SILICON VALLEY
DATA SCIENCE

Data Science and Software Product Development

John Akred | @BigDataAnalysis

DSAA 2018

To receive a copy of these slides, please send me a direct message on Twitter (@BigDataAnalysis)

MY INTREPID COLLEAGUES

WE DO DATA RIGHT.

- We work in cross-functional teams made up of data scientists, engineers, and solutions architects.
- We combine enterprise know-how with custom methods derived from Silicon Valley best practices.
- We use an agile development approach to make iterative progress against difficult problems.
- We focus on delivering business value as early as possible, while iterating toward the larger goal.

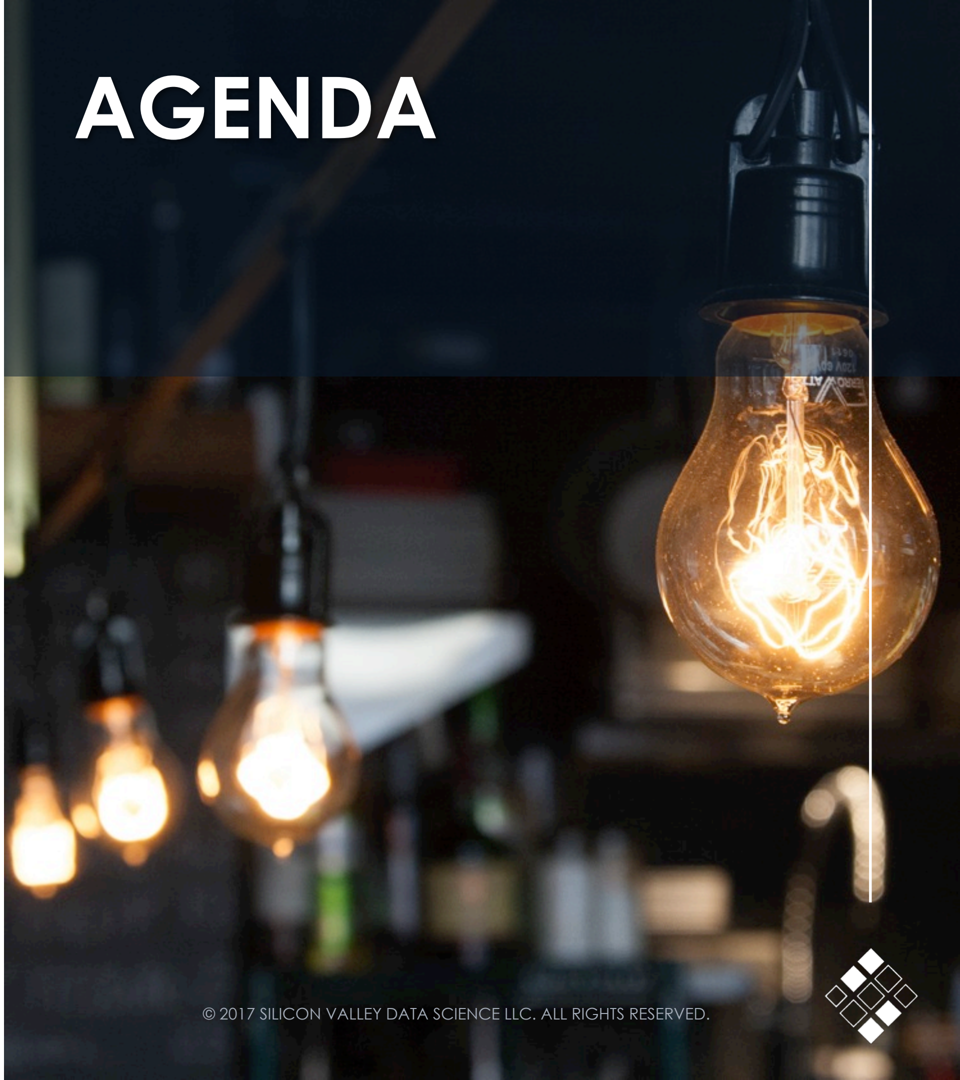
OUR PHILOSOPHY

- Prioritize for highest business value when innovating with technology
- Design with outcomes in mind
- Be agile: share intermediate outputs, incorporate feedback
- Collaborate constantly with stakeholders and partners

@SVDataScience

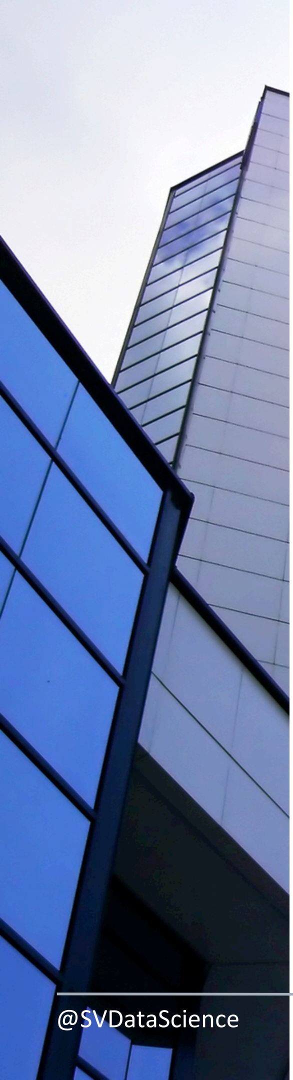
AGENDA

- Challenges of Integrating Data Science and Software Development
- Challenges in the Enterprise Environment
- Methods: What do we have to work with?
- A Method for Integrating Data Science with Agile Software Development
- Opportunities for research and best practices



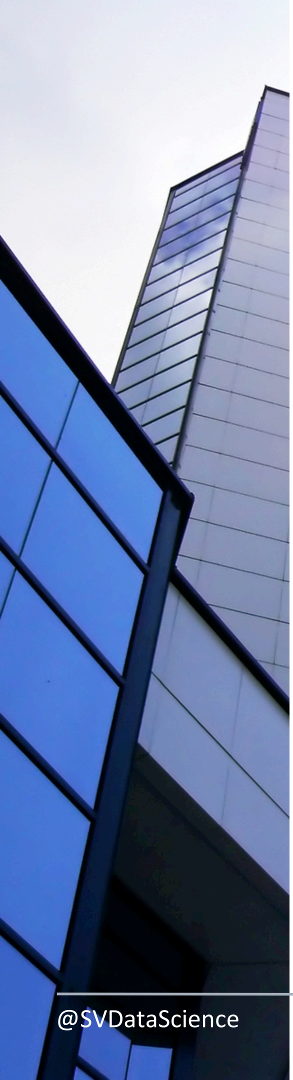
@SVDataScience

Challenges of Integrating Data Science and Software Development



BUSINESS INTELLIGENCE AND DATA SCIENCE

Business Intelligence	Data Science
Information in dashboards	Guided decision-making
How much churn was there?	How might I reduce churn rate?
Current and historical	Future-looking
What?	Why? How?
Business-focused skills	Math-focused skills
Proprietary tools	Open source tools
Tactical	**Strategic**



DATA SCIENCE & ENGINEERING

ABSTRACT MODEL
(theory)

↑ compare

CONCRETE DESCRIPTION
(data)

↑ measure

PHYSICAL SYSTEM
(object of study)

ABSTRACT MODEL
(design concept)

↓ design

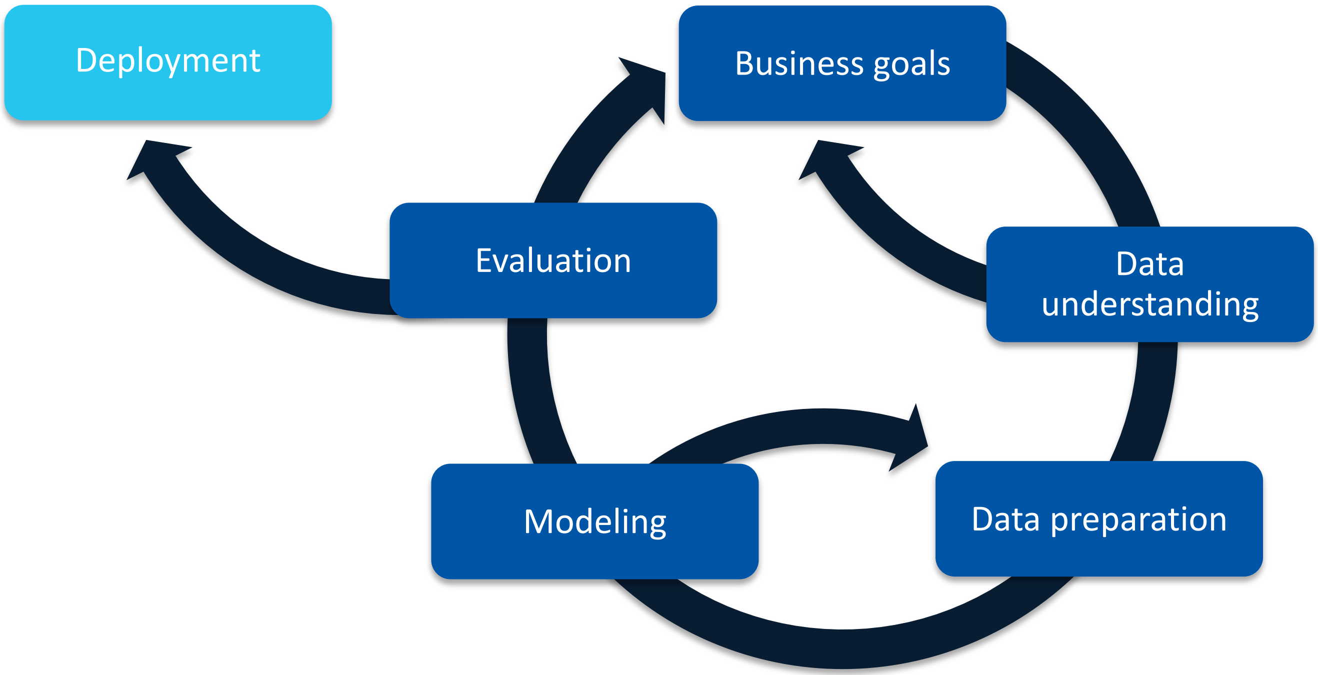
CONCRETE DESCRIPTION
(specification)

