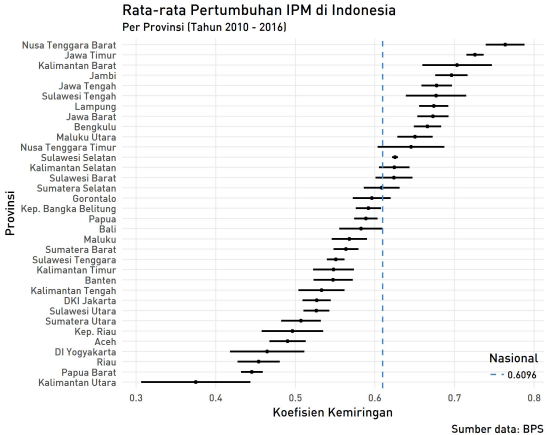
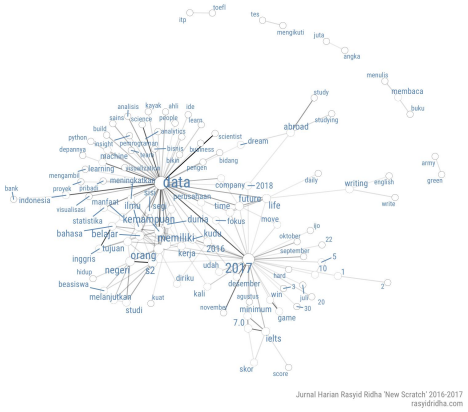
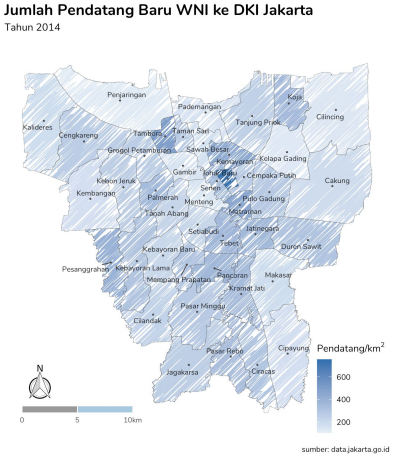
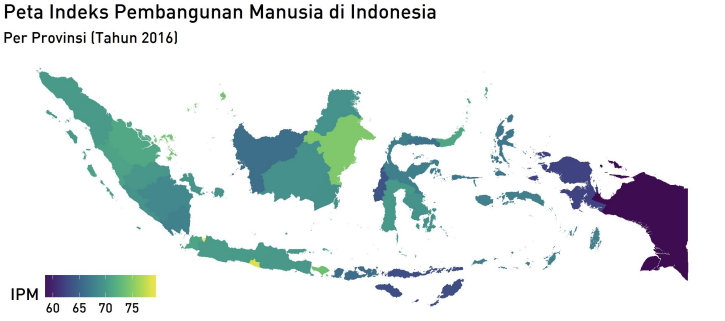
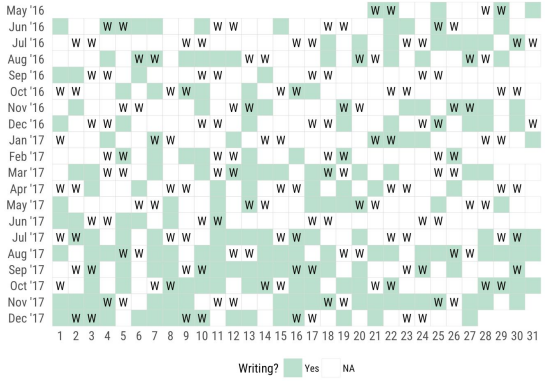










Have fun with R

Muhammad Rasyid Ridha

rasyidstat@gmail.com
github.com/rasyidstat
rasyidridha.com

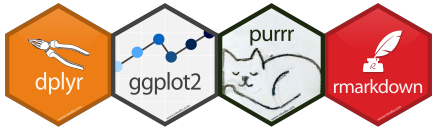
useR! Jakarta, 19-01-19

Outline

1. My R Journey
2. R for Daily Works
    a. My Daily Works using R
    b. Building R packages
3. Won Finhacks 2018 using R
    a. What is Finhacks 2018?
    b. Framework and solution details

