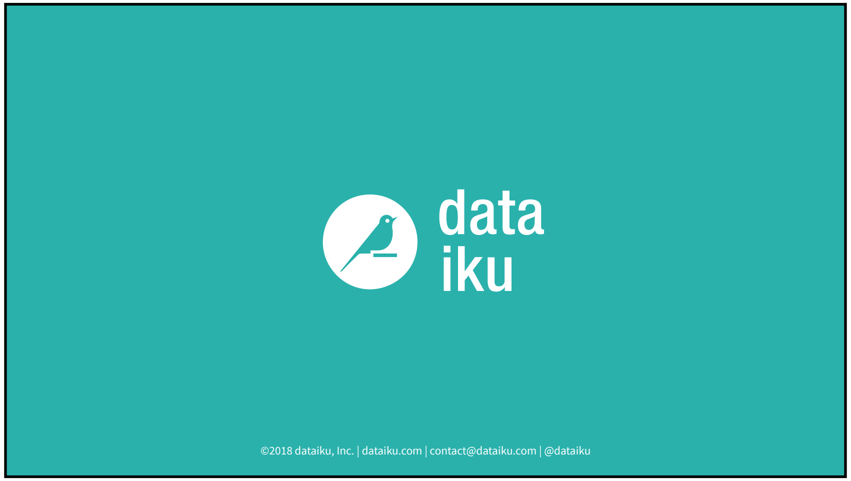
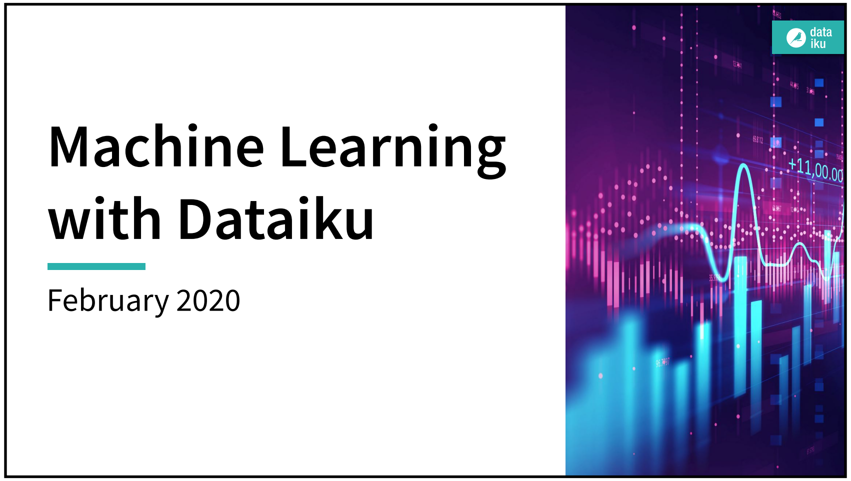




Machine Learning with Dataiku

February 2020

Agenda

- Sign attendance sheet
- Dataiku introduction
- Demo Time
- Start Building your first project
- Q&A
- Lunch
- Deep Learning with Dataiku
- Getting Certified on Dataiku DSS

About Us

data iku

Key Company Milestones

A Brief History of Dataiku

2013 2014 2015 2016 2017 2018

COMPANY

Dataiku Created

20th employee

Office in New York

$4M raised

$14M Series A

Office in London

$28M Series B

200th employee

$100M Series C

CUSTOMERS

First customer

First 10+
users customers

50th
customer

First 100+
users customers

EGG
CONFERENCE

100th
customer

First 500+
users customers

200th
customer

1st CAB
Customer Advisory Board

SOFTWARE

Beta

version 1
1st tool integrating data prep and ML

version 2
Real-time collaboration Spark integration

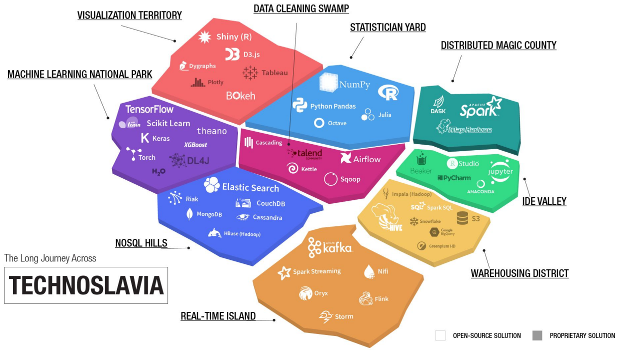
version 3
Scenarios, automation and version control

