

# End2End ML

Lior Sidi & Noa Lubin



# What you will learn today

Formulating a business problem to a machine learning one  
Best practices in data science

Creating real impact with your models by scoping it

Evaluating and hypertunning models  
Understand your models and improving them

How to become a Lego-Master-Builder



# What you will **NOT** learn today

Coding best practices (but you might see some nice python tricks)

MLOps and model deployment

Scaling learning on big datasets

Learn how to do nice plots

New algorithm



# You are a...

A Data scientist who works in a big electronic online retailer (i.e Best Buy).  
Millions of users enter the company sites daily.

KPI is to improve sales on the site:

- Improve experience
- Improve offerings



# You build superpowers

- Improve experience
  - Product recommendation
  - Chatbot & Sentiment analysis
  - Improve sites navigation
- Improve offerings
  - Optimize price with discounts, coupons and bundles
  - Identify new products to sell
  - Inventory prediction



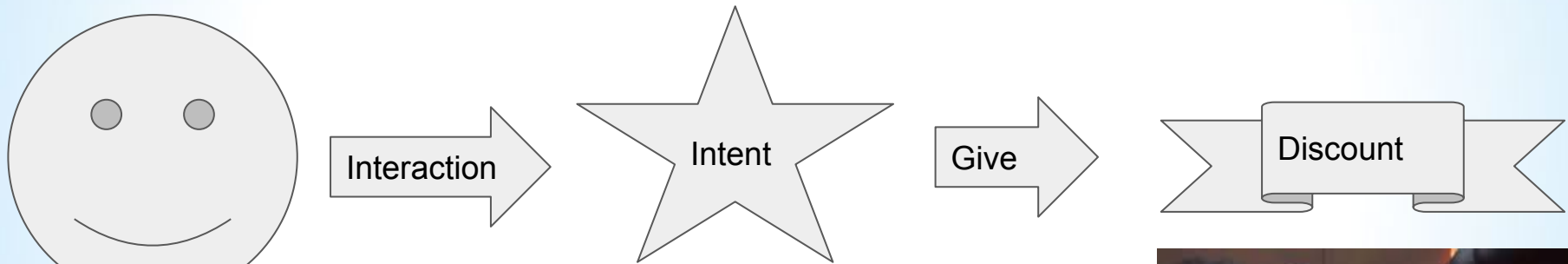
# Why personalization is important?

Each customer has different:

- Products needs
- Budget
- Intent



# What is intent?



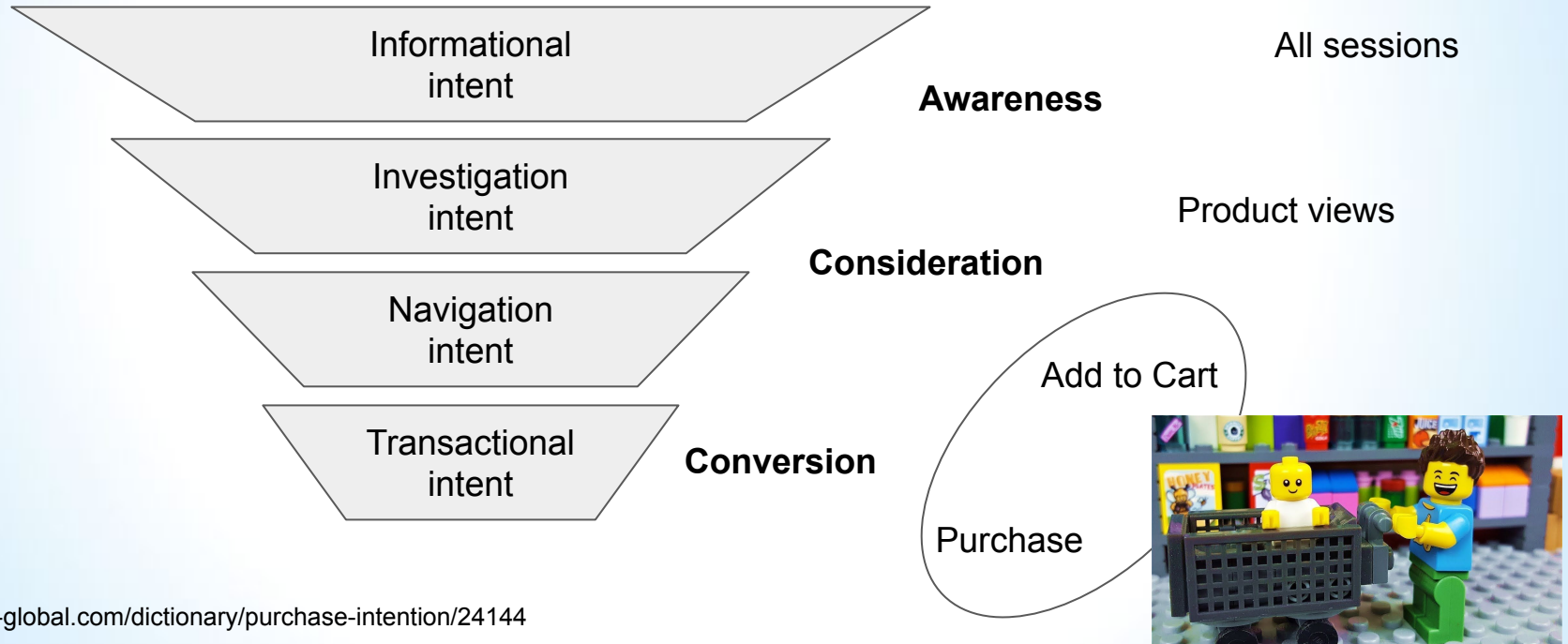
Customer





# Customer Purchase Intent

*The willingness of a customer to buy a product or service in a certain condition\**



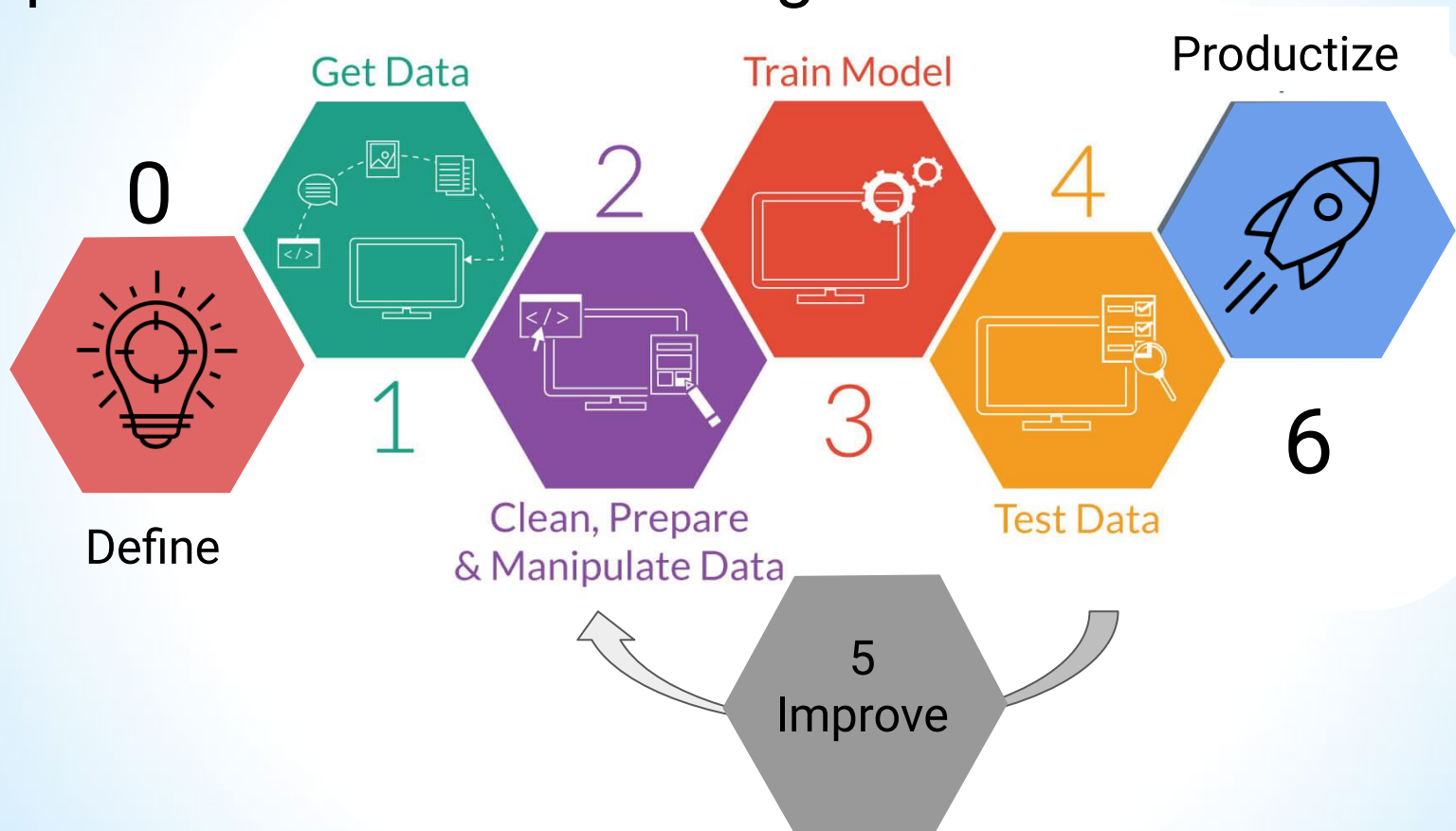
\*<https://www.igi-global.com/dictionary/purchase-intention/24144>



