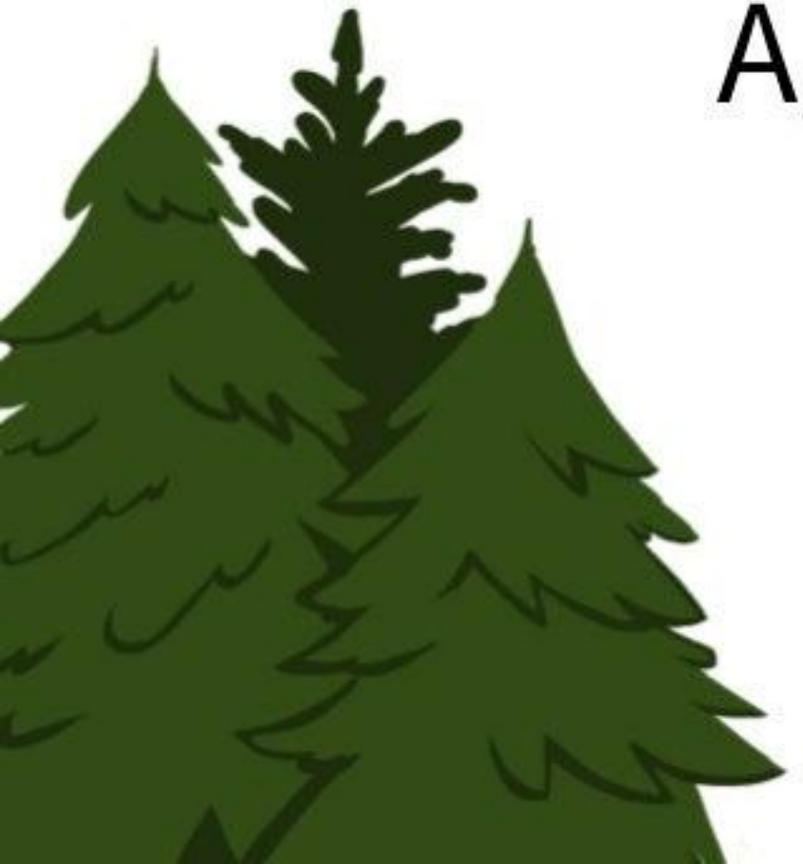
Alpine Data Labs



Agile Data Science

INFORMS BIG DATA CONFERENCE
6/23/2013

Presented by Joel S Horwitz
Follow me @JSHorwitz

Agile Manifesto

We are uncovering better ways of developing software by doing it and helping others do it. Through this work we have come to value:

- Individuals and **interactions** over processes and tools
- Working software over comprehensive documentation
- Customer **collaboration** over contract negotiation
- Responding to change over following a plan

That is, while there is value in the items on the right, we value the items on the left more.

