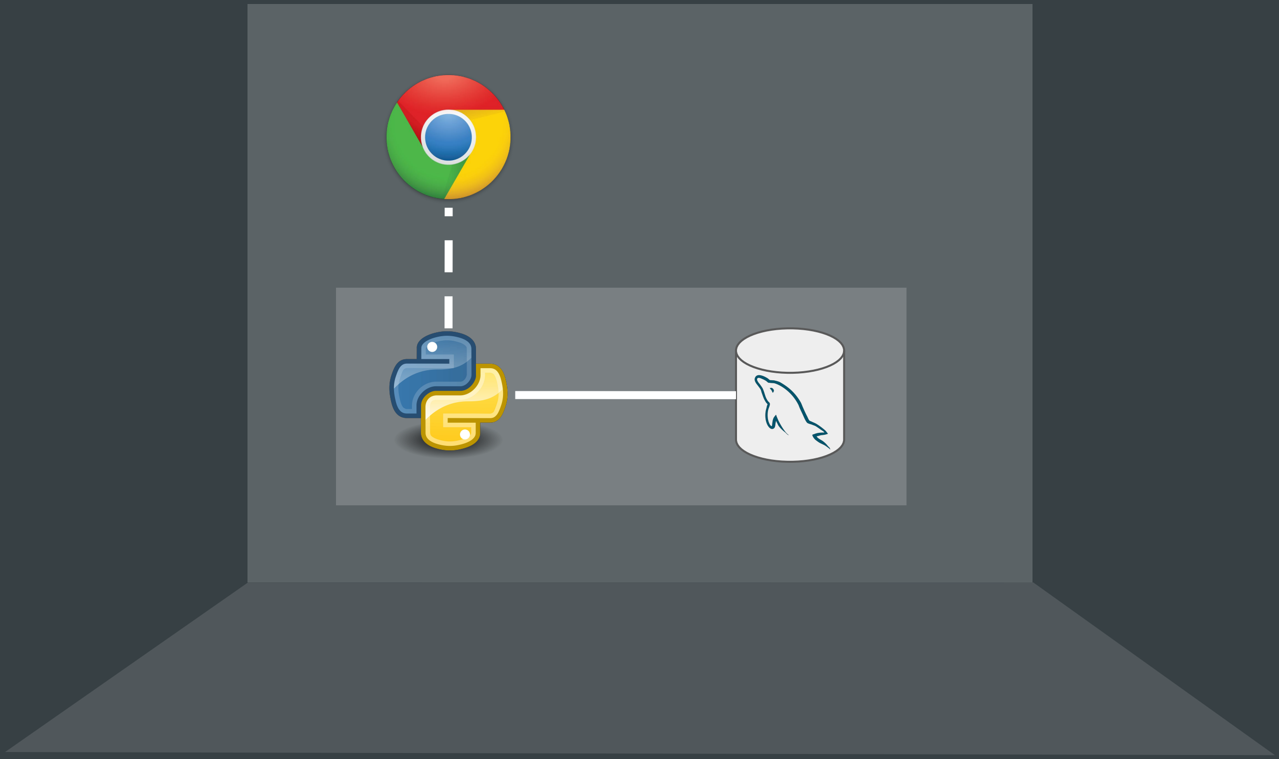
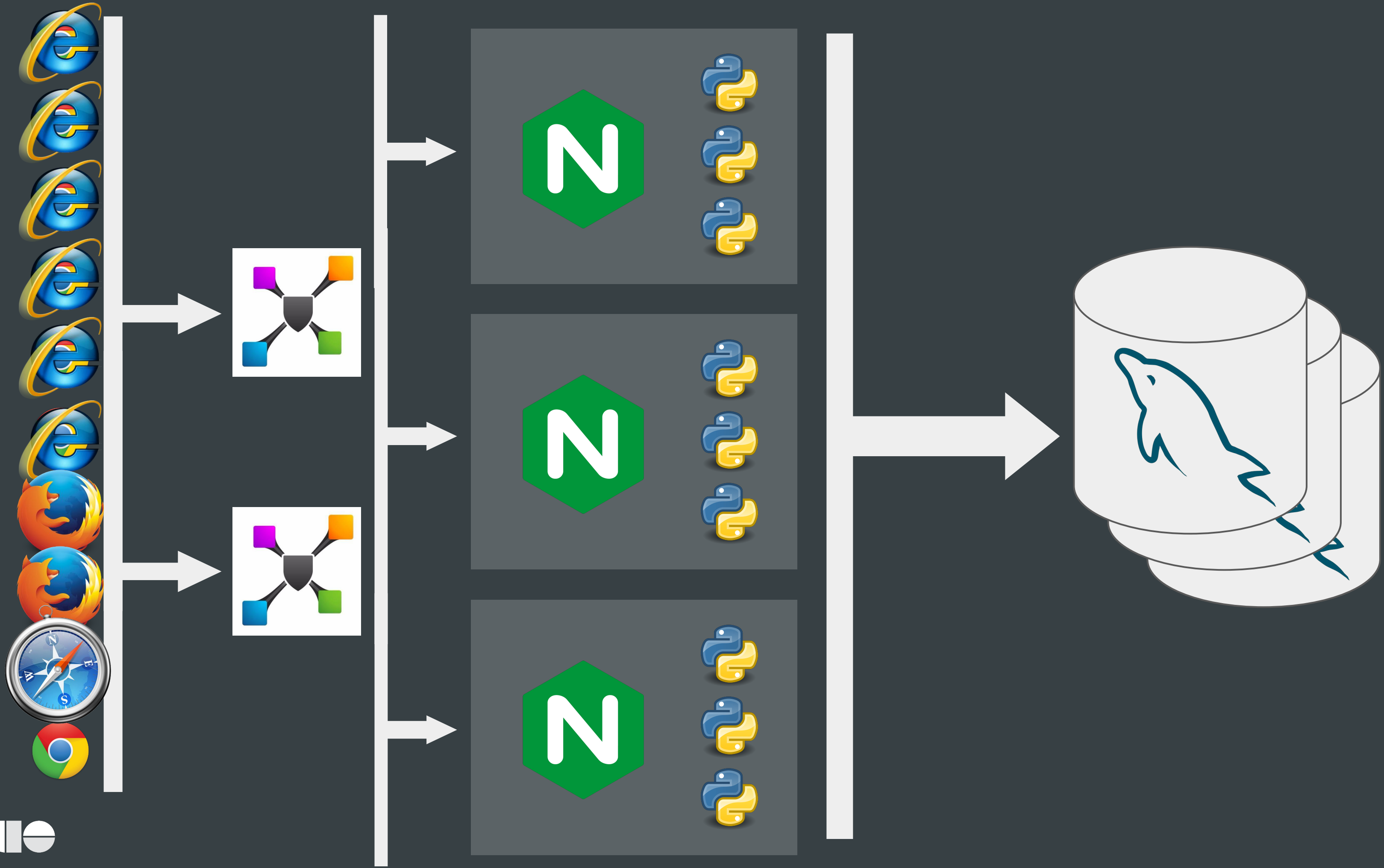
Big(ish) Data