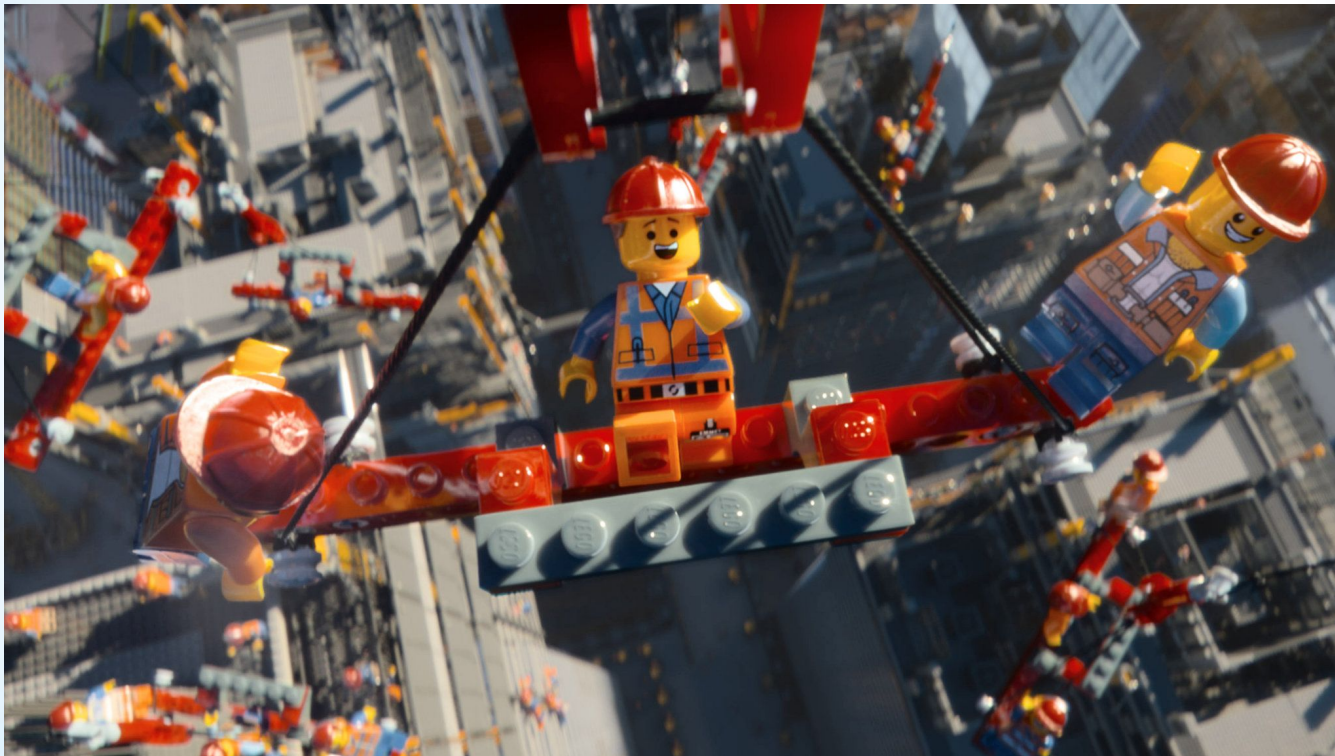
# How Customer interaction looks like?

user_session	user_id	event_time	category_code	product_id	brand	price	event_type
CxMKMQDRAN	1515915625610973155	2021-02-28 19:13:22 UTC	computers.peripherals.printer	3829912	pantum	122.86	view
CxMKMQDRAN	1515915625610973155	2021-02-28 19:15:40 UTC	computers.peripherals.printer	3829913	pantum	116.05	view
CxMKMQDRAN	1515915625610973155	2021-02-28 22:56:17 UTC	computers.peripherals.printer	3829913	pantum	116.05	view
CxMKMQDRAN	1515915625610973155	2021-02-28 23:01:14 UTC	computers.peripherals.printer	500058	pantum	67.00	view
CxMKMQDRAN	1515915625610973155	2021-02-28 23:02:38 UTC	computers.peripherals.printer	500058	pantum	67.00	view
CxMKMQDRAN	1515915625610973155	2021-02-28 23:03:42 UTC	computers.peripherals.printer	500058	pantum	67.00	cart
CxMKMQDRAN	1515915625610973155	2021-02-28 23:08:57 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:20:48 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:23:11 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:26:07 UTC	computers.peripherals.printer	500058	pantum	67.00	purchase
CxMKMQDRAN	1515915625610973155	2021-02-28 23:43:24 UTC	computers.peripherals.printer	3829912	pantum	122.86	view

# Steps to Predictive Modeling



# Let's Start!!!!



# Why this is happening?

- “Kaggle mindset” - ready data, just fit predict
- Focus on modeling and not data and product
- Non skeptical about problem and data

# Why this is happening?

- “Kaggle mindset” - ready data, just fit predict
- Focus on modeling and not data and product
- Non skeptical about problem and data

## How to solve it?

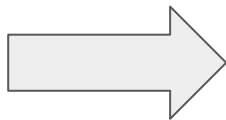
- Gain EXPERIENCE
- Be skeptical
- Work with Product definition

# Why this is happening?

- “Kaggle mindset” - ready data, just fit predict
- Focus on modeling and not data and product
- Non skeptical about problem and data

## How to solve it?

- Gain EXPERIENCE
- Be skeptical
- Work with Product definition



**Do Exploratory Data Analysis**

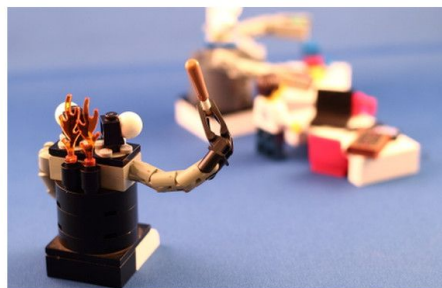


# What we want to achieve in EDA

Understanding what and when we calling the model (Prediction point)

Understanding how that data and features behave - do we have signal?

