# Bike2Vec: Vector Embedding Representations of Road Cycling Riders and Races

Ethan Baron

University of Toronto & Zelus Analytics

June 26, 2023

# My Collaborators

Bram Janssens, PhD


Matthias Bogaert, PhD


Le
de TOUR
France

# Road Cycling

# Cycling Analytics

| Source | $y$ | $f$ | $X$ |
|---|---|---|---|
| De Spiegeleer 2019 | Grand tour stages | Various | Historical results, terrain, weather, rider characteristics |
| Mortirolo 2019 | Race outcome probabilities | Bayesian Additive Regression Trees | Historical and recent results, rider ratings, team indicators |
| Kholkine, De Schepper, et al. 2020 | Tour of Flanders top 10 | XGBoost | Historical and recent results, weather, team indicators |
| Kholkine, Servotte, et al. 2021 | One-day races top 10 | XGBoost | Historical results, rider age |
| Demunter 2021 | Rider rankings in races | Various | Historical and recent results, race cluster |
| Van Bulck, Vande Weghe, and Goossens 2021 | PCS points in first 3 years of U23 prospects | Linear regression, random forest | Summaries of U23 performance |
| Janssens, Bogaert, and Maton 2022 | PCS points in first 2 years of U23 prospects | Random forest | Results in particular U23 races, imputed non-participated race results |

# Motivation

Q: Can we build a generalized prediction algorithm which does not rely on domain knowledge and manual feature engineering?

# Motivation

Q: Can we build a generalized prediction algorithm which does not rely on domain knowledge and manual feature engineering?

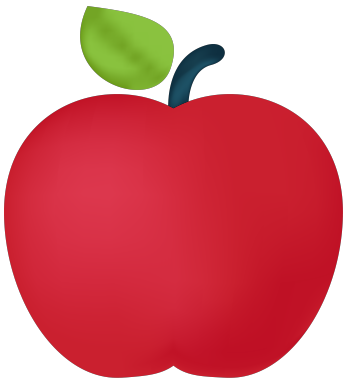
A: Yes, by using vector embeddings!

# Vector Embeddings

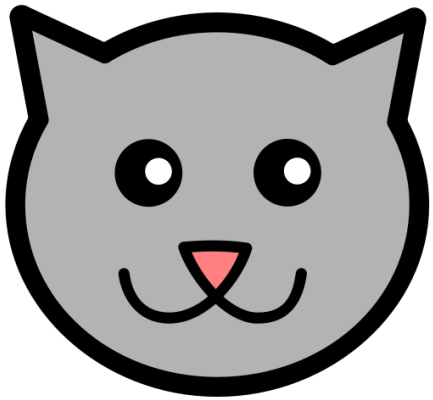
- Encode objects (e.g. words, movies) as vectors of certain length

# Vector Embeddings

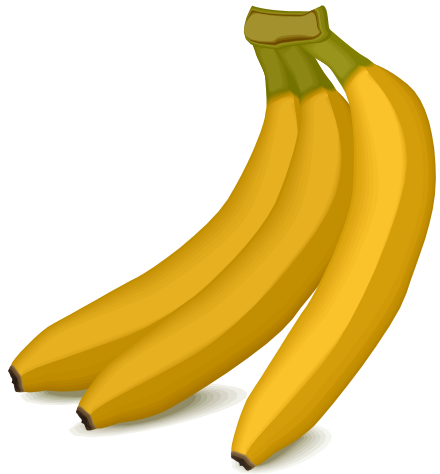
- Encode objects (e.g. words, movies) as vectors of certain length



"apple"
$(0.1, 0.5, 0.2)$

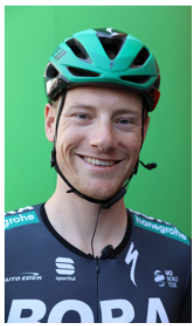


"cat"
$(-0.5, -0.2, 0.4)$



"banana"
$(0.2, 0.4, 0.2)$

# Idea



Sam Bennett

| | | | | |
|---|---|---|---|---|
| 0.49 | -2.94 | -0.14 | 3.12 | -4.48 |

•


| | | | | |
|---|---|---|---|---|
| -1.29 | -0.36 | 1.14 | -0.44 | 0.79 |

= -4.6

Nairo Quintana

| | | | | |
|---|---|---|---|---|
| 1.10 | 2.94 | 0.04 | 0.43 | 1.72 |

•


| | | | | |
|---|---|---|---|---|
| -1.29 | -0.36 | 1.14 | -0.44 | 0.79 |

= -1.3

# Learning Embeddings

Q: How to learn/train the vector embeddings?

