Multi-scale Deep Tensor Factorization Learns a Latent Representation of the Human Epigenome

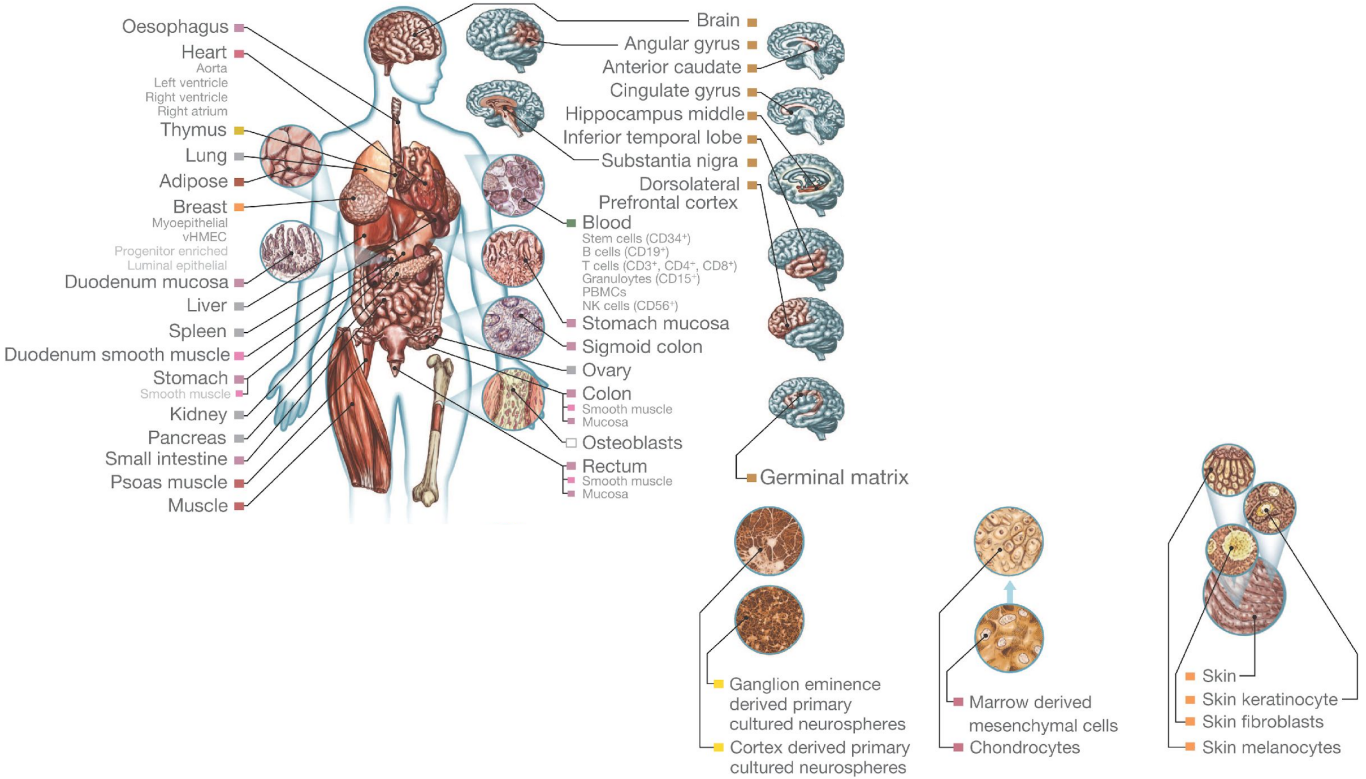
Jacob Schreiber
Paul G. Allen School of Computer Science and Engineering
University of Washington

jmschreiber91

@jmschrei

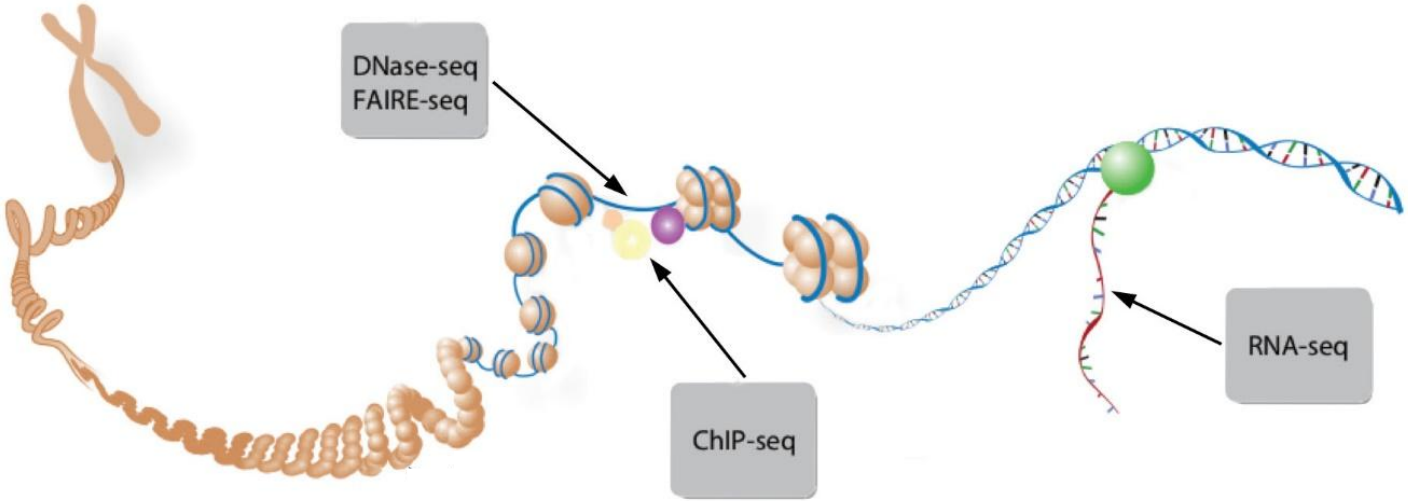
@jmschreiber91

The sequence of the human genome cannot explain the diversity of human cell types



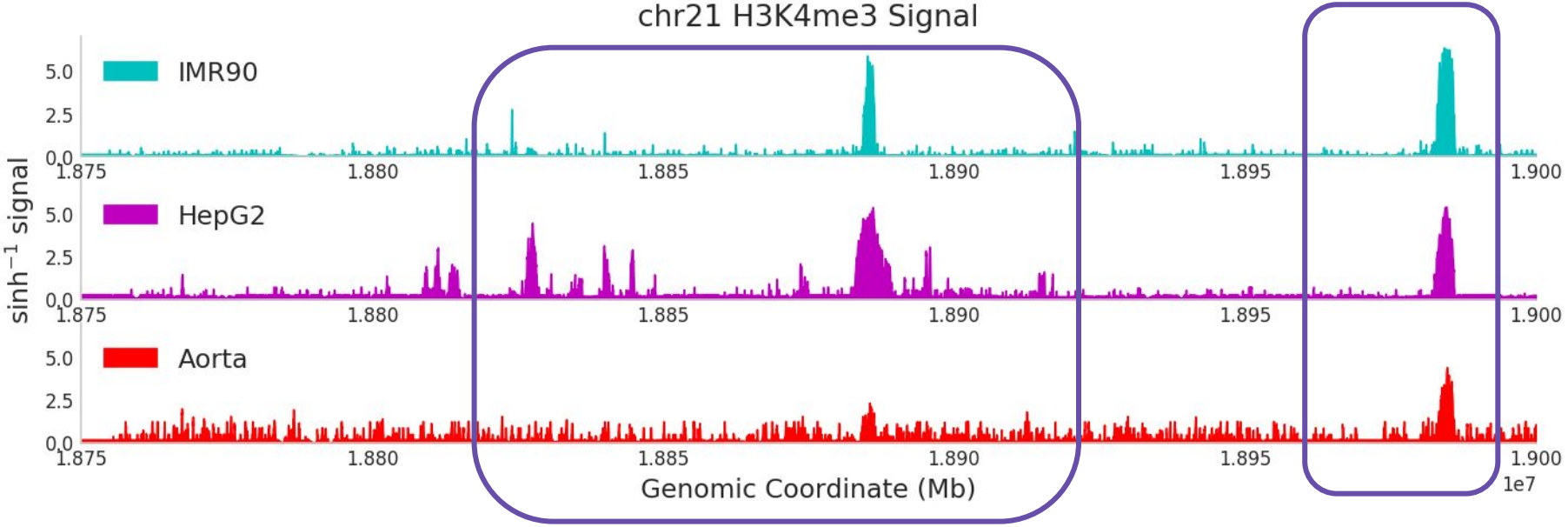
Credits: Roadmap Consortium, Nature (2015)

