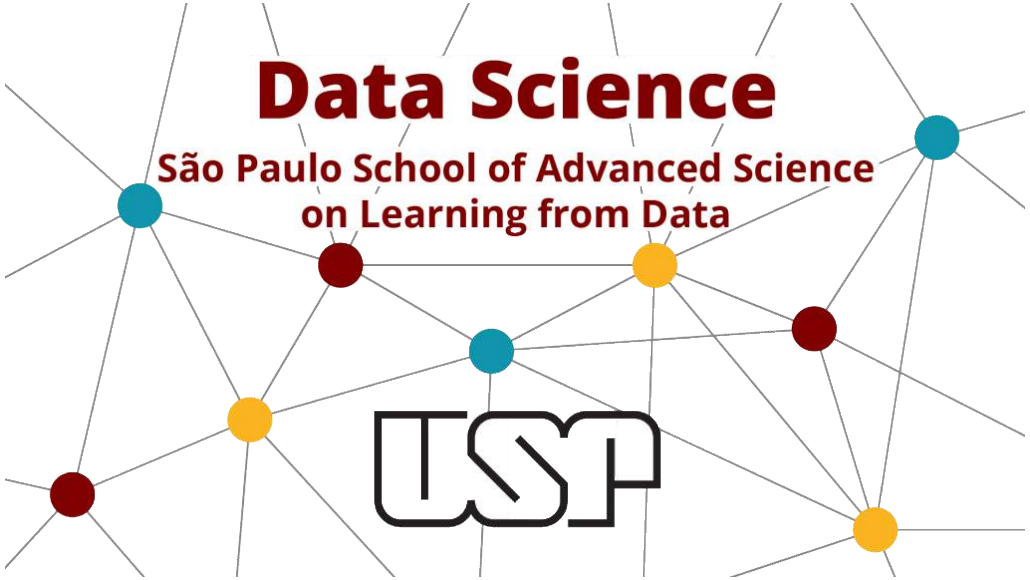


FAPESP

Machine Learning Life-Cycle

Fabio Porto

(fporto@lncc.br),

LNCC – MCTIC

DEXL Lab (http://dexl.lncc.br)

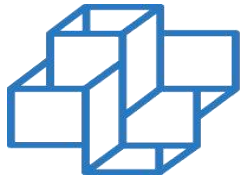
Laboratório Nacional de Computação Científica

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES

BRASIL GOVERNO FEDERAL

DEXL LAB
DATA EXTREME LAB

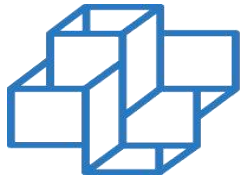
Agenda



- DEXL Presentation
- Intro: Need for Data and Model Management in ML
- Data Preparation
- Model Construction
- Model Serving
- Final Comments

DEXL LAB
DATA EXTREME LAB

Data EXtreme Lab (DEXL)



- Mission:
  - To excel in research, development and Innovation in Data Science, AI and Big Data
- Team
  - Coordinator: Fabio Porto
  - Researchers
    - Artur Ziviani
    - Luiz Gadelha
    - Marcel Pedroso (visiting - FIOCRUZ)
  - PosDocs
    - Yania Molina Souto
    - Douglas de Oliveira
    - Klaus Wehmuth
    - Felipe S. Abrahão
  - PhDs
    - Hermano L. Lustosa
    - Daniel Gaspar
    - Rocio Milagros Z. Coz
    - Daniel Ramos da Silva
    - Maria Luiza Mondelli
    - Matheus Ribeiro F, de Mendonça
    - Yasmin Cortes Martins
    - Claudio Tenório
  - MSc
    - Rafael Pereira Silva
    - Henrique Matheus F. da Silva
    - Gustavo Carnivali
    - Haron C. Fantecele
    - Juliana Z. G. Mascarenhas
  - Technical Support
    - Adolfo Simões
    - Carlos Cardoso
    - Enver Choque Cayo
    - João Guilherme N. Rittmeyer
  - Trainees
    - Raquel Junqueira
    - Andre Demori
    - Viviane Matioli

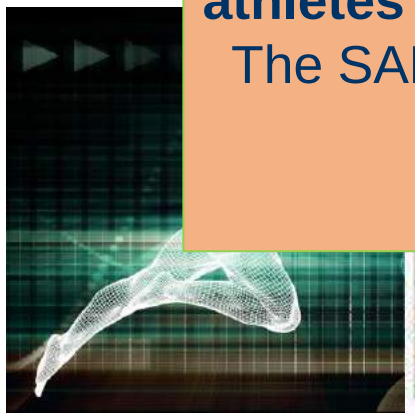
DEXL LAB
DATA EXTREME LAB

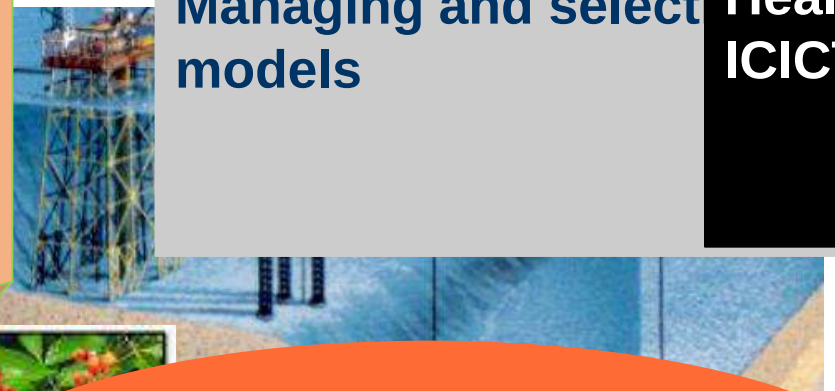
Ongoing Projects

Follow-Up and Prediction of athletes performance
The SAHA system

Gypscie System for Managing and select models

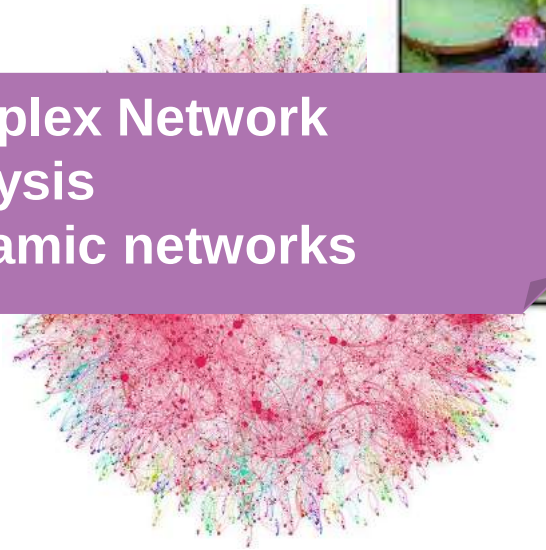
Big Data Platform Health(PcDaS) – FIOCRUZ ICICT





DEXL group

Complex Network Analysis
dynamic networks



Detection of ecology niches; Reproducibility of experiments– Jardim Botânico

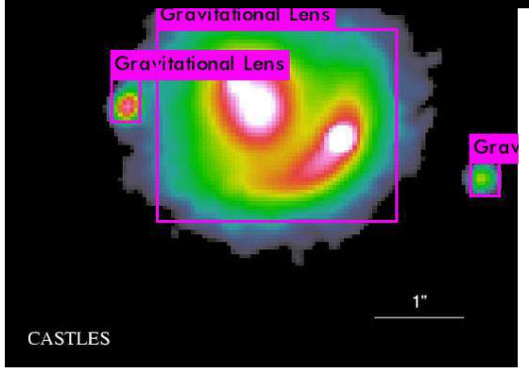
Search for Gravitational lensing: Apache Spark + DES/SDSS - LineA

