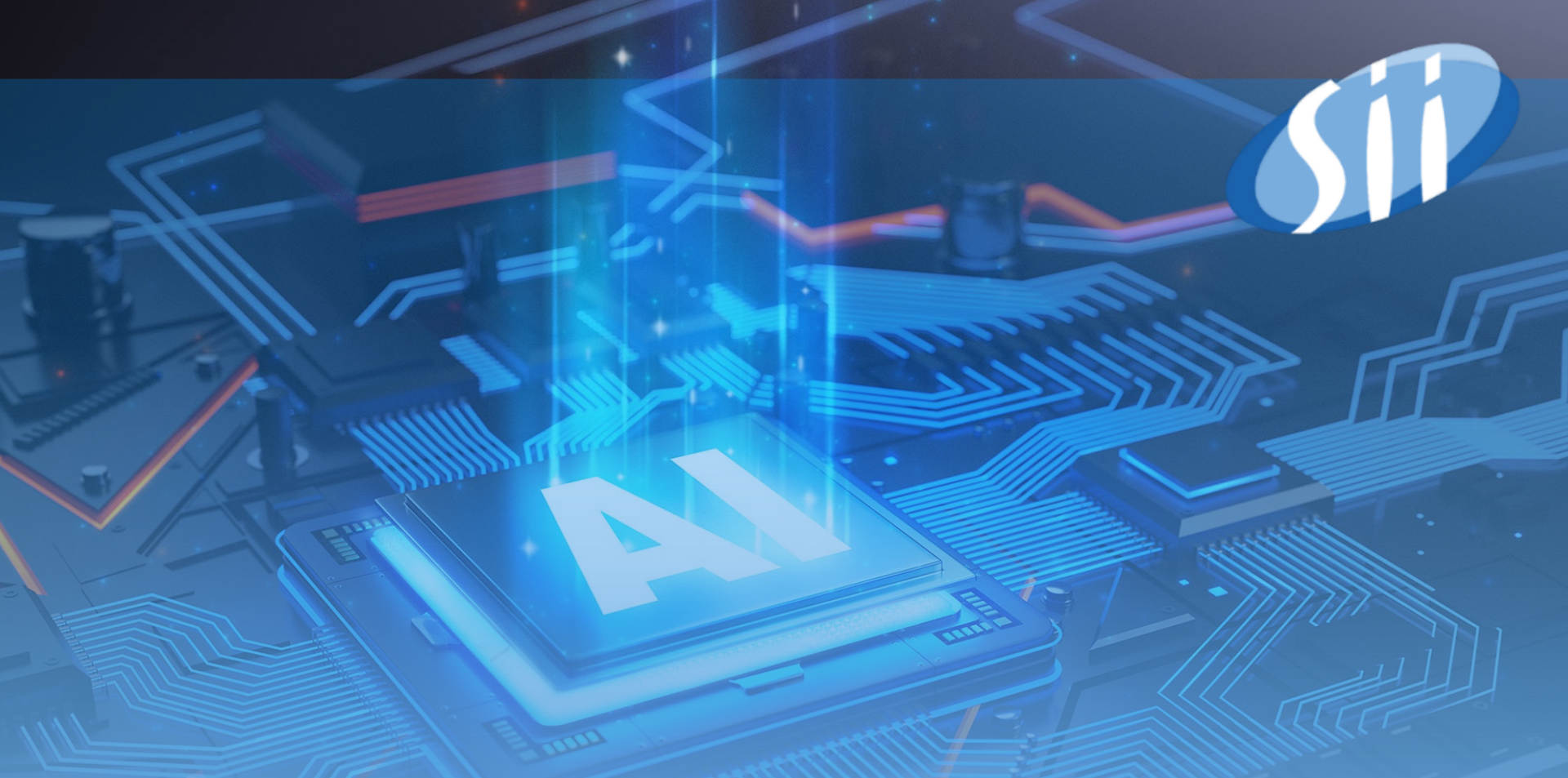


DEEP NEURAL DEDUPLICATION

Marcin Mosiołek
AI Architect

PyData Global

AGENDA

1. Business Case
2. Contrastive Learning
3. Our Solution
4. Reducing Search Space
5. Results
6. Next Steps
7. Summary

BUSINESS CASE

BUSINESS CASE

DODGE DATA & ANALYTICS

DODGE DATA & ANALYTICS | THE BLUE BOOK NETWORK

Solutions That Grow Your Business

Boost your spec rate.

Find, get specified and installed in more projects.

Bid better. Win more.

Find, bid and win the most profitable projects.

Find products for your projects.

Download details on 100,000+ products

Link to Dodge. Nourish your business.

Pipe Dodge data directly into your business.

Uncover unique insights. Gain an edge.

Leverage our expert analysts and researchers.

BUSINESS CASE

ARTICLES DEDUPLICATION

www.123.com

The Board of Construction, New York is requesting sealed bids for Empire State Building Renovation. The deadline for the bids has been set for December 30th at 2:00 PM in the owner officess located at the 123 5th Avenue New York. Plans might be viewed at the building administration office at 34th St, New York or at the construction manager office Joe Doe, 16th St. New York. Alternatively they might be delivered by email…

www.abc.ue

Empire State building requires Rennovation, as decided by the Board of Construction. The plans of the work might be found at the office of Joe Doe company's. Please submit the bids by the end of December. The project value is estimated at $5-10M...

DEDUPLICATION: THE ALGORITHMIC APPROACH

- A rule-based heuristics to identify duplicates but not very accurate
- Data stewards to approve found duplicates or identify manually
- A large dataset of labeled data

https://www.adultswim.com/videos/rick-and-morty

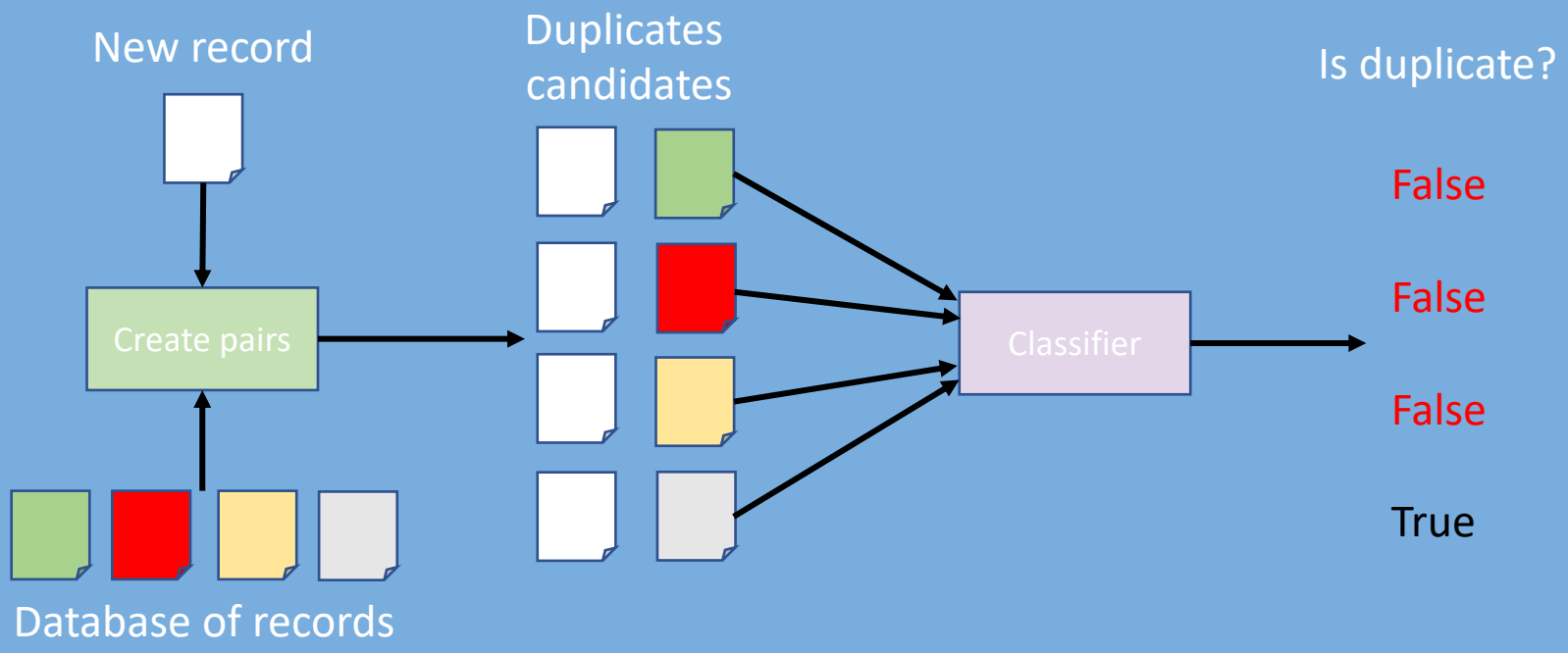
DATA OVERVIEW

- Millions of data records
- Duplicates assigned
- Records are dictionaries of named entities
- Named entities extracted from free text articles
- Variety of data fields including text, continuous and categorical types


```
(…)
"project_title": "Rennovation of the Empire State building",
"project_address": "34th Street",
"project_city": "New York",
"project_min_value": "$5M",
"project_stage": "bidding",
"architect_name": "Joe Doe",
"project_category": "rennovation",
"bid_date": "25th December",
(...)
```

