



## DEVOPS ENTERPRISE SUMMIT

Las Vegas  
October 28-30, 2019

# Your Data Nerd Friends Need You!

*How the world of data analytics, science and insights is failing and how the principles from Agile, DevOps, and Lean are the way forward. #DataOps*

*October 30, 2019*



# What is this talk about?



A BIG PROBLEM  
THAT CAN USE  
YOUR HELP



BY HAVING  
EMPATHY FOR A  
GROUP OF PEOPLE



THAT ARE  
SUFFERING



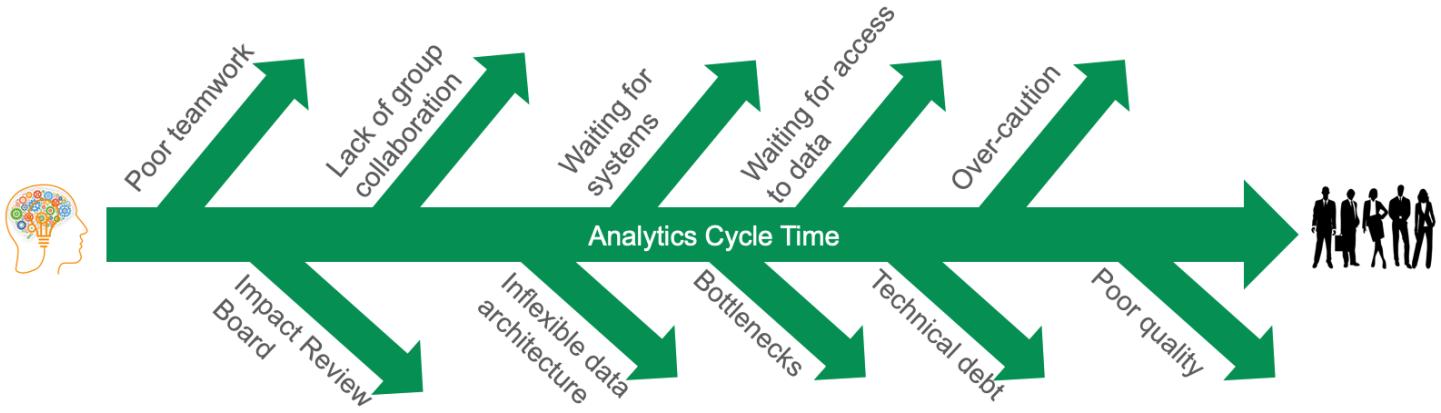
AND WHAT YOU  
KNOW CAN HELP  
THEM

# It's Big (Data) ...



- ‘**New Oil**’ amount of data increasing fast
- **Buzz:** Big Data, Data Science, Data Lakes, Machine Learning, AI
- **\$189.1 Billion Market**, Double-Digit Annual Growth Through 2022.
  - \$7.5B for GitHub, \$15.7B for Tableau
- **10s millions of people** creating insight from data
  - More than software developers.
  - 1 of 25 workers full time, significant part time.

# It's a big problem that can use your help



- 87% of data science projects never make it into production.
- Data analytics investment up, yet “data driven” organizations down 37% to 31% since 2019.
- 80% of AI projects resemble alchemy
- 60% of all data analytic projects fail
- 79% of data projects have too many errors
- ... “They’re not even using version control!”

# Walk down the hall to your data analytics group and observe

---

- Poor quality, high errors
- Minor changes take months to implement, manual processes
- 75 percent of the day is hijacked by unplanned work
- Oversubscribed resources limit overall productivity.

..... Sound familiar?



# Agenda



**A BIG PROBLEM  
THAT CAN USE  
YOUR HELP**



**THAT ARE  
SUFFERING**



**BY HAVING  
EMPATHY FOR A  
GROUP OF PEOPLE**



**AND WHAT YOU  
KNOW CAN HELP  
THEM**

# Who Are These People?



## DATA AND ANALYTICS MANAGER

### DATA SCIENCE TEAM LEADER

**Role**  
Manages a team of analysts and

**Mindset**  
Data Wizard's Cheerleader

## DATA SCIENTIST

### AS RARE AS UNICORNS

**Role**  
Cleans, massages and organizes (big) data

**Mindset**  
Curious data wizard

**Languages**  
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

**Skills & Talents**

- Distributed computing
- Predictive modeling
- Story-telling and visualizing
- Math, Stats, Machine Learn

## DATA ANALYST

### DATA DETECTIVE

**Role**  
Collects, processes and performs statistical data analyses

**Mindset**  
Intuitive data junkie with high "figure-it-out" quotient

**Languages**  
R, Python, HTML, Javascript, C/C++, SQL

**Skills & Talents**

- Spreadsheet tools (e.g. Excel)
- Database systems (SQL and NO SQL based)
- Communication & visualization
- Math, Stats, Machine Learning

## BUSINESS ANALYST

### CHANGE AGENT

**Role**  
Improves business process as intermediary between business and IT

**Mindset**  
Resilient project juggler

**Languages**  
SQL

## DATA ENGINEER

### SOFTWARE ENGINEERS BY TRADE

**Role**  
Develops, constructs, tests and maintains architectures (such as databases and large scale processing systems)

**Mindset**  
All-purpose everyman

## DATA ARCHITECT

### THE CONTEMPORARY DATA MODELLER

**Role**  
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

**Mindset**  
Inquiring ninja with a love for data architecture design patterns

**Languages**  
SQL, XML, Hive, Pig, Spark

**Skills & Talents**

- Data warehousing solutions
- In-depth knowledge of database architecture
- Extraction Transformation and Load(ETL), spreadsheet and BI tools
- Data modeling
- Systems development

## STATISTICIAN

### HISTORIC LEADERS OF DATA

**Role**  
Collects, analyzes and interprets qualitative as well as quantitative data with statistical theories and methods

**Mindset**  
Logical and enthusiastic stats genius

**Languages**  
R, SAS, SPSS, Matlab, Stata, Python, Perl, Hive, Pig, Spark, SQL

**Skills & Talents**

- Statistical theories & methodology
- Data mining & machine learning
- Distributed Computing (Hadoop)
- Database systems (SQL and NO SQL based)
- Cloud tools

## DATABASE ADMINISTRATOR

### DATABASE CARETAKER

**Role**  
Ensures that the database is available to all relevant users, is performing properly and is being kept safe

**Mindset**  
Master of Disaster Prevention

**Languages**  
SQL, Java, Ruby on Rails, XML, C#, Python

**Skills & Talents**

- Backup & recovery
- Data modeling and design
- Distributed Computing (Hadoop)
- Database systems (SQL and NO SQL based)
- Data security
- ERP & business knowledge

# They took a different door ...

---

- Talk like you, look like you
- But early in their career they took the data analytics door, not the software door
- Complex toolchain
- 50+ tools in each category
- People love their tools
- Some code, some configure



# They work in Teams



- **Data Engineer Team**



- **Data Science Team**



- **Self Service Team**



- **Data Governance Team**

# They work in teams together



- **Data Engineer Team**
  - Source data
  - Create a database table
  - Load data



- **Data Science Team**
  - Use data to create model
  - Add a column to data with results of model (batch)



- **Self Service Team**
  - Visualize Data and Model results
  - Add More Calculations to data (Alteryx)



- **Data Governance Team**
  - Catalog data, model results

Name	Sales
joe	\$1234.56
kelly	\$4567.89

Name	Sales	Segment
joe	\$1234.56	Lo Value
kelly	\$4567.89	Hi Value

# They work in teams together



- **Data Engineer Team**
  - Source data
  - Create a database table
  - Load data



- **Data Science Team**
  - Use data to create model
  - Add a column to data with results of model (batch)



- **Self Service Team**
  - Visualize Data and Model results
  - Add More Calculations to data (Alteryx)



