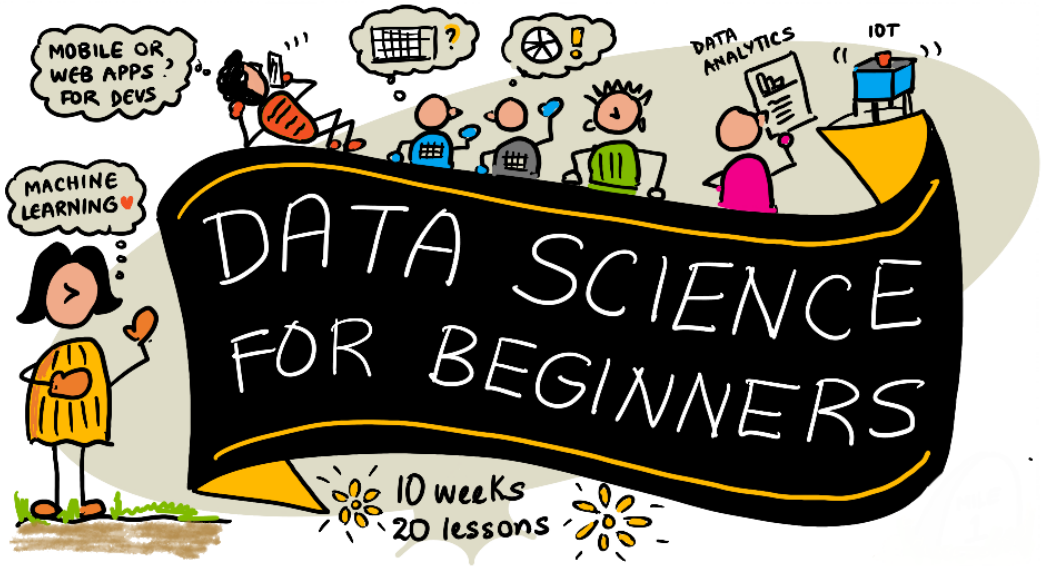
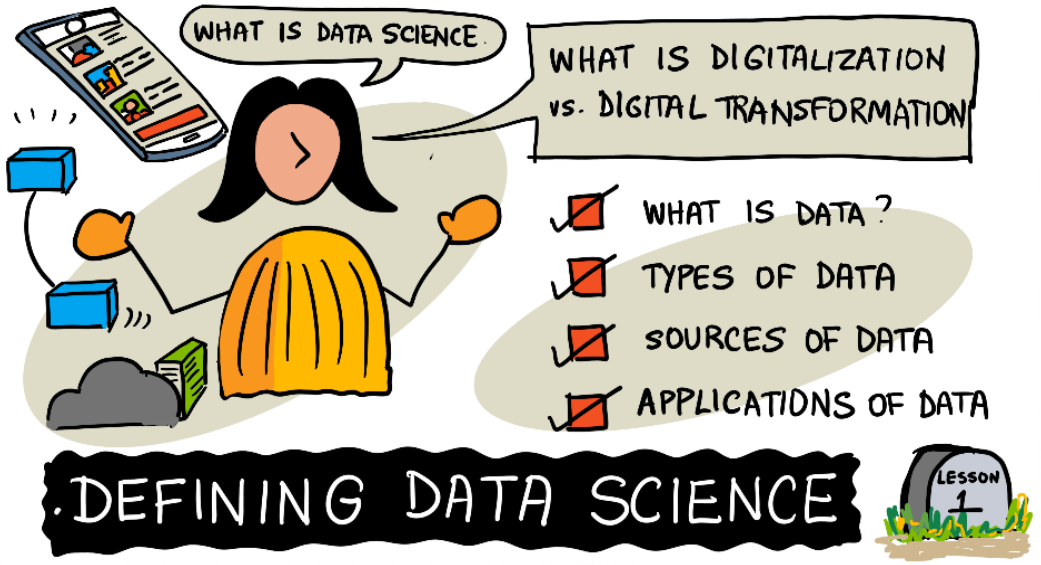
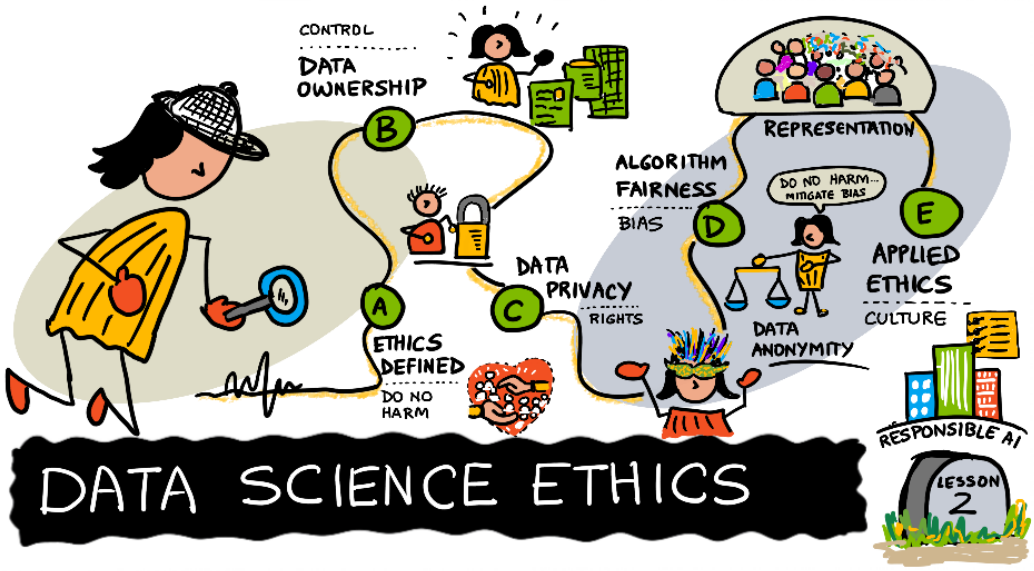
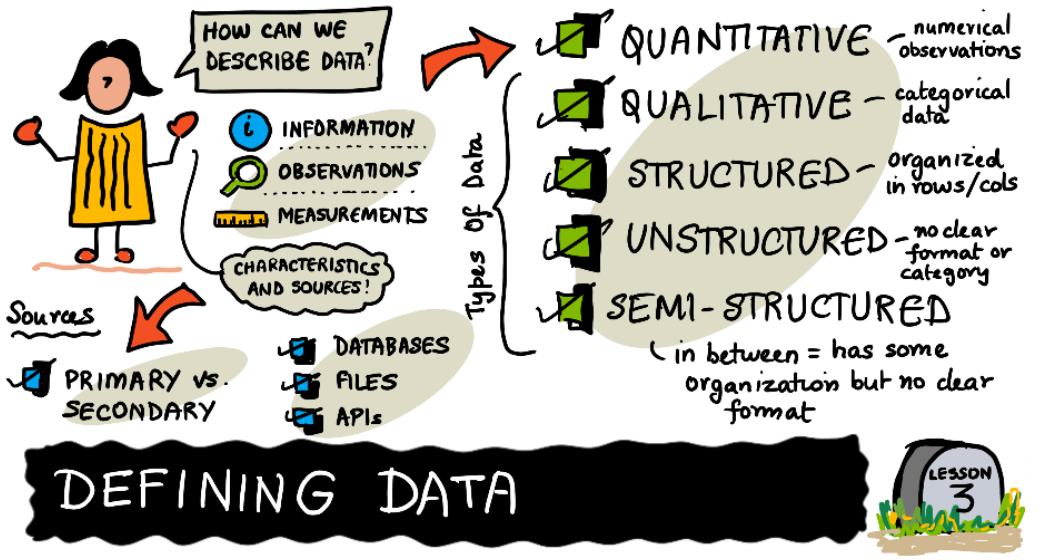
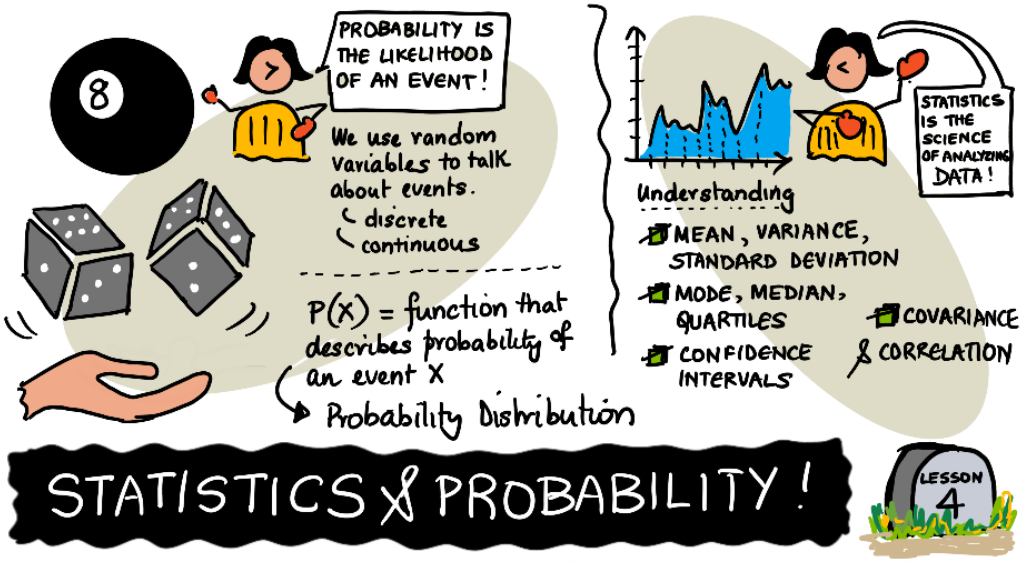
Data Science For Beginners Sketch notes