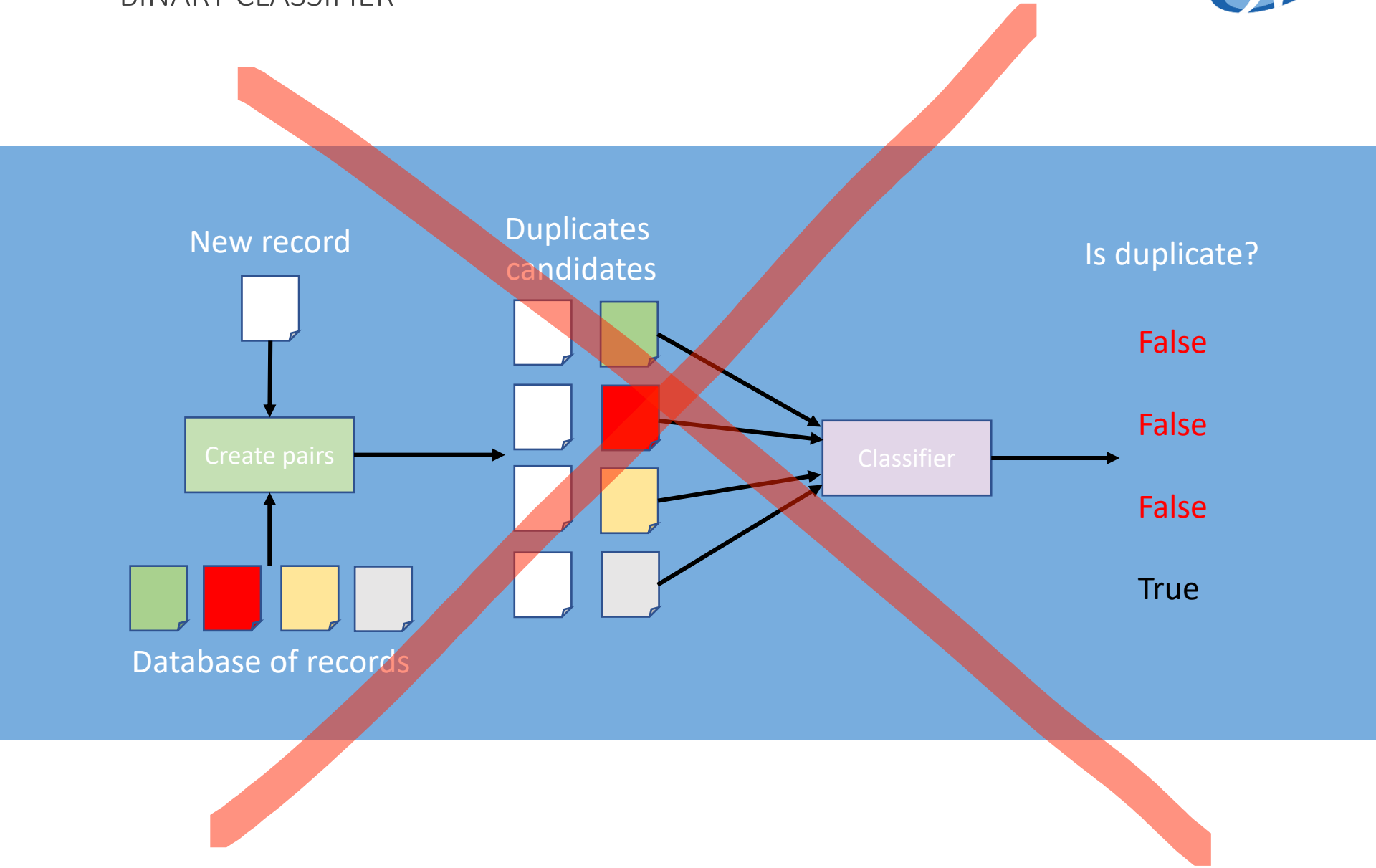
MACHINE LEARNING

BINARY CLASSIFIER



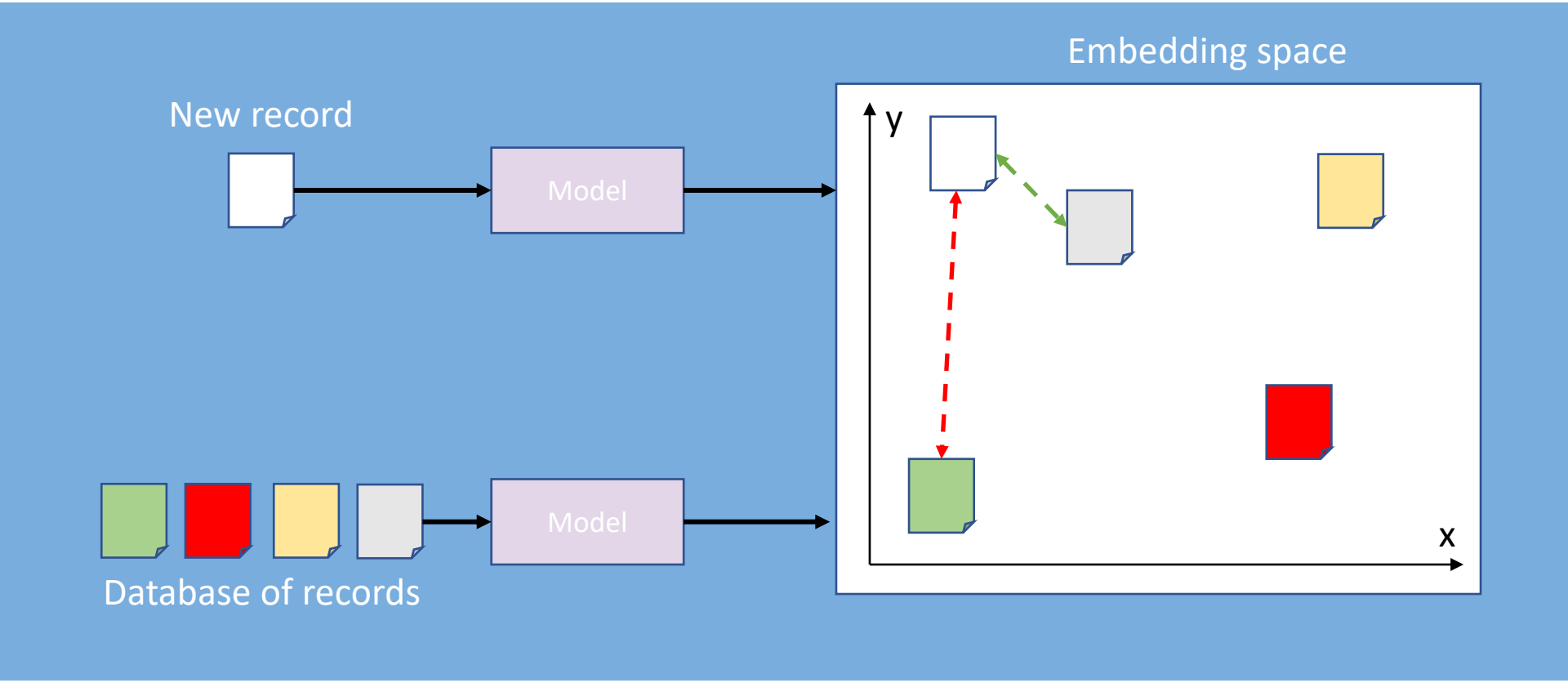
MACHINE LEARNING

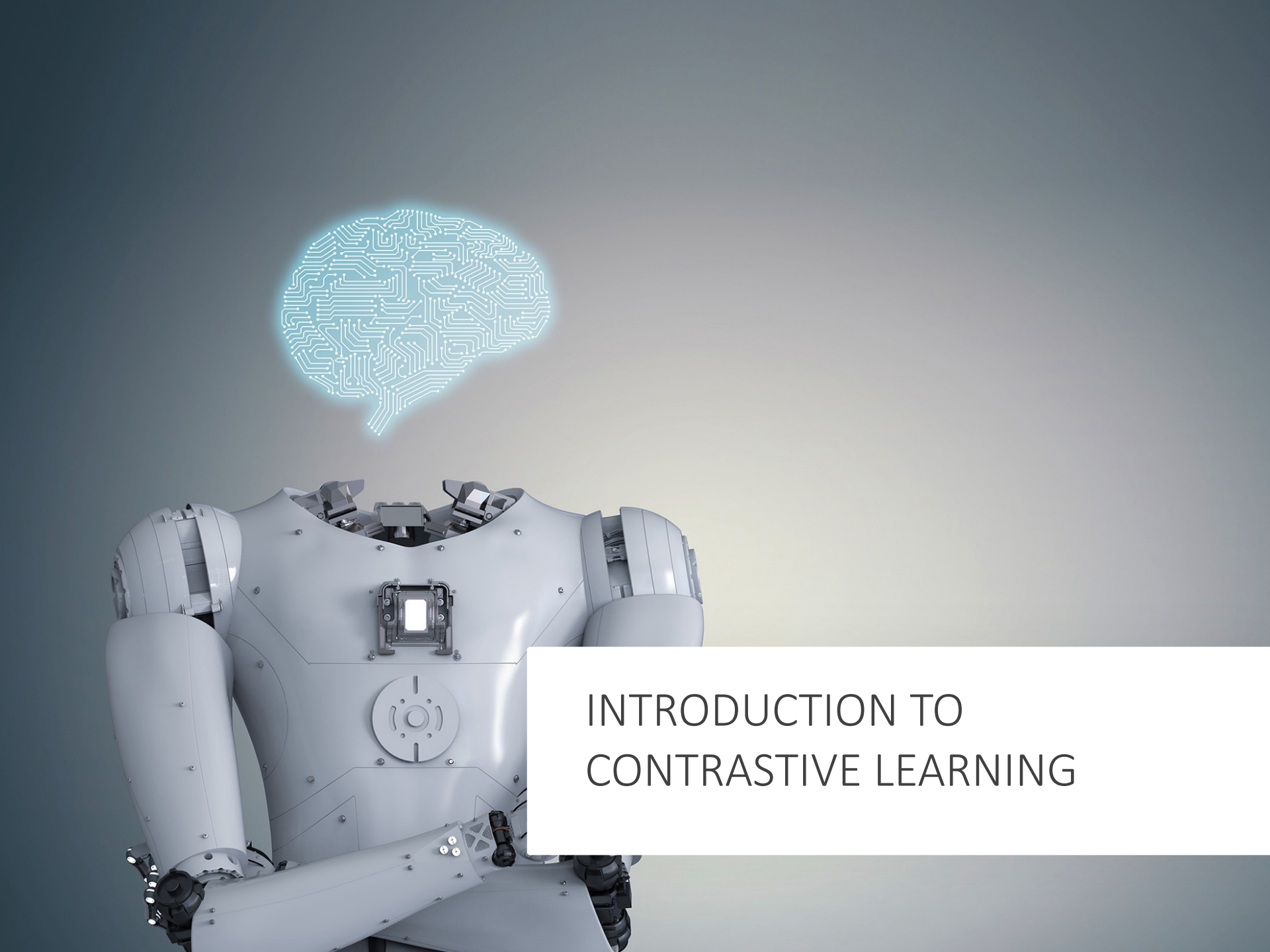
BINARY CLASSIFIER



MACHINE LEARNING

REPRESENTATION LEARNING





INTRODUCTION TO CONTRASTIVE LEARNING

CONTRASTIVE LEARNING

Learning a Similarity Metric Discriminatively, with Application to Face Verification

Sumit Chopra Raia Hadsell Yann LeCun

Courant Institute of Mathematical Sciences
New York University
New York, NY, USA
{sumit, raia, yann}@cs.nyu.edu

FaceNet: A Unified Embedding for Face Recognition and Clustering

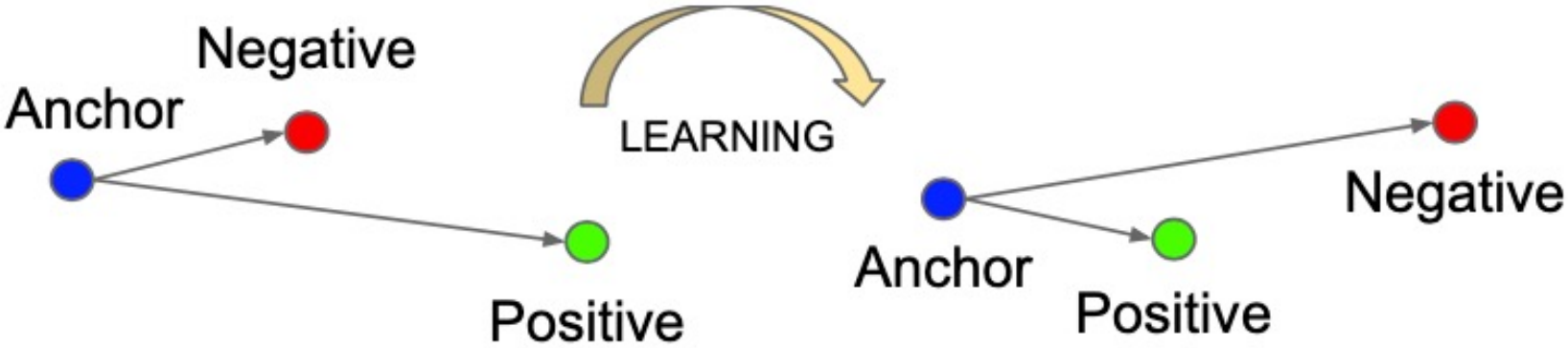
Florian Schroff
fschroff@google.com
Google Inc.

Dmitry Kalenichenko
dkalenichenko@google.com
Google Inc.

James Philbin
jphilbin@google.com
Google Inc.

