DATA KITCHEN

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

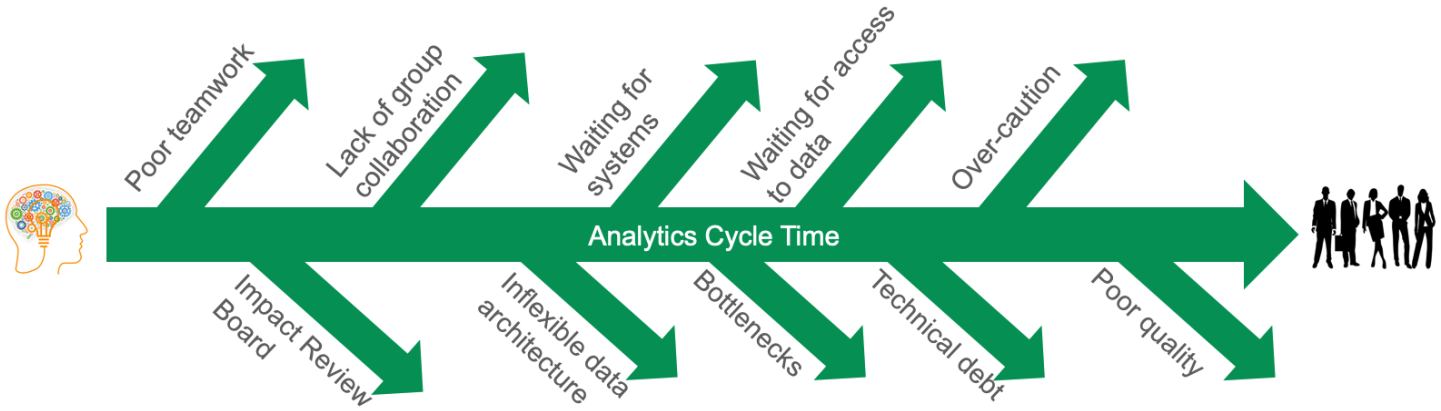
THAT ARE **SUFFERING**

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

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

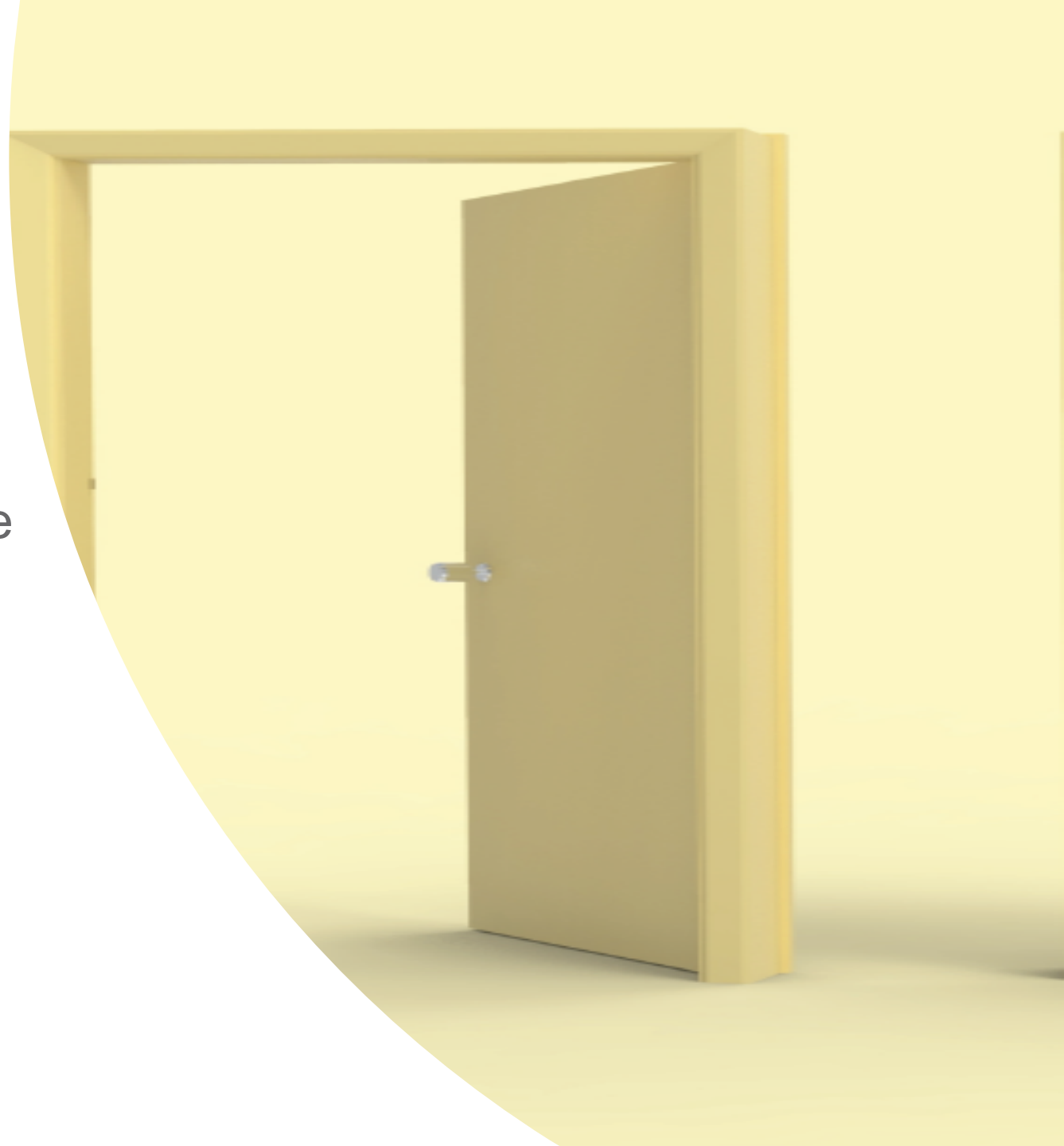
THAT ARE **SUFFERING**

AND **WHAT YOU KNOW CAN HELP** THEM

Who Are These People?

