DATA KITCHEN

DAMA Phoenix

The DataOps Transformation

www.datakitchen.io

Gartner COOL VENDOR 2019

Speaker: co-Founder of DataKitchen

Gil Benghiat, Founder, VP of Products

🏷️ *A series of data centric software projects*

🎓 Brown, Stanford

🏢 Bell Labs, Sybase, PhaseForward, LeapFrogRx

✉️ gil@datakitchen.io

DataKitchen DataOps Software Platform

Automatically deliver and operate all your data analytic pipelines - simply, collaboratively, and at enterprise scale - with DataKitchen's DataOps Platform.

Iterate to Innovate.

Main Features

1. **Orchestrate** complex data pipelines
2. **Test** and monitor quality
3. **Create** Analytic environments
4. **Deploy** new ideas to production

Topics

Why DataOps Is Essential

Agile in a Nutshell

Seven Steps to DataOps

Bonus Steps to DataOps

Next Steps With DataOps

Topics

Why DataOps Is Essential

Agile in a Nutshell

Seven Steps to DataOps

Bonus Steps to DataOps

Next Steps With DataOps

Exercises

The world changed February 2005

2 day shipping ➔

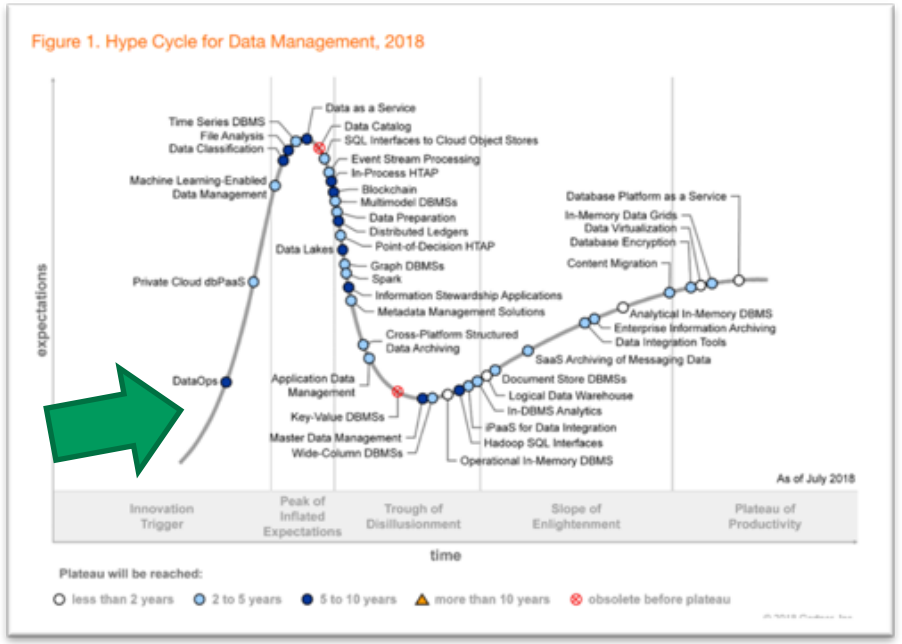
“I am competing with Amazon”

-- James Royster

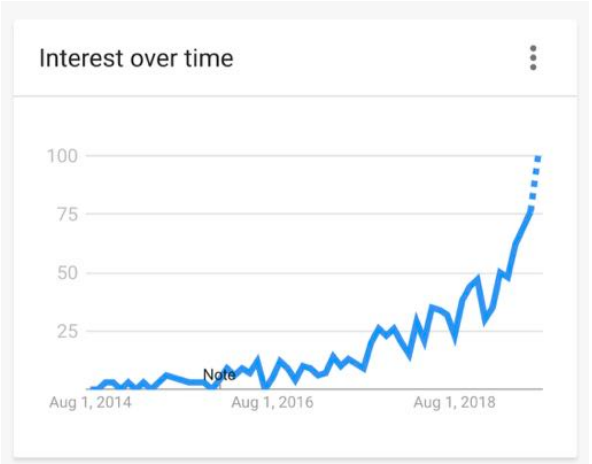
Senior Director, Data Strategy and Operations

Celgene, Inc.

Strategic Trend: DataOps

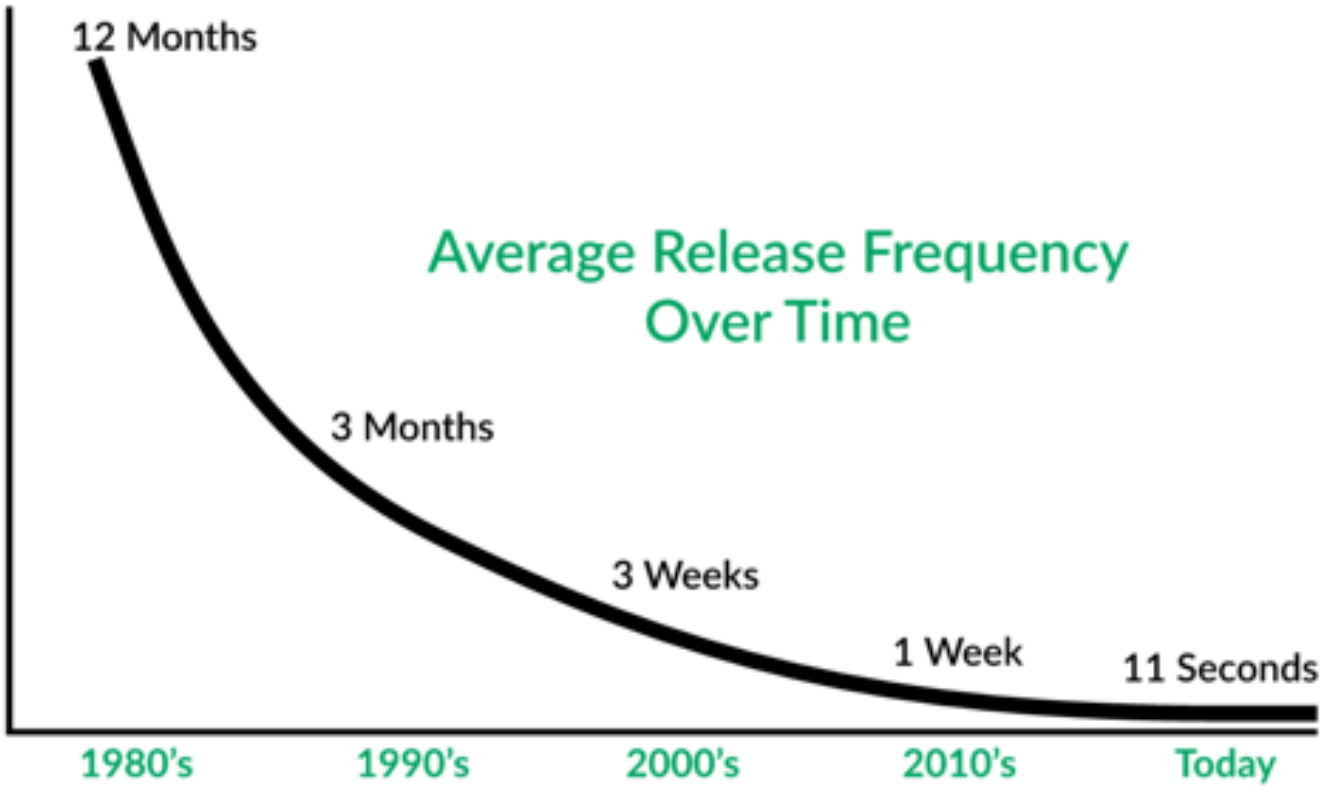


- Increased rate of market adoption of DataOps principles by leaders of data and analytic teams
- Gartner Hype Cycle in late 2018
- Increased Analysts Coverage



DevOps has resulted in a transformative improvement in Software Development

- High-performing IT organizations **deploy 200 times more frequently**
- They have 24 times faster recovery times and three times lower change failure rates
- And they spend 22 percent less time on unplanned work and rework

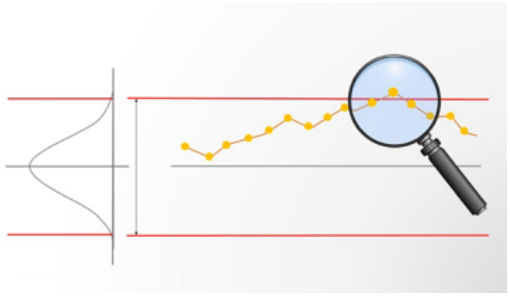


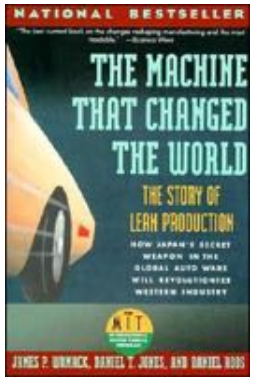
Source: State of DevOps Report

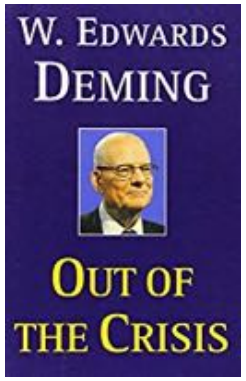
Lean has resulted in a transformative improvement in manufacturing

- Lean manufacturing improves efficiency, reduces waste, and increases productivity.

- The benefits are manifold:
  - Increased product quality
  - Reduces rework
  - Employee satisfaction
  - Higher profits

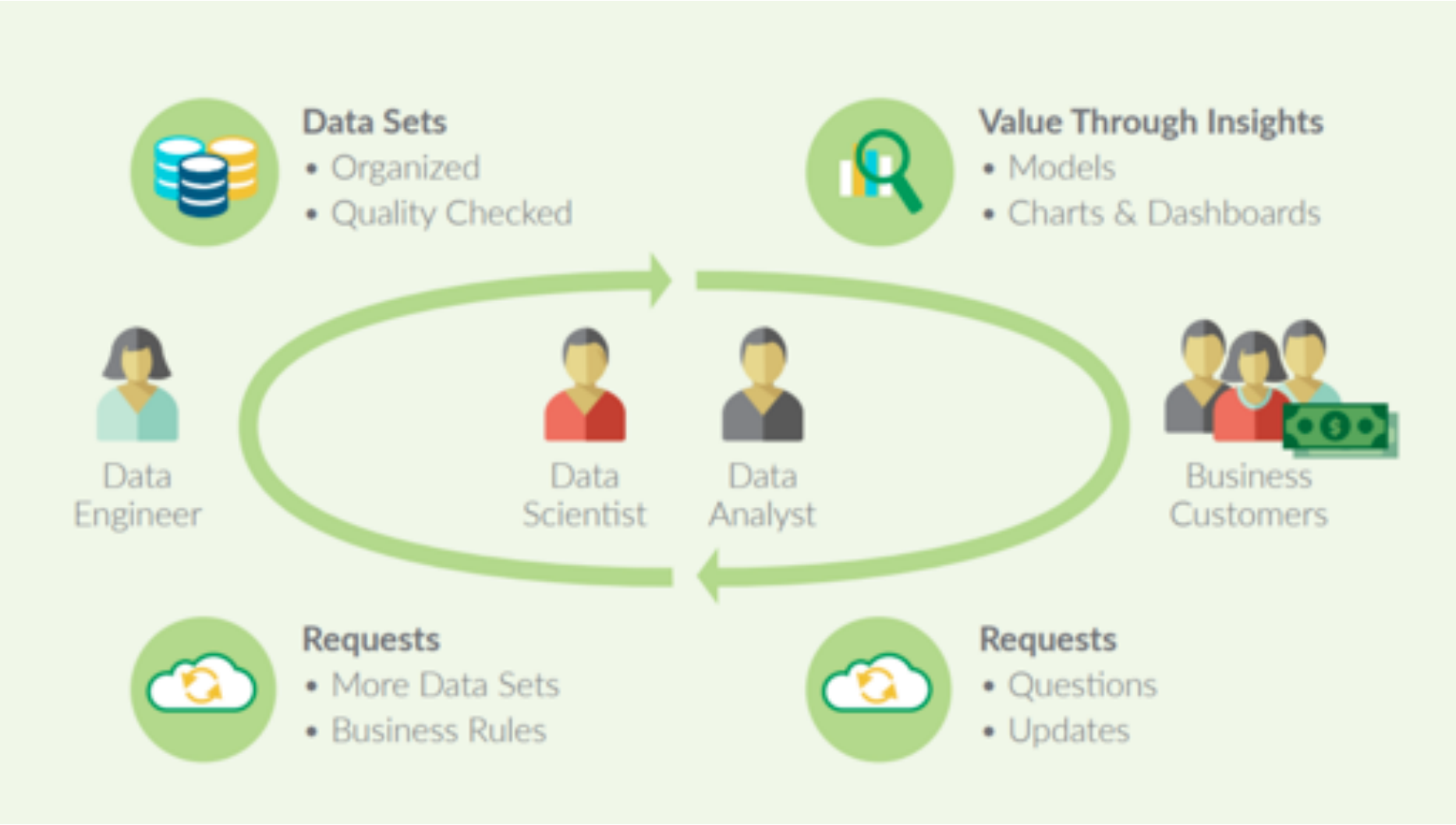






Successful projects generate lots of requests

ANSWERS MORE QUESTIONS MORE CHANGES = ITERATION



Why are world class, game changing data analytics so difficult?

DATA KITCHEN

Key Observation: Innovation Requires Iteration, And Iteration Is Hard

Challenges:

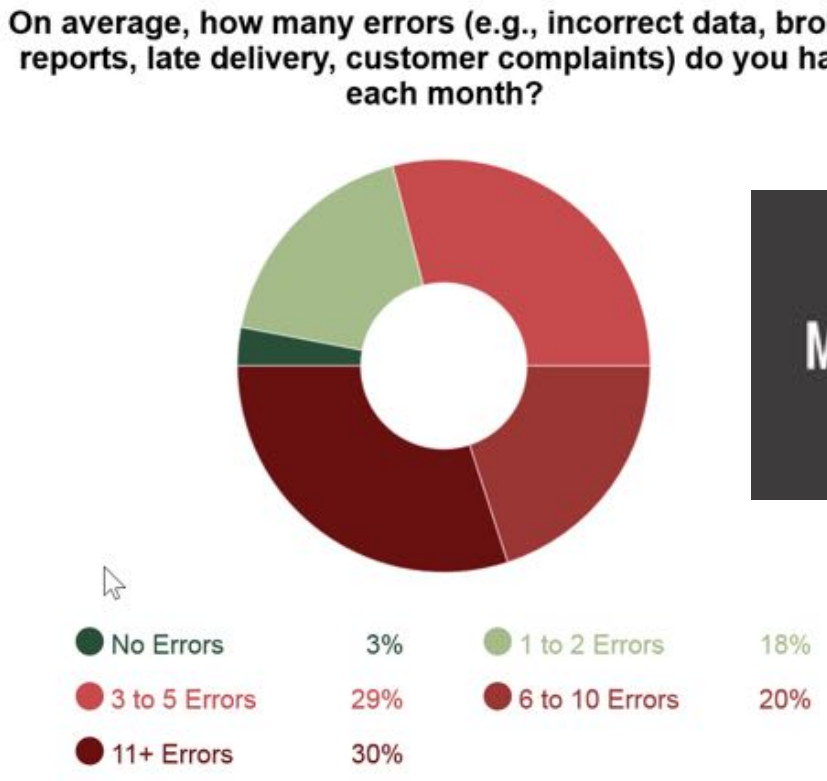
- Innovative insights cannot be delivered at the speed of business
- Data Quality can be compromised
- New innovative technologies are difficult to test, deploy and leverage

Dichotomies:

- Development & protect production
- Experimentation & reproducibility
- Central control & self service
- Group sharing & individual control
- Reuse & isolation

Currently, Teams Have High Errors

DataKitchen/Eckerson Survey (May 2019)



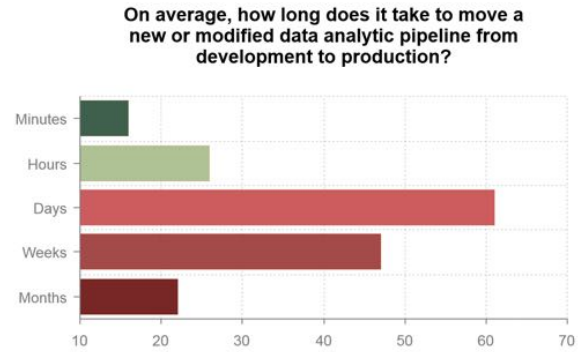
Forthcoming DataKitchen / Eckerson Research Survey of Medium – Large Companies US And Abroad

Currently, Teams Struggle to Deploy

DataKitchen/Eckerson Survey (May 2019)