TRIPLET LOSS

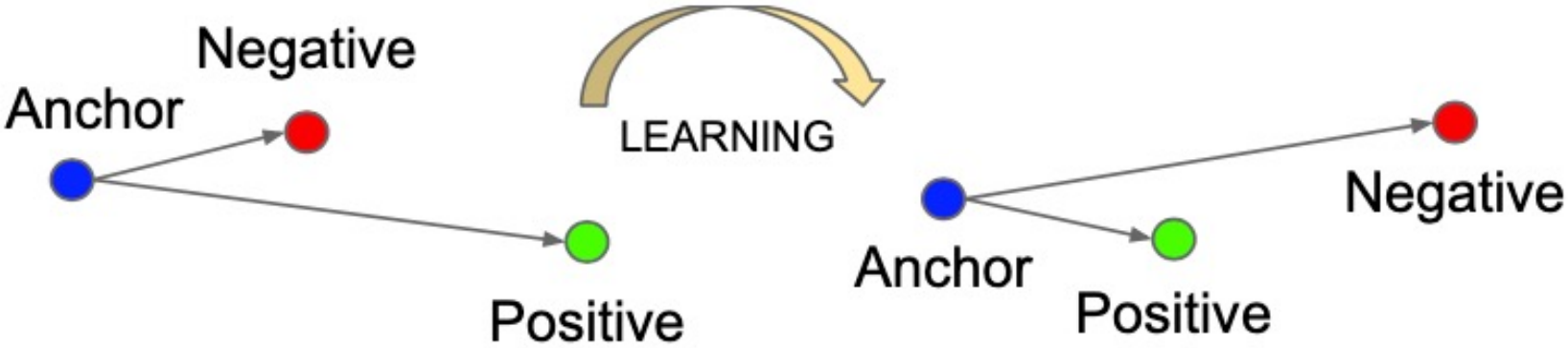
THREE INPUTS



$$d(A, P) - d(A, N) < 0$$

TRIPLET LOSS

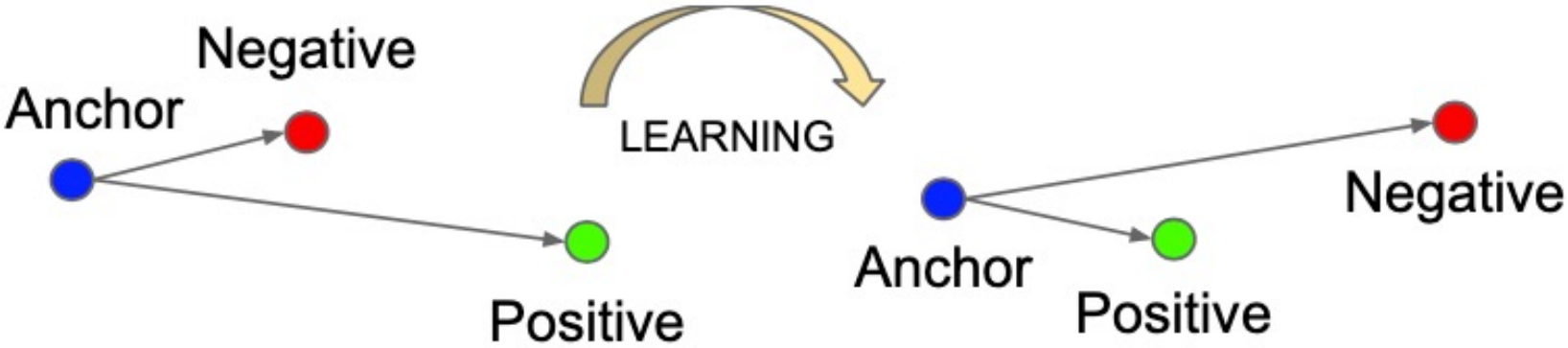
THREE INPUTS



$$L(A, P, N) = max(d(A, P) - d(A, N), 0)$$

TRIPLET LOSS

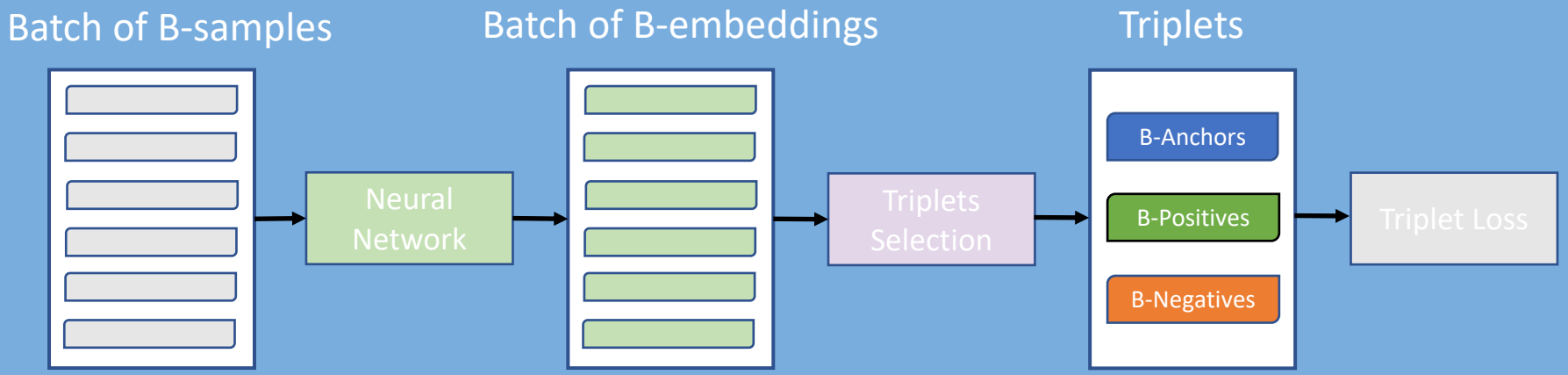
THREE INPUTS



$$L(A, P, N) = max(d(A, P) - d(A, N) + margin, 0)$$

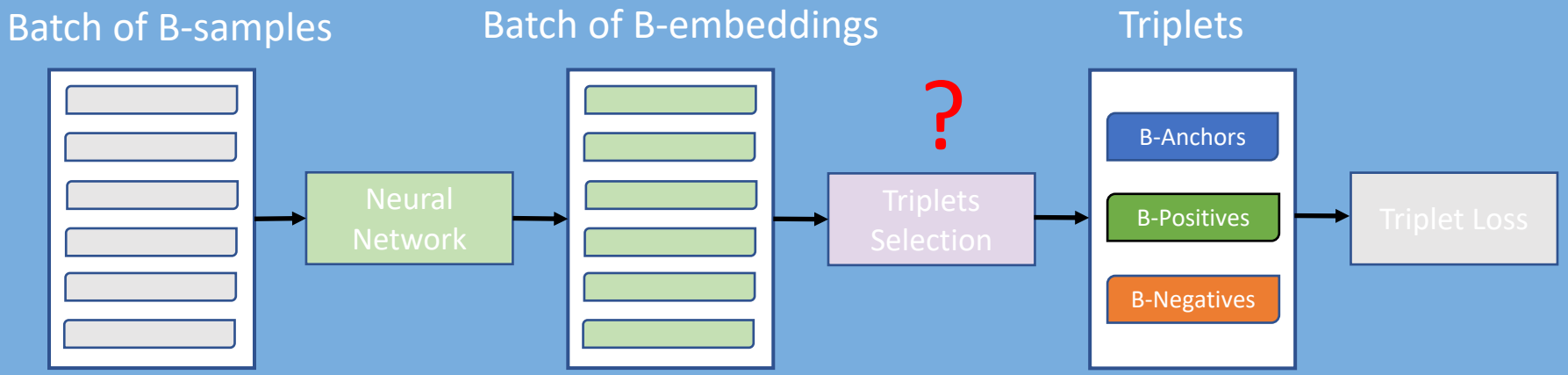
TRIPLET LOSS

HOW TO IMPLEMENT TRIPLETS SELECTION?



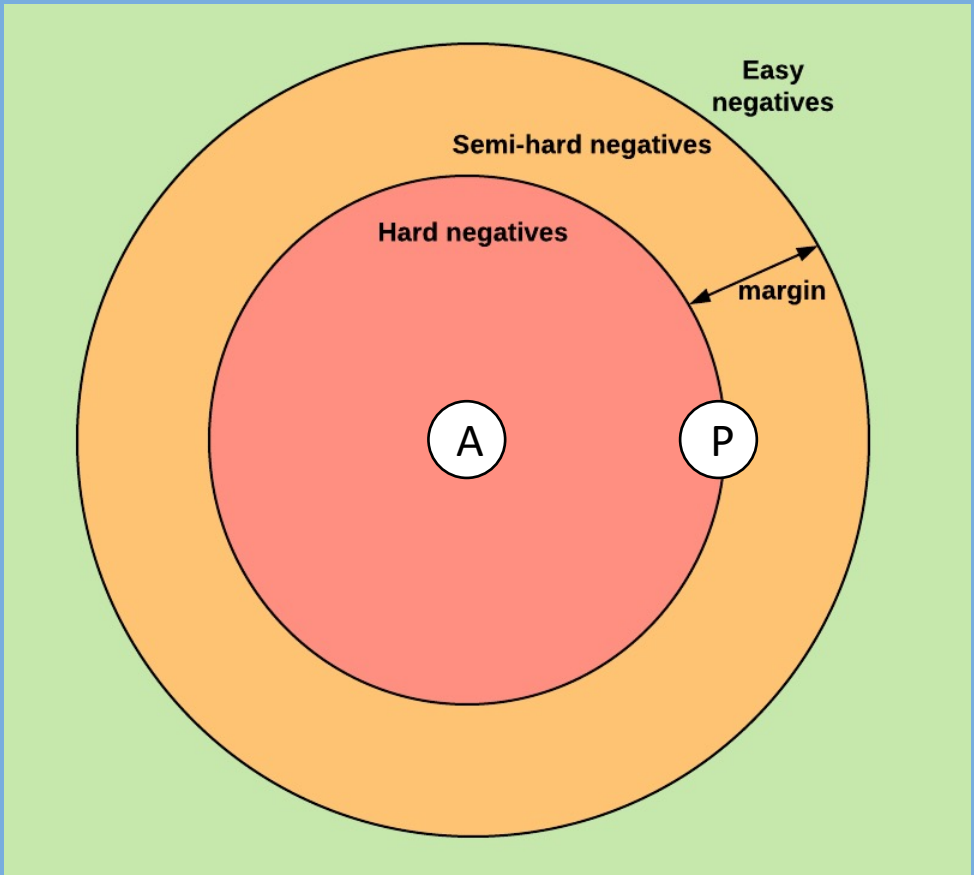
TRIPLET LOSS

HOW TO IMPLEMENT TRIPLETS SELECTION?



HOW TO SELECT TRIPLETS?

THREE KINDS OF NEGATIVES



https://omoindrot.github.io/triplet-loss

Easy negatives

$d(A, P) + \text{margin} < d(A, N)$

Hard negatives

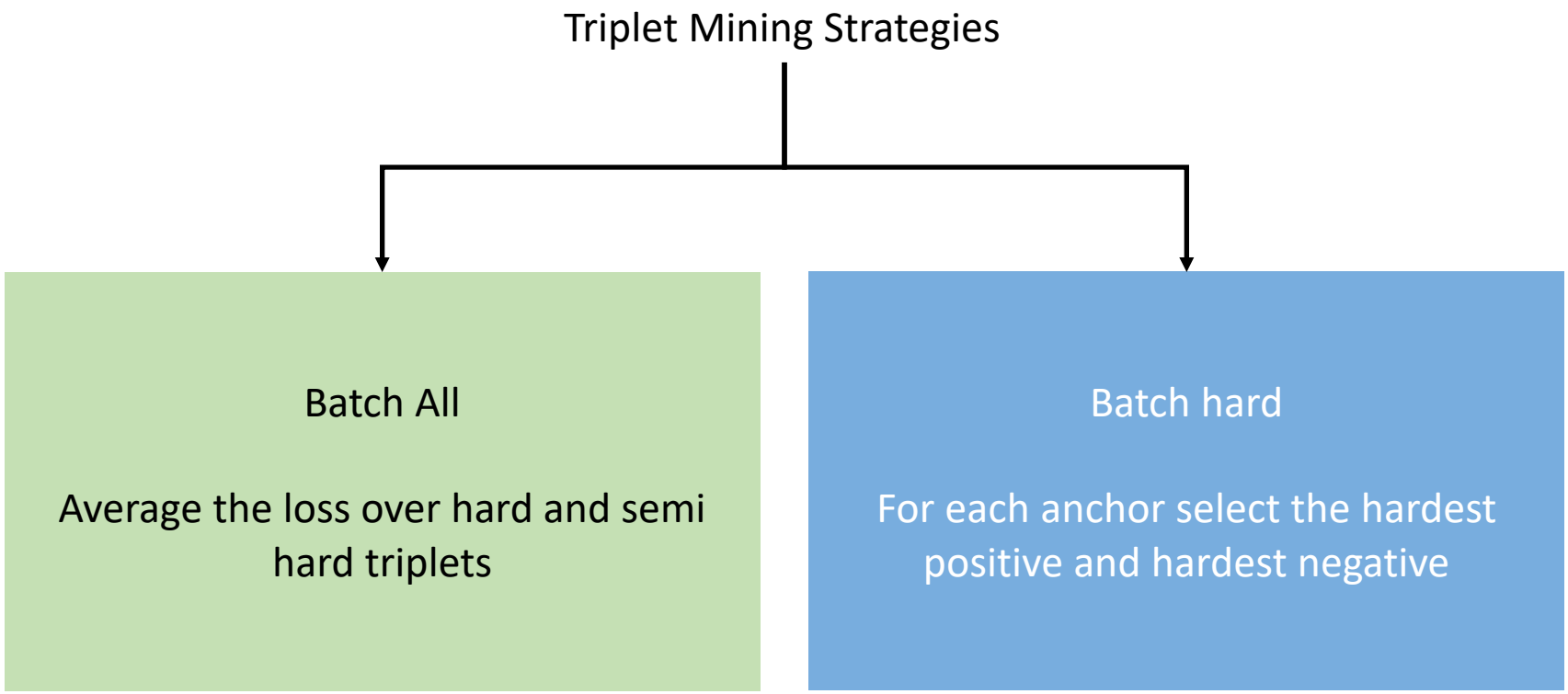
$d(A, N) < d(A, P)$

Semi-hard negatives

$d(A, P) < d(A, N) < d(A, P) + \text{margin}$

HOW TO SELECT TRIPLETS?

