Data Science and Analytics
-Competing in a Data-driven World-

Cesar Acosta Ph.D.

Department of Industrial and Systems Engineering
University of Southern California

OVERVIEW

- Data-driven
- Data Science
- Data Analytics
- What is the difference?

DATA SCIENCE

What is Data-driven?

DATA SCIENCE

Being data-driven means

Making business decisions
Managing processes

based on facts / insights derived from data

DATA SCIENCE

Why
Data-driven?

DATA SCIENCE

Most companies
have information systems
but are not data-driven

DATA SCIENCE

These companies
have information systems

- ERP system
- CRM module
- Accounting system

DATA SCIENCE

But not aware if the data is

- accurate
- up-to-date

DATA SCIENCE

Moreover, they may not know

how to make

the best use of the data

DATA SCIENCE

They may be

doing transactions and

making decisions

- with not accurate
- with not up-to-date

data

DATA SCIENCE

They may be

doing transactions and

making decisions

- with not accurate
- with not up-to-date, or
- ignoring the data

DATA SCIENCE

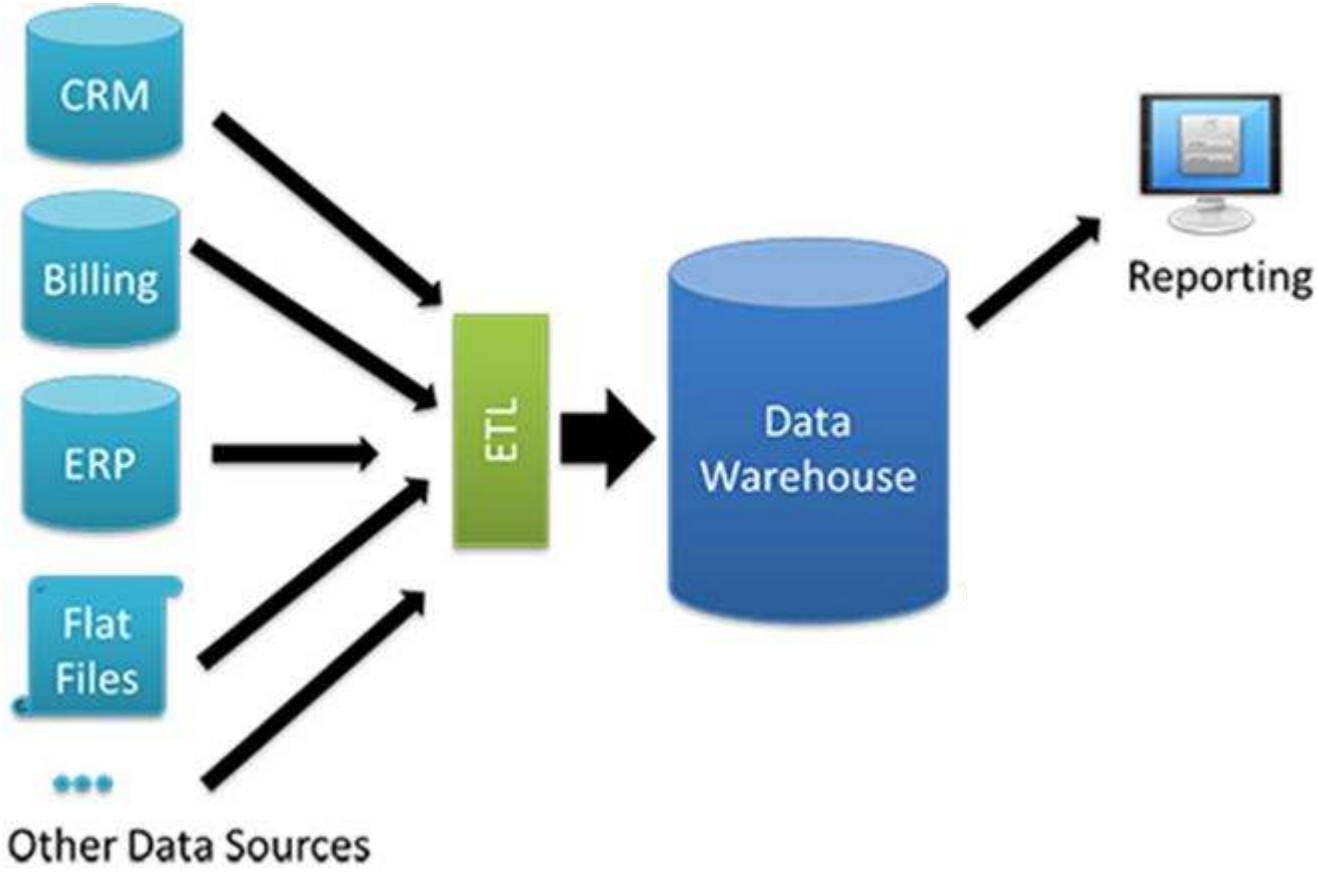
Making decisions ignoring the data?

DATA SCIENCE

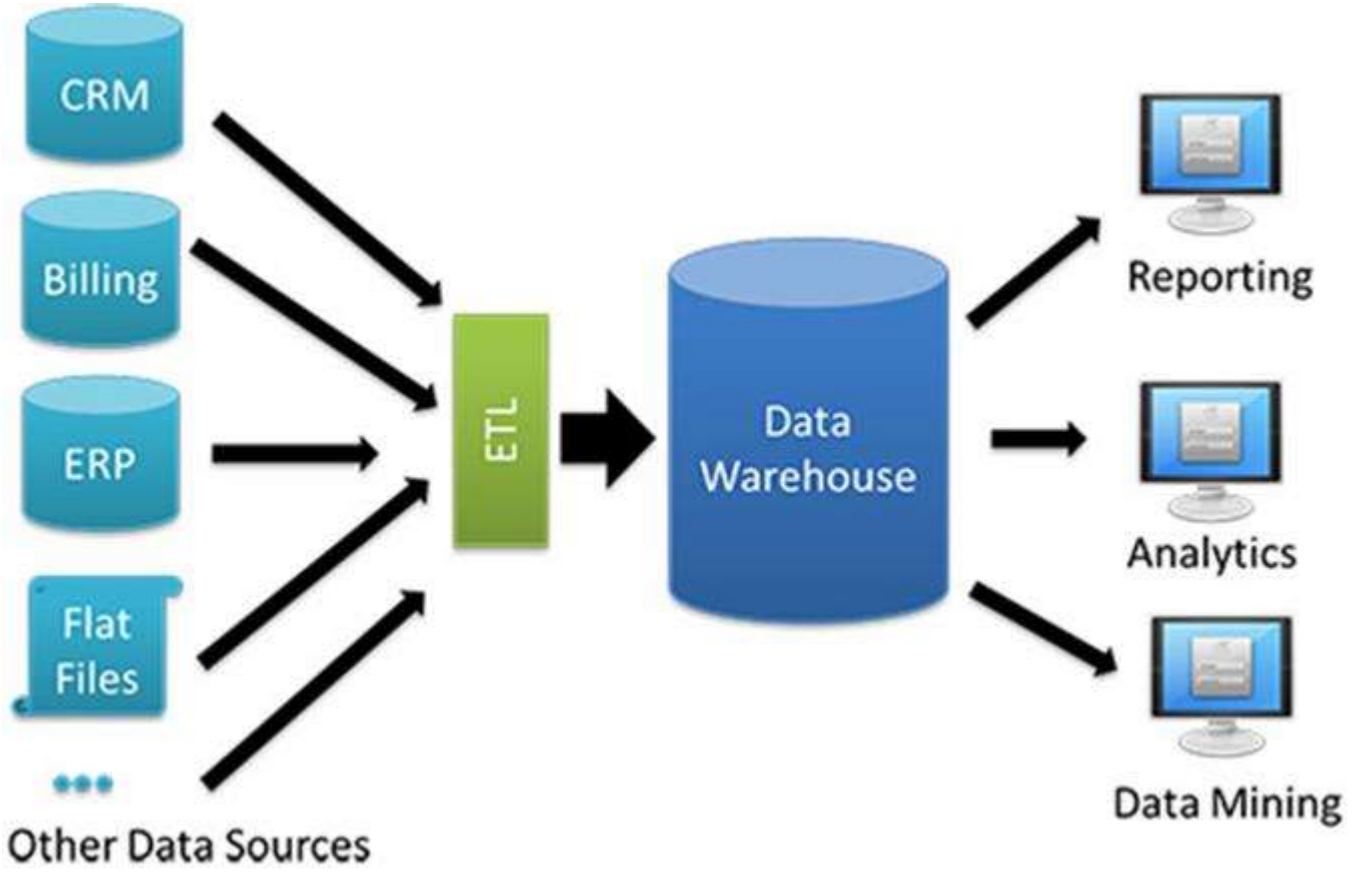
Making decisions ignoring the data?