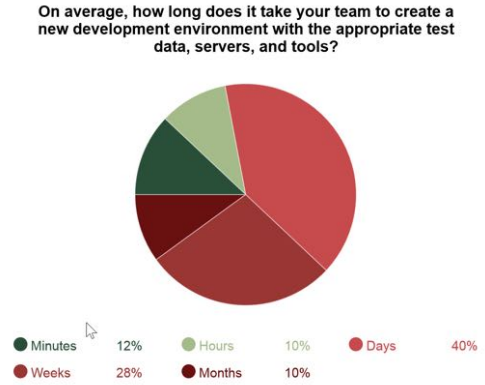
DataKitchen Interpretation

MOST COMPANIES ARE TOO SLOW TO DEPLOY CHANGES INTO PRODUCTION



DataKitchen Interpretation

MOST COMPANIES ARE VERY SLOW CREATING NEW DEVELOPMENT ENVIRONMENTS



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

Hidden Technical Debt in Machine Learning Systems

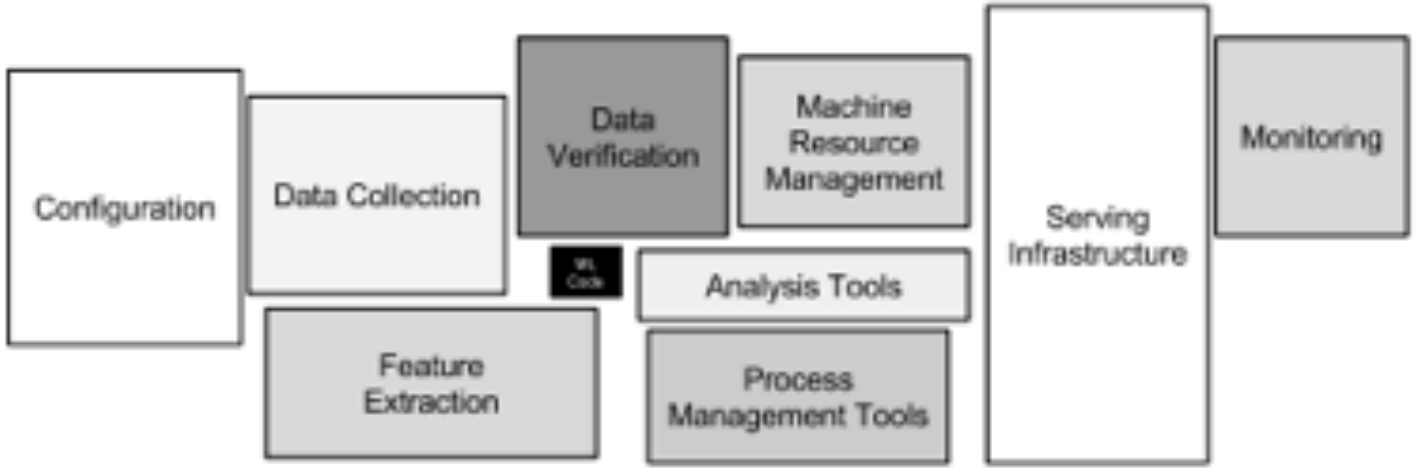
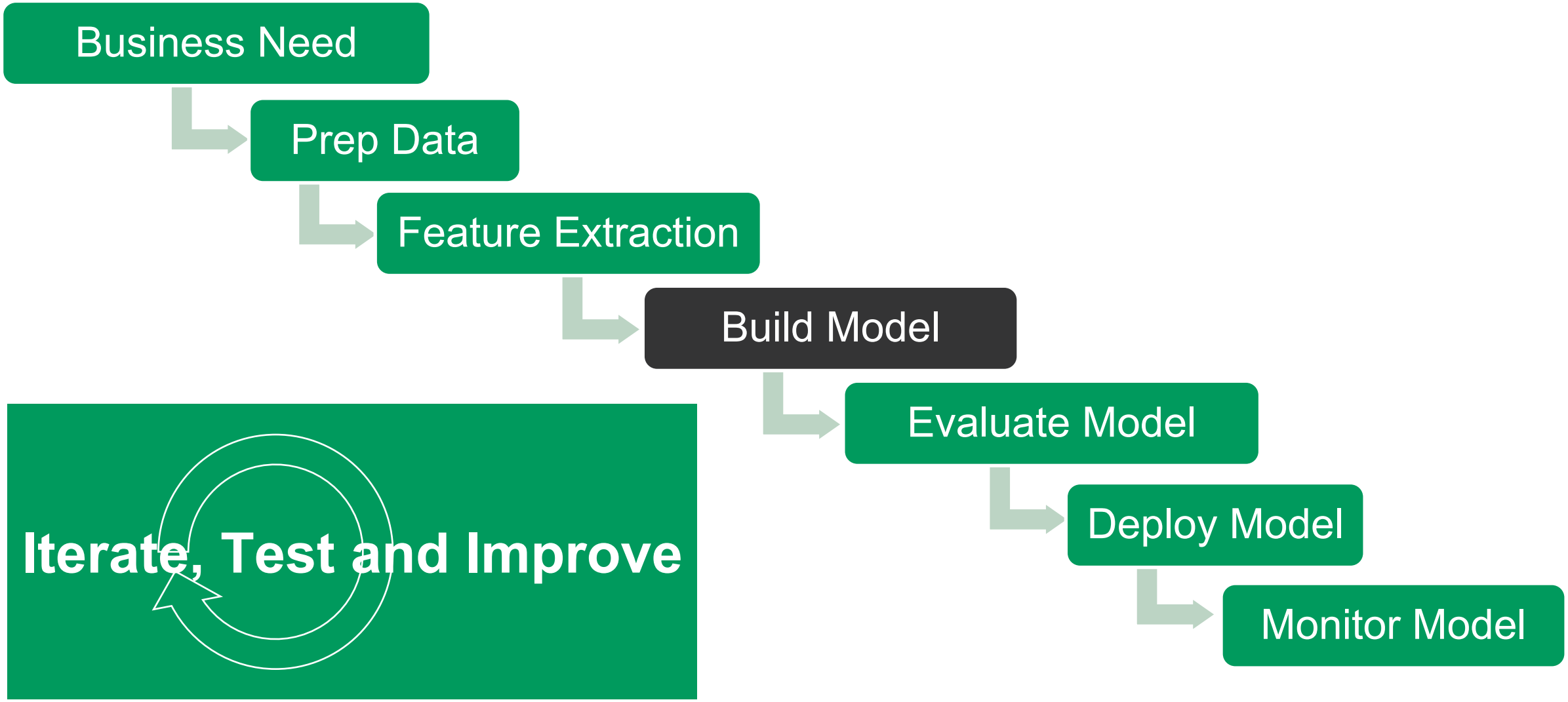


Figure 1: Only a small fraction of real-world ML systems is composed of the **ML code**, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Google
Advances in Neural Information Processing Systems 28 (NIPS 2015)

Model building



How To Succeed?

A Mindset Change to DataOps…

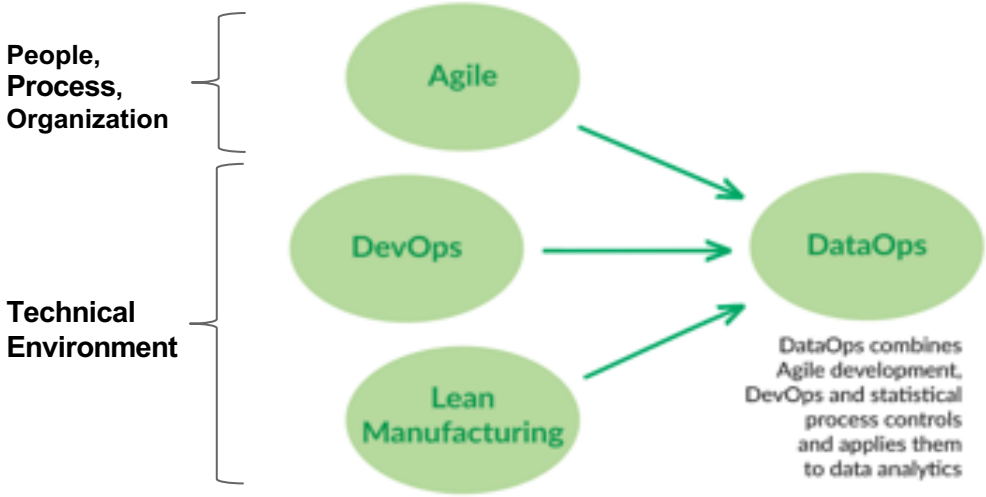
From	To
Change Fear	**Change Velocity**
Manual Operations	**Automated Operations**
Hope For Quality	**Integrated Quality**
Hero Mentality	**Repeatable Processes**
Tool Centric	**Code Centric**
Vendor Lock-In	**Diverse Tools**

…to power your highly agile data culture.

DataOps – Transformative to Data Analytics

DataOps – Continuous Delivery of Analytics

- Delivery insights faster
- Ensure high quality
- Add features at the speed of business
- Automate, orchestrate complex environment of people and technology



"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 2018

Source: Gartner

Keep this question in mind

What can I take from this session and apply tomorrow?

Topics

Why DataOps Is Essential

Agile in a Nutshell

Seven Steps to DataOps

Bonus Steps to DataOps

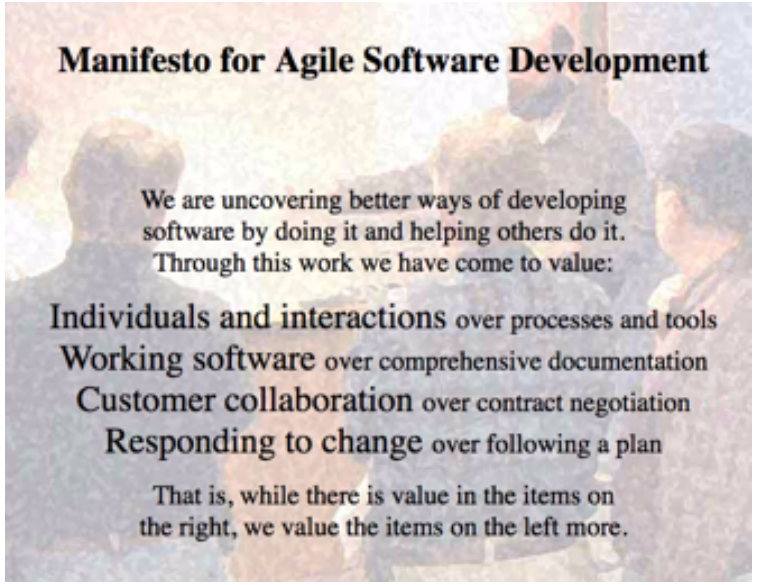
Next Steps With DataOps

Agile development is not a method, it is a *mindset*

1. Release frequently (“get it in the bank”, most important first)
2. Get feedback on your releases
3. Adjust your behavior to become more effective

