SwRI
SOUTHWEST RESEARCH INSTITUTE

SEARCH

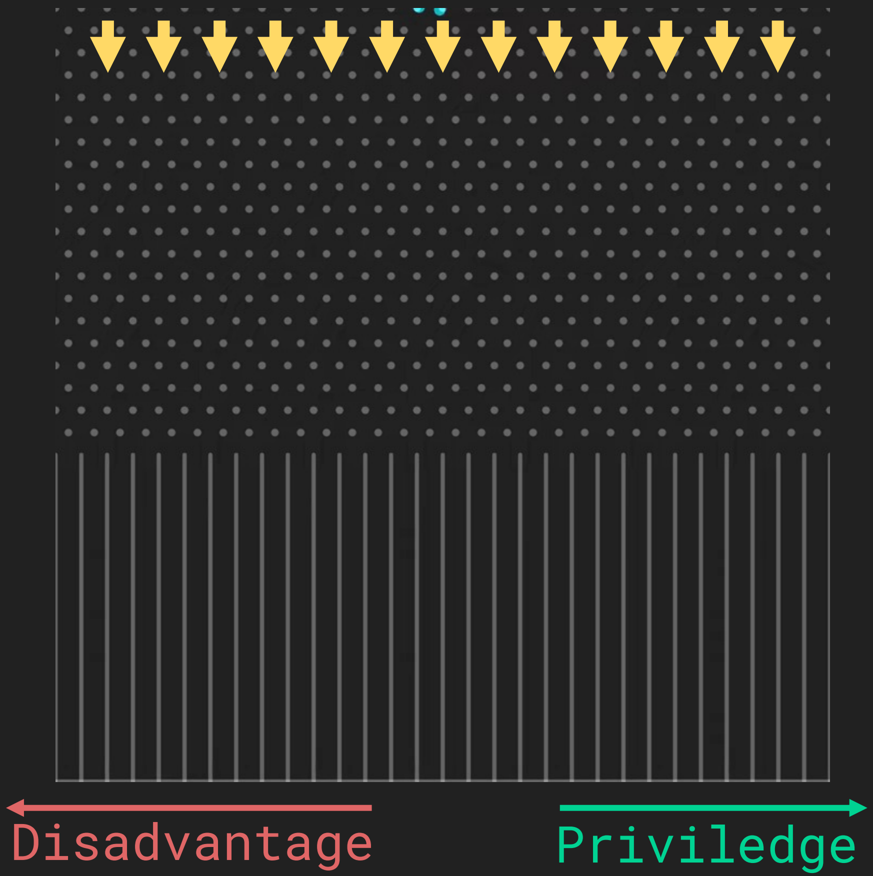
SDO Exploration And Research Community for Heliophysics

An experiment on using applied AI research to educate a diverse workforce

Subhamoy Chatterjee, Nadia Ahmed, & Andrés Muñoz-Jaramillo
andres.munoz@swri.org

Funded by NASA-HITS

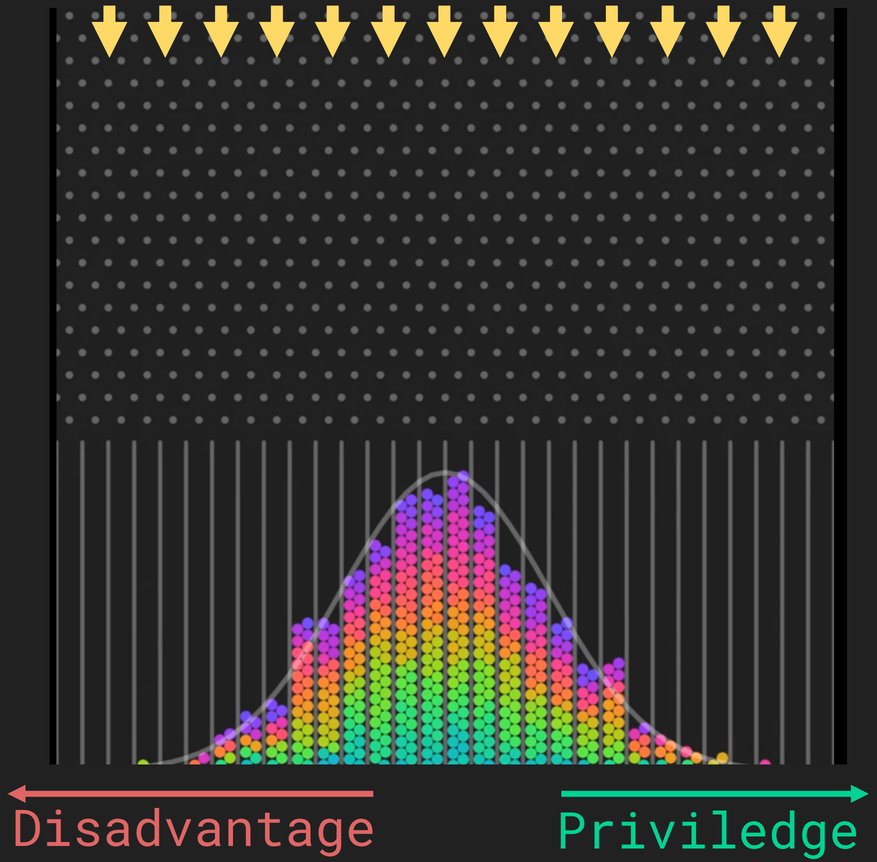
Why SEARCH?



Our lives are very random. They are **NOT** the deterministic outcome of the choices we make.

Our birth has an outsize influence on the rest of our lives. We do **NOT** start with the same circumstances and opportunities.

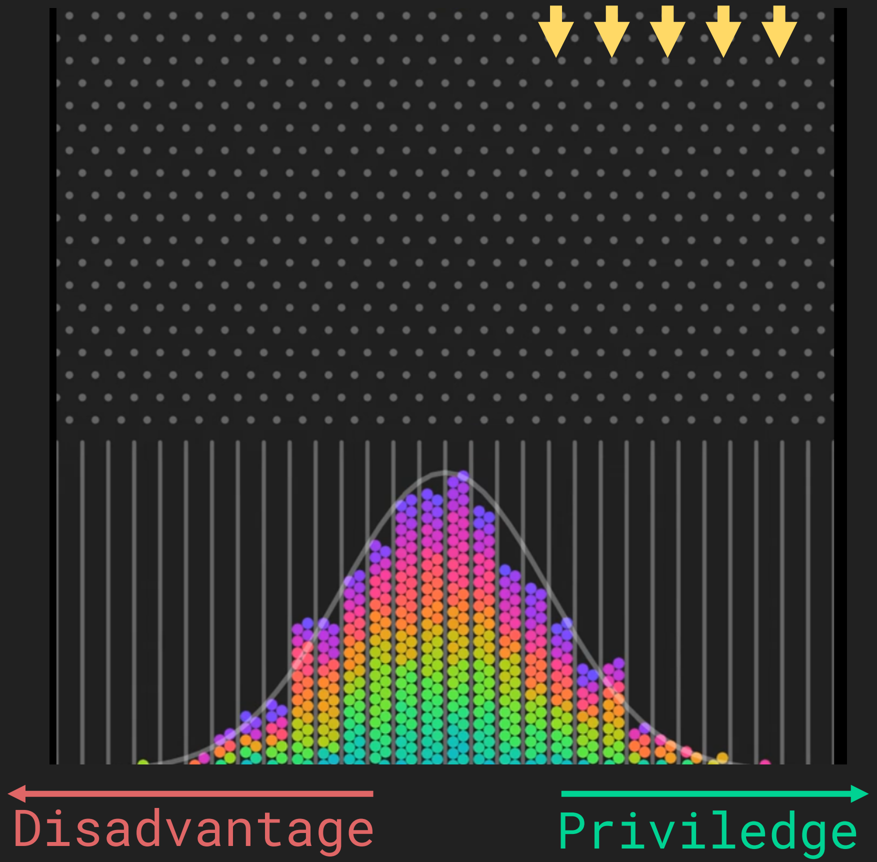
Why SEARCH?



Education is one of the main ways we have to help disadvantaged populations.

Most educational oportunities are gated, which tends to help those that are ahead get farther ahead.

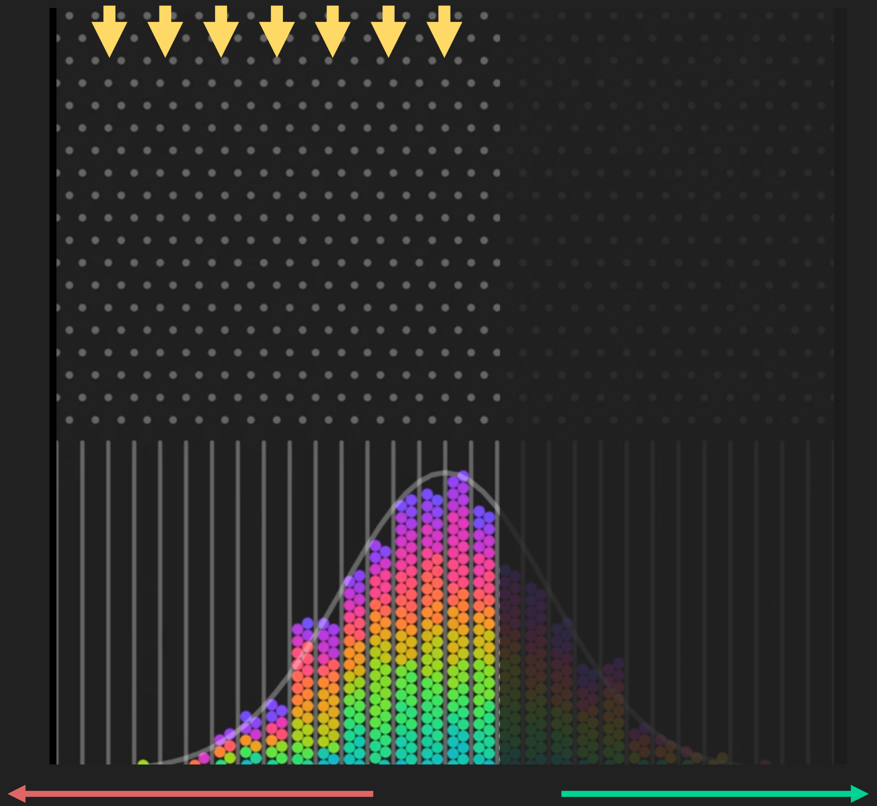
Why SEARCH?



Education is one of the main ways we have to help disadvantaged populations.

Most educational oportunities are gated, which tends to help those that are ahead get farther ahead.

Why SEARCH?

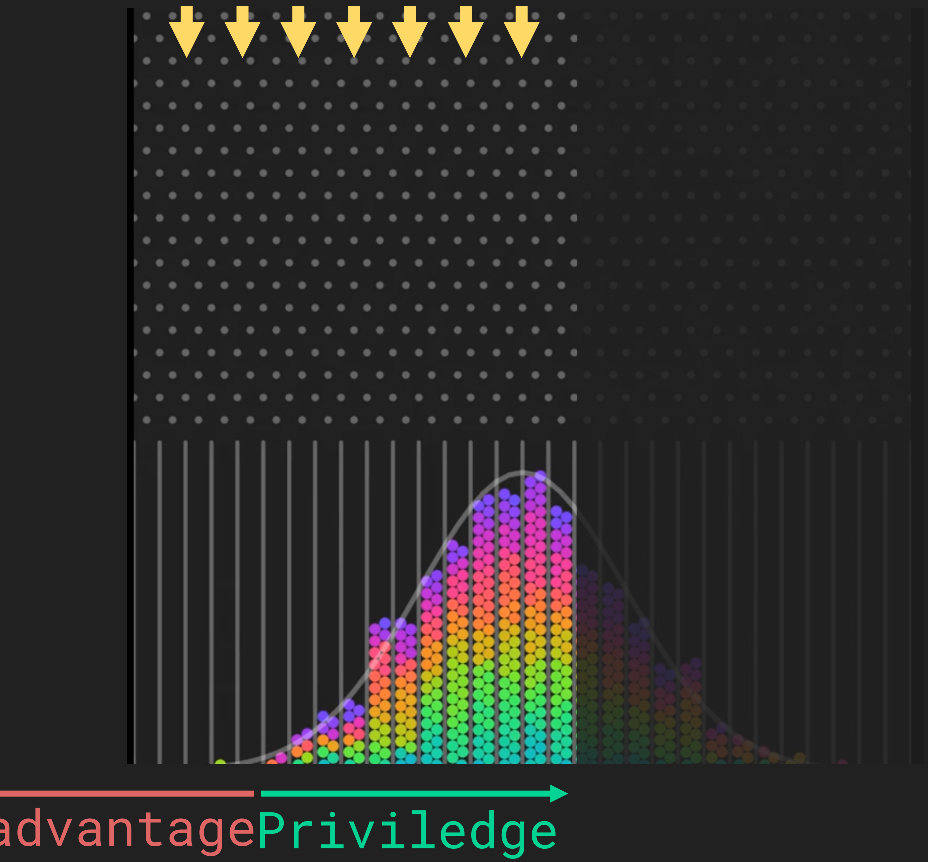


Education is one of the main ways we have to help disadvantaged populations.

