BOS DATA FESTIVAL

2015

Agile Data In Seven Shocking Steps!

Christopher Bergh

@OPENDATASCI

#ODSC

AGENDA

What Is The Problem?

A Look At Agile Through Data Lens

How To Do Agile Data In Seven Shocking Steps

WHO AM I

Algorithm Nerd

Columbia, MIT, NASA-Ames; ATC Automation

Into In 1990

Fuzzy Logic, Neural Networks, Constraint Satisfaction; Unix/C

Software Nerd

CTO, Dir Engineering, VP Product Management

Into In 2000

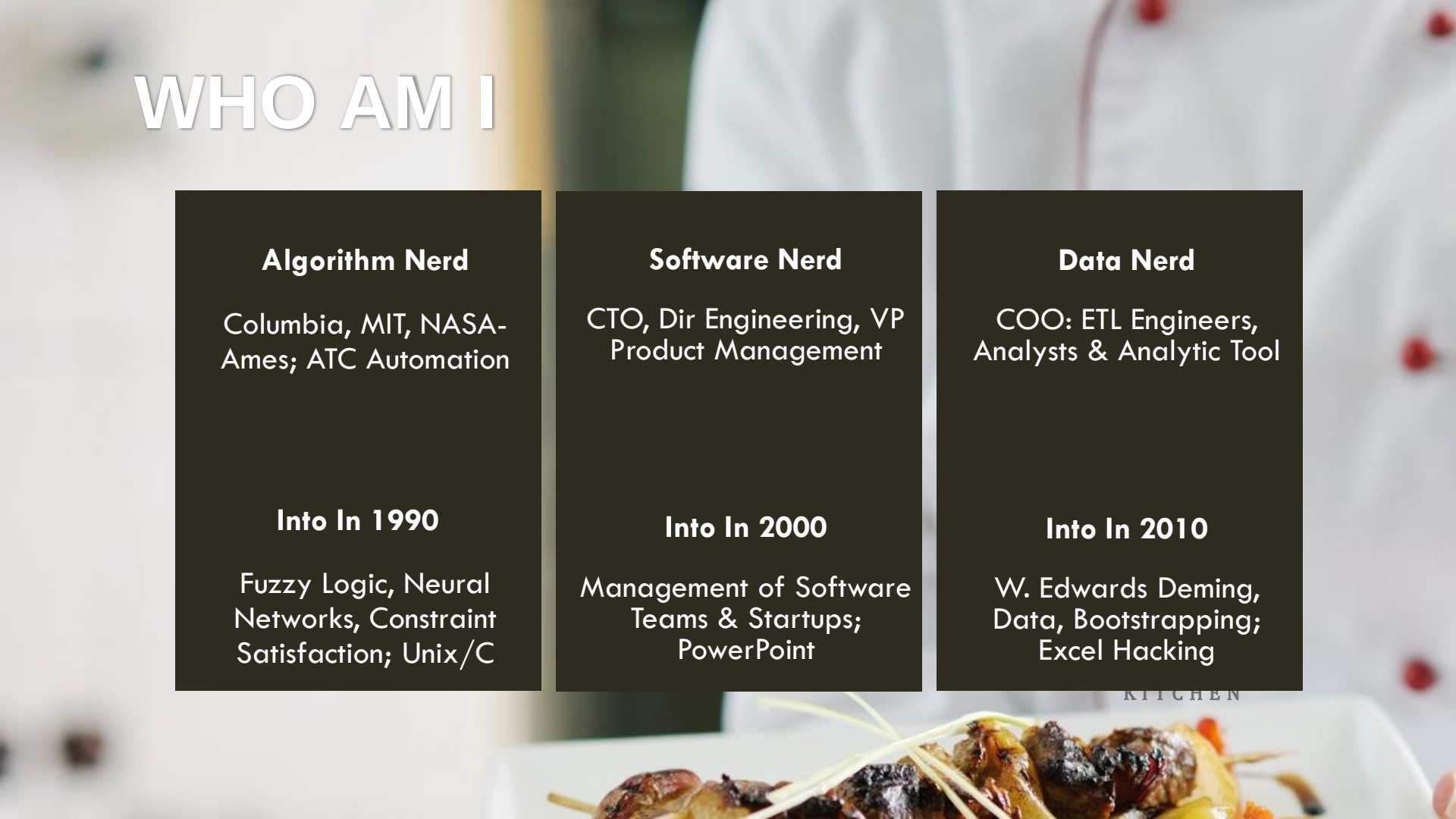
Management of Software Teams & Startups; PowerPoint

