

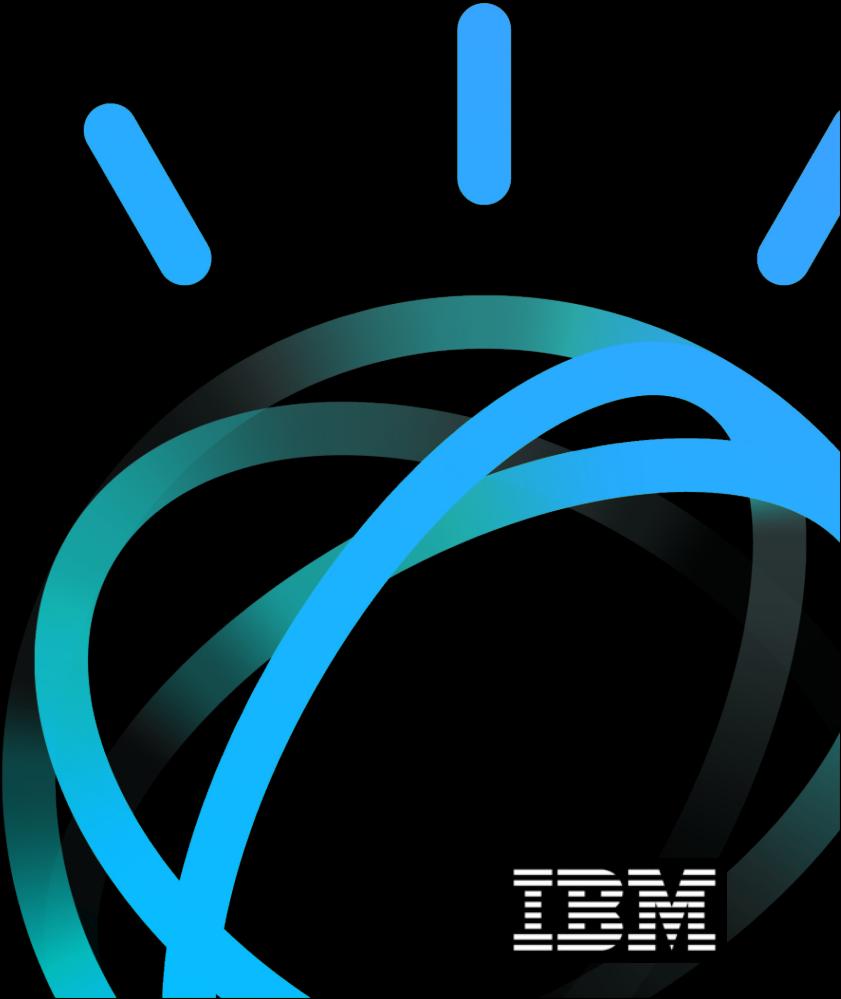
Watson Studio

Data Management and Data Visualization

Emmanuel Génard – genard@fr.ibm.com

Cloud Developer Advocate Europe &
Data Scientist

October 2019





Introduction

I am Emmanuel Génard

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



@manuGenard



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

Emmanuel GENARD

Architecte IT & Leader Technique
Business Analytics and Optimization

—
IBM



Stand IBM: 23

AI is shaping the future of work

Geisinger Caring

Predict and shape future outcomes

Revenue Increase

Woodside

Empower people to do higher value work

72%

How AI
pioneers
see value

experian.