Most educational oportunities are gated, which tends to help those that are ahead get farther ahead.

So we restrict them, and it helps, up to a point.

Why SEARCH?



Education is one of the main ways we have to help disadvantaged populations.

Most educational oportunities are gated, which tends to help those that are ahead get farther ahead.

So we restrict them, and it helps, up to a point.

Because priviledge is relative.

Sources of priviledge we wanted to offset

- Socio-economic status, wealth, and place of origin:
    - Limited access to educational opportunities.
    - Significant difference in skillsets.
    - Limited access to computational resources.

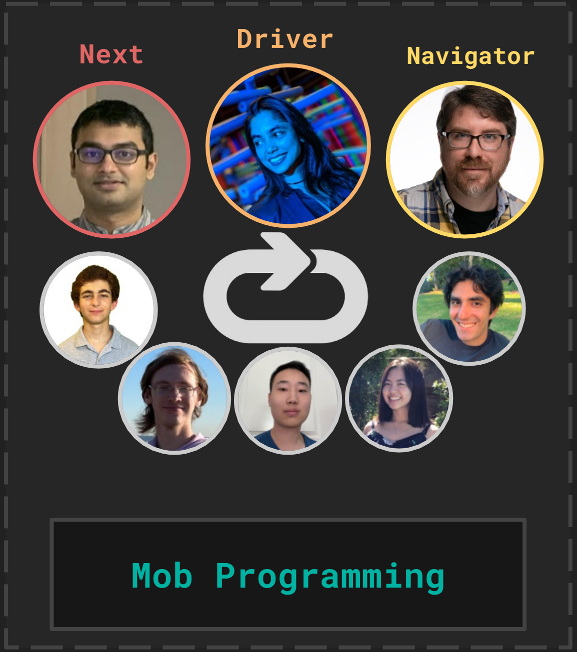
- Gender, race, ethnicity, religion, and sexual orientation:
    - Subconcious and concious bias.
    - Harder access to opportunities.

- Age, employment, family responsibility, veteran status:
    - Scientific education is focused on students.
    - Limited time availability to learn.
    - Reduced access to new techniques and technologies.

Program design

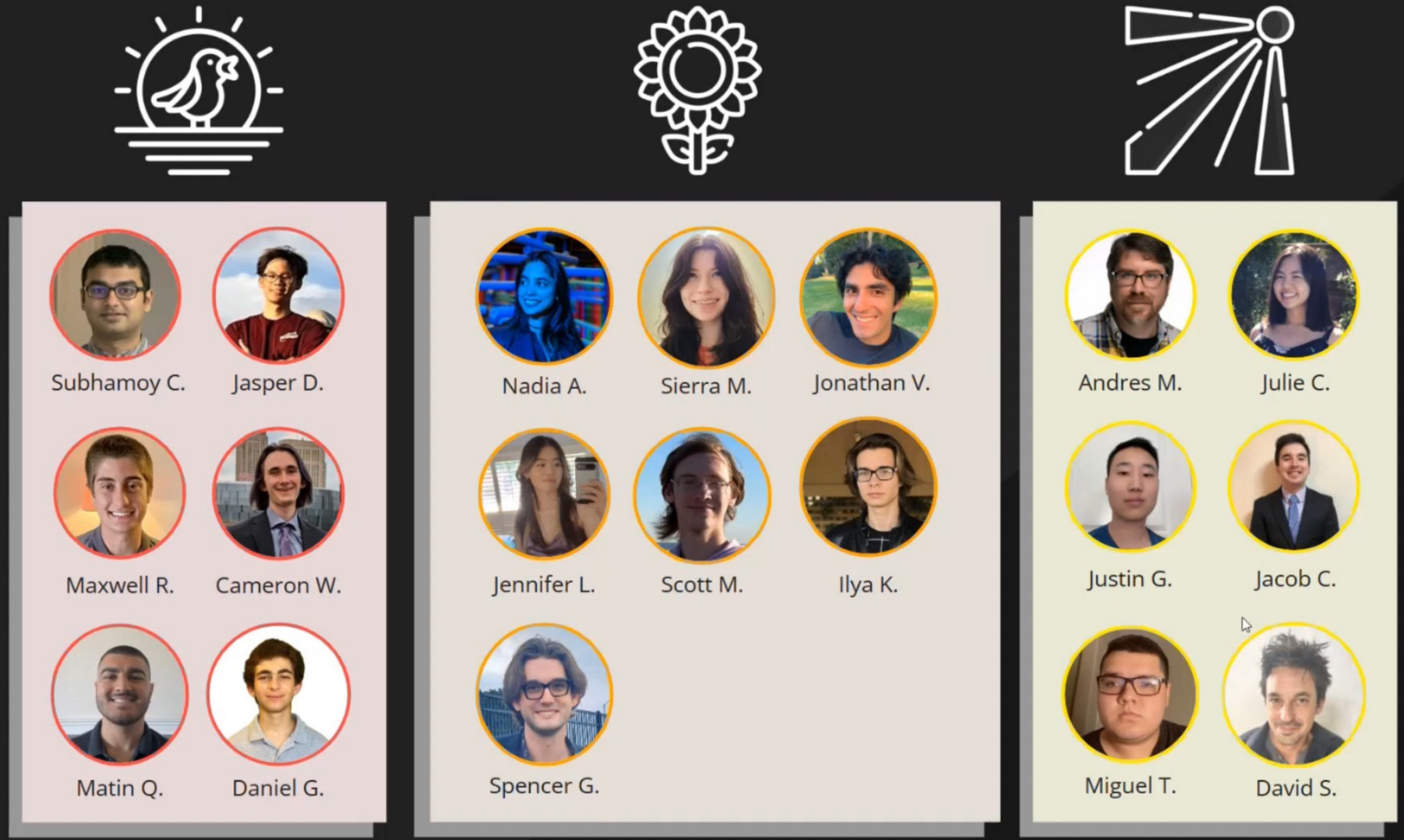
Significant difference in skillsets. Reduced access to new techniques and technologies.	Collaborative programing (mob-coding)

- Equal sharing of development responsibility.
- Short, low-stakes rotations.
- Simultaneous development and documentation of code.
- Mob efficienly finds solutions to obstacles.
- Effective learning environment for equalizing skillsets.



Program design

Significant difference in skillsets. Reduced access to new techniques and technologies.	Collaborative programing (mob-coding)
Limited time availability to learn	4 hours/week commitment, no assignments
Limited access to computational resources.	Development uses Google colab and visual studio code liveshare sessions
Limited access to educational opportunities. Harder access to opportunities. Scientific education is focused on students.	Anyone can join as long as they show up and actively engage in collaborative programming.


Subhamoy C.
Jasper D.
Maxwell R.
Cameron W.
Matin Q.
Daniel G.
Nadia A.
Sierra M.
Jonathan V.
Jennifer L.
Scott M.
Ilya K.
Spencer G.
Andres M.
Julie C.
Justin G.
Jacob C.
Miguel T.
David S.

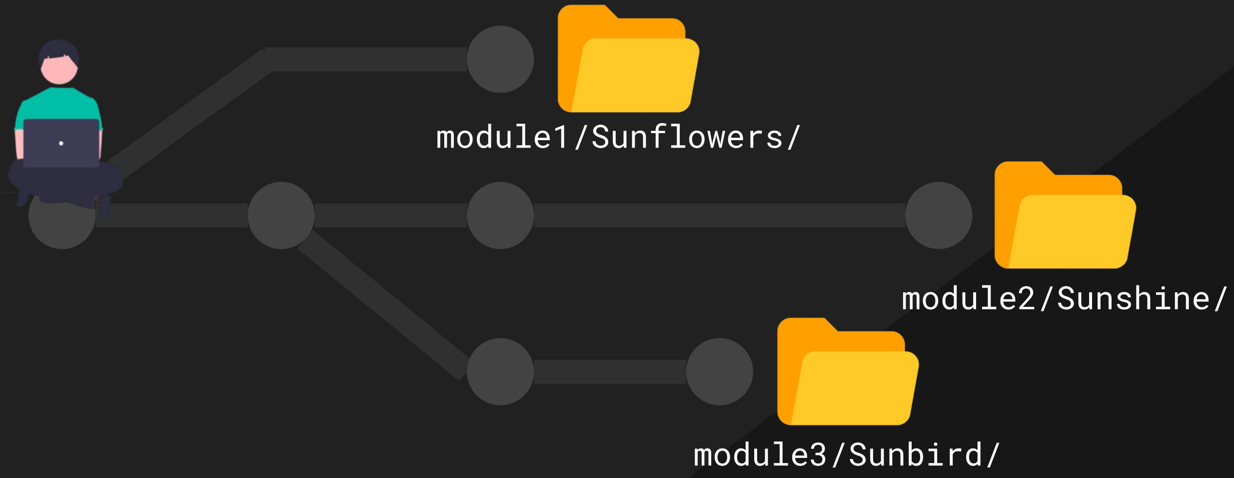
Opening: SEARCH objectives

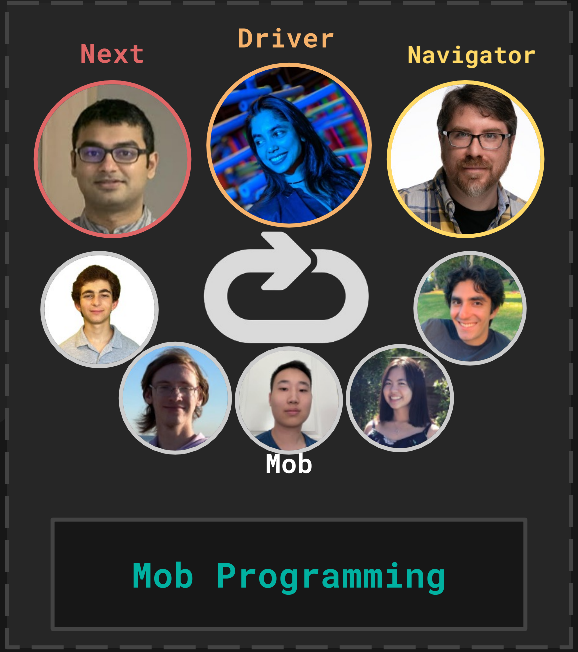
- ❑ To perform cutting edge Machine Learning (ML) research in heliophysics involving a diverse group of people without requiring *any prior ML technical expertise*.
- ❑ To create a space where all are welcome and where all participants level up and work together towards common technical goals.

Opening: Team Organization and Workflow

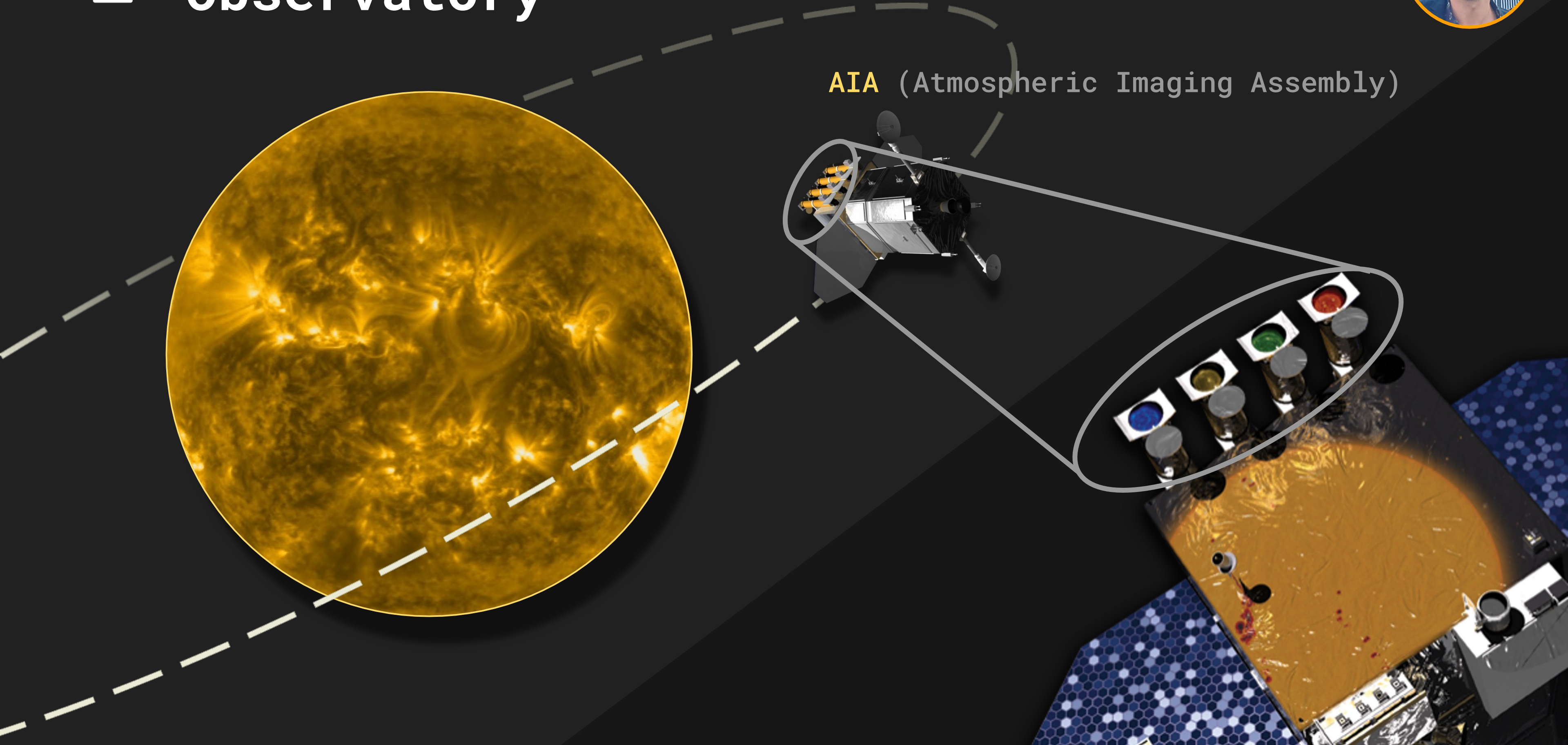
- VSCode Liveshare
- Google Colab
- Github