- **Data Governance Team**
  - Catalog data, model results

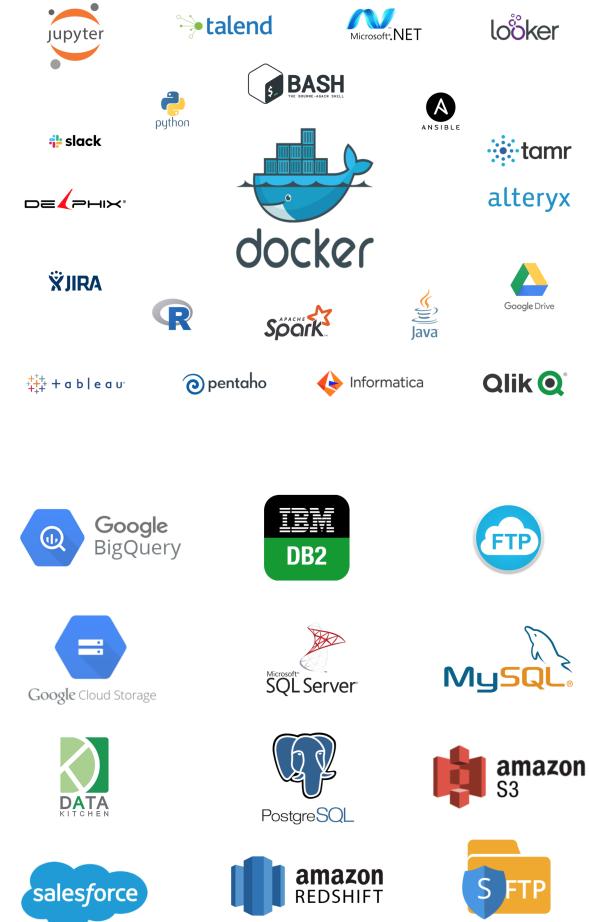
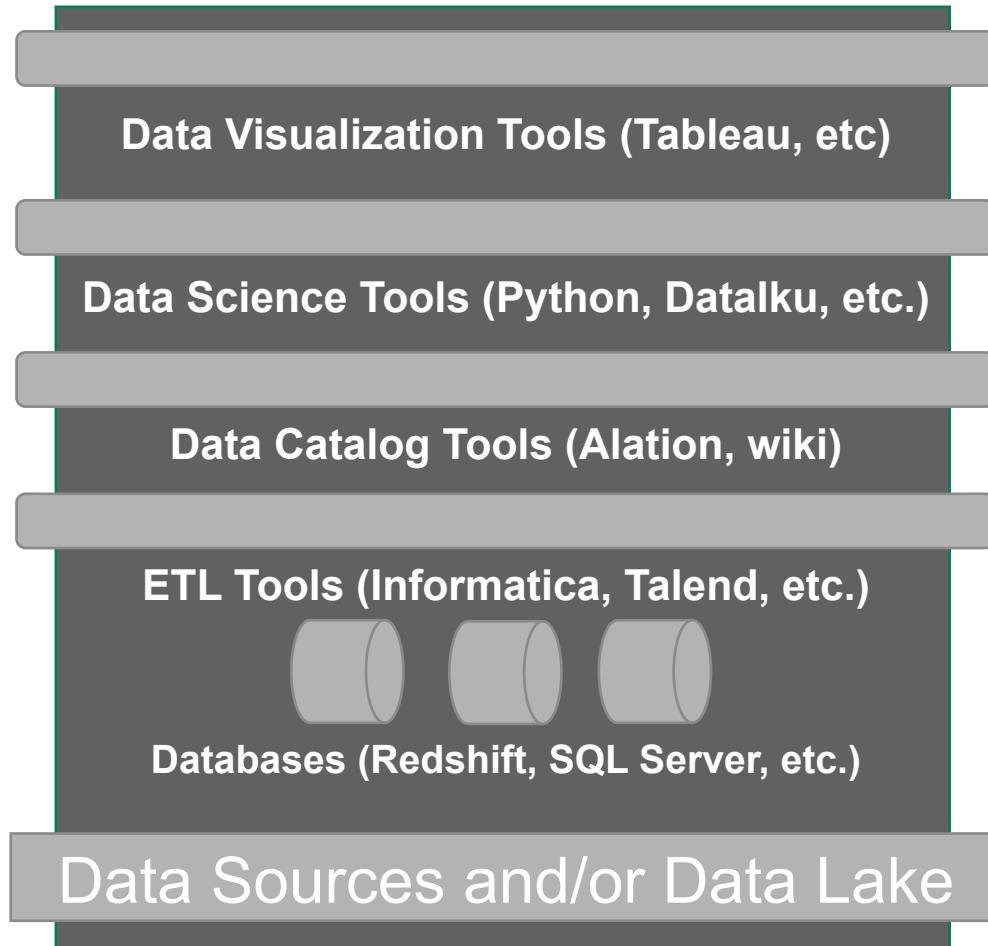
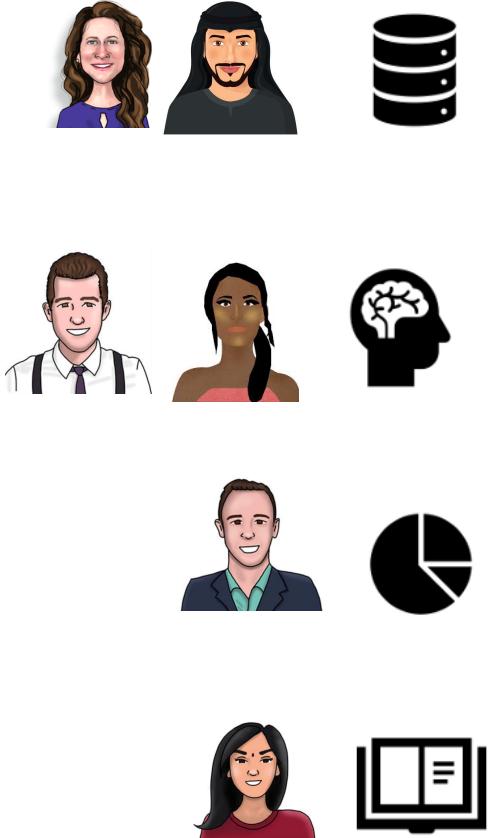
Name	Sales
joe	\$1234.56
kelly	\$4567.89

Name	Sales	Segment
joe	\$1234.56	Lo Value
kelly	\$4567.89	Hi Value

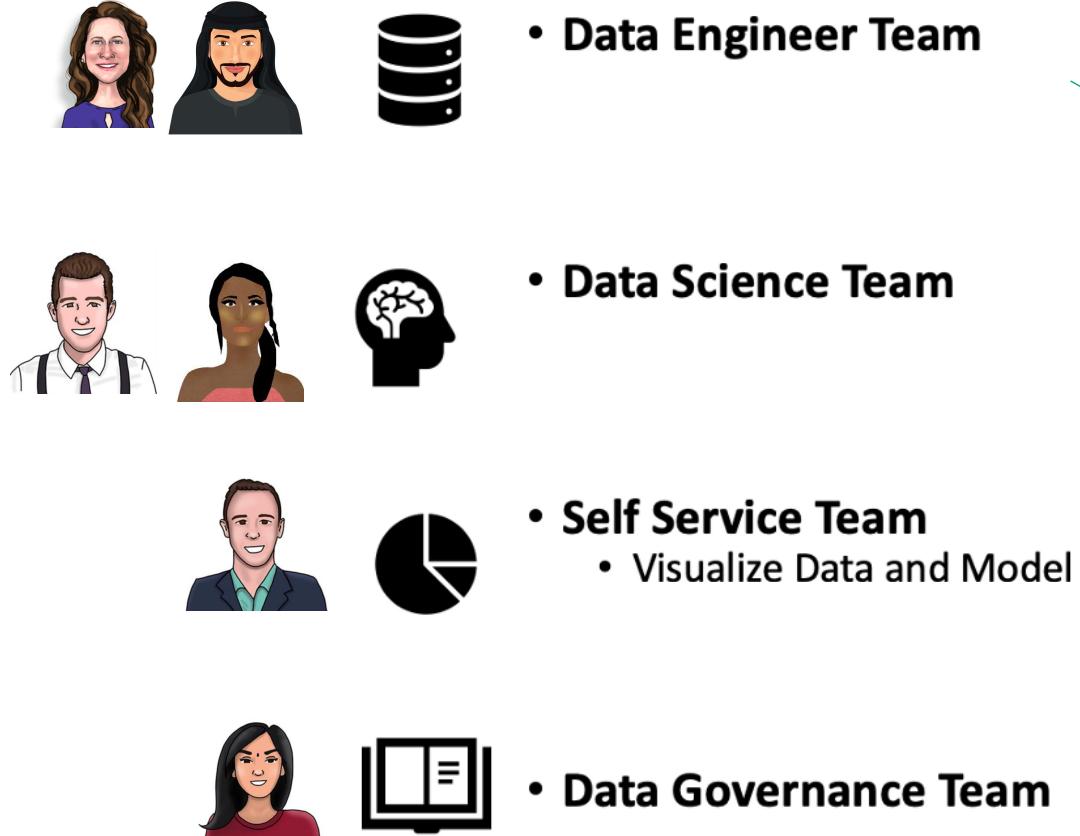
Name	Sales	Segment	Owner
joe	\$1234.56	Lo Value	West Team
kelly	\$4567.89	Hi Value	East Team

Column	Description	Source
Name	...	Raw data (data eng)
Sales	..	Raw data (data eng)
Segment	.....	Data Science
Owner	.....	Self - Service

