stuff
use
post
going
see
million
web
machine
Nice
cluster
year
forensics
know
learning
source
Internet
today
system
nice
tools
research
@TheSuggmeister
book
presentation
visualizations
code
using
need
released
get
video
list
TellVicEverything
bigdata
network
law
go
bad
last
Security
blog
help
much
@ebellis
python
find
via
make
data
Infosec
Python
take
next
tables
just
Google
good
look
Facebook
search
link
Need
Vic
website
BigData
better
Just
access
analytics
getting
Data
sure
Now
PCAP
DFIR
twitter
Vancouver
crawford
@wordpressdotcom
analysis
day
password
one
@s
http
love
available
build
online
Hey
@kylemaxwell
sta
read
paper
way
found
work
got
new
based
Using
big
Like
Microsoft
videos
used
Analytics
give
Hadoop
Visualization
visualization
Social
Open
infosec
Machine
Network
mining
passwords
security
Twitter
like
people
really
social
malware
Reading
Analysis
Great
police
Big
think
privacy
want
media
years
time
Learning
Thanks
Intelligence
New
software
open
now
Metasploit
talk
looking
script
tool
Gephi
rainbow

Data Science, Big Data and Analytics: Present and Future

Perspectives from Academia, Industry and Consulting

Zoran Bursac, PhD, MPH

Josh Callaway, MS, MPH

Fernando Lopez, MS, PhD Candidate

Data

Health care, business, technology -> data

Big data -> voluminous data sets (structured or unstructured)

Produced every day all around us

Analytics -> examining data to detect patterns

Big Data Analytics and Data Science

- Different sources, different sizes
- High variety, volume, velocity
- Online networks, web pages, audio/video, social media, logs
- Techniques-> machine learning, data mining, natural language processing, statistics
- Extraction, preparation, storage/warehousing, blending, analytics

Real-time Benefits

- Healthcare
- Banking
- Energy
- Tech
- Consumer
- Education

- Data sourcing -> Hadoop HDFS
- Data storing -> Oracle, MySql
- Data conversion -> Sqoop
- Data transformation -> Hive
- Exploratory analysis -> Elastic search, knime
- Model building -> R, SAS, Python, Julia
- Visualization -> Tableau, ggplot2
- Model execution -> Hadoop, Java, Spark, C#
- Version control -> Git, BitBucket
- IDE -> Rstudio
- Text for coding -> Jupyter Notebook, R Shiny

Current Trends and Common Data Science Tools
Process, Perform and Visualize

Free Open Source

- Hadoop->distributed processing of large data across clusters
- Hive->warehouse to manage large data in distributed SQL storage
- Kafka->real time pipeline of streaming data
- Pig->large data analytics
- R, Rstudio, ggplot-> analytics and data visualization
- Python, Julia -> high level programming with efficient algorithms and speed for large data processing
- Jupyter notebook -> manage documents such as code, explanatory and shared
- RapidMiner -> data preparation, machine learning and model deployment

Do you need to know all? NO

Hadoop

R

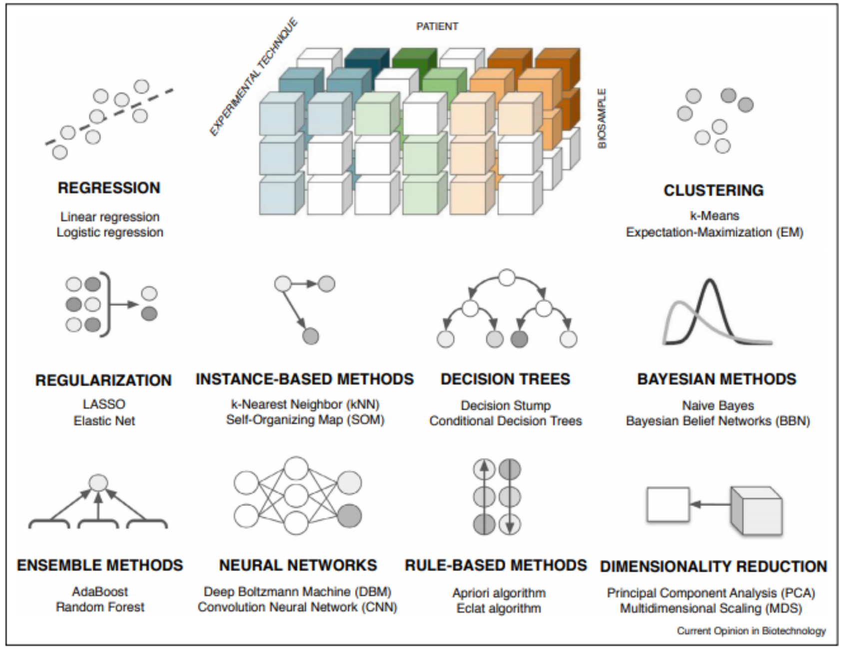
SQL

Python

Hive

Pig etc…

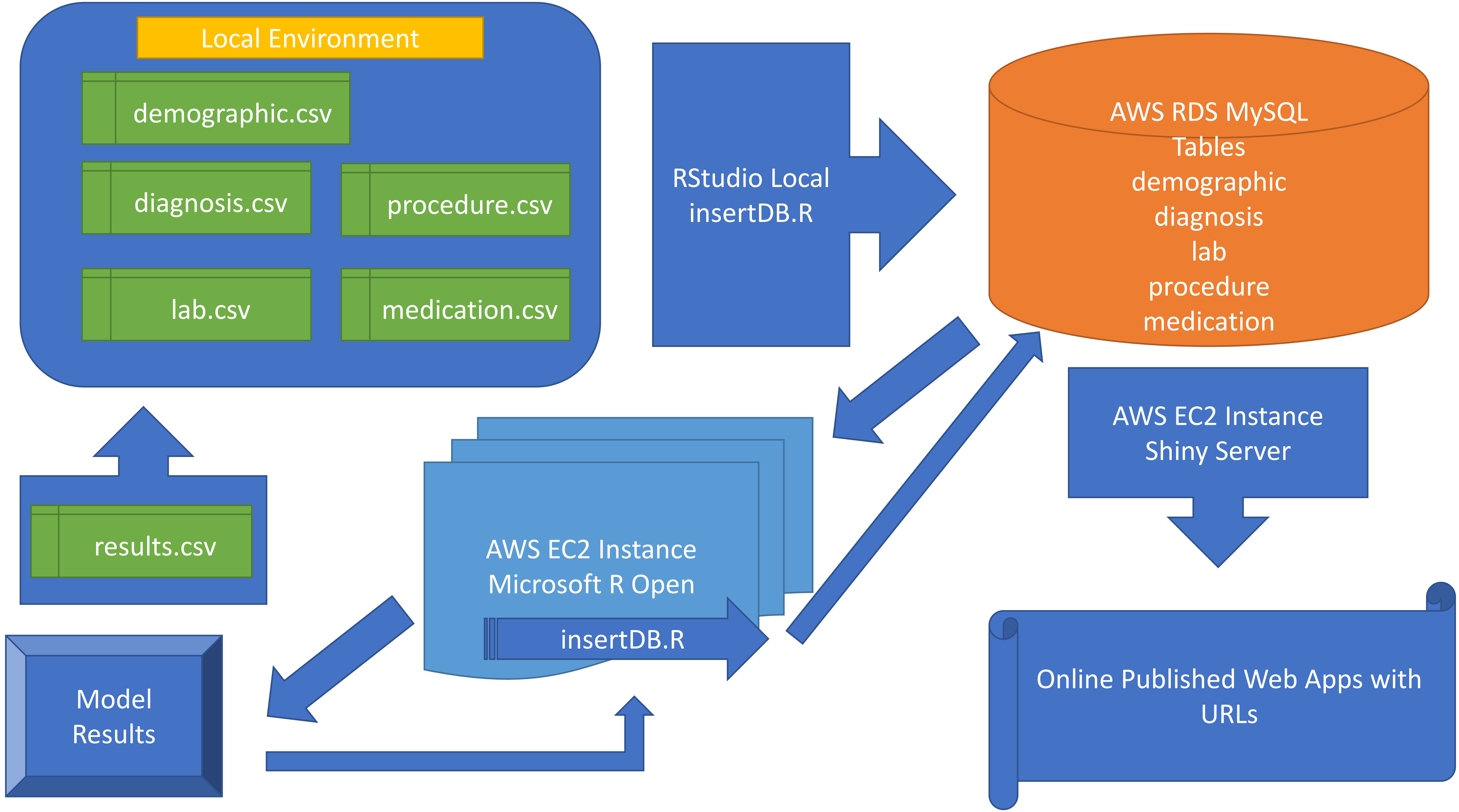
Big data analytics for personalized medicine and pharmacogenomics