↓ produce

PHYSICAL SYSTEM
(useful product)

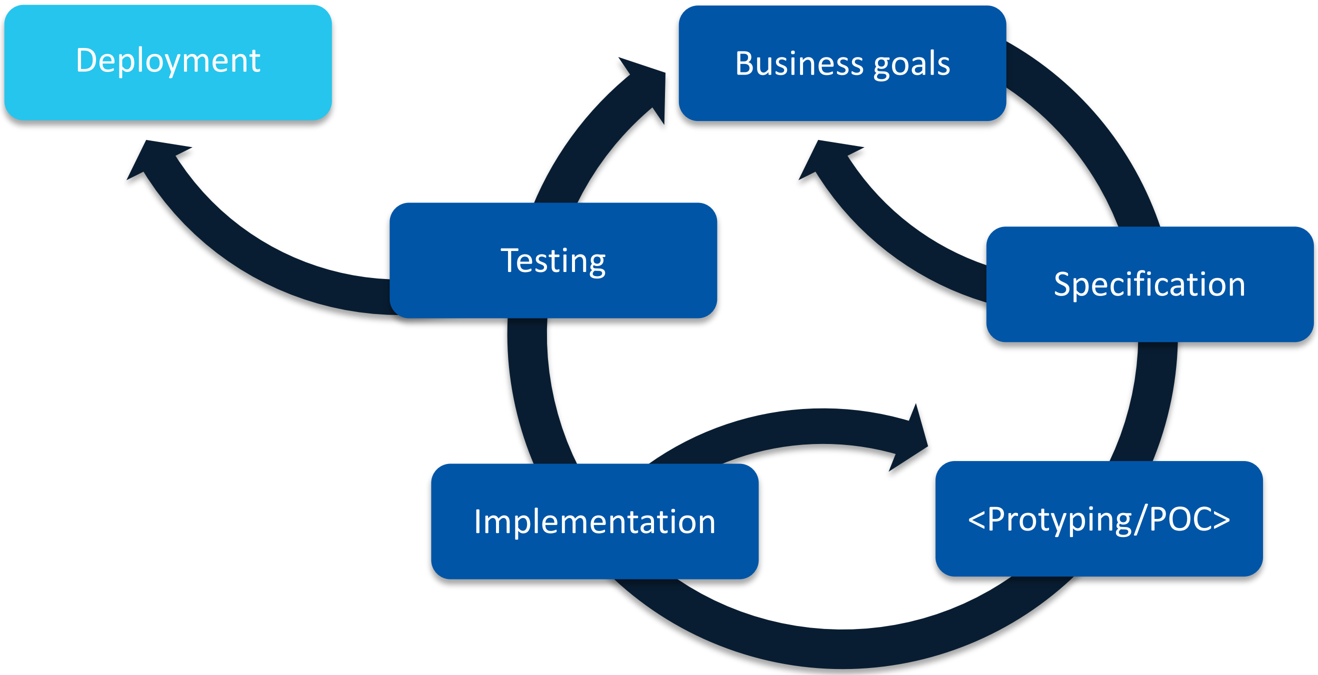
ADAPTED FROM: https://www.farnamstreetblog.com/2013/07/the-difference-between-science-and-engineering/

CRISP-DM



@SVDataScience

Software Development

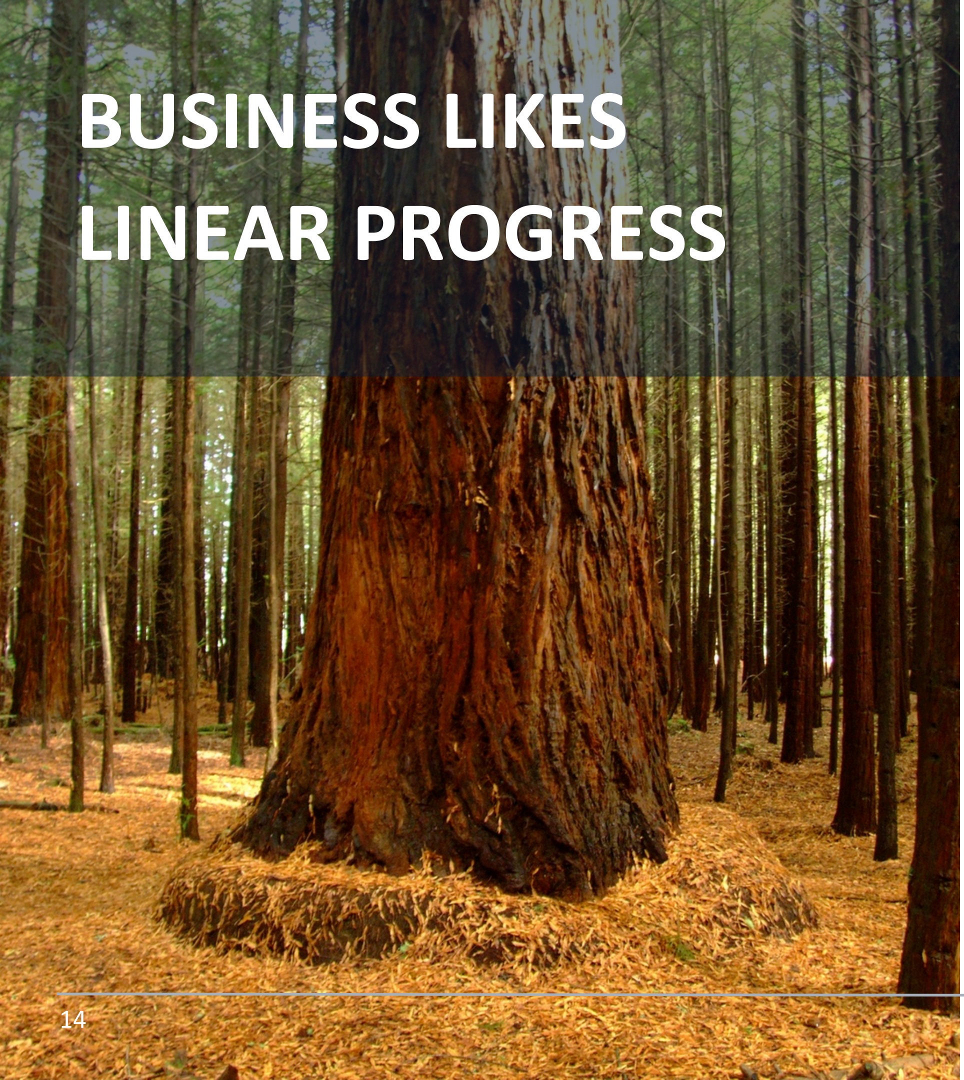


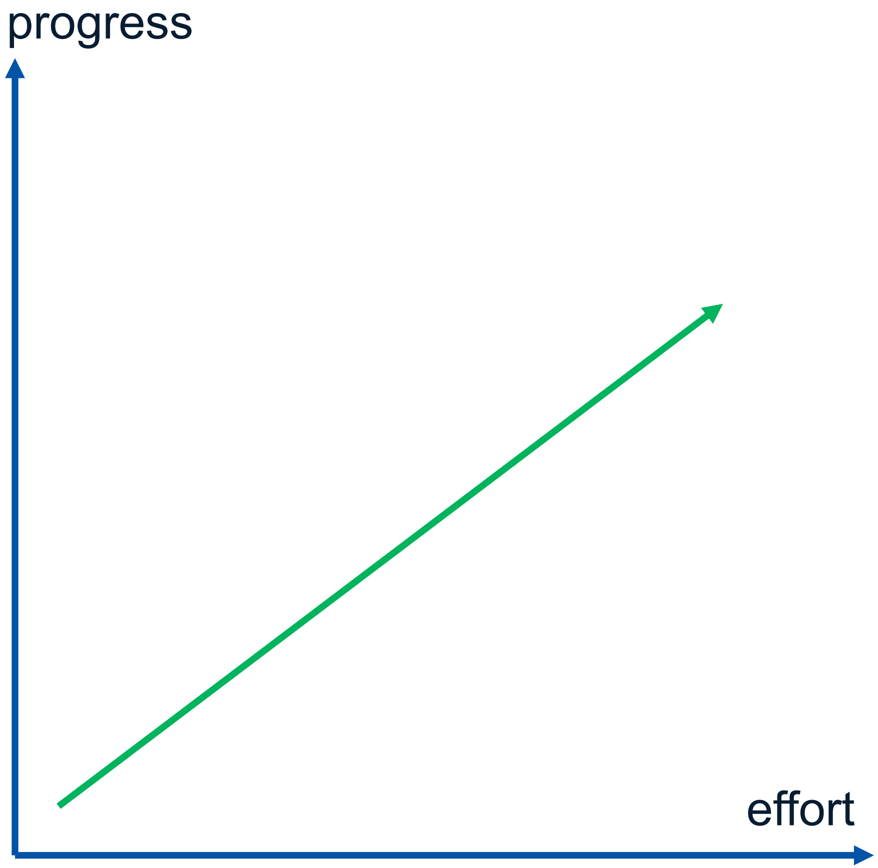
WANT SOME?