# With a massive, fragmented toolchain



# They may work for the same boss



Chief Data Officer  
Chief Analytics Officer

# Or not



- Data Engineer Team



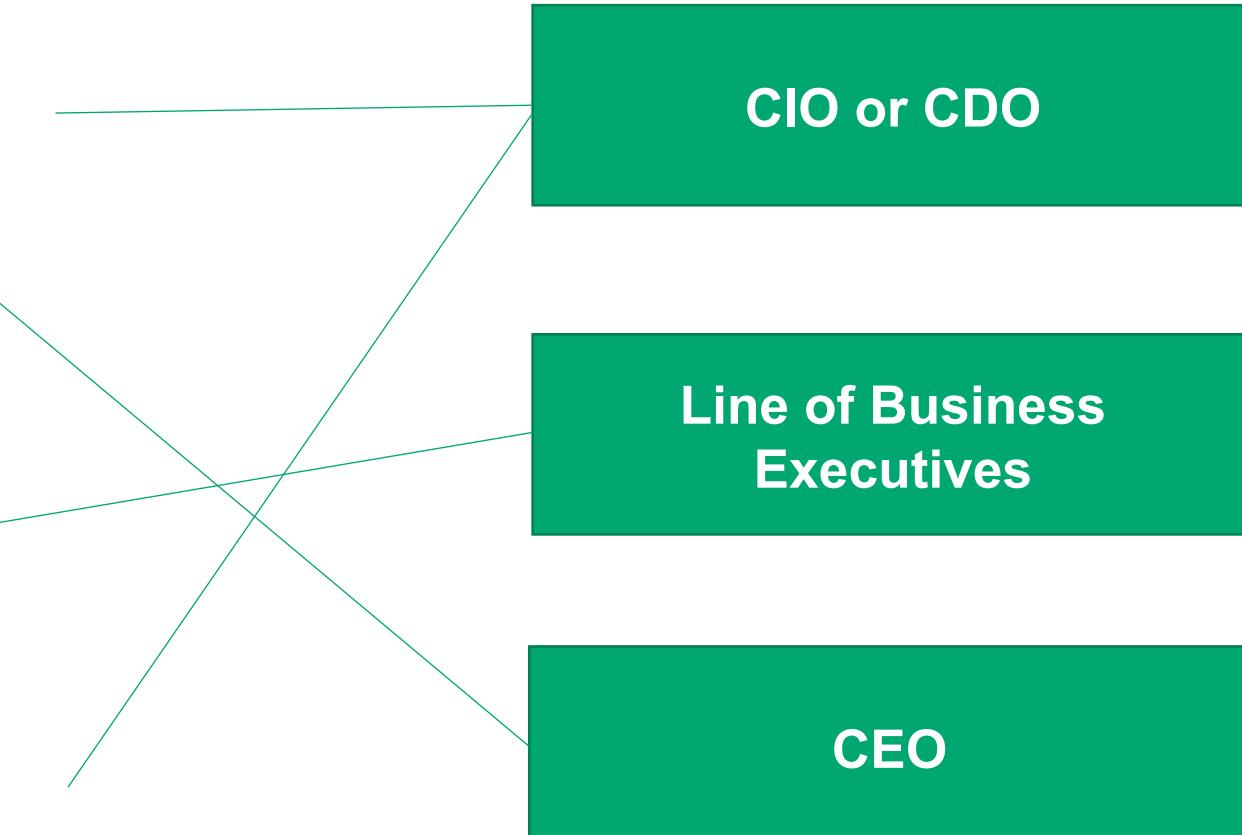
- Data Science Team



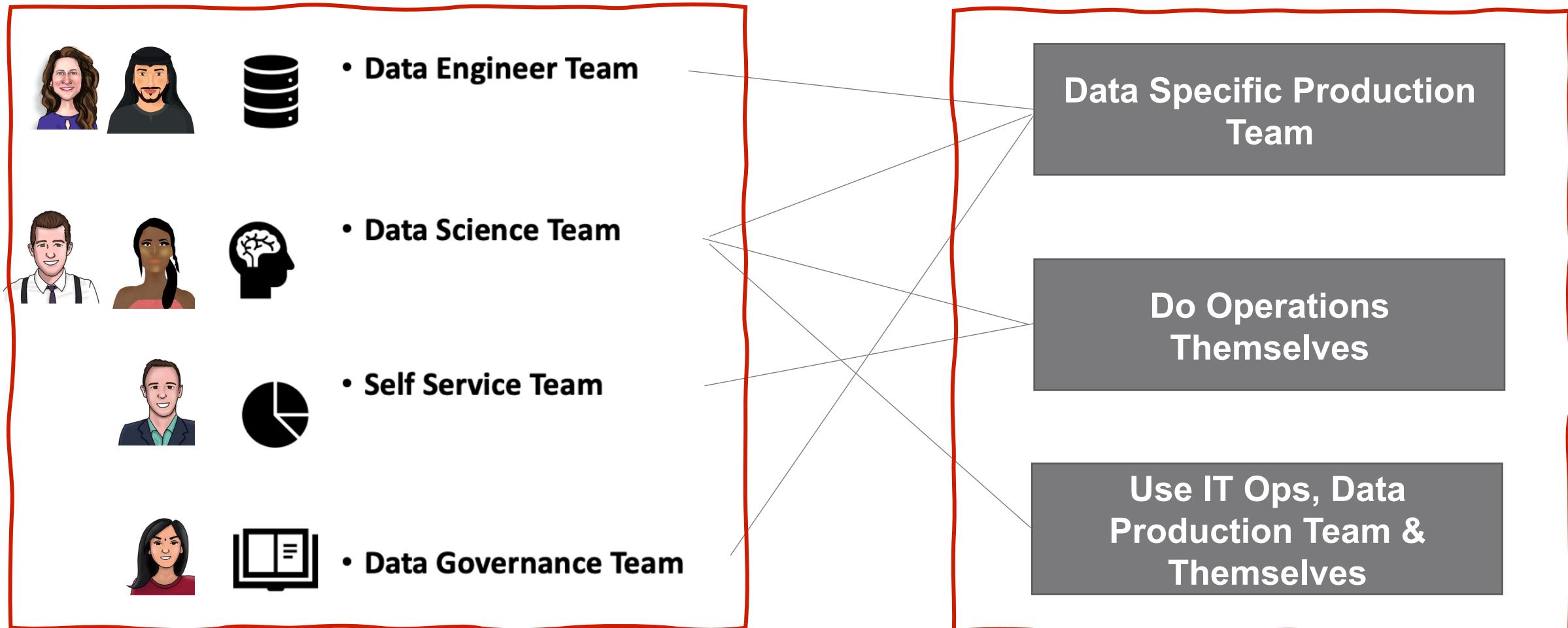
