#ibmdevconnect

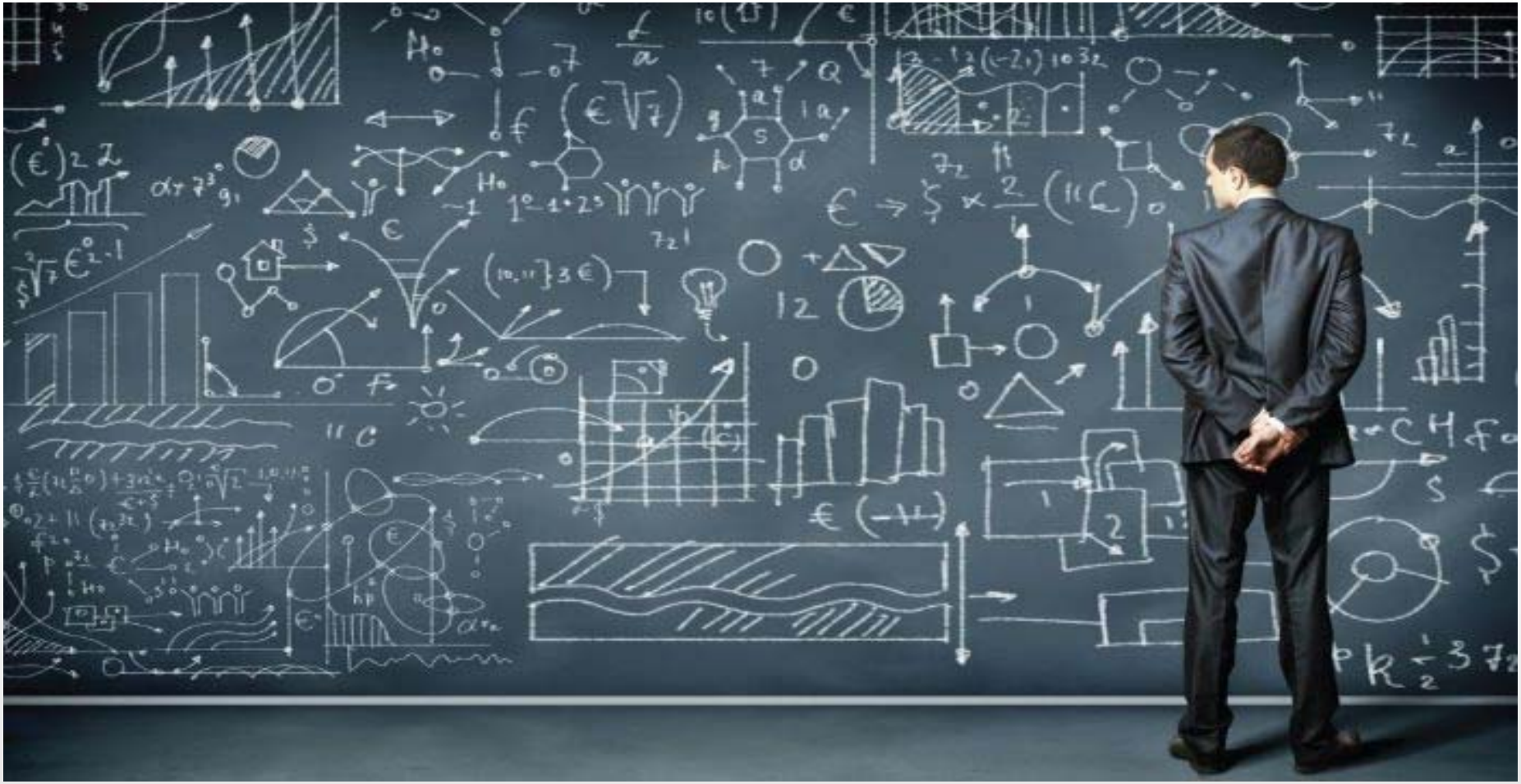
Data Science Experience

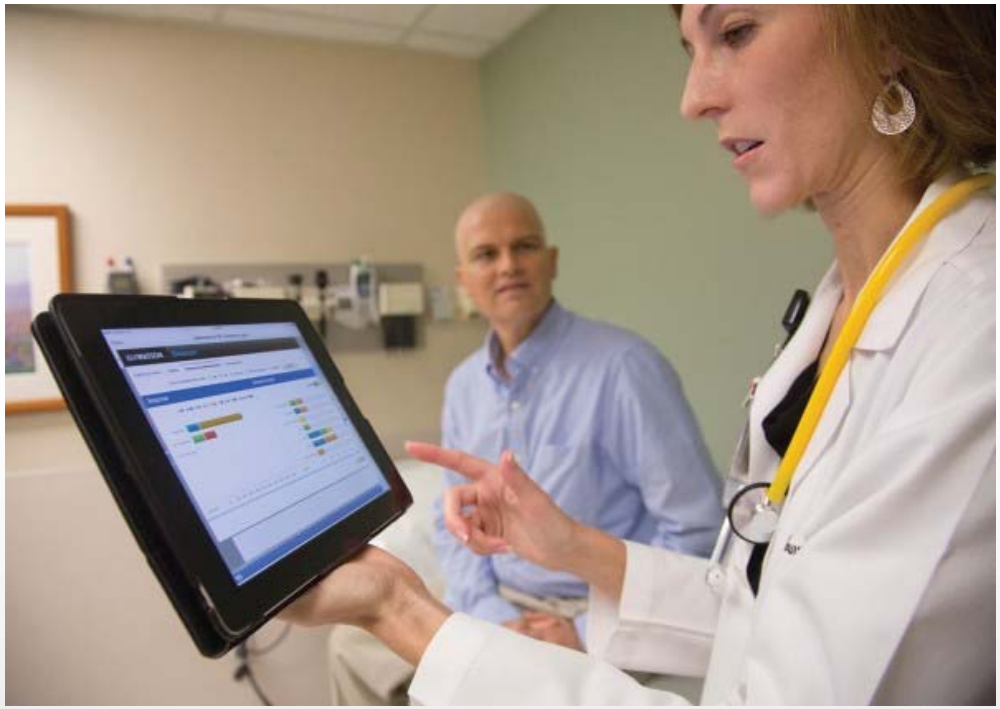
Putting Data to work !!

