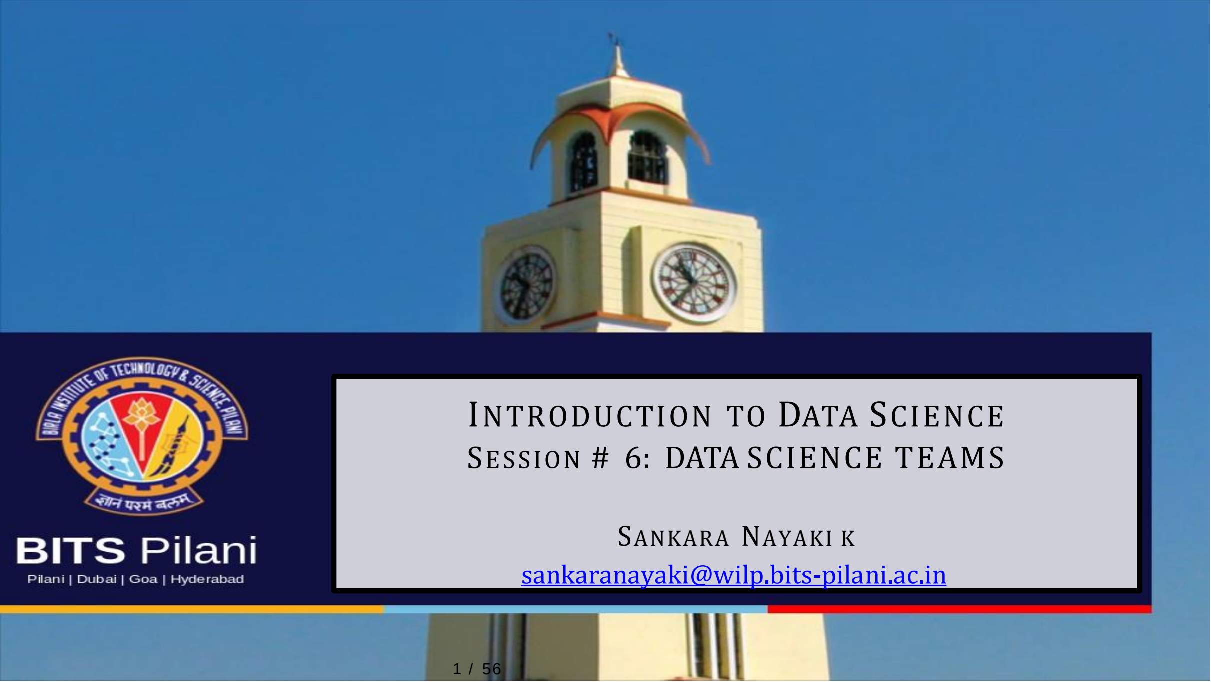


BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science

Session # 6: DATA SCIENCE TEAMS

Sankara Nayaki K

sankaranayaki@wilp.bits-pilani.ac.in

The instructor is gratefully acknowledging the authors who made their course materials freely available online.

References:

- Introducing Data Science by Cielen, Meysman and Ali
- Storytelling with Data by Cole Nussbaumer Knaflic; Wiley
- Introduction to Data Mining by Tan, Steinbach and Vipin Kumar
- The Art of Data Science by Roger D Peng and Elizabeth Matsui
- Python Data Science Handbook: Essential tools for working with data by Jake VanderPlas

Table of Contents

Course Handout

M1 Introduction to Data Science

M2 Data Analytics

M3 Data Science Process

M4 Data Science Teams

M5 Data and Data Models

M6 Data wrangling and Feature Engineering

M7 Data visualization

M8 Storytelling with Data

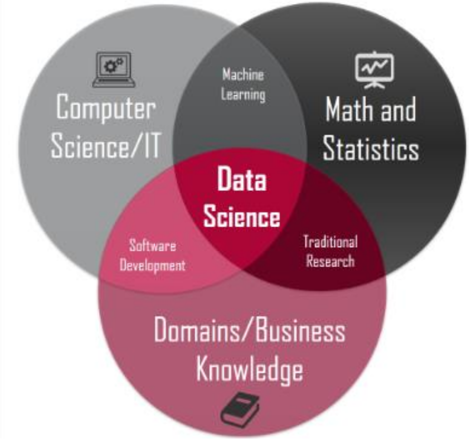
M9 Ethics for Data Science

Table of Contents

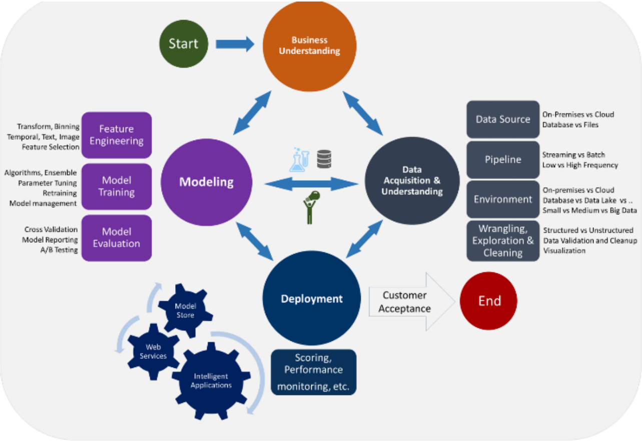
1 Course Handout

2 Data Science Teams

Introduction



DATA SCIENCE LIFECYCLE



Start

Business Understanding

Data Acquisition & Understanding

Data Source — On-Premises vs Cloud, Database vs Files

Pipeline — Streaming vs Batch, Low vs High Frequency

Environment — On-premises vs Cloud, Database vs Data Lake vs .., Small vs Medium vs Big Data

Wrangling, Exploration & Cleaning — Structured vs Unstructured, Data Validation and Cleanup, Visualization

Modeling

Feature Engineering — Transform, Binning, Temporal, Text, Image, Feature Selection

Model Training — Algorithms, Ensemble, Parameter Tuning, Retraining, Model management

Model Evaluation — Cross Validation, Model Reporting, A/B Testing

Deployment

Scoring, Performance monitoring, etc.

Model Store

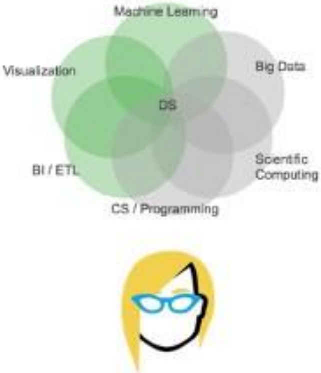
Web Services

Intelligent Applications

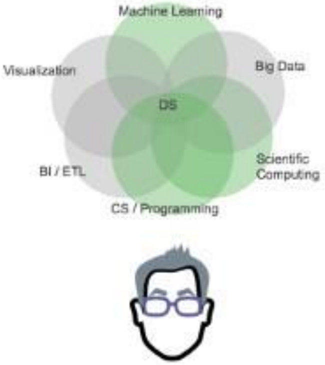
Customer Acceptance

End

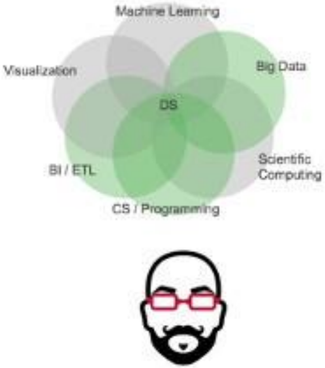
DATA TEAM



Statistician / Analyst

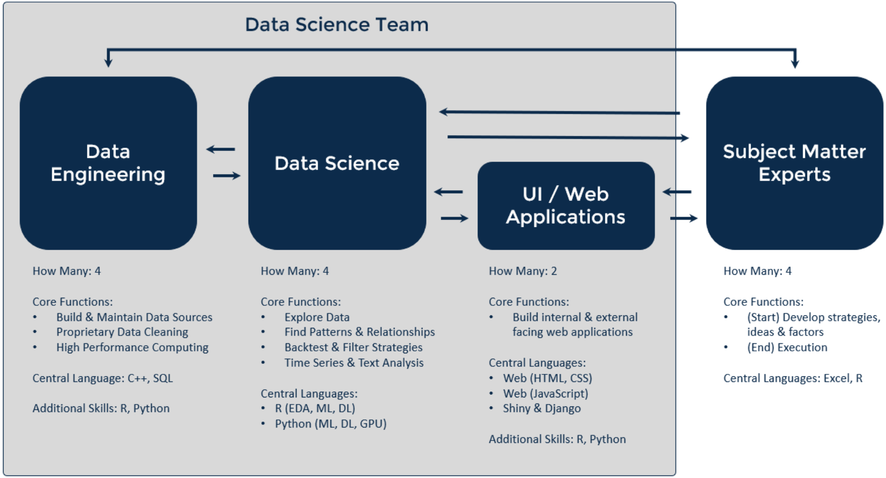


Research / Computational Scientist



Developer / Engineer

DATA SCIENCE TEAM



Data Science Team

Data Engineering

Data Science

UI / Web Applications

Subject Matter Experts

How Many: 4

Core Functions:

- Build & Maintain Data Sources
- Proprietary Data Cleaning
- High Performance Computing

Central Language: C++, SQL

Additional Skills: R, Python

How Many: 4

Core Functions:

- Explore Data
- Find Patterns & Relationships
- Backtest & Filter Strategies
- Time Series & Text Analysis

Central Languages:

- R (EDA, ML, DL)
- Python (ML, DL, GPU)

How Many: 2

Core Functions:

- Build internal & external facing web applications

Central Languages:

- Web (HTML, CSS)
- Web (JavaScript)
- Shiny & Django

Additional Skills: R, Python

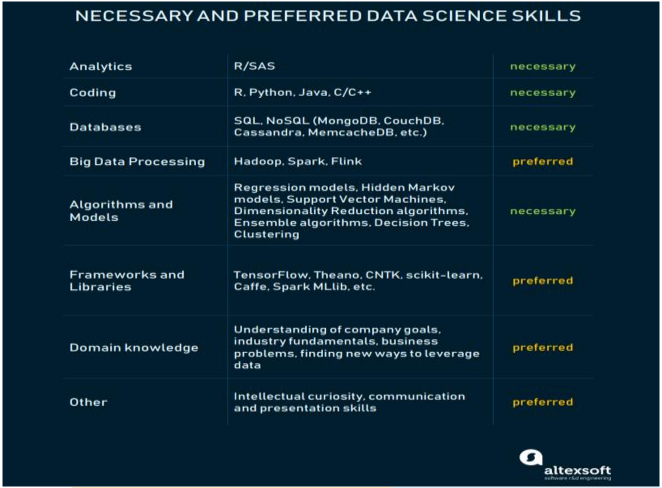
How Many: 4

Core Functions:

- (Start) Develop strategies, ideas & factors
- (End) Execution

Central Languages: Excel, R

DATA SCIENCE – SKILL SET

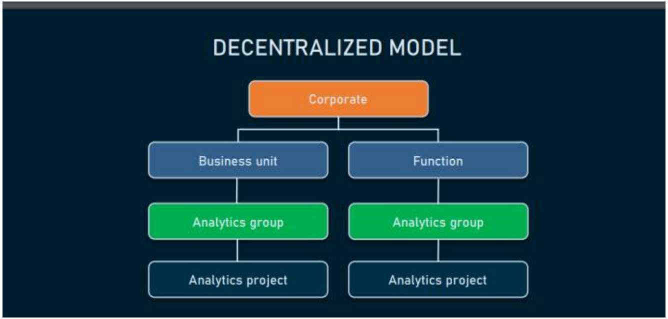


NECESSARY AND PREFERRED DATA SCIENCE SKILLS

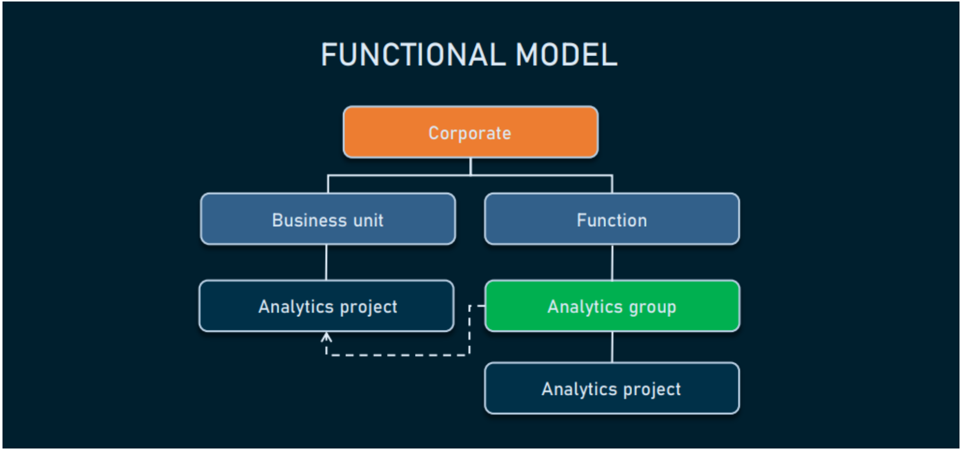
Analytics	R/SAS	necessary
Coding	R, Python, Java, C/C++	necessary
Databases	SQL, NoSQL (MongoDB, CouchDB, Cassandra, MemcacheDB, etc.)	necessary
Big Data Processing	Hadoop, Spark, Flink	preferred
Algorithms and Models	Regression models, Hidden Markov models, Support Vector Machines, Dimensionality Reduction algorithms, Ensemble algorithms, Decision Trees, Clustering	necessary
Frameworks and Libraries	TensorFlow, Theano, CNTK, scikit-learn, Caffe, Spark MLlib, etc.	preferred
Domain knowledge	Understanding of company goals, industry fundamentals, business problems, finding new ways to leverage data	preferred
Other	Intellectual curiosity, communication and presentation skills	preferred

altexsoft

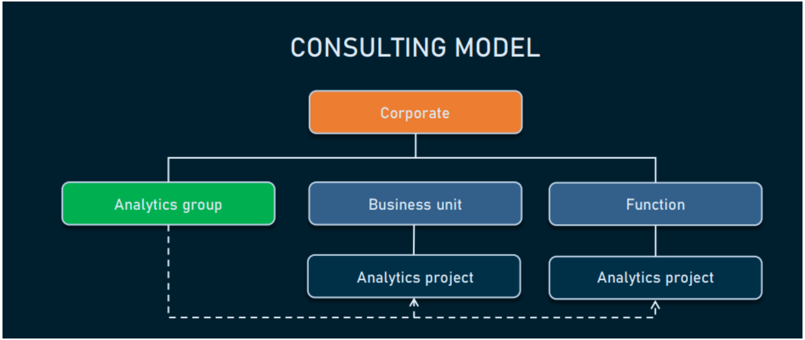
MODELS


DECENTRALIZED MODEL
Corporate
Business unit
Function
Analytics group
Analytics group
Analytics project
Analytics project

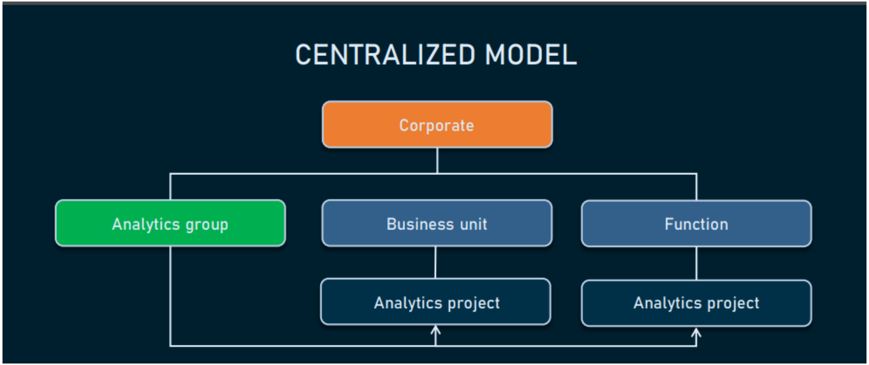
MODELS


FUNCTIONAL MODEL
Corporate
Business unit
Function
Analytics project
Analytics group
Analytics project

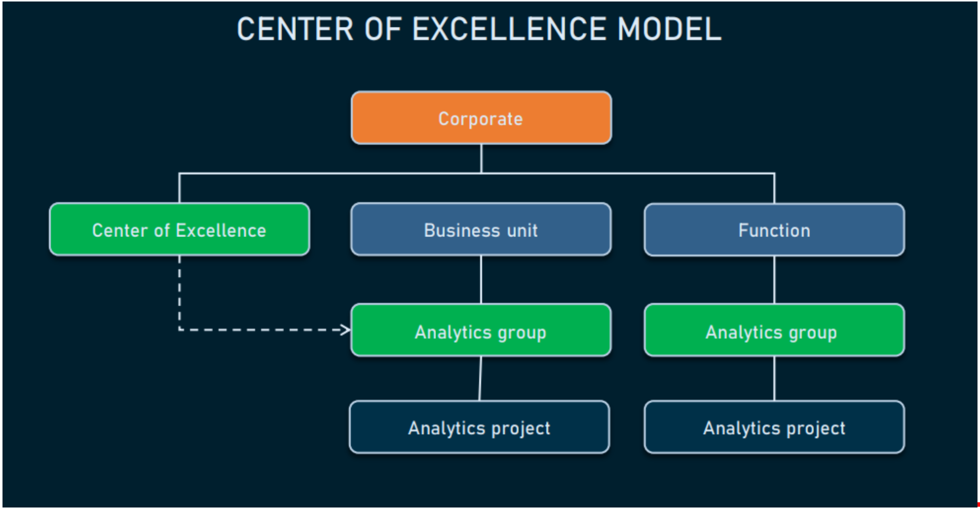
MODELS


CONSULTING MODEL
Corporate
Analytics group
Business unit
Function
Analytics project
Analytics project

