Engineering, Test & Technology
Boeing Research & Technology

Machine Learning and Natural Language Processing for Cost Analysis

Karen Mourikas, Jose Lemus, Enrique Serrot

International Cost Estimating & Analysis (ICEAA) Workshop May 2019

Machine Learning Track ML05

Machine Learning and Natural Language Processing for Cost Analysis

Machine Learning and Natural Language Processing

- Machine Learning Overview
- Introduction to Natural Language Processing

Application

- Process-based Cost Analysis

Going Forward

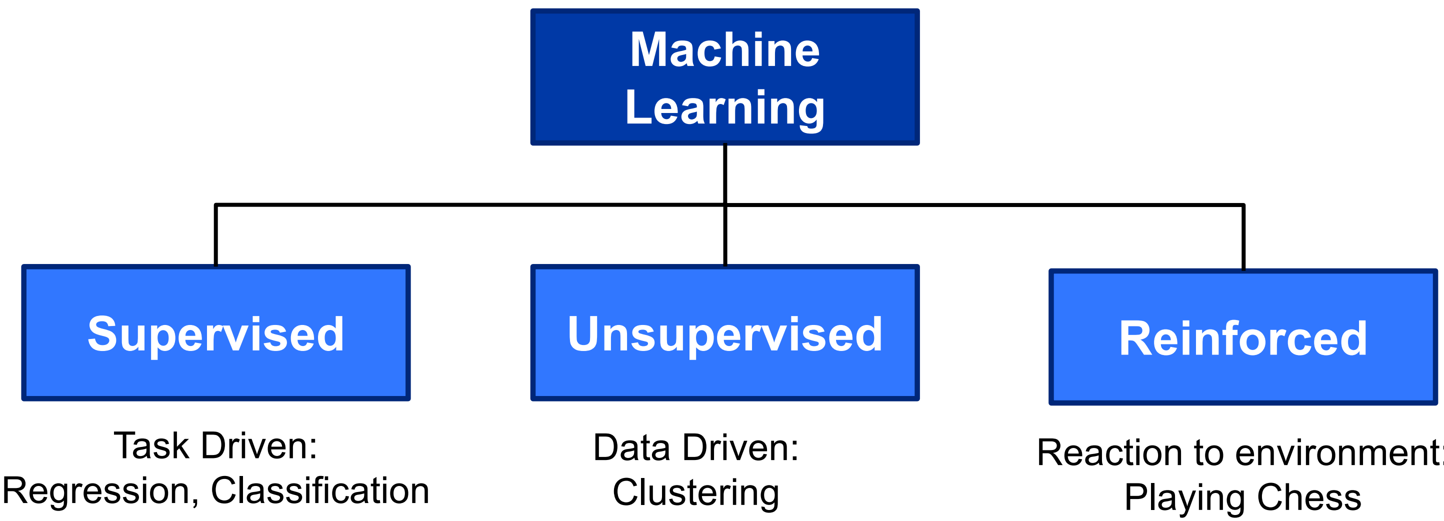
- Challenges

What is Machine Learning?

Simply, **when a machine mimics "cognitive" functions such as "learning" and "problem solving" ***

Machine Learning (ML) is a method in which algorithms …

- teach themselves to grow (i.e. learn) from data
- learn without being explicitly programmed