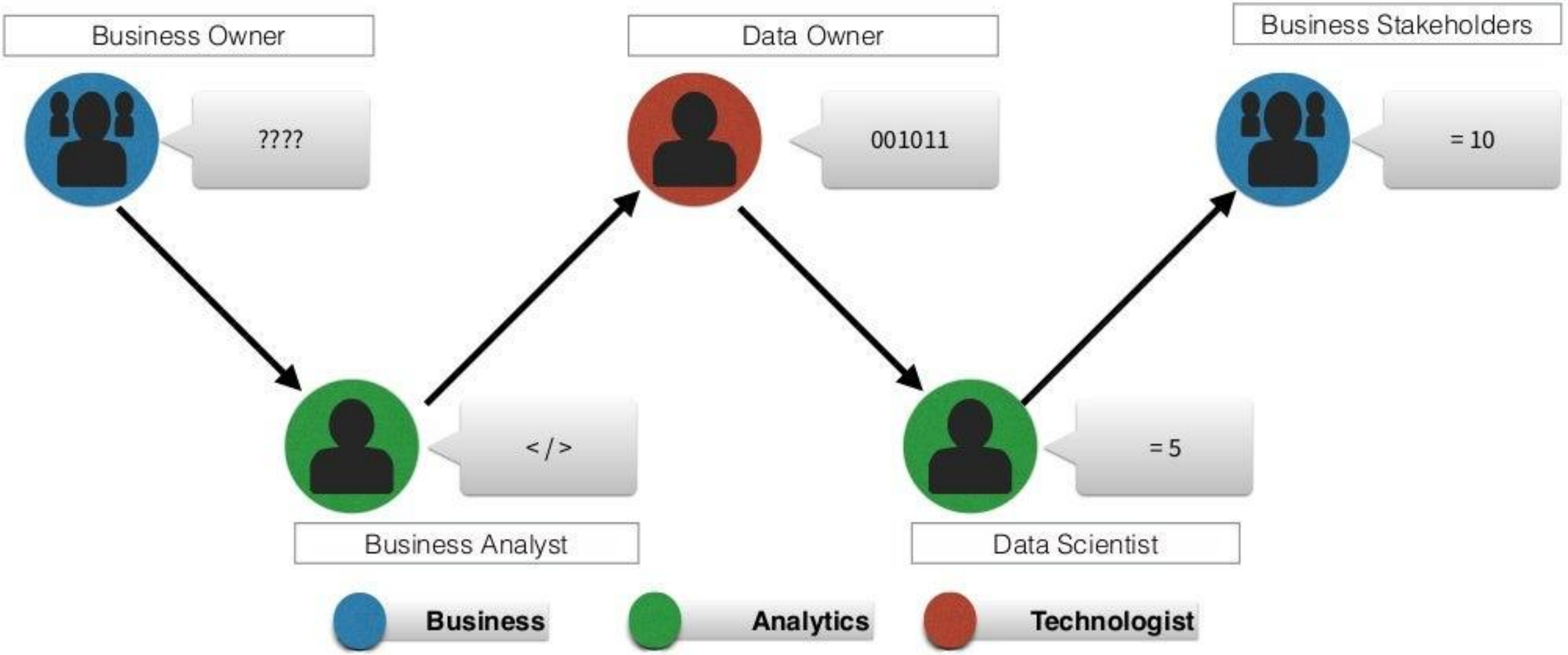
Agile Manifesto

We are uncovering better ways of developing ~~software~~ **models** by doing it and helping others do it. Through this work we have come to value:

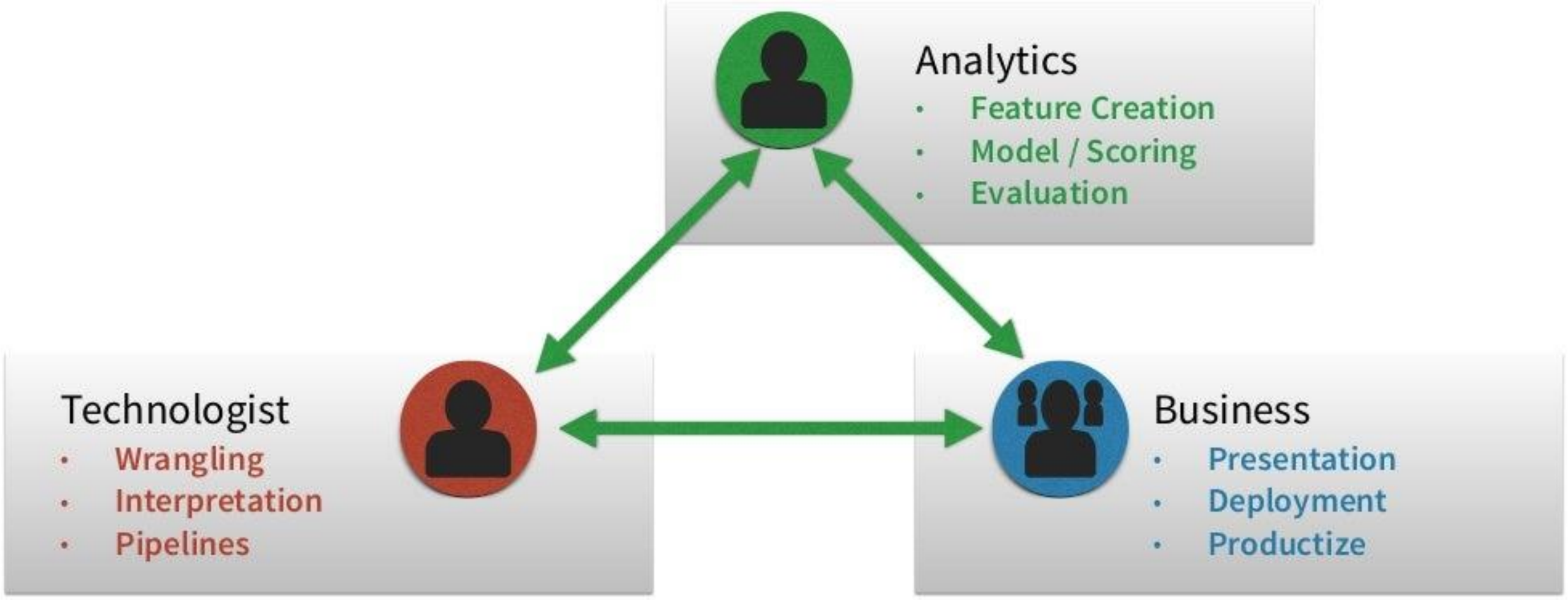
- Individuals and **interactions** over processes and tools
- Working ~~software~~ **models** over comprehensive documentation
- Customer **collaboration** over contract negotiation
- Responding to change over following a plan