# Learning Embeddings

Q: How to learn/train the vector embeddings?

A: Use historical results!

Specifically, minimize loss between predicted and actual points scored by riders in races

# Loss Function

$$L(y, R, S) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\sigma(R_{r(i)} \cdot S_{s(i)})) + (1 - y_i) \log(1 - \sigma(R_{r(i)} \cdot S_{s(i)}))$$

- $R$ Rider embeddings matrix
- $S$ Race embeddings matrix
- $y$ Observed results
- $N$ Total # of results
- $r(i)$ Index of rider for result $i$
- $s(i)$ Index of race for result $i$
- $\sigma(x)$ Sigmoid activation

# Loss Function

$$L(y, R, S) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\sigma(R_{r(i)} \cdot S_{s(i)})) + (1 - y_i) \log(1 - \sigma(R_{r(i)} \cdot S_{s(i)}))$$

- $R$ Rider embeddings matrix
- $S$ Race embeddings matrix
- $y$ Observed results
- $N$ Total # of results
- $r(i)$ Index of rider for result $i$
- $s(i)$ Index of race for result $i$
- $\sigma(x)$ Sigmoid activation

# Loss Function

$$L(y, R, S) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\sigma(R_{r(i)} \cdot S_{s(i)})) + (1 - y_i) \log(1 - \sigma(R_{r(i)} \cdot S_{s(i)}))$$

- $R$ Rider embeddings matrix
- $S$ Race embeddings matrix
- $y$ Observed results
- $N$ Total # of results
- $r(i)$ Index of rider for result $i$
- $s(i)$ Index of race for result $i$
- $\sigma(x)$ Sigmoid activation

# Loss Function

$$L(y, \tilde{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\tilde{y}_i) + (1 - y_i) \log(1 - \tilde{y}_i)$$

