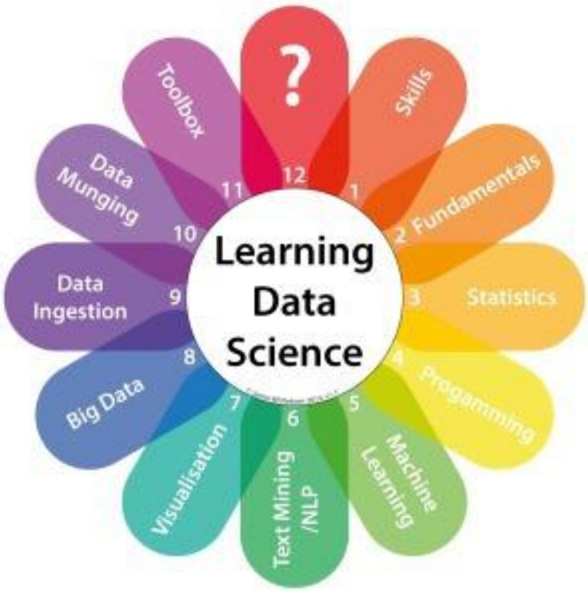
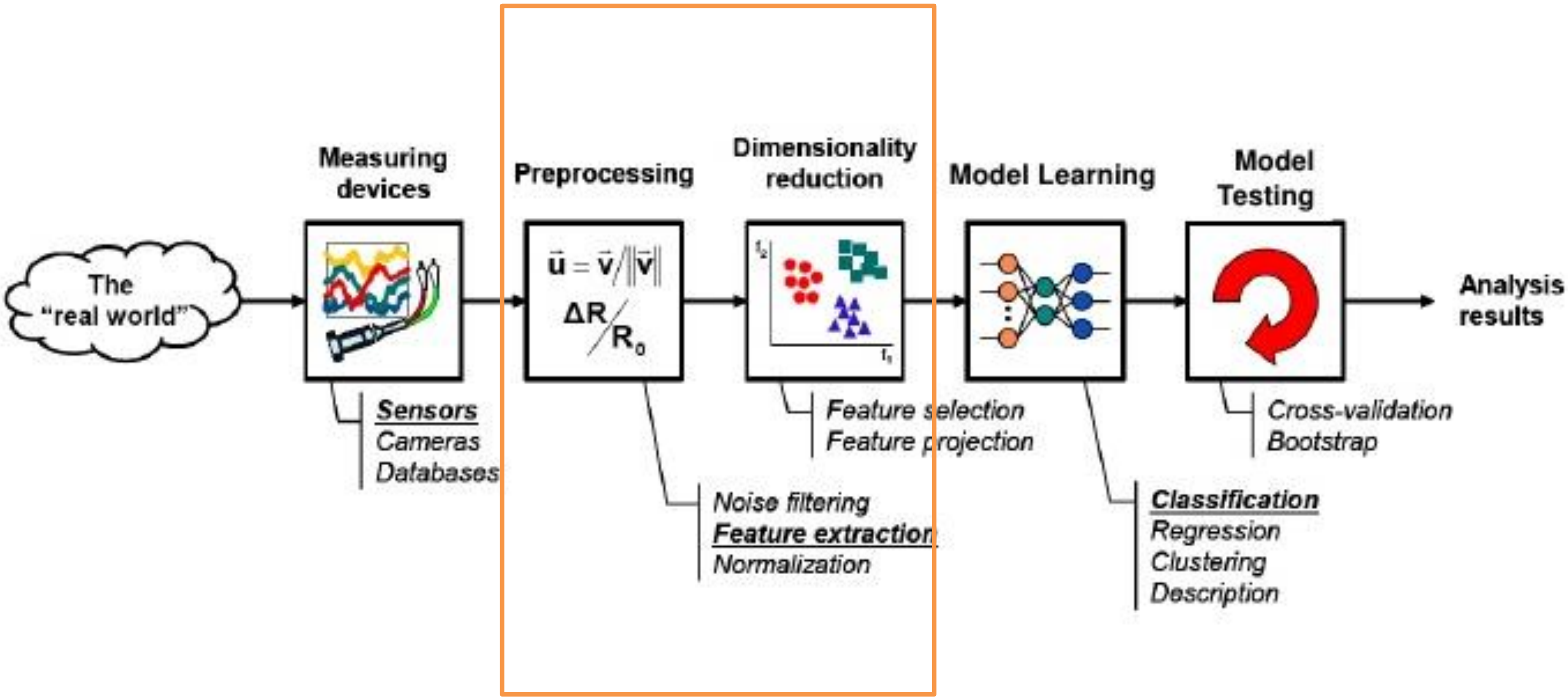


Feature Engineering

NYDSSG Demo By: Susan Sun
February 1, 2016 @Dstillery

Feature Engineering vs. Feature Selection



What is Feature Engineering?