Rajesh K Jeyapaul
Advocate, startup Mentor & Solution Architect
IBM

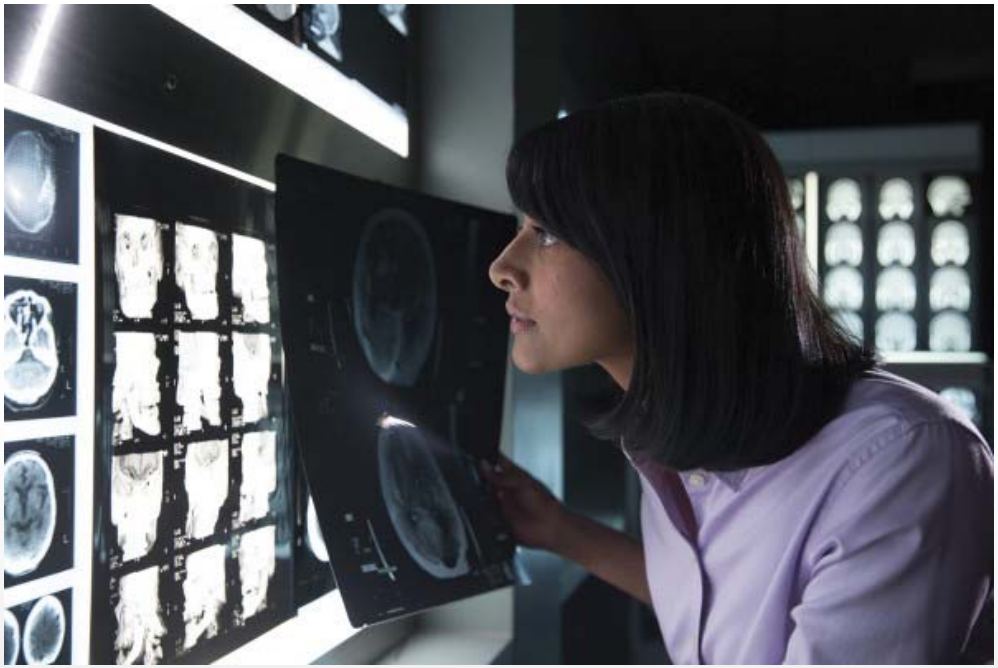
Agenda

- Understand the eco system around Data
- Role of a Data Scientist – What is a Data Model
- Importance of Machine Learning and Deep Learning





A leukemia doctor at M.D. Anderson, Courtney DiNardo, used IBM's Watson system while consulting with a patient



IBM claims that Watson’s diagnostic capabilities would be boosted by data obtained from Merge Healthcare, a medical imaging management company that IBM bought for about $1 billion.

Now Data-Driven Professionals Are At The Forefront

- Business professionals
- App developers
- Data engineers
- Data scientists

How to collect and Where to Store ?

How to get a meaningful Insight ?

How can I take quick decision with the Data?

What would be the recommended approach to (Big) Data Analytics ?

IBM

3 Basic steps:

Prepare

Store

Analyze

IBM

Prepare your data

- Access to Data
- Connectors to load from external resource
- Migrate from on-premise to cloud

IBM

Store your data

- RDBMS to every type of NSQL

Store - Database Option

- When to use SQL and when to use NSQL ?
- What is their difference ?
- Can you name some open source databases ?
  - Firebird (relational)
  - CUBRID (relational)
  - MySQL (relational)
  - MongoDB (NSQL)
  - Cassandra (NSQL)

IBM

Analyse your data

- Visualize - Quality of data
- Statistics
- Find Pattern and create Model
- Leave it to system to Identify and Predict for further Actions

How To ?

Prepare

Store

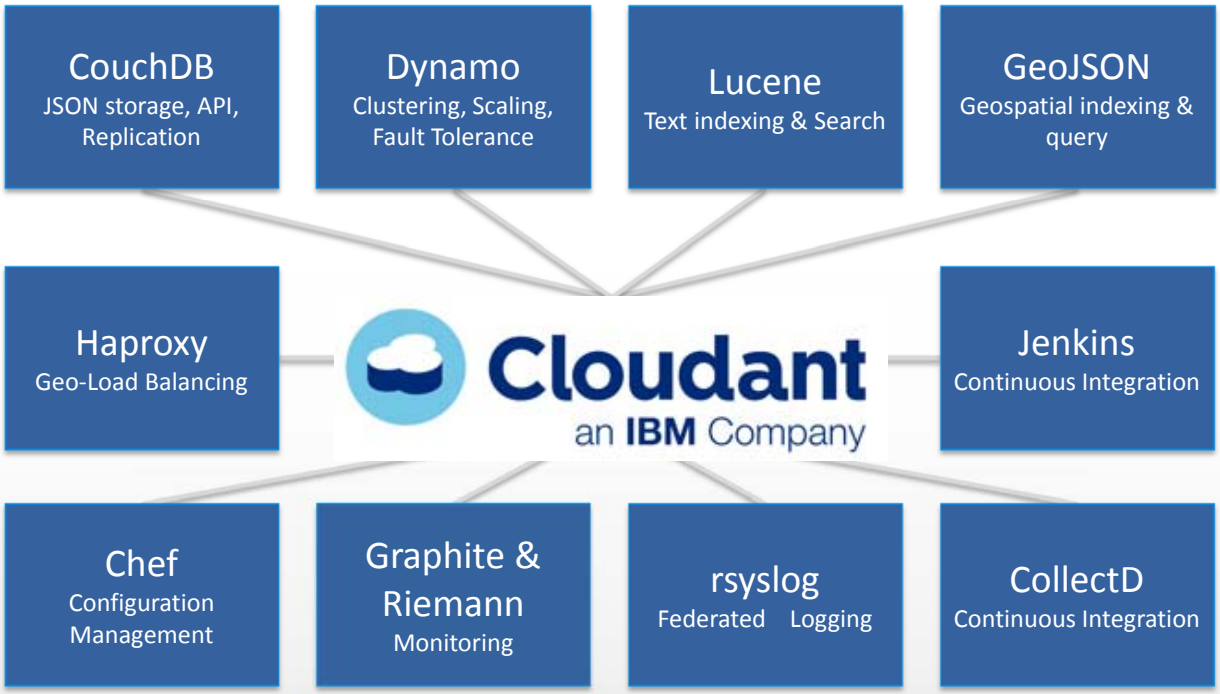
Analyze

dashDB (SQL) and Cloudant (NSQL)

- dashDB – transactional and Analytical

- *Data warehouse* - dashDB include features such as in-memory data processing and columnar tables for online analytical processing (OLAP).
- *Relational database* - The managed service transactional plans deliver fast query processing with enterprise-level performance and capabilities for online transactional processing (OLTP).

Cloudant's DNA



Cloudant combines the **best Open Source technology & thinking** to create the most **scalable**, **flexible**, **always-on DBaaS** for **big mobile** and the **Internet of Things**

IBM Bluemix Lift

- Lift makes it easy to quickly, securely, and reliably migrate your data from on-premises sources to the cloud
- It eliminates source database downtime by capturing changes during migration and automatically applying them to your target database.
- Lift CLI currently supports the following source and target combinations:
  - Migration from IBM PureData® for Analytics to IBM dashDB™
  - Migration from CSV files to IBM dashDB
  - Migration from IBM DB2® on premises to DB2 on Cloud
  - Migration from CSV files to DB2 on Cloud

Data Connect

- Research from Forrester found that 68 percent of simple BI requests take weeks, months or longer to derive insight from the data.
  - So this entails that the enterprises must find ways to transform line of business professionals into skilled data workers, taking some of the burden off of IT.
  - It means business users should be empowered work with data from many sources

- Data Connect enables you to find data, shape it, and deliver it to applications and systems.
- Allow technical and non-technical users to draw value from data quickly and easily.
- Ensure data quality with simple data preparation and movement services in the cloud.
- Integrate with leading cloud data services to create a seamless data management platform.