Data Nerd

COO: ETL Engineers, Analysts & Analytic Tool

Into In 2010

W. Edwards Deming, Data, Bootstrapping; Excel Hacking



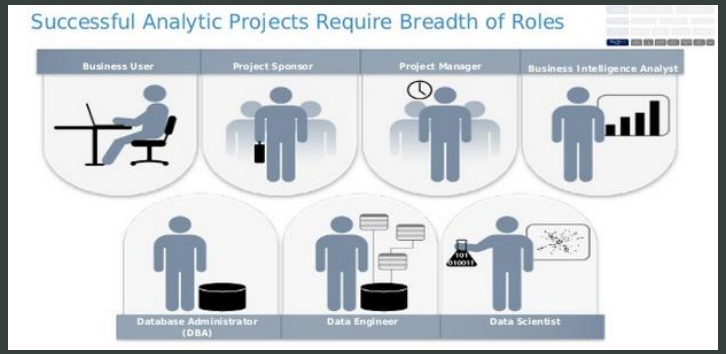
SO WHAT IS THE PROBLEM?

In one word ….

LOTSA

LOTSA

People and Roles In Analytic Teams



DATA SCIENTIST

ETL ENGINEER

Data Governance

DATABASE ARCHITECT

REPORTING ANALYST

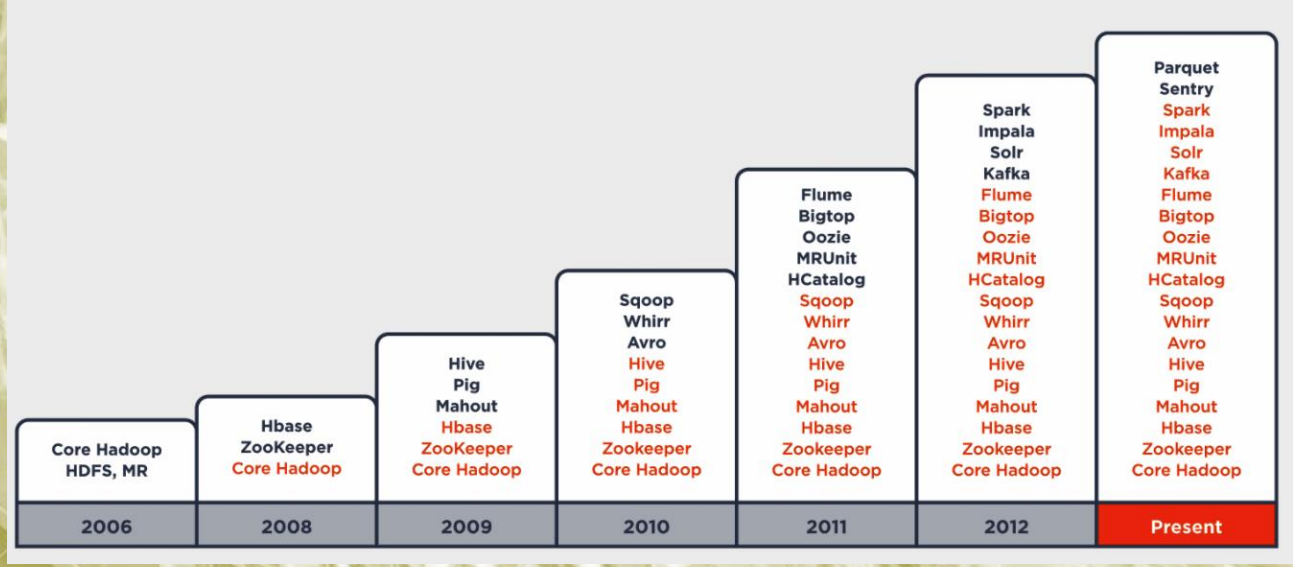
DEV OPS ENGINEER

Manager

Graphic Source: EMC Education -- Building Data Science

