Microsoft

MLADS

MACHINE LEARNING, AI,
AND DATA SCIENCE CONFERENCE

November 14 -18

1

What are we doing today? Lecture on;

- Introduction to Data Science
- Introduction to Statistics
- Introduction to Python
- Python and Machine Learning Packages
- Data Engineering and the Machine Learning Process
- Machine Learning

2

2

Labs

- Bing SQLSHEP GITHUB
- https://github.com/sqlshep/MLADS112022
- We will do the labs through out the day

3

3

Microsoft

Introduction to Data Science and Machine Learning

Shep Sheppard
shep.sheppard@microsoft.com
https://www.linkedin.com/in/sqlshep/
@sqlshep

4

Microsoft

Introduction to Data Science and Machine Learning

Shep Sheppard
Customer Engineer
Fast Track ISV

5

Session goals

- Gain an understanding of what Data Science and Machine Learning
- Try out some hands on labs with Jupyter Notebooks
- Gain and understanding of the data science process
- Execute a few models

6

6

Agenda

- What is Data Science?
- History in one slide
- How did we get here?
- What Skills are required?
- The Data Science Process
- Data Science Gone Wild
- How do I Data Science?

7

7

History in one slide

8

8

History in one slide

- The term Data Science was anecdotally coined around 2008.
- Statistical foundations date to 1700.
- Data-Driven Science is an interdisciplinary field about scientific methods.
- Data Science derived from a 30 year old term Dataology.
- Data Science has become a popular moniker in the last ten years after HBR Article "The Sexiest Job of the 21st Century".

9

9

What is Data Science

10

10

What is Data Science

Data Science

Interdisciplinary field about scientific methods to extract knowledge from data

Data science is a "concept to unify statistics, data analysis and their related methods" in order to "understand and analyze actual phenomena" with data.

Statistical Learning

To Predict or Infer Y based on X

Artificial intelligence

Colloquially, is applied when a machine mimics "cognitive" functions that humans associate with other human minds

Machine Learning

Learn without being explicitly programed.

Closely related and often overlaps with computational statistics

Deep Learning

At the heart of AI, combines supervised and unsupervised learning, multiple layers of cascading algorithms.

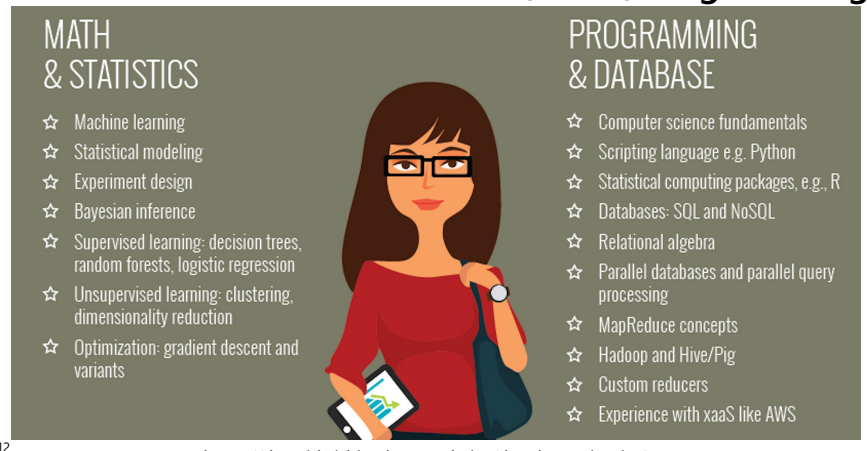
Neural Networks

Inspired by biological neural networks, collection of small computing units

11

11

The Modern Data Scientist Math, Stats, Programming



12

https://thegrid.ai/big-data-analytics/the-data-scientist/

12

The Modern Data Scientist, Domain and Communication

13

https://thegrid.ai/big-data-analytics/the-data-scientist/

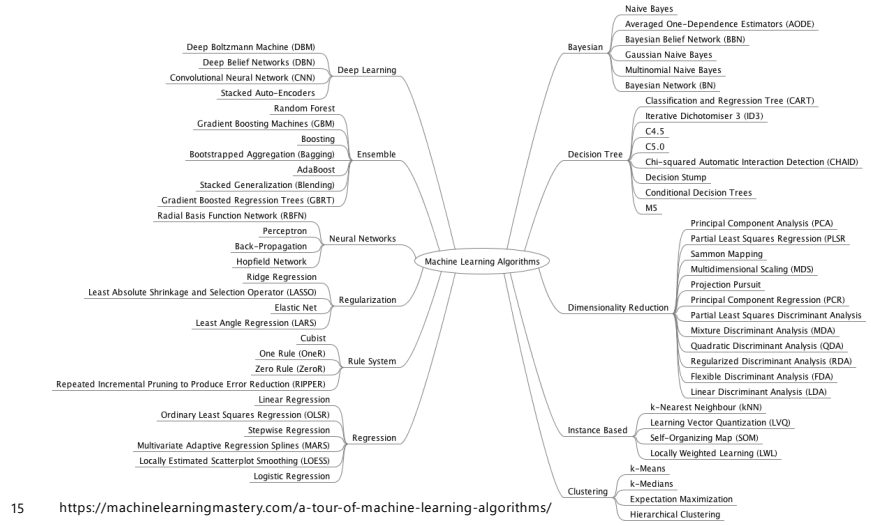
13

Gentle Intro to a few common algorithms

- In the next few slides we will discuss a few of the most common algorithms you may see and hear about
- Mind map of most of them
- Supervised, Unsupervised learning and PCA
- Trees and Forests
- Neural Networks

14

14



15 https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

15

Supervised and Unsupervised Learning

- Machine learning is broadly categorized into two main categories
- Supervised learning
  - Is a learning function that maps and input to an output based on sample input
  - For instance taking a sample of actual vehicle weight, horse power, engine size and mpg and using this data to estimate an unknown mpg based on weight, horse power, engine size
- Unsupervised learning
  - Is a model that learns from data that has not been labeled.
  - For instance, passing census or shopping data into a clustering algorithm and having it determine similarities in the data with no guidance from an operator.