Automate decisions, processes, experiences

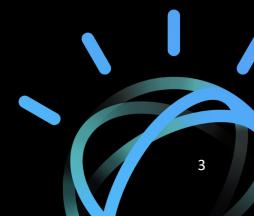
28%

MIT Sloan

LEGALMATION™

Reimagine new business models

Cost Savings



However, AI is not magic



Achieved a 40% call deflection rate with virtual agents



Mercedes-Benz

Cognitive car manual explaining increased vehicle complexity



Identifies gaps in terms in complex RFPs



Predict power demand by for renewable energy



Predict fraud across their web & mobile banking system

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



Visually categorize damage & instantly issues quote



Optimize cardiac care in high volume remote regions



Predict and target first-time buyers in the US



Surface hidden insights to optimize fantasy football outcomes



AI-powered advertising engagement

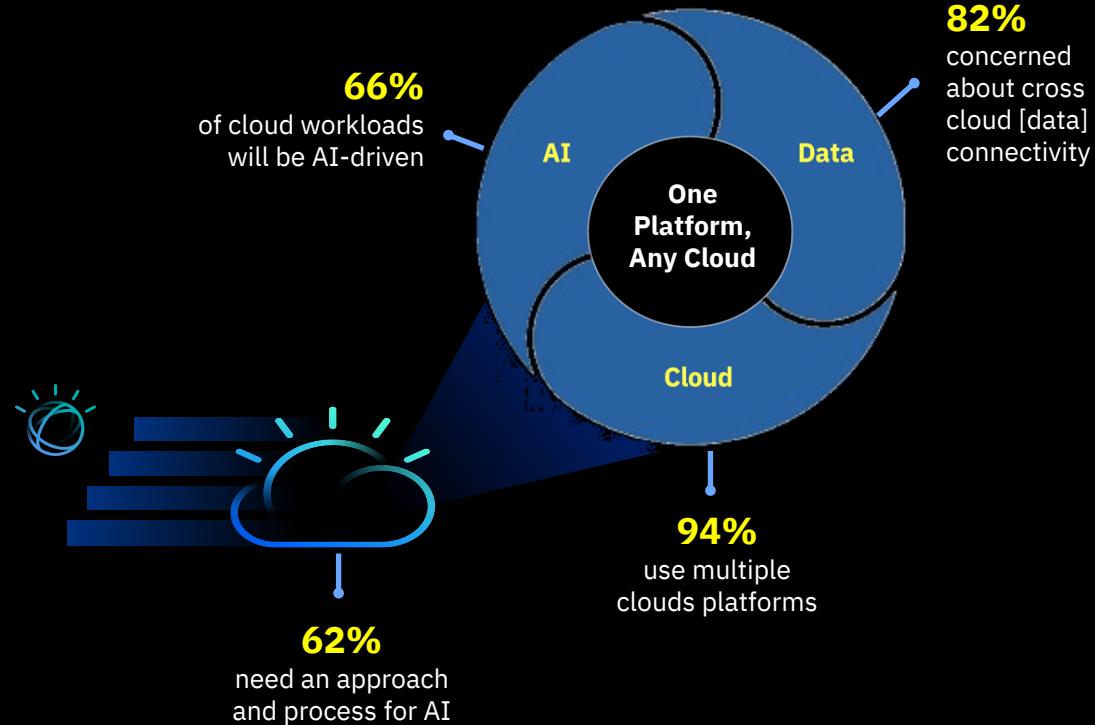
An Information Architecture for AI

Eliminate data silos, connect all data, on any cloud(s)

Automate and govern a unified data & AI lifecycle

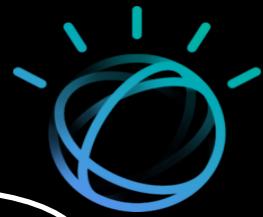
Operationalize AI with trust and transparency

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

Use Case

Articulate Use Case - Source of Value

Data

Unlock Data &
Break Down
Silos

Skills

Build an open,
collaborative and
Data Science team

Tools

Apply latest AI
Technologies
and Techniques

Agile

Create an Agile process to
iterate use case development,
Winning with AI is based on
Rate and Pace of projects

Integration & Trust

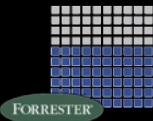
Integrate AI in your
business workflow and
Applications



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%

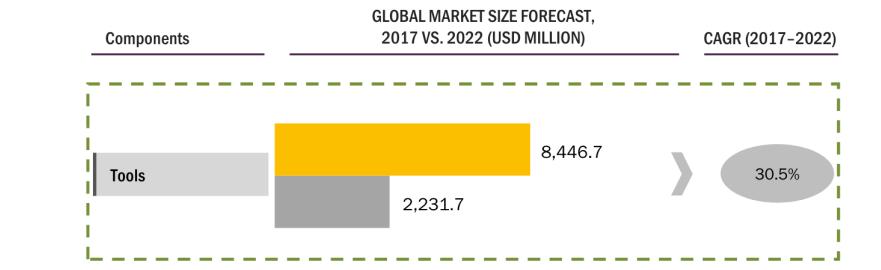
find operationalizing, sustaining and scaling AI challenging

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

FIGURE 6 ARTIFICIAL INTELLIGENCE PLATFORM MARKET SNAPSHOT, BY COMPONENT (2017 VS. 2022)



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



Decision Automation



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 constraints

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.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants



DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

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

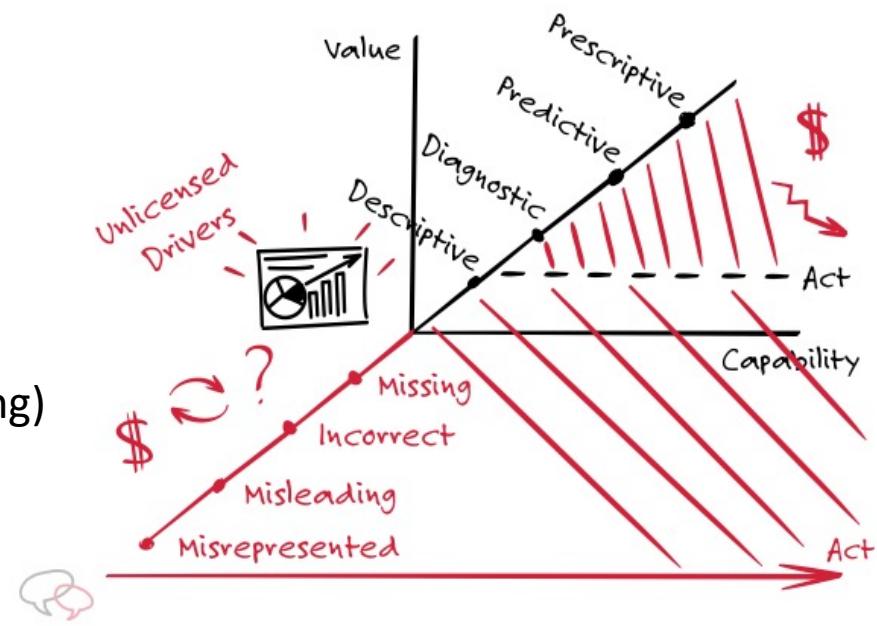
Machine Learning (incl. Deep Learning)

Prescriptive:

What should we do ?

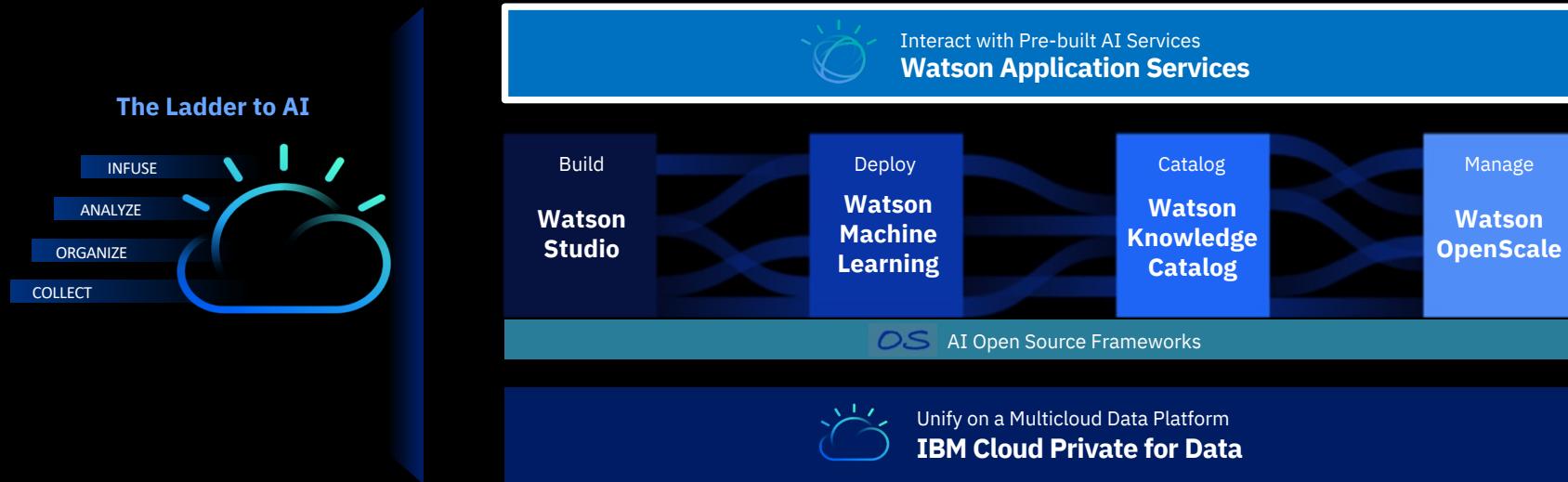
Actionable Insights

Optimization (Operational Research, Constraint Programming)



