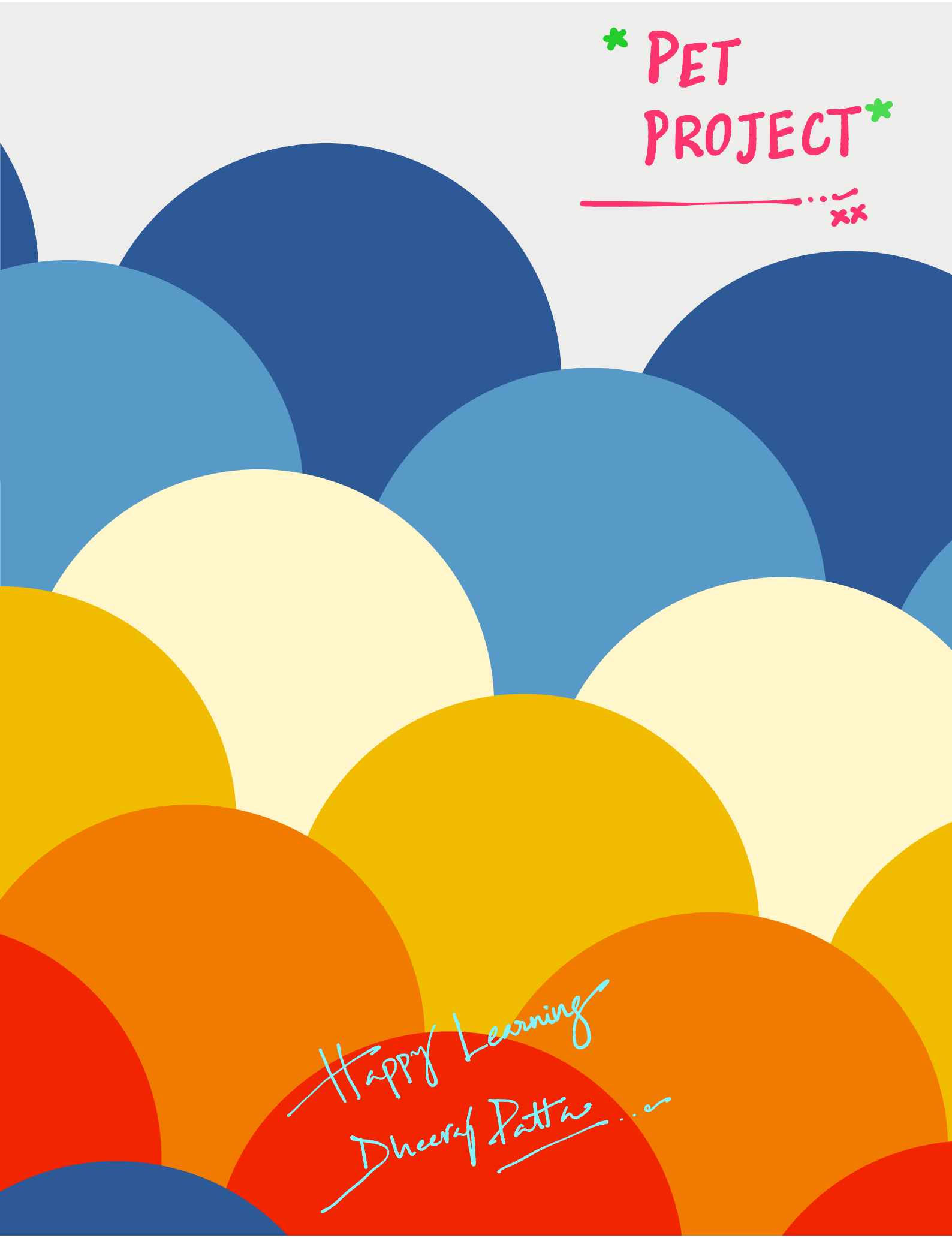
* PET PROJECT *



3 INGREDIENTS OF SUCCESSFUL PET-PROJECT

- YOUR **MINDSET** [IT'S YOUR LIFE. PLAN]
- YOUR **STRATEGY** [WHERE YOU WANT TO GO]
- YOUR **EXECUTION** [HOW YOU WILL GET THERE]

WHY PET-PROJECTS IN FIRST PLACE?

- ☑ EXPERIMENTATION
- ☑ NEW LEARNING
- ☑ INSPIRATION
- ☑ NEW SKILLS
- ☑ YOUR STORY
- ☑ GREAT THING TO TALK ABOUT
- ☑ JOB
- ☑ SHOWCASE
- ☑ STRETCH GOALS
- ☑ GAIN WORK EXPERIENCE
- ☑ FUN

WHAT SHOULD BE YOUR GOAL?

- MAKE IT YOUR **VOICE** EVERYWHERE YOU GO (SOFT SKILL - PRACTISE)
- YOUR **BRAND IDENTITY** (STRATEGIZE. KNOW WHO YOU WANT TO BE)
- FOCUS ON CREATING **VALUE** & **SOLVING NEEDS** (KEY)
- HAVE **STRONG FOUNDATION** (TECHNICAL AND BEHAVIORAL)
- BUILD TANGIBLE AND SCALABLE **PRODUCTS** (CAN YOU BUILD A BUSINESS OUT OF YOUR PROJECT IDEA?)
- ADD YOUR **UNIQUE** PERSPECTIVE
- CREATE TOUCHPOINTS TO **TALK & SHOWCASE** (IN INTERVIEWS)
- TELL A **COMPELLING STORY** (EMOTIONAL CONNECT)
- HAVE **FUN** & **LOVE** WHAT YOU DO
- BRINGING **EVERYTHING TOGETHER**

LESSONS FROM MY **JOURNEY LEARNING EXPERIENCE**

#1. DEFINE PROBLEM*

- CLEAR & CONCISE
- ACTIONABLE
- MEANINGFUL
- MEASURABLE
- STRUCTURED APPROACH

2* SIMPLE WAYS TO DEFINE

POINT OF VIEW STATEMENT [POV]

_____ (USER) NEEDS A WAY TO _____ (VERB) BECAUSE _____ (COMPELLING INSIGHT)

WHY-HOW LADDERING

- ASK "WHY" TO GET MORE ABSTRACT STATEMENTS
- ASK "HOW" TO GET MORE SPECIFIC STATEMENTS

PROBLEM SOLVING TECHNIQUES / WAYS

- ZOOM IN & OUT OF THE PROBLEM (5 WHY'S → GO DEEPER LEVELS)
- REDUCE PROBLEM TO ONE-WORD AND FIND RELATED CONCEPTS (MIND MAPS)
- SWITCH DIFFERENT ROLES AND PERSPECTIVES (ROLE PLAYING)
- DEFINE CAUSE-AND-EFFECT AND EXPAND ON SUB-CAUSES (FISHBONE DIAGRAM)
- GATHER SIMILAR ITEMS, TYPES, NEEDS, FEATURES (AFFINITY DIAGRAM)
- CONTINOUS IMPROVEMENT OF PROCESSES, PRODUCTS (DEMING'S WHEEL - PDCA)
  PLAN - DO - CHECK - ACT

MY CHECKLIST

- ☐ PROBLEM STATEMENT*
- ☐ PROBLEM MINDMAPS
- ☐ STAKEHOLDER MAPS
- ☐ ASSUMPTIONS & QUESTIONS
- ☐ POSSIBILITIES LIST*
  (ALL POSSIBLE USECASES FOR YOUR PROJECT)
- ☐ RETURN ON INVESTMENT*
  (ROI FOR BUSINESS)
- ☐ NEXT BEST ALTERNATIVE
- ☐ EARLY HYPOTHESIS*
- ☐ POTENTIAL SOLUTION
- ☐ INTEGRATION INTO BUSINESS*
  (HOW TO SCALE)

* INTERVIEW TOUCHPOINTS

LESSONS FROM MY JOURNEY LEARNING EXPERIENCE

#2 CHOOSE A METHODOLOGY*

BRING THE BEST IN YOUR PET PROJECTS xx

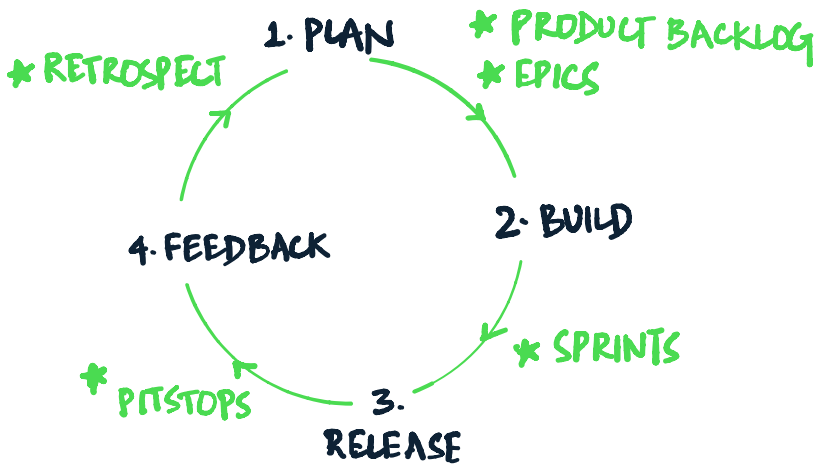
WHY?
- LIFECYCLE OF ANY PIECE OF SOFTWARE
- PROCESS BRINGS STRUCTURE & CLARITY
- CONTROL DEVELOPMENT OF ANY INFORMATION SYSTEM
- FRAMEWORK FOR SUCCESSFUL PRODUCTS/PROJECTS/SERVICES