Machine Learning is a type of Artificial Intelligence

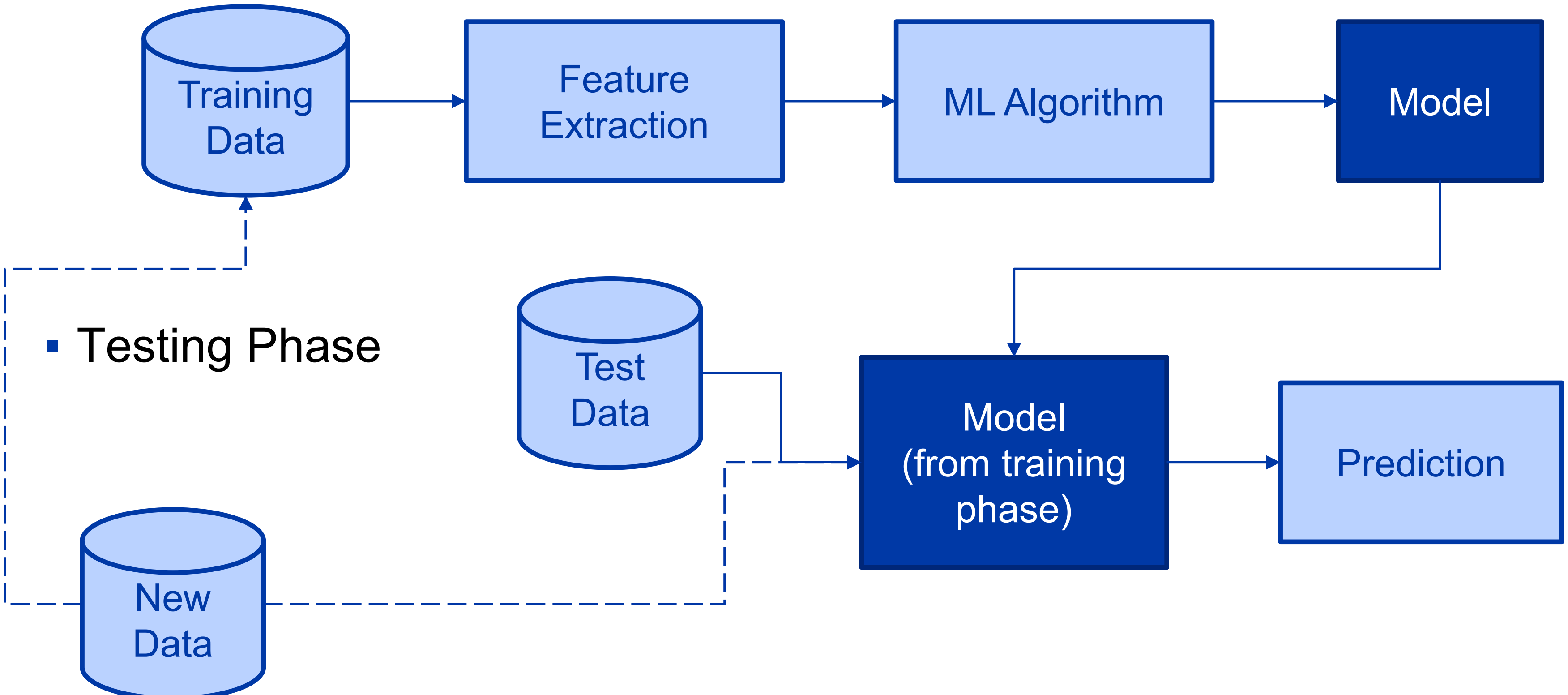
* Russell, Stuart J.; Norvig, Peter ; Artificial Intelligence: A Modern Approach, 2003 & 2009

How does Machine Learning Work?

Typically consists of two stages

- Training phase

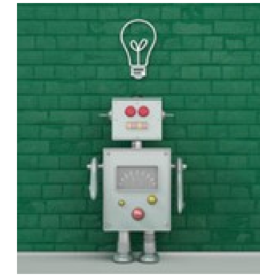
- Testing Phase



Note similarity with Statistical Analysis

General Machine Learning Process

Machine Learning Vs. Statistics – Pt 1

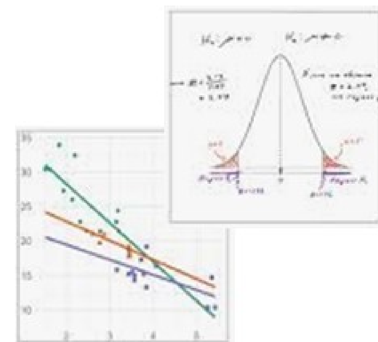


"Machine learning is … "

"glorified statistics"

"statistics scaled up to big data"

"statistics minus any checking of models and assumptions"



"Machine learning is for Computer Science majors …

who couldn't pass a Statistics course"

"There is no difference"

"I don't know what Machine Learning will look like in ten years, but whatever it is …

I'm sure Statisticians will be whining that they did it earlier and better"

"Machine Learning is teaching computers to do Statistics with tons of data"

"The difference…is not one of algorithms or practices but of *goals* and *strategies*."

Public opinion

Machine Learning Vs. Statistics – Pt 2

Highly related, with similar mechanisms

but … with different purposes

- Machine Learning
  - Make most accurate predictions possible
- Statistics
  - Make inferences about relationships between variables

Both attempt to make sense of data

- "The math is the same, but the point of view is completely different"

	Machine Learning	Statistics
Characteristics	Accurate Predictions	Inferences / Relationships
	Computer Science / AI	Mathematics
	Real-world algorithms for practical problem	Mathematical foundation for scientific research
	Test (new) data	Diagnostic tests
	Not needed	Prior Assumptions
	More data / higher dimensions	Less data / lower dimensions
	Matlab / Python	R
Terminology	Inputs/outputs	Data points
	Features	Variables
	Label	Response
	Feature Creation	Transformation

More Public Opinion

Similar but with different points of view

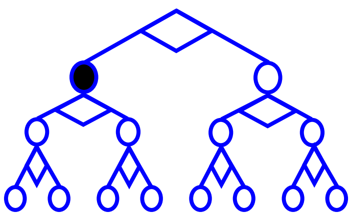
Natural Language Processing

Natural Language Processing (NLP) is a method in which computers …

- Analyze large amounts of “natural” language data
- Understand how humans communicate
- Make sense of human language