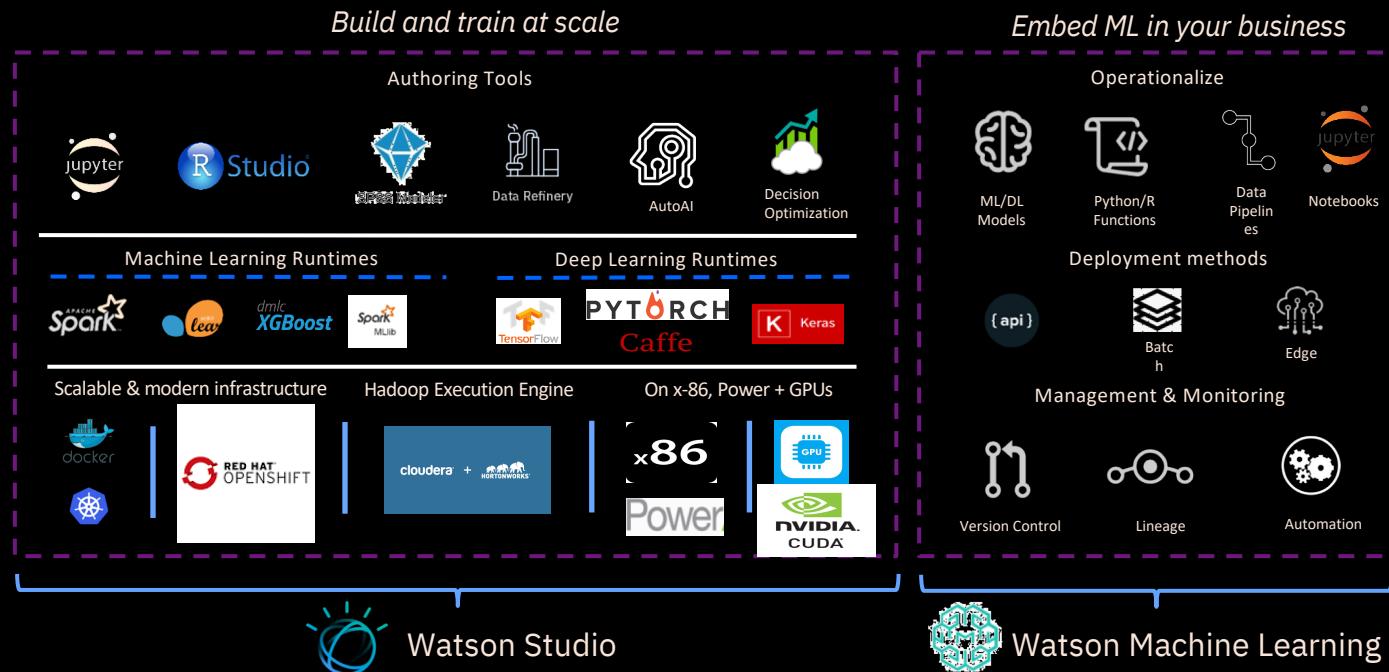
IBM AI Portfolio

Everything you need for Enterprise AI, on any cloud



Watson Studio

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



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

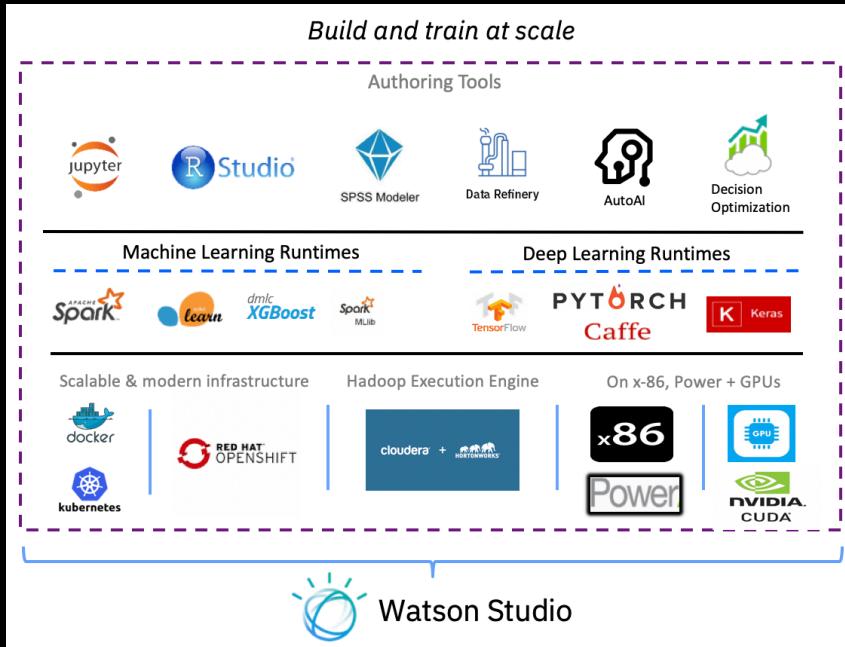
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