That is, while there is value in the items on the right, we value the items on the left more.

Linear workflows in non-agile culture

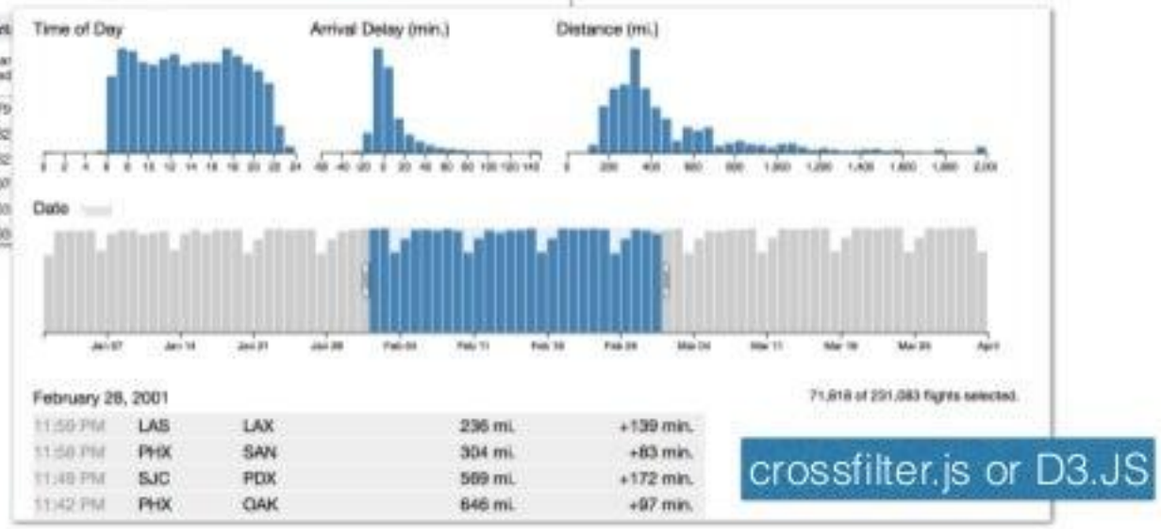


Agile is about *continuous* interactions



Minimally Viable Data Products (MVDP)