Algorithms convert unstructured text into computer-readable data

- Early use: “Hard-coded” If-then rules or patterns
- More recently, ML enabling Generalization



Analysis of words and phrases … by a computer

Natural Language Processing Tasks

Unstructured Text to Usable (Formatted) Data

- Tokenization – break text into pieces that a computer can understand
- Part of Speech (PoS) Tagging – label words (noun, verb, adjective, …)
- Parsing – break sentence into grammatical phrases
- Sentence Chaining – connect related sentences to a topic

Preprocessing the data / Cleansing the text

- Remove punctuation
- Remove Stopwords
- Determine Stem / Root form
- Vectorize (Bag-of-Words, n-grams, Term Frequency)
- Perform Feature Creation

"I washed his mouth out with soap"

Preprocessing, a necessary, but time-consuming effort

Natural Language Processing for Cost Analysis?

Provide context to cost data

- Datasets may contain free-form textual data
  - Descriptions, operations, caveats, sequences, notes
  - More than categorical pre-defined terms
- Examples show different formats
  - With varying degrees of detail

Examples of Cost Reduction Ideas*

Idea	Savings
Alternative fabric	5%
Change material of housing	2%
Use of updated xyz transistors	12%
Could design be simpler with fewer components? Specifically, the bars	15%

*notional

Manufacturing Operations Costs*

Deep draw	1.89
Cut out bottom	0.16
Drill holes	0.95
Inspect	4.75
Cut sheet metal to size	0.17
Deep draw height	13.78
Cut out bottom	10.34
Inspect	2.01
Rough mill	1.02
Rough mill slot	0.29
Finish mill	21.35
Drill 6 holes & deburr	20.95

*notional

Natural Language Processing provides additional context to cost analysis

Machine Learning & Natural Language Processing

Machine Learning *and* Natural Language Processing

- To identify & tag parts of speech
- To determine sentiment (positive, negative, neutral)
- To categorize or cluster data into similar groups

Supervised and Unsupervised Methods

- Supervised: Labeled inputs & outputs
  - Text documents are tagged and used to train a model
  - Categories often pre-determined
- Unsupervised: No labels
  - Data grouped into clusters to extract meaning
  - Categories not specified

ML & NLP to generate insights from unstructured text data

Application: Process-based Cost Analysis

Objective

- Predict production costs of products based on manufacturing data

Data

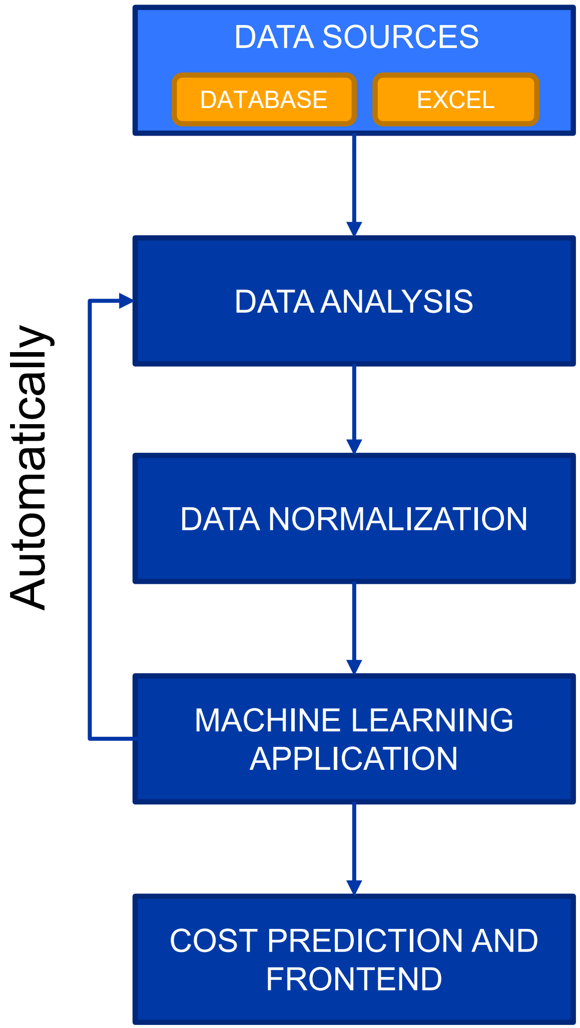
- 40,000 data points as categorical, free-form, and quantitative data
  - Process steps, material types, machines, cycle times, cost, operators
  - Lots of free text
    - Description of processes or materials, specifications, standards, limitations, qualifiers, build/assembly plans, tech notes, additional comments
    - Inconsistencies: Typos, abbreviations, styles, categories, groupings