- https://www.youtube.com/watch?v=Q6Tuo48qG3w

And We Provide a Bridge to the Watson Data Platform for your Existing Investments

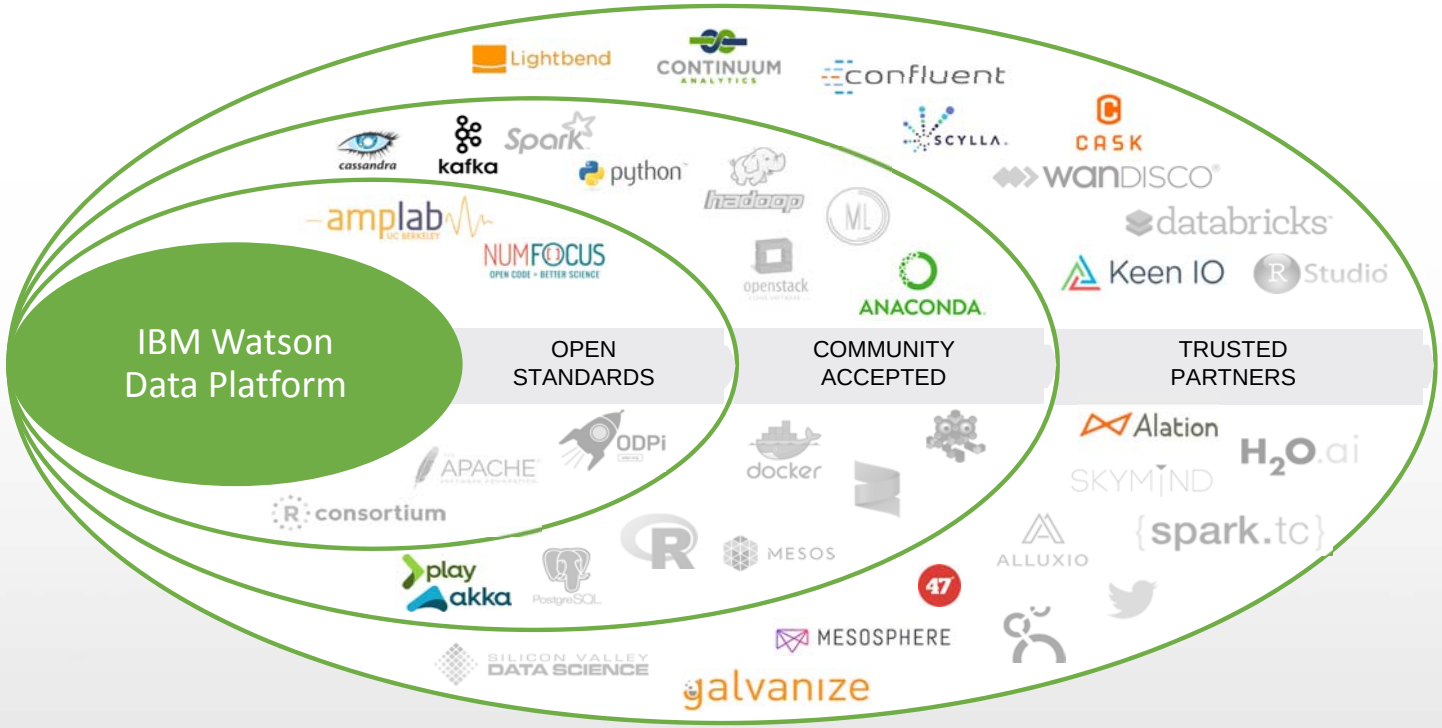
Managed Public Cloud Service	Software-defined	Appliance	Custom Deployable Software	Hadoop / Spark Environment
dashDB	dashDB Local	PureData for Analytics	DB2	BigInsights, BigSQL

Built on a common and fluid analytics SQL engine
enabling true hybrid analytic data stores with portability

#ibmdevconnect

IBM Watson Data Platform Partner Ecosystem

The Open Community To Innovate Faster With Data



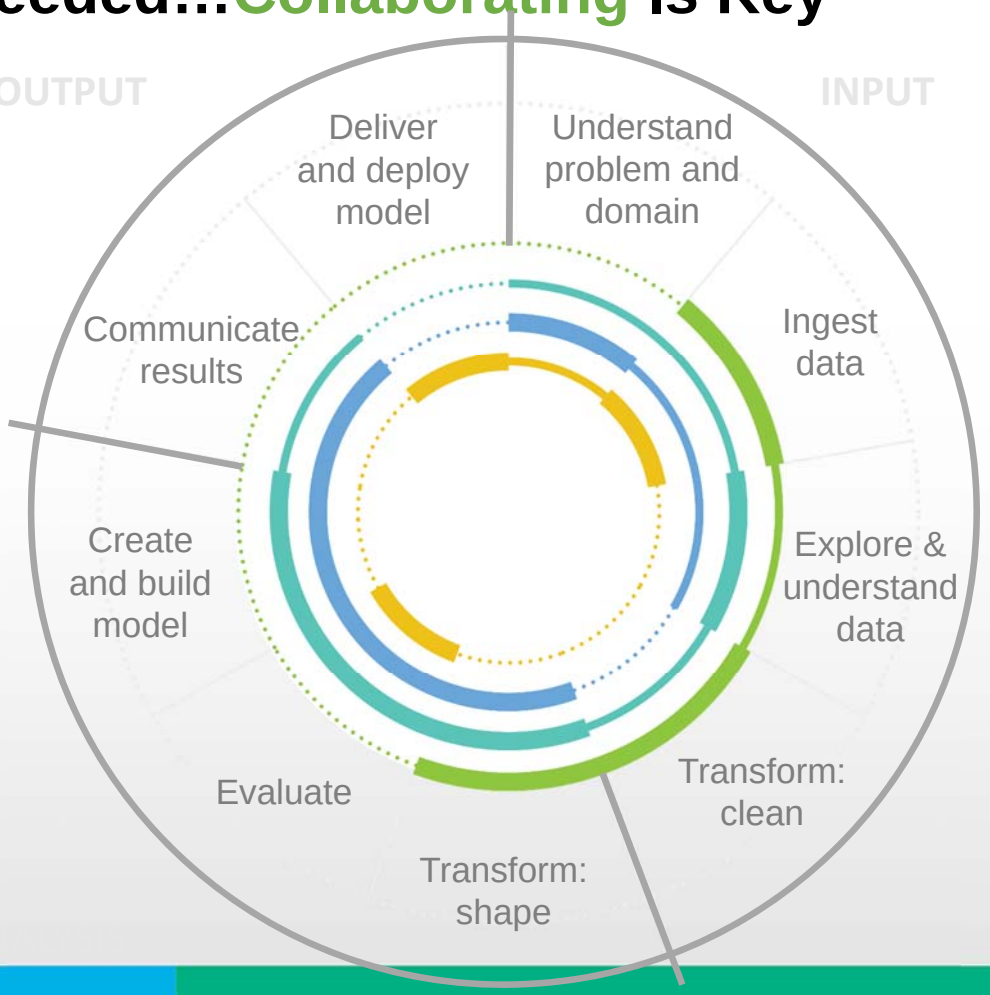
IBM

Importance of Machine Learning and Role of Data Scientist

- Who does what ?
  - Business & data Analyst , Data Scientist , Developer
- Role of a Data Scientist ?