MODELS


CENTRALIZED MODEL
Corporate
Analytics group
Business unit
Function
Analytics project
Analytics project

MODELS



CENTER OF EXCELLENCE MODEL

Corporate

Center of Excellence

Business unit

Function

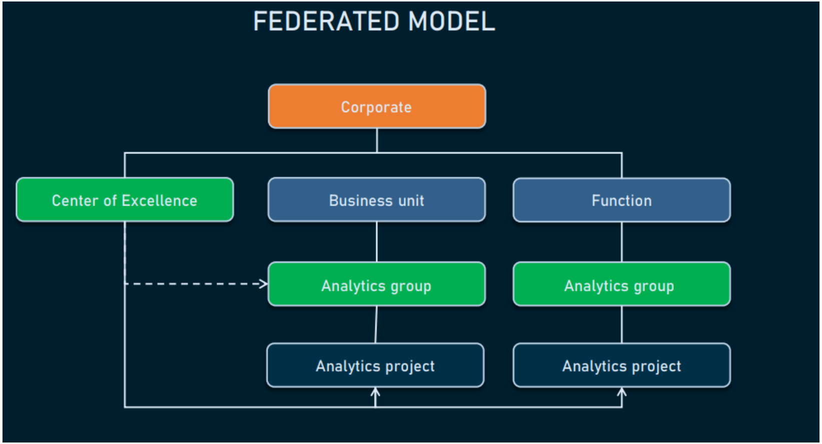
Analytics group

Analytics group

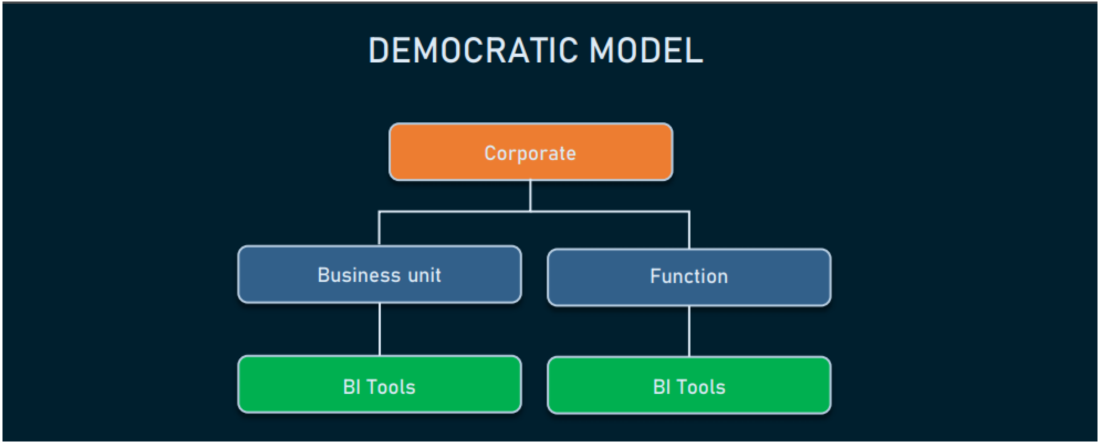
Analytics project

Analytics project

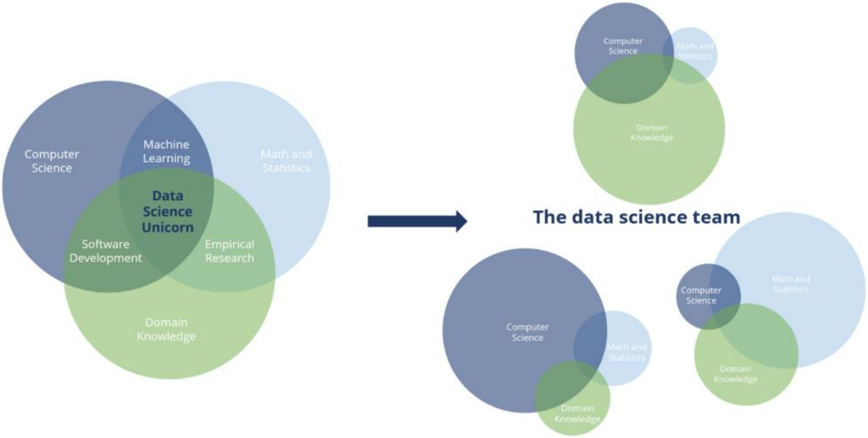
MODELS


FEDERATED MODEL
Corporate
Center of Excellence
Business unit
Function
Analytics group
Analytics group
Analytics project
Analytics project