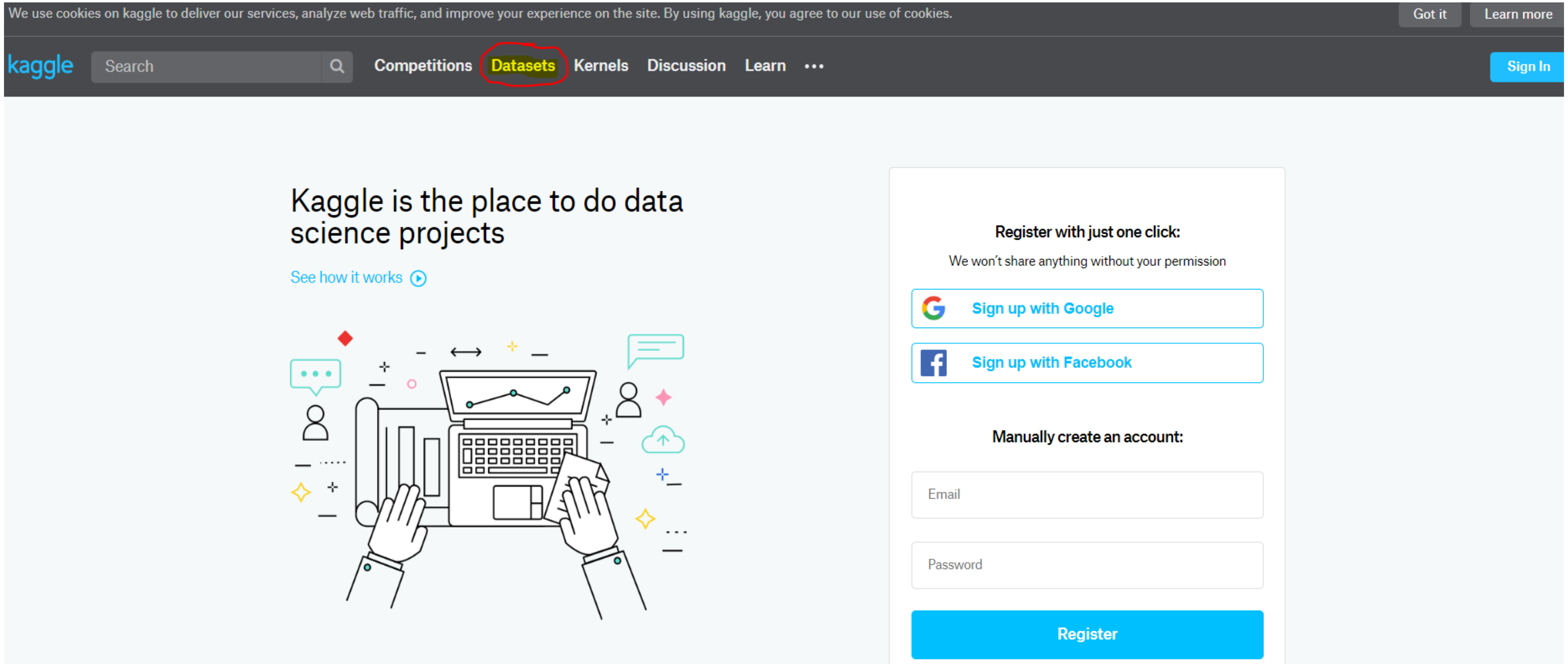


- Cirillo and Valencia (2019). Machine learning algorithms for multi-view data analysis. Data from multiple sources (genomic, proteomic, metabolomic) used to identify associations within and between multiple sets of patients, and generate models for patient clustering.

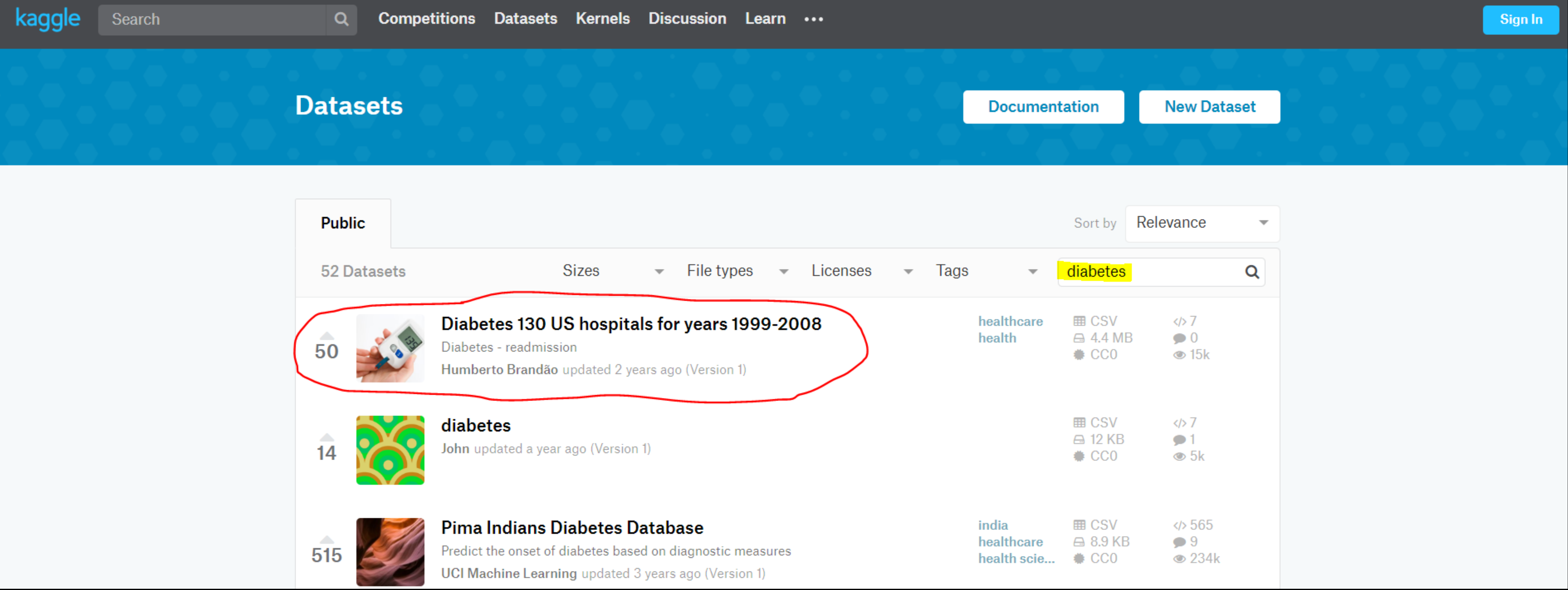
The Latest Buzzwords

Data science	Artificial intelligence	Machine learning	Data mining	Big data
Data warehouse	Data lake	Cloud computing	Hadoop	Internet of things


Local Environment
demographic.csv
diagnosis.csv
procedure.csv
lab.csv
medication.csv
RStudio Local
insertDB.R
AWS RDS MySQL
Tables
demographic
diagnosis
lab
procedure
medication
AWS EC2 Instance
Shiny Server
results.csv
AWS EC2 Instance
Microsoft R Open
insertDB.R
Model
Results
Online Published Web Apps with
URLs



www.Kaggle.com


kaggle
Search
Competitions
Datasets
Kernels
Discussion
Learn
Sign In
Datasets
Documentation
New Dataset
Public
Sort by
Relevance
52 Datasets
Sizes
File types
Licenses
Tags
diabetes
50
Diabetes 130 US hospitals for years 1999-2008
Diabetes - readmission
Humberto Brandão updated 2 years ago (Version 1)
healthcare
health
CSV
4.4 MB
CC0
7
0
15k
14
diabetes
John updated a year ago (Version 1)
CSV
12 KB
CC0
7
1
5k
515
Pima Indians Diabetes Database
Predict the onset of diabetes based on diagnostic measures
UCI Machine Learning updated 3 years ago (Version 1)
india
healthcare
health scie...
CSV
8.9 KB
CC0
565
9
234k

Diabetes

Step 1. Local Environment

Name	Date modified	Type	Size
diabeticShinyDashboard	4/8/2019 7:31 PM	File folder	
Dropbox Downloads	4/3/2019 3:08 PM	File folder	
S3 Downloads	4/3/2019 3:08 PM	File folder	
db_diabetic.accdb	4/3/2019 1:58 PM	Microsoft Access ...	75,320 KB
admission.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	4,304 KB
demographic.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	5,759 KB
diabetic_data.csv	3/30/2019 11:25 A...	Microsoft Excel Co...	18,711 KB
diabetic_data_numeric.csv	4/4/2019 11:59 AM	Microsoft Excel Co...	14,901 KB
diagnosis.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	10,276 KB
lab.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	15,028 KB
medication.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	2,699 KB
payment.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	3,459 KB
procedure.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	2,047 KB
ref_encounter.csv	4/2/2019 10:21 AM	Microsoft Excel Co...	2,735 KB
db_diabetic.xlsx	4/2/2019 10:51 AM	Microsoft Excel W...	30,930 KB
mlDBdiabetic.R	4/4/2019 12:09 PM	R File	12 KB
mlDBdiabeticDraft2.R	4/4/2019 6:24 PM	R File	15 KB
mlDBdiabeticDraft3.R	4/9/2019 1:24 PM	R File	19 KB
prepDataForML.R	4/4/2019 11:59 AM	R File	7 KB
server.R	4/8/2019 3:51 PM	R File	23 KB
shinyDBdiabetic.R	4/8/2019 2:38 PM	R File	7 KB
shinyMLdiabetic.R	4/8/2019 6:16 PM	R File	12 KB
ui.R	4/8/2019 3:48 PM	R File	7 KB
db_diabetic.RData	4/2/2019 10:24 AM	R Workspace	8,748 KB
diabeticMachineLearningEnsemble.RData	4/9/2019 12:06 PM	R Workspace	235,850 KB
.Rhistory	4/8/2019 5:34 PM	RHISTORY File	17 KB