01 My R Journey

+6 years with R



Joined **GOJEK** and built R package, rojek;
Applied master degree scholarship and solve the test using R;
Won my first data science competition, **Finhacks 2018**, using R

Graduated;
Got my first job;
Learn dplyr, ggplot2 and tidyverse

2013

Start using R (base R)

2015

First time use RStudio;
Thesis about time series using R

2016

2017

Built an R package for the first time at **Smartfren** (smartr, nusantr);
Learn git;
Won my first competition using R (EDA competition)

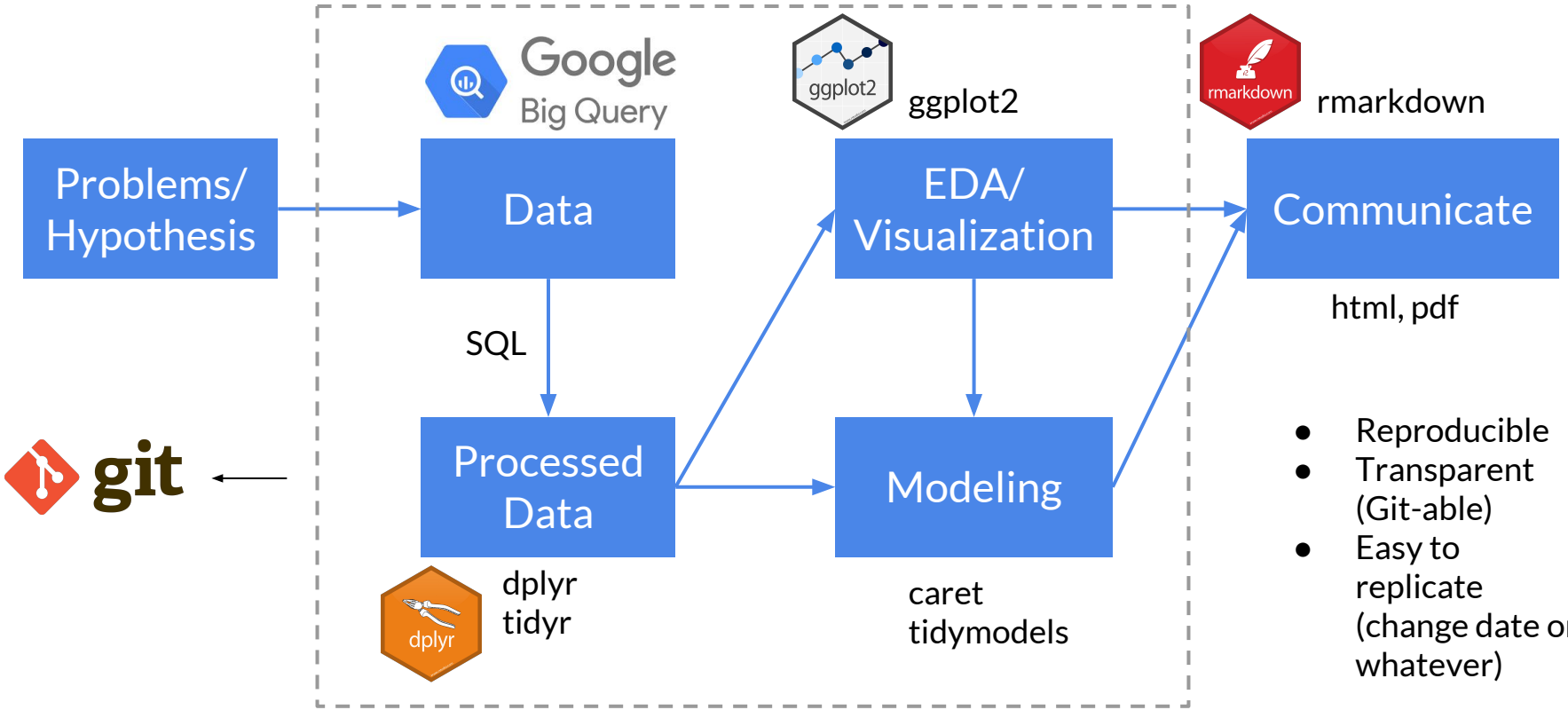
2018

2019

What next?
Publish an R package to CRAN? Attend R international workshop?