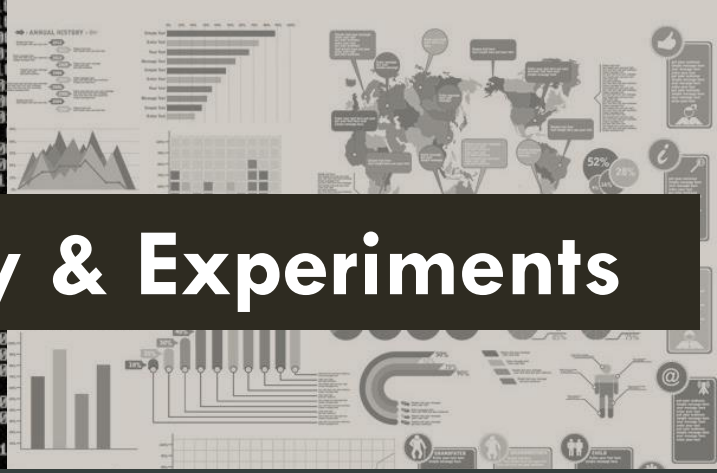
LOTSA

Technologies in Analytics

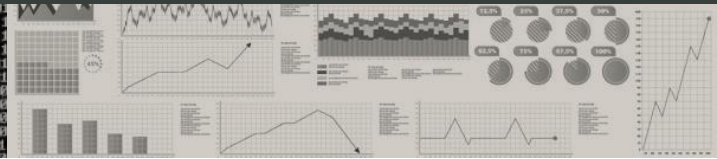


LOTSA

Discovery & Experiments



> *Rapid expansion of data discovery as an alternative to traditional, centrally managed BI delivery models has shifted significant power ”*
>
> Gartner



CREATING Complexity

Another Field, ***Software Development,*** *Ran into the Same Problems With Complexity …*

… They Used Something Called 'Agile' To Solve The Problem

What Is The Problem?