Feature engineering is the process of using **domain knowledge** of the data to create features that make machine learning algorithms work.

Why Talk about Feature Engineering?

Feature engineering is fundamental to the application of machine learning. It is **manual**, it is **slow**, it requires a lot of human brain power, and it **makes a big difference**, especially in competitive machine learning.

Let’s Feature Engineer a Model for the Question…

Should I have another cup of tea?

Numerical Data

Person	Date & Time	Amount of Caffeine in My System (in mg)
Susan	Day 1	400
...	Day 2	600
...	Day 3	400
...	Day 4	2400
...	Day 5	400
Person B	Day 1	600
...	Day 2	400

Person	Average Caffeine Amount (in mg)
Susan	840
Person B	500

Numerical Data

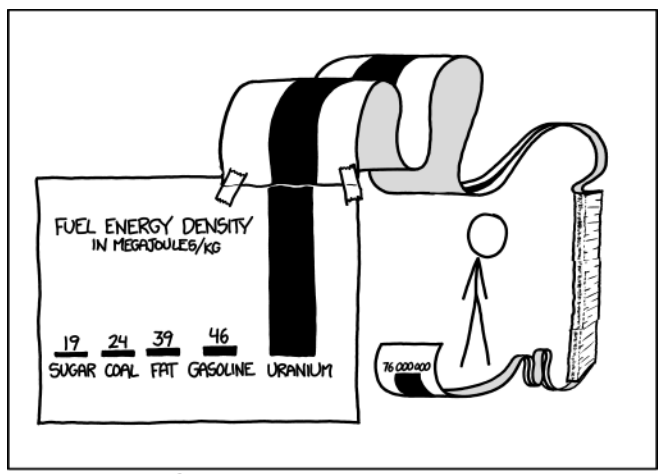
Instead of just the **AVERAGE**, also consider:

Descriptive Statistics	Min, Max, Median, Mode, Variance
Transformations	Square, Cube, Log, Inverse
Standardizations	Capping, Binning, Normalization, Ratio

Numerical Data – Food for Thought

1. Instead of transforming each numerical variable with every transformation methodology, it helps to stop and think which transformations make sense.

2. Mean and variance are the 1st and 2nd "moments" in descriptive statistics. Transformations and standardizations are used for controlling skewness, which is the 3rd "moment."



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY.*

Categorical Data

Person	Date & Time	Allergic to Caffeine?	Was it Decaf?	What Kind of Tea?
Susan	Day 1	No	No	Earl Grey
...	Day 2	No	No	Green
...	Day 3	No	No	Chamomile
...	Day 4	No	No	Earl Grey
...	Day 5	No	No	Earl Grey
Person B	Day 1	Yes	Yes	Chamomile
...	Day 2	Yes	Yes	Chamomile
Person C	Day 1	Yes	Yes	Chamomile
...	Day 2	Yes	No	Earl Grey

Notice the interaction of the variables here. Should someone allergic to caffeine be drinking caffeinated tea?

Person	Tea_EarlGrey	Tea_Green	Tea_Chamomile	Flag_Allergic	Flag_DrankDecaf	Flag_Allergic_DrankDecaf
Susan	3	1	0	0	0	0
Person B	0	0	3	0	1	0
Person C	1	0	1	1	1	1

Categorical Data Summary

Tea_*	Example of aggregation of categories
Flag_*	Example of dummy variables
Allergic_DrankDecaf	Example of interaction variables

Categorical Data – Food for Thought

1. If the data is ordinal instead of nominal, find some way to preserve the order information.

2. Missing data can also be transformed into binary variables.

3. Always do a check of how many distinct values the categorical variable has, before performing interactions with other categorical variables, to prevent variable explosion.

Machine Learning: The Art and Science of Algorithms that Make Sense of Data by Peter Flach

p.304

Table 10.1: Kinds of feature

Kind	*Order*	*Scale*	*Tendency*	*Dispersion*	*Shape*
Categorical	×	×	mode	n/a	n/a
Ordinal	√	×	median	quantiles	n/a
Quantitative	√	√	mean	range, interquartile range, variance, standard deviation	skewness, kurtosis

Kinds of feature, their properties and allowable statistics. Each kind inherits the statistics from the kinds above it in the table. For instance, the mode is a statistic of central tendency that can be computed for any kind of feature.