Understanding business problem , so that the relevant data can be acquired ?
Preparing the data for Analytics ?
Estimating the quality of Data ?
Deriving statistical information out of Data ?
Model the Data ?

Multiple Skills Needed…Collaborating Is Key

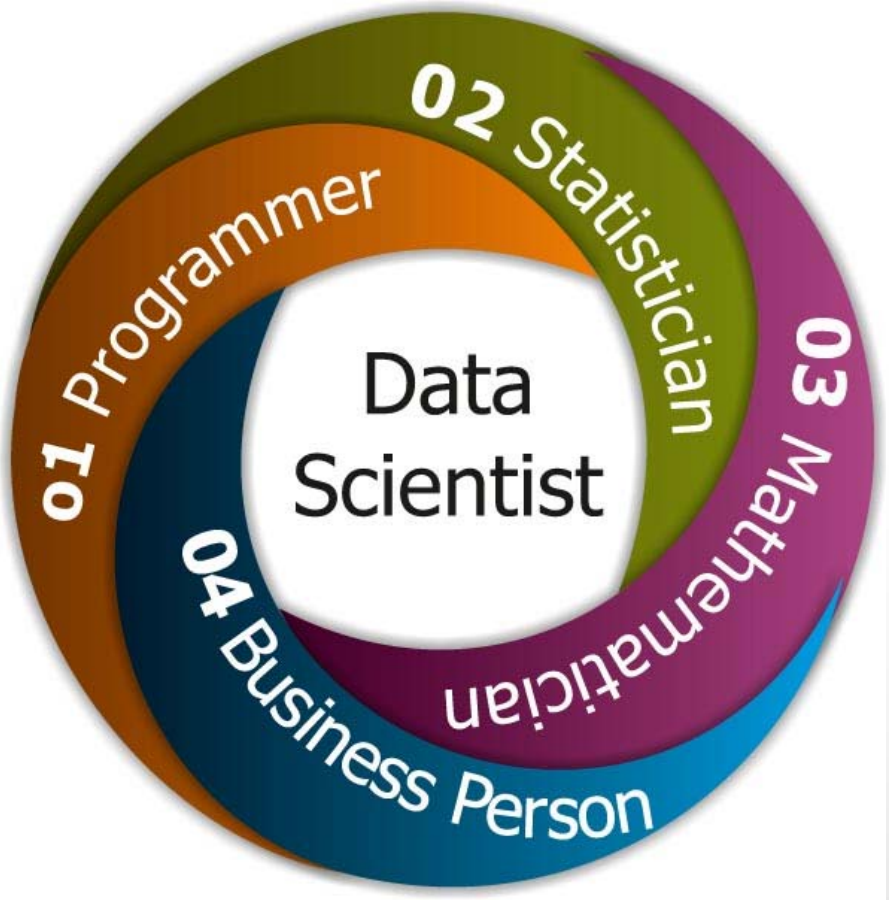


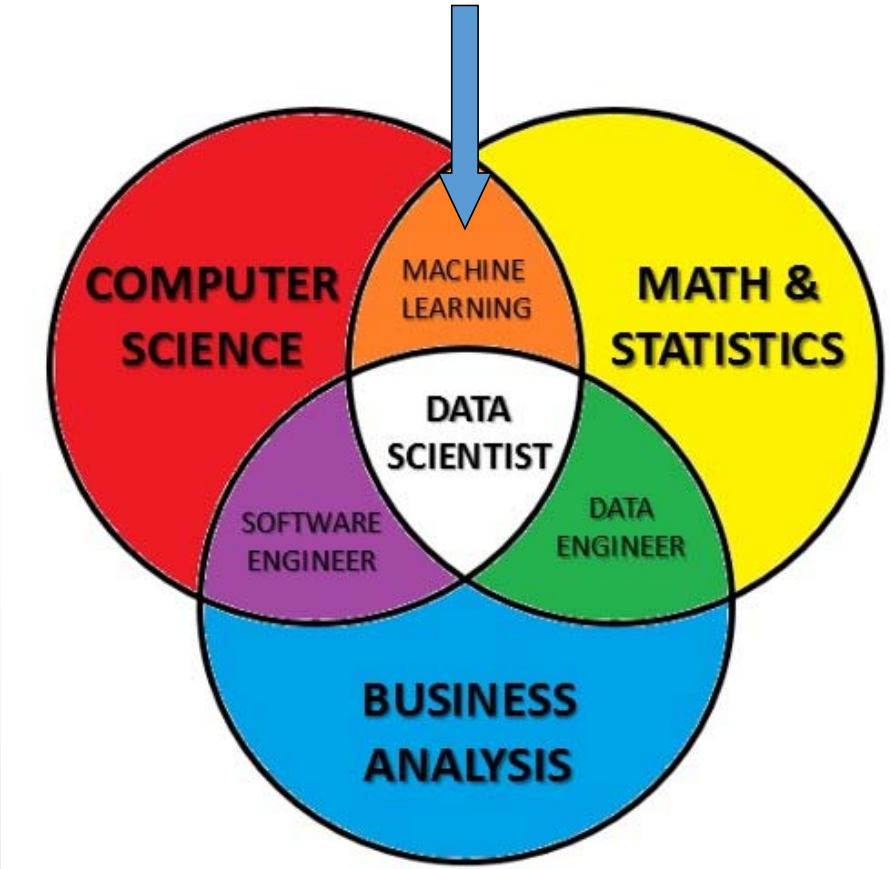

Data Engineers


Data Scientists


Business Analysts


App Developers


01 Programmer
02 Statistician
03 Mathematician
04 Business Person
Data Scientist


COMPUTER SCIENCE
MACHINE LEARNING
MATH & STATISTICS
DATA SCIENTIST
SOFTWARE ENGINEER
DATA ENGINEER
BUSINESS ANALYSIS

Machine Learning – Key for Data Scientist

IBM

Categories of Machine Learning

Supervised
Machine needs to be told what The correct label for a particular input
"Here is a spammy email"
Label - Spam

UnSupervised
Machine identifies similar examples In the dataset without knowing the labels
"news.google.com"

Semi Supervised
Only some examples have labels

"Detecting lawbreakers"