- Self Service Team



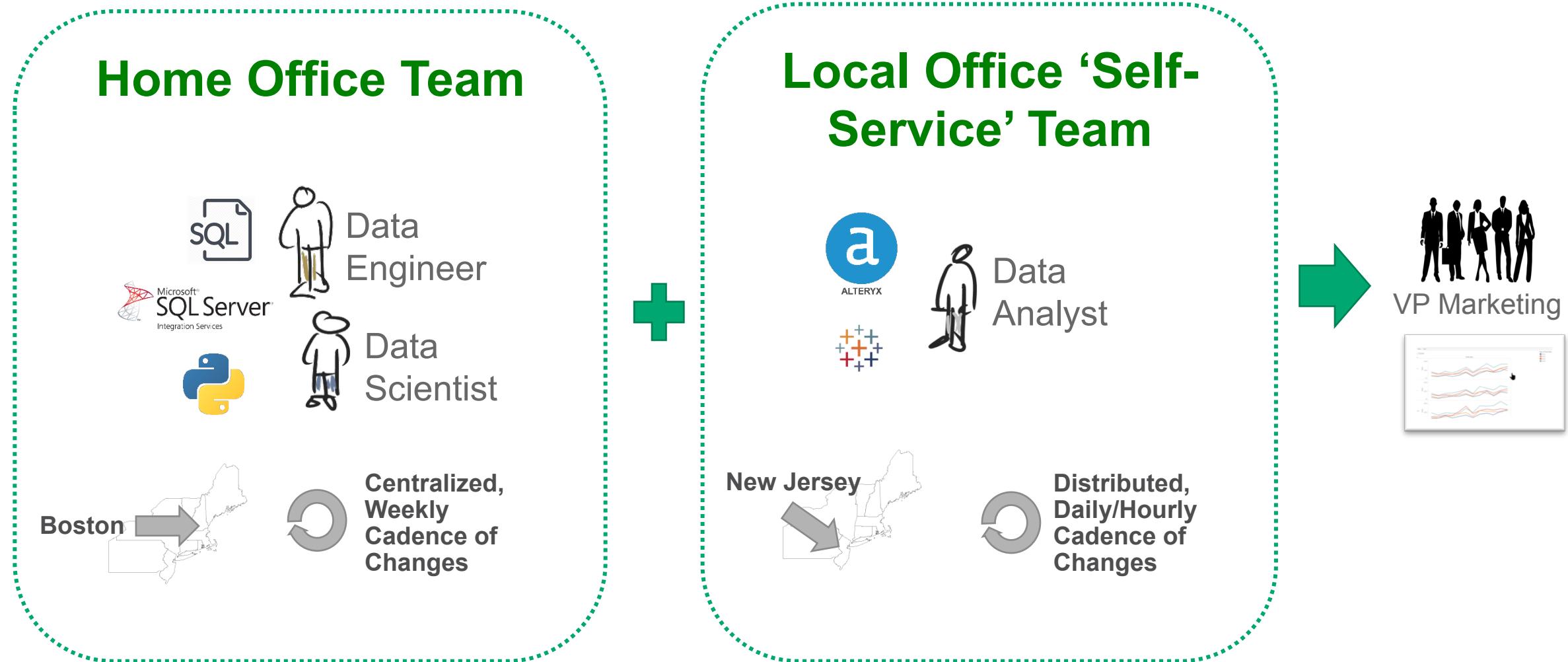
- Data Governance Team



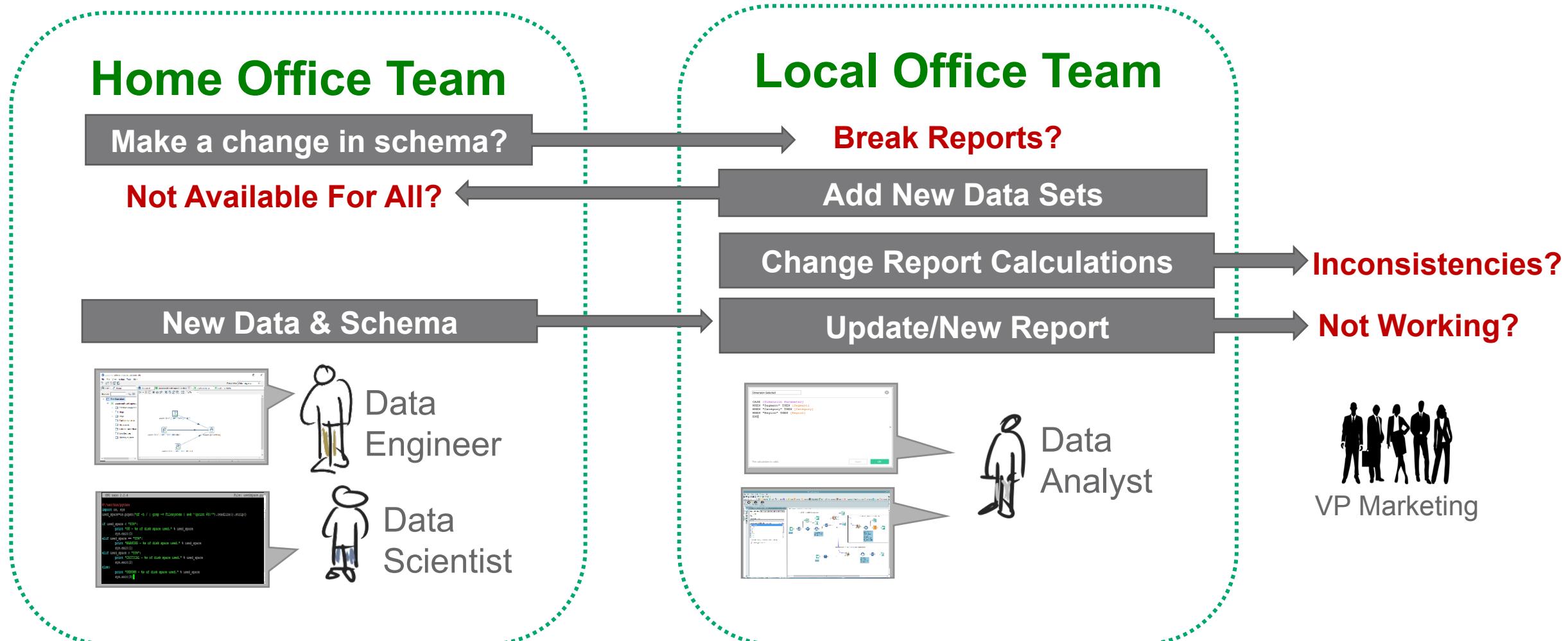
# A many to many dev/ops relationship



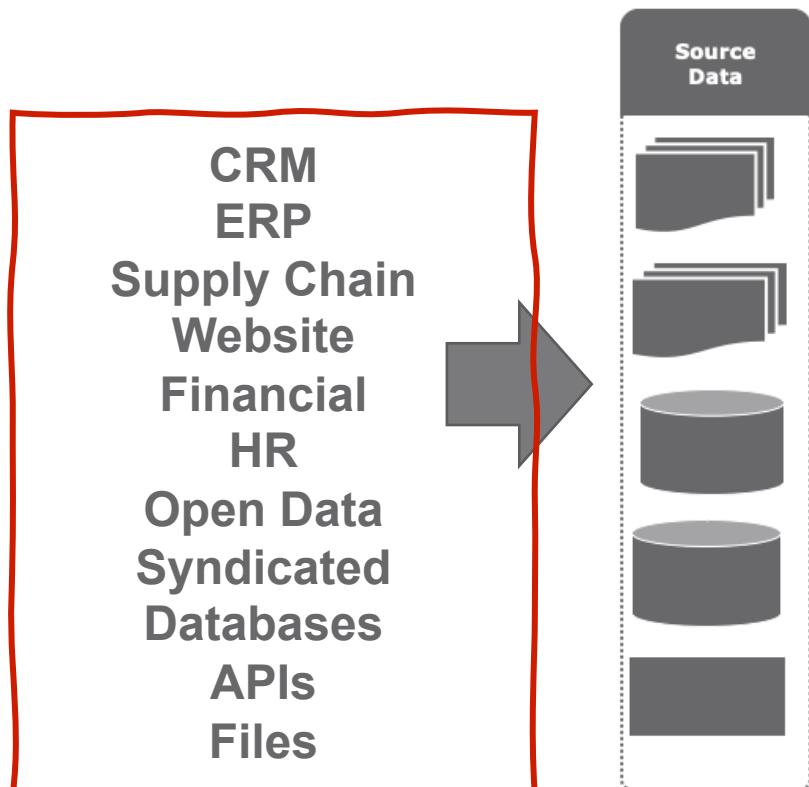
# Example: Coordination of Two Teams, Two Locations, Two Ops, Many Tools



