Overcoming the Five Dysfunctions of a Data Science Team

How to Foster Healthy Data Science Collaboration

Anaconda Enterprise Webinar

March 28, 2017

INTRODUCTION

Stephen Kearns - Product Marketing Manager at Continuum Analytics

Prior to Continuum Analytics, Steve was director for the portals & collaboration practice area of a boutique consulting firm in Toronto.

Steve holds an undergraduate degree in computer engineering from the University of Waterloo.

Data Science Collaboration

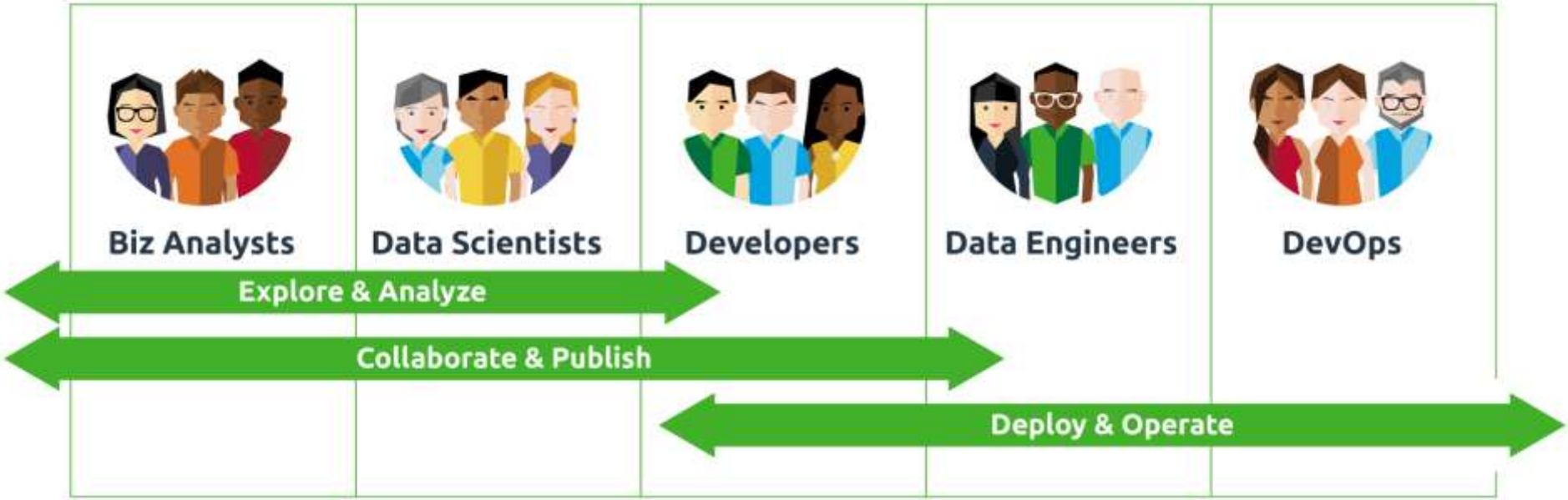
> *Data science is about exploring numbers to identify business challenges and identify solutions — initiatives that require **cross functional teams of storytellers**, programmers, statisticians, designers and accountants*