Many measurements can be gathered in addition to nucleotide sequence



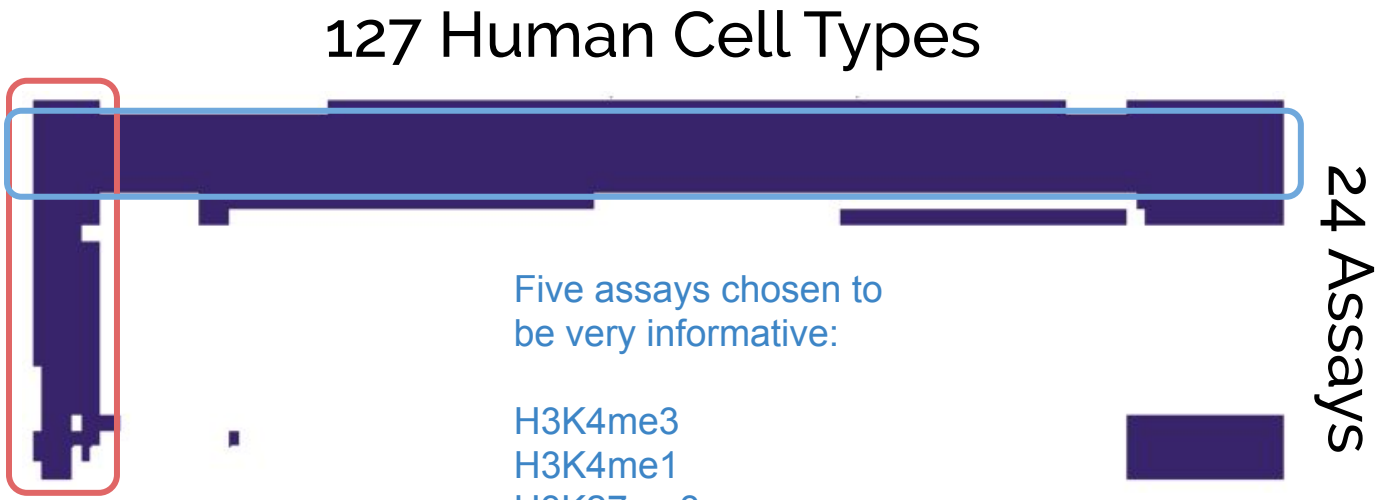
Credits: Darryl Leja (NHGRI), Ian Dunham (EBI)

The signal of epigenomic assays vary across cell types



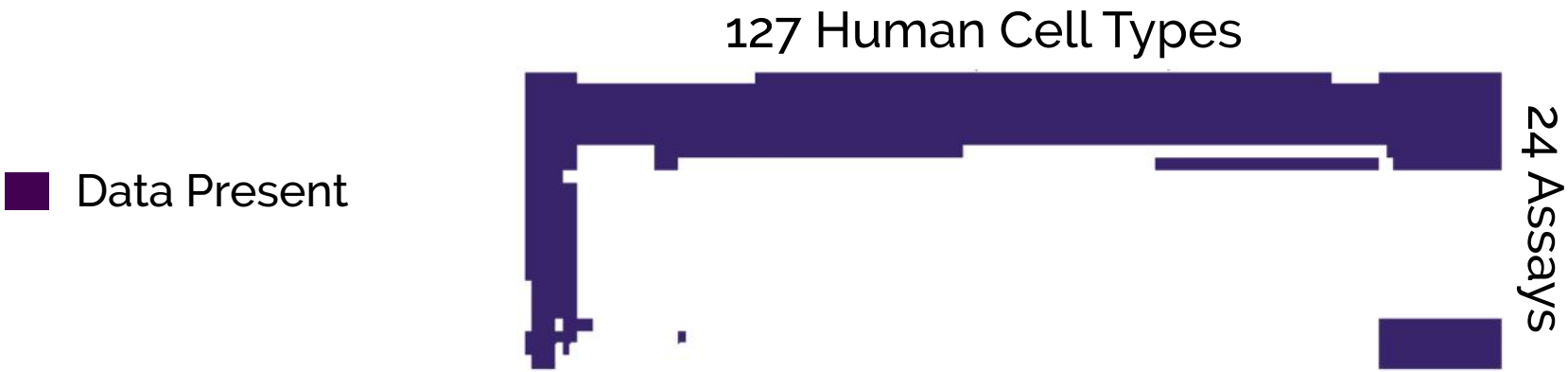
Many experiments have been performed, but still only a fraction of possible experiments

Data Present



1,014 experiments performed out of a possible 3,048

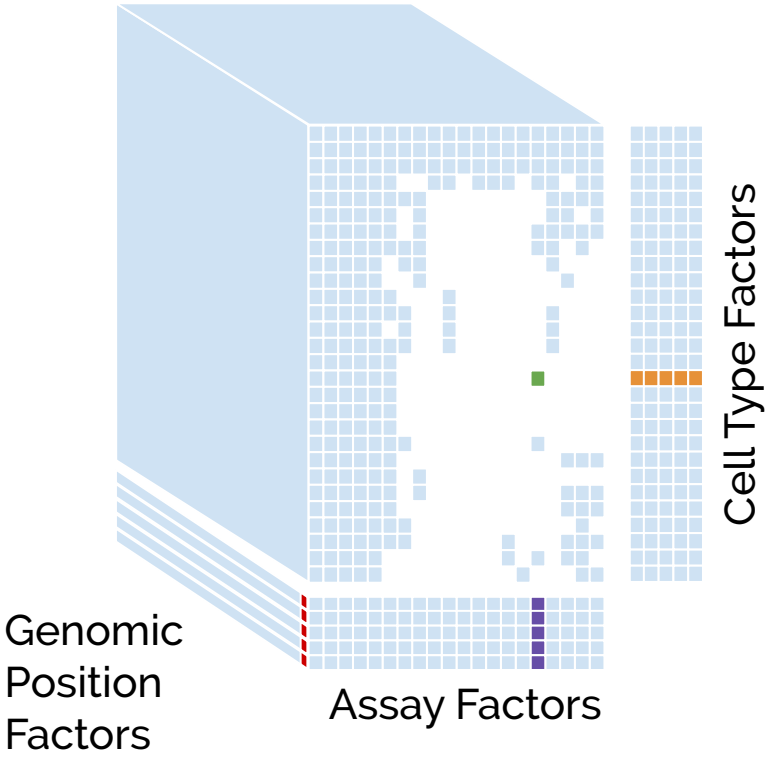
Have we characterized the human epigenome yet?

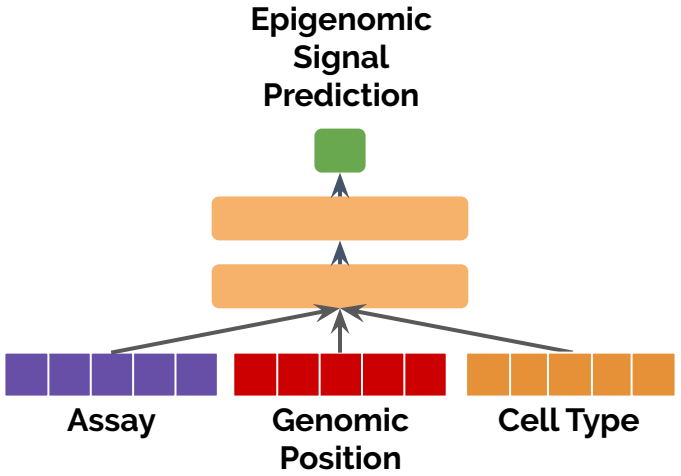


- Previous work sought to fully characterize the epigenome through imputing all potential experiments (ChromImpute[1], PREDICTD[2])
- Can we characterize the epigenome through distilling the available measurements into an informative latent representation?