TRIPLET MINING STRATEGY



CONTRASTIVE LEARNING

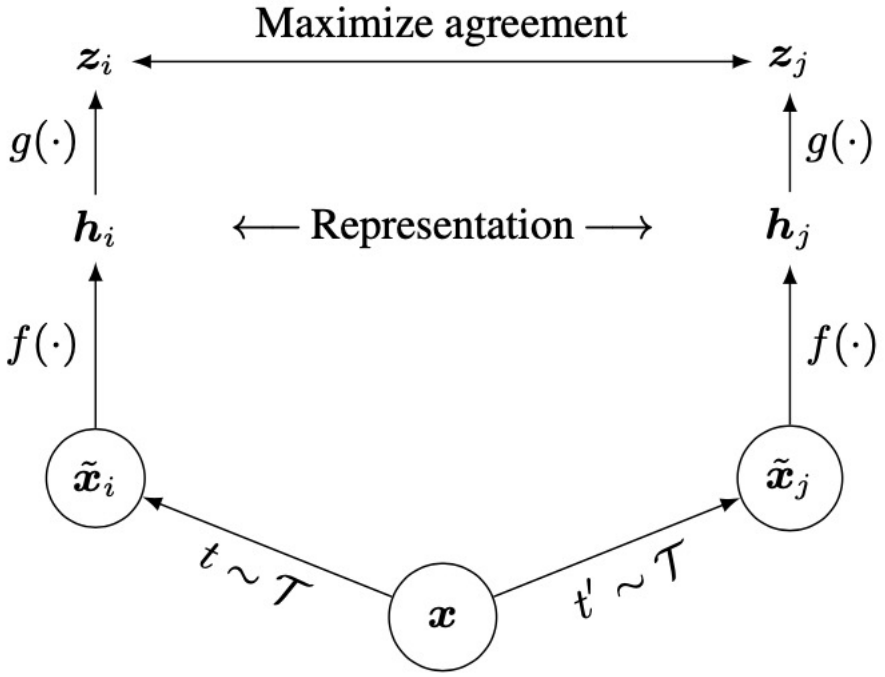
Is it possible to use contrastive learning without labeled data?

CONTRASTIVE LEARNING

UNSUPERVISED

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] **Simon Kornblith** [1] **Mohammad Norouzi** [1] **Geoffrey Hinton** [1]



CONTRASTIVE LEARNING

UNSUPERVISED

A Simple Framework for Contrastive Learning of Visual Representations

Ting Chen [1] **Simon Kornblith** [1] **Mohammad Norouzi** [1] **Geoffrey Hinton** [1]