Re-inforced
Decision to maximize rewards

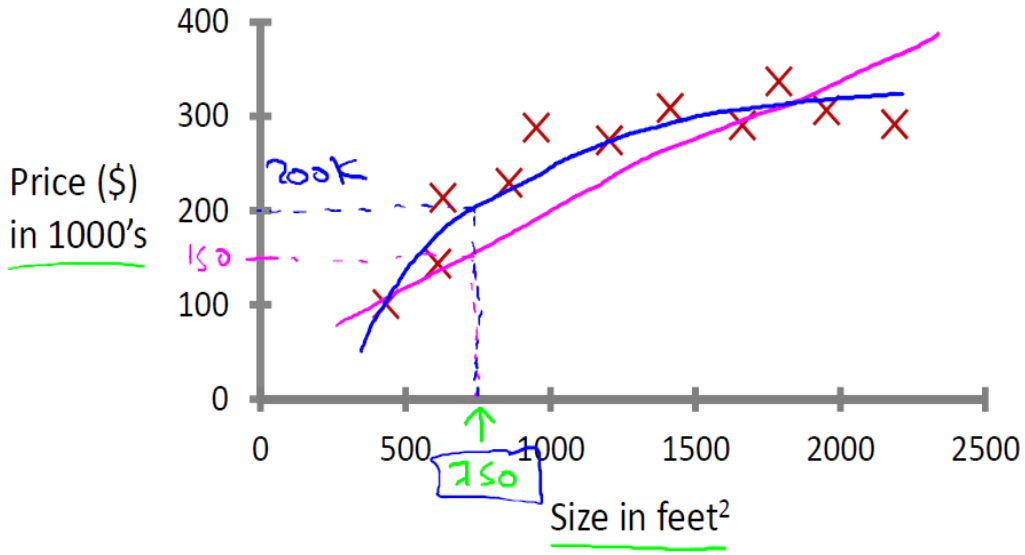
"AlphaGo"

IBM

Supervised Learning

$Y=f(x)$

Housing price prediction.



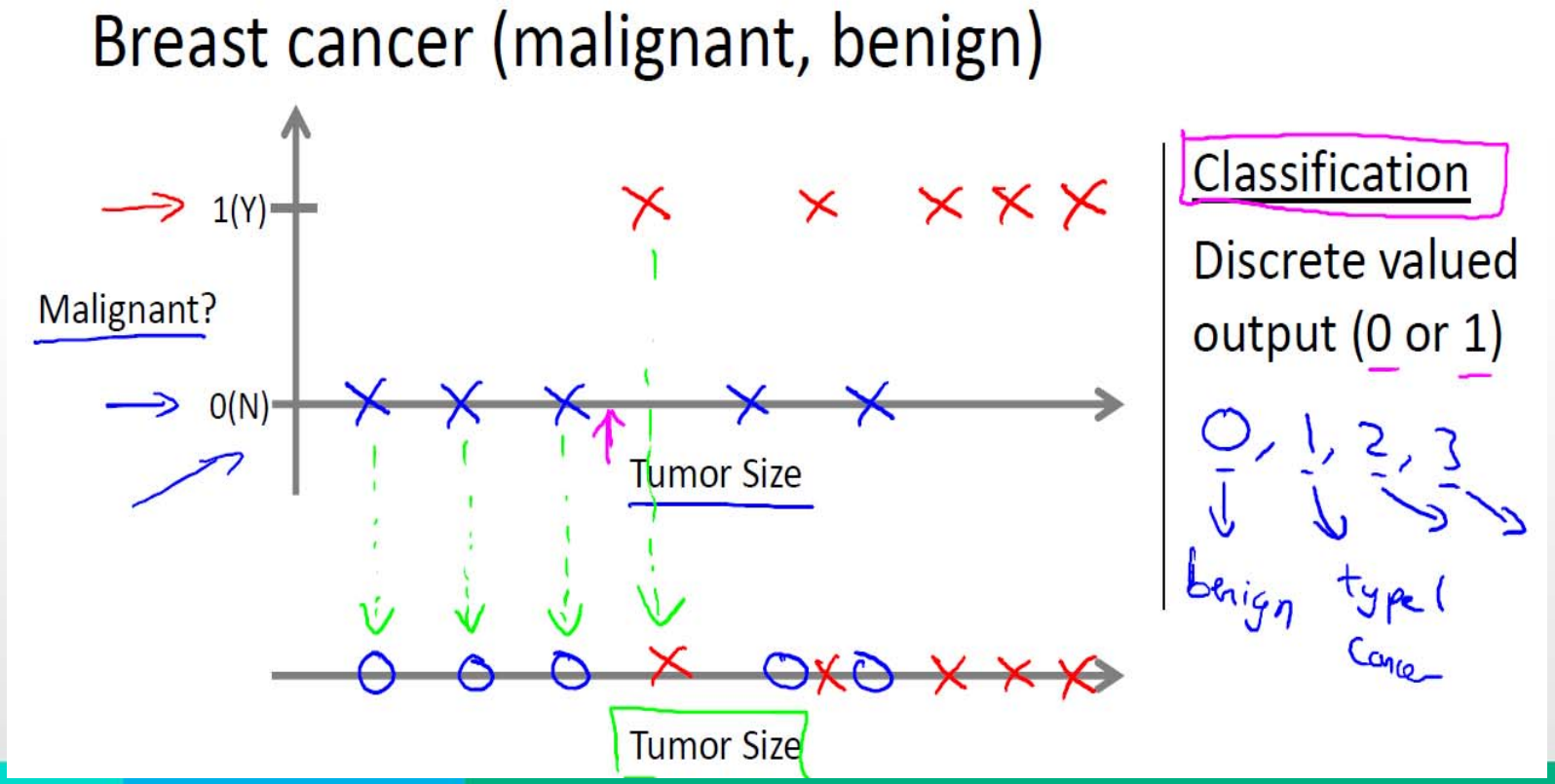
Supervised Learning

"right answers" given

Regression: Predict continuous valued output (price)

IBM

Supervised Learning



You're running a company, and you want to develop learning algorithms to address each of two problems.

1000's

Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.

Problem 2: You'd like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

0 – not hacked

1 – hacked

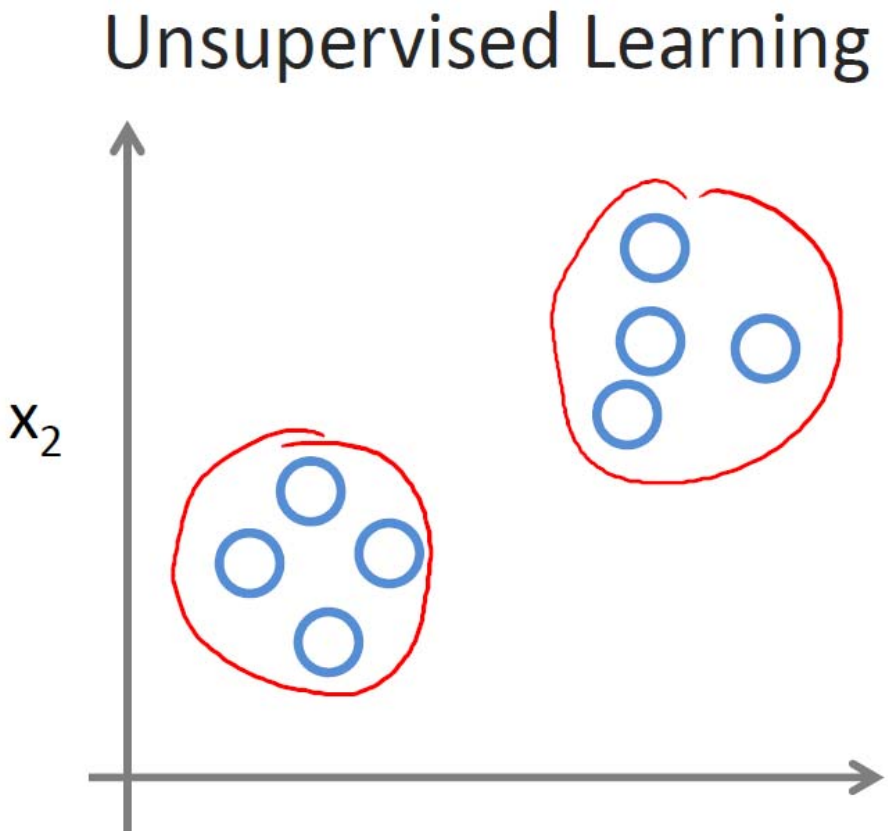
Should you treat these as classification or as regression problems?