02 R for Daily Works

My Daily Works using R



Building R packages

https://medium.com/airbnb-engineering/using-r-packages-and-education-to-scale-data-science-at-airbnb-906faa58e12d

Using R packages and education to scale Data Science at Airbnb

AirbnbEng Follow
Mar 29, 2016 · 9 min read

By Ricardo Bion

Start building your own R packages from scratch http://r-pkgs.had.co.nz/

R packages that I have built

- smartr (at Smartfren)
- nusantr (Indonesia R package)
- mrsq (My personal R package)
- rojek (at GOJEK)

For what? ggplot2 theme, database connection, data reference, utility functions, etc.

03 Won Finhacks 2018 Using R

What is Finhacks 2018?

Finhacks 2018 Data Challenge is the first national data science competition in Indonesia, bringing machine learning problems in banking industry.
The opportunity was huge since there were 15 finalists who would be selected and awarded with total prize, IDR480,000,000.

Fraud Detection

- +10K data
- Classification
- Imbalance dataset
- High cardinal variables
- AUC

Credit Scoring

- +10K data
- Classification
- Imbalance dataset
- AUC

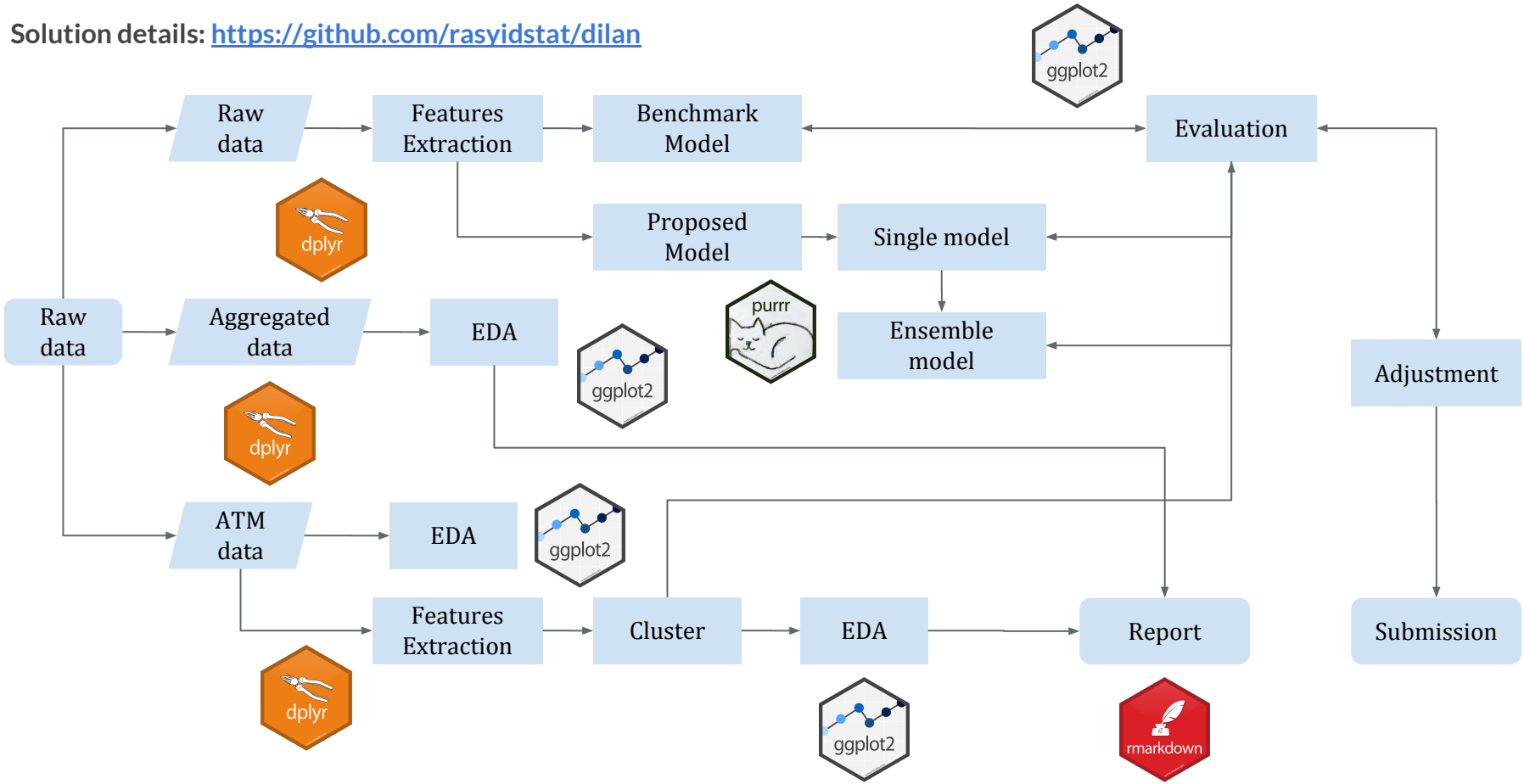
I focused 100% on this