- Based on feelings
- Based on personal experiences

DATA SCIENCE



DATA SCIENCE



DATA SCIENCE

A data-driven organization

puts data at the core of their business processes

using facts, insights derived from data

to drive their decision-making

DATA SCIENCE

A data-driven organization

moves from guessing and assumptions

to using data and analytics

to make faster and better decisions

DATA SCIENCE

Relevant and accurate data are

at the core

of a data-driven organization

DATA SCIENCE

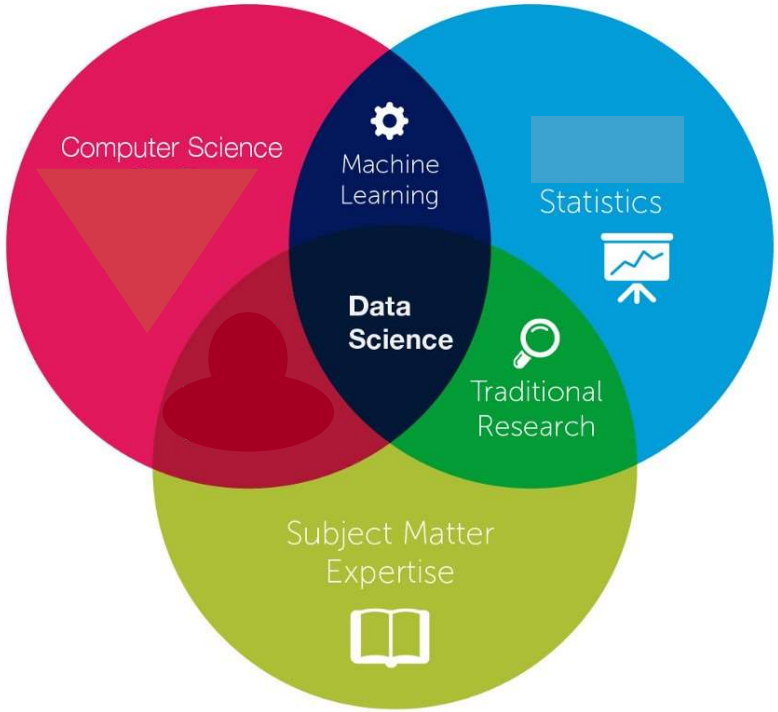
What is Data Science?

DATA SCIENCE

Data Science is the field of study that combines

- computer science
- statistics
- business

to find useful information from raw data.



DATA SCIENCE

Why Statistics?

DATA SCIENCE

Statistics = data analysis

DATA SCIENCE

Statistics is data

- collection
- cleaning
- organization
- visualization
- analysis
- modeling
- presentation

DATA SCIENCE

Statistics is data

- collection
- cleaning
- organization
- visualization
- analysis
- modeling
- presentation

DATA SCIENCE

data modeling

- Generalized linear models
- Bayesian modeling
- cluster analysis
- time series modeling
- principal components
- partial least squares
- spatial analysis

DATA SCIENCE

Statistical tools to understand, analyze the data

- Random variables
- density functions
- Outliers
- Covariance, correlation
- Probabilities
- Bootstrapping
- Confidence and Prediction Intervals

DATA SCIENCE

Why
Computer Science?

DATA SCIENCE

Computer Science tools

to collect, process, store the data

- Data Wrangling (unstructured to structured data)
- Data Warehousing (repo of structured data)
- Cloud computing
- Big data
- Machine learning models
- Web developing (front-end)

DATA SCIENCE

Why Business?

DATA SCIENCE

Business domain knowledge

to make the right questions about

- Customer needs
- Products
- Processes
- Variables
- KPIs
- Environment variables

DATA SCIENCE

Business domain = Industry

- Retail
- Health care
- Financial
- Manufacturing
- Government
- Services

DATA SCIENCE

Business domain = Science

- Biology
- Medicine
- Physics
- Materials science
- Chemistry

DATA SCIENCE

What is Data Analytics?

DATA ANALYTICS

Data Analytics professional is someone whose focus is on

- collecting
- summarizing
- analyzing

} data

to find answers to business questions

DATA ANALYTICS

- Who is the Business Analyst?
- What are the Business questions?

DATA ANALYTICS

B.A.

Decision Maker (questions) ➡ Data Analyst (solutions)

DATA ANALYTICS

Business questions

- What happened?
- What will happened?

DATA ANALYTICS

What happened? -business case-

- Which products underperformed?
- Which were more profitable?
- Did our market share change?
- What is our retention rate?
- Who are our most valuable customers?

DATA ANALYTICS

What will happen? -business case-

- What is the expected growth?
- Who are potential customers?
- Most promising product lines?
- What will be our market share?
- What new competitors may arise?

DATA ANALYTICS

What will happen? -new product-

- What is the probability of success?
- What is the risk of failure?
- What is the market acceptance rate?
- Will it outperform current best product?

DATA ANALYTICS

What will happen? -investment-

- What is the expected return?
- What is the probability of a loss?
- If there is a loss, how large can it be?
- What scenarios are possible?
- Major external risk in our sector?

DATA SCIENCE

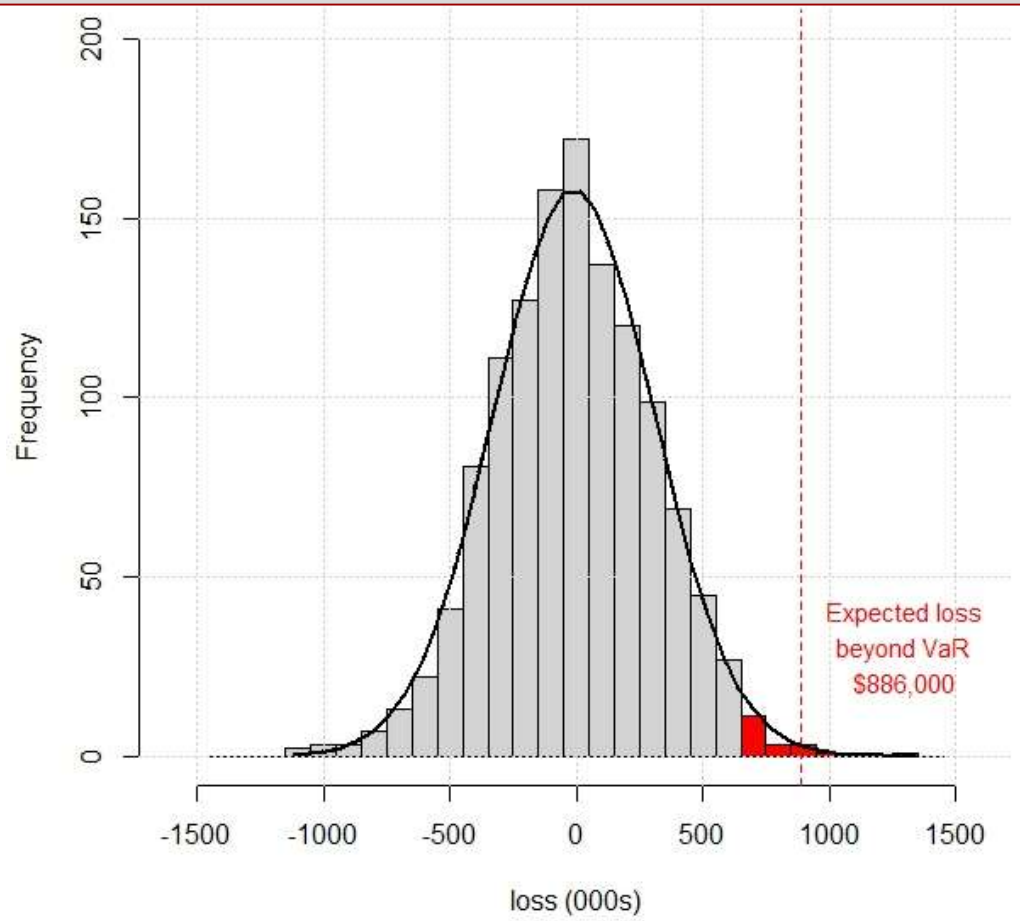
How does the Data Analyst answer these questions?

DATA ANALYTICS

Example: If there is 95% probability of a loss, how large can it be?

- Collect data
- Find distribution of daily losses
- Find 95% quantile of daily losses
- Find expected loss beyond that quantile (VaR)