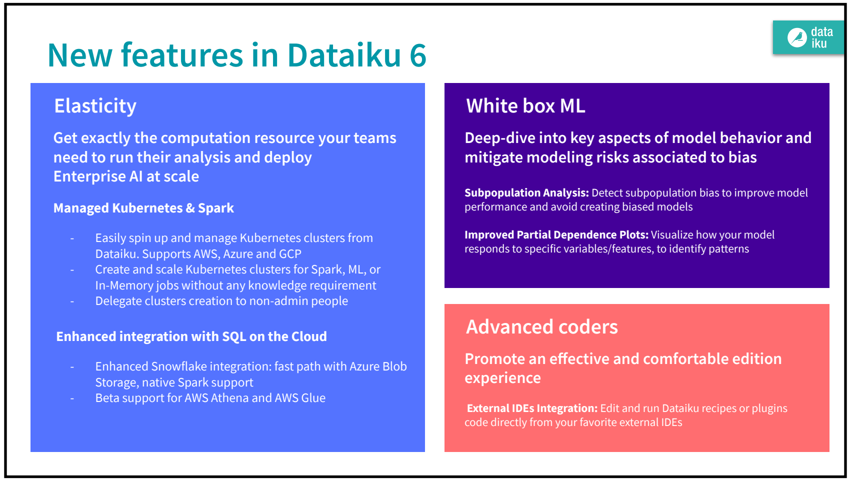
version 4
Advanced scalability, cloud support

version 5
Deep learning, Containerization, Project wikis

data iku

The Long Journey





New features in Dataiku 6

Elasticity

Get exactly the computation resource your teams need to run their analysis and deploy Enterprise AI at scale

Managed Kubernetes & Spark

- Easily spin up and manage Kubernetes clusters from Dataiku. Supports AWS, Azure and GCP
- Create and scale Kubernetes clusters for Spark, ML, or In-Memory jobs without any knowledge requirement
- Delegate clusters creation to non-admin people

Enhanced integration with SQL on the Cloud

- Enhanced Snowflake integration: fast path with Azure Blob Storage, native Spark support
- Beta support for AWS Athena and AWS Glue

White box ML

Deep-dive into key aspects of model behavior and mitigate modeling risks associated to bias

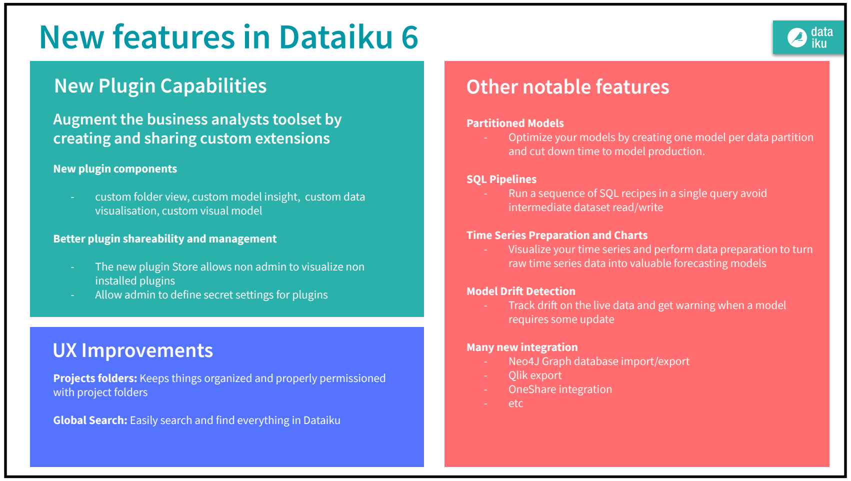
Subpopulation Analysis: Detect subpopulation bias to improve model performance and avoid creating biased models

Improved Partial Dependence Plots: Visualize how your model responds to specific variables/features, to identify patterns

Advanced coders

Promote an effective and comfortable edition experience

External IDEs Integration: Edit and run Dataiku recipes or plugins code directly from your favorite external IDEs



New features in Dataiku 6

New Plugin Capabilities

Augment the business analysts toolset by creating and sharing custom extensions

New plugin components

- custom folder view, custom model insight, custom data visualisation, custom visual model

Better plugin shareability and management

- The new plugin Store allows non admin to visualize non installed plugins
- Allow admin to define secret settings for plugins

UX Improvements

Projects folders: Keeps things organized and properly permissioned with project folders

Global Search: Easily search and find everything in Dataiku

Other notable features

Partitioned Models

- Optimize your models by creating one model per data partition and cut down time to model production.