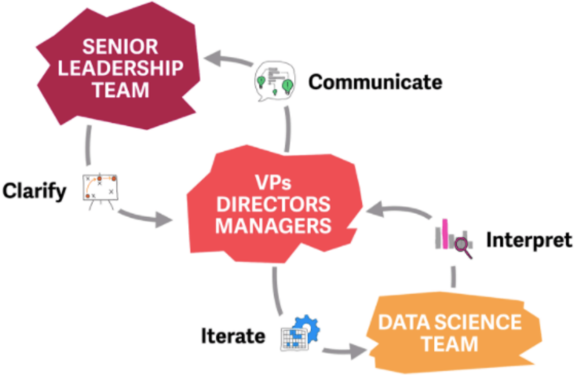
MODELS


DEMOCRATIC MODEL
Corporate
Business unit
Function
BI Tools
BI Tools

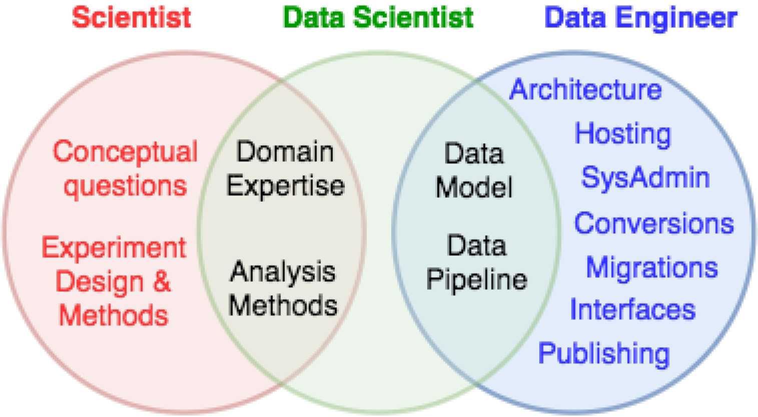
DEFINING DATA SCIENCE TEAM



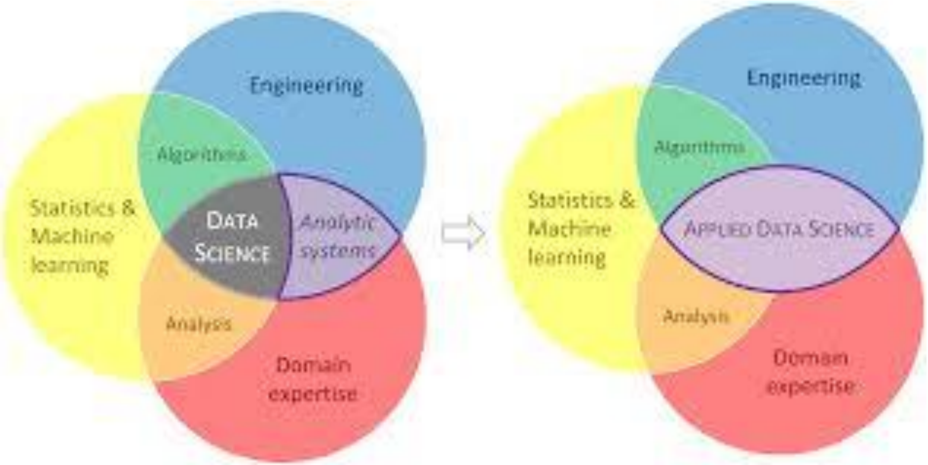
DATA SCIENCE - ROLES



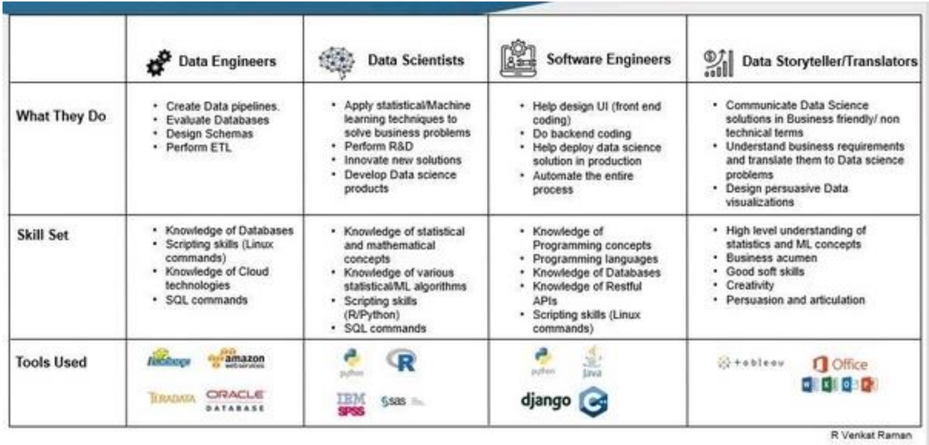
DATA SCIENTIST – JOB ROLE



UNDERSTANDING DATA SCIENCE



DATA SCIENCE TEAM EFFORT



	Data Engineers	Data Scientists	Software Engineers	Data Storyteller/Translators
What They Do	• Create Data pipelines. • Evaluate Databases • Design Schemas • Perform ETL	• Apply statistical/Machine learning techniques to solve business problems • Perform R&D • Innovate new solutions • Develop Data science products	• Help design UI (front end coding) • Do backend coding • Help deploy data science solution in production • Automate the entire process	• Communicate Data Science solutions in Business friendly/ non technical terms • Understand business requirements and translate them to Data science problems • Design persuasive Data visualizations
Skill Set	• Knowledge of Databases • Scripting skills (Linux commands) • Knowledge of Cloud technologies • SQL commands	• Knowledge of statistical and mathematical concepts • Knowledge of various statistical/ML algorithms • Scripting skills (R/Python) • SQL commands	• Knowledge of Programming concepts • Programming languages • Knowledge of Databases • Knowledge of Restful APIs • Scripting skills (Linux commands)	• High level understanding of statistics and ML concepts • Business acumen • Good soft skills • Creativity • Persuasion and articulation
Tools Used	hadoop amazon web services TERADATA ORACLE DATABASE	python R IBM SPSS sas	python java django C	tableau Office

R Venkat Raman

Maturity Levels with Data

Phases	Data Engineering			Data Science			Data as 'Culture'
Maturity	Data Collection	Data Storage	Data Transformatio n	Reporting	Insights	Consumption	Decisions
Activitie	Int/External Logs, IOT Stage/Strea m	Un/Structured SQL, Spark.. Data lake..	ETL Cleaning Preparation	Metrics/KPI Aggregates Reports	EDA ML AI	Narrative Info Design Data Stories	Change Mgmt Workflows Actions

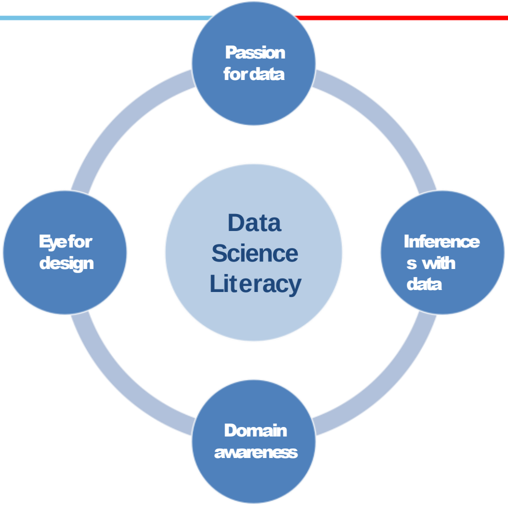
https://techcrunch.com/2019/12/13/when-and-how-to-build-out-your-data-science-team/

MATURITY LEVELS WITH DATA

	Data Engineering			Data Science			Data as 'Culture'
Maturity Phases	Data Collection	Data Storage	Data Transformation	Reporting	Insights	Consumption	Decisions
Activities	Int/External Logs, IOT Stage/Stream	Un/Structured SQL, Spark.. Data lake..	ETL Cleaning Preparation	Metrics/KPII Aggregates Reports	EDA ML AI	Narrative Info Design Data Stories	Change Mgmt Workflows Actions

https://techcrunch.com/2019/12/13/when-and-how-to-build-out-your-data-science-team/

PRE-REQUISITE FOR EVERY ROLE IN DATA SCIENCE



Data Science Literacy

Passion for data

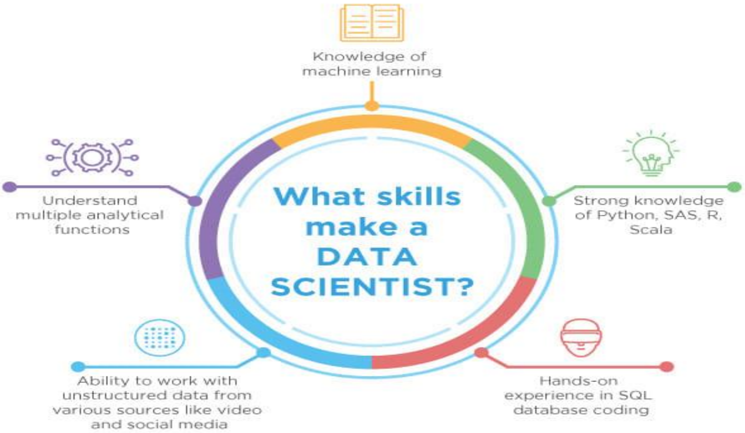
Inferences with data

Domain awareness

Eye for design

https://towardsdatascience.com/whats-the-secret-sauce-to-transforming-into-a-unicorn-in-data-science-94082b01c39d

SKILLS



What skills make a DATA SCIENTIST?

Knowledge of machine learning

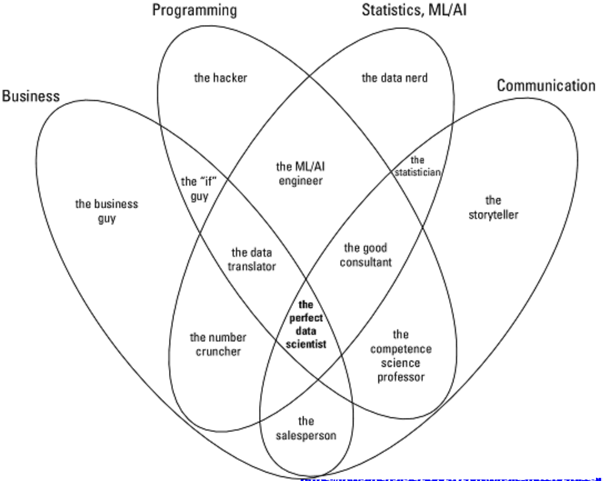
Strong knowledge of Python, SAS, R, Scala

Hands-on experience in SQL database coding

Ability to work with unstructured data from various sources like video and social media

Understand multiple analytical functions

PERFECT DATA SCIENTIST



https://towardsdatascience.com/whats-the-secret-sauce-to-transforming-into-a-unicorn-in-data-science-94082b01c39d

5 ROLES & SKILLS IN DATA SCIENCE

	1. Data Translator	2. Data Scientist
Responsibilities	- Own from inception to adoption - Translate across domain & data - Act as a glue in the team	- Devise analytics approach - Analyze data & identify insights - Build ML models
Skills	- Domain expertise - Business analysis & solutioning - Interpersonal & mentoring skills	- Statistics and machine learning - Identify & interpret insights - Scripting skills
Closest role	Business analyst, Domain experts	Statistician, ML experts

https://techhq.com/2019/12/a-complete-data-science-team-requires-more-than-just-data-scientists/

5 ROLES & SKILLS IN DATA SCIENCE

	3. Information Designer	4. ML Engineer
Responsibilities	- Ensure consumption of insights - Design information architecture - Understand user, drive adoption	- Package data science solution - Productionizing, DevOps - Data pipelines/integration
Skills	- Information design - User centered design - Aspects of interface/visual design	- Software engineering - Data handling - Front-end / Back-end coding
Closest role	UX Designer, Interaction designer	Software engineer, Data architect

https://techhq.com/2019/12/a-complete-data-science-team-requires-more-than-just-data-scientists/

5 ROLES & SKILLS IN DATA SCIENCE

5. Data Science Manager

Responsibilities

- Identify roadmap & scale maturity
- Ensure biz value from data science
- Drive a culture of data