DATA ANALYTICS



DATA ANALYTICS

Example: Medicine

Business Question

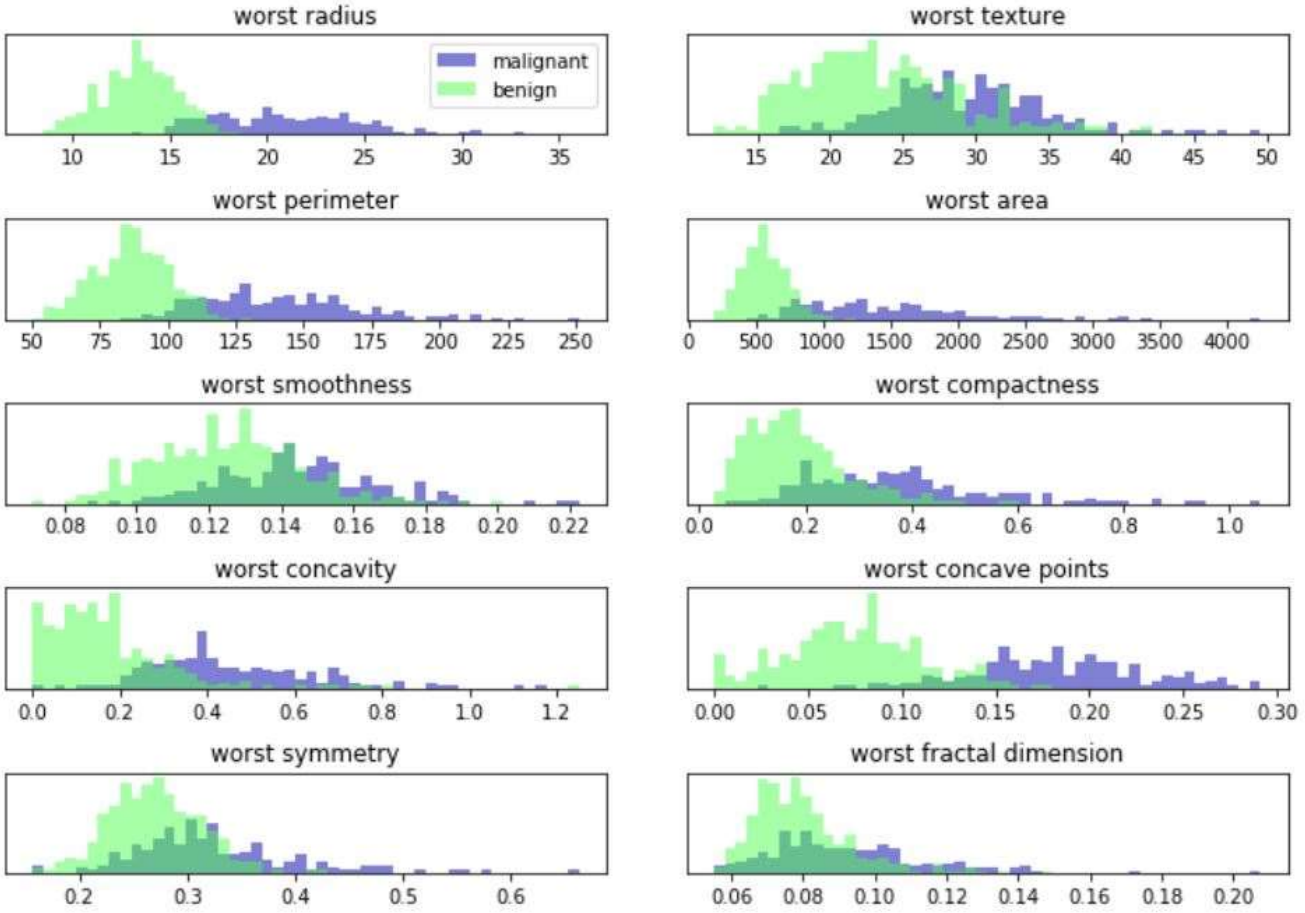
Predict tumor outcome (benign or malign)
based on tissue measurements

- Collect lab data about variables related to cancer tumors
- Build classification model

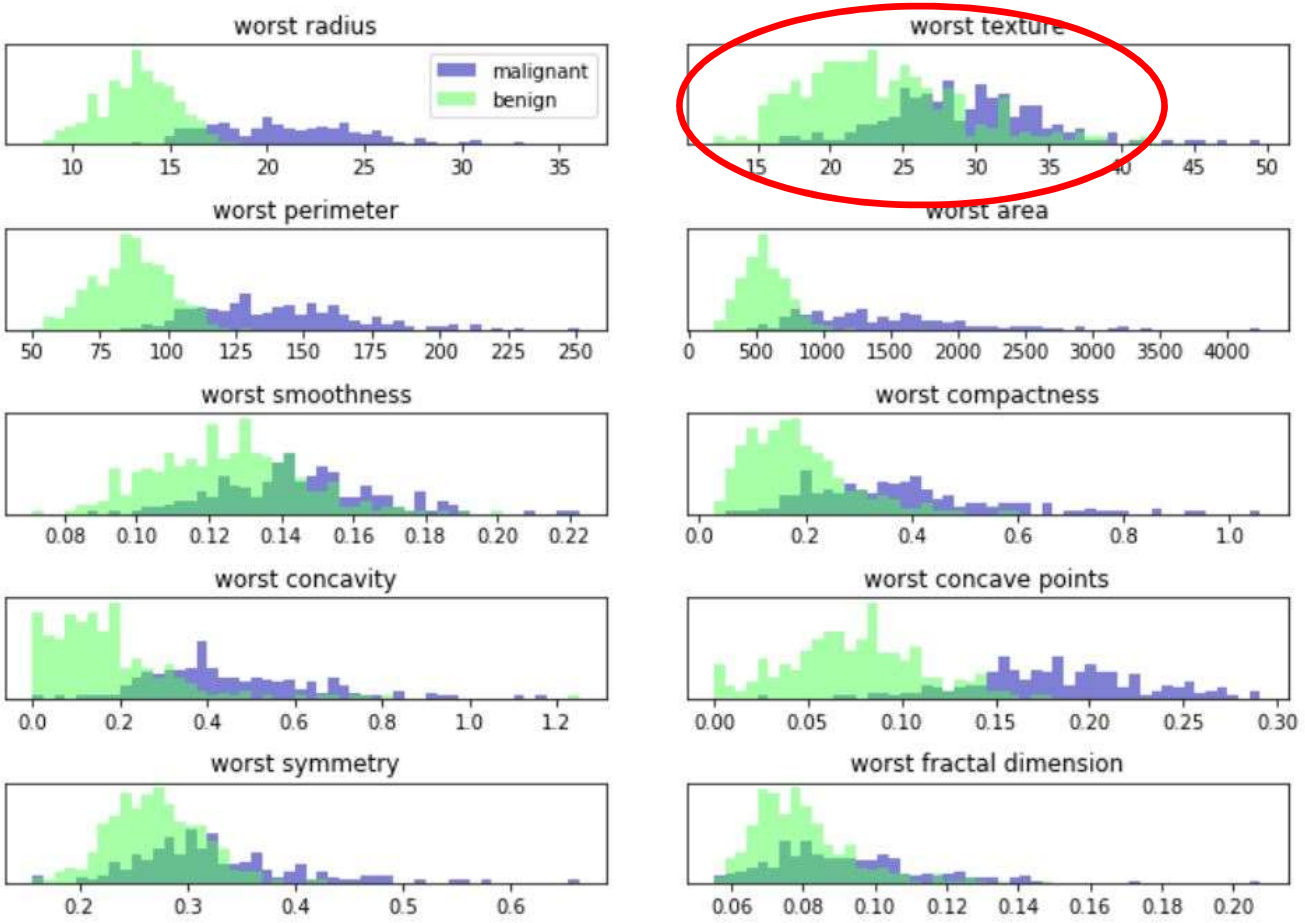
DATA VISUALIZATION

	<-------	----------	----------	-------	average	values	----------	----------	----------	---------->	<-------	--------	----------	--------	worst	values	----------	----------	----------
out	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry	fractal_dir	radius	texture	perimeter	area	smoothness	compactness	concavity	concave p	symmetry
M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	25.38	17.33	184.6	2019	0.1622	0.6656	0.7119	0.2654	0.4601
M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	24.99	23.41	158.8	1956	0.1238	0.1866	0.2416	0.186	0.275
M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	23.57	25.53	152.5	1709	0.1444	0.4245	0.4504	0.243	0.3613
M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	14.91	26.5	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638
M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	22.54	16.67	152.2	1575	0.1374	0.205	0.4	0.1625	0.2364
M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	15.47	23.75	103.4	741.6	0.1791	0.5249	0.5355	0.1741	0.3985
M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	22.88	27.66	153.2	1606	0.1442	0.2576	0.3784	0.1932	0.3063
M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	17.06	28.14	110.6	897	0.1654	0.3682	0.2678	0.1556	0.3196
M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	15.49	30.73	106.2	739.3	0.1703	0.5401	0.539	0.206	0.4378
M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	15.09	40.68	97.65	711.4	0.1853	1.058	1.105	0.221	0.4366
M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	19.19	33.88	123.8	1150	0.1181	0.1551	0.1459	0.09975	0.2948
M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	20.42	27.28	136.5	1299	0.1396	0.5609	0.3965	0.181	0.3792
M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	20.96	29.94	151.7	1332	0.1037	0.3903	0.3639	0.1767	0.3176
M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	16.84	27.66	112	876.5	0.1131	0.1924	0.2322	0.1119	0.2809
M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	15.03	32.01	108.8	697.7	0.1651	0.7725	0.6943	0.2208	0.3596
M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077	17.46	37.13	124.1	943.2	0.1678	0.6577	0.7026	0.1712	0.4218
M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922	19.07	30.88	123.4	1138	0.1464	0.1871	0.2914	0.1609	0.3029
M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356	20.96	31.48	136.8	1315	0.1789	0.4233	0.4784	0.2073	0.3706
M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395	27.32	30.88	186.8	2398	0.1512	0.315	0.5372	0.2388	0.2768
B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766	15.11	19.26	99.7	711.2	0.144	0.1773	0.239	0.1288	0.2977
B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811	14.5	20.49	96.09	630.5	0.1312	0.2776	0.189	0.07283	0.3184
B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905	10.23	15.66	65.13	314.9	0.1324	0.1148	0.08867	0.06227	0.245
M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032	18.07	19.08	125.1	980.9	0.139	0.5954	0.6305	0.2393	0.4667
M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278	29.17	35.59	188	2615	0.1401	0.26	0.3155	0.2009	0.2822