Analysis Approach

- Combine Machine Learning and Natural Language Processing
  - To cleanse, analyze, preprocess, train, and predict

Cost Analysis using Machine Learning & Natural Language Processing

Machine Learning Data Analysis Approach



Extract the data

- Various Sources – Need for aggregation
- Python NLP Toolkit

Understand the data

- Cleanse the data

Prepare the data for Machine Learning

- Preprocess data: convert to computer-readable format
- Identify Candidate Features & Targets

Apply Machine Learning

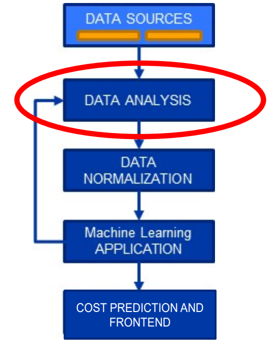
- Select algorithm(s); Eliminate non-significant features
- Train, test and validate model

Predict in real-time

- With user-friendly front end

Data Analysis & Normalization steps extremely important

Understand the Data (1 of 2)

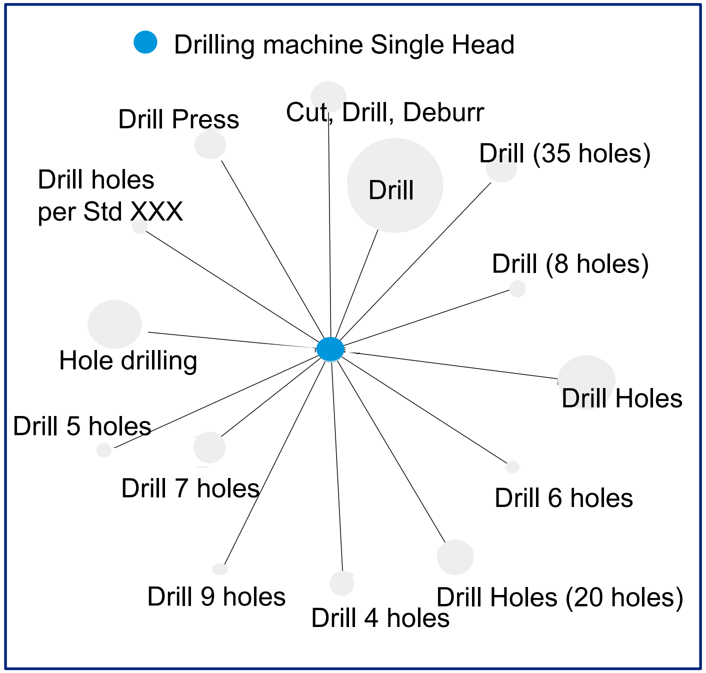


What does our Text Data* look like?

Manufacturing Operations

Text Inconsistencies Examples	
“ Visual Inspection * ”	
Visual Insp	
Visial Inspect	
AbRasiVe CLeaN oPeRation	
Drill vs	Drill 5 holes (D-.541, depth-.45)
Check vs	Verify
Two vs	2
Mask	Mask - Unmask
Unmask	

Operations Grouping



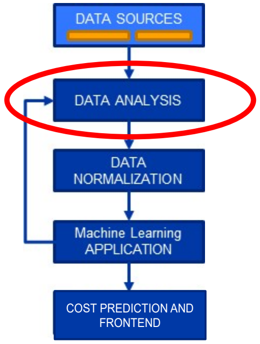
What data preparation needed?

- Data preparation huge effort
  - Typically 80% of total effort
  - Manual approach not feasible for large datasets

* Presentation focuses on Manufacturing operations data

Need for automation to cleanse and prepare text data in large datasets

Understand the Data (2 of 2)



Cleanse data

- Numeric, Categorical, Free-form Text Data
- Punctuation, Stop words, Roots, Vectorization

Group individual parameters to create new features*

- Materials & quantity used associated with each Manufacturing Process Step

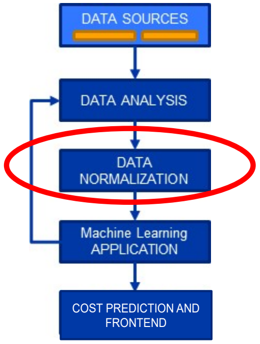
Determine significance & context of text fields

- Additional information extracted from free-text fields
  - Material standards, Specifications
  - Restrictions, Tech notes

*ML terminology

Data cleansing & Preparation key

Data Preparation (1 of 3)



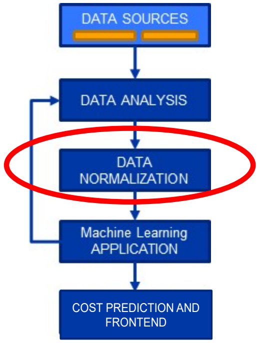
Prepare data for the Machine Learning model