SQL Pipelines

- Run a sequence of SQL recipes in a single query avoid intermediate dataset read/write

Time Series Preparation and Charts

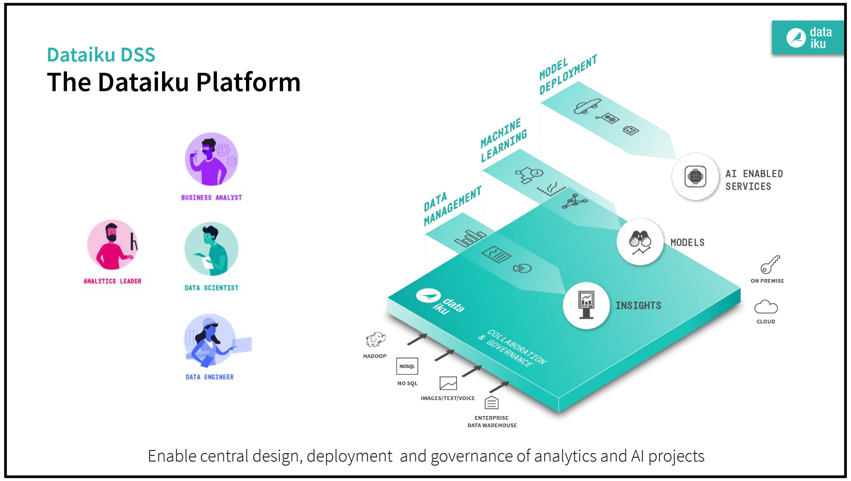
- Visualize your time series and perform data preparation to turn raw time series data into valuable forecasting models

Model Drift Detection

- Track drift on the live data and get warning when a model requires some update

Many new integration

- Neo4J Graph database import/export
- Qlik export
- OneShare integration
- etc



Dataiku DSS

The Dataiku Platform

BUSINESS ANALYST

ANALYTICS LEADER

DATA SCIENTIST

DATA ENGINEER

MODEL DEPLOYMENT

MACHINE LEARNING

DATA MANAGEMENT

AI ENABLED SERVICES

MODELS

INSIGHTS

COLLABORATION & GOVERNANCE

ON PREMISE

CLOUD

HADOOP

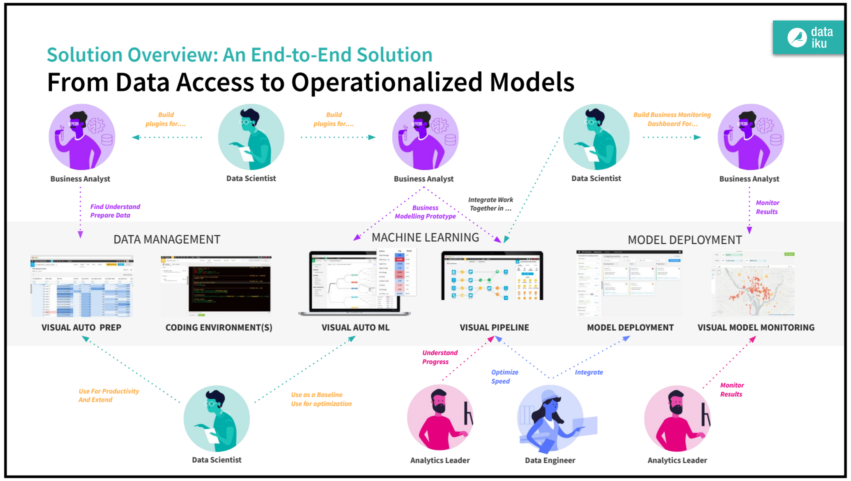
NOSQL

NO SQL

IMAGES/TEXT/VOICE

ENTERPRISE DATA WAREHOUSE

Enable central design, deployment and governance of analytics and AI projects



Solution Overview: An End-to-End Solution

From Data Access to Operationalized Models

Business Analyst

Build plugins for....