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



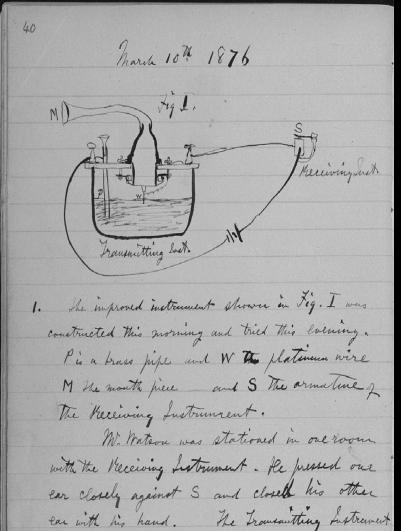
What is a Notebook?



Pen and Paper

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

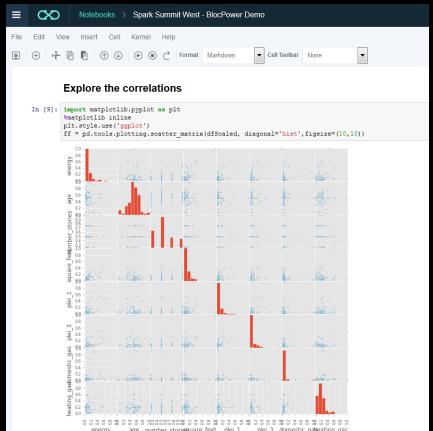
- Expressive
- Cumulative
- Collaborative



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

The screenshot shows the IBM Watson Studio interface. At the top, there's a dark header bar with the text "IBM Watson Studio". Below it, the main area has a light blue background. On the left, there's a vertical sidebar with some icons and text, though its content is mostly illegible due to blurring. The main content area starts with a section titled "New connection". Under this, there are two main categories: "IBM services" and "Third-party services", each containing a list of various data sources with their respective icons.

Category	Service	Icon
IBM services	BigInsights HDFS	Cloud storage icon
	Cognos Analytics	Analytics cube icon
	Db2 Big SQL	Db2 icon
	Db2 on Cloud	Db2 icon
	PureData for Analytics	IBM logo icon
Third-party services	Amazon Redshift	Amazon S3 icon
	Dropbox	Dropbox icon
	Hortonworks HDFS	HDFS icon
	Microsoft SQL Server	Microsoft SQL Server icon
	Amazon S3	Amazon S3 icon
FTP	FTP icon	
Looker	Looker icon	
MySQL	MySQL icon	

- 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

The screenshot shows a Jupyter Notebook titled "Train and deploy a heart disease prediction model using XGBoost and IBM Watson Machine Learning APIs". The notebook includes a diagram of a neural network with a heart icon in the center, and text explaining the process of training and deploying the model using Python 3.5, XGBoost 0.6, and Scikit-Learn 0.17. A sidebar on the right shows a conversation between ARMAND RUIZ and MANISH GOVAI.

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

The screenshot shows the SPSS Modeler Lab interface, which is a visual tool for data mining. It displays a complex network of nodes and connections, including various modeling techniques like Association Rules, Auto Classifier, and Bayesian Network, all interconnected by lines representing data flow.