Identify possible pitfalls - evaluation, leakage, noise, imbalance



DATA



SORTED



ARRANGED



PRESENTED  
VISUALLY

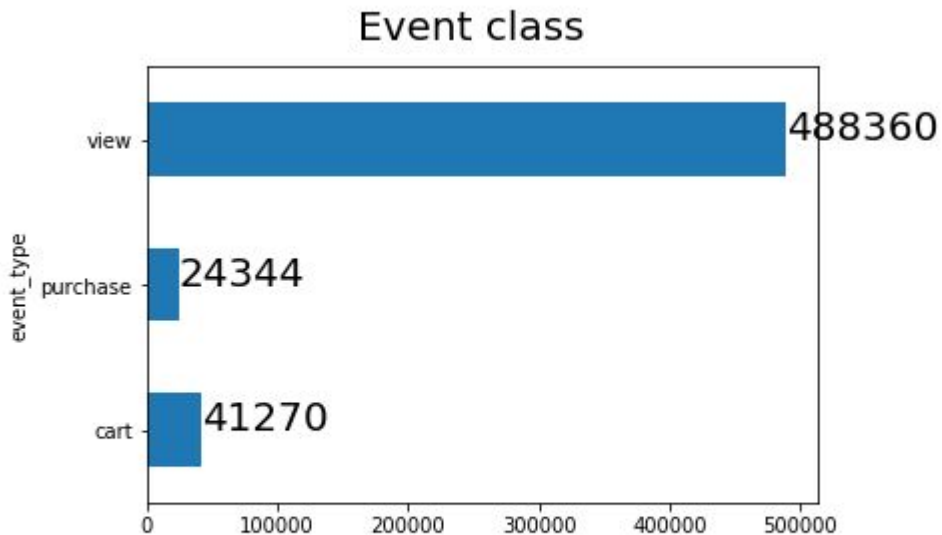




# Back to reality - EDA



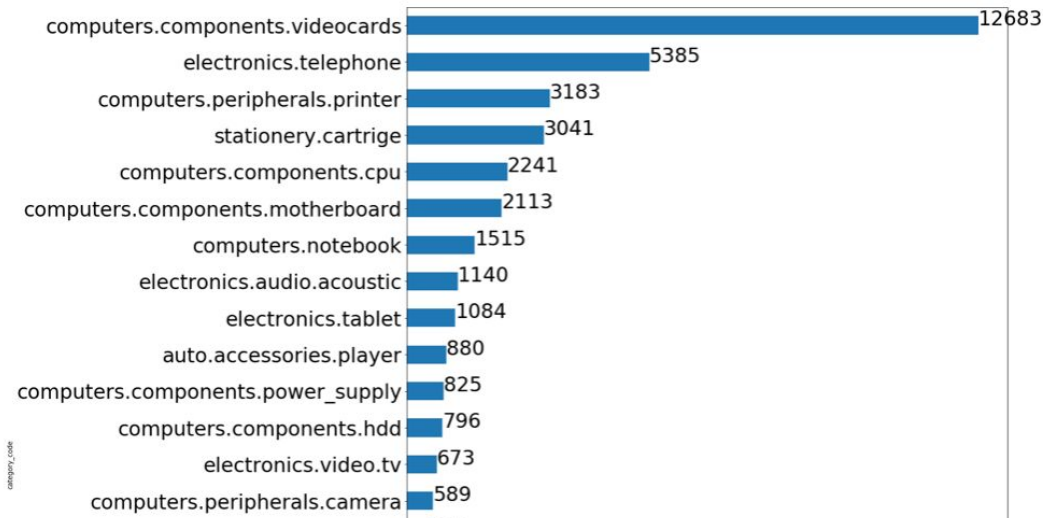
# Label distribution



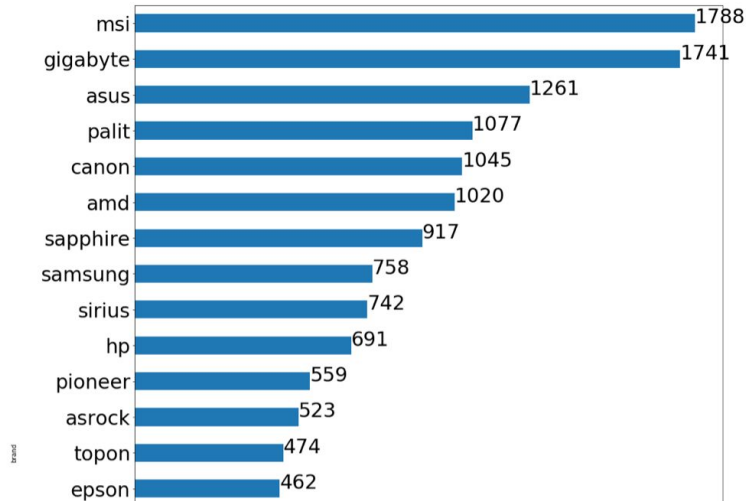
**Labeling insights:** there is imbalance, also might consider using cart over purchase

# Categories correlation with label

Added to cart Categories



Purchased brands



## Categories add to cart ratio

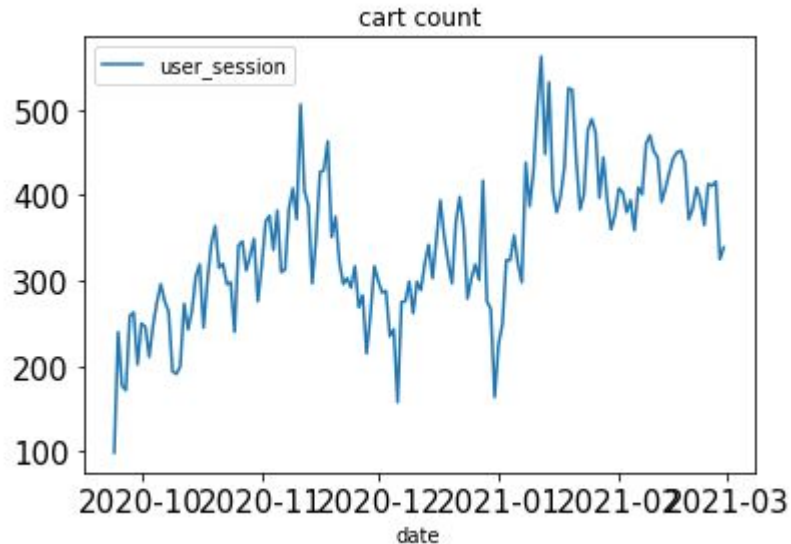
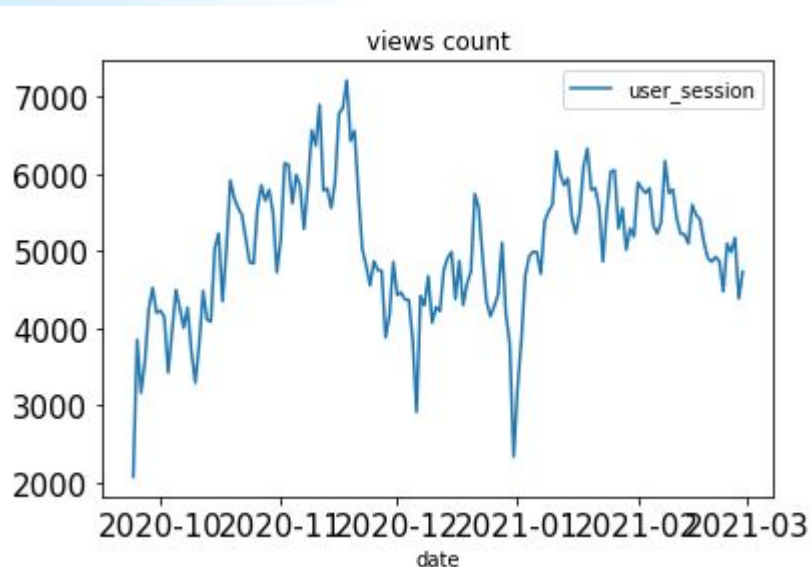
category_code	view_count	cart_count	cart_ratio
computers.peripherals.camera	4369	589.0	0.134813
computers.components.videocards	97141	12683.0	0.130563
computers.peripherals.scanner	1600	200.0	0.125000
computers.components.hdd	7549	796.0	0.105444
computers.components.cpu	21314	2241.0	0.105142
computers.components.power_supply	8050	825.0	0.102484
computers.ebooks	2827	268.0	0.094800
stationery.cartrige	32939	3041.0	0.092322
computers.components.motherboard	23221	2113.0	0.090995
electronics.video.projector	1372	118.0	0.086006
computers.peripherals.printer	37479	3183.0	0.084928
construction.tools.painting	451	35.0	0.077605
electronics.video.tv_remote	930	68.0	0.073118
computers.peripherals.wifi	6235	455.0	0.072975
electronics.audio.music_tools.piano	370	27.0	0.072973
electronics.telephone	74839	5385.0	0.071954

## Products add to cart ratio

product_id	view_count	cart_count	cart_ratio
623426	106	51.0	0.481132
1586461	78	30.0	0.384615
1856480	181	66.0	0.364641
4013214	328	119.0	0.362805
4171147	167	52.0	0.311377
8093	178	54.0	0.303371
3581576	221	64.0	0.289593
1038724	194	55.0	0.283505
672145	99	28.0	0.282828
4013582	428	111.0	0.259346
866570	110	28.0	0.254545
4171037	143	36.0	0.251748
886023	169	42.0	0.248521
3829374	280	69.0	0.246429
665345	268	66.0	0.246269
3606492	394	93.0	0.236041
3699150	137	32.0	0.233577
841972	160	37.0	0.231250
893196	2866	662.0	0.230984
885572	299	69.0	0.230769
821773	126	29.0	0.230159
821628	158	36.0	0.227848

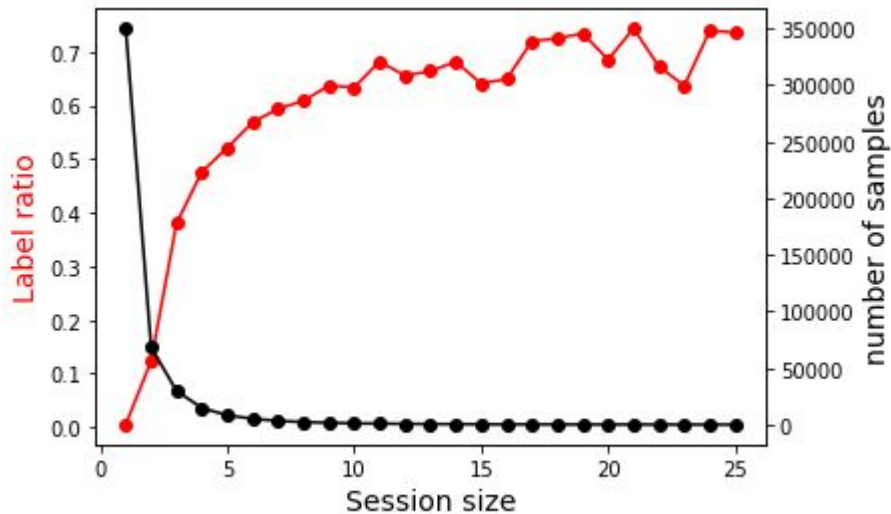
**Categories insights:** categories do matter for classification, need to incorporate them correctly in the features. like using the prior.