DEXL LAB
DATA EXTREME LAB

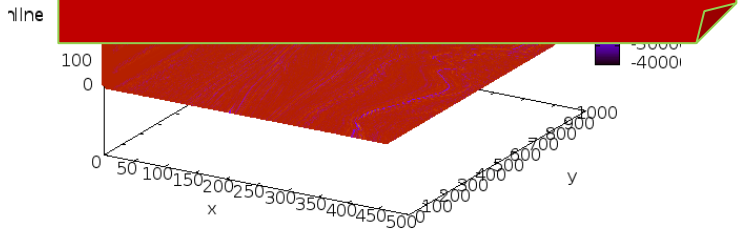
Deep Learning Models

Predict streets and roads

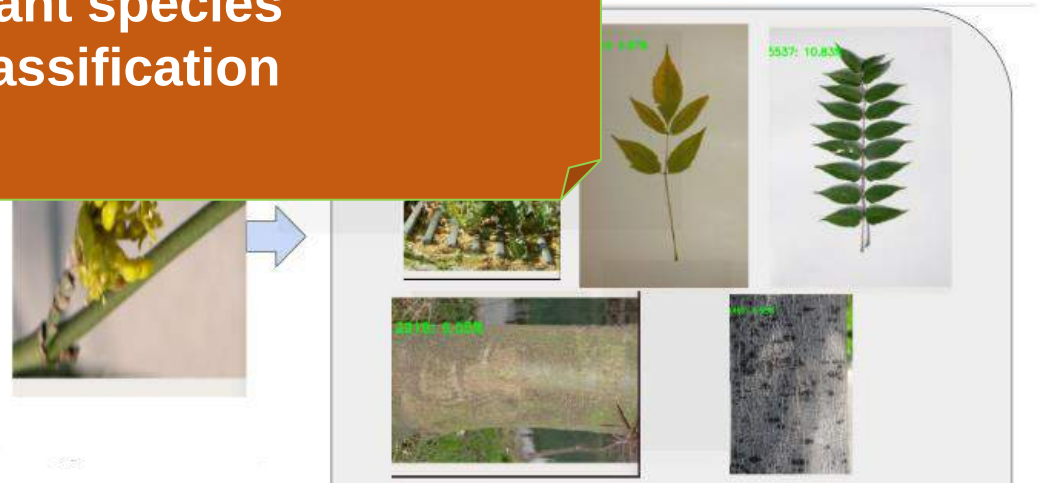
Automatic Detection of lensing effect

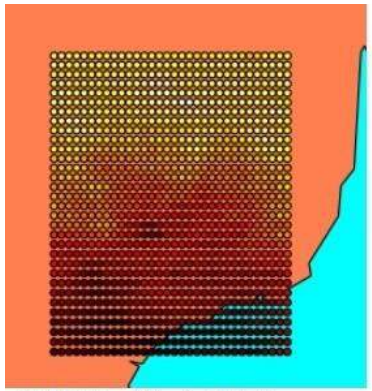


Prediction of amplitude in borders of seismic cube



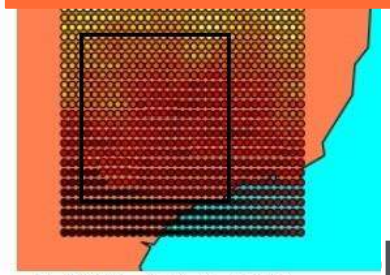
Plant species classification





a) MAE médio de 3.503 mm

Prediction of temperature and rainfall



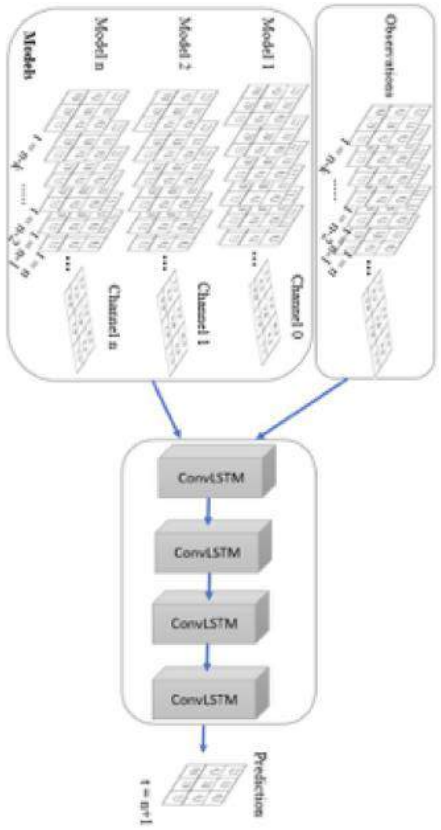
b) MAE médio de 3.917 mm

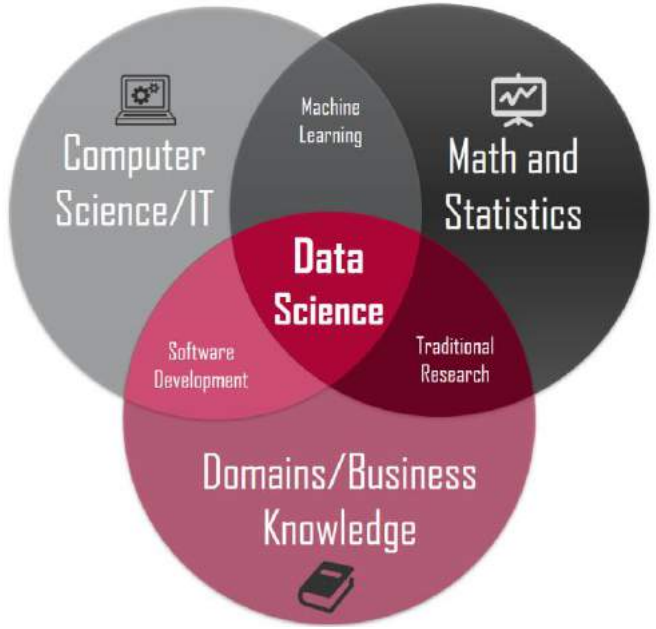
EXL LAB
DATA EXTREME LAB

ARTIFICIAL INTELLIGENCE

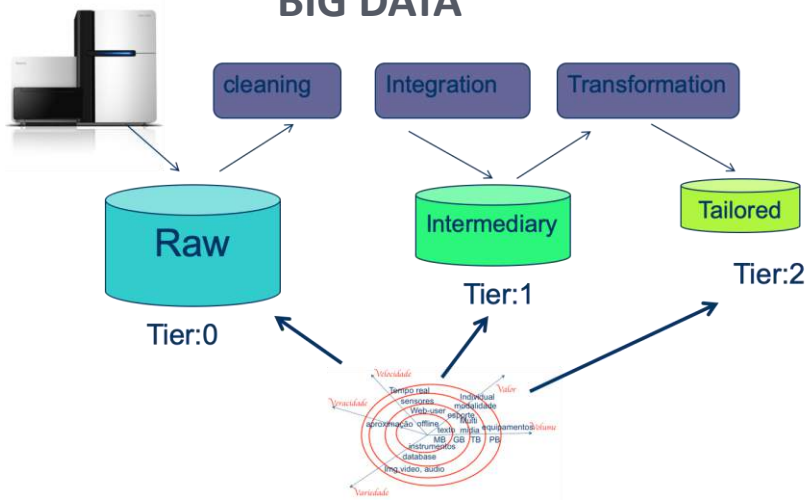
DATA SCIENCE

DEXL group





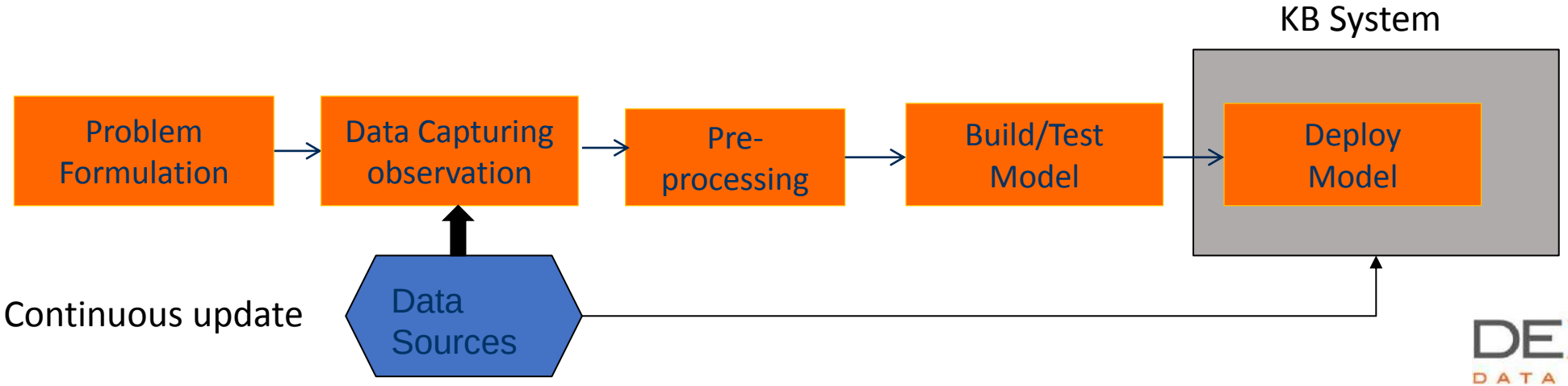
BIG DATA



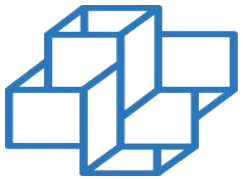
DEXL LAB
DATA EXTREME LAB

Introduction

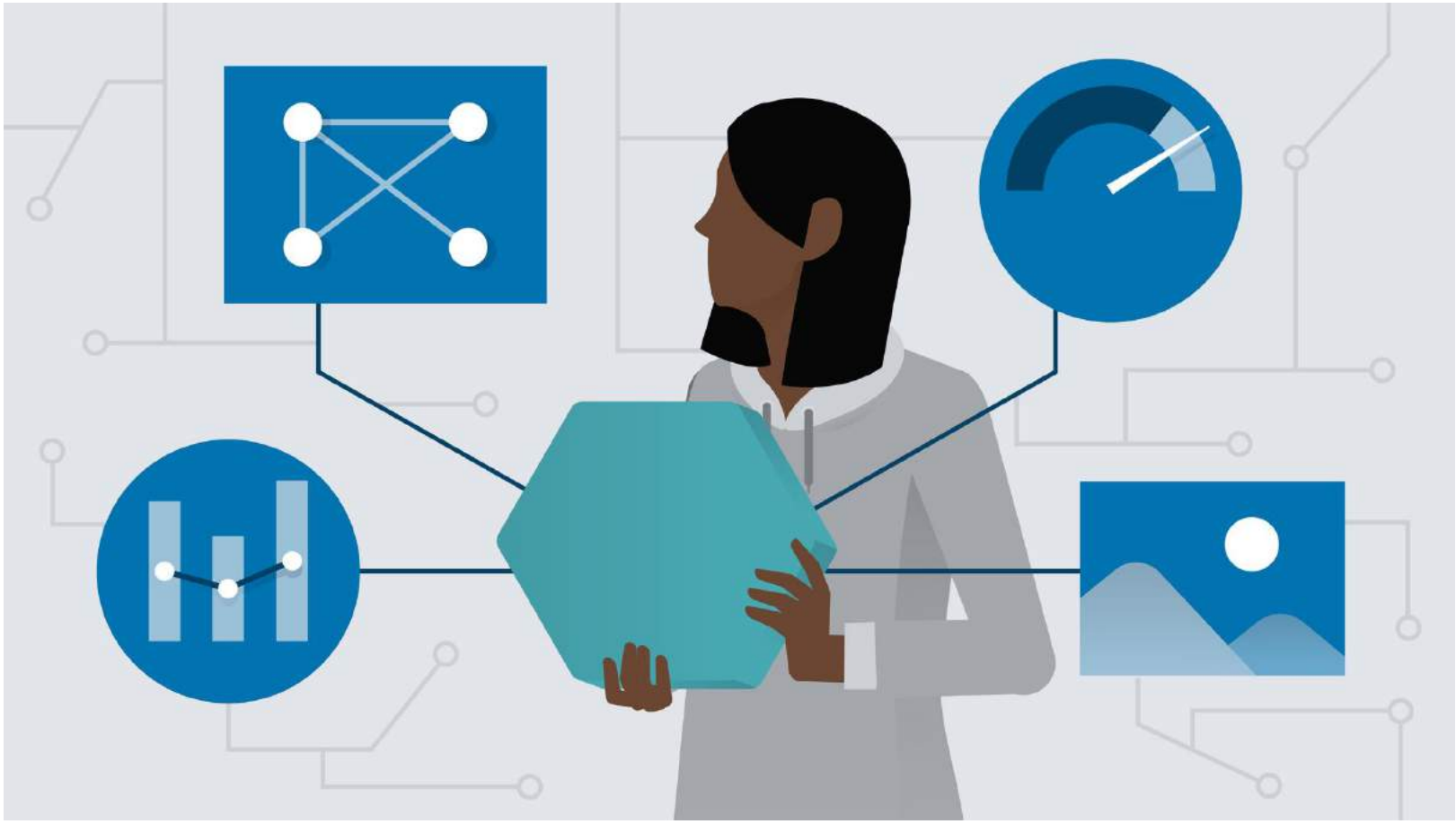
Data Science – Information System Context



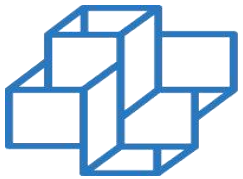
DEXL LAB
DATA EXTREME LAB



The ***wealth*** of ***Data*** and ***Models*** should be managed !!



DEXL LAB
DATA EXTREME LAB

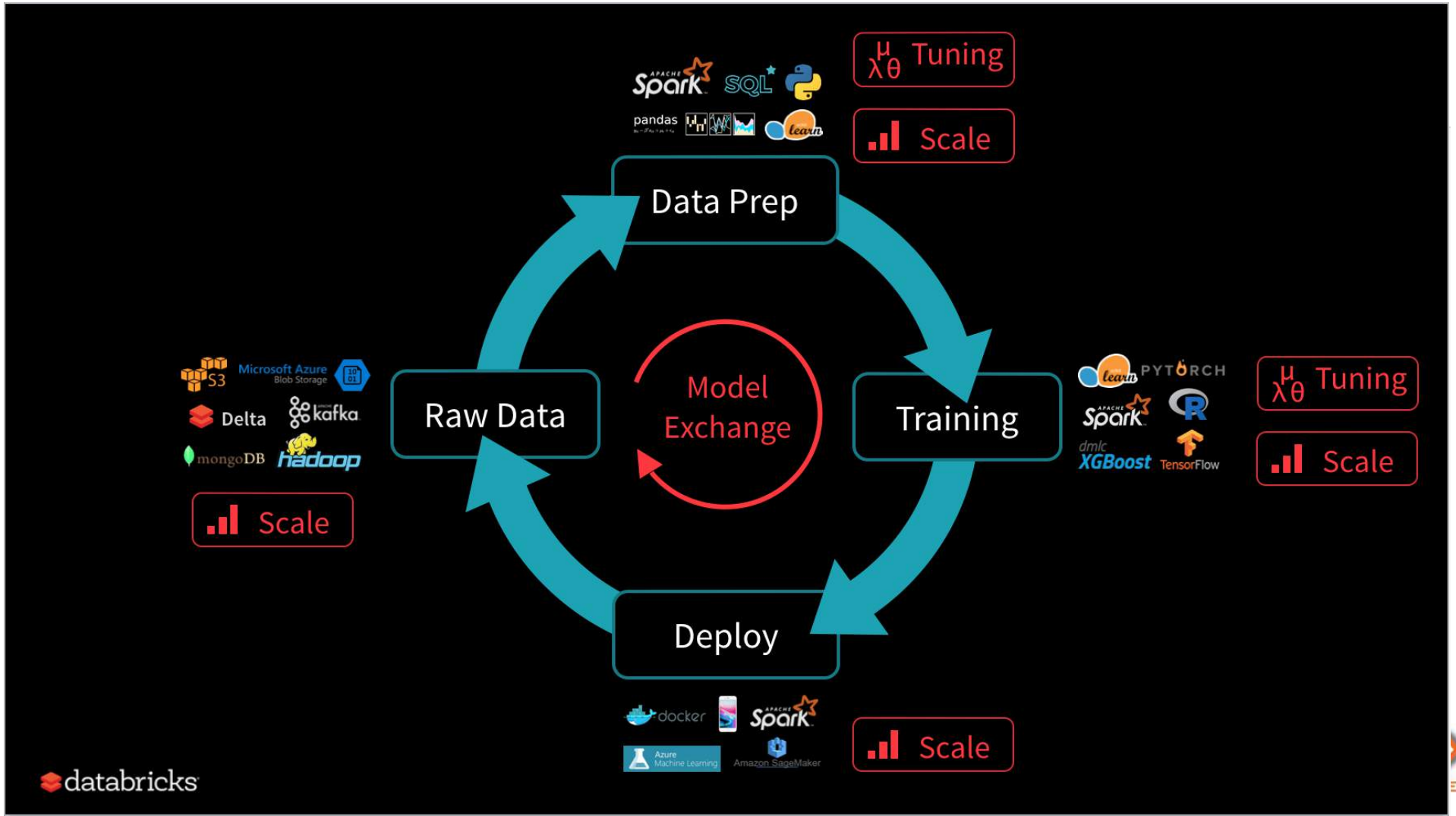


"Analysts face major practical bottlenecks in using ML that slow down the analytics lifecycle. To understand these bottlenecks, we spoke with analysts at several enterprise and Web companies. Unanimously, they mentioned that choosing the right features and appropriately tuned ML models were among their top concerns"

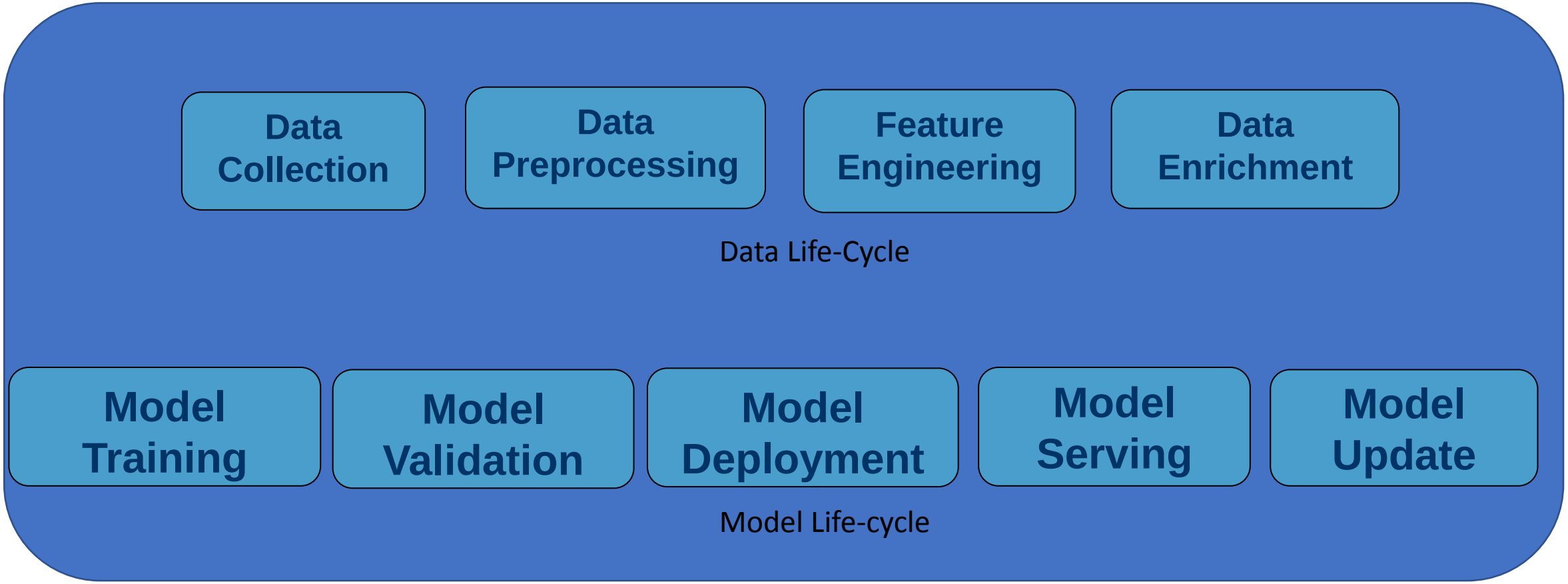
Arun Kumar et al, "Model Selection Management System: The Next Frontier of Advanced Analytics, SIGMOD Records 2016

DEXL LAB
DATA EXTREME LAB

Machine Learning (ML) Life-cycle



ML Life-Cycle

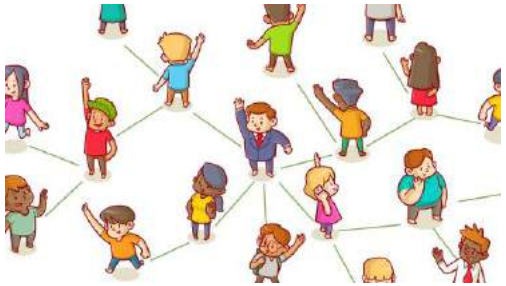


Data Preparation

Data in different formats and generated from heterogeneous data sources- Preprocessing