3 DEVELOPMENT METHODOLOGIES

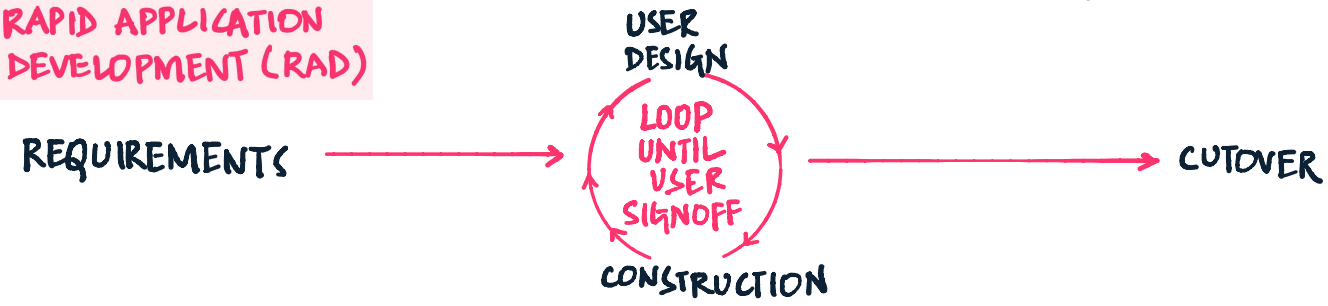
TRADITIONAL

SEQUENTIAL
- REQUIREMENTS
  - ANALYSIS
    - DESIGN
      - IMPLEMENTATION
        - VALIDATION (V&V)
          - DEPLOYMENT/ MAINTENANCE

AGILE/SCRUM



RAPID APPLICATION DEVELOPMENT (RAD)



MY RECOMMENDATION [BASED ON MY WORK EXPERIENCE]

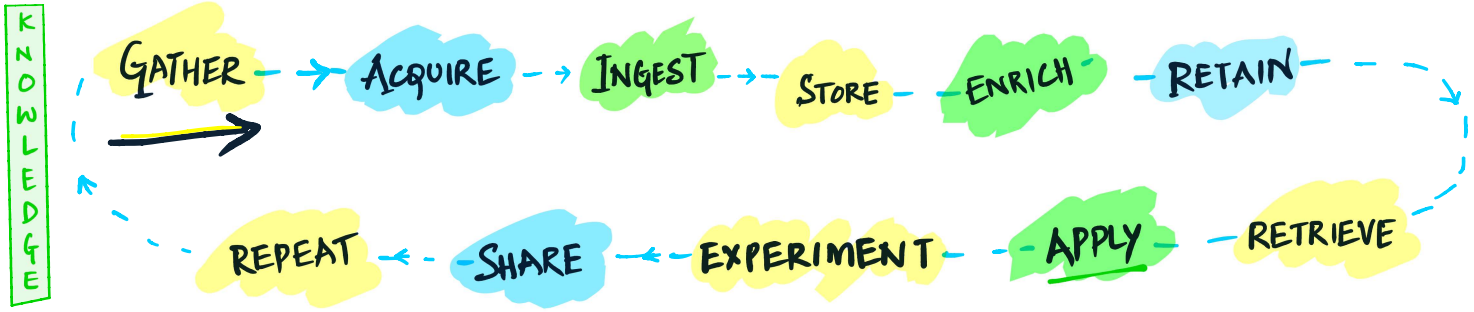
- ☐ IMPROVING PROCESS? → SIX SIGMA/LEAN
- ☐ BUILDING PRODUCT (INCREMENTAL UPDATES)? → AGILE/SCRUM
- ☐ LONG-TERM & FIXED PROJECT? → TRADITIONAL
- ☐ HIGH STAKES & HIGH VISIBILITY? → HYBRID (AGILE + ITERATIVE)

* MAKE A POINT TO DISCUSS ABOUT YOUR METHODOLOGY IN YOUR INTERVIEWS *

#3 GATHER KNOWLEDGE & IDEATE

KNOWLEDGE PHASES

ELEMENT OF THOUGHT
• VISUAL • CONCRETE • ABSTRACT



IDEA MANAGEMENT

GENERATE
→ COLLECT (INCUBATE)
→ EVALUATE
→ SELECT
→ REFINE (REVISION)
→ IMPLEMENT*

MY IDEA GENERATION TECHNIQUES

- MINDMAPS*
- DOODLING
- DESIGN THINKING*
- 5WHY'S
- REVERSE ENGINEER
- COLLABORATION*

MY CHECKLIST [FOR PET PROJECTS]

- ☐ ELEVATOR PITCH* (15 SECS TO EXPLAIN YOUR PET PROJECT)
- ☐ TAG LINE (SINGLE LINE)
- ☐ IDEATION TECHNIQUES*
- ☐ LANDSCAPING (IDENTIFY GAPS) (USED WHILE GENERATING IDEAS FOR PATENTS)
- ☐ IDEA POOL (LIST OF ALL IDEAS GOOD, BAD, OBVIOUS)
- ☐ USECASES*
- ☐ IDEA STACKING* (STACK MULTIPLE IDEAS TO FORM BIGGER AND BETTER IDEAS)
- ☐ MOODBOARDS (FOR INSPIRATION)
- ☐ OPPORTUNITY STATEMENT
- ☐ OSBORN CHECKLIST (GOOGLE)
- ☐ PRODUCT OR BUSINESS IDEAS

MAKE A POINT TO DISCUSS YOUR IDEATION TECHNIQUES IN YOUR INTERVIEWS

#4 DATA ESSENTIALS 101*

- INDIVIDUAL PIECES OF FACTUAL INFORMATION
- BASIS OF REASONING AND INTERPRETATIONS
- QUINTESSENTIAL COMPONENT OF DECISION SCIENCE

DATA

- STRUCTURED (TABULAR)
- QUASI STRUCTURED (WEB, CLICKSTREAM)
- SEMI STRUCTURED (XML, HTML)
- UNSTRUCTURED (AUDIO, VIDEO, IMAGES ETC)

DATA

- QUANTITATIVE* (NUMERICAL)
  - CONTINOUS
    - INTERVAL
    - RATIO
  - DISCRETE
- QUALITATIVE (CATEGORICAL)
  - ORDINAL
  - NOMINAL
  - BINOMIAL

* G IS YOUR BEST FRIEND *

DATA LIFE CYCLE

COLLECTION → PROCESSING → CERTIFICATION → STORE → ACCESS → ANALYSIS → PUBLISH → SHARE → REUSE → PRESERVE → ARCHIVE → DESTROY (END OF LIFE)

DATA COLLECTION METHODS

DATA

- EXPERIMENTS
  - CONTROLLED OBSERVATIONS
- FOCUS GROUP DISCUSSIONS
- PUBLIC RESEARCH
- DELPHI TECHNIQUE
- SURVEYS QUESTIONAIRES POLLS
- INTERVIEWS
- PRIMARY DATA SOURCES (UPSTREAM SYSTEMS)

DATA PROCESSING TYPES

- BATCH PROCESSING
- REAL-TIME
- STREAMING
- MULTI PROCESSING
- TIME SHARED
- AUTOMATIC PROCESSING

TO BE CONTINUED XX

#5 DATA ESSENTIALS 101*

DATA STORAGE & RETRIEVAL POLICIES

- **HOT** → HIGH-PERFORMANCE*
  FREQUENTLY ACCESSED (IN-MEMORY/EDGE/SSD)
  FASTER STORAGE
- **WARM** → CAPACITY-OPTIMIZED
  LESS FREQUENTLY ACCESSED (DATA WAREHOUSES/DATA MARTS)
  SLIGHTLY SLOWER STORAGE
- **COLD** → LONG-TERM ARCHIVE (DATA LAKES/CLOUD)
  VERY RARELY ACCESSED (HISTORICAL DATA)
  SLOWEST STORAGE (HADOOP, VORA)

DATA SECURITY

3 STATES

- DATA AT REST → ENCRYPTION / PASSCODES / KEYS
- DATA IN USE → ACCESS PREVILIGES / USER SECURITY
  SSO AUTHENTICATION
- DATA IN MOTION → TRUSTED NETWORKS/ ENCRYPTION

ENTERPRISE DATA PIPELINE [GENERIC & APPLICABLE TO MOST ORGANIZATIONS]

UPSTREAMS*
SOURCE SYSTEMS

- OLTP (TRANSACTIONAL)
- CRM
- ERP SYSTEMS
- APPLICATIONS
- WEB/CLOUD
- MAINFRAMES
- DATABASES
- LOGS
- SAAS
- MACHINE DATA
- CONNECTED DEVICES (IOT)