Tableau Dashboards



crossfilter.js or D3.JS

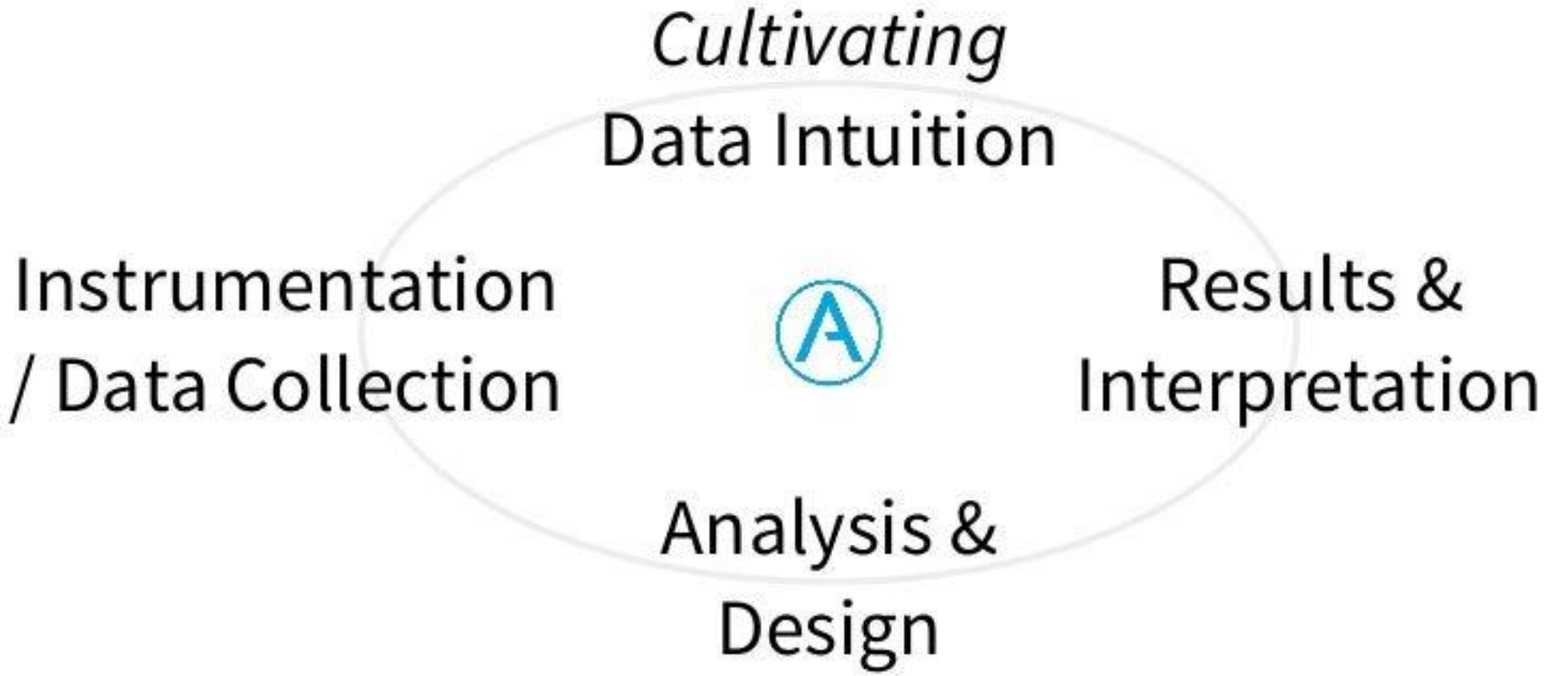
Product Recommendations

Also, Product Recommendations

People Recommendations

One model, many use cases.

Agile Data Science Feedback Loop



What do you need?

1. Business Champion
2.Integrated Environment
3. Analytics Ninjas

1. Business Champion(s): Defining the Problem

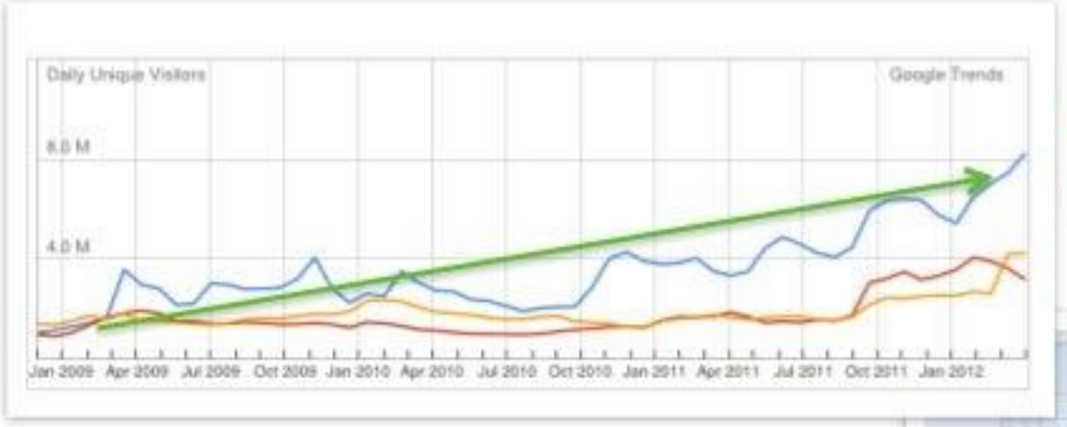
Executive Sponsor(s) who have a vested interest... ready to take action from results, has an impact to their business goals.