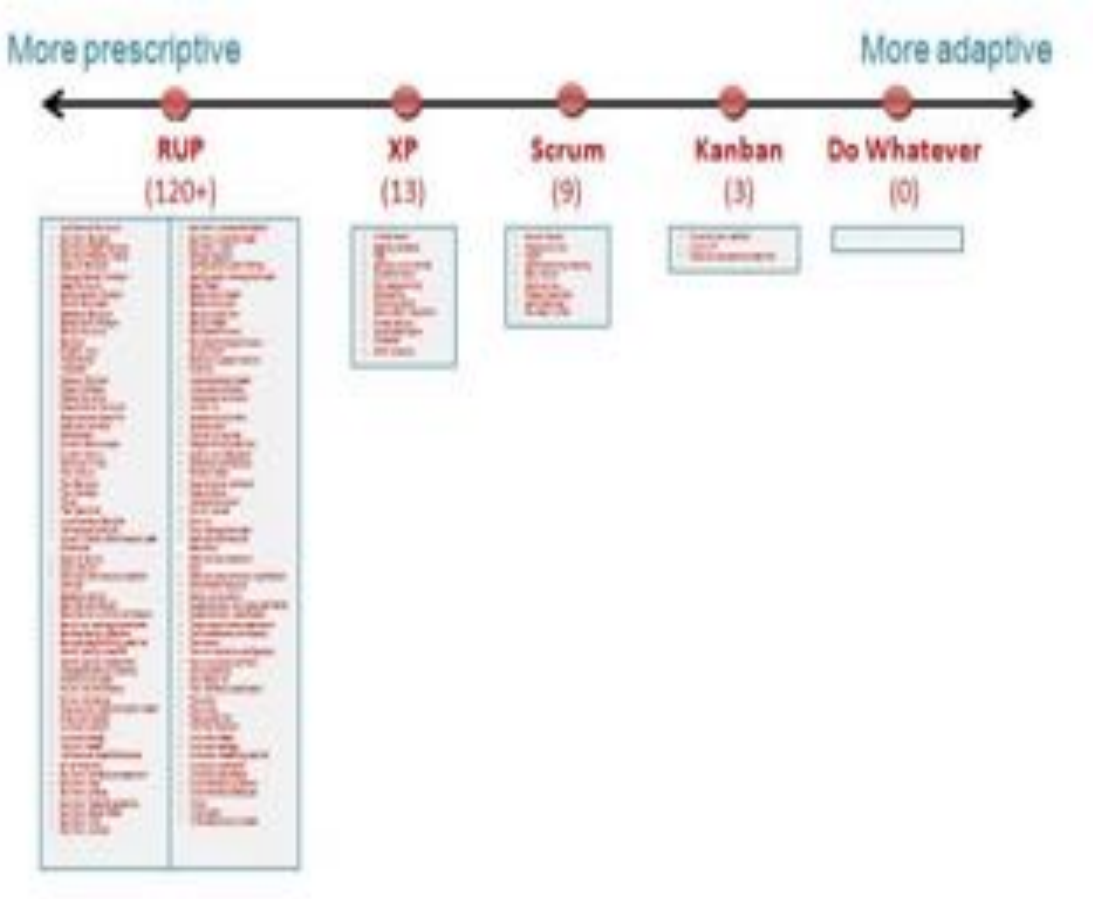
It is a focus on value

4 Values
12 principles



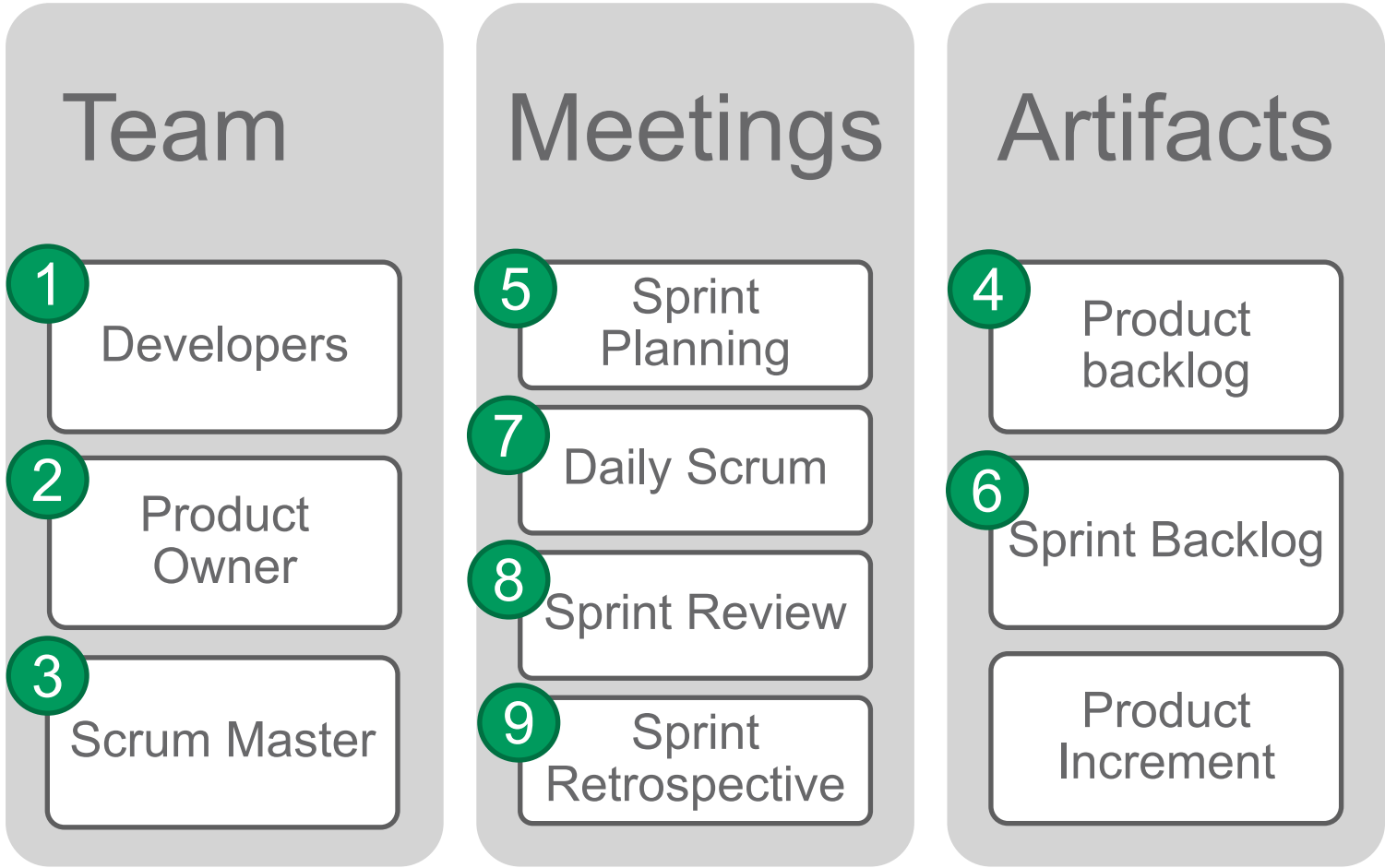
Be Pragmatic not Dogmatic

There are many agile frameworks available



http://www.crisp.se/

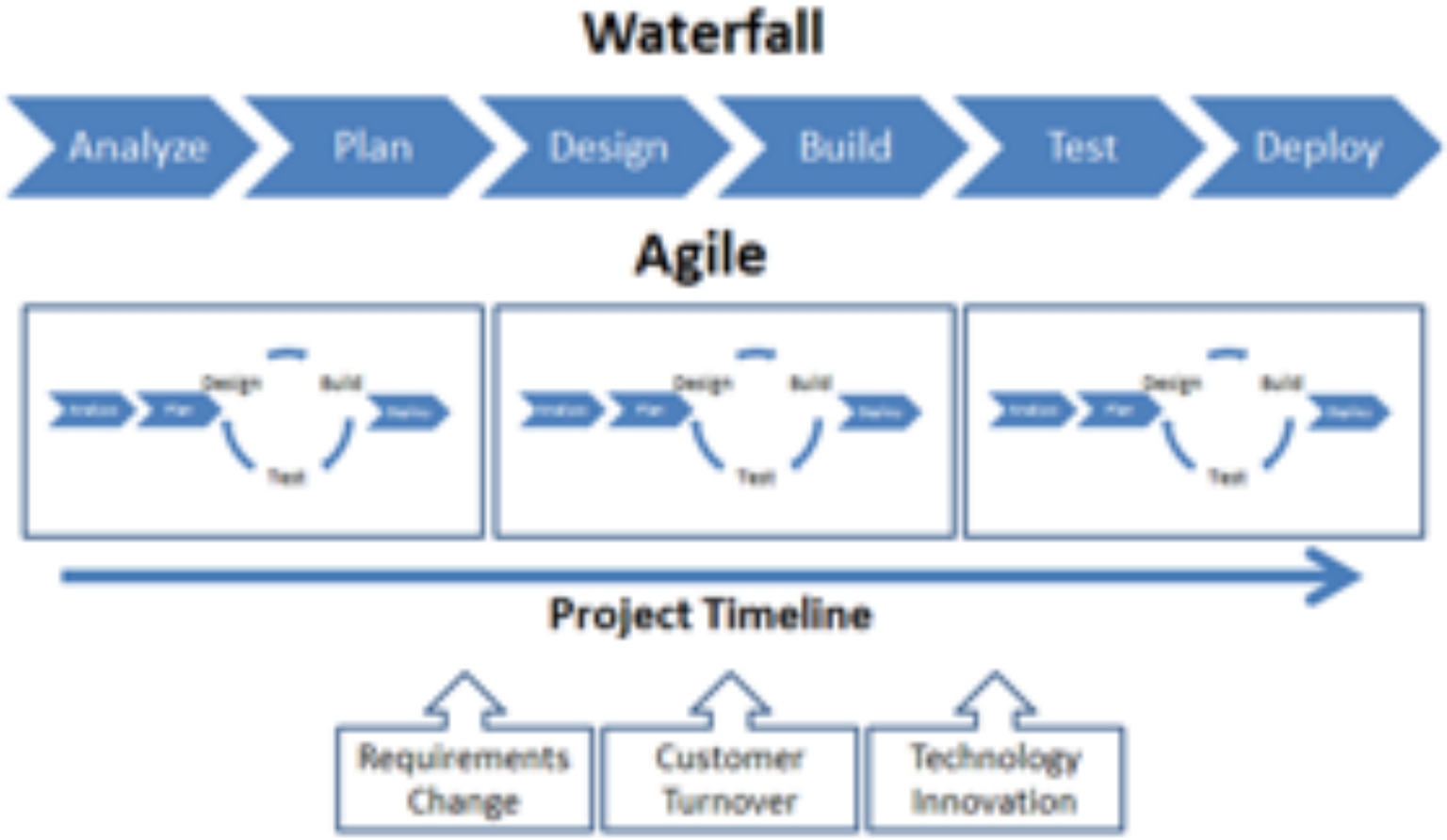
Get started with Scrum in 9 easy steps



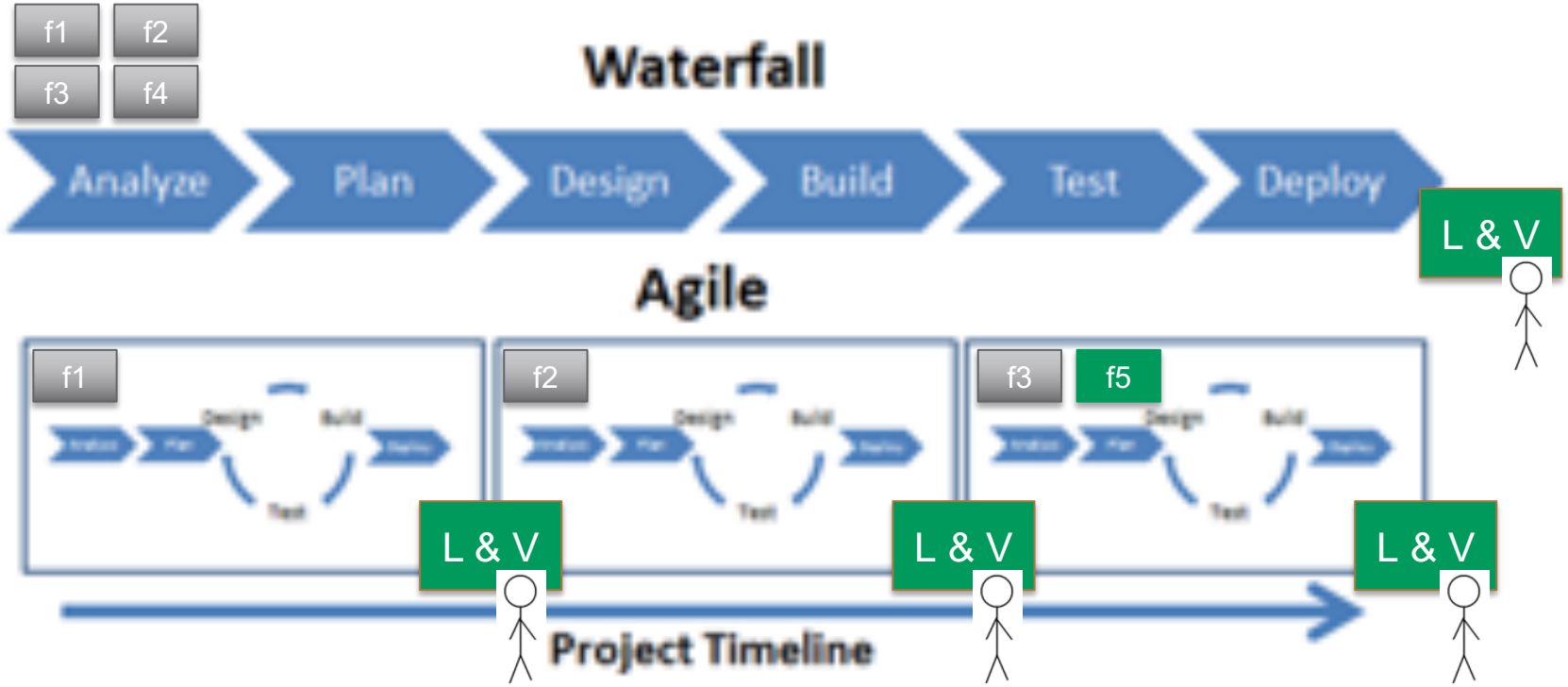
https://www.scrumguides.org/

What do frequent deliveries look like?

> **03** Deliver working software frequently, from a couple of weeks to a couple of months, with a preference to the shorter timescale.



What are some benefits? Learning & Value (L&V)



Stories

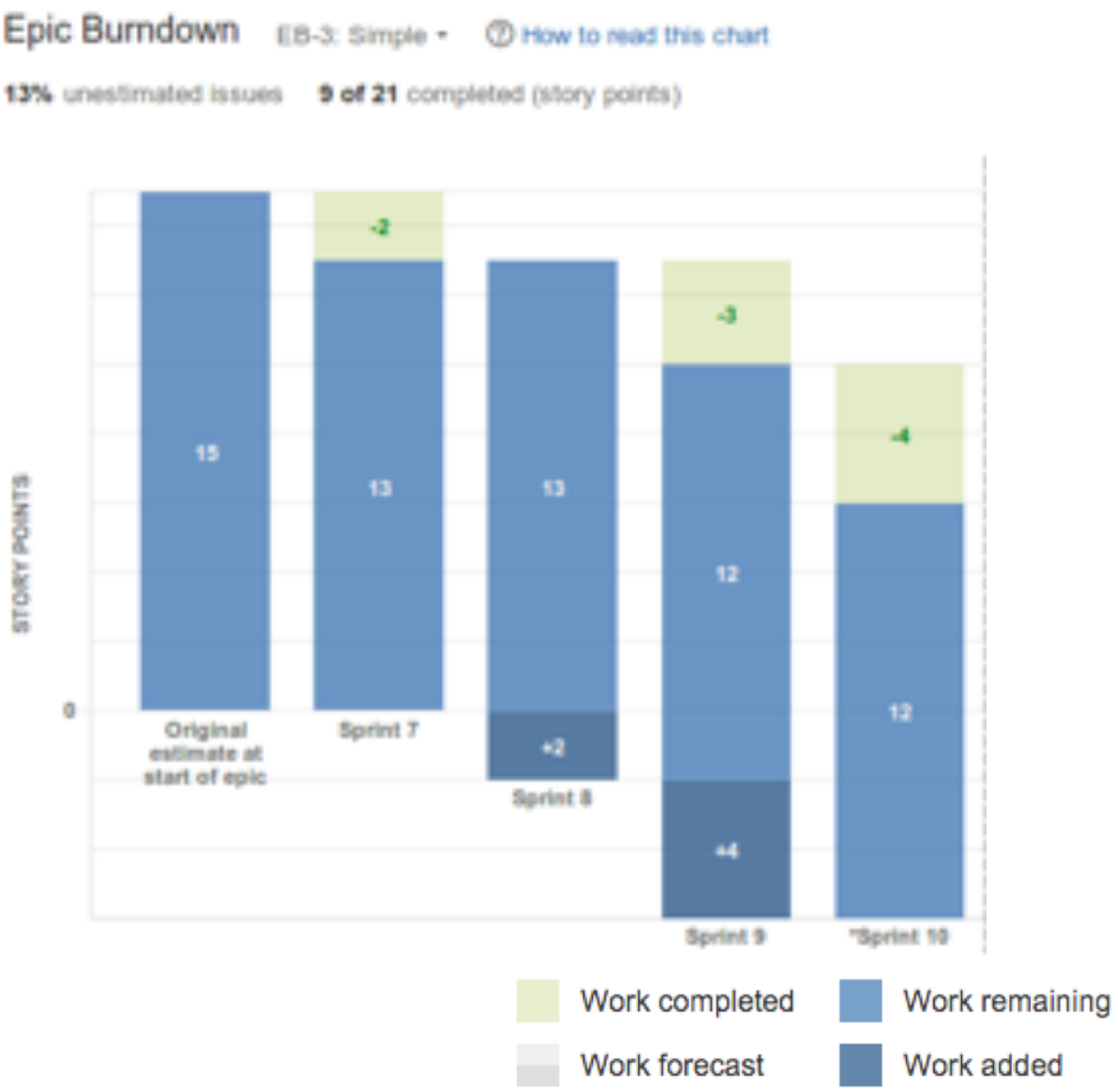
As a <role>, I want <result>, so that <context/benefit>

As a sales person, I want to see a list of doctors ranked by Lipitor sales, so that I can visit the top prescribers and make sure they are happy.

- No *How*, just *What*
- Actionable / Implementable / Complete
- Small – **fits in a sprint**
- Testable

Epics

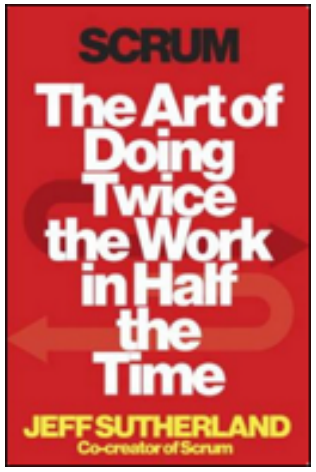
- An *epic* captures a large body of work.
- It is essentially a large user story that can be broken down into a number of smaller stories.
- It may take several sprints to complete an epic.



WIP = Work In Progress

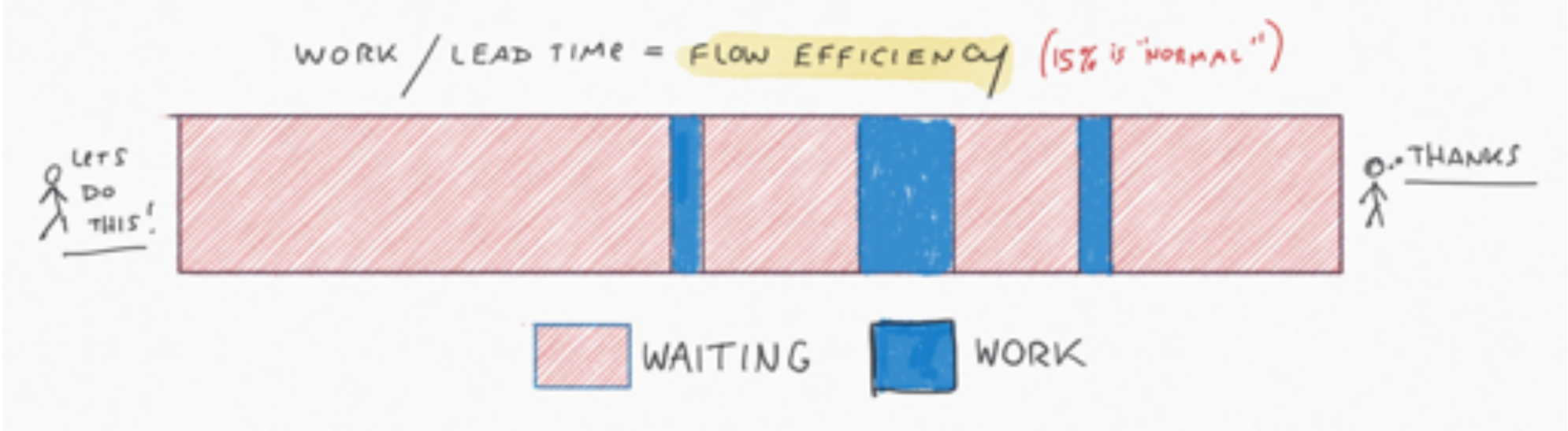
- Kanban practice to limit the amount of **Work In Progress**
- Benefits
  - Less context switching (more velocity)
  - Focus on getting stories done (more value)
  - Less chance to abandon a task (less waste)
  - Make blockers and bottlenecks visible (fewer delays)
- Pick a number for each person or the whole team
- Know that you will get done and move to the next task

Do one thing at a time



Similar to the Kanban idea of limiting work in progress (WIP)

Look at wait time in your system



15% is normal
45% is best in class

https://hackernoon.com/why-isnt-agile-working-d7127af1c552

Agile applies to diverse situations

DESTINATION IMAGINATION

Exercise

Which idea from agile would be most helpful in your organization?