DATA PIPELINES*

- MASS INGESTION
- ETL/LT/ELT
- DATA REPLICATION
- DATA STREAMING
- API + APPLICATION DATA

NO SQL

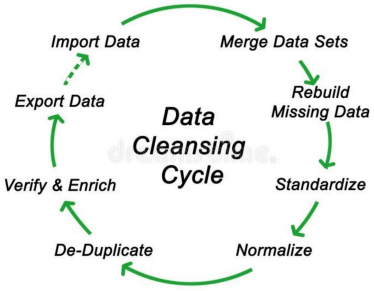
- DATA LAKES
- CLOUD
- DATA WAREHOUSES

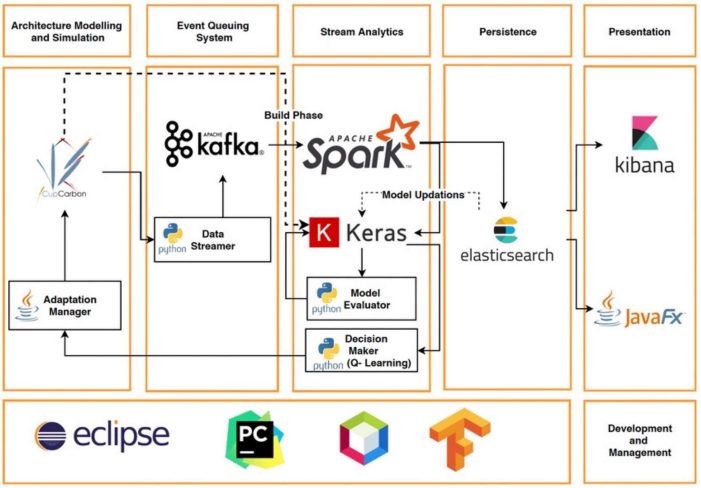
INSIGHTS*

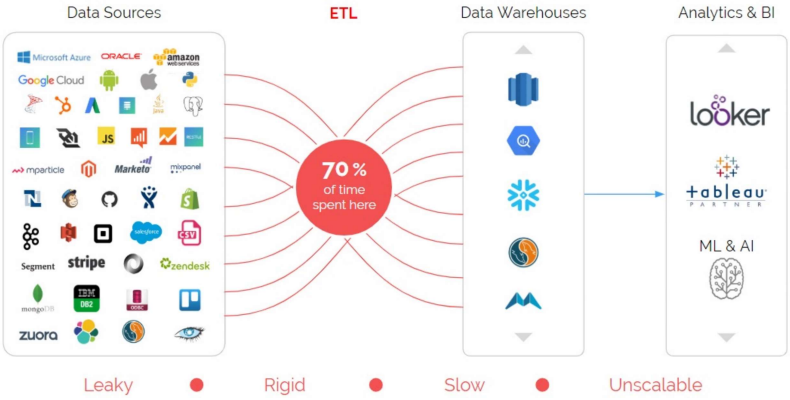
- ADVANCED ANALYTICS
- BUSINESS INTELLIGENCE
- DATA SCIENCE
- MACHINE LEARNING
- DECISION SCIENCE
- INSIGHTS/APPS
- VISUAL STORYTELLING

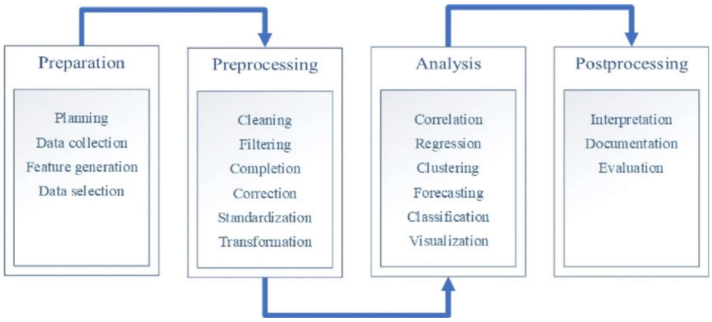
TO BE CONTINUED XX

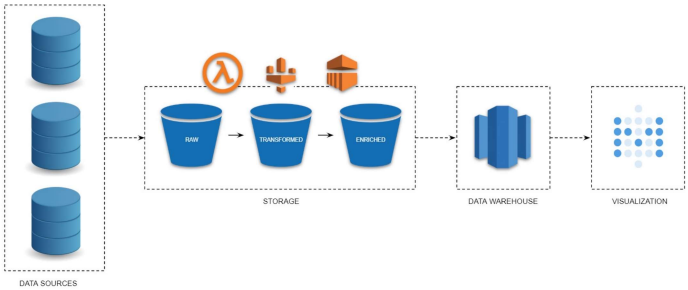
#6 KNOW YOUR DATA [IN THE CONTEXT OF PET PROJECT]

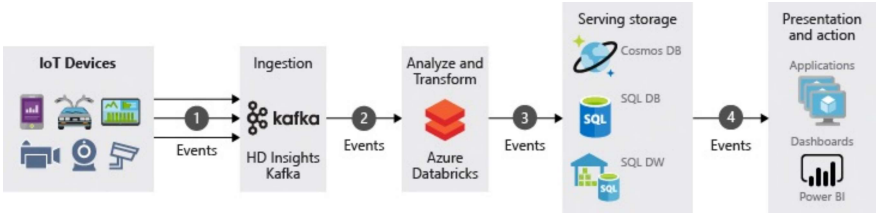


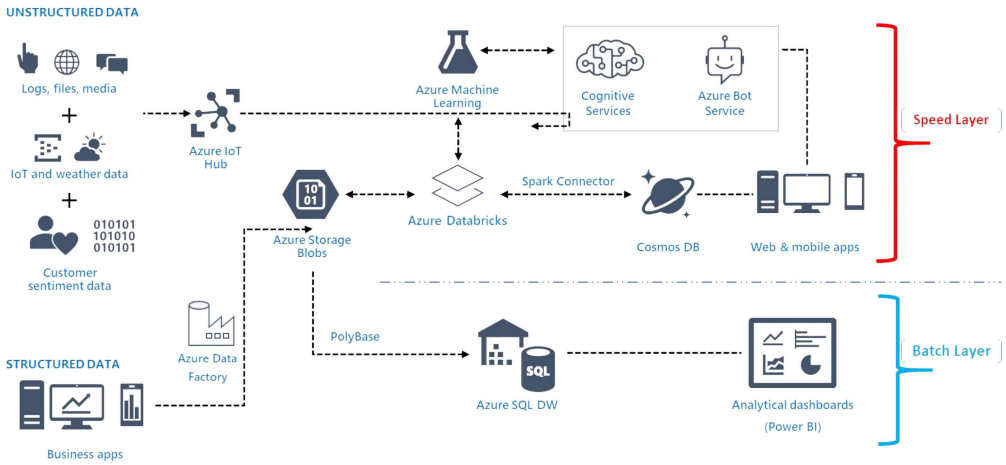












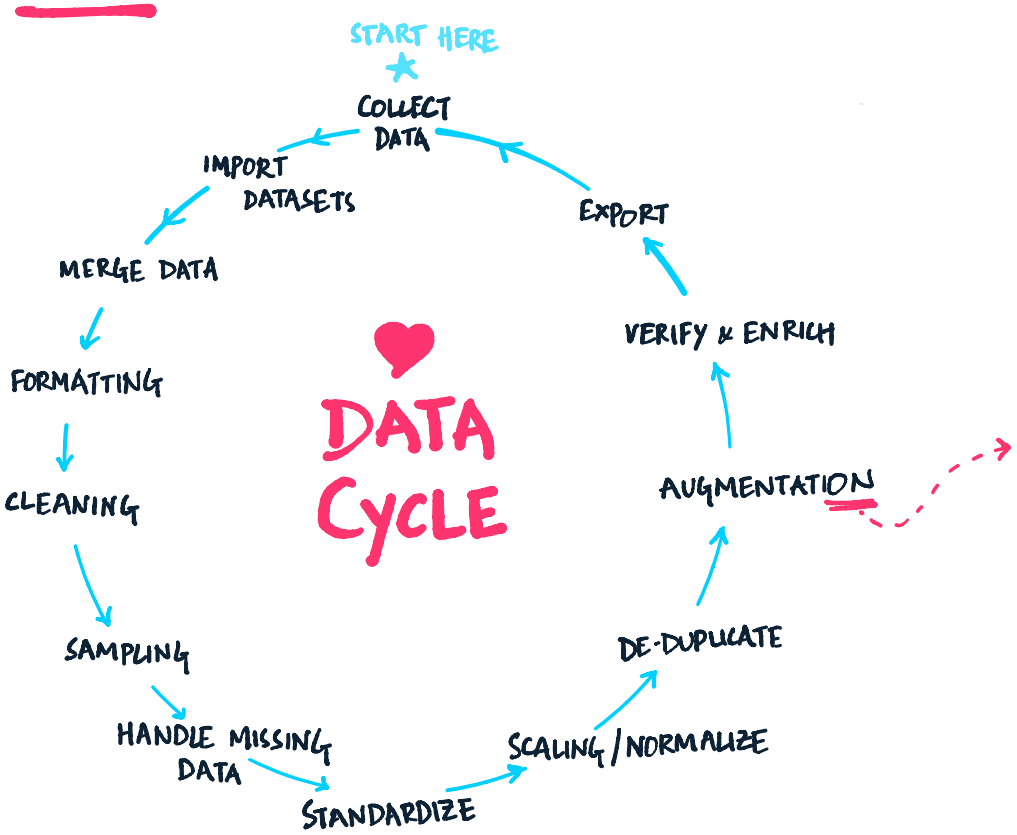
#6 KNOW YOUR DATA [IN THE CONTEXT OF PET PROJECT]

3 STEP DATA MANAGEMENT

DATA COLLECTION

DATA PRE-PROCESSING

DATA TRANSFORMATION



DATA AUGMENTATION TECHNIQUES

(CREATE NEW USEFUL DATA)

IMAGE

- CROPPING • SCALING
- PADDING • FLIPPING
- KERNAL FILTERS
- COLOR SPACE TRANSFORMATION

AUDIO

- CHANGE PITCH • SPEED
- ADD NOISE • SHIFT TIME

TEXT

- USE SYNONYMS • THESARUS
- RANDOM SWAP • DELETION
- MASKED LANGUAGE MODEL
- BACK TRANSLATION • NOISE

NUMERICAL DATA

- PCA • LOG TRANSFORMATIONS
- AUTO ENCODERS

EXPLORATORY DATA ANALYSIS

- GAIN MAXIMUM INSIGHT INTO A DATASET
- DISCOVER PATTERNS, TEST HYPOTHESIS
- SUMMARIZE MAIN CHARACTERISTICS OF DATA VISUALLY
- FIND ANAMOLIES, CHECK ASSUMPTIONS