- Target* (response) variables
  - Costs of Machinery, Tooling, Batch Setup
- Initial features* (Feature engineering)
  - 65 Potential Predictors
  - Requires domain knowledge
- Transformation into ML-readable fields
  - Categorical variables
    - One-hot encoded, binned or binarized
  - Obscure numerical variables
    - Categorize

*ML terminology

Data Preparation is time consuming – Allow enough time

Data Preparation (2 of 3)



Normalization: Matching Terms

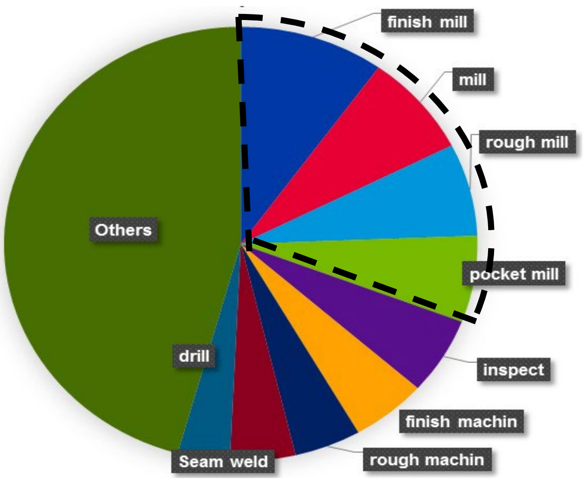
- Sequence matching from description
  - Sequence Matcher
    - 90% coincidence ~ 15% reduction
    - 80% coincidence ~ 23% reduction

Parent	Children
Visual Inspection	Visual Inspection
	Visial Inspection
	Visual Inspect
	Final Inspection
	Final visual inspection

- Stem extraction from description
  - Find the root of the words
    - ~ 4% reduction

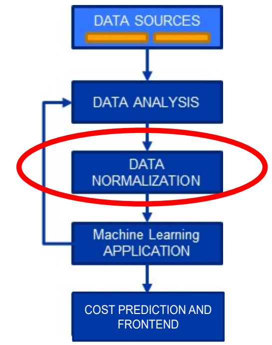
Parent	Children
visual inspect	Visual Inspection
	Visual Inspect
part mark	Part Mark
	Part marking

- Cognitive Grouping



What terms belong together?

Data Preparation (3 of 3)



Normalization: Feature Transformation

- One-hot encode: each word represented as unique binary vector
  - 283 materials → becomes 283 categorical variables

Material
'Aluminum','Primer'
'Titanium'

1	2	3	…283
Material: Aluminum	**Material: Primer**	**Material: Titanium**	
1	1	0	
0	0	1	

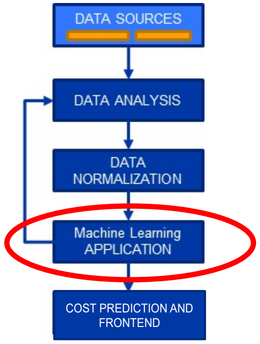
- Binarize encoding: each word represented as a binary bitstring
  - 128 machines → becomes 7 categorical variables ($2^7 = 128$)

Machine	BIN
ANODIZING LINE	000
ASSEMBLY WORKPLACE, LARGE	001
AUTOCLAVE FURNACE 1M³	010

1	2	3	…7
Machine: bit2	**Machine: bit1**	**Machine: bit0**	
0	0	0	
0	0	1	
0	1	0	

Transform data to optimize analysis

Machine Learning Application (1 of 4)



Resulting Dataset after cleaning & pre-processing

- Reduced from ~40K to ~15K observations (inputs/outputs)*
- Data shortage (!!!) limits the predictive capability of the model

Feature* Selection

- 65 Potential Predictors converted into 425 features
  - Reduced to 15 most relevant features (human-understandable)
- Initial selection process time consuming (> 24 hrs)
  - Subsequent updates < 1 hr