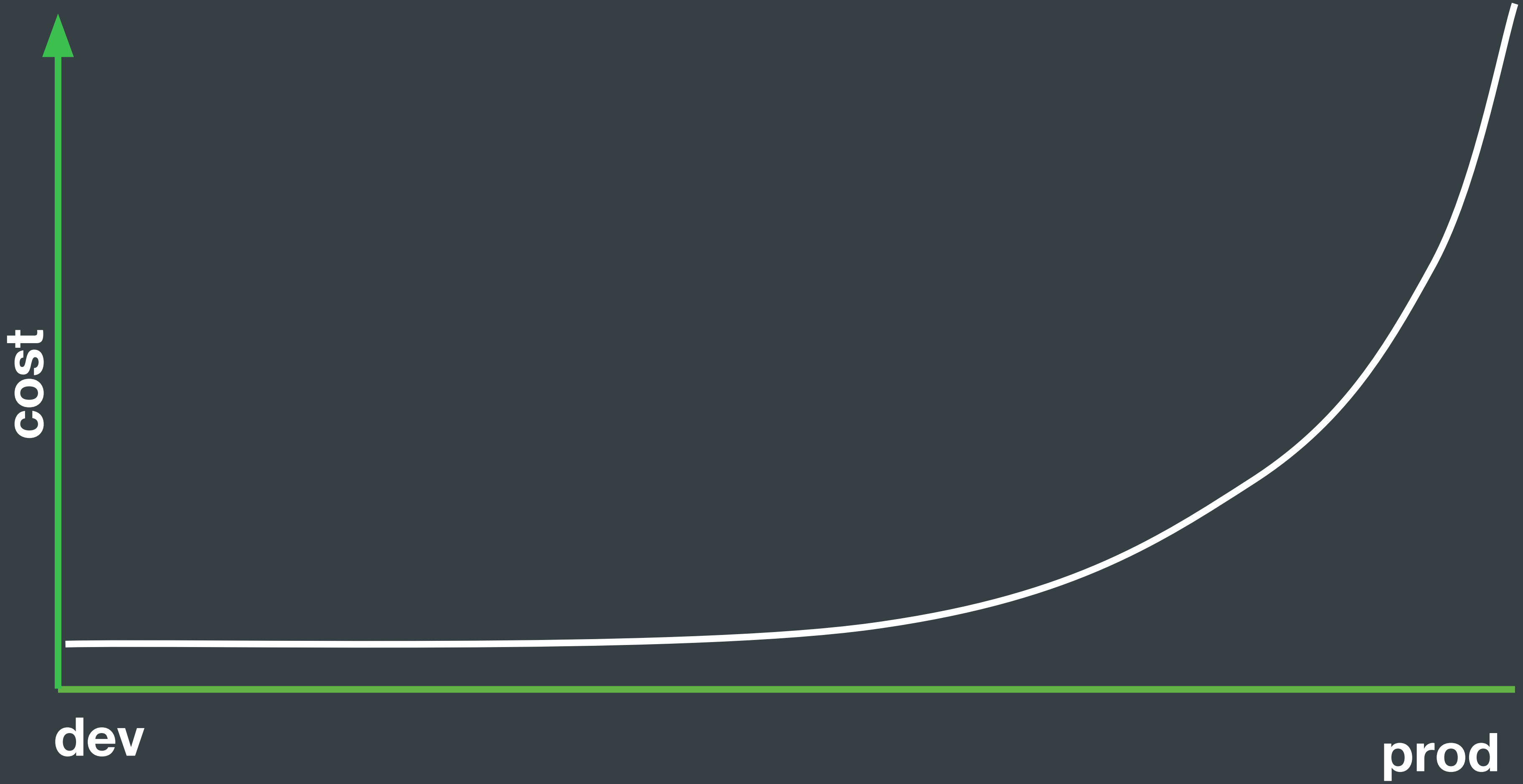
Scaling up customer-facing logs

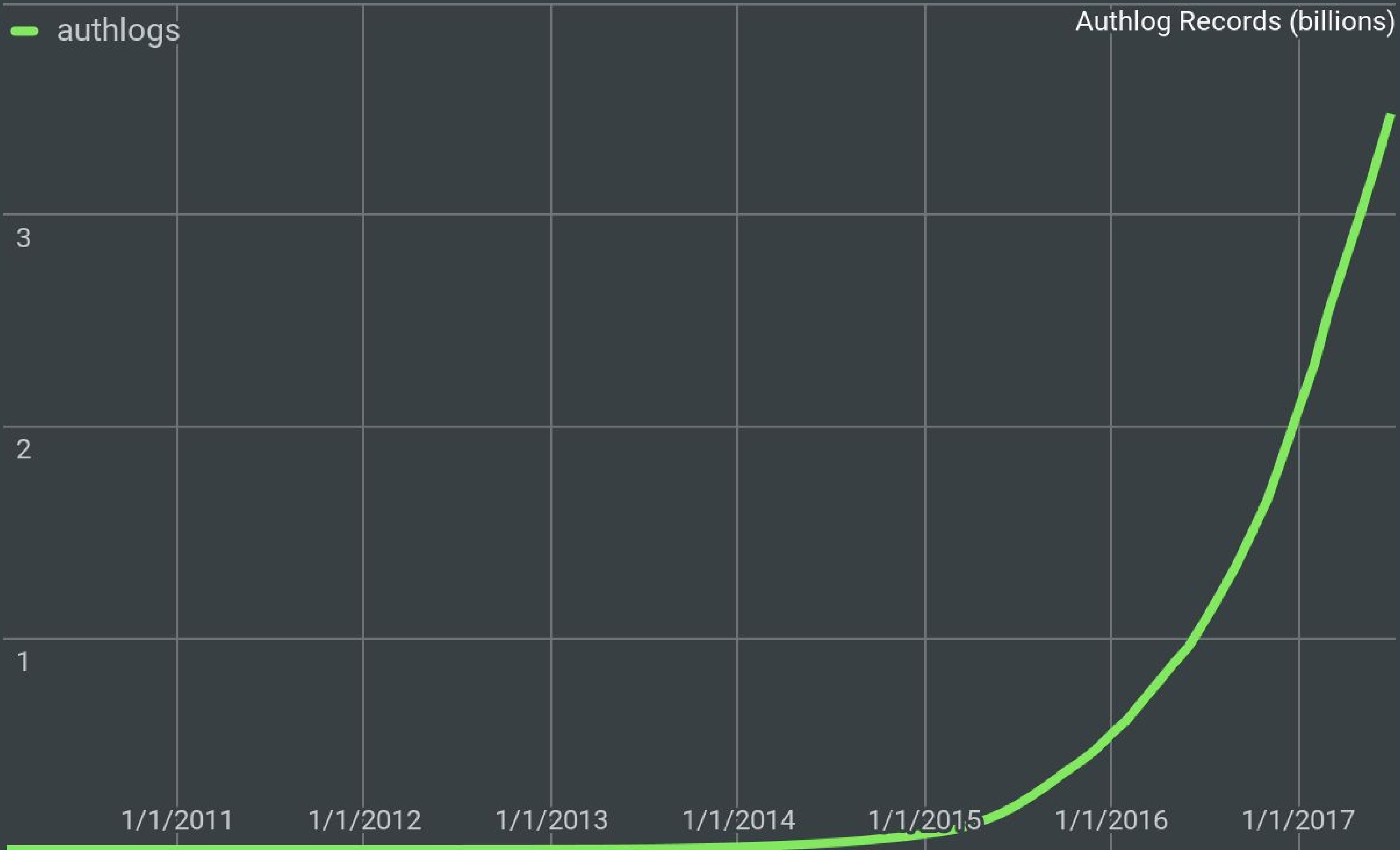
Bryan Witherspoon (@uoodsq)
Principal SE, Duo Security

Victory has defeated you





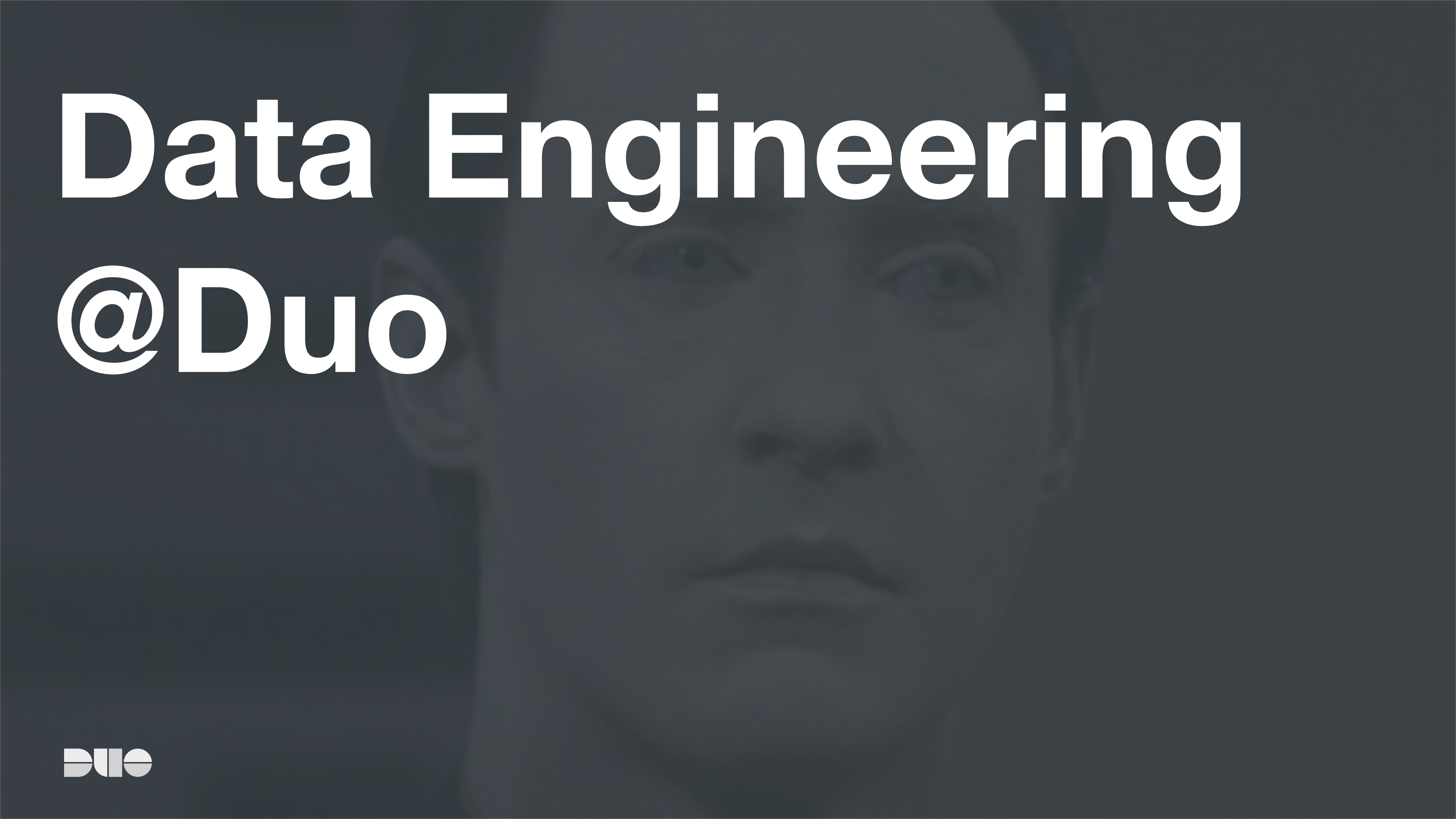

cost
dev
prod


authlogs
Authlog Records (billions)
3
2
1
1/1/2011
1/1/2012
1/1/2013
1/1/2014
1/1/2015
1/1/2016
1/1/2017