# Challenges With Coordination

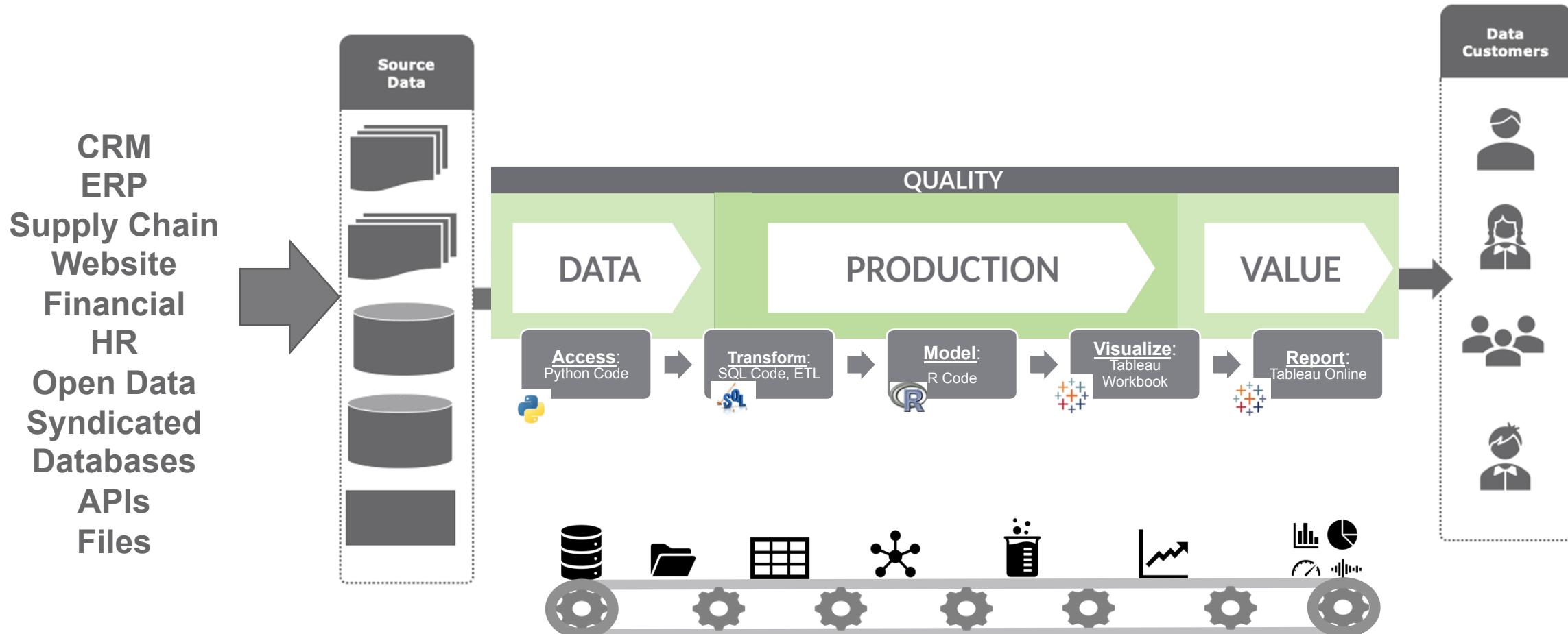


# And they source data from internal and external system

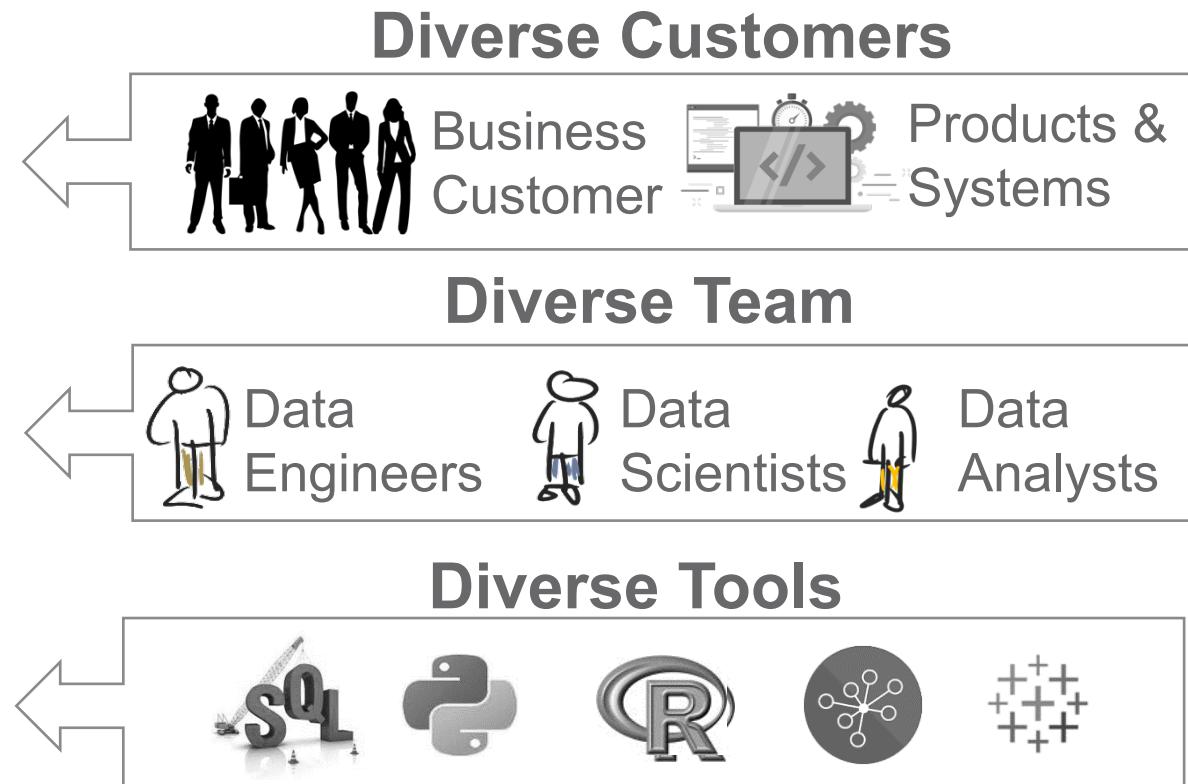
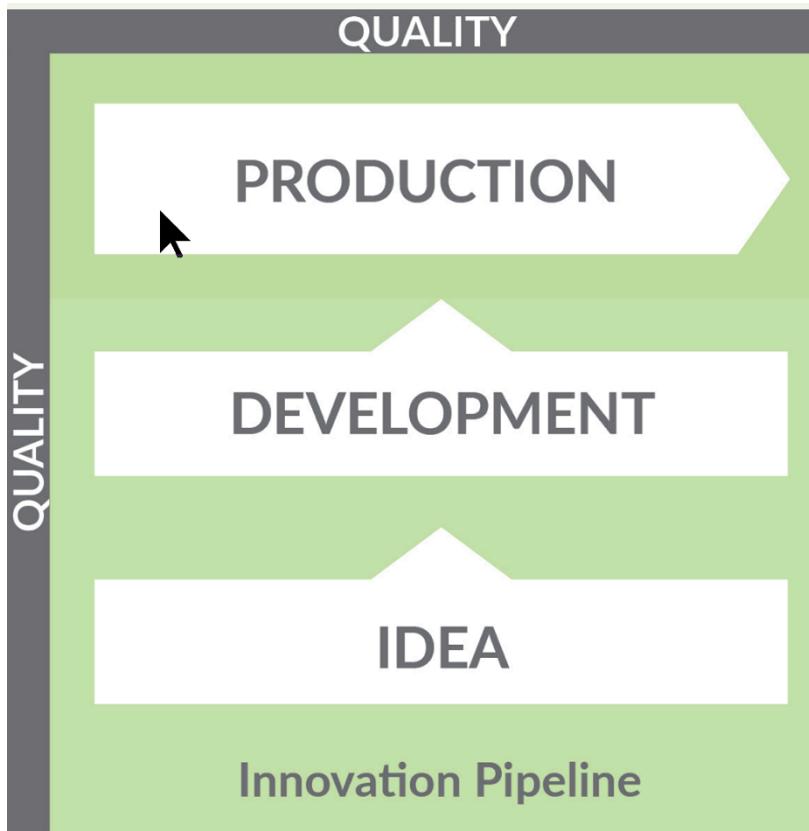