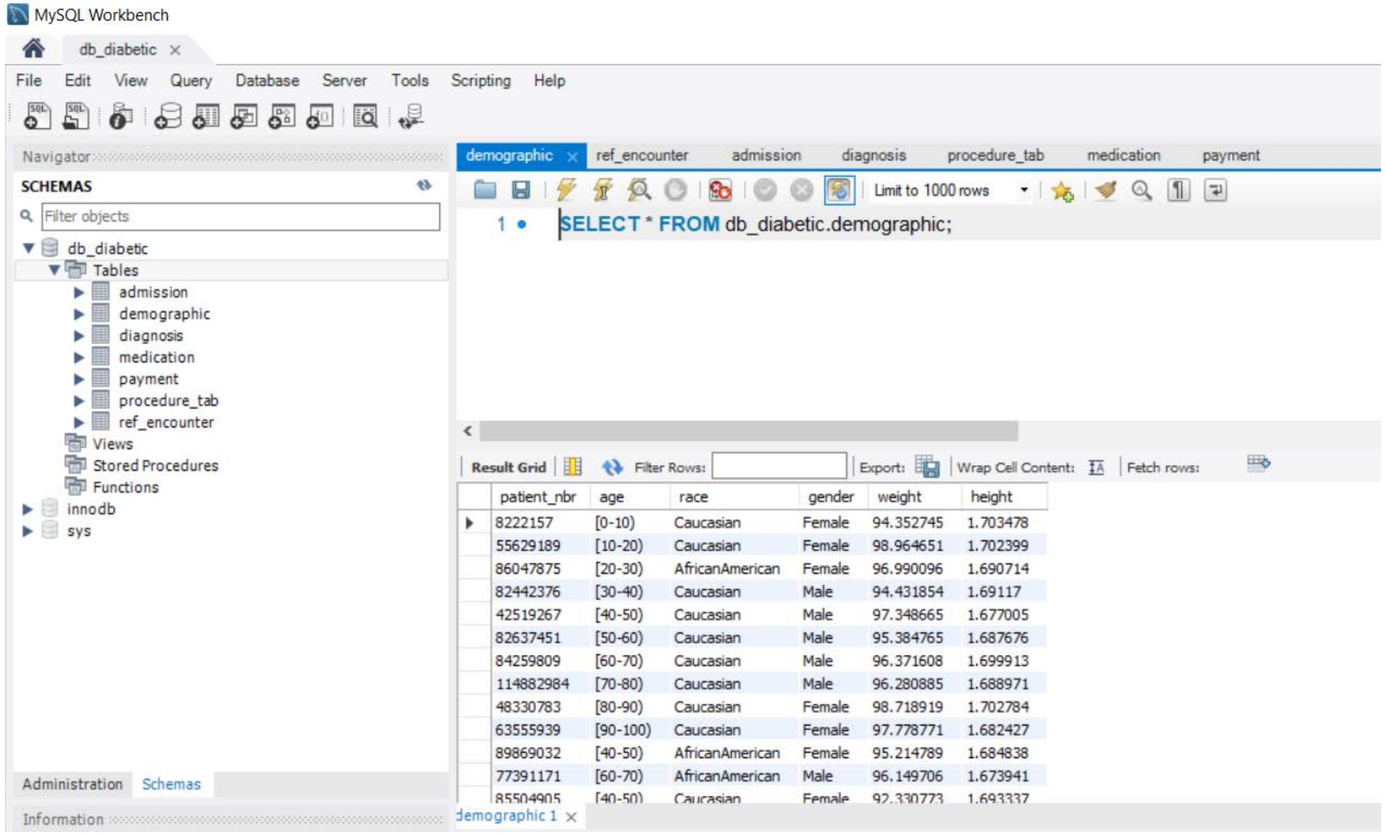
Step 2.
Insert into Remote Database

```r
db_diabetic <- dbPool(
  RMySQL::MySQL(),
  dbname = "db_diabetic",
  host = "globalhealth.cunhdm3u6041.us-east-2.rds.amazonaws.com",
  username = "db_diabetic",
  password = "uh29vawh4tnWHFHcJ"
)
```

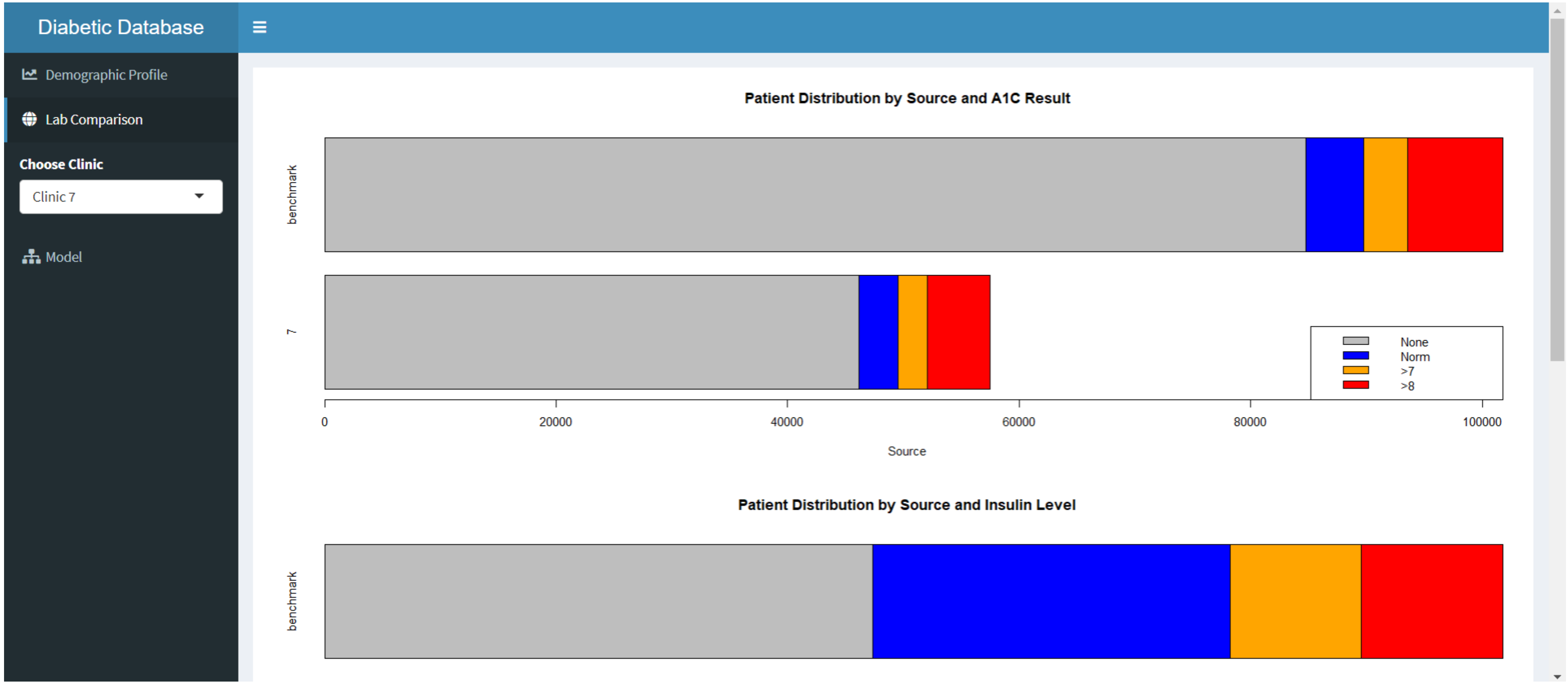
```r
demographic <- read_csv("demographic.csv")
ref_encounter <- read_csv("ref.encounter.csv")
admission <- read_csv("admission.csv")
diagnosis <- read_csv("diagnosis.csv")
lab <- read_csv("lab.csv")
procedure <- read_csv("procedure.csv")
medication <- read_csv("medication.csv")
payment <- read_csv("payment.csv")
```

```r
query_ref_encounter1 <- paste(
  "CREATE TABLE ref_encounter (
  patient_nbr DOUBLE,
  encounter_id DOUBLE
);"
)
# query2 <- "INSERT INTO Product_Names
#    VALUES (?, ?);"

dbExecute(db_diabetic, query_ref_encounter1)

# Begin the query
query_ref_encounter2 <- "INSERT into ref_encounter (patient_nbr, encounter_id) VALUES"

# Finish it with
query_ref_encounter3 <- paste0(
  query_ref_encounter2,
  paste(sprintf("('%s', '%s')",
                ref_encounter$patient_nbr,
                ref_encounter$encounter_id
  ),
  collapse = ",")
)

dbExecute(db_diabetic, query_ref_encounter3)
```

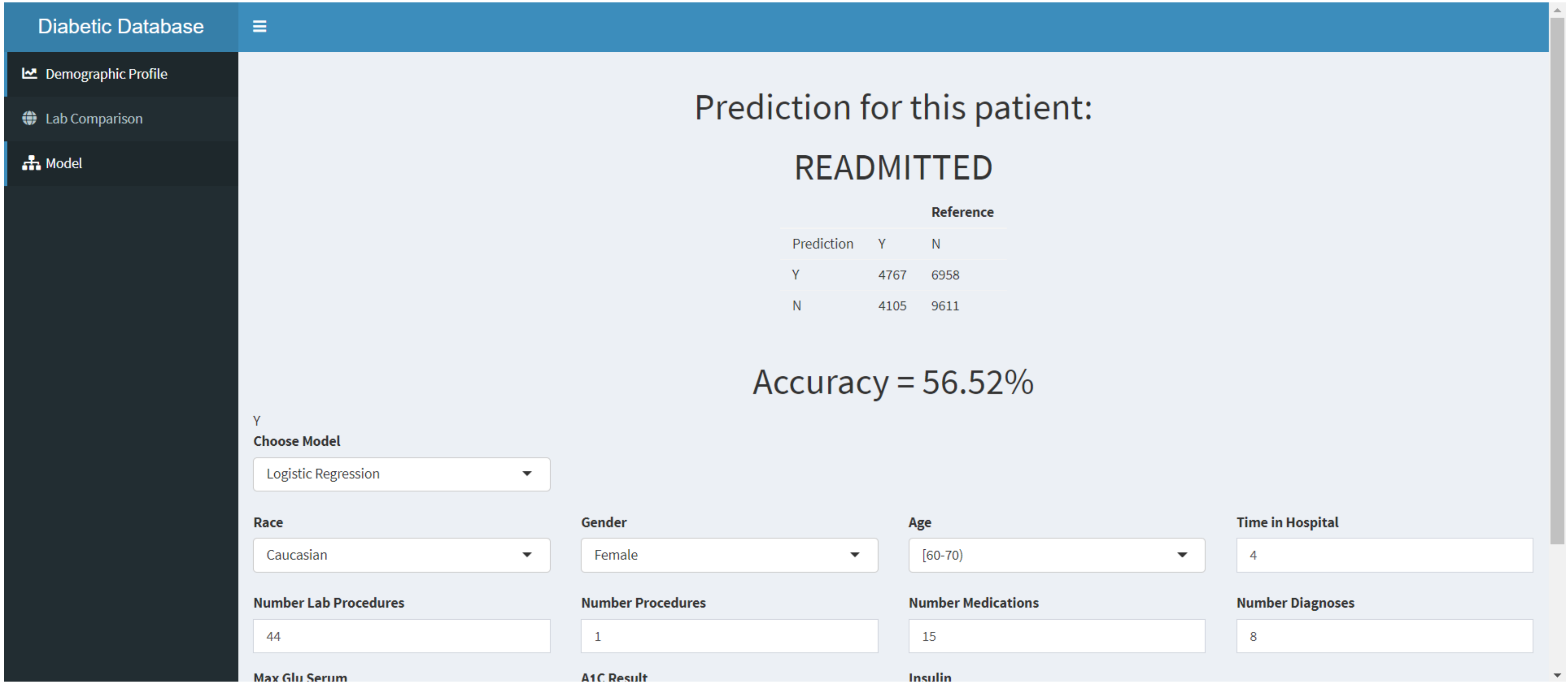
Step 3.
SQL Database



Step 4.
R Shiny Web App



Step 4.
R Shiny Web App



Local Storage

Text Files (.txt)