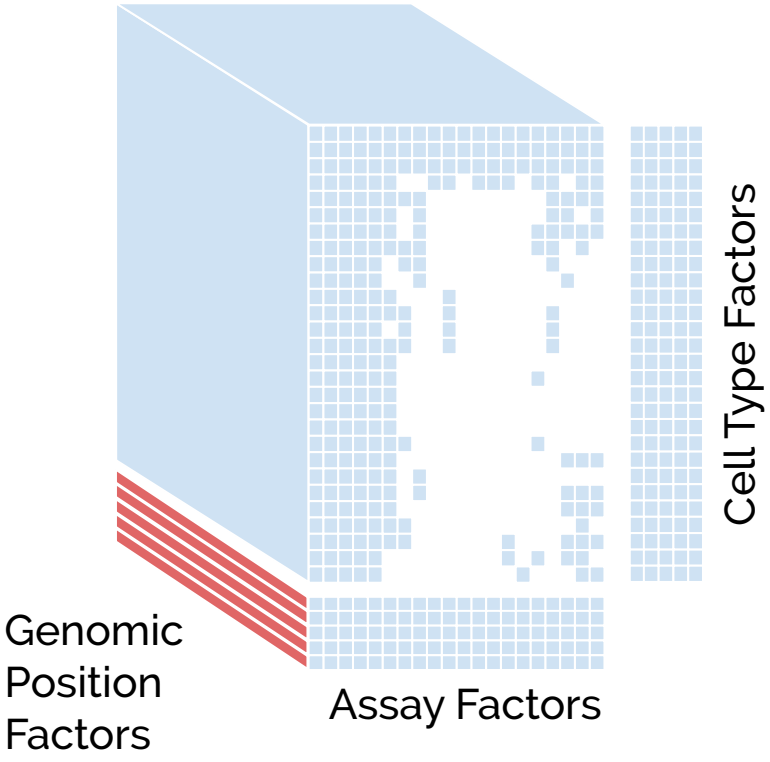
1. Ernst, et al. *Nature Methods, 2015*
2. Durham, et al. *Nature Communications, 2018*

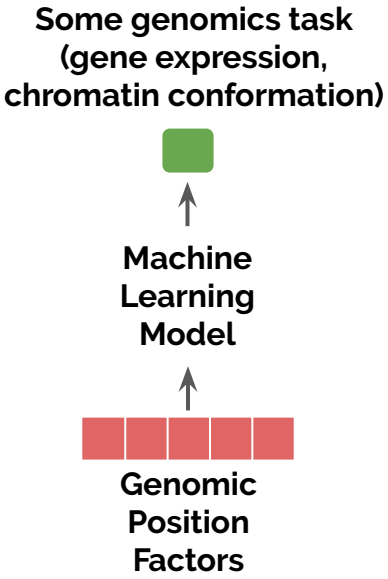
Avocado is a deep tensor factorization approach



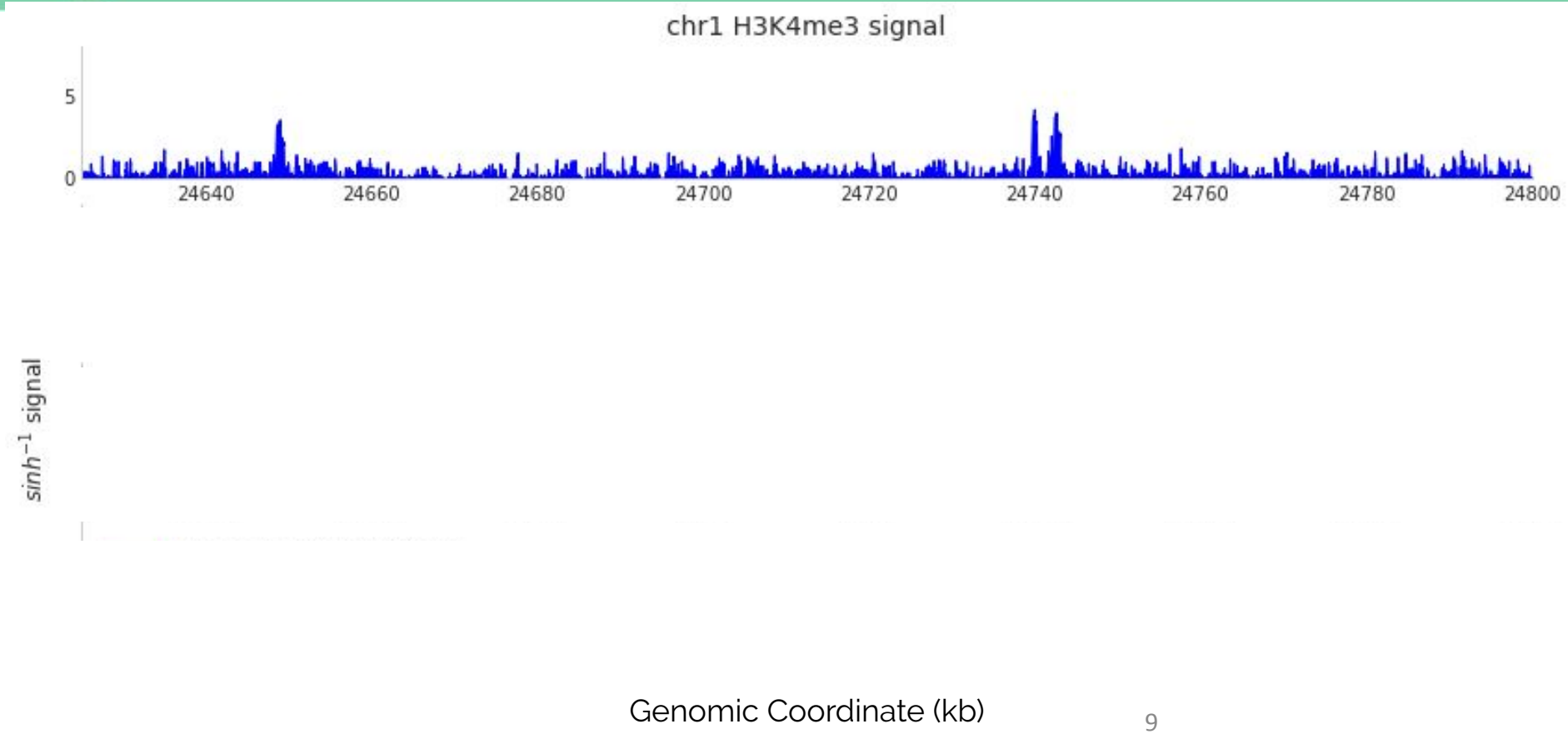


Our goal is to use the genomic latent factors for other tasks





Initial inspection of the imputations suggest that Avocado performs well



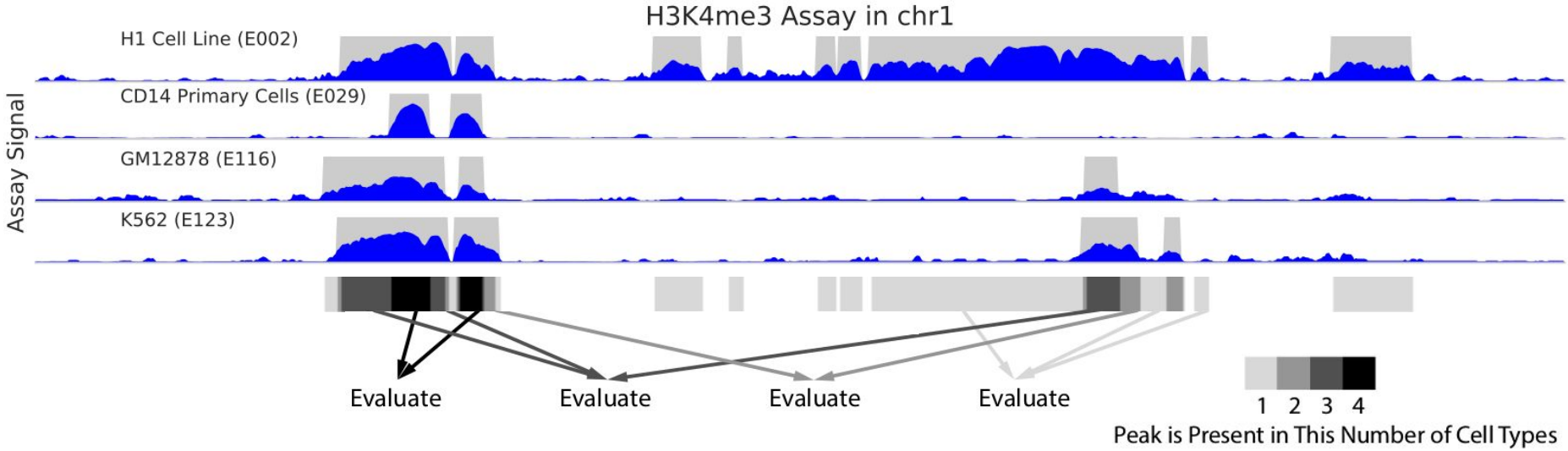
Avocado continues to perform well genome-wide

???