**'DevOps' Governed Systems**

# They run a ‘Factory’ of Insight

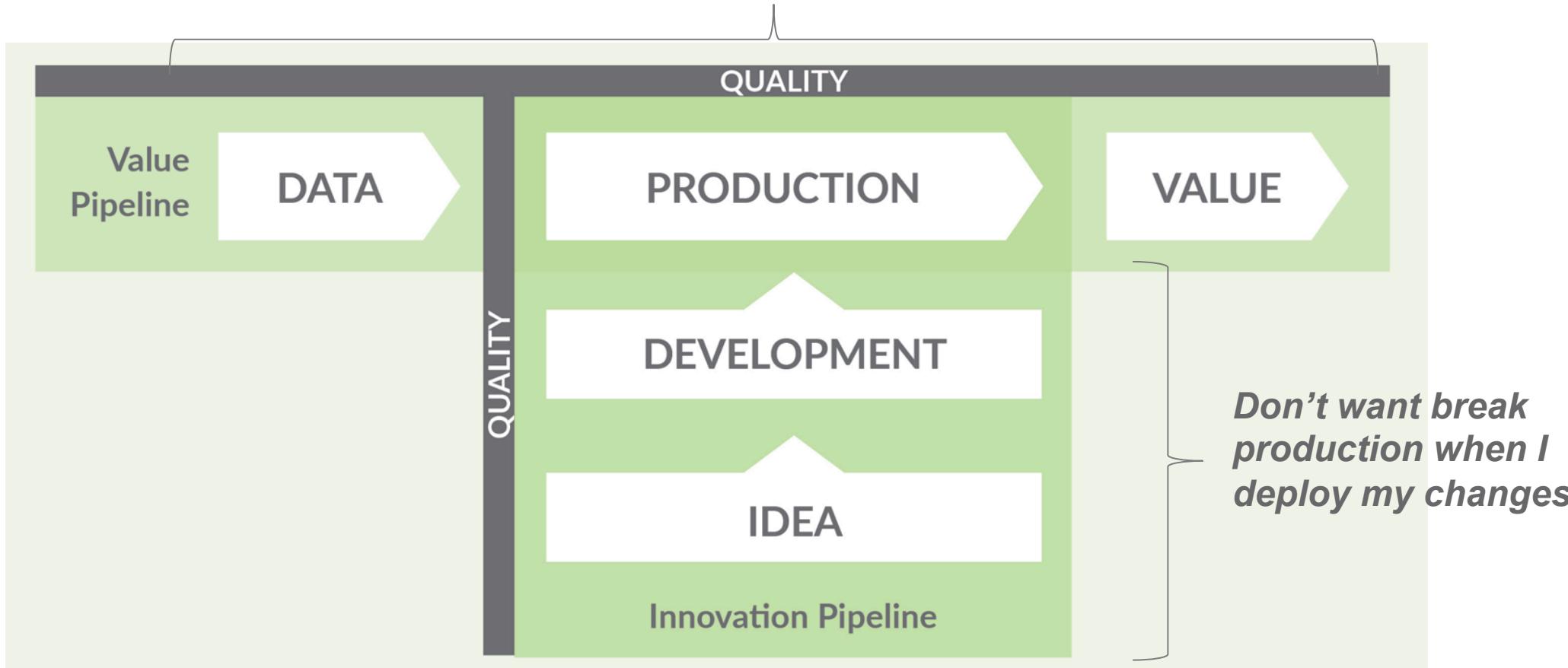


# And need to deploy quickly from dev to production



# And need to do both simultaneously

*Don't want to learn about data quality issues from my customers*



# Agenda



A BIG PROBLEM  
THAT CAN USE  
YOUR HELP



BY HAVING  
EMPATHY FOR A  
GROUP OF PEOPLE



THAT ARE  
SUFFERING

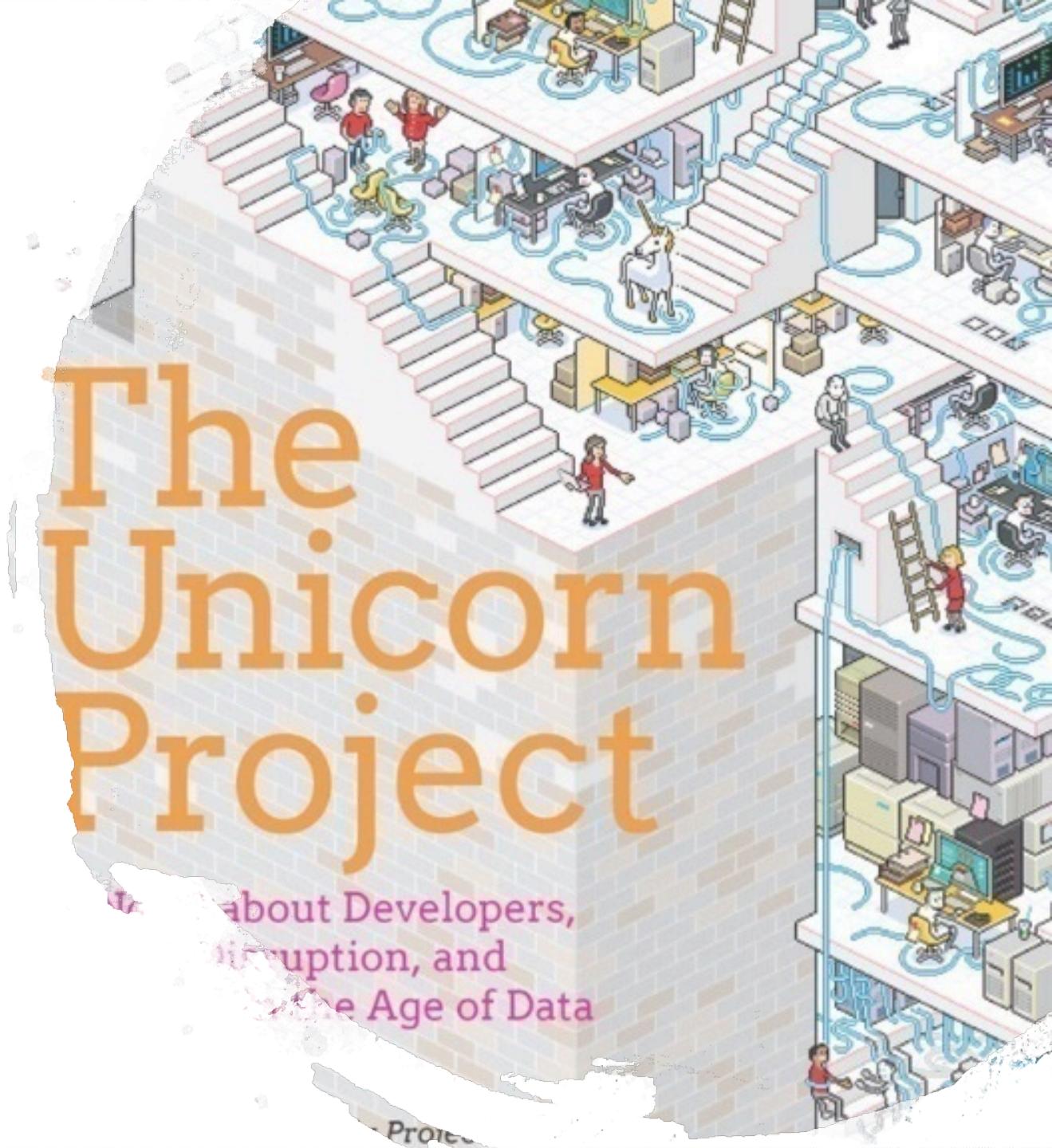


AND WHAT YOU  
KNOW CAN HELP  
THEM

# Your Data Nerd Friends Are Suffering

- Hero culture
- Fear culture
- Insanely high error rate
- Complete lack of automated testing
- Deploy to product rates of months
- Lots of hope, heroism and fear.
- Technology Review Boards

**Project Panther! It's a subplot in Gene's new book for a reason**

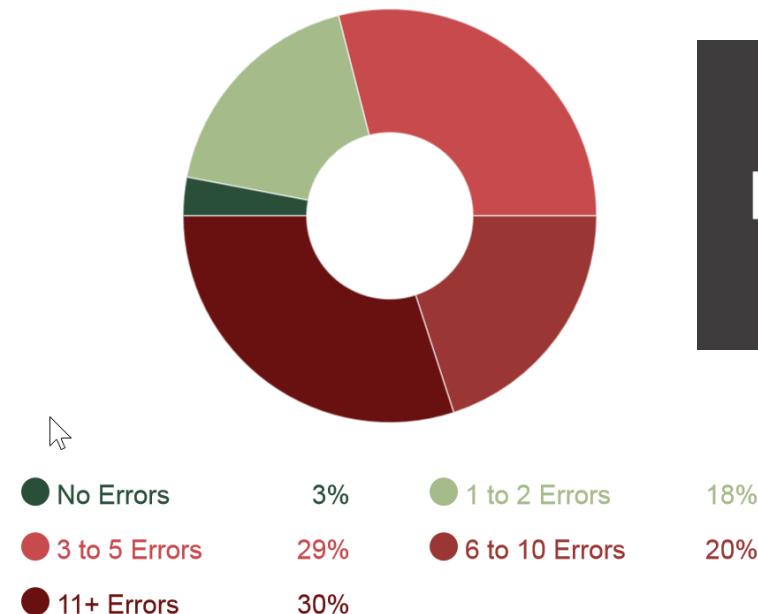


# Currently, Teams Have High Errors

## DataKitchen/Eckerson Survey (May 2019)



On average, how many errors (e.g., incorrect data, broken reports, late delivery, customer complaints) do you have each month?

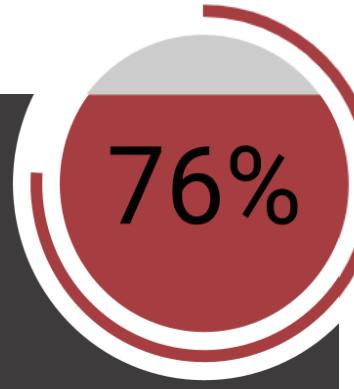


*DataKitchen / Eckerson Research Survey of Medium – Large Companies US And Europe*

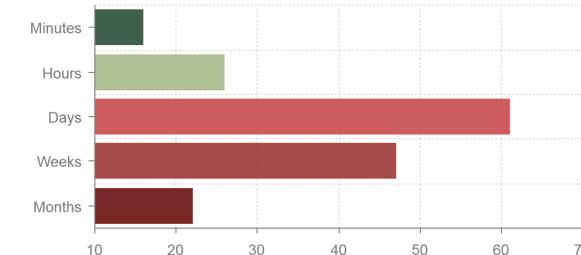
# Currently, Teams Struggle to Deploy



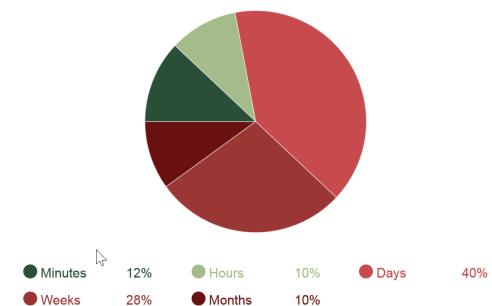
## DataKitchen/Eckerson Survey (May 2019)



On average, how long does it take to move a new or modified data analytic pipeline from development to production?



