Watson Studio

Data Management and Data Visualization

Emmanuel Génard – *genard@fr.ibm.com*
Cloud Developer Advocate Europe &
Data Scientist

October 2019

Introduction

Emmanuel GENARD

Architecte IT & Leader Technique
Business Analytics and Optimization
–
IBM

I am Emmanuel Génard

You can find me at *genard@fr.ibm.com*

@manuGenard

http://fr.linkedin.com/in/egenard

Stand IBM: 23

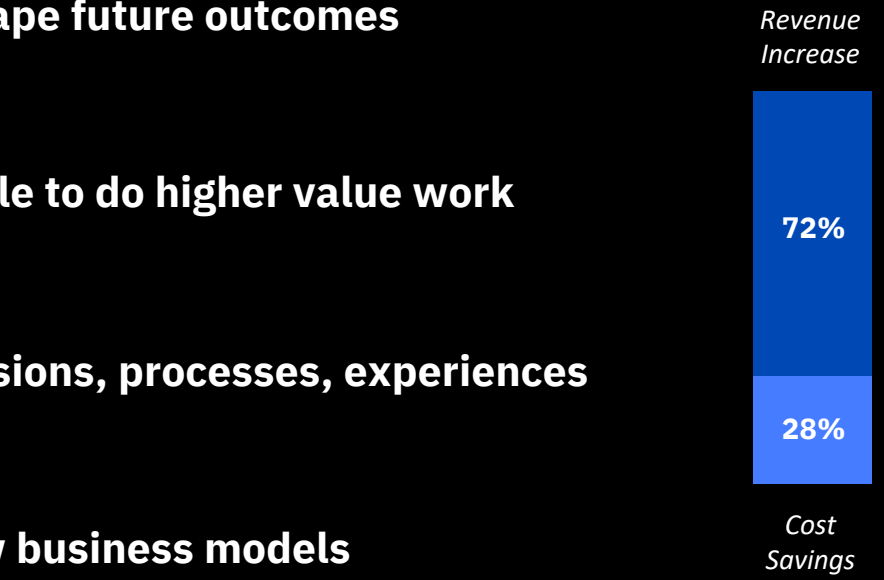
AI is shaping the future of work

Geisinger Caring — **Predict and shape future outcomes**

Woodside — **Empower people to do higher value work**

experian. — **Automate decisions, processes, experiences**

LEGALMATION — **Reimagine new business models**



How AI pioneers see value

MITSloan

However, AI is not magic

RBS The Royal Bank of Scotland — Achieved a 40% call deflection rate with virtual agents

Mercedes-Benz — Cognitive car manual explaining increased vehicle complexity

MITSUBISHI HEAVY INDUSTRIES, LTD. — Identifies gaps in terms in complex RFPs

RED ELÉCTRICA DE ESPAÑA — Predict power demand by for renewable energy

Our learnings from experience in helping thousands of enterprises put AI to work

AUTOGLASS® — Visually categorize damage & instantly issues quote

NEDBANK — Predict fraud across their web & mobile banking system

iKure — Optimize cardiac care in high volume remote regions

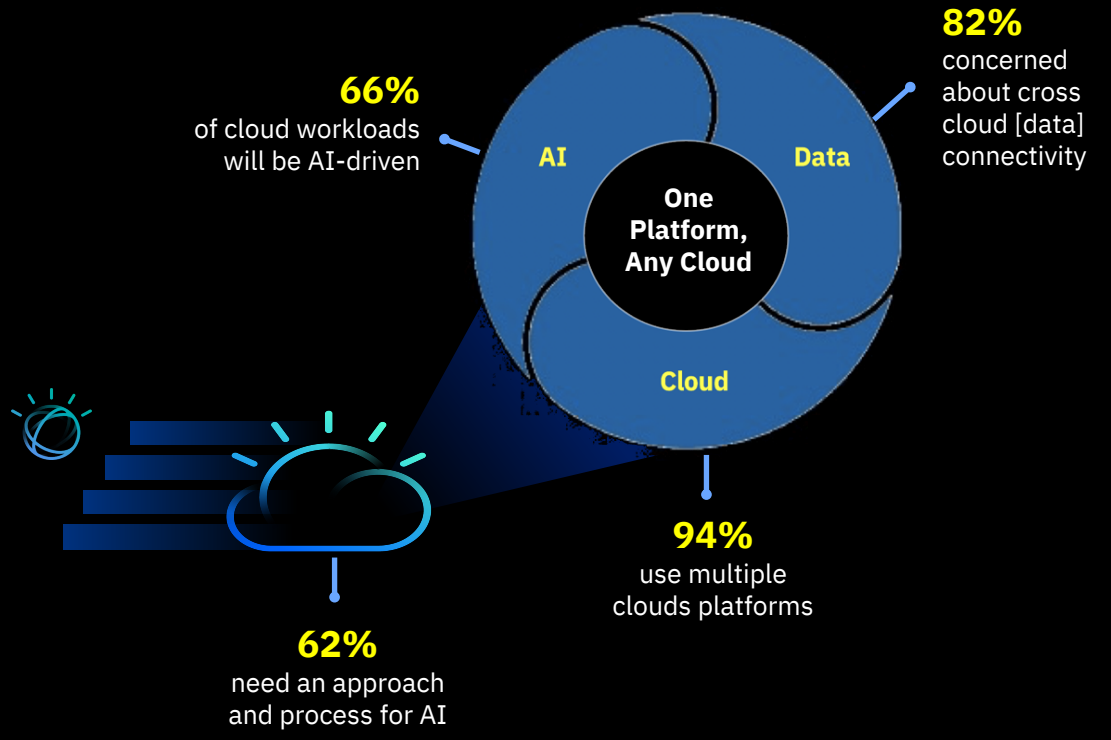
MOTOR HARLEY-DAVIDSON CYCLES — Predict and target first-time buyers in the US