MSE-	global	1obs	1imp	Prom	Gene	Enh
ChromImpute	0.113	**0.941**	1.09	0.3246	0.1494	0.3164
PREDICTD	**0.1**	1.76	0.897	0.2576	**0.1295**	0.267
Avocado	**0.1**	1.66	**0.845**	**0.249**	**0.1295**	**0.26**

MSE-global: Mean squared error (MSE) across the full length of the genome
MSE-1obs: MSE at the top 1% of genomic positions ranked by experimental signal
MSE-1imp: MSE at the top 1% of genomic positions ranked by imputed signal
MSE-Prom: MSE at promoter regions defined by GENCODE
MSE-Gene: MSE at gene bodies defined by GENCODE
MSE-Enh: MSE at enhancer regions defined by FANTOM5

How well can these approaches recover cell type specific peaks?



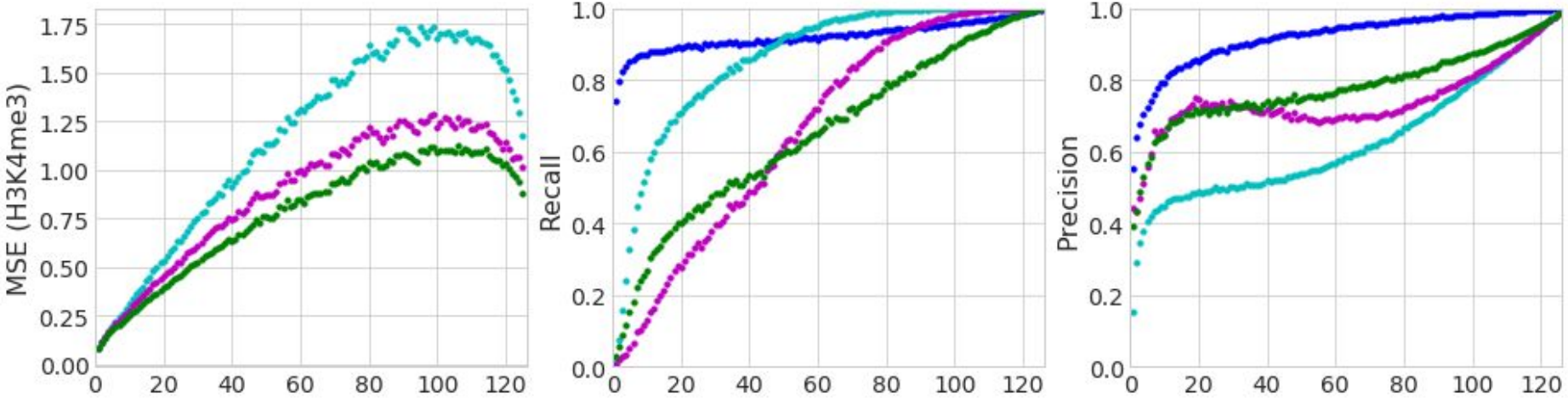
Evaluate by calculating:

(1) MSE

(2) Recall (thresholding the imputed signal at 1.44)

(3) Precision (thresholding the imputed signal at 1.44)

How well can these approaches recover cell type specific peaks?



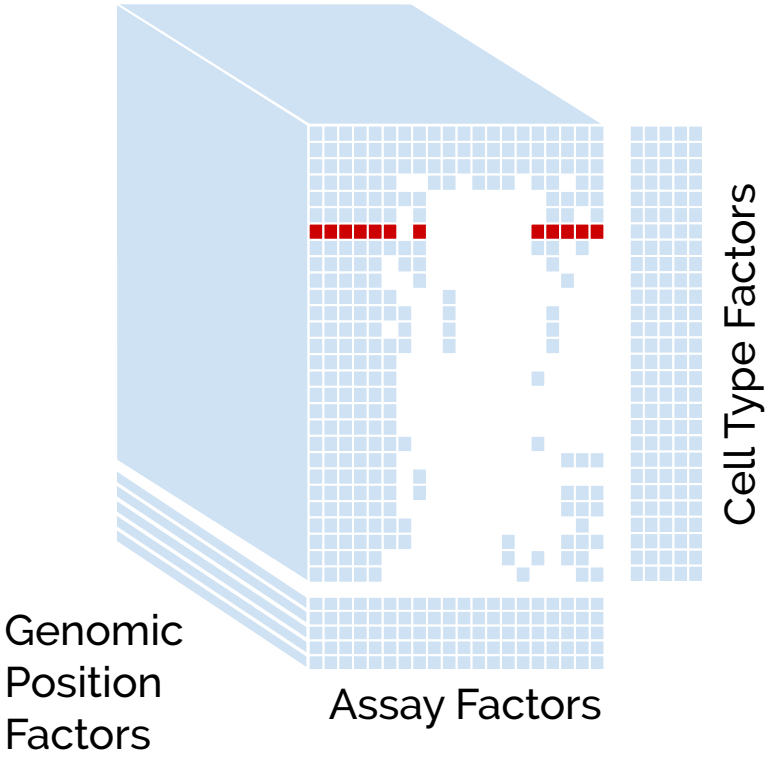
We evaluated our learned representation in many contexts

STEP 1:
Choose a Prediction Task
- Gene Expression
- Promoter-Enhancer Interactions
- Frequently Interacting REgions (FIREs)
- Topologically Associating Domain (TAD) boundaries

STEP 2:
Choose a Cell Type
- Task dependant

STEP 3:
Choose a Feature Set
- Available epigenomic tracks from the chosen cell type
- Full set of ChromImpute imputed marks for that cell type
- Full set of PREDICTD imputed marks for that cell type
- Full set of Avocado imputed marks for that cell type
- Avocado latent factors
- Full Roadmap compendium

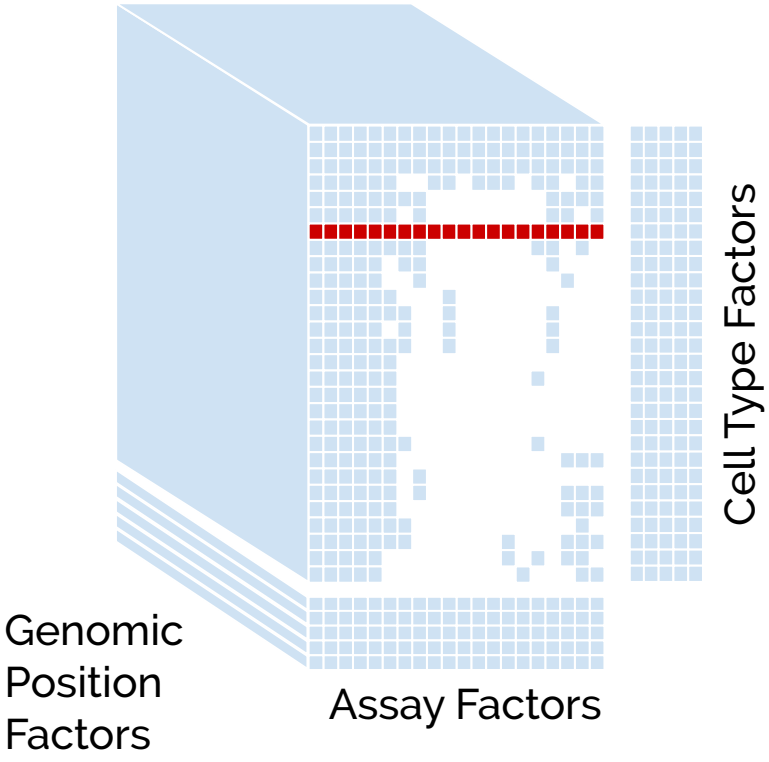
We evaluated our learned representation in many contexts



STEP 3:
Choose a Feature Set

- **Available epigenomic tracks from the chosen cell type**
- Full set of ChromImpute imputed marks for that cell type
- Full set of PREDICTD imputed marks for that cell type
- Full set of Avocado imputed marks for that cell type
- Avocado latent factors
- Full Roadmap compendium