Mob programming -
Team member rotation -
Weekly Code review -
Agile workflow -



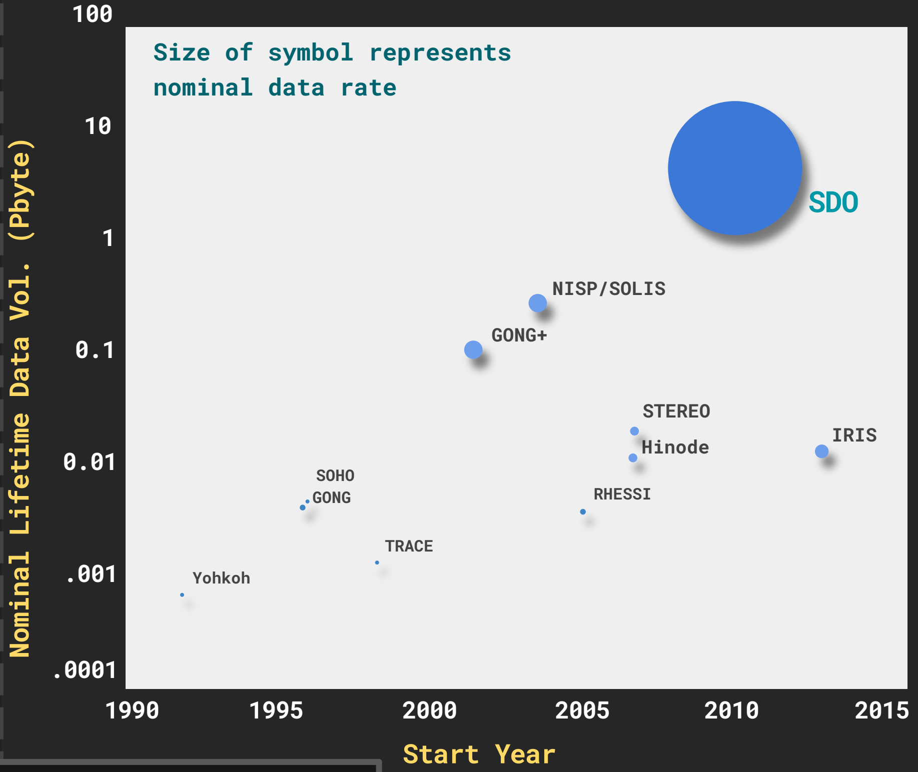


Introduction: Solar Dynamics Observatory



Introduction: AIA Data Volume & Motivation

- Petabytes of data
- Data must be manually searched through



Solar Data Volume Growth

2012, J.B Gurman

Program Phases

EDUCATION

DATA

ML

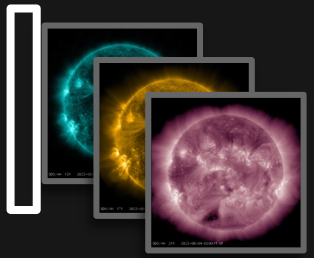
USER INTERFACE

Components: Overall Pipeline

SDO Downloader

```
sdate: (str)
edate: (str)
instrument: (str)
wavelength: (list)
cadence: (str)
grayscale: (bool)
multiwavelength: (bool)
```

Stanford Joint Science Operations Center (JSOC)



Solar Images

Dataset (AIA)

...

...

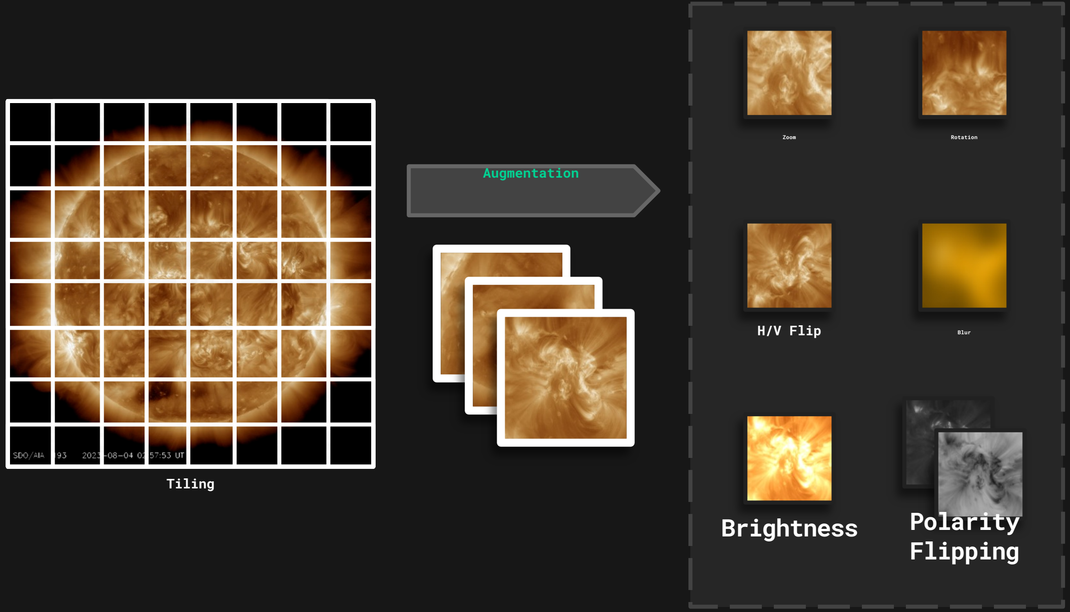
Solar Dynamics Observatory Downloader

AI/ML Ready Packager

Similarity Search Engine

Components: Overall Pipeline

. . .



. . .

SDO Downloader

AI/ML Ready Packager

Similarity Search Engine

Components: Overall Pipeline

...

...

SDO Downloader

AI/ML Ready Packager

Similarity Search Engine

Introduction to Self-Supervised Learning:
General Overview (Model-Agnostic)

Trained by comparing:

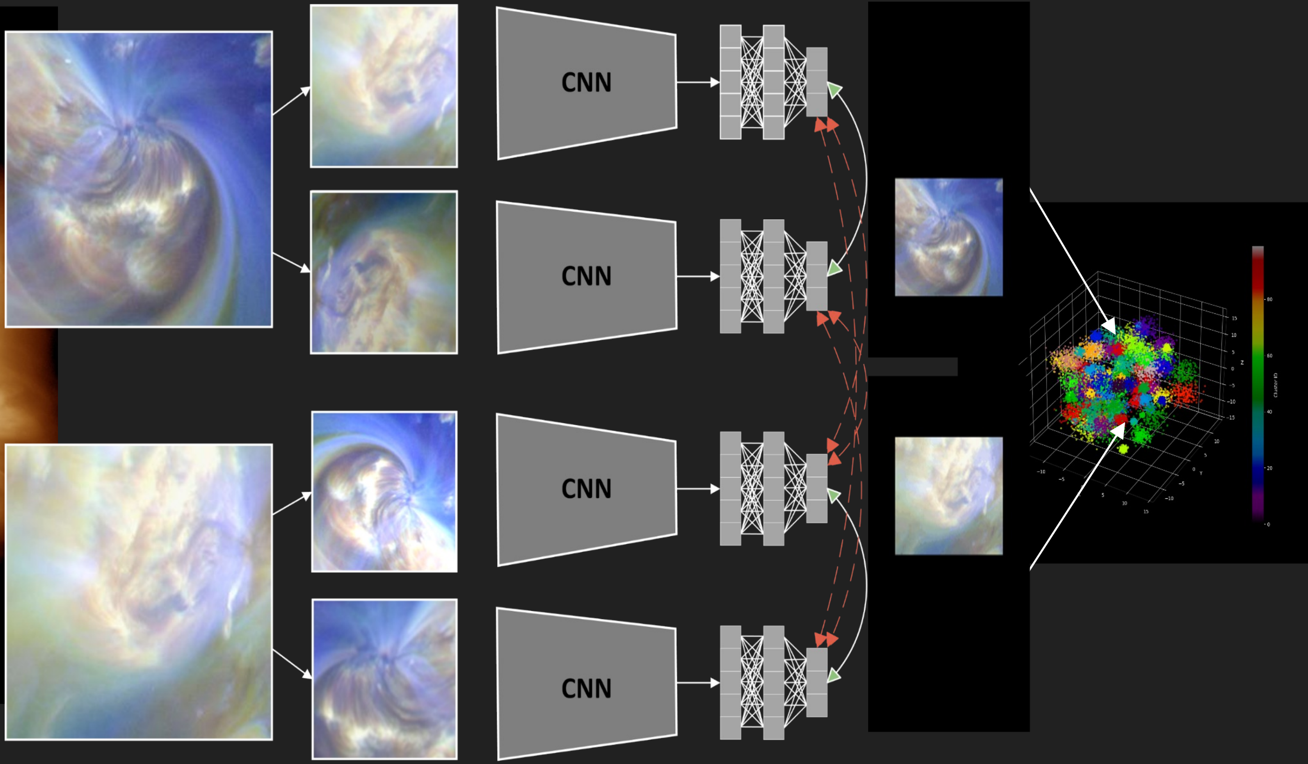
Image vs. transformed image (and/or a contrasting image)

Allows us to discern similar & dissimilar images.

Returns:

Similar embeddings for an image & its transformed counterpart.

For dissimilar images the embeddings should be very different.

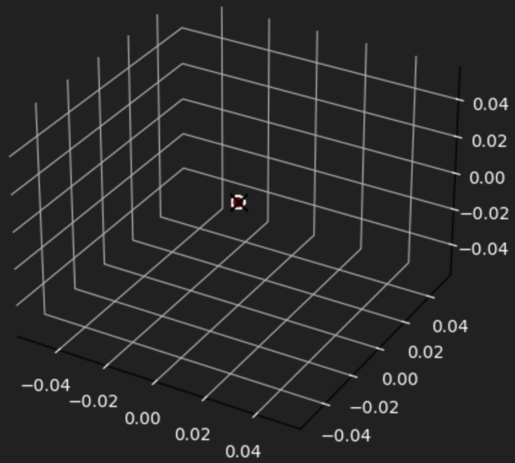


Introduction to Self-Supervised Learning:
General Overview (Model-Agnostic)

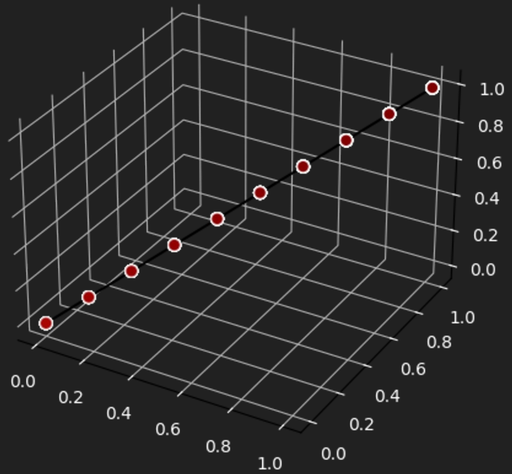
Collapse:

Solution collapse can occur when a model fails to learn meaningful feature representations from the data and converges on a single embedding or limited embedding space.