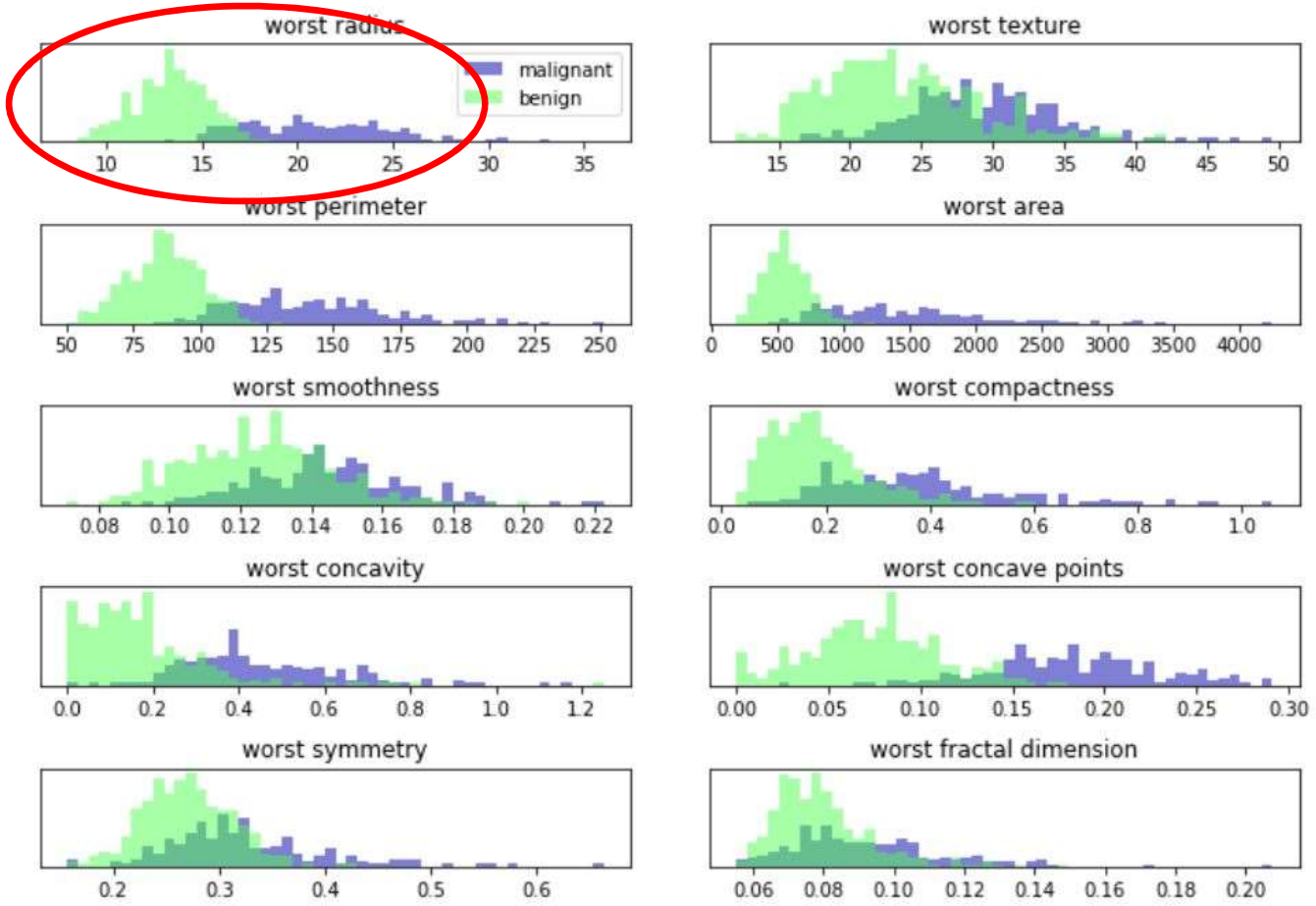
DATA VISUALIZATION



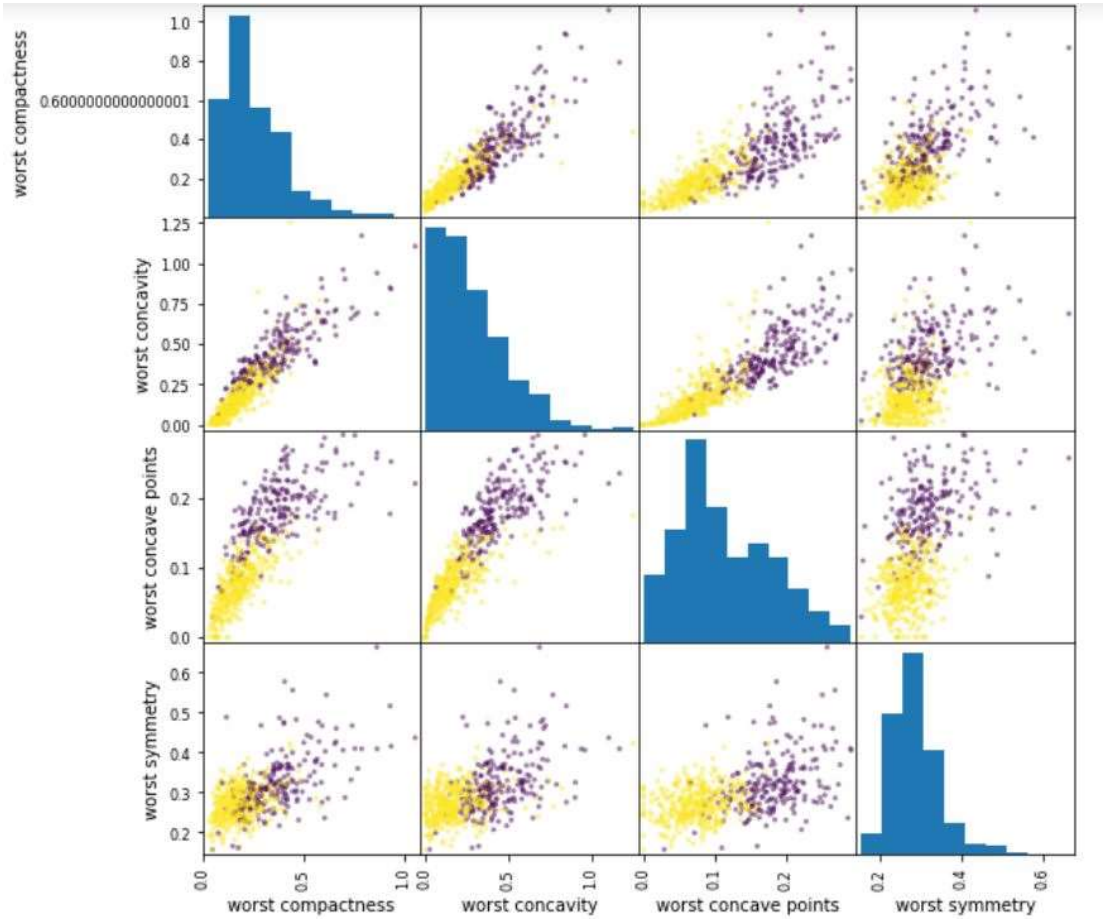
DATA VISUALIZATION



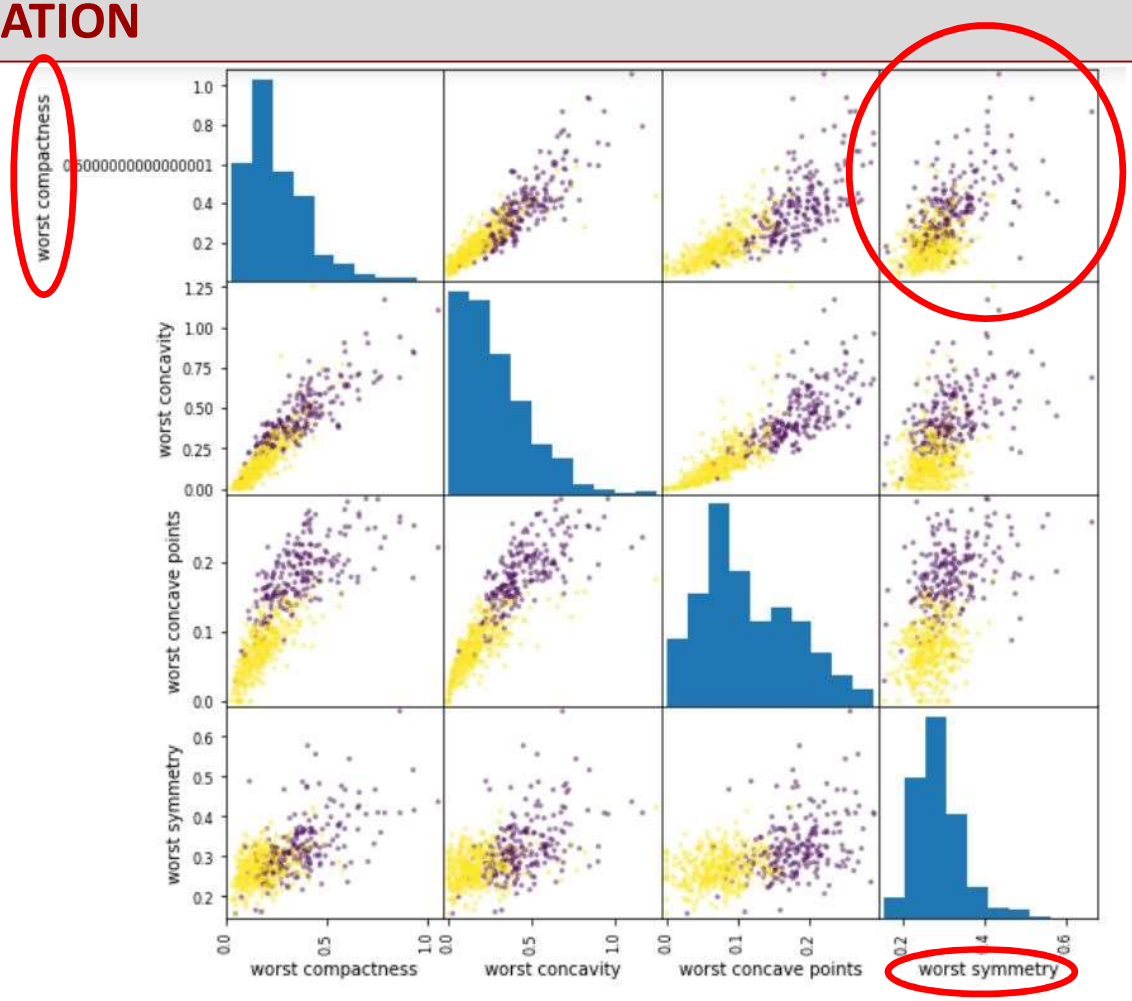
DATA VISUALIZATION



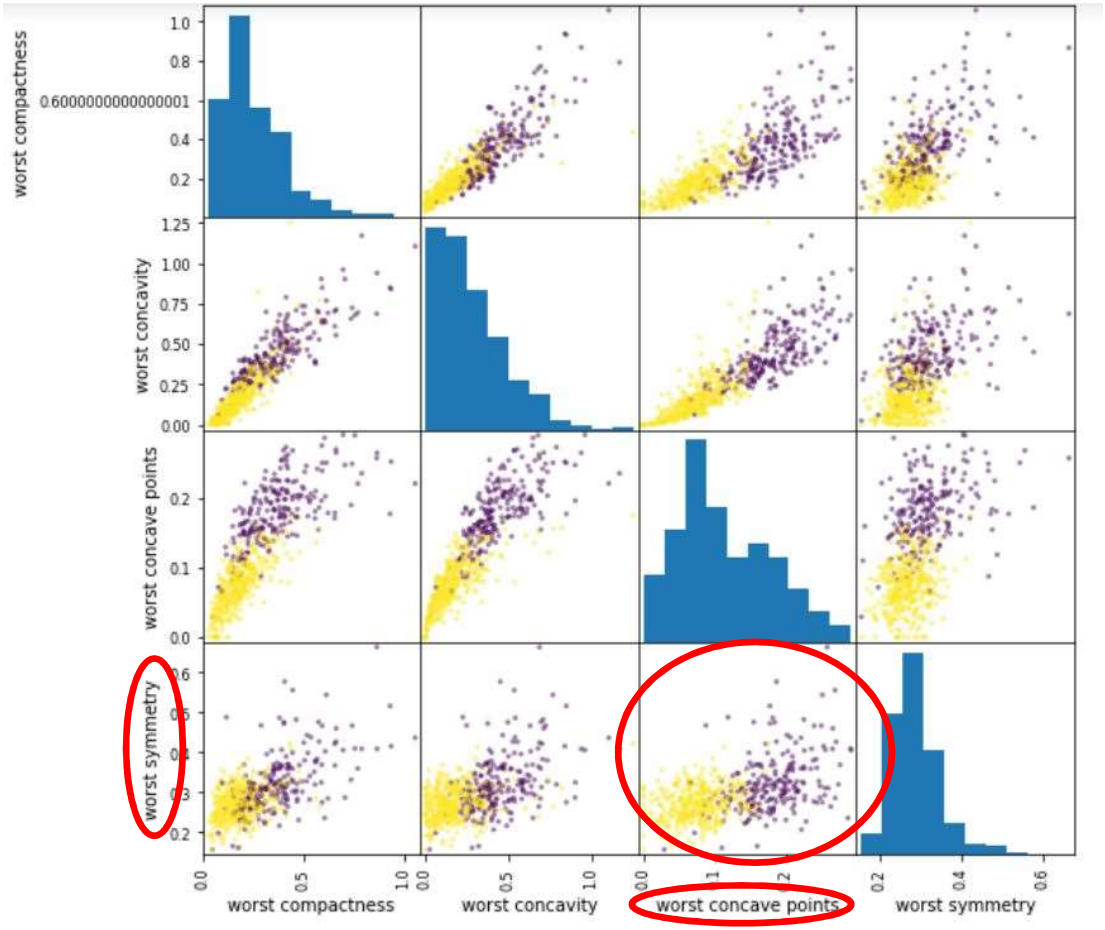
DATA VISUALIZATION



DATA VISUALIZATION



DATA VISUALIZATION

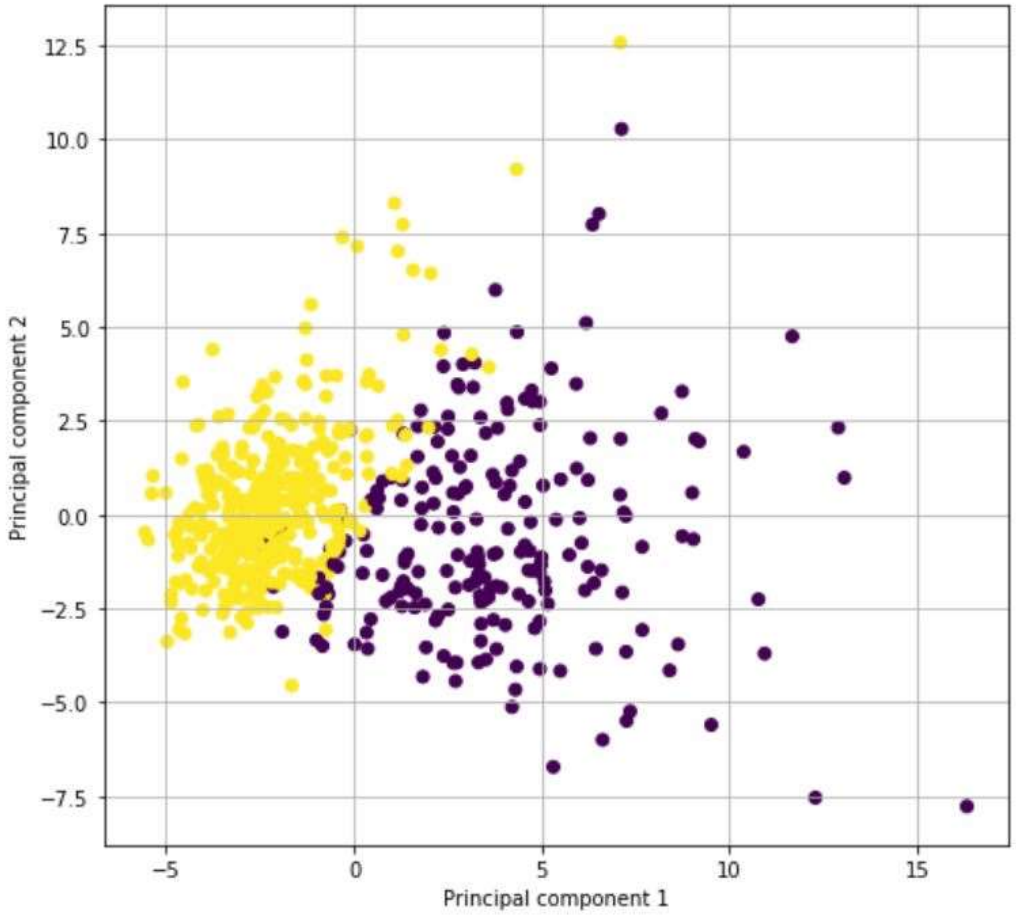


DATA ANALYTICS

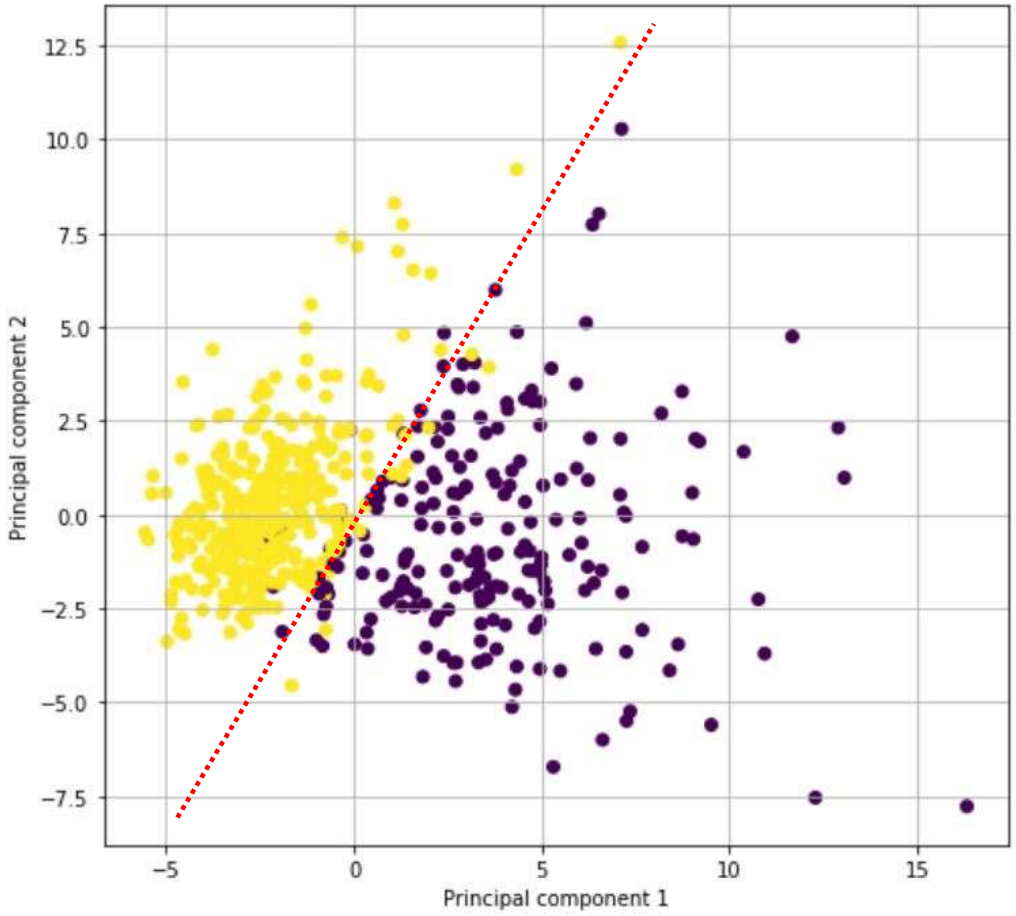
Data Analyst

- Searches subsets of variables to identify malign cancer
- Use PCA plot to verify if the variables are able to identifying cancer
- Develop a decision boundary

DATA VISUALIZATION



DATA VISUALIZATION



DATA VISUALIZATION

DATA ANALYTICS

Data Analytics focus is on answering **business questions**

- What happened?
- What will happen?