Topics

Why DataOps Is Essential

Agile in a Nutshell

Seven Steps to DataOps

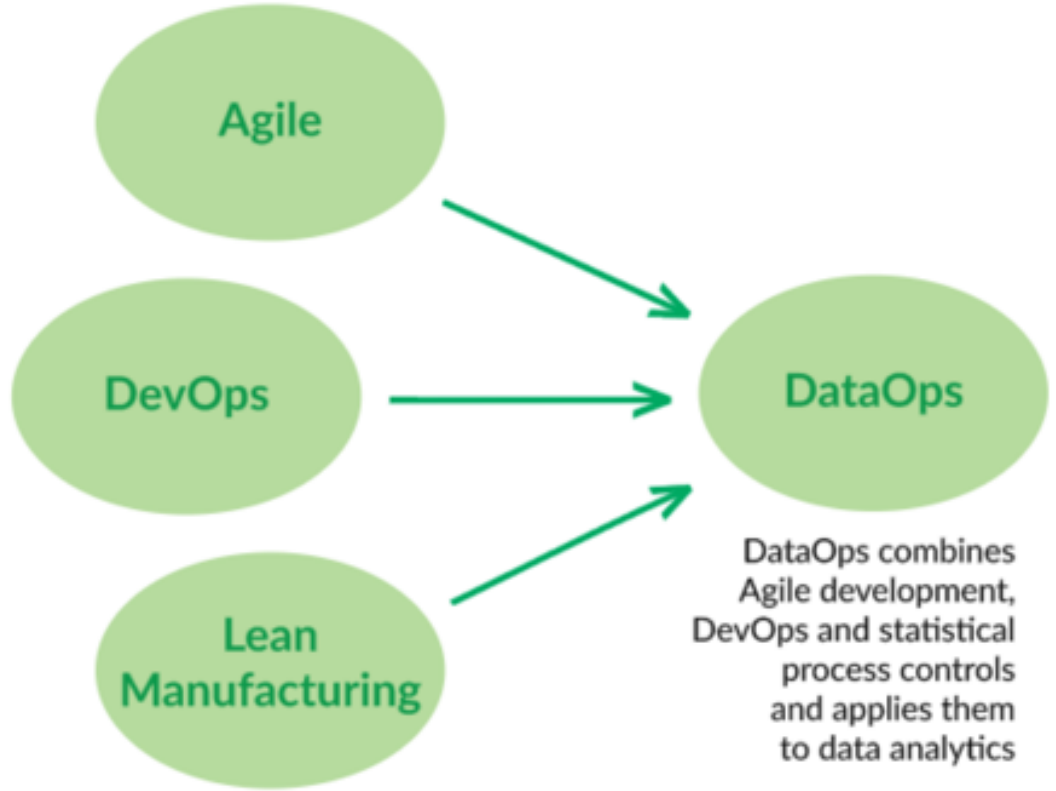
Bonus Steps to DataOps

Next Steps With DataOps

Seven Steps to DataOps

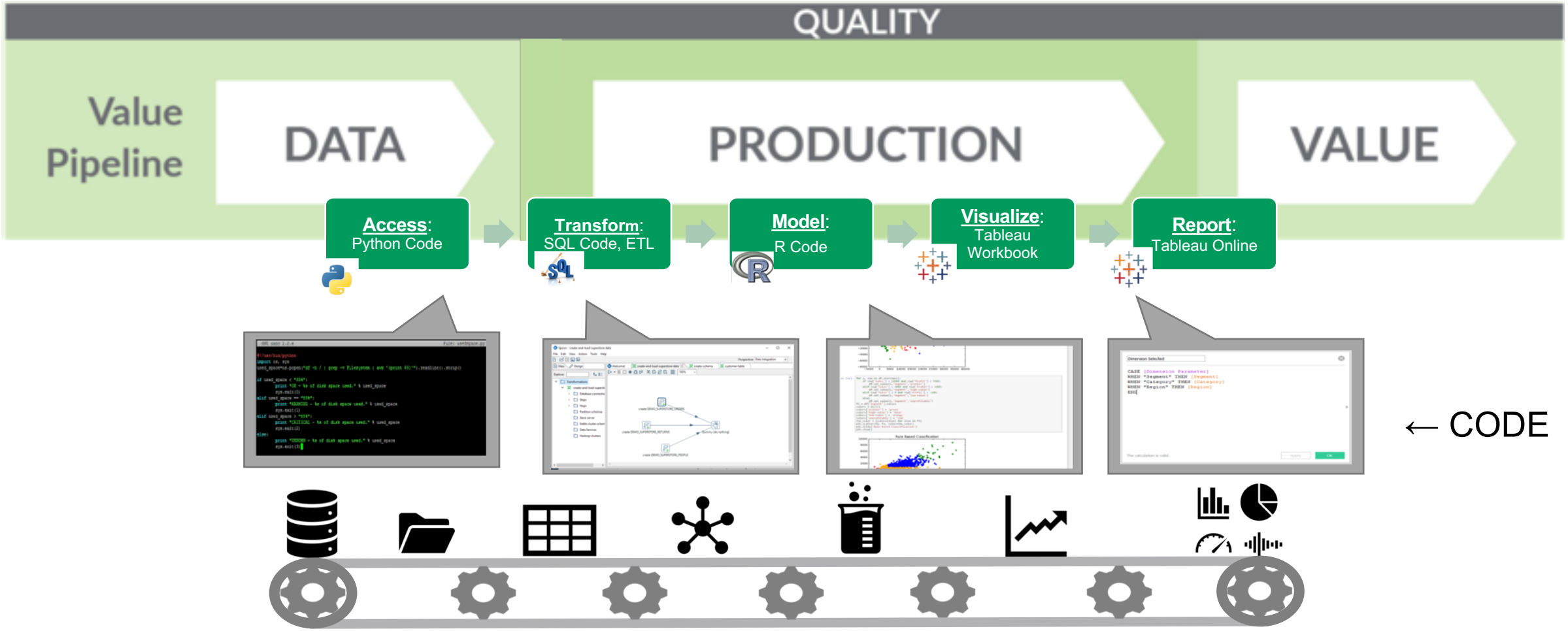
1. Orchestrate Two Journeys
2. Add Tests And Monitoring
3. Use a Version Control System
4. Branch and Merge
5. Use Multiple Environments
6. Reuse & Containerize
7. Parameterize Your Processing

Technical Environment



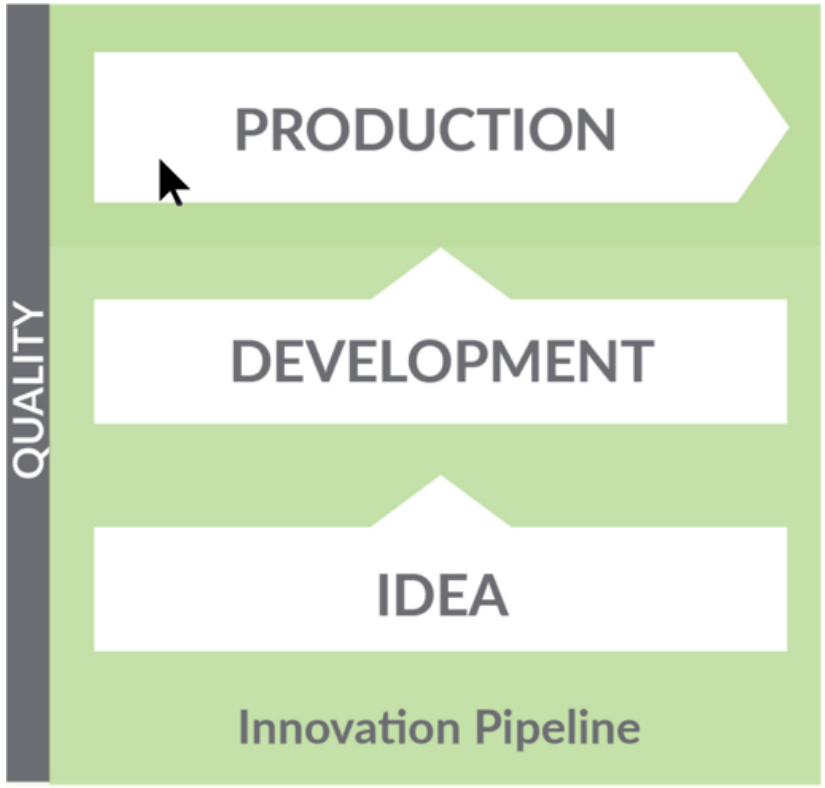
1 Orchestrate data to customer value

Analytic process are like manufacturing: materials (data) and production outputs (refined data, charts, graphs, model)



1 Speed deployment to production

Analytic processes are like software development: deliverables continually move from development to production



Diverse Customers

Business Customer

Products & Systems

Diverse Team

Data Engineers

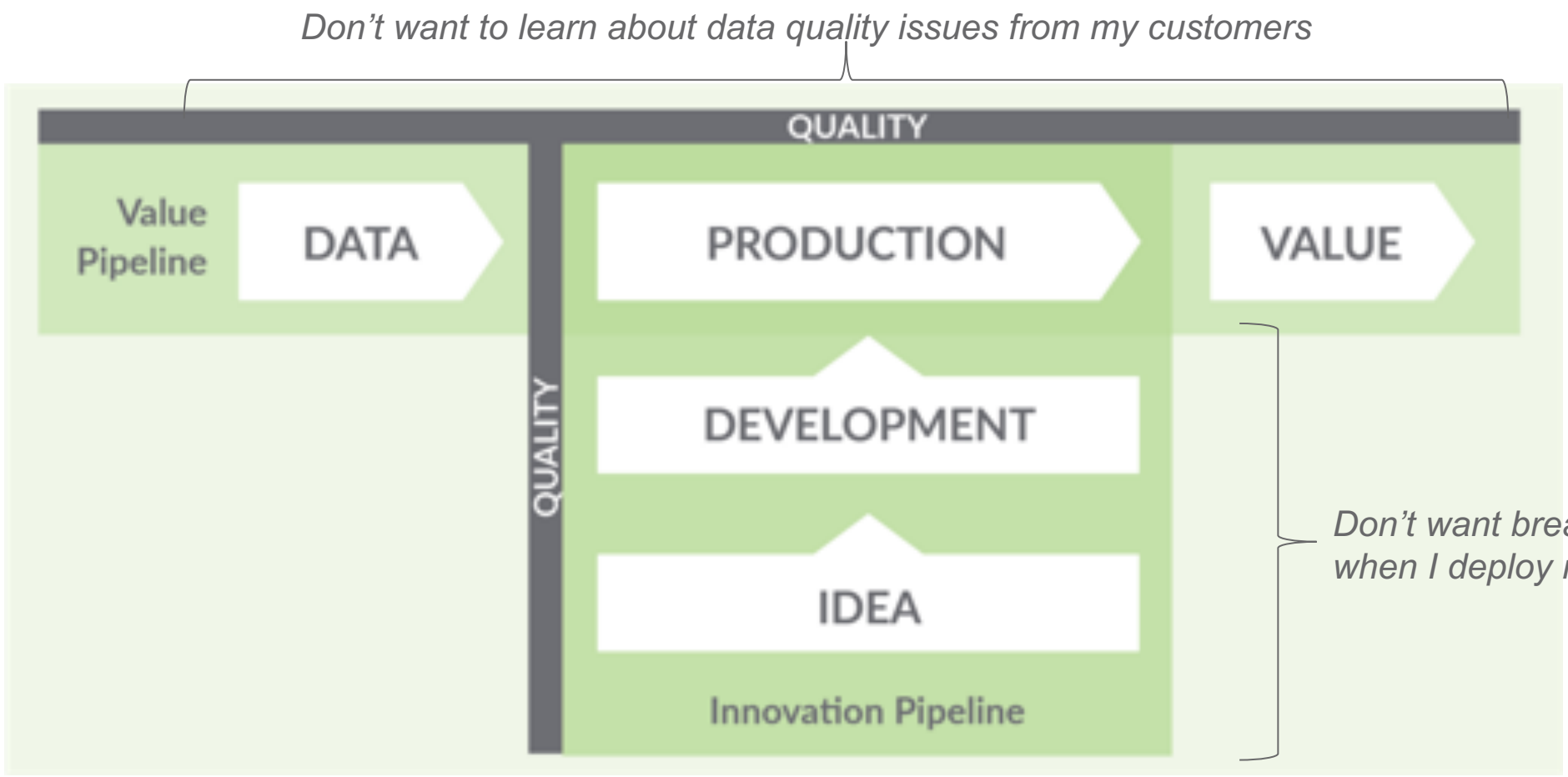
Data Scientists

Data Analysts

Diverse Tools

1 Innovation and Value Pipeline Together

Focus on both orchestration and deployment while automating & monitoring quality

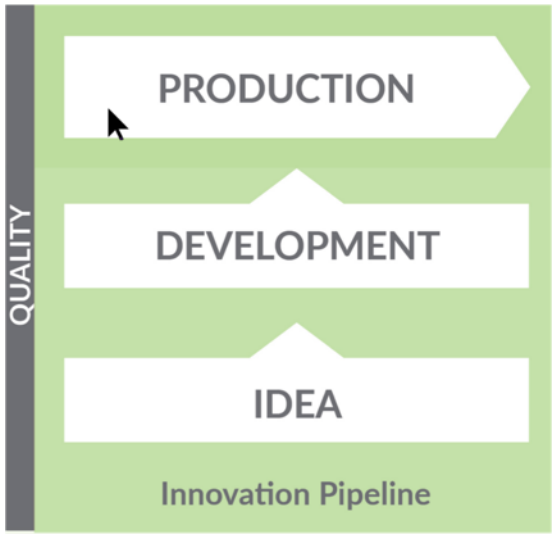


2 Add Automated Monitoring And Tests

Move Fast and Count Things (Facebook Analytics)

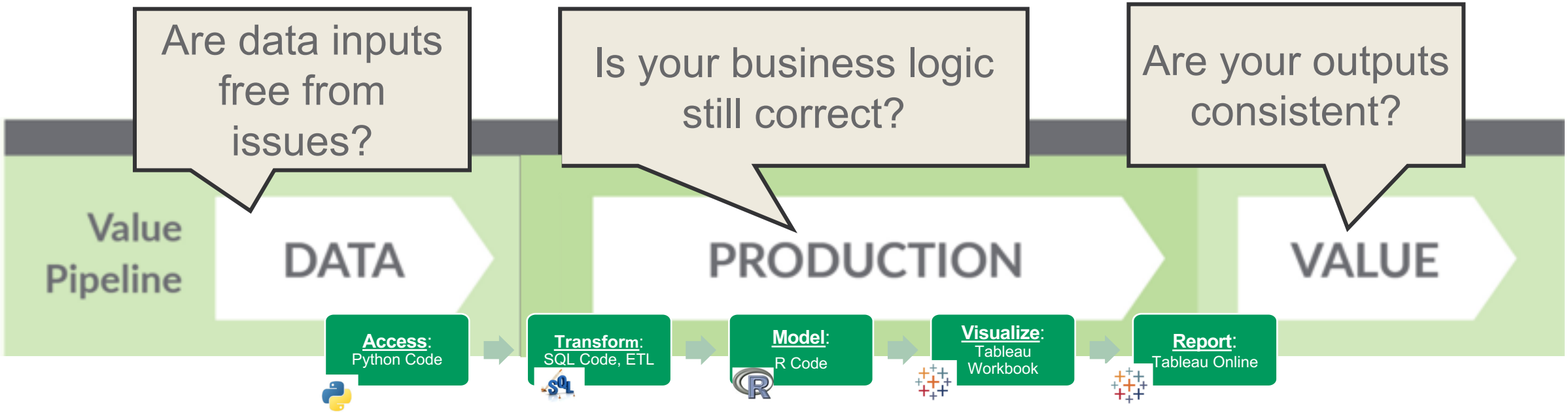
Monitoring: To ensure that during in the **Value Pipeline**, the data quality remains high.

Tests: Before promoting work, running new and old tests gives high confidence that the change did not break anything in the **Innovation Pipeline**



2 Automate Monitoring & Tests In Production

Test Every Step And Every Tool in Your Value Pipeline



And Save Test Results!