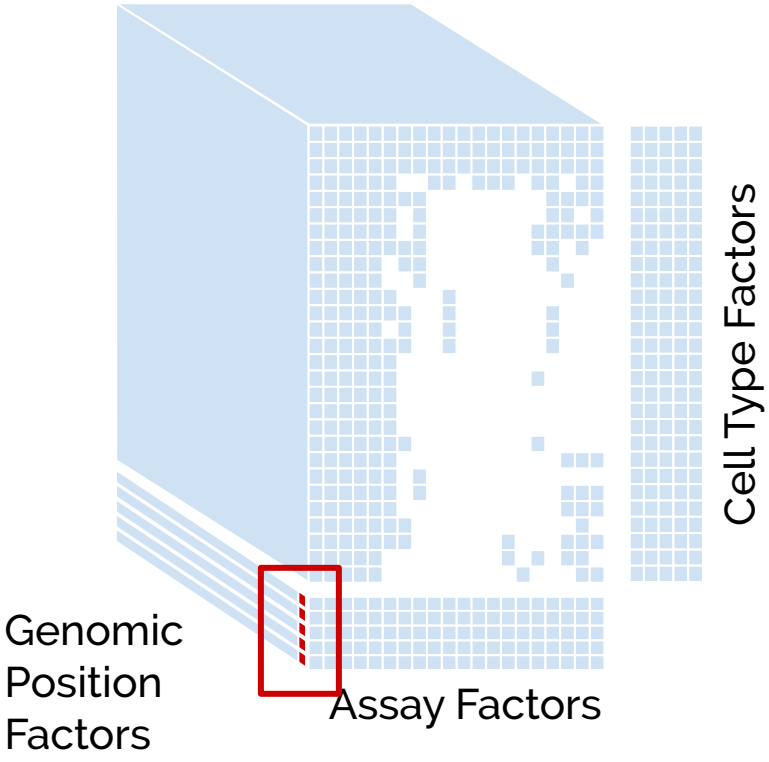
We evaluated our learned representation in many contexts



STEP 3:
Choose a Feature Set

- Available epigenomic tracks from the chosen cell type
- **Full set of ChromImpute imputed marks for that cell type**
- **Full set of PREDICTD imputed marks for that cell type**
- **Full set of Avocado imputed marks for that cell type**
- Avocado latent factors
- Full Roadmap compendium

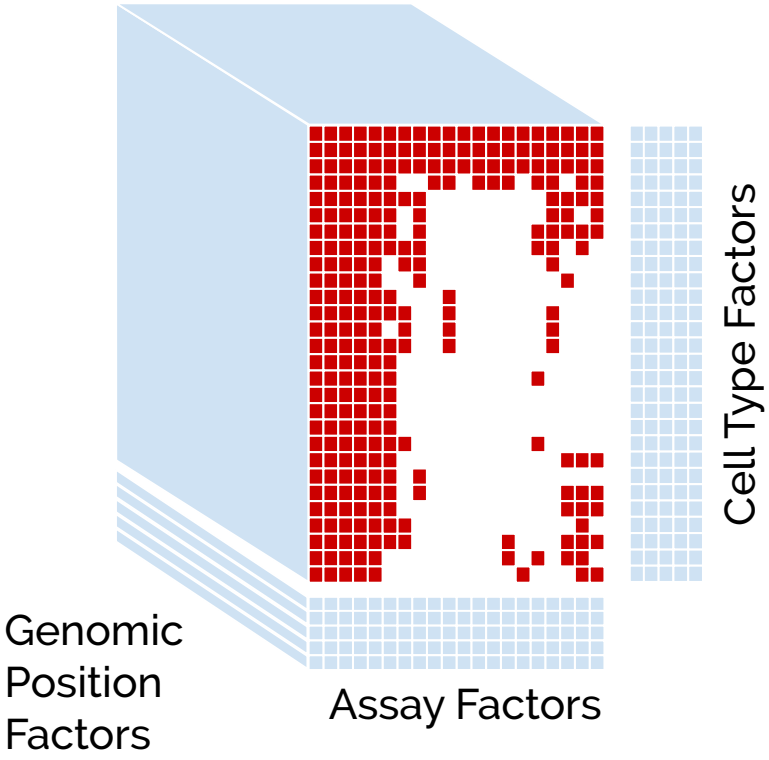
We evaluated our learned representation in many contexts



STEP 3:
Choose a Feature Set

- Available epigenomic tracks from the chosen cell type
- Full set of ChromImpute imputed marks for that cell type
- Full set of PREDICTD imputed marks for that cell type
- Full set of Avocado imputed marks for that cell type
- **Avocado latent factors**
- Full Roadmap compendium

We evaluated our learned representation in many contexts



STEP 3:
Choose a Feature Set

- Available epigenomic tracks from the chosen cell type
- Full set of ChromImpute imputed marks for that cell type
- Full set of PREDICTD imputed marks for that cell type
- Full set of Avocado imputed marks for that cell type
- Avocado latent factors
- **Full Roadmap compendium**

We evaluated our learned representation in many contexts

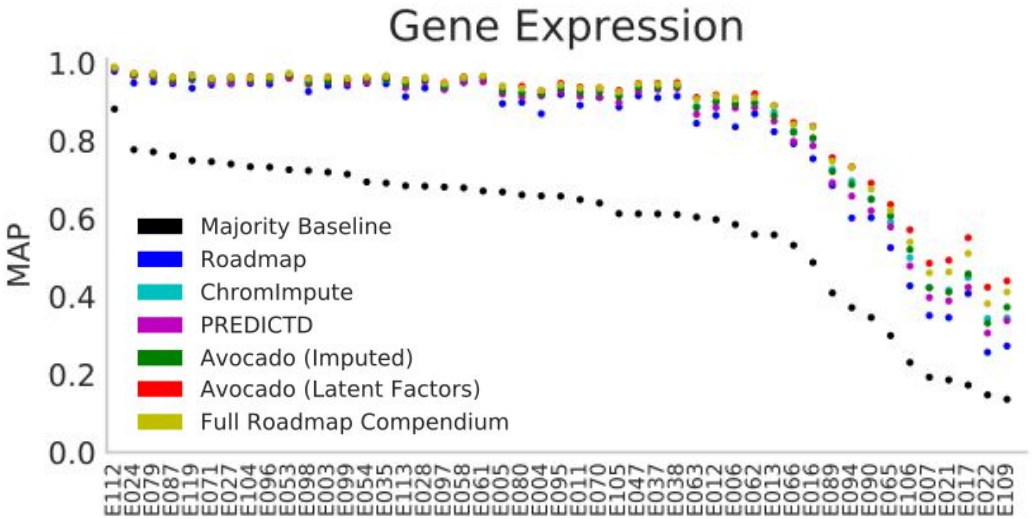
STEP 1:
Choose a Prediction Task
- Gene Expression
- Frequently Interacting REgions (FIREs)
- Topologically Associating Domain (TAD) boundaries

STEP 2:
Choose a Cell Type
- Task dependant

STEP 3:
Choose a Feature Set
- Available epigenomic tracks from the chosen cell type
- Full set of ChromImpute imputed marks for that cell type
- Full set of PREDICTD imputed marks for that cell type
- Full set of Avocado imputed marks for that cell type
- Avocado latent factors
- Full Roadmap compendium

STEP 4:
Run 5 fold CV on data set using a gradient boosting machine classifier and calculate the mean average precision (MAP) over all five folds

Avocado latent factors can predict gene expression



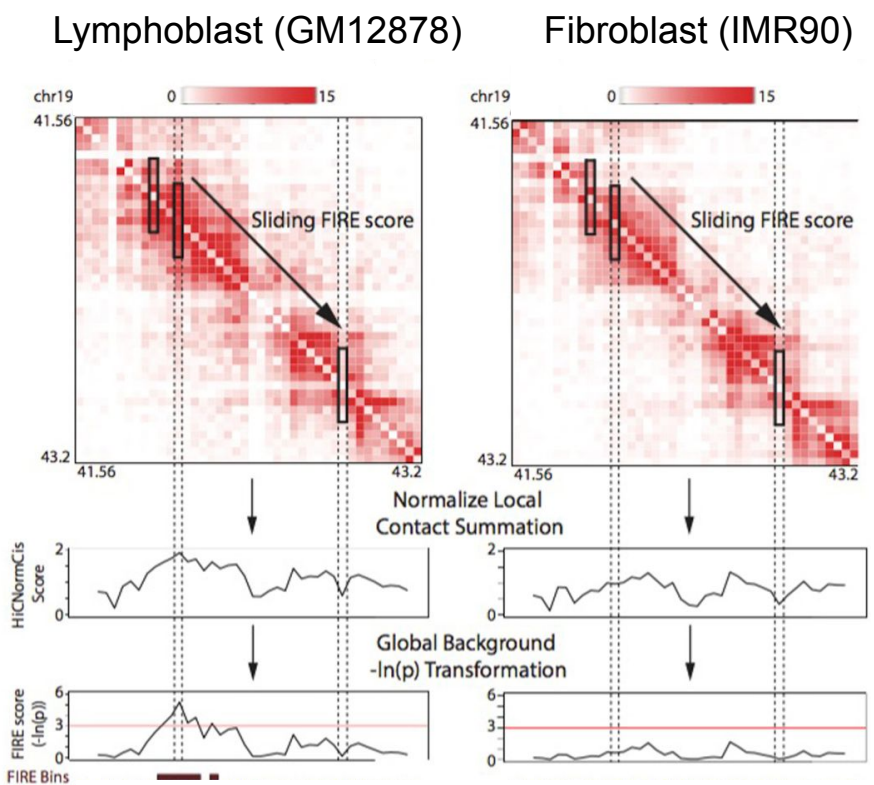
Avocado > Epigenomic Measurements

- All cell types
- By an average of 0.144 MAP
- By an average of 0.167 MAP on the 7 most difficult cell types