Data Scientist

Build plugins for....

Business Analyst

Data Scientist

Build Business Monitoring Dashboard For...

Business Analyst

Find Understand Prepare Data

Business Modelling Prototype

Integrate Work Together in ...

Monitor Results

DATA MANAGEMENT

MACHINE LEARNING

MODEL DEPLOYMENT

VISUAL AUTO PREP

CODING ENVIRONMENT(S)

VISUAL AUTO ML

VISUAL PIPELINE

MODEL DEPLOYMENT

VISUAL MODEL MONITORING

Use For Productivity And Extend

Use as a Baseline Use for optimization

Understand Progress

Optimize Speed

Integrate

Monitor Results

Data Scientist

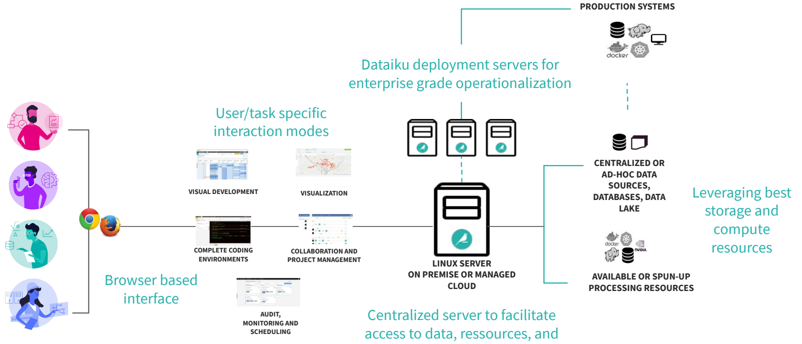
Analytics Leader

Data Engineer

Analytics Leader

Solution Overview: Architecture

Leverage your full stack and skills



PRODUCTION SYSTEMS

Dataiku deployment servers for enterprise grade operationalization

User/task specific interaction modes

VISUAL DEVELOPMENT

VISUALIZATION

COMPLETE CODING ENVIRONMENTS

COLLABORATION AND PROJECT MANAGEMENT

AUDIT, MONITORING AND SCHEDULING

Browser based interface

LINUX SERVER ON PREMISE OR MANAGED CLOUD

CENTRALIZED OR AD-HOC DATA SOURCES, DATABASES, DATA LAKE

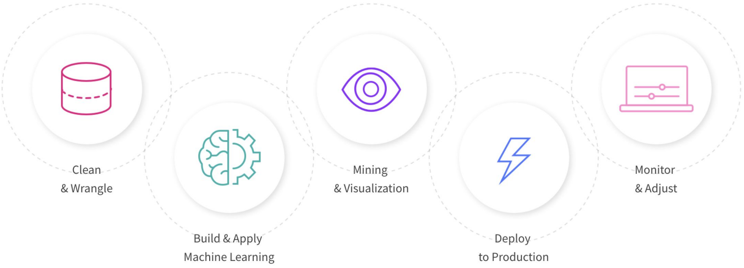
AVAILABLE OR SPUN-UP PROCESSING RESOURCES

Leveraging best storage and compute resources

Centralized server to facilitate access to data, ressources, and foster collaboration

Dataiku DSS

End-to-end Software



Clean & Wrangle

Build & Apply Machine Learning

Mining & Visualization

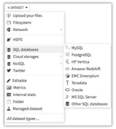
Deploy to Production

Monitor & Adjust

Flashlight on Dataiku DSS
Data Access and Processing

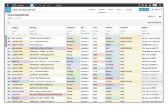
CONNECT TO YOUR (MANY) DATA SOURCES

- Click based connection to your datalake, databases, flat files or any other source
- Native connectors for most common technologies (SQL, Hadoop, Spark, NoSQL…)
- Find relevant data with the catalog



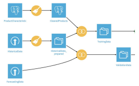
UNDERSTAND YOUR DATA

- Directly navigate through data samples and compute the key
- Build immediate visualisation for self- or shared- use



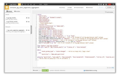
DEVELOP REUSABLE/MAINTAINABLE DATA FLOWS

- Simple representation of overall data processing despite complexity of underlying operations
- Portability of in-built recipes to various execution engines