DUO

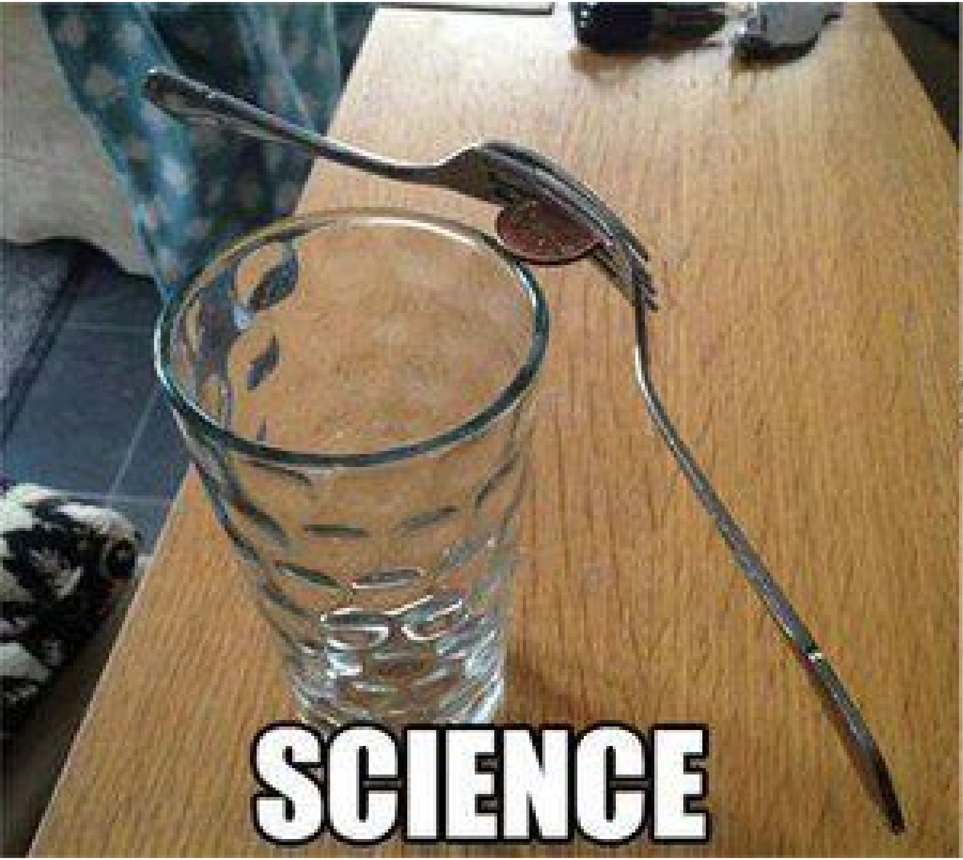
cHRome!
Firefox!
GLUE
Never heard of me?
Okay
shoze
shoze.blogspot.com

Data Engineering @Duo



Data Engineering vs Data Science

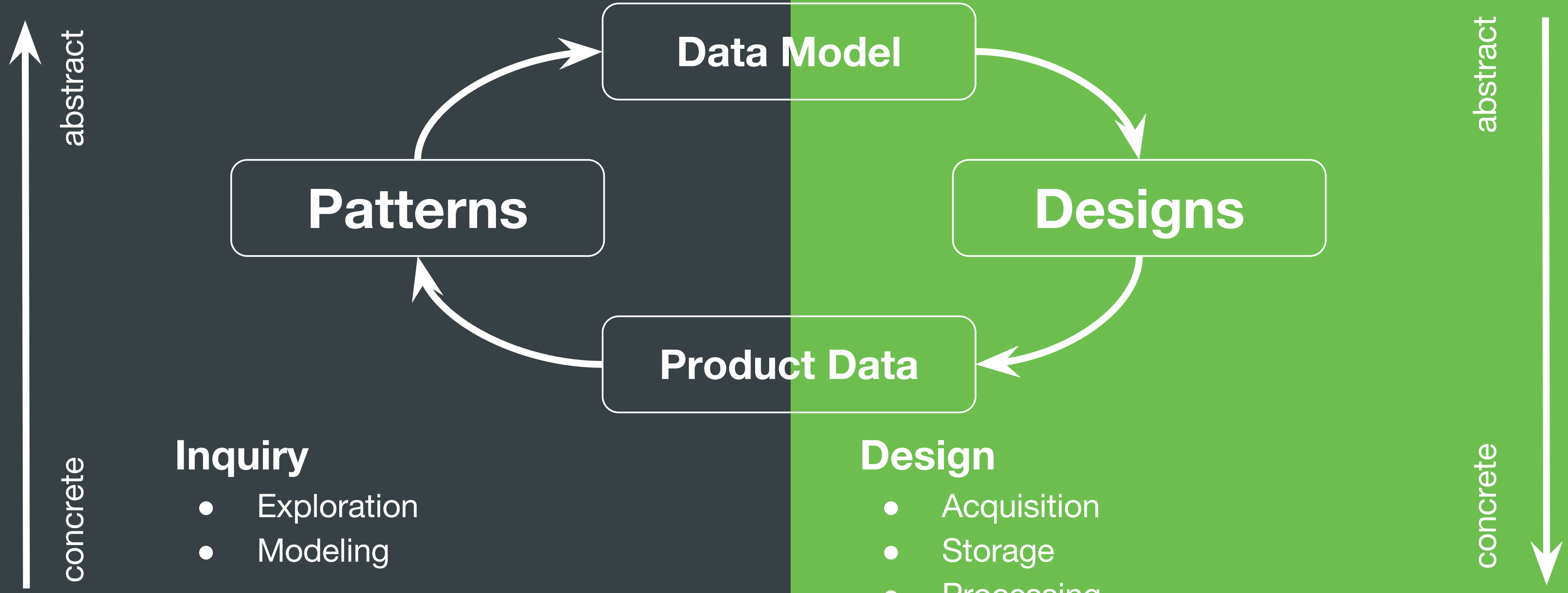
Engineering
vs
Science


SCIENCE

ENGINEERING

Data Science

Data Engineering



Inquiry

- Exploration
- Modeling

Design

- Acquisition
- Storage
- Processing

https://www.farnamstreetblog.com/2013/07/the-difference-between-science-and-engineering/

Data Science

- Modeling
- Machine Learning

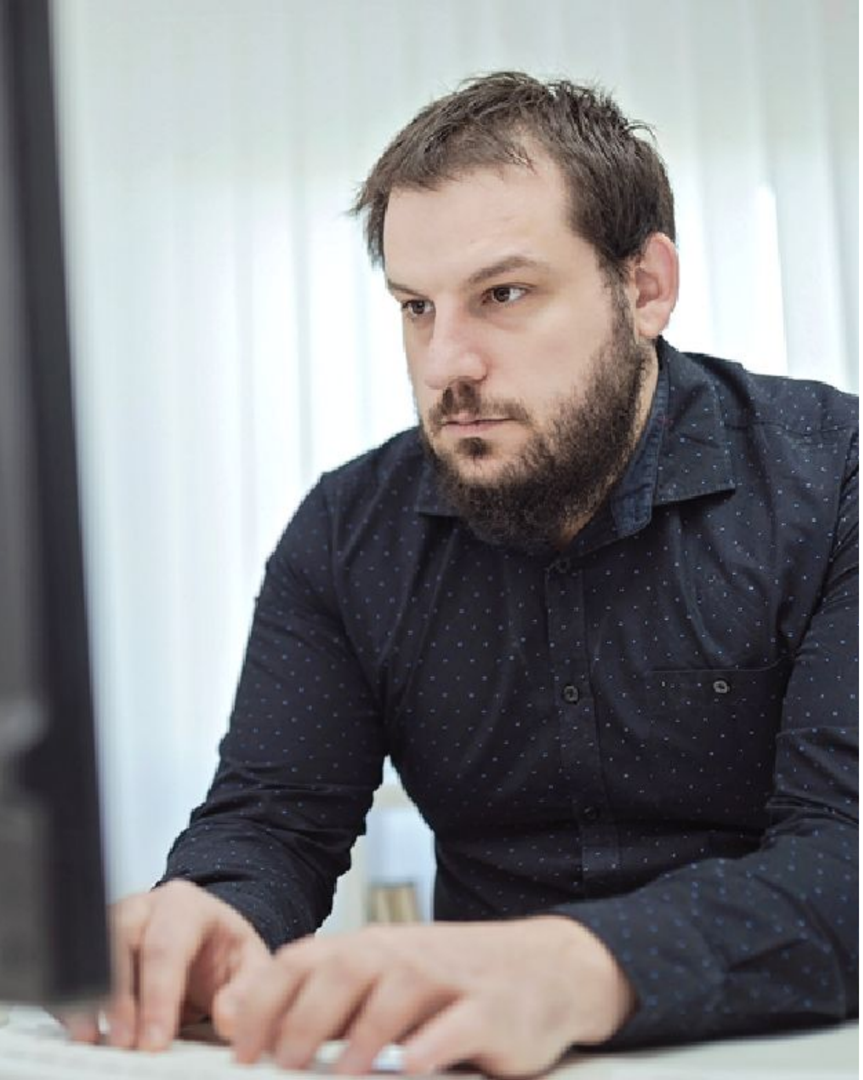
Data Engineering

- Storage
- Pipeline
- Visualization

users

devices

applications



Gary

IT administrator

- Data security is just portion of his job
- Improves technology and IT services
- Wants to focus on exceptions: "what's out of the ordinary"
- Needs to maintain reputation for recommending good technology