Skills

- Project management
- Business analysis, solutioning
- Team handling

Closest role

Project manager, Business analyst

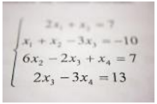
https://techhq.com/2019/12/a-complete-data-science-team-requires-more-than-just-data-scientists/

5 CORE ROLES IN DATA SCIENCE

Data Science Manager

Data Translator

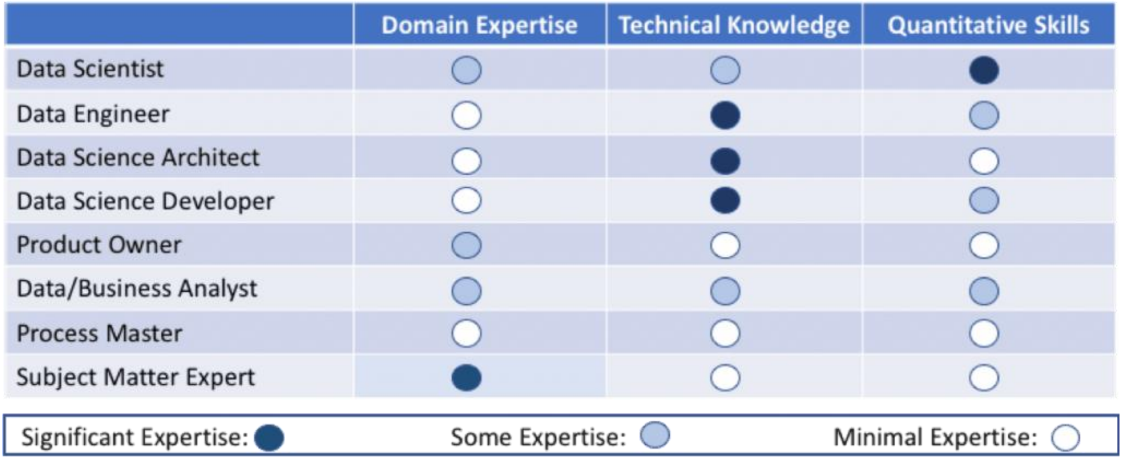
Data Scientist



Information Designer

ML Engineer

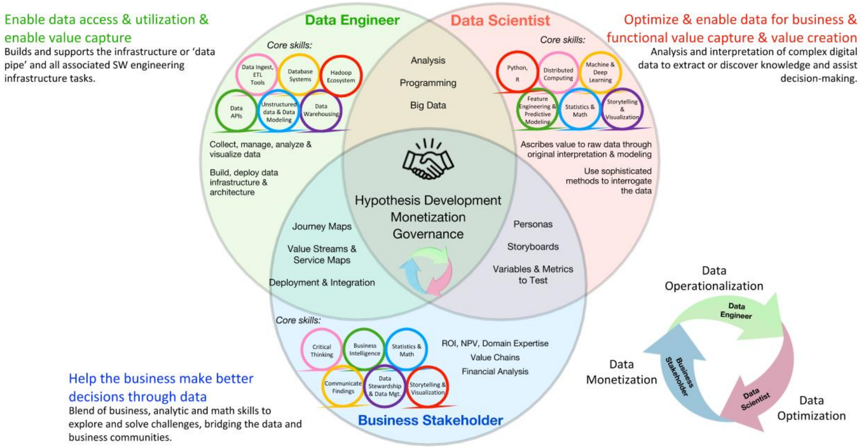
ROLES – EXTENDED WITH KNOWLEDGE



	Domain Expertise	Technical Knowledge	Quantitative Skills
Data Scientist	◍	◍	●
Data Engineer	○	●	◍
Data Science Architect	○	●	○
Data Science Developer	○	●	◍
Product Owner	◍	○	○
Data/Business Analyst	◍	◍	◍
Process Master	○	○	○
Subject Matter Expert	●	○	○

Significant Expertise: ● Some Expertise: ◍ Minimal Expertise: ○

INTERACTION AMONG ROLES



Enable data access & utilization & enable value capture

Builds and supports the infrastructure or 'data pipe' and all associated SW engineering infrastructure tasks.

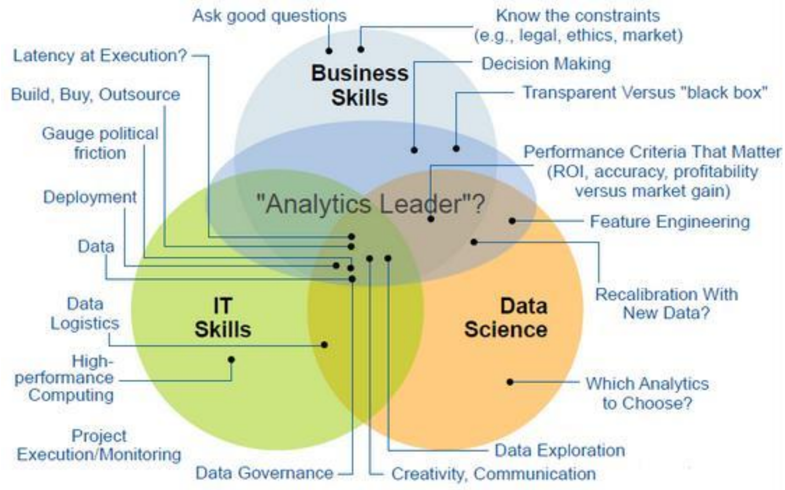
Optimize & enable data for business & functional value capture & value creation

Analysis and interpretation of complex digital data to extract or discover knowledge and assist decision-making.

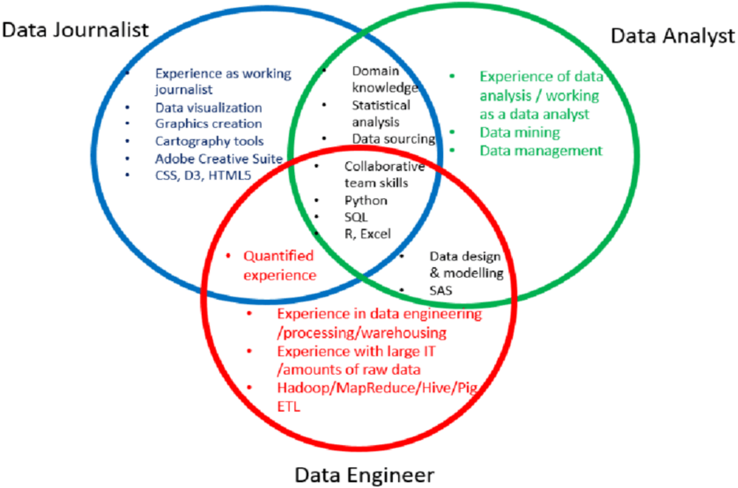
Help the business make better decisions through data

Blend of business, analytic and math skills to explore and solve challenges, bridging the data and business communities.

DATA SCIENCE & ANALYTICS


Ask good questions
Know the constraints (e.g., legal, ethics, market)
Latency at Execution?
Business Skills
Decision Making
Build, Buy, Outsource
Transparent Versus "black box"
Gauge political friction
Performance Criteria That Matter (ROI, accuracy, profitability versus market gain)
Deployment
"Analytics Leader"?
Feature Engineering
Data
Recalibration With New Data?
Data Logistics
IT Skills
Data Science
High-performance Computing
Which Analytics to Choose?
Project Execution/Monitoring
Data Exploration
Data Governance
Creativity, Communication

DE vs DA vs DJ



Data Journalist

- Experience as working journalist
- Data visualization
- Graphics creation
- Cartography tools
- Adobe Creative Suite
- CSS, D3, HTML5

- Domain knowledge
- Statistical analysis
- Data sourcing

Data Analyst

- Experience of data analysis / working as a data analyst
- Data mining
- Data management

- Collaborative team skills
- Python
- SQL
- R, Excel

- Quantified experience

- Data design & modelling
- SAS

- Experience in data engineering /processing/warehousing
- Experience with large IT /amounts of raw data
- Hadoop/MapReduce/Hive/Pig ETL

Data Engineer

DATA SCIENTIST

DATA SCIENTIST
"AS RARE AS UNICORNS"

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents
- ✓ *Distributed computing*
- ✓ *Predictive modeling*
- ✓ *Story-telling and visualizing*
- ✓ *Math, Stats, Machine Learning*