Ritika Puri

The Next Web Insider

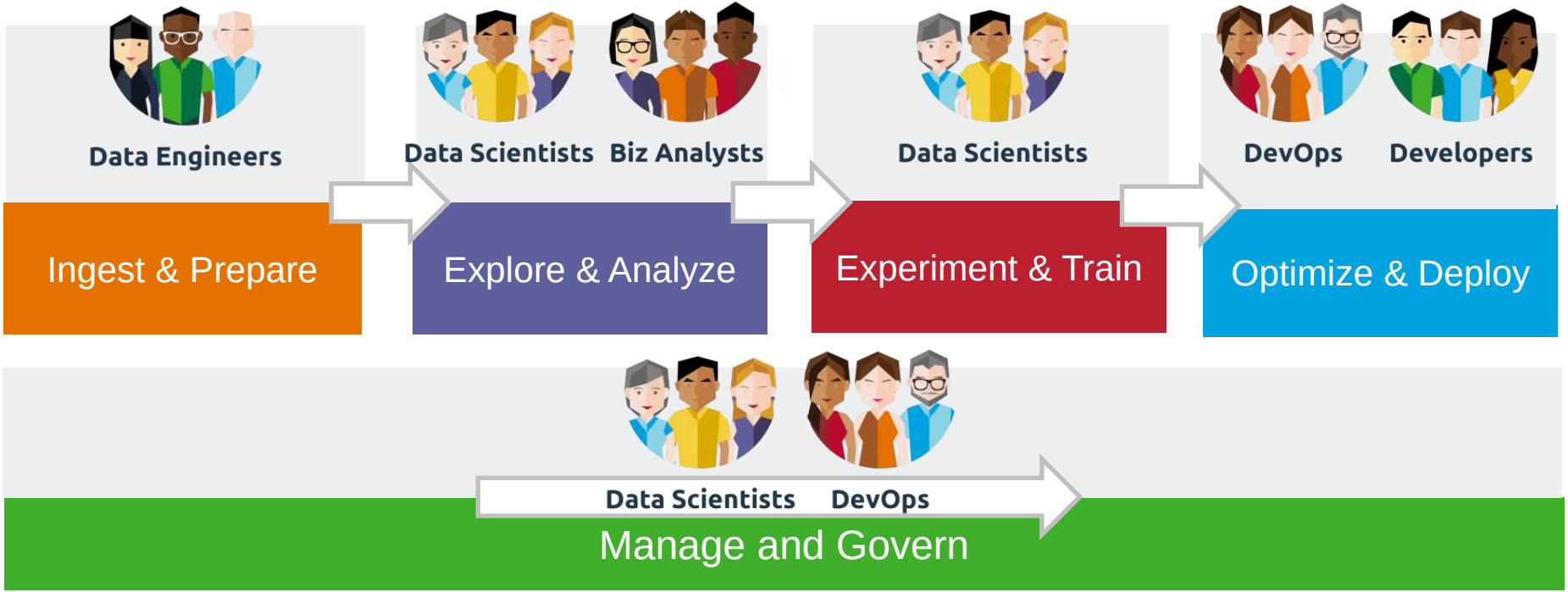
August 2015

DATA SCIENCE AS A TEAM SPORT



ENTERPRISE DATA SCIENCE

INHERENTLY A TEAM SPORT



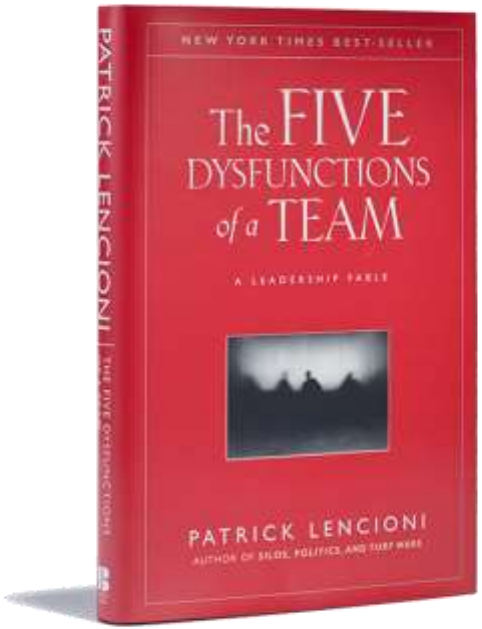
Views on Collaboration

> *We strongly believe that having **people** from different backgrounds **collaborating around a problem is more important** than selecting some fancy algorithms…*

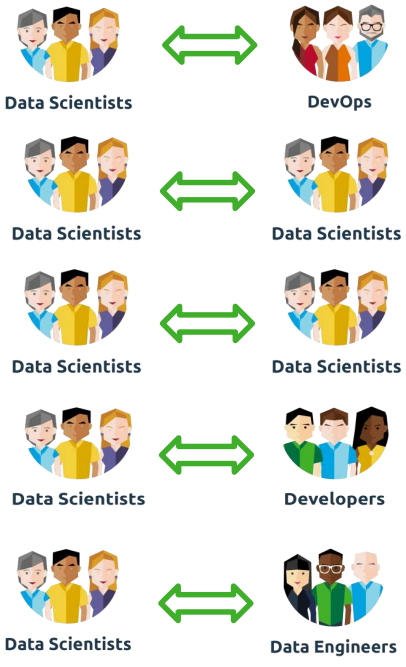
Steven Hillion
Wall Street Journal
March 2017