ESPN FANTASY — Surface hidden insights to optimize fantasy football outcomes

Lufthansa — AI-powered advertising engagement

An Information Architecture for AI

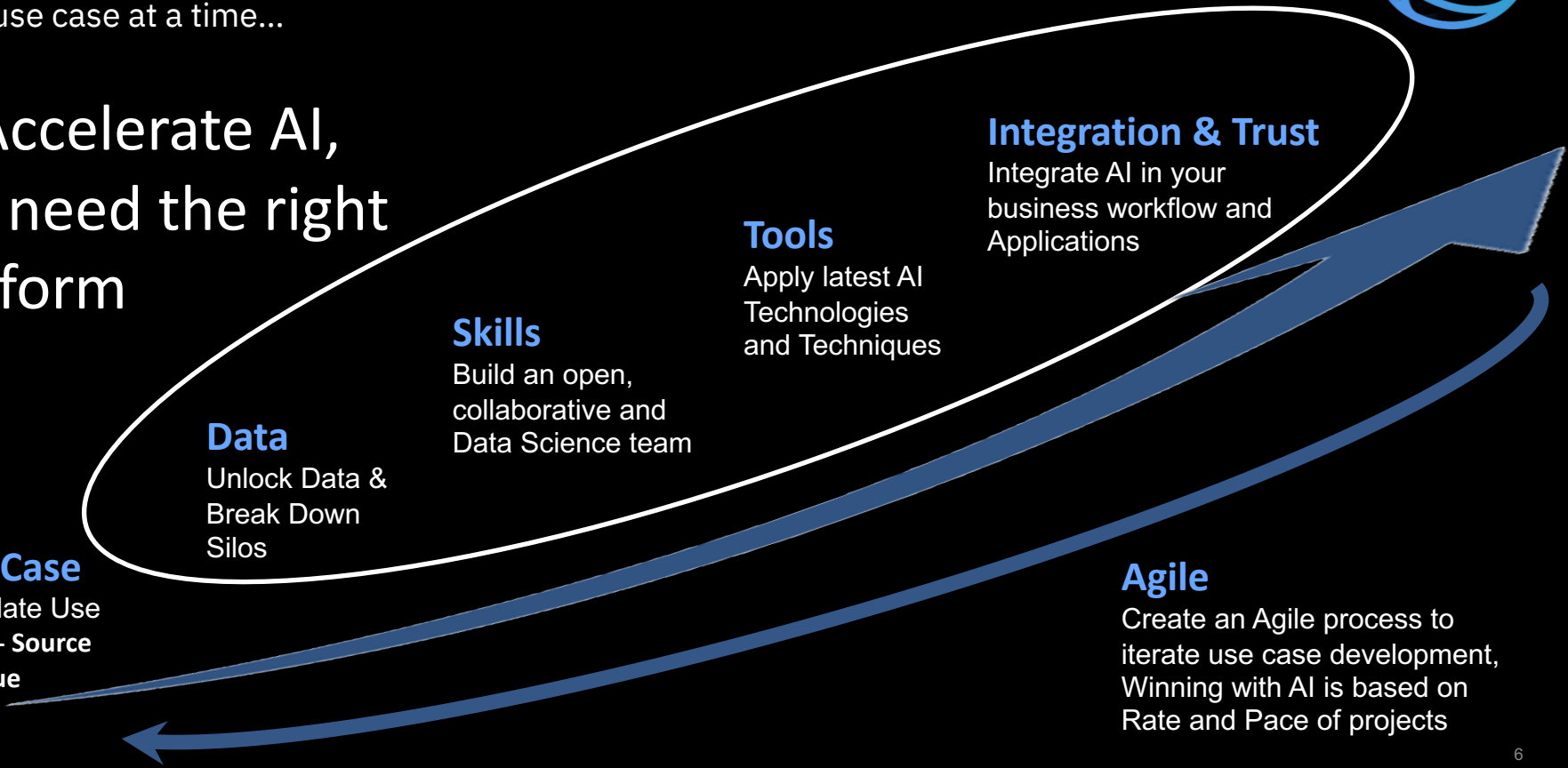
- Eliminate data silos, connect all data, on any cloud(s)
- Automate and govern a unified data & AI lifecycle
- Operationalize AI with trust and transparency
- Avoid lock-in, run anywhere with agility



Putting AI to Work for Business

One use case at a time…

To Accelerate AI, You need the right Platform



Turning AI aspirations into outcomes

DATA

The lifeblood of AI, but complexity slows progress

60%
Are challenged in managing data quality

TALENT

AI skills are rare and in high demand

62%
Are challenged to acquire talent [and build skills]

TRUST

Skepticism of AI systems & processes

62%
Need an approach to AI production readiness

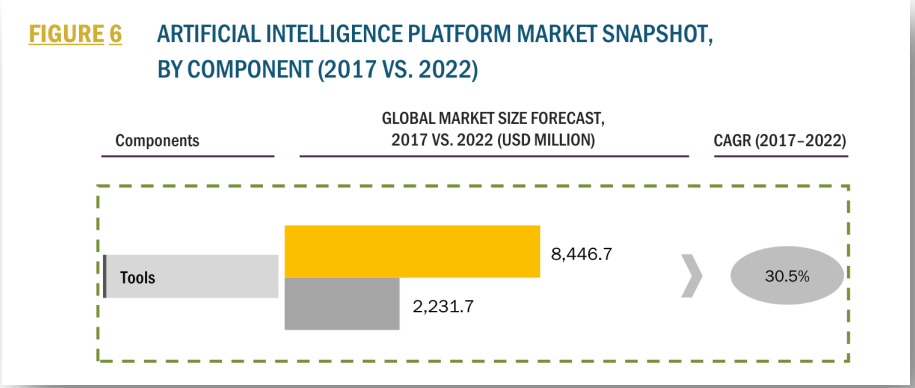
Stuck in Experimentation | **51%** find operationalizing, sustaining and scaling AI challenging

Based on 2019 Forrester "Challenges That Hold Firms Back From Achieving AI Aspirations"

Skills Problem?

Rapidly Growing Market

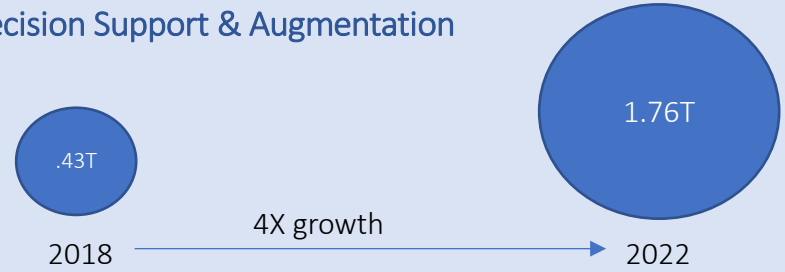
- AI Platform Market expected 30.5% CAGR from 2017 to 2022
- Investment in building models and productizing them to generate this revenue will be to the tune of several 100 Billions



The above research estimations were done using bottoms up analysis – and excluded services

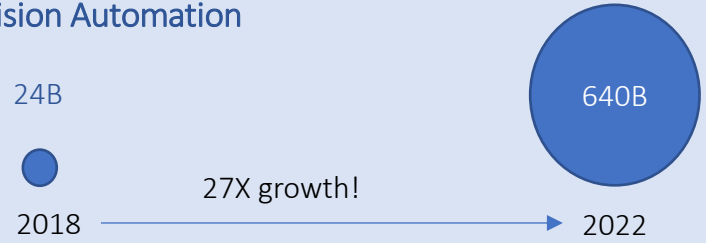
https://ibm.northernlight.com/document.php?trans=view&docid=IB20171121610000024&datasource=IBM&context=BNES

Decision Support & Augmentation



Models for enhanced insights and decision support tools

Decision Automation



Models to deploy in production workflows

There are three major challenges for developing AI applications today:

Challenge

Enterprise data is fragmented
- Moving data is costly, risky, and slow
- Duplication can lead to different outcomes

Data Scientists are a scarce resource
- Diversity in skill levels, and preferences for different open source frameworks
- Computationally intense workloads (Big data, Deep Learning) constrain productivity
- Citizen data scientists emerging

Operationalizing AI is hard
- Hard to trust how models perform in production.
- Data scientists difficulty with integration into engineering and support efforts
- Operationalizing introduces security, scalability, governance constrains

IBM's Strategy

Move data science to the data
- Multi-cloud support
- Pushing model training to Cloud, Hadoop, Mainframe, GPU-supported infrastructure