- Treat both as classification problems.
- Treat problem 1 as a classification problem, problem 2 as a regression problem.
- Treat problem 1 as a regression problem, problem 2 as a classification problem.
- Treat both as regression problems.

IBM

Unsupervised Learning

- Only X , No Y



- people that buy X also tend to buy Y
- grouping customers by purchasing behavior

Getting started with Machine Learning – 7 steps

- Define
- Get Data
- Explore
- Choose Techniques
- Get Tools
- Model
- Deploy

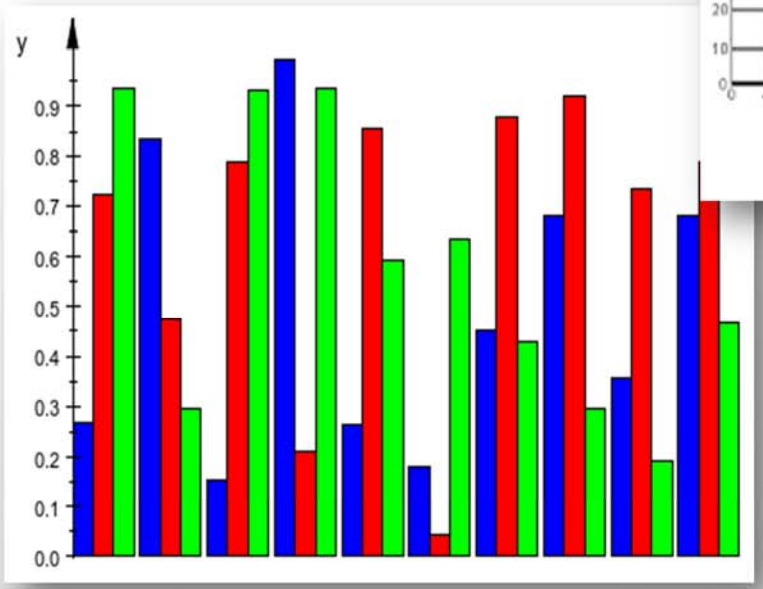
M
O
N
I
T
O
R

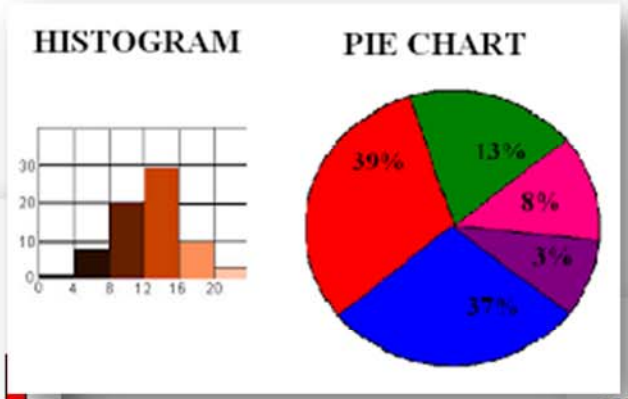
IBM

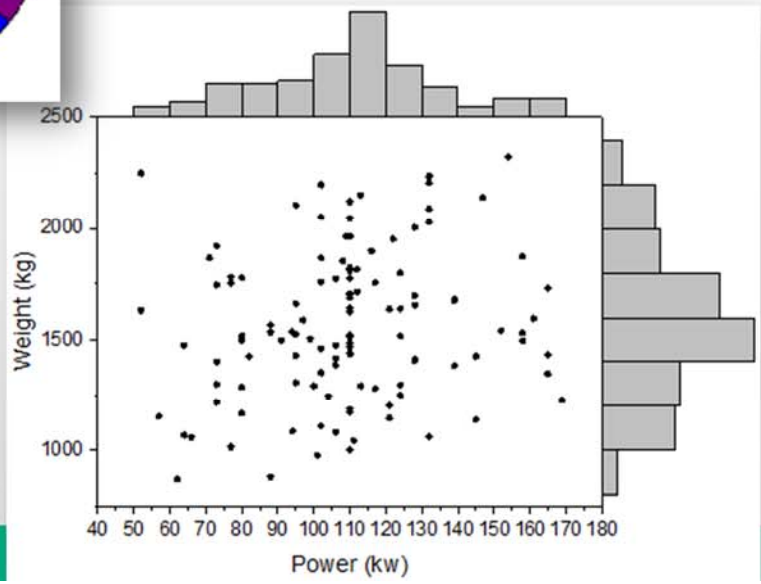
Getting started with ML - Define

- Scenarios : Derive Feature Vectors
  - Understanding Food Inflation in India
  - Understanding Diabetics based on Urban and Rural statistics
  - Finance – Fraud Detection
  - Healthcare – Predicting lifestyle based disease outcome
  - Heart Emotion – Stress or normal

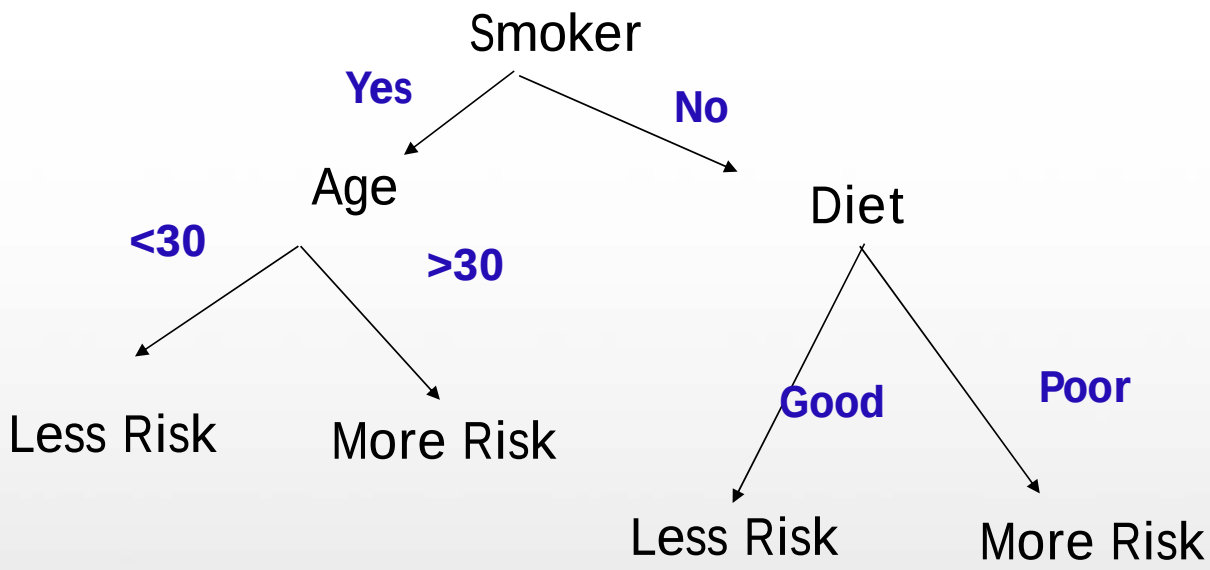
Getting started with ML – Collect & Explore