Comma Separated Value Files (.csv)

Excel Database (.xlsx)

Microsoft Access Database (.accdb)

Remote Storage

- Dropbox
- Google Sheets
- AWS S3 Bucket
- Cloud Database
  - MySQL
  - MongoDB
  - PostgreSQL
  - NoSQL
  - Oracle

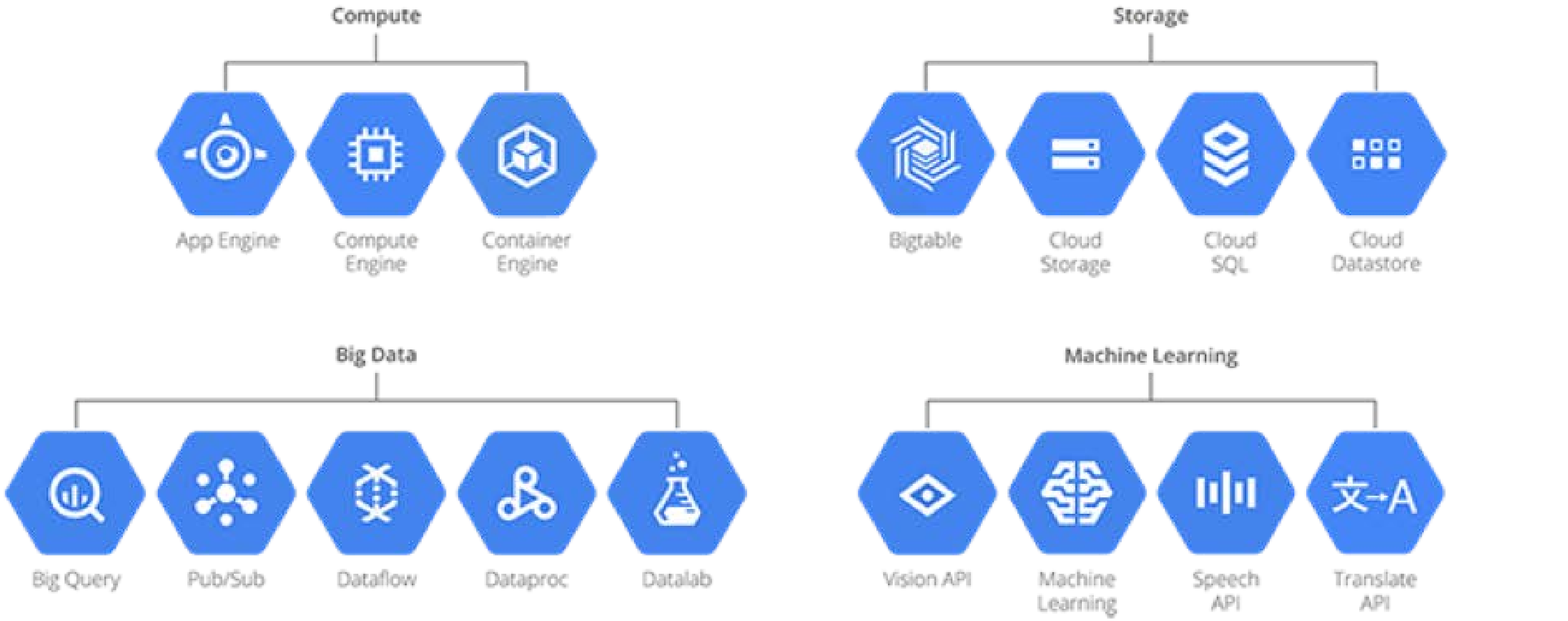
Cloud Platforms


amazon web services
Amazon EC2
Amazon RDS
AWS Direct Connect
Amazon EBS
Amazon S3
Elastic Load Balancing
Amazon Route 53
Amazon VPC
Elastic IP

Devices	Device Connectivity	Storage	Analytics	Presentation & Action
	Event Hubs	SQL Database	Machine Learning	App Service
	Service Bus	Table/Blob Storage	Stream Analytics	Power BI
	External Data Sources	DocumentDB	HDInsight	Notification Hubs
		External Data Sources	Data Factory	Mobile Services
				BizTalk Services