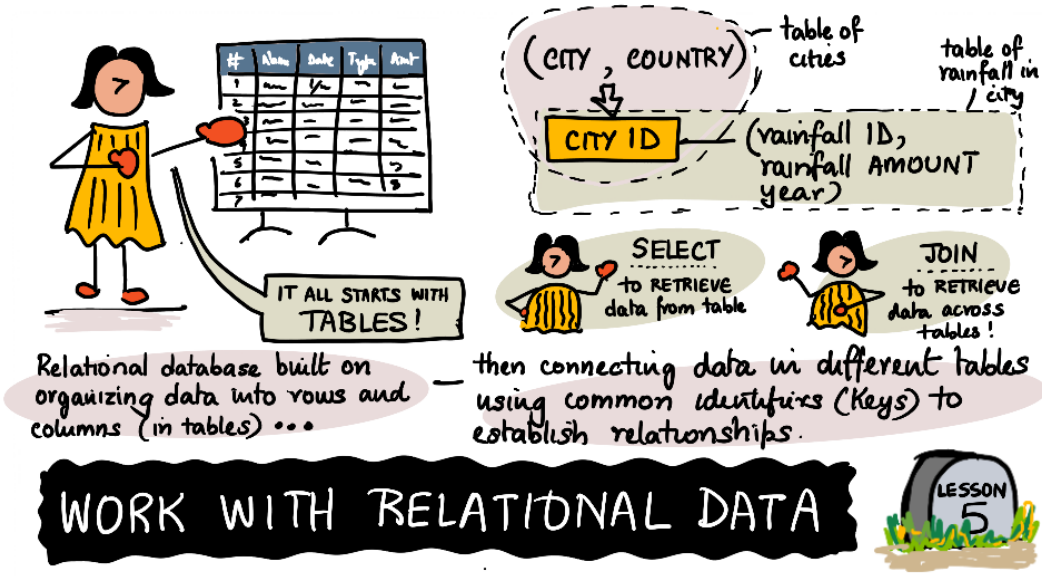


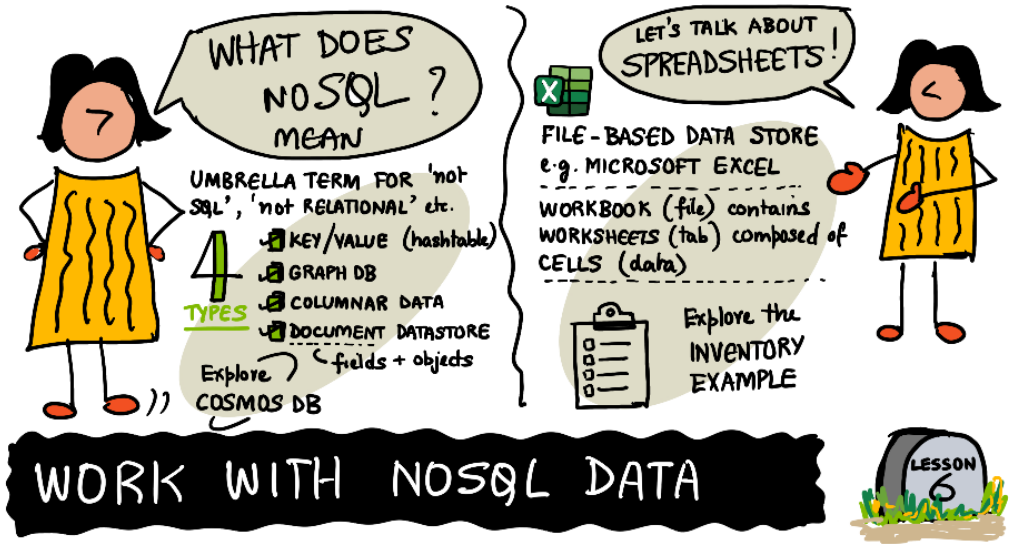


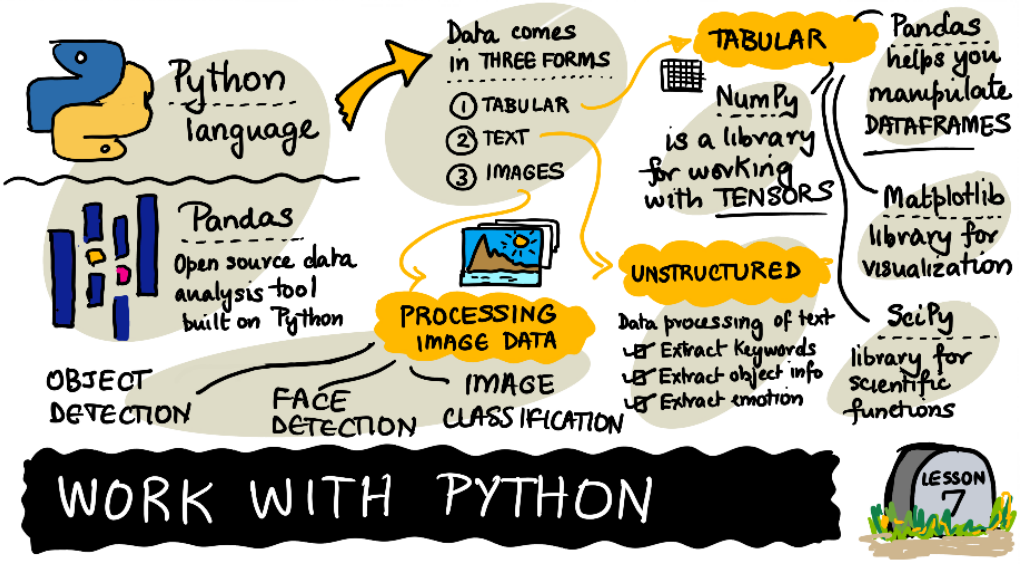

CONTROL
DATA OWNERSHIP
B
REPRESENTATION
ALGORITHM FAIRNESS
BIAS
DO NO HARM... MITIGATE BIAS
D
E
APPLIED ETHICS
CULTURE
DATA PRIVACY
RIGHTS
A
C
ETHICS DEFINED
DO NO HARM
DATA ANONYMITY
RESPONSIBLE AI
DATA SCIENCE ETHICS
LESSON 2