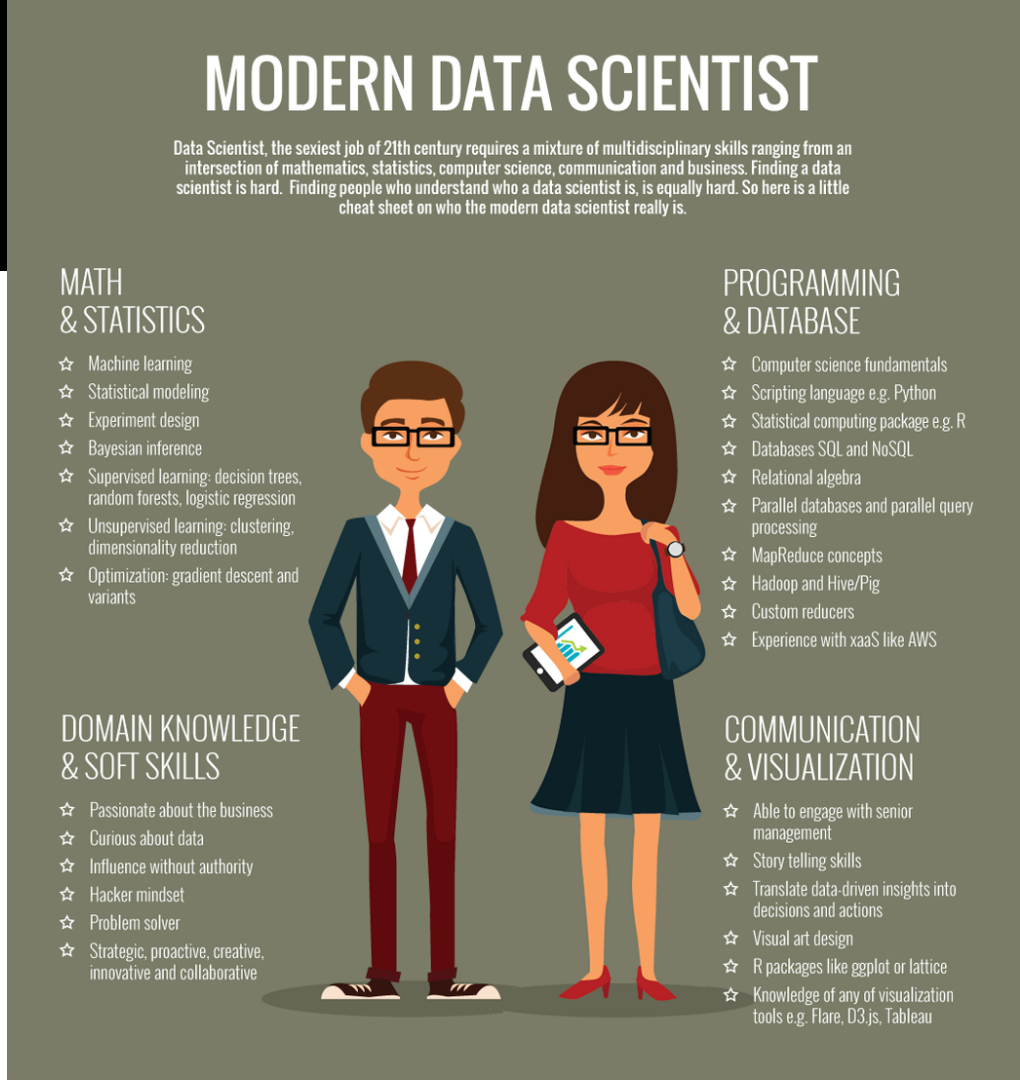
Enable Expert & Citizen Data Scientists
- Reduce barriers to distribution of work
- Visual productivity tools
- Intelligent management of AI infrastructure
- High performance HW/SW acceleration
- Automating Data Science

Full AI Lifecycle, including Open Source
- Ease & flexibility of deployment
- Security, compliance and governance
- Advanced model management capabilities
- Extend to Deep Learning and AI applications

What is a Data Scientist?

A **data scientist** is a professional responsible for collecting, analyzing and interpreting large amounts of data to identify ways to help a business improve operations and gain a competitive edge over rivals.

Data Scientist is the sexiest job of the 21th century.



Flavors of Data Science – Turning data into Information, Insights and Actions

Descriptive and Diagnostics:

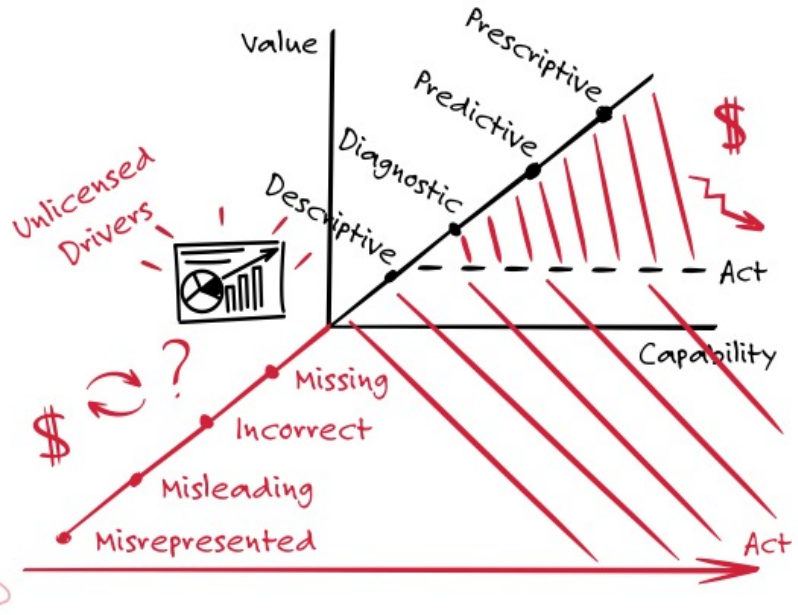
- **WHAT** happened**?**
  - Visualization, Reporting, Dashboards
- **WHY** did it happen **?**
  - Data Exploration, Analysis

Predictive:

- What ***WILL*** happen next**?**
  - Forecasts and previsions
  - Machine Learning (incl. Deep Learning)

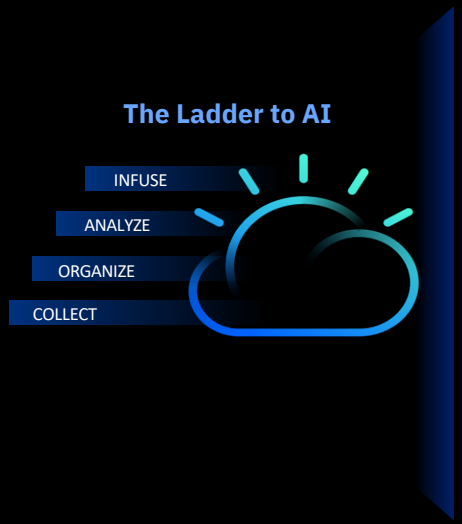
Prescriptive:

- What should we do **?**
  - Actionable Insights
  - Optimization (Operational Research, Constraint Programming)