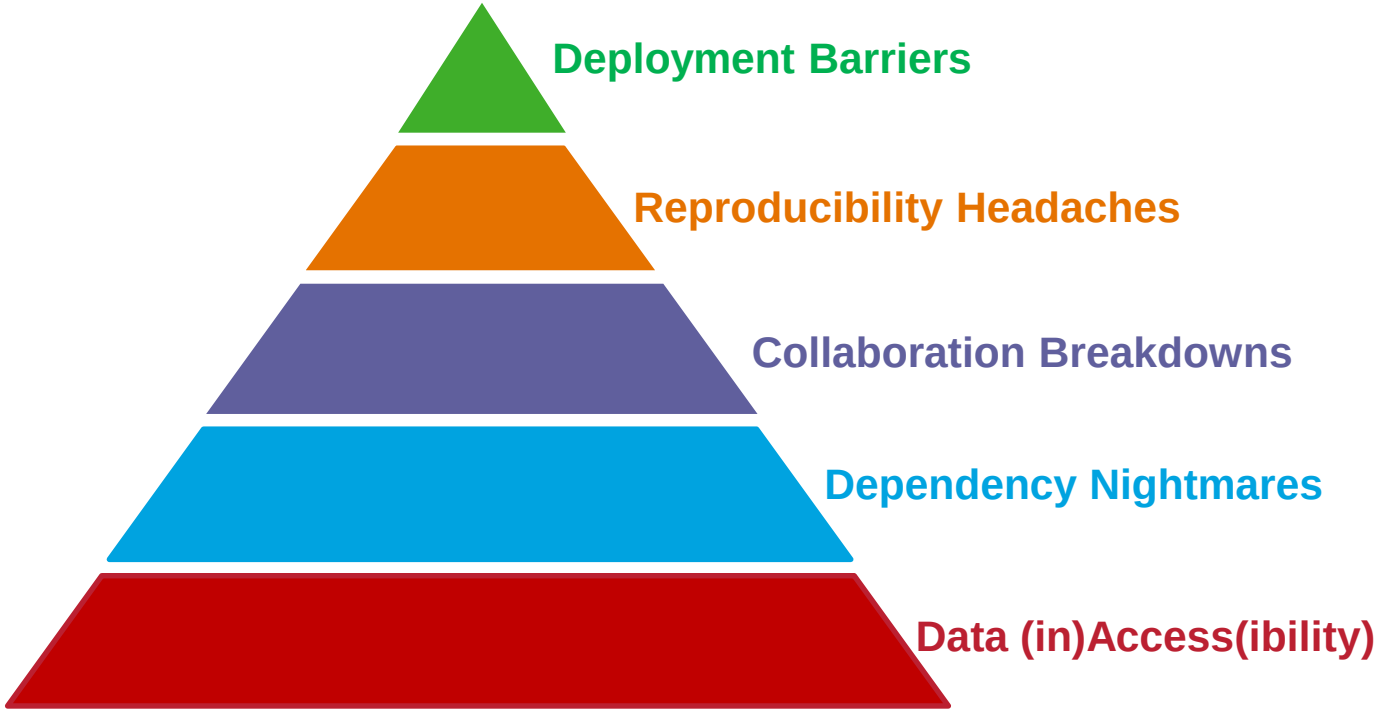
LENCIONI'S FIVE DYSFUNCTIONS©

Table Group



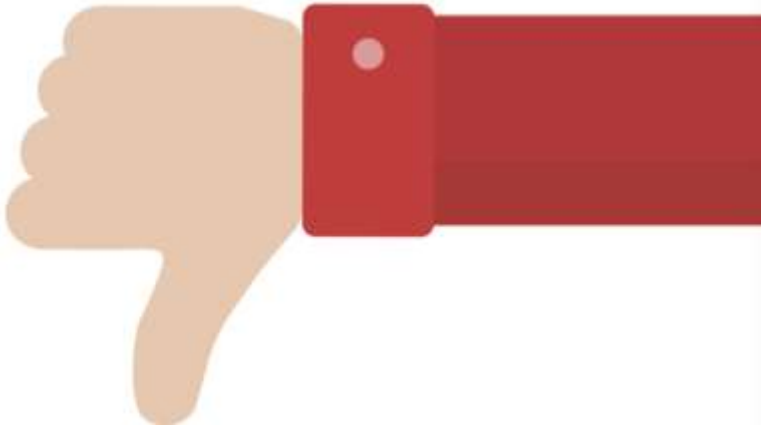
FIVE DYSFUNCTIONS OF A DATA SCIENCE TEAM





THESE ARE REAL PROBLEMS…

The cost to dysfunctional data science teams



Duplication of effort
- Searchability & discoverability
- Lack of reuse

Team frustrations
- Onboarding costs
- Synchronization
- Versioning

Governance and compliance issues
- Data loss
- Data access challenges

BUT WHEN WE SOLVE THESE DYSFUNCTIONS

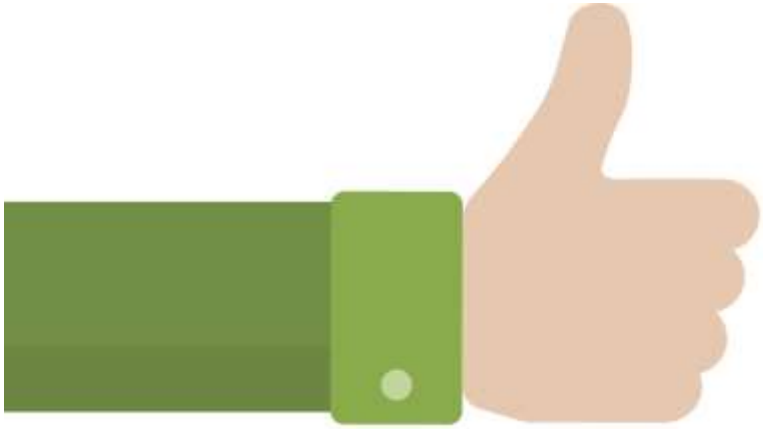
Benefits of healthy collaboration

Increased data science **workflow velocity**

Increased **retention** with **happier** data science teams

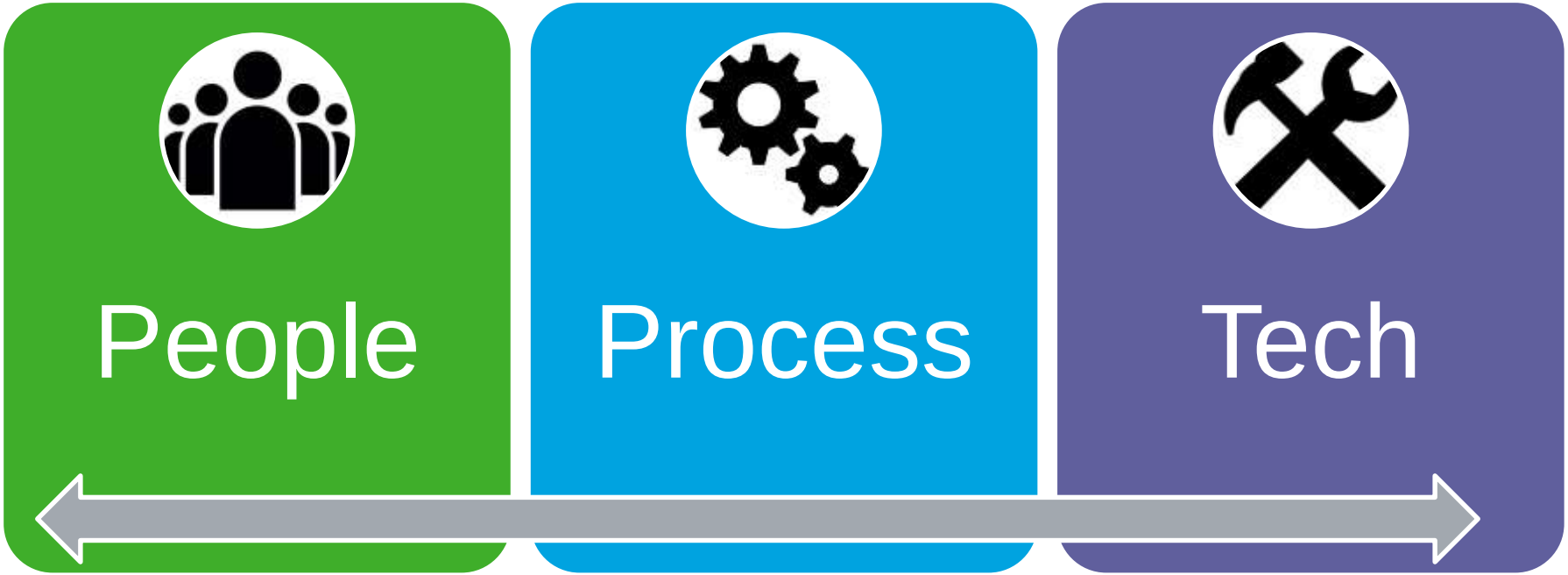
- Streamlined sharing reduces communication overhead
- Less friction accessing sensitive data
- Teams are automatically synchronized (esp. geo-dispersed)
- Centralized hosting, discovery and search empower reuse
- Easier to ramp up on new projects and packages