Avocado > Full Roadmap Compendium

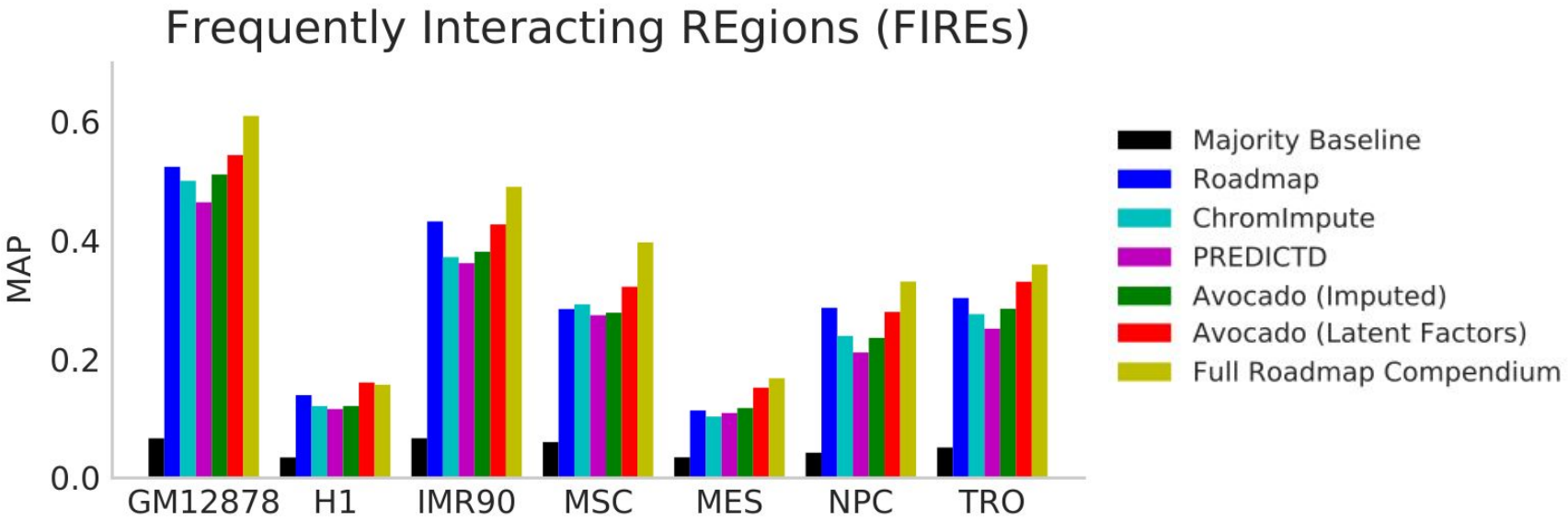
- 36 / 47 cell types
- By an average of 0.006 MAP
- By an average of 0.03 MAP on the 7 most difficult cell types

Avocado latent factors can predict FIREs

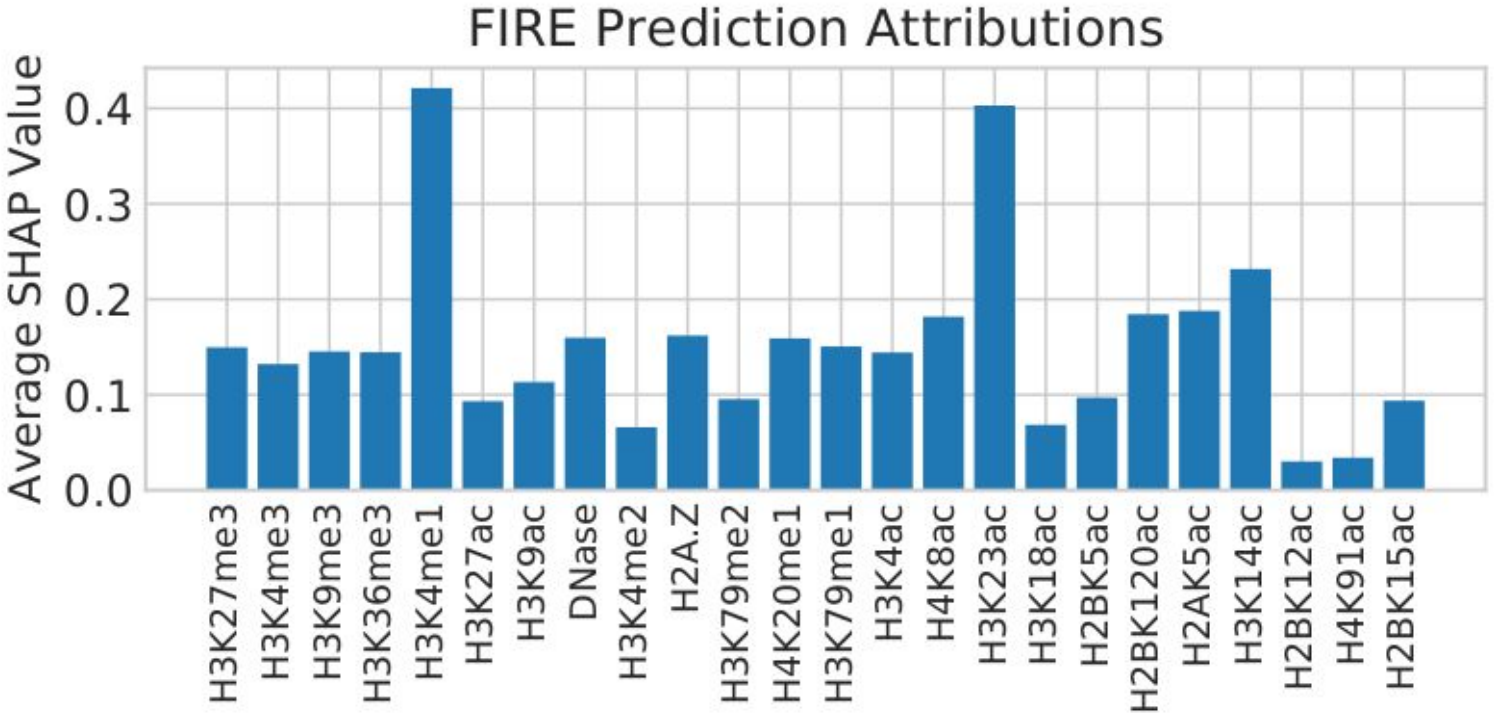


Schmitt et al, 2016

Avocado latent factors can predict FIREs



Feature attribution methods reveal two important marks



Review

- Avocado is a deep tensor factorization approach for modeling the human epigenome
- After being trained to impute epigenomic marks, it yields more accurate imputations than previous work
- Avocado's genome latent factors serve as a useful input for machine learning models on downstream genomics tasks, outperforming using epigenomic measurements themselves
- Using the entirety of the Roadmap compendium appears to be a stronger baseline than expected suggesting that measurements in many cell types can aid the prediction for a single cell type

Preprint, model, and GitHub repo are online now!

45 commits | 1 branch | 0 releases | 1 contributor | View license

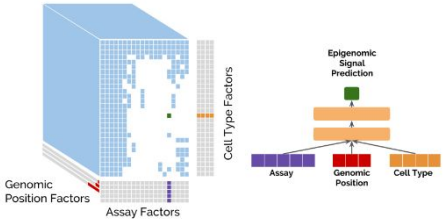
Branch: master | New pull request | Create new file | Upload files | Find file | Clone or download

jmschrei Update README.md — Latest commit 7551492 22 days ago

avocado	v0.3.0	23 days ago
data	Initial commit	2 months ago
figures	Add files via upload	27 days ago
Avocado Downstream Task Demo.ipynb	Initial commit	2 months ago
Avocado Training Demo.ipynb	ADD new tutorial	27 days ago
LICENSE	Initial commit	2 months ago
README.md	Update README.md	22 days ago
setup.py	v0.3.0	23 days ago

README.md

avocado



Avocado is a multi-scale deep tensor factorization model that is used to learn a latent representation of the human epigenome. The purpose of this model is two fold; first, to impute epigenomic experiments that have not yet been performed, and second, to learn a latest representation of the human epigenome that can be used as input for machine learning models in the place of epigenomic data itself. The project page with links to the full set of imputations and model parameters can be found at https://noble.gs.washington.edu/proj/avocado/. The manuscript is currently under review and the preprint can be found here.

