The 3 Key Barriers Keeping Companies from Acting Upon the Possibilities That Big Data has to Offer
A little bit about me…

• Born & raised in Palo Alto, California
• BA in European History From Columbia University
• Masters in Marketing & Communication from Sciences Po Paris
• Director of Marketing at Dataiku
• Currently living in Paris
Shift from
Uomo Universale

Excel at all things:
• Intellect
• Mathematics
• Science
• Art
• Social
• Physical

« A man can do all things if he will. »
-Leon Battista Alberti (1404-72)
Shift from *Uomo Universale* To Expert
Data Science Superstar

Required Assets:
• Hacker mindset
• Logic
• Statistics
• Polyglot Programmer
• Mathematics
• Algorithmics
• Engineering
• Databases
• Machine Learning
• Strong creativity
• Strategical thinker
• Business understanding
• Strong communication skills
• Project management
Are we in some sort of Renaissance Era of Big Data?
If so, what’s next?
Investigation Part 1: What is a Data Product?
Data Products
Data Products
Data Products
Data Products
Data Products
Data Products
Data Products
Data Products
Data Products
Data Products
Data Products

= 

Data + Technology + *Data Scientist*? + End User
Investigation Part 2: What goes on behind the scenes?
Building a Data Product

Stream / Real-time → Preprocessing → Query → Data → User Interface
Building a (Predictive) Data Product

Historical Data → Preprocessing → Machine Learning → Model → Predicted Data
Pre-processing & cleaning data alone can take up to 80% of the time spent on a data project.
Building a (Predictive) Data Product

Historical Data → Preprocessing → Machine Learning → Model → Predicted Data
Building a (Predictive) Data Product
Running a (Predictive) Data Product
Investigation Part 3: Who does what?
Data Product = Business Incentive
Data Product = Business Incentive

Mathematics / statistics / Business
Data Product = Business Incentive
Data Engineers

Data Scientist

mongoDB
TERADATA
VERTICA
Spark
MLlib
Greenplum
amazon web services

«Data Scientist»
Data Engineers
What I’ve Learned
Fact 1: The Skill Sets Exist

- **Build**
  - Business
  - Statistics
  - Math

- **Maintain**
  - Data Engineering
Fact 1: The Skill Sets Exist (& your company probably already has them)

- Business Analyst
- Data Engineer
- Mathematician / Statistician
- Build
- Maintain

(dataiku)
Fact 2: The Technologies Exist
(and some are free!)
Fact 3: The Data Exists

WE’VE DECIDED TO TAKE BIG DATA TO THE NEXT LEVEL...

HUMONGOUS DATA

© D.Fletcher for CloudTweaks.com
Why is Production & Industrialisation of (Predictive) Data Products Important?
Those who win are those who deliver new data products continuously.
Those who win are those who deliver data products

Data products are supposed to deliver business value... if you don’t deploy them, where’s the long term value?
Those who win are those who deliver data products.

No Industrialisation = Limited ROI

It’s like building your dream house but never moving in.
No Industrialisation = Limited ROI

Those who win are those who deliver data products

It’s like building your dream house but never moving in. Absurd!
So Why Aren’t More Companies Deploying (Predictive) Data Products?
A Data Product must be business focused (ROI) & mathematically accurate (RELIABLE)
1° Business Analytic & Algorithmic Minds Are Different…
1° Business Analytic & Algorithmic Minds Are Different…

The Business Analysts Brain

Patterns. Patterns. Patterns.
1° Business Analytic & Algorithmic Minds Are Different…

...But Your Data Product Needs Both

1° Business Analytic & Algorithmic Minds Are Different…

Patterns, patterns, patterns
Performance, Truth, Anomaly
MINDSET
• Project alignment from conception to execution – install team mindset with common goal – even if the paths to get there are different

FRAMEWORK
• One common platform with enough flexibility for both mindsets to fully exercise their individual skill and expertise on a common project
2° So Many Technologies, Languages, and Needs
2° So Many Technologies, Languages, and Needs
2° So Many Technologies, Languages, and Needs
2° So Many Technologies, Languages, and Needs