$y$ Observed results

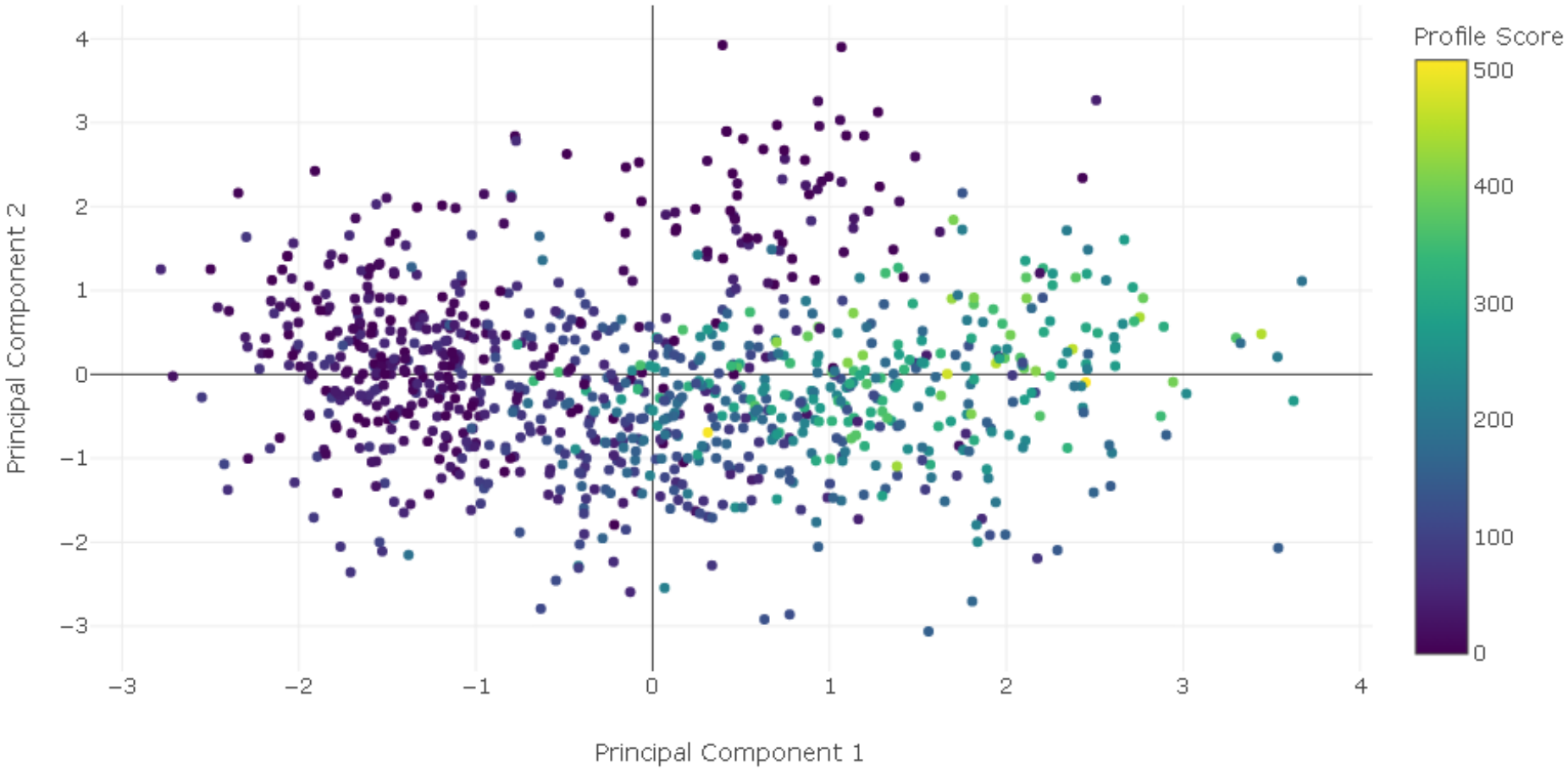
$\tilde{y}$ Predicted results

$N$ Total # of results

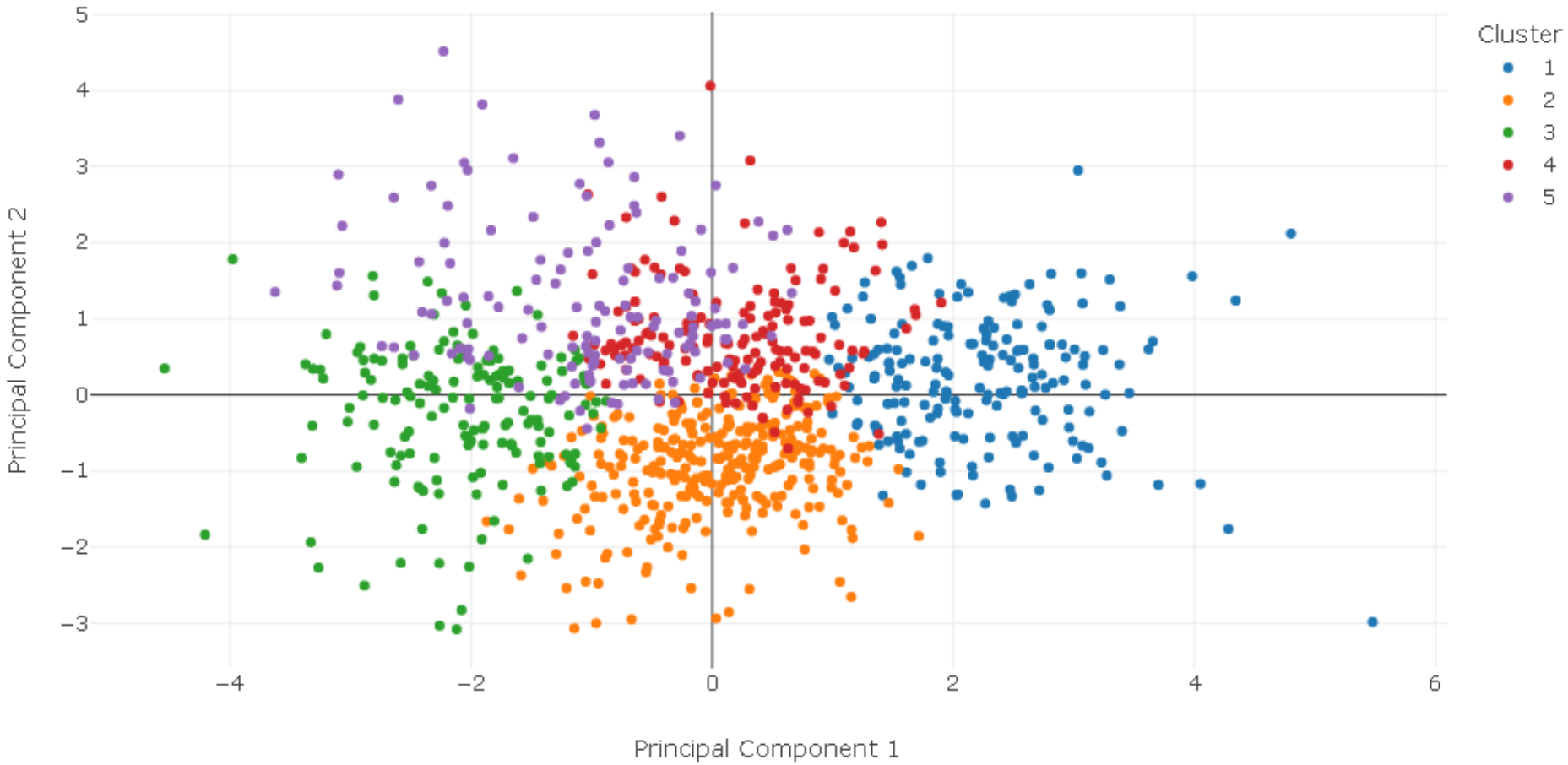
# Implementation

- Collect results from 2016-2022 UCI World Tour (`procyclingstats.com`)
- Train embeddings for 958 riders and 973 races
- Embeddings dimension $= 5$
- Adam optimizer, learning rate $= 0.001$, 100 epochs
- Reproducible code at `https://github.com/baronet2/Bike2Vec`

# Validating Race Embeddings



# Validating Rider Embeddings



# Validating Rider Embeddings

| Cluster | Examples of Riders |
|---|---|
| 1 | Sagan, Kristoff, Viviani, Ewan, Bennett |
| 2 | Van Avermaet, Colbrelli, Naesen, Mohoric |
| 3 | Quintana, Valverde, Roglic, Pogacar |
| 4 | Van Aert, Matthews, Stuyven, Kwiatkowski |
| 5 | Van der Poel, Gilbert, Lampaert, Stybar |

# Rider Similarities

| Rider 1 | Rider 2 |
|---|---|
| POGAČAR Tadej | ROGLIČ Primož |
| SAGAN Peter | COLBRELLI Sonny |
| ALAPHILIPPE Julian | HIRSCHI Marc |
| YATES Simon | BARDET Romain |
| EVENEPOEL Remco | ALMEIDA João |
| QUINTANA Nairo | ZAKARIN Ilnur |
| VIVIANI Elia | GREIPEL André |
| DENNIS Rohan | CAVAGNA Rémi |

# Applications

1. Race prediction
   - Incorporate rider/race/team features
   - Use race elevation profile in embeddings
2. Talent identification
   - Develop time-varying rider embeddings
   - Predict rider development over multiple years
3. Team construction
   - Identify types of riders needed to win
   - Develop evaluation method for teammates

# Conclusion

1. Cycling analytics is hard: each race/rider is different
2. Vector embeddings can automatically capture key characteristics
3. Can use embeddings for analysis without feature engineering