Improved **compliance and governance**



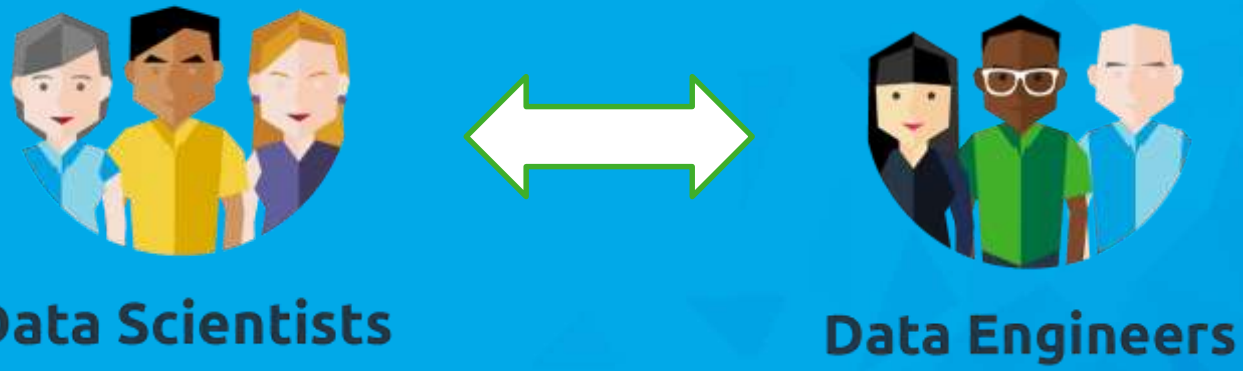
REQUIREMENTS FRAMEWORK

Breaking the challenges down



DYSFUNCTION #1

Data (in)Access(ibility)



DATA (IN)ACCESS(IBILITY)

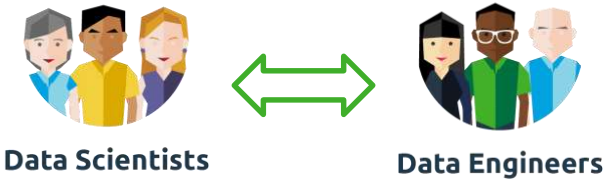
Trusting the data sources

How it affects collaboration

- Assumptions made by the data engineer can be silently impacting the results prepared by the data scientist
- Data scientists mistrust the data, and the data engineer without either having direct access or some way of viewing the provenance of the transformed data
- Using different tools reduces visibility of what's been done to the data

DATA (IN)ACCESS(IBILITY)

Connecting to diverse sets of data

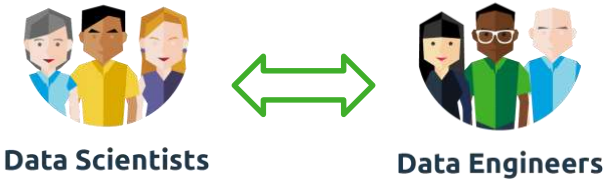


Data Science Team Requirements

- **People**: Data engineers who relentlessly question/expose their own assumptions about the data; willing to write their ingest and ETL scripts in a language the data scientists use can build trust
- **Tech**: Which language ecosystem is going to facilitate data access across many data sources in the fastest, easiest way, and has the critical mass to be ubiquitous?

DATA (IN)ACCESS(IBILITY)