2 Support Multiple Types Of Tests

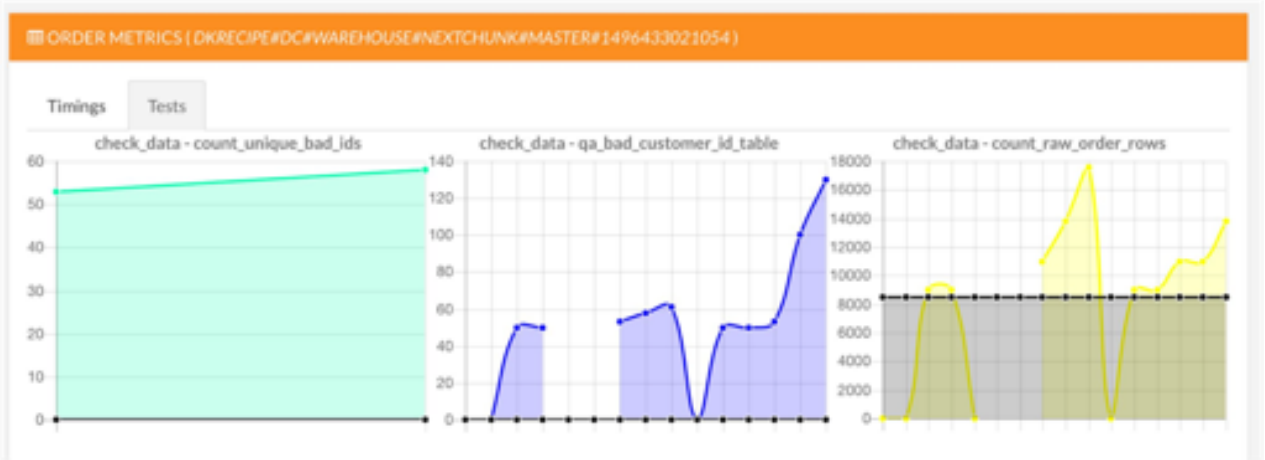
Testing Data Is Not Just Pass/Fail in Your Value Pipeline

Support Test Types

- **Error** – stop the line
- **Warning** – investigate later
- **Info** – save values

Keep Test History

- Statistical Process Control



2 Example Tests (Basic)

Inputs	**Verifying the inputs to an analytics processing stage** Count Verification - Check that row counts are in the right range, ... Conformity - US Zip5 codes are five digits, US phone numbers are 10 digits, ... History - The number of prospects always increases, ... Balance - Week over week, sales should not vary by more than 10%, ... Temporal Consistency - Transaction dates are in the past, end dates are later than start dates, ... Application Consistency - Body temperature is within a range around 98.6F/37C, ... Field Validation - All required fields are present, correctly entered, ...
Business Logic	**Checking that the data matches business assumptions** Customer Validation - Each customer should exist in a dimension table Data Validation - 90 percent of data should match entries in a dimension table
Output	**Checking the result of an operation, for example, a cross-product join** Completeness - Number of customer prospects should increase with time Range Verification - Number of physicians in the US is less than 1.5 million

2 Example Tests Simple

DATA KITCHEN

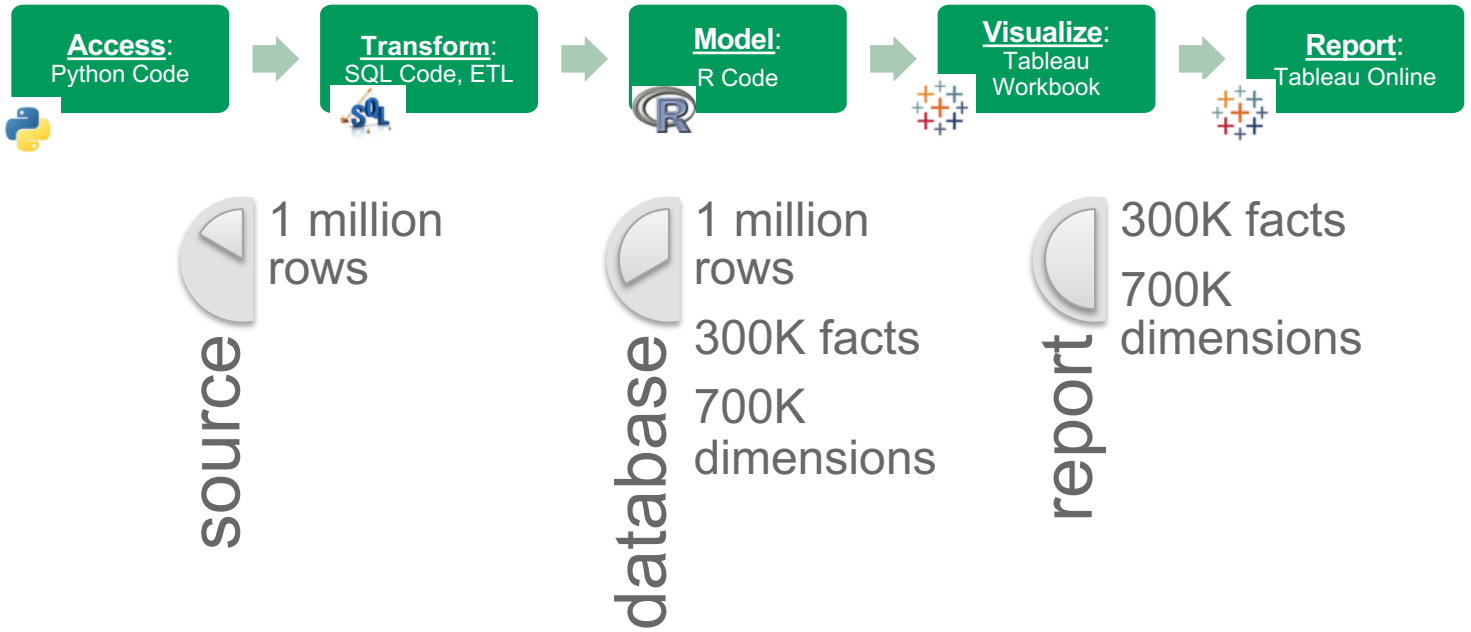
```
1  select count(*) from customers :: is above a threshold
2
3  select count(*) from customers :: is larger or equal to than last time
4
5  select count(*) from raw_specialty_pharmacy where len(zip) != 5 :: is zero
6
```

2 Example Test More Complex

Make sure all table counts are the same in the production and development environment

```
IF (EXISTS(SELECT *
           FROM {{sql_database}}.INFORMATION_SCHEMA.TABLES
           WHERE TABLE_NAME = 'qa_bad_customer_id'))
BEGIN
  select max(c) from  (
    select case when
    (SELECT count(*) from {{sql_database}}.dbo.chunker)
    =
    (SELECT count(*) FROM {{sql_database}}_dev.dbo.chunker)
    then
    0
    else
    1
    end as c

    union all

    select case when
    (SELECT count(*) FROM {{sql_database}}.dbo.chunk_num) =
    (SELECT count(*) FROM {{sql_database}}_dev.dbo.chunk_num)
    then
      0
    else
      1
    end as c

    union all

    select case when
    (SELECT count(*) FROM {{sql_database}}.dbo.raw_orders) =
    (SELECT count(*) FROM {{sql_database}}_dev.dbo.raw_orders)
    then
      0
    else
      1
    end as c

    union all

    select case when
    (SELECT count(*) FROM {{sql_database}}.dbo.qa_bad_customer_id) =
    (SELECT count(*) FROM {{sql_database}}_dev.dbo.qa_bad_customer_id)
    then
      0
    else
      1
    end as c
  ) as m
END
```

2 Example Test (Location Balance)



2 Example Test (Historical Balance)

Production Data, Pipeline & Environment

SKU	Product	Product Group	Volume
SKU1	P1	G1	100
SKU2	P1		50
SKU3	P2		75
SKU4	P3	G2	125
SKU5	P4		200
SKU6	P5		25
			575

Histbal	
G1	225
G2	350

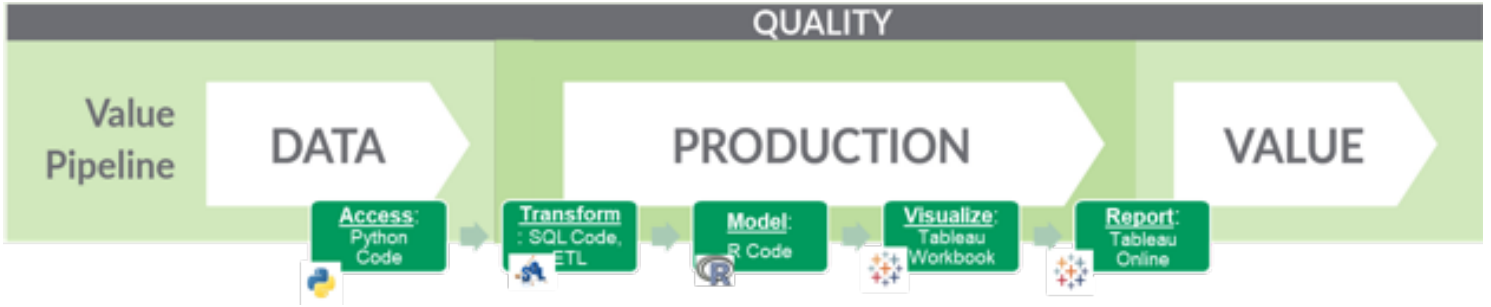
Pre-Production Data, Pipeline & Environment

SKU	Product	Product Group	Volume
SKU1	P1	G1	101
SKU2	P1		55
SKU3	P2		76
SKU4	P3		126
SKU5	P4	G2	200
SKU6	P5		29
			587

Histbal	
G1	358
G2	229

Production Testing and Monitoring

Lower Your Error Rates and Embarrassment!



[HOW] Test and Monitor:

- Automatically, in production
- Across the entire tool chain
- Send alerts / notification
- Keep track of history
- Make it easy to create tests

[WHAT] Test Types:

- Traditional Data Quality
- Statistical Process Control
- Location Balance Test
- Historic Balance Test
- Business Based Tests

2 Duality of Tests

DATA KITCHEN

Quality Your Customer Receives = $f(data, code)$

Automated 'Tests' Serve a Dual Purpose:

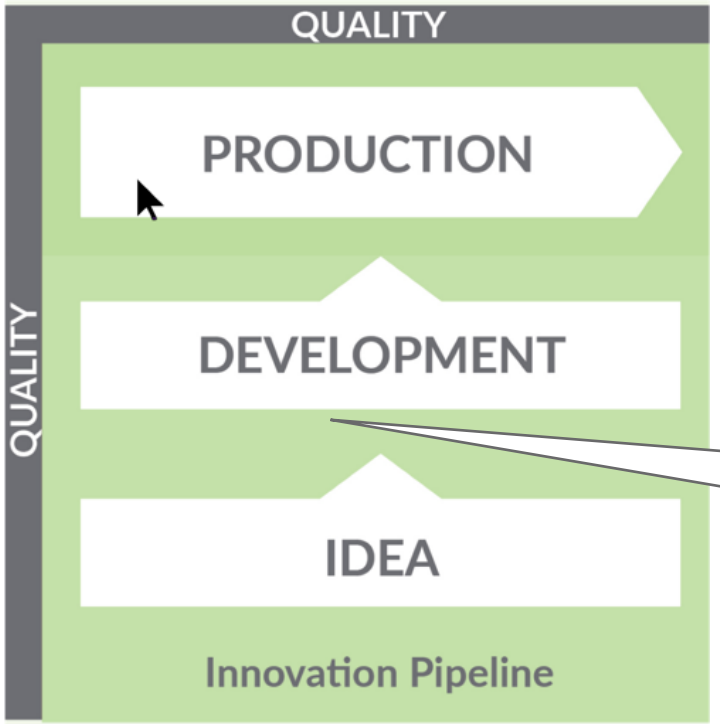
1. Data Tests and Monitoring in Production
2. Regression, Functional and Performance Tests in Development

	Data Fixed	Data Variable
Code Fixed		Value Pipeline
Code Variable	Innovation Pipeline	

https://medium.com/data-ops/disband-your-impact-review-board-automate-analytics-testing-42093d09fe11