Installation

Avocado can be installed using pip.

```
pip install avocado-epigenome
```

https://github.com/jmschrei/avocado

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | AB | CHANNEL

Search

New Results

Multi-scale deep tensor factorization learns a latent representation of the human epigenome

Jacob Schreiber, Timothy J Durham, Jeffrey Bilmes, William Stafford Noble

doi: https://doi.org/10.1101/364976

This article is a preprint and has not been peer-reviewed [what does this mean?].

Abstract | Info/History | Metrics | Supplementary material | Preview PDF

Abstract

The human epigenome has been experimentally characterized by measurements of protein binding, chromatin acessibility, methylation, and histone modification in hundreds of cell types. The result is a huge compendium of data, consisting of thousands of measurements for every basepair in the human genome. These data are difficult to make sense of, not only for humans,

Acknowledgements

eScience Institute
ADVANCING DATA-INTENSIVE DISCOVERY IN ALL FIELDS

NSF National Science Foundation
WHERE DISCOVERIES BEGIN

The Netflix Challenge was a similar problem!

Users (m) ⟶

Movies (n) ⟶

≈

Movie Factors (n by k)

M

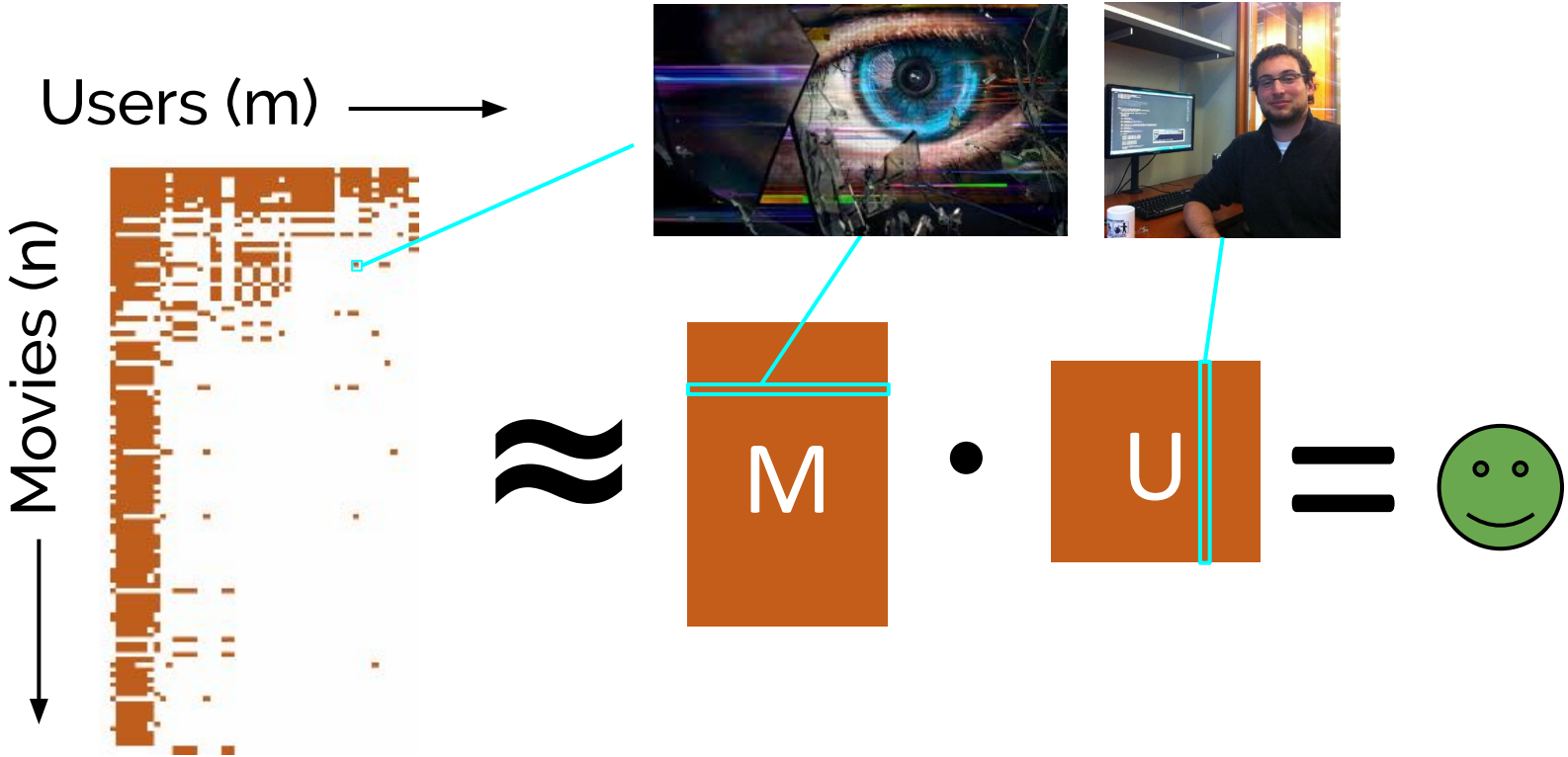
•

User Factors (k by m)