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

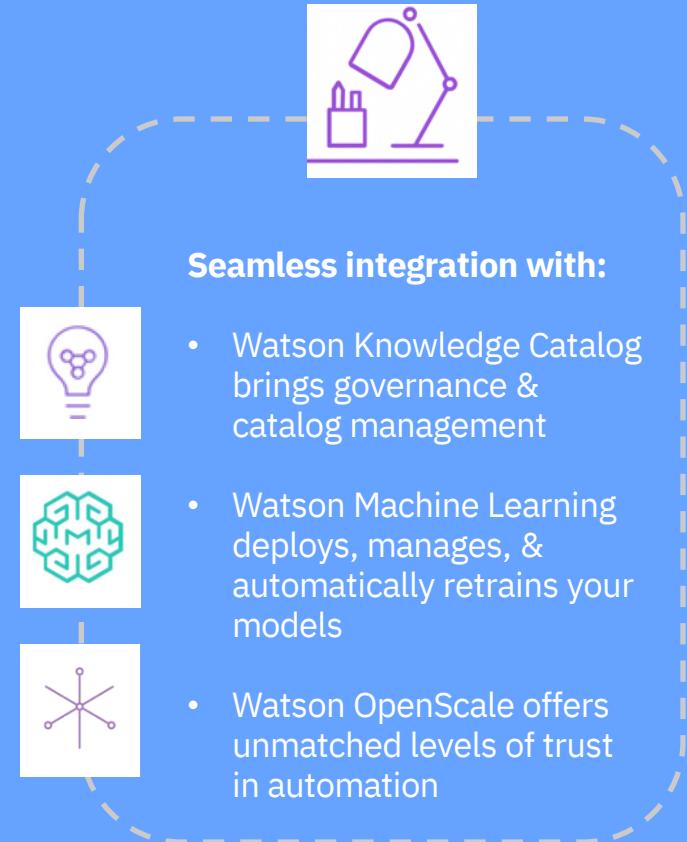
Connect to and analyze your data without moving a single byte through dozens of connectors and multiple deployment options

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



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...

The screenshot shows the IBM Watson Studio Data Refinery interface. At the top, there's a navigation bar with 'IBM Watson Studio', 'Upgrade', a notification bell, and 'Emmanuel GENARD's Account'. Below the header, the project path is 'My Projects / EDHEC_BBA / cars.csv'. The main area has tabs for 'Preview' (which is selected), 'Profile', and 'Lineage'. A message indicates 'Schema: 9 Columns' and 'Preview: 406 rows - Last refresh: 45 seconds ago' with a 'Refresh' button. On the right, there's a 'Refine' button. The data preview table has columns: mpg, cylinders, engine, horsepower, weight, acceleration, year, origin, and name. The first three rows of data are visible:

mpg	cylinders	engine	horsepower	weight	acceleration	year	origin	name
Type: String								
18	8	307	130	3504	12	70	American	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	American	buick skylark 320
18	8	318	150	3436	11	70	American	plymouth satellite



Data Refinery



IBM Watson Studio

My Projects / EDHEC_BBA / cars.csv / Refine data

Operation Code an operation to cleanse and shape your data

Search operations Convert column type

	mpg	cylinders	engine
Filter	Decimal	Integer	Decimal
Math	18	8	307
Remove	15	8	350
Rename	18	8	318
Sort ascending	16	8	304
Sort descending	17	8	302
Substitute	15	8	429
Text	14	8	454
CLEANSE	14	8	440
Convert column value to missing	14	8	455
Extract date or time value	15	8	390
Remove duplicates	NaN	8	350
Remove empty rows	NaN	8	351
Replace missing values	NaN	8	383
Replace substring	NaN	8	360
ORGANIZE	15	8	383
Aggregate	14	8	400
Concatenate	22	6	455
Conditional replace	18	6	198
Join	21	6	200
Sample	10	8	199
Split column	10	8	360
	11	8	307
	11	8	318

1 STEPS

Details Help

Edit

Data Source cars.csv

Convert column type AUTOMATIC

Automatically converted one or more columns to inferred data types.

DATA REFINERY FLOW DETAILS

LOCATION EDHEC_BBA

DATA REFINERY FLOW NAME cars.csv_flow

Enter a description of the Data Refinery flow

STEPS 1

DATA REFINERY FLOW OUTPUT

LOCATION EDHEC_BBA/Data assets

DATA SET NAME cars.csv_shaped.csv

More ▾

File format: CSV



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

My Projects / EDHEC_BBA / RefineCitibike

RefineCitibike
No description

Scheduled to run Edit Environment definition Edit Default Data Refinery XS

Associated Asset DATA REFINERY FLOW 201701-citibike-tripdata.csv_flow 23 Steps

INPUT 201701-citibike-tripdata.csv delimited OUTPUT 201701-citibike-tripdata_cleansed.parquet parquet

Runs

Start Time	Status	Duration	Started By	Action
Oct 9, 2019, 11:41:07 AM	Completed	2 minutes 44 seconds	Emmanuel GENARD	