On average, how long does it take your team to create a new development environment with the appropriate test data, servers, and tools?



*DataKitchen / Eckerson Research Survey of Medium – Large Companies US And Europe*



# **My Story**

---

# Agenda



A BIG PROBLEM  
THAT CAN USE  
YOUR HELP



BY HAVING  
EMPATHY FOR A  
GROUP OF PEOPLE

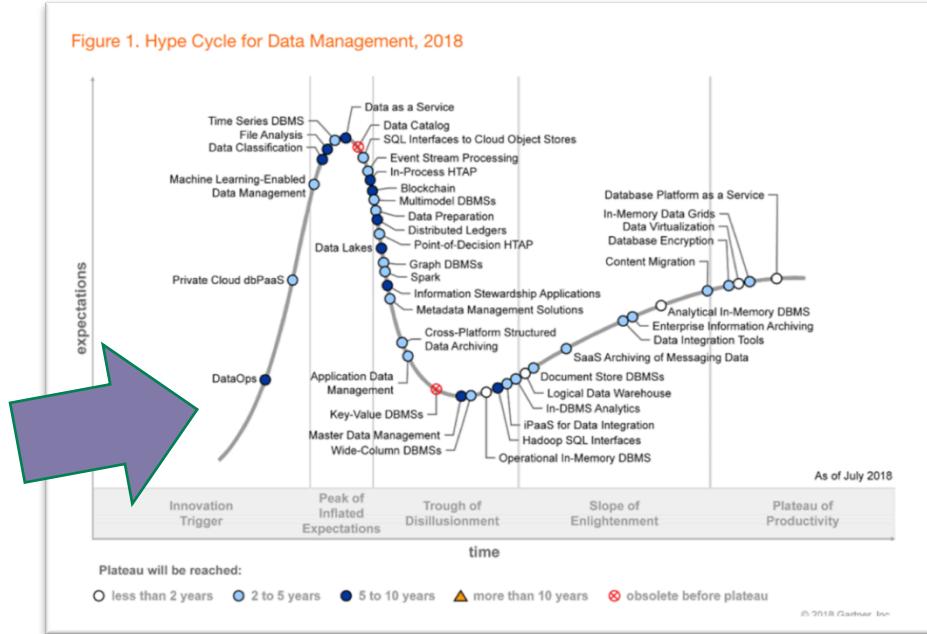


THAT ARE  
SUFFERING

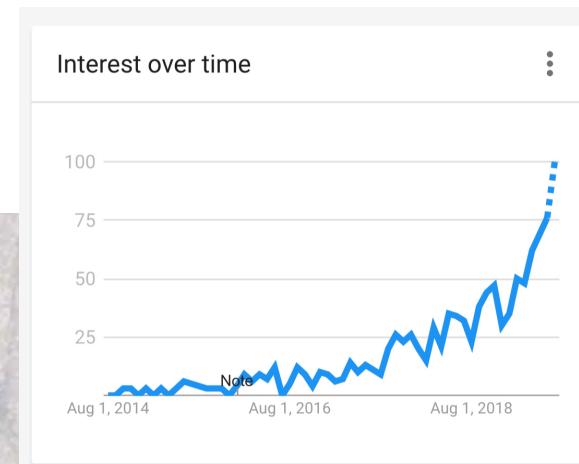


AND WHAT YOU  
KNOW CAN HELP  
THEM

# DataOps is having a moment



- DataOps Manifesto 2017
  - 6000 signatures
- Gartner Hype Cycle in late 2018
- Increased market adoption of DataOps principles by leaders of data and analytic teams in 2019



## The DataOps Manifesto

Through firsthand experience working with data across organizations, tools, and industries we have uncovered a better way to develop and deliver analytics that we call DataOps.

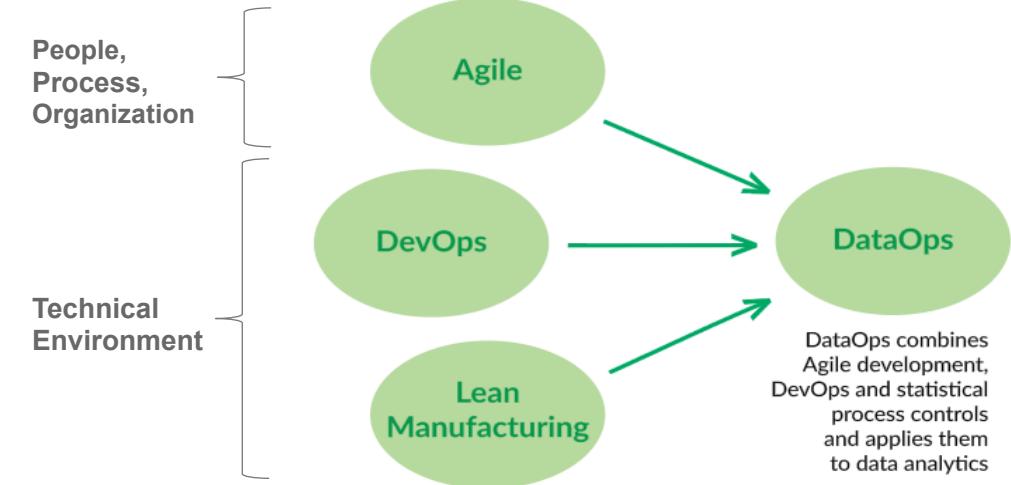


# DataOps – Transformative to Data Analytics



DataOps is a set of technical practices, cultural norms, and architecture that enable:

- Rapid experimentation and innovation for the fastest delivery of new insights to our customers
- Low error rates
- Collaboration across complex sets of people, technology, and environments
- Clear measurement and monitoring of results



*“Organizations that adopt a DevOps- and DataOps-based approach are more successful in implementing end-to-end, reliable, robust, scalable and repeatable solutions.”*

*Sumit Pal, Gartner, November 2019*

# How To Succeed?

A Mindset Change to ...



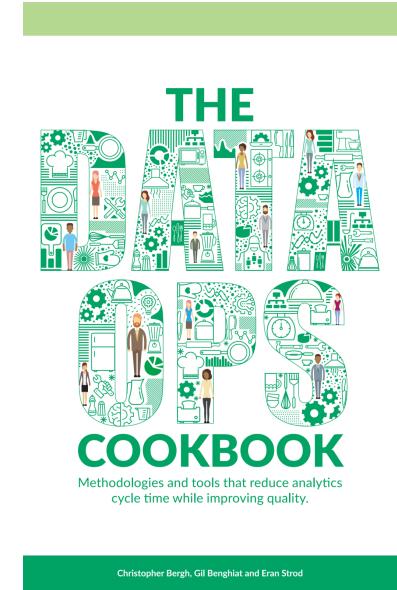
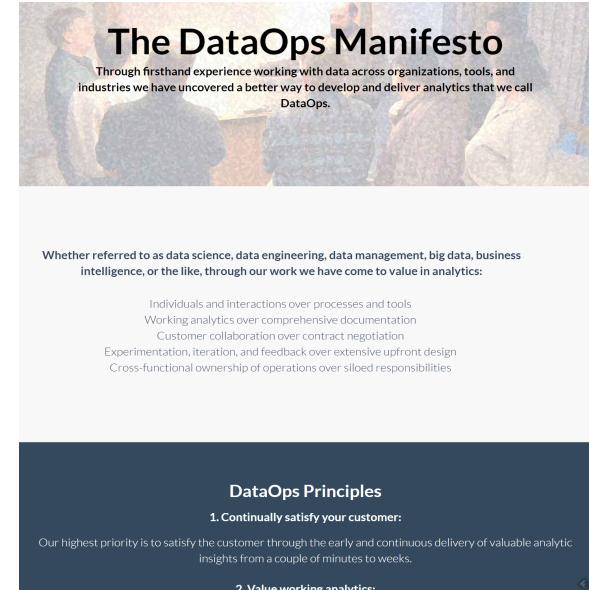
From	To
Change Fear	Change Velocity
Manual Operations	Automated Operations
Hope For Quality	Integrated Quality
Hero Mentality	Repeatable Processes
Heads Down	Collaboration
Vendor Lock-In	Diverse Tools