Laborious process to connect to diverse sets of data

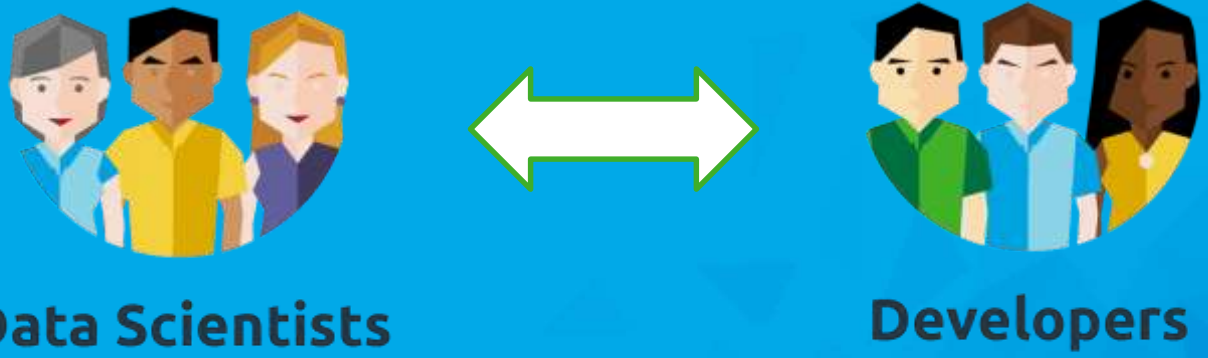


How people are solving it today

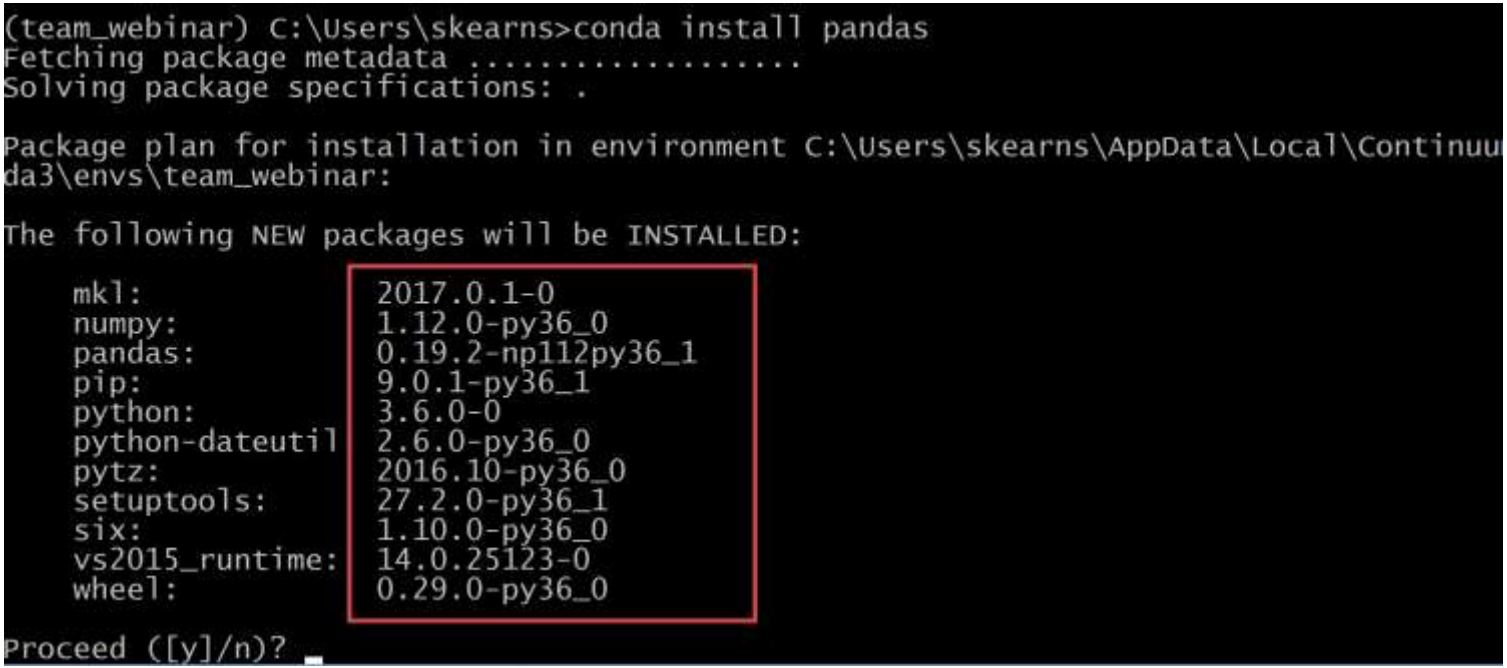
- **Languages**: Python, R, Java, C/C++
- **Open Source:** Pandas, NumPy, Blaze, others
- **Enterprise Ready:** Anaconda Enterprise

DYSFUNCTION #2

Dependency Nightmares



DEPENDENCY NIGHTMARES



Down the rabbit hole we go…

DEPENDENCY NIGHTMARES

Large effort to set up and maintain list of required packages

How it affects collaboration

- Extra time spent curating packages, specific versions, and handling updates
- Teams trying to collaborate across different projects may have different dependency needs which can cause conflicts when they go to share tools and packages

DEPENDENCY NIGHTMARES

Large effort to set up and maintain list of required packages

Data Science Team Requirements

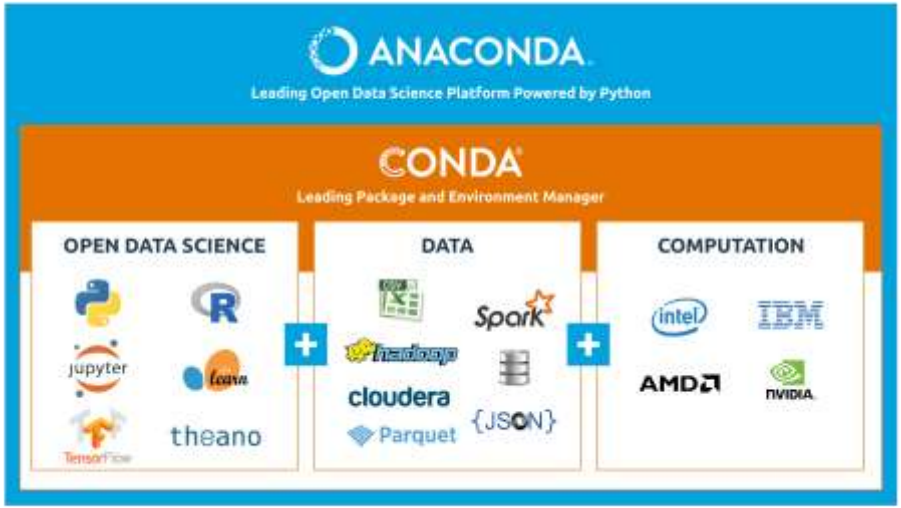
- **Process**: Sourcing packages & mirroring (inc. behind firewall & airgap), whitelisting/blacklisting, publishing new packages
- **Tech**: Support for multiple languages AND multi-operating system support; strong dependency solver; integration with source control, CI/CD