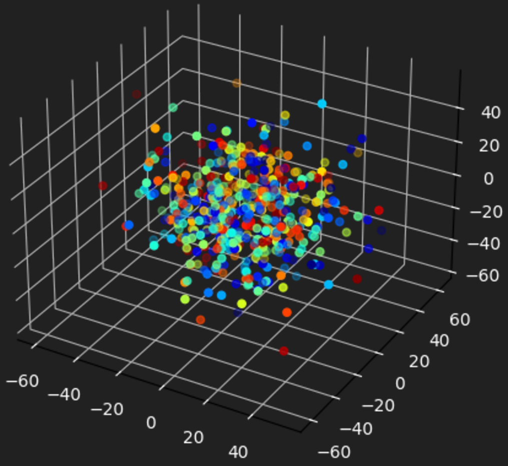
Complete Collapse



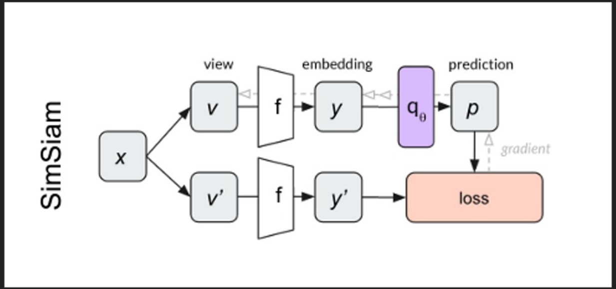
Dimensional Collapse



No Collapse

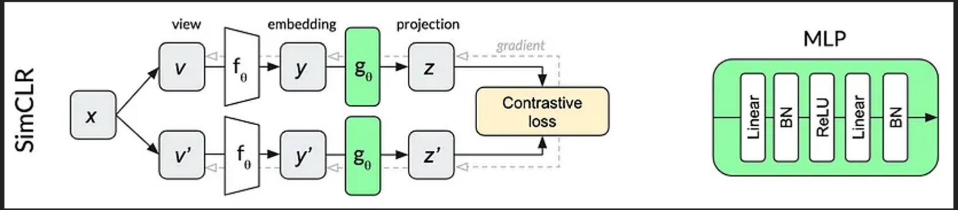


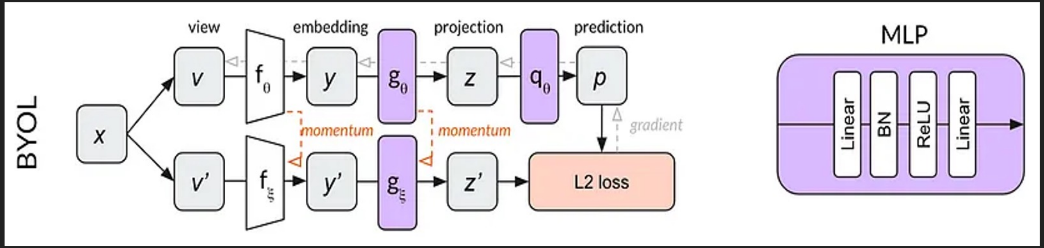
Introduction to Self-Supervised Learning:
General Overview (Model-Agnostic)



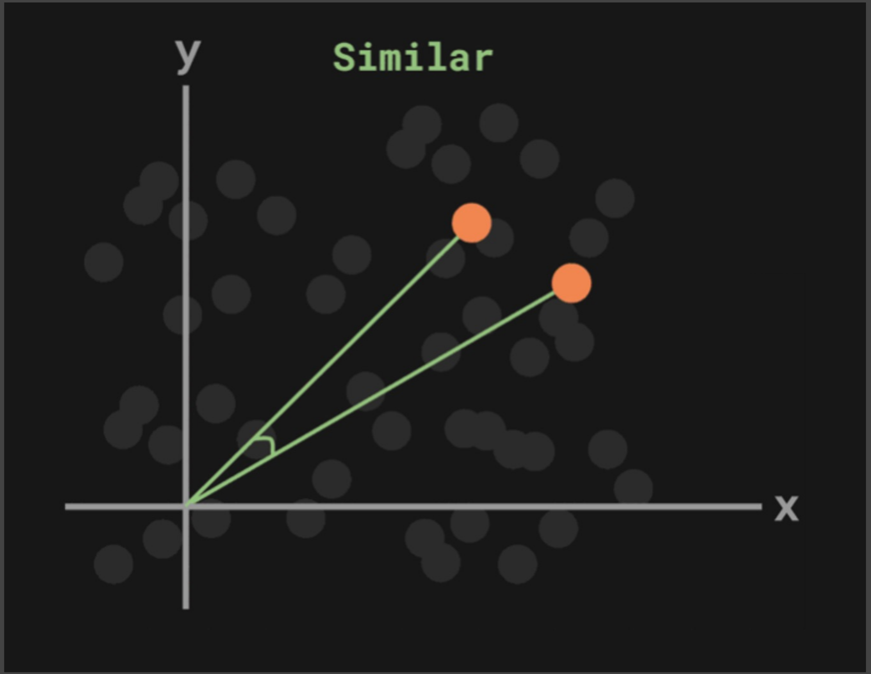
ML algorithms we used:

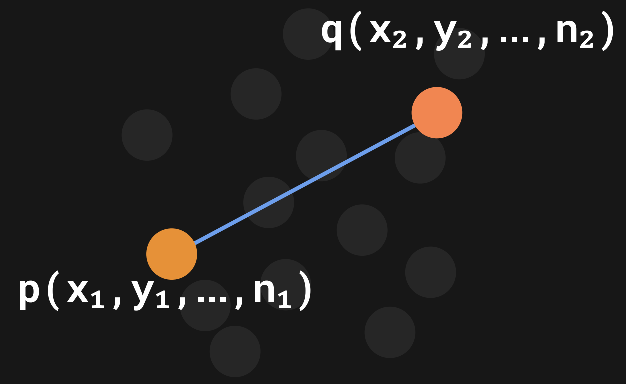
SimSiam; SimCLR; BYOL ("Bootstrap Your Own Latent").





Introduction to Self Supervised Learning:
Metrics and Losses





$$d(p,q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

Cosine Similarity

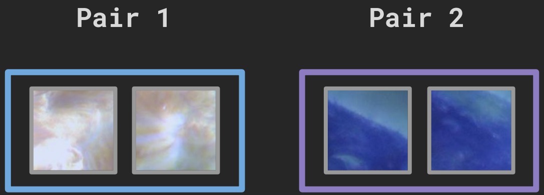
Euclidean Distance

Introduction to Self Supervised Learning:
Metrics and Losses

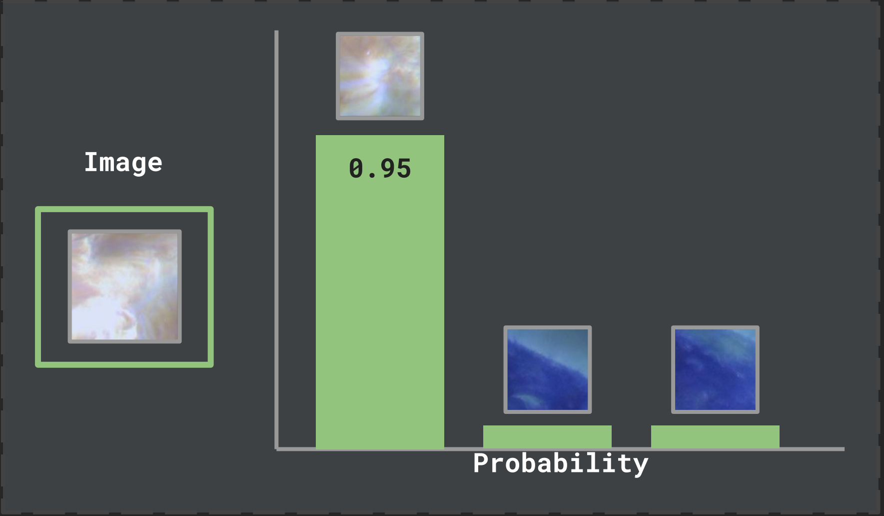
Contrastive Loss

Measures how well the model can contrast between similar and dissimilar data points

1) Augmented Image in Batch



2) Softmax =



Chen, Kornblith, Norouzi, Hinton, A Simple Framework for Contrastive Learning of Visual Representations (2020)

Introduction to Self Supervised Learning:
Metrics and Losses

3) Noise Contrastive Estimation (NCE)

l(,)) = -log(Softmax(,))

l(,)) = -log(Softmax(,))

$$l(i,j) = -\log\left(\frac{\exp(s_{i,j})}{\sum_{k=1}^{2N} l_{[k!=i]}\exp(s_{i,j})}\right)$$

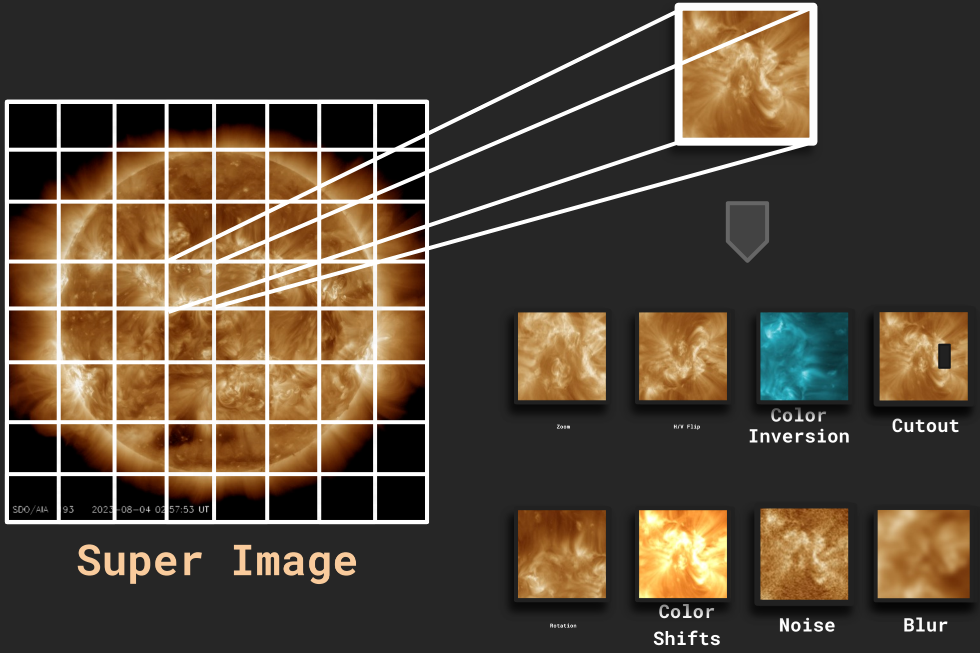
4) Compute Loss

L = [l(,)) + l(,))] + [l(,)) + l(,))] / 2*N

$$L = \frac{1}{2N}\sum_{k=1}^{N}[l(2k-1, 2k) + l(2k, 2k-1)]$$

Chen, Kornblith, Norouzi, Hinton, A Simple Framework for Contrastive Learning of Visual Representations (2020)

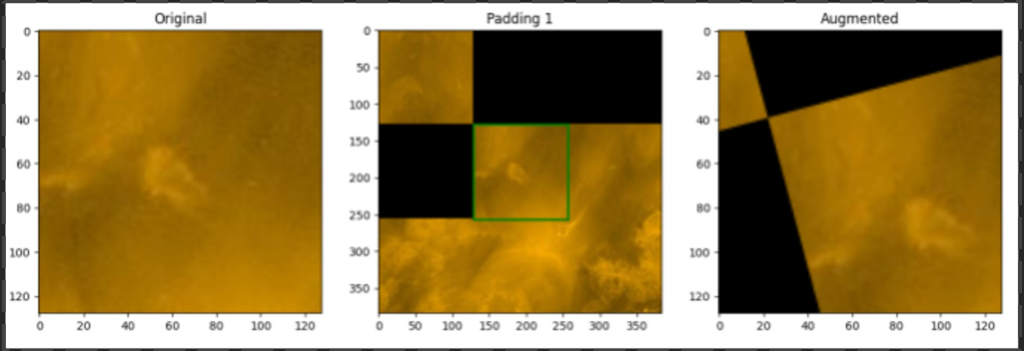
Introduction to Self Supervised Learning:
Augmentations



Augmenting Data

- ✓ Introduce Variations
- ✓ Robustness to Noise
- ✓ Generalizability
- ✓ Prevents Data Scarcity
- ✓ Reduces Overfitting