Chief Technology Officer

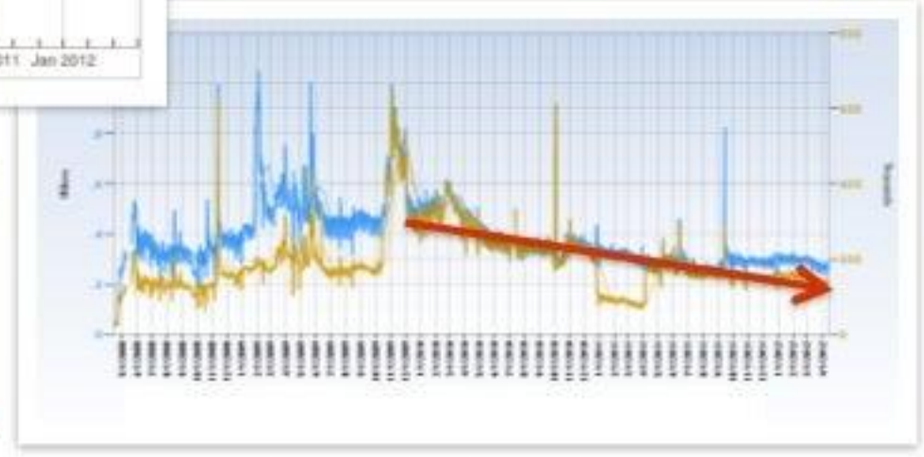
SVP of Product

Chief Marketing Officer

VP of Sales



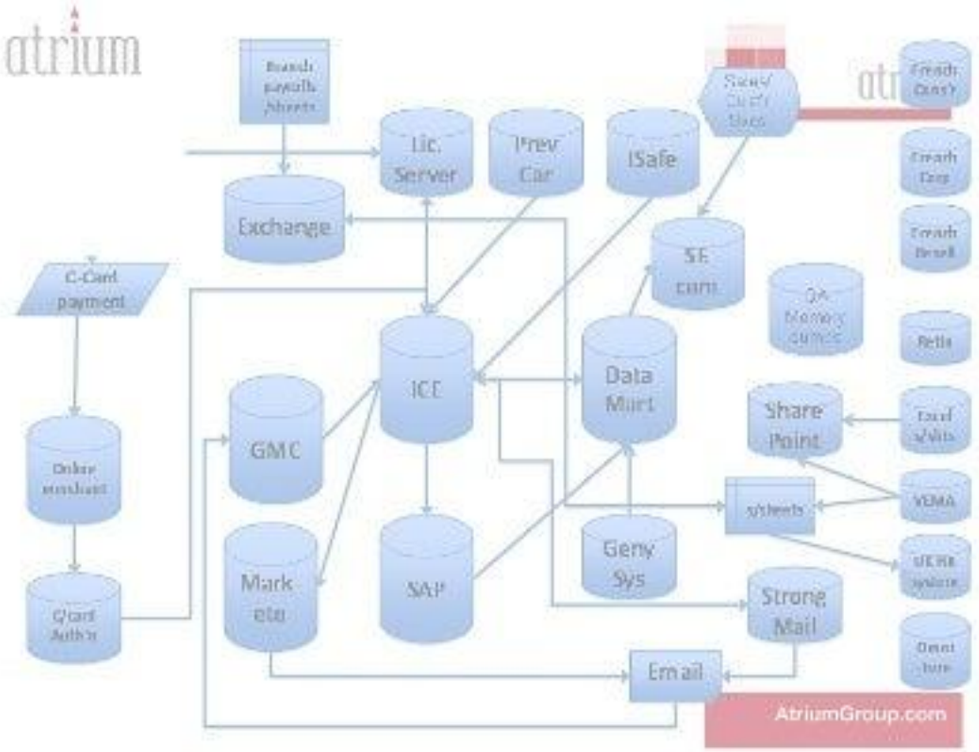
Problem statement... Monthly active user count growth was stagnating. Evidence of where to look... acquisition funnel, user engagement, and customer loyalty.



Question to answer... How do I connect each part of the acquisition funnel?