CODE… OR CLICK

- Code in your language of choice (R, Python, SQL…)
- Use the 100 in-built processors to perform advancecd operations in a few click
- Package your specific processes in reusable click based plugins for wide usage



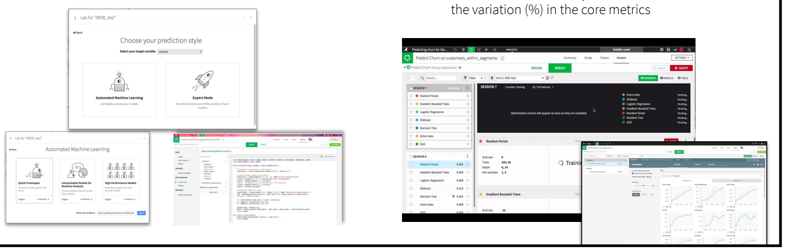
LEVERAGE VARIOUS COMPUTE RESOURCES

- Push process execution in your databases or computation clusters
- Deploy via containers or spin up compute resources

Flashlight on Dataiku DSS
Machine Learning

AUTO-ML OR CODE BASED

- Create, improve, and automatically compare multiple machine learning models in a few clicks
- Leverage the best ML libraries in the way most suited to your users


Lab for "IMDB_test"
Back
Choose your prediction style
Select your target variable
Automated Machine Learning
Let Dataiku create your models.
Expert Mode
Have full control over of the creation of your models.
Automated Machine Learning
Interpretable Models for Business Analysts
High Performance Models
Training

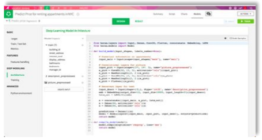
FULL TRAINING MANAGEMENT

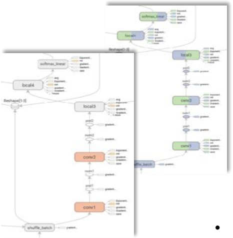
- Track model performance
- Historize trainings, setup and sessions
- Select best performers or revert to past designs
- Update your model with ease: add new data, create new variables in your models, and check the variation (%) in the core metrics

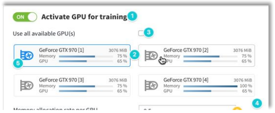
Flashlight on Dataiku DSS
Deep Learning

VISUAL DEEP LEARNING

- Write your own TensorFlow / Keras code
- Extensive library of code samples
- Deploy using the API deployer
- GPU Support






ON Activate GPU for training 1

Use all available GPU(s) □ 3

GeForce GTX 970 [1]		3076 MiB
Memory		75 %
GPU		65 %
GeForce GTX 970 [2]		3076 MiB
Memory		75 %
GPU		65 %
GeForce GTX 970 [3]		3076 MiB
Memory		75 %
GPU		65 %
GeForce GTX 970 [4]		3076 MiB
Memory		100 %
GPU		65 %

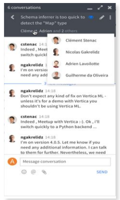
BENEFITS

- Learn and create Deep Learning models
- Once created, deploy in the flow or in a rest API and benefit from the Dataiku DSS model deployment framework (versioning, rollback, scalability, high availability)

Flashlight on Dataiku DSS
Full stack collaborative Data Science

DISCUSSIONS

- Conversations are associated to items in Dataiku
- Comment, answer comments and close threads.
- Subscribe and get notified.
- Centralized Inbox for all conversations


6 conversations

Schema inferrer is too quick to detect the "Map" type

Clément, Adrien and 2 others

Clément Stenac
Nicolas Gakrelidz
Adrien Lavoilotte
Guilherme da Oliveira

cstenac 14:1
Indeed , Meet
switch quickl

ngakrelidz 1
I'm on versio
need any add

3

ngakrelidz 14:18
Don't expect any kind of fix on Vertica ML - unless it's for a demo with Vertica you shouldn't be using Vertica ML.

cstenac 14:18
Indeed , Meetup with Vertica :-). Ok , i'll switch quickly to a Python backend ...

ngakrelidz 14:18
I'm on version 4.0.5. Let me know if you need any additional information. I can talk to them for further. Nevertheless, we need

WIKI FOR TEAM DOCUMENTATION

- Create team documentation
- Centralize and organize the shared resources in a hierarchical manner
- Create a project new entry point with structured documentation