Getting started with ML - Techniques



Supervised

- Decision Tree
- Random Forest
- Neural Network

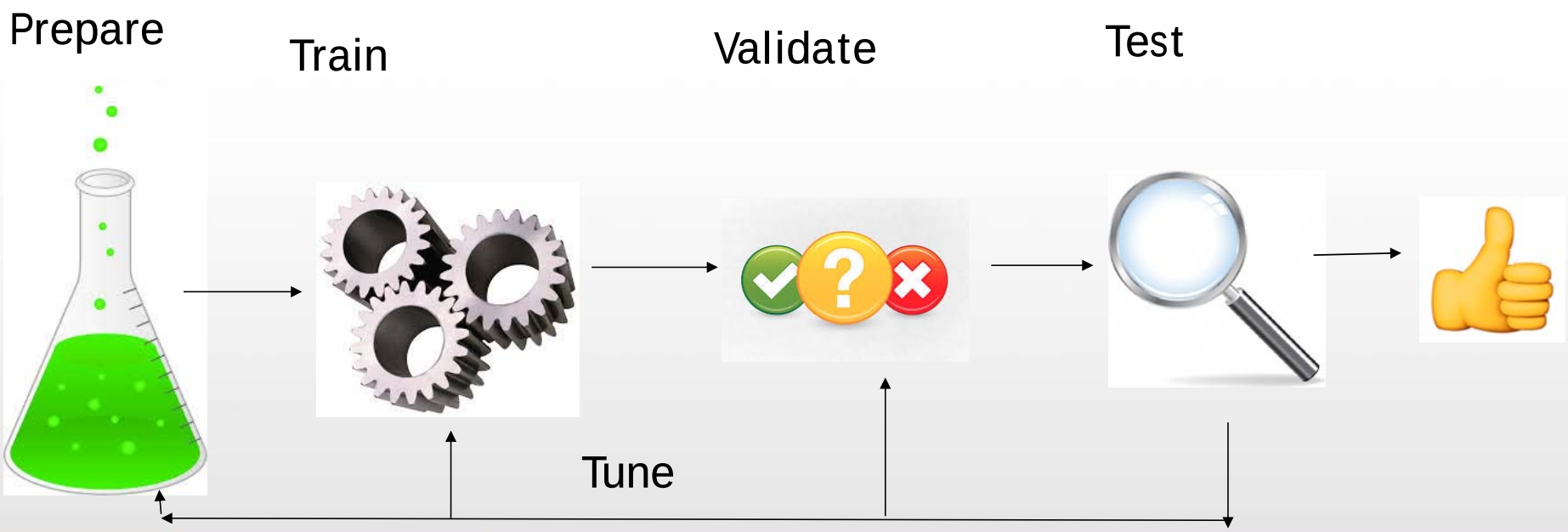
ML Algorithmns

Algorithms	Tasks
Clustering	Genre classification, spam labeling
Decision trees	Semantic type (Entity & Event) or ontological (inter relationship of entities) class assignment, coreference resolution (Ramesh visited Delhi.He went around parliament campus)
Naïve Bayes	Sentiment classification, semantic type or ontological class assignment
Maximum Entropy (MaxEnt)	Sentiment classification, semantic type, or ontological class assignment
Structured pattern induction (HMMs(hidden M Model), CRFs (cond. Randon field), etc.)	POS tagging, sentiment classification, word sense disambiguation

Getting started with ML - Tools

Type	Tools / Techniques
Supervised and Unsupervised Learning	Expander , Graph Mining Tools
Data Analysis and Interpretable Models	LPH , Glassbox , DataConnect , Data Lift
ML Platforms	Data Science Experience (DSX) , tensorflow
Hyper-parameter Optimization	Vizer

Getting started with ML - Model



Heart Emotion - Demo

Bluemix Data Science Experience

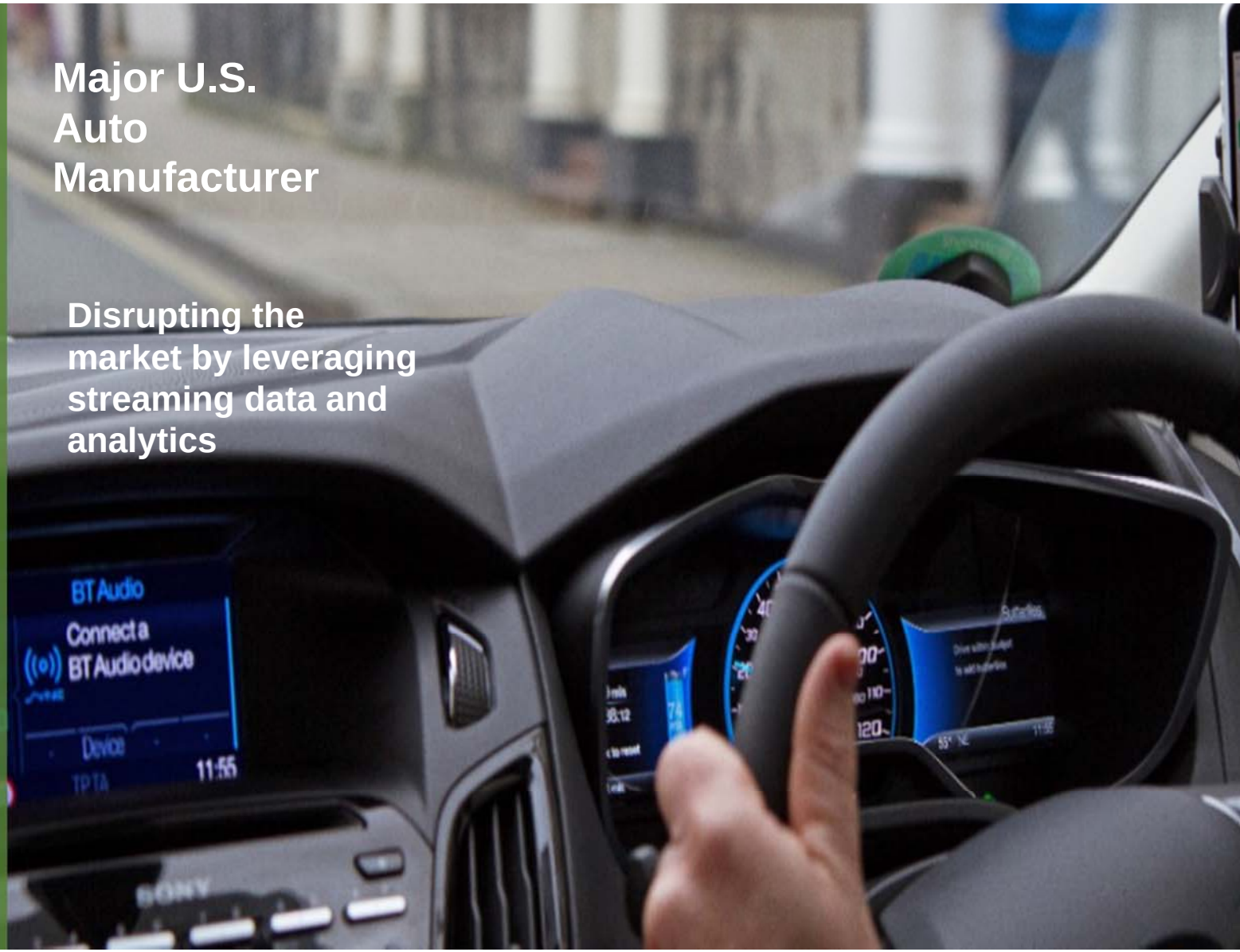
- Enables vital insight into where people go and what they buy, driving smarter optimization in crowd safety and service provision. Extreme scalability enables deeper analysis of near-real-time data.
- **Quote**
  - *“dashDB, SPSS and Watson Analytics enabled us to process, store and analyze huge volumes of data.”—Per Østergaard Jacobsen, External Lector and Project Manager, Copenhagen Business School*
- **Solution components**