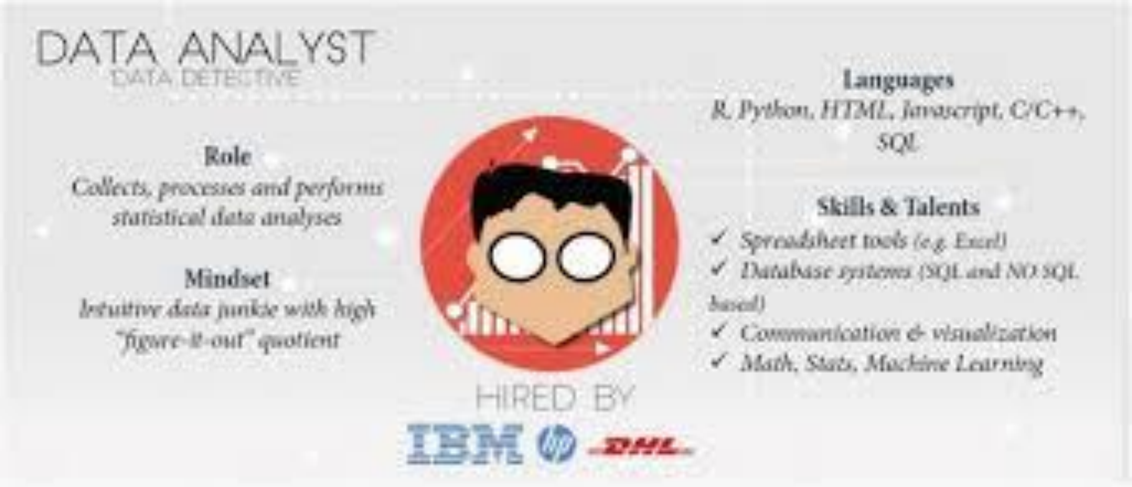
Role
Cleans, massages and organizes (big) data

Mindset
Curious data wizard

HIRED BY

Google
Microsoft
Adobe

DATA ANALYST



DATA ANALYST

DATA DETECTIVE

Role

Collects, processes and performs statistical data analyses

Mindset

Intuitive data junkie with high "figure-it-out" quotient

Languages

R, Python, HTML, Javascript, C/C++, SQL

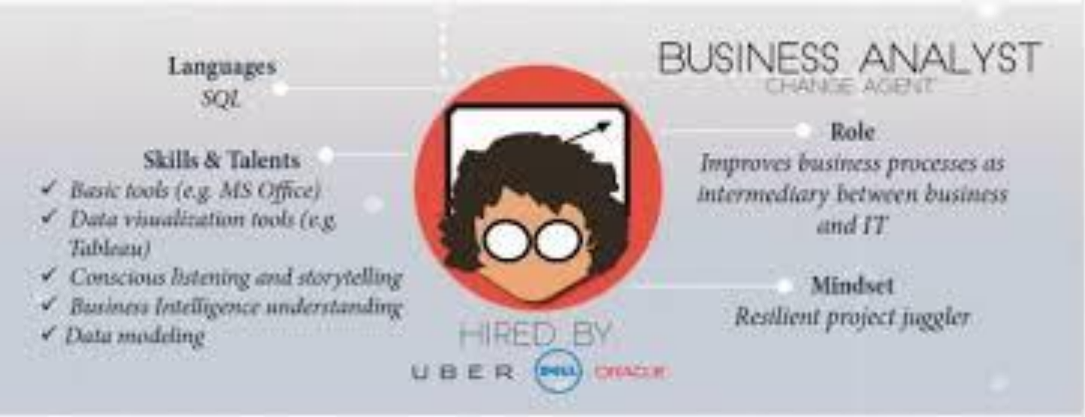
Skills & Talents

- ✓ *Spreadsheet tools (e.g. Excel)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Communication & visualization*
- ✓ *Math, Stats, Machine Learning*

HIRED BY

IBM hp DHL

BUSIESS ANALYST



BUSINESS ANALYST

CHANGE AGENT

Languages

SQL

Skills & Talents

- ✓ *Basic tools (e.g. MS Office)*
- ✓ *Data visualization tools (e.g. Tableau)*
- ✓ *Conscious listening and storytelling*
- ✓ *Business Intelligence understanding*
- ✓ *Data modeling*

Role

Improves business processes as intermediary between business and IT

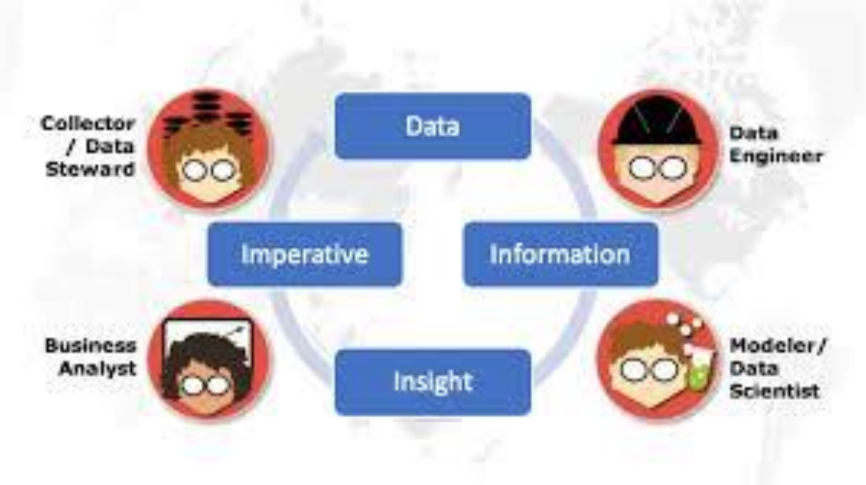
Mindset

Resilient project juggler

HIRED BY

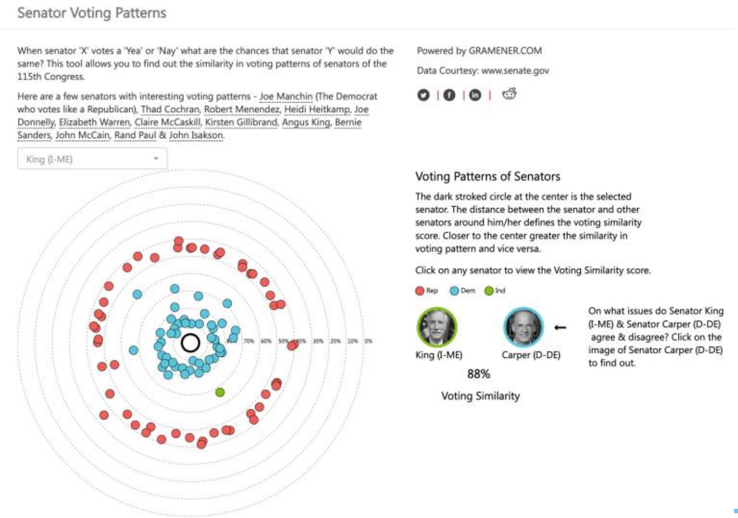
UBER DELL ORACLE

ROLES AND RESPONSIBILITIES



THREE EMERGING ROLES

1. DATA STORYTELLER



Senator Voting Patterns

When senator 'X' votes a 'Yea' or 'Nay' what are the chances that senator 'Y' would do the same? This tool allows you to find out the similarity in voting patterns of senators of the 115th Congress.

Here are a few senators with interesting voting patterns - Joe Manchin (The Democrat who votes like a Republican), Thad Cochran, Robert Menendez, Heidi Heitkamp, Joe Donnelly, Elizabeth Warren, Claire McCaskill, Kirsten Gillibrand, Angus King, Bernie Sanders, John McCain, Rand Paul & John Isakson.

King (I-ME)

Powered by GRAMENER.COM

Data Courtesy: www.senate.gov

Voting Patterns of Senators

The dark stroked circle at the center is the selected senator. The distance between the senator and other senators around him/her defines the voting similarity score. Closer to the center greater the similarity in voting pattern and vice versa.

Click on any senator to view the Voting Similarity score.

● Rep ● Dem ● Ind

King (I-ME) Carper (D-DE) ←

88%

Voting Similarity

On what issues do Senator King (I-ME) & Senator Carper (D-DE) agree & disagree? Click on the image of Senator Carper (D-DE) to find out.

Role Highlights

- **Dashboards are NOT data stories**
- **Stories=visual+context+narrative**
- **Fields: Journalism, creative arts**

https://gramener.com/playground/senate/similarity

2. BEHAVIORAL PSYCHOLOGIST

Role Highlights

- **Human side of data insights**
- **More practical, 'accurate' results**
- **Fields: Social sciences**

Gramener Telecom case study

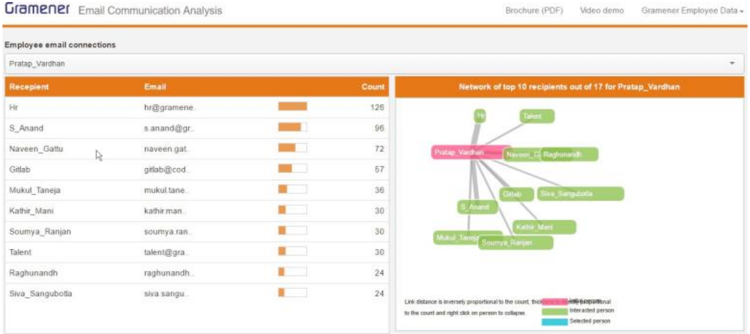
3. DATA ETHICIST

A top BFSI player wanted a scientific way to identify peers, for employee feedback.