- Solid data strategy
- Functioning platform
- Tolerance for failure
- Ability to act on insights

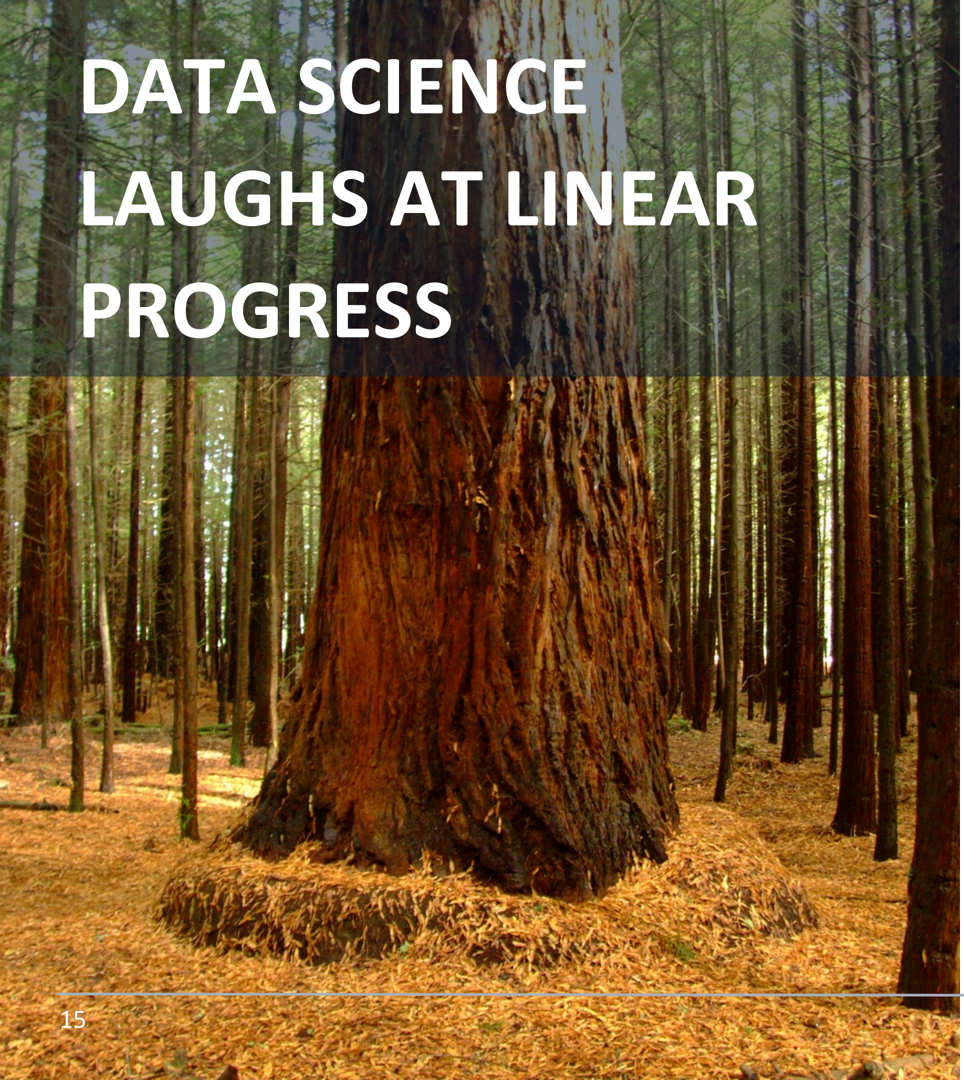
CHALLENGES IN THE ENTERPRISE ENVIRONMENT

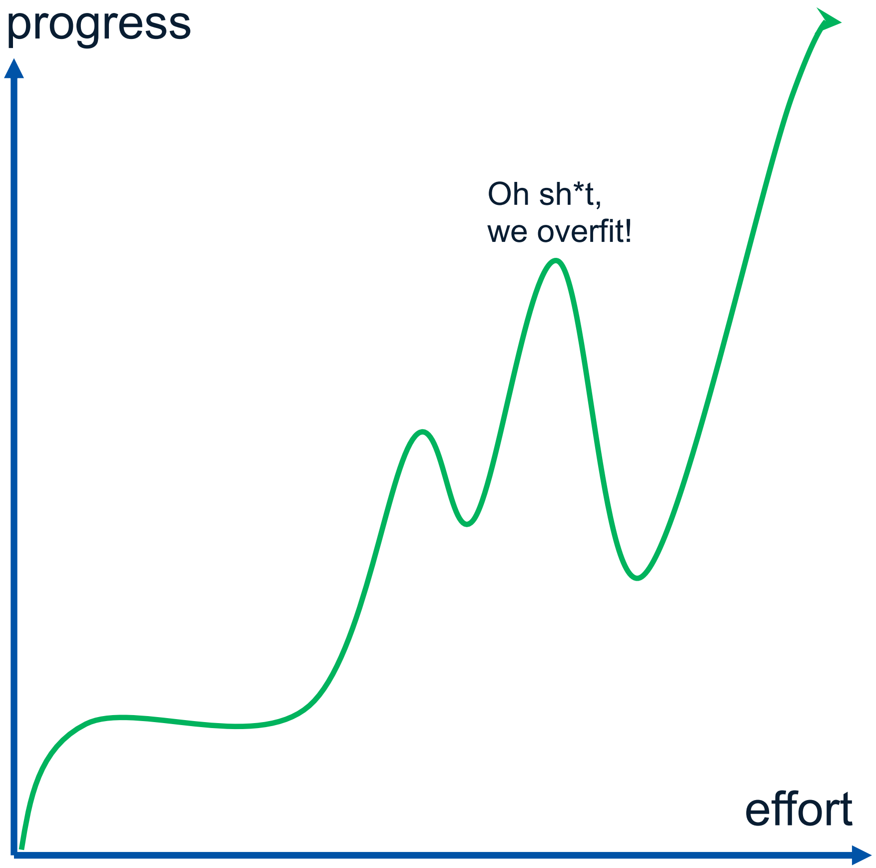
BUSINESS LIKES LINEAR PROGRESS





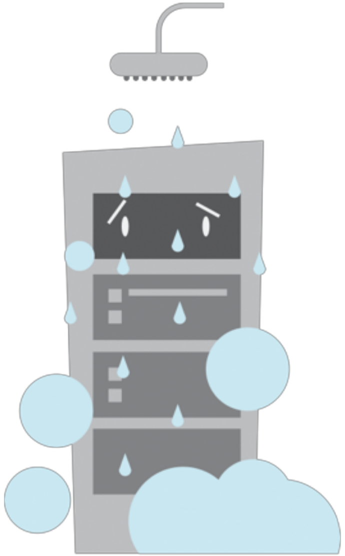
DATA SCIENCE LAUGHS AT LINEAR PROGRESS



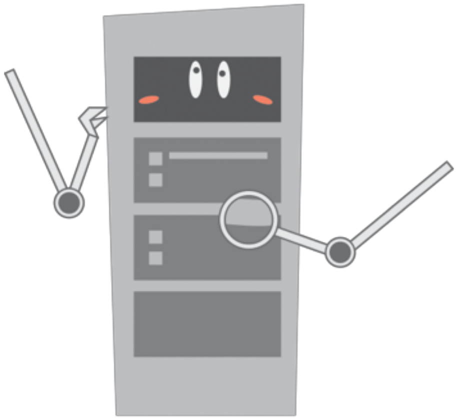


CONVENTIONAL DATA STRATEGY

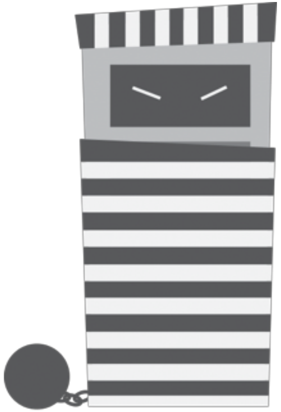
“WHAT YOU DO ***TO*** DATA”