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

Name	Sales
joe	$1234.56
kelly	$4567.89



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

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


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


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

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

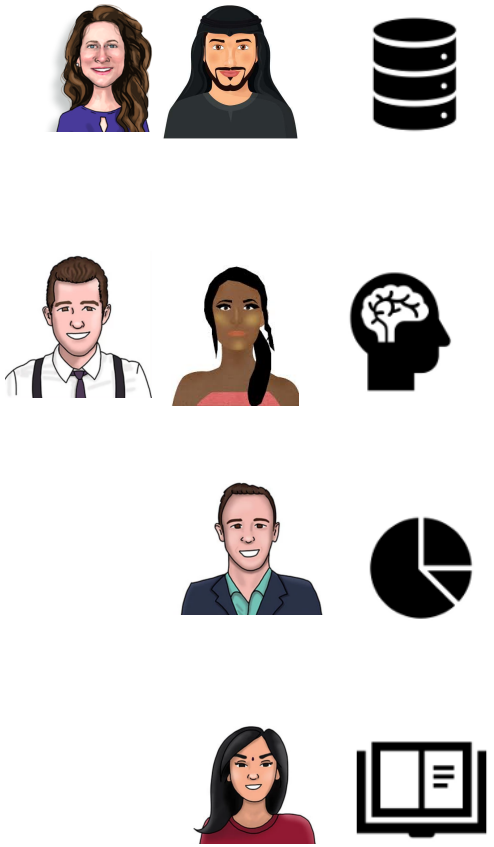
Name	Sales
joe	$1234.56
kelly	$4567.89

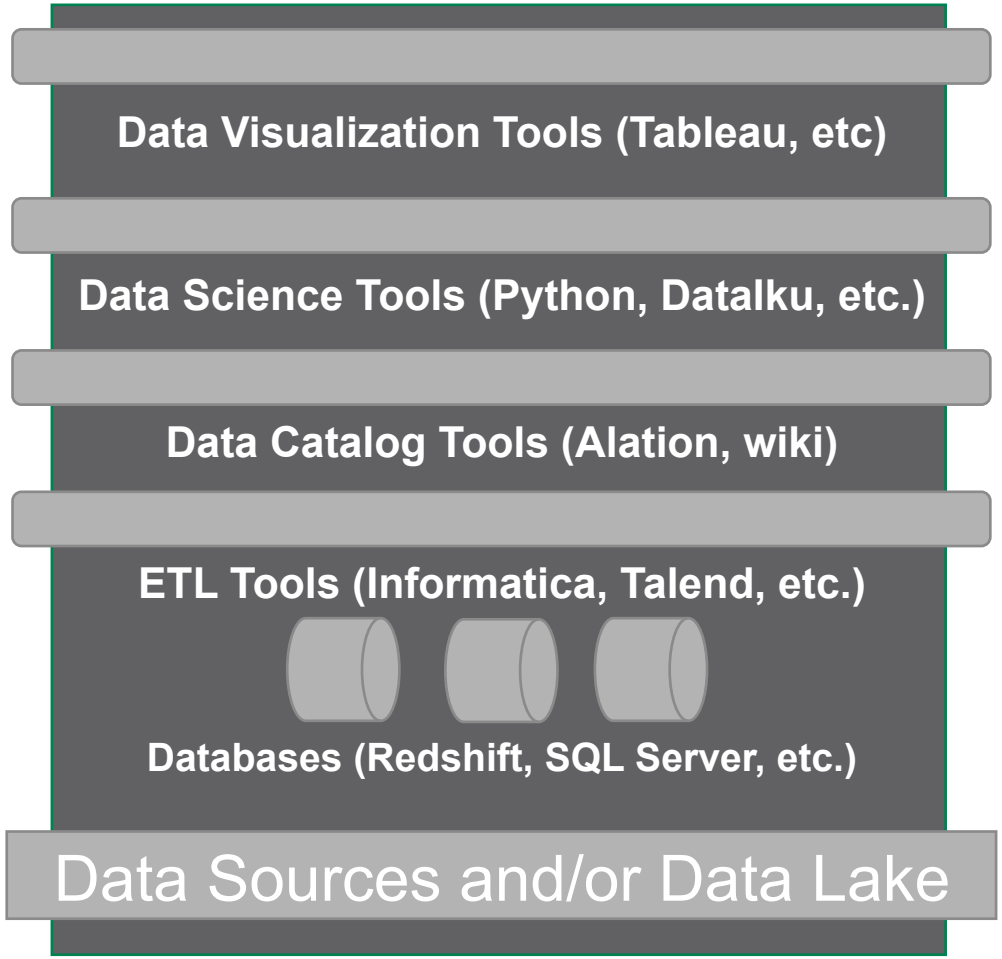
Name	Sales	Segment
joe	$1234.56	Lo Value
kelly	$4567.89	Hi Value

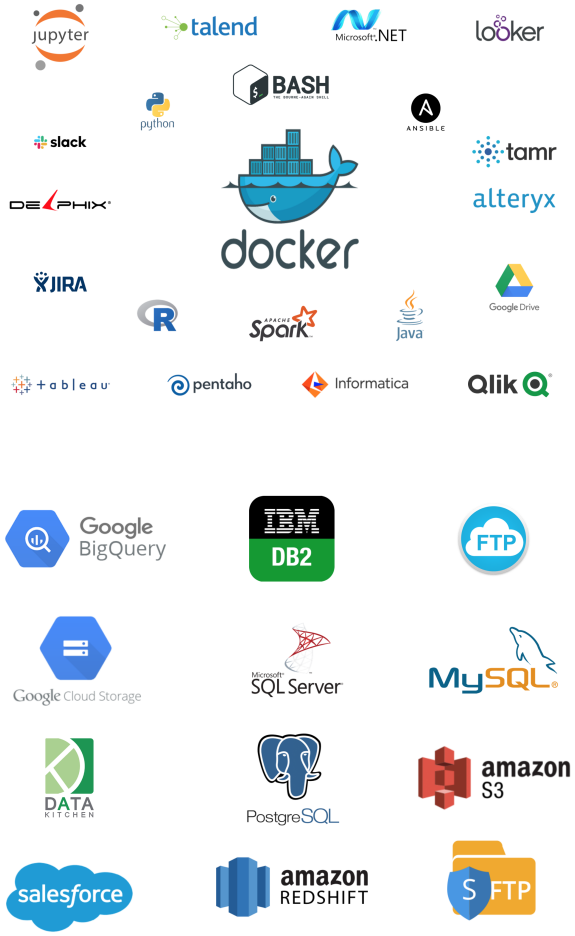
Name	Sales	Segment	Owner
joe	$1234.56	Lo Value	West Team
kelly	$4567.89	Hi Value	East Team