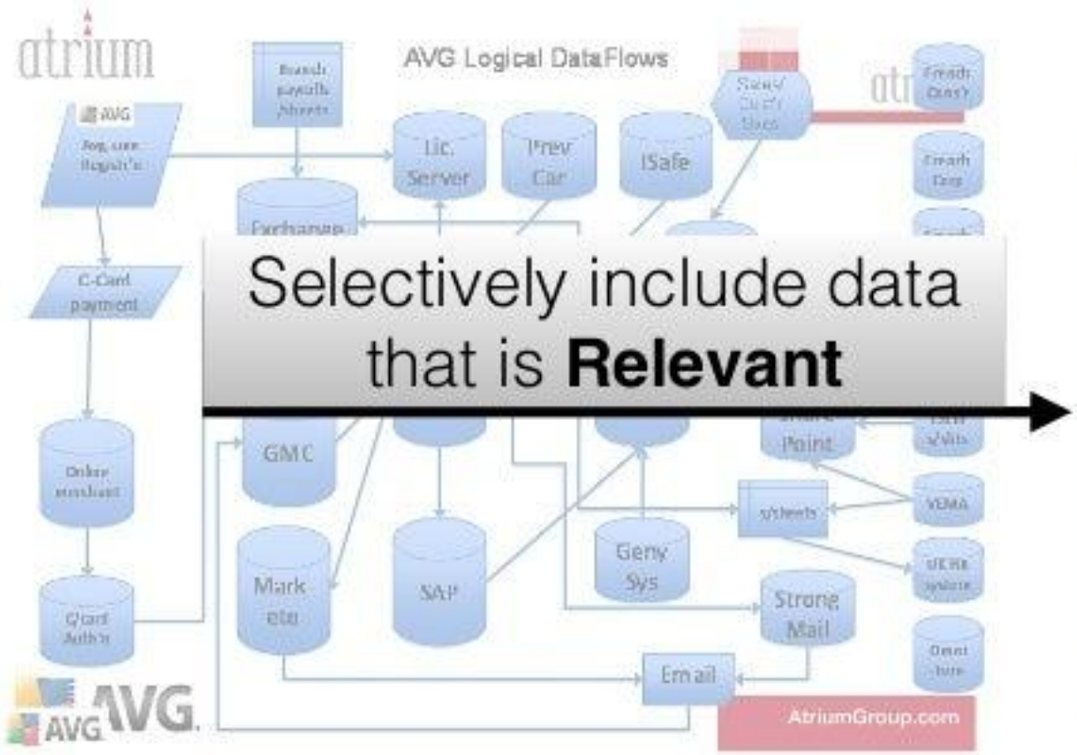
We started with the acquisition funnel... web visits, downloads, installs, and activations.

2. Environment: Before



- Many data silos and technology limitations
- Data definitions not well defined.
- Multiple data formats.
- No place to build MV**D**Ps at scale.

2. Environment: After



Analytics Sandbox

Open Source Technologies

mahout

HIVE

Apache

Flume, Sqoop, Oozie, and others

Relational Database

Minimally Viable *Data* Products

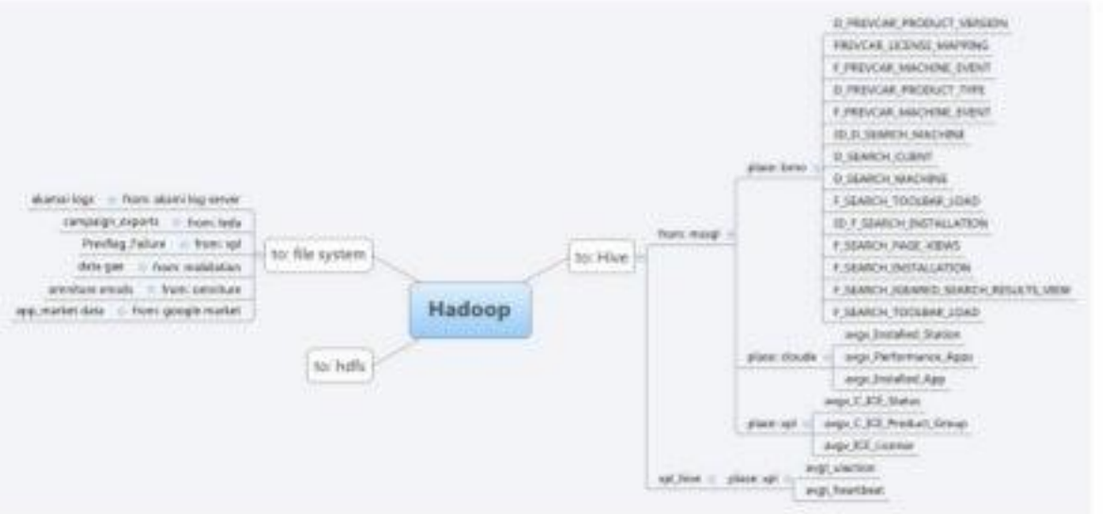
How was our analytics platform deployed and connected to the network of systems.