2 For the Innovation Pipeline

Tests Are For Also Code: Keep Data Fixed

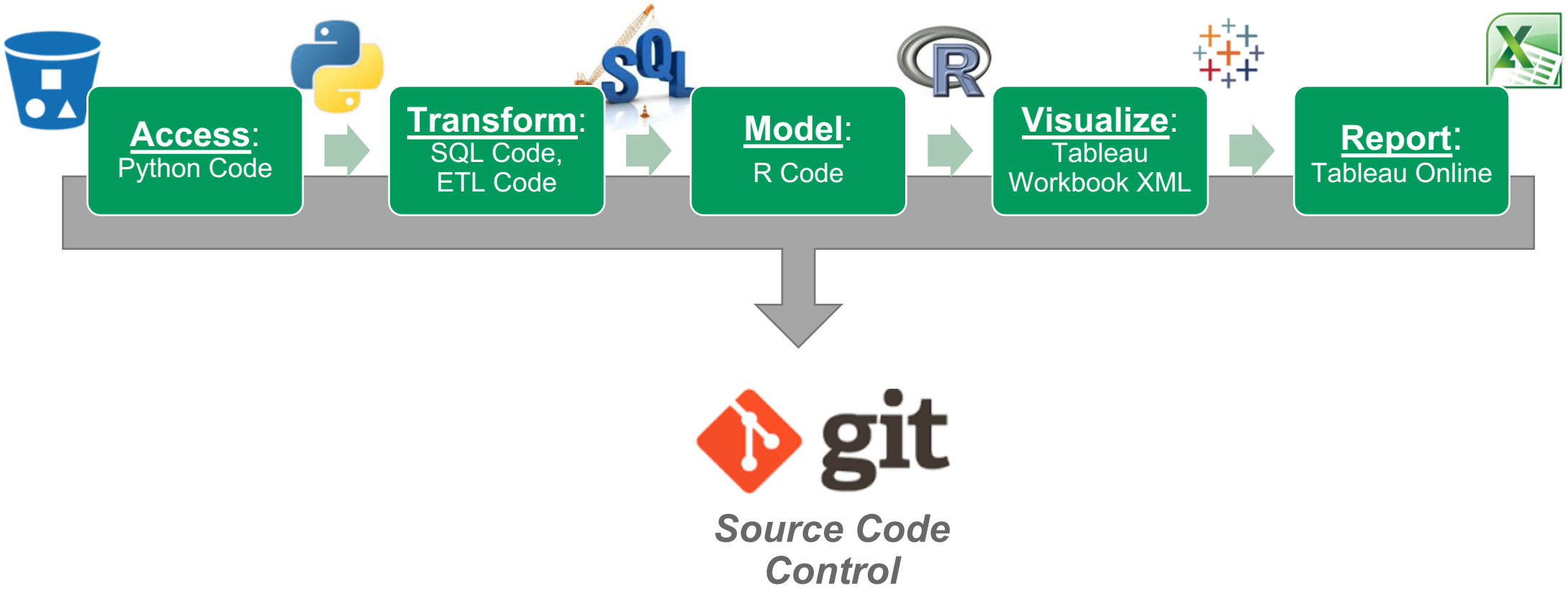


Run all tests here before promoting

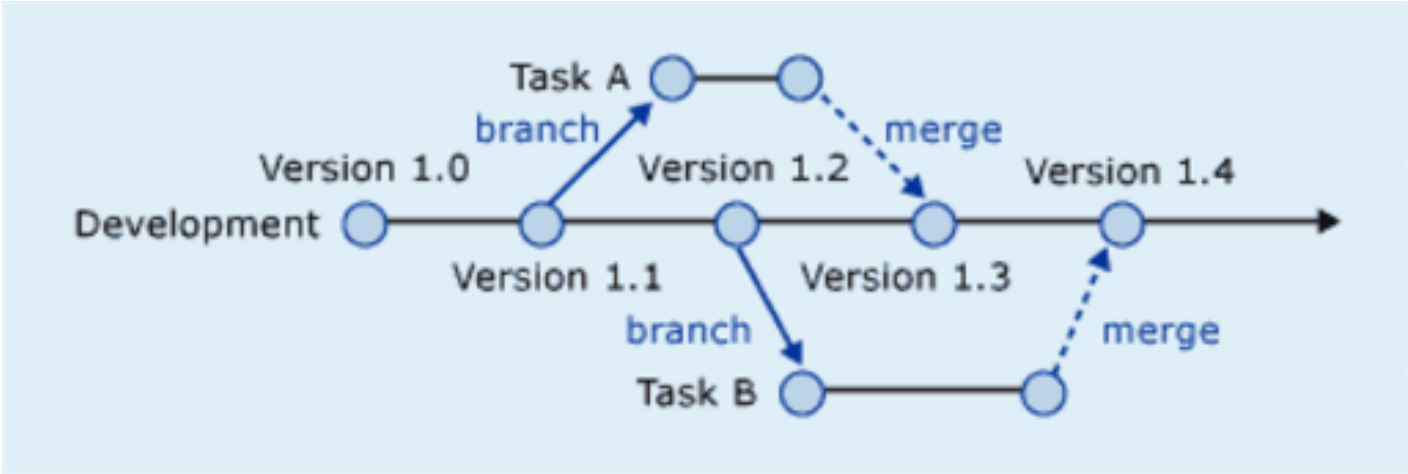
Deploy Feature

3 Use a Version Control System

At The End Of The Day, Analytic Work Is All Just Code



4 Branch & Merge

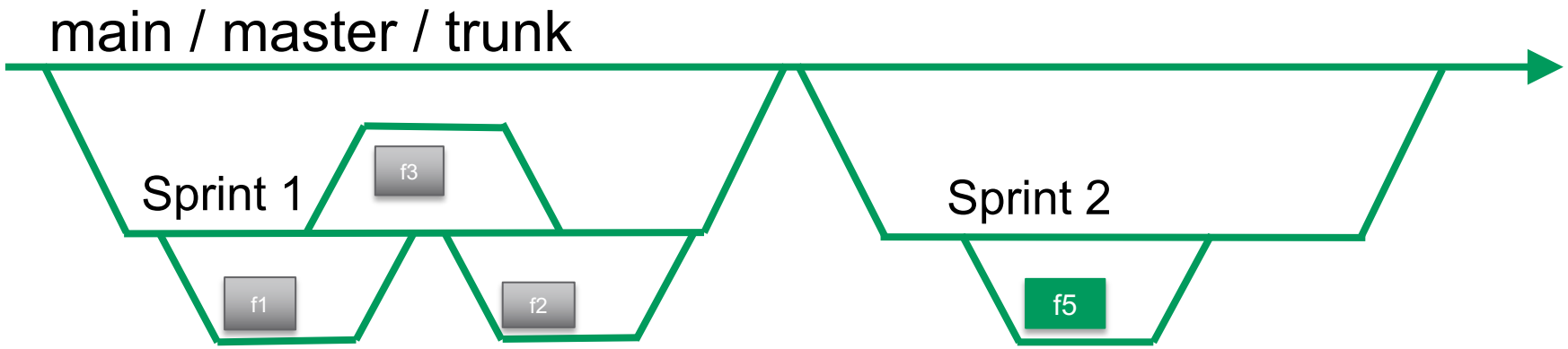


Branching & Merging enables people to safely work on their own tasks

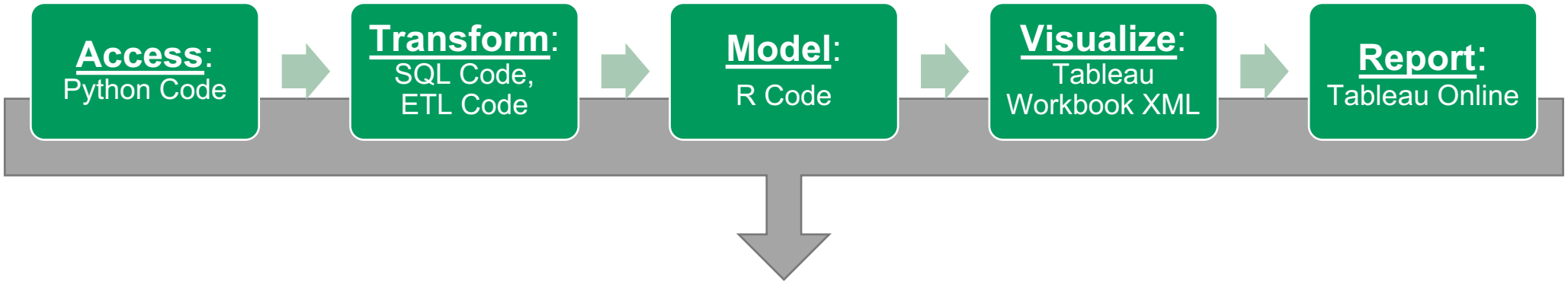
git

Source Code Control

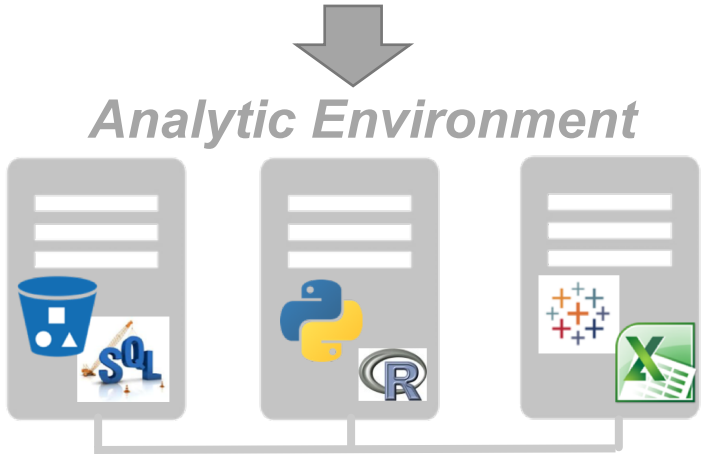
4 Example Branch And Merge Pattern



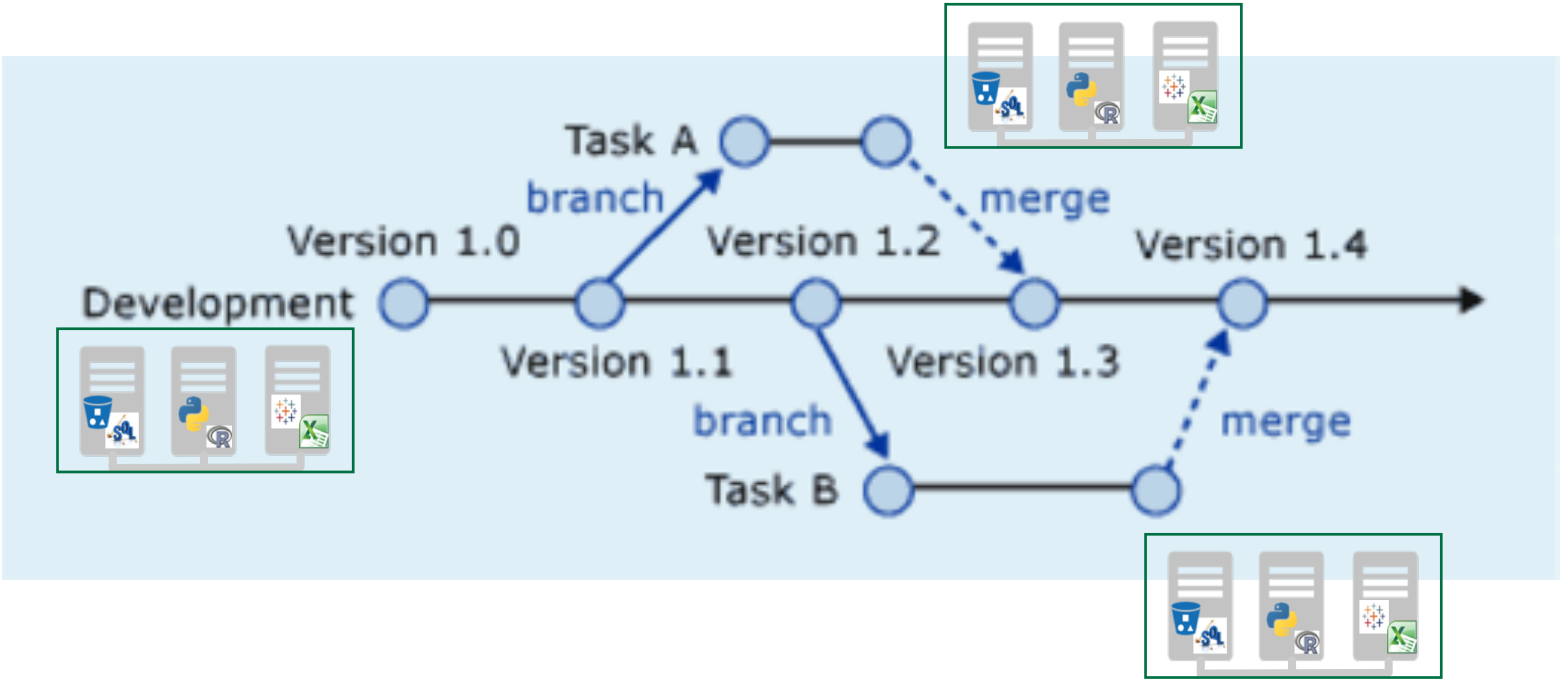
5 Use Multiple Environments



Your Analytic Work Requires Coordinating Tools And Hardware



5 Use Multiple Environments



Provide an Analytic Environment for each branch

- Analysts need a controlled environment for their experiments
- Engineers need a place to develop outside of production
- Update Production only after all tests are run!

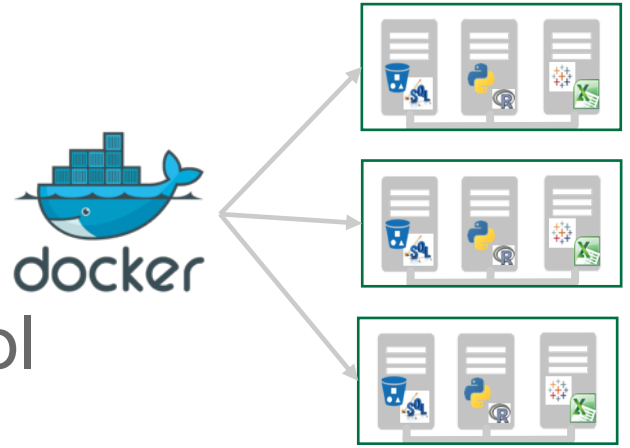
6 Reuse & Containerize

Reuse

1. Do not create one ‘monolith’ of code
2. Reuse the code and results

Containerize

1. Manage the environment for each component (e.g. Docker, AMI)
2. Practice Environment Version Control



7 Parameterize Your Processing

Think Of Your Value Pipeline Like A Big Function

- Named sets of parameters will increase your velocity
- With parameters, you can vary”
  - Inputs
  - Outputs
  - Steps in the workflow
- You can make a time machine
- Secure storage for credentials

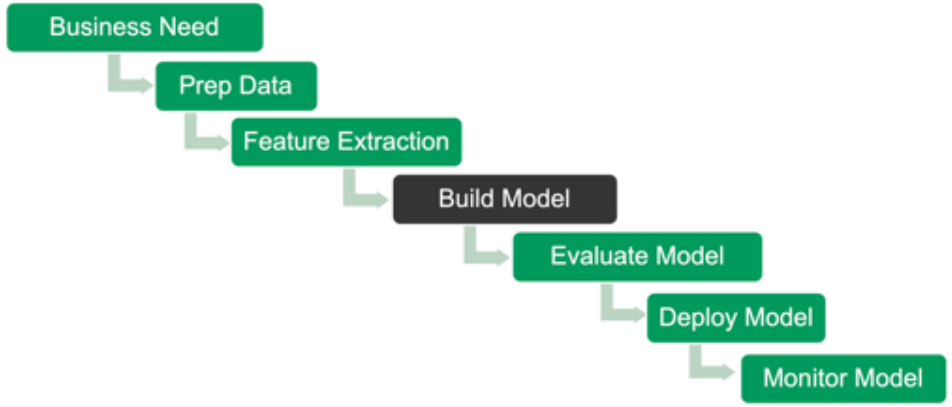
The Seven Steps In Action

1. Select story
2. Create branch
3. Create environment
4. Implement feature
5. Write new tests
6. Run new and existing tests
7. Check in to branch
8. Merge to parent
9. Delete environment

When sprint ends

- Deliver all completed features to customer
- Merge sprint branch to master
- Roll un-merged features into the next sprint

The 7 Steps and Data Science



		Journeys	Tests	Version Control	Branch and Merge	Environments	Reuse / Containerize	Parameterize
Business Need	Agile							
Prep Data		X	X	X	X	X	X	X
Feature Extraction		X	X	X	X	X	X	X
Build Model		X	X	X	X	X	X	X
Evaluate Model			X					
Deploy Model		X	X	X	X	X	X	X
Monitor Model			X					

Exercise

Which of the seven steps would give you the most benefit?

Topics

Why DataOps Is Essential

Agile in a Nutshell

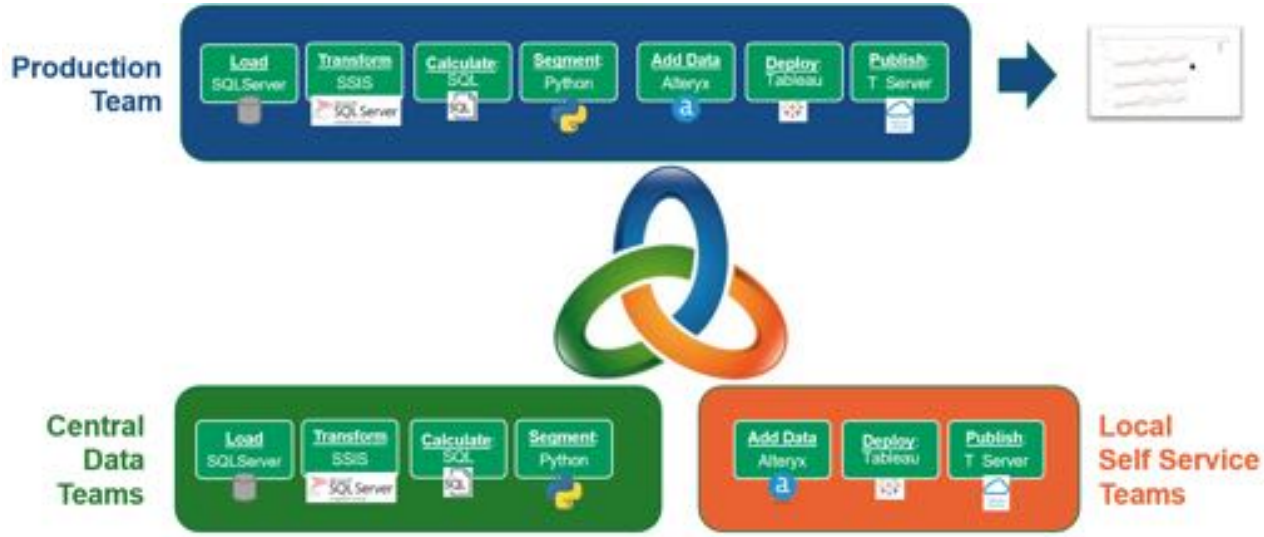
Seven Steps to DataOps

Bonus Steps to DataOps

Next Steps With DataOps

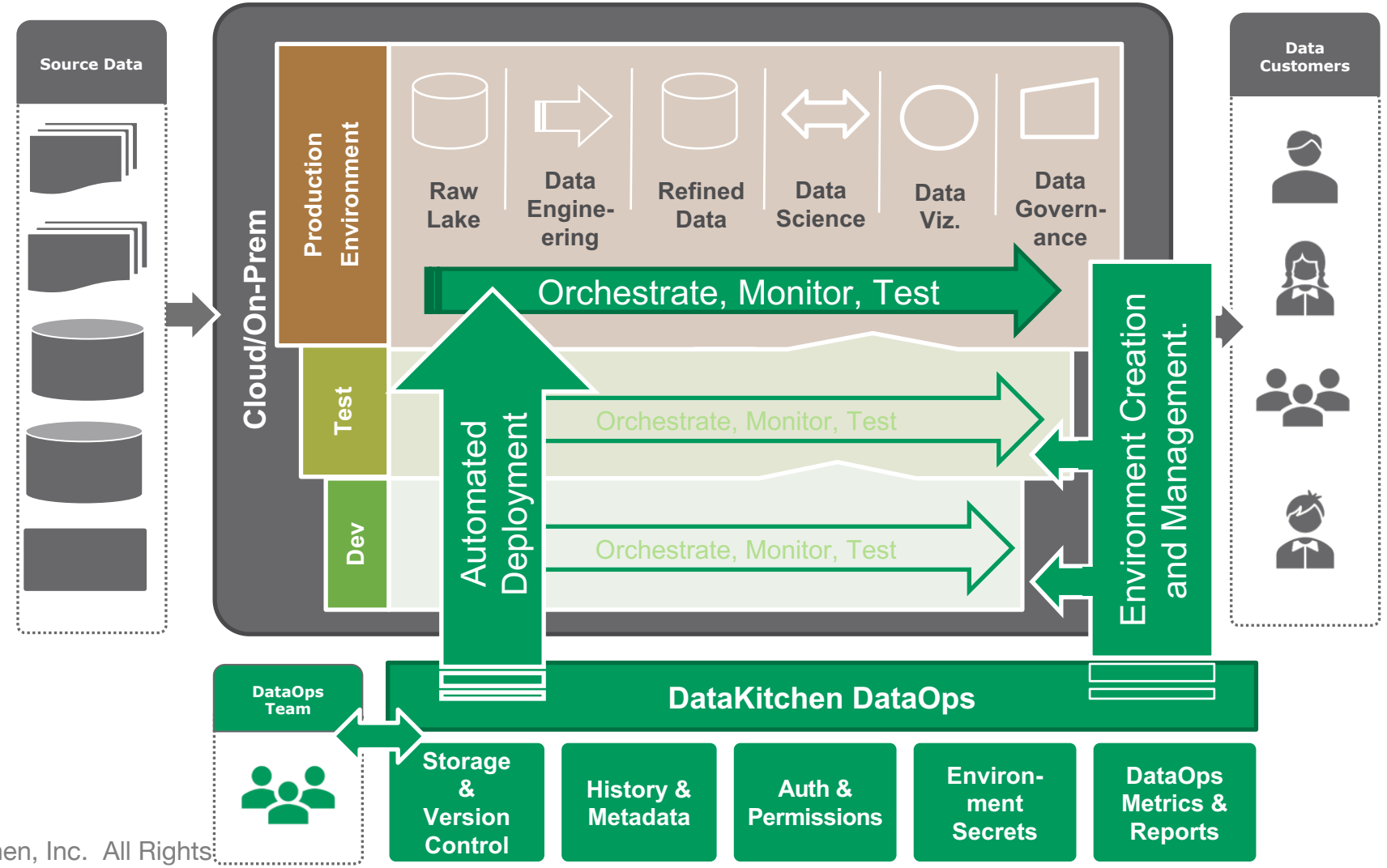
Four Bonus Steps!