Social Network Analysis Applications

Friendship relationship



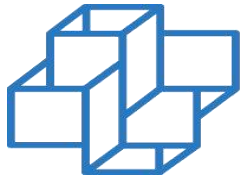
Visited pages

Comm

Data Preprocessing to prepare features for model training

DEXL LAB
DATA EXTREME LAB

Data Preparation Activities



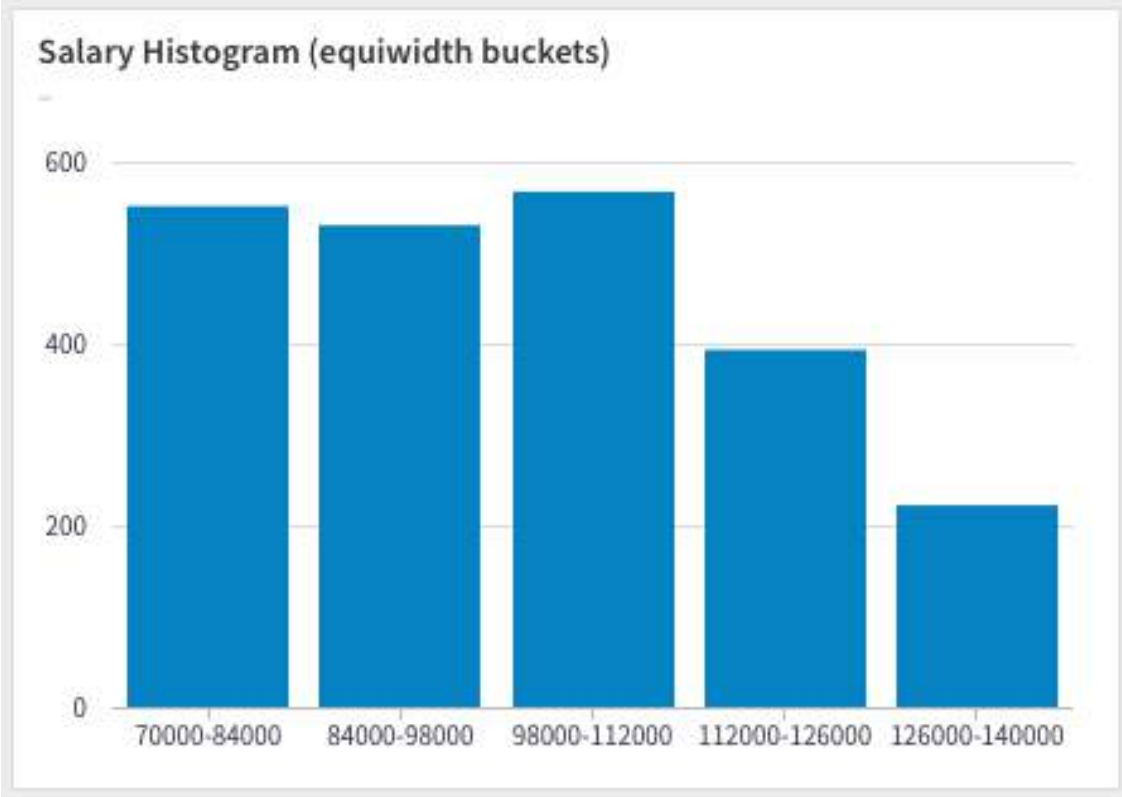
- Feature Extraction
  - Source Identification
  - Data Capture
  - Format preparation for training
- Data Cleaning
  - Missing data (ex: imputation)
  - Duplicate identification
  - Value correction

  Ihab Ylias, Holoclean - https://github.com/HoloClean - A Machine Learning System for Data Enrichment,
  *Rafael Silva Pereira, Fabio Porto, **Dealing with categorical missing data using CleanerR, BRESCI, 2019***

- Reduction, selection and transformation
  - Dimensionality reduction
  - Sampling
  - Feature selection

DEXL LAB
DATA EXTREME LAB

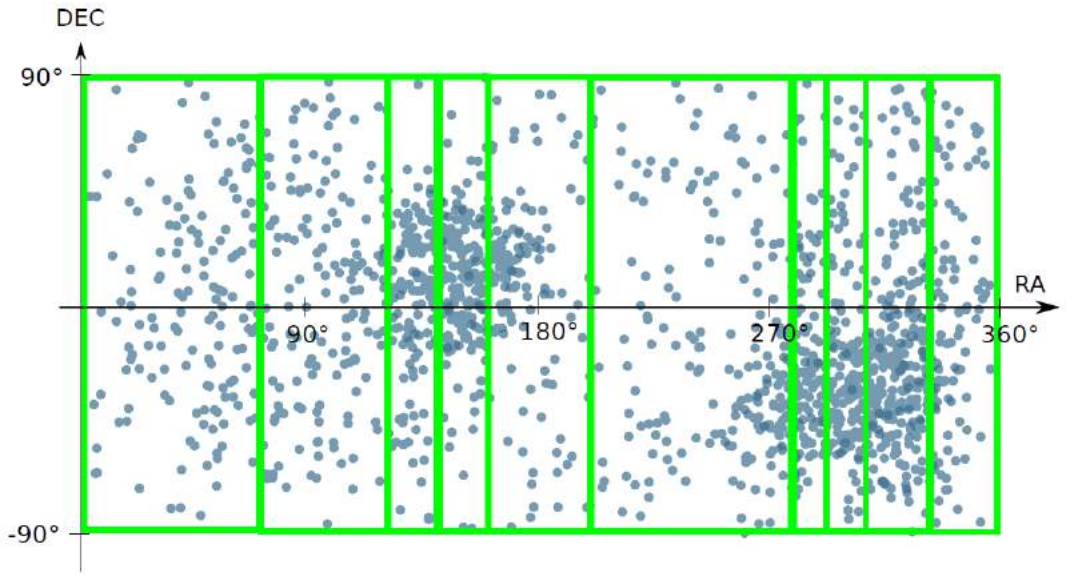
Data Preparation: Histograms Equi-width



DEXL LAB
DATA EXTREME LAB

Data Preparation: Histograms Equi-Depth

- Spatial data unevenly distributed
  - Each interval holds approximately the same number of points
  - Each interval of different width

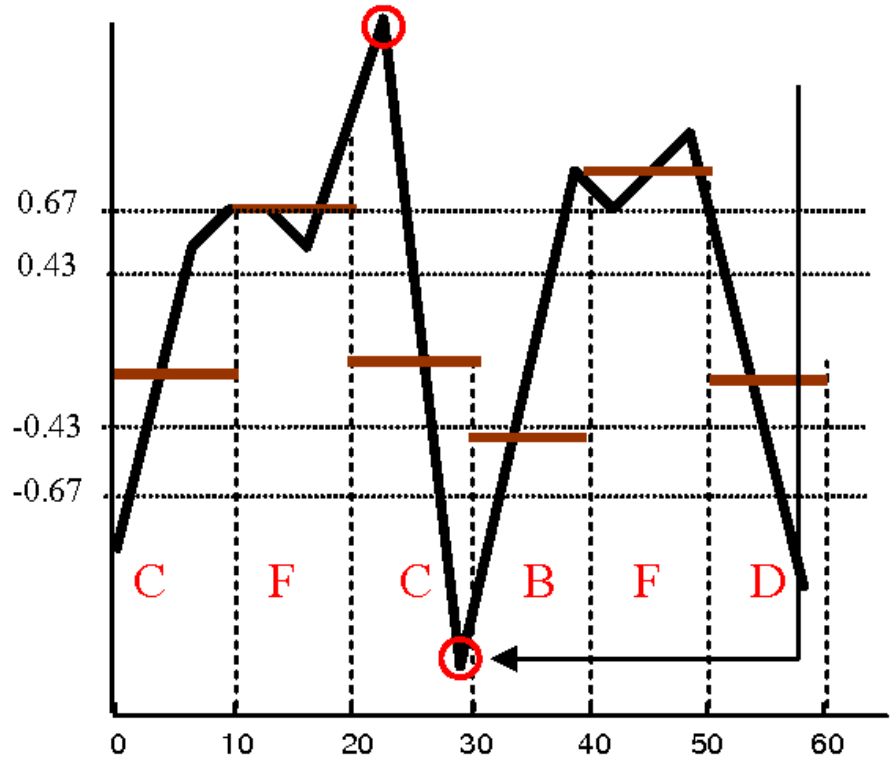


France – Algorithm: equi-depth

DEXL LAB
DATA EXTREME LAB

Data Preparation: Temporal Series–> Discrete Sequences

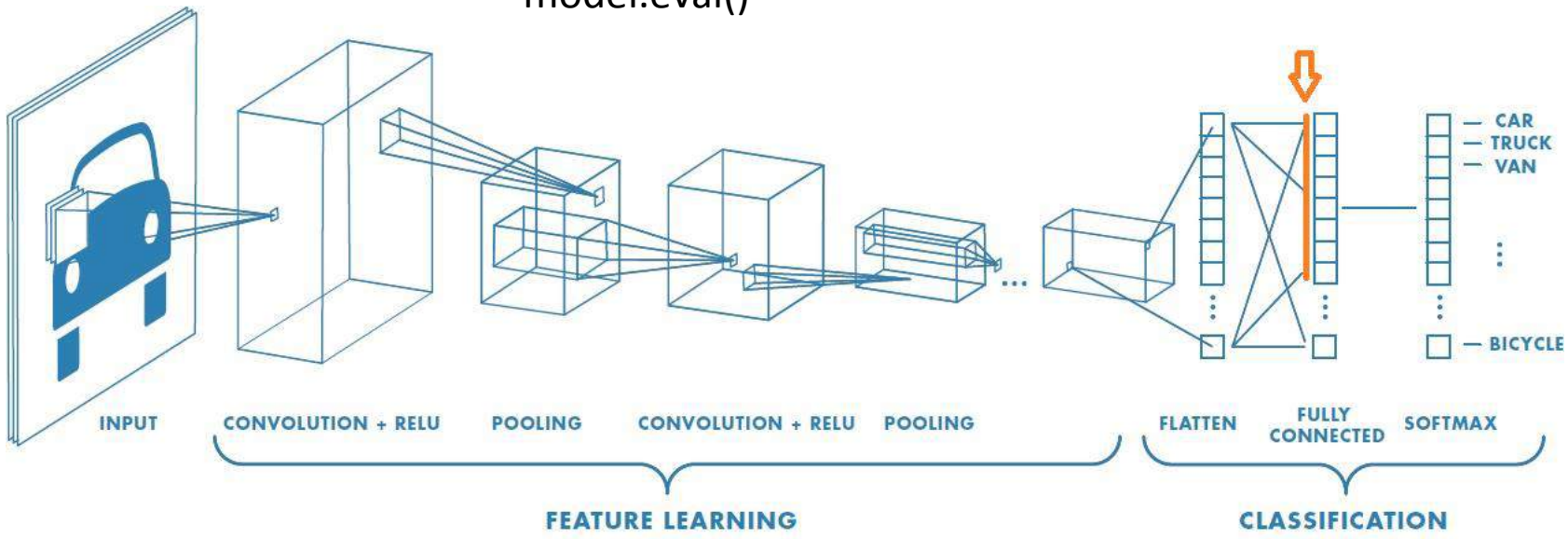
- SAX – Symbolic Aggregate Approximation
  - Series split into sequences(eg. equi-depth)
  - Compute avg value at each interfal



Data Preparation: Image -> Characteristics Vector

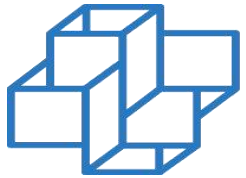
- Use of Deep NN to generate characteristic vector
- For instance, get a pretrained network
  - Ex: Res-net-18 , trained using *ImageNet*

```
# pretrained model
model = models.resnet18(pretrained=True)
 # Use the model object to select the desired layer
layer = model._modules.get('avgpool’)
model.eval()
```



DEXL LAB
DATA EXTREME LAB

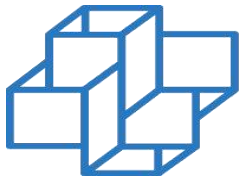
Data Preparation



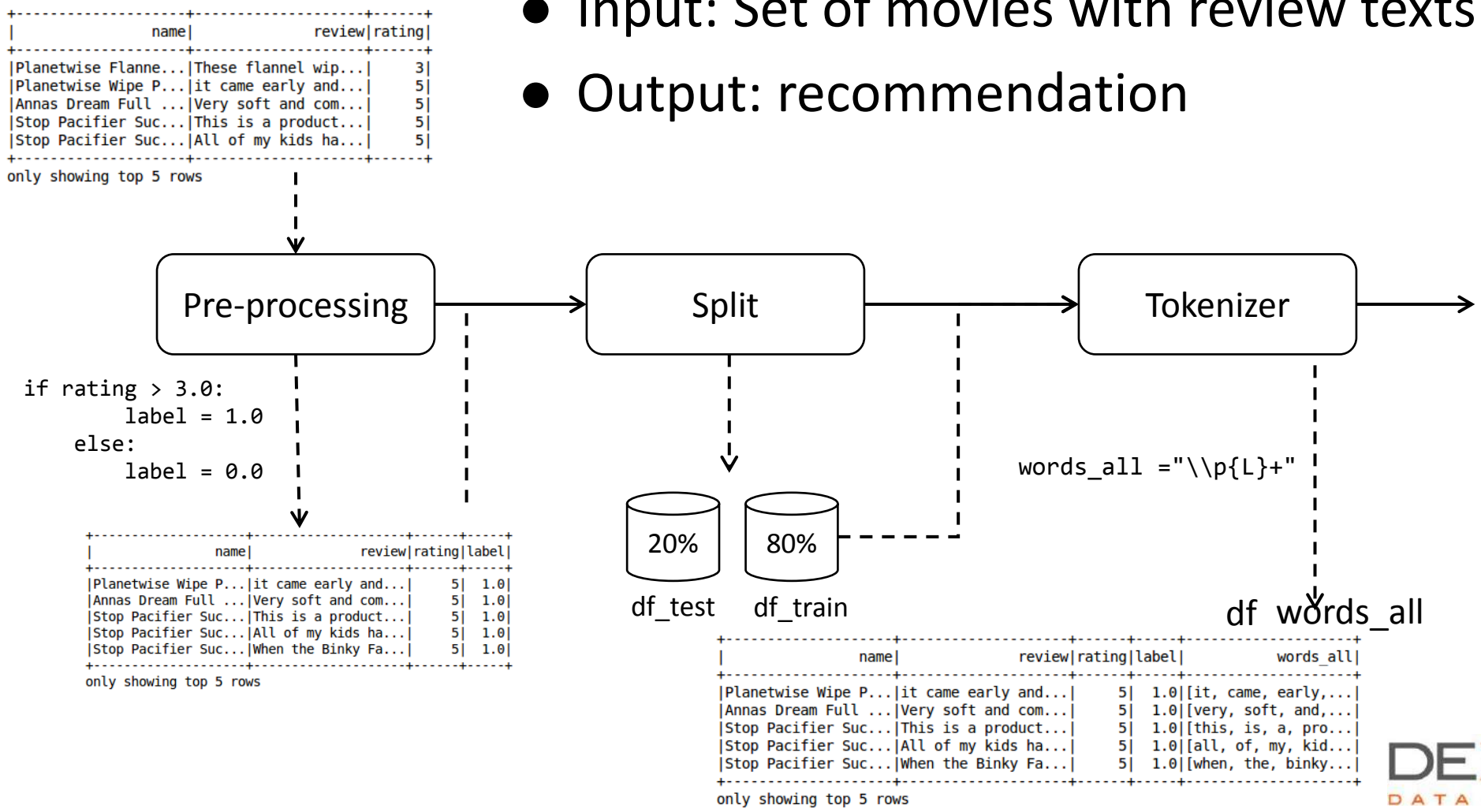
- Machine Learning models learn with examples;
- Learning a phenomenon may require thousands to millions of examples -> Data distribution
- Fine-tuning model parameters may require a number of iterations over the training samples (epochs)
- Data Preparation and Training become a Data Intensive Computing Problem
  - Use Big Data Frameworks – Apache Spark, Flink, etc…