CLEAN



VALIDATE



CONTROL

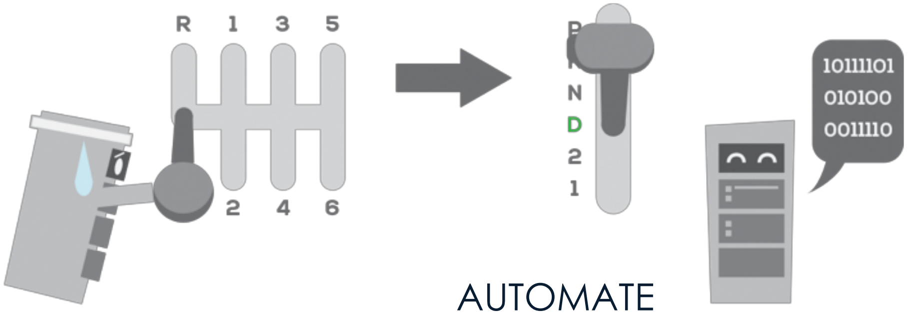
PROTECT

MODERN DATA STRATEGY

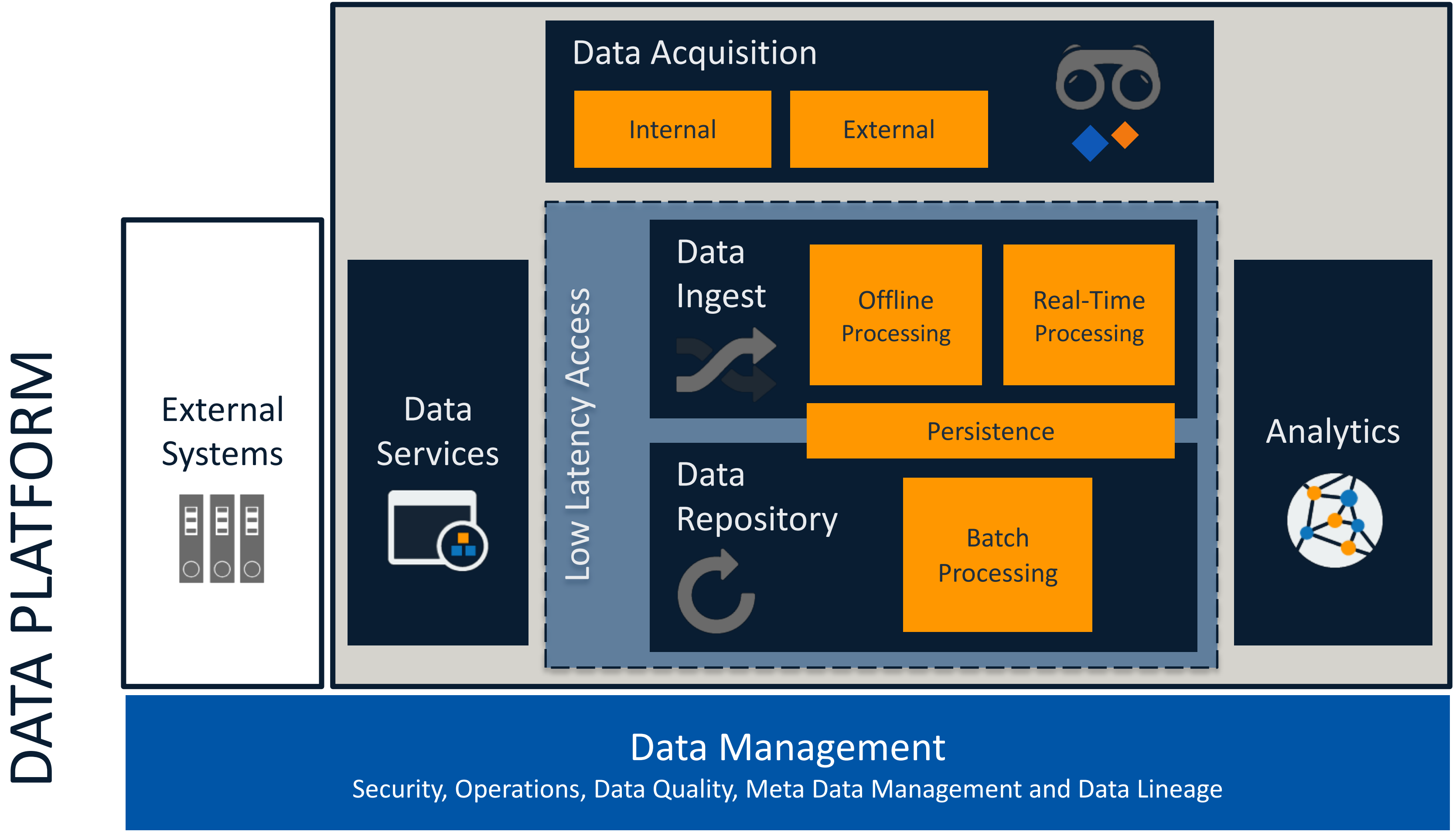
“WHAT YOU DO ***WITH*** DATA”

ATTRACT NEW CUSTOMERS

TARGET VIP CUSTOMERS



AUTOMATE


DATA PLATFORM
Data Acquisition
Internal
External
External Systems
Data Services
Low Latency Access
Data Ingest
Offline Processing
Real-Time Processing
Persistence
Data Repository
Batch Processing
Analytics
Data Management
Security, Operations, Data Quality, Meta Data Management and Data Lineage

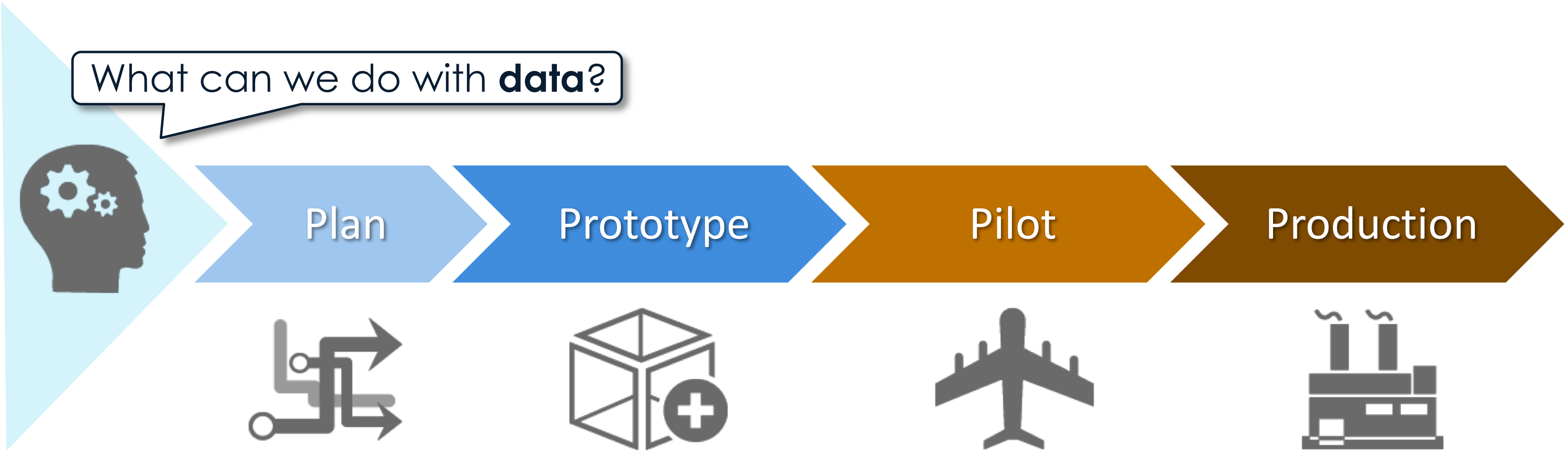
THE DATA VALUE CHAIN

from raw data to data-driven product

Discover Ingest Process Persist Integrate Analyze Expose

FROM IDEA TO PRODUCTION

We identify the business goals, distill those into hypotheses, and then work in short, iterative cycles to achieve tangible gains.



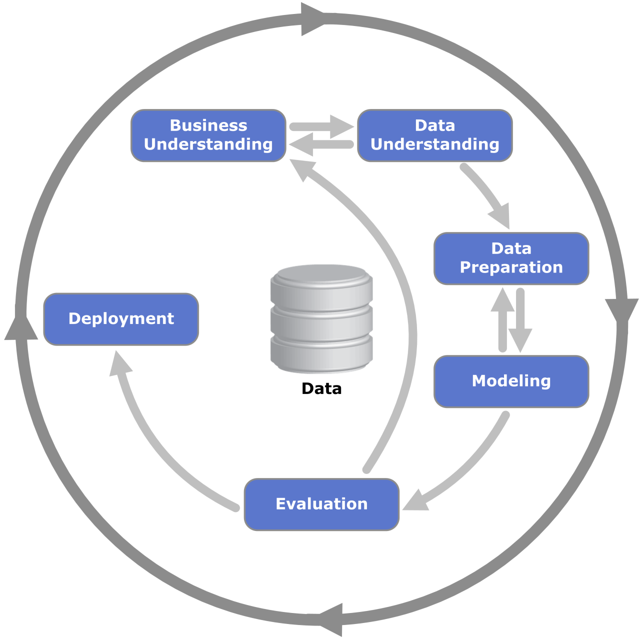
What are our options?

Methodologies for Data Science

Methods for Data Science

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]
2014 poll 2007 poll

	2014 poll	2007 poll
CRISP-DM (86)	43%	42%
My own (55)	27.5%	19%
SEMMA (17)	8.5%	13%
Other, not domain-specific (16)	8%	4%
KDD Process (15)	7.5%	7.3%
My organizations' (7)	3.5%	5.3%
A domain-specific methodology (4)	2%	4.7%
None (0)	0%	4.7%