16

16

A Few Common Supervised Models

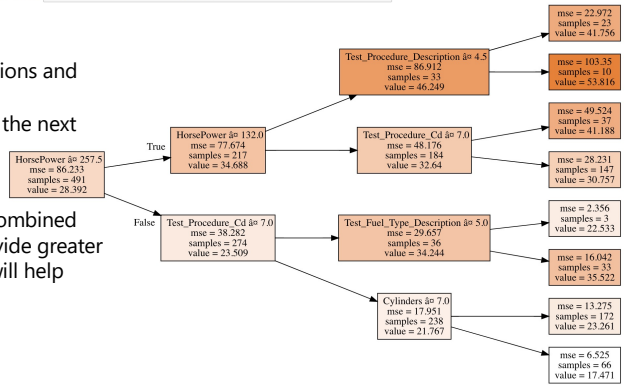
- Linear Regression
  - To infer Y based on X, or many X's
  - For instance, you know the weight(X_1), engine size(X_2), and horsepower(X_3) of a car, can you infer the MPG (Y)?
- Logistic Regression or Logit
  - Used to determine a binary outcome, yes or no, or the probability of yes or now
  - Will it rain today? What is the likelihood of rain? Can be used to determine odds as well.

17

17

Trees and Forests

- Decision Tree
  - A tree like model of decisions and consequences.
  - Each branch takes you to the next decision
- Decision Forest
  - Multiple Decision Trees combined together that should provide greater predictive accuracy and will help account for outliers.



18

18

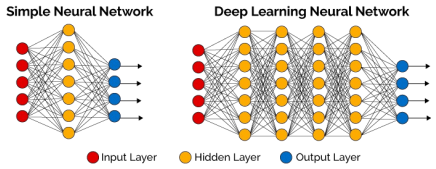
Unsupervised models

- K-Nearest Neighbor (KNN)
  - Is as the name suggest find items that are near and alike
- K-Means (Clustering)
  - Based on a specified number of clusters divide the data into like clusters based on distance
- Principal Component Analysis (PCA)
  - This is typically used for Dimensionality Reduction or
  - When you have hundreds or thousands of variables in you dataset, PCA can help eliminate the nonessential ones
  - It will identify linearly correlated and uncorrelated features

19

19

Neural Networks (Deep Learning)



- Neural Network
  - The neural network itself is not an algorithm, but rather a framework for many different machine learning algorithms to work together and process complex data input
  - Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules.

20

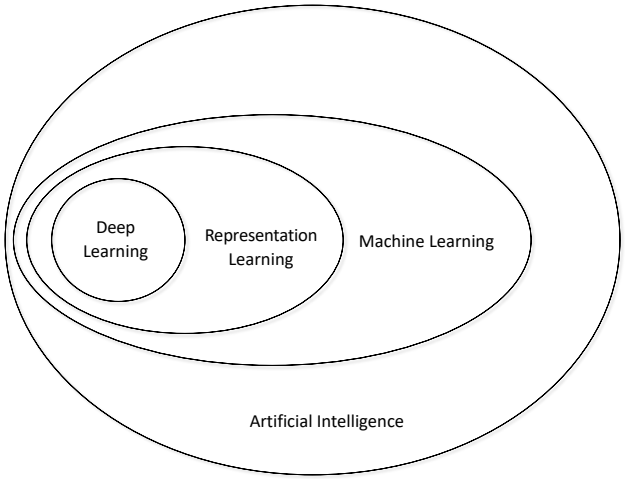
20

Reinforcement Learning

- Reinforcement Learning relies on providing a machine learning model with rules and constraints
- Thus allowing the model to learn how to achieve its goals
- Define the state of the desired goal, allowed actions and constraints
- Such as, do not touch a flame, it is hot and will hurt you

21

21



22

How did we get here

23

23

24

Archive and Purge used to be a thing

- Storage has become cheaper and faster
  - Cloud Storage is a race to the bottom to see who can eventually offer it for free
- Databases have become faster (For Reals!).
- Column and Table compression has become mainstream
  - Columnar compression is capable of 98%-99% compression in specific cases.
  - Every database technology offers some form of compression
- Expectation of value in the data, though little knowledge of how to derive it.
- Massively parallel scale out became real, 1000+ nodes as a PaaS service

25

25

Data Spewing Devices

- Light Bulbs
- Belts
- Wine Bottle
- BBQ Grill
- Trash can
- Tortilla Maker
- Water Bottle
- Bluetooth Smart Fork
- YOU

26

26

I present to you the Smalt

Smart Home meets Smart Dining

THE WORLD'S FIRST INTERACTIVE CENTERPIECE AND SMART SALT DISPENSER
DESIGNED TO ENHANCE YOUR DINING EXPERIENCE

27

27

Data Science in the Real World

28

28

What's it used for

- Hand writing recognition
- Predicative maintenance
- Machine Translation, real time
- Natural Language Processing (NLP)
- Predictive text
- Financial Fraud Detection
- Search, Google, Bing
- Money ball Scenarios
- Athletic Performance

Optimization

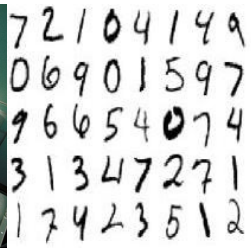
- Fitness Trackers
- Suicide Prevention
- Police Adverse Interactions
- Medical Diagnosis – IBM Watson
- Autonomous Cars
- Tracking YOU
- Recurrent Neural Network

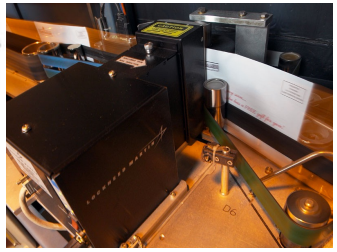
29

29

Hand Writing Recognition

- Post office has used this for decades to read city/state/zip
- First MPLSM OCR installed November 1965