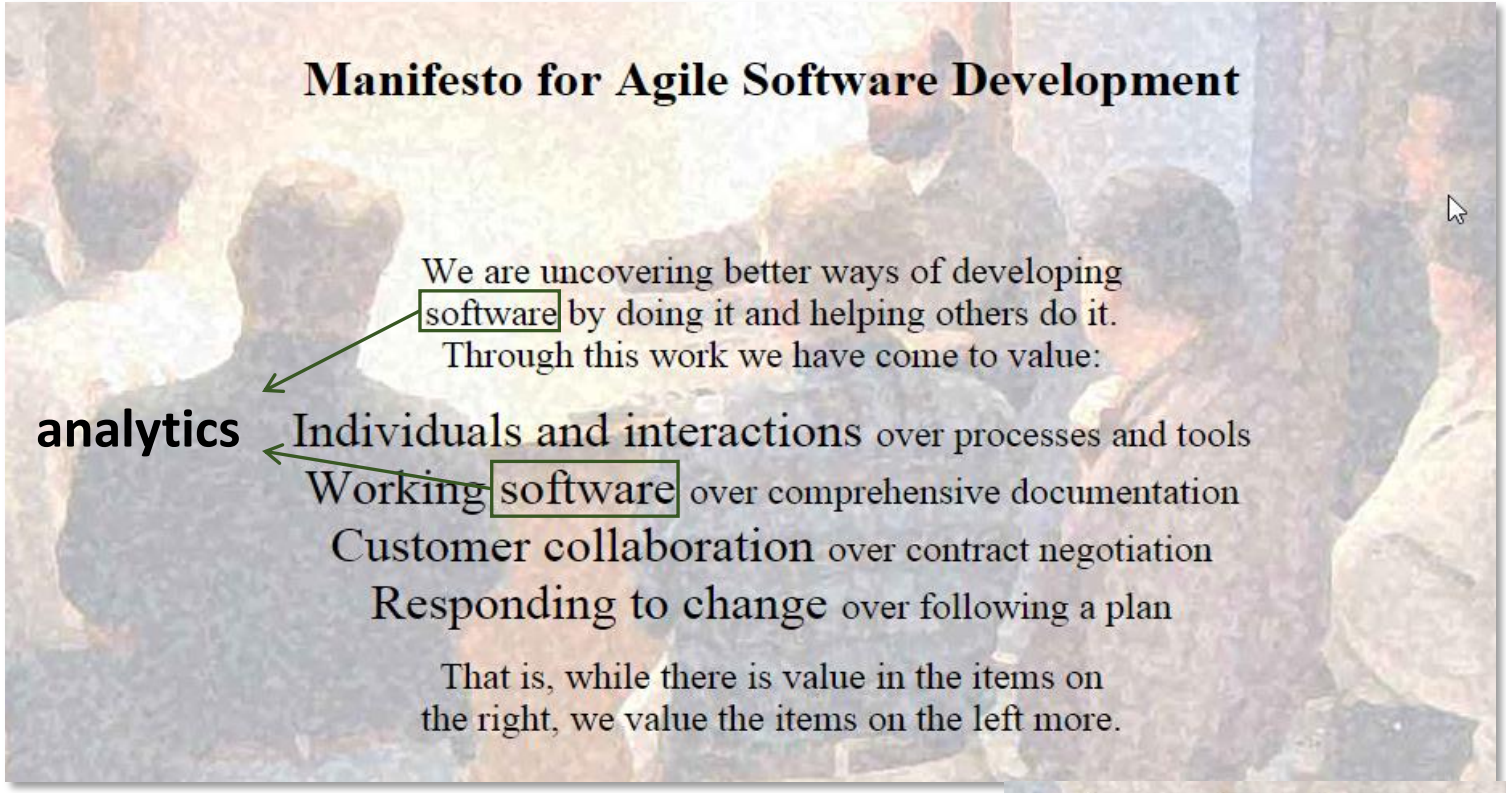
A Look At Agile Through Data Lens

How To Do Agile Data In Seven Shocking Steps

AGENDA

Agile … A Solution To That Complexity

agilemanifesto.org



Twelve Principles of Agile Software

DATA KITCHEN

Some Agile Practices Are Easier To Apply

- Development Sprints
- User Stories
- Daily Meetings
- Defined Roles
- Retrospectives
- Pair Programming
- Burn Down Charts

DATA KITCHEN

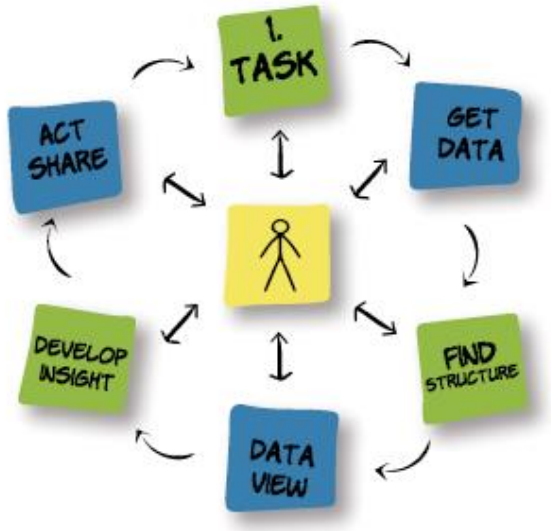
Some practices have been difficult to apply

- Individual Development Environments
- Test Driven Development
- Branching And Merging
- Refactoring
- Small Releases
- Frequent Or Continuous Integration
- Experimentation For Learning