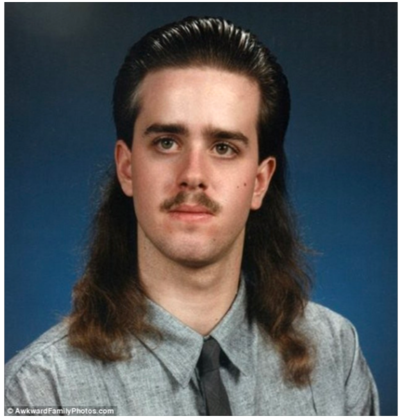


http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

SOFTWARE ENGINEERING METHODS

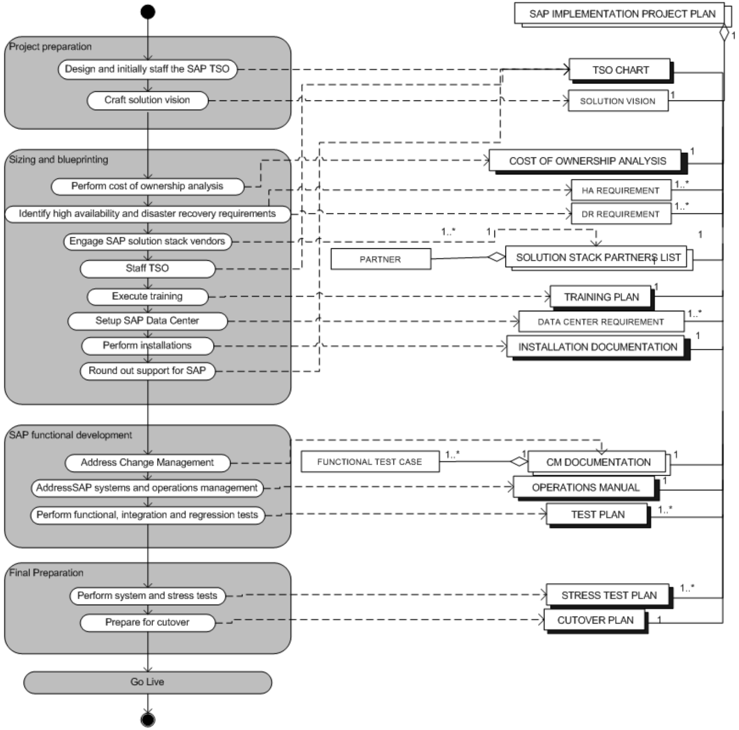
SCRUMFALL

The "Mullet" of Methodologies

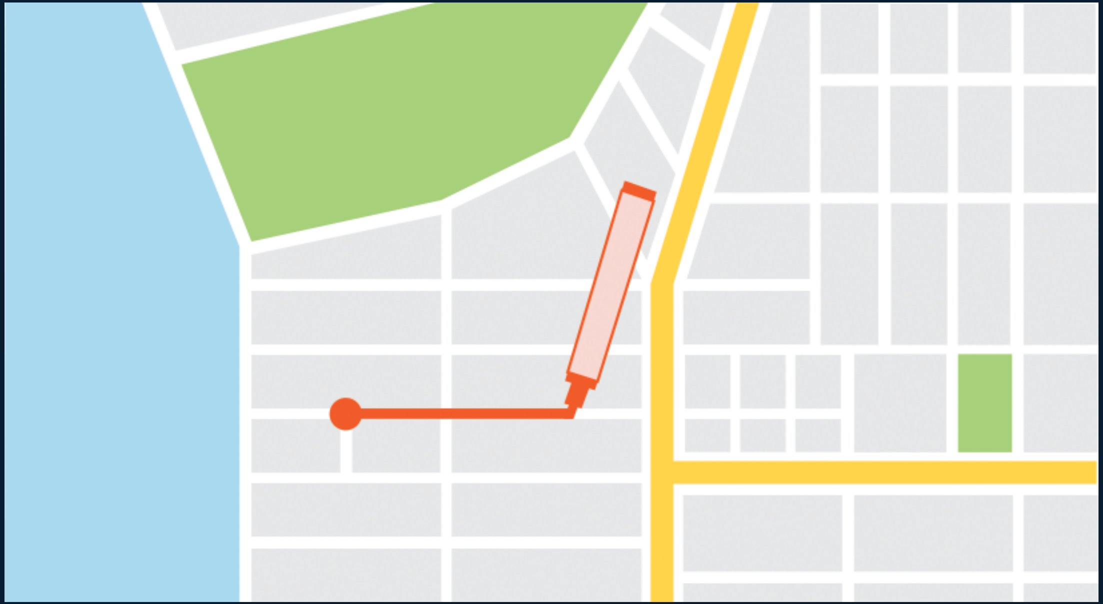


WATERFALL

Great in theory, sometimes in practice

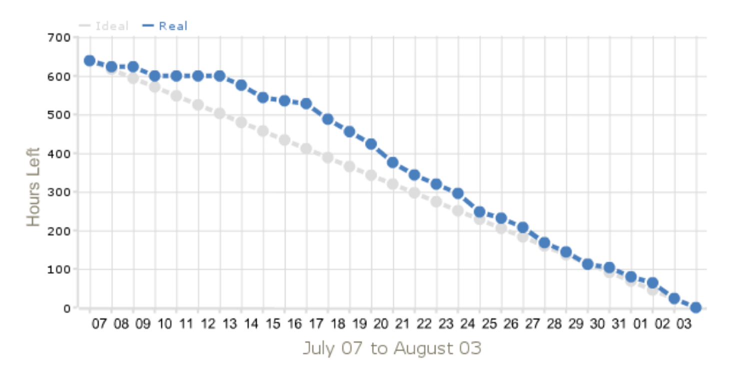


Where to?



Common Objections

- Software development emphasizing shipping product
- Data Science is non-linear



DEFINING SUCCESS

- ✓ Incremental revenue
- ✓ Time to market
- ✓ Economic functional implementation
- ✓ Cost avoidance
- ✓ Brand benefit
- ✓ Goodwill

HOW DO WE GET THERE?

Charter

Investigation Themes

Epics

Stories

Agile Data Science Basics

- The **Project Charter** identifies the desired end point, and *expected* timeline for getting there
- There is a **plan** for the overall project which charts the investigation themes that will be focused on over the expected timeline
- The project is organized and run in **sprints** – (typically) two week increments of work
- Work is organized into **stories** – specific tasks necessary for reaching the goal which can be reasonably expected to be completed in the sprint
- Each sprint has a regular cadence of coordination and feedback meetings
    - **Kickoff** – populate the backlog with the stories selected for the current sprint, assign responsibility for each story
    - **Standups** – daily, BRIEF coordination meetings
    - **Retrospective** – the time to show what's been accomplished and steer the project based on lessons learned, "product" feedback, etc.

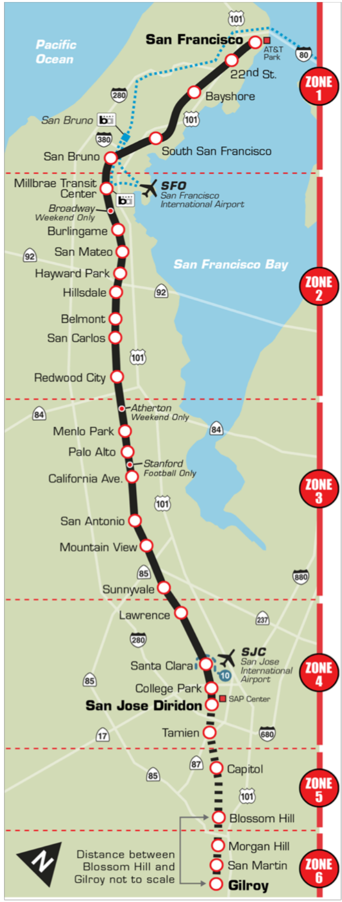
SILICON VALLEY
DATA SCIENCE

Charter:

Why is a data science & engineering consulting company building its own Caltrain app?

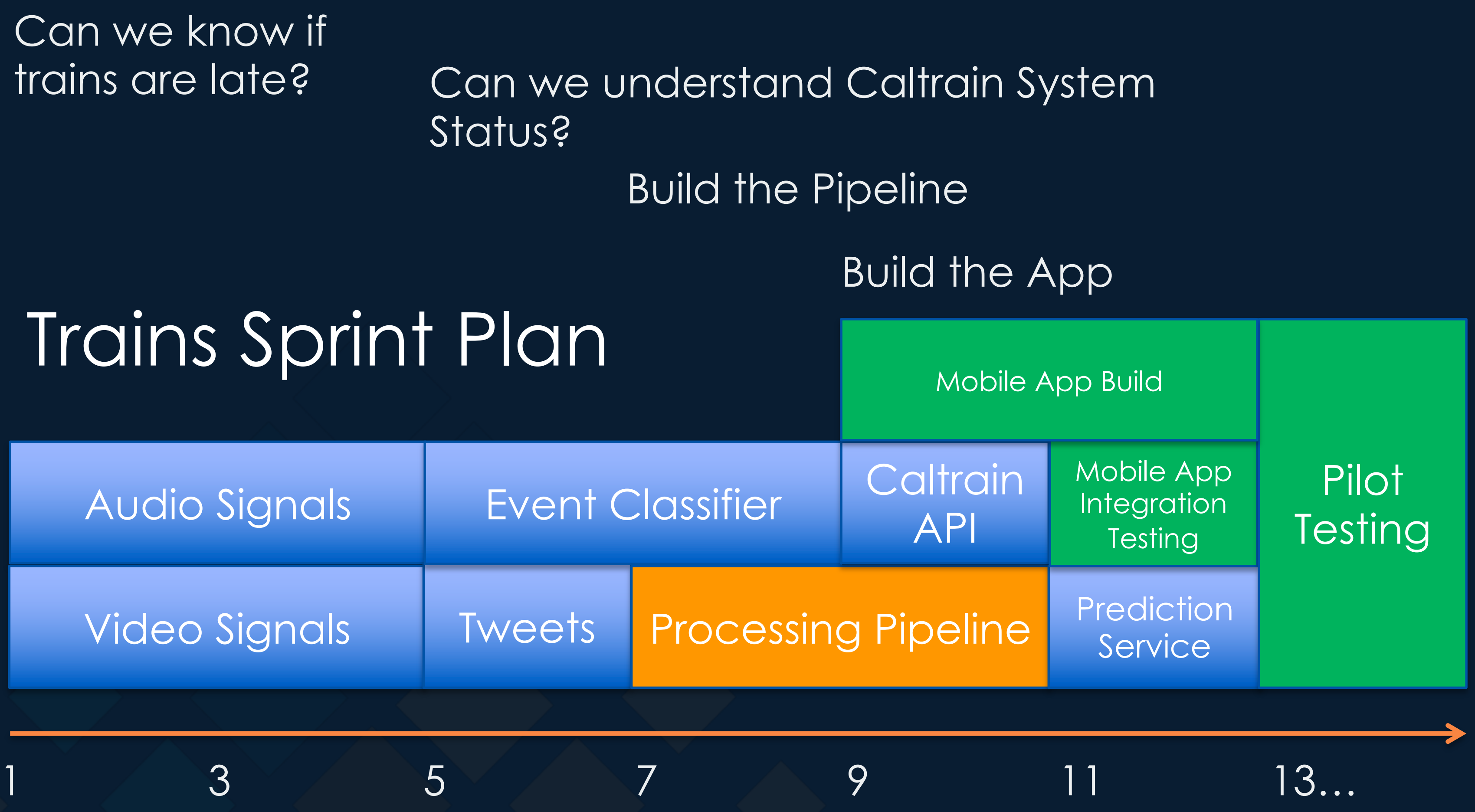
Caltrain Rider
CR
START
END
Burlingame
San Francisco
Local
Limited
Bullet
$5.25 one way
($0.50 less with Clipper, $10.50 day pass)
My Commute
Time to train
Scheduled times
Riding time
Riding this train
3:38PM - 4:04PM
#257 - Limited
26
Riding this train
4:07PM - 4:40PM
#159 - Local
33
40
4:53PM - 5:31PM
#263 - Limited
38
5:17PM - 5:43PM

- Commuter rail between San Francisco and San Mateo and Santa Clara counties ~30 stations
- 118 passenger cars
- 60% >=30 years old
- 2014 weekday ridership is 52,019 people
- *On-time performance is about 92%?*
- **No reliable real-time status information**
- **API outage between April 5th and June 2nd**



SPRINT PLANS

AGILE DOES NOT MEAN AN ABSENCE OF INTENT


Can we know if trains are late?
Can we understand Caltrain System Status?
Build the Pipeline
Build the App
Trains Sprint Plan
Mobile App Build
Audio Signals
Event Classifier
Caltrain API
Mobile App Integration Testing
Pilot Testing
Video Signals
Tweets
Processing Pipeline
Prediction Service
1
3
5
7
9
11
13…

Investigation Themes

INVESTIGATION THEME:

UNDERSTANDING SCHEDULE VARIANCE: HOW DO WE KNOW IF THE TRAIN IS LATE?