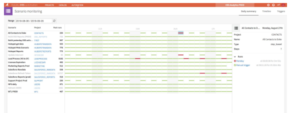
USER HOME PAGE

- Only show items that are the most likely relevant to you
- Favorite projects and Dashboards
- Learning Documentation for new users
- Group projects into folders for better organization

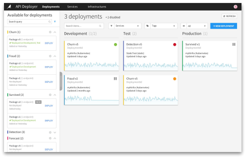
Flashlight on Dataiku DSS
Project Deployment

FULL DATA FLOW ORCHESTRATION

- Deploy and monitor Data Projects execution in production environments
- Define execution scenarios



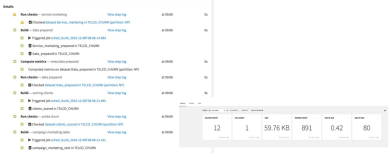
API DEPLOYER



- Self-service deployment of models by Data Scientist / Analyst
- Models Lifecycle Management (challenger, dev, test, prod)
- Models quality monitoring and rollback management
- Scalable and HA deployment of ML API services on container fleets (Docker/Kubernetes either on premise or using cloud services)
- Deployment health monitoring

CHECKS AND METRICS

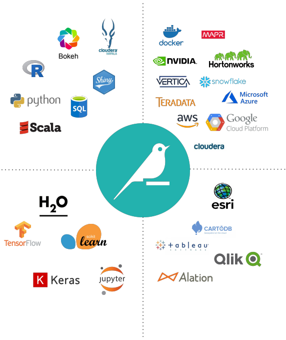
- Set-up IT, Data, Model performance checks and metrics to ensure optimal performance in production environment
- Automate warning or retraining in case of drift



Future proof your data effort
Get Results Today, Build for Tomorrow

Leverage existing skills and ensure availability of expertise

Combination of different modes of interaction (visual data prep, full code, auto ML…) and most popular programming languages



Use your current infrastructure and be ready for tomorrow’s

Support an increasingly wide array of storage/processing technologies and facilitate decoupling of data project from selected infrastructure

Maximise usage of most up-to-date technologies

Full integration of the most popular Data and Machine Learning libraries

Extend based on current and future specific requirements

Connect to a full ecosystem of solutions and customize and manage your own connectors, processors, code, processes, libraries as needed

What we have had the opportunity to see

Enterprise AI: Unique Journeys, Shared Challenges

A wide array of use cases

Production Improvements

Predictive Maintenance

Market Analysis

Fraud detection

Risk Analysis

Product Recommendation

Logistics optimization

Pricing

Churn Prediction

And many, many, many more

Across many industries

Technology GE SAMSUNG	Banking Morgan Stanley UBS
Heavy Industry SOLVAY HYUNDAI STEEL	Consumer Goods Unilever SEPHORA
Semiconductors SK hynix NXP	Consulting pwc BCG
Transportation DELTA DAIMLER	Media 21ST CENTURY FOX UBISOFT
Healthcare HCSC Pfizer	Insurance ZURICH AVIVA

Enabling self-service analytics

For fortune 500 companies

Our typical Fortune 500 deployment would comprise:

ANALYTICS LEADER

15 Projects Leaders

Scale their team to deliver

10x Projects / Briefs / Models / ….