# Macro understanding



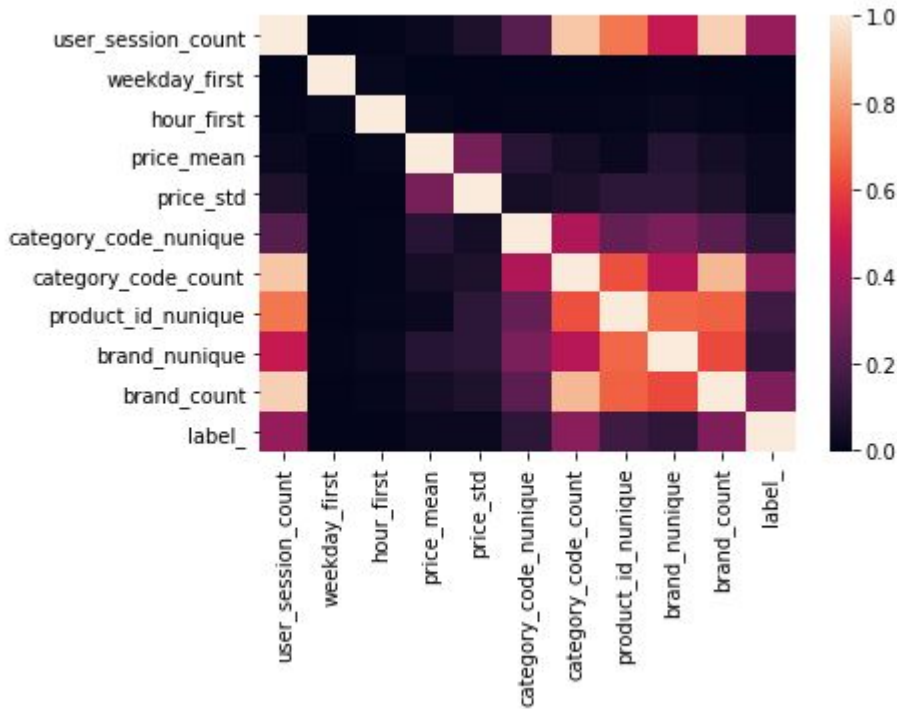
**Macro insights** : there is seasonality in the data which requires from us to evaluate properly by time.

# Session size analysis



**Session size insights:** we can use 3 events that hold 72% target labels.

# Features understanding



**Features insights:** there is a small correlation between the session count, barnd, and category code with the label. other features are not correlate and can be used together



# What did we learn?

The problem is imbalance

Some categories and product has strong prior

There is seasonality in the data

The session size is critical for framing the solution

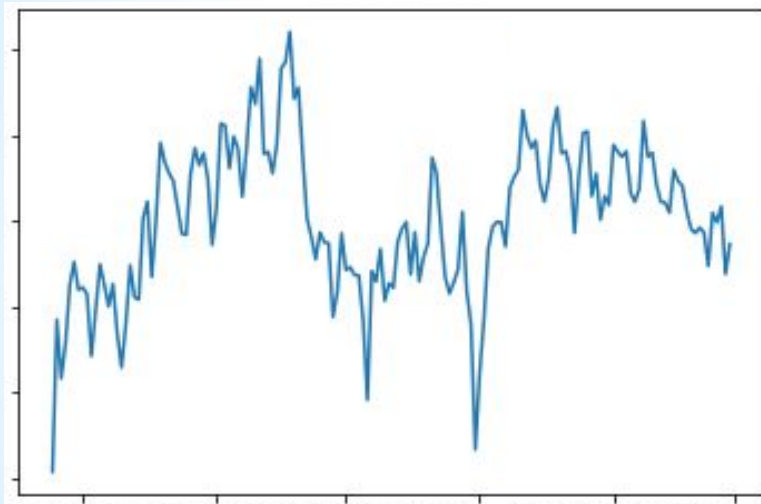
## **Also Try Pandas GUI -**

<https://towardsdatascience.com/pandasgui-analyzing-pandas-dataframes-with-a-graphical-user-interface-36f5c1357b1d>