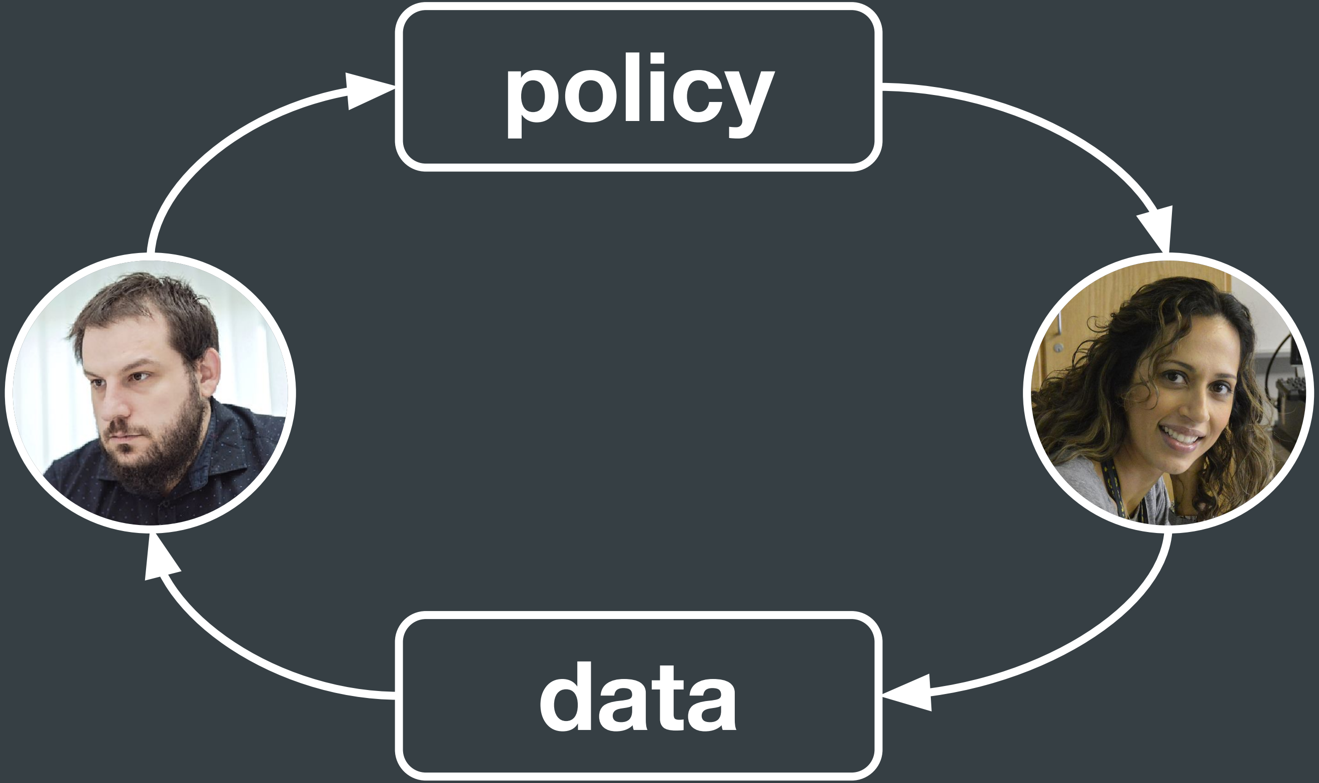
> The company counts on me to keep technology running, keep it secure, and keep people happy.

Eve

End user

- Wants to use applications and data whenever she needs to
- Frustrated when security stuff gets in her way or adds friction
- Not an expert at security or technology

I know security is important, but honestly, it’s never a top thing on my mind.
At best, it’s an afterthought.



policy

data

auth

lotsa data

👤

- name
- groups

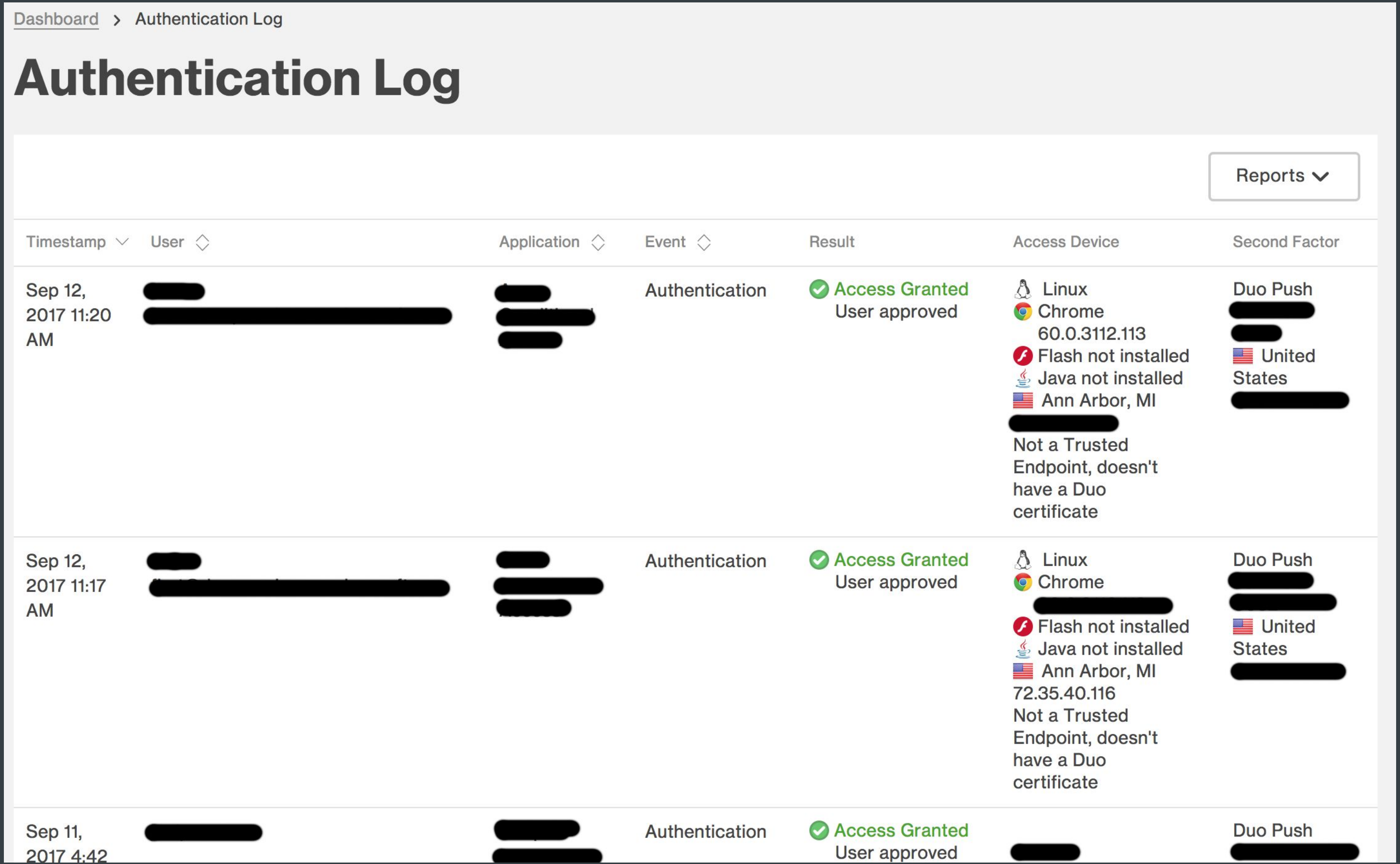
📱 🖥

- operating system
- browser
- plugins
- security features
- ip address → geo + rep