- Direct observation
  - We can hear the train horn
  - We can see the train when it goes by
- Purpose-built systems:
  - We can use Caltrain API's (when working)
- Other signals
  - We can check Twitter for delay info or rider comments

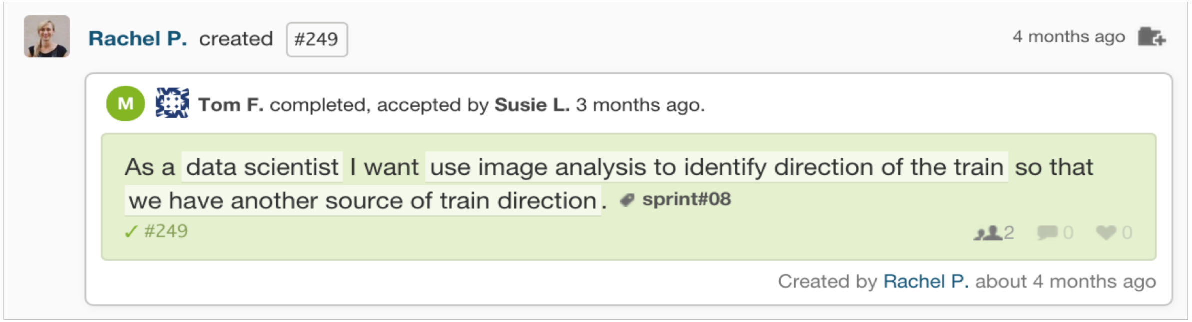
SPRINTS

EPIC HYPOTHESIS: WE CAN CLASSIFY A PASSING TRAIN INTO “LOCAL” OR “EXPRESS”

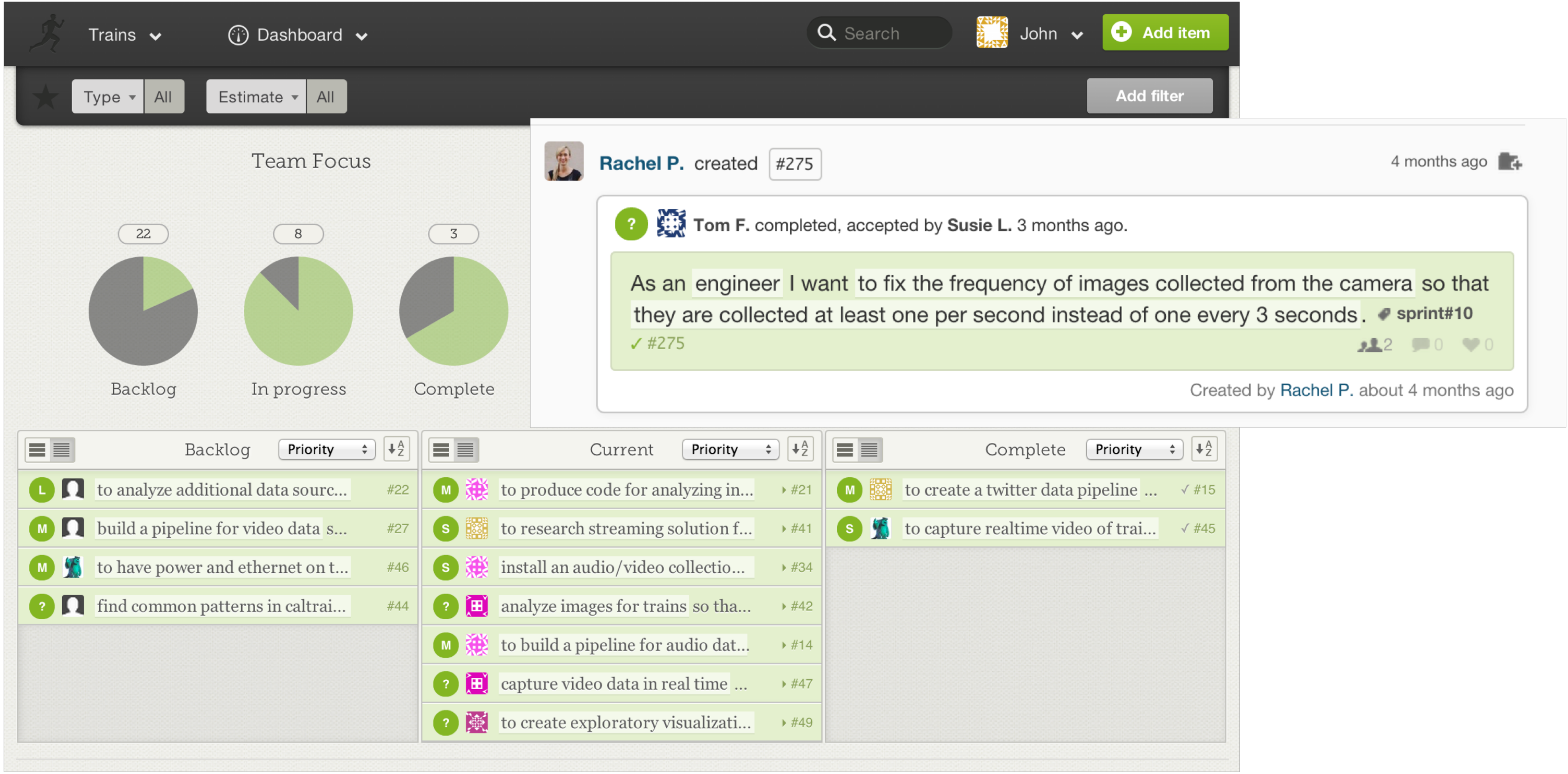
- Define a candidate approach and technical method:
  - I’m going to compare the beginning and ending fundamental frequencies of a sound to determine how fast a train is moving
  - I’m going to build a classifier based on that derived difference between starting and ending frequency to identify local vs. express—essentially trying to observe the Doppler effect on fast-moving express trains

STORIES & EPICS: WHICH WAY IS THAT TRAIN GOING, AND HOW FAST?

- **Stories** are the unit of work. They should identify activities that can reasonably be expected to be completed in a sprint.
- **Epics** are collections of stories that comprise all or part of an investigation theme



THE BACKLOG


Trains
Dashboard
Search
John
Add item
Type All
Estimate All
Add filter
Team Focus
22
8
3
Backlog
In progress
Complete
Rachel P. created #275
4 months ago
Tom F. completed, accepted by Susie L. 3 months ago.
As an engineer I want to fix the frequency of images collected from the camera so that they are collected at least one per second instead of one every 3 seconds. sprint#10
✓ #275
Created by Rachel P. about 4 months ago
Backlog Priority
to analyze additional data sourc... #22
build a pipeline for video data s... #27
to have power and ethernet on t... #46
find common patterns in caltrai... #44
Current Priority
to produce code for analyzing in... #21
to research streaming solution f... #41
install an audio/video collectio... #34
analyze images for trains so tha... #42
to build a pipeline for audio dat... #14
capture video data in real time ... #47
to create exploratory visualizati... #49
Complete Priority
to create a twitter data pipeline ... ✓ #15
to capture realtime video of trai... ✓ #45

So called out of a desire for **brevity** and **accountability**

DO:

- Collect your thoughts briefly in advance
- Keep it snappy!
- Stay on point:
  - Yesterday
  - Today
  - Blockers

DON'T:

- Prepare
- Try to solve problems on the spot
- Allow long discussions
- Get hung up on blame or responsibility for blockers
- Leave without understanding what everyone is working on

S	M	T	W	R	F	S
	t-2 Meeting Invites	*t*-1 s-1 Retro, Sprint Review	*t* Planning Kickoff 1	standup 2	standup 3	
	standup 4	standup 5	standup 6	standup 7	standup 8	
	Holiday	Standup Retro Prep 9	*t*+s Retro, Sprint Review 10			

example sprint

SPRINT REVIEWS & RETROSPECTIVES

SPRINT REVIEW AGENDA:

1. HIGHLIGHTS
2. REVIEW STORIES
3. DEMO
4. LESSONS LEARNED
5. RECOMMENDATIONS

- About Us section is complete
- Began exploratory analysis on Caltrain API/website
- Debugged the image script logs
- Updated UI for app