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

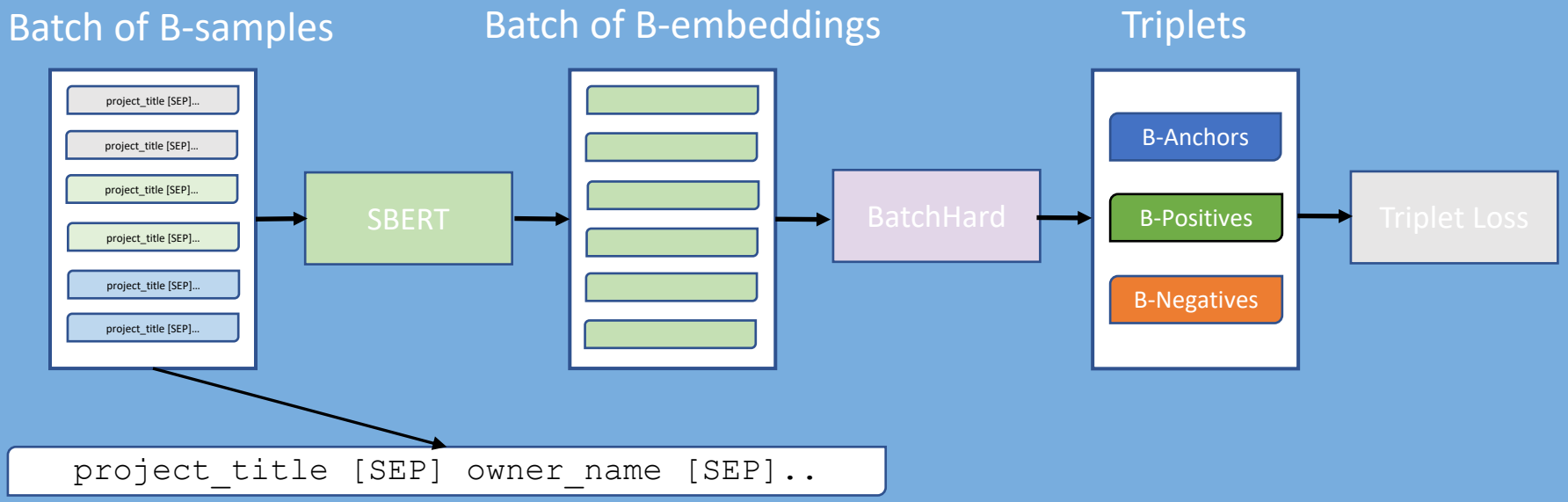
CONTRASTIVE LEARNING

We didn’t even scratch the surface…

OUR SOLUTION

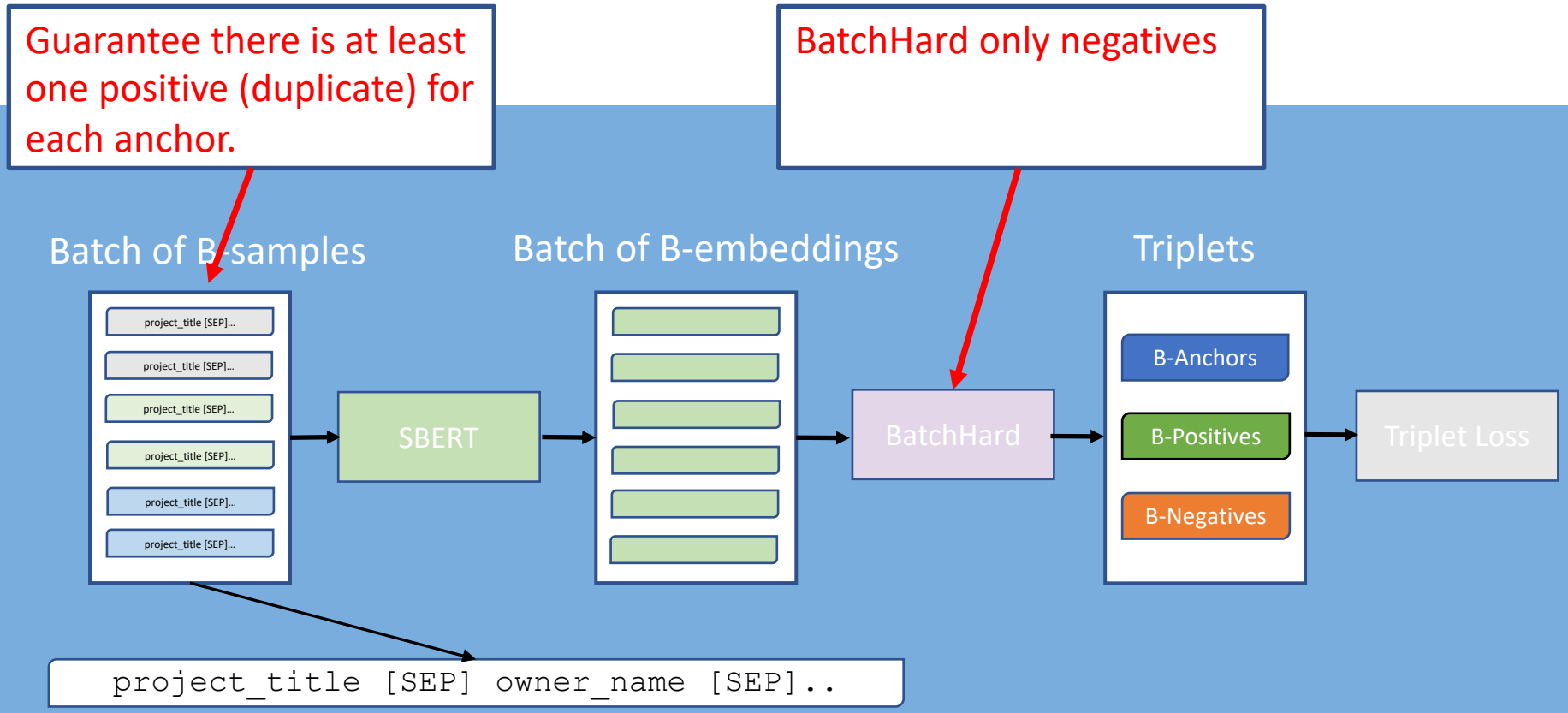
OUR APPROACH

TEXT FEATURES