UNIVARIATE • ONE VARIABLE OF INTEREST

MULTI VARIATE • MULTIPLE VARIABLES OF INTEREST

- ☐ CENTRAL TENDENCY (MEAN, MEDIAN, MODE)
- ☐ FREQUENCY DISTRIBUTION (BINS, HISTOGRAM/BOX)
- ☐ VARIANCE (QUARTILES, STDDEV)
- ☐ CO-LINEARITY
- ☐ IDENTIFY RELATIONSHIPS
- ☐ SPATIAL DEPENDENCIES
- ☐ OUTLIERS
- ☐ SKEWNESS
- ☐ KURTOSIS (PEAKS)
- ☐ CLUSTER ANALYSIS
- ☐ DIMENSIONALITY

#7 KNOW YOUR DATA [IN THE CONTEXT OF PET PROJECT]

WHERE TO FIND BEST PUBLIC DATASETS

FiveThirtyEight · BuzzFeed · kaggle · Quandl · DATA.GOV · UCI Machine Learning Repository · WIKIPEDIA The Free Encyclopedia

data.world · Socrata · reddit · Academic Torrents · VISUALDATA.IO

amazon web services · Google Dataset Search · Carnegie Mellon University Libraries · GitHub · awesome · FOURSQUARE · yelp

THE WORLD BANK · CDC Centers for Disease Control and Prevention · OSDC Open Science Data Cloud · LINCOLN LABORATORY Massachusetts Institute of Technology · SNAP · unicef · United States Census Bureau · UNdata A world of information · Berkeley DeepDrive

SPECIFIC*

- COMPUTER VISION
  - xVIEW
  - KINETICS-700
  - IMAGENET
  - VISUAL DATA
  - GOOGLE OPEN IMAGES
- SENTIMENT
  - SENTIMENT 140
  - IMDB
  - YELP
  - STANFORD TREEBANK
  - LEXICODER
  - TWITTER US AIRLINE
- TEXT
  - 20 NEWSGROUP
  - REUTERS NEWS
  - WORDNET
  - WIKI QA CORPUS
  - UCI SPAM BASE
- NLP
  - BLOGGER CORPUS
  - JEOPARDY
  - SMS SPAM
  - GUTENBURG
  - HOTSPOT QA
  - AMAZON REVIEWS
  - ROTTEN TOMATOES
  - GOOGLE BOOKS N-GRAMS
- AUDIO
  - LIBRI SPEECH
  - TIMIT
  - 2000 HUB5 ENGLISH
  - SPOKEI JIKI CORPUS
  - FREE MUSIC ARCHIVE (FMA)
  - MILLION SONG DATASET
- SELF-DRIVING
  - WAYMO OPEN DATA
  - BOSCH SMALL TRAFFIC
  - BERKELEY DEEP DRIVE
  - LARA TRAFFIC
  - WPI DATASETS

#8 DATA ....→ IN A NUT-SHELL [IN THE CONTEXT OF PET PROJECT]

PLANNING ---→ PRE-PROCESSING ---→ ANALYSIS

- → COLLECTION
- → DATA MERGING
- → GENERATE NEW DATA
- → FEATURE GENERATION
- → DATA SELECTION

- → PROFILING
- → FORMATTING
- → CLEANING
- → FILTERING
- → COMPLETION
- → SAMPLING
- → CORRECTION
- → STANDARDIZE
- → NORMALIZE
- → TRANSFORM
- → AUGMENT
- → VALIDATE

- → VISUALIZATION
- → CORRELATION
- → SKEWNESS
- → DISTRIBUTION
- → CLUSTERING
- → COHORTS
- → DIMENSIONALITY
- → OUTLIERS/NOISE
- → TRENDS/PATTERNS
- → STORY POINTS
- → INTERPRETATION
- → DOCUMENTATION

ADDITIONAL STEPS REAL-WORLD ENTERPRISE DATA SETUP

- AVAILABILITY (ON-PREM/CLOUD)
- AUTOMATED DATA WORKFLOWS (DATA LOADS)
- GOVERNANCE
- SECURITY
- INTEGRITY
- OWNERSHIP
- CERTIFICATION
- INFRA STRUCTURE
- STORAGE & COMPRESSION

MY ARTIFACT CHECKLIST

→ FOR DATA PHASE OF PET PROJECTS

- ☐ DATA SOURCES SUMMARY *
- ☐ PROCESS FLOWS/ DATA WORKFLOW
- ☐ DATA ASSUMPTIONS *
- ☐ DATA PREP * TECHNIQUES
- ☐ LIST OF ENRICHED/ NEW DATA/FEATURES *
- ☐ DATA QUALITY ISSUES LOG
- ☐ QUICK SUMMARY OF TOOLS/PACKAGES
- ☐ LIST OF DEPENDENCIES
- ☐ EXPLORATORY DATA ANALYSIS SUMMARY
- ☐ EDA * KEY FINDINGS
- ☐ CONCLUSION ON DATASET
- ☐ NEXT STEPS

* MAKE A POINT TO DISCUSS YOUR DATA & ANALYSIS TECHNIQUES IN YOUR INTERVIEWS *

#9 DESIGN YOUR ARCHITECTURE

HIGH-LEVEL (HLD)
- OVERALL SYSTEM DESIGN, MAIN COMPONENTS AND INTERFACES
- RELATION BETWEEN MODULES AND FUNCTIONS
- OVERALL FLOW, DATA FLOW, PROCESS FLOW, FEATURES FLOW

LOW-LEVEL (LLD)
- DETAILED-LEVEL EXTENSION OF HIGH-LEVEL DESIGN
- COMPONENT LEVEL DESIGN PROCESS
- METHODS, CLASSES, FUNCTIONS, DEPENDENCIES, PACKAGES ETC.

DESIGN/SOLUTION ARCHITECTURE

DATA | SYSTEM | MODEL | ANALYTICS | INTELLIGENCE

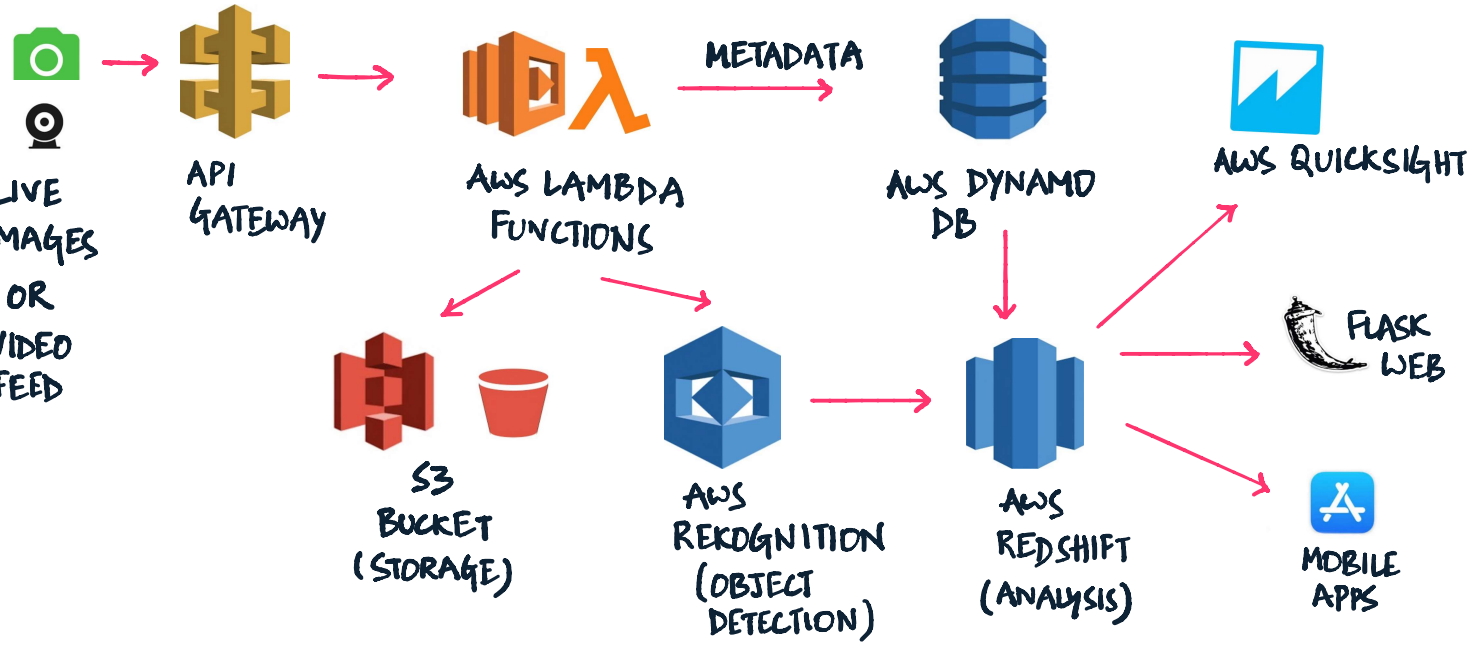
TRANSLATES TO

IMPLEMENTATION

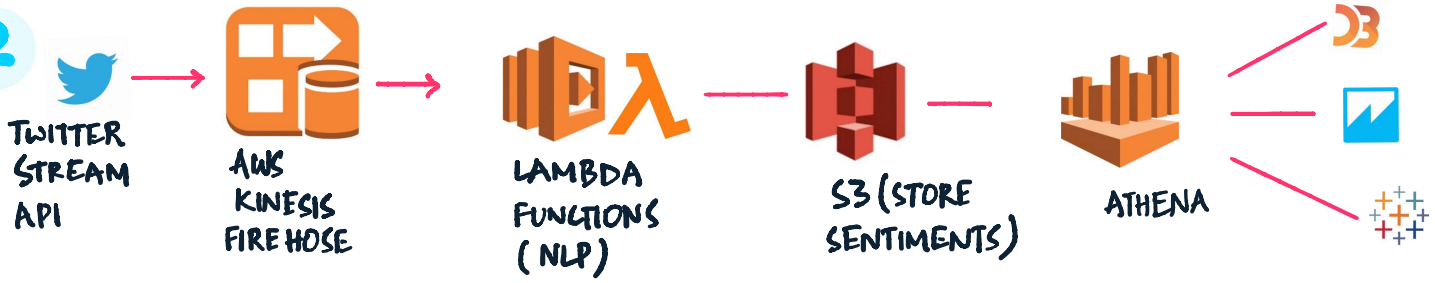
PRODUCT | SERVICE | INSIGHTS | PREDICTIONS | PROCESS