DATA KITCHEN

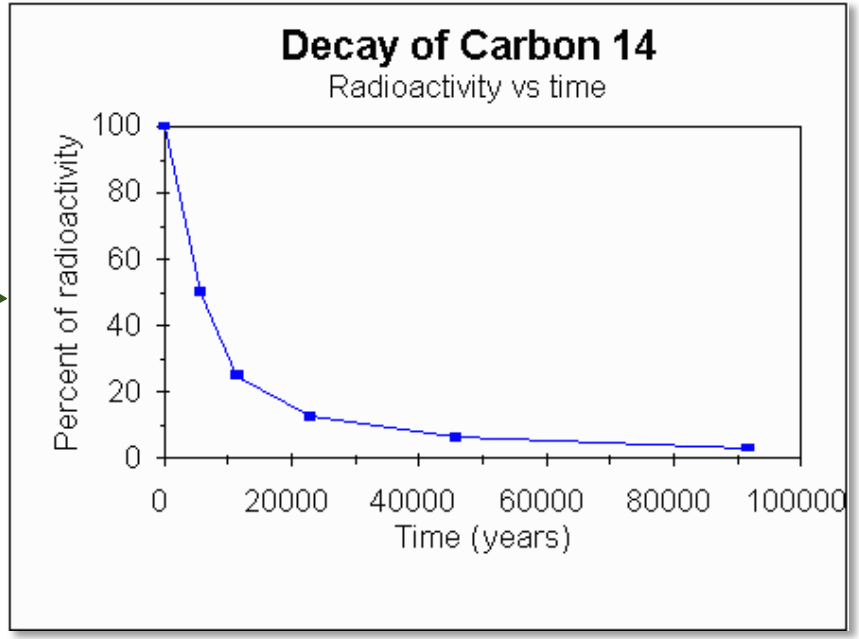
Agile – What Is Unique To Analytics?

PUT THE ANALYST (New Role) AT THE CENTER



Agile – What Is Unique To Analytics?

ANALYICS PERCIEVED VALUE DECAY CURVE



What Is The Problem?

A Look At Agile Through Data Lens

How To Do Agile Data In Seven Shocking Steps

AGENDA

Seven Key Steps To Agile In Analytics

1. Add Tests
2. Modularize & Containerize
3. Do Branching & Merging
4. Use Multiple Environments
5. Give Your Analyst ‘Knobs’
6. Use Simple Storage
7. Support Three Workflows

DATA KITCHEN

❶ Add tests

Types

1. **Error** – stop the line
2. **Warning** – investigate later
3. **Info** – list of changes

Examples

1. Input file row count way below a critical threshold
2. Input file row count a little below a threshold
3. These customers changed territories

And keep adding them with each feature developed!

DATA KITCHEN

❷ Modularize & Containerize

Modularize

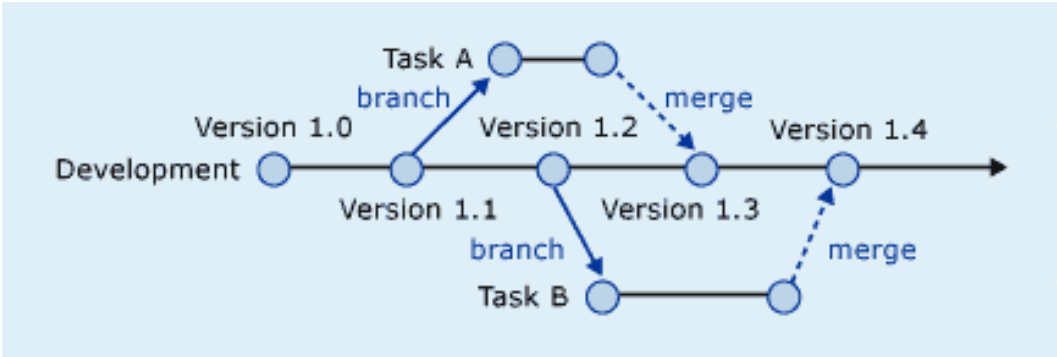
1. Do not create on ‘monolith’ of code
2. Break up you work into component modules