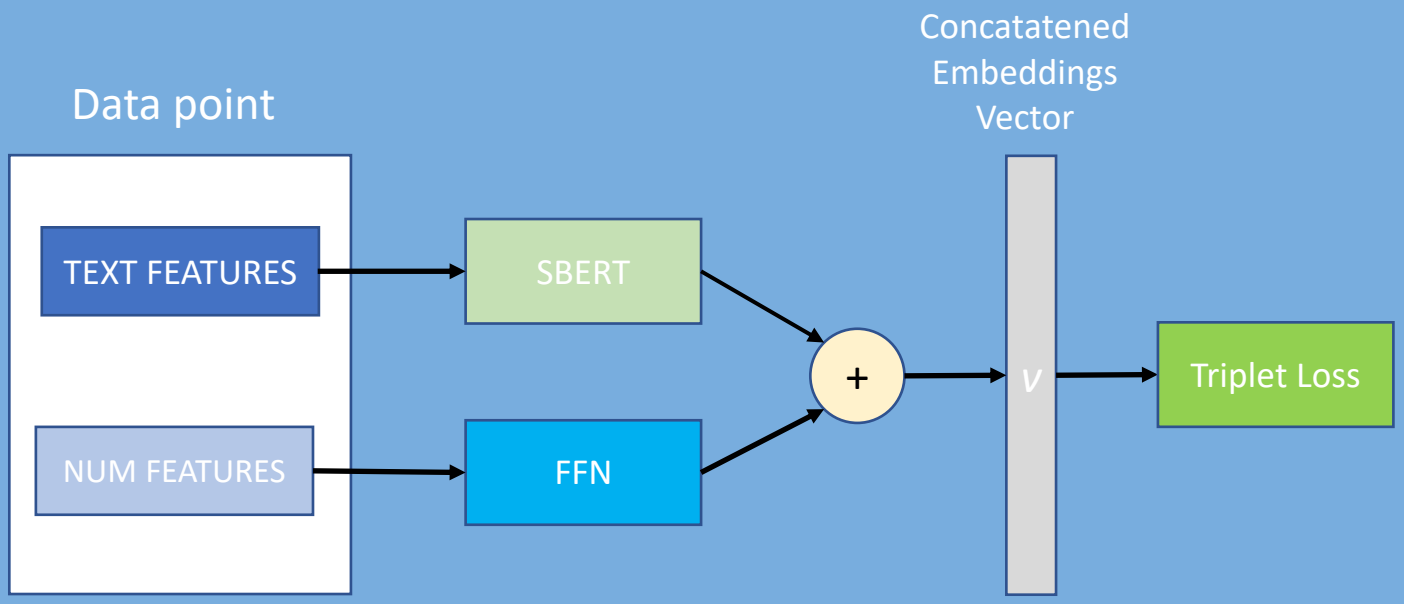
OUR APPROACH

TEXT FEATURES



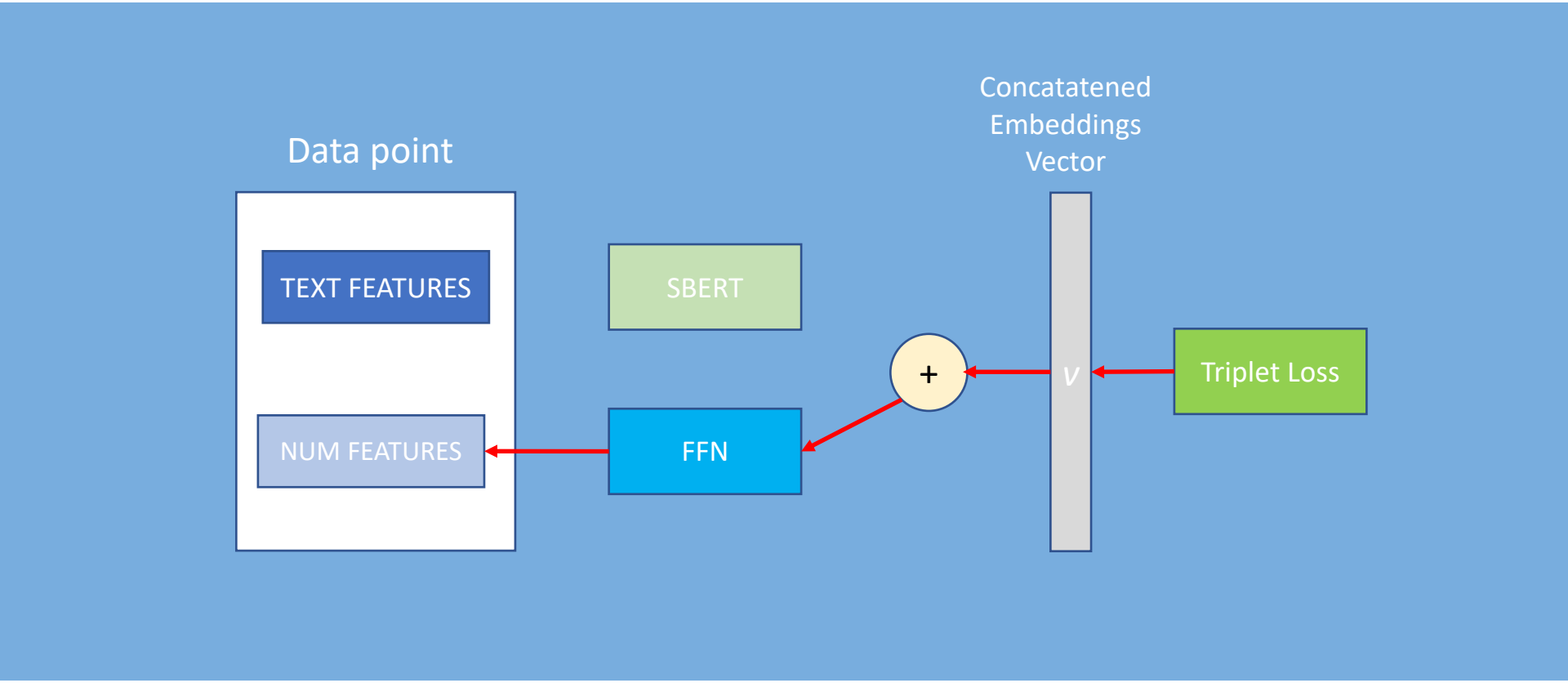
OUR APPROACH

ALL FEATURES



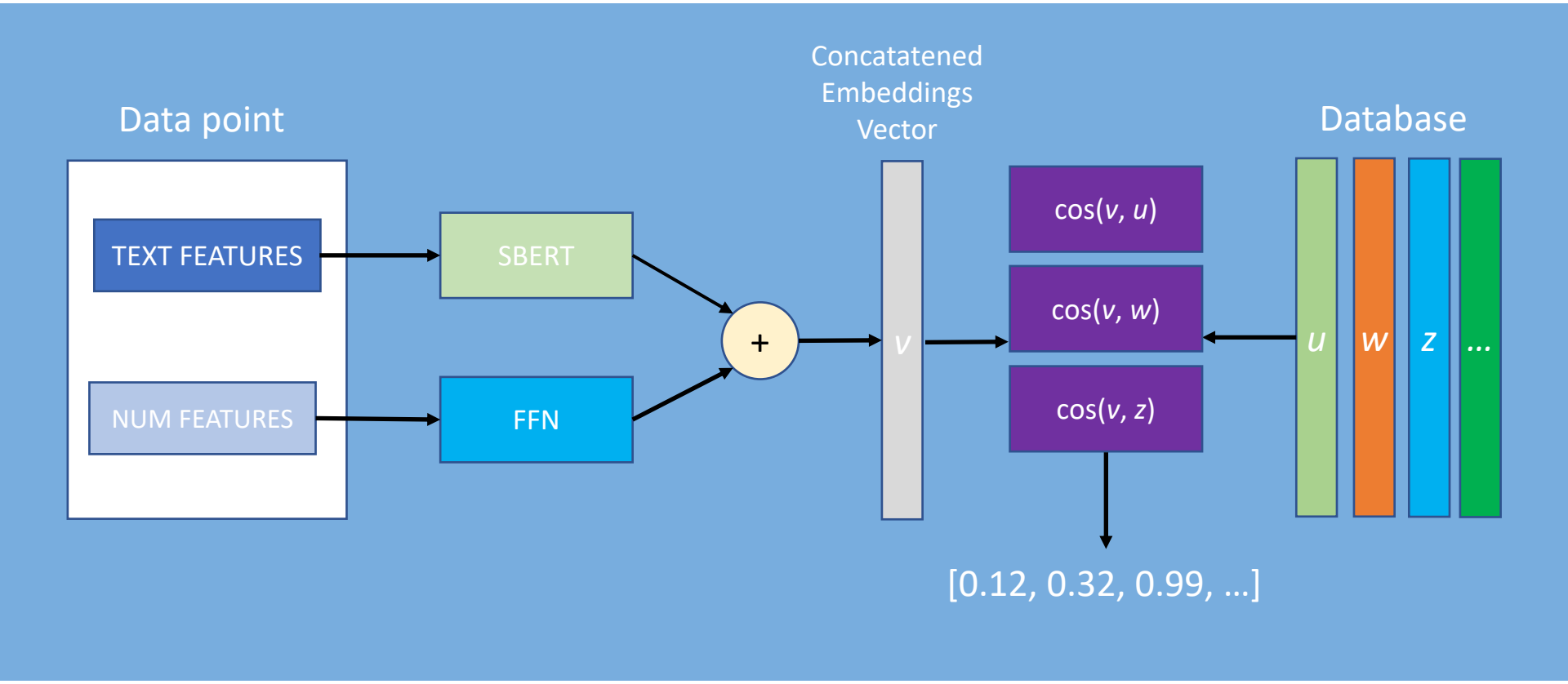
OUR APPROACH

ALL FEATURES



OUR APPROACH

INFERENCE



HOW TO SCALE IT UP?