# Back to reality - Modeling

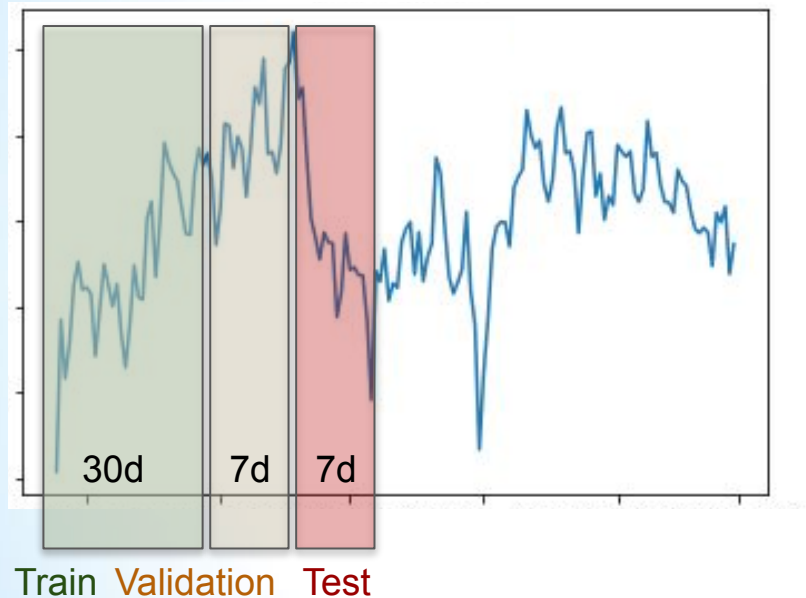


# Temporal cross validation

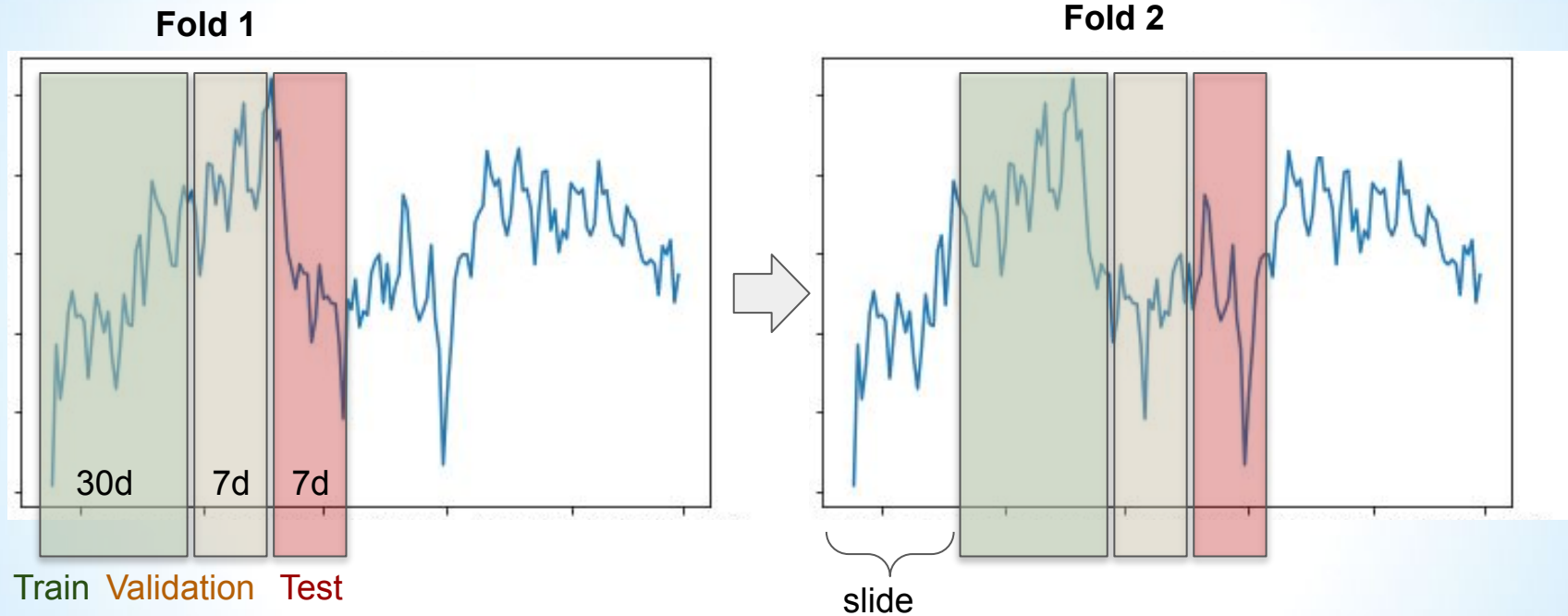


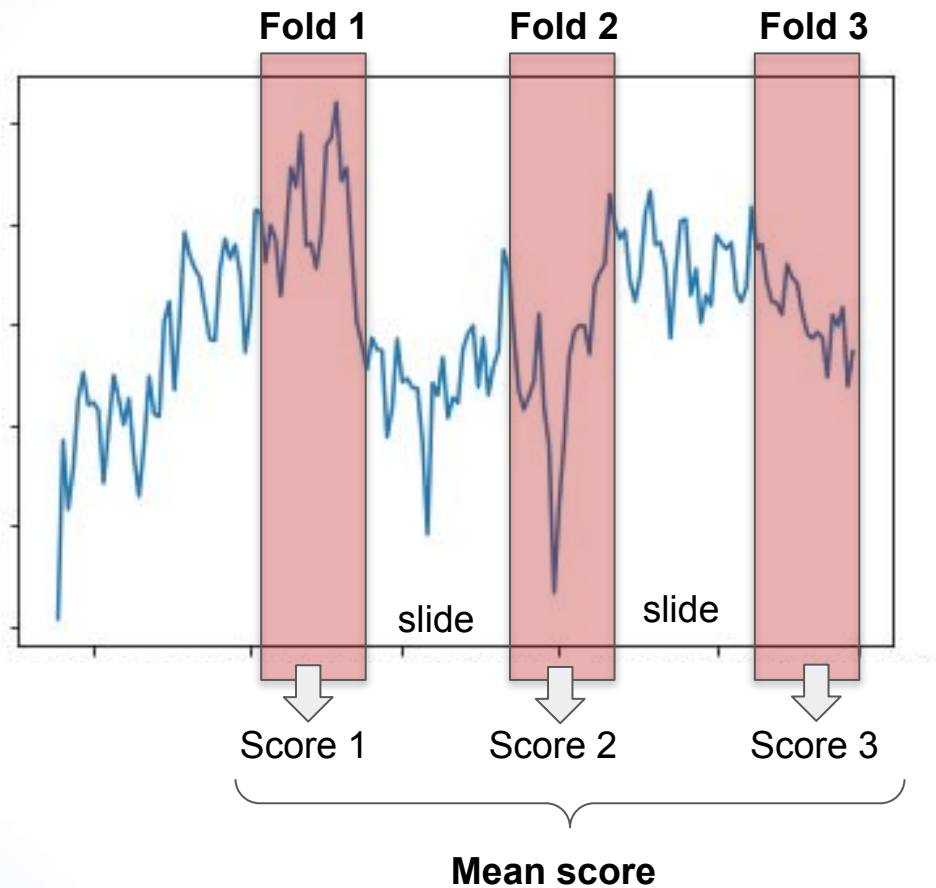
# Temporal cross validation

Fold 1

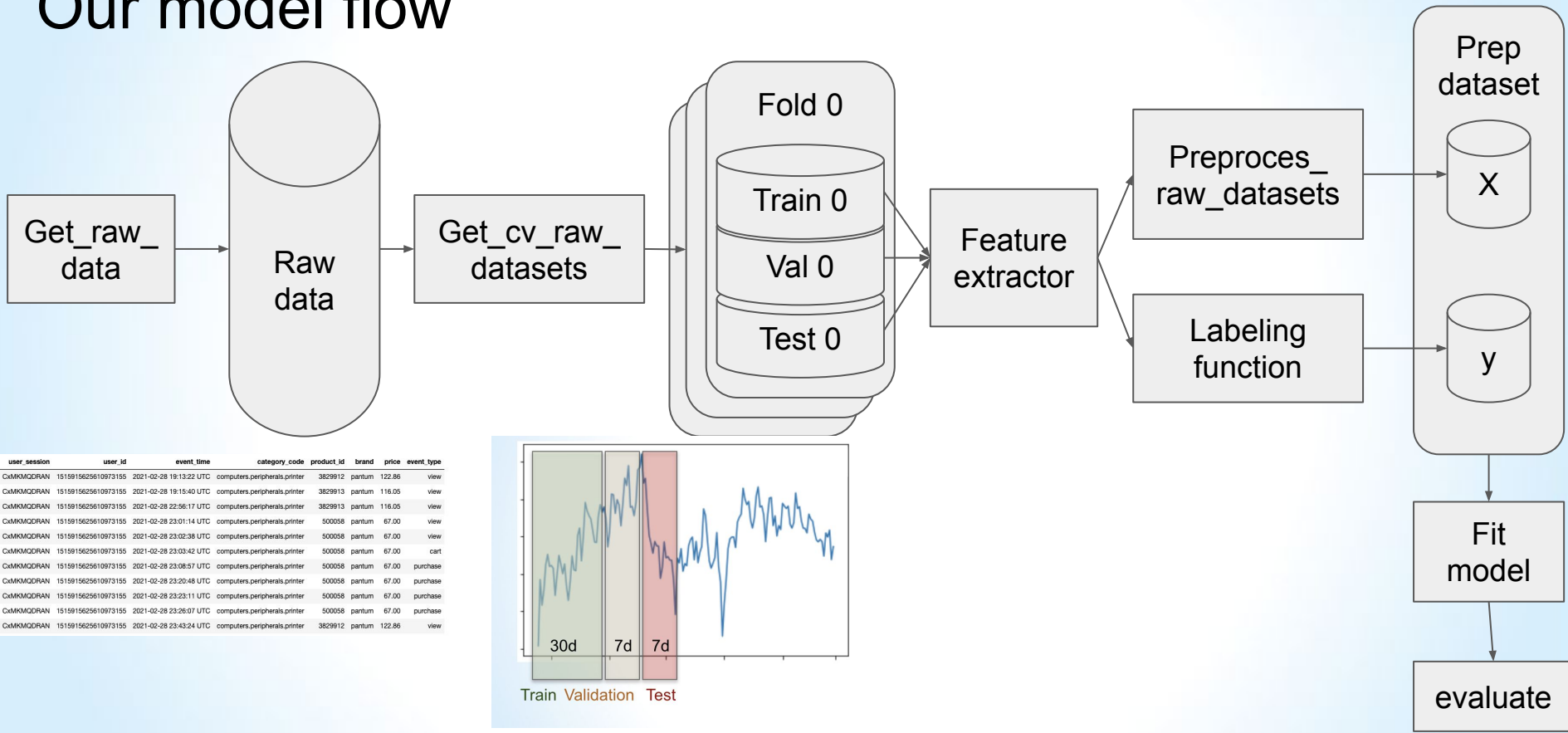


# Temporal cross validation





# Our model flow





# build a baseline

What do you think would be the simplest baseline?

Who can build it?

How to evaluate it?

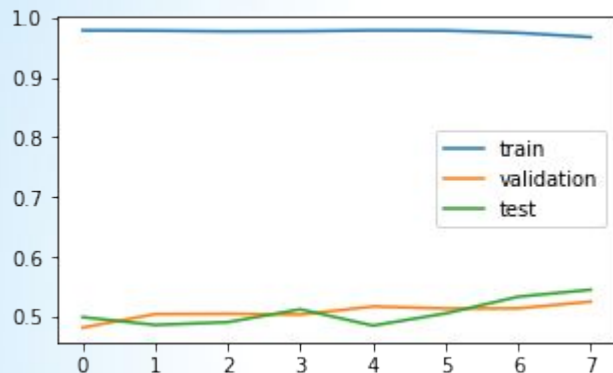


# Fitting our model

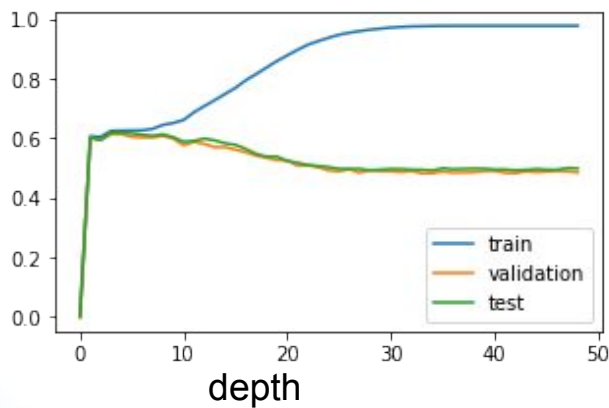
- Good for decision modeling
- Easy to understand
- Non-linear
- Fast train and predict
- Apply feature selection
- Robust to skewed features