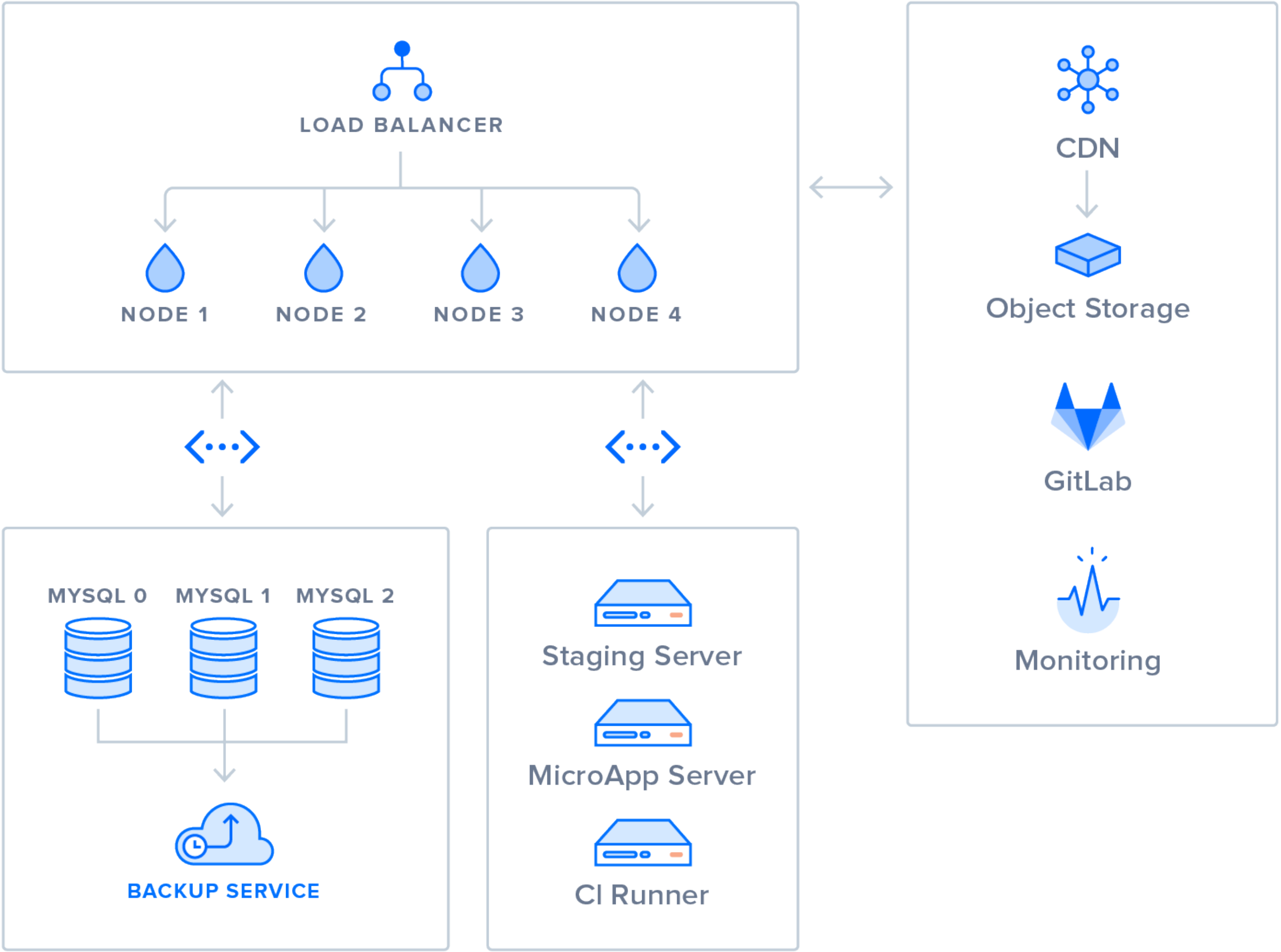
Microsoft Azure

Google Cloud Platform



DigitalOcean

Current Content Ignite Infrastructure



Distributed Computing Technology

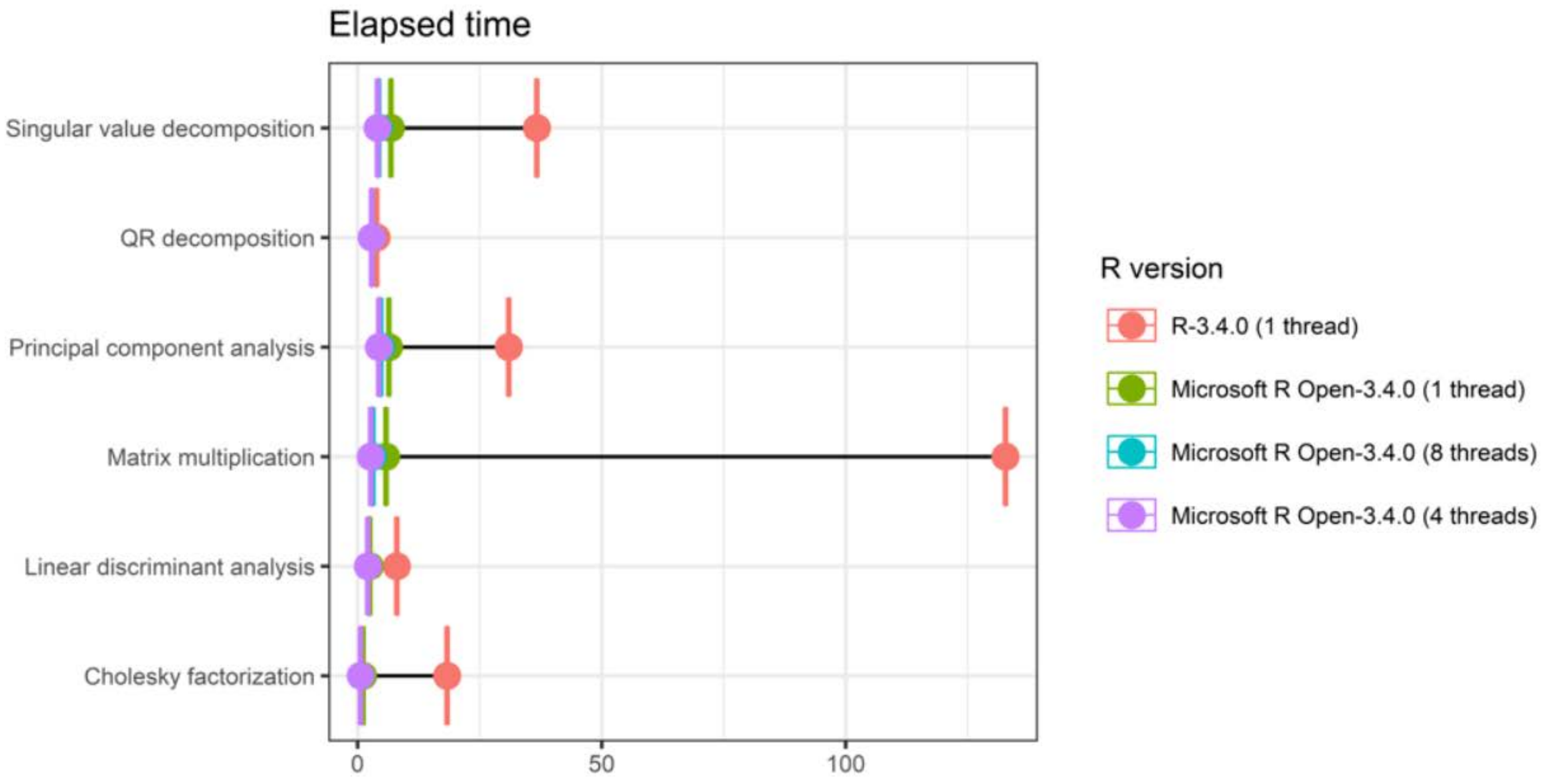
All processing jobs (scripts; i.e. R, Python, Scala, etc.) are divvied up among all available processing units (computers, cores, threads, etc.)