# References


De Spiegeleer, Emiel (June 2019). "Predicting Cycling Results using Machine Learning". MA thesis. Ghent: Ghent University. URL: https://libstore.ugent.be/fulltxt/RUG01/002/785/834/RUG01-002785834%5C%5F2019%5C%5F0001%5C%5FAC.pdf.

Demunter, Jarne (June 2021). "Predicting Ranking Multientrant Races: Road Cycling". MA thesis. Ghent: Ghent University. URL: https://libstore.ugent.be/fulltxt/RUG01/003/010/353/RUG01-003010353%5C%5F2021%5C%5F0001%5C%5FAC.pdf.

Janssens, Bram, Matthias Bogaert, and Mathijs Maton (Jan. 2022). "Predicting the next Pogačar: a data analytical approach to detect young professional cycling talents". In: *Annals of Operations Research*, pp. 1–32. DOI: https://doi.org/10.1007/s10479-021-04476-4.

Kholkine, Leonid, Tom De Schepper, et al. (Sept. 2020). "A Machine Learning Approach for Road Cycling Race Performance Prediction". In: *Machine Learning and Data Mining for Sports Analytics*. Ed. by Ulf Brefeld et al. Communications in Computer and Information Science. Springer International Publishing, pp. 103–112. ISBN: 978-3-030-64912-8. DOI: 10.1007/978-3-030-64912-8_9.

Kholkine, Leonid, Thomas Servotte, et al. (Oct. 2021). "A Learn-to-Rank Approach for Predicting Road Cycling Race Outcomes". In: *Frontiers in Sports and Active Living* 3, p. 714107. ISSN: 2624-9367. DOI: 10.3389/fspor.2021.714107.

Mortirolo (July 2019). *Cycling prediction method*. URL: https://mortirolo.netlify.app/post/cycling-prediction-method/ (visited on 05/06/2023).

Van Bulck, David, Arthur Vande Weghe, and Dries Goossens (Oct. 2021). "Result-based talent identification in road cycling: Discovering the next Eddy Merckx". In: *Annals of Operations Research*. DOI: 10.1007/s10479-021-04280-0.