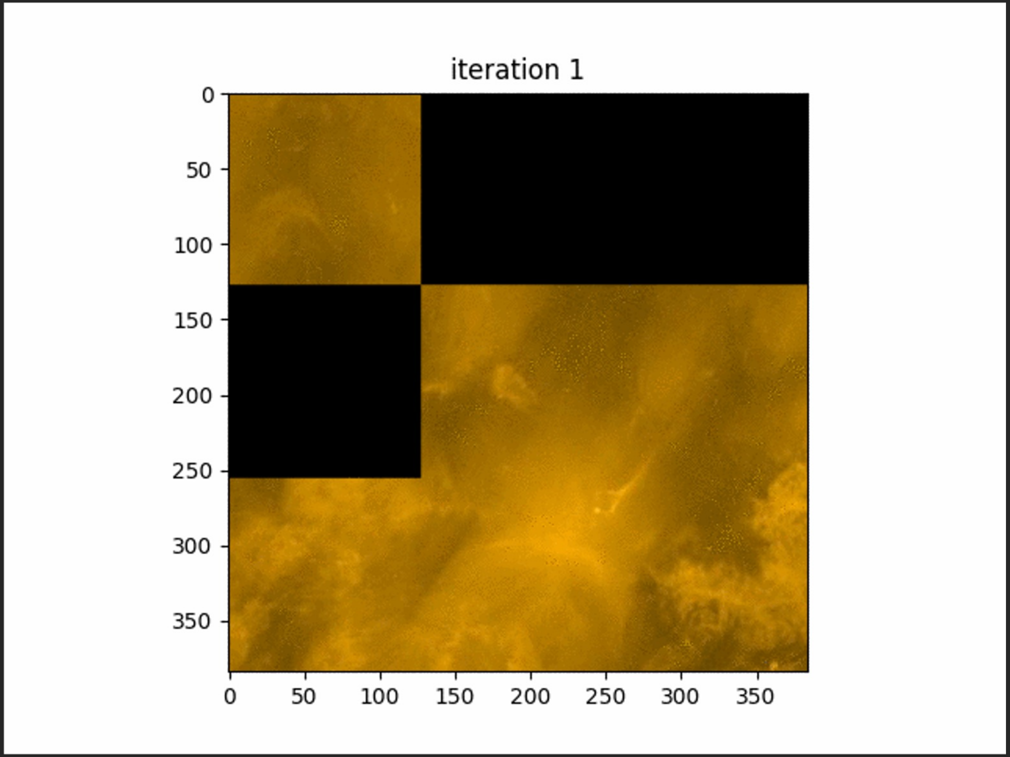
Introduction to Self Supervised Learning:
Augmentations



Augmenting Data

- ✓ Augmentations, and train/val split creates gaps
- ✓ Nearest neighbor VS Iterative filling
- ✓ Iterative filling creates improvement in terms of collapse

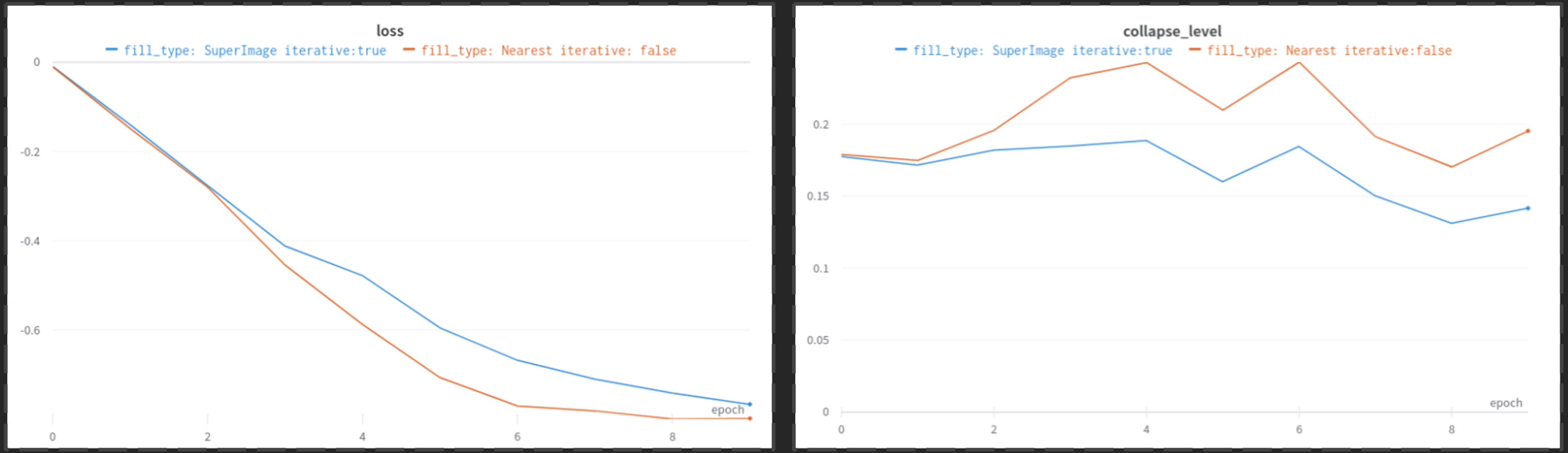
Introduction to Self Supervised Learning:
Augmentations



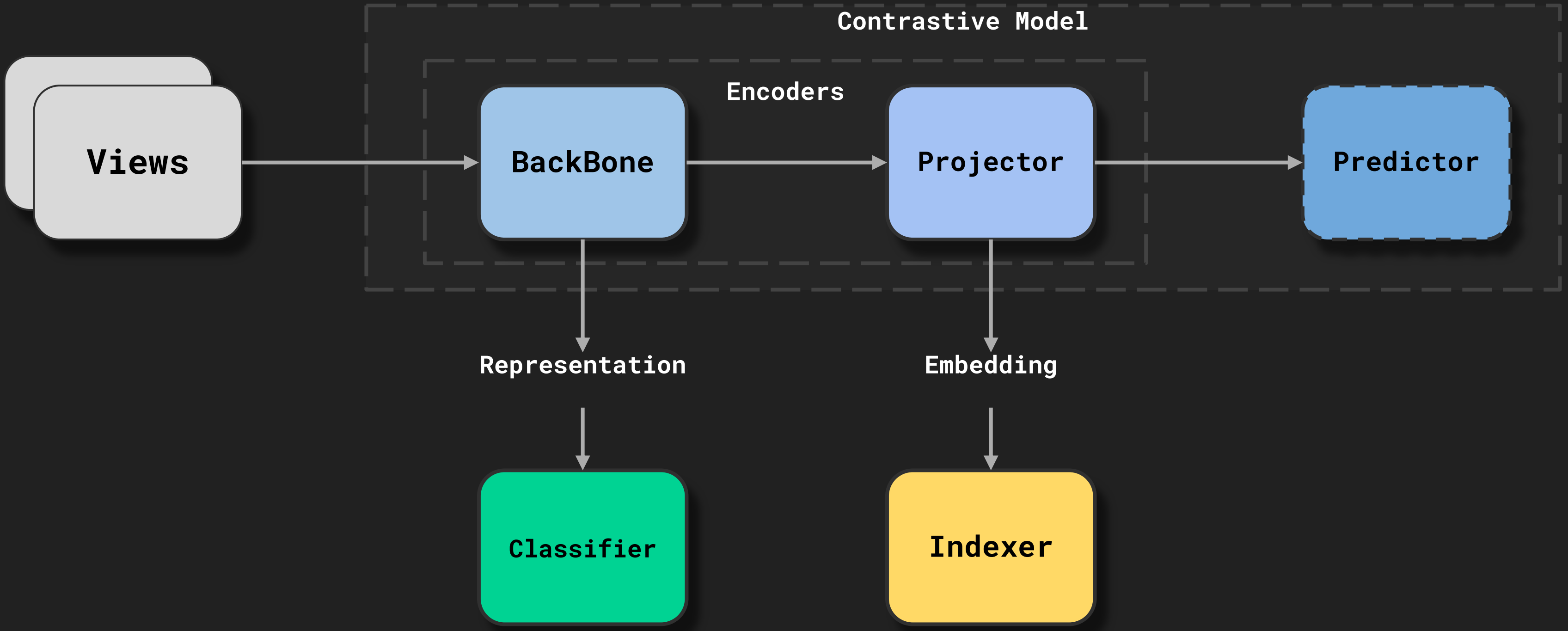
Augmenting Data

- ✓ Augmentations, and train/val split creates gaps
- ✓ Nearest neighbor VS Iterative filling
- ✓ Iterative filling creates improvement in terms of collapse

Introduction to Self Supervised Learning:
Augmentations



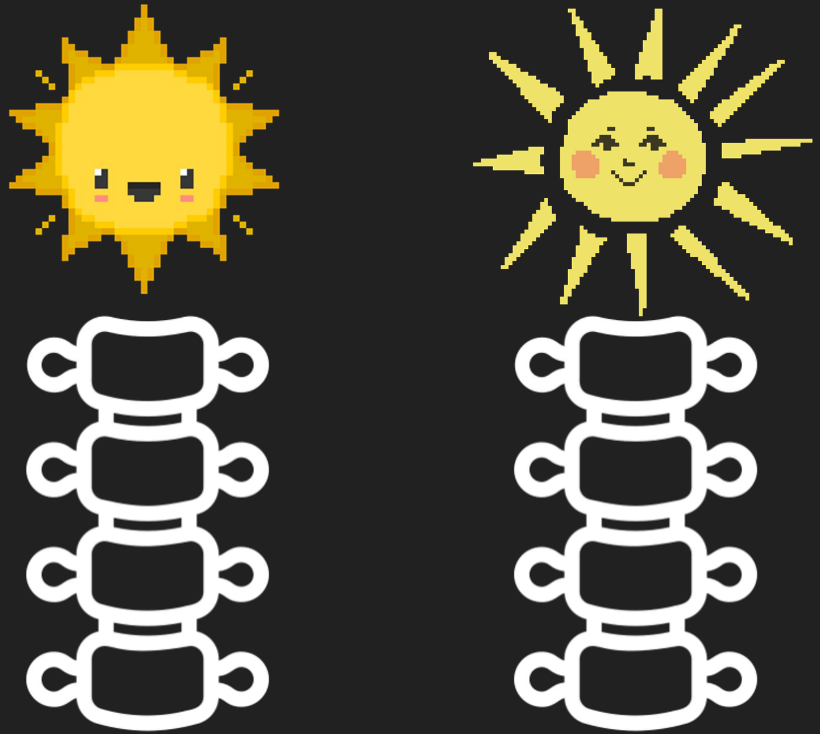
Introduction to Self Supervised Learning:
Embeddings from backbone vs. projection head



Introduction to Self Supervised Learning:
Embeddings from backbone vs. projection head