🚀

- name
- type

- id (monotonically increasing + uuid)
- timestamp
- customer id
- data…

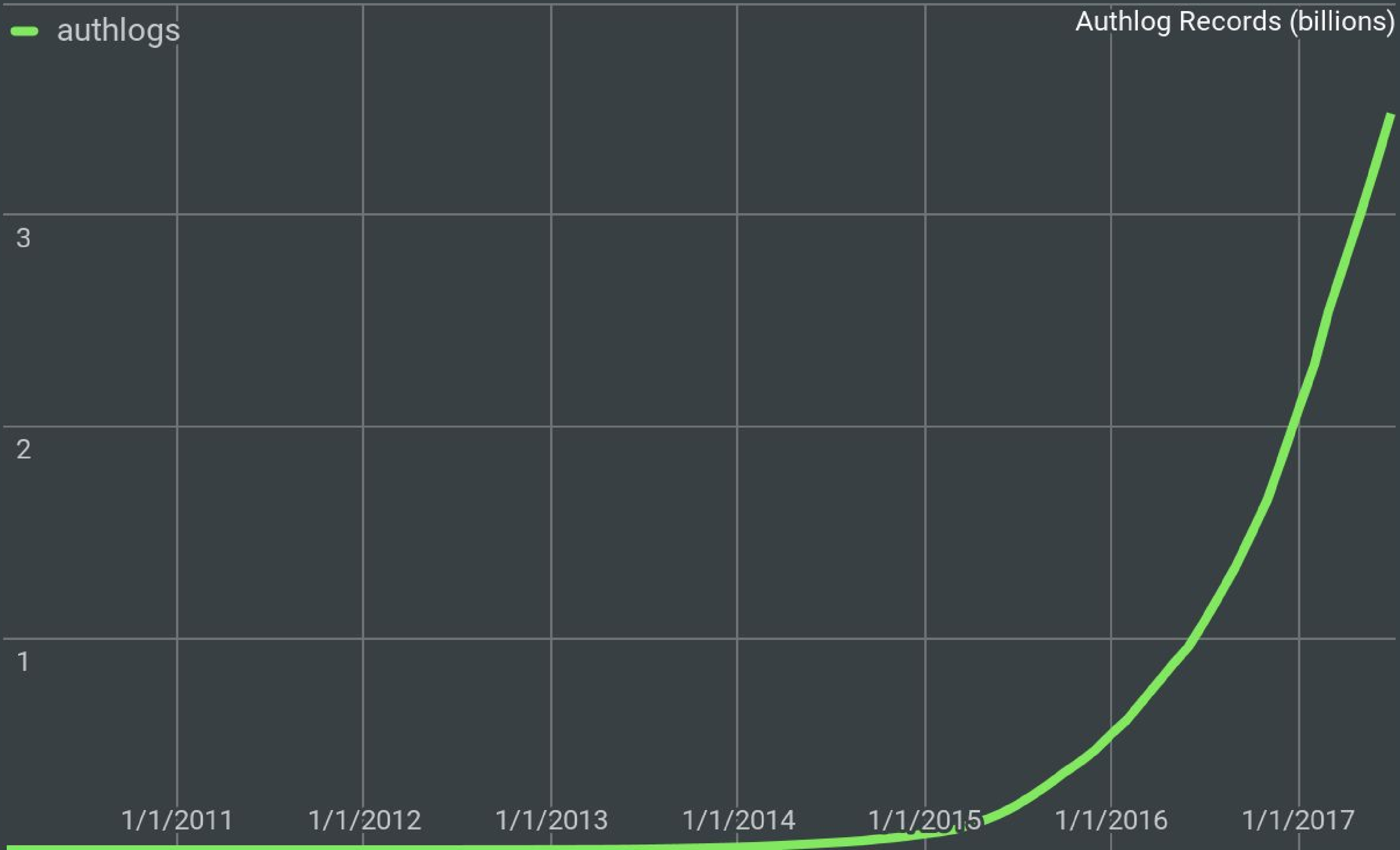


Dashboard > Authentication Log

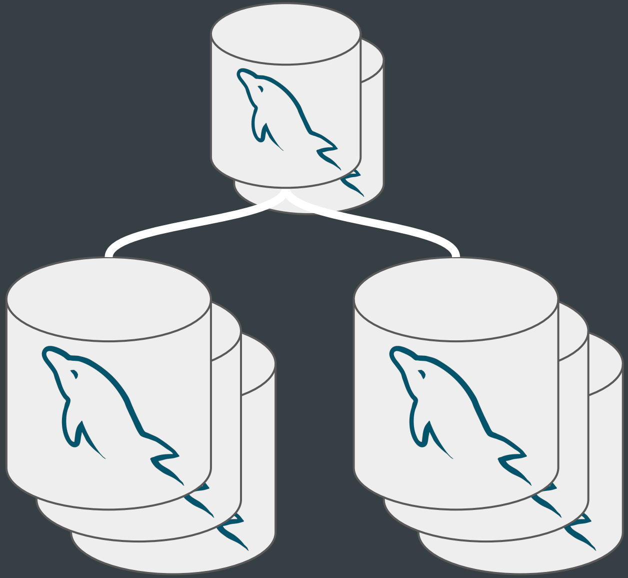
Authentication Log

Timestamp	User	Application	Event	Result	Access Device	Second Factor
Sep 12, 2017 11:20 AM	[redacted]	[redacted]	Authentication	✓ Access Granted User approved	Linux Chrome 60.0.3112.113 Flash not installed Java not installed Ann Arbor, MI [redacted] Not a Trusted Endpoint, doesn't have a Duo certificate	Duo Push [redacted] United States [redacted]
Sep 12, 2017 11:17 AM	[redacted]	[redacted]	Authentication	✓ Access Granted User approved	Linux Chrome [redacted] Flash not installed Java not installed Ann Arbor, MI 72.35.40.116 Not a Trusted Endpoint, doesn't have a Duo certificate	Duo Push [redacted] United States [redacted]
Sep 11, 2017 4:42	[redacted]	[redacted]	Authentication	✓ Access Granted User approved	[redacted]	Duo Push [redacted]

```
SELECT * FROM authlog WHERE
user_name = ?
```


authlogs
Authlog Records (billions)
3
2
1
1/1/2011
1/1/2012
1/1/2013
1/1/2014
1/1/2015
1/1/2016
1/1/2017