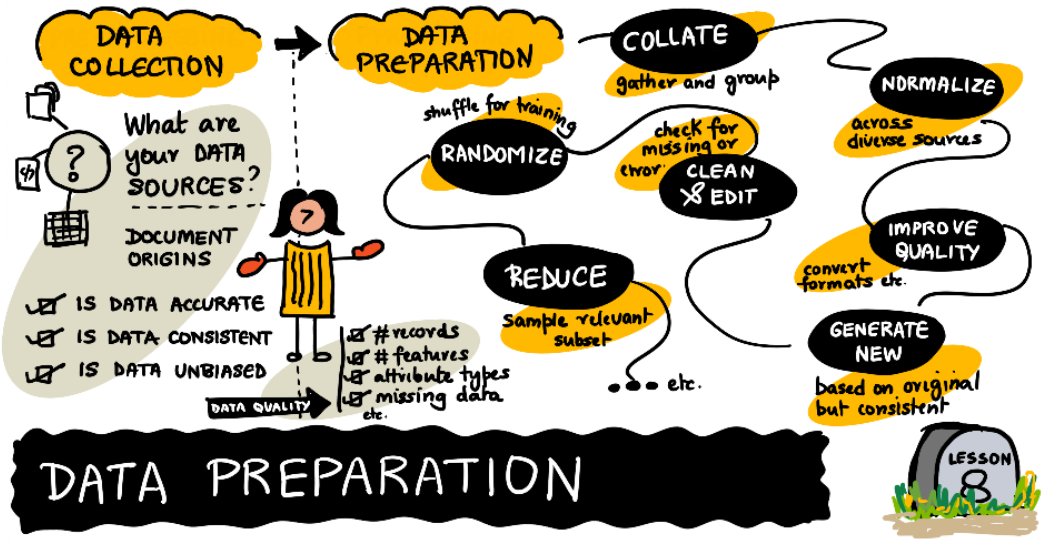

HOW CAN WE DESCRIBE DATA?
INFORMATION
OBSERVATIONS
MEASUREMENTS
CHARACTERISTICS AND SOURCES!
Sources
PRIMARY vs. SECONDARY
DATABASES
FILES
APIs
Types of Data
QUANTITATIVE - numerical observations
QUALITATIVE - categorical data
STRUCTURED - organized in rows/cols
UNSTRUCTURED - no clear format or category
SEMI-STRUCTURED
(in between = has some organization but no clear format
DEFINING DATA
LESSON 3

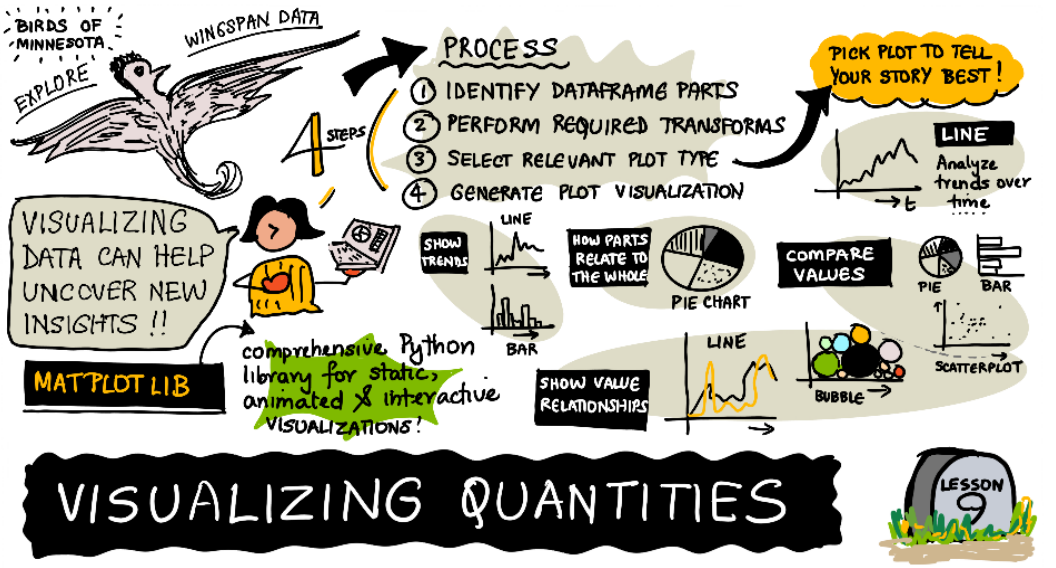

PROBABILITY IS THE LIKELIHOOD OF AN EVENT!
We use random variables to talk about events.
discrete
continuous
P(X) = function that describes probability of an event X
Probability Distribution
STATISTICS IS THE SCIENCE OF ANALYZING DATA!
Understanding
MEAN, VARIANCE, STANDARD DEVIATION
MODE, MEDIAN, QUARTILES
CONFIDENCE INTERVALS
COVARIANCE & CORRELATION
STATISTICS & PROBABILITY!
LESSON 4

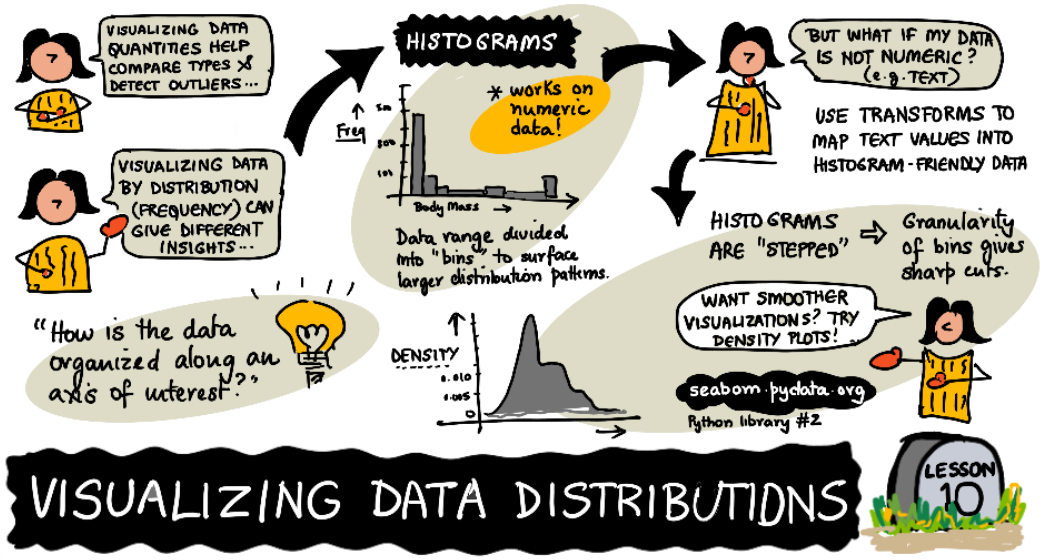

IT ALL STARTS WITH TABLES!
(CITY, COUNTRY)
table of cities
table of rainfall in city
CITY ID
(rainfall ID, rainfall AMOUNT year)
SELECT
to RETRIEVE data from table
JOIN
to RETRIEVE data across tables!
Relational database built on organizing data into rows and columns (in tables) ...
then connecting data in different tables using common identifiers (Keys) to establish relationships.
WORK WITH RELATIONAL DATA
LESSON 5