IBM AI Portfolio

Everything you need for Enterprise AI, on any cloud



Interact with Pre-built AI Services
Watson Application Services

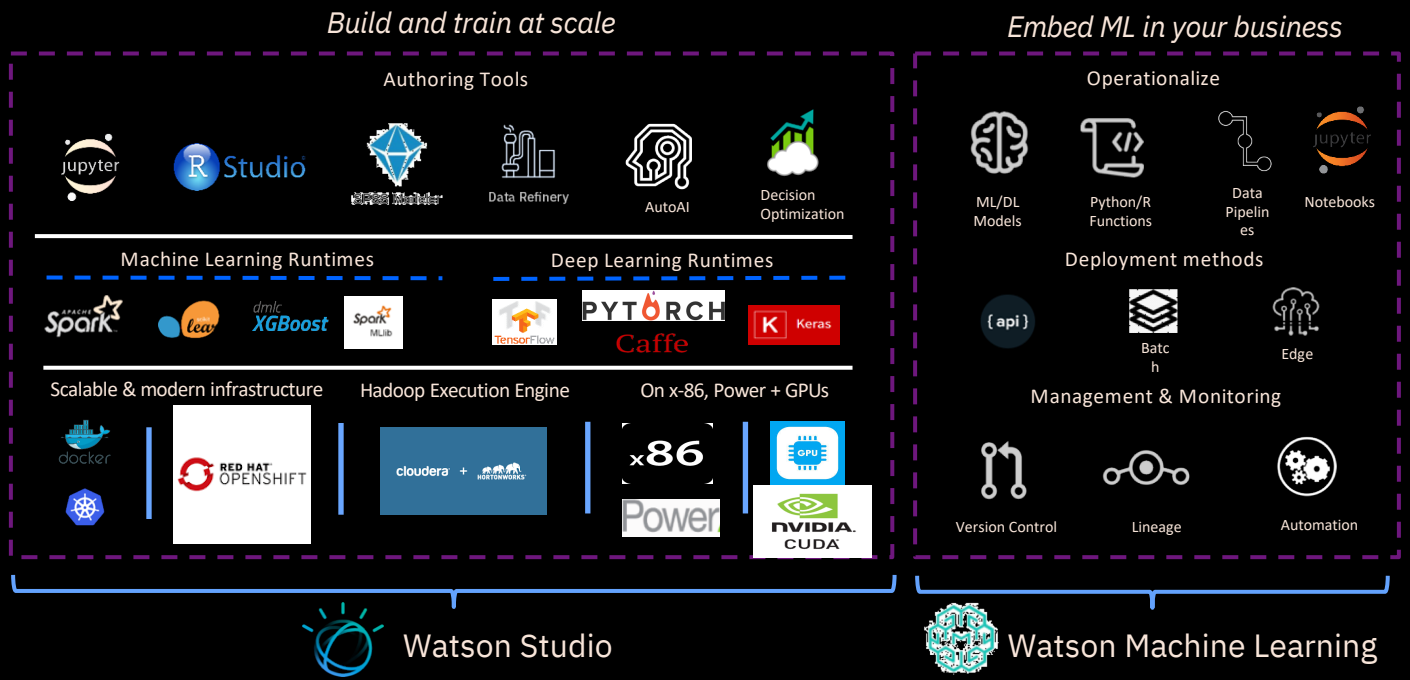
Build	Deploy	Catalog	Manage
Watson Studio	**Watson Machine Learning**	**Watson Knowledge Catalog**	**Watson OpenScale**

OS AI Open Source Frameworks

Unify on a Multicloud Data Platform
IBM Cloud Private for Data

Watson Studio

Watson Studio and Watson Machine Learning inject AI firepower into your business



IBM SOFTLAYER · On-prem · amazon web services · Azure

Mix and Match your deployment

- ✓ Cloud – IBM Cloud, Azure, AWS
- ✓ On Premise / Private Data center
- ✓ Desktop

IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

Analyze any data, no matter where it lives

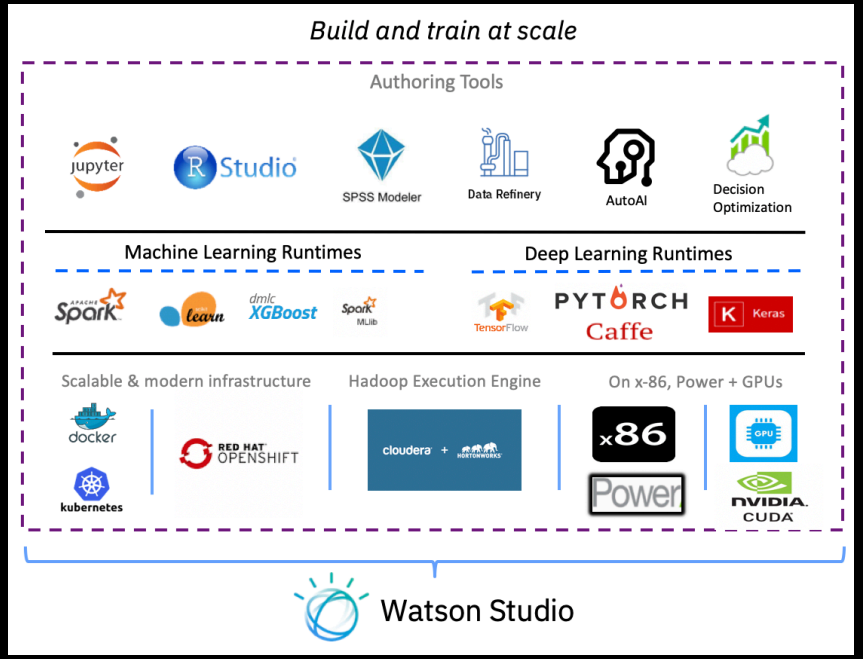
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

Empower your entire organization with notebooks, visual productivity, and automation tools

Leverage your entire organization with a variety of tools in a single integrated platform

One platform to rule them all from discovery to production

Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale

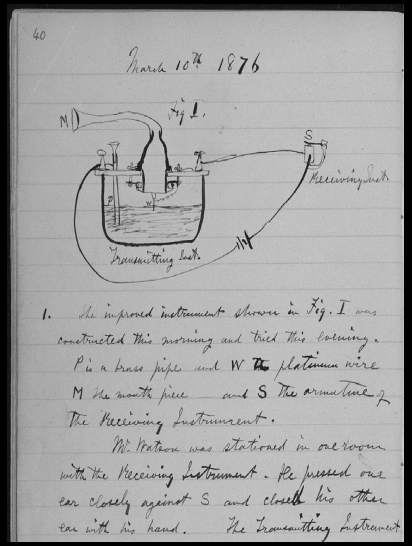


What is a Notebook?