DATA SCIENTIST

75 Data Scientists

Focus On Complex Data Processing

Deliver Code and Plugins for **Reuse**

BUSINESS ANALYST

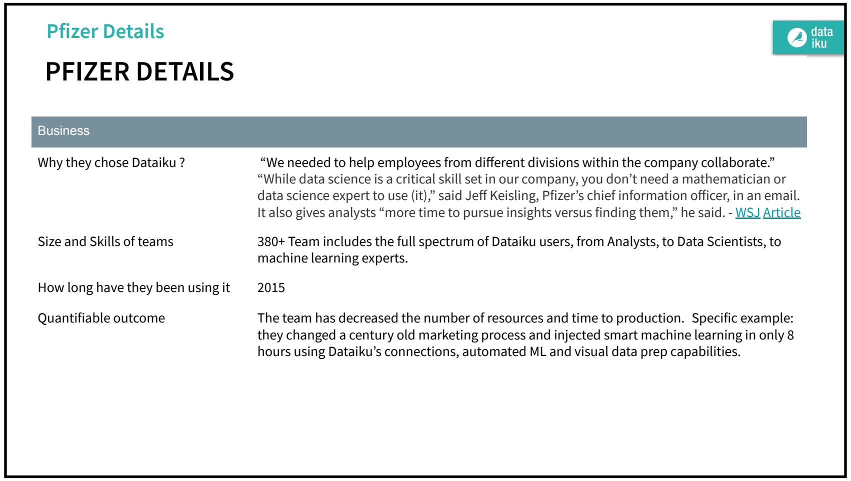
290 Business Analysts

Leverage Large and Complex Data Sources

Independent to Deliver New Projects

More than 200 billion data points (and counting)
Large, established data team
Data Scientists
Data Analysts
Large, established data ecosystem
100s of staff in 150+ countries
Physical + technological distance between data teams
2017
2021
Long-term company vision
Stable & structured processes
Challenges
Disconnect between data and business teams
Problems in finding, using, and interpreting data
Lack of agility and quick iteration
Too much time prototyping

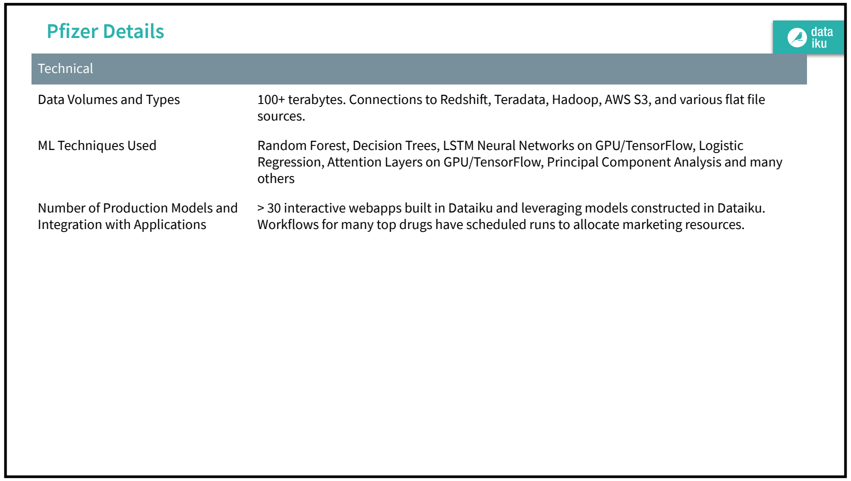
Dataiku as the cornerstone of the new Analytics Workbench initiative at Pfizer Global Business Analytics and Insights Team



Pfizer Details

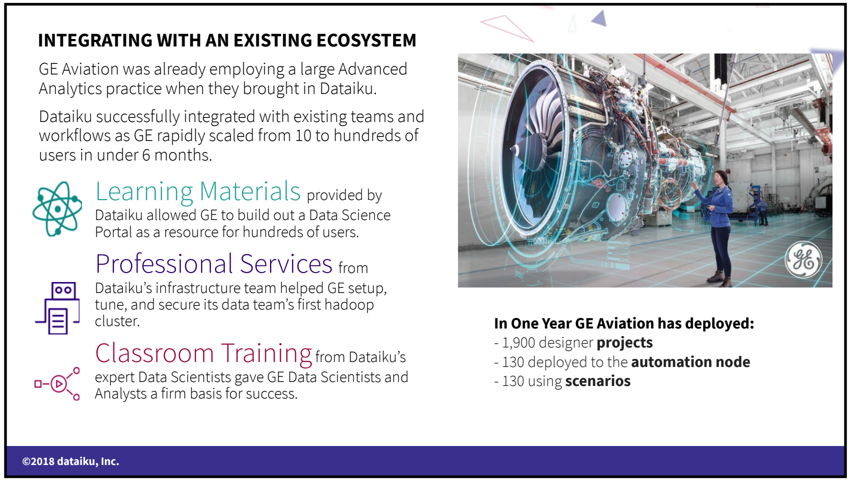
PFIZER DETAILS

Business	
Why they chose Dataiku ?	"We needed to help employees from different divisions within the company collaborate." "While data science is a critical skill set in our company, you don't need a mathematician or data science expert to use (it)," said Jeff Keisling, Pfizer's chief information officer, in an email. It also gives analysts "more time to pursue insights versus finding them," he said. - WSJ Article
Size and Skills of teams	380+ Team includes the full spectrum of Dataiku users, from Analysts, to Data Scientists, to machine learning experts.
How long have they been using it	2015
Quantifiable outcome	The team has decreased the number of resources and time to production. Specific example: they changed a century old marketing process and injected smart machine learning in only 8 hours using Dataiku's connections, automated ML and visual data prep capabilities.