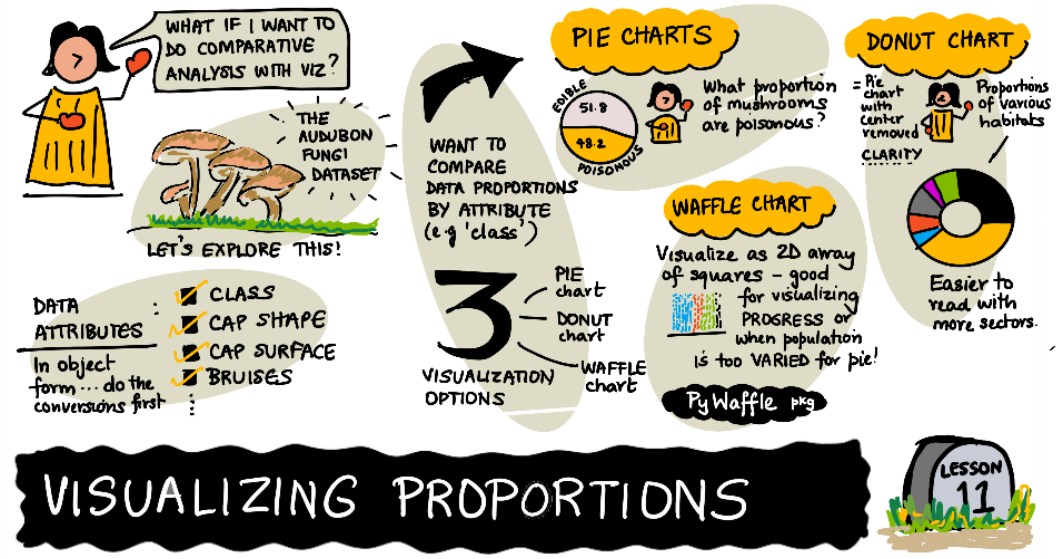

WHAT DOES NOSQL ? MEAN
UMBRELLA TERM FOR 'not SQL', 'not RELATIONAL' etc.
4 TYPES
KEY/VALUE (hashtable)
GRAPH DB
COLUMNAR DATA
DOCUMENT DATASTORE
fields + objects
Explore COSMOS DB
LET'S TALK ABOUT SPREADSHEETS!
FILE-BASED DATA STORE
e.g. MICROSOFT EXCEL
WORKBOOK (file) contains WORKSHEETS (tab) composed of CELLS (data)
Explore the INVENTORY EXAMPLE
WORK WITH NOSQL DATA
LESSON 6

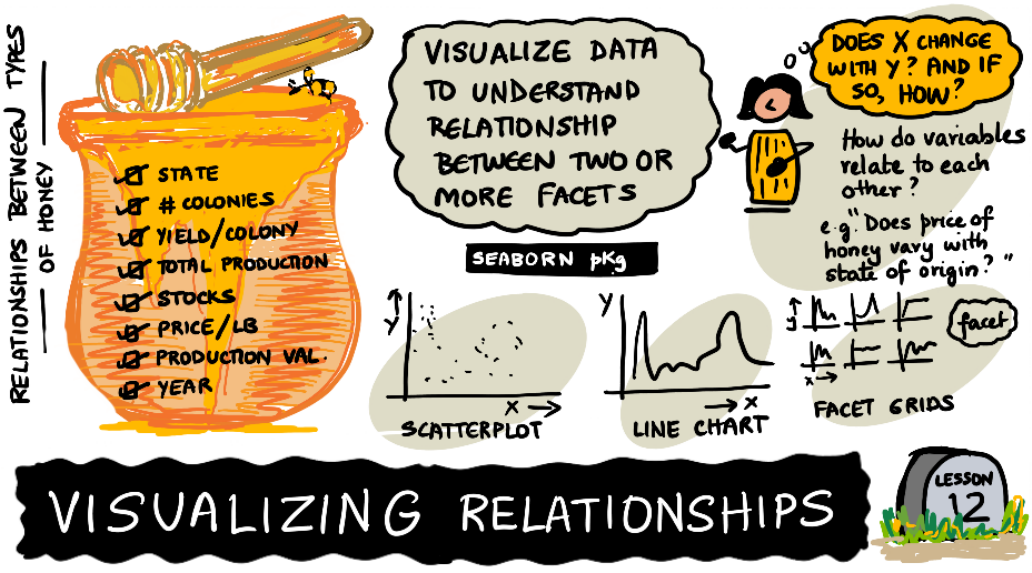

Python language
Pandas
Open source data analysis tool built on Python
Data comes in THREE FORMS
1 TABULAR
2 TEXT
3 IMAGES
TABULAR
NumPy is a library for working with TENSORS
Pandas helps you manipulate DATAFRAMES
Matplotlib library for visualization
SciPy library for scientific functions
UNSTRUCTURED
Data processing of text
Extract Keywords
Extract object info
Extract emotion
PROCESSING IMAGE DATA
OBJECT DETECTION
FACE DETECTION
IMAGE CLASSIFICATION
WORK WITH PYTHON
LESSON 7

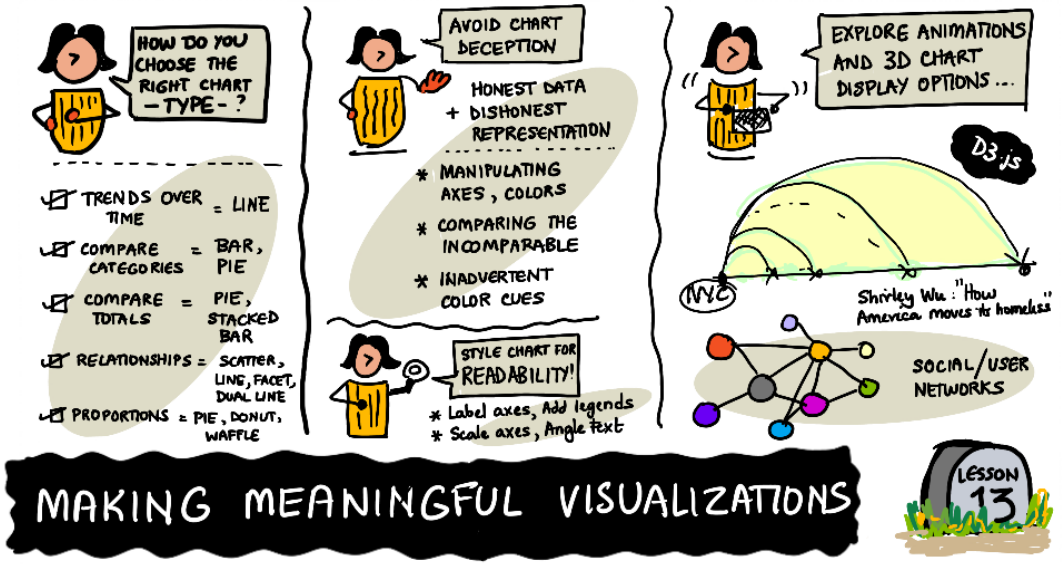

DATA COLLECTION
What are your DATA SOURCES?
DOCUMENT ORIGINS
IS DATA ACCURATE
IS DATA CONSISTENT
IS DATA UNBIASED
DATA QUALITY
#records
#features
attribute types
missing data
etc.
DATA PREPARATION
COLLATE
gather and group
NORMALIZE
across diverse sources
shuffle for training
RANDOMIZE
check for missing or error
CLEAN & EDIT
IMPROVE QUALITY
convert formats etc.
REDUCE
Sample relevant subset
etc.
GENERATE NEW
based on original but consistent
DATA PREPARATION
LESSON 8