30

30

Predictive Maintenance

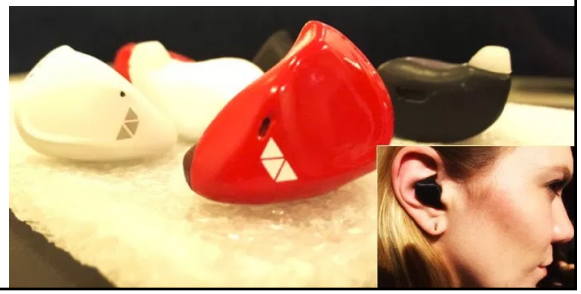
- Future impact based on past behavior
- The old days, ATA codes on an aircraft that are known to cause delays
- Sensors monitoring components near failure
- A380 had 10,000 sensors per wing…
- Can generate up to 2.5Tb of data per day

31

31

Machine Translation, Real Time

- Microsoft Skype Translator, 10 languages Voice, 60 for text
- Waverly Labs Pilot, 20 languages, connected to your phone
- Modern day Babel fish or Rosetta stone



32

32

Natural Language Processing

- NLP
- Anti Spam
  - Reads emails and looks for spam like patterns
- Create Ads and New Spam
  - Read your emails then target you with offers
- Summarize lots of stuff
  - Twitter
  - Knowledge Extraction
- Q&A –
  - Imagine with no users responding just the corpus of past interactions being fed up by algorithm.
  - The Chat Bot

33

33

Predictive Text

- Predictive text (Hidden Markov, T9)
- N-Grams (NLP)
- Good-Turing Smoothing (N-Gram)
- Autocorrect and failed Autocorrect

34

34

Financial Fraud Detection

- Credit Card used where it was not supposed to be
- Money moved in a nefarious way
- Improper purchases by employee
- Forensic Accounting
- Earnings Manipulation
- Some use a several machine learning techniques depending on the complexity

35

35

Search

- Bing,
- Google
- RankBrain
- Complex combination of Machine learning, and AI used for scoring, ranking web pages, and interpreting queries and determining intent.
- And most importantly, serving ads up to you based on search history

Google bing YAHOO! Aol. Ask DuckDuckGo

36

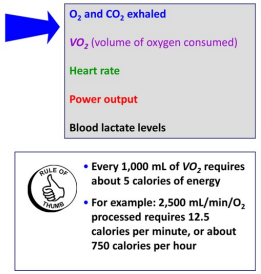
DS in the Real World – Money Ball

- Boston Red Sox (Money Ball Scenario)
- World Series 1903, 1912, 1915, 1916, 1918… then Nothing
- Oakland Athletics started winning using data– Made famous by the Michael Lewis book and movie Moneyball about data driven decision making.
- 2003 John Henry bought Red Sox, Henry anecdotally know as a data driven guy
- Using Data Driven Methodology, won World Series 2004, 2007, 2013
- They then abandon the methodology to go with old school recruiting

37

37

Athletic Performance Optimization



- There is a "*large*" market for professional sports data science
- Teams spend a lot of time and money finding the next Tom Brady
- Imagine the athlete as an IoT device
- Then attempt to govern the device and predict the next one…

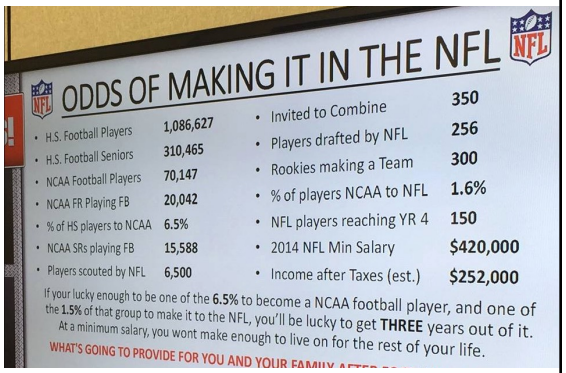
38

38

Athletic Performance Prediction, NFL

Why it matters?

89.8% Retirement rate by year 4!



39

39

Fitness Trackers

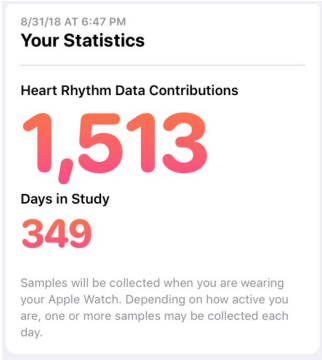
- Constant battle to try and track and predict the next big thing.
- HIPAA issues with tracking and uploading some data.
- They are unable to solve some issues with machine learning, some are waiting on the magical AI to solve their problems.
- Microsoft Band failed at it, discontinued fall 2016.
- Cannot be a diagnostic tool due to FDA regulations, not even a heart rate monitor. Though Apple is working with the FDA on this.
- Anecdotal evidence of users claiming it saved their lives.

40

40

Stanford Medicine Apple Heart Study

- The Apple Heart Study app uses data from Apple Watch to identify irregular heart rhythms, including those from potentially serious heart conditions such as atrial fibrillation.
- Apple is conducting this research study in collaboration with Stanford Medicine to improve the technology used to detect and analyze irregular heart rhythms, like atrial fibrillation - a leading cause of stroke.



41

41

Suicide Prevention - DSFSG

- Florida State University - FSU Psychology researcher Dr. Jessica Ribeiro
- “Studies show about 60-90 percent of people who die by suicide had visited their medical provider within the past year and the clinician never saw it coming. ”
- ”machine learning — a future frontier for artificial intelligence — can predict with 80-90 percent accuracy whether someone will attempt suicide as far off as two years into the future. ”
- “giving clinicians the ability to predict who will attempt suicide up to two years in advance with 80 percent accuracy.”