SEARCH
Projection
Head

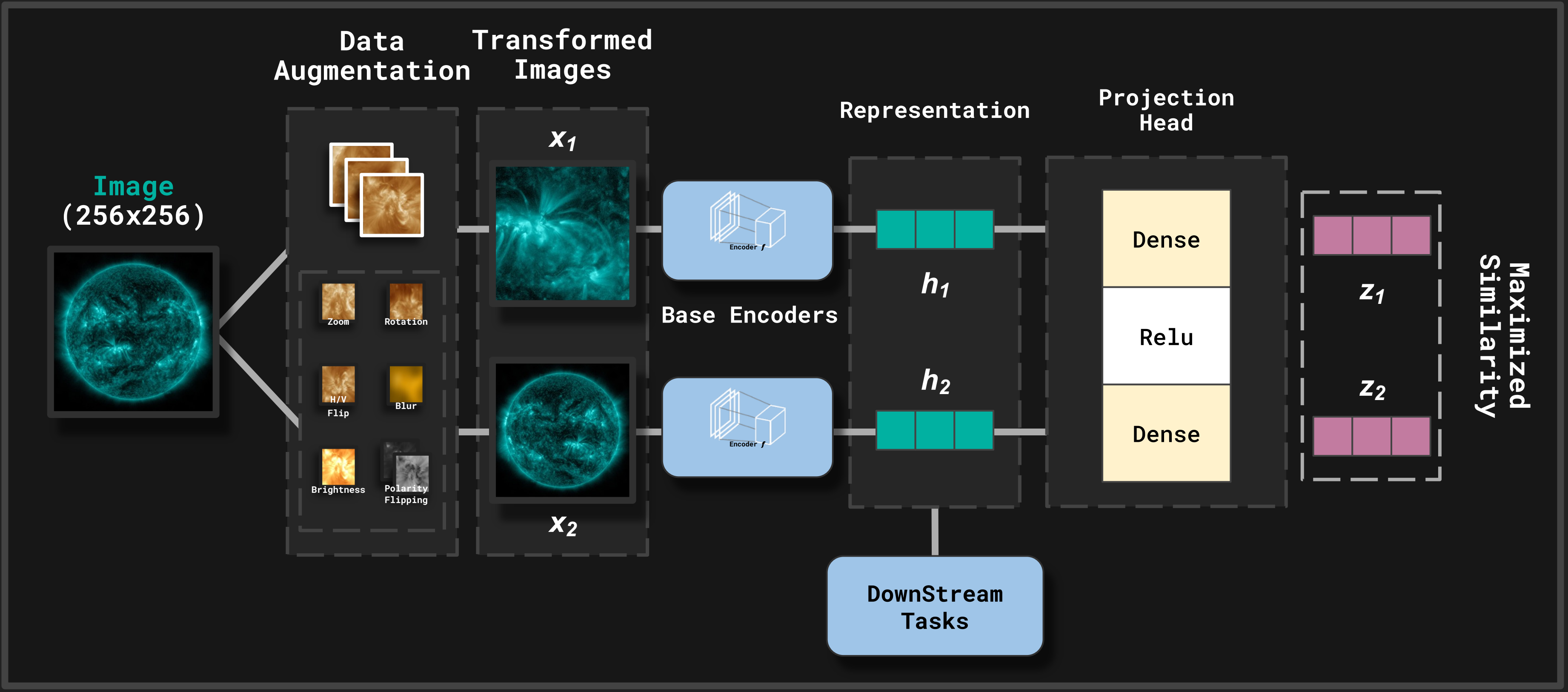
SEARCH
Backbone



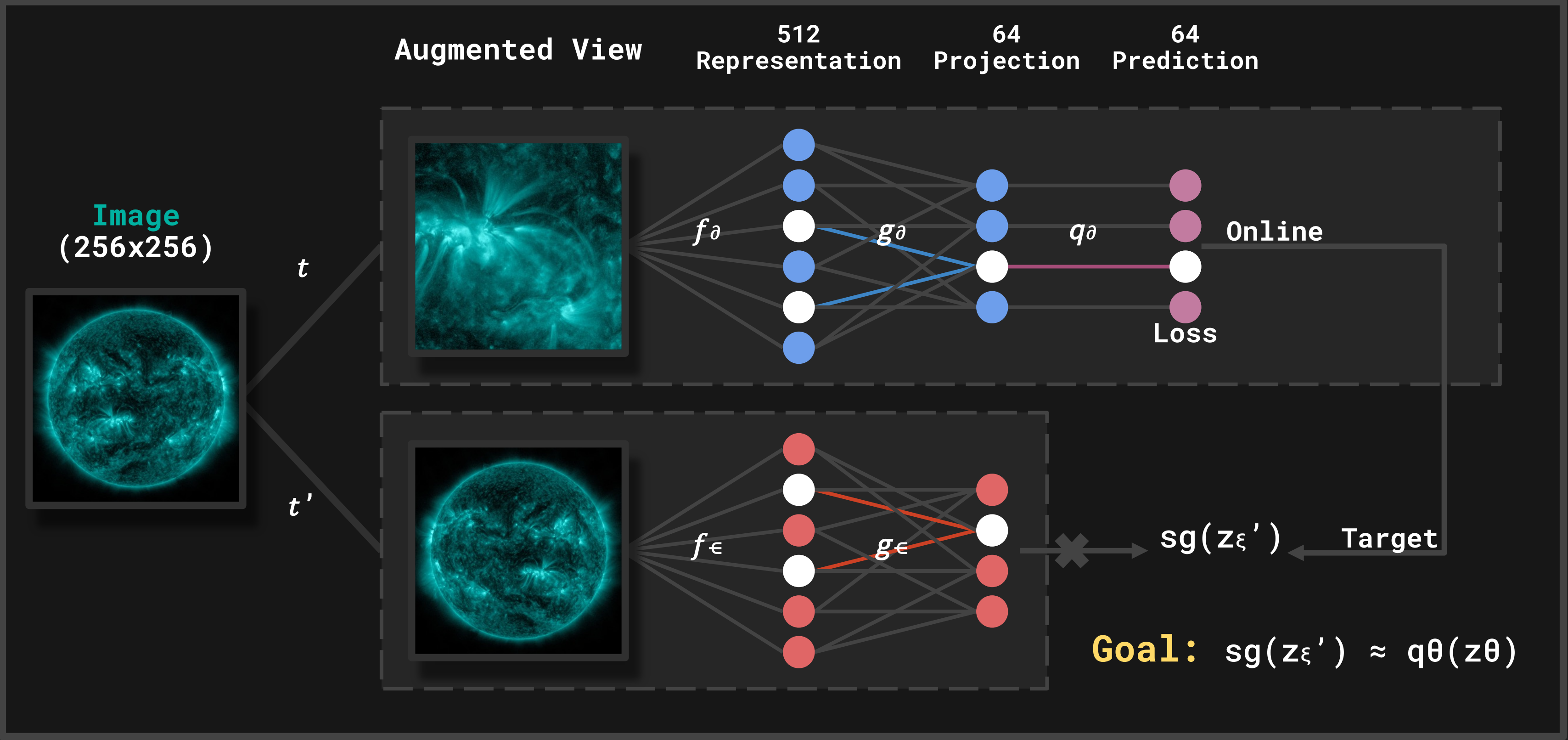
NEW
Projection
Head

SEARCH
Backbone

Introduction to Self Supervised Learning:
SimCLR

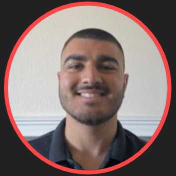


Introduction to Self Supervised Learning:
BYOL: Bootstrap Your Own Latent



Goal: $sg(z_\xi') \approx q_\theta(z_\theta)$

Introduction to Self Supervised Learning:
SimSiam



SimSiam

✓ Identical encoders

✓ Stop-gradient operation is sufficient to prevent collapse

✓ Back-propagation through only one branch

BYOL

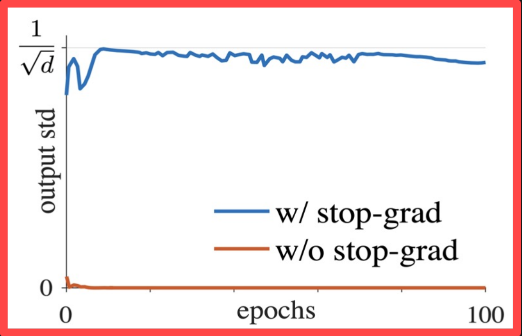
✗ Target branch is created as a moving average of online branch

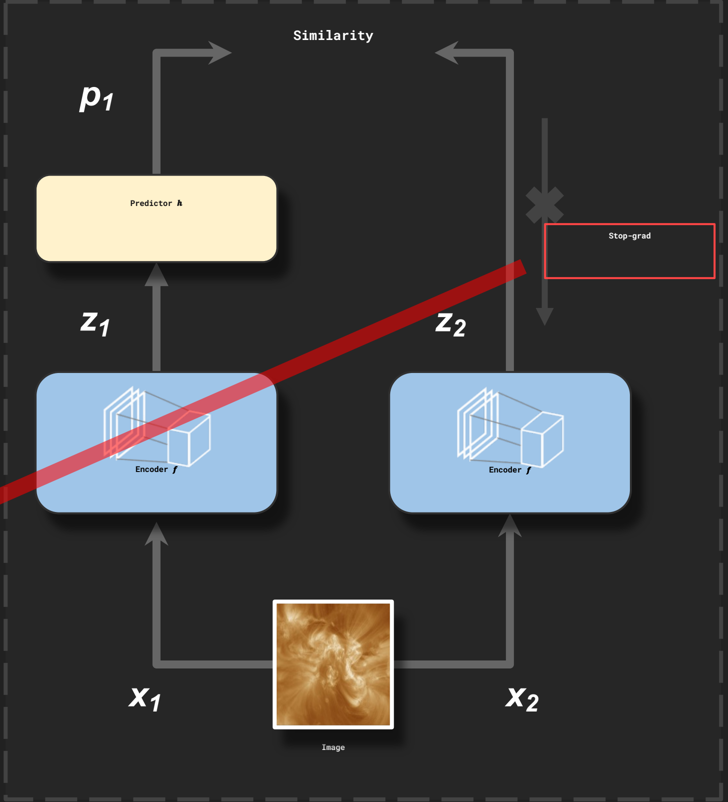
✗ Utilize various architectural asymmetries to prevent collapse

✓ Back-propagation through only one branch

+ Stop-gradient operation is critical for preventing collapse

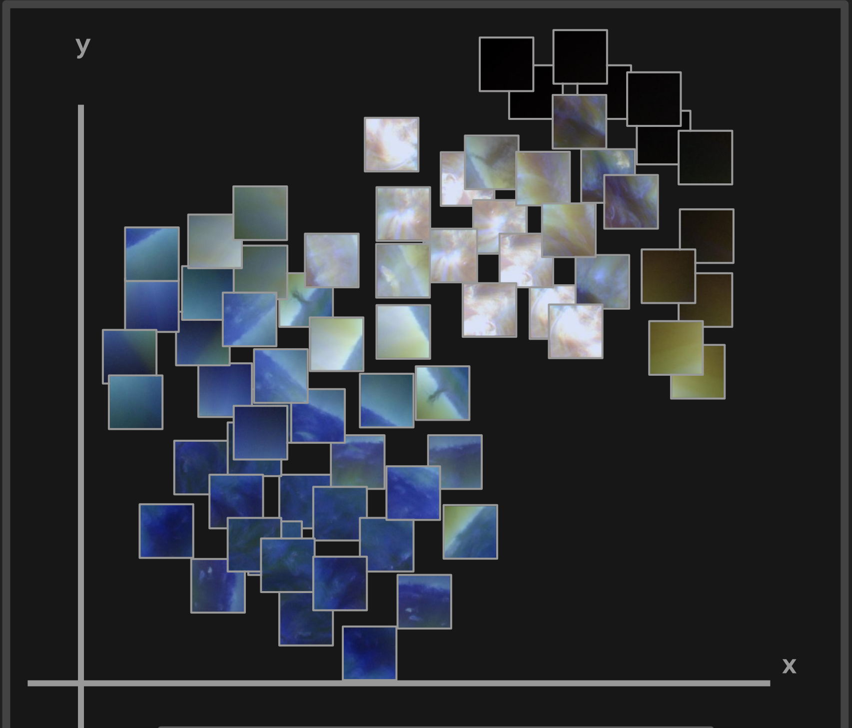
+ Collapse level measured through standard deviation of vectors.



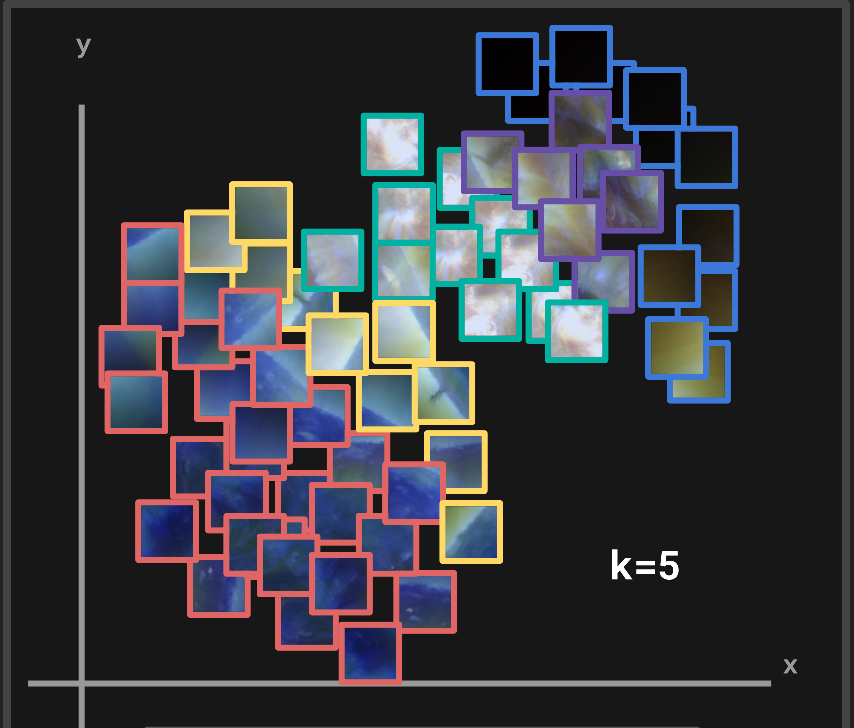


Chen & He, Exploring Simple Siamese Representation Learning (2020)