DATA ANALYTICS

What happened?

- Descriptive Stats
- Summary Tables (crosstabs, pivot tables)
- Data visualization
- Dashboards

DATA ANALYTICS

What happened? Descriptive Analytics

- Descriptive Stats
- Summary Tables (crosstabs, pivot tables)
- Data visualization
- Dashboards

DATA ANALYTICS

What happened? Descriptive Analytics

- Descriptive Stats
- Summary Tables (crosstabs, pivot tables)
- Data visualization
- Dashboards

What may happen?

- Prediction Models
- Classification Models
- Clustering methods

DATA ANALYTICS

What happened? Descriptive Analytics

- Descriptive Stats
- Summary Tables (crosstabs, pivot tables)
- Data visualization
- Dashboards

What may happen? Predictive Analytics

- Prediction Models
- Classification Models
- Clustering methods

DATA ANALYTICS

What happened? Descriptive Analytics

- Descriptive Stats
- Summary Tables (crosstabs, pivot tables)
- Data visualization
- Dashboards

Why did it happen?

What may happen? Predictive Analytics

- Prediction Models
- Classification Models
- Clustering methods

DATA ANALYTICS

What happened? — Descriptive Analytics

- Descriptive Stats
- Summary Tables (crosstabs, pivot tables)
- Data visualization
- Dashboards

Why did it happen? — Diagnostic Analytics

What may happen? — Predictive Analytics

- Prediction Models
- Classification Models
- Clustering methods

DATA ANALYTICS

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics

DATA ANALYTICS

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

INTRODUCTION

Past performance Historical data	Today observe & predict	Future performance results

INTRODUCTION

Past performance Historical data	Today observe & predict	Future performance results
What happened?		What may happen?

INTRODUCTION

Past performance Historical data	Today observe & predict	Future performance results
What happened? Describe/summarize data		What may happen? scenarios

INTRODUCTION

Past performance Historical data	Today observe & predict	Future performance results
What happened?		What may happen?
Describe/summarize data		scenarios
<u>Descriptive Stats</u>		
Barplots, scatterplots, boxplots		<u>Prediction Models</u>
Line charts, Histograms		prediction models
Averages, std. deviations		classification models
correlations		clustering methods
Descriptive Analytics		Predictive Analytics

INTRODUCTION

Past performance Historical data	Today Decisions & predictions	Future performance results
What happened?	What we want to happen?	What may happen?
Describe/summarize data		scenarios
<u>Descriptive Stats</u>		
Barplots, scatterplots, boxplots		<u>Prediction Models</u>
Line charts, Histograms		prediction models
Averages, std. deviations		classification models
correlations		clustering methods
Descriptive Analytics		Predictive Analytics

INTRODUCTION

Past performance Historical data	Today Decisions & predictions	Future performance results
What happened?	What we want to happen?	What may happen?
Describe/summarize data	What decisions are needed	scenarios
<u>Descriptive Stats</u>	to make things happen?	
Barplots, scatterplots, boxplots		<u>Prediction Models</u>
Line charts, Histograms		prediction models
Averages, std. deviations		classification models
correlations		clustering methods
Descriptive Analytics		Predictive Analytics

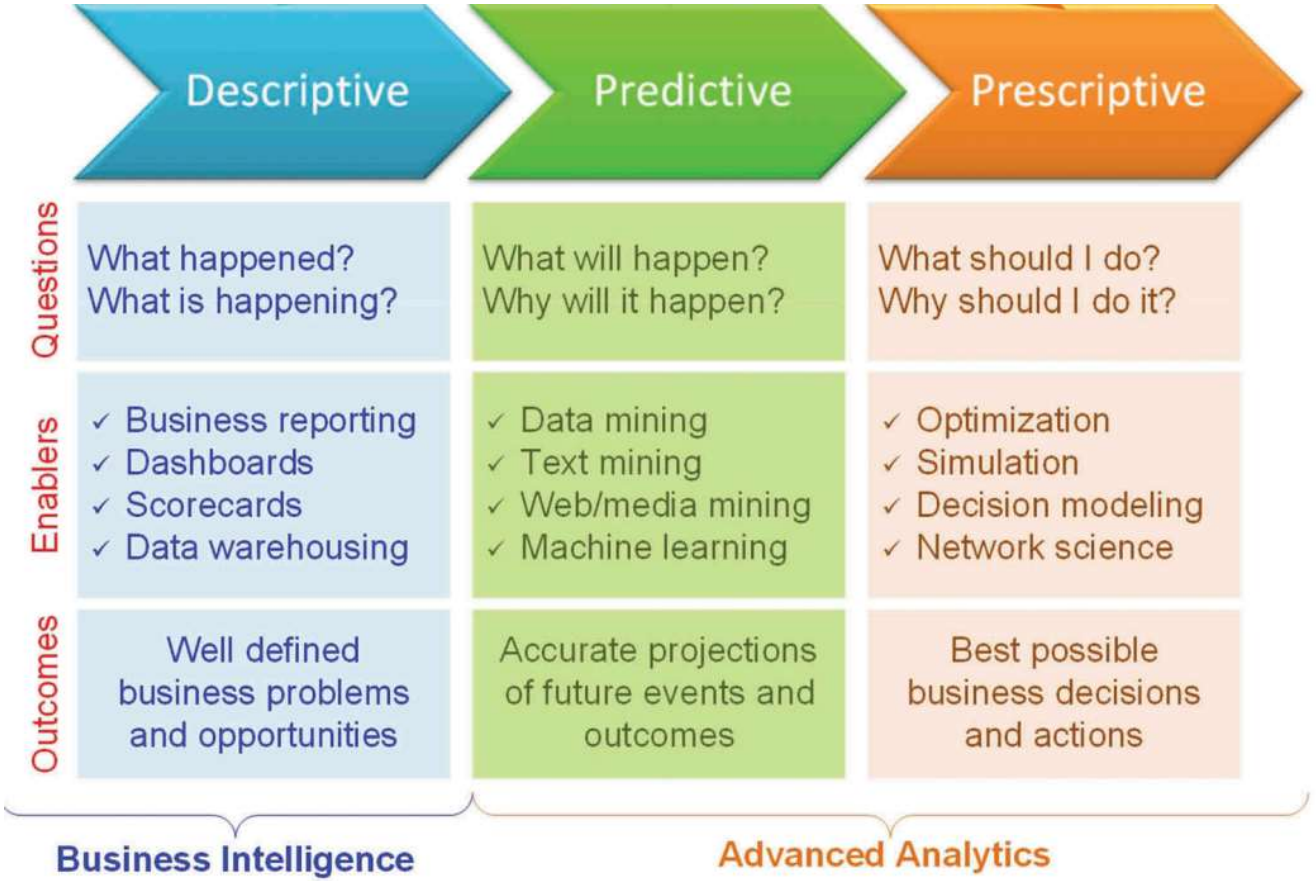
INTRODUCTION

Past performance Historical data	Today Decisions & predictions	Future performance results
What happened?	What we want to happen?	What may happen?
Describe/summarize data	What decisions are needed	scenarios
<u>Descriptive Stats</u>	to make things happen?	
Barplots, scatterplots, boxplots		<u>Prediction Models</u>
Line charts, Histograms	Simulation models	prediction models
Averages, std. deviations	Optimization models	classification models
correlations		clustering methods
Descriptive Analytics		Predictive Analytics