Hadoop

Spark

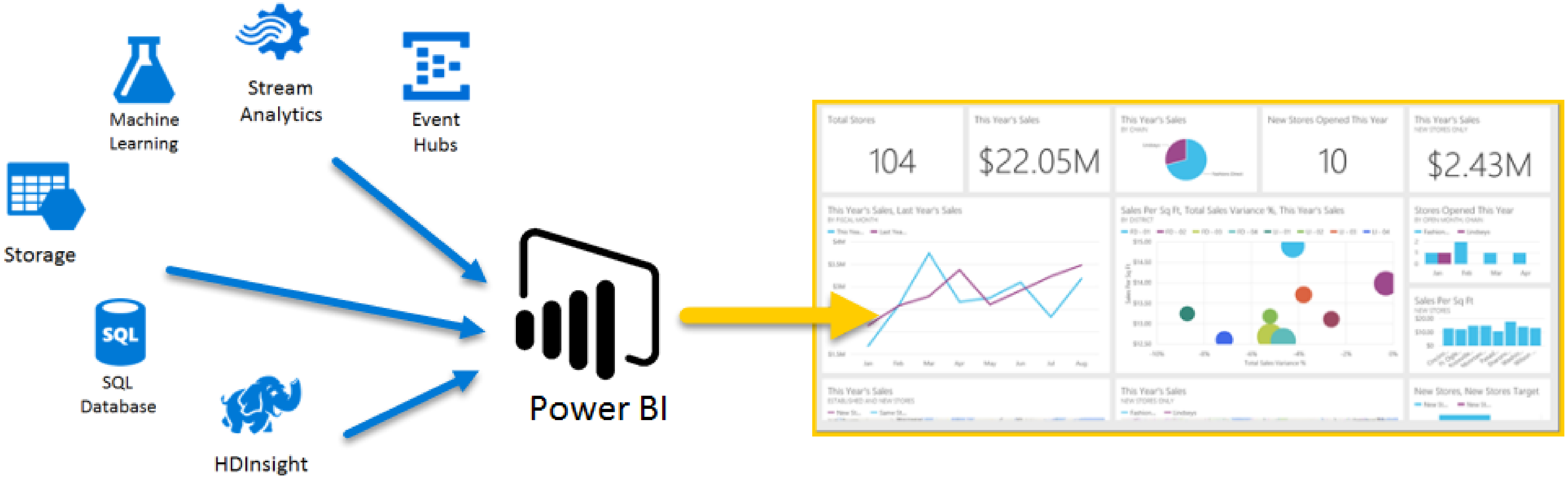
Microsoft R Open

Microsoft R Open Multithreaded Performance

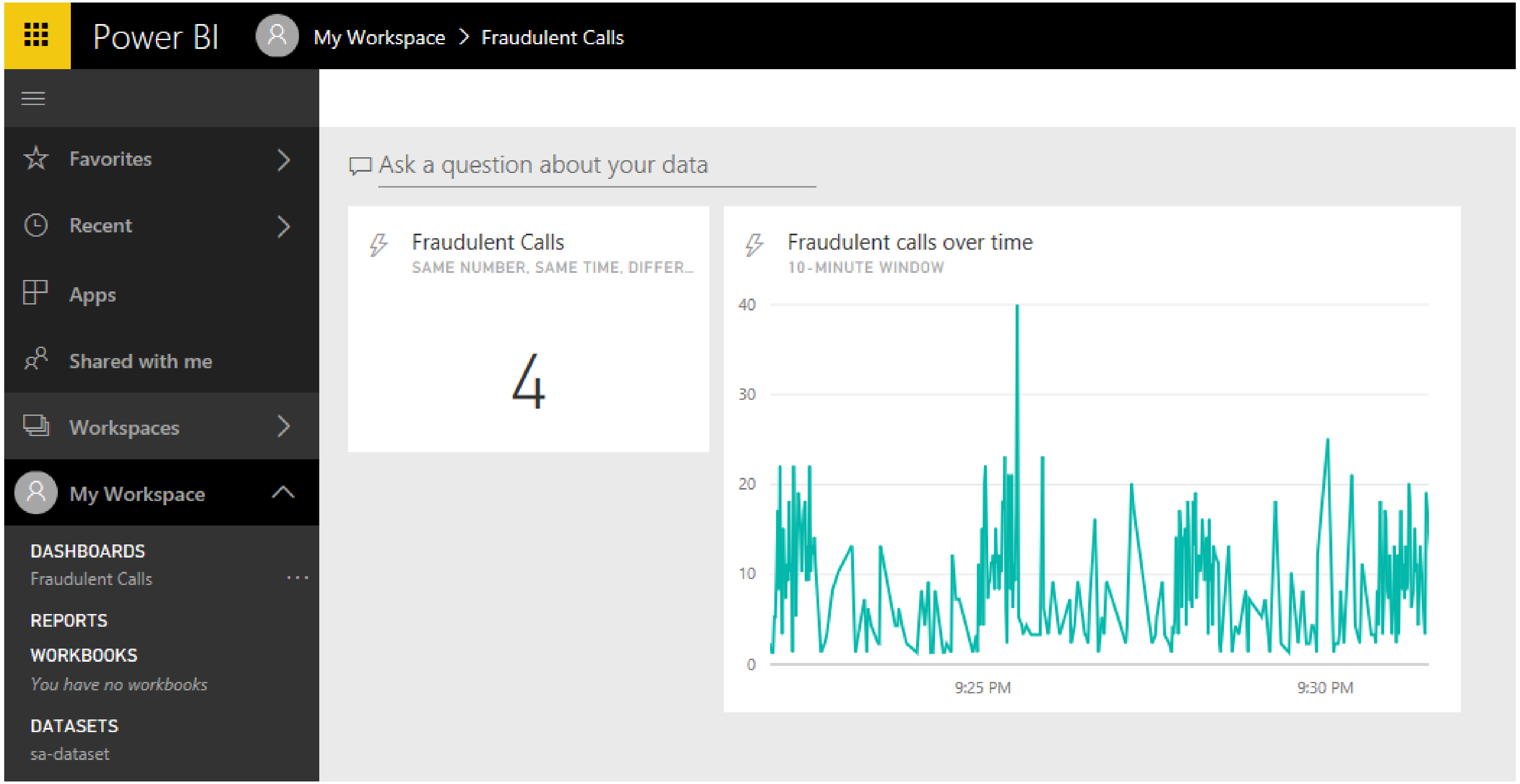


Machine Learning

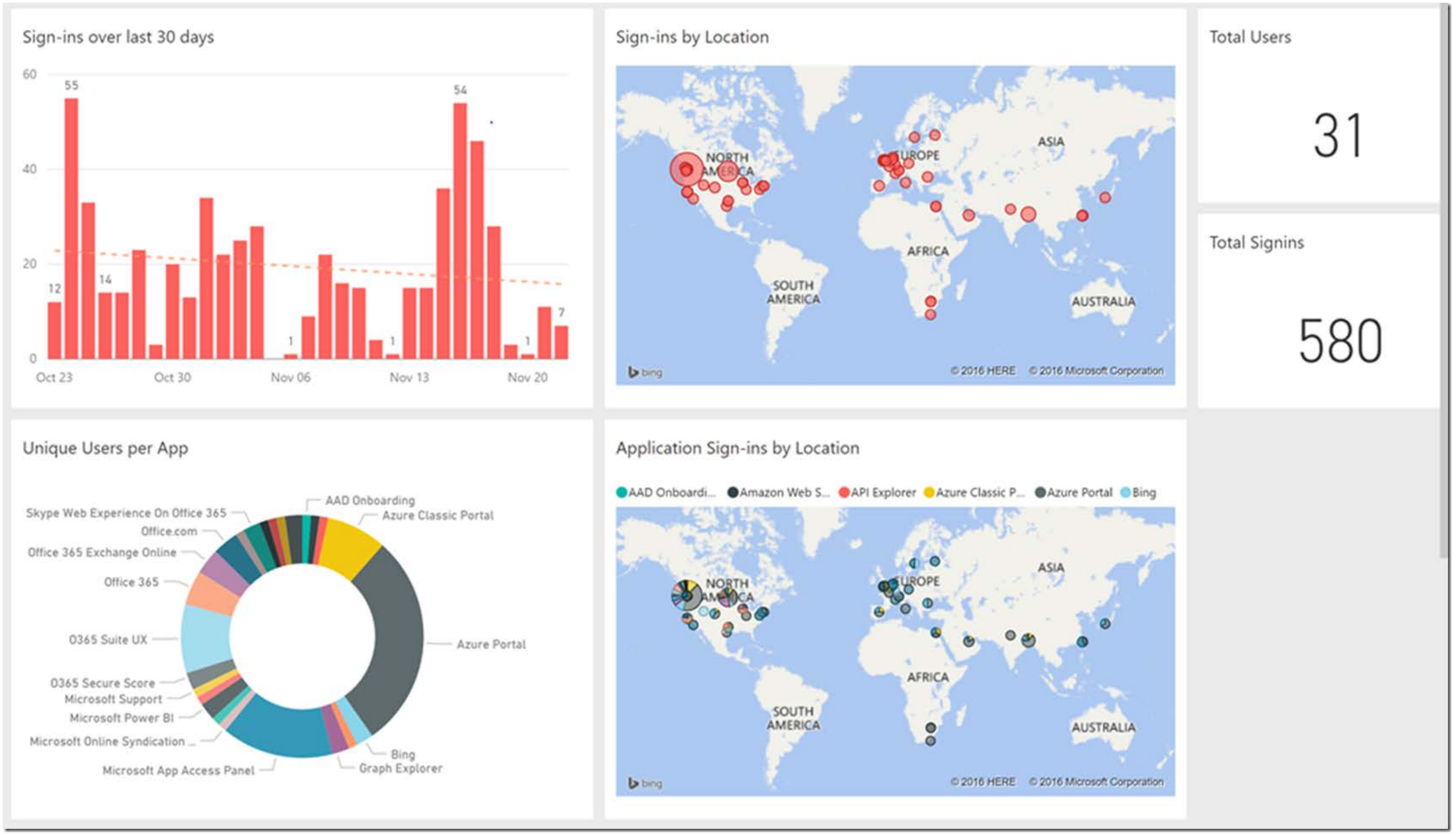
Microsoft Azure Power BI



Microsoft Azure Power BI

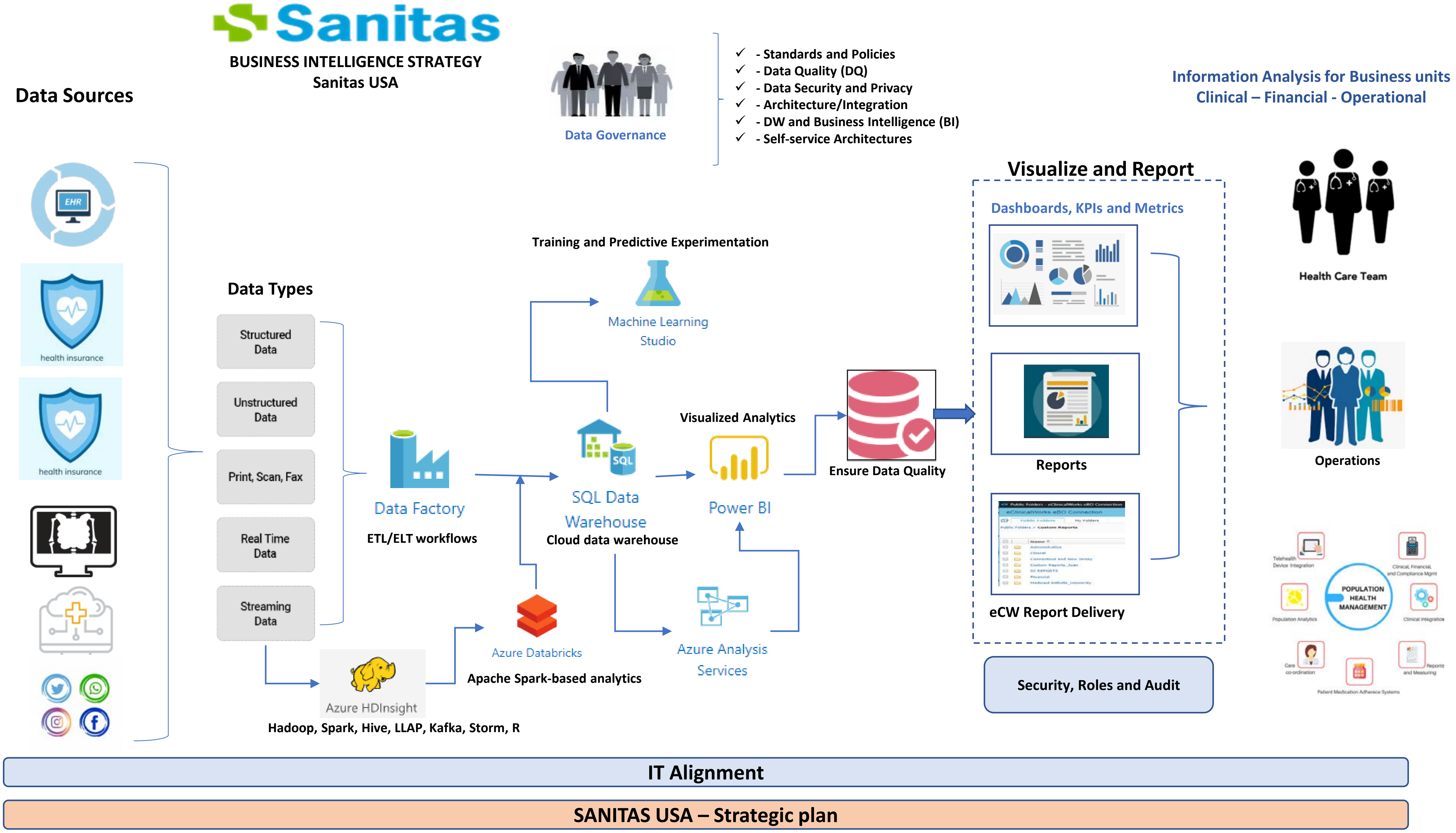


Microsoft Azure Power BI

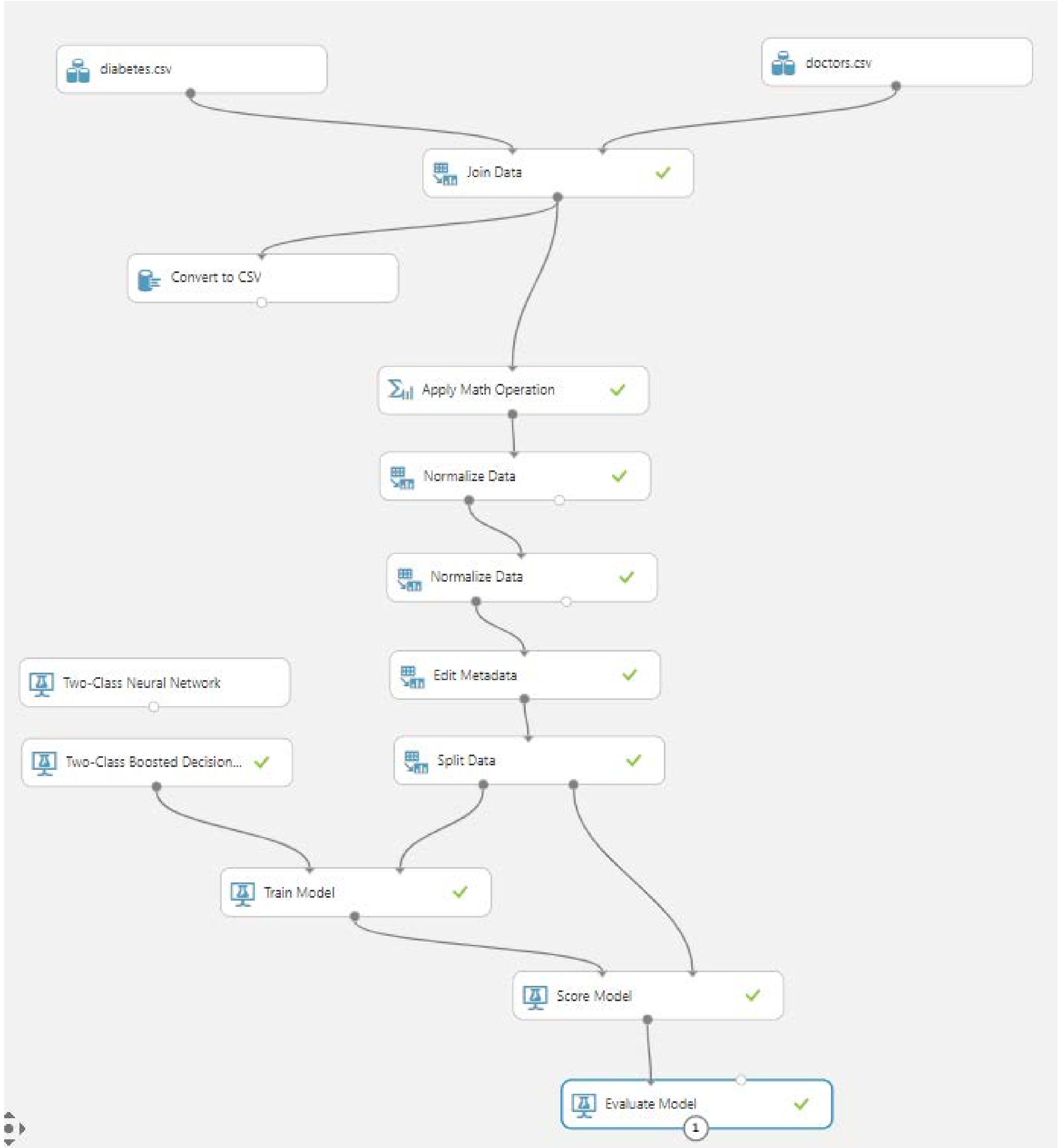


Solution Architectures at Sanitas

- Analytics on big data
- Data warehouse
- Real-time analytics
- Population Health Management for Healthcare