First there was FTP… dump raw log files from web analytics, content delivery network, and application log files.

Second there was MS SQL… Most of the data was in Microsoft SQL databases.

Third there was Hadoop (with Hive)… Engineering and Development backup to Hadoop.

Now there is web based analytics…



What is Hadoop?

- **Overview**: Apache Hadoop is a framework for running applications on large cluster built of commodity hardware.
- **Storage**: Hadoop Distributed File System (HDFS™) is the primary storage system used by Hadoop applications. HDFS creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, extremely rapid computations.
- **Applications:** Apache Hive is a large scale Data Warehouse system and Apache Mahout is a machine learning system.

3. Analytics Ninjas

Build a TIGER Team… subject matter expert, modeler, technologist, and storyteller

- Curiosity
- Willingness to learn
- Resourceful
- Risk Takers

Schedule Standups… daily is best depending on the project it can vary

- Take notes!
- Open dialogue (peer review)
- Share knowledge

Centralize and Version Control Your work… files, code, knowledge, and data lineage are key to success

- Create a Wiki
- Work backwards from the end goal to the analysis to the data
- Share, share, and share! Stand on the shoulder of Giants! Why start new, there is plenty of boilerplate to go around.

What is Analytics?

*Analytics is the application of **computer_technology, operational research, and statistics** to solve problems in business and industry.*

*Historically, Analytics was heavily used in **banking** for portfolio assessment using social status, geographical location, net value, and many other factors.*

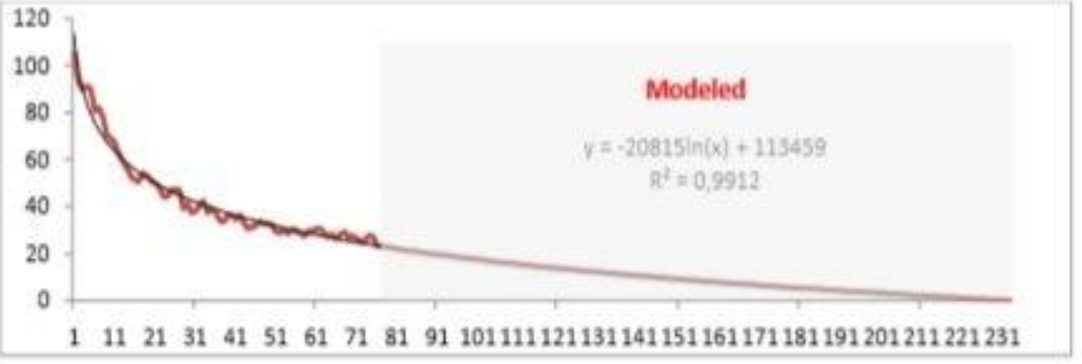
*Today, Analytics is applied to a vast number of industries and is **re-emerging** due to the phenomenal explosion of data from our connected world.*

*Big Data consists of data sets that grow so **large and complex** that they become awkward to work with using on-hand database management tools*

*McKinsey Global Institute estimates that big data analysis could save the American health care system **$300 billion** per year and the European public sector **€250 billion**.*

Analytics Example: Platform Monetization

1. Business Champion
   - Sales & Product
2. Integrated Environment
   - Web analytics / app data and SQL Database.
3. Analytics Ninjas
   - Search engine marketers, business analysts, and statisticians.