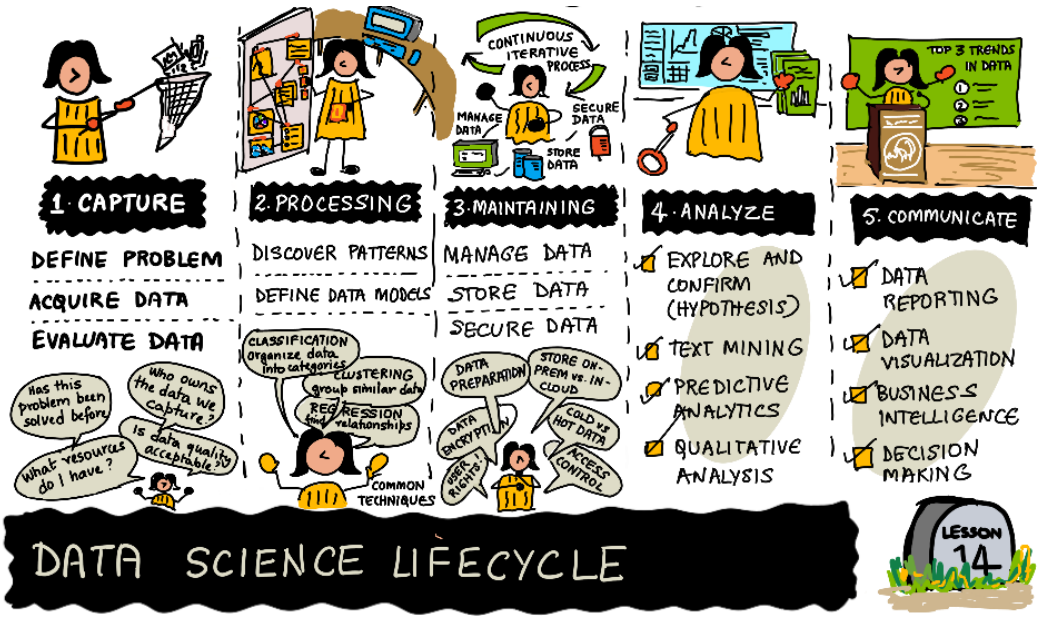

BIRDS OF MINNESOTA
EXPLORE
WINGSPAN DATA
4 STEPS
PROCESS
1 IDENTIFY DATAFRAME PARTS
2 PERFORM REQUIRED TRANSFORMS
3 SELECT RELEVANT PLOT TYPE
4 GENERATE PLOT VISUALIZATION
PICK PLOT TO TELL YOUR STORY BEST!
LINE
Analyze trends over time
VISUALIZING DATA CAN HELP UNCOVER NEW INSIGHTS !!
SHOW TRENDS
LINE
BAR
HOW PARTS RELATE TO THE WHOLE
PIE CHART
COMPARE VALUES
PIE
BAR
SCATTERPLOT
MATPLOTLIB
comprehensive Python library for static, animated & interactive VISUALIZATIONS!
SHOW VALUE RELATIONSHIPS
LINE
BUBBLE
VISUALIZING QUANTITIES
LESSON 9

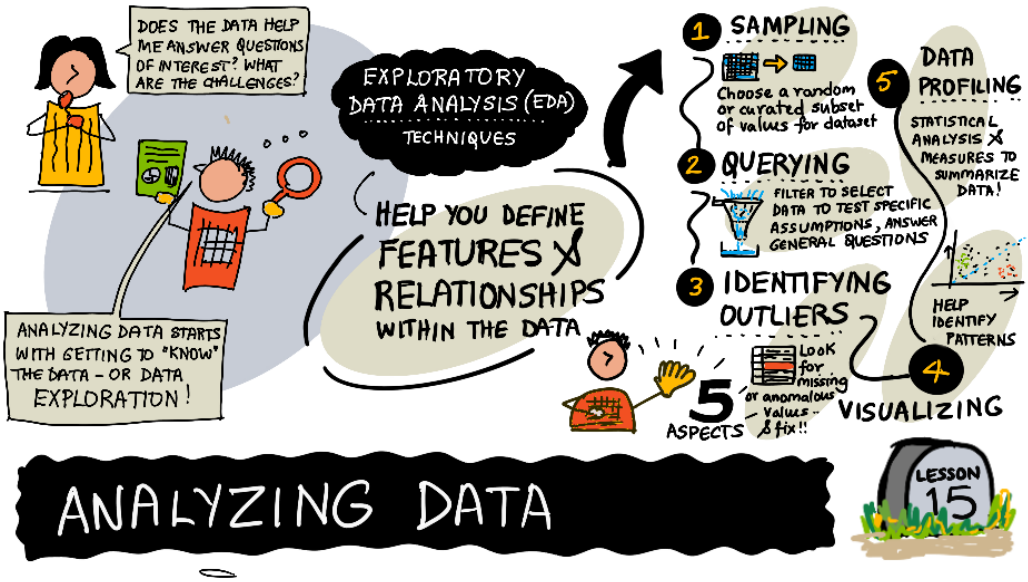

VISUALIZING DATA QUANTITIES HELP COMPARE TYPES & DETECT OUTLIERS...
VISUALIZING DATA BY DISTRIBUTION (FREQUENCY) CAN GIVE DIFFERENT INSIGHTS...
"How is the data organized along an axis of interest?"
HISTOGRAMS
* works on numeric data!
Freq
Body Mass
Data range divided into "bins" to surface larger distribution patterns.
BUT WHAT IF MY DATA IS NOT NUMERIC? (e.g. TEXT)
USE TRANSFORMS TO MAP TEXT VALUES INTO HISTOGRAM-FRIENDLY DATA
HISTOGRAMS ARE "STEPPED"
Granularity of bins gives sharp cuts.
WANT SMOOTHER VISUALIZATIONS? TRY DENSITY PLOTS!
DENSITY
seaborn.pydata.org
Python library #2
VISUALIZING DATA DISTRIBUTIONS
LESSON 10

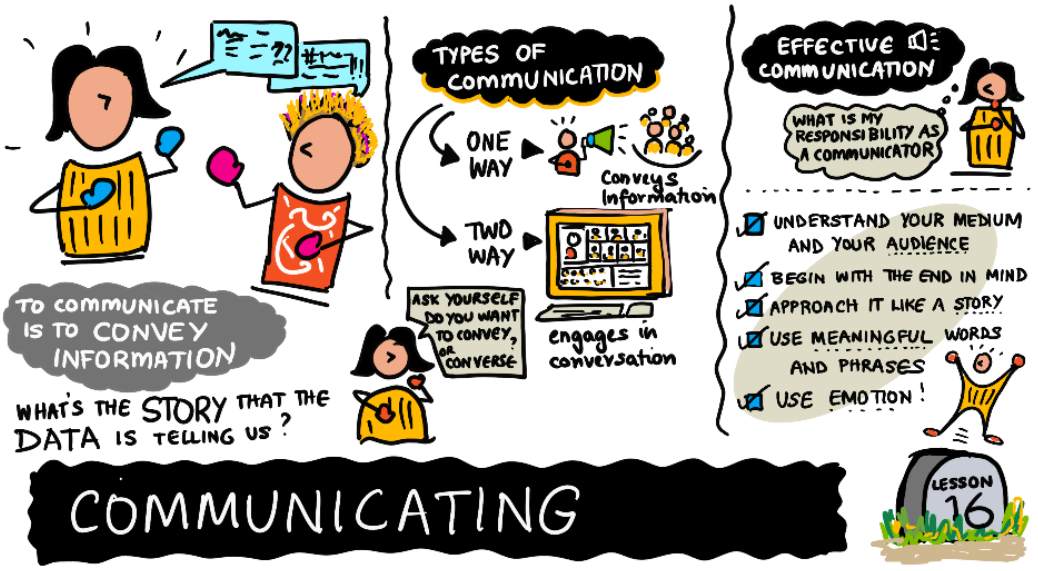

WHAT IF I WANT TO DO COMPARATIVE ANALYSIS WITH VIZ?
THE AUDUBON FUNGI DATASET
LET'S EXPLORE THIS!
DATA ATTRIBUTES
In object form... do the conversions first
CLASS
CAP SHAPE
CAP SURFACE
BRUISES
WANT TO COMPARE DATA PROPORTIONS BY ATTRIBUTE (e.g 'class')
3
VISUALIZATION OPTIONS
PIE chart
DONUT chart
WAFFLE chart
PIE CHARTS
EDIBLE
51.8
48.2
POISONOUS
What proportion of mushrooms are poisonous?
DONUT CHART
= Pie chart with center removed
Proportions of various habitats
CLARITY
Easier to read with more sectors.
WAFFLE CHART
Visualize as 2D array of squares - good for visualizing PROGRESS or when population is too VARIED for pie!
PyWaffle pkg
VISUALIZING PROPORTIONS
LESSON 11