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

- **Data Engineer Team**

- **Data Science Team**

- **Self Service Team**
  - Visualize Data and Model

- **Data Governance Team**

Chief Data Officer
Chief Analytics Officer

Or not

- Data Engineer Team



- Data Science Team


- Self Service Team

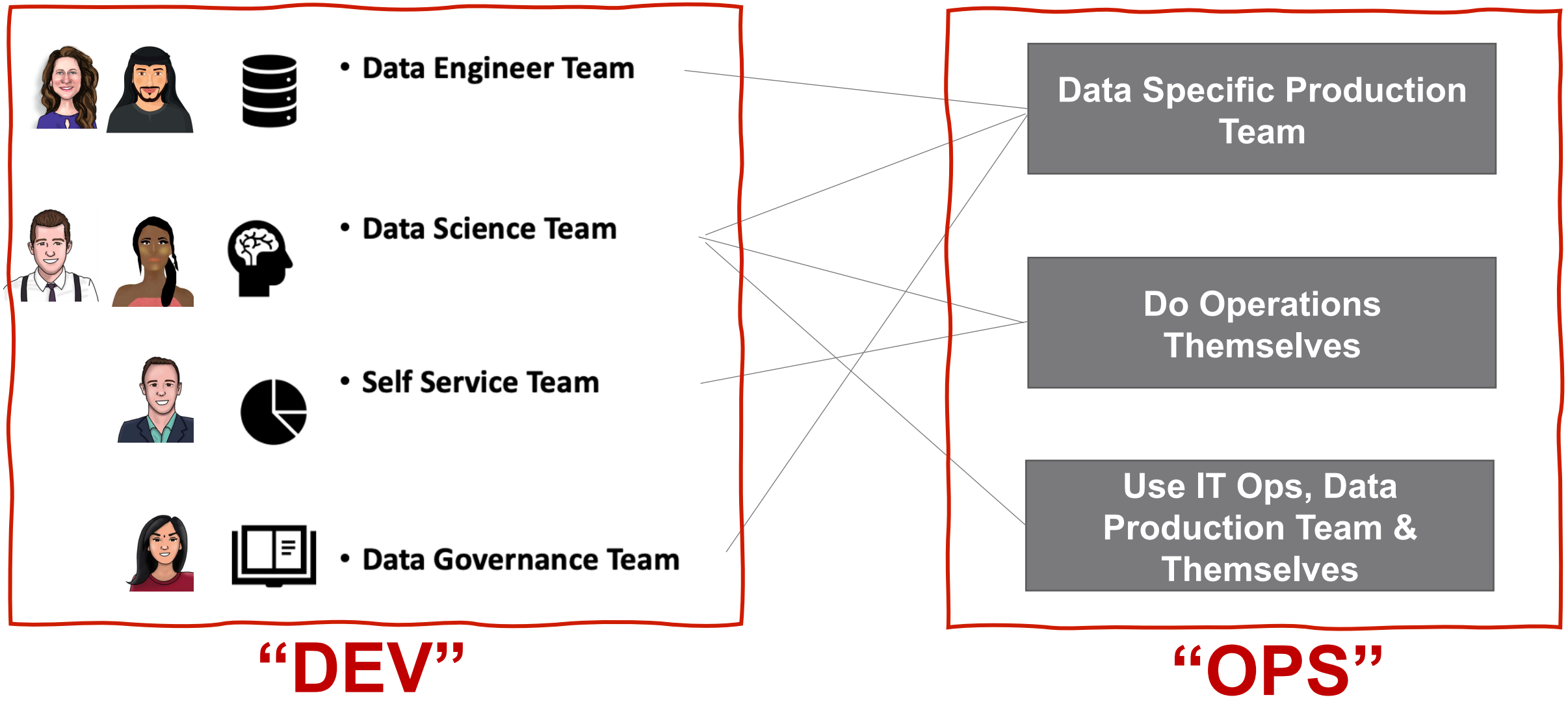

- Data Governance Team

CIO or CDO

Line of Business Executives

CEO

A many to many dev/ops relationship



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

DATA KITCHEN

Home Office Team

Data Engineer

Data Scientist

Boston

Centralized, Weekly Cadence of Changes

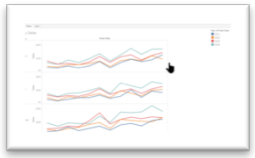
Local Office 'Self-Service' Team

Data Analyst

New Jersey

Distributed, Daily/Hourly Cadence of Changes

VP Marketing



Challenges With Coordination

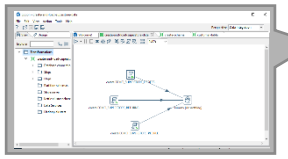
DATA KITCHEN

Home Office Team

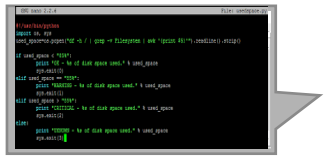
Make a change in schema?

Not Available For All?

New Data & Schema



Data Engineer



Data Scientist

Local Office Team

Break Reports?

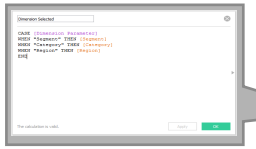
Add New Data Sets

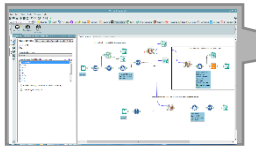
Change Report Calculations

Inconsistencies?

Update/New Report

Not Working?