DATA REFINERY FLOW DETAILS

LOCATION EDHEC_BBA

DATA REFINERY FLOW NAME 201701-citibike-tripdata.csv...
Enter a description of the Data Refinery flow

STEPS 23

DATA REFINERY FLOW OUTPUT

LOCATION EDHEC_BBA/Data assets

DATA SET NAME 201701-citibike-tripdata_cle...
More ▾

File format: PARQ

Start_Station_ String
W 82 St & Central
Cooper Square & I
Central Park West
Broadway & W 60
Broadway & W 37
York St & Jay St
Central Park West
2 Ave & E 31 St
5 Ave & E 29 St
W 43 St & 6 Ave
E 15 St & 3 Ave
W 43 St & 10 Ave
W 74 St & Columb...
W 43 St & 10 Ave
W 54 St & 9 Ave
Washington Pl & E
E 72 St & Park Ave
Peck Slip & Front...
E 2 St & Avenue B
MacDougal St & P

Rename column
Renamed column Bike ID to Bike_ID

Rename column
Renamed column User Type to User_Type

Rename column
Renamed column Birth Year to Birth_Year

Calculate
Subtracted 2019 from Birth_Year into Age

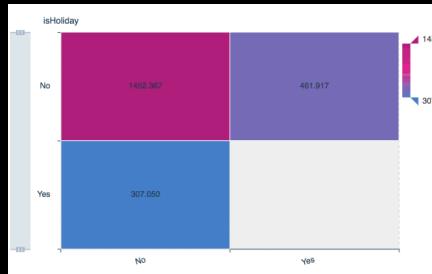
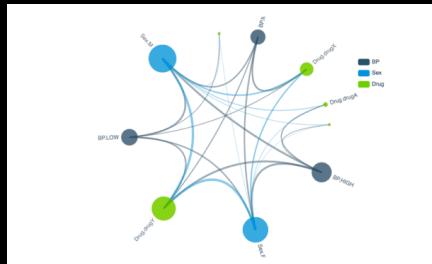
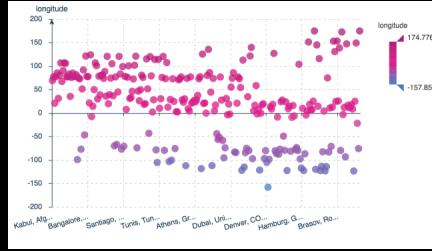
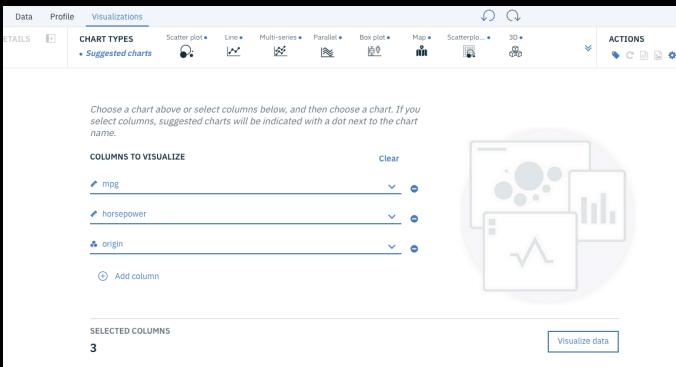
Math
Got the absolute value of Age

Extract date or time value
Extracted date from Start_Time column into Start_Date

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: Matplotlib quick review, R & Shiny
 - 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)



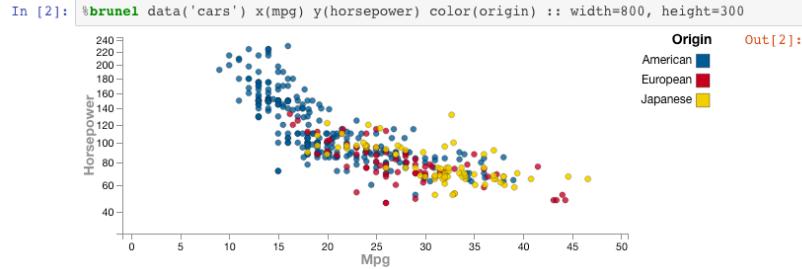
Brunel visualizations

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 visualizations can be zoomed and scrolled but are not actionable