Was there an alternative to manually screening for peer review?

This visual shows the network of email exchanges between people.

Look for the closest neighbors. The distance is a function of email exchange.



Gramener Email Communication Analysis

Brochure (PDF) Video demo Gramener Employee Data ▾

Employee email connections

Pratap_Vardhan

Recepient	Email	Count
Hr	hr@gramene...	126
S_Anand	s.anand@gr...	96
Naveen_Gattu	naveen.gat...	72
Gitlab	gitlab@cod...	57
Mukul_Taneja	mukul.tane...	36
Kathir_Mani	kathir.man...	30
Soumya_Ranjan	soumya.ran...	30
Talent	talent@gra...	30
Raghunandh	raghunandh...	24
Siva_Sangubotla	siva.sangu...	24

Network of top 10 recipients out of 17 for Pratap_Vardhan

Link distance is inversely proportional to the count, thic[illegible] proportional to the count and right click on person to collapse.

Role Highlights

- Ensure trust & fairness
- Act as a collective conscience
- Fields: Law, Humanities

https://gramener.com/emailnetwork/

3 EMERGING SKILLS IN DATA SCIENCE

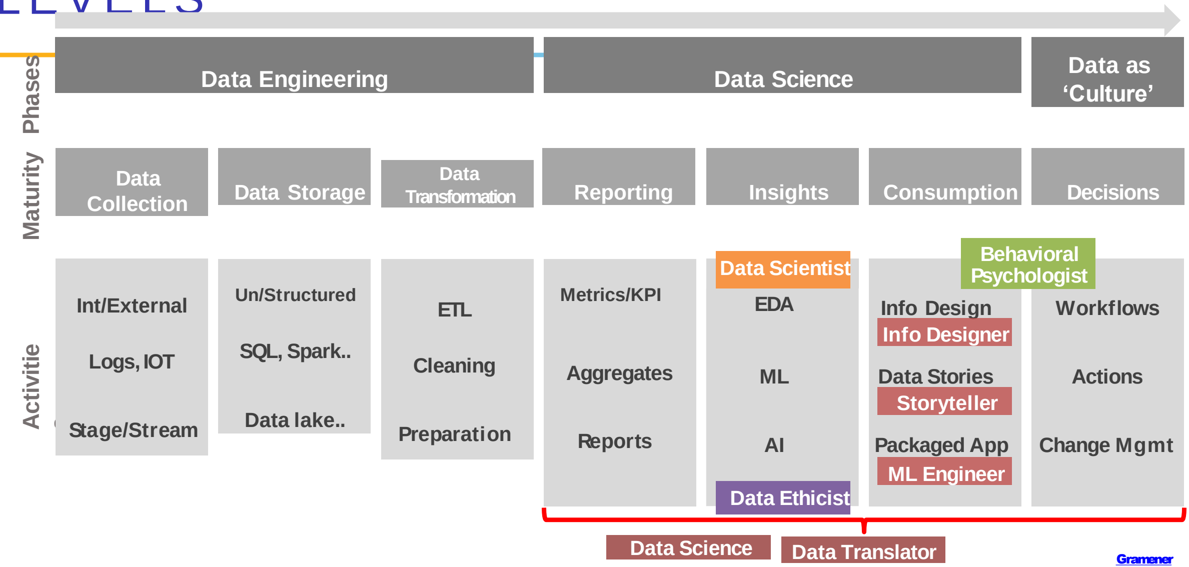
1. Data Storyteller

2. Behavioral Psychologist

3. Data Ethicist

https://towardsdatascience.com/the-3-missing-roles-that-every-data-science-team-needs-to-hire-97154cc6c365

DATA SCIENCE ROLES ACROSS MATURITY LEVELS



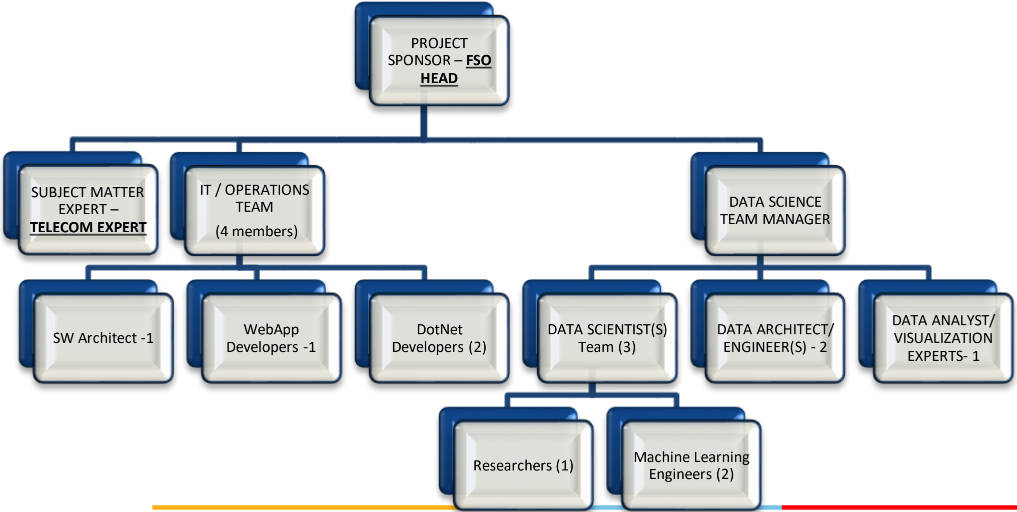
Phases	Data Engineering			Data Science			Data as 'Culture'
Maturity	Data Collection	Data Storage	Data Transformation	Reporting	Insights	Consumption	Decisions
Activitie	Int/External Logs, IOT Stage/Stream	Un/Structured SQL, Spark.. Data lake..	ETL Cleaning Preparation	Metrics/KPI Aggregates Reports	Data Scientist EDA ML AI Data Ethicist	Info Design Info Designer Data Stories Storyteller Packaged App ML Engineer	Behavioral Psychologist Workflows Actions Change Mgmt

Data Science

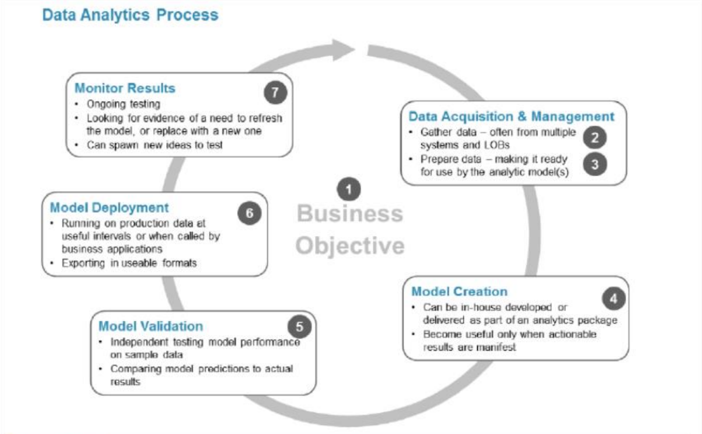
Data Translator

Gramener

DATA SCIENCE PROJECT HIERARCHY


PROJECT SPONSOR – **FSO HEAD**
SUBJECT MATTER EXPERT – **TELECOM EXPERT**
IT / OPERATIONS TEAM
(4 members)
DATA SCIENCE TEAM MANAGER
SW Architect -1
WebApp Developers -1
DotNet Developers (2)
DATA SCIENTIST(S) Team (3)
DATA ARCHITECT/ ENGINEER(S) - 2
DATA ANALYST/ VISUALIZATION EXPERTS- 1
Researchers (1)
Machine Learning Engineers (2)

DATA ANALYTICS PROCESS



Data Analytics Process

1 Business Objective

Data Acquisition & Management
- Gather data – often from multiple systems and LOBs 2
- Prepare data – making it ready for use by the analytic model(s) 3

Model Creation 4
- Can be in-house developed or delivered as part of an analytics package
- Become useful only when actionable results are manifest