Pen and Paper

Pen and paper has long provided the rich experience that scientists need to document progress through notes and drawings:

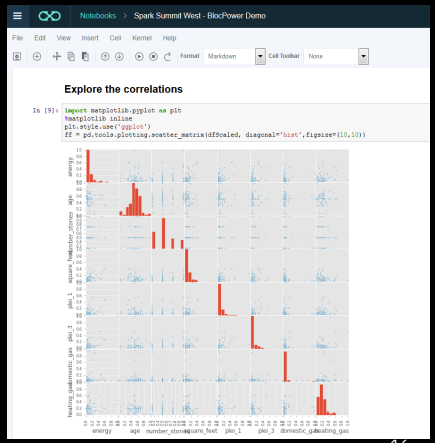
- Expressive
- Cumulative
- Collaborative



Electronic Notebooks

Notebooks are the digital equivalent of the “pen and paper” lab notebook, enabling data scientists to document reproducible analysis:

- Markdown and visualization
- Iterative exploration
- Easy to share



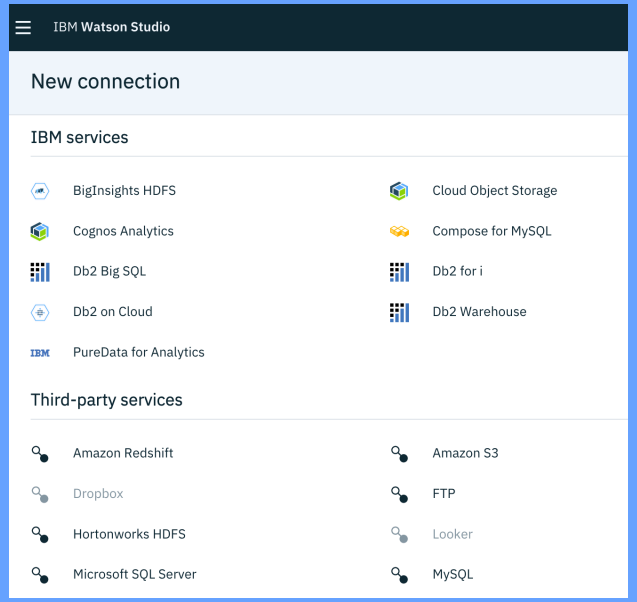
IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

Analyze any data, no matter where it lives
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

Empower your entire organization with notebooks, visual productivity, and automation tools
Leverage your entire organization with a variety of tools in a single integrated platform

One platform to rule them all from discovery to production
Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale



- IBM Services like **Cognos** & **DB2**
- 3rd Party Services like **Amazon S3**, **Hadoop**, & **Microsoft SQL Server**
- We have **Public Cloud**, **Private Cloud**, & **Desktop/Server** deployment options

IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

Analyze any data, no matter where it lives

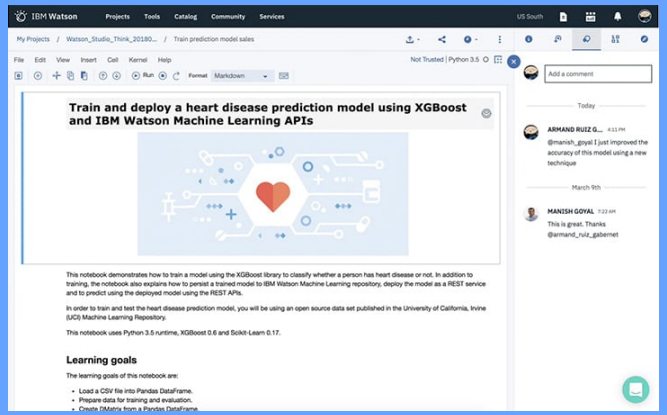
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

Empower your entire organization with notebooks, visual productivity, and automation tools

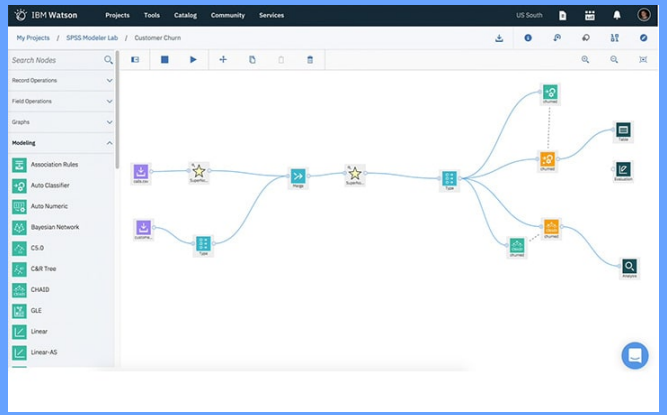
Leverage your entire organization with a variety of tools in a single integrated platform

One platform to rule them all from discovery to production

Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale



Super charged Jupyter Notebooks & R Studio as most popular IDEs for data scientists well integrated with data connectors and rich set of default environments



Visual tools such as SPSS Modeler, Data Refinery, & AutoAI for non coders to analyze data and build models

IBM Watson Studio

Enterprise Data Science platform that helps your team work together to build models to make better data driven decisions for your business

Analyze any data, no matter where it lives
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

Empower your entire organization with notebooks, visual productivity, and automation tools
Leverage your entire organization with a variety of tools in a single integrated platform

One platform to rule them all from discovery to production
Analyze data, build predictive models, and seamlessly integrate Watson Machine Learning to deploy at scale

Seamless integration with:

- Watson Knowledge Catalog brings governance & catalog management

- Watson Machine Learning deploys, manages, & automatically retrains your models

- Watson OpenScale offers unmatched levels of trust in automation

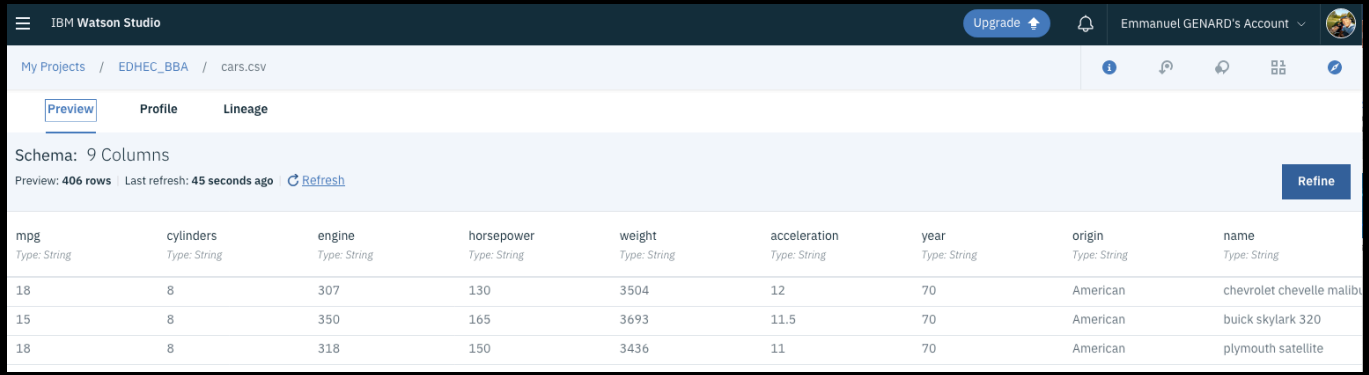
Watson Studio

Data Refinery

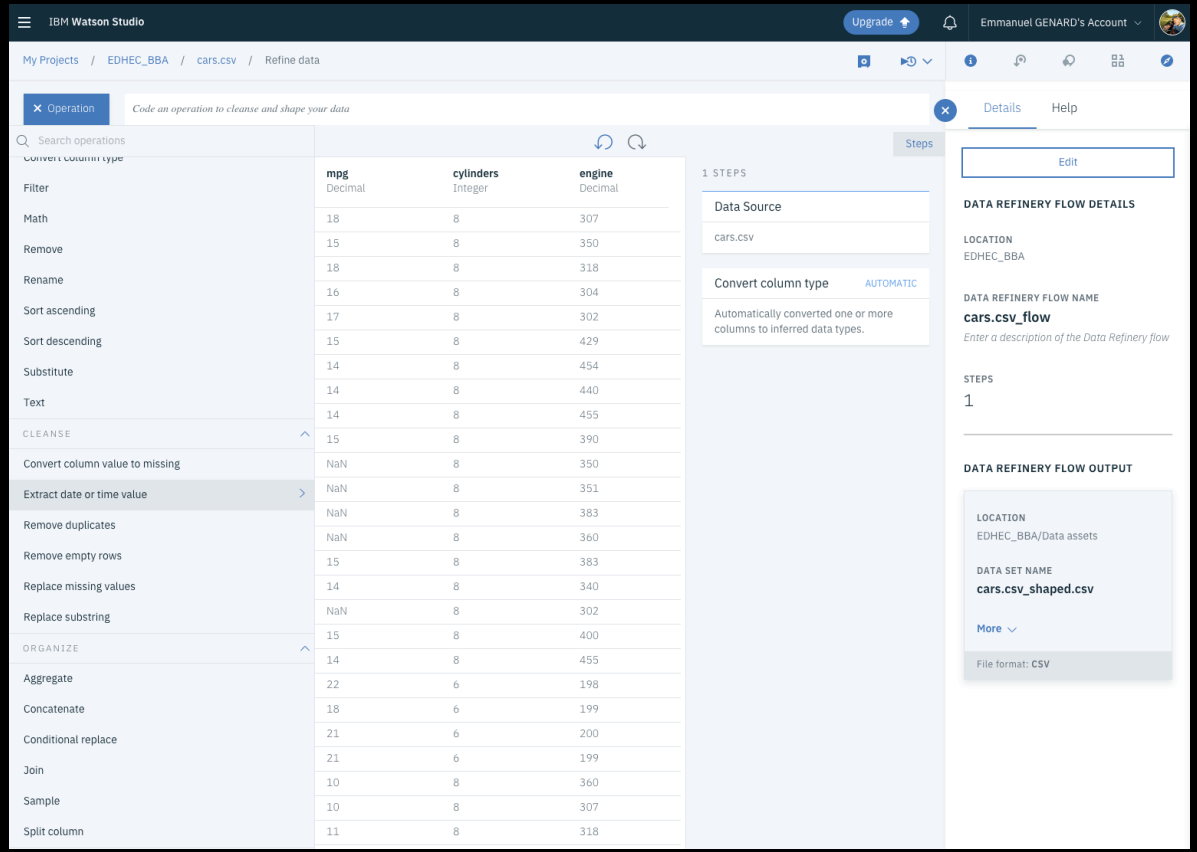
Data Refinery

Allows low-code implementation of Data Processing pipelines

- Data processing steps are composed through a simple GUI
  - Uses ***dplyr*** as the underlying language
- Creates a ‘Data Flow’, which can run against same-shape data sources
- Can read and write several formats, DB connection, CSV, Parquet, Avro…



Data Refinery



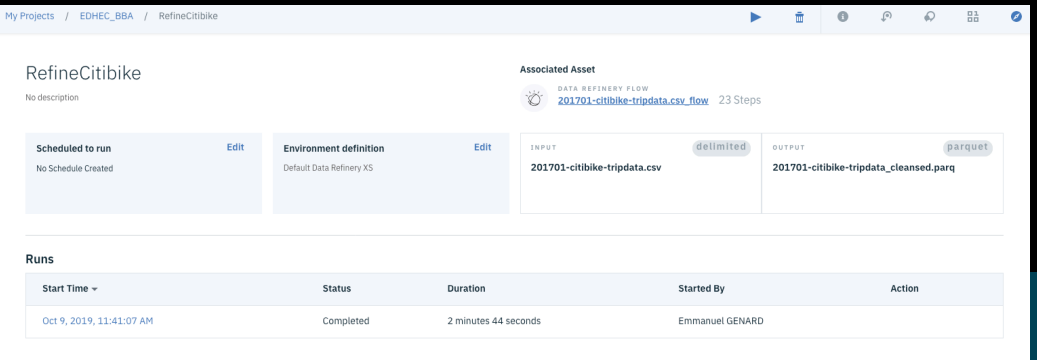
Data Refinery

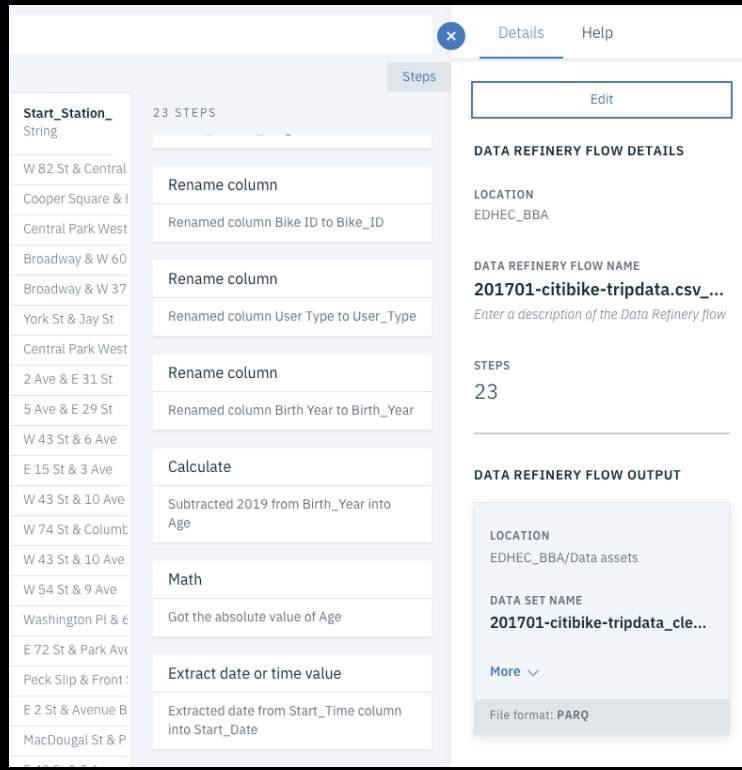
Data Flows can be deployed as Jobs

- Jobs are running in a managed runtime
  - dplyr code is translated to Spark for optimized execution
- Job execution triggers:
  - Manual or through API
  - Scheduled to run at a given time or periodically

