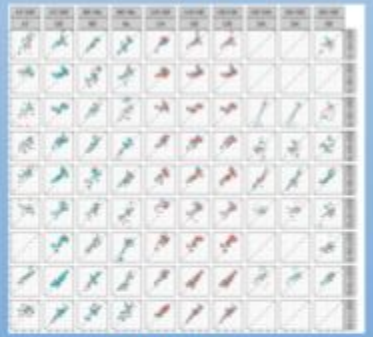
STOP VISUALIZING DATA

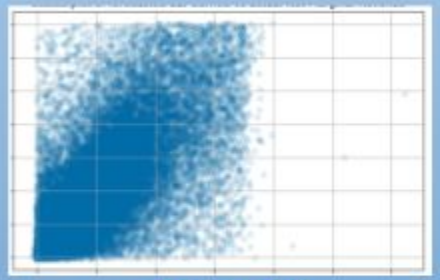
- NUMBERS ARE NOT IMAGES
- YEARS OF PLOTTING yet NO INSIGHT
- COLOR-BLIND palettes?? How about REALITY-BLIND!!

Look at what data scientists have been demanding your respect for all this time, with all the sns.displot and geom_errorbar we've built for them

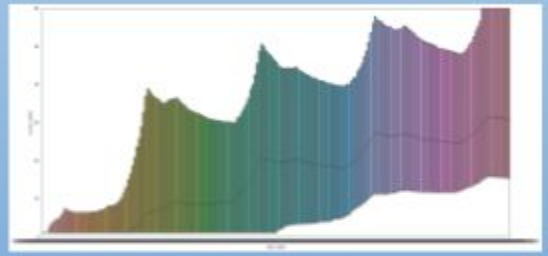
(these are REAL plots, made by REAL data scientists)



???



??????



????????