DEPENDENCY NIGHTMARES

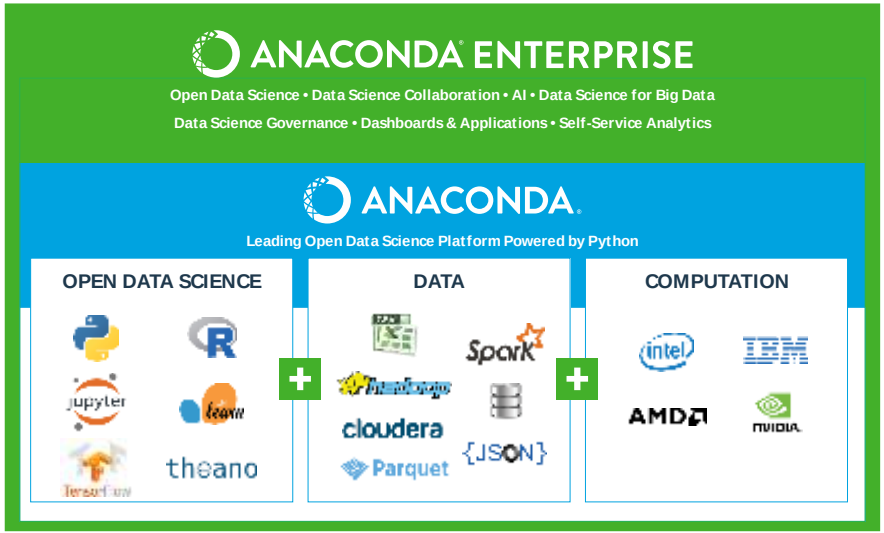
Package Management

How people are solving it today

- **Open Source:** conda, Pip (Python), install.packages (R), Maven (Java), Composer (PHP)
- **Enterprise Ready:** Anaconda Enterprise



- Intelligent dependency management
- Cross-platform, cross-language



DYSFUNCTION #3

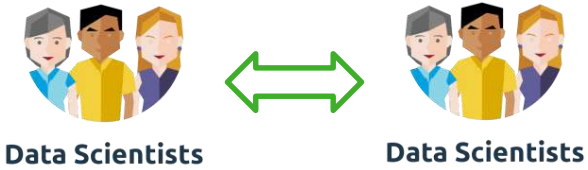
Collaboration Breakdowns

Data Scientists

Data Scientists


Email: Where Information Goes to Die

COLLABORATION BREAKDOWNS

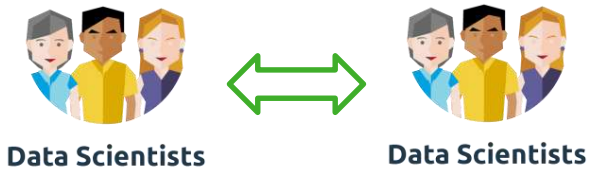


Email is where information goes to die

How it affects collaboration

- Recipients-only → not discoverable → duplication of work
- Terrible version control system → working with incorrect information

COLLABORATION BREAKDOWNS

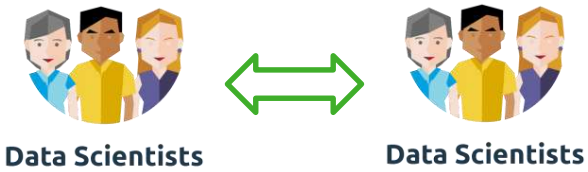


Email is where information goes to die

Data science team requirements

- **People:** value information sharing
- **Process:** minimize collaboration tax – simple, close to working local as possible
- **Tech:** Security and permissions, discoverable, searchable, seamless integration working from local workstation to hosted collaboration environment to cluster – has to work anywhere your data lives

COLLABORATION BREAKDOWNS



Email is where information goes to die

How people are solving it today

- **Open Source:** JupyterHub, Git/Github
- **Enterprise Ready:** Anaconda Enterprise Notebooks