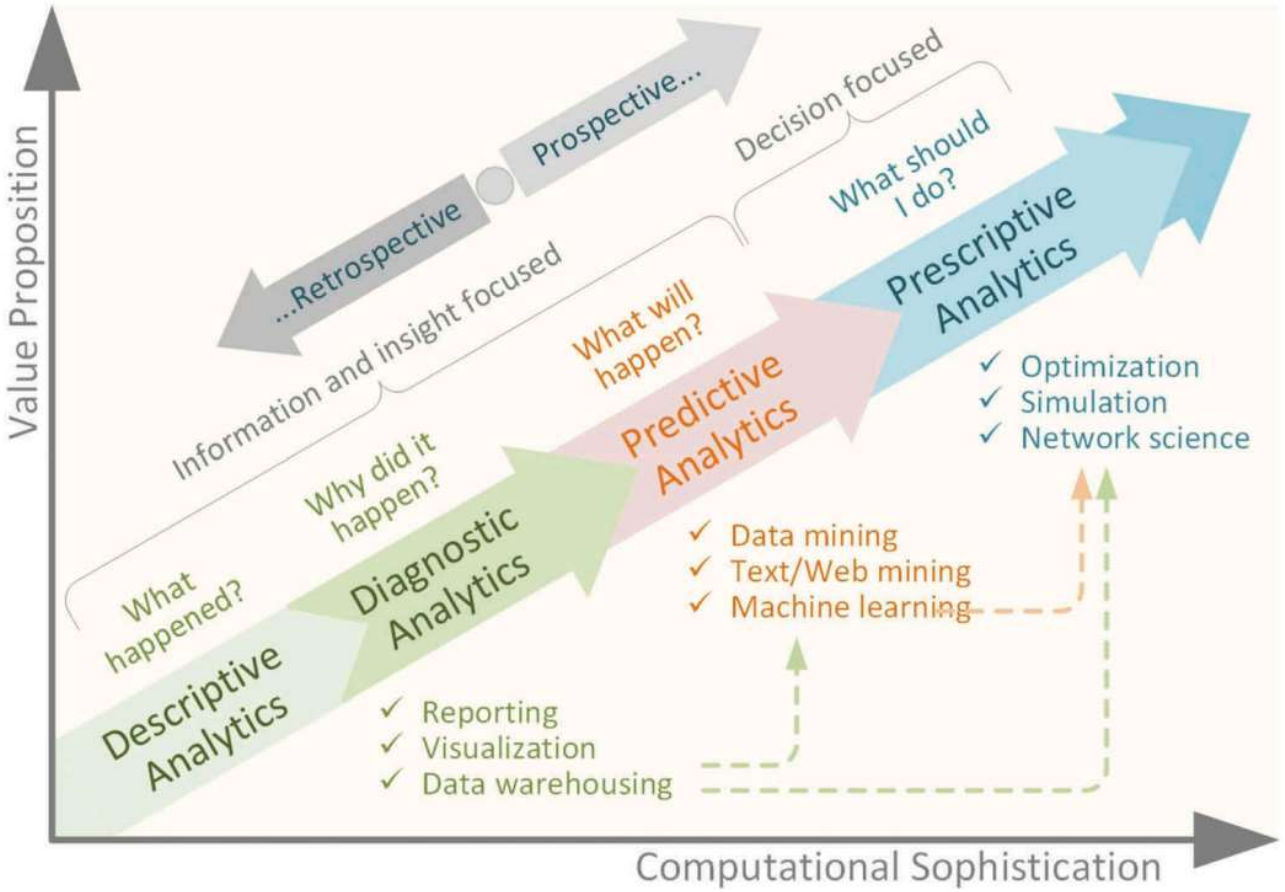
INTRODUCTION

Past performance Historical data	Today Decisions & predictions	Future performance results
What happened?	What we want to happen?	What may happen?
Describe/summarize data	What decisions are needed	scenarios
<u>Descriptive Stats</u>	to make things happen?	
Barplots, scatterplots, boxplots		<u>Prediction Models</u>
Line charts, Histograms	Simulation models	prediction models
Averages, std. deviations	Optimization models	classification models
correlations		clustering methods
Descriptive Analytics	Prescriptive Analytics	Predictive Analytics

DATA ANALYTICS



DATA ANALYTICS



DATA SCIENCE

Data Scientist
vs
Data Analyst

DATA ANALYTICS

Both

Collect data

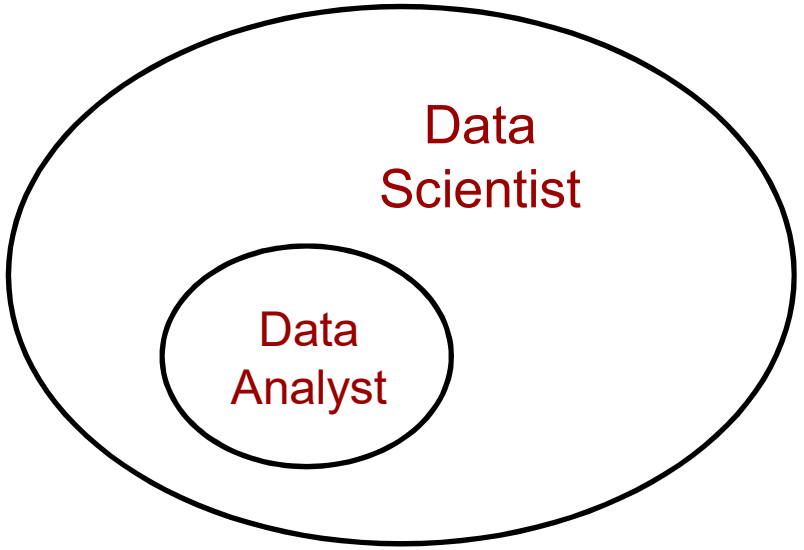
- Data wrangling (organizing, sorting, merging data sets)
- Cleaning
- Converting data formats
- Querying databases

Summarize data

- Tables (crosstabs, pivot tables)
- Data visualisation
- Dashboards

Build prediction/classification models

DATA SCIENCE

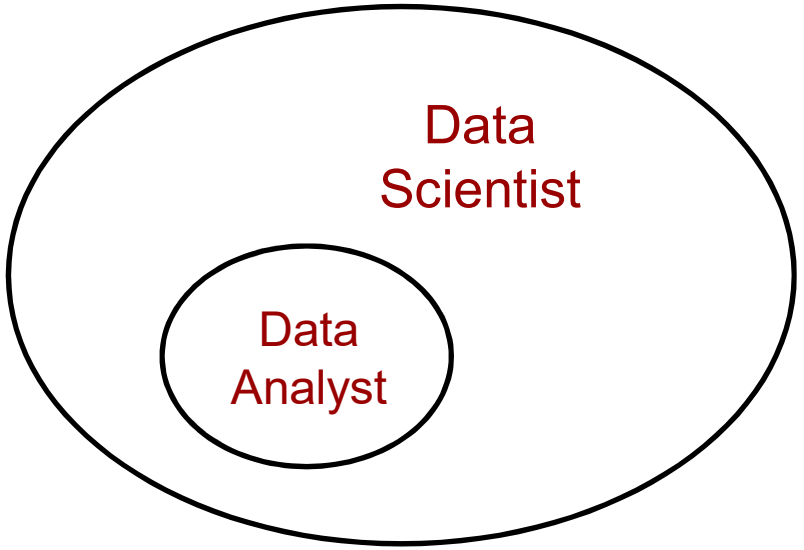


DATA SCIENCE

Data Analyst

- Data query
- Data wrangling
- Data Discovery

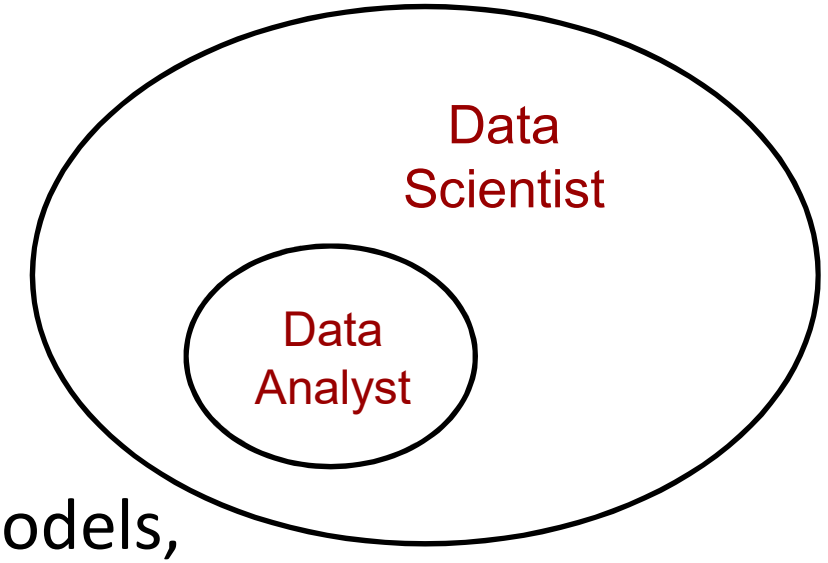
(pivot tables,
Regression models,
Classification models)



DATA SCIENCE

Data Scientist

- Web-data app
- Cloud-data app
- Big data tools
- ML (Adv. Regression models,
  Adv. Classification models
  Deep learning)



DATA SCIENCE

Data Scientists have an increasing need of becoming proficient in Web Applications and AI

- Big data Analytics
- Web data visualization
- Online dash boarding
- Text analytics
- Image data analytics
- IoT

SOFTWARE

Data Analytics+

- Tableau, Power BI
- Python, R
- SQL, NoSQL
- HTML, CSS

Data Science

- Linux
- AWS, GCP, Azure
- Java Script
- Hadoop, Spark, Scala

DATA SCIENCE

Data Science and Data Analytics -Job Market-

DATA ANALYSTS & DATA SCIENCE - JOBS MARKET

Analytical Rigor

DSA Framework Category	Functional Role	Sample Occupations
Data Scientists & Advanced Analytics	Create sophisticated analytical models used to build new datasets and derive new insights from data	Data Scientist
Data Analysts	Leverage data analysis and modeling techniques to solve problems and glean insight across functional domains	Data Analysts Business Intelligence Analyst
Data Systems Developers	Design, build and maintain and organization's data and analytical infrastructure	Systems Analyst Database Administrator
Analytics Managers	Oversee analytical operations and communicate insights to executives	Chief Analytics Officer Marketing Analystics Manager
Functional Analysts	Utilize data and analytical models to inform specific functions and business decisions	Business Analyst Financial Analyst
Data-Driven Decision Makers	Leverage data to inform strategic and operational decisions	IT Project Manager Marketing Manager