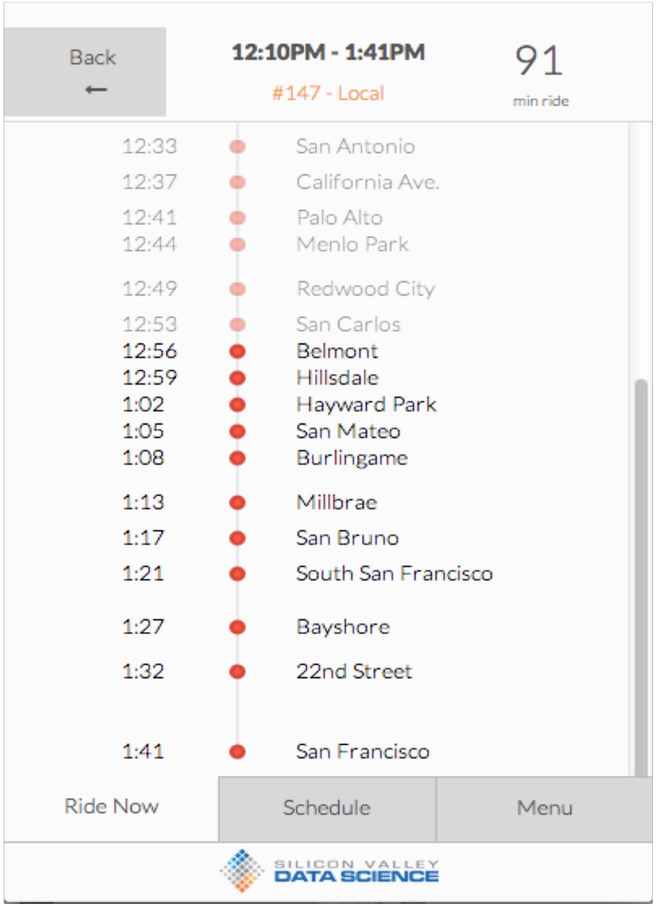
SPRINT REVIEW AGENDA:

1. HIGHLIGHTS
2. **REVIEW STORIES**
3. DEMO
4. LESSONS LEARNED
5. RECOMMENDATIONS

☑ / ☐	#	Story
☑	250	Write about section for app
☑	261	Update flume to grab rows with magic numbers from image logs
☑	260	Add magic number to the logs output from image script
☑	259	Basic evaluation platform from predicting ETAs
☑	262	Run PCA on Caltrain API data
☑	256	Trace through basic Caltrain day/examples
☑	248	Create graphics that will be used in the app
☑	252	Create how to use the app page
☑	270	Set up test flight
☑	274	Set up Android developer tools

SPRINT REVIEW AGENDA:

1. HIGHLIGHTS
2. REVIEW STORIES
3. DEMO
4. LESSONS LEARNED
5. RECOMMENDATIONS



RETROSPECTIVE AGENDA:

1. HIGHLIGHTS
2. REVIEW STORIES
3. DEMO
4. LESSONS LEARNED
5. RECOMMENDATIONS

- Hidden Markov Models are a pain in the neck
- Simple decision tree is better at classifying train direction and speed

RETROSPECTIVE AGENDA:

1. HIGHLIGHTS
2. REVIEW STORIES
3. DEMO
4. LESSONS LEARNED
5. **RECOMMENDATIONS**

We’re in a good spot to move forward with the basic classifier from combined video and audio pipelines

SUMMARY

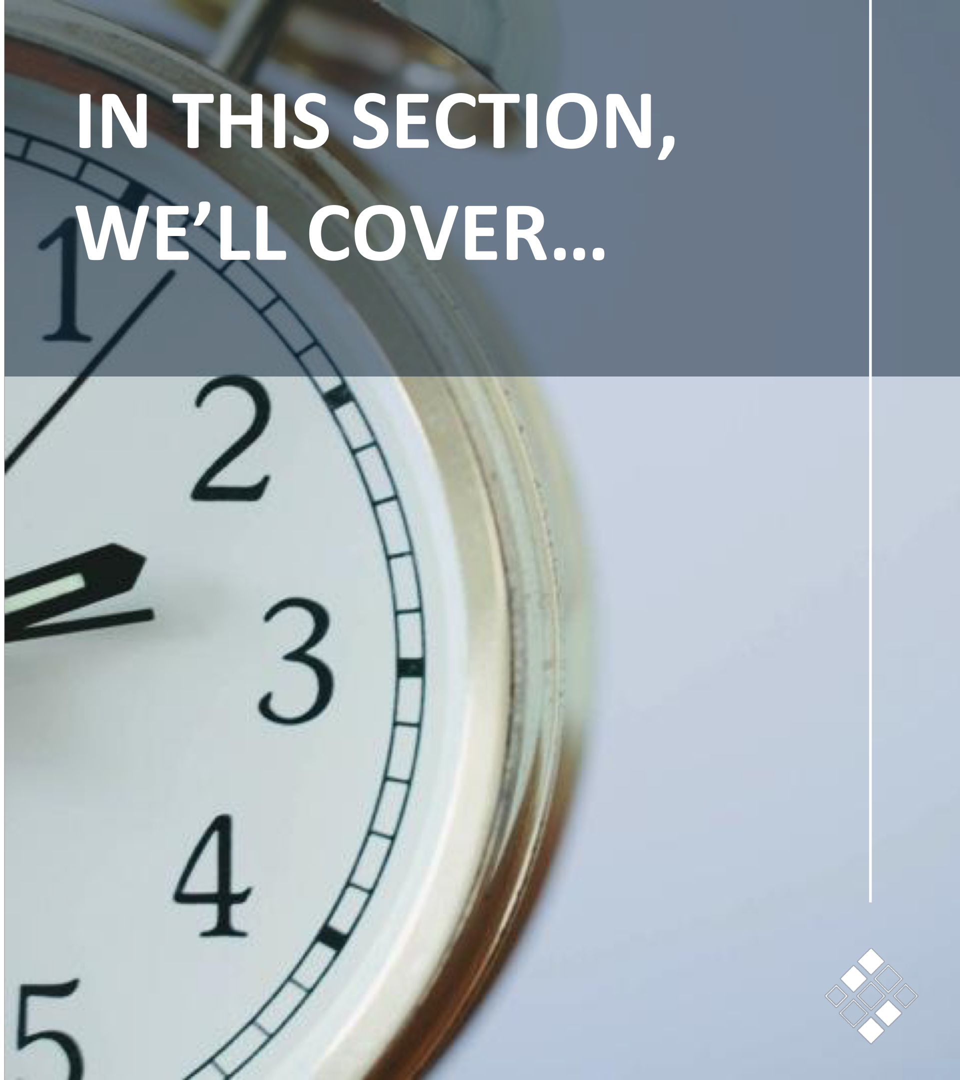
- Methods for *doing* data science != methods for running data science projects
- Running projects is about calling the shot, managing expectations, and being able to deliver as much value as possible
- We can successfully adapt scrum and other aspects of “agile” methods to our data science projects

FROM THE LAB

to the Factory

IN THIS SECTION, WE'LL COVER…

- A helpful framework
- Deploying data science
  - Insight deployment
  - Product deployment
- Regulatory complexity
- Operational complexity
- Model management
- Summary



What decisions do you make that, if informed by data and analysis, could more be more reliable or drive more valuable outcomes