Scatter plots

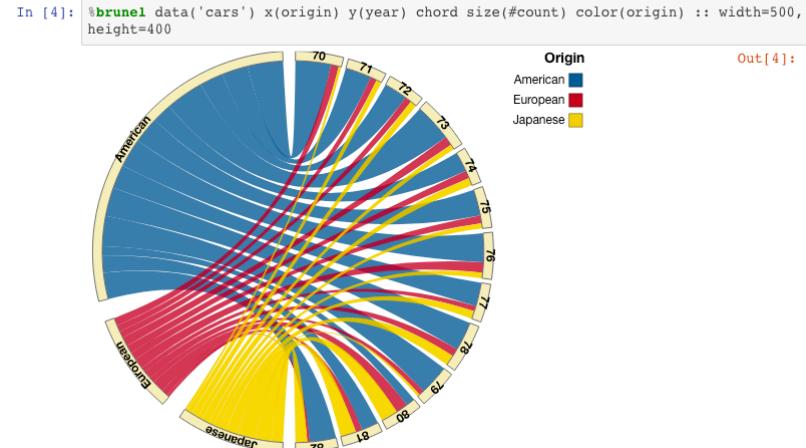
Run the next cell to show the relationship between the miles per gallon and the horsepower of the vehicles in a scatter plot. The color identifies the origin of the vehicles.



Put your cursor over the chart and scroll to zoom in and out. When you zoom in, you can pan across the chart by clicking and dragging.

Chord plot

Run the next cell to show a chord plot that correlates the origin and number of cars produced per year. The x and y commands specify that the origin is mapped to the year of manufacture. The size of the segments is based on the number of cars. The color is based on the origin of the cars.



PixieDust

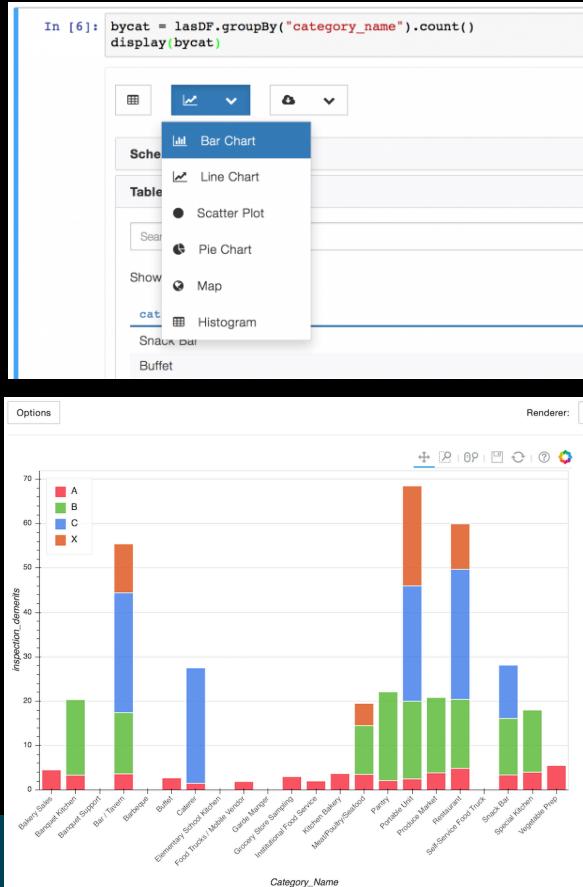


PixieDust is a visual data exploration framework for Jupyter and Spark

- 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

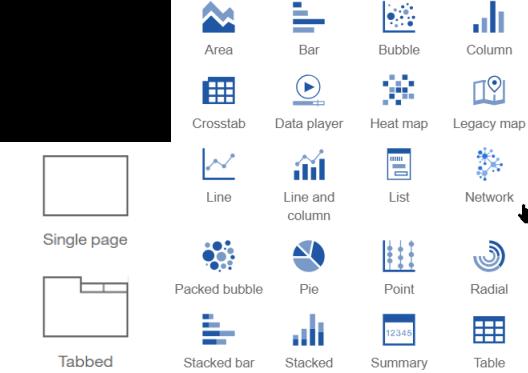
- Tooling part of Watson Studio
- Leverages Watson Studio Data Sources

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 Cognos API





demo

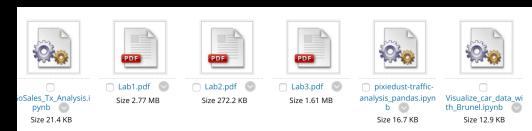
Time for Lab 2: Data Refinery and Visualizations

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!