LARGE SCALE NEAREST NEIGHBOR SEARCH

COMPLEXITY OF EXHAUSTIVE SEARCH

- To idenfity duplicates we need to find nearest neighbors
- The simplest way is to perform an exhaustive search
- It means we need to perform O(N*D) operations:
  - 1) We have a big number of records (N=4.5M)
  - 2) We have a high dimensional vectors (D~800)

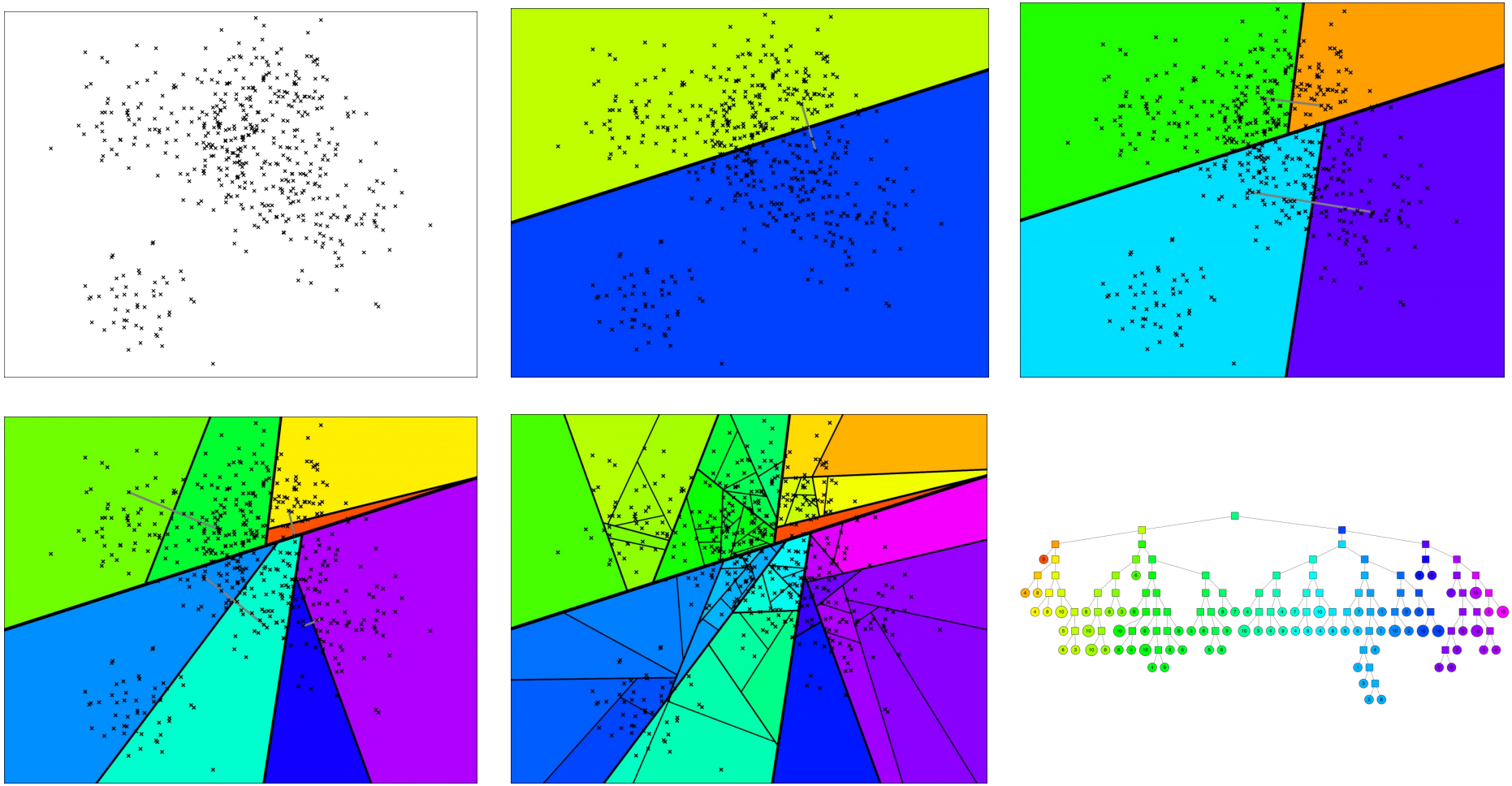
LARGE SCALE NEAREST NEIGHBOR SEARCH

COMPLEXITY OF EXHAUSTIVE SEARCH

- To idenfity duplicates we need to find nearest neighbors
- The simplest way is to perform an exhaustive search
- It means we need to perform O(N*D) operations:
  - 1) We have a big number of records (N=4.5M)
  - 2) We have a high dimensional vectors (D~800)

Might be handled with approximate nearest neighbor

APPROXIMATE NEAREST NEIGHBOR SEARCH



https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html

REDUCING SEARCH SPACE

... we used ElasticSearch as it supports meta data filtering.

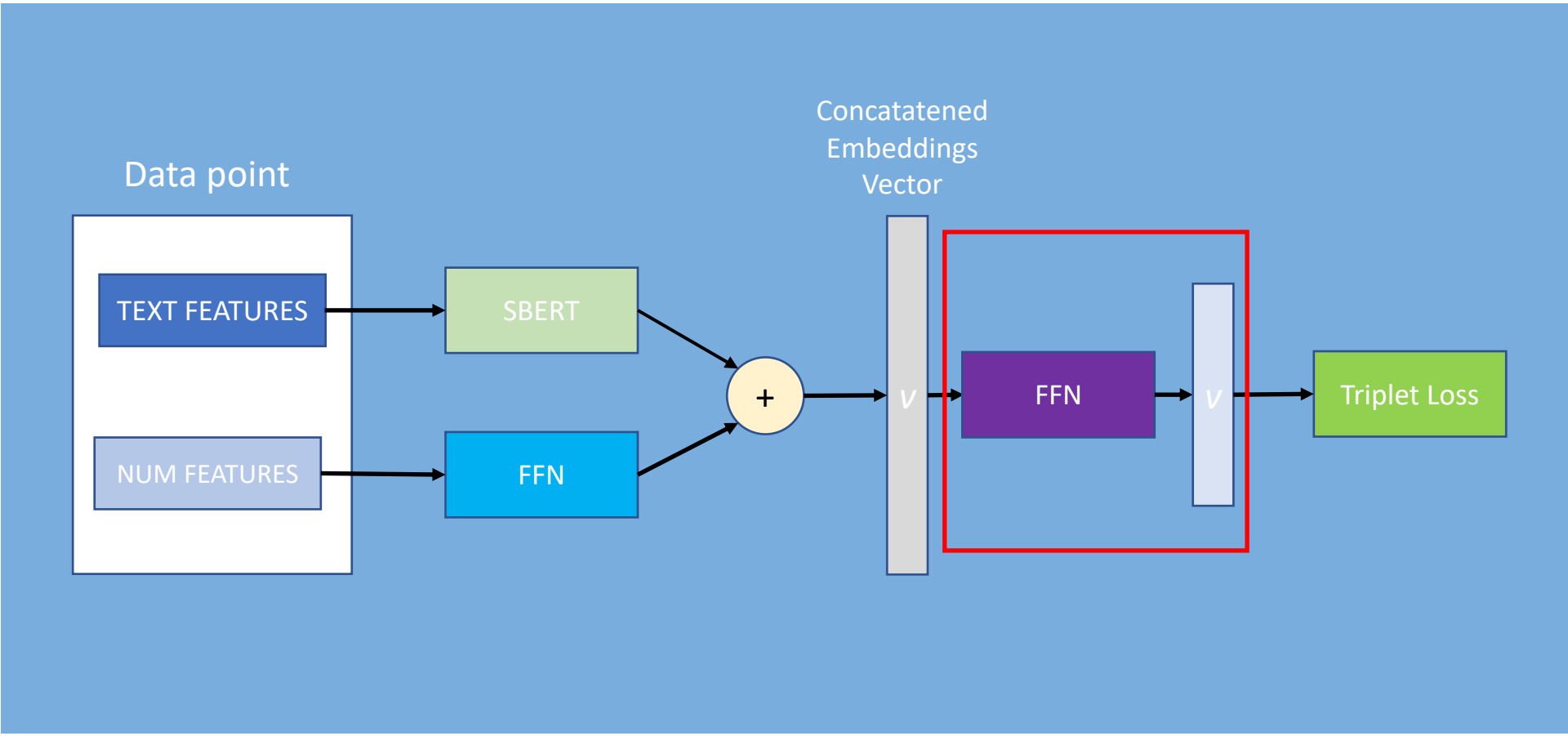
LARGE SCALE NEAREST NEIGHBOR SEARCH

COMPLEXITY OF EXHAUSTIVE SEARCH

- To idenfity duplicates we need to find nearest neighbors
- The simplest way is to perform an exhaustive search
- It means we need to perform O(N*D) operations:
  - 1) We have a big number of records (N=4.5M)
  - 2) We have a high dimensional vectors (D~800)

Might be handled with PCA or learning for smaller embeddings.

LARGE SCALE NEAREST NEIGHBOR SEARCH



RESULTS

OUR APPROACH

TEXT FIELDS: INITIAL RESULTS (RAW)

Method	TOP-1 ACC	TOP-5 ACC	TOP-10 ACC
Baseline	0.431		
CL-text-only	0.637	0.718	0.745
CL-all-features	**0.752**	**0.812**	**0.859**

NEXT STEPS

NEXT STEPS

IMPROVE DATA QUALITY

Let the model run in the production environment with a human in the loop at first. As the experts are provided with better suggestions, the duplicates labeling should improve as well. Thus, after few months, the model is to be retrained on new data.