Dataflows

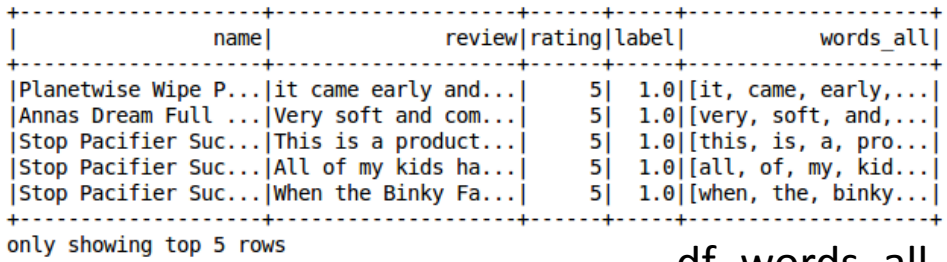
Example: Movie Recommendation



- Dataflow: movie rating;
- Input: Set of movies with review texts
- Output: recommendation



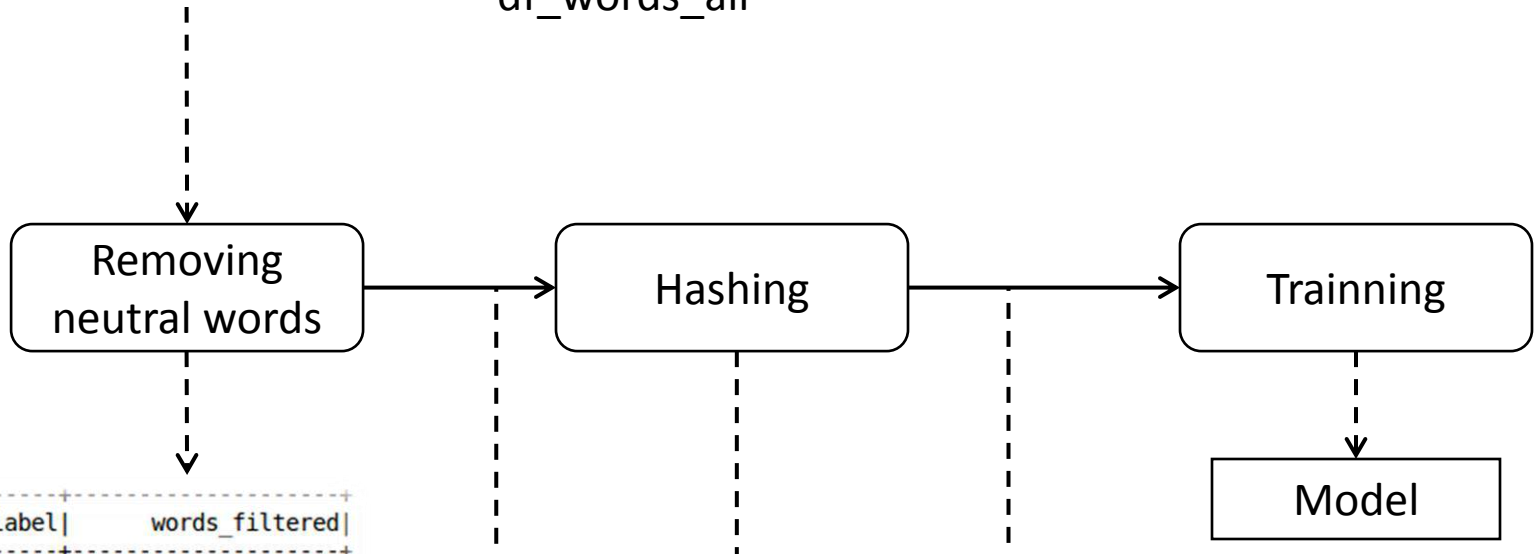
DEXL LAB
DATA EXTREME LAB

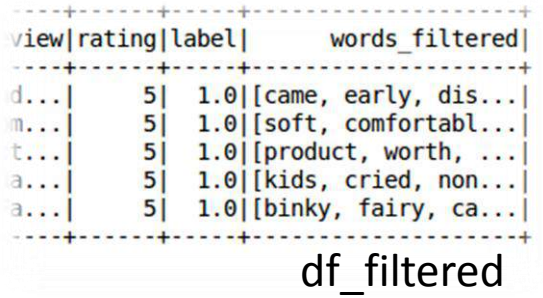


name	review	rating	label	words_all
Planetwise Wipe P...	it came early and...	5	1.0	[it, came, early,...]
Annas Dream Full ...	Very soft and com...	5	1.0	[very, soft, and,...]
Stop Pacifier Suc...	This is a product...	5	1.0	[this, is, a, pro...]
Stop Pacifier Suc...	All of my kids ha...	5	1.0	[all, of, my, kid...]
Stop Pacifier Suc...	When the Binky Fa...	5	1.0	[when, the, binky...]

only showing top 5 rows

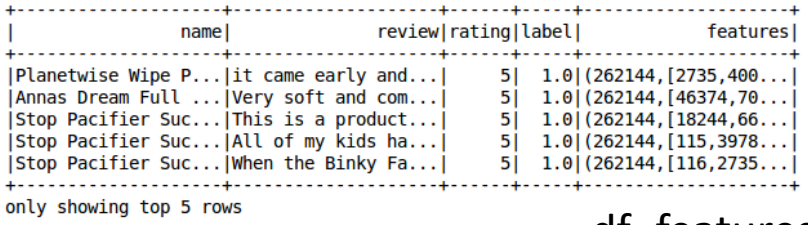
df_words_all





view	rating	label	words_filtered
d...	5	1.0	[came, early, dis...]
m...	5	1.0	[soft, comfortabl...]
t...	5	1.0	[product, worth, ...]
a...	5	1.0	[kids, cried, non...]
a...	5	1.0	[binky, fairy, ca...]

df_filtered



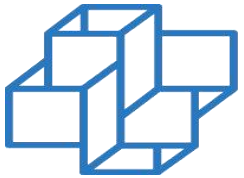
name	review	rating	label	features
Planetwise Wipe P...	it came early and...	5	1.0	(262144,[2735,400...]
Annas Dream Full ...	Very soft and com...	5	1.0	(262144,[46374,70...]
Stop Pacifier Suc...	This is a product...	5	1.0	(262144,[18244,66...]
Stop Pacifier Suc...	All of my kids ha...	5	1.0	(262144,[115,3978...]
Stop Pacifier Suc...	When the Binky Fa...	5	1.0	(262144,[116,2735...]

only showing top 5 rows

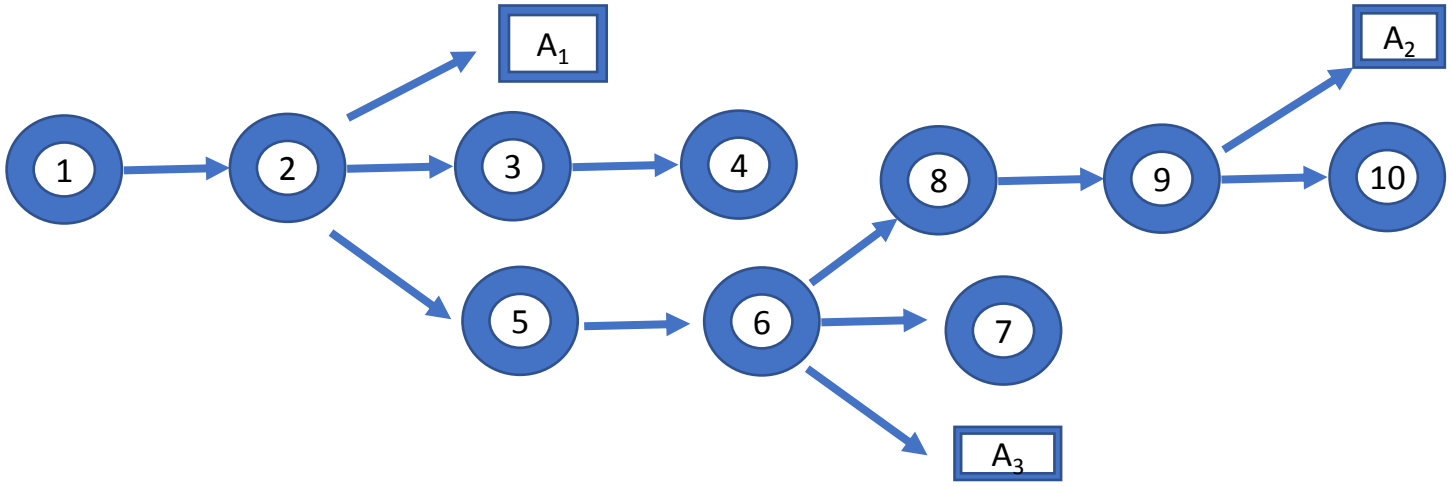
df_features

DEXL LAB
DATA EXTREME LAB

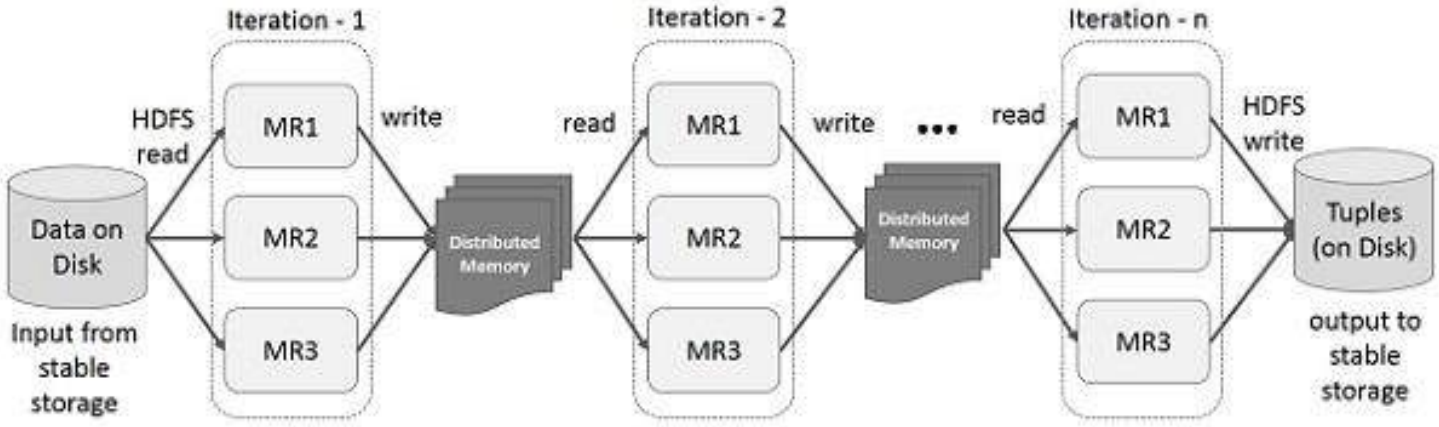
Dataflow model for data pre-processing



- Nodes correspond to pre-processing activities
- Edges define the communication path between activities: input/output files

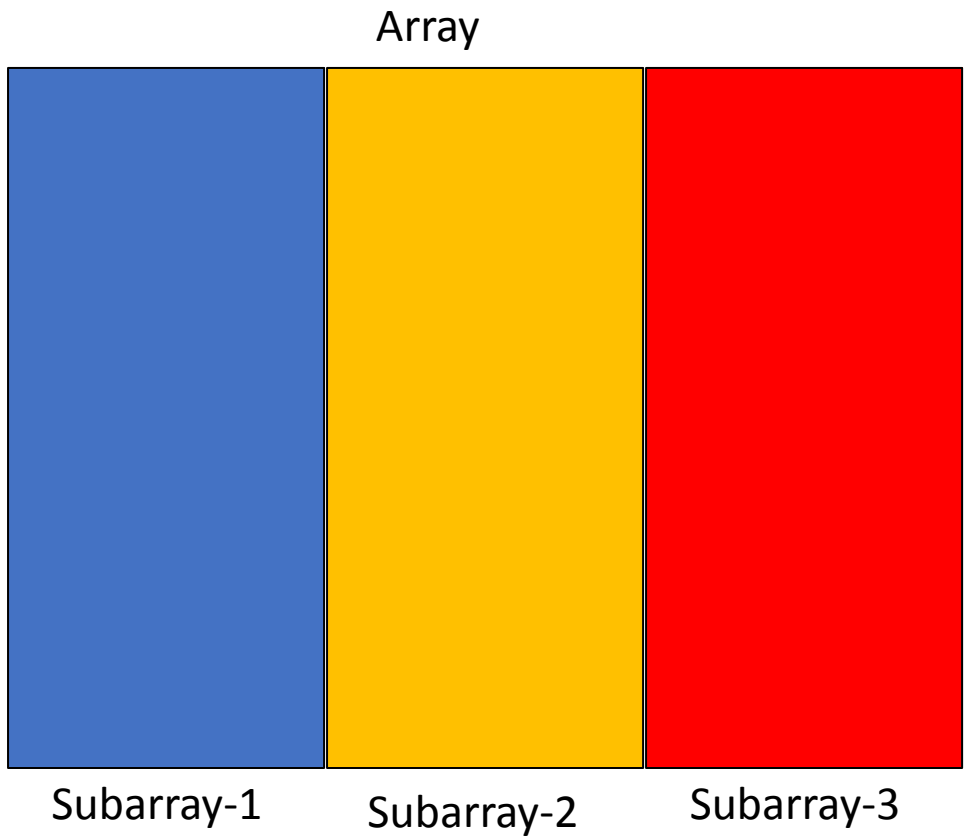


Apache Spark based in-memory dataflow Model

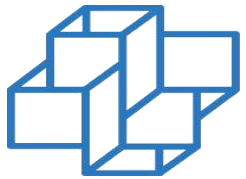


DEXL LAB
DATA EXTREME LAB

Spark – Data Locality with Partitioning

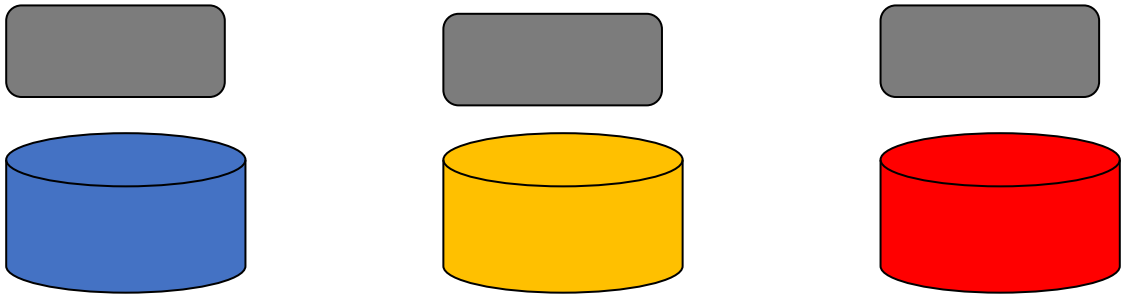


Data Locality

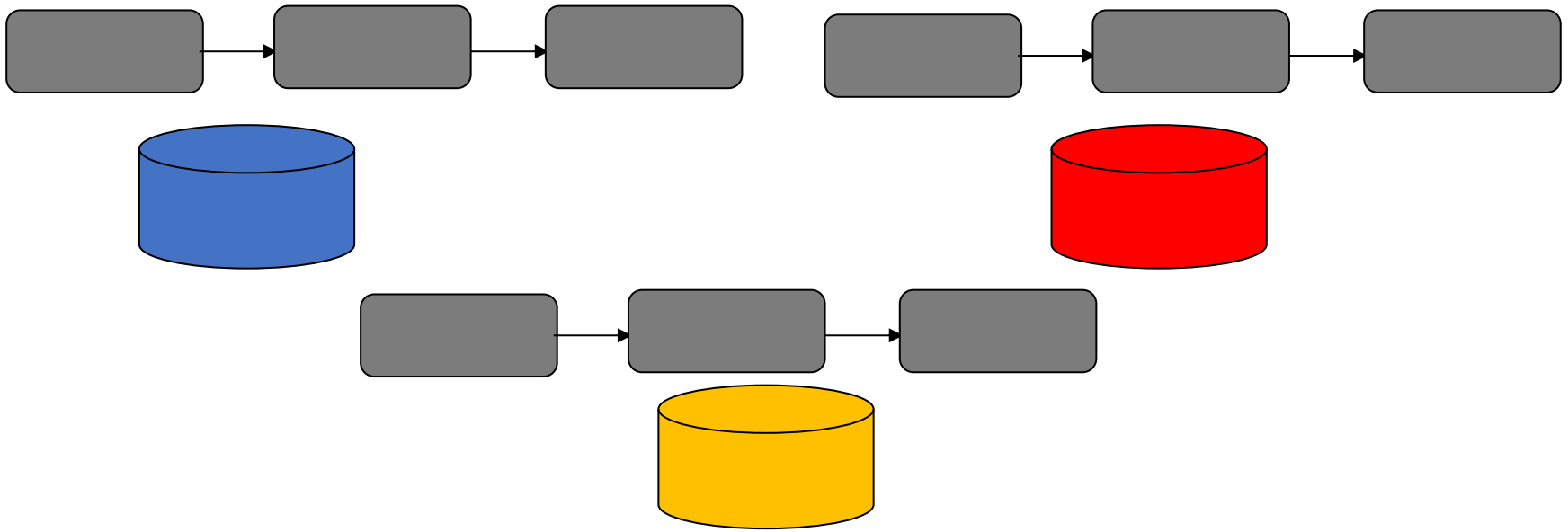


- In Big Data, na eficiente processing model shall reduce the cost involved in data transfer;
  - Considering:
    - a data set partitioned into chuncks
    - Chunks distributed through shared-nothing nodes
  - Data Locality
    - To schedule processing code to node containing the Chuck to be processed;
    - Minimize data movement.