Data Flows are used to implement recurring data engineering tasks

- Data Extraction, Filtering and Transformation

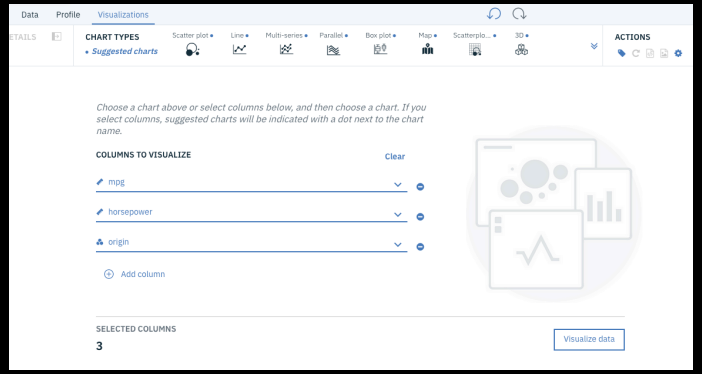


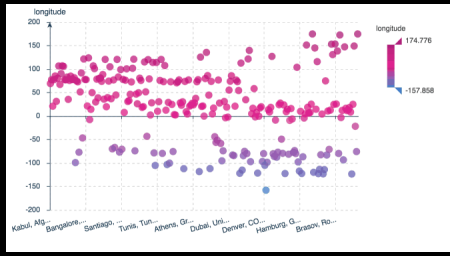


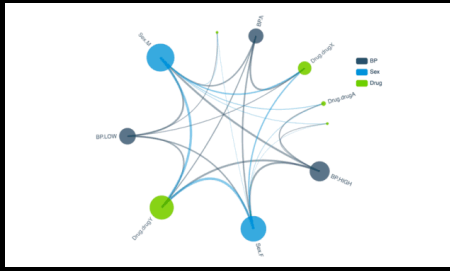
Visualize your data in Data Refinery

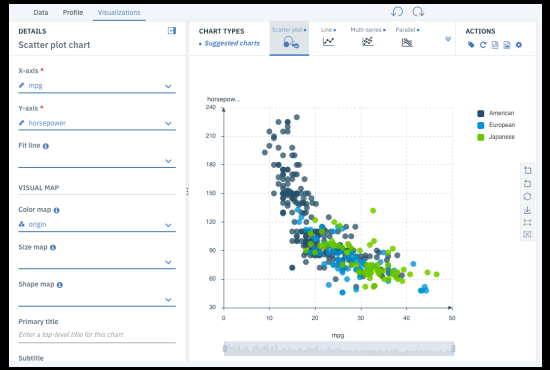
Data Refinery provides support for the following charts:

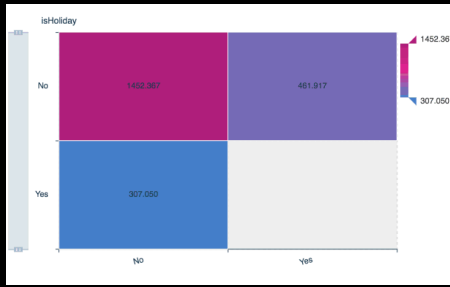
•3D charts, Bar charts, Box plot, Candlestick, Dual Y-axes, Error bars, Heat maps, Histogram charts, Line charts, Map charts, Multi-series charts, Parallel coordinate charts, Pie charts, Population Pyramid charts, Quantile-quantile plot charts, Relationship charts, Scatterplots, Word Cloud…











Watson Studio

Data Visualization

Data Visualization with Watson Studio

- **Introduction to Data Visualization in notebooks**.
  - Standard notebook visualisation: Mathplotlib quick review, R & Shinny
  - IBM visualization packages: Brunel and PixieDust overview
  - Publishing visualizations and graphs
- **Hans-on-Labs:** Getting started with Brunel & Pixie Dust
  - A three-part lab
    - Charting Pandas with Brunel
    - Exploring DataFrames with PixieDust
    - Building dashboards with Cognos Dashboard Embedded service

Descriptive Analytics introduction

Descriptive Analytics is usually the first step towards understanding your data shaping and structure

Main way to conduct descriptive analytics is to build visualizations

- For data exploration, performed by the data scientist as an iterative quest, using notebook visualization libraries
- For reporting, through dashboards that are built for business stakeholders and external consumption

Usually, dashboarding is the final stage, as it requires a proper understanding of the data and some feature engineering to arrange the data in a suitable way

- Aggregations to reduce volume, derived attributes, ...

Visualization technologies in notebooks

The major library used in Jupyter notebooks is called **matplotlib**

- Provides a very extensive set of graphical visualizations
  - https://matplotlib.org/gallery/index.html
- It is a library, which requires Python code to create visualizations
  - Very well supported by Pandas dataframes (df.plot())
  - Spark dataframes need to be converted to pandas
- Other libraries exist, such as seaborn, bokeh (interactive), plot.ly, …

R Studio has its own library called Shiny

- *Also a very extensive scope of graphical visualizations*
  - *https://shiny.rstudio.com/gallery/*
- *Provides the ability to publish interactive graphical output as 'Shiny Apps'*

IBM contribution to Notebook visualization

IBM has a long-standing background in data visualization technologies

- Promoter of the RAVE client-side visualization framework
- The goal is to provide higher-level non-programmatic visualization capabilities

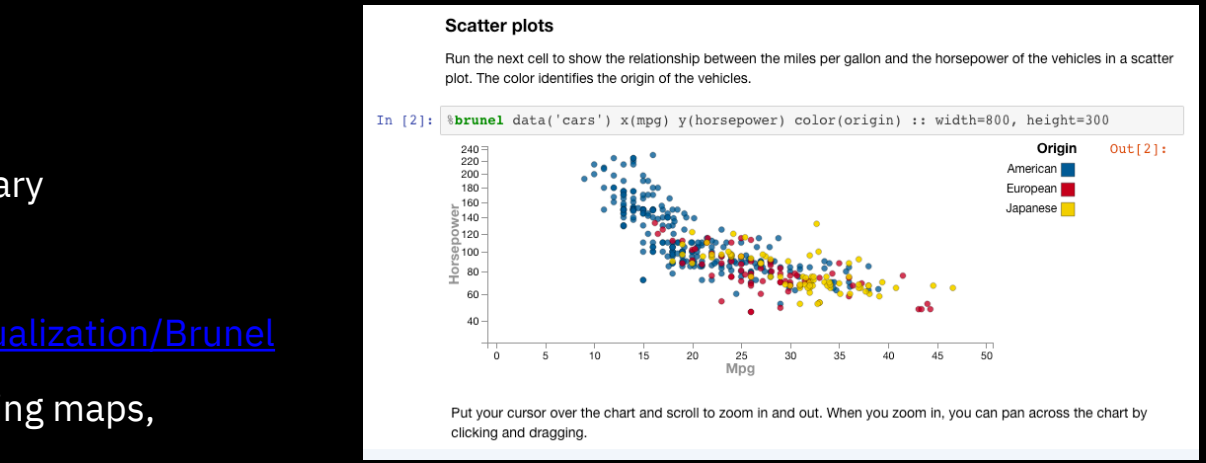
IBM has worked on two libraries made available as Open-source:

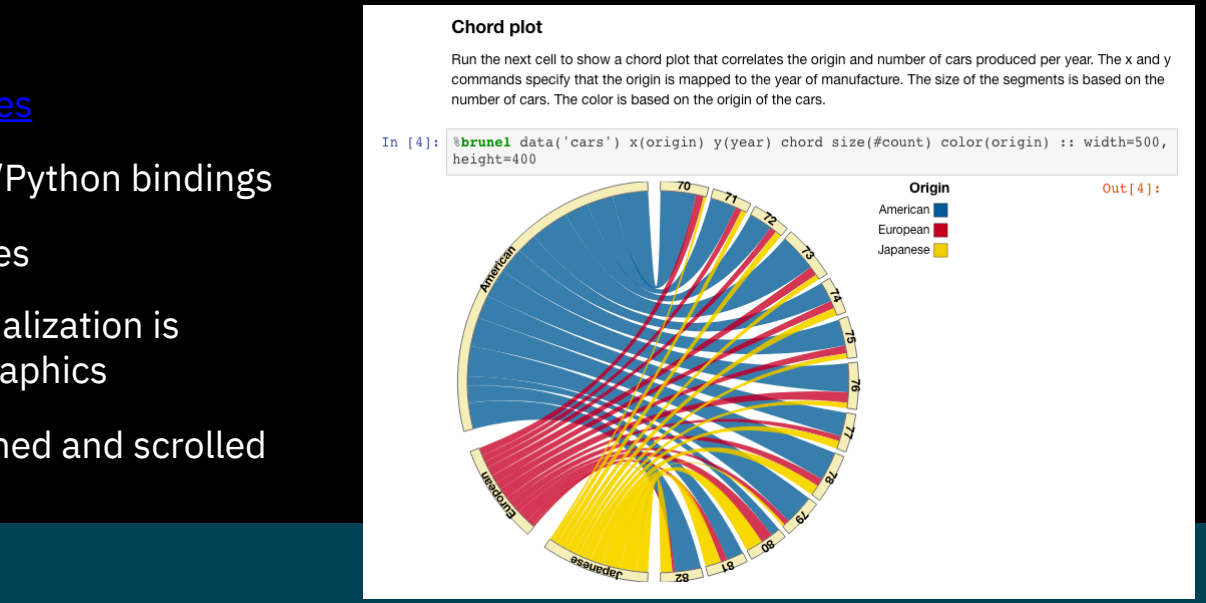
- **Brunel**: https://github.com/Brunel-Visualization/Brunel
  - Brunel defines a highly succinct and novel language that defines interactive data visualizations based on tabular data.
    - *Zero to Visualization in Sixty Seconds*
  - Supports Jupyter, R and Spark
- **PixieDust**: https://github.com/ibm-watson-data-lab/pixiedust
  - Designed for interactive data visualization to support data exploration
  - Supports Jupyter and Spark (Scala)

Brunel visualizations

Brunel is a high-level visualization library

- Open-sourced in Github
  - https://github.com/Brunel-Visualization/Brunel
- Extensive types of graphs, including maps, tagclouds, chords,
  - https://github.com/Brunel-Visualization/Brunel/wiki#samples
- Language-agnostic, with Jupyter/Python bindings
  - Operates on Pandas DataFrames
  - One-line description of the visualization is sufficient to produce complete graphics
- Brunel vizualizations can be zoomed and scrolled but are not actionable





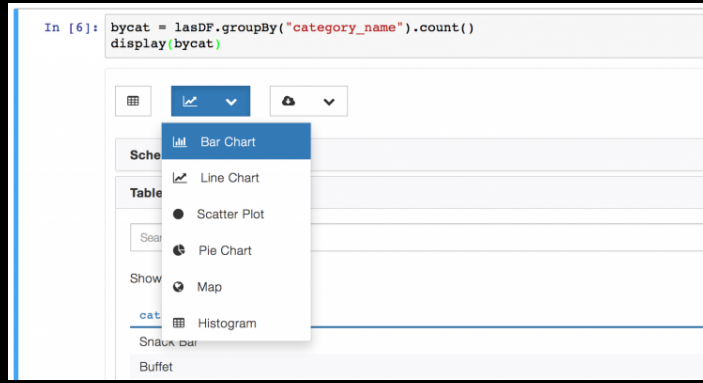
PixieDust

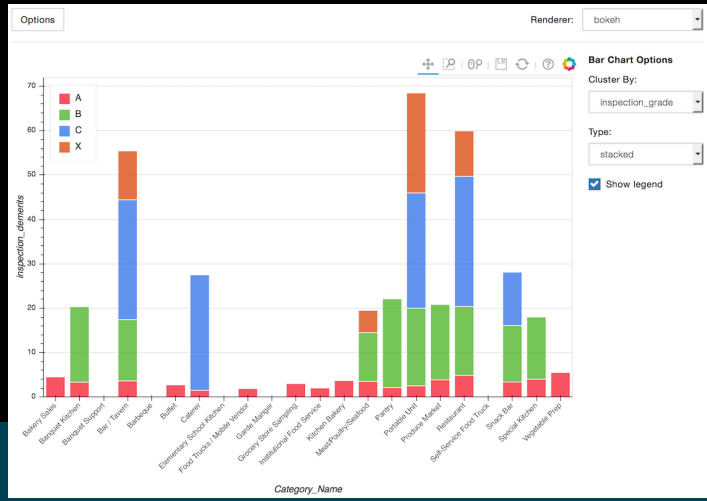
PixieDust is a visual data exploration framework for Jupyter and Spark

- An IBM-contributed open-source project
  - https://github.com/ibm-watson-data-lab/pixiedust
- Goal is to provide visual insights that are extremely interactive
  - User can dynamically change graph types and contents
  - Many types of graphs, including tabular, maps, ...
- PixieDust operates on Pandas or Spark DataFrames
  - Minimal API: `display(df)`

PixieApps

- *Similarly to R's Shiny, live dashboards can be published as Web Apps*
  - *https://ibm-watson-data-lab.github.io/pixiedust/pixieapps.html*





Watson Studio Analytics Dashboards

Interactive UI-driven dashboard builder

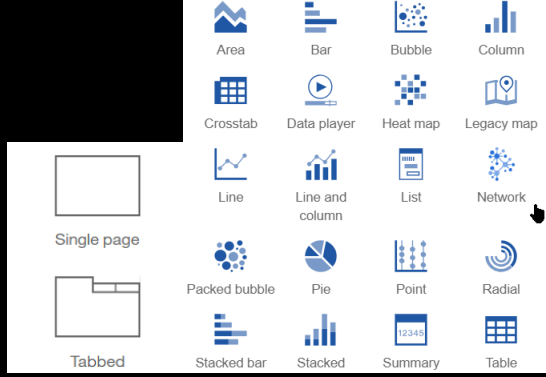
- Tooling part of Watson Studio
- Leverages Watson Studio Data Sources

Advanced dashboarding capabilities

- Multi-tab layouts with linked graphical widgets
- Many visualization types

Application integration

- Dashboard can be published for external viewing
- Dashboards can be integrated in an outside web application
  - Using the client-side JavaScript CognosAPI




demo

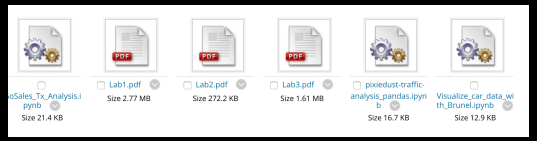
Time for Lab 2: Data Refinery and Visualizations

This lab is multipart

1. Data Preparation using Watson Studio Data Refinery
2. Data Visualization Experiment with Brunel visualizations
3. Experiment with PixieDust interactive visualization
4. Build a Dashboard using Watson Studio's Cognos Dashboard service

Lab material at https://github.com/Azzoz06/EDHEC_BBA_19_20 or on Blackboard

Folder TP_Consumers_Insights_and_Big_data\Day2



Thank You!