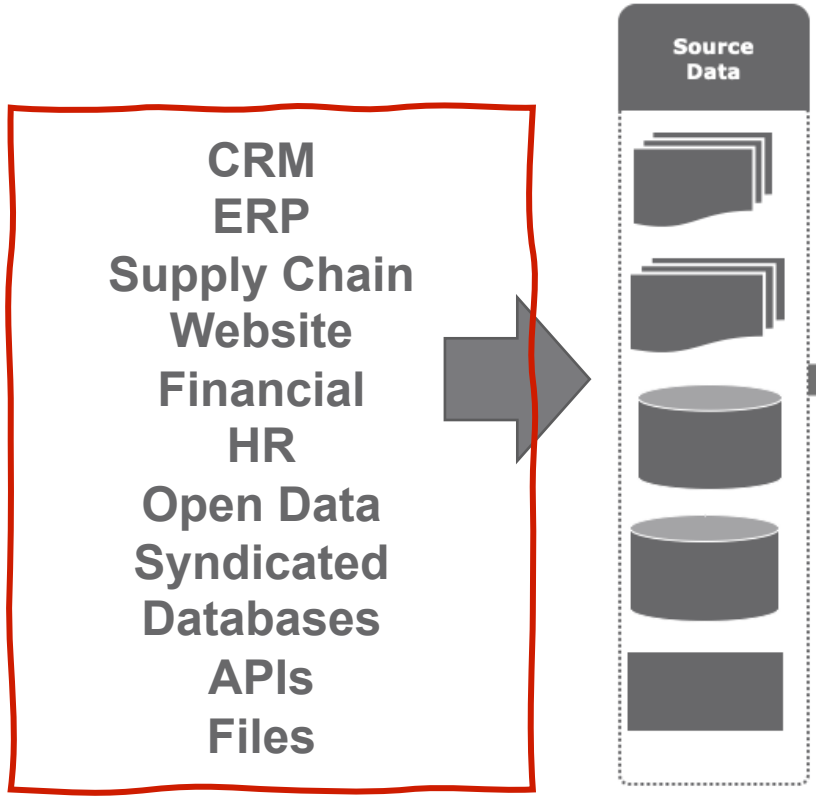




Data Analyst

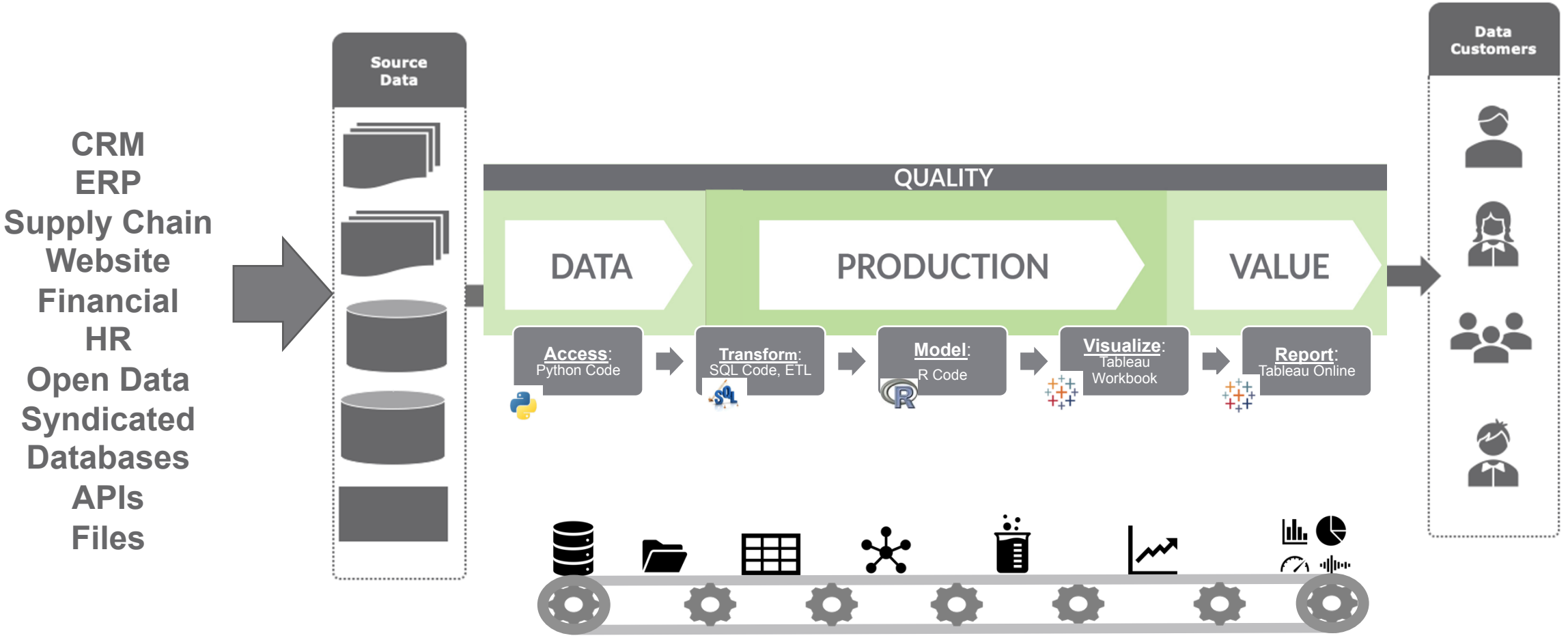
VP Marketing

And they source data from internal and external system

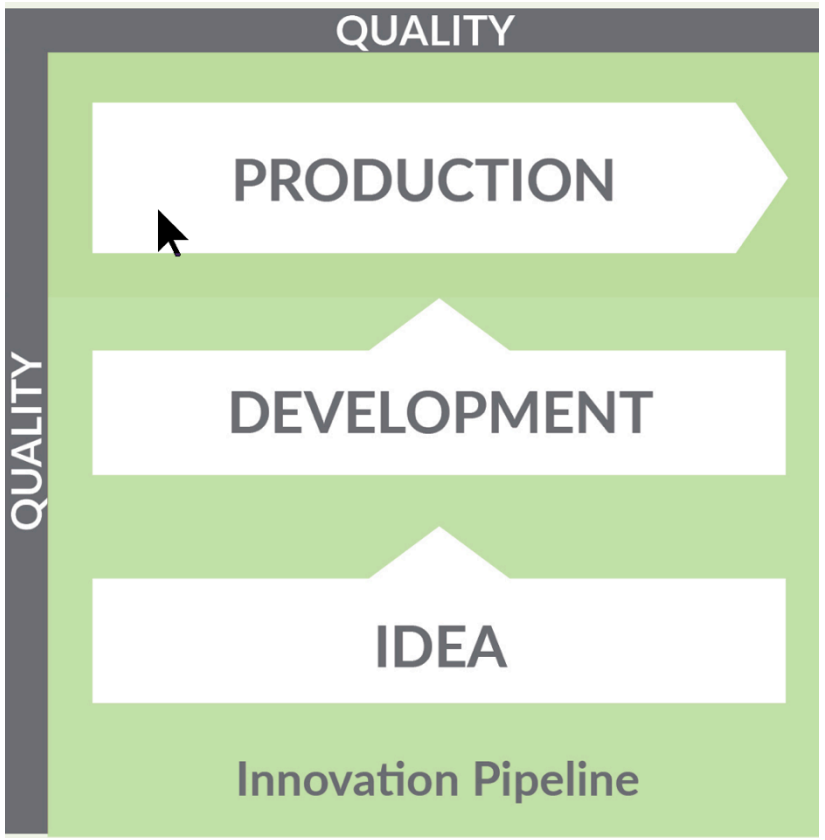


‘DevOps’ Governed Systems

They run a ‘Factory’ of Insight



And need to deploy quickly from dev to production



Diverse Customers

Business Customer

Products & Systems

Diverse Team

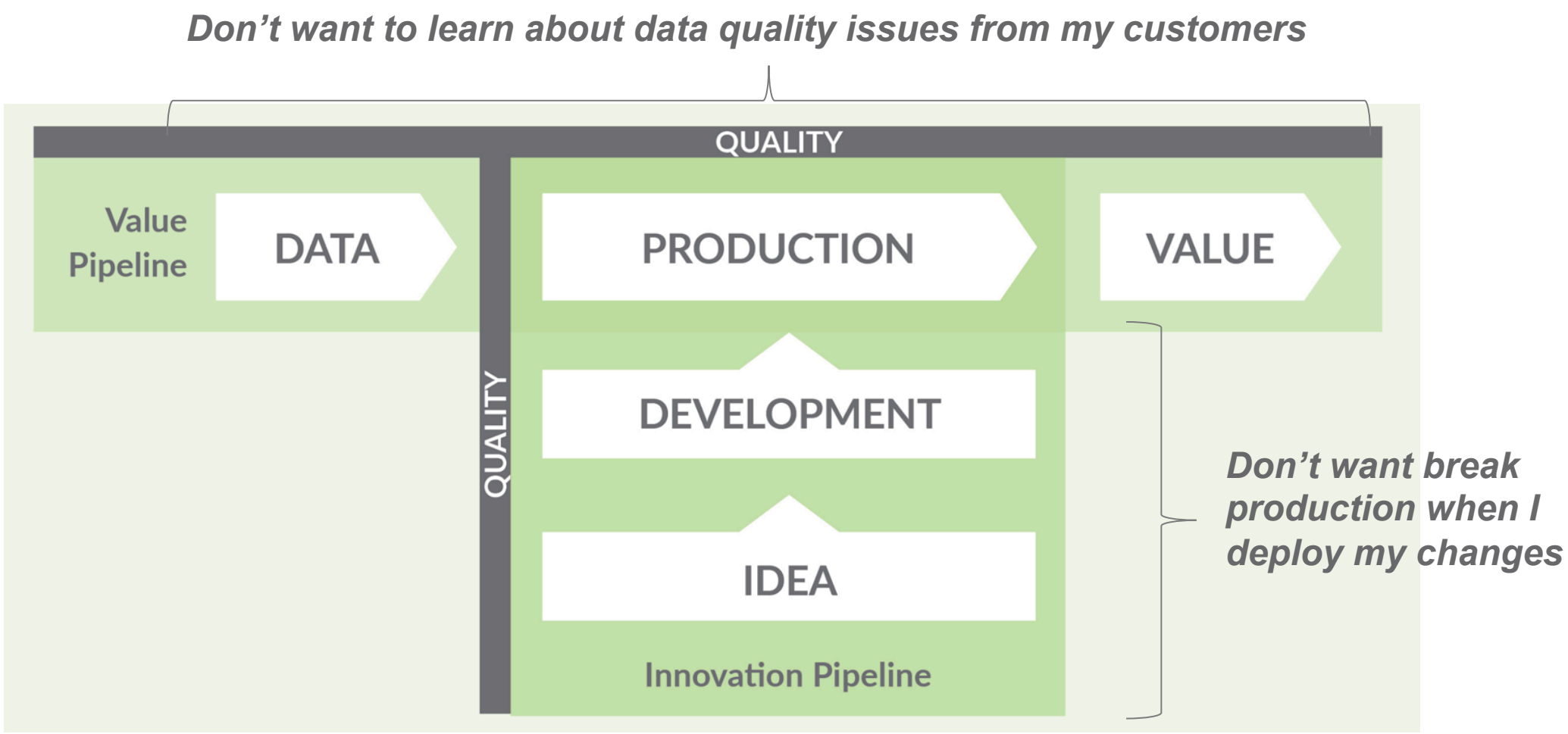
Data Engineers

Data Scientists

Data Analysts

Diverse Tools

And need to do both simultaneously



Agenda