One task many data chunks



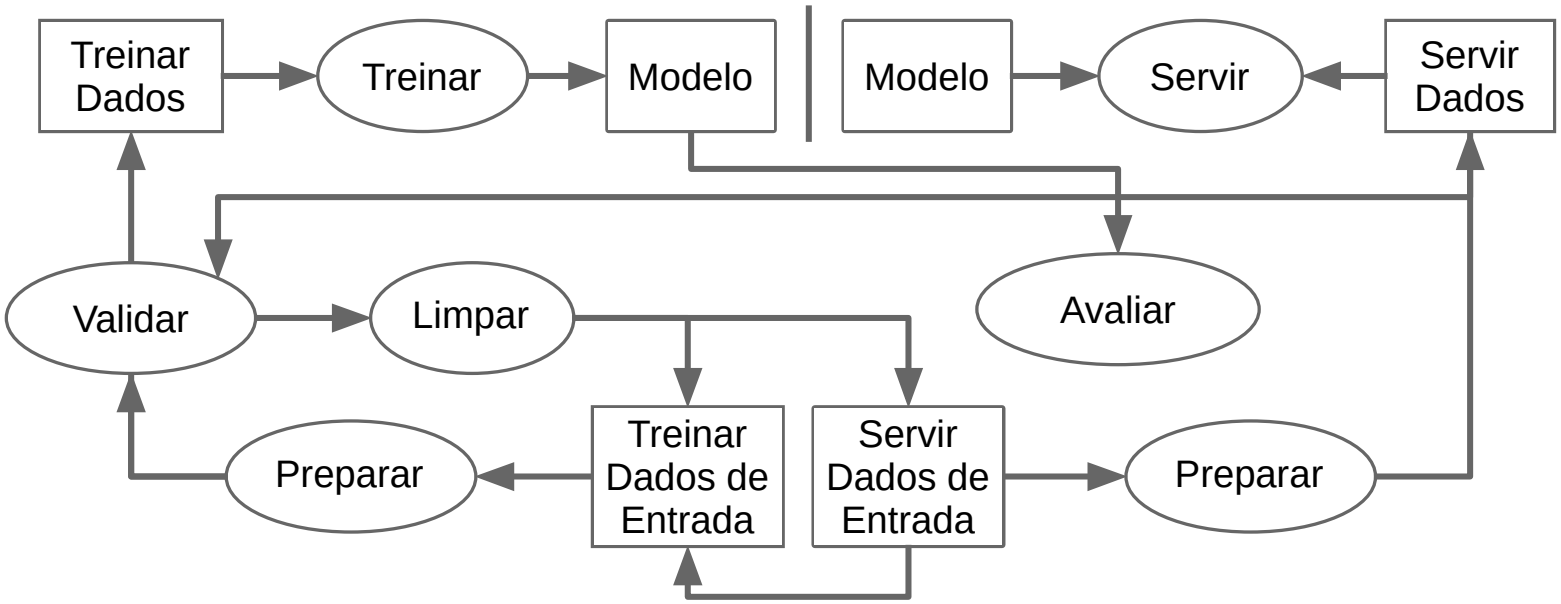
One dataflow (Pipeline) various data chunks



Using Big Data Frameworks to solve problems

- Problems solved using the framework
  - Unsupervised learning – eg. K-means
  - Statistical Models – Linear Regression
  - Data Preparation – Join datasets
- Problem representation in the form of dataflows
  - In its simplest form: a sequence of Map-Reduce
- Operations:
  - Map, join, union, filter, groupbykey
  - Reduce

Data Life-cycle in ML

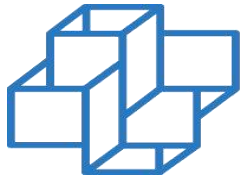


Data Lifecycle Challenges in Production Machine Learning: A Survey, SIGMOD Record, V(47),N.2, 2018

DEXL LAB
DATA EXTREME LAB

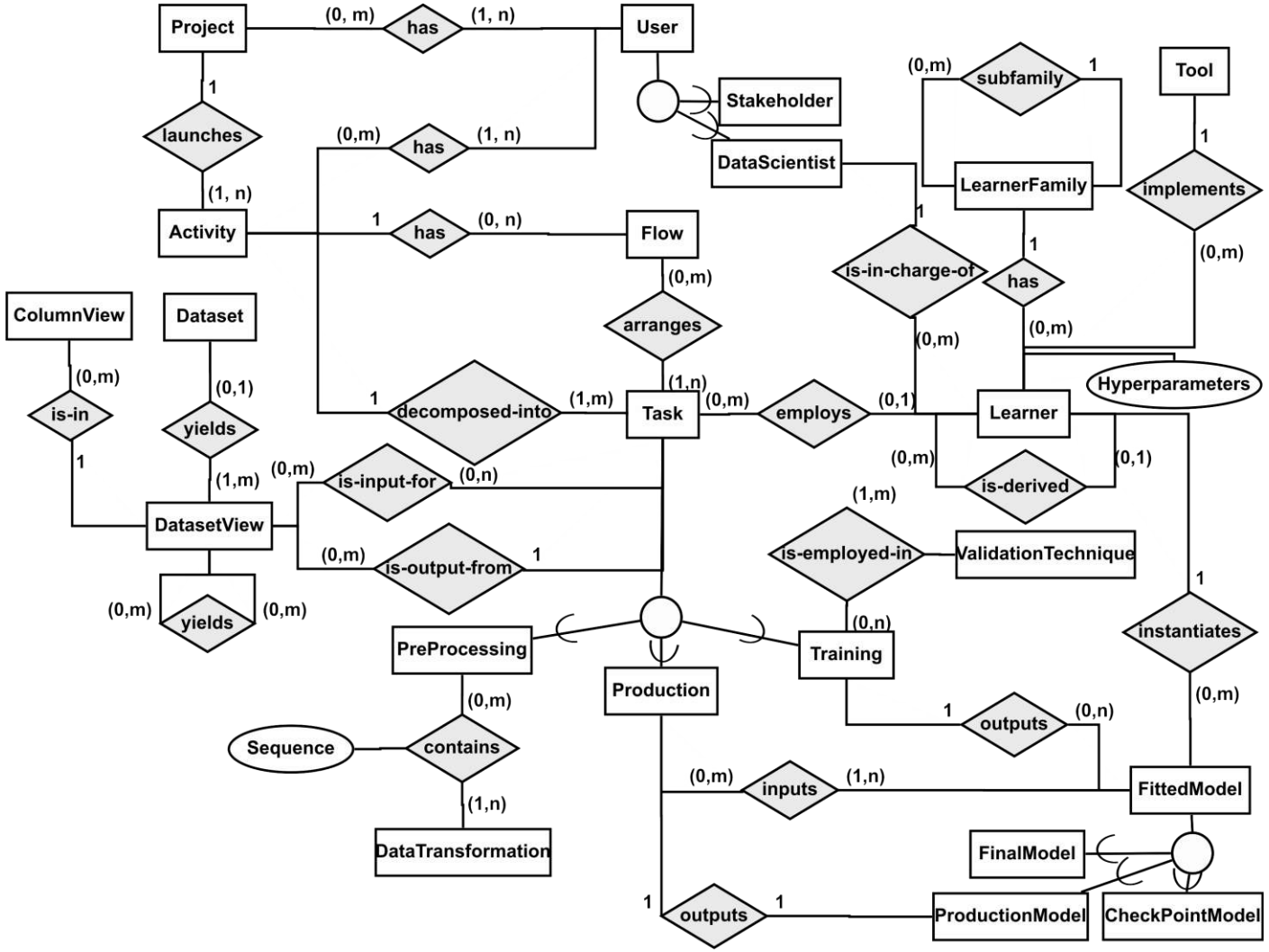
Model Construction

ML life-cycle: Two Important phases



- Model Construction
  - Data Preparation
  - Model Training
    - Parameter and hyperparameter tuning
  - Model Validation
- Model Serving
  - Homologation
  - Deployment
  - Serving
  - Update
- Different tools supporting the Life-cycle phases

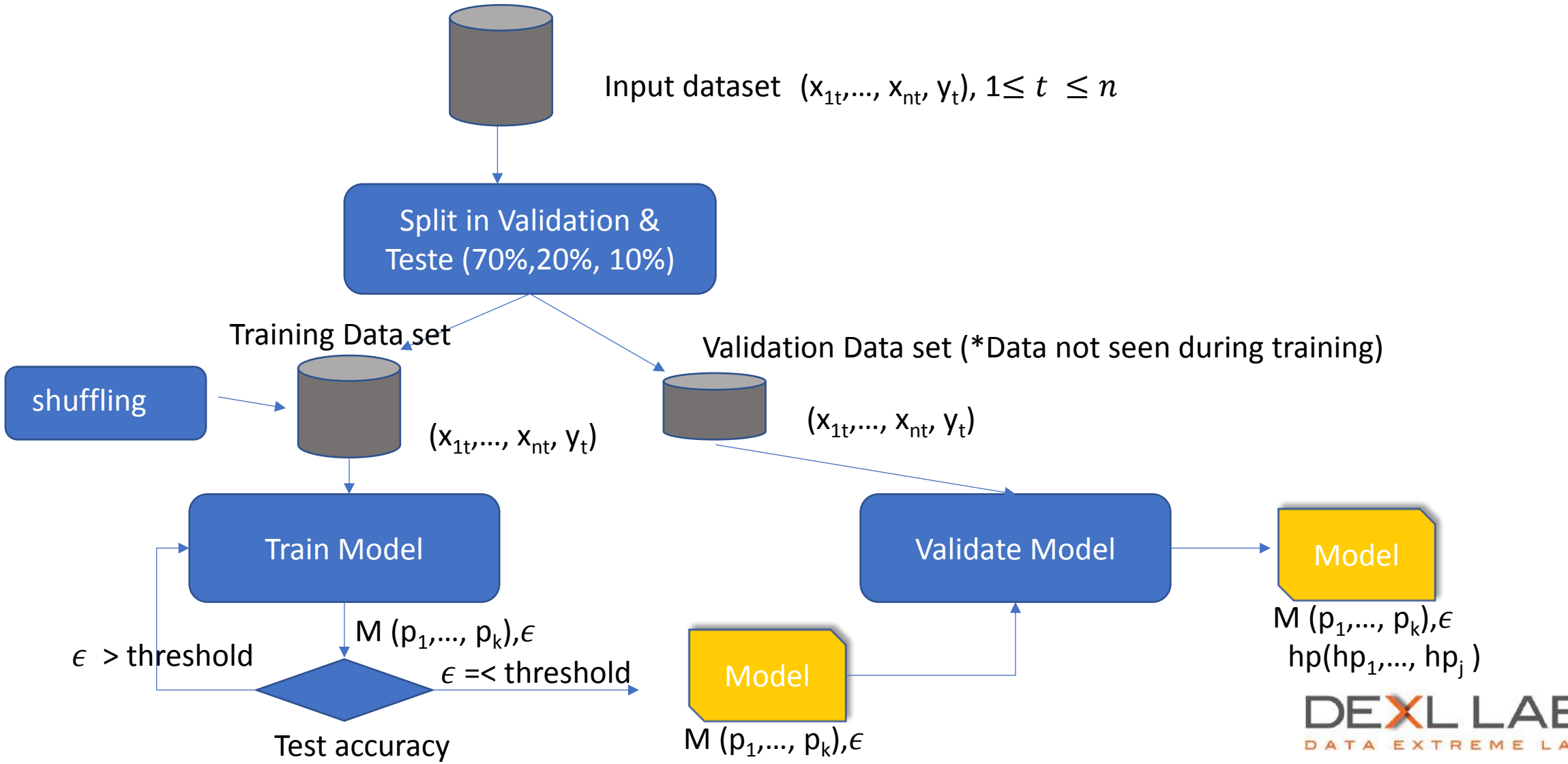
Conceptual Model



Daniel Ramos Silva et al., A conceptual vision toward the management of Machine Learning models, ER 2019 (submitted)

