Ensemble Learning

Ali Akbar Septiandri

Universitas Al-Azhar Indonesia

aliakbars@live.com

November 22, 2019

Selayang Pandang

1. Bagging
2. Boosting
3. Stacking

Bahan Bacaan

1. VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media. https://jakevdp.github.io/PythonDataScienceHandbook/05.08-random-forests.html
2. Besbes, A. (2016, August 10). How to score 0.8134 in Titanic Kaggle Challenge [Blog post]. Retrieved from http://ahmedbesbes.com/how-to-score-08134-in-titanic-kaggle-challenge.html
3. Grover, P. (2017, December 9). Gradient Boosting from scratch [Blog post]. Retrieved from https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d
4. Parr, T. & Howard, J. (2018). How to explain gradient boosting. explained.ai. Retrieved from https://explained.ai/gradient-boosting/index.html

Occam's Razor

Definisi

Given two models with the same generalization errors, the simpler model is preferred over the more complex model.

Bagging

Bagging

1. Definisikan jumlah iterasi T
2. Untuk setiap t dalam T:
    1. Lakukan pengambilan sampel dengan pengembalian (sampling with replacement)
    2. Latih model t dengan menggunakan sampel tersebut
3. Agregasi hasil dari prediksi dari setiap iterasi

Random Forest

- Membuat K pohon keputusan yang berbeda:
  - memilih subset acak S_r
  - membuat pohon keputusan penuh T_r (tanpa *pruning*)
  - repetisi untuk $r = 1...K$
- Jika diberikan data baru X:
  - klasifikasi dengan setiap pohon $T_1...T_K$
  - Gunakan *majority vote*
- Salah satu metode yang paling efektif (*state-of-the-art*)

Bagging vs. Voting

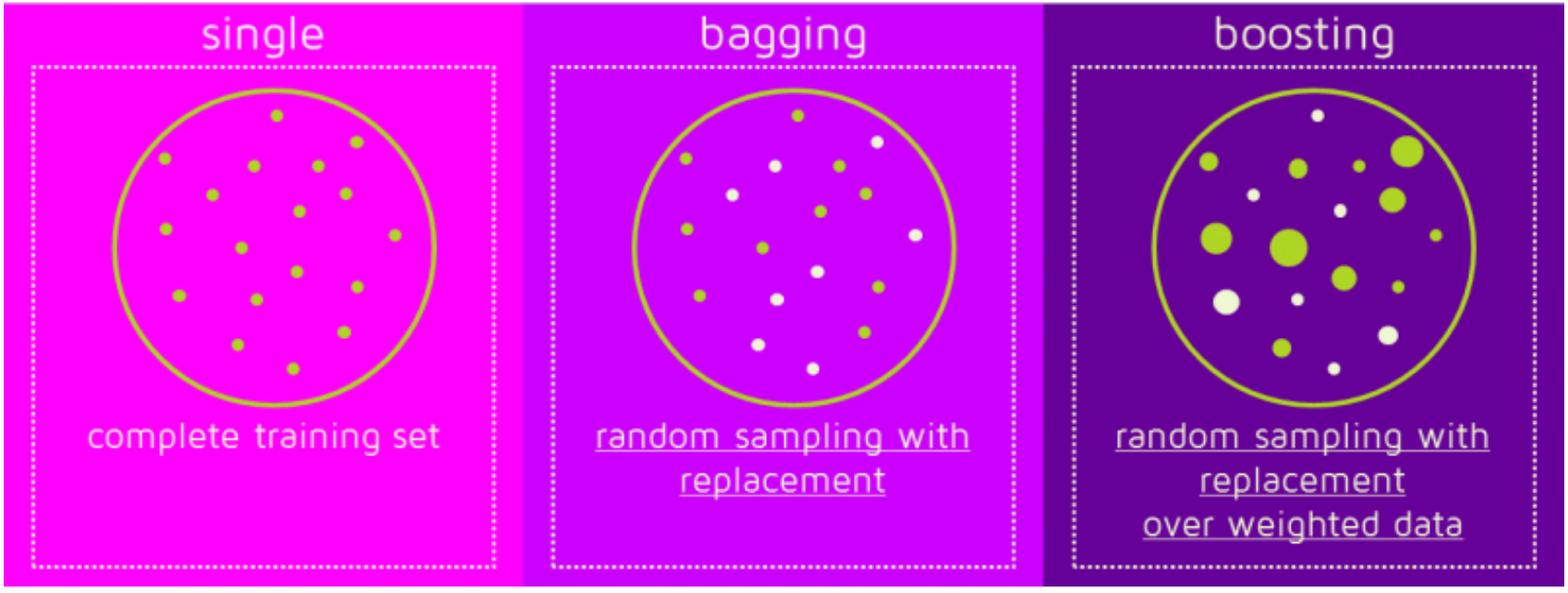
- Bagging
  - Jenis algoritmanya hanya 1
  - Menggunakan *bootstrap sampling*
- Voting
  - Bisa ada N jenis algoritma
  - Mengagregasi hasil dari berbagai model