IBM Burning Glass Tech Report

DATA ANALYSTS & DATA SCIENCE - JOBS MARKET

Analytical Rigor

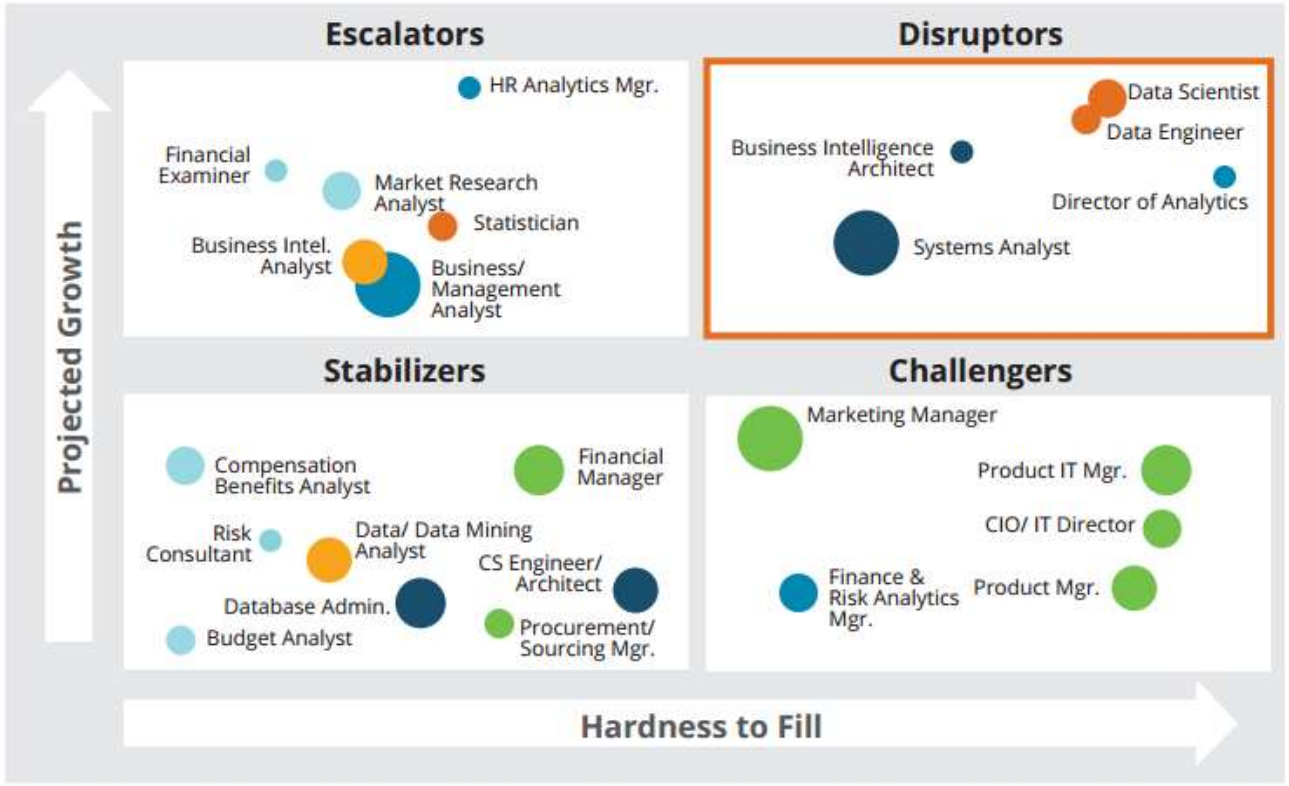
DSA Framework Category	Functional Role	Sample Occupations
Data Scientists & Advanced Analytics	Create sophisticated analytical models used to build new datasets and derive new insights from data	Data Scientist
Data Analysts	Leverage data analysis and modeling techniques to solve problems and glean insight across functional domains	Data Analysts Business Intelligence Analyst
Data Systems Developers	Design, build and maintain and organization's data and analytical infrastructure	Systems Analyst Database Administrator
Analytics Managers	Oversee analytical operations and communicate insights to executives	Chief Analytics Officer Marketing Analystics Manager
Functional Analysts	Utilize data and analytical models to inform specific functions and business decisions	Business Analyst Financial Analyst
Data-Driven Decision Makers	Leverage data to inform strategic and operational decisions	IT Project Manager Marketing Manager

IBM Burning Glass Tech Report

DATA ANALYSTS & DATA SCIENCE – JOBS CLASSIFICATION

IBM Burning Glass Tech Report

DATA ANALYSTS & DATA SCIENCE - CURRENT JOB MARKET



Key:

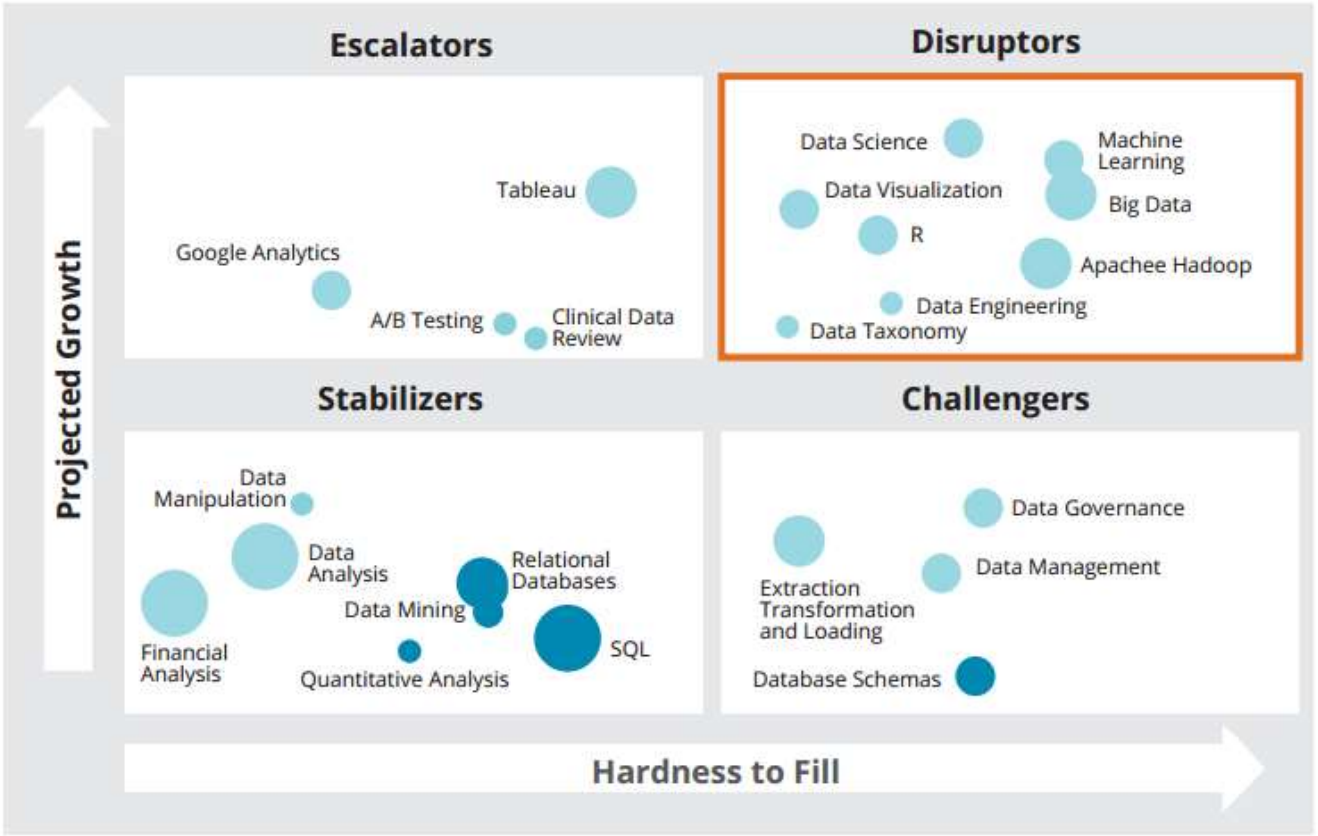
- Data Driven Decision Makers
- Functional Analysts
- Data System Developers
- Data Analysts
- Data Scientists & Adv. Analysts
- Analytics Managers

Postings:

- Less than 5,000
- 5,000 - 10,000
- 10,000 - 50,000
- 50,000- 150,000
- +150,000

IBM Burning Glass Tech Report

DATA ANALYSTS & DATA SCIENCE – SKILLS NEEDED - JOB MARKET



IBM Burning Glass Tech Report

DATA ANALYSTS DATA SCIENCE – Highest Paying Analytical Skills (7000+ postings)

Skill Name	Average Salary
MapReduce	$115,907
PIG	$114,474
Machine Learning	$112,732
Apache Hive	$112,242
Apache Hadoop	$110,562
Big Data	$109,895
Data Science	$107,287
NoSQL	$105,053
Predictive Analytics	$103,235
MongoDB	$101,323

IBM Burning Glass Tech Report

DATA SCIENCE

Cesar Acosta
acostame@usc.edu
MS Analytics - Director
University of Southern California
Department of Industrial and Systems Engineering

https://github.com/cesar-acosta

Data Science and Analytics

Thank you!

Data Science and Analytics

Questions?