- read-only replicas
- sharding
- ~~indexing~~

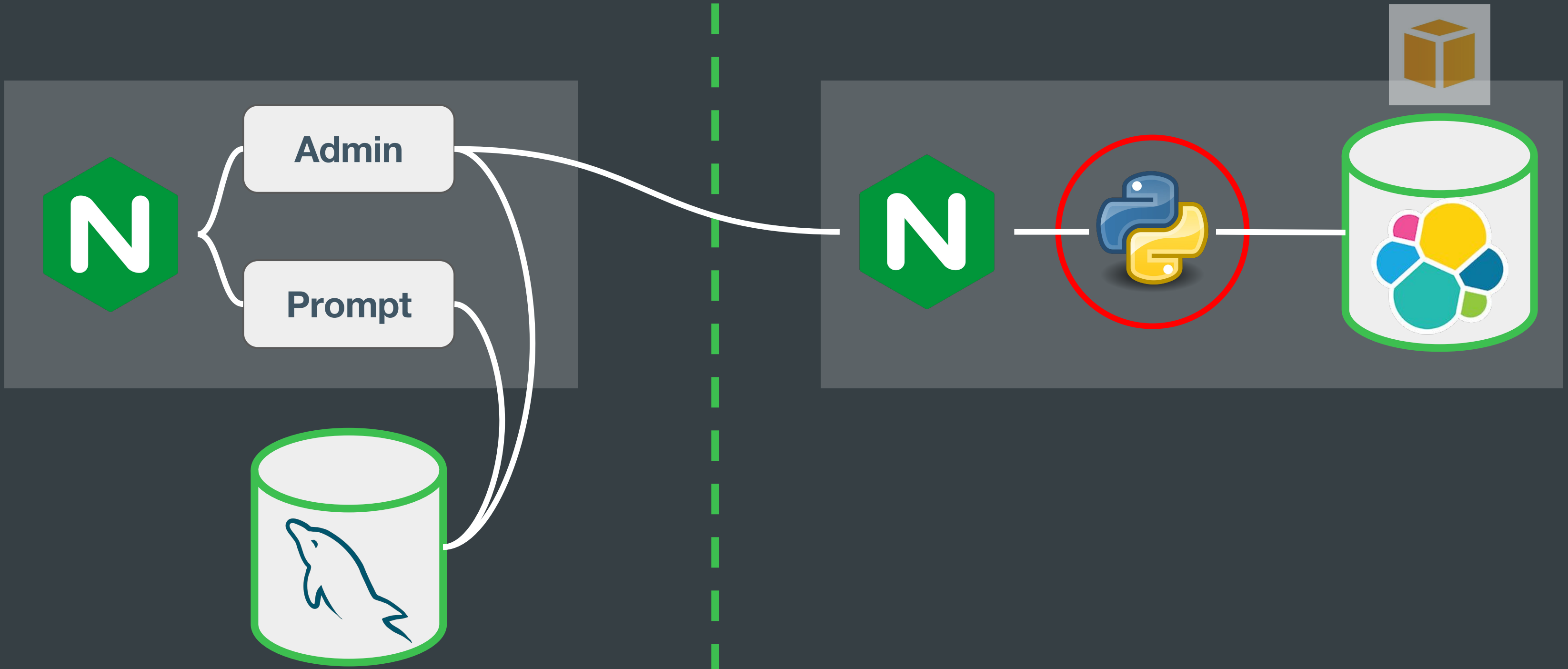


1. **create read replica**
2. **migrate the replica**
3. **switch over**

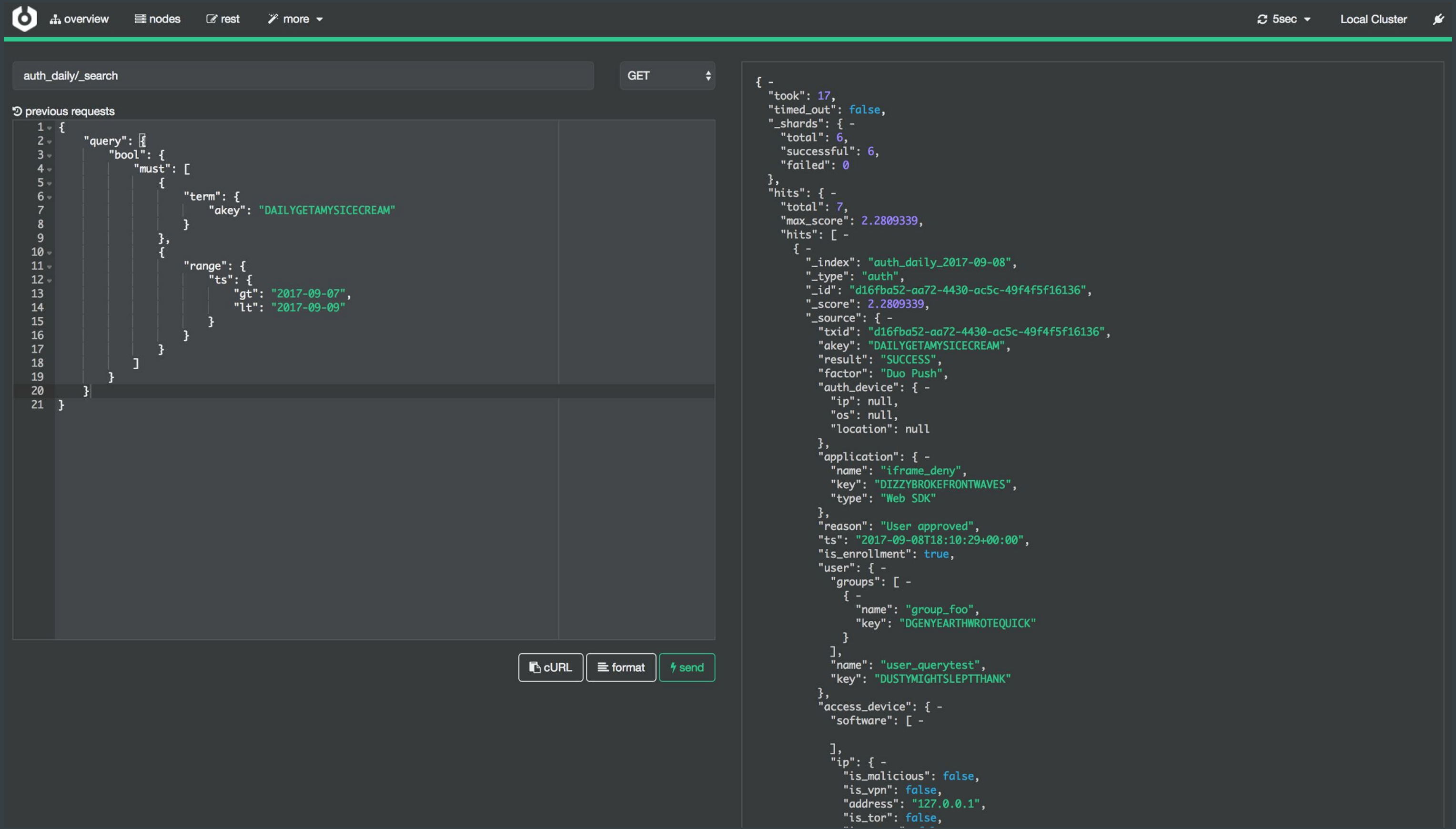
Goal 1:
Improve storage

- Flexible, RESTish API
- Zero-downtime migrations
- In-house operational experience

- Security yikes
- Brittle


Admin
Prompt

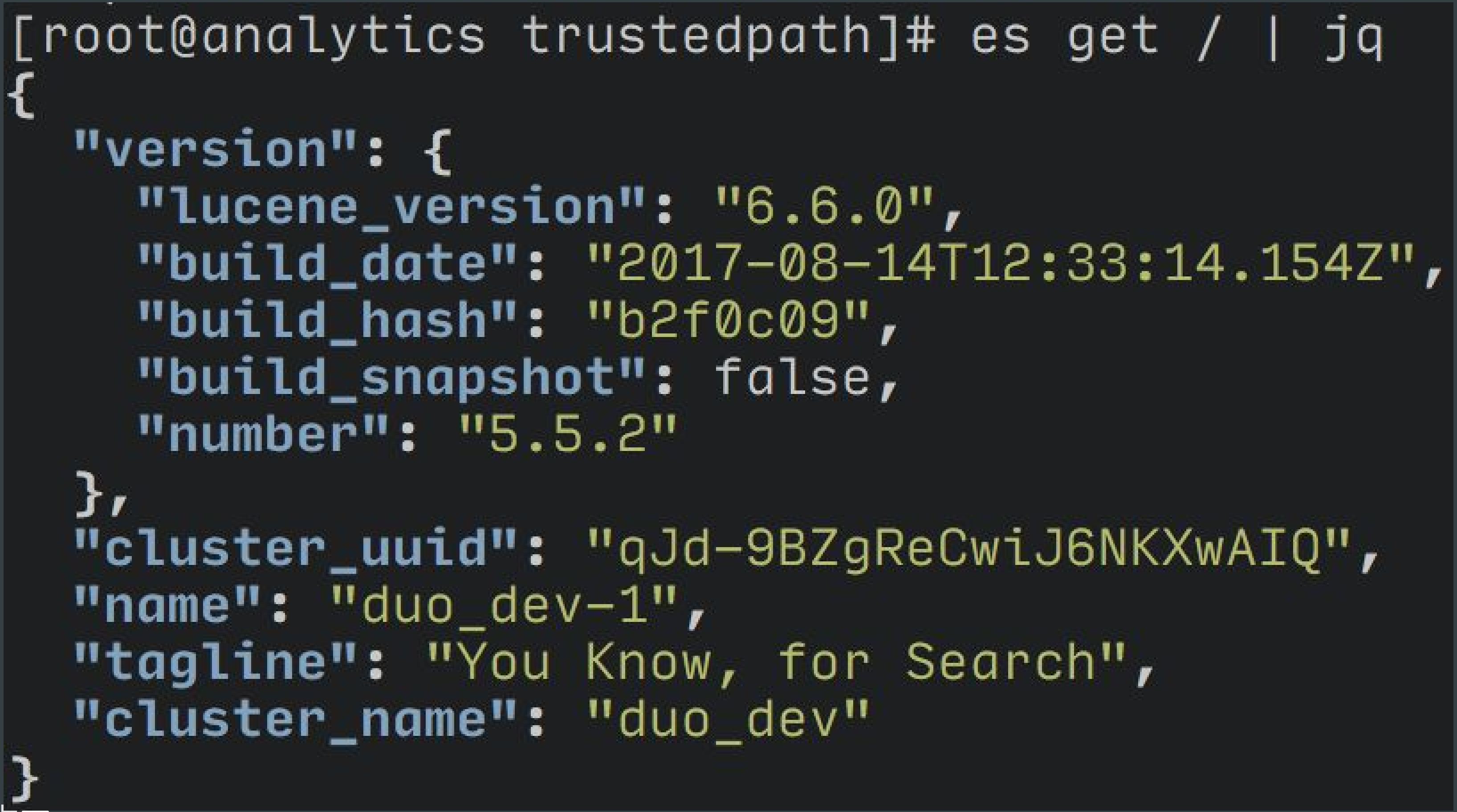
- control all the things!
- python2 + twisted
  - event driven, like node or asyncio
  - good for web applications
  - steep learning curve