Boosting

Adaptive Boosting (AdaBoost)

1. Definisikan jumlah iterasi T
2. Inisiasi nilai bobot tiap sampel sama besar
3. Untuk setiap t dalam T:
    1. Lakukan pengambilan sampel dengan pengembalian (sampling with replacement)
    2. Latih model t dengan menggunakan sampel tersebut
    3. Hitung loss per sampel dan rata-rata loss-nya
    4. Hitung ulang bobot sampel dan bobot model t menggunakan loss

Bagging vs Boosting



Gambar: Proses pembuatan subsets [QuantDare, 2016]

Gradient Boosting

1. Definisikan jumlah iterasi T
2. Untuk setiap t dalam T:
    1. Latih model t
    2. Hitung residual dari prediksi model
    3. Latih model dengan residual tersebut sebagai targetnya untuk model $t + 1$

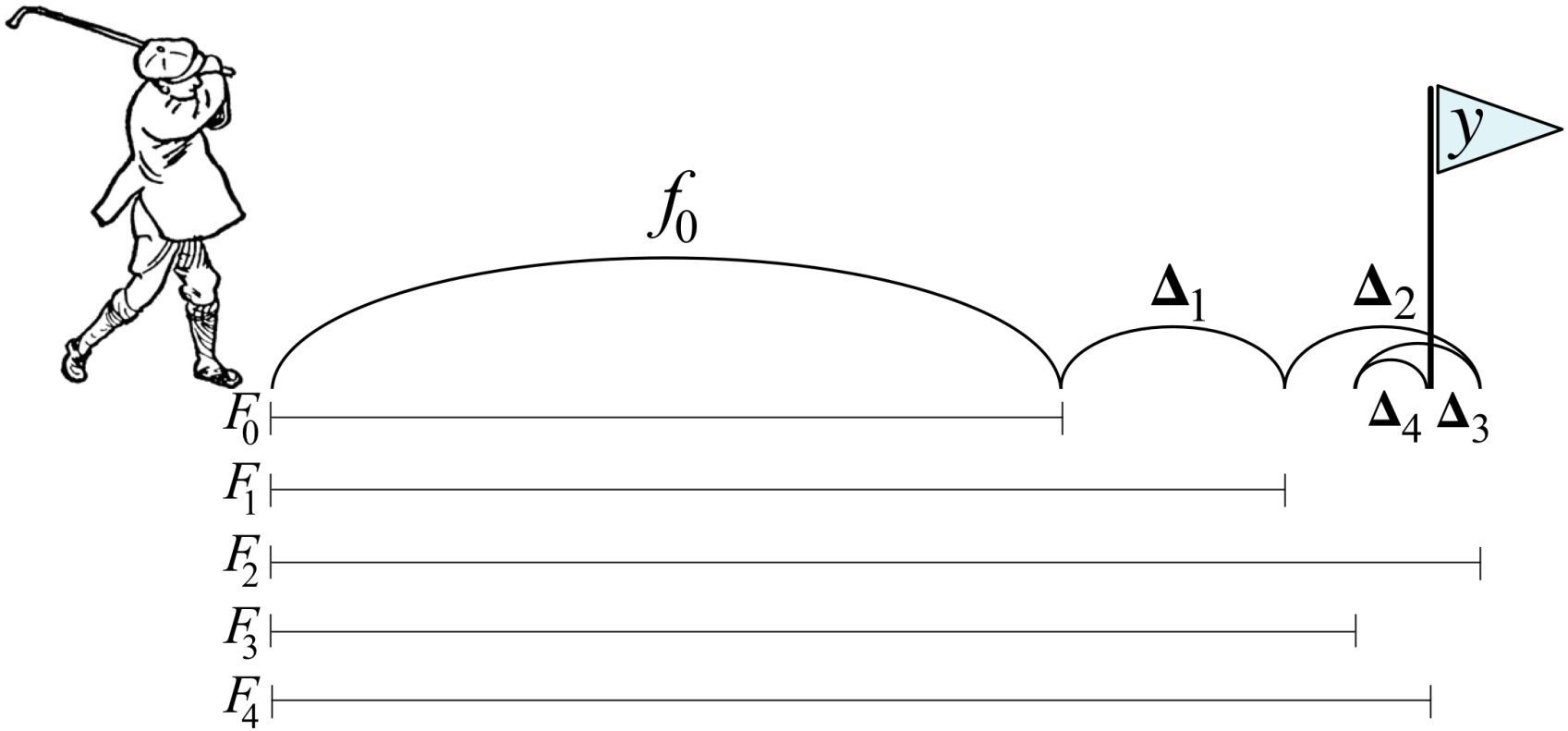
Prediksi dengan Gradient Boosting

$$\begin{aligned}\hat{y} &= f_0(x) + \Delta_1(x) + \Delta_2(x) + ... + \Delta_M(x) \\ &= f_0(x) + \sum_{m=1}^{M} \Delta_M(x) \\ &= F_M(x)\end{aligned}$$

atau dengan rekurens

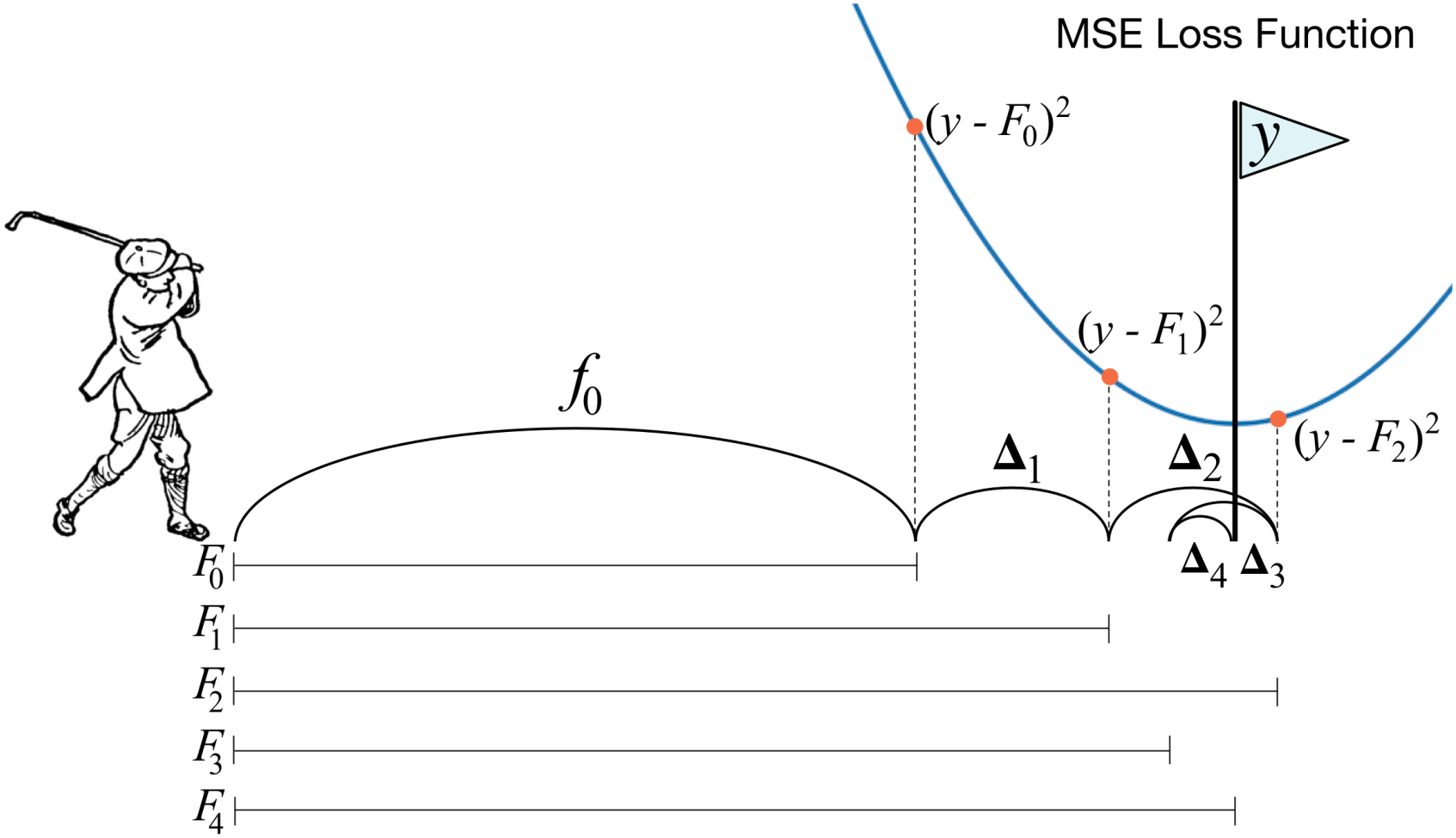
$$\begin{aligned}F_0(x) &= f_0(x) \\ F_m(x) &= F_{m-1}(x) + \Delta_m(x)\end{aligned}$$