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

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

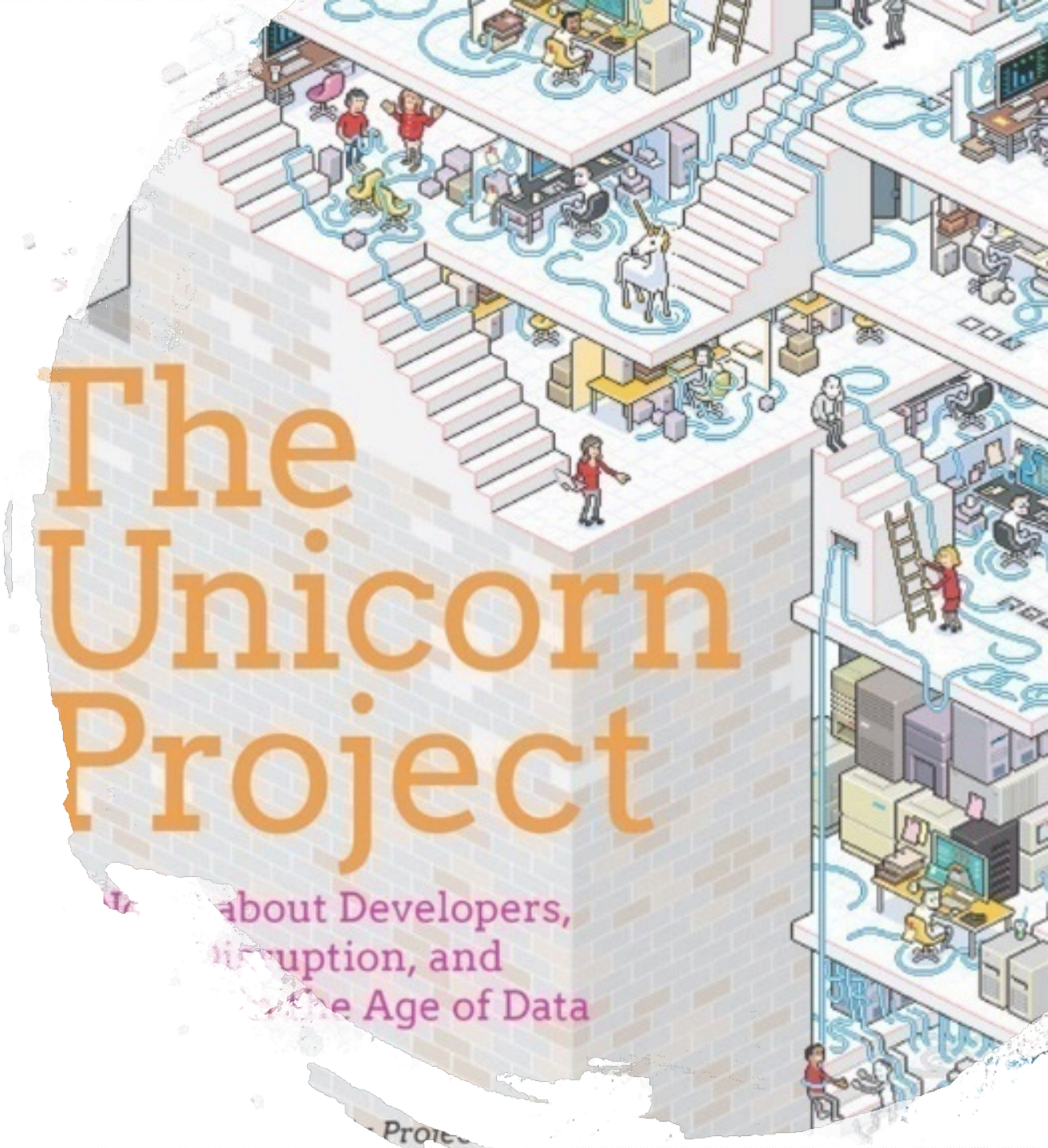
THAT ARE **SUFFERING**

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

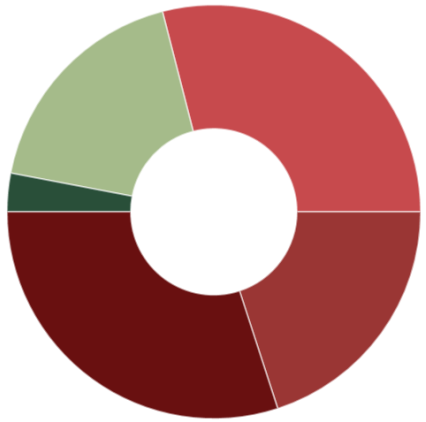
Project Panther! It's a subplot in Gene's new book for a reason



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?



No Errors	3%	1 to 2 Errors	18%
3 to 5 Errors	29%	6 to 10 Errors	20%
11+ Errors	30%		

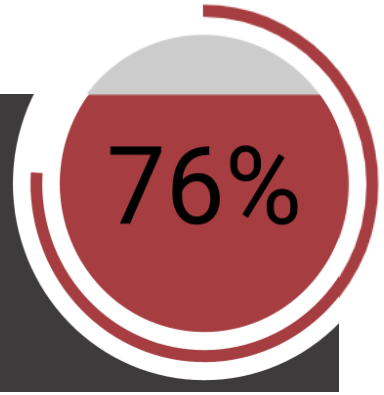