Search Lifetime Value

- 3rd Party Distribution
- SEM
- Organic

Traffic Quality Analysis

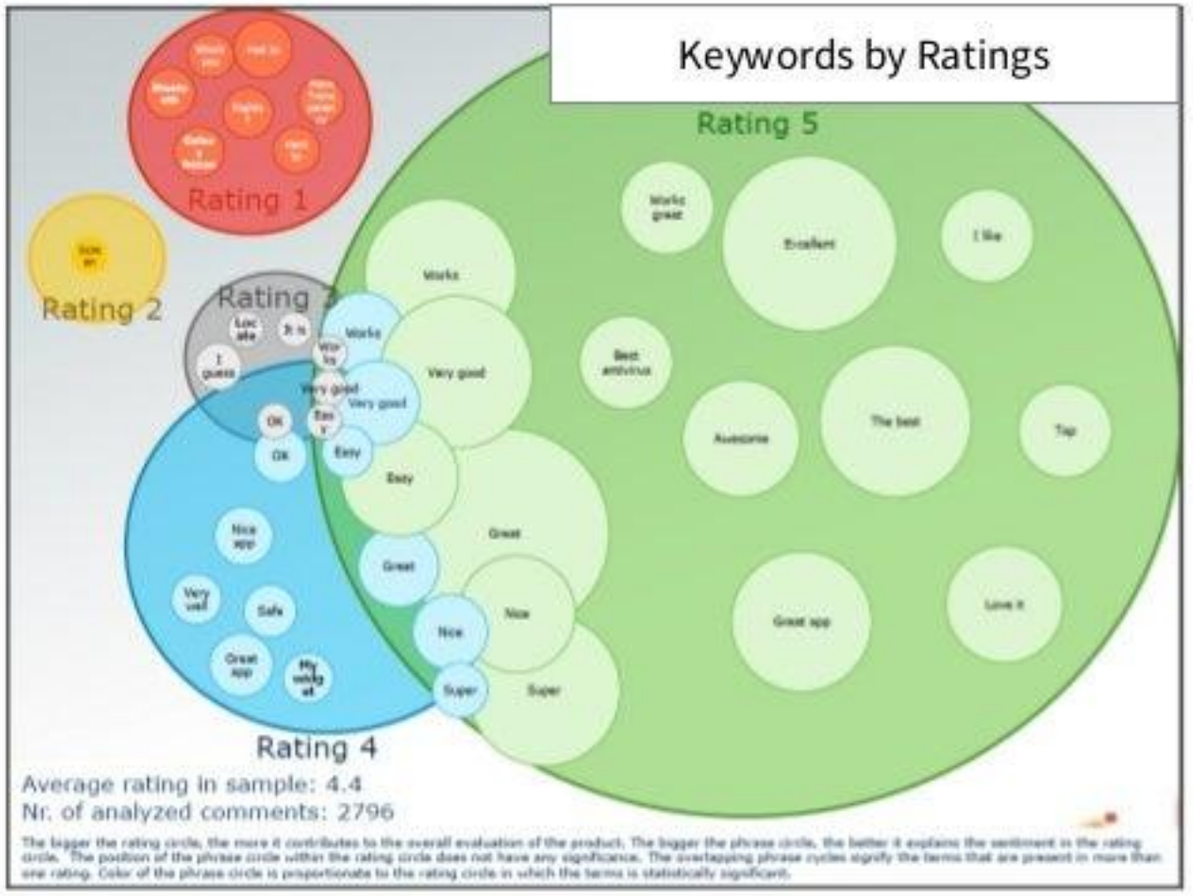
- PPC
- PPD
- PPI
- PPA

Examples of Analytics: Campaign Management

1. Business Champion
   - Sales, Product, Marketing, and Project Managers.
2. Integrated Environment
   - Data Warehouse, SQL DB, Hadoop, and Tableau
3. Analytics Ninjas
   - Web analytics, product managers, business analysts, and business intelligence.

Examples of Analytics: Customer Sentiment

1. Business Champion
   - Product and Marketing
2. Integrated Environment
   - Mobile analytics, app logfiles, and Hadoop.
3. Analytics Ninjas
   - Web analytics, product managers, business analysts, mobile developers.



What to avoid

Its all about plugins for web analytics… Google Analytics, App Stores (iOS, Android, others), Social (Twitter, Facebook FQL, others?).

Analysis lifecycle…

1. **Give me data…** import data and ETL it into submission.
2. **What does this data mean…** Crunch, Blend, Join, Pivot, Predict, Count, Map, or whatever
3. **Show and Tell…** Static (powerpoint, excel, etc.) or dynamic (trends, filtering, drilldown). (No more dashboards)...

Interestingness rocks! Tell me where to look (I'm feeling lucky…)

known knowns = data puking (dashboards)

Viva la revolucion! Data democracy!

get people to "make love to data" to make actual good use of it

Thank you! Any Questions?

Want to jump start your Agile Data Science project? Head over to **http://start.alpinenow.com**

Follow me on @JSHorwitz