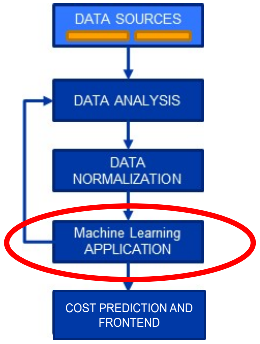
Data Splitting

- 80% Train & Validate the model
- 20% Test the model

*ML terminology

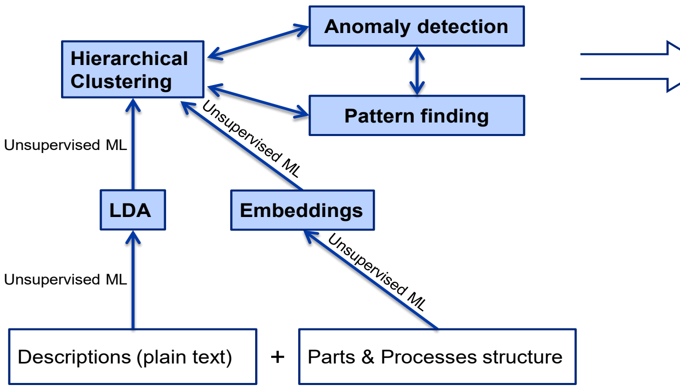
Resulting dataset for ML training

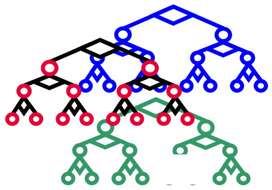
Machine Learning Application (2 of 4)



Machine Learning Methods

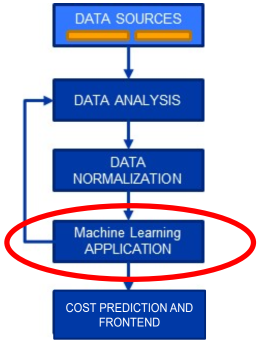
- Hybrid approach to analyze our data
- Unsupervised: new information based on data relations
  - Latent Dirichlet Allocation for Text analytics & grouping
  - Clustering for aggregation of operations, material, and machines
  - Association for sequences of operations
  - Anomaly detection for inconsistencies or deviations
- Supervised: new information on labeled data
  - Random Forest for variable importance
  - XGBoost for cost prediction





Different ML methods for different analytical purposes

Machine Learning Application (3 of 4)



ML Algorithm Selection for cost prediction

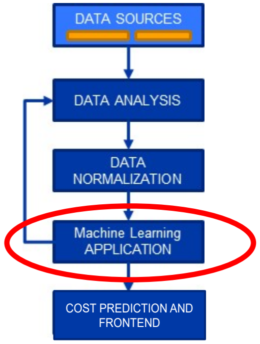
- Linear, non-linear, ensemble regression models
  - 700 different "models" evaluated via
    - Mean Square Error (MSE), Mean Absolute Error (MAE), and R^2

Linear	Non-linear	Ensembles
LinearRegression	KNeighborsRegressor	AdaBoostRegressor
Lasso, Ridge	DecisionTreeRegressor	BaggingRegressor
ElasticNet	ExtraTreeRegressor	RandomForestRegressor
HuberRegressor	SVR(kernel='linear')	ExtraTreesRegressor
Lars, LassoLars	SVR(kernel='poly')	GradientBoostingRegressor
PassiveAggressiveRegressor		XGBRegressor
RANSACRegressor		
SGDRegressor		
TheilSenRegressor		

  - XGBoost Regressor selected as best fit
    - Faster run time and better accuracy

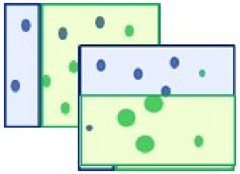
Choice of many algorithms – May use more than one

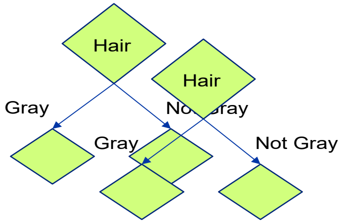
Machine Learning Application (4 of 4)



XGBoost: eXtreme Gradient Boosting

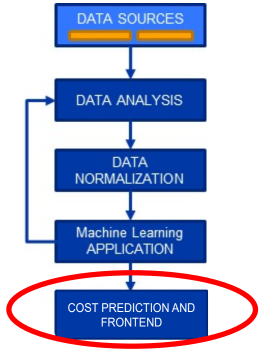
- Gradient Boosted Tree
  - Very fast implementation
- Ensemble technique of simple models
  - New models correct the errors of existing models
  - Each subsequent tree trained to improve upon the errors (residuals) of the previous tree(s)
- Resulting model is the "final" tree
  - Combines weak models into a single strong model iteratively