CUSTOMER PERSPECTIVE

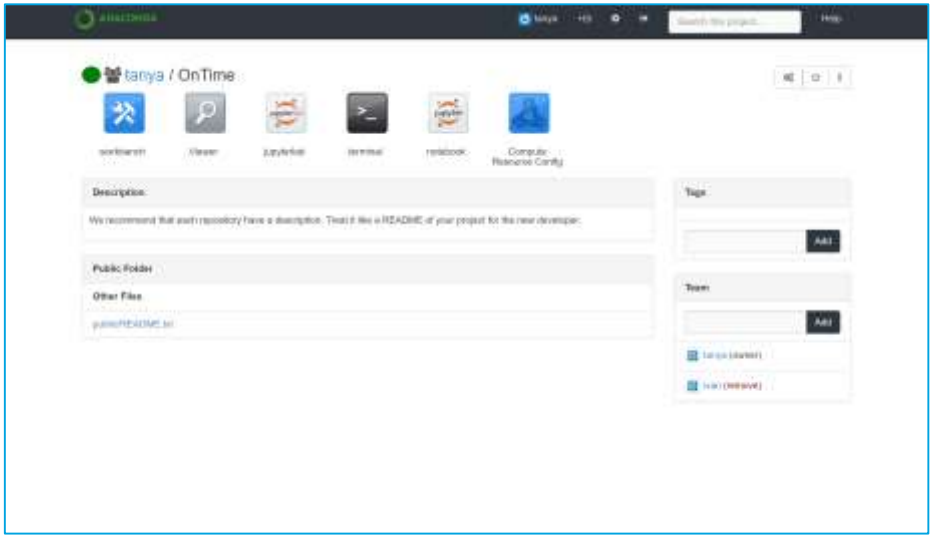
Girish Modgil

Senior Director of Data & Analytics

General Electric

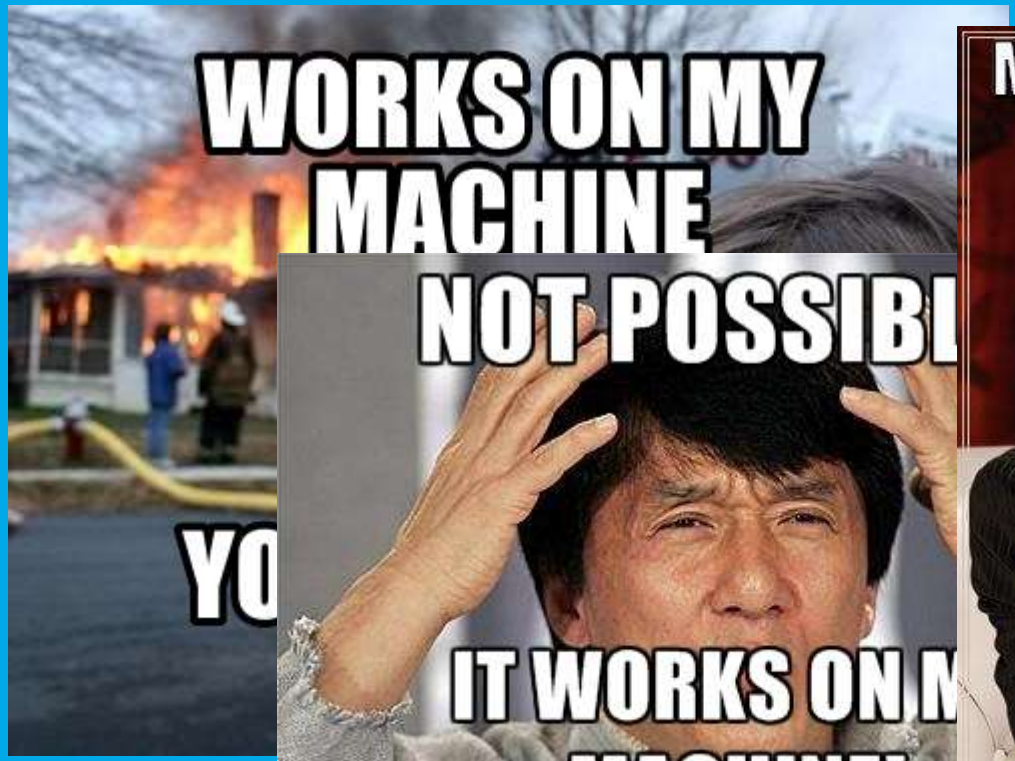
ANACONDA ENTERPRISE DEMO

- Shared notebooks
- Locking
- Versions & differences

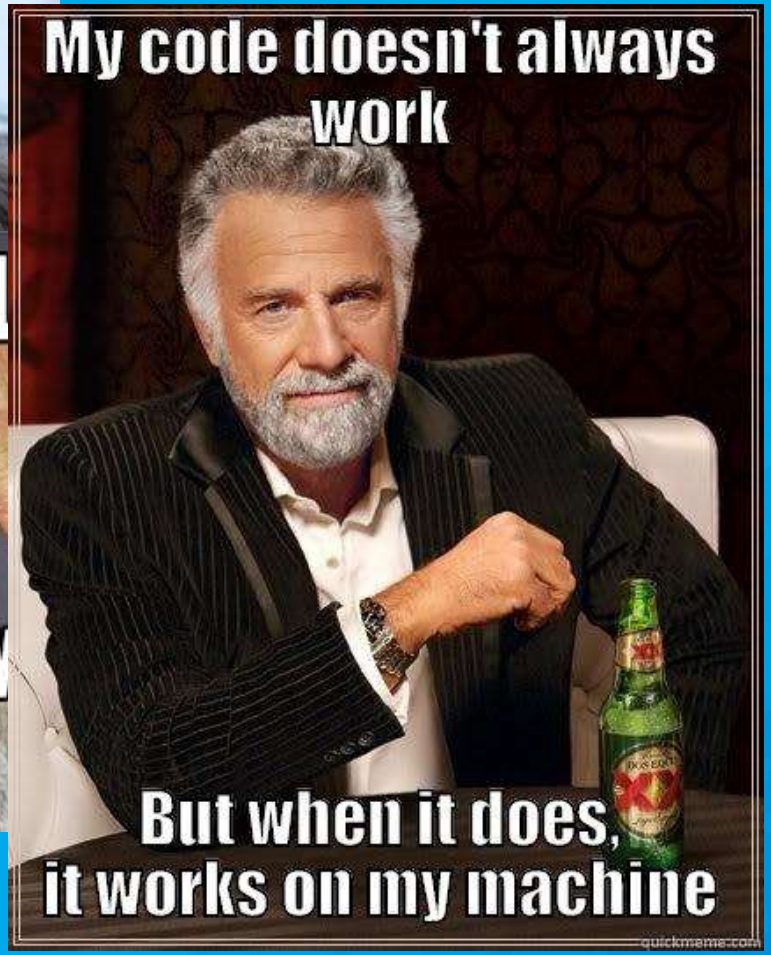


DYSFUNCTION #4

REPRODUCIBILITY HEADACHES



Data Scientists



REPRODUCIBILITY HEADACHES

"Works on my machine!?!"

How it affects collaboration

- Slows exchange of ideas as time is spent reconstructing exact source environments
- Other environment constraints may prevent you from running a new project at all in your current environment which then triggers the requirement for a whole new environment

REPRODUCIBILITY HEADACHES

"Works on my machine!?!"

Data science team requirements

- **Tech:** Isolation, lightweight, cross-platform, easy-to-use

REPRODUCIBILITY HEADACHES

"Works on my machine!?!"

How people are solving it today

- **Open Source:** Anaconda Project (see Blog Post), conda (w/ conda Environments), VirtualBox (VM), Docker
- **Enterprise Ready:** Anaconda Enterprise

DYSFUNCTION #5

Deployment Barriers

Data Scientists

DevOps

DEPLOYMENT BARRIERS

"You can't put THAT into prod!?!"

How it affects collaboration

- Data Scientists and IT/DevOps have very different areas focus; models aren't often built to operationalize in existing production systems

DEPLOYMENT BARRIERS

"You can't put THAT into prod!?!"