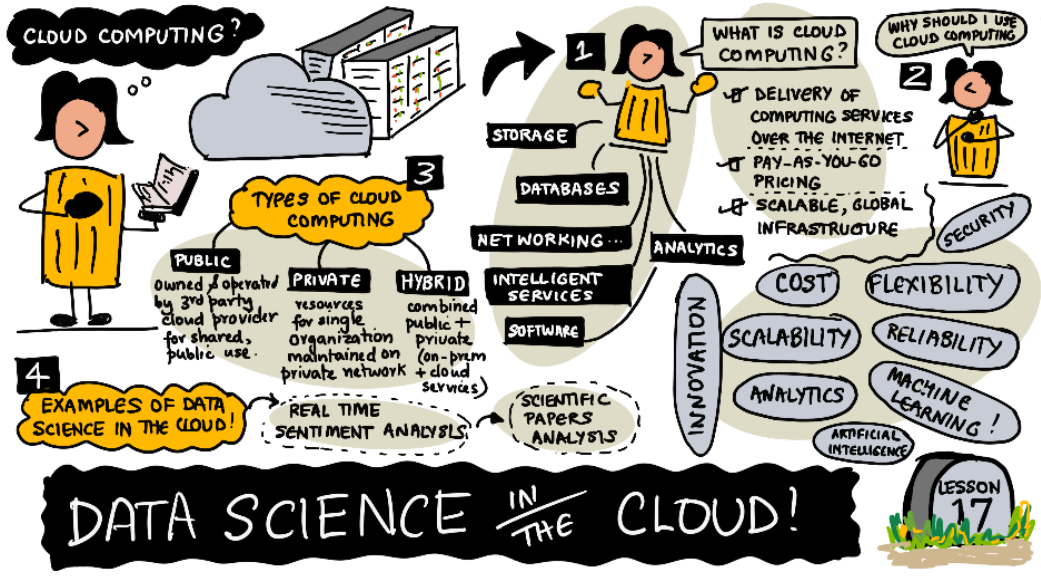

RELATIONSHIPS BETWEEN TYPES OF HONEY
STATE
COLONIES
YIELD/COLONY
TOTAL PRODUCTION
STOCKS
PRICE/LB
PRODUCTION VAL.
YEAR
VISUALIZE DATA TO UNDERSTAND RELATIONSHIP BETWEEN TWO OR MORE FACETS
SEABORN pkg
DOES X CHANGE WITH Y? AND IF SO, HOW?
How do variables relate to each other?
e.g. "Does price of honey vary with state of origin?"
SCATTERPLOT
LINE CHART
facet
FACET GRIDS
VISUALIZING RELATIONSHIPS
LESSON 12

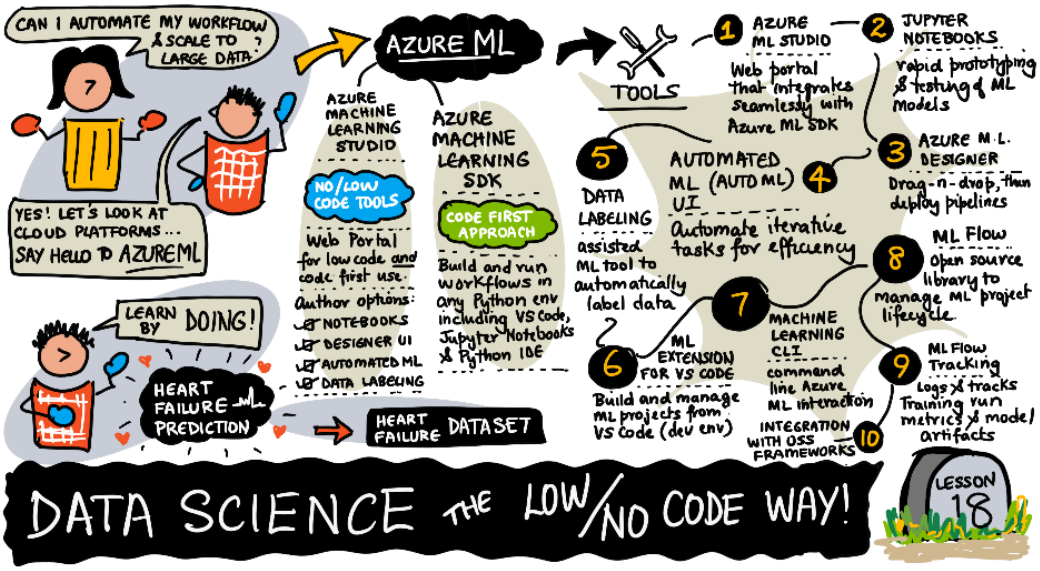

HOW DO YOU CHOOSE THE RIGHT CHART -TYPE-?
TRENDS OVER TIME = LINE
COMPARE CATEGORIES = BAR, PIE
COMPARE TOTALS = PIE, STACKED BAR
RELATIONSHIPS = SCATTER, LINE, FACET, DUAL LINE
PROPORTIONS = PIE, DONUT, WAFFLE
AVOID CHART DECEPTION
HONEST DATA + DISHONEST REPRESENTATION
* MANIPULATING AXES, COLORS
* COMPARING THE INCOMPARABLE
* INADVERTENT COLOR CUES
STYLE CHART FOR READABILITY!
* Label axes, Add legends
* Scale axes, Angle text
EXPLORE ANIMATIONS AND 3D CHART DISPLAY OPTIONS...
D3.js
NYC
Shirley Wu: "How America moves its homeless"
SOCIAL/USER NETWORKS
MAKING MEANINGFUL VISUALIZATIONS
LESSON 13

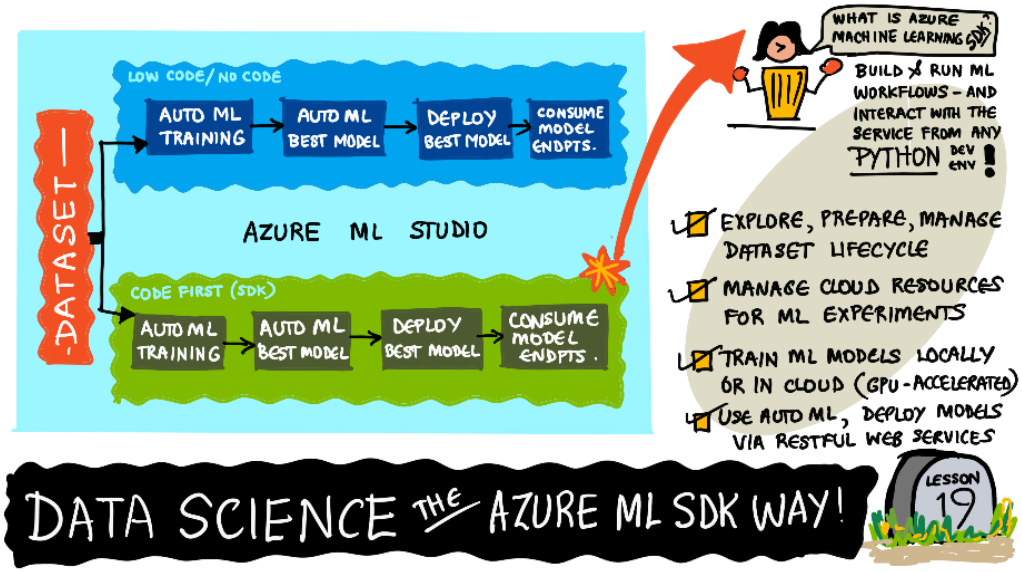

1 · CAPTURE
DEFINE PROBLEM
ACQUIRE DATA
EVALUATE DATA
Has this problem been solved before
Who owns the data we capture?
What resources do I have?
Is data quality acceptable?
2. PROCESSING
DISCOVER PATTERNS
DEFINE DATA MODELS
CLASSIFICATION organize data into categories
CLUSTERING group similar data
REGRESSION find relationships
COMMON TECHNIQUES
CONTINUOUS ITERATIVE PROCESS
MANAGE DATA
SECURE DATA
STORE DATA
3 · MAINTAINING
MANAGE DATA
STORE DATA
SECURE DATA
DATA PREPARATION
STORE ON-PREM VS. IN-CLOUD
DATA ENCRYPTION
COLD VS HOT DATA
USER RIGHTS
ACCESS CONTROL
4 · ANALYZE
EXPLORE AND CONFIRM (HYPOTHESIS)
TEXT MINING
PREDICTIVE ANALYTICS
QUALITATIVE ANALYSIS
TOP 3 TRENDS IN DATA
5. COMMUNICATE
DATA REPORTING
DATA VISUALIZATION
BUSINESS INTELLIGENCE
DECISION MAKING
DATA SCIENCE LIFECYCLE
LESSON 14