42

42

Predicting Police Adverse Interaction - DSFSG

- Data Science for Social Good Project at U Chicago
  - Check out their website https://dssg.uchicago.edu/ &
  - White House Police Data Initiative
- Tested with *Charlotte-Mecklenburg Police District and* Metro Nashville Police Department
- *Correctly* Predict when a police officer is at increased risk of an adverse interaction
- The model used a full cadre of data, demographics, join date, arrests, dispatches, training, IA activity, weather, quality of life surveys.
- Model correctly flagged 10—20% more officers that were later involved in an adverse interaction over EIS. (eventually 80% total prediction using Random Forest)
- Reduced the Type 1 errors(false positive), by 50%

43

43

Autonomous Vehicles

- Otto by Uber (Shuttered)
- Tesla, Google
- Fully Autonomous Race car
- 18 Wheelers Soon
- Common tool is Simultaneous localization and mapping (SLAM)
- SLAM is a complete system of Algorithms - AI

44

44

Customer Tracking

- Loyalty Apps for Malls, Shopping Centers.
- Connect to WiFI, tracks your position in a shopping center.
- The dirty little secret, they are tracking you if you do not have the loyalty app installed.
- Wifi routers capture mac address, two or more routers can triangulate your location.
- Used to track where you are to provide recommendations to future customers.
- Visit length, location every second, frequency of visits, if you are in a pack, which stores you visit.

45

45

Merchandise Tracking

- What to do with a smart Button?
- Smart Washing Machines
- Have a store inventory itself
- Walk out without stopping at a register
- Track you when you come back in?

46

46

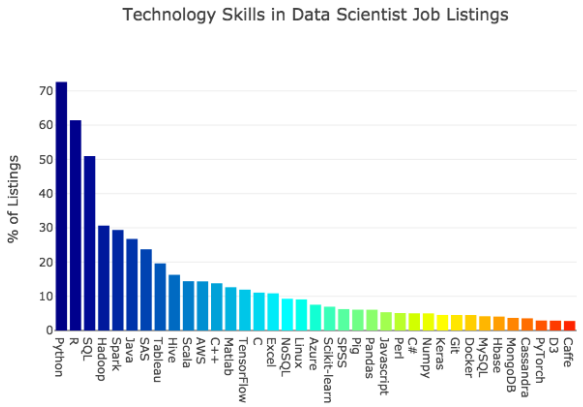
What skills are required

47

47

Popular Data Science Skills, KDNuggets

- The chart shows an even bigger list of the most in demand languages, frameworks, and other data science software tools.

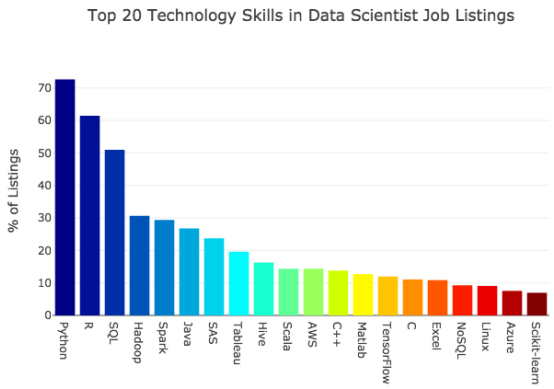


48

48

Popular Languages, libraries, and tools, KDNuggets

- The top 20 specific languages, libraries, and tech tools employers are looking for data scientists to have experience with.



49

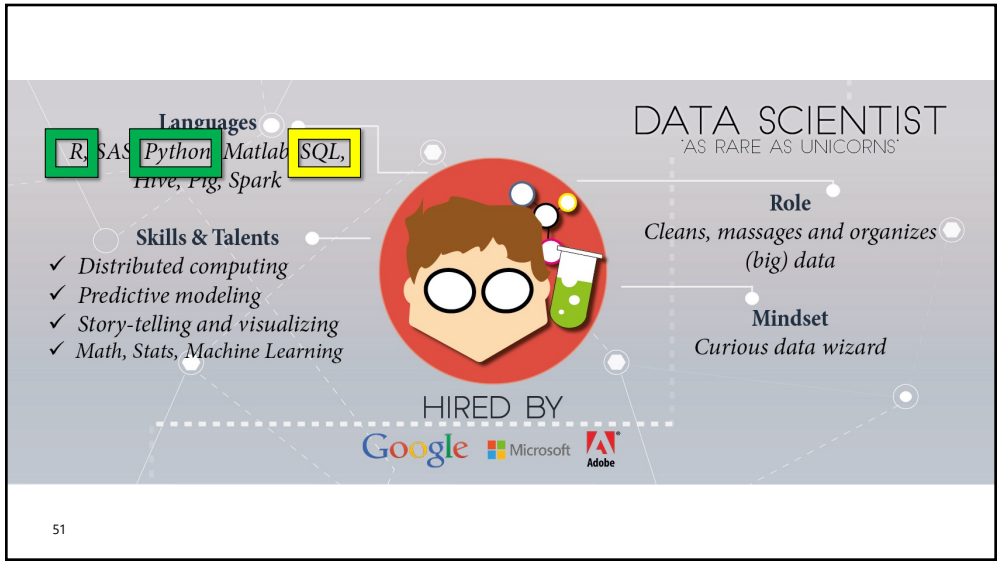
49

What do the data skills have in common?