EXAMPLE SOLUTION DESIGNS (USING AWS)

1



2



* DISCUSS YOUR DESIGN & ARCHITECTURE IN YOUR INTERVIEWS *

RECAP + UPCOMING ❤

Pet-Projects Post Series

Post 0 - Pet Project Intro

Post 1 - 3 Ingredients of a successful pet-project

Post 2 - Define Problem

Post 3 - Choose a Methodology

Post 4 - Gather Knowledge

Post 5 - Data Essentials

- Post 5.1 - Data Essentials I
- Post 5.2 - Data Essentials II
- Post 5.3 - Know your Data
- Post 5.4 - Know your Data - Public Datasets
- Post 5.5 - Data in a Nut-shell

Post 6 - Implementation

- Post 6.1 - Design Architecture - Using AWS examples (** we are here **)
- Post 6.2 - Dig Deeper into Design (Architecture)
- Post 6.3 - Model Know-how's - Essentials 101
- Post 6.4 - Model Engineering - Compute (GPU/TPU), ML Frameworks, AutoML
- Post 6.5 - Feature Engineering (An Art)
- Post 6.6 - Coding Essentials
- Post 6.7 - Model Artifacts / My checklist
- Post 6.8 - Deployment / Scale / CI-CD (Continuous Integration / Deployment pipelines)
- Post 6.9 - Model Inferences / Model Serving / End Points / Drift / Monitoring / Logging

Post 7 - Deliver Insights / Inferences - Visualizations / End User Interfaces - Interaction Touch points

BONUS - Bringing it all together - Final Pet-project Report* (Showcase / Git / Portfolio)

#10 ❤ SYSTEMS

PARTS · STEPS · COMPONENTS · MODULES · PROCESS · INTERFACES · ARCHITECTURE · INTERACTIONS

HOLISTIC SYSTEM ARCHITECTURE [TOP-DOWN]

DATA

DATA
- COLLECT
- ACQUIRE (BUY)
- GENERATE
- HARVEST
- AUGMENT

INGESTION & INTEGRATION
- EXTRACT
- TRANSFORM
- LOAD
- METADATA
- RULES
- BATCH
- REAL-TIME
- STREAM
- TIME-BASED
- REPLICATION

AVAILABILITY & DELIVERY PLATFORMS
- RELATIONAL
  - EDW (WAREHOUSES)
  - DATA MARTS
  - IN-MEMORY DB
- DISTRIBUTED
  - DATA LAKES
  - NO-SQL
  - HADOOP (LOW-COST)
  - CLOUD

DATA GOVERNANCE

QUALITY · POLICIES · SECURITY · PROFILING · BACKUP · ARCHIVE
STANDARDS · AUDITING · AUTHENTICATION · ROLES · RETRIEVAL · PURGING

DATA PREP & ANALYSIS

→ DATA PRE-PROCESSING → DATA LABELING → DATA SEGREGATION → FEATURE EXTRACTION → EDA (ANALYSIS)
- HYPOTHESIS
- QUESTIONS
- VIZ

→ DATA ENRICHMENT → DATA SELECTION

DEV TEST VALIDATE

FEEDBACK

MODEL ENGINEERING — DEPLOY — MONITOR

EXPERIMENTS

MODEL 1 · MODEL 2 · MODEL 3 · MODEL N

DATA
- FEATURES
- TRAIN
- BASELINE
- CRITERIA
- AUTO ML

CANDIDATE EVALUATION
- INFERENCE
- TUNING
- HYPER PARAMS

SAVE –
- MODEL
- WEIGHTS
- PARAMS
- EXPERIMENTS
- VERSIONS

PACKAGING
JAR · PKL · ONNX · HDF5

CLOUD
- COMPUTE
- DOCKER
- LOAD WEIGHTS
- TAGGING

DRIFT
- LOGGING
- RE-RUNS
- PERFORMANCE
- REGISTRY
- VERSIONING
- DECAY

MODEL SERVING (ML OPS)

REST END POINTS · WEB SERVICES · STORE INFERENCES · API (HTTP & POST)
DOCKER · KUBERNETES (K8S) · KUBEFLOW · CLOUD/ STORAGE

INTELLIGENCE/INSIGHTS

SELF-SERVICE REAL-TIME INSIGHTS (APPS) · DISCOVERY REPORTING · ANALYTICS MOBILE · BUSINESS INTELLIGENCE WEB (JS, FLASK, REACT) · ADVANCED INSIGHTS APPLICATIONS

#11 MODEL ESSENTIALS 101

G IS YOUR BEST FRIEND

ARTIFACT TRAINED TO DETECT CERTAIN TYPE OF PATTERNS IN DATA

ALGORITHMS THAT IMPROVE AUTOMATICALLY THROUGH EXPERIENCE

KEY ELEMENTS OF ML

- REPRESENTATION — KNOWLEDGE
- OPTIMIZATION — SEARCH PROCESS
- EVALUATION — CANDIDATE MODELS (HYPOTHESIS/RECIPES)

MODEL

CHOOSE MODEL FAMILY
- SIZE
- TYPE OF PROBLEM
- RANDOMNESS
- PARAMETERS

EXPLAINABILITY
REPRODUCE

MODEL PARAMETERS
- FEATURES
- LEARNED DIRECTLY FROM TRAINING

OPTIMIZERS
- ADAM
- GRADIENT DESCENT
- MOMENTUM
- RMSPROP
- SGD
- ADAGRAD

HYPER PARAMETERS
- LEARNING RATE
- EPOCHS / PASSES
- DECIDED BEFORE LEARNING

ACTIVATION FUNCTIONS
- RELU
- TANH
- GAUSSIAN
- SIGMOID
- IDENTITY
- BINARY STEP

TUNING
- CROSS-VALIDATION (K-FOLD)
- GRID SEARCH

ETHICS
BIAS

EVALUATION METRICS (BASED ON MODEL FAMILY)
- AUC — AREA UNDER ROC CURVE
- MICRO AVERAGED F1 SCORE
- ACCURACY, PRECISION, RECALL
- LOG LOSS / ERROR
- MSE, RMSE, MAE, R-SQUARED
- CUT-OFF

OVER-FITTING
- DROP-OUT
- REGULARIZATION (L1 & L2)
- DATA SHUFFLING

PREDICTIONS
- REAL-TIME
- BATCH
- ONE-OFF / LOCAL

#11 MODEL ESSENTIALS 101*

- ARTIFACT TRAINED TO DETECT CERTAIN TYPE OF PATTERNS IN DATA
- ALGORITHMS THAT IMPROVE AUTOMATICALLY THROUGH EXPERIENCE

4 TYPES OF AI

DL ML AI

- REACTIVE MACHINES
  - PERCEIVE THE WORLD DIRECTLY AND ACT ON WHAT IT SEES.
  - NO ABILITY TO FORM MEMORIES OR USE PAST EXPERIENCES
- LIMITED MEMORY
  - TYPE-II MACHINES THAT CAN LOOK INTO PAST
  - MEMORIES AREN'T SAVED AS PART OF LIBRARY OF EXPERIENCE
- THEORY OF MIND
  - THOUGHTS, EMOTIONS, CO-EXISTANCE THAT EFFECT THEIR OWN BEHAVIOR
  - TYPE-III SENTIENT MACHINES (FUTURISTIC)
- SELF-AWARENESS
  - FORM REPRESENTATIONS ABOUT THEMSELVES
  - CONSCIOUS BEINGS — AWARENESS OF THEIR STATE & OTHERS FEELINGS