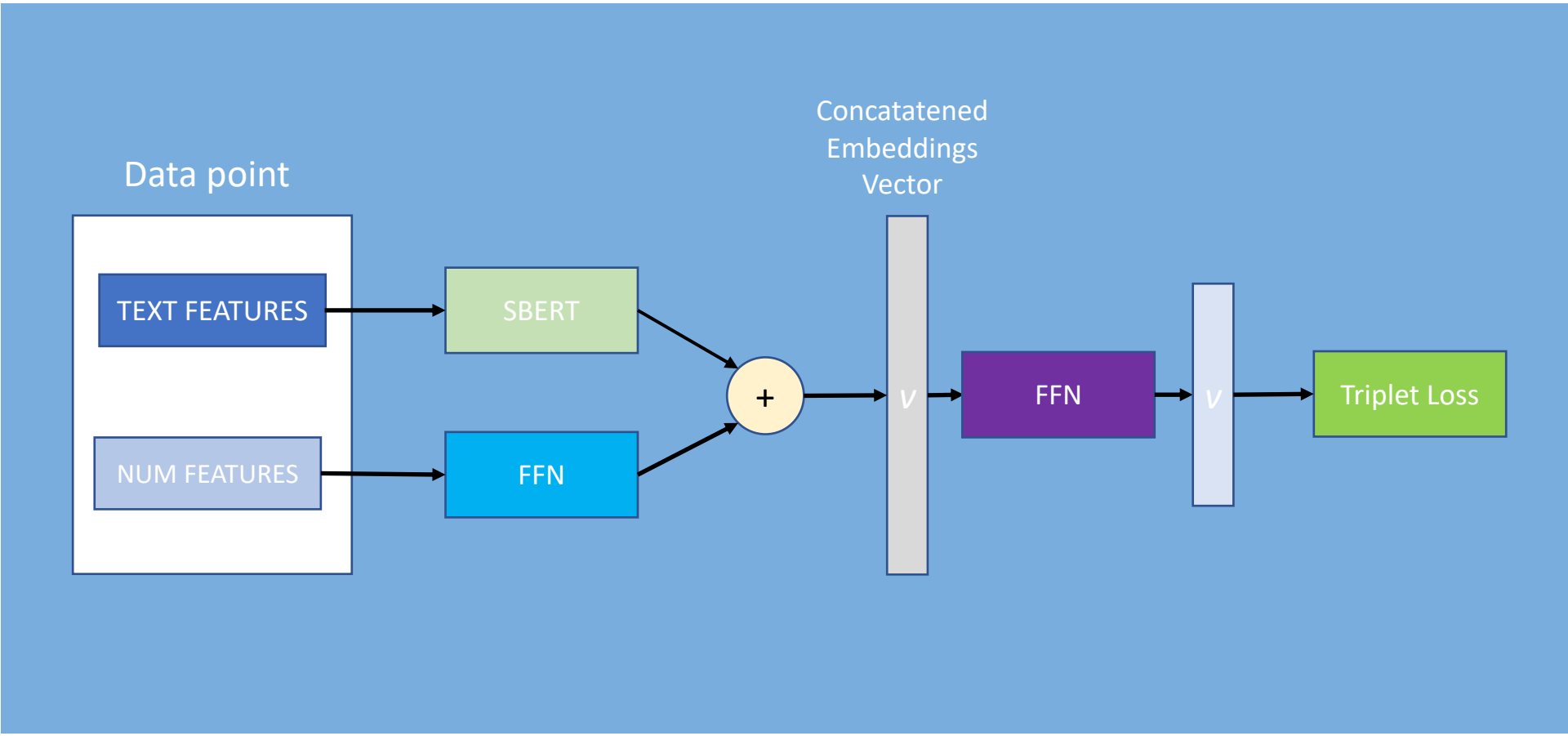
NEXT STEPS

GET MORE DISCRIMINATIVE DATA FIELDS

BID NOTICE SpawGlass Contractors Inc (constructionManager.companyName) . Construction Manager-at-Risk for the Alamo Colleges District (owner.companyName) - Northwest Vista College (owner.companyName) (NVC (owner.companyName)) Cypress Campus Center Building Capital Improvement Project (CIP) (projectTitle) will be accepting bids from qualified contractors for the project located at Northwest Vista College (projectLocation.address1) 3535 N. Ellison Drive (projectLocation.address1) San Antonio (projectLocation.city) TX (projectLocation.state) 78251 (projectLocation.zipCode) . This solicitation is for bidding this project in its entirety. The project consists of the gut and renovation/expansion of (projectDescription) 65000 (totalSquareFeet) SF (approx.) student center facility comprised of a single (projectDescription) 2-story (numberOfStoriesAboveGrade) building with North and South Wings surrounding a central courtyard on a pond. The facility will provide classrooms laboratories student lounge and server/cafeteria and associated site work (projectDescription) all located on the Northwest Vista College (owner.companyName) campus in San Antonio (projectLocation.city) Texas (projectLocation.state) . Bids will be due Friday December 4 2020 (bid.Date) @ 2:00 PM (bid.Time) CST via email or fax. Email to saestimating@spawglass.com (constructionManager.contact.email) or fax (210) 651-4300. Plans and specifications are available electronically via www.isqft.com or Virtual Builders Exchange plan room. Please contact Brad Fielden (constructionManager.contact.fullName) for access at saestimating@spawglass.com (constructionManager.contact.email) or (210) 651-9000 (constructionManager.contact.phoneNumber) . MBE/SBE/WBE/AABE/HUB firms are highly encouraged to submit proposals on this project. SpawGlass Contractors Inc (constructionManager.companyName) . is an equal opportunity (EEO) employer. There will be a non-mandatory but highly encouraged Pre- Proposal Conference / Diversity Outreach at the Holiday Inn-Seaworld 10135 TX-151 (projectLocation.address1) San Antonio (projectLocation.city) TX (projectLocation.state) 78251 (projectLocation.zipCode) on Thursday November 19 2020 (preBid.Date) @ 10:00 AM (preBid.Time) . A follow- up site visit will not be occurring due to current COVID restrictions on campus.

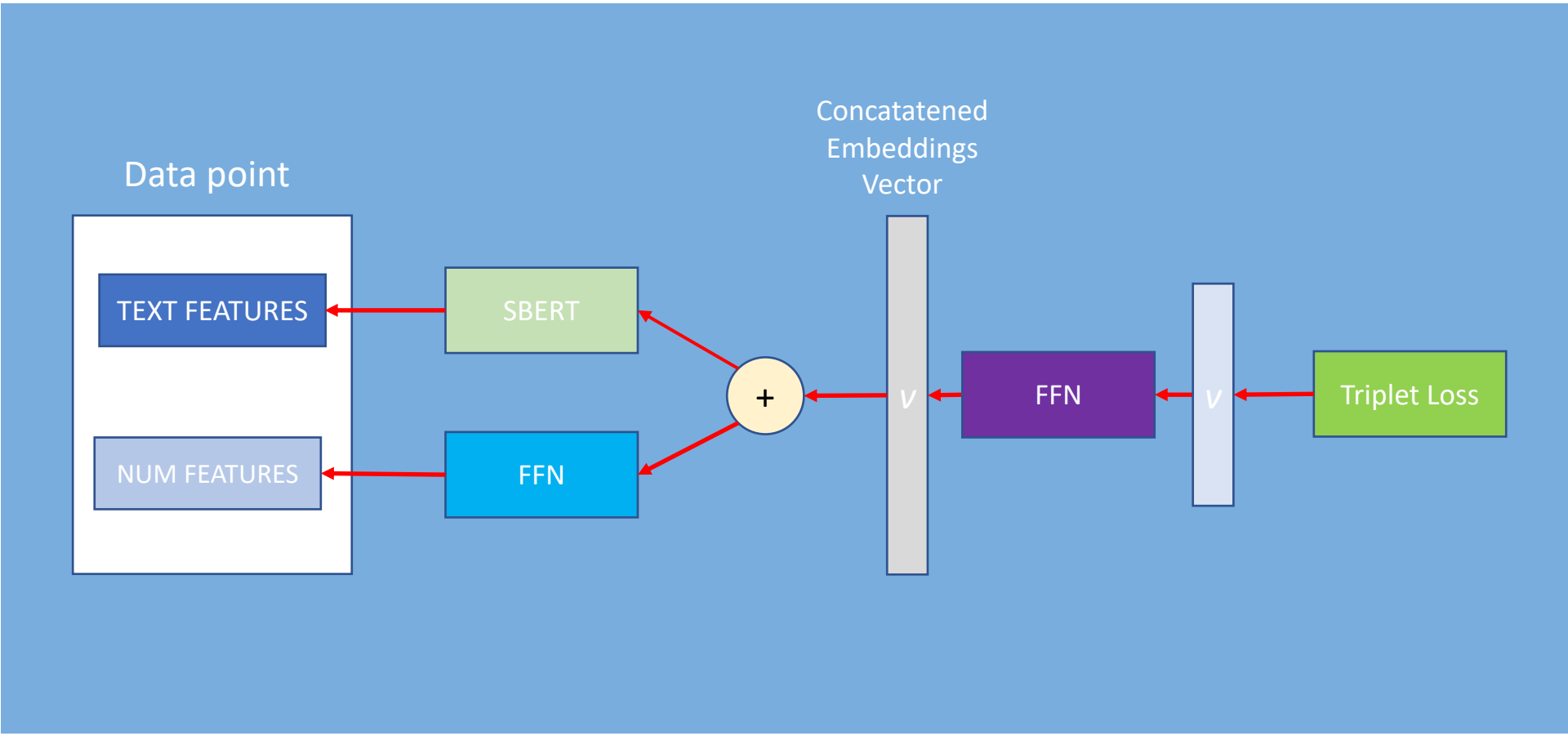
NEXT STEPS

REDUCE DIMENSIONALITY



NEXT STEPS

TRAIN END TO END



SUMMARY

KEY TAKEAWAYS

- Contrastive learning is a representation learning method that learns such and embedding space where similar records are located near each other and far away from dissimilar ones
- It might be applicable in supervised and unsupervised settings
- To perform efficient search among embeddings you can rely on approximate methods and reduce the dimensionality of your vectors
- Applying ANN is not always necessary
- Selecting the right ANN method depends on your task and should be performed wisely


Thank you!