Gradient Boosting



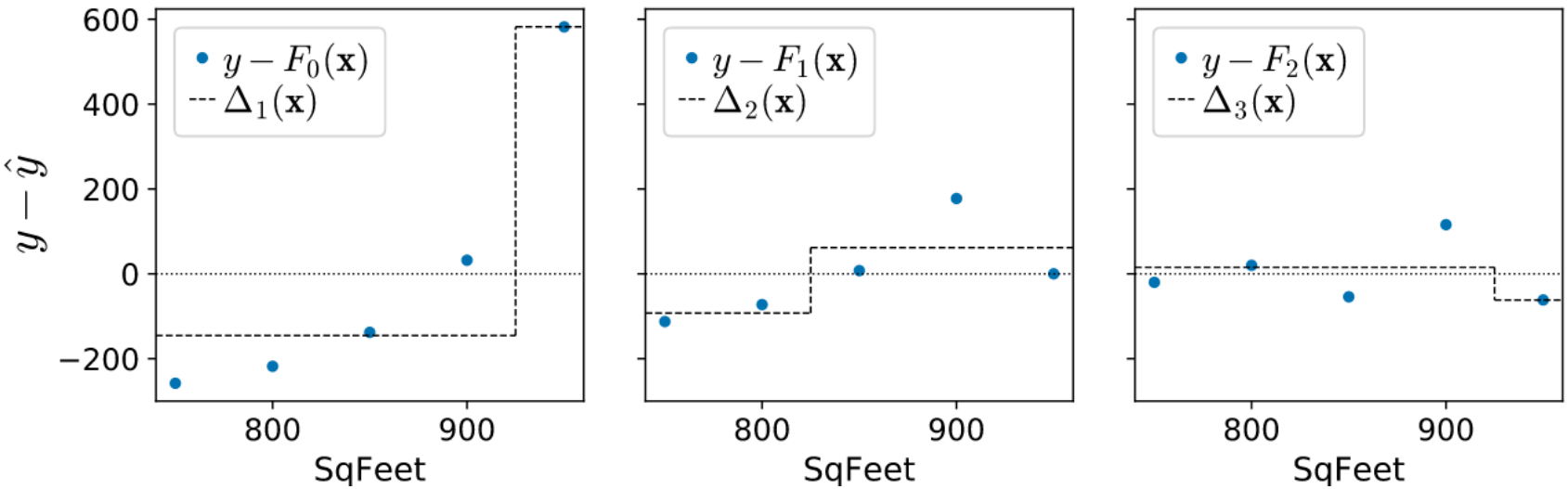
Gambar: Ilustrasi gradient boosting (Parr & Howard, 2018)