DataKitchen / Eckerson Research Survey of Medium – Large Companies US And Europe

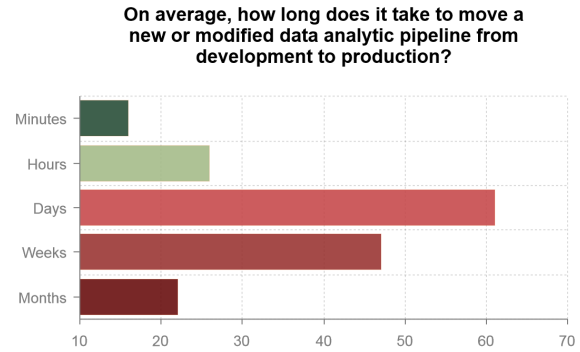
Currently, Teams Struggle to Deploy

DataKitchen/Eckerson Survey (May 2019)

DataKitchen Interpretation

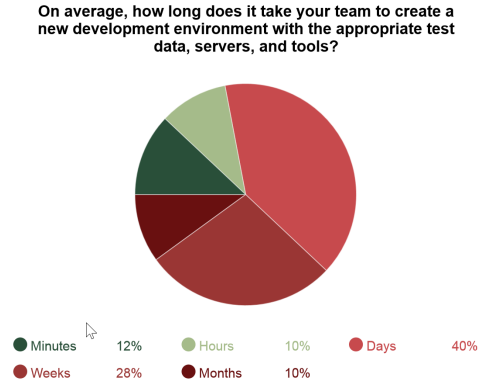
MOST COMPANIES ARE TOO SLOW TO DEPLOY CHANGES INTO PRODUCTION





DataKitchen Interpretation

MOST COMPANIES ARE VERY SLOW CREATING NEW DEVELOPMENT ENVIRONMENTS



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

My Story

Agenda

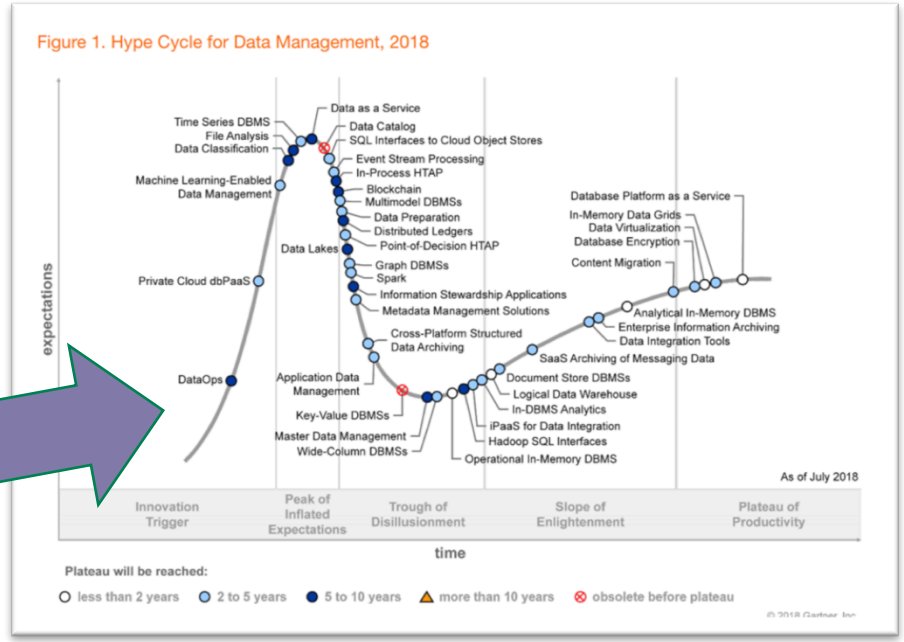
A **BIG PROBLEM** THAT CAN USE YOUR HELP