Sanitas
BUSINESS INTELLIGENCE STRATEGY
Sanitas USA
Data Sources
Data Governance
- Standards and Policies
- Data Quality (DQ)
- Data Security and Privacy
- Architecture/Integration
- DW and Business Intelligence (BI)
- Self-service Architectures
Information Analysis for Business units
Clinical – Financial - Operational
EHR
health insurance
health insurance
Data Types
Structured Data
Unstructured Data
Print, Scan, Fax
Real Time Data
Streaming Data
Training and Predictive Experimentation
Machine Learning Studio
Data Factory
ETL/ELT workflows
SQL Data Warehouse
Cloud data warehouse
Visualized Analytics
Power BI
Ensure Data Quality
Azure Databricks
Apache Spark-based analytics
Azure HDInsight
Hadoop, Spark, Hive, LLAP, Kafka, Storm, R
Azure Analysis Services
Visualize and Report
Dashboards, KPIs and Metrics
Reports
eCW Report Delivery
Security, Roles and Audit
Health Care Team
Operations
POPULATION HEALTH MANAGEMENT
Telehealth Device Integration
Clinical, Financial, and Compliance Mgmt
Population Analytics
Clinical Integration
Care co-ordination
Reports and Measuring
Patient Medication Adherence Systems
IT Alignment
SANITAS USA – Strategic plan

ML Studio Experiment



ML Studio Dataset

Diabetes Analysis › diabetes.csv › dataset

rows 15000

columns 10

view as

PatientID	Pregnancies	PlasmaGlucose	DiastolicBloodPressure	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic
1354778	0	171	80	34	23	43.509726	1.213191	21	0
1147438	8	92	93	47	36	21.240576	0.158365	23	0
1640031	7	115	47	52	35	41.511523	0.079019	23	0
1883350	9	103	78	25	304	29.582192	1.28287	43	1
1424119	1	85	59	27	35	42.604536	0.549542	22	0
1619297	0	82	92	9	253	19.72416	0.103424	26	0
1660149	0	133	47	19	227	21.941357	0.17416	21	0
1458769	0	67	87	43	36	18.277723	0.236165	26	0
1201647	8	80	95	33	24	26.624929	0.443947	53	1
1403912	1	72	31	40	42	36.889576	0.103944	26	0
1943830	1	88	86	11	58	43.225041	0.230285	22	0
1824483	3	94	96	31	36	21.294479	0.25902	23	0
1848869	5	114	101	43	70	36.49532	0.07919	38	1
1669231	7	110	82	16	44	36.089293	0.281276	25	0
1683688	0	148	58	11	179	39.192076	0.160829	45	0
1738587	3	109	77	46	61	19.847312	0.204345	21	1
1884264	3	106	64	25	51	29.044573	0.589188	42	1
1485251	1	156	53	15	226	29.786192	0.203824	41	1

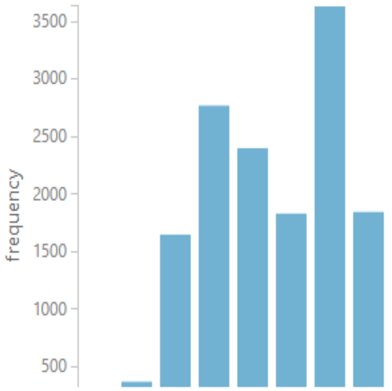
Statistics

Mean	71.2207
Median	72
Min	24
Max	117
Standard Deviation	16.7587
Unique Values	90
Missing Values	0
Feature Type	Numeric Feature

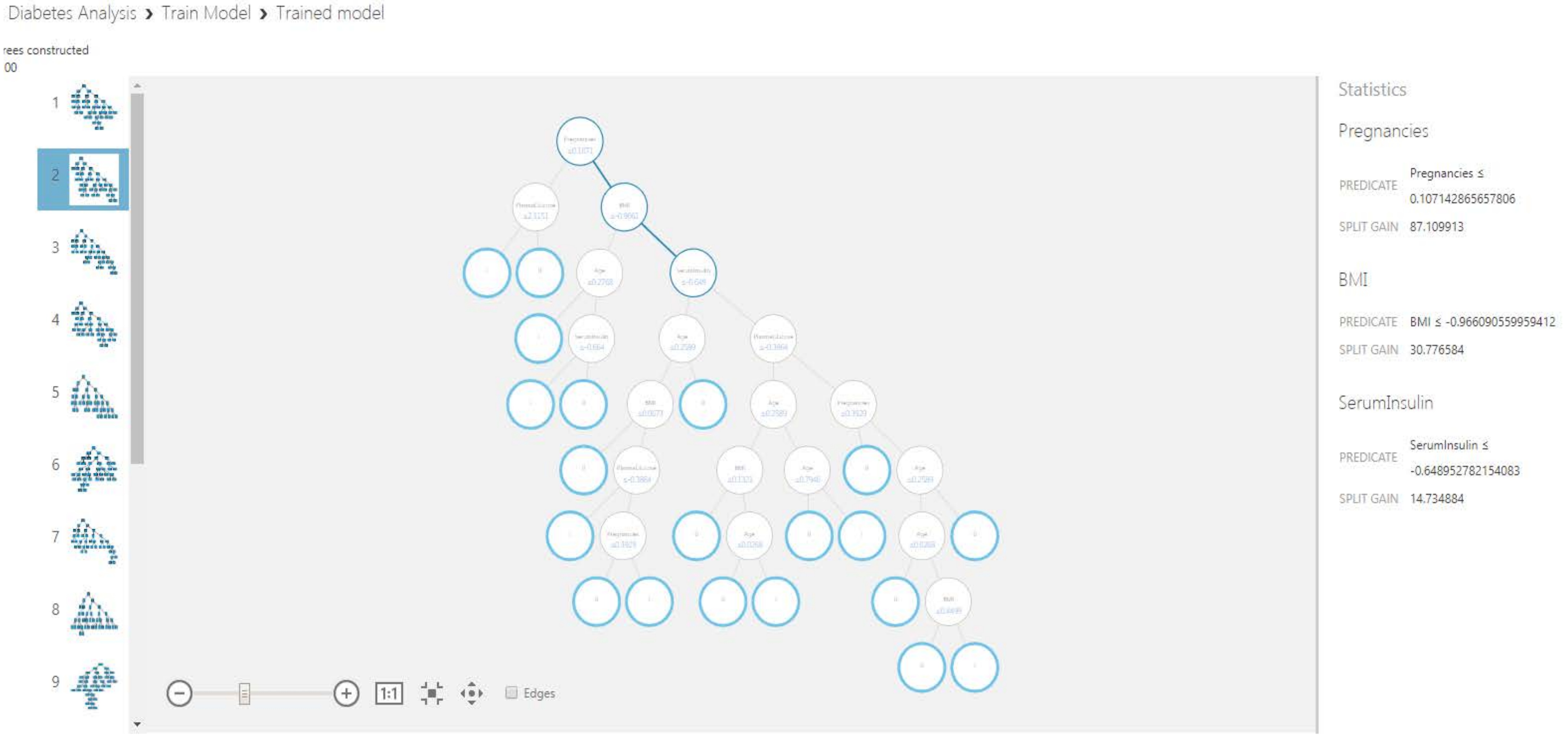
Visualizations

DiastolicBloodPressure

Histogram



ML Studio
Model
Train



ML Studio Model Score

Diabetes Analysis › Score Model › Scored dataset

rows	columns
4500	14

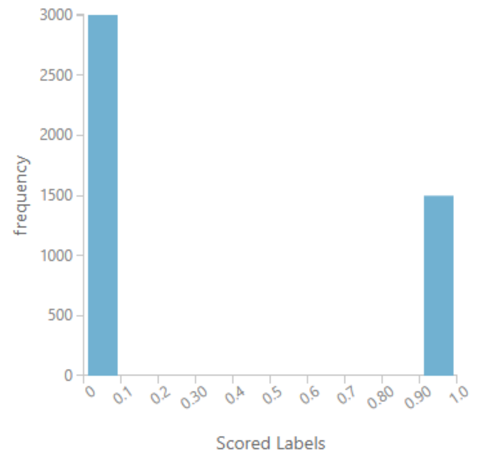
se	DiastolicBloodPressure	TricepsThickness	SerumInsulin	BMI	DiabetesPedigree	Age	Diabetic	Physician	Ln(Age)	Scored Labels	Scored Probabilities
	0.643229	-0.399444	1.586817	1.439866	0.476071	0.017857	1	Wade Munger	0.035804	1	0.998918
	-0.55022	1.11204	1.812272	-0.156933	0.006973	0.017857	1	Niew Leekpai	0.035804	1	1
	0.583557	-1.292594	-0.660226	0.775076	0.225251	0.017857	0	Roman Pilcher	0.035804	0	0.123397
	-0.967927	-0.742963	-0.772953	0.664622	0.388068	0.035714	0	Mara Rasmussen	0.070017	0	0.000004
	-0.192185	-0.949075	0.271658	-0.662061	0.021152	0.017857	1	Ethan Rincon	0.035804	1	0.999995
	1.060936	-0.33074	-0.742893	-1.338164	0.033424	0.035714	0	Vaughn Oquendo	0.070017	0	0.000011
	-1.683996	0.56241	-0.667741	1.764616	0.019228	0.232143	0	Neandro Baeza	0.370849	0	0.000025
	-0.908254	0.287594	1.368876	-1.073757	0.039045	0.178571	0	Delmar Pelchat	0.299754	0	0.000011
	1.299626	-1.22389	-0.412224	0.696152	0.040121	0.321429	1	Nazzareno Piccio	0.476447	1	0.998053
	-0.729237	0.974632	-0.562528	0.016523	0.202185	0.464286	1	Billie Stonge	0.620054	1	0.999978
	0.941591	-1.430001	-0.915742	-1.275964	0.293586	0.285714	0	Jimmie Turman	0.435929	0	0.000005
	1.478643	2.692228	3.525736	0.4355	0.304365	0.160714	1	Deanna Ball	0.274517	1	0.999623
	0.822246	-1.22389	-0.833075	-1.212273	0.056504	0	0	Jenny Norgaard	0	0	0.000028
	1.180281	0.837225	-0.65271	0.701326	0.260088	0.232143	0	Daitaro Ishida	0.370849	0	0.000919