Containerize

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

DATA KITCHEN

❸ Branch & Merge

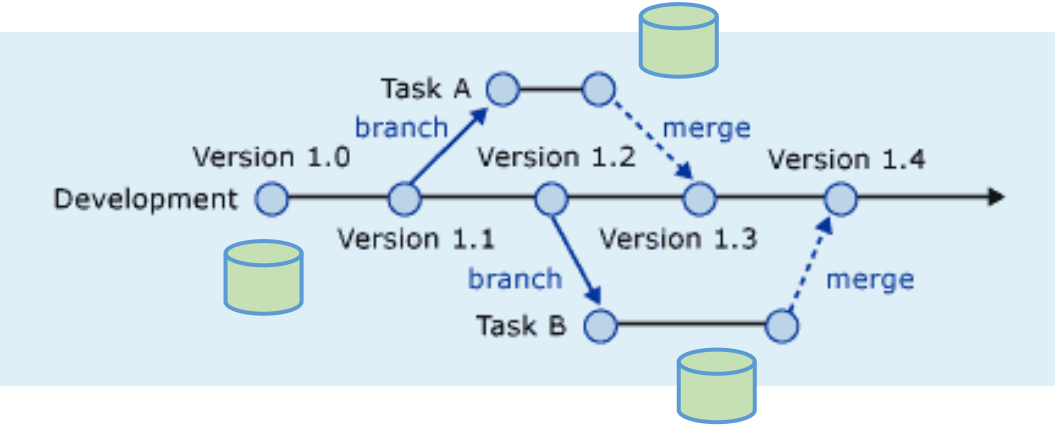


Manage your work like code

- Why? Your work is just code: models, transforms, etc.
- Use a source code control system (like GIT) to enable Branching & Merging

DATA KITCHEN

❹ Use Multiple Environments



Provide a data environment for each branch

- Analysts need a controlled environment for their experiments
- Engineers need a place to develop outside of production

DATA KITCHEN

❺ Give Your Analysts Knobs

- Give you analysts and data scientists the ability to edit the Production Database safely
- Why?
  - *"Best-in-class companies **take 12 days** to **integrate new data sources** into their analytical systems; industry **average** companies take* **60** *days; and*, **laggards** *average* **143** *days"*

Figure out how to do this in minutes

Source: Aberdeen Group: Data Management for BI: Fueling the analytical engine with high-octane information

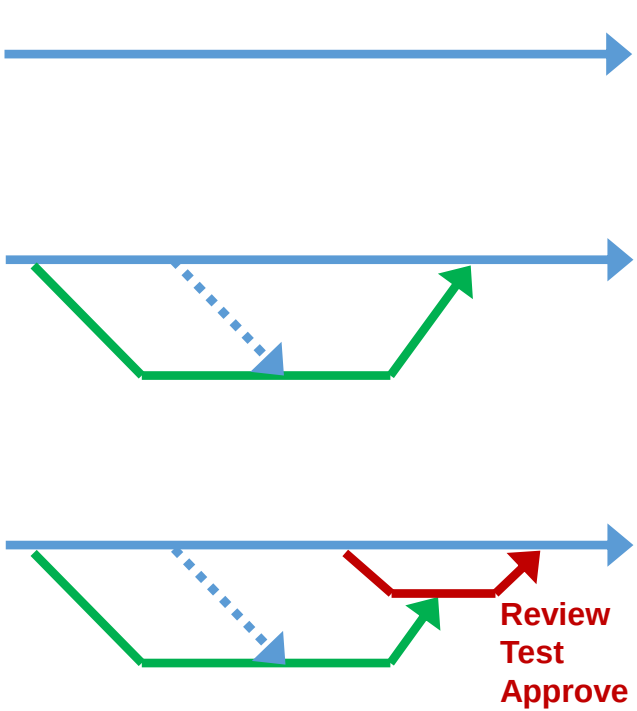
DATA KITCHEN

❻ Use Simple Storage

- Keep copies of all your raw data in simple, cheap storage (s3, HFDS)
  - ➢ Data Lake
- Create parameterized variations of your process that allow you to assemble data for experimentation, development, and production
  - ➢ Data Marts
- Be able to back up and restore your databases easily
  - ➢ “My own database”

DATA KITCHEN

❼ Support Three Workflows



Small Team

- Promote directly to production

Feature Branch

- Merge back to production branch

Data Governance

- 3rd party verification before production merge

DATA KITCHEN

CONCLUSION


It Takes A Whole Village to Raise A Child
www.peaceproject.com 888-822-7075 (#MS10F)
- African Proverb

CONCLUSION