Introduction to Self Supervised Learning: Clustering

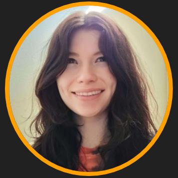


AIA 171 Embeddings



K-means Clusters

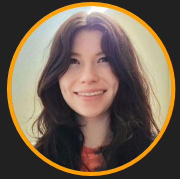
SimCLR: Similarity Search 🌻

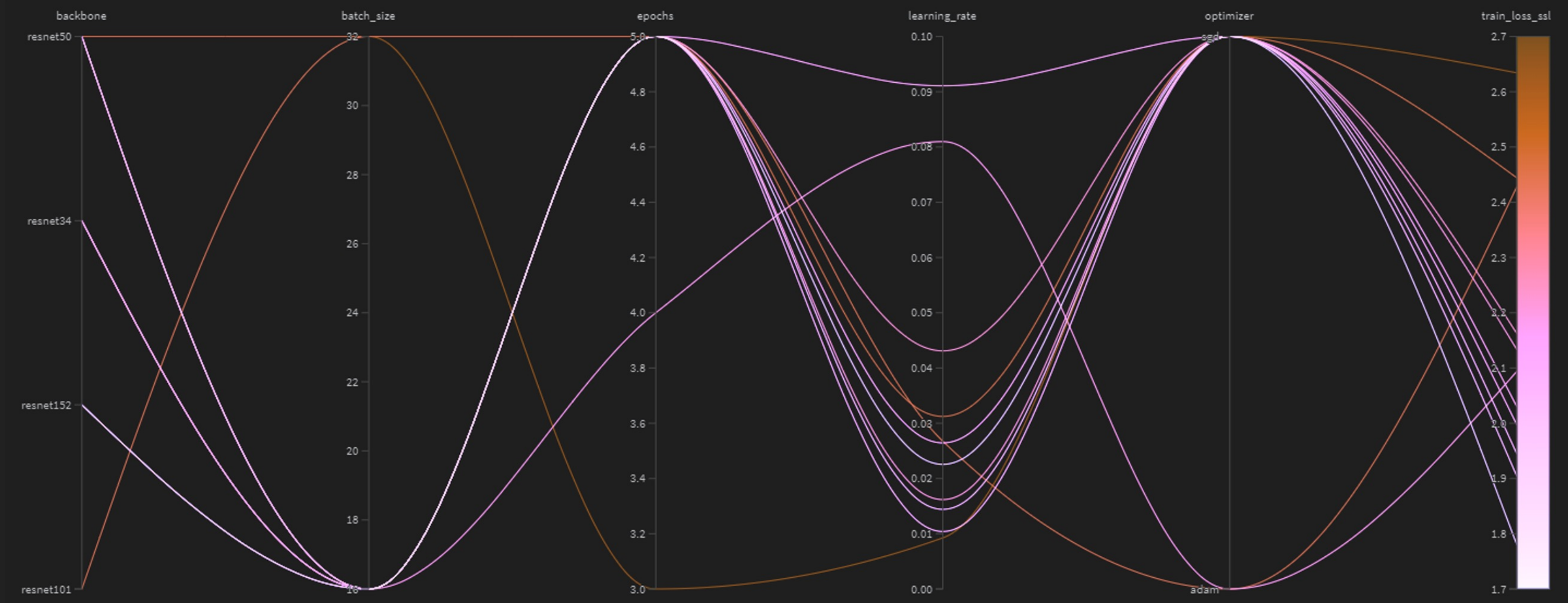


```
(main)
$ python model_run.py
Global seed set to 1
wandb: Currently logged in as: martinsierrasue (search-byol). Use `wandb login --relogin` to force r
elogin
wandb: wandb version 0.15.11 is available!  To upgrade, please run:
wandb:  $ pip install wandb --upgrade
wandb: Tracking run with wandb version 0.15.8
wandb: Run data is saved locally in D:\source\repos\hits-sdo-similaritysearch\search_simclr\visualiz
ations\simclr_knn\wandb\run-20230929_143018-27148n46
wandb: Run `wandb offline` to turn off syncing.
wandb: Syncing run sweet-paper-175
wandb:  View project at https://wandb.ai/search-byol/search_simclr
wandb:  View run at https://wandb.ai/search-byol/search_simclr/runs/27148n46
Not using wandb config
100%|██████████████████████████████████████████████| 10000
/10000 [00:00<00:00, 4996788.18it/s]
100%|██████████████████████████████████████████████| 800
0/8000 [00:00<00:00, 3996002.38it/s]
100%|██████████████████████████████████████████████| 200
0/2000 [00:00<00:00, 1996336.98it/s]
```

```
completed  350  /  500 epochs
Epochs completed:  76%|
completed  400  /  500 epochs
Epochs completed:  87%|
completed  450  /  500 epochs
Epochs completed: 100%|
```

SimCLR: Sweep

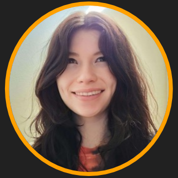


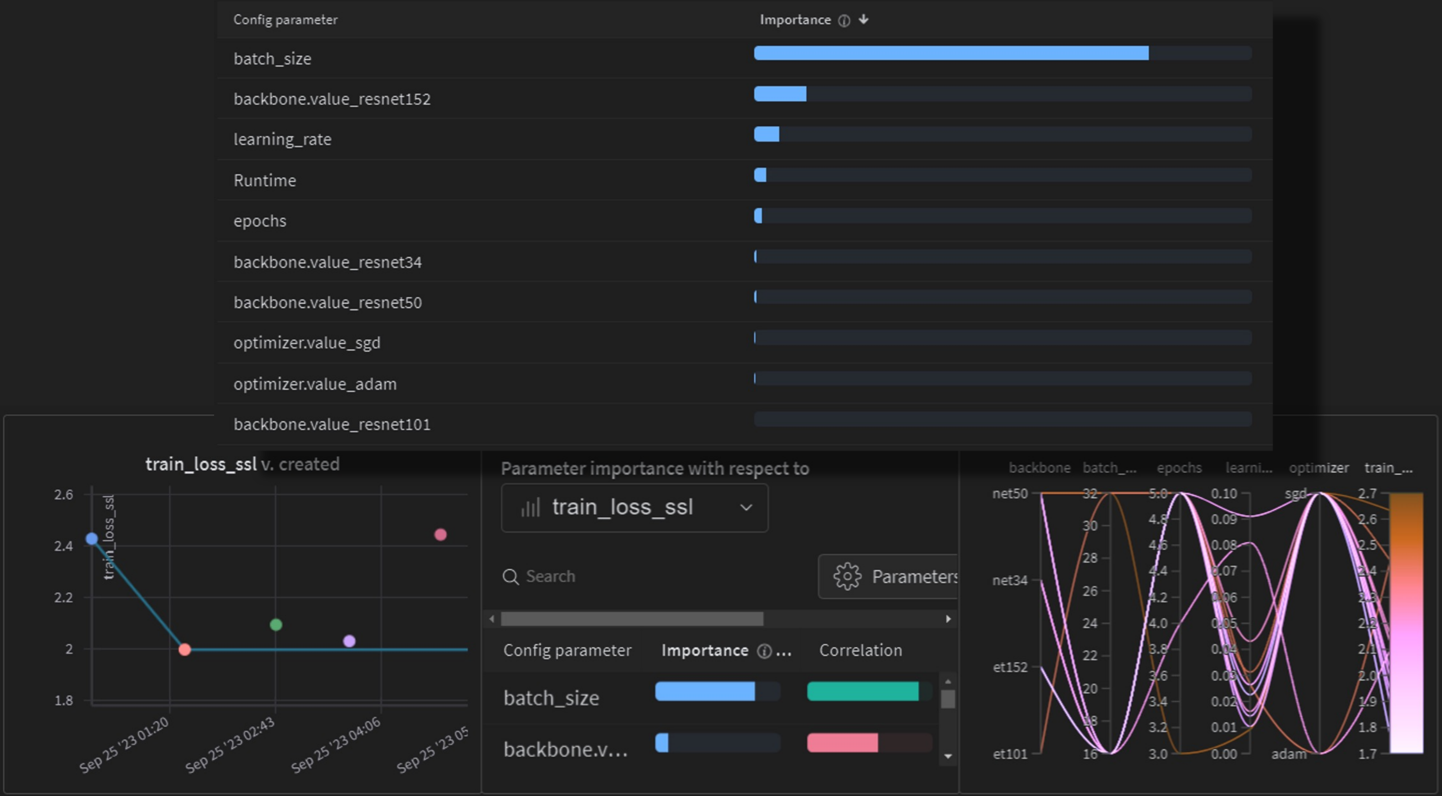


Features:

- “NT-Xent loss” (Normalized Temperature-Scaled Cross-Entropy Loss)
- Additional augmentations

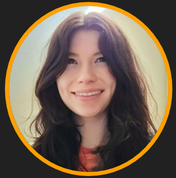
SimCLR: Hyperparameters



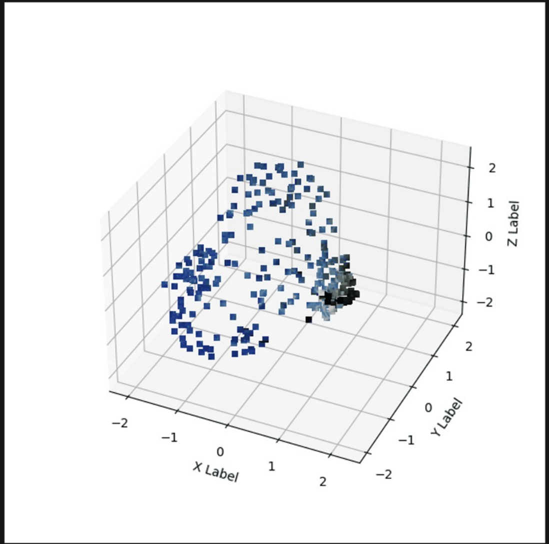


Config parameter	Importance
batch_size	
backbone.value_resnet152	
learning_rate	
Runtime	
epochs	
backbone.value_resnet34	
backbone.value_resnet50	
optimizer.value_sgd	
optimizer.value_adam	
backbone.value_resnet101	