Gradient Boosting



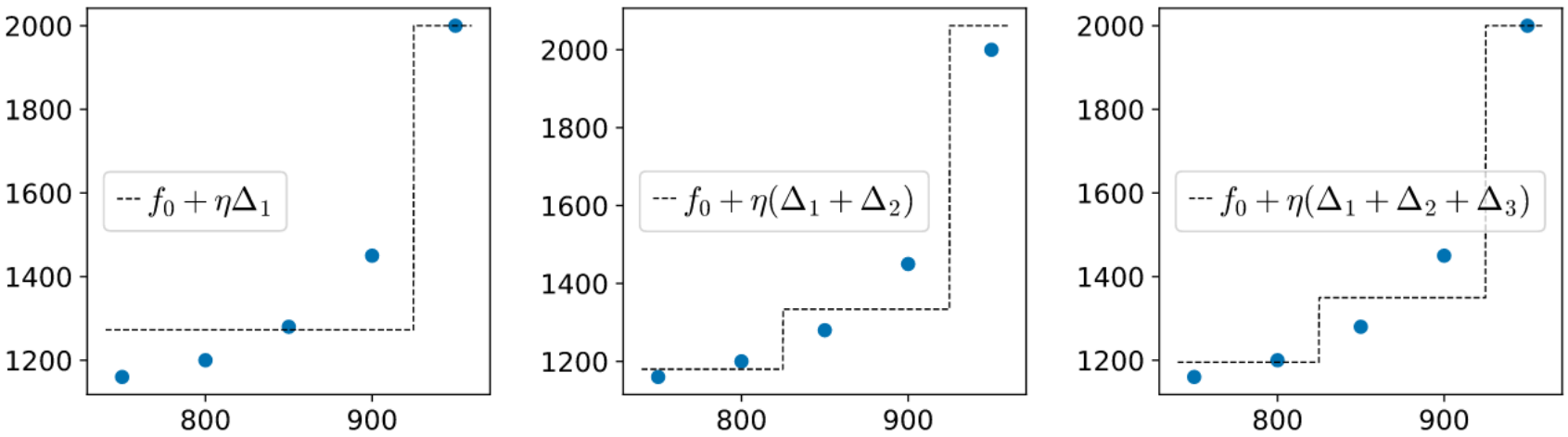
Gambar: Ilustrasi gradient boosting (Parr & Howard, 2018)

Gradient Boosting



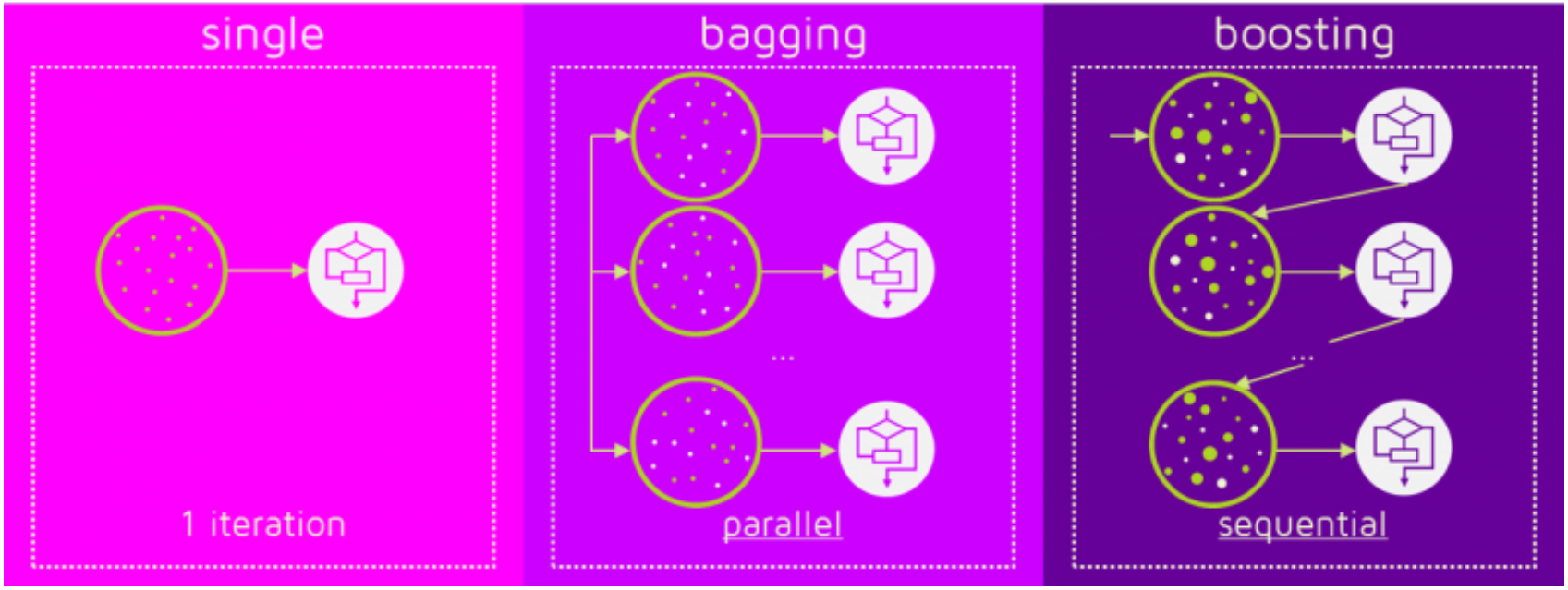
Gambar: Hasil penjumlahan pohon dengan gradient boosting (Parr & Howard, 2018)