auth_daily/_search

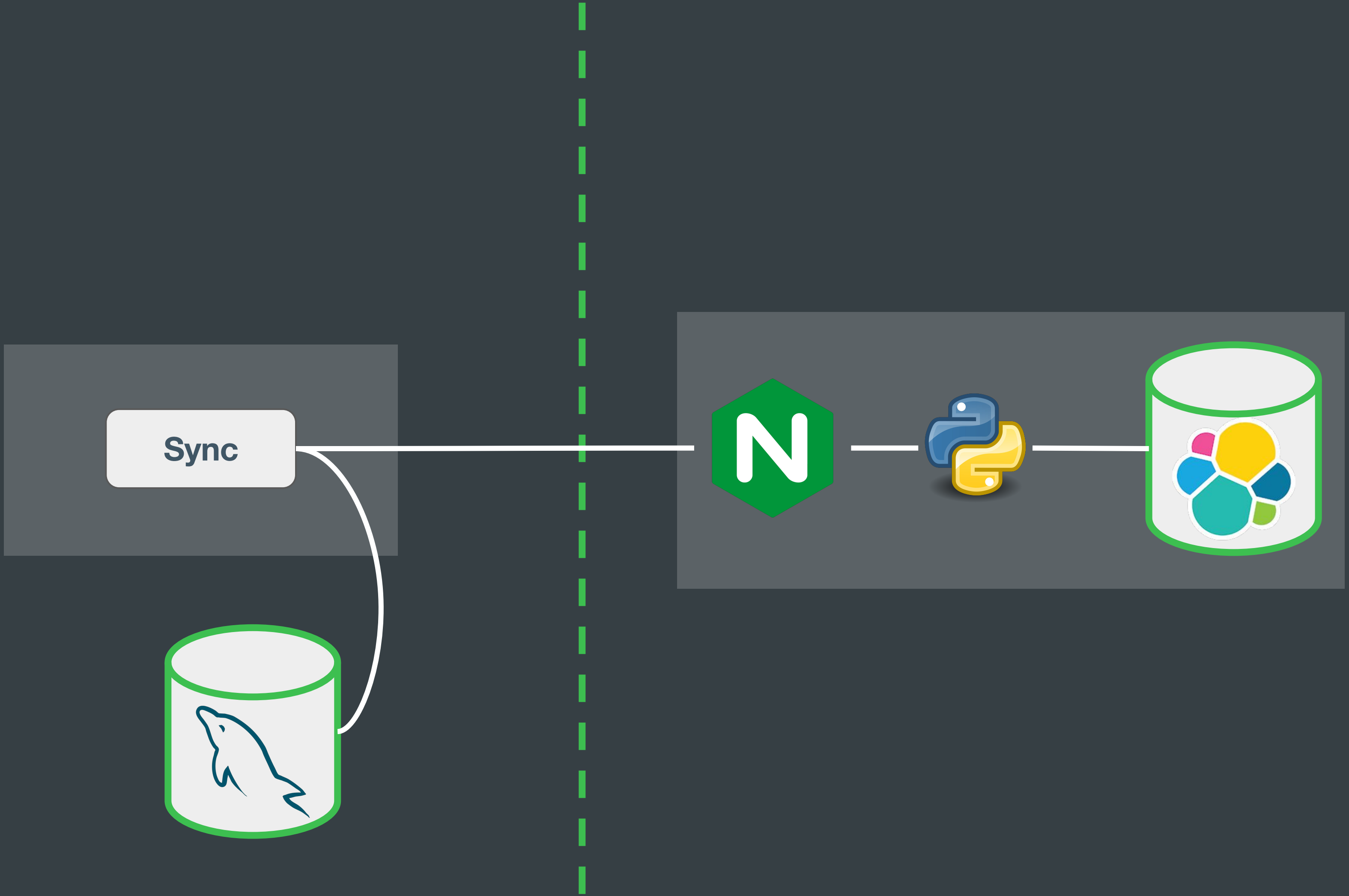
```
{
    "query": {
        "bool": {
            "must": [
                {
                    "term": {
                        "akey": "DAILYGETAMYSICECREAM"
                    }
                },
                {
                    "range": {
                        "ts": {
                            "gt": "2017-09-07",
                            "lt": "2017-09-09"
                        }
                    }
                }
            ]
        }
    }
}
```

```
{ -
  "took": 17,
  "timed_out": false,
  "_shards": { -
    "total": 6,
    "successful": 6,
    "failed": 0
  },
  "hits": { -
    "total": 7,
    "max_score": 2.2809339,
    "hits": [ -
      { -
        "_index": "auth_daily_2017-09-08",
        "_type": "auth",
        "_id": "d16fba52-aa72-4430-ac5c-49f4f5f16136",
        "_score": 2.2809339,
        "_source": { -
          "txid": "d16fba52-aa72-4430-ac5c-49f4f5f16136",
          "akey": "DAILYGETAMYSICECREAM",
          "result": "SUCCESS",
          "factor": "Duo Push",
          "auth_device": { -
            "ip": null,
            "os": null,
            "location": null
          },
          "application": { -
            "name": "iframe_deny",
            "key": "DIZZYBROKEFRONTWAVES",
            "type": "Web SDK"
          },
          "reason": "User approved",
          "ts": "2017-09-08T18:10:29+00:00",
          "is_enrollment": true,
          "user": { -
            "groups": [ -
              { -
                "name": "group_foo",
                "key": "DGENYEARTHWROTEQUICK"
              }
            ],
            "name": "user_querytest",
            "key": "DUSTYMIGHTSLEPTTHANK"
          },
          "access_device": { -
            "software": [ -

            ],
            "ip": { -
              "is_malicious": false,
              "is_vpn": false,
              "address": "127.0.0.1",
              "is_tor": false,
```



```
[root@analytics trustedpath]# es get / | jq
{
  "version": {
    "lucene_version": "6.6.0",
    "build_date": "2017-08-14T12:33:14.154Z",
    "build_hash": "b2f0c09",
    "build_snapshot": false,
    "number": "5.5.2"
  },
  "cluster_uuid": "qJd-9BZgReCwiJ6NKXwAIQ",
  "name": "duo_dev-1",
  "tagline": "You Know, for Search",
  "cluster_name": "duo_dev"
}
```

Goal 1a:
Moving bits

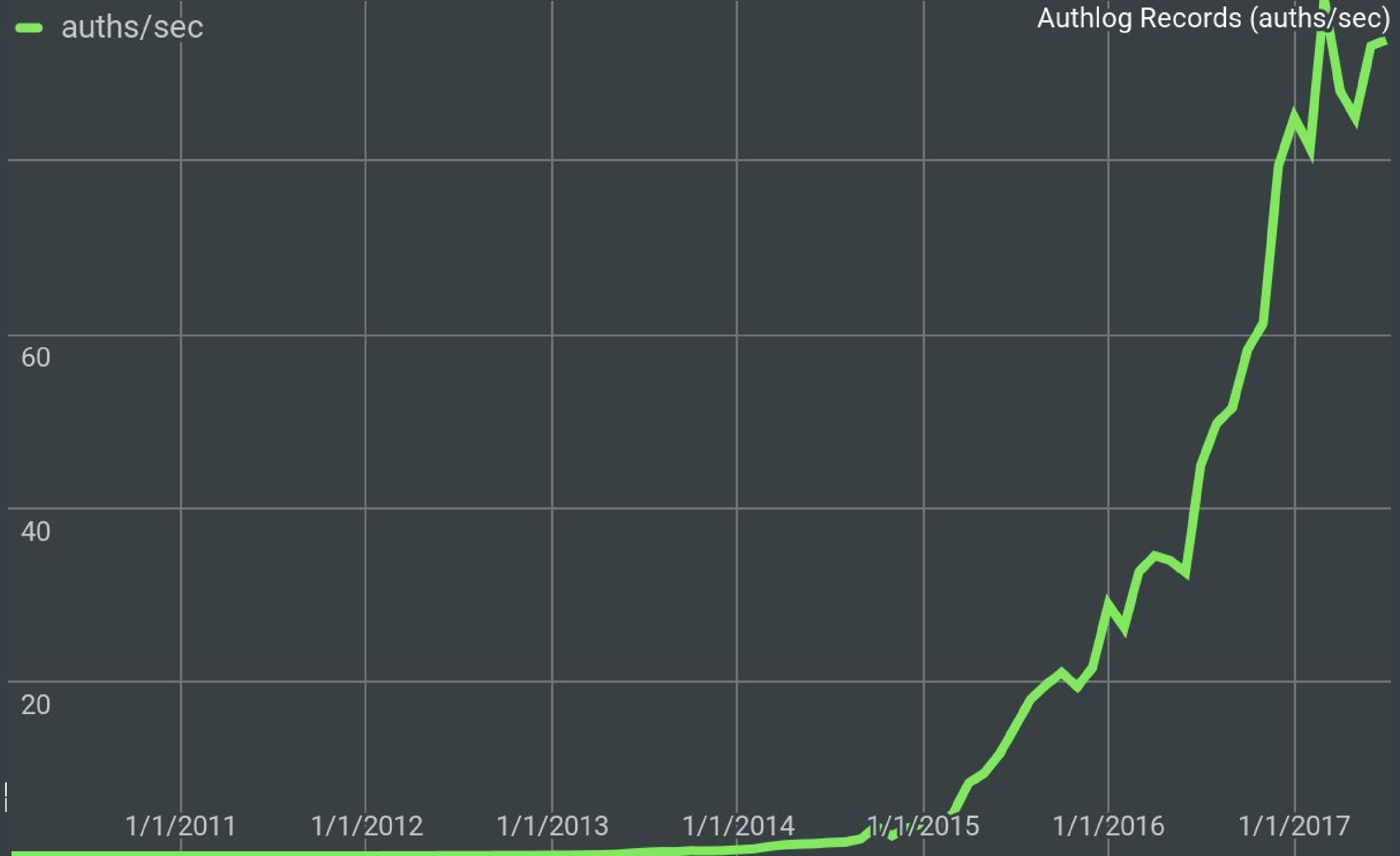

Sync

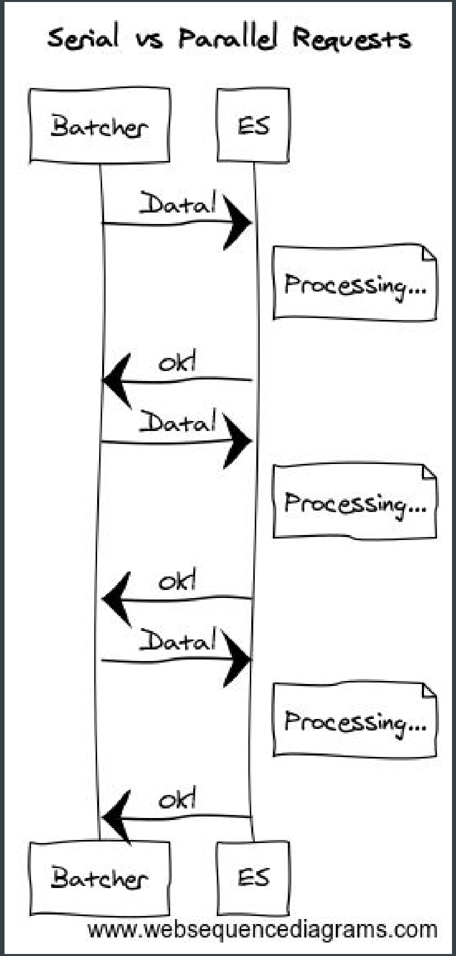
v1

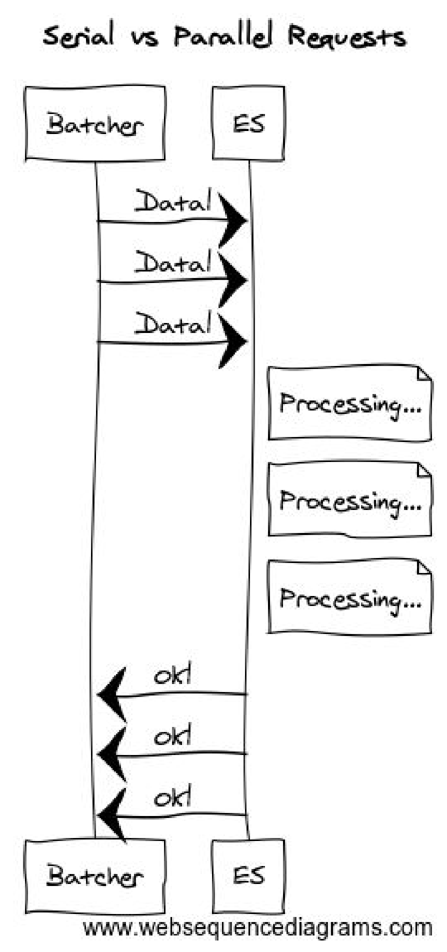
- batching process
- cursor per-db shard