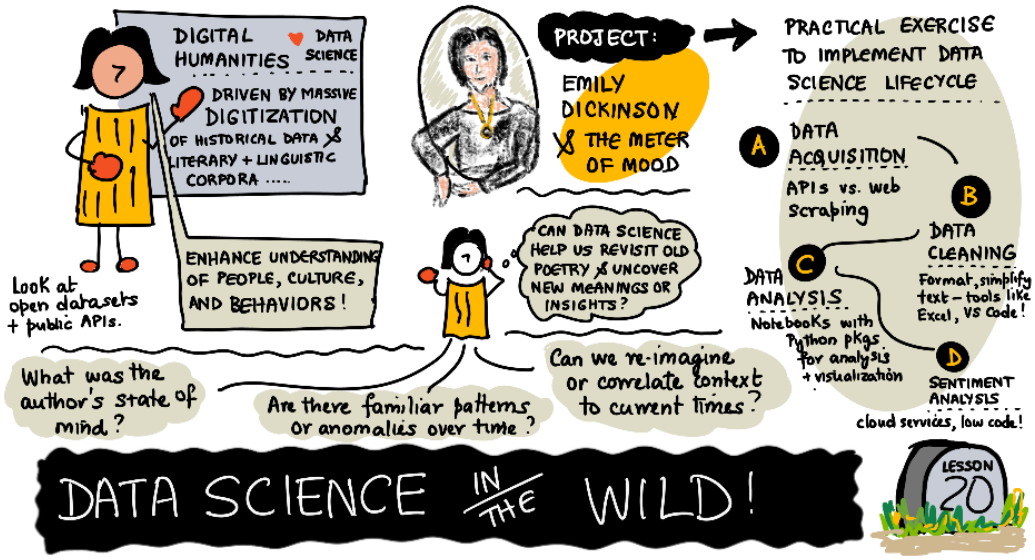

DOES THE DATA HELP ME ANSWER QUESTIONS OF INTEREST? WHAT ARE THE CHALLENGES?
ANALYZING DATA STARTS WITH GETTING TO "KNOW" THE DATA - OR DATA EXPLORATION!
EXPLORATORY DATA ANALYSIS (EDA) TECHNIQUES
HELP YOU DEFINE FEATURES & RELATIONSHIPS WITHIN THE DATA
1 SAMPLING
Choose a random or curated subset of values for dataset
2 QUERYING
FILTER TO SELECT DATA TO TEST SPECIFIC ASSUMPTIONS, ANSWER GENERAL QUESTIONS
3 IDENTIFYING OUTLIERS
Look for missing or anomalous values & fix!!
4 VISUALIZING
HELP IDENTIFY PATTERNS
5 DATA PROFILING
STATISTICAL ANALYSIS & MEASURES TO SUMMARIZE DATA!
5 ASPECTS
ANALYZING DATA
LESSON 15

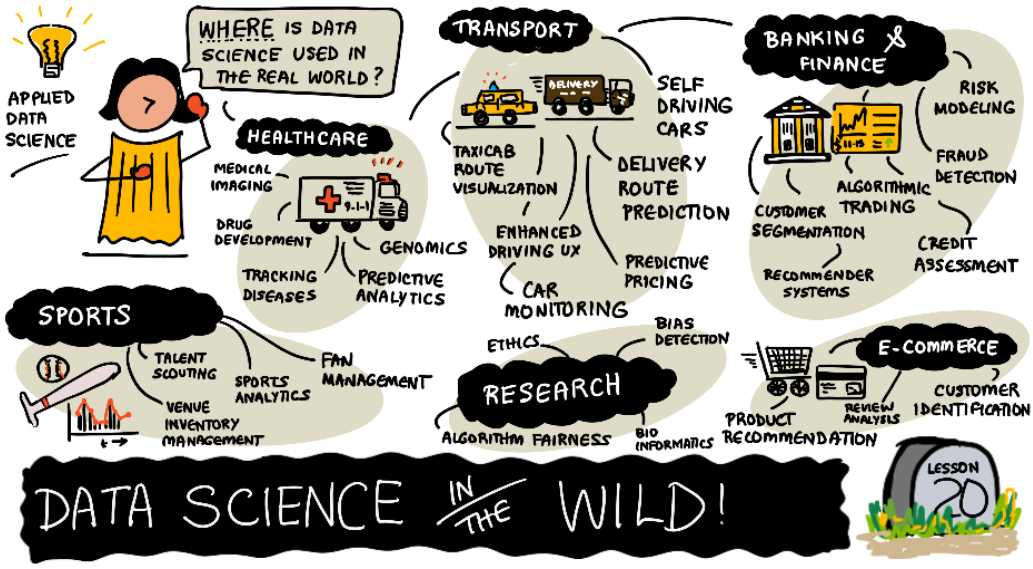

TYPES OF COMMUNICATION
ONE WAY
Conveys Information
TWO WAY
engages in conversation
ASK YOURSELF DO YOU WANT TO CONVEY, OR CONVERSE
TO COMMUNICATE IS TO CONVEY INFORMATION
WHAT'S THE STORY THAT THE DATA IS TELLING US?
EFFECTIVE COMMUNICATION
WHAT IS MY RESPONSIBILITY AS A COMMUNICATOR
UNDERSTAND YOUR MEDIUM AND YOUR AUDIENCE
BEGIN WITH THE END IN MIND
APPROACH IT LIKE A STORY
USE MEANINGFUL WORDS AND PHRASES
USE EMOTION!
COMMUNICATING
LESSON 16

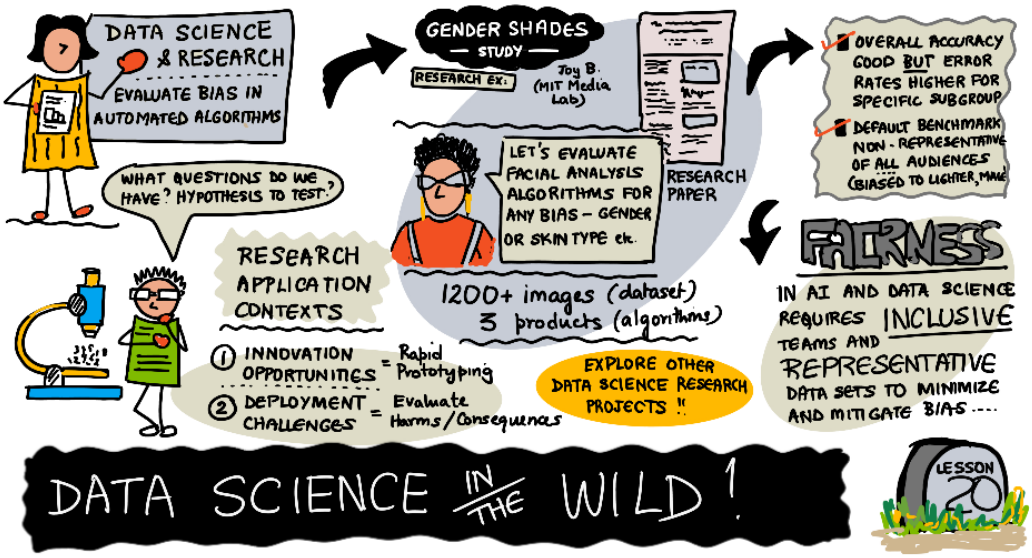

CLOUD COMPUTING?
1
WHAT IS CLOUD COMPUTING?
DELIVERY OF COMPUTING SERVICES OVER THE INTERNET
PAY-AS-YOU-GO PRICING
SCALABLE, GLOBAL INFRASTRUCTURE
STORAGE
DATABASES
NETWORKING
ANALYTICS
INTELLIGENT SERVICES
SOFTWARE
2
WHY SHOULD I USE CLOUD COMPUTING
SECURITY
COST
FLEXIBILITY
SCALABILITY
RELIABILITY
ANALYTICS
MACHINE LEARNING!
ARTIFICIAL INTELLIGENCE
INNOVATION
3
TYPES OF CLOUD COMPUTING
PUBLIC
Owned & operated by 3rd party cloud provider for shared, public use.
PRIVATE
resources for single organization maintained on private network
HYBRID
combined public + private (on-prem + cloud services)
4
EXAMPLES OF DATA SCIENCE IN THE CLOUD!
REAL TIME SENTIMENT ANALYSIS
SCIENTIFIC PAPERS ANALYSIS
DATA SCIENCE IN THE CLOUD!
LESSON 17