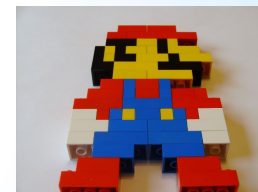
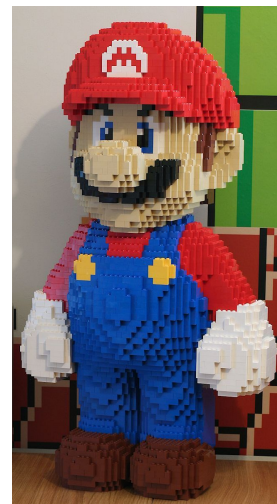
# Bias variance tradeoff



folds



depth



# Error Analysis

Qualitative analysis

Vs

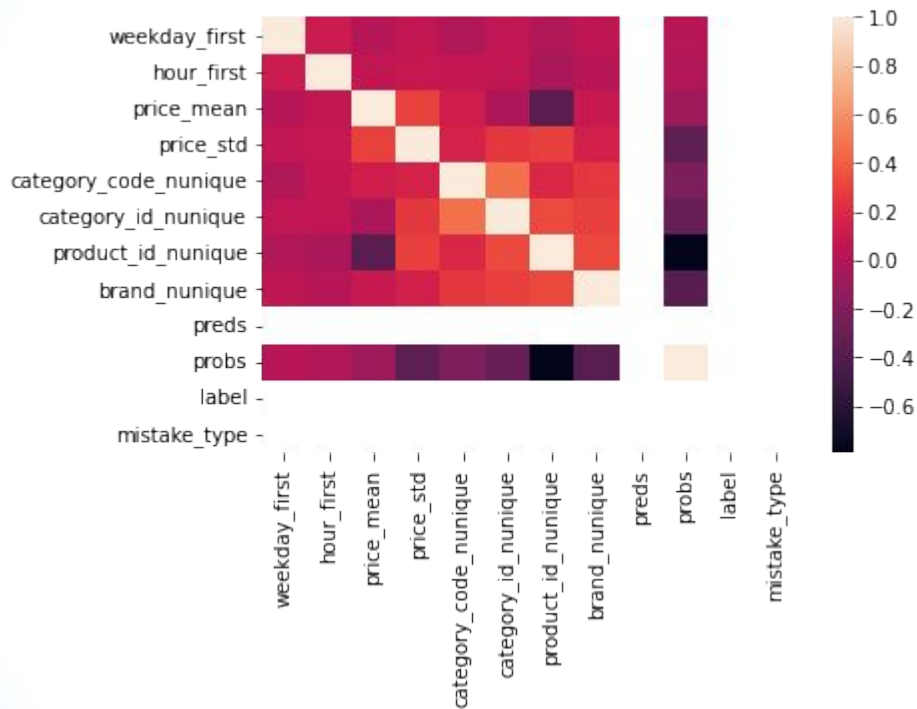
Quantitative analysis



		weekday_first	hour_first	price_mean	price_std	category_code_nunique	category_id_nunique	product_id_nunique	brand_nunique	preds	probs	label
user_id	user_session											
11526401	mmnAEtLCQQ	2	12	9.32	0.0	0	1	1	1	1	0.734783	0
13018078	AkRAnIT5FI	6	7	5.03	0.0	0	1	1	0	1	0.734783	0
11667939	urcCEPULgl	2	19	10.48	0.0	0	1	1	0	1	0.734783	0
13284340	Rad3ycN4qv	6	15	10.56	0.0	0	1	1	0	1	0.734783	0
11498410	cby5RmEUS0	2	11	10.24	0.0	0	1	1	1	1	0.734783	0
12209581	DQkZZaOwVa	4	9	6.79	0.0	0	1	1	0	1	0.734783	0
12214708	jZM6AUFEUj	4	9	1.57	0.0	0	1	3	1	1	0.734783	0
18895688	dAPqKdd6JE	2	18	10.83	0.0	0	1	1	0	1	0.734783	0
12233134	hd0U0MpJv2	4	10	9.48	0.0	0	1	1	0	1	0.734783	0
12607583	XerFJigByA	5	11	11.62	0.0	0	1	1	0	1	0.734783	0
11760735	j4iZRjkHK6	3	5	10.95	0.0	0	1	1	1	1	0.734783	0
11447262	3s9QOamVa9	2	9	9.44	0.0	0	1	1	0	1	0.734783	0
11903737	QR6SgCJCat	3	12	11.62	0.0	0	1	1	0	1	0.734783	0
10755682	Dz7dBcv4xo	0	14	1.57	0.0	0	1	1	0	1	0.734783	0
16328989	2zgeCaOupE	0	15	6.79	0.0	0	1	1	0	1	0.734783	0
11398044	wFG2cS7OP3	2	7	11.57	0.0	0	1	1	1	1	0.734783	0
11868216	SmWBn1ve6D	3	10	10.21	0.0	0	1	1	0	1	0.734783	0
11667939	J5VhIbXfYw	2	19	10.48	0.0	0	1	1	0	1	0.734783	0

user_id	user_session	weekday_first	hour_first	price_mean	price_std	category_code_nunique	category_id_nunique	product_id_nunique	brand_nunique	preds	probs	label
531710707	dSfKNrgzVH	3	0	647.290000	0.000000	1	1	1	1	0	0.000000	1
529899183	oc7L76LywG	0	2	462.760000	0.000000	0	1	1	1	0	0.000000	1
531604323	Vle7A9hmT8	2	16	35.896667	3.958615	0	1	3	3	0	0.046602	1
530412518	OgF81GfmNO	0	9	11.220000	1.052758	0	1	3	0	0	0.046602	1
530678688	BtHsbYkdqU	4	11	76.940000	29.438582	0	1	3	2	0	0.046602	1
533336721	YkVC64jSwb	6	17	33.686667	22.216981	0	1	3	1	0	0.046602	1
533336721	YkVC64jSwb	5	18	92.696667	16.629896	0	1	3	1	0	0.046602	1
532737428	SS3RP3Wjvt	0	8	33.526667	0.386825	0	1	3	1	0	0.046602	1
529620948	EeS1kWTCEg	0	20	50.366667	2.932939	0	1	3	2	0	0.046602	1
530882770	GutBHFrNZ7	4	13	149.270000	67.008773	0	1	3	0	0	0.046602	1
532303130	G3sHaeo9Dz	6	19	114.773333	107.483912	0	1	3	2	0	0.046602	1
533373717	R6b5xbZSwA	3	10	697.336667	803.610913	0	1	3	1	0	0.046602	1
528266021	VkxxpxfqzK	1	17	512.663333	55.676288	1	1	3	2	0	0.102213	1
530874328	HohNoH4Xsz	3	10	176.140000	92.905811	1	1	3	2	0	0.102213	1
530874328	HohNoH4Xsz	4	3	153.213333	75.766882	1	1	3	2	0	0.102213	1
531858190	1V4Pd2cukW	3	23	91.463333	65.473939	1	1	3	0	0	0.102213	1