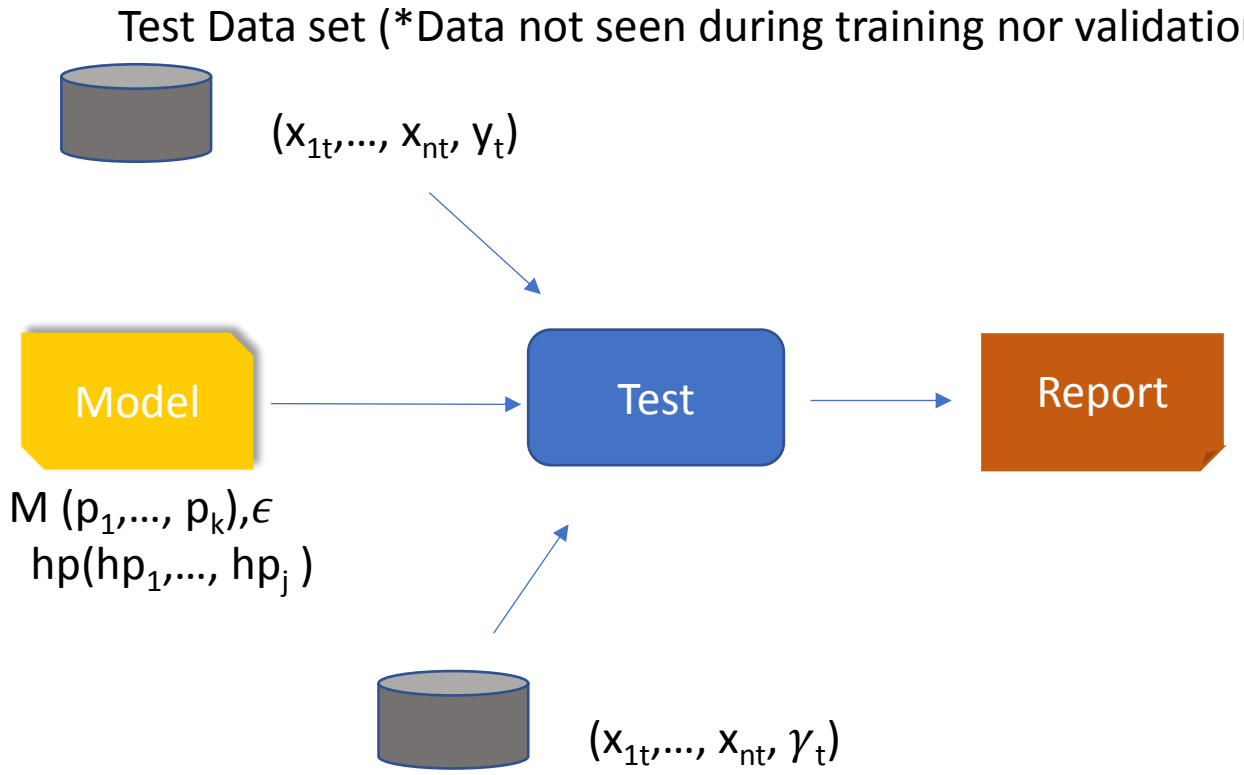
DEXL LAB
DATA EXTREME LAB

Model Construction



DEXL LAB
DATA EXTREME LAB

Model Testing

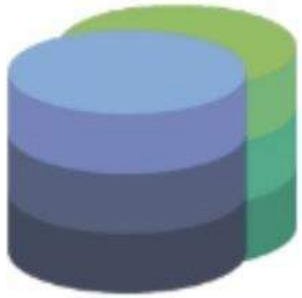


System supporting Model Construction

ModelDB: an end-to-end model management system

Ingest models, metadata

track



Model artifact Storage & Versioning

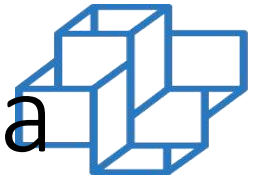
store & index

Query

Collaboration, Reproducibility

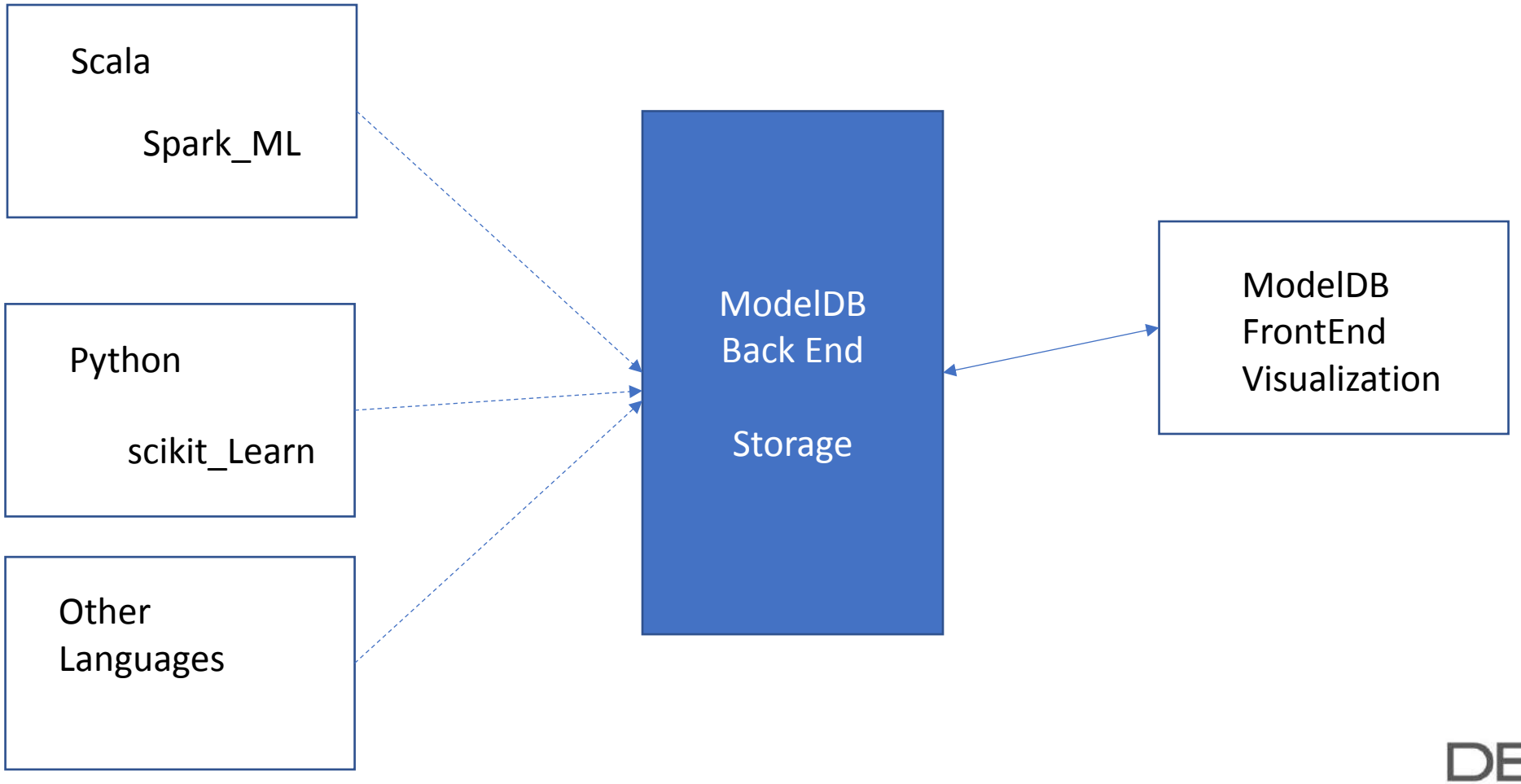
query, reproduce++

ModelDB – MIT –San Madden, Matei Zaharia



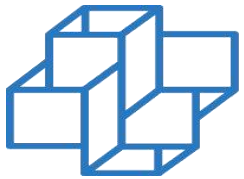
- Manages the ML life-cycle
  - Logs information about models, parameters and hyperparameters
    - Reproducibility
    - Add some adapted functions to store data/metadata about operations:
      - read_csv_sync; fit_sync; train_test_split_sync;predict_sync
  - Each modeling run = version
  - Supports models trained with (Python -scikit-learn) and (Scala –Spark MLIB)
  - Supports learners: Logistic Regression, Random Forest, Decision Tree
  - Limited functionality for model selection (comparison of models)
    - Help through visualization on the results of models
    - Model search
  - Run Queries in Models and Data
    - Where is that logistic regression model I created last week with feature X?
  - Enable Model annotation, model review
    - Keeps track of data transformation pipelines

ModelDB Architecture



DEXL LAB
DATA EXTREME LAB

mlflow



Languages (python, R, Java) APIs (REST, CLI,..)

Project

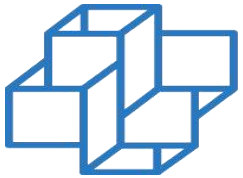
Tracking

Model

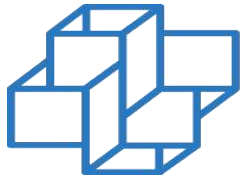
Ml libraries

Environments (Notebook, Spark Cluster,..)

DEXL LAB
DATA EXTREME LAB



- Platform for managing model life cycle:
  - Experiment; Reproductibility and Deployment
  - Enables the execution of models using Python or Scikit-learn
- MLFlow Tracking; MLFlow Project; MLFlow Models
- Tracking
  - Records artefacts used during model execution:
    - Parameters, hyperparameters, input datasets and results;
    - Performance metrics
    - Model code version
  - User adds functions to the code to capture metadata
- Project
  - Packaging ML code to be reused, preproduced to be shared with users or passed to production
- Model
  - Standard format for storing models and interoperability of models among platforms
  - Supports different execution flavors:
    - Python, h2o, Keras, Mleap, Pytorch, Scikit-Learn, Spark-Mlib, TensorFlow
  - Enables Model Selection
    - Visual comparison of models
    - Highlighting performance metric differences



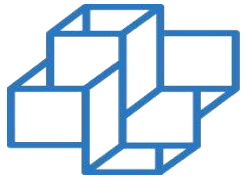
- Tracker API
  - mlflow.create_experiment()
  - mlflow.set_experiment()
  - mlfow.start_run()
  - mlfow.end_run()
  - mlfow.log_param(key, value) ; mlfow.log_params(params)
  - mlfow.log_metric(key, value, step)

  - Example:
    - **with** mlflow.start_run():
      for epoch **in** range(0, 3):
      mlflow.log_metric(key="quality", value=2*epoch, step=epoch)

DEXL LAB
DATA EXTREME LAB

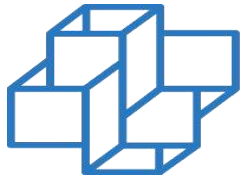
Model Serving

Model Serving



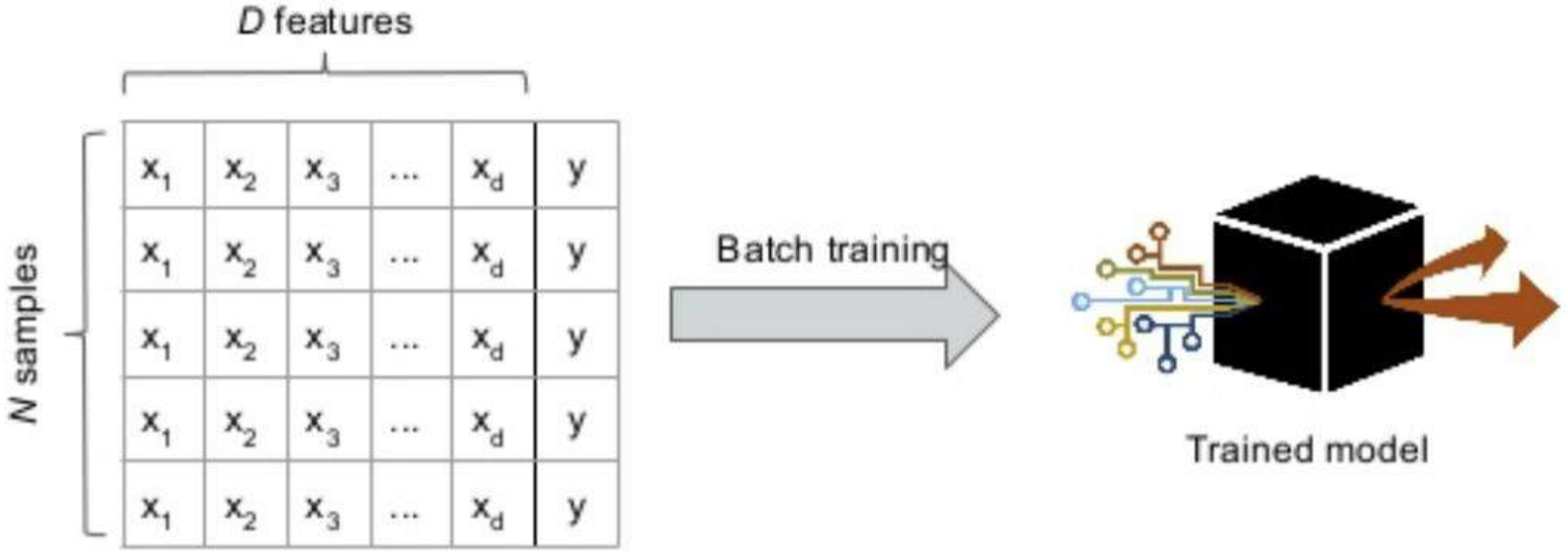
- Once a model has been validated and homologated it can be deployed to production
- A deployed model can be seen as a UDF
  - UDF (X):-> $\hat{Y}$
    - X – is the input set of features
    - $\hat{Y}$ - is the prediction
    - Loss(Y, $\hat{Y}$) determines the prediction error
- Given a pair (X,Y) there may be a set of trained models M that receive a
  - $X' \sqsubseteq X$ – as input, and produce
  - $\hat{Y}$ as a prediction
- The best model may depend on the context (accuracy, latency, user or region)

Online Learning

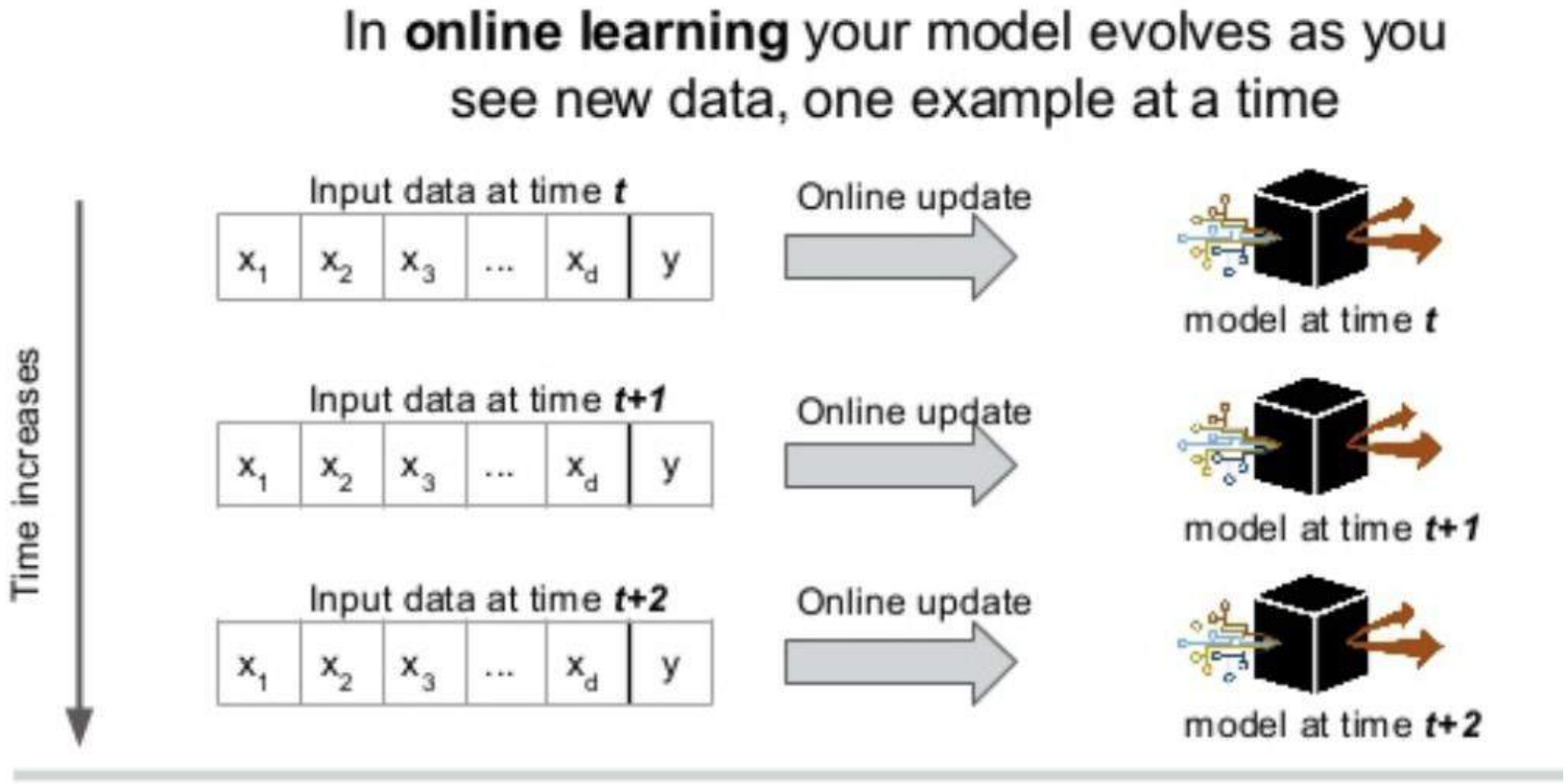


- Data instances are input one-by-one
  - There exists a temporal dependency among data instances;
  - The learner is continuously updated with new observations
    - For each observation a prediction is made and compared against a real observation
    - The error indicates the update to be applied to the model
  - The model accuracy may vary continuously
  - Requires only the update to the current model and the new observation
  - Objective is to react in real-time to data distribution variation