They have played us for absolute fools

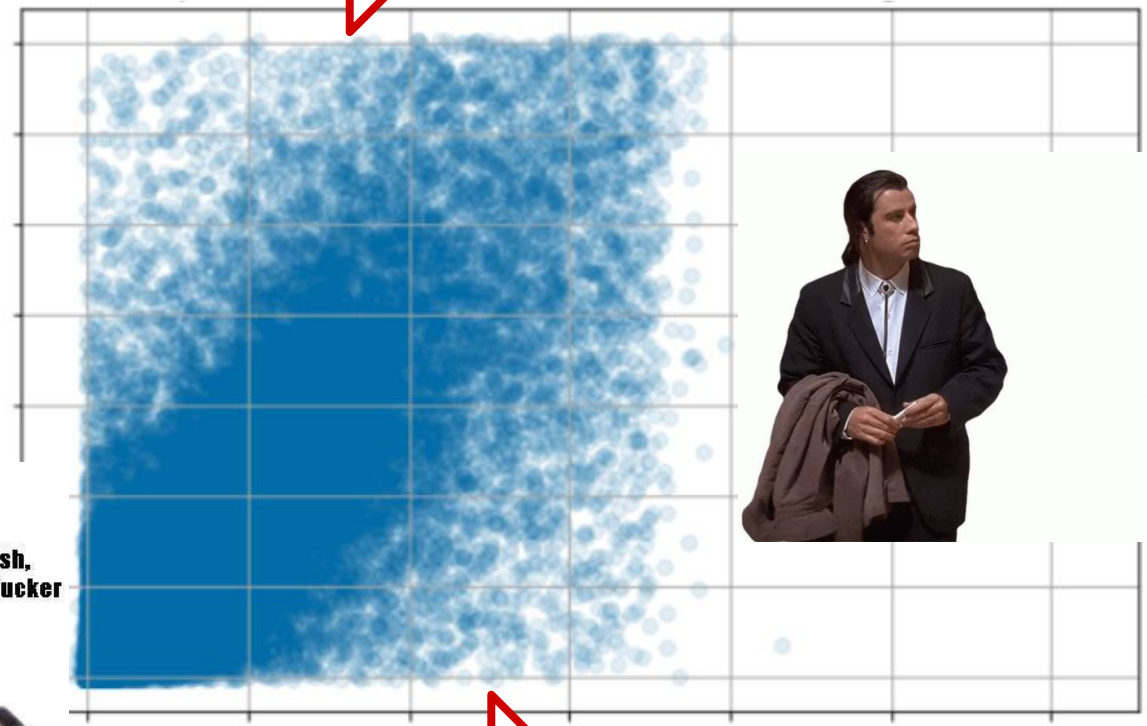
Tales of data science soft skills

Lars Roemheld, http://mdl.fit/
pyData Zurich, 25 Jan 2024

Where fashion meets tech and convenience.

Do as I say,
not as I do.

SO WHAT??



??

??

Ownership

Our job is to solve ambiguous problems end-to-end.

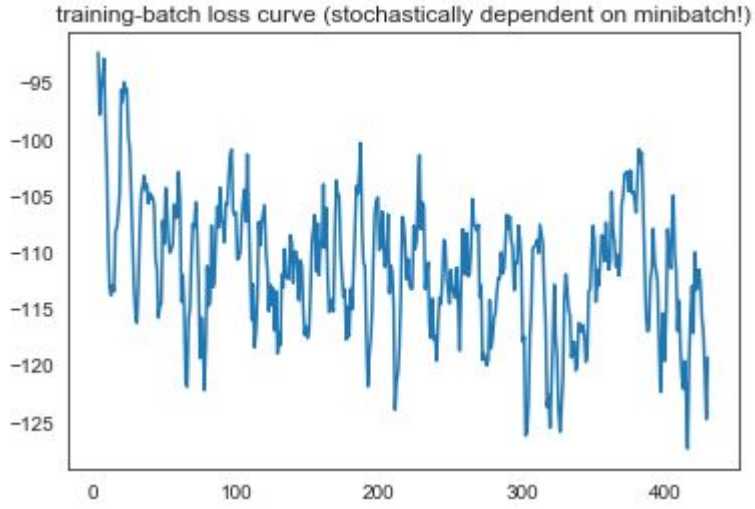
International Journal of Forecasting
Volume 36, Issue 3, July–September 2020, Pages 1181-1191

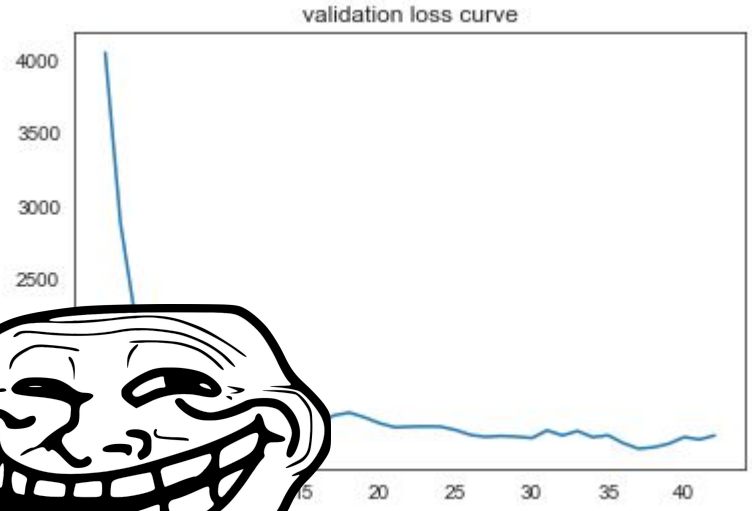
DeepAR: Probabilistic forecasting with autoregressive recurrent networks