ATM Cash Demand Forecasting

- +800K data
- +10K ATMs
- Regression (time series)
- Data train: 83 days
- Data test: 7 days
- % of percentage error below 25 %

Pipeline

Solution details: https://github.com/rasyidstat/dilan



Learning from the Champion

- Kaggle Web Traffic (2017)
  - 1st winner: RNN
  - 2nd winner: deep learning + xgboost
  - 3rd winner: linear combination of median 7, 28, 49 and 365 days
- M4 Competition (2018)
  - 1st winner: Hybrid ETS-RNN
  - 2nd winner: ARIMA + ETS + tbats + Theta + naive + seasonal naive + NN + LSTM -> xgboost (weighted average)
  - 3rd winner: weighted average, pool based on time series characteristics
  - 4th winner: combination of statistical method
  - 5th winner: ARIMA + ETS + Theta (weighted average)
  - 6th winner: ETS + CES + ARIMA + Theta (median)

Features Engineering

Lag 7 days

Dummy variables / regressor

$$Y_t = Y_{t-7} + \textit{Main Features} + \textit{Payday Effect}$$

- Week-of-the-month
- Day-of-the-week
- Public holiday
  - Mon, 1 January 2018 (New Year)
  - Fri, 16 February 2018 (Chinese New Year)
  - Sat, 17 March 2018 (Nyepi Day)
  - Fri, 30 March 2018 (Good Friday)