per-deployment fuzzy numbers

- max - ~1.7 billion
- mean - ~60 million
- median - ~3 million


auths/sec
Authlog Records (auths/sec)
60
40
20
1/1/2011
1/1/2012
1/1/2013
1/1/2014
1/1/2015
1/1/2016
1/1/2017


Serial vs Parallel Requests
Batcher
ES
Data!
Processing...
ok!
Data!
Processing...
ok!
Data!
Processing...
ok!
Batcher
ES
www.websequencediagrams.com


Serial vs Parallel Requests
Batcher
ES
Data!
Data!
Data!
Processing...
Processing...
Processing...
ok!
ok!
ok!
Batcher
ES
www.websequencediagrams.com

```
body = ''
for record in records:
    body += json.dumps(record) + '\n'
```

```
body = []
for record in records:
    body.append(json.dumps(record) + '\n')
body = ''.join(body)
```

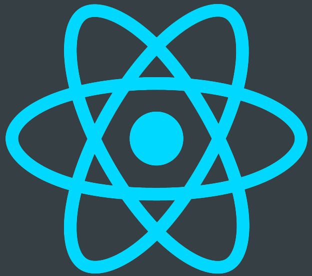
Takeaways

- Python+Twisted was ok!
- Dev speed + prior art worth the performance optimization costs
- Twisted and string gotchas
- Use realistic dev data

Goal 2:
Better data viz

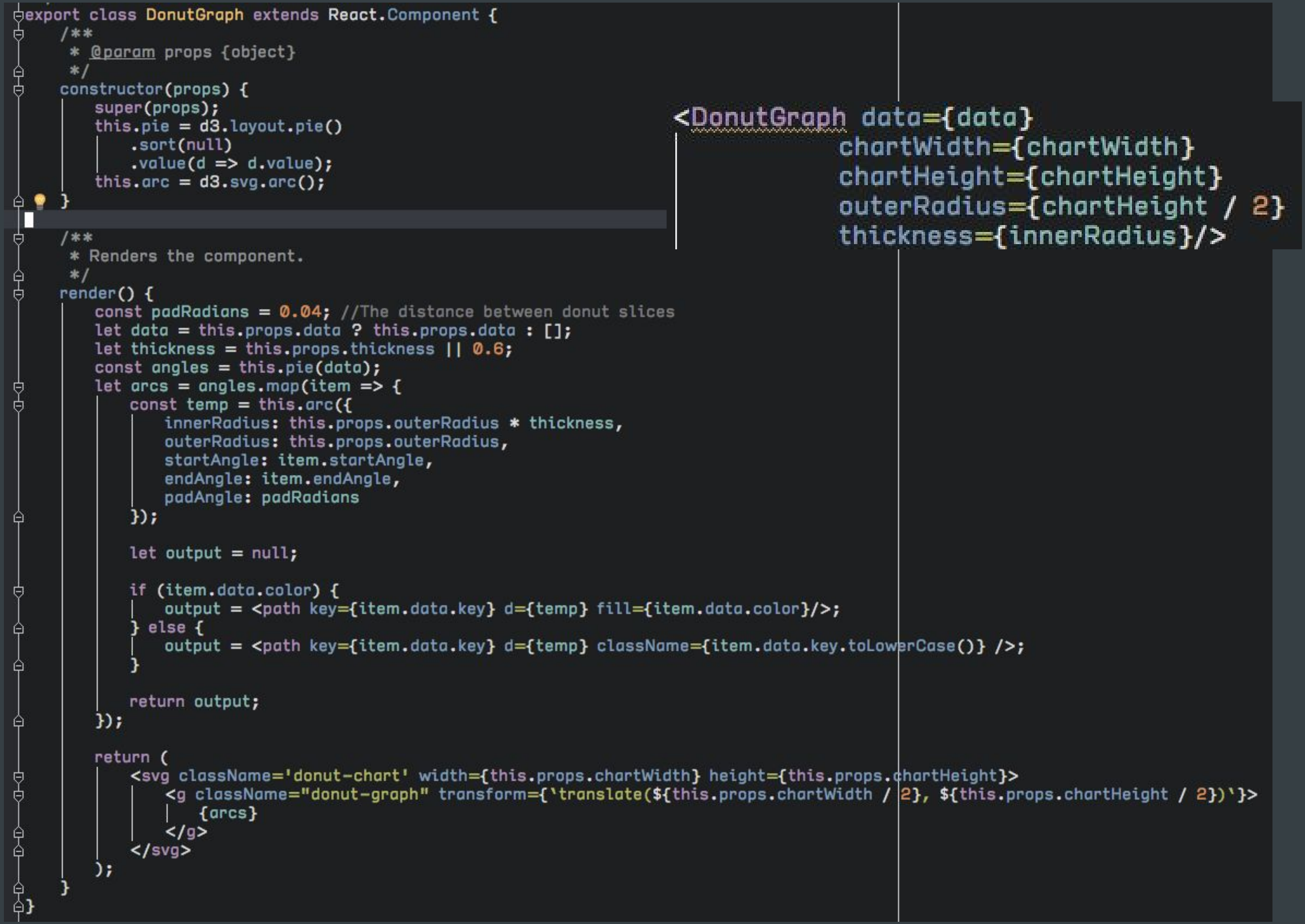
```
DataTableHelper.prototype.settings = {
    lengthChange: true,
    pageLength: 25,
    pagingType: 'full_numbers',
    deferRender: true,
    processing: false,
    dom: '<"top-control-wrapper" <"loading"> <"button-row" f>r>tlip',
    stateSave: true,
    fixedHeader: {
        header: false,
        footer: true
    },
    language: {
        search: '',
        info: '_START_&ndash;_END_ of _TOTAL_ total',
        infoEmpty: '0 items ',
        infoFiltered: ' (filtered from _MAX_)',
        lengthMenu: 'Show _MENU_ items',
        paginate: {
            first: '&#x25C5;&#x25C5;',
            last: '&#x25BB;&#x25BB;',
            next: '&#x25BB;',
            previous: '&#x25C5;'
        }
    },

    initComplete: function () { ... },

    drawCallback: function () { ... },

    stateSaveParams: function (settings, data) { ... },

    // Don't save the sorting method
    stateLoadParams: function (settings, data) { ... },

    stateSaveCallback: function (settings, data) { ... },
```

React + D3 = 💚



- Component-oriented architecture
- Does most of the work

- Number crunching
- Axes
- Stacked layouts
- Arcs/paths



```
export class DonutGraph extends React.Component {
    /**
     * @param props {object}
     */
    constructor(props) {
        super(props);
        this.pie = d3.layout.pie()
            .sort(null)
            .value(d => d.value);
        this.arc = d3.svg.arc();
    }

    /**
     * Renders the component.
     */
    render() {
        const padRadians = 0.04; //The distance between donut slices
        let data = this.props.data ? this.props.data : [];
        let thickness = this.props.thickness || 0.6;
        const angles = this.pie(data);
        let arcs = angles.map(item => {
            const temp = this.arc({
                innerRadius: this.props.outerRadius * thickness,
                outerRadius: this.props.outerRadius,
                startAngle: item.startAngle,
                endAngle: item.endAngle,
                padAngle: padRadians
            });

            let output = null;

            if (item.data.color) {
                output = <path key={item.data.key} d={temp} fill={item.data.color}/>;
            } else {
                output = <path key={item.data.key} d={temp} className={item.data.key.toLowerCase()} />;
            }

            return output;
        });

        return (
            <svg className='donut-chart' width={this.props.chartWidth} height={this.props.chartHeight}>
                <g className="donut-graph" transform={`translate(${this.props.chartWidth / 2}, ${this.props.chartHeight / 2})`}>
                    {arcs}
                </g>
            </svg>
        );
    }
}
```

```
<DonutGraph data={data}
            chartWidth={chartWidth}
            chartHeight={chartHeight}
            outerRadius={chartHeight / 2}
            thickness={innerRadius}/>
```

What's next?

- NRT and RT analysis
- Modeling fraudulent activity
- Moar standard data viz

fin

- tweets @uoodsq
- duo.com/jobs