Data science team requirements

- **People**: Ability to balance other needs in the system
- **Process**: Sketch out the "road to production" blending stability and speed; consider enterprise support requirements – security for open source, assurance, indemnification
- **Tech**: Ability to run original data scientist code in production at required performance levels; APIs and interoperate between applications and projects; also security, scalability, high-availability and version control → alleviate these infrastructure concerns from the Data Scientist

- This blog post provides an in-depth review of processes and tools required to deploy enterprise data science and meet the stringent requirements of enterprise DevOps

DEPLOYMENT BARRIERS

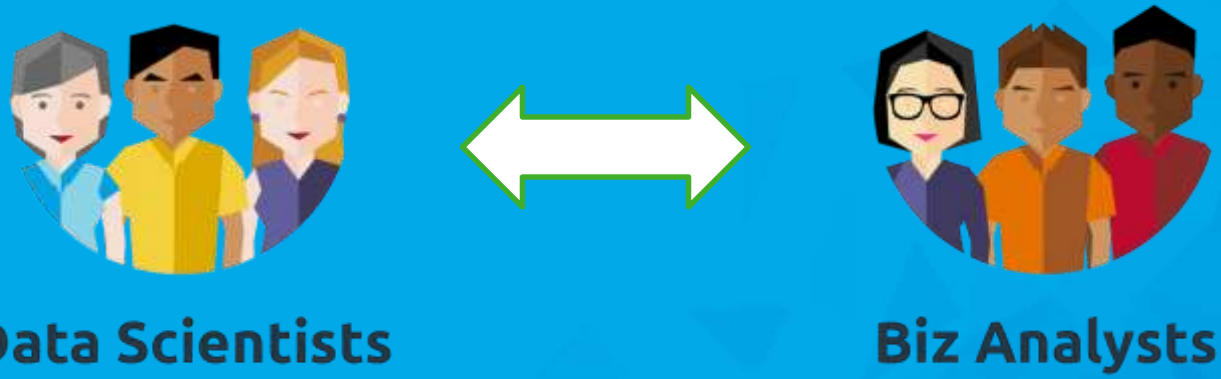
"You can't put THAT into prod!?!"

How people are solving it today

- Convert from R or Python into Java, C/C++ or .Net
- **Open Source:** conda, Java, C/C++, .Net
- **Enterprise Ready:** Anaconda Enterprise (Join Innovators Program Now)

BONUS - DYSFUNCTION #6

Excel as Data Science Tool?




What Excel Looks Like
to a Data Scientist

EXCEL® AS A DATA SCIENCE TOOL

Bridging the divide between Business Analysts and Data Scientists

How it affects collaboration

- No common ground to share tools and exchange information
- Broken workflow

EXCEL® AS A DATA SCIENCE TOOL

Bridging the divide between Business Analysts and Data Scientists

Data science team requirements

- **Process:** Allow Data Scientist to use Data Science tools, don't make Business Analysts have to learn a programming language
- **Tech:** bridge the gap to provide Business Analyst support for basic use cases as well as machine learning, visualization and connecting to Big Data from inside Excel

EXCEL® AS A DATA SCIENCE TOOL

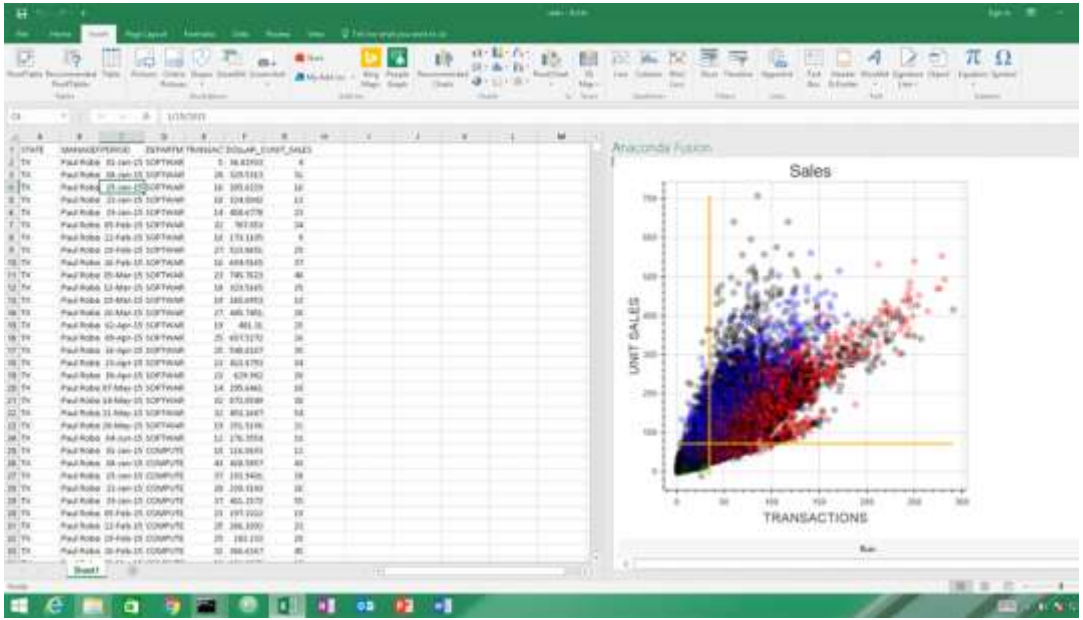
Bridging the divide between Business Analysts and Data Scientists

How people are solving it today

- **Open Source:** XLWings
- **Enterprise Ready:** Anaconda Fusion

ANACONDA FUSION DEMO

- Pull data from data source and querying
- Machine learning
- Visualization example



TAKEAWAYS

Optimizing for collaboration can increase data science workflow velocity

People: Team members who value collaboration

Process: Minimize the collaboration tax

Tech: A platform that enables healthy data science collaboration

Result: engaged, productive teams,
working towards their shared goals

Next Steps

DOWNLOAD Breaking Data Science Open eBook
Go.continuum.io/download-ebook-breaking-data-science-open

CHECKOUT Anaconda Fusion Trial
Go.continuum.io/2017-anaconda-fusion-trial

EXPERIENCE Anaconda Enterprise on your own
Know.continuum.io/Anaconda-Enterprise-Test-Drive.html

Q&A