LEARNING

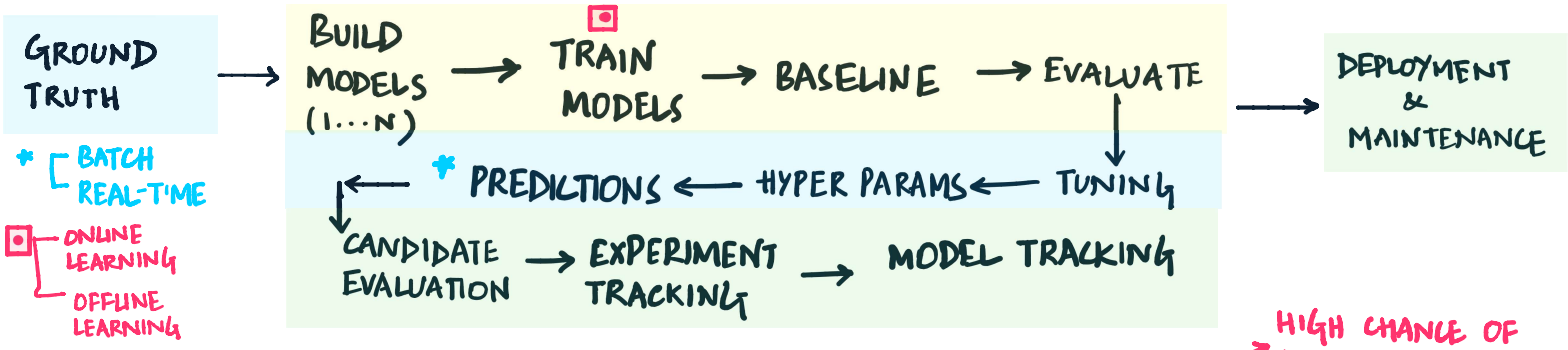
- TECHNIQUES
  - MULTI-TASK LEARNING
  - ACTIVE LEARNING
  - TRANSFER LEARNING
  - ENSEMBLE LEARNING
- PROBLEMS
  - SUPERVISED
  - UNSUPERVISED
  - REINFORCEMENT
- HYBRID
  - SEMI-SUPERVISED
  - SELF-SUPERVISED
  - MULTI-INSTANCE
- STATISTICAL INFERENCE
  - INDUCTIVE LEARNING
  - DEDUCTIVE LEARNING
  - TRANSDUCTIVE LEARNING

4 TYPES OF AI

DL ML AI

- REACTIVE MACHINES
  - NO MEMORY / PAST EXPERIENCE
- LIMITED MEMORY*
  - CAN LOOK INTO PAST.
- THEORY OF MIND
  - THOUGHTS, BEHAVIOR
- SELF-AWARENESS
  - CONSCIOUS · SENTIENT

#12 MODEL* ENGINEERING* AT SCALE



BASELINE

- TEST ERROR ... REDUCE
- VALIDATION SET ERROR
- TRAINING ERROR
- AUTO ML
- HUMAN ERROR
- BAYESIAN ERROR

HYPER PARAM SEARCH*

HIGH CHANCE OF SUCCESS

AUTOMATED
HYPERAS
SIGOPT
W&B

- LEARNING RATE*
- MODEL SIZE
- NUMBER OF EPOCHS*
- WEIGHTS
- MODEL DEPTH*
- LAYER PARAMS
- OPTIMIZER CHOICE
- BATCH SIZE
- LOSS FUNCTION*
- REGULARIZATION WEIGHTS

RESOURCES (COMPUTE)

- ON-PREM (CUDA - GPU/LAMBDA)
- CLOUD* (AWS, GCP, AZURE)
- DISTRIBUTED TRAINING (MULTIPLE NODES, GPUS, CPU)
- PARALLEL DATA
- MODEL PARALLESIM

EXPERIMENTS

- RECORD & TRACK ALL EXPERIMENTS
- SYNCH & MANAGE (ML FLOW, TENSORBOARD)

* MAKE A POINT TO DISCUSS ABOUT SCALING IN YOUR INTERVIEWS *

IN THE CONTEXT OF PET PROJECTS*

- [ ] GET BASELINE METRICS
- [ ] AUTO ML FRAMEWORKS
- [ ] BUILD MULTIPLE MODELS FROM DIFFERENT FAMILIES
- [ ] EVALUATE USING STANDARD METHODS (PRECISION, RECALL, CONFUSION MATRIX, LOSS ETC.)
- [ ] BIAS-VARIANCE TRADE OFF & DISTRIBUTION SHIFT
- [ ] ERROR ANALYSIS
- [ ] FEATURE SELECTION & ENGINEERING
- [ ] HYPER PARAMETER OPTIMIZATION
- [ ] ENSEMBLES (BAGGING BOOSTING, STACKING)
- [ ] PICK THE BEST MODEL BASED ON YOUR SUCCESS CRITERIA
- [ ] USE COLAB/GPU SUPPORTED ENVIRONMENT

#13 FEATURE ENGINEERING 101

G IS YOUR BEST FRIEND

WHY?

- MAKES THE MODEL UNDERSTAND THE INTRICACIES OF THE PROBLEM
- REPRESENTS THE MOST MEANINGFUL ATTRIBUTES/RELATIONS/LEARNING

DATA → FEATURE EXTRACTION ↔ FEATURE GENERATION
FEATURE CONSTRUCTION ↔ FEATURE LEARNING
FEATURE IMPORTANCE ↔ FEATURE SELECTION
FEATURE STABILITY ↔ FEATURE EVALUATION
→ MODEL

FEATURES

TYPES

- IMAGE
- TEXT
- CATEGORICAL
- NUMERICAL
- SPATIAL
- TEMPORAL

TYPES

- INTERACTION FEATURES
- FEATURE REPRESENTATION
- INDICATOR VARIABLES
- EXTERNAL DATA

TECHNIQUES

- LABEL ENCODING*
- UNIVARIATE CREDIBILITY
- GROUPING CLASSES
- STANDARDIZATION
- SUMS, DIFF, PRODUCT, RATIOS*
- TOPIC MODELING
- DATE & TIME HANDLING*
- COUNT & FREQUENCY*
- VARIABLE TRANSFORMATION
- MISSING VALUE IMPUTATION*
- OUTLIERS*
- ONE-HOT ENCODING
- ROUNDING
- FEATURE SPLIT*
- SCALING
- BINNING
- SCALING*
- QUANTIZATION
- BINARIZATION

TO BE CONTINUED XX

#14 FEATURE ENGINEERING DEEP DIVE

TECHNIQUES

G IS YOUR BEST FRIEND

1 MISSING DATA IMPUTATION

- COMPLETE CASE ANALYSIS
- MEAN/MEDIAN/MODE
- RANDOM SAMPLE
- MISSING VALUE INDICATOR
- REPLACEMENT BY ARBITRARY VALUE
- MULTIVARIATE IMPUTATION

2 CATEGORICAL ENCODING

- ONE-HOT ENCODING
- COUNT & FREQUENCY
- ORDINAL ENCODING
- TARGET ENCODING
- WEIGHT OF EVIDENCE
- RARE LABEL ENCODING
- BASE N
- FEATURE HASHING

3 OUTLIERS

- REMOVAL
- TREAT AS NAN
- CAPPING
- WINSORIZATION (TOP/BOTTOM/ZERO CODING)

4 IMAGES

- PIXELS
- LINES
- EDGES
- THRESHOLDS

5 VARIABLE TRANSFORMATION

- BOX-COX
- LOG - $\log(x)$
- RECIPROCAL $1/x$
- SQRT $\sqrt{}$
- EXPONENTIAL
- YEO-JOHNSON

6 FEATURE SCALING

- STANDARDIZATION
- MIN-MAX SCALING
- MEAN SCALING
- MAX ABSOLUTE SCALING
- UNIT-NORM SCALING

7 TEXT

- TOPIC EXTRACTION
- BAG OF WORDS
- TFI-DF
- N-GRAMS
- WORD2VEC

8 DISCRETISATION

- EQUAL FREQUENCY
- EQUAL LENGTH
- WITH TREES
- WITH CHIMERGE

9 DATE & TIME

- DAYS WEEKS
- MONTHS QUARTERS
- YEARS
- WEEKDAYS
- LEAP YEARS
- TIME DELTA
- DAY OF MONTH
- CALENDAR/FISCAL
- TIMESTAMP/SPECIAL

10 FEATURE CREATION (GROUP OF FEATURES)

- SUM
- MINUS
- PRODUCT
- MEAN
- MIN
- QUOTIENT
- ABS
- RATIOS

11 FEATURE STABILITY

- PEARSON'S COEFFICIENT
- JACCARD'S INDEX
- SYMMETRICAL UNCERTAINITY
- SPEARMAN'S RANK
- CANBERRA DISTANCE

ORIGINAL WORK BY SOLEDAD GALLI

#15 WRITE WORLD-CLASS CODE*

FULLY FUNCTIONAL

- ZERO TO MIN BUGS
- EXCEPTION HANDLING
- SOLVES THE PURPOSE
- SINGLE POINT OF FOCUS
- SIMPLIFY, SPLIT, KISS

READABILITY

- INDENTATION
- STYLE GUIDES
- STANDARDS
- NAMING
- LINTING
- CODING PATTERNS
- TOOLS/IDE/AUTOMATION

CLARITY OF THOUGHT/FLOW

- WORKFLOWS
- CONNECT THE DOTS
- SYSTEMS WORKING COHESIVELY

SELF-DOCUMENTING

- MINIMIZE OVERHEAD
- STANDARDIZE COMMENTS
- EXPRESSIVE

MODULARITY (PLUG & PLAY)

- SELF-CONTAINED

REFACTORIZATION

MAINTAINABILITY

- DECODABLE
- UNDERSTANDABLE
- WELL-STRUCTURED
- LESS DEPENDENCIES
- LONG-LIFE

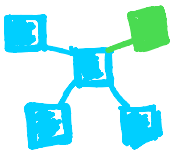
DEBUGGING/LOGGING

- LOGGER MODULES
- CHECKPOINTS
- COMMON ERRORS/ KNOWN ISSUES

TESTABILITY

- UNIT TESTING
- INTEGRATION
- REGRESSION
- SMOKE TESTS

VERSIONING & BACKUPS (LOCAL & CLOUD)