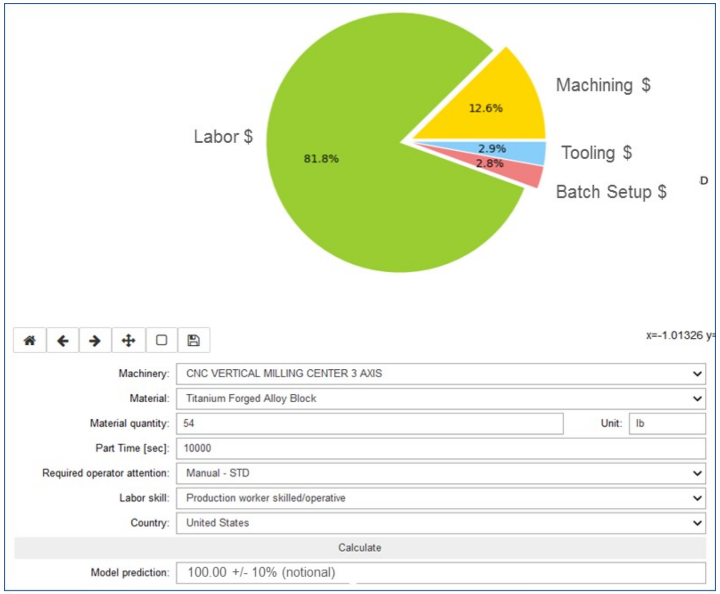
Training the model

Cost Prediction and Frontend Step



Automate Processes for training, updating, and running model

- Include in interactive Jupyter notebook
  - Data (raw and processed), python code
  - Analysis description (objectives, methods) and results (graphs, tables, numbers)
- Ability to execute code
  - Clean, Normalize, Preprocess, Binarize
  - Split Data, Train, Test, Validate, Predict
- Easy to train new model within notebook



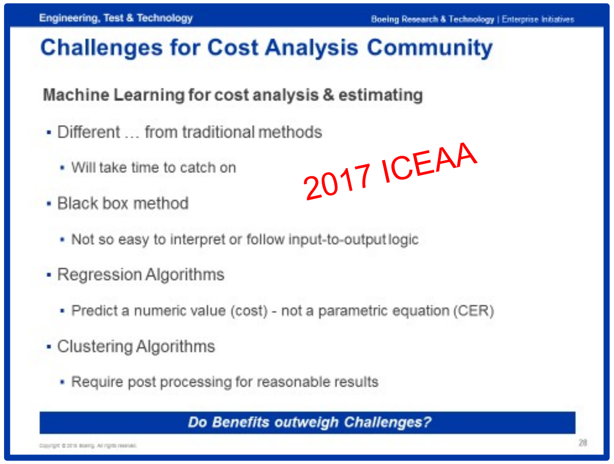
User Front End

- Embedded in the same Jupyter notebook
- Limited inputs
- Predicts costs with immediate results

Easy-to-use Cost Prediction Methodology

Challenges for Cost Analysis Community

Previously identified challenges



- Different from "traditional" approaches
- Black box method
- Requires pre and post processing

… plus more

Misconceptions about Machine Learning & Natural Language Processing

- It's the answer to all data-related questions
- Why does it take so long just to prepare the data?
- Anybody can do it

It's our job to educate & promote new Cost Analysis methods

Authors

Karen Mourikas is an Associate Technical Fellow at The Boeing Company specializing in Operations Analysis, Affordability, and Systems Optimization. Her current work includes Production Systems Cost & MBSE modeling, Product Teardown & Optimal-cost analyses, involving machine learning and natural language processing, and Affordability analyses. Karen has MS degrees in Applied Math and in Operations Research Engineering from the University of Southern California. Karen is a life-time member of ICEAA, has presented at several ICEAA & ISPA/SCEA conferences and was the recipient of the ICEAA 2018 Technical Achievement of the Year Award.

Jose Lemus is a Systems Engineer at The Boeing Company specializing in electronics engineering. His current work includes Industrial Computed Tomography Dimensional analysis and Product Teardown and Optimal cost analyses, involving machine learning and natural language processing. Jose has a MS degree in Telecommunication Engineering from the Polytechnic University of Madrid.

Enrique Serrot is a Systems Engineer at Boeing Research & Technology Europe. His current assignment includes Model Based Production Engineering and Data Analytics in the field of Optimal cost. Enrique has a bachelor degree in Aeronautical Engineering from the Polytechnic University of Madrid.

Contributors:

- Sergio Rodil, Enrique Garcia, Lucas Mengual, Clarisa Martinez: Tessella, Altran Group, Madrid Spain

Note: Graphics are from Getty Images or internally developed

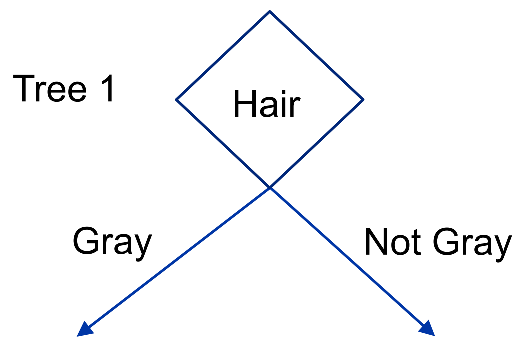
karen.mourikas@boeing.com *jose.l.lemus@boeing.com* *enrique.e.serrot@boeing.com*