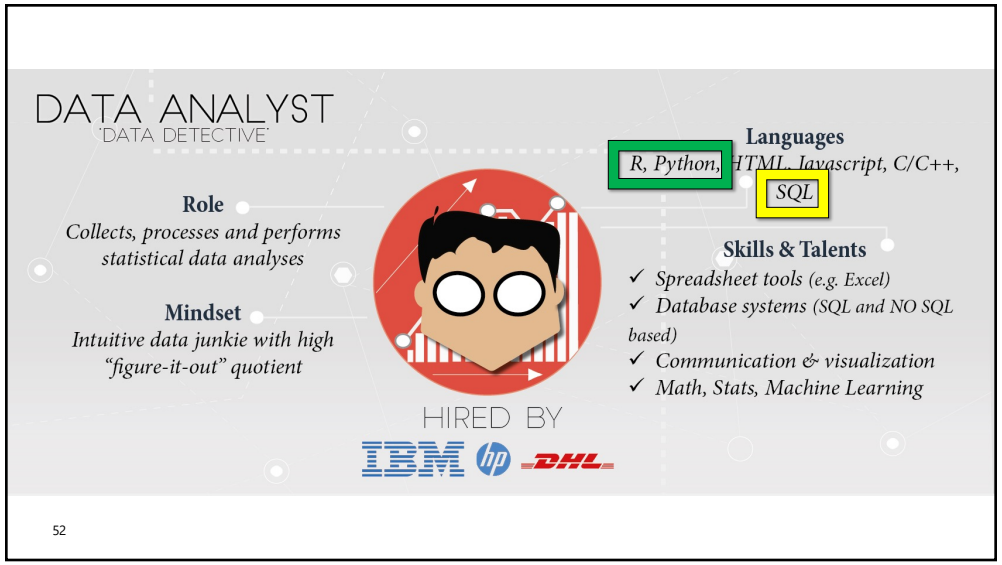
- Data Scientist
- Data Analyst
- Data Architect
- Data Engineer
- Statistician
- Database Administrator
- Business Analyst
- Data and Analytics Manager