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

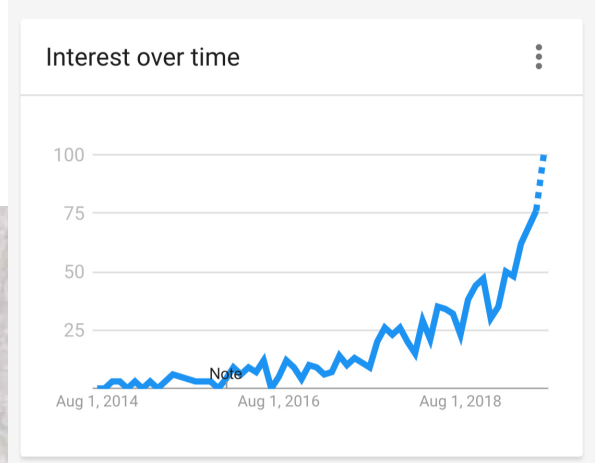
THAT ARE **SUFFERING**

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

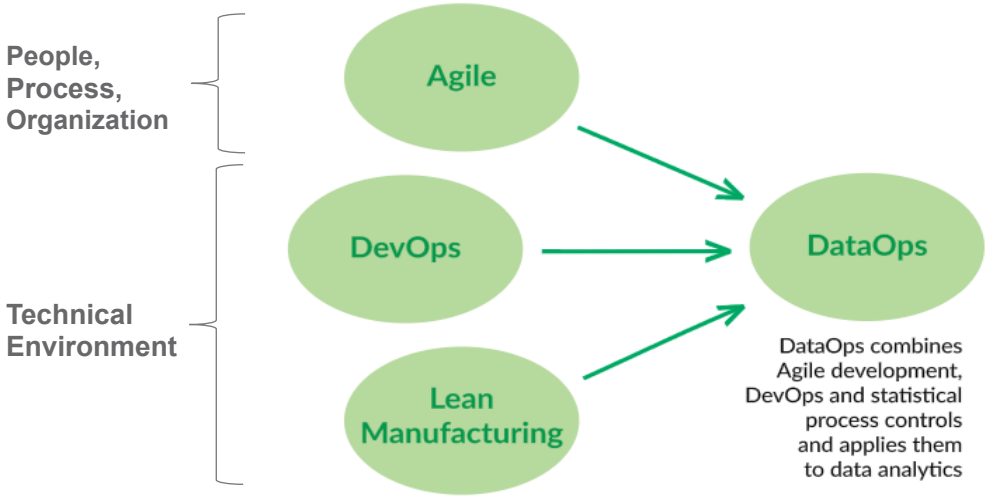


Gartner COOL VENDOR 2019 DATA KITCHEN

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

Source: Gartner

How To Succeed?

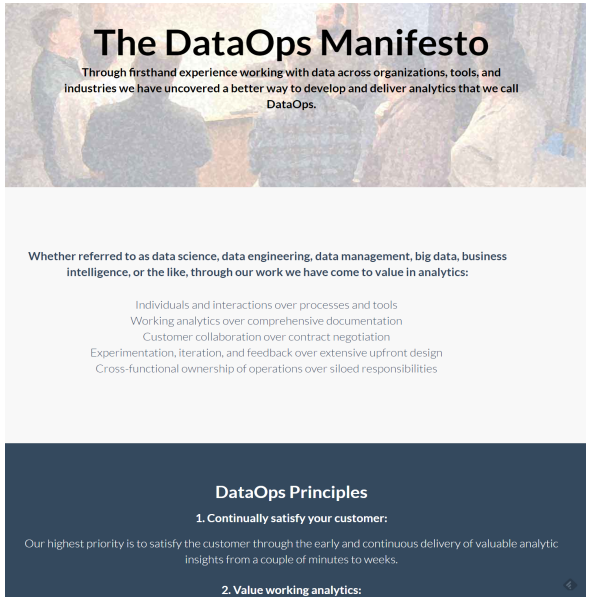
A Mindset Change to …

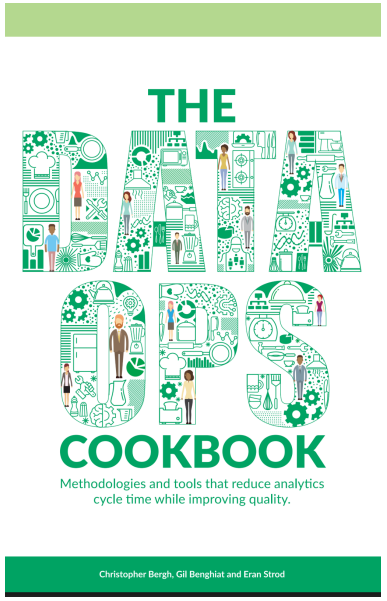
From	To
Change Fear	**Change Velocity**
Manual Operations	**Automated Operations**
Hope For Quality	**Integrated Quality**
Hero Mentality	**Repeatable Processes**
Heads Down	**Collaboration**
Vendor Lock-In	**Diverse Tools**