Statistics

Visualizations

Scored Labels

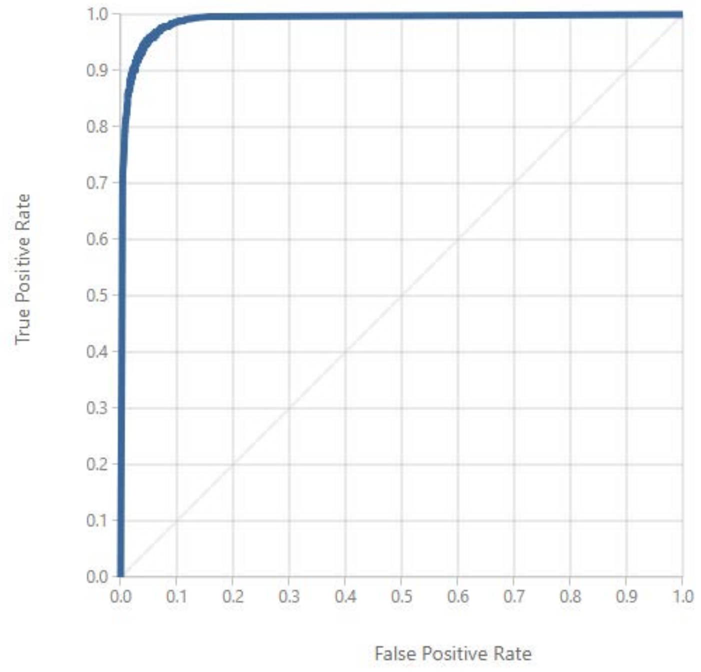
Histogram



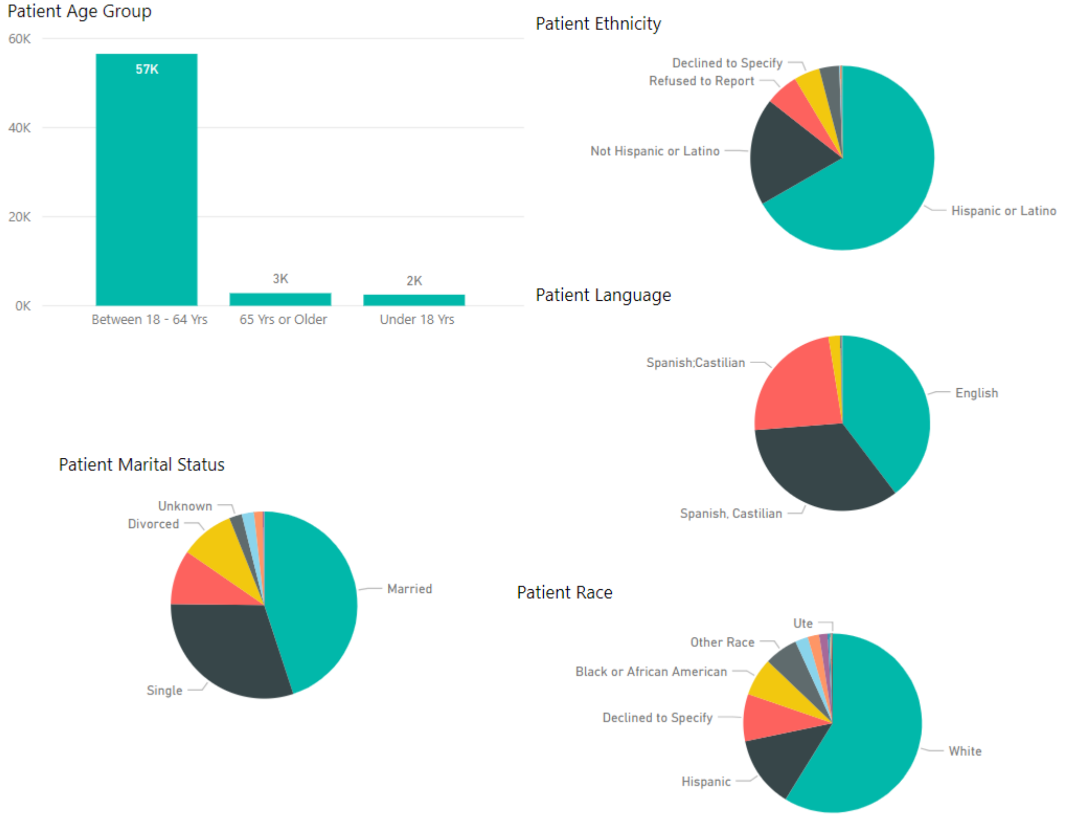
ML Studio Model Evaluation

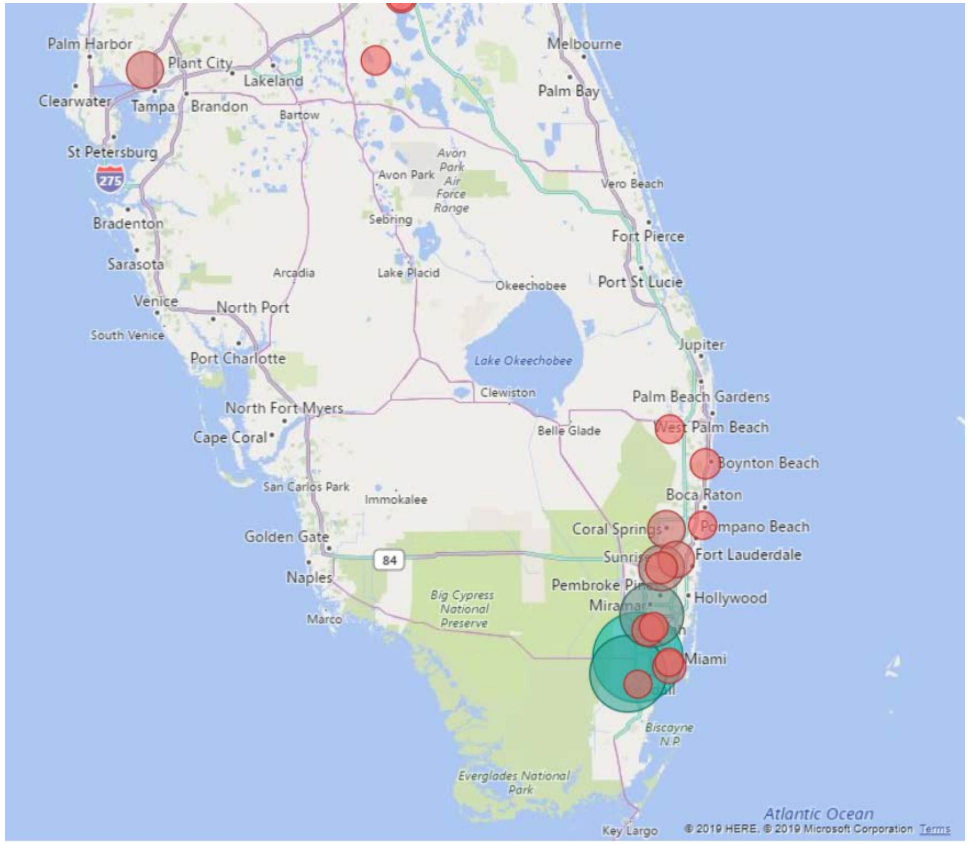
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
1399	108	0.954	0.934	0.5	0.991
False Positive	True Negative	Recall	F1 Score		
99	2894	0.928	0.931		
Positive Label	Negative Label				
1	0				

ROC PRECISION/RECALL LIFT



Power BI


Patient Age Group
60K
40K
20K
0K
57K
3K
2K
Between 18 - 64 Yrs
65 Yrs or Older
Under 18 Yrs
Patient Ethnicity
Declined to Specify
Refused to Report
Not Hispanic or Latino
Hispanic or Latino
Patient Language
Spanish;Castilian
English
Spanish, Castilian
Patient Marital Status
Unknown
Divorced
Married
Single
Patient Race
Ute
Other Race
Black or African American
Declined to Specify
Hispanic
White


Palm Harbor
Plant City
Lakeland
Clearwater
Tampa
Brandon
Bartow
St Petersburg
Melbourne
Palm Bay
Avon Park
Sebring
Vero Beach
Fort Pierce
Port St Lucie
Bradenton
Sarasota
Arcadia
Lake Placid
Okeechobee
Venice
North Port
South Venice
Port Charlotte
Lake Okeechobee
Jupiter
Clewiston
North Fort Myers
Cape Coral
Belle Glade
Palm Beach Gardens
West Palm Beach
Boynton Beach
San Carlos Park
Immokalee
Boca Raton
Golden Gate
Coral Springs
Pompano Beach
Fort Lauderdale
Naples
Big Cypress National Preserve
Pembroke Pines
Hollywood
Miami
Marco
Biscayne N.P.
Everglades National Park
Atlantic Ocean
Key Largo
© 2019 HERE, © 2019 Microsoft Corporation Terms

Rejoinder

- Data -> our biggest asset — Emphasis on “good” data
- Processing streams -> wrangling, carpenting
- Storage –> lakes, warehousing, data bases
- Analytics -> mining, machine learning
- Output -> new knowledge, information, inferences
- Feed back to the users -> gather more data

Question to the global audience

- What are your needs and where are you currently with respect to?
  - Data collection, quality, storage, analytical and computing power
    - Where is data coming from, single or multiple sources
    - Who is maintaining data quality and fidelity
    - Do you have adequate storage with proper security; planning for the future
    - Are you investing in resources and trained personnel with data science skills