6 different features engineering scenario

f5: payday at 1,25,28,31

f51: payday at 1,25,28,31 + pre and after effect during saturday

f52: payday at 1,2,5,25,26,27,28,29,30,31

f53: payday at 1,2,5,25,26,27,28,29,30,31 + pre and after effect during saturday

f54: payday at 1,25,28,31 + separated pre and after effect during saturday

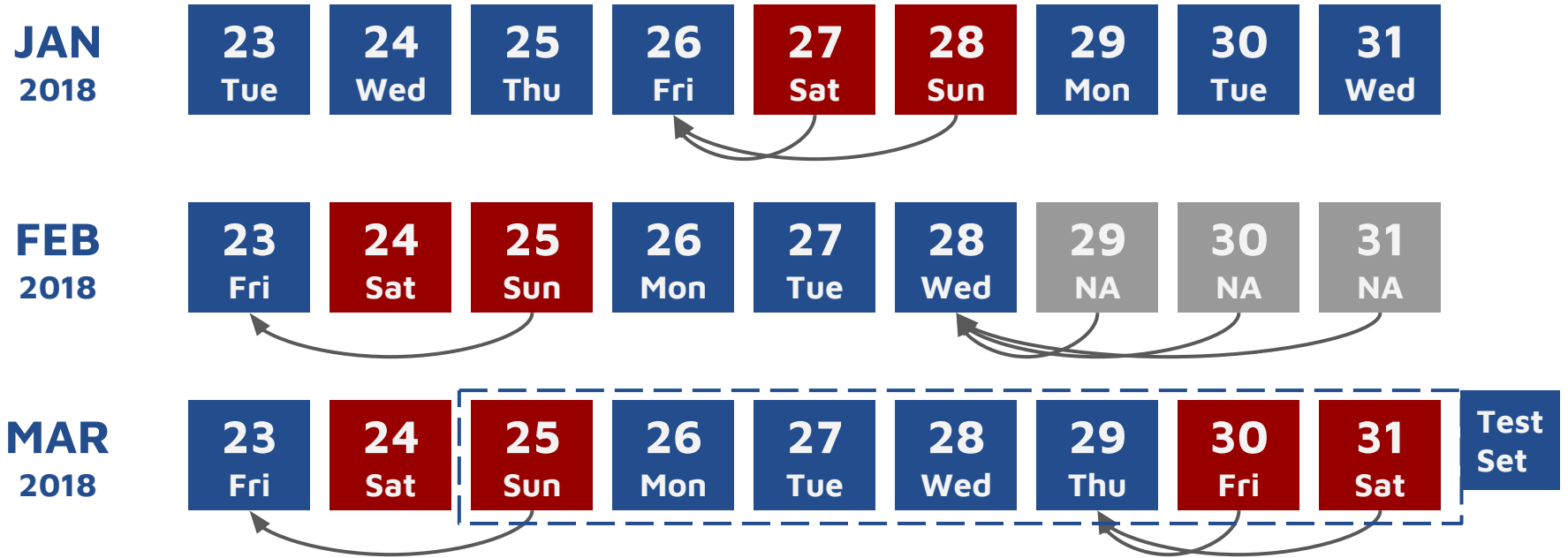
f55: payday at 1,2,5,25,26,27,28,29,30,31 + separated pre and after effect during saturday

Total: 18-26 features

Features Engineering: Model the Real World

When a payday falls in weekend/holiday, it moves to the previous working day

None of the finalists used this assumption as the general knowledge to the model

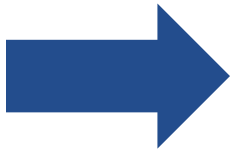


Model Framework

The final model is the model combination based on best model selection of 1 Naive + 4 XGBoost models with different features engineering (f5, f52, f53, f55) and 1.02 correction (0.95 correction for Good Friday, 30 March 2018 and no correction for Thursday, 29 March 2018) trained using full dataset.

Single Model

- **Naïve**
  `naive()`
- Exponential Smoothing
  `ets()`
- ARIMA
  `auto.arima()`
- Linear Regression
  `lm()`
- **XGBoost** (learning rate = 0.01, 6 different features set)
  `xgboost()`



Train models for every ~10K ATMs

Iteration using `purrr`, functional programming in R

Model Combination

- Ensemble averaging
  Average prediction result from chosen models
- **Best model selection**
  (Bucket of models technique)
  Select best model for each ATM based on majority of best evaluation metrics in 4 cross-validation sets. If the result is tie, default model will be selected

In a Nutshell

- Learn basic R
- Learn tidyverse (dplyr, ggplot2, rmarkdown, etc.)
- Learn git
- Learn machine learning
- Use R for real case problems in your work
- Build an R package in your work
- Join a competition and win

Thank you! Let's play with data!

Muhammad Rasyid Ridha

rasyidstat@gmail.com
github.com/rasyidstat
rasyidridha.com