…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

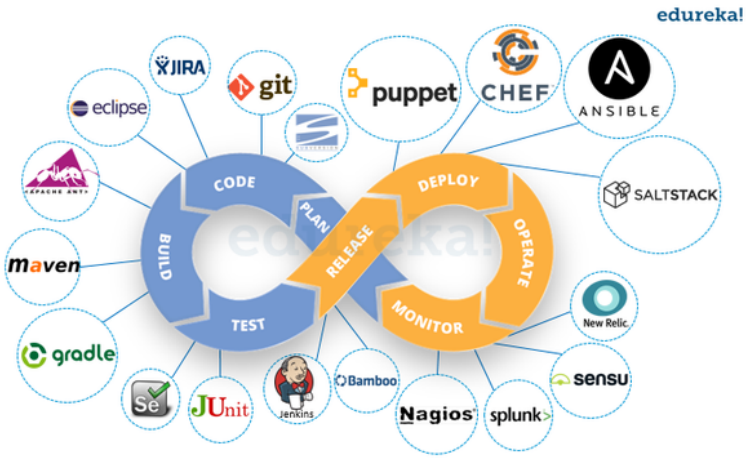




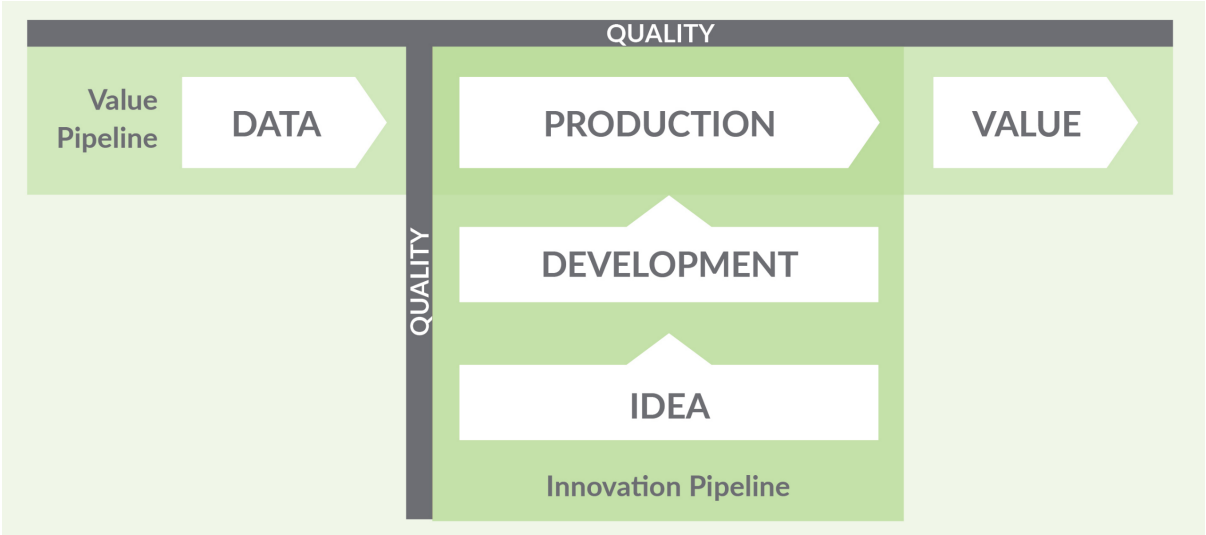
+ Three (Architecture, Metrics and Inter/Intra Team Collaboration)

DevOps vs DataOps:

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

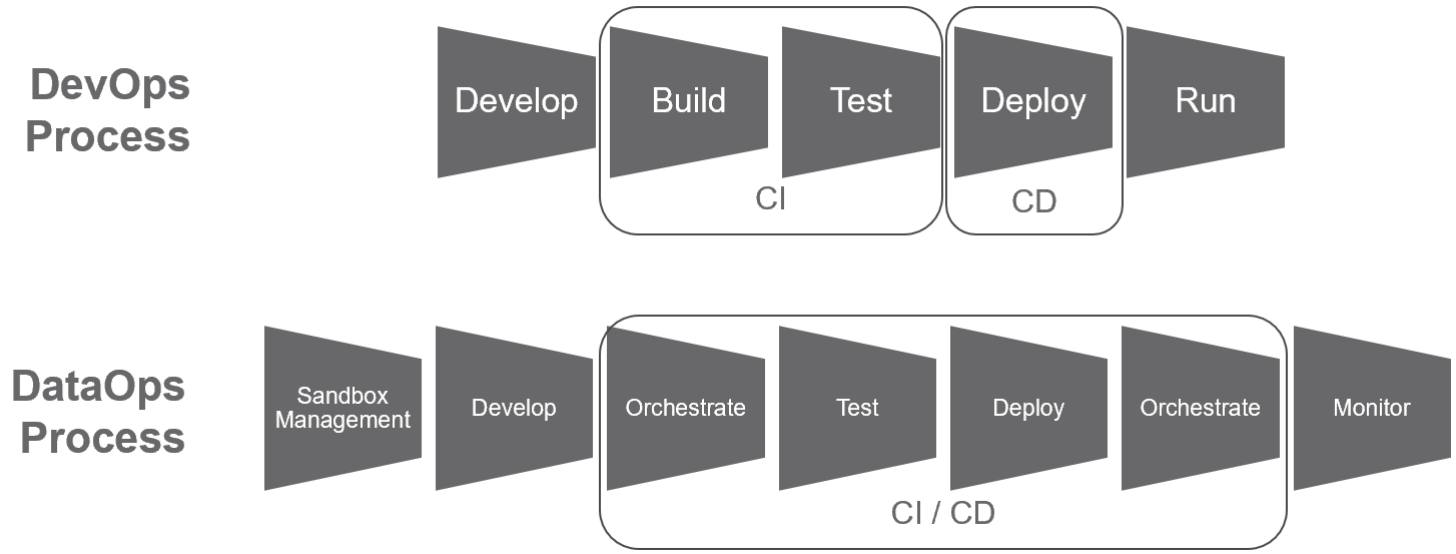


DevOps



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

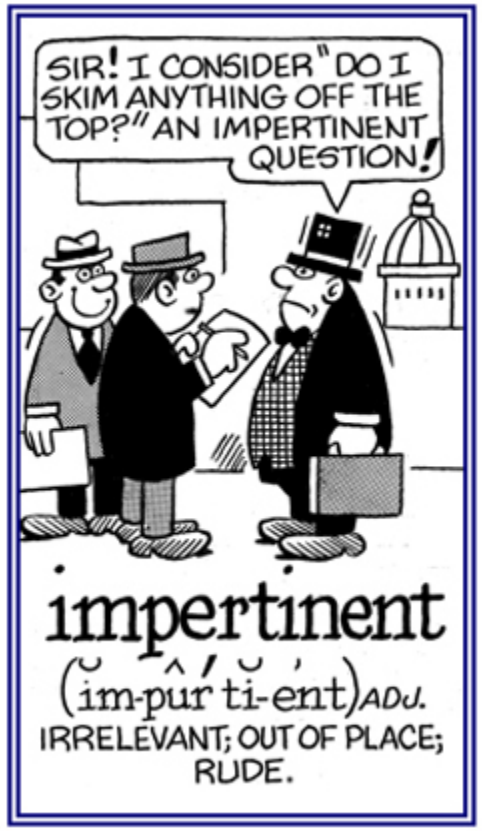
Las Vegas
October 28–30, 2019

DATA KITCHEN

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 to you have in production?
- Do you have regression, functional or unit tests for you 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