Machine Learning: The Art and Science of Algorithms that Make Sense of Data by Peter Flach

p.307

Table 10.2: Feature transformations

↓ to, from →	*Quantitative*	*Ordinal*	*Categorical*	*Boolean*
Quantitative	**normalisation**	**calibration**	**calibration**	**calibration**
Ordinal	**discretisation**	**ordering**	**ordering**	**ordering**
Categorical	**discretisation**	**unordering**	**grouping**	
Boolean	**thresholding**	**thresholding**	**binarisation**	

An overview of possible feature transformations. **Normalisation and calibration** adapt the scale of quantitative features, or add a scale to features that don't have one. **Ordering** adds or adapts the order of feature values without reference to a scale. The other operations abstract away from unnecessary detail, either in a deductive way (**unordering**, **binarisation**) or by introducing new information (**thresholding**, **discretisation**).

Date and Time

Person	Date & Time of Caffeination	Year	Month	Day	Season	Time	Sunset?
Susan	2015-12-20 02:00:00 EST	2015	12	20	Winter	2:00 AM EST	Yes
...	2015-12-20 03:30:00 EST	2015	12	20	Winter	3:30 AM EST	Yes
...	2016-01-01 04:15:00 EST	2016	1	1	Winter	4:15 AM EST	Yes
...	2016-01-27 01:00:00 EST	2016	1	27	Winter	1:00 AM EST	Yes
...	2016-01-28 02:00:00 EST	2016	1	28	Winter	2:00 AM EST	Yes

Person	Metric 1	Metric 2	Metric 3	...
Susan				

Date and Time

“Duration since last action”
e.g. How many days / hours / seconds since last action [tea]

“Gap measure”
e.g. How many days / hours / seconds between two actions?

“Seasonality”
e.g. Does the time of day / time of month / time of year affect how frequently the action takes place?

Date and Time – Food for Thought

Be conscious of how the interaction of geography and timestamps will affect your calculations. (e.g. February in South America is NOT the same season as February in North America)

Text

Recorded Conversation

Man, I really want some caffeine right now.

I want some tea. Do you want some tea? Let me get you some tea.

Ugh! No more tea, I'm so caffeinated I'm vibrating.

I'm going to die if I don't have some caffeine. Who the hell broke into my tea stash?!!

TEAAAAAAA!!!!!!!!!!!!

Person	Count # of Times "Tea" was Mentioned	Count # of Times "Caffeine" was Mentioned
Susan	...	...

Text

Recorded Conversation

Man, I really want some caffeine right now.

I want some tea. Do you want some tea? Let me get you some tea.

Ugh! No more tea, I'm so caffeinated I'm vibrating.

I'm going to die if I don't have some caffeine. Who the hell broke into my tea stash?!!

TEAAAAAAA!!!!!!!!!!!!

⇩

Stemming & Lemmatization

"Caffeine"

Variations on the Same Word

caffeine, caffeination, caffeinated

Text

Recorded Conversation

Man, I really want some caffeine right now.

I want some tea. Do you want some tea? Let me get you some tea.

Ugh! No more tea, I'm so caffeinated I'm vibrating.

I'm going to die if I don't have some caffeine. Who the hell broke into my tea stash?!!

TEAAAAAAA!!!!!!!!!!!!

N-grams	Count # of Times [x] was Mentioned
1-gram	"tea"
2-gram	"more tea", "some tea"
3-gram	"no more tea"

Text – Food for Thought

1. How to handle typos or deliberate misspellings? (Damerau-Levenshtein distance for typos, Double Metaphone for spelling)

2. Instead of manually picking keywords, off-the-shelf packages for tokenization can also produce passable results (Python's NLKT package).

Other Types of Data

Categorical data with too many levels

a. Convert to numerical whenever possible (Zip code: use Google Geocoding API or Yahoo! PlaceFinder)

b. Aggregate to higher levels (Zip code: aggregate to province by taking first few numbers)

Other Types of Data

Use model outputs for your model's input

a. Image processing

b. Including trend analysis as an input

What is Good Feature Engineering?

A well-behaved feature should be…

Reusable: You should be able to reuse features in different models, applications, and teams.

Transformable: Besides directly reusing a feature, it should be easy to use a transformation of it (e.g. log(f), max(f))

Interpretable: In order to do any of the previous, you need to be able to understand the meaning of features and interpret their values.

Reliable: It should be easy to monitor and detect bugs/issues in features

And Finally…

Feature engineering is an art form, not a science. (aka: Don't torture your data!)

Thank You!

@susanweisun

https://www.linkedin.com/in/susanwsun