50

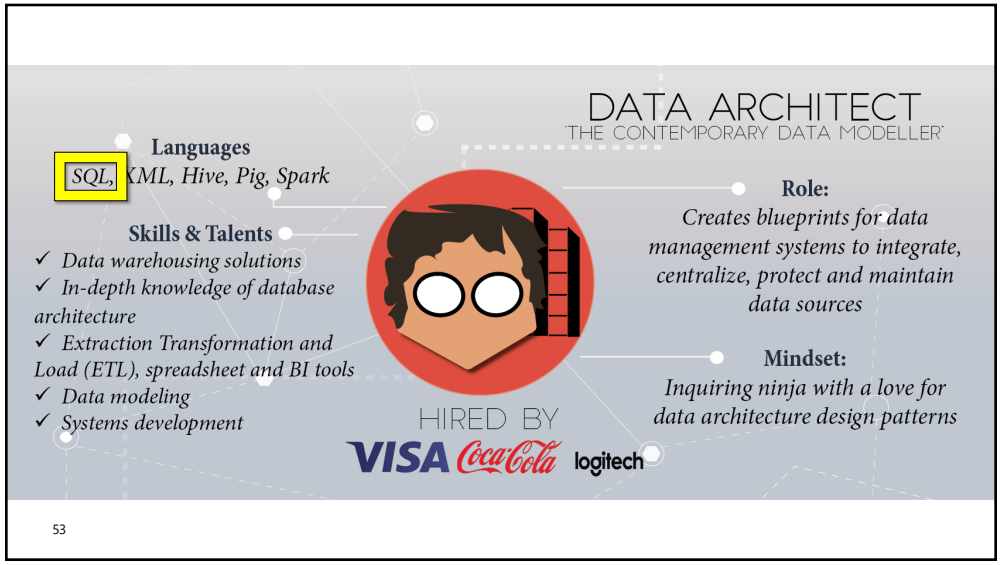
50


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
51

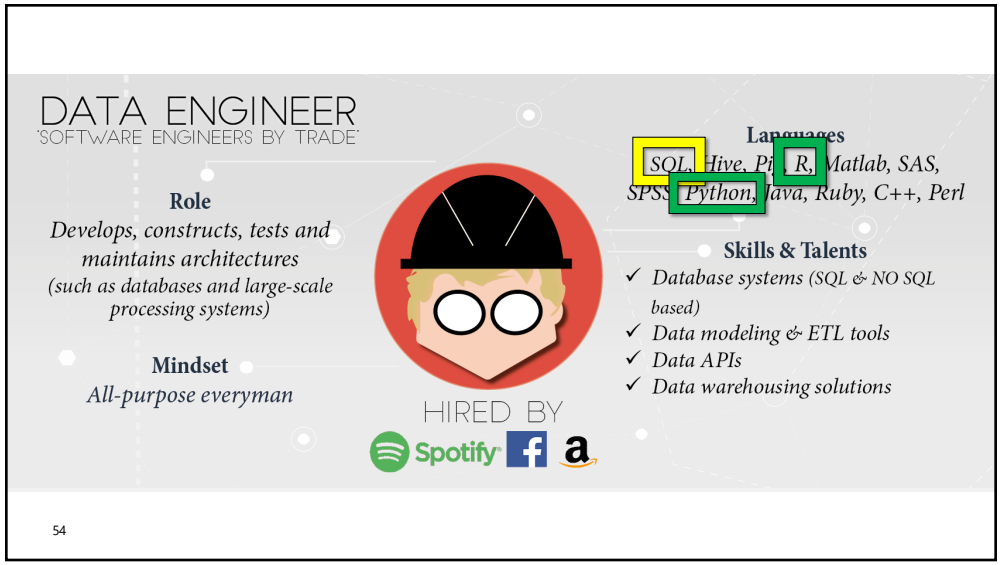
51


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
52

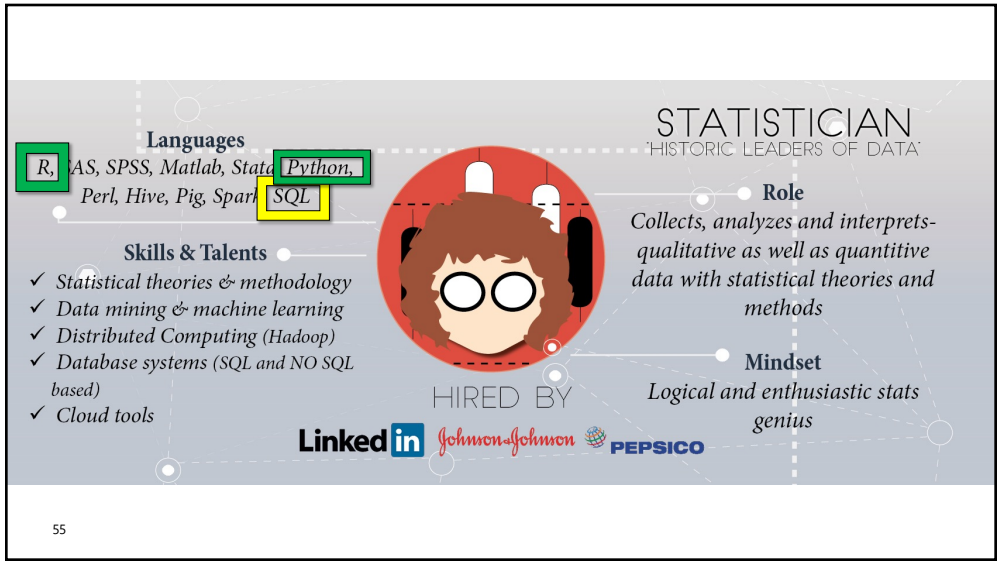
52


DATA ARCHITECT
'THE CONTEMPORARY DATA MODELLER'
Languages
SQL, XML, Hive, Pig, Spark
Skills & Talents
✓ Data warehousing solutions
✓ In-depth knowledge of database architecture
✓ Extraction Transformation and Load (ETL), spreadsheet and BI tools
✓ Data modeling
✓ Systems development
Role:
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources
Mindset:
Inquiring ninja with a love for data architecture design patterns
HIRED BY
VISA
Coca-Cola
logitech
53

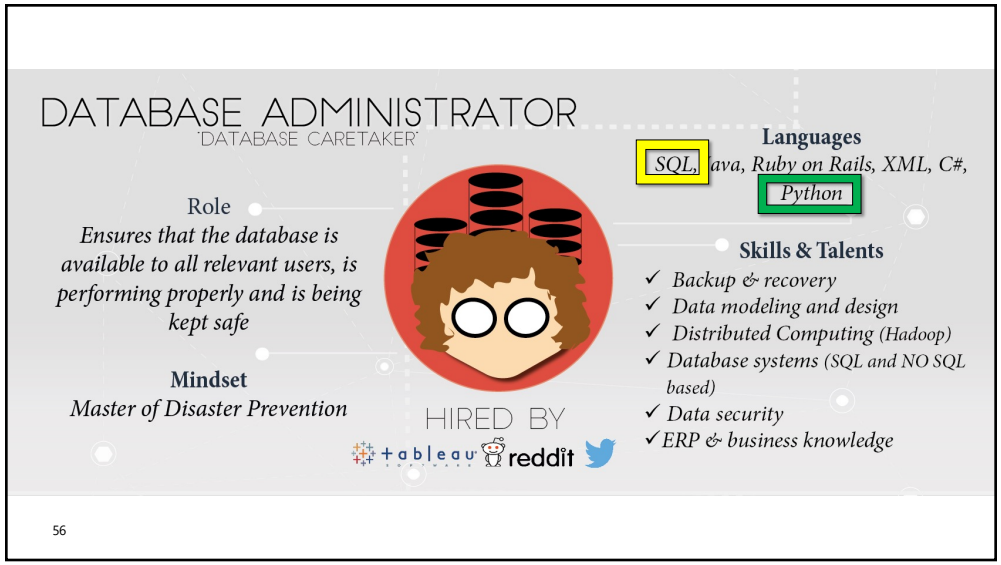
53


DATA ENGINEER
'SOFTWARE ENGINEERS BY TRADE'
Role
Develops, constructs, tests and maintains architectures
(such as databases and large-scale processing systems)
Mindset
All-purpose everyman
Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl
Skills & Talents
✓ Database systems (SQL & NO SQL based)
✓ Data modeling & ETL tools
✓ Data APIs
✓ Data warehousing solutions
HIRED BY
Spotify
54

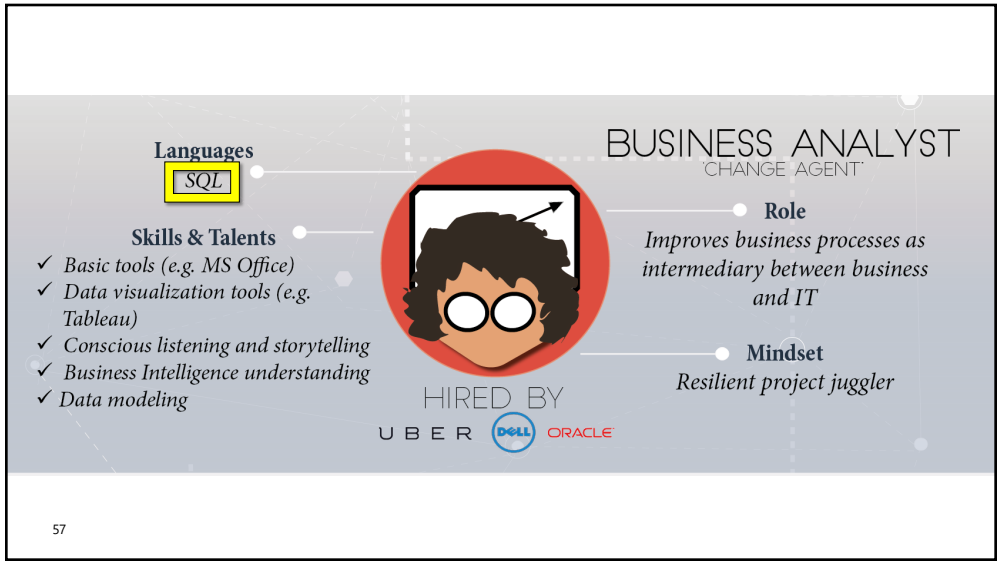
54


STATISTICIAN
'HISTORIC LEADERS OF DATA'
Languages
R, SAS, SPSS, Matlab, Stata, Python, Perl, Hive, Pig, Spark, SQL
Skills & Talents
✓ Statistical theories & methodology
✓ Data mining & machine learning
✓ Distributed Computing (Hadoop)
✓ Database systems (SQL and NO SQL based)
✓ Cloud tools
Role
Collects, analyzes and interprets- qualitative as well as quantitive data with statistical theories and methods
Mindset
Logical and enthusiastic stats genius
HIRED BY
LinkedIn Johnson&Johnson PEPSICO
55