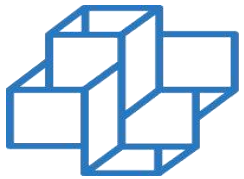
Offline Learning



Online Learning



Online Learning



For t = 1, 2, ..

retrives new observation $x_t \in \chi$

predict $p_t \in \mathcal{D}$

obtain the correct answer $y_t \in \gamma$

Computes error(p_t, y_t)

Once the error is computed

Model is updated

And used to predict new observation

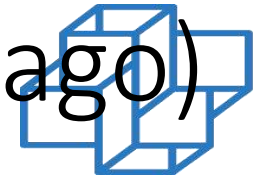
each t represents an observation

X_t is an observation at time t

P_t is the prediction for x_t

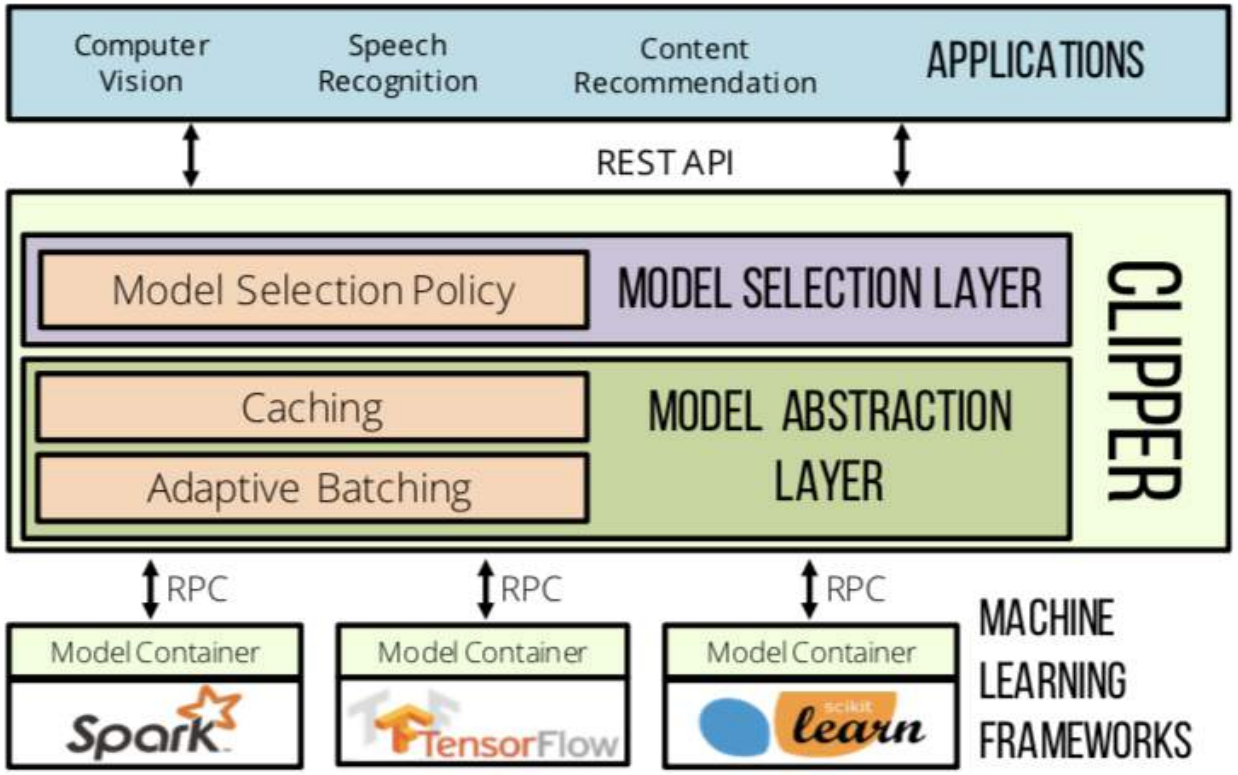
This is the difference between the prediction and the right label

Clipper – Model Serving (UC Berkeley, Univ. Chicago)

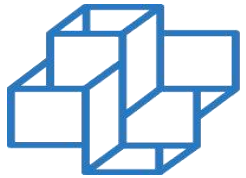


- Clipper is a model serving system
- Enables the invocation of various models for answer to a prediction query;
- Implements Model prediction caching
  - Reduce prediction latency
- Adaptive batching
  - Aggregate point queries in a queue to increase throughput
- Enables model selection
  - Selects a best model using “multi-armed bandit” algorithm
  - Enables combining models
    - Linear ensemble of models

Clipper Architecture



Clipper API



- Predict(m: ModelId, x: X) -> y: Y
- Batch Interface:
  - **interface** Predictor<X,Y> {

    List<List<Y>> pred_batch(List<X> inputs);

    }
- Selection :
  - **interface** SelectionPolicy<S, X, Y> {

    S init();

    List<ModelId> select(S s, X x);

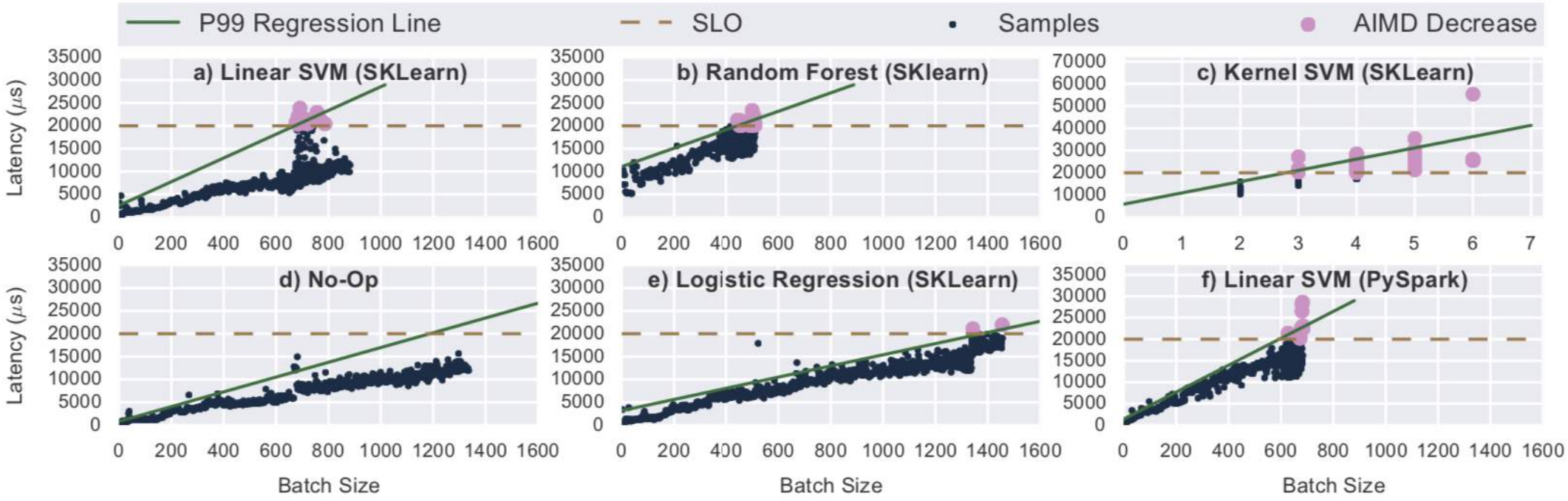
    pair<Y, double> combine(S s, X x, Map<ModelId, Y> pred);

    S observe(S s, X x, Y feedback, Map<ModelId, Y> pred);

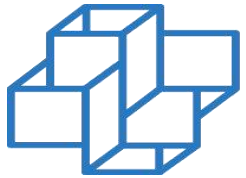
  }

DEXL LAB
DATA EXTREME LAB

Maximum Batchsize with Latency 20ms

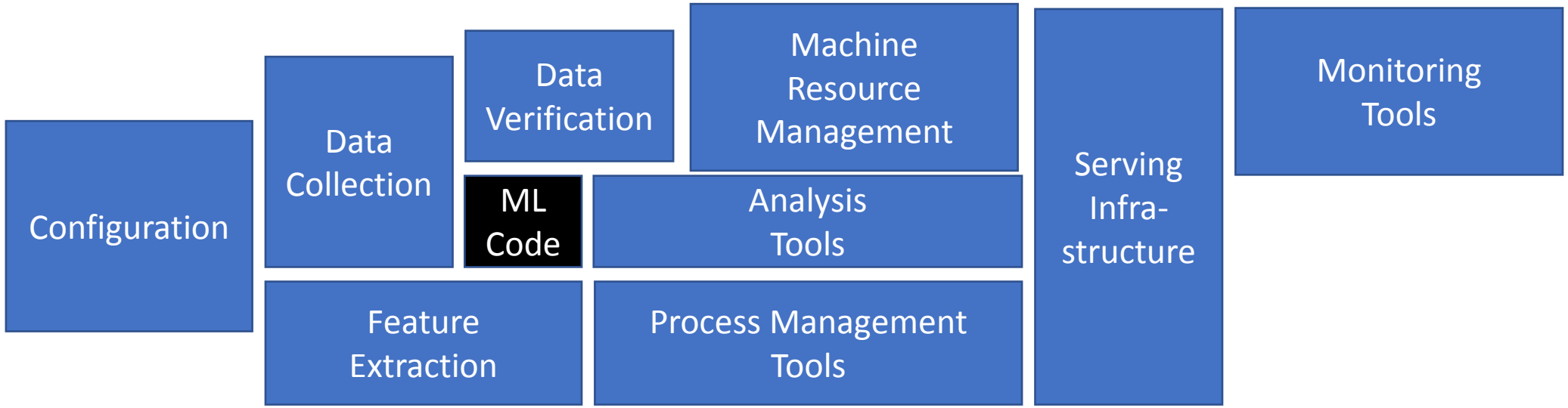


Other Models serving systems



- TensorFlow serving
  - Coupled with the Tensorflow model may better tune the model for efficient execution. For instance, make efficient use of GPUs for a given model;
- Mlflow

ML System Perspective

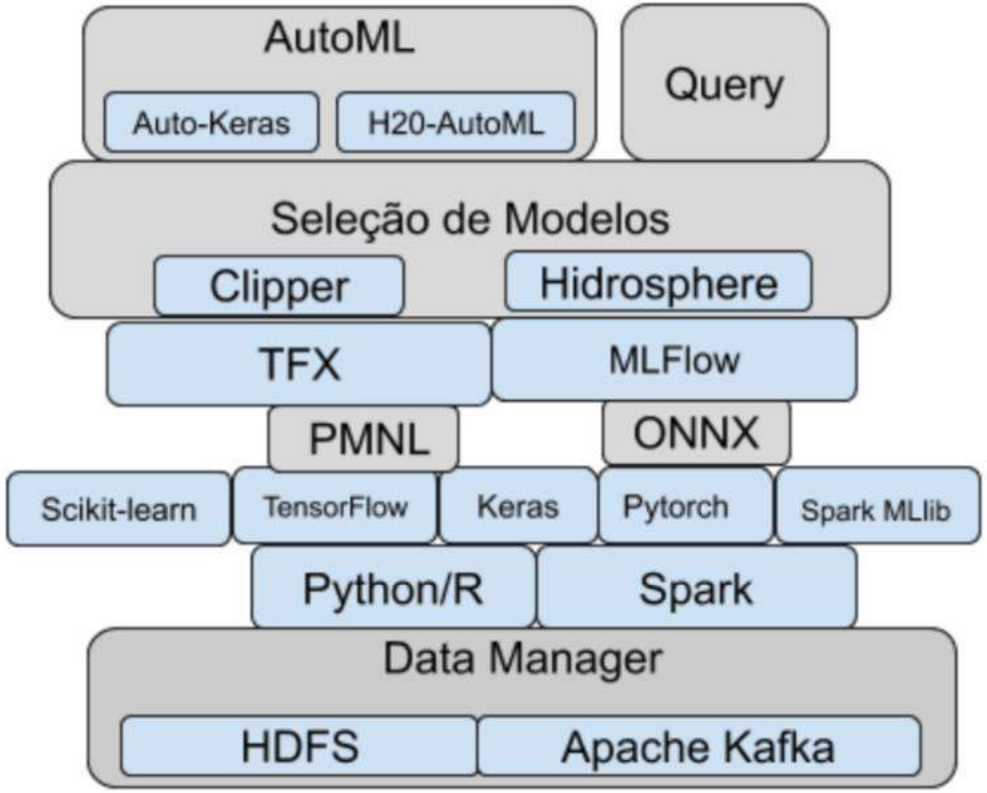


Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

D.Sculley et al., Hidden Technical Debt in Machine Learning Systems, HILDA 2016

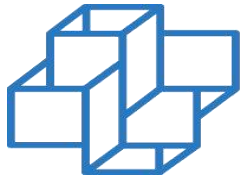
DEXL LAB
DATA EXTREME LAB

Gypscie – Managing ML Models



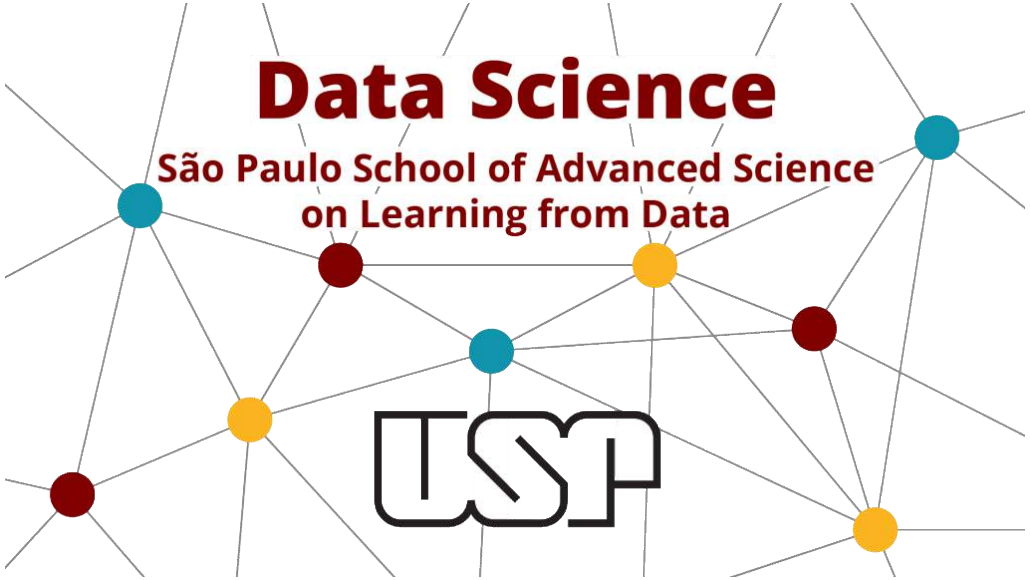
DEXL LAB
DATA EXTREME LAB

Final Comments



- The models and data are assets for institutions in the XXI century (Industry 4.0)
- Understanding the complexity involved in the heterogeneous environment is paramount
  - Languages
  - Learners
  - Formats
  - Libraries
  - Etc..
- ML Management Systems foster a principled approach for ML life-cycle