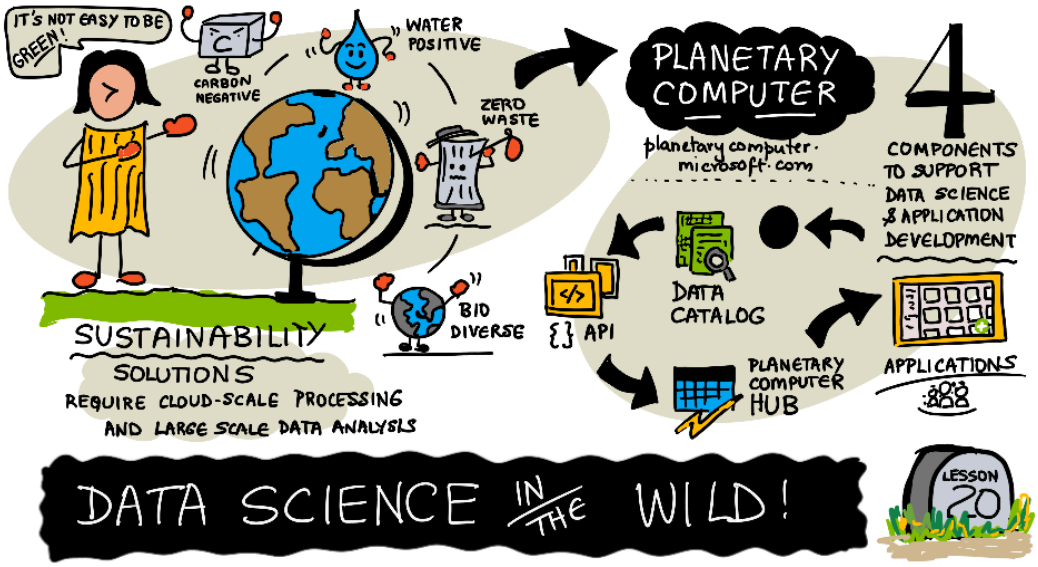

CAN I AUTOMATE MY WORKFLOW & SCALE TO ? LARGE DATA
YES! LET'S LOOK AT CLOUD PLATFORMS... SAY HELLO TO AZUREML
AZURE ML
AZURE MACHINE LEARNING STUDIO
NO/LOW CODE TOOLS
Web Portal for low code and code first use.
Author options:
NOTEBOOKS
DESIGNER UI
AUTOMATED ML
DATA LABELING
AZURE MACHINE LEARNING SDK
CODE FIRST APPROACH
Build and run workflows in any Python env including VS Code, Jupyter Notebooks & Python IDE
TOOLS
1 AZURE ML STUDIO
Web portal that integrates seamlessly with Azure ML SDK
2 JUPYTER NOTEBOOKS
rapid prototyping & testing of ML models
3 AZURE M.L. DESIGNER
Drag-n-drop, then deploy pipelines
4 AUTOMATED ML (AUTO ML) UI
Automate iterative tasks for efficiency
5 DATA LABELING
assisted ML tool to automatically label data
6 ML EXTENSION FOR VS CODE
Build and manage ML projects from VS Code (dev env)
7 MACHINE LEARNING CLI
command line Azure ML interaction
INTEGRATION WITH OSS FRAMEWORKS
8 ML Flow
open source library to manage ML project lifecycle
9 ML Flow Tracking
Logs & tracks Training run metrics & model artifacts
10
LEARN BY DOING!
HEART FAILURE PREDICTION
HEART FAILURE DATASET
DATA SCIENCE THE LOW/NO CODE WAY!
LESSON 18


DATASET
LOW CODE/NO CODE
AUTO ML TRAINING
AUTO ML BEST MODEL
DEPLOY BEST MODEL
CONSUME MODEL ENDPTS.
AZURE ML STUDIO
CODE FIRST (SDK)
AUTO ML TRAINING
AUTO ML BEST MODEL
DEPLOY BEST MODEL
CONSUME MODEL ENDPTS.
WHAT IS AZURE MACHINE LEARNING SDK
BUILD & RUN ML WORKFLOWS - AND INTERACT WITH THE SERVICE FROM ANY PYTHON DEV ENV
EXPLORE, PREPARE, MANAGE DATASET LIFECYCLE
MANAGE CLOUD RESOURCES FOR ML EXPERIMENTS
TRAIN ML MODELS LOCALLY OR IN CLOUD (GPU-ACCELERATED)
USE AUTO ML, DEPLOY MODELS VIA RESTFUL WEB SERVICES
DATA SCIENCE THE AZURE ML SDK WAY!
LESSON 19


DIGITAL HUMANITIES ♥ DATA SCIENCE
DRIVEN BY MASSIVE DIGITIZATION OF HISTORICAL DATA & LITERARY + LINGUISTIC CORPORA .....
ENHANCE UNDERSTANDING OF PEOPLE, CULTURE, AND BEHAVIORS!
Look at open datasets + public APIs.
PROJECT:
EMILY DICKINSON & THE METER OF MOOD
PRACTICAL EXERCISE TO IMPLEMENT DATA SCIENCE LIFECYCLE
A DATA ACQUISITION
APIs vs. web Scraping
B DATA CLEANING
Format, simplify text – tools like Excel, VS Code!
C DATA ANALYSIS
Notebooks with Python pkgs for analysis + visualization
D SENTIMENT ANALYSIS
cloud services, low code!
CAN DATA SCIENCE HELP US REVISIT OLD POETRY & UNCOVER NEW MEANINGS OR INSIGHTS?
What was the author's state of mind?
Are there familiar patterns or anomalies over time?
Can we re-imagine or correlate context to current times?
DATA SCIENCE IN THE WILD!
LESSON 20


APPLIED DATA SCIENCE
WHERE IS DATA SCIENCE USED IN THE REAL WORLD?
HEALTHCARE
MEDICAL IMAGING
DRUG DEVELOPMENT
GENOMICS
TRACKING DISEASES
PREDICTIVE ANALYTICS
TRANSPORT
SELF DRIVING CARS
TAXICAB ROUTE VISUALIZATION
DELIVERY ROUTE PREDICTION
ENHANCED DRIVING UX
PREDICTIVE PRICING
CAR MONITORING
BANKING & FINANCE
RISK MODELING
FRAUD DETECTION
CUSTOMER SEGMENTATION
ALGORITHMIC TRADING
CREDIT ASSESSMENT
RECOMMENDER SYSTEMS
SPORTS
TALENT SCOUTING
SPORTS ANALYTICS
FAN MANAGEMENT
VENUE INVENTORY MANAGEMENT
RESEARCH
ETHICS
BIAS DETECTION
ALGORITHM FAIRNESS
BIO INFORMATICS
E-COMMERCE
CUSTOMER IDENTIFICATION
PRODUCT RECOMMENDATION
REVIEW ANALYSIS
DATA SCIENCE IN THE WILD!
LESSON 20





***Sketch Credit*: Nitya Narasimhan**

- https://github.com/microsoft/Data-Science-For-Beginners/tree/main/sketchnotes

Follow Aspersh Upadhyay for more like this. Read my free articles on Medium.

Get free learning resources on Bits of Data Science related to data science, Artificial Intelligence, Machine Learning, Python and more.

Bits of Data science

Welcome Data Pioneers. Here you will get all possible resources related to AI | ML | DL | PYTHON | and more. Insta:

https://t.me/bitsofds