# False negative Qualitative Analysis

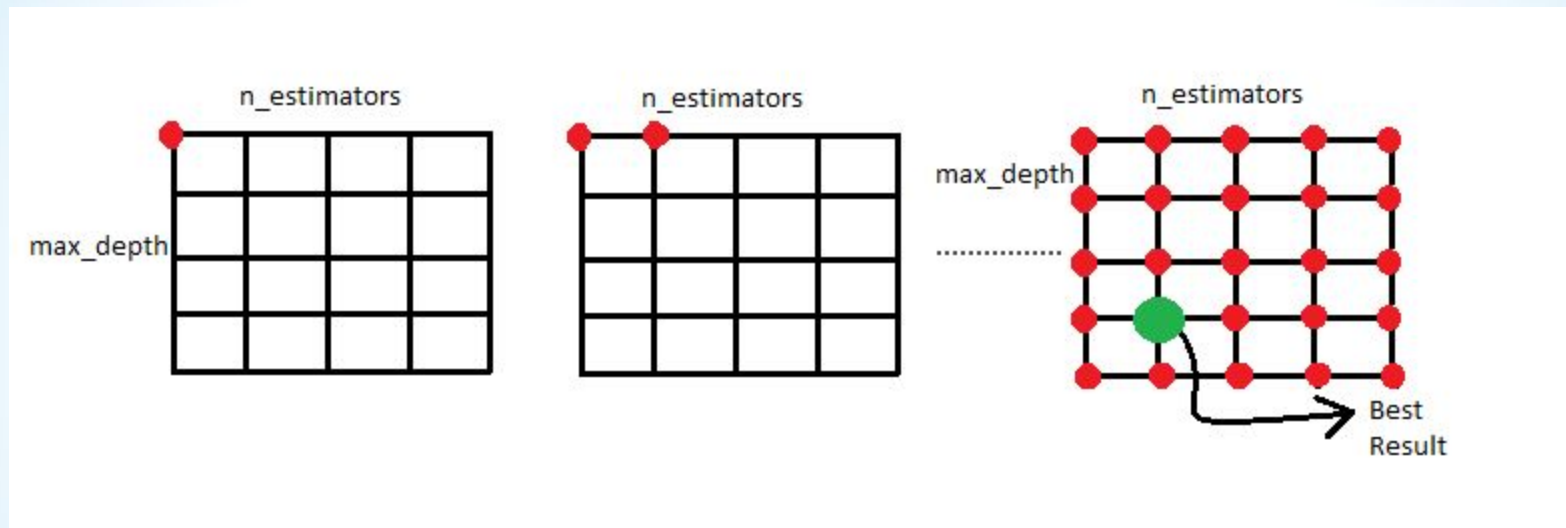




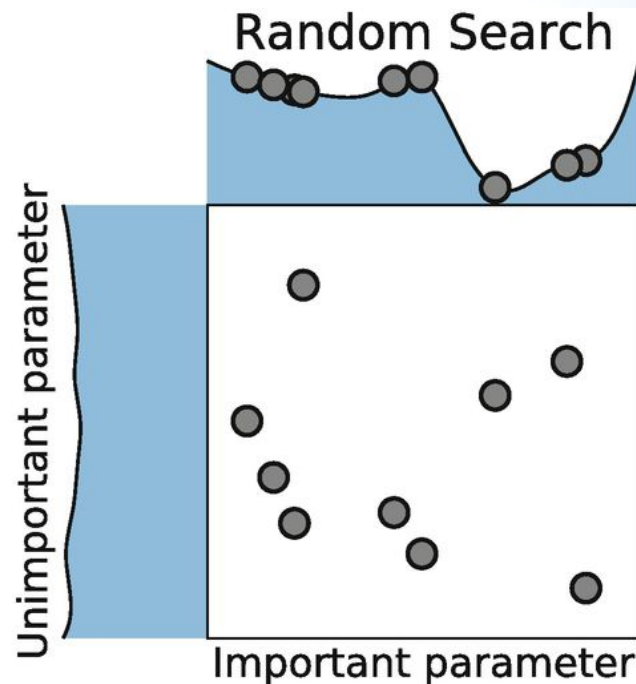
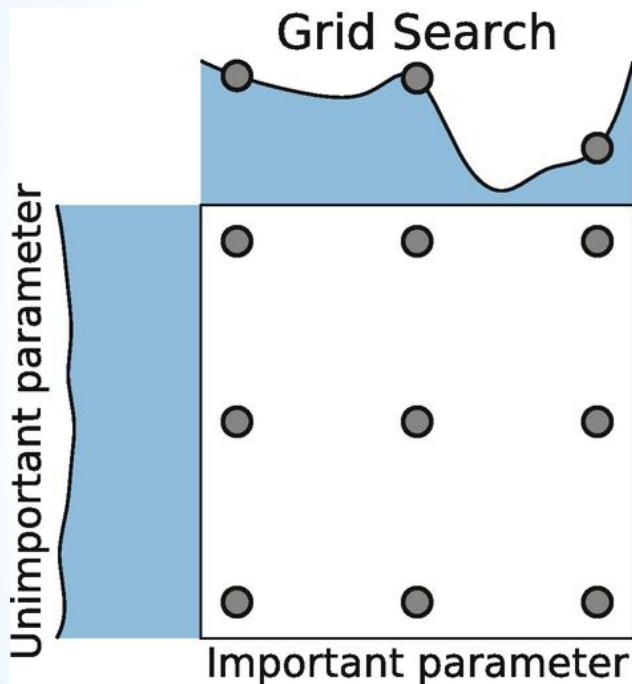
# Optimizing