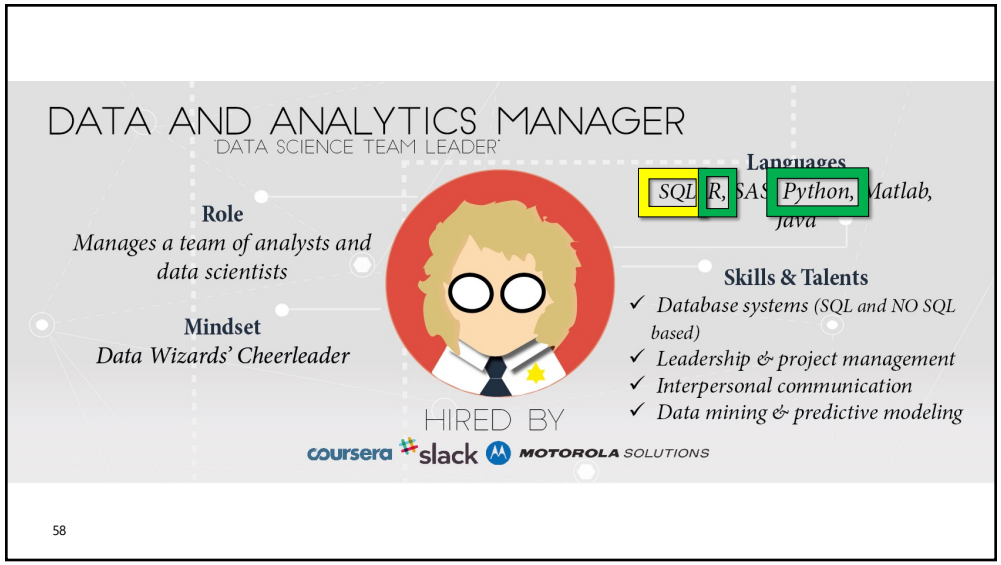
55


DATABASE ADMINISTRATOR
'DATABASE CARETAKER'
Role
Ensures that the database is available to all relevant users, is performing properly and is being kept safe
Mindset
Master of Disaster Prevention
Languages
SQL, Java, Ruby on Rails, XML, C#, Python
Skills & Talents
✓ Backup & recovery
✓ Data modeling and design
✓ Distributed Computing (Hadoop)
✓ Database systems (SQL and NO SQL based)
✓ Data security
✓ ERP & business knowledge
HIRED BY
tableau reddit
56

56


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
57

57


DATA AND ANALYTICS MANAGER
DATA SCIENCE TEAM LEADER
Role
Manages a team of analysts and data scientists
Mindset
Data Wizards' Cheerleader
Languages
SQL
R, SAS
Python, Matlab, Java
Skills & Talents
✓ Database systems (SQL and NO SQL based)
✓ Leadership & project management
✓ Interpersonal communication
✓ Data mining & predictive modeling
HIRED BY
coursera
slack
MOTOROLA SOLUTIONS
58

58

The data science process

60

60

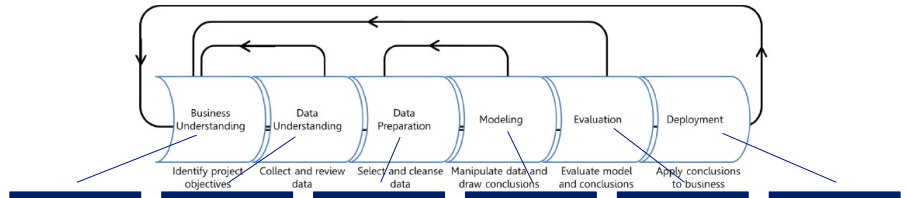
The Data Science Process

- Exploratory Data Analysis (EDA)
- Data cleansing and preparation
- Model feature engineering
- Model training and evaluation
- Model deployment
- The specialized roles in the data science process

61

61

The Data Science Process



Problem: The company loses sales because inventory is exhausted.

Can we stock shelves ahead of time with the correct inventory, i.e. can we predict demand?

Data Scientists explore data available to find patterns and relationships to the business problem, i.e. what drives demand?

Data Scientists cleanse and transform the data as needed to support modeling. Some columns may need to be aggregated and others reformatted.

Features, the attributes that can be used to predict the outcome, are identified and extracted, potential machine learning models are selected and trained.

The trained models are evaluated and compared to find the one that performs the best, i.e. the one that most reliably predicts sales.

The best performing model is deployed to an environment where it can be used by applications to make predictions.

62

62

Data Sceince Gone WIld

63

63

Artificial Intelligence

- Colloquially, the term "artificial intelligence" is applied when a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving"

64

64

The Turing Test

- Via text only test a machines ability to mimic human interaction.
- If the human is unable to distinguish between computer and human, the computer is said to have past the test.

65

65

Will Data Science and AI destroy the world?

66

66

Probably not

67

67

In Fantasy?

- ""In short, success in creating AI could be the biggest event in the history of our civilization," ***Prof Hawking said.***
- "But it could also be the last unless we learn how to avoid the risks.
- "Alongside the benefits, AI will also bring dangers – like powerful autonomous weapons, or new ways for the few to oppress the many." "

68

68

Has it happened yet?