DataOps Collaboration
   1. Intra Team
   2. Inter Team
3. DataOps Process Analytics
4. Data Categorization

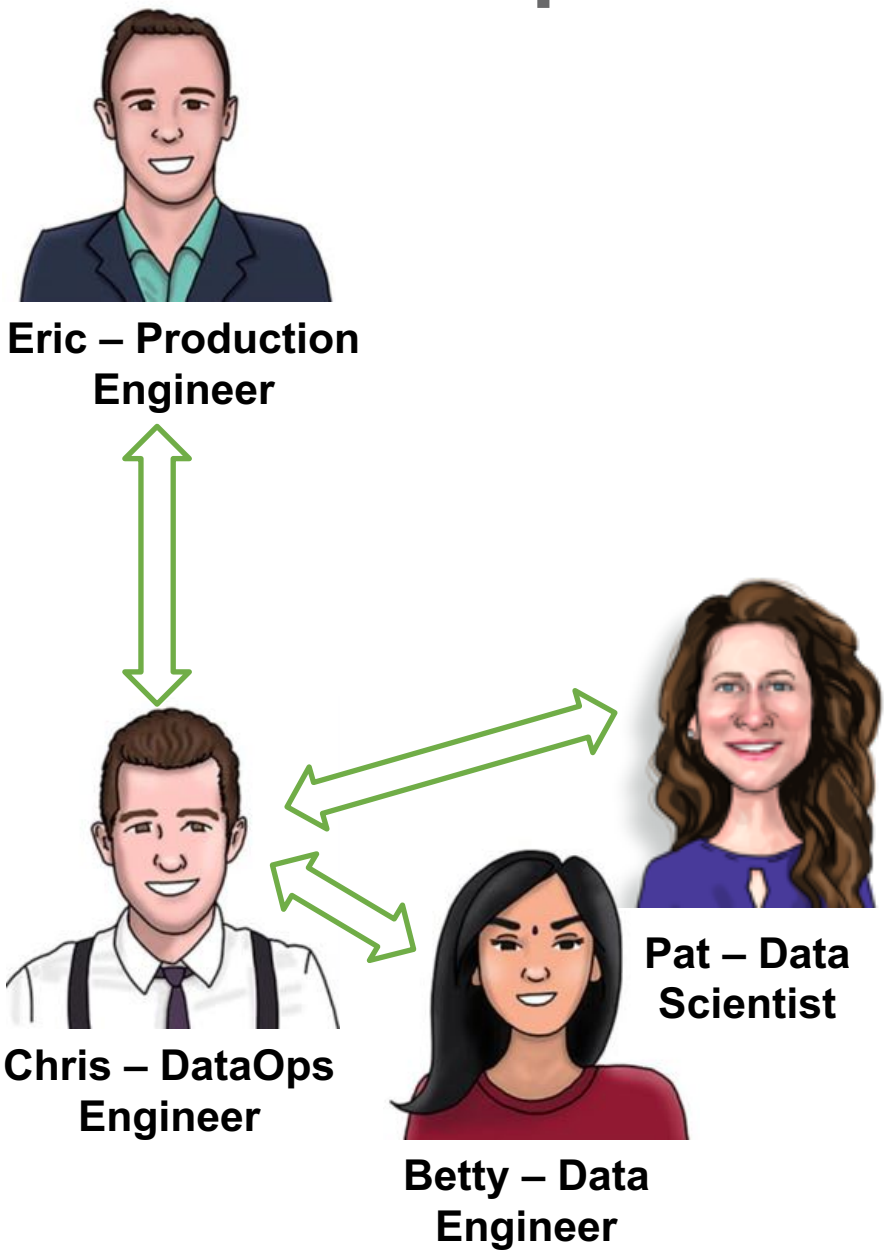


“Home Office” Intra-team Coordination

Spans tool chain & environments



DataOps and Intra-Team Coordination



This Is A Multi Step, Multi Person, Multi Environment Process To Make this Request a Reality

Challenges:

- How to leverage best practices and re-use?
- How to collaborate and coordinate work?
- How to ease movement between team members with many tools and environments?
- How to maintain security?
- How to automate work and reduce manual errors?

PRODUCTION

Intra-Team Coordination

Eric – Production Engineer

Production Environment:

- Separate Hardware/Software Environment
- Secure
- No Access By Developers
- Managed by Eric
- Separate Credentials

DEVELOPMENT

Chris – DataOps Engineer

Pat – Data Scientist

Betty – Data Engineer

Development Environment:

- Separate Hardware/Software Environment
- Secure
- Access By Data Engineers, Data Scientists, Analysts and DataOps Engineers
- Setup by Chris (DataOps Engineer)
- Separate Credentials

Inter-Team Coordination

Also spans locations

- **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)

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

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

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

VP Marketing

Inter-Team Coordination: Two Locations, Multiple Tools

DATA KITCHEN

Home Office Team

SQL

Microsoft SQL Server Integration Services

Data Engineer

Data Scientist

Boston

Centralized, Weekly Cadence of Changes

Local Office 'Self-Service' Team

ALTERYX

Data Analyst

New Jersey

Distributed, Daily/Hourly Cadence of Changes

VP Marketing

Challenges With Coordination

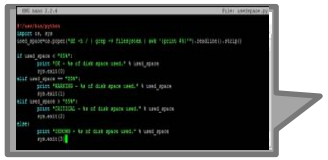
Home Office Team

Make a change in schema?

Not Available For All?

New Data & Schema

Data Engineer



Data Scientist

Local Office 'Self-Service' Team

Break Reports?

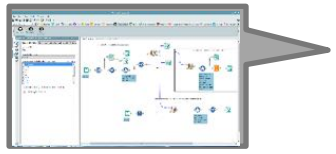
Add New Data Sets

Change Report Calculations

Inconsistencies?

Update/New Report

Not Working?

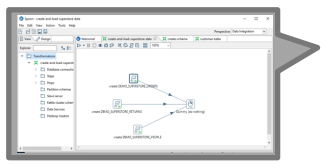


Data Analyst

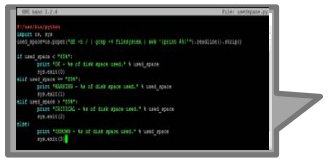
VP Marketing

Shared Result, Separate Responsibilities

Home Office Team



Data Engineer



Data Scientist

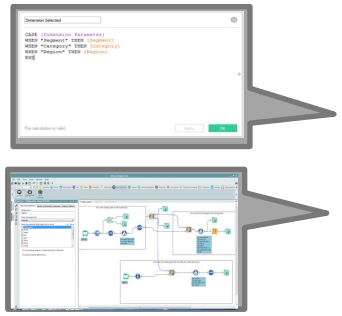
Load SQLServer

Transform SSIS

Calculate: SQL

Segment: Python

Local Office 'Self-Service' Team



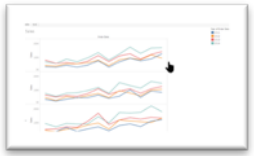
Data Analyst

Add Data Alteryx

Deploy: Tableau

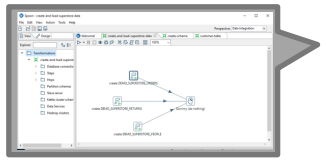
Publish: T Server

VP Marketing

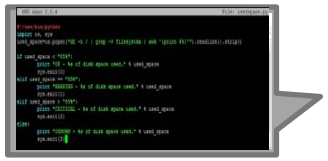


Overall Orchestration

Home Office Team



Data Engineer



Data Scientist

From-S3-To-Stage → From-Stage-To-Star → Run-Star-Tests → Calculate-Profitability → Segment-Customers

Load SQLServer

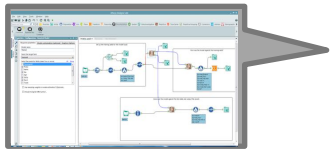
Transform SSIS

Calculate: SQL

Segment: Python

Local Office 'Self-Service' Team

Data Analyst



Target-Customers → Run-Target-Tests → Publish-Tableau-Datasource

Add Data Alteryx

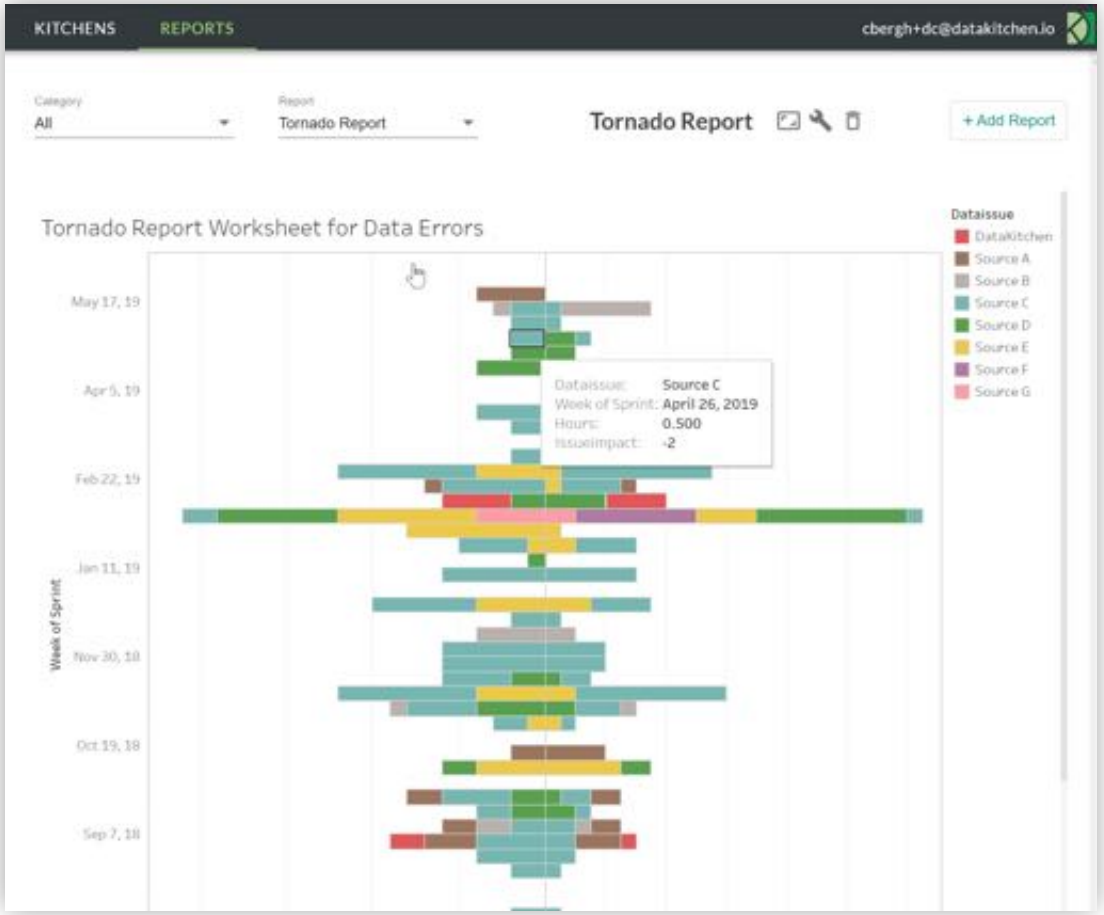
Deploy: Tableau

Publish: T Server

VP Marketing

run-home-ingredient → run-local-ingredient

DataOps Process Analytics

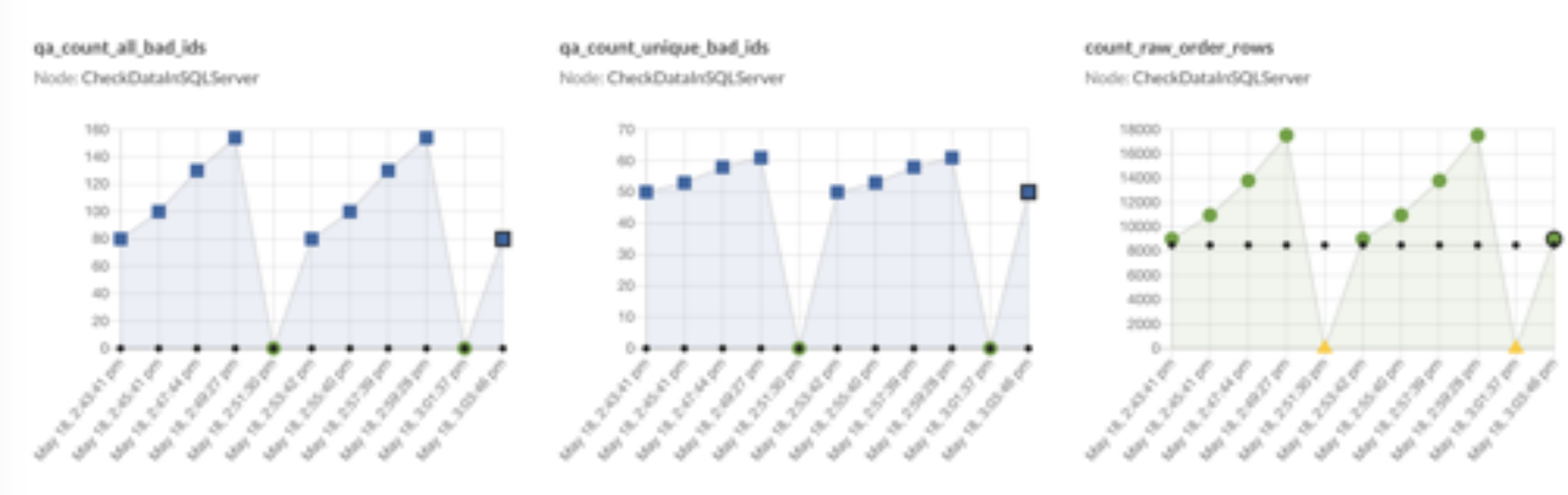


Analytic teams are not very analytic about measuring and improving their internal work

- Prove your teams' value, measure:
  - Team and individual productivity
  - Production error rates
  - Data provider error rates
  - SLAs
  - Production deployment rates
  - Release environments
  - Tests Coverage
- Customizable with data export to fit your company's needs

DataOps: data about data

Statistical process control graphs monitoring “bad IDs” and raw row counts



Team Collaboration Increased

Error Rates Decline in Production

Productivity: Recipe Work Increasing

Per Project Analytics

Deploys Between Environments Increased

Number of Automated Tests Increasing

On Time Delivery within SLA, & decreasing build time

In addition to **data quality**, focus on **errors**

Data Categorization

Background: Nature of Analytics & Data

- Nature of analytics:
  - Focus is on answering business questions;
  - Experimental, iterative, value is hard to determine up front, always changing
  - Analysts are the ‘tip of the spear’

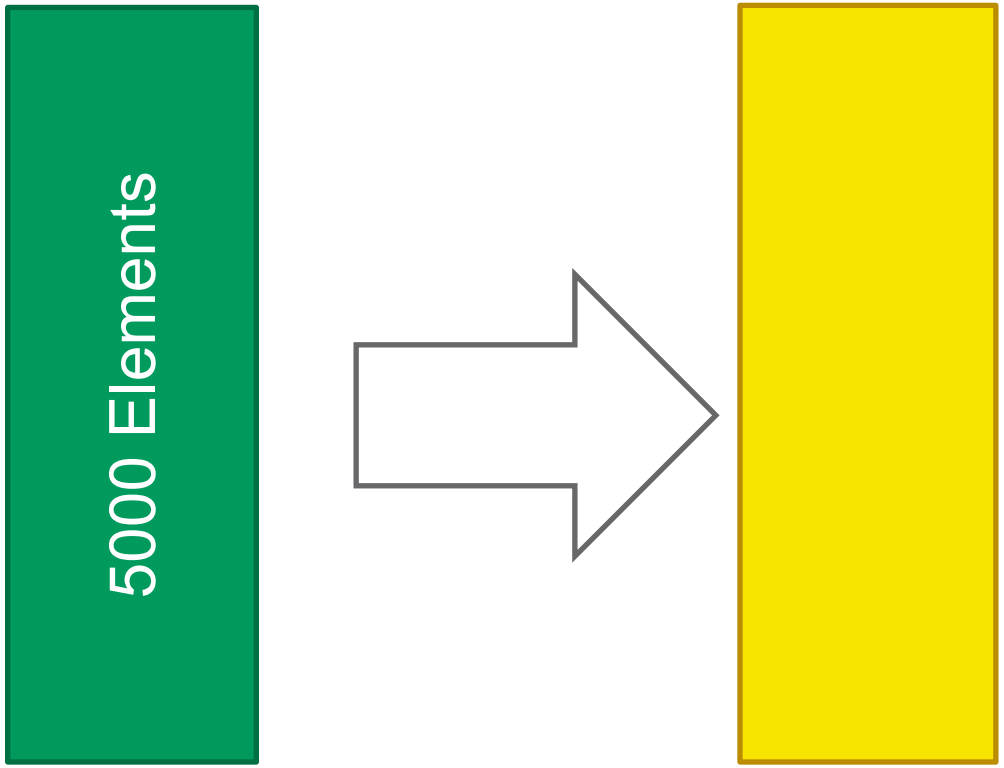
- Three types of data for Analytics:
  - Directionally correct data
  - Gold Standard: As good as it gets data
  - Not needed at all

Scenario: 5000 New Data Elements

Data	Treatment of New Data
5000 Elements	?