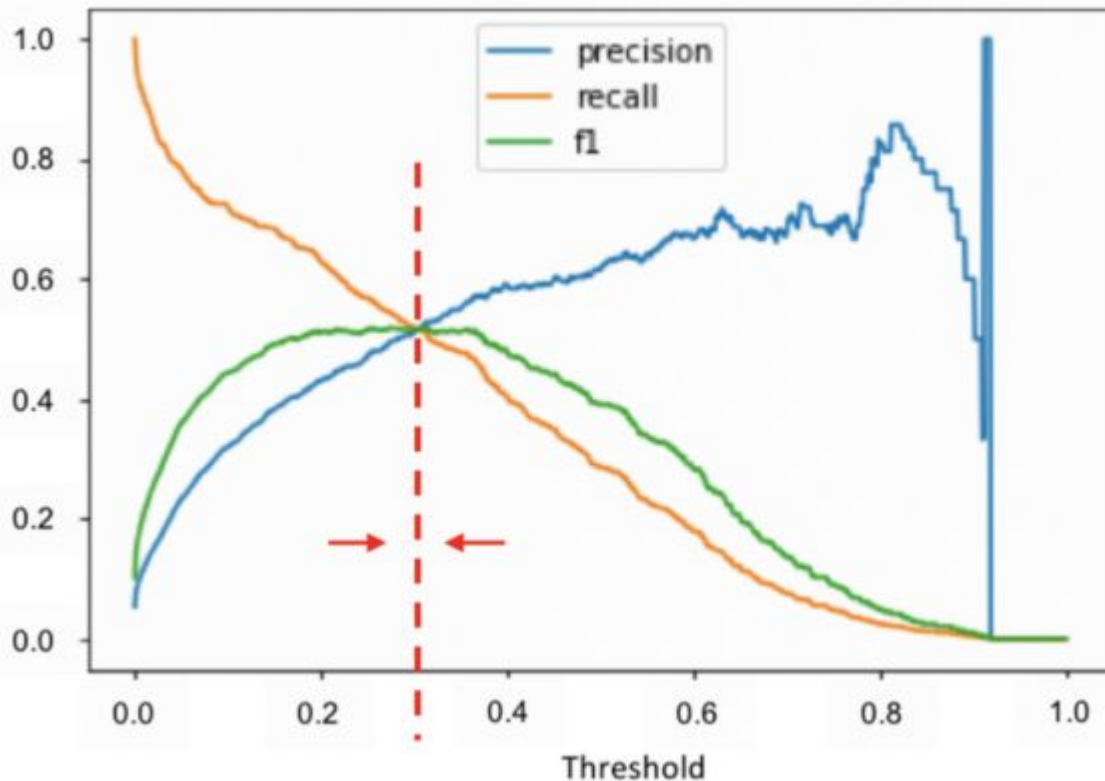
# Grid Search



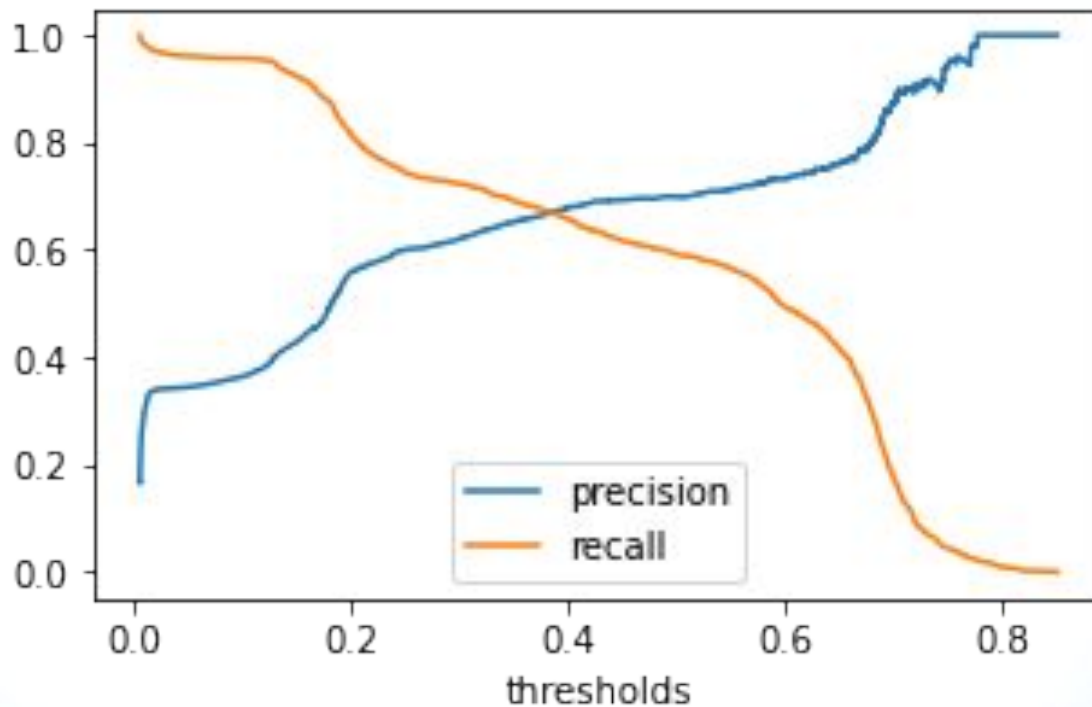
# Hyper parameter tuning



# Threshold optimization



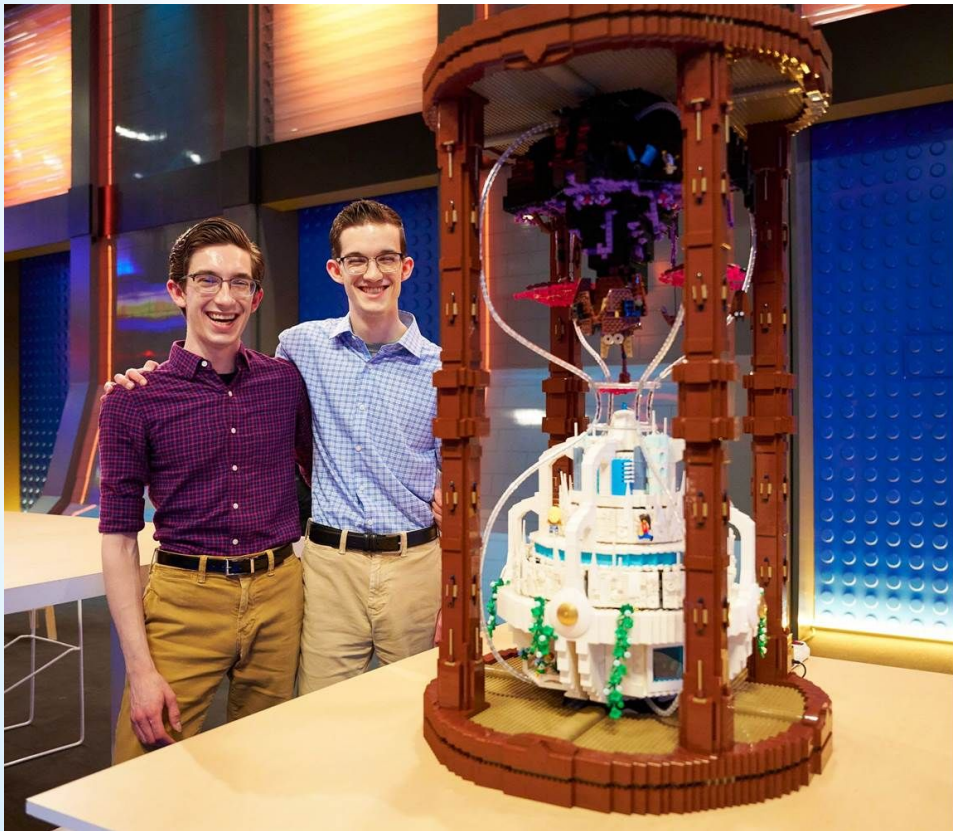
# Threshold optimization



# Batch Evaluation

batch	model_name	f1_score	f1_score_optimized
2020-10-31 00:00:00	BernoulliNB	0.645727	0.645445
2020-10-31 00:00:00	DecisionTreeClassifier	0.619484	0.688262
2020-10-31 00:00:00	LogisticRegression	0.695299	0.000000
2020-10-31 00:00:00	RandomForestClassifier	0.642916	0.704264
2020-11-14 00:00:00	BernoulliNB	0.635775	0.639161
2020-11-14 00:00:00	DecisionTreeClassifier	0.599850	0.679037
2020-11-14 00:00:00	LogisticRegression	0.689349	0.697489
2020-11-14 00:00:00	RandomForestClassifier	0.629699	0.691538

# Show your model to the world!



## Streamlit

<https://streamlit.io/gallery>



# Back to reality - Error Analysis

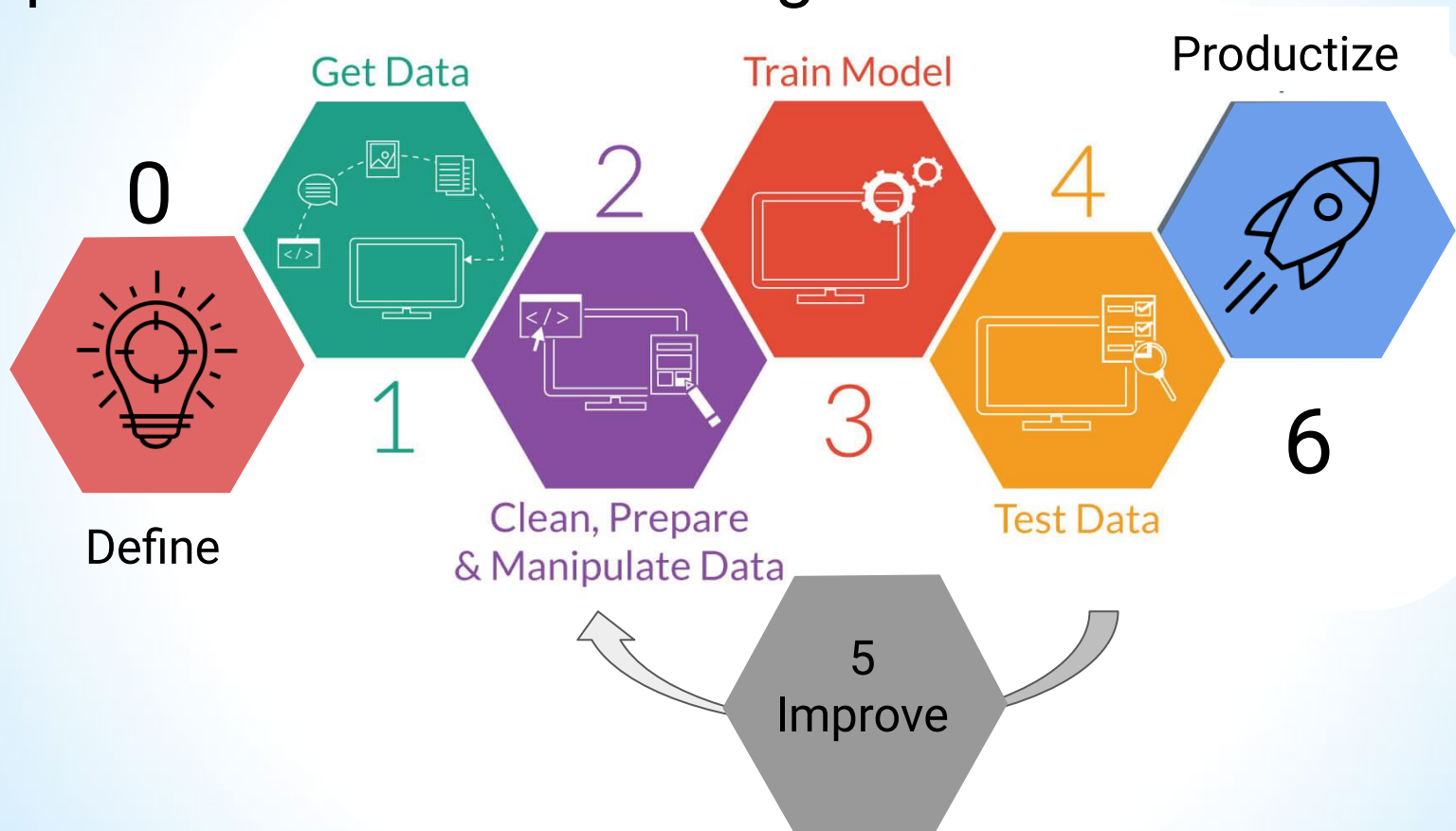




# Back to reality - Improve your signal!



# Steps to Predictive Modeling



# Code Tricks Summary

Save prepared data and load if not `os.path.isfile`.

Work on small portion of data

Use constants

Don't waste time on nice graphs

Use kwargs - `class(**kwargs_dict)`

Use classes to separate code ares

Use static functions when possible

Use list of dictionaries to build dataframe

# How to Dev?

## Jupyter

- Analysis and demoing
- In memory context - fast but cause mistakes with variables

## Pycharm

- Organized code
- Debuggable
- Auto-complete
- Argument with bash

## Combine

- Use Classes developed in pycharm at jupyter

```
In [1]: %load_ext autoreload
        %autoreload 2

In [2]: from my_code.my_cool_utils import my_sum
        my_sum(x=5,y=5) # Should return the sum of x+y

Out[2]: 10

In [3]: # Change my_sum to print some wrapping text and then return the sum
        my_sum(x=5,y=5)

        The sum of 5 and 5 is:

Out[3]: 10
```



# Future steps

More Error analysis - going back the the raw data and aggregations

Extract session features: duration, average time between views

Reduce Features with Chi2 selection or PCA

Undersample / oversample the data

# Q&A

Pandas GUI -

<https://towardsdatascience.com/pandasgui-analyzing-pandas-dataframes-with-a-graphical-user-interface-36f5c1357b1d>