XGBoost Tree Illustrative Example

Back

We want to predict a person’s age based on the data below

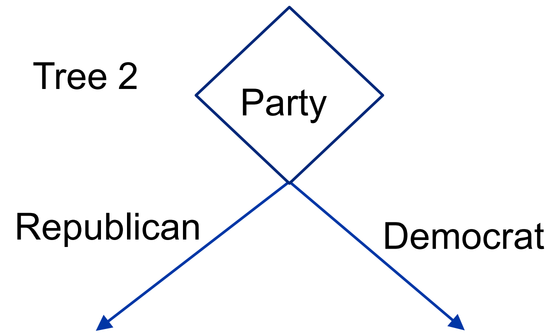
- Let’s build some trees



{44,56,61,73,78} {18,27,33,39,55,57}

62 ← Predicted age → 38

Age	Gray Hair	Party	Residual Tree1
18	N	D	-20
27	N	R	-11
33	N	D	-5
39	N	D	1
44	G	R	-18
55	N	R	17
56	G	D	-6
57	N	D	19
61	G	D	-1
73	G	R	11
78	G	R	16



{27,44,55,73,78} {18,33,39,56,57,61}

55 ← Predicted age → 44

{-11,-18,17,11,16} {-20,-5,1,-6,19,-1}

3 ← Predicted Residual → 2

New Observation:
Gray Hair, Republican

Average from each tree 58.5

Combined Results from all trees 65

Assumptions:
1 Level Deep
2 Trees total

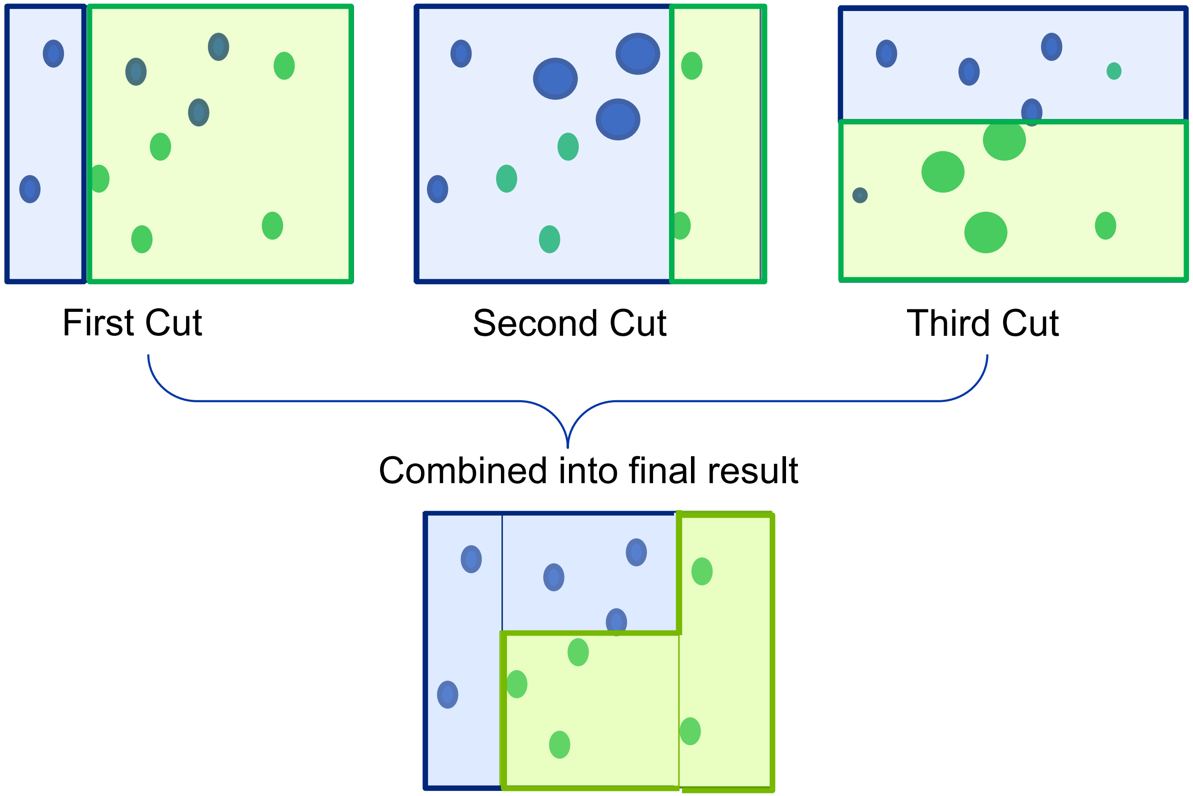
XGBoost: subsequent trees based on residuals of previous trees

Boosting Illustrative Example

Back

Goal: Separate Dots by Color

- Incorrectly classified dots have higher weighting in the next round
- Correctly classified dots have lower weighting



New models correct errors of existing models