David Salinas, Valentin Flunkert, Jan Gasthaus, Tim Januschowski

```
class DeepAR(nn.Module):
    def __init__(self, n_lstm_layers, n_lstm_hidden, lstm_direct
        '''
        Instantiates a DeepAR model.
        n_lstm_layers: number of LSTM cells stacked in sequence
        n_lstm_hidden: number of hidde units per LSTM cell (i.e
        lstm_direct_features: list of p as-colu
                              Note: these sh uld
                              of ~[-6, 6] due t
        lstm_categorical_features: dictionary of
        y_nonnegative: indicates if predicted val
        '''
        super(DeepAR, self).__init__()
        self.n_lstm_layers = n_lstm_layers
        self.n_lstm_hidden = n_lstm_hidden
        self.lstm_direct_features = lstm_direct_features
        self.lstm_categorical_features = lstm_categorical_featur
        self.y_nonnegative = y_nonnegative

        # build network
        self.cat_embeddings_total_dim = 0
        self.cat_embedding_layers = {} # TODO use nn.moduledict
        for c, embed_config in lstm_categorical_features.items()
            self.cat_embedding_layers[c] = nn.Embedding(embed_co
```





Cool loss bro. Xgboost yet?

Measure twice, cut once

In exploratory work, explicit planning prevents getting lost.

Demand forecast: hypothesis tree

Problem: forecast demand in e-commerce for pricing

1. Problem: seasonality, trend
    1.1. fbprophet/nixtla/…
    1.2. ARIMA from a package
    1.3. Exponential seasonal smoothing w/ xgboost/lgbm 🤦
    1.4. …
    1.5. …
    1.6. That cool SOTA paper I saw on xitter

Demand forecast: hypothesis tree

2. Problem: different scales of timeseries
    2.1. StandardScaler()
    2.2. Log scales
    2.3. Normalization w/ moving average
    2.4. …
    2.5. …
    2.6. That latest scale-invariant transformer architecture

Demand forecast: hypothesis tree

3. Problem: censored data on stockouts
    - 3.1. NULL value and no forecast
    - 3.2. …
    - 3.3. …
    - 3.4. Self-consistent hallucination-interpolation

Algorithm 1 Hypothesis Trees

Require: Clear problem statement X

Ensure: Envisioned solution solves original problem

```
while Problem X is unsolved do
    if X is purely empirical then
        Find solution to X in data
        return                                          ▷ Phew!
    end if
    C ← {breadth-first brainstormed approaches, incl. off-the-shelf}
    ∀c ∈ C : Ê(Validation-effort(c))
    C ← sorted(C)
    X ← C_0
end while
```

Complete tree of validation effort breadth-first search *before* first line of code!

Efficient curiosity

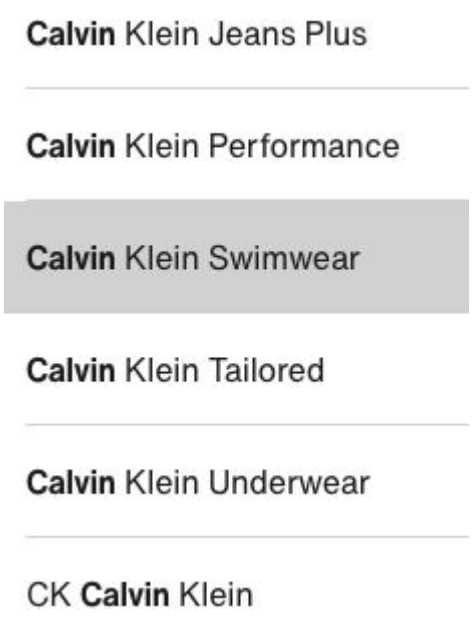
We like to learn. Ideally more efficiently than through brute-force experiments.

Build your scientific intuition

- Meta-reasoning and mentors: *why* did someone choose the approach they did?
- Fundamentals
  - Databases 101 (e.g., Harvard CS50)
  - Stats & learning theory (e.g., Taddy - Business Data Science)
- Curiosity and similar problems

Iterative results

Frequent feedback reduces ambiguity.



- Prefix/suffix
- ~~Deep NLP model~~
- Get senior mentorship for breadth-first search
- Levenshtein distance
- Find string tools library, experiment
- Max substring, min edit distance
- Get PR review, ship it!

mdl.fit