Pfizer Details

Technical	
Data Volumes and Types	100+ terabytes. Connections to Redshift, Teradata, Hadoop, AWS S3, and various flat file sources.
ML Techniques Used	Random Forest, Decision Trees, LSTM Neural Networks on GPU/TensorFlow, Logistic Regression, Attention Layers on GPU/TensorFlow, Principal Component Analysis and many others
Number of Production Models and Integration with Applications	> 30 interactive webapps built in Dataiku and leveraging models constructed in Dataiku. Workflows for many top drugs have scheduled runs to allocate marketing resources.



INTEGRATING WITH AN EXISTING ECOSYSTEM

GE Aviation was already employing a large Advanced Analytics practice when they brought in Dataiku.

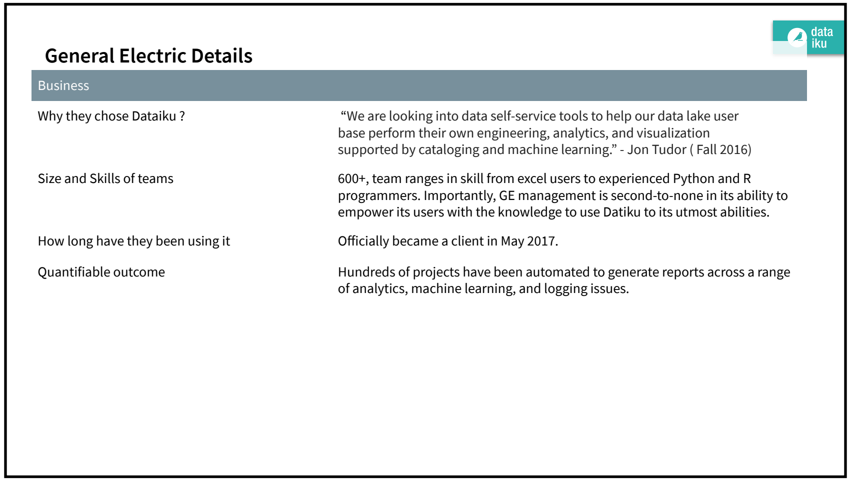
Dataiku successfully integrated with existing teams and workflows as GE rapidly scaled from 10 to hundreds of users in under 6 months.

Learning Materials provided by Dataiku allowed GE to build out a Data Science Portal as a resource for hundreds of users.

Professional Services from Dataiku's infrastructure team helped GE setup, tune, and secure its data team's first hadoop cluster.

Classroom Training from Dataiku's expert Data Scientists gave GE Data Scientists and Analysts a firm basis for success.

In One Year GE Aviation has deployed:
- 1,900 designer **projects**
- 130 deployed to the **automation node**
- 130 using **scenarios**



General Electric Details

Business	
Why they chose Dataiku ?	"We are looking into data self-service tools to help our data lake user base perform their own engineering, analytics, and visualization supported by cataloging and machine learning." - Jon Tudor (Fall 2016)
Size and Skills of teams	600+, team ranges in skill from excel users to experienced Python and R programmers. Importantly, GE management is second-to-none in its ability to empower its users with the knowledge to use Datiku to its utmost abilities.
How long have they been using it	Officially became a client in May 2017.
Quantifiable outcome	Hundreds of projects have been automated to generate reports across a range of analytics, machine learning, and logging issues.

General Electric Details

Technical	
Data Volumes and Types	Hundreds of Terabytes. Spread across Hadoop, Greenplum and Teradata. The client is moving to cloud but currently has a mix of on-prem and cloud systems.
ML Techniques Used	Due to contracts with US Military the client is not allowed to release information on any specific techniques.
Number of Production Models and Integration with Applications	- 1,900 designer projects - 130 deployed to the automation node - 130 using scenarios - “scenarios and automation are 1 to 1 for us”

Demo!