- IBM® Bluemix®
- IBM dashDB™
- SoftLayer®
- IBM Machine Learning
- IBM Watson™ Analytics

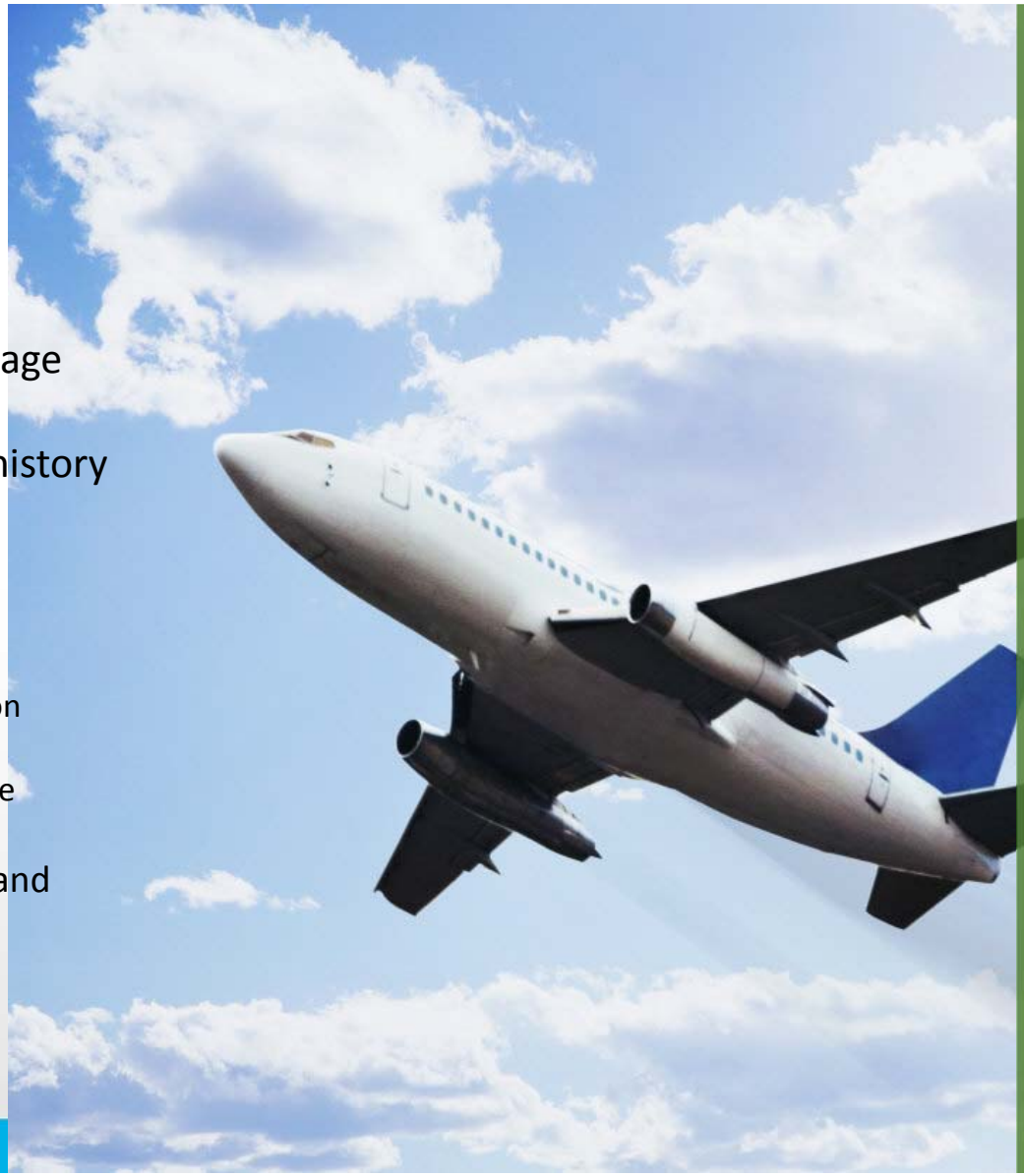
#ibmdevconnect

Major U.S. Auto Manufacturer

Disrupting the market by leveraging streaming data and analytics



Low-fare US Airline



Enhancing online booking conversions with real-time personalization

Presenting special offers, page designs, etc based on user demographic or purchase history

- Original architecture, based on MongoDB & SQLServer, was unable to respond in real-time
- Replaced MongoDB & SQLServer with Cloudant and DB2 on Cloud

Please give us your valuable feedback
http://ibm.biz/devconnect17

#ibmdevconnect

Stay Connected and continue coding !

Code & instructions available here
https://github.com/IBMDevConnect17/

Join developerWorks India Community
https://developer.ibm.com/in/

Check out the cool developer journeys
https://developer.ibm.com/code/

Join our Slack team and stay in touch with the experts
https://ibmdevconnect.slack.com
Send in your request to -
http://ibm.biz/slackrequest

Meetup

Join our Meetup groups
Mumbai :
https://www.meetup.com/Cloud-Mumbai-Meetup/
Hyderabad:
https://www.meetup.com/Hyderabad-Cognitive-with-Cloud
Bangalore :
https://www.meetup.com/IBMDevConnect-Bangalore