Harvard Business Review

OPERATIONS

To Work with Data, You Need a Lab and a Factory

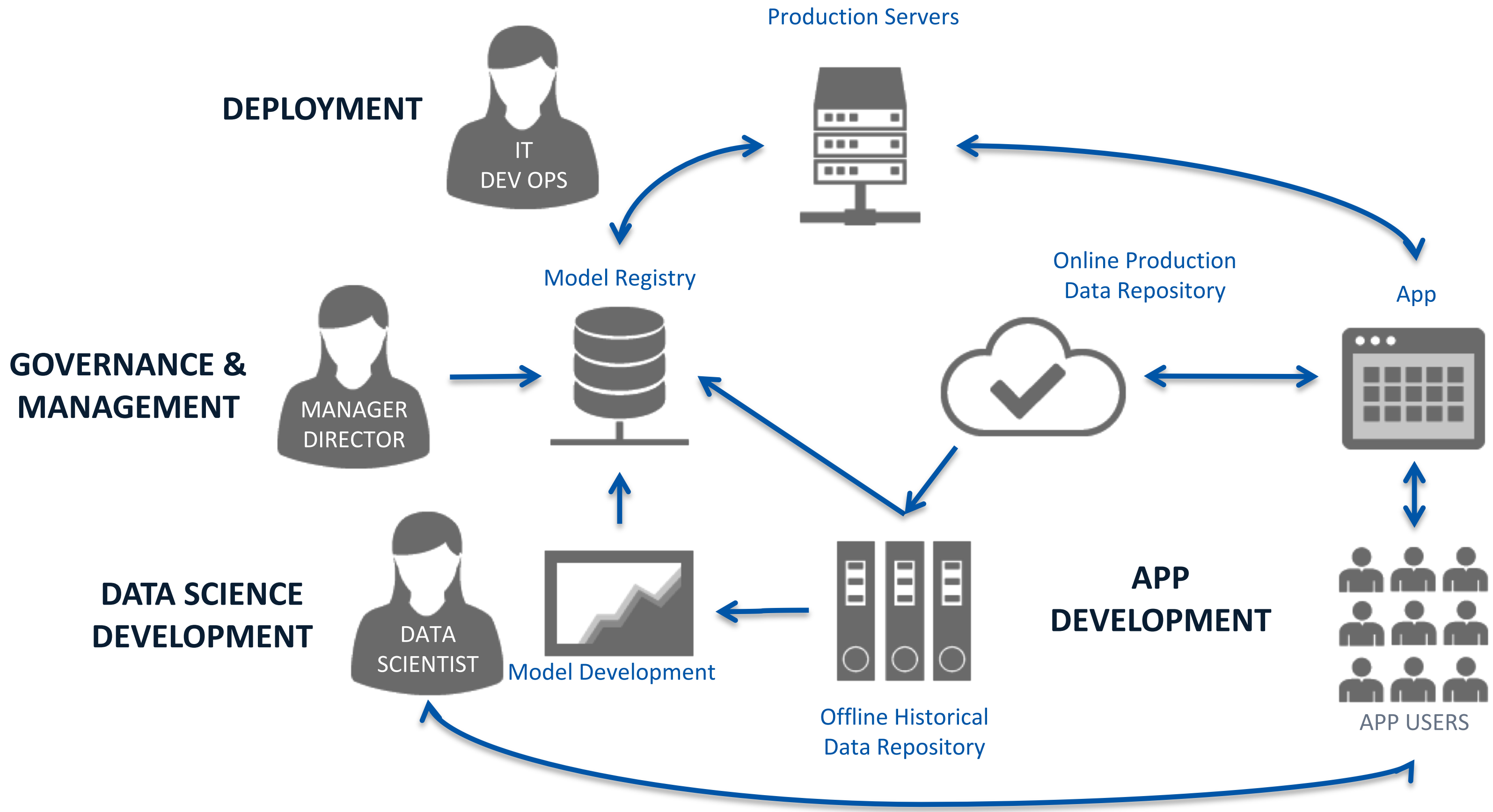
by Thomas C. Redman and Bill Sweeney

APRIL 24, 2013

SAVE | SHARE | 0 COMMENT | TEXT SIZE | PRINT

Companies that aim to score big over the long term with big data must do two very different things well. They must find interesting, novel, and useful insights about the real world in the data. And they must turn those insights into products and services, and deliver those products and services at a profit.

https://hbr.org/2013/04/two-departments-for-data-succe/


Production Servers
DEPLOYMENT
IT
DEV OPS
Model Registry
Online Production
Data Repository
App
GOVERNANCE &
MANAGEMENT
MANAGER
DIRECTOR
DATA SCIENCE
DEVELOPMENT
DATA
SCIENTIST
Model Development
Offline Historical
Data Repository
APP
DEVELOPMENT
APP USERS

THE LAB AND THE FACTORY

	Lab	Factory
People	Data scientists and engineers	Platform/data engineers
Process & Systems emphasizes	• Speed • Collaboration • Exploration • Reproducibility	• Stability & robustness • Governance • max(Value – Cost)

THE LAB AND THE FACTORY

The Hard Part: two functions in one platform

1. Hire the right people
2. Architect for agility
3. Integrated data/engineering culture

DATA INSIGHTS AND DATA PRODUCTS

- **Data insights** — “why?” and “what if?”
- **Data products** — rely on user and/or company data to carry out primary function
- On a spectrum rather than either/or

DATA INSIGHT DEPLOYMENT

Business question: “What is customer LTV by segment? What causes customers to leave?”

Data science project: Churn model by customer segment incorporating customer behavior

DATA INSIGHT DEPLOYMENT

	Requirements
People	Data-driven culture
Process	Agile product development Reproducible data science
Systems	A/B testing (experimentation) Robust data science tooling

DATA PRODUCT DEPLOYMENT

Business need: Identify and prevent churn by targeted retention offers to valuable customers.

Data product: On-demand customer churn prediction

REGULATORY COMPLEXITY

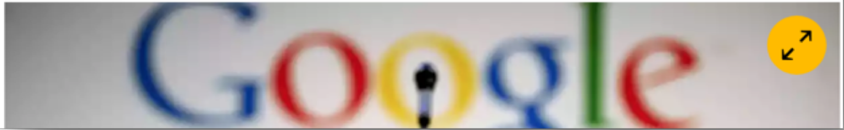
European parliament approves tougher data privacy rules

'Groundbreaking' changes strengthen EU privacy protections, enshrine right to be forgotten and give regulators wide-reaching powers

The Guardian, April 14, 2016

Google takes right to be forgotten battle to France's highest court

Company is appealing against decision by French data protection authority to apply search-results ruling to all its domains



The Guardian, May 19, 2016

Hard to Explain

The regulations prohibit any automated decision that "significantly affects" EU citizens. This includes techniques that evaluate a person's "performance at work, economic situation, health, personal preferences, interests, reliability, behavior, location, or movements." At the same time, the legislation provides what Goodman calls a "right to explanation." In other words, the rules give EU citizens the option of reviewing how a particular service made a particular algorithmic decision.

Wired, July 2016

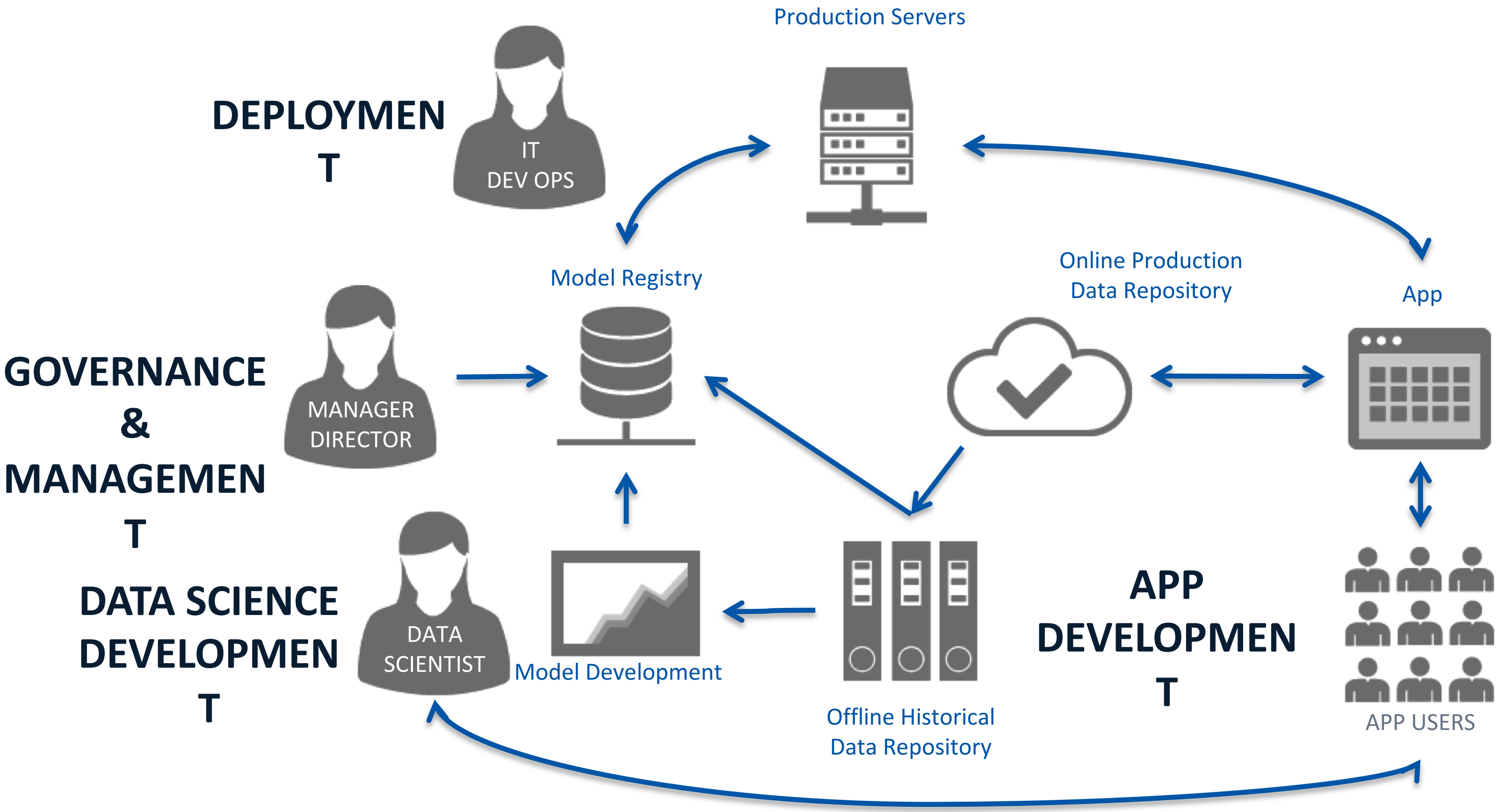
OPERATIONAL COMPLEXITY

- How do you know when your production model needs to be retrained?
- What happens when some data is no longer collected, or a new data source becomes available?
- How can you ensure the inputs themselves are valid, and your model is robust to violations of these assumptions?
- What happens when your production model breaks? Would you even know?

WHAT HAPPENS WHEN DATA PRODUCTS DEPEND ON EACH OTHER?

MODEL MANAGEMENT

Data Governance	Model Development	Model Deployment
Versioning	Versioning	Versioning
Lineage	Containerization	Lab and factory
Ownership	Automation	Auditability
Discoverability	Monitoring	Monitoring
Auditability	Auditability	Feedback loop


Production Servers
DEPLOYMEN
T
IT
DEV OPS
Model Registry
Online Production
Data Repository
App
GOVERNANCE
&
MANAGEMEN
T
MANAGER
DIRECTOR
DATA SCIENCE
DEVELOPMEN
T
DATA
SCIENTIST
Model Development
Offline Historical
Data Repository
APP
DEVELOPMEN
T
APP USERS

SUMMARY

- The Lab and the Factory are in fact not distinct, but inter-related
- There's also a spectrum between insights and products
- There are critical feedback loops in the lifecycle
- In order to productionalize your insights/products, you must understand and manage various complexities

THE EXPERIMENTAL ENTERPRISE

Data science allows us to observe our experiments and respond to the changing environment.

We need to both support investigative work and build a solid layer for production.

The foundation of the experimental enterprise focuses on making infrastructure readily accessible.


John Akred
@BigDataAnalysis