Option #1

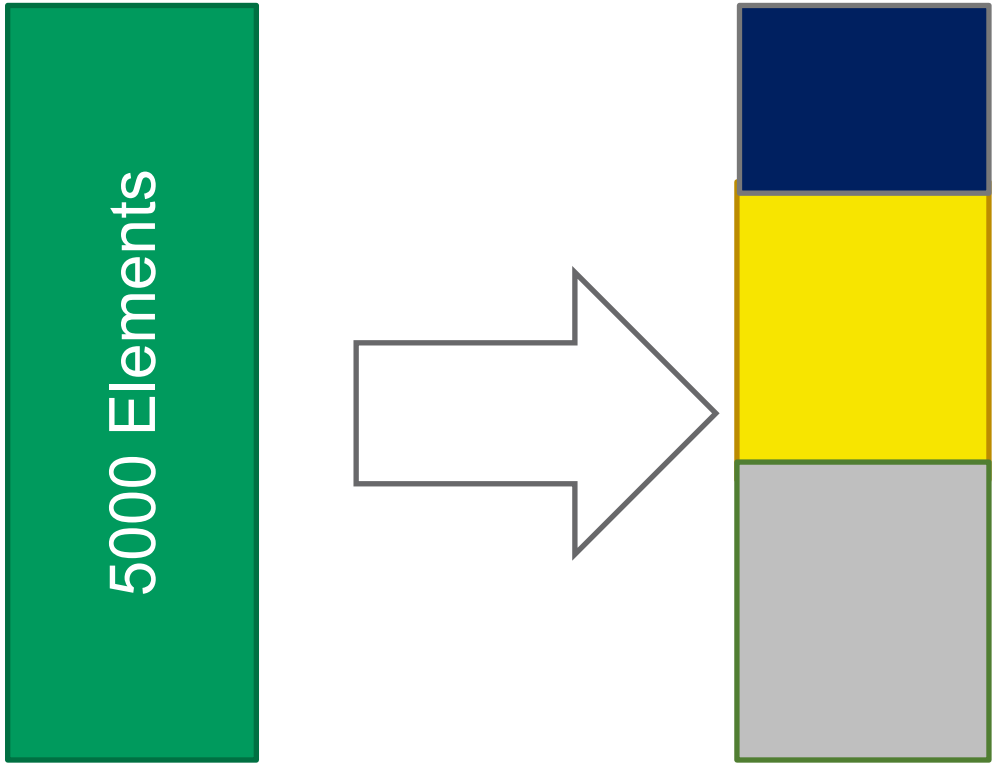
Data	Treatment of New Data



All elements are useful for analytics and other enterprise systems

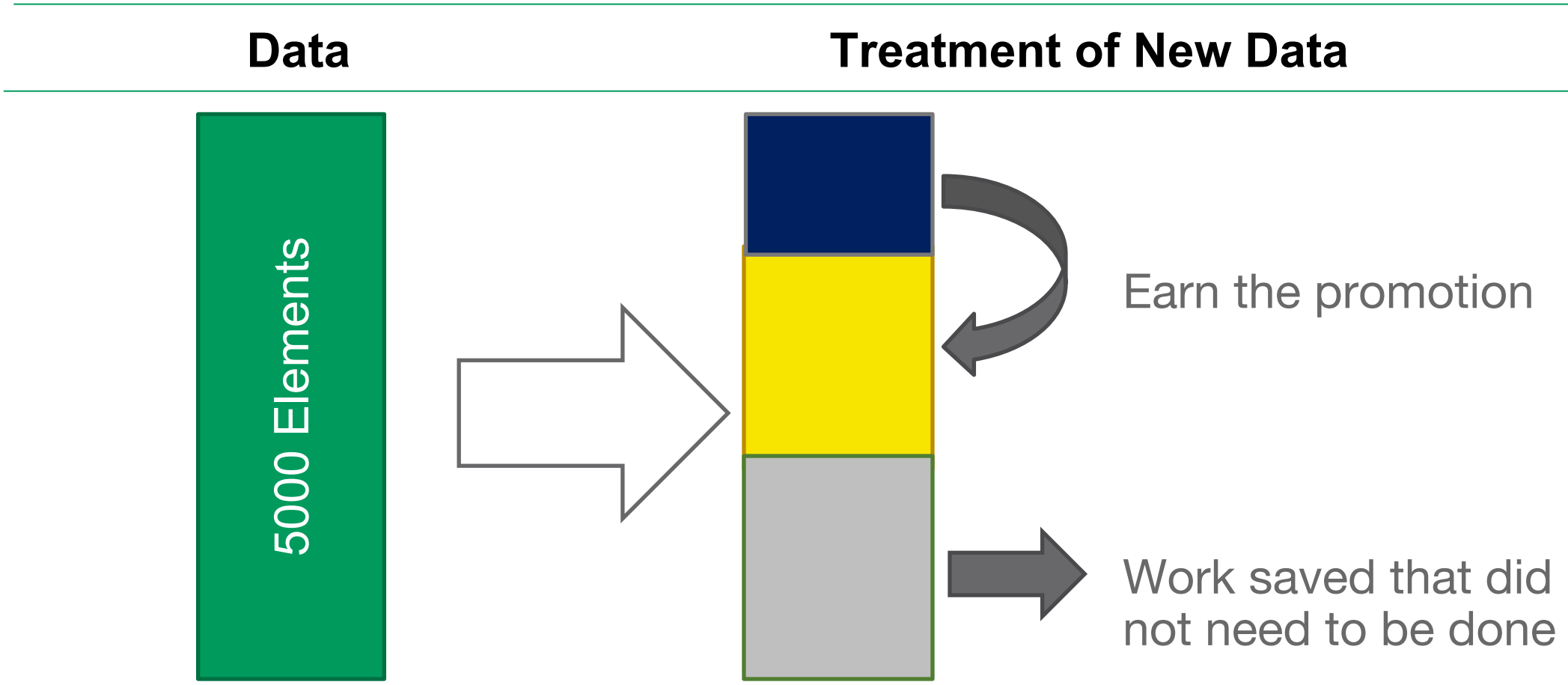
Option #2

Data	Treatment of New Data



- Some are just useful for analytics
- Some are useful for analytics and other enterprise systems
- Some are not useful for either

Option #2



Exercise

What can you use from the bonus steps?

DataOps Collaboration

1. Intra Team
2. Inter Team
3. DataOps Process Analytics
4. Data Categorization

Topics

Why DataOps Is Essential

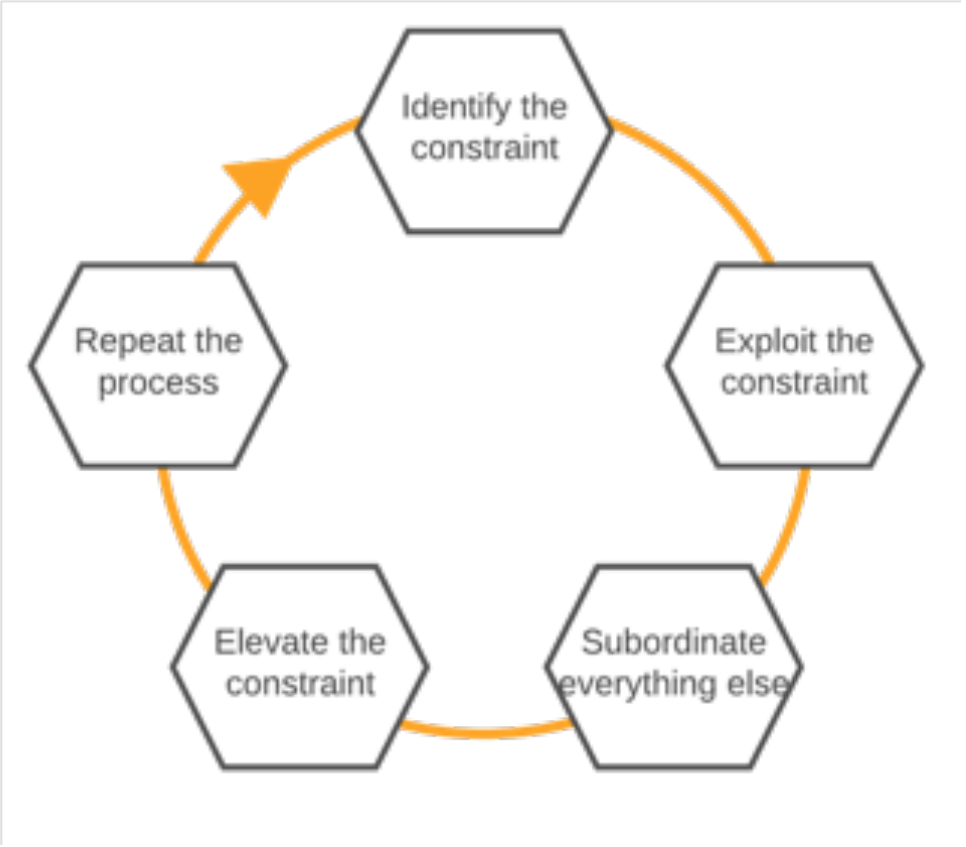
Seven Steps to DataOps

Next Steps With DataOps

Where to Start With DataOps?

Look to manufacturing/DevOps ‘Theory of Constraints’

- Where are ‘bottlenecks’ (or constraints in your data science or analytic process?
- What impedes from creating new insight for you customers?
- Iterate & improve



Example Constraints

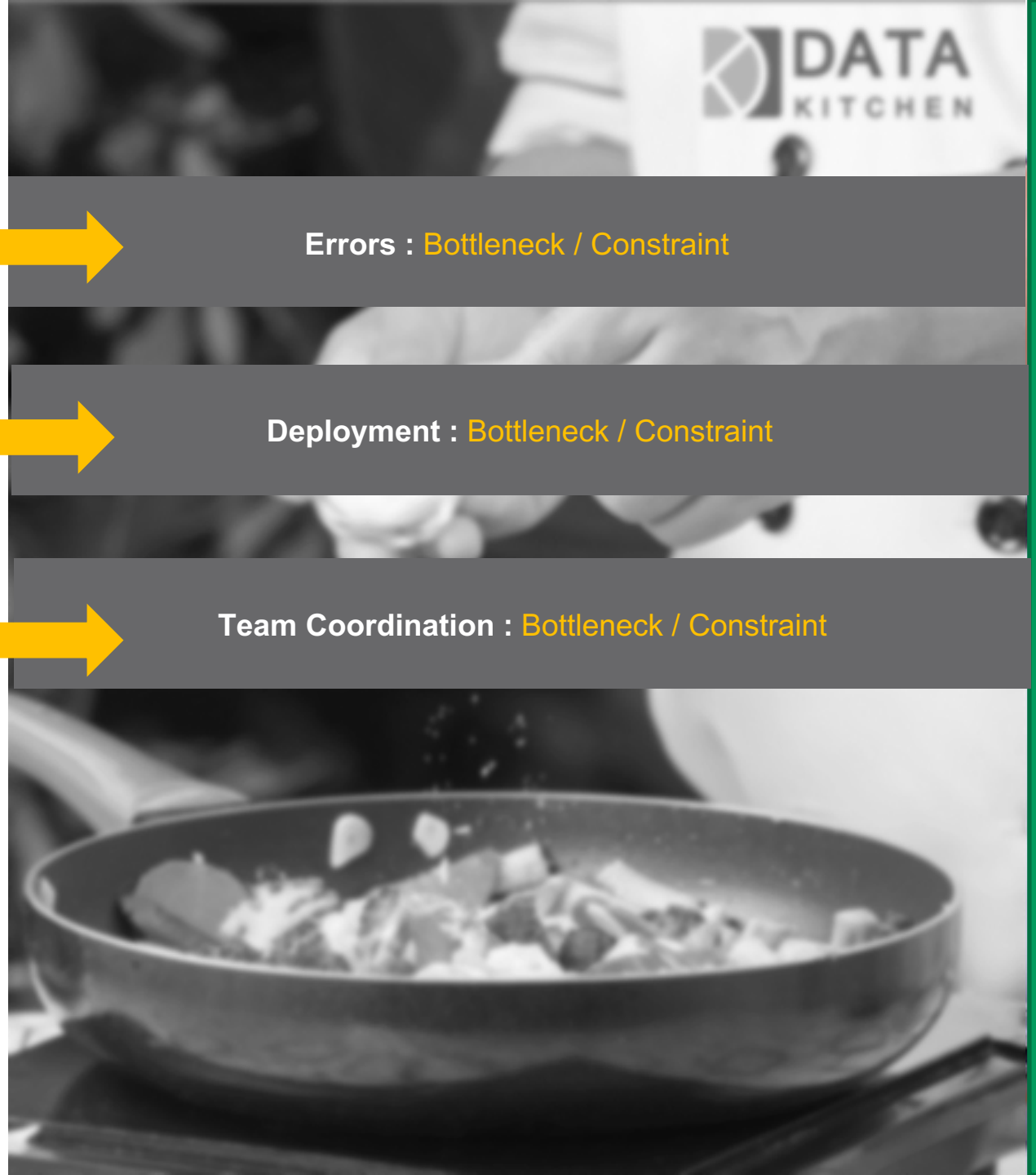
Part 1: [Factory] "I don't want to learn about data quality issues from my customers"

Part 2: [Flow] "I don't want break production when I deploy my changes"

Part 3: [Intra-team coordination] "I don't want my team to struggle working together"

Part 4: [Inter-team coordination] "I don't like the Hatfields vs Mccoys war between different analytic teams"

Part 5: [Management] "How to measure team progress with and show results to leadership?"



Example Constraints

Part 1: [Factory] "I don't want to learn about data quality issues from my customers"

Part 2: [Flow] "I don't want break production when I deploy my changes"

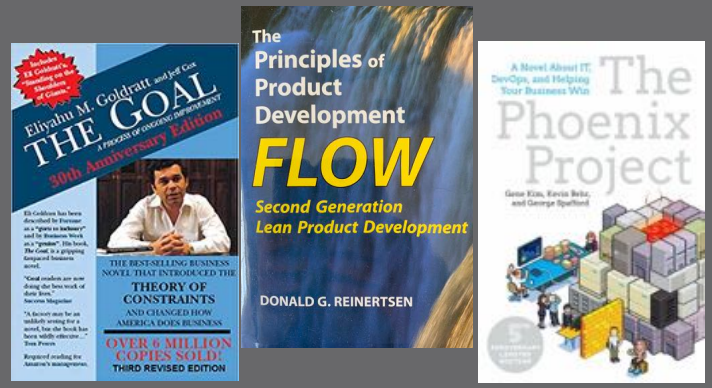
Part 3: [Intra-team coordination] "I don't want my team to struggle working together"

Part 4: [Inter-team coordination] "I don't like the Hatfields vs Mccoys war between different analytic teams"

Part 5: [Management] "How to measure team progress with and show results to leadership?"

Errors, Deployment, and Team Coordination Are All Bottlenecks That

GOAL: Flow of Innovation



DataKitchen Software Platform

Our cloud platform orchestrates data to customer value, speeds features to production, and automates quality.

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

\+ Bonus Steps

Kitchens

Recipes & Tests

Ingredients

Orders

For More Information

- For These Slides, Contact Me:
  - gil@datakitchen.io
- DataOps Manifesto:
  - http://dataopsmanifesto.org
- DataOps Blog:
  - http://medium.com/data-ops
- Follow Twitter:
  - #DataOps

Write down your answer

DATA KITCHEN

What can I take from this session and apply tomorrow?

Exercise

How can I get started?