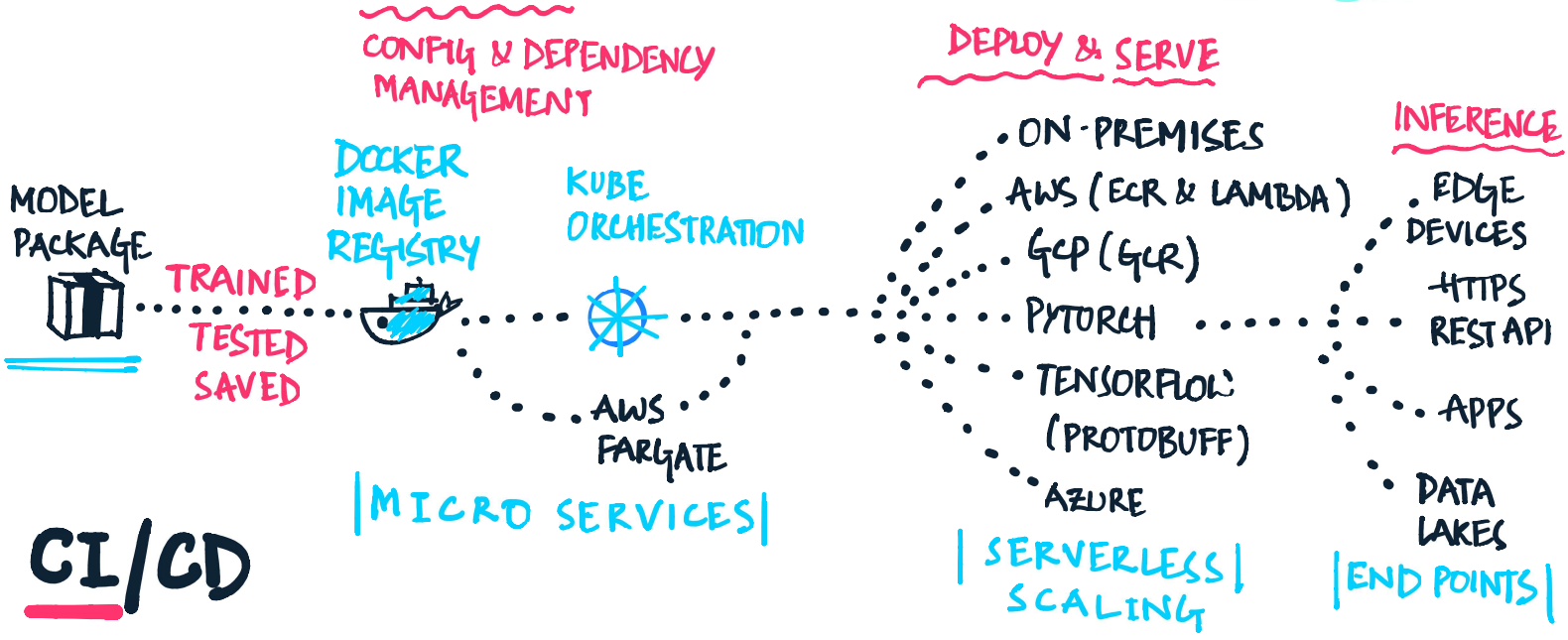


EXTENSIBILITY

CODE

THINK OF ART. BEAUTY, FLOW, SERENITY, CALMNESS, CANDID. COMPOSED, PLEASURE & HEAVEN — THAT'S HOW YOUR CODE SHOULD FEEL — SPECIAL*

#16 DEPLOY FOR SCALE ..... IS THE "MAGIC BULLET"



CI/CD

DATA

- SKEW TESTS
- QUALITY TESTS
- AVAILABILITY TESTS
- FEATURE TESTS

MODEL BUILD TESTING [AUTOMATED*]

- UNIT TESTING
- INTEGRATION TESTING
- INFRASTRUCTURE TESTING
- RELIABILITY TESTING
- PERFORMANCE TESTING
- SYSTEM TESTING
- EVALUATION TESTS
- BIAS/FAIRNESS

CONTINOUS INTEGRATION

DEV → QA → STAGING → PROD*

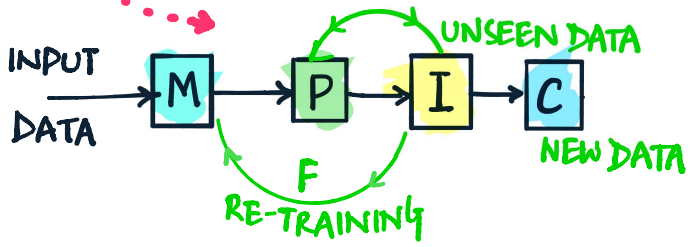
- PACKAGING
- DEPENDENCIES
- VERSIONING
- IMAGE/CONFIG CONTAINERS
- ORCHESTRATORS
- SCHEDULING / EVENT TRIGGERS
- REPRODUCABILITY TESTING
- TEST SCORING (READINESS)
- SANITY CHECK POINTS
- AUTO SCALING
- FAULT TOLERANCE / FALLBACK

CONTINOUS DELIVERY

KEYWORDS

SERVERLESS CONTAINERS SERVING REGISTRY LOAD BALANCING VERSIONS COST MICROSERVICES ORCHESTRATION REFITTING ACCELARATORS TAGGING SCORING

#17 SERVE MODEL*



M - MODEL P - PREDICTIONS F - FEEDBACK
I - INSIGHTS C - CONSUMPTION

TYPES OF SERVING (ML OPS*)

MODEL-AS-A-SERVICE

ACCESSED AS A CALLABLE SERVICE
REST API / WSDL

MODEL-AS-A-DEPENDENCY

INTEGRATED WITHIN A SOFTWARE APPLICATION.
PACKAGE DEPENDENCIES

PRECOMPUTE

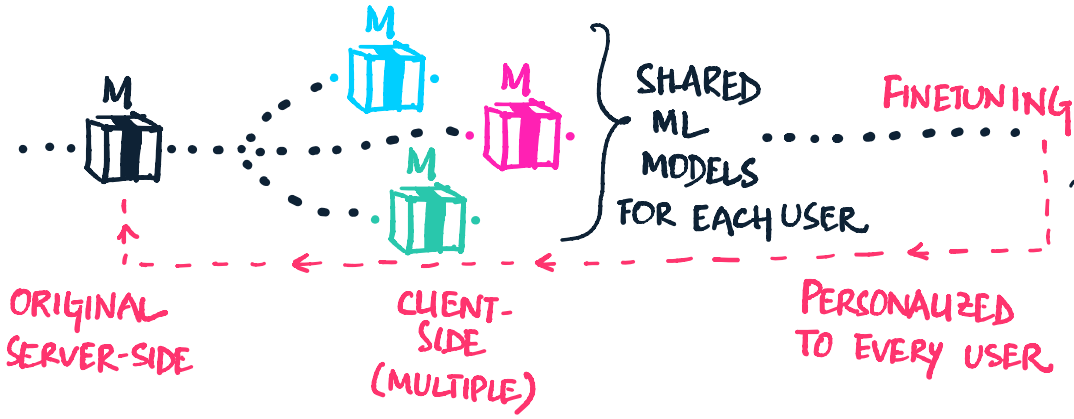
STORE & QUERY THE DATABASE FOR PREDICTIONS
(AWS REDSHIFT, BIG QUERY)

MODEL-ON-DEMAND (EVENT/TRIGGER DRIVEN)

MESSAGING - BROKER ARCHITECTURE

REQUEST ROUTING AT RUN-TIME

HYBRID (FEDERATED LEARNING)



VERY POWERFUL WAY OF LEARNING WITHOUT LOT OF DATA AND ONLY SPECIFIC TO THAT USER

* LEARN TO FIT YOUR PET-PROJECTS IN ONE OF THE ABOVE. INTERVIEW TOUCHPOINT*

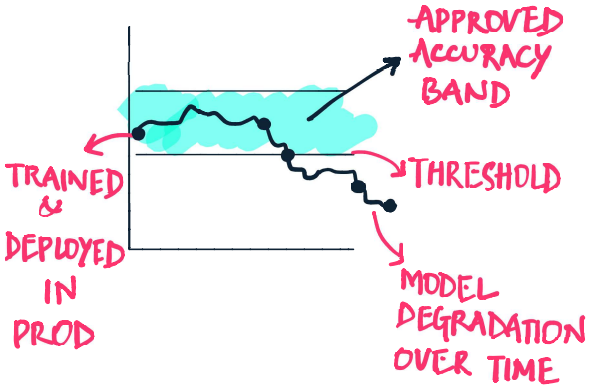
#18 CONTROL DRIFT

CHANGE IN THE DIRECTION OF THE PERFORMANCE OF MODEL

DEGRADATION OVER TIME

TYPES OF DRIFT

- CONCEPTUAL / CONTEXTUAL
  - → TRAINING DATA NO LONGER REPRESENT THE PREDICTIONS OR CORRECT LEARNING
  - → CHANGE OF CONTEXT / APPLICABILITY OF PREVIOUS TRAINING RULES
- DATA
  - → CHANGE IN THE BEHAVIOR OF DATA (AS MORE DATA GETS ADDED)
  - → DATA QUALITY ISSUES
  - → CHANGE OF ENVIRONMENT, POLICIES OR ANYTHING THAT IMPACTS DATA



- FEATURE / COVARIATE DRIFT
  - → CHANGE TO THE UNDERLYING FEATURES
  - → NUMERICAL VARIANCE
  - → DISTRIBUTION SHIFT
- DUAL DRIFT
  [COMBINATION OF ABOVE]

MODEL MONITORING [DETECT DRIFT]