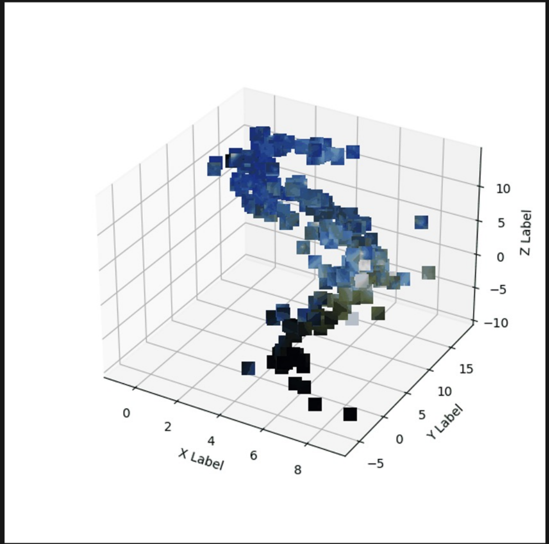
SimCLR: Plots



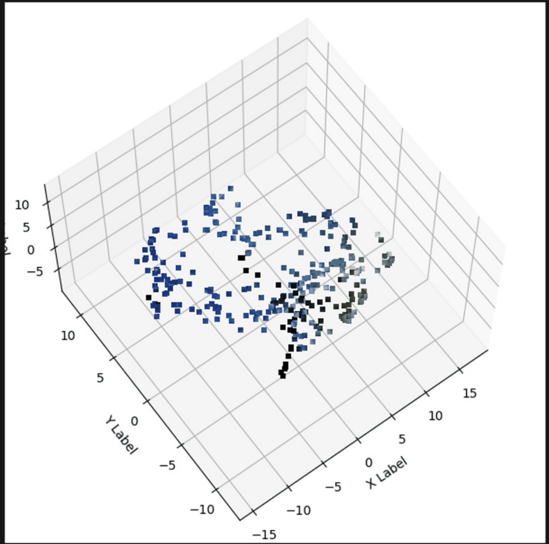
PCA



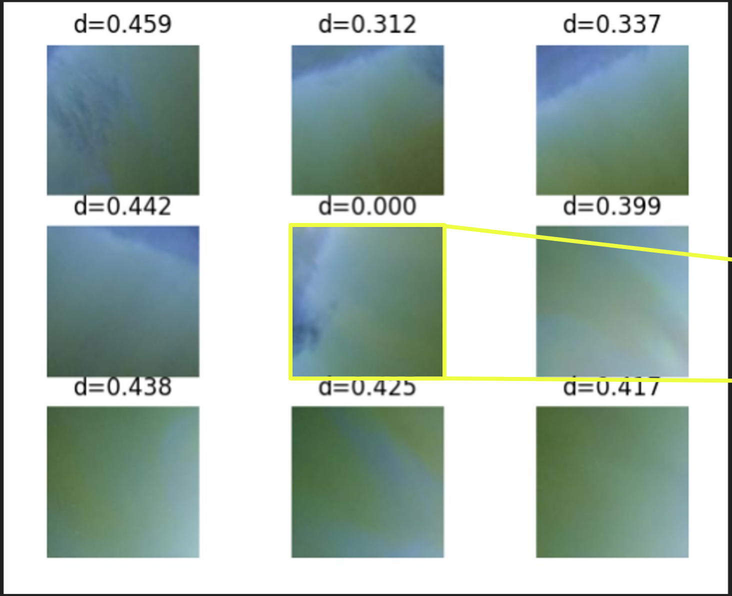
UMAP

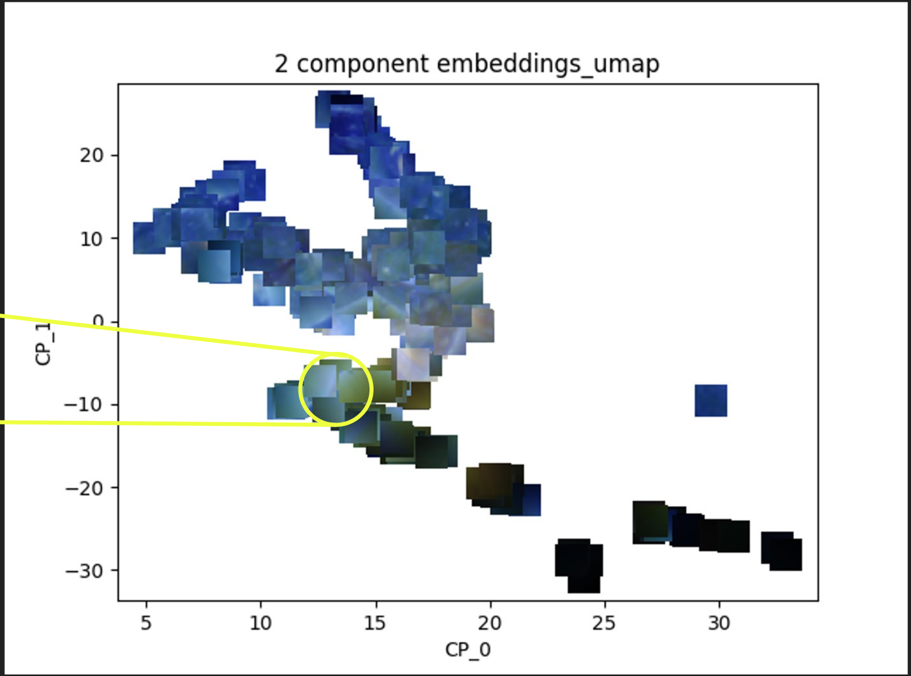


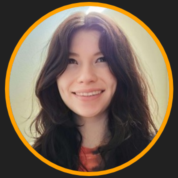
TSNE



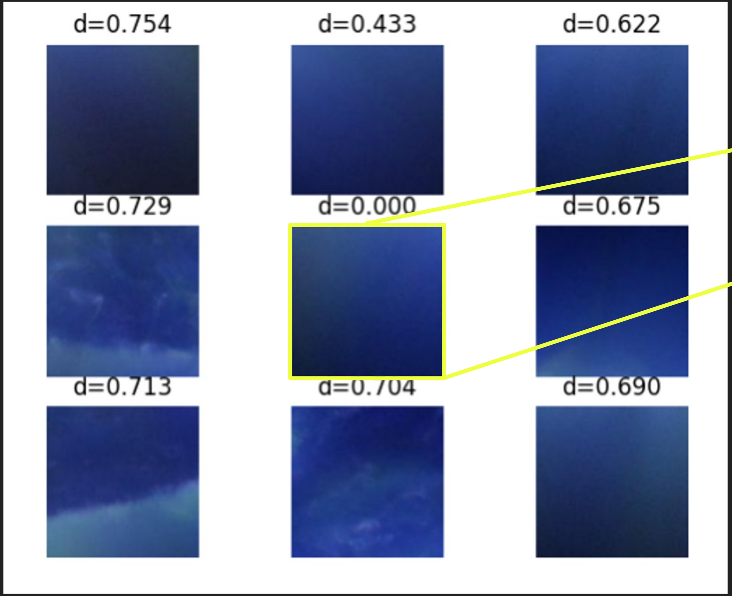
SimCLR: K-Nearest Neighbors

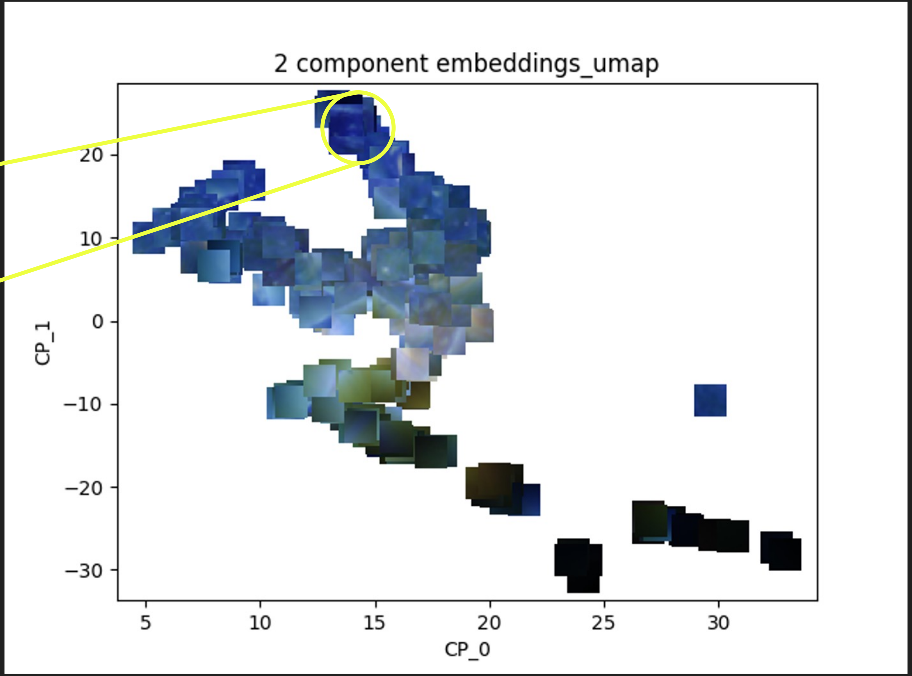


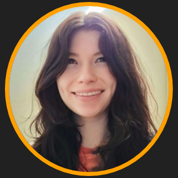




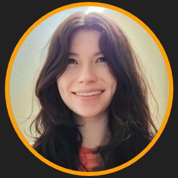
SimCLR: K-Nearest Neighbors

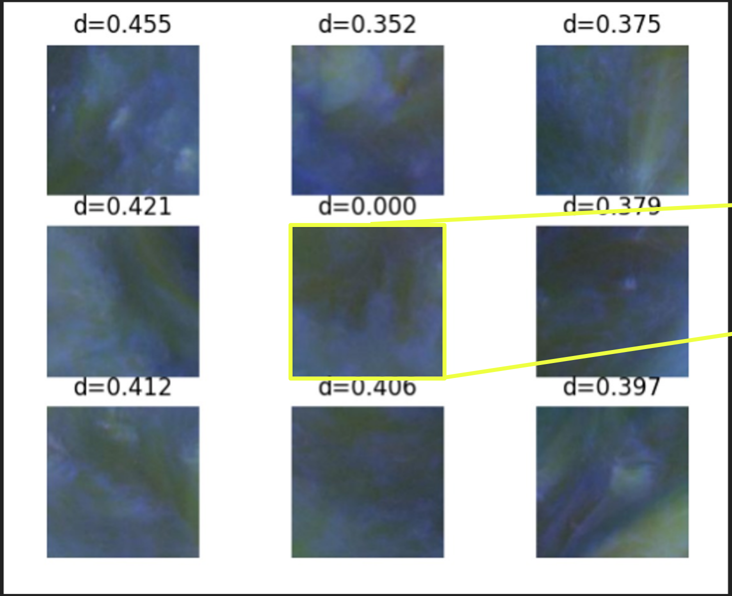


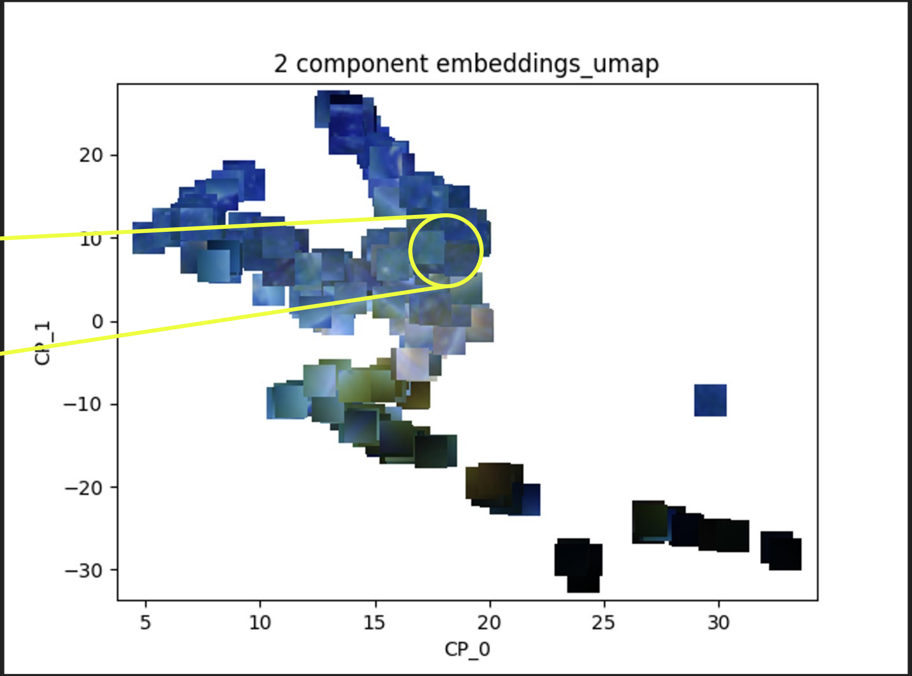




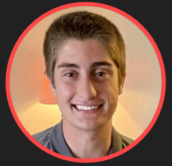
SimCLR: K-Nearest Neighbors







Team Highlights: Diverse sampling and Iterative training



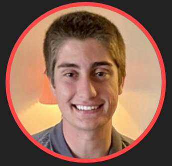
Diverse Sampler's Distance-Based Approach:

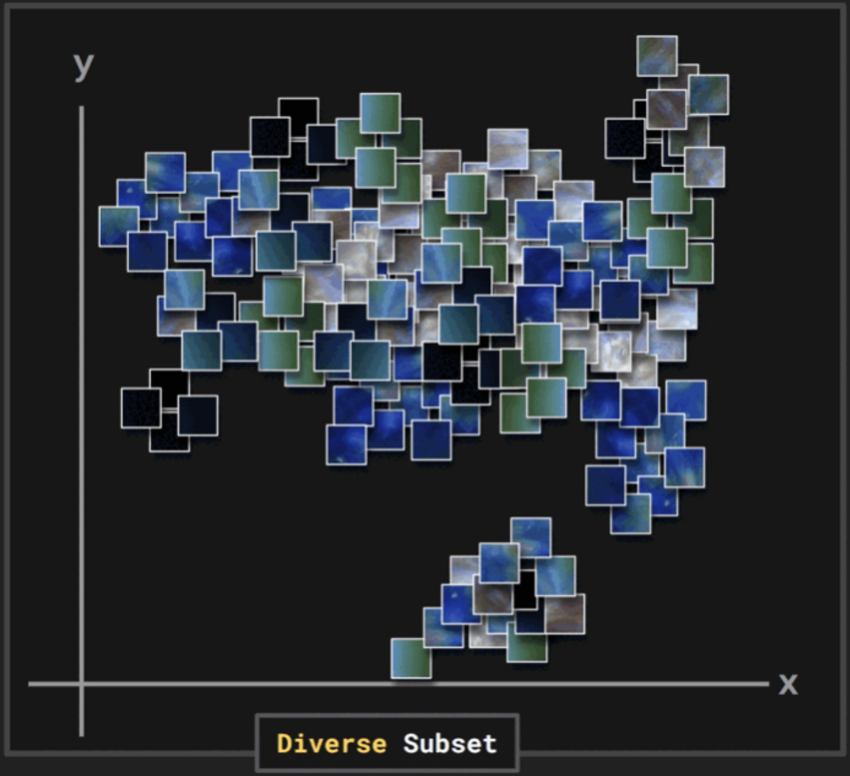
- Maximize point distance selection.
- Ensure diverse embedding coveraged in space regions.

Diverse Subset

Patel, Gao, Koul, Ganju & Kasam, Scalable Data Balancing for Unlabeled Satellite Imagery (2021)

Team Highlights: Diverse sampling and Iterative training



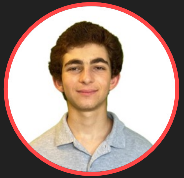


Enhanced Training with Diverse Sampling:

- Custom training loop integration.
- Iteratively update dataset with diverse samples.
- Broaden model exposure to capture data variations.

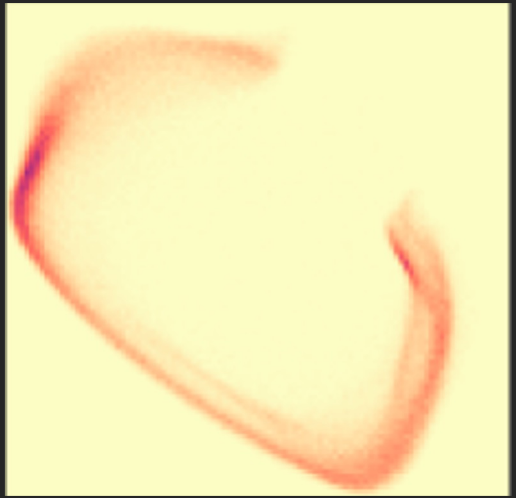
Patel, Gao, Koul, Ganju & Kasam, Scalable Data Balancing for Unlabeled Satellite Imagery (2021)

Team Highlights: Contrastive vs. Non-contrastive losses on embedding space



2D Histogram of Embedding Spaces

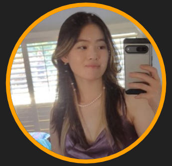
Non-Contrastive Loss



Contrastive Loss

Self-Supervised Search Demos involving a User Query

Concluding Remarks: Implications and Potential



Collaborative Learning

Automation

Advances in Heliophysics

Concluding Remarks: Future Work

- Fine-Tuning
  - Using *Labeled Data*
  - Different *Algorithms* and *Architecture*
- Improving JSOC Usability & Accessibility
- Optimization & Expansion
- ML Democratization outside Educational Institutions & Open Research Accessibility
- Continue Experimentation

Concluding Remarks: Conclusion

- We have brought together passionate individuals from diverse backgrounds in machine learning and heliophysics

- From humble beginnings, we have built advanced data and machine learning systems that have made the intricate world of scientific research a bit more approachable

- We are also proud of what we have done and we look forward to accomplishing more in the future