- Code-free
- SQL
- R / Python
2° So Many Technologies, Languages, and Needs

- Code-free
- SQL
- Hadoop
- R / Python
…Or As My Boss Calls It: **Technoslavia**

Florian Douetteau  
Dataiku CEO
…Or As My Boss Calls It: **Technoslavia**
…Or As My Boss Calls It: **Technoslavia**

Florian Douetteau
Dataiku CEO
…Or As My Boss Calls It: Technoslavia

Florian Douetteau
Dataiku CEO
Tip!

Living in Harmony with Technoslavia

OPTION 1:
Enterprise dictatorship
Tip!

Living in Harmony with Technoslavia

OPTION 2 (my personal favorite):
Accept a polyglot approach
3° Production and Industrialisation is Complex
3° Production and Industrialisation is Complex

Technological Complexity

Data reliability is hard to guarantee
3° Production and Industrialisation is Complex

Technological Complexity

The assumption that production will identically reproduce the analysis phase is a hard promise to make.
Technological Complexity

Monitoring a predictive model’s life cycle is a tedious and continuous task
3rd Production and Industrialisation is Complex

Human & Organisational Complexity

BUILDING ≠ MAINTAINING
3° Production and Industrialisation is Complex

Human & Organisational Complexity

BUILDING ≠ MAINTAINING

Business Analyst

• patterns
3° Production and Industrialisation is Complex

Human & Organisational Complexity

**BUILDING**

Business Analyst / Algorithmic
- patterns
- performance
- truth

≠

**MAINTAINING**
3° Production and Industrialisation is Complex

Human & Organisational Complexity

Building

- Business Analyst / Algorithmic
  - patterns
  - performance
  - truth

MAINTAINING

- Data engineers
  - stability
  - reliability
  - cost of ownership
Human & Organisational Complexity

BUILDING

- Business Analyst / Algorithmic
  - patterns
  - performance
  - truth

MAINTAINING

- Data engineers
  - stability
  - reliability
  - cost of ownership
Tip!

Making Complexity Work for You

TIP #1: Invest in a platform where development and production are the same
TIP #2: Invest in monitoring capabilities & strategies
Tip!

Making Complexity Work for You

TIP #3: Name your Data Engineer(s) Wisely & Define Responsibilities

Data Engineers must have visibility and understanding of the key business metrics
Making Complexity Work for You

TIP #3: Name your Data Engineer(s) Wisely & Define Responsibilities

Data Engineers must know if (and when) a model is diverging
TIP #3: Name your Data Engineer(s) Wisely & Define Responsibilities

Data Engineer must be responsible for quality of service
Making Complexity Work for You

TIP #4: It’s Not a One Man Show

Differentiate builders of new data products from those that maintain them.
What To Expect?
From the Renaissance of Big Data…

Where the Data Science Superstar is one person that excels at all skill sets…

…and where actual data products are rarely deployed and maintained
To the Enlightenment of (Big) Data

Where the Data Science Superstar is a team of complimentary skill sets...

... and where data products are designed, built, tested, and deployed by a team of skilled individuals that each have a distinct role.
SPOTLIGHT on the
Data Science Team Manager
The Data Science Team Manager must understand the stakeholders’ needs, translate them into a business need that can be answered with a data product…
The Rise of the Data Science Team Manager

The Data Science Team Manager must permit and enable collaboration between business analysts, statisticians, & engineers...

Collaboratively design, build, & deploy Data Products
The Rise of the
Collaborative Data Science Team

…all the while maintaining the distinction between each individual role and each individual skill set.
The Secret to Building and Industrializing Data Products is Collaboration.

Today, collaboration between different skill sets, technologies, and data is finally possible.
Data Science Studio: One Platform for Development and Industrialization
Thank You!

Pauline Brown
Dataiku, Director of Marketing
Pauline.brown@dataiku.com
@pauline8brown
www.dataiku.com