**...to power your highly agile data culture.**

# Education: Seven Steps to DataOps (+3)



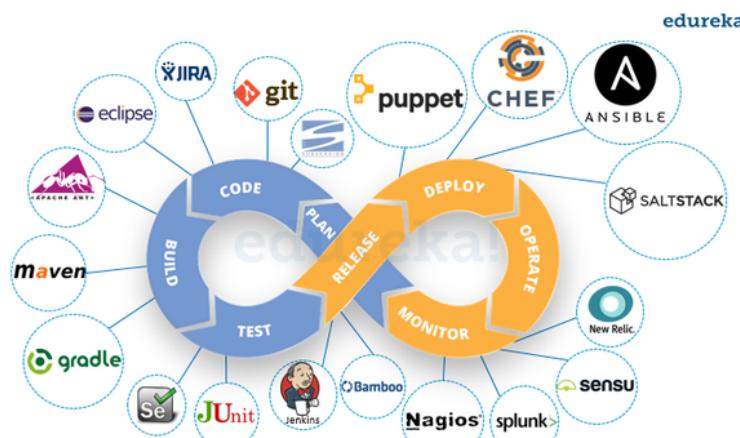
1. Orchestrate Two Journeys
2. Add Tests And Monitoring
3. Use a Version Control System
4. Branch and Merge
5. Use Multiple Environments
6. Reuse & Containerize
7. Parameterize Your Processing



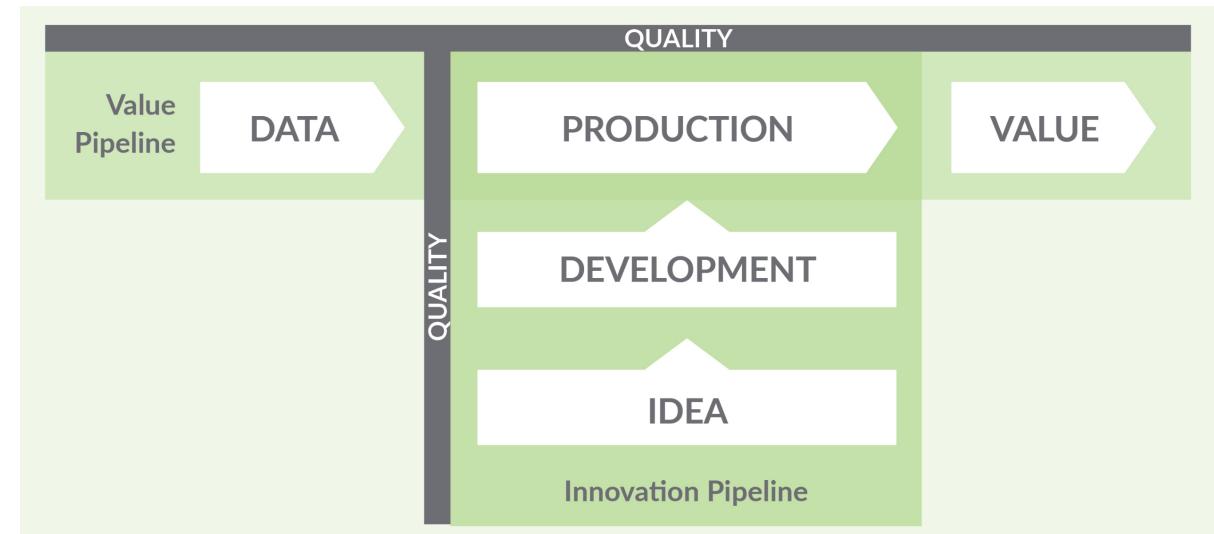
+ Three (Architecture, Metrics and Inter/Intra Team Collaboration)

# DevOps vs DataOps:

**DataOps contains many of the same concepts as software development and many unique to data analytics**

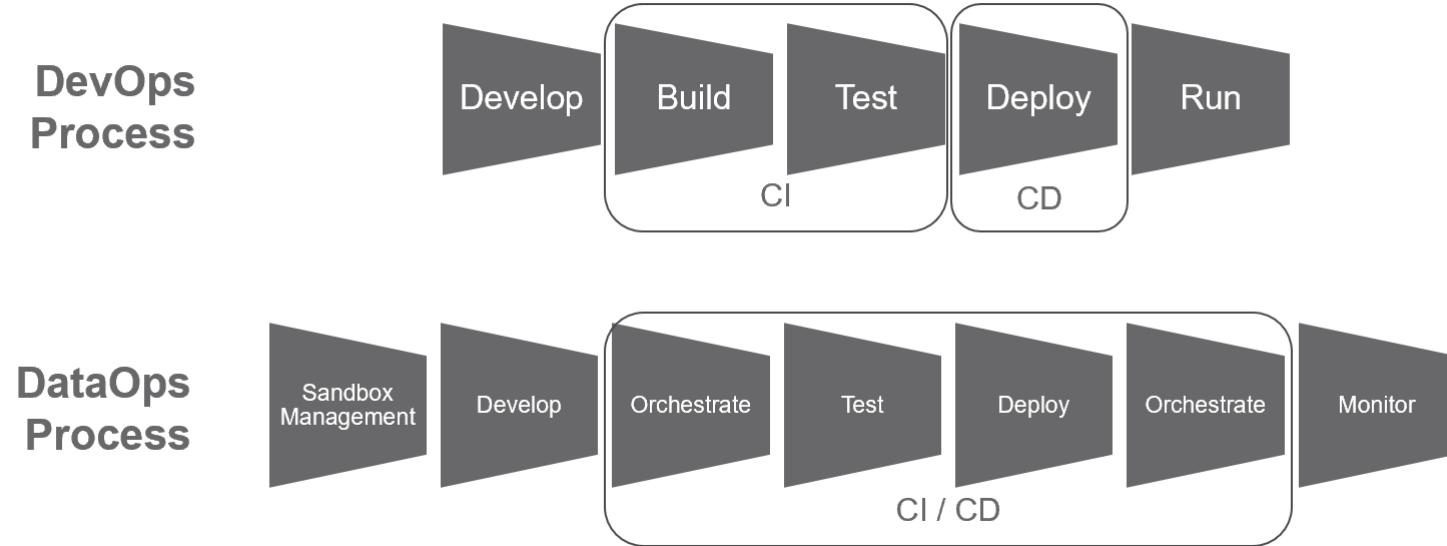


**DevOps**



**DataOps**

# DevOps vs DataOps:



- 1. Different Process, Different People and Expectations**
- 2. DevOps 1:1 DataOps Many:Many**
  - Multiple 'Dev' and 'Ops' groups
- 3. DataOps Views Data Analytics as 'Factory'**
  - Multi-Tool Orchestration Testing, Monitoring and Statistical Process Control
- 4. DataOps Has Additional Development Complexities**

**DEVOPS  
ENTERPRISE  
SUMMIT**

Las Vegas  
October 28-30, 2019

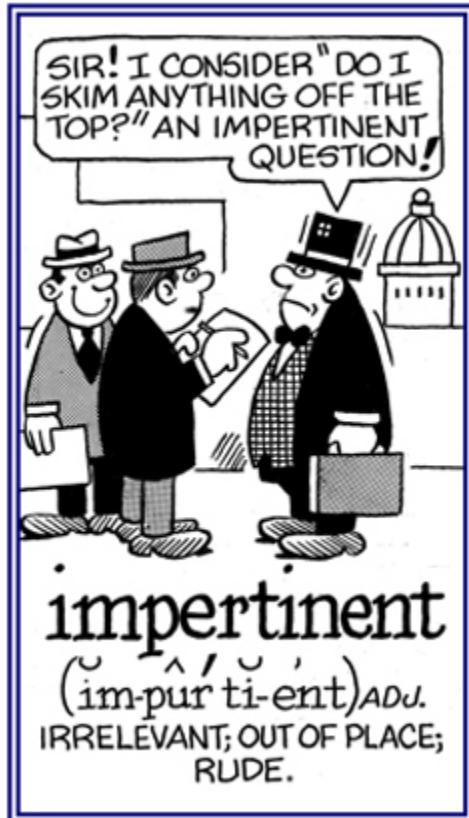


# DevOps Leaders



## Join the DataOps Movement

# Ask your data analytic teams impertinent questions



- Are you using source control for your work?
- How many automated tests do you have in production?
- Do you have regression, functional or unit tests for your work?
- How long does it take to deploy ETL/models/BI report from development to production?
- Do you have automated deployment?
- How up to date is your development environment?
- How often are your business users finding errors in the data?

# Learn More



- For these slides, contact me:
  - cbergh@datakitchen.io
- DataOps Manifesto:
  - <http://dataopsmanifesto.org>
- Free DataOps Cookbook:
  - <https://www.datakitchen.io/dataops-cookbook-main.html>
- Excerpt from Gene's Unicorn Project Book on DataOps
  - <https://www.datakitchen.io/unicorn-project.html>