U

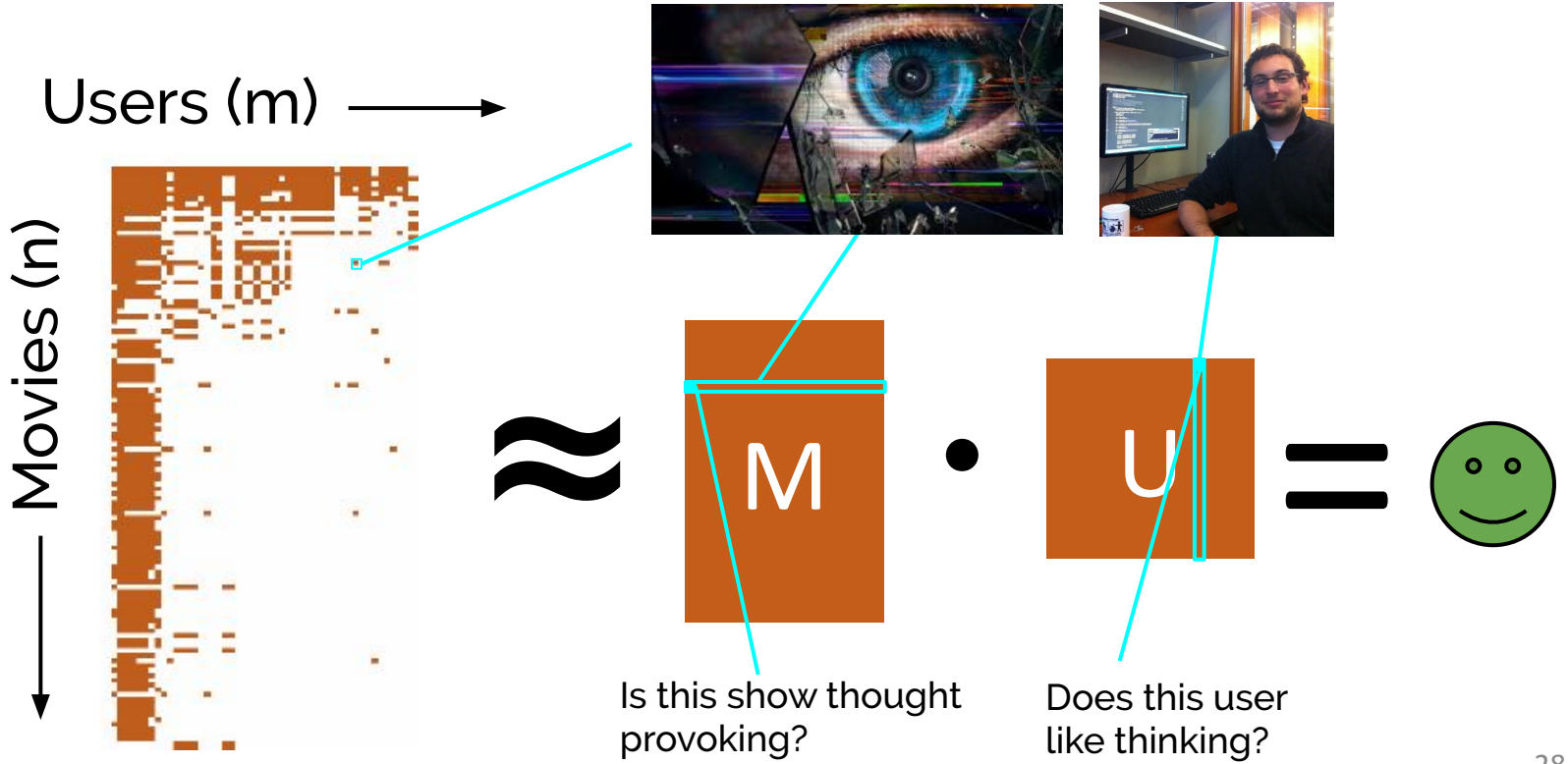
Credits: Tim Durham

Train ‘Black Mirror’ factors and ‘Jacob’ factors based on my rating of the show



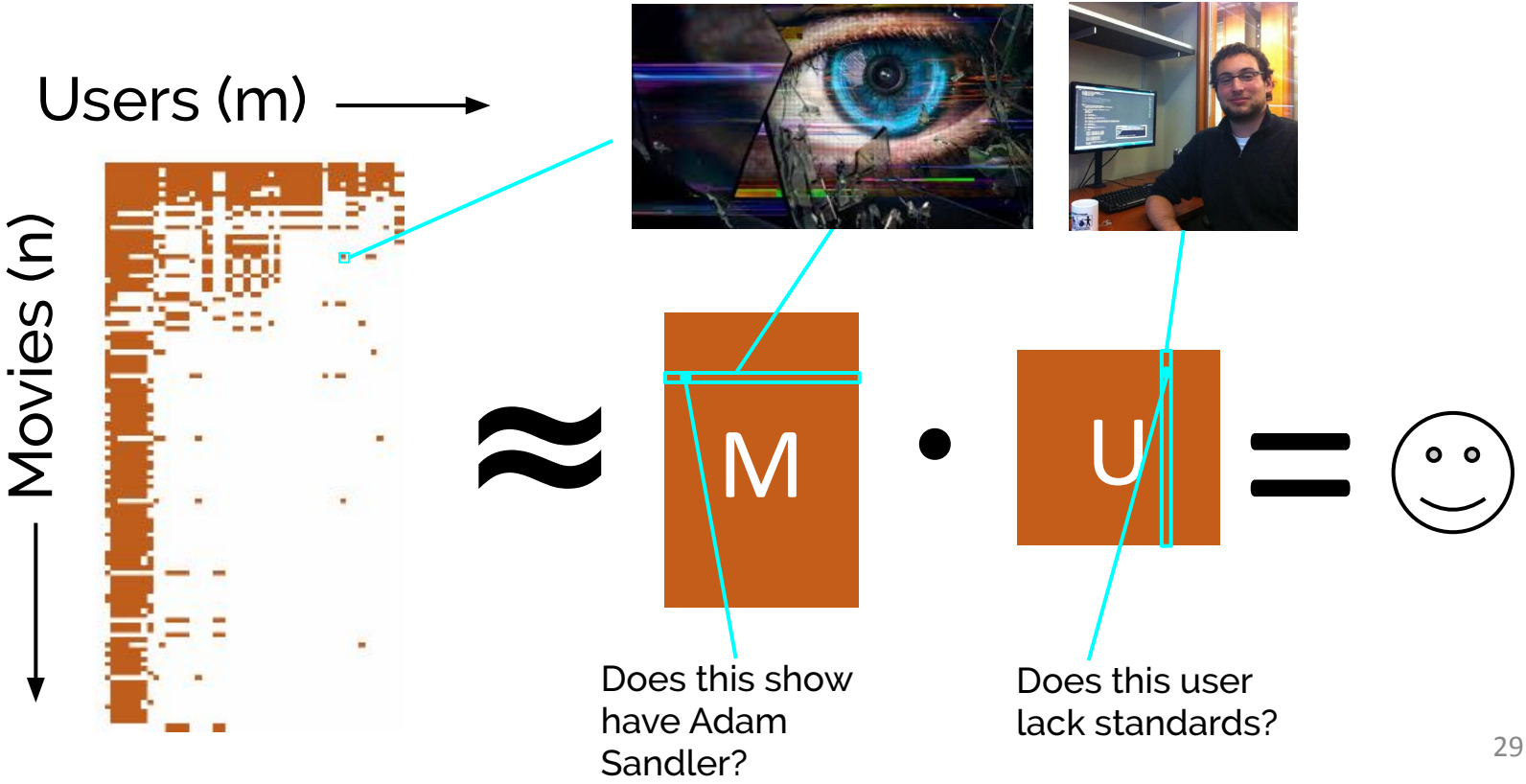
Credits: Tim Durham

Train ‘Black Mirror’ factors and ‘Jacob’ factors based on my rating of the show



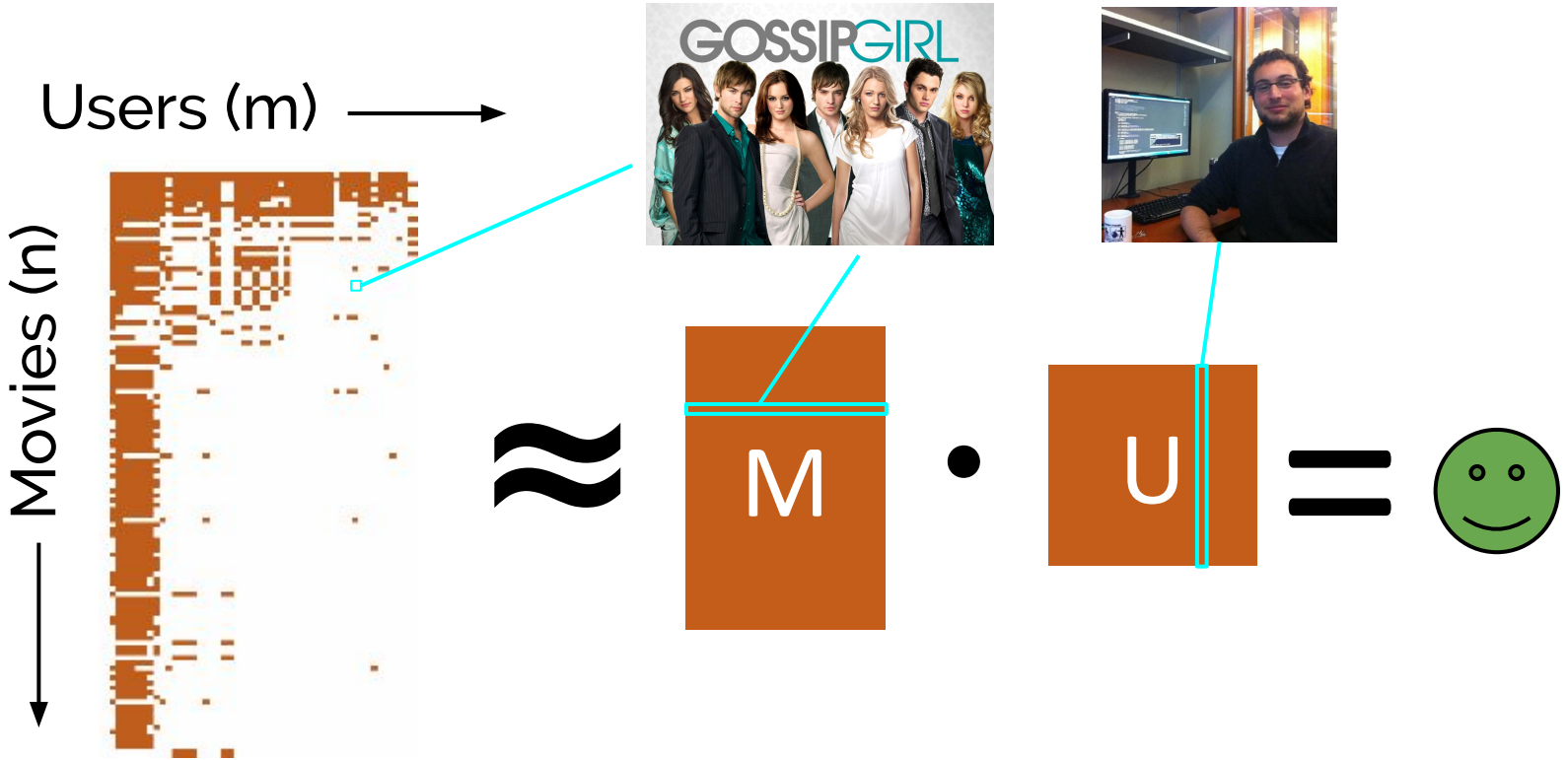
Credits: Tim Durham

Train ‘Black Mirror’ factors and ‘Jacob’ factors based on my rating of the show



Credits: Tim Durham

Based on these learned factors, predict user ratings for all other shows



Credits: Tim Durham