- HYPOTHESIS TESTS
- CHI-SQUARED, P-VALUE, T-TESTS
- DISTANCE MEASURES
- KULLBACK-LIEBNER (KL) MEASURE
- JENSEN-SHANNON DIVERGENCE
- COMPARE WITH GROUND TRUTH
- SUMMARY STATS & INPUT DISTRIBUTIONS
- ROC CURVE, MANN-WHITNEY U & GINI
- DATA IN TRAIN VS. DATA IN PROD

MODEL MAINTENANCE [KEEP IT GOING]

- AUDITING
- LOGGING
- PROFILING
- ALERTING
- EVENT/TRIGGERS
- A/B TESTS
- SPLIT TESTS

* FAMOUS OPEN-SOURCE MONITORING METRICS TOOLS ARE PROMETHEUS & GRAFANA.
CLOUD OFFERING AWS SAGEMAKER MODEL MONITOR

#19 DELIVER INTELLIGENCE

ENDPOINTS

MOBILE/ SMART INTERFACES

IOT SENSORS EDGE DEVICES DRONES

API/REST CALLABLE WEBSERVICE

ENTERPRISE CLOUD

QUEUES/EVENTS

WEB/NATIVE APPLICATIONS

INSIGHTS

EDGE/ CLIENT SIDE

ENTERPRISE DATAWAREHOUSE

DASHBOARDS

ANALYSIS

REAL TIME

SCORECARDS

PREDICTIONS

REPORTS

MODEL

BATCH

TRENDS

EXECUTIVE OVERVIEWS

OFFLINE

DATABASE/ DATAMARTS (SQL/NO-SQL)

HISTORICAL PATTERNS

ONE-OFF LOCAL

MOBILE APPS

INFERENCES/ PREDICTIONS

ALERTS THRESHOLDS

DATALAKES

INTEGRATION INTO NATIVE APPS

DOCUMENTS/MEDIA

CONVERSATIONS

CHATBOTS (NLP AGENTS)

BUSINESS PROCESSES

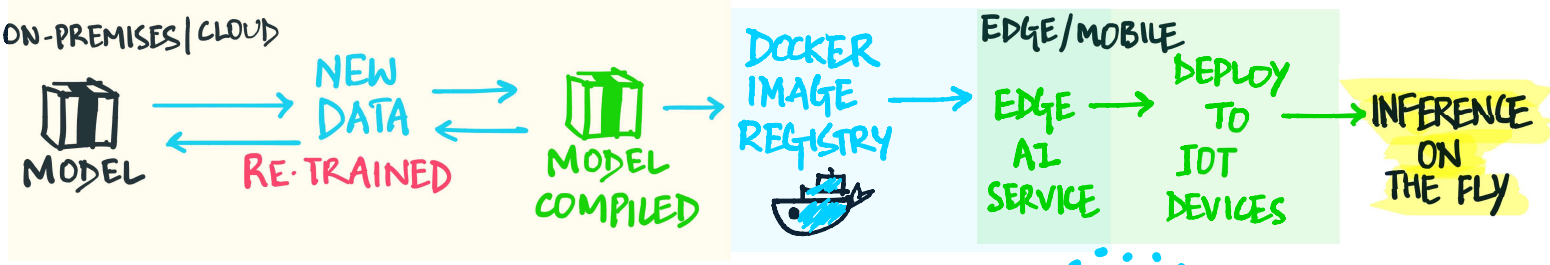
AR/VR FUTURISTIC

* BUILD "INTELLIGENCE" END USER INTERFACE AS PART OF YOUR PET-PROJECT *

#20 GO LITE (EDGE)

DEPLOY YOUR PROJECT ON EDGE/MOBILE FOR ON-THE FLY PREDICTIONS

- INEXPENSIVE
- CONNECTIVITY
- LOW POWER DEVICES
- NO LATENCY
- PRIVACY & SECURE
- LIMITED PROCESSING POWER



EDGE AI FRAMEWORKS

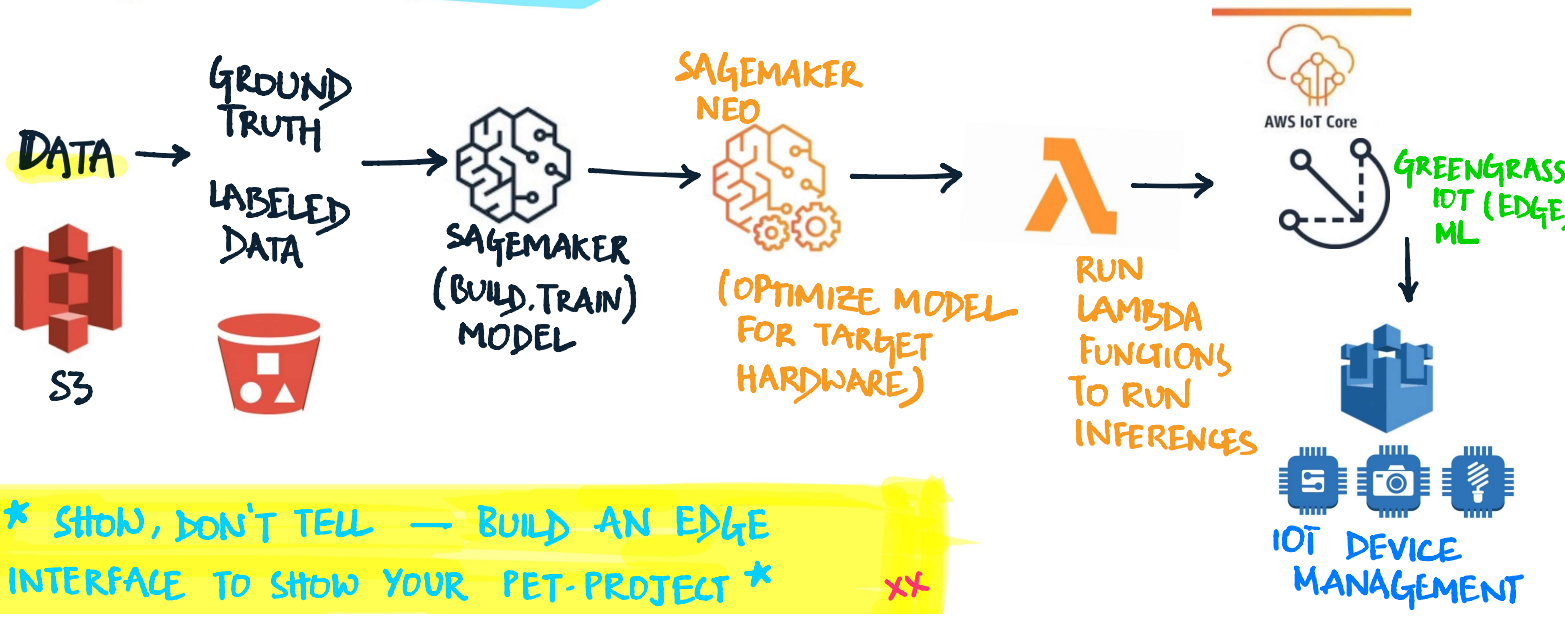
TensorFlow Lite · ML Kit for Firebase · PYTORCH MOBILE · APPLE CORE ML · APACHE mxnet · Microsoft Embedded Learning Library

JETSON NVIDIA · JOS ANDROID · MICRO CONTROLLERS · RASPBERRY PI ARDUINO · ARM ETHOS INTEL

TF WORKFLOW

PRE-TRAINED MODELS (TF, KERAS) → TENSORFLOW LITE CONVERTER (LITE FORMAT) → TENSORFLOW INTERPRETER (IOT FRIENDLY) WITH API's → MODEL OPTIMIZATION TOOLKIT (QUANTIZATION) → PREDICTIONS/ INFERENCES ON IOT/ EDGE DEVICES

AWS WORKFLOW



* SHOW, DON'T TELL — BUILD AN EDGE INTERFACE TO SHOW YOUR PET-PROJECT *

XX

#21 GO WEB (APPS, API's)

GIVE A FACE TO YOUR AI/ML PROJECT
BUILD A USABLE WEB APP INTERFACE

START HERE

M

TF
PYTORCH
PYTHON
KERAS
SPARK...

TRAINED

- ARCHITECTURE + WEIGHTS + OPTIMIZED STATE (PICKLE, HDF5, PROTOBUFF...)
- ONLY WEIGHTS — JSON / YAML

DOCKER/KUBERNETES — CONFIGURE IMAGE REGISTRY

S E R V E

WEB FRAMEWORKS

- FLASK*
- DJANGO
- REACT/ANGULAR
- GOOGLE APP ENGINE
- HTML+CSS+JS
- BOOTSTRAP*
- NODE.JS
- TF.JS

GET PUT POST DELETE

FLASK

REST API

CREATE CALLABLE API

2 MAIN COMPONENTS

- APP.PY*
- MAIN/INDEX.HTML

BOOTSTRAP/HTML UI TEMPLATE

POSTMAN

(TEST THE END POINT WHICH IS EXPOSED VIA FLASK API)

* STREAMLIT / DASH

BUILD WEB APP / VISUALIZATION APP (IGNORE IF GOING FLASK + BOOTSTRAP UI)

DEPLOY TO HEROKU USING GEMFURY OR *GITHUB* REPO OR GUNICORN

PROCFILE (FOR BINDING)

HEROKU

DYNO(512MB)

REQUIREMENTS.TXT (FREEZE PACKAGES)

(PLATFORM AS A SERVICE)

SERVE PREDICTIONS

END-USER (BROWSER, MOBILE, APP)