DEXL LAB
DATA EXTREME LAB



FAPESP

Thank you

Fabio Porto

fporto@lncc.br

dexl.lncc.br

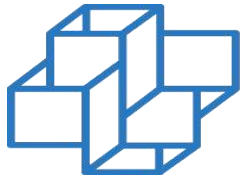
Laboratório Nacional de Computação Científica

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES

BRASIL GOVERNO FEDERAL

DEXL LAB
DATA EXTREME LAB

Avaliação de Modelos



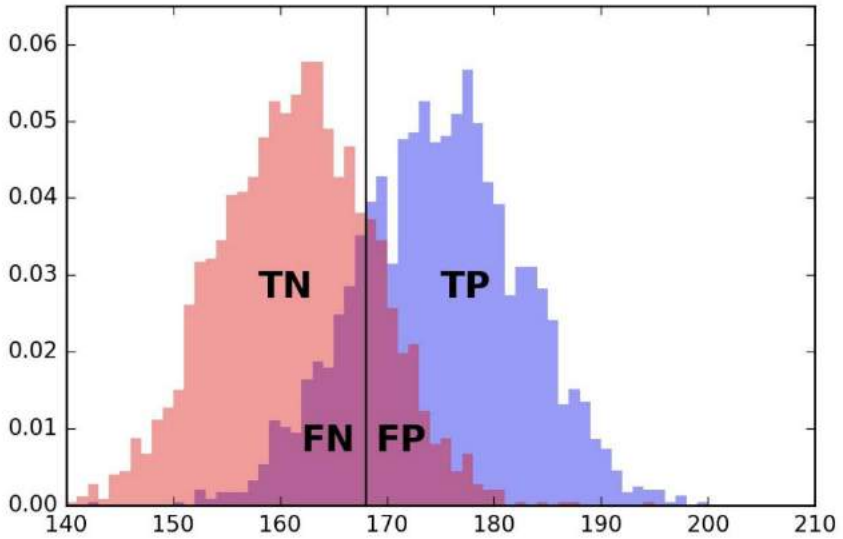
- O processo de avaliação de modelos procura responder algumas questões:
  - Quão boa é a predição oferecida pelo modelo?
  - Quão preciso espera-se que o modelo deva ser?
  - Qual o ganho de uma pequena melhora na acurácia do modelo?
- A medida de acurácia é melhor quando se refere a dados nunca vistos
  - Modelos de alta acurácia em dados de treinamento podem levar à sobreajustes

Avaliando Classificadores

- Binários:
  - Escolha um valor para ser positivo;
  - TP – True Positive – o classificador prediz positivo um rótulo que é positivo;
  - TN – True Negative – o classificador prediz negativo um rótulo que é negativo;
  - FP – False Positive – o classificador prediz positivo um rótulo negativo;
  - FN – False Negative – o classificador prediz negativo um rótulo positivo;

Matriz de Confusão

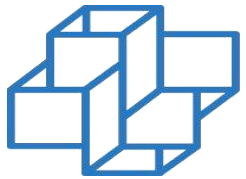
		Valor Predito	
		Posit.	Negativo
Valor Real	Posit.	TP	FN
	Negat.	FN	TN



Preditor:
se $\geq 1{,}68$
Masculino

DEXL LAB
DATA EXTREME LAB

Matriz de Confusão multi-classe



Digitos Preditos

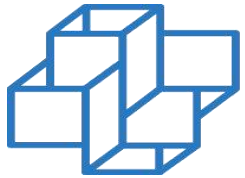
Digitos Reais

Digits	0	1	2	3	4	5	6	7	8	9
0	351	0	5	4	2	7	2	1	6	0
1	0	254	0	0	2	0	0	1	1	2
2	1	1	166	4	5	1	3	2	2	1
3	1	2	4	142	0	5	0	1	4	0
4	3	3	8	1	180	3	2	5	4	4
5	0	0	3	11	0	140	3	0	7	1
6	0	2	2	0	4	0	158	0	1	0
7	0	0	2	2	1	0	0	132	2	1
8	2	1	8	0	0	0	2	1	137	1
9	1	1	0	2	6	4	0	4	2	167

Predição de Digitos de CEP

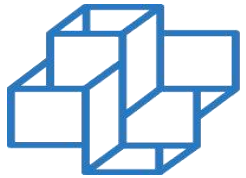
DEXL LAB
DATA EXTREME LAB

Extração de características: Conversão de Tipos de dados



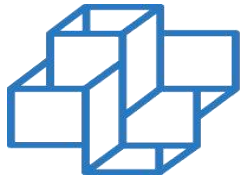
Tipo de Dado fonte	Tipo de dado destino	Método
Numérico	Categórico	Discretização
Categórico	Numérico	Binarização
Texto	Numérico	Latent Semantic Analysis (LSA)
Séries Temporais	Séries discretas	SAX
Séries Temporais	Numérico Multi-dimensional	DWT, DFT
Imagem	Vetor de características	Redes profundas

Discretização



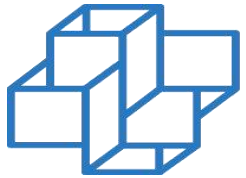
- Particiona o conjunto de valores de um atributo numérico em *n* intervalos
- O atributo passa a ser descrito por *n* categorias
- Exemplo:
  - Atributo: Idade
    - [0-11] Infância - símbolo : A
    - [12-18] Adolescência- símbolo: B
    - [19-60] Adulto – símbolo : C
    - [60 -] Idoso – símbolo: D
- Problema de predição passa a ser de classificação

Discretização – Equi-width



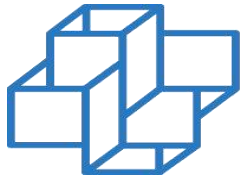
- Considera atributo com valores apresentando uma distribuição uniforme;
  - Considere um atributo com *N* valores, possivelmente duplos
  - Considere ainda o tamanho *B* de intervalos
    - Manter a ordem numérica
    - Total de elementos em cada intervalo => E= N/B
  - Ordenar o conjunto
  - Varrer o conjunto
    - Para cada elemento associá-lo ao intervalo corrente
    - Se total de elemento = E ; alterar intervalo corrente; zerar total elemento

Limpeza de Dados



- Fontes de distúrbio nos dados:
  - Dados capturados por sensores inerentemente imprecisos
  - Dados extraídos por scanners a partir de técnicas de reconhecimento de caracteres
  - Dados privados não fornecidos pelos usuários
  - Dados produzidos manualmente
  - Dificuldade na obtenção de todos os dados planejados

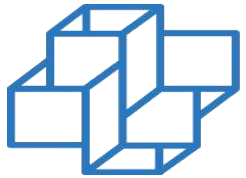
Limpeza de Dados



- Esses cenários levam aos seguintes problemas:
  - Tratamento de dados faltantes
    - Estimativa dos dados faltantes (imputação de dados)
  - Tratamento de dados incorretos
    - Tratamento de inconsistências entre fontes de dados heterogêneas
    - Utilização de informações complementares do domínio do atributo com dados faltantes
      - Por exemplo: altura máxima de uma pessoa
      - Detecção de *outliers*
  - Tratamento de escala e normalização
    - Dados expressos em escalas muito diferentes (ex. Idade e salário) levando a pesos enviesados entre as características
      - Importância em normalizer os dados
- *Holoclean: Holistic Data Repairs with Probabilistic Inference, Theodoros Rekatsinas, Ilhab Ilyas, Cris Ré*

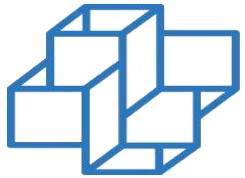
DEXL LAB
DATA EXTREME LAB

Tratamento de Dados Faltantes



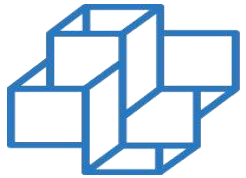
- Alternativas de tratamento
  - Eliminação de características com percentual de dados faltantes acima de um limite
  - Estimação ou Imputação dos valores faltantes
    - Ex: identificação de atributos correlacionados e uso da distribuição de valores do atributo correlacionado para geração dos valores faltantes
    - Adoção do mesmo processo de aprendizado por classificação. A coluna com valores faltantes passa a ser uma coluna alvo com valores a serem preditos
  - Defnição de estratégias analíticas e de predição que funcionem sob dados incompletos

Tratamento de Dados Incorretos e Inconsistentes



- Deteção de Inconsistência
  - Caso comum: dados fornecidos por múltiplas fontes autônomas
    - Deteção de duplicidade e inconsistência
    - Uso de pesos sobre credibilidade das fontes
  - Uso de conhecimento do Domínio
    - Por exemplo:
      - Relação de País e Cidade/Estado
      - Limítes de valores de atributos, por ex: Idade, altura, peso...
  - Métodos orientados a dados
    - Deteção de *outliers* a partir de informações estatísticas sobre a distribuição dos dados
      - *Outliers* nem sempre são problema. Podem ser comportamentos raros mas importantes a serem tratados
      - Técnicas: Dados Extremos, Clusterização; Distâncias (KNN maior que a de outros pontos; Teoria-da-Informação (limita o desvio a partir da distribuição esperada

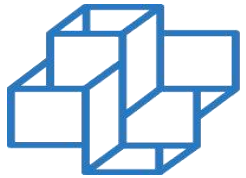
Tratamento de Escala e Normalização



- Em muitos casos a distribuição de valores entre atributos apresenta diferentes escalas de referência.
  - Ex: idade e salário
  - Medidas de distância entre os pontos multi-dimensionais serão enviesadas em função da variação na escala dos atributos
  - Padronização
    - $z_i^j = \frac{x_i^j - \mu_j}{\sigma_j}$
  - Escala Min-Max , mapeia atributos no interval [0,1]
    - $y_i^j = \frac{x_i^j - min_j}{max_j - min_j}$

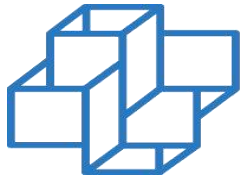
DEXL LAB
DATA EXTREME LAB

Redução de Dados e Transformação



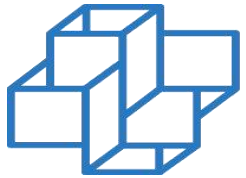
- Reduzir o volume de dados permite o uso de algoritmos mais complexos
- Pode-se reduzir as linhas (instâncias/observações) ou as colunas (características – dimensões)
- Redução de dados pode causar perda de informação
- Tipos de redução de dados
  - Amostragem
  - Seleção de características
  - Redução de dados por rotação de eixos: PCA (Análise de Componentes principais, …)
  - Transformações – DWT (Transformação discreta de wavelet)

Redução de Dados: Amostragem



- Considerando-se um dataset *D,* uma amostragem corresponde a um subconjunto de tamanho *n* extraído de *D*.
  - Amostragem com viés
    - Algumas partes dos dados são mais interessantes do que outras. Por exemplo, a importância dos dados decai com o tempo. A função de amostragem é aplicada sobre uma janela mais recente;
  - Amostragem por estratificação
    - Partes dos dados apresentam menos instâncias devido a sua raridade. Neste caso pode-se separar os estratos conhecidos e aplicar a amostragem internamente aos estratos.
      - Por exemplo, no treinamento para classificação de plantas a partir de imagens, algumas espécies raras aparecem com muito poucos exemplos. Estratificando os dados por Família ou Gênero pode-se fazer a amostragem nos grupos, evitando a ausência de exemplos de grupos menos frequentes.
  - Viés por densidade,....

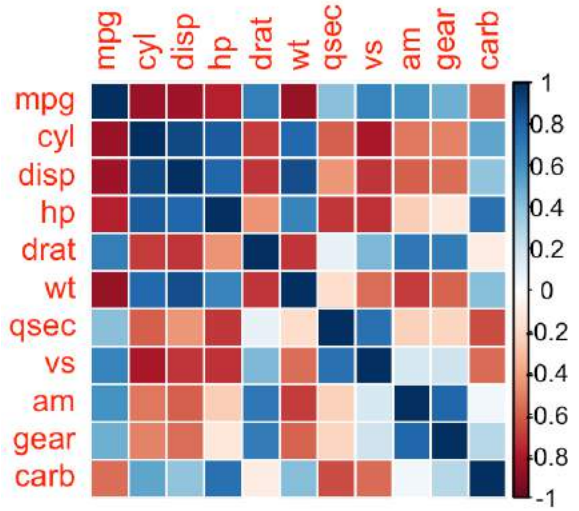
Seleção de observações iid



- i.i.d – exemplos a serem usados como amostragem parta treinamento devem ser Independentes e indivualmente distribuidos

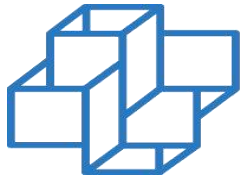
Redução de Dados: Seleção de subconjunto de Características

- Algumas características podem ser descartadas quando se mostram irrelevantes:
  - Eliminação de *features* reduz a complexidade do processo de treinamento
  - Deteção de correlação entre as características. Pode-se produzir matriz de correlação e eliminar características derivadas



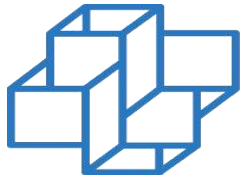
DEXL LAB
DATA EXTREME LAB

Redução: PCA



- Rotação dos dados para eixos em que a variância é observada com redução da dimensão dos dados

Comentários Finais



- A preparação dos dados para entrada do processo de treinamento influencia no resultado do treinamento
- Existem várias técnicas de pre-processamento
- O processo de preprocessamento é adhoc e evolve o conhecimento dos dados e sua análise
- Vejam o notebook o classroom da disciplina