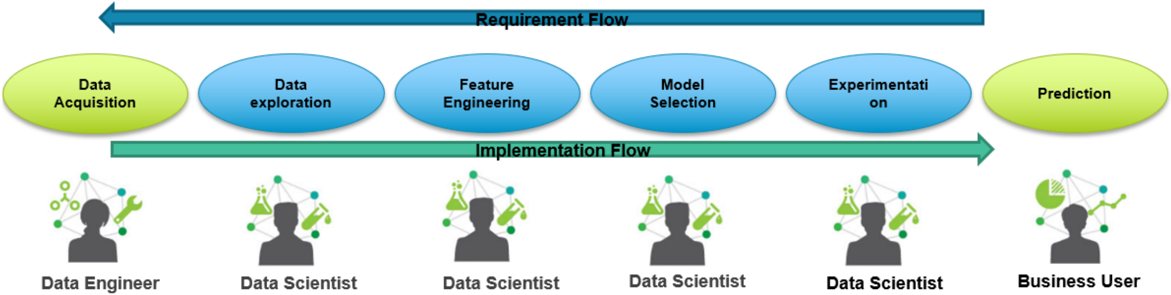
Model Validation 5
- Independent testing model performance on sample data
- Comparing model predictions to actual results

Model Deployment 6
- Running on production data at useful intervals or when called by business applications
- Exporting in useable formats

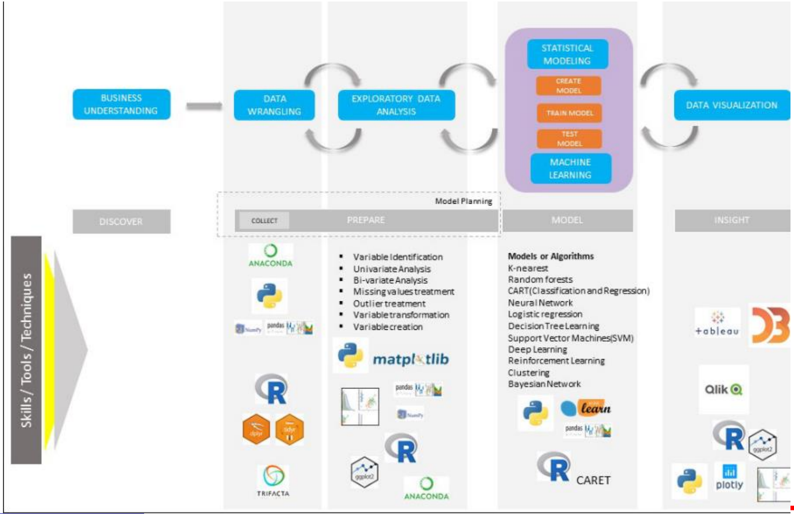
Monitor Results 7
- Ongoing testing
- Looking for evidence of a need to refresh the model, or replace with a new one
- Can spawn new ideas to test

Source: Celent

DATA ANALYTICS PROCESS – JOB ROLES



DATA ANALYTICS PROCESS – JOB ROLES & SKILLS



BUSINESS UNDERSTANDING

DATA WRANGLING

EXPLORATORY DATA ANALYSIS

STATISTICAL MODELING

CREATE MODEL

TRAIN MODEL

TEST MODEL

MACHINE LEARNING

DATA VISUALIZATION

Model Planning

DISCOVER

COLLECT

PREPARE

MODEL

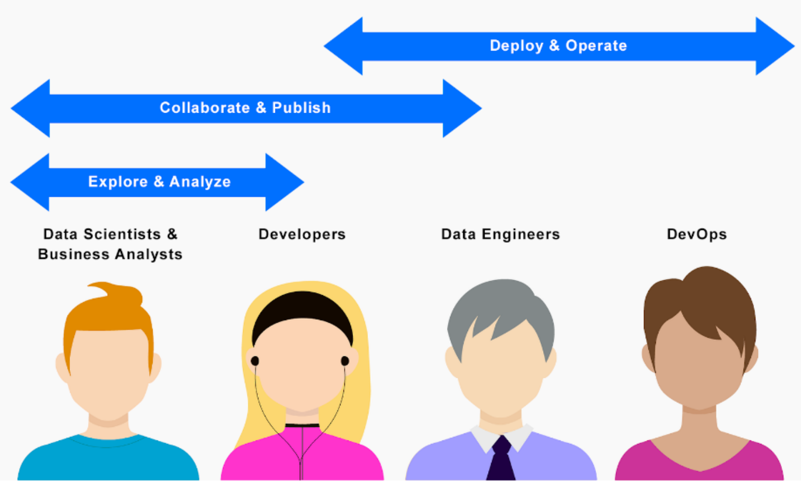
INSIGHT

Skills / Tools / Techniques

- Variable Identification
- Univariate Analysis
- Bi-variate Analysis
- Missing values treatment
- Outlier treatment
- Variable transformation
- Variable creation

Models or Algorithms
K-nearest
Random forests
CART(Classification and Regression)
Neural Network
Logistic regression
Decision Tree Learning
Support Vector Machines(SVM)
Deep Learning
Reinforcement Learning
Clustering
Bayesian Network

ROLES IN TIMELINE



Managing Data Science Team

- **Knowledge Management**
- **Attracting Top Talent**
- **Hiring Process**
- **Onboarding**
- **Retention and Management**

Focus –

- On boarding and evaluating the success of team
- Working with other teams

BITS Pilani, Deemed to be University under Section 3 of UGC Act, 1956

HOW TO CREATE AN ENVIRONMENT FOR TEAM SUCCESS?

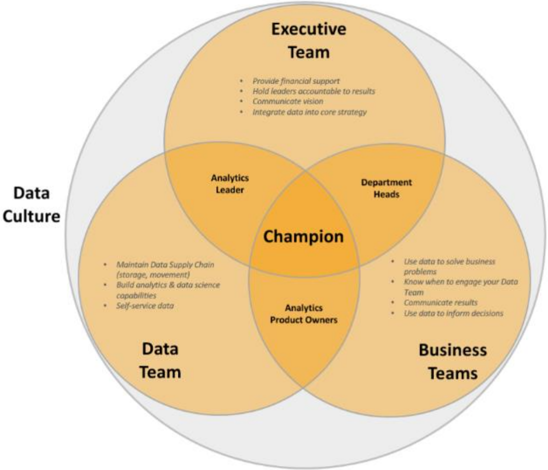
- Promote collaboration within teams
- Align closely with business users
- Measure outcomes through business value
- Adopt process frameworks for consistency
- Up-skill continuously on tech skillsets
- Nudge cross-training across disciplines

Common Challenges

1. Hiring a balanced data science team - build a cross-functional data science team that enables your organization to get insights from data and build production ready models.(Data Scientist, machine learning Engineer, Data Architect/Engineer, SW Developer, Research Scientist)
2. Retaining the team and Growing the team
3. Translating the business goals to smaller chunks of tasks, and defining measurable KPIs for the Data Science Team to work on achieving these KPIs.
4. Transforming Data Science team output/deliverables to a business understandable form, with key focus on Data Visualization. Hence try to bridge the gap between the Business Teams who relatively less/non-technical and the very technical Data Science Team
5. Engage and keep team motivated during the failures, and also keep the Senior leadership aligned with the fact that Data Science Projects are not like any SW Engineering project which can very Agile and give results every 7 days.

BITS Pilani, Deemed to be University under Section 3 of UGC Act, 1956

DATA CULTURE



Data Culture

Executive Team

- *Provide financial support*
- *Hold leaders accountable to results*
- *Communicate vision*
- *Integrate data into core strategy*

Analytics Leader

Department Heads

Champion

Data Team

- *Maintain Data Supply Chain (storage, movement)*
- *Build analytics & data science capabilities*
- *Self-service data*

Analytics Product Owners

Business Teams

- *Use data to solve business problems*
- *Know when to engage your Data Team*
- *Communicate results*
- *Use data to inform decisions*

Data driven decision making

- **Definition** – When it is data and not instincts that drives the business decisions.
- **Examples** – Fraud detection in Loans, Credit Cards (Cibil scores); Insurance, Six sigma projects to improve efficiency; Target advertising in e-commerce; Product Roadmap planning, Team planning
- 6 Steps to Data Driven decision making-
  1. Strategy – Define clear Business goals
  2. Identifying key data focus areas – Data is everywhere, flowing from multiple sources. Based on domain knowledge define key focus data sources which seem to impact the most, easier to access, reliable and clean
  3. Data Collection & Storage – Defining data architecture to collect, store, archive i.e. manage data. Connect multiple data sources, clean, prepare and organize
  4. Data Analytics – Analyzing the data and derive key insights
  5. Turning insights to Actions – business actions to be taken based on the findings from key insights from data
  6. Operationalize and Deploy – Using IT systems, automate the data collection, storage, analysis and presenting the key highlights

BITS Pilani, Deemed to be University under Section 3 of UGC Act, 1956

Thank You