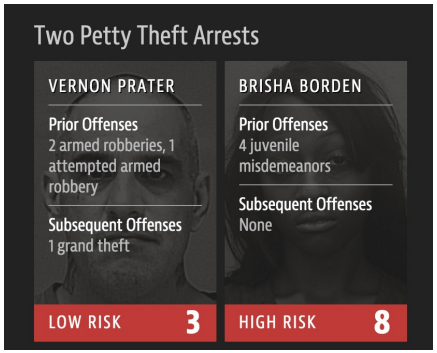
4Chan trolled Microsoft's Twitter AI chat bot Tay and taught it to be bad AI in less than a day

69

69

Bias in Recidivism

- ProPublica did a multi year investigation on recidivism and the machine learning tools used to predict it.
- Their determination was that there exists significant bias in the models, though the creator of the software denies it.
- **The point: The model represents what you put into it.**



70

70

Michigan Unemployment System

- Michigan Integrated Data Automated System (Midas)
- Michigan unemployment agency made over 20,000 false fraud accusations, 93% rejection rate. (Type 2 Error)
- Automated system erroneously accused claimants in 93% of cases, state review finds: 'It's balancing the books on the backs of the poorest,' lawyer says
- Bankruptcy petitions filed as a result of unemployment insurance fraud also increased during the timeframe when Midas was in use.

71

71

Meet Sophia, Saudi Arabia has become the first country to give a robot citizenship.

- "I want to use my artificial intelligence to help humans live a better life, like design smarter homes, build better cities of the future."

72

72

73

73


Dmitry Rogozin
@DRogozin
Follow
Shooting exercises is a method of teaching the robot to set priorities and make instant decisions. We are creating AI, not Terminator

74

75

Ethics

"not worried about artificially intelligent death bots"

"Algorithms help courts set bail, determine which news stories appear on Facebook users' feeds, and sometimes decide who will be given a line of credit from a bank."

"One popular misconception is that if it's an algorithm, then it's unbiased—it has some kind of inherent objectivity,"

"Algorithms are not neutral. They are maximizing parameters that were chosen by the people that designed [them]."

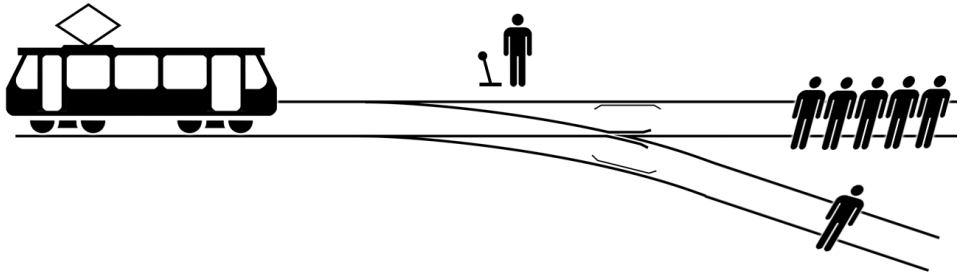
Urs Gasser,
Berkman Klein Center for Internet and Society
Harvard Law School

76

76

The Trolley Problem

· Who to kill? Self driving cars are the new trolley problem.



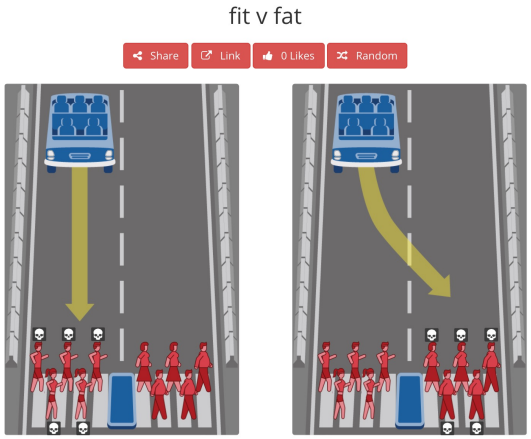
. .

77

The Moral machine

- Human perspectives on moral decisions made by machine intelligence

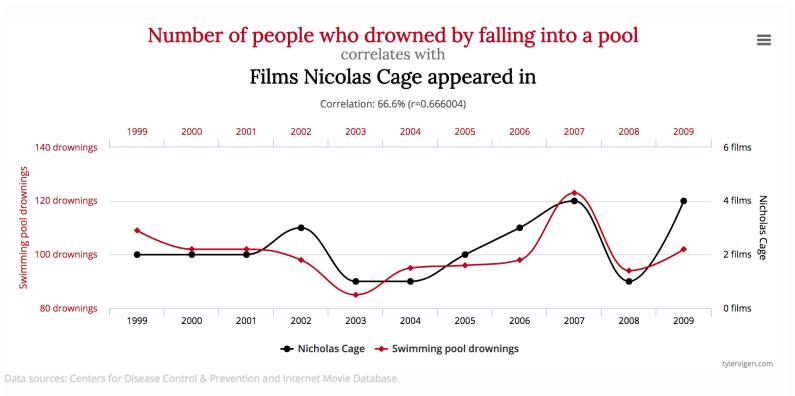
http://moralmachine.mit.edu



78

78

Final Thoughts



79

79

Thank you for attending the MLADS Conference and helping to build a strong community

To find recordings, presentations, and other resources from the event, go to: http://aka.ms/november2022mlads

More information about the AI & ML Connected Community: http://aka.ms/aiml-cc

We want your feedback!

We read each evaluation and incorporate your comments into future MLADS events. Both sessions and event evaluations will be available within the MyMLADS event portal during the conference dates—to access them, navigate to the **Evaluations tab**.

We request that session evaluations are submitted following each day of programming. Evaluations for both weeks of the conference will be open until **11:59 PM PT on Friday, June 17th**

80

80

Q&A

This slide is required—do not delete—please read the notes for this slide, then delete this text box

81

81

82