Gradient Boosting



Gambar: Hasil penjumlahan pohon dengan gradient boosting (Parr & Howard, 2018)

Parallel vs Sequential

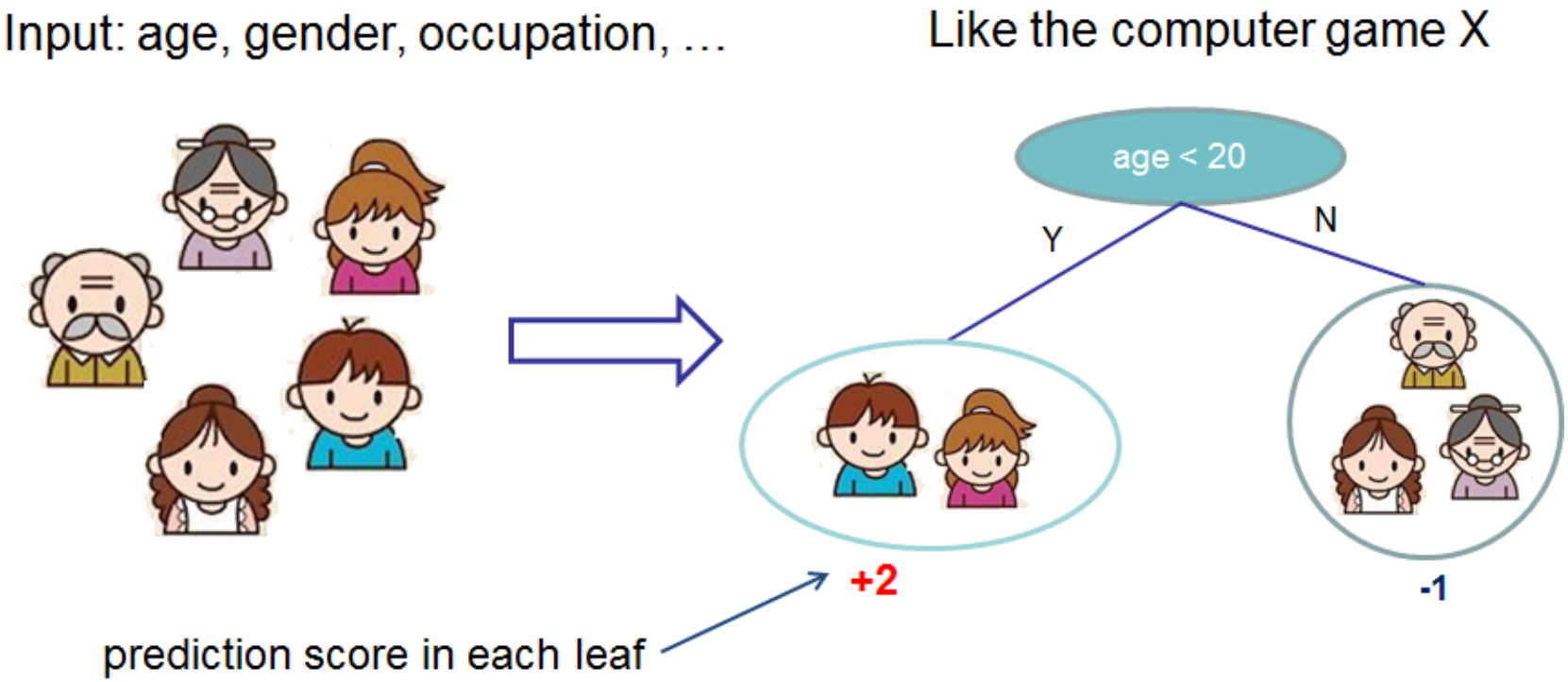


Gambar: Pembuatan model secara paralel dan sekuensial [QuantDare, 2016]

XGBoost

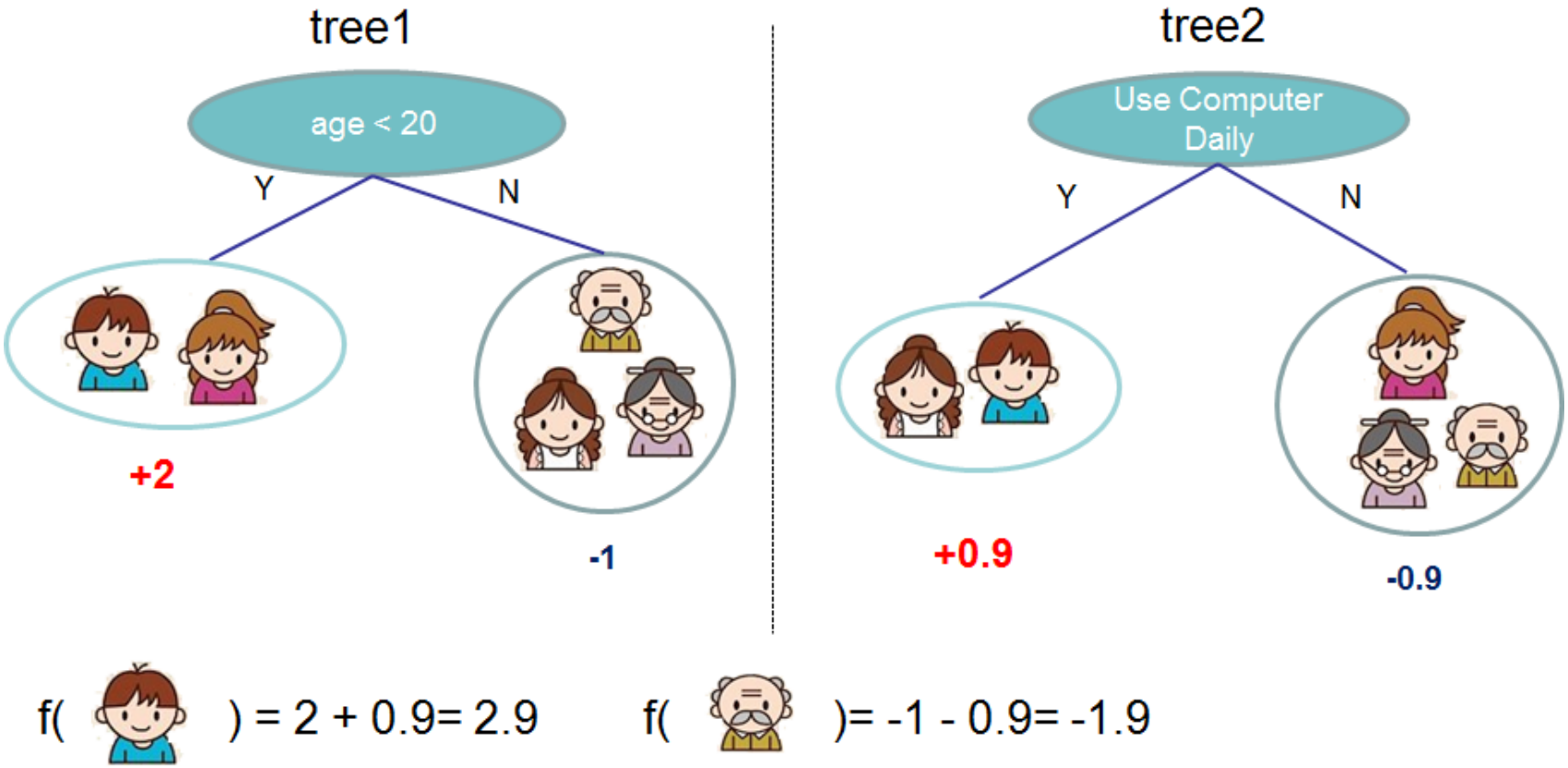
Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM.

Prediksi dengan XGBoost



Gambar: Ilustrasi prediksi

Prediksi dengan XGBoost



Gambar: Ilustrasi prediksi

Stacking

Stacking

1. Ada N model (*strong learner*)
2. Untuk setiap model:
    1. Latih model
    2. Prediksi data dengan model yang sudah dilatih $\rightarrow \hat{y}_n$
3. Gunakan setiap prediksi $\hat{y}_n$ sebagai metafitur
4. Latih model baru dengan menggunakan metafitur ini

References

QuantDare (2016)

What is the difference between Bagging and Boosting?

https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/

Kumpulkan dataset pekan depan
untuk capstone project

Terima kasih