HistoEncoder

Towards a foundation model for digital pathology

Joona **Pohjonen**, Antti **Rannikko**, Tuomas **Mirtti**, Esa **Pitkänen**

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

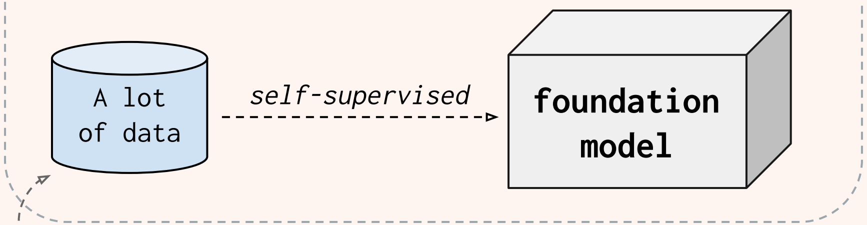
+ FIMM

What is a foundation model?

LUMI

pre-training

Centralises information from vasts amounts of data.



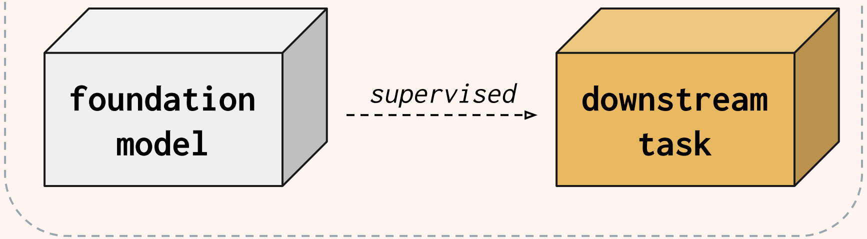
A lot of data

self-supervised

foundation model

adaptation

One model adapts easily to a wide range of downstream tasks.

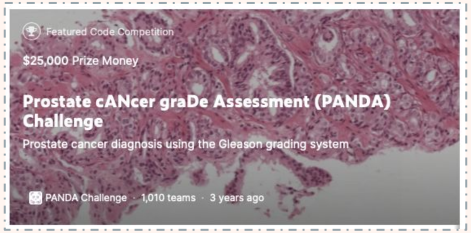


foundation model

supervised

downstream task

Prostate tissue samples from Helsinki University Hospital



Featured Code Competition

$25,000 Prize Money

Prostate cANcer graDe Assessment (PANDA) Challenge

Prostate cancer diagnosis using the Gleason grading system

PANDA Challenge · 1,010 teams · 3 years ago

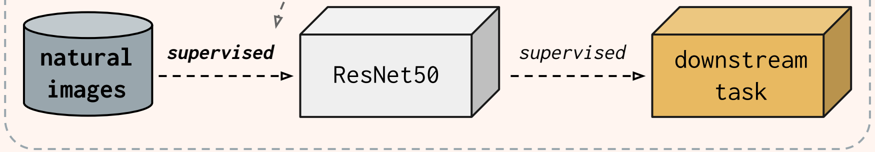
5.8 million tile images

898.4 million tile images

Dogs and cats differ widely from histological samples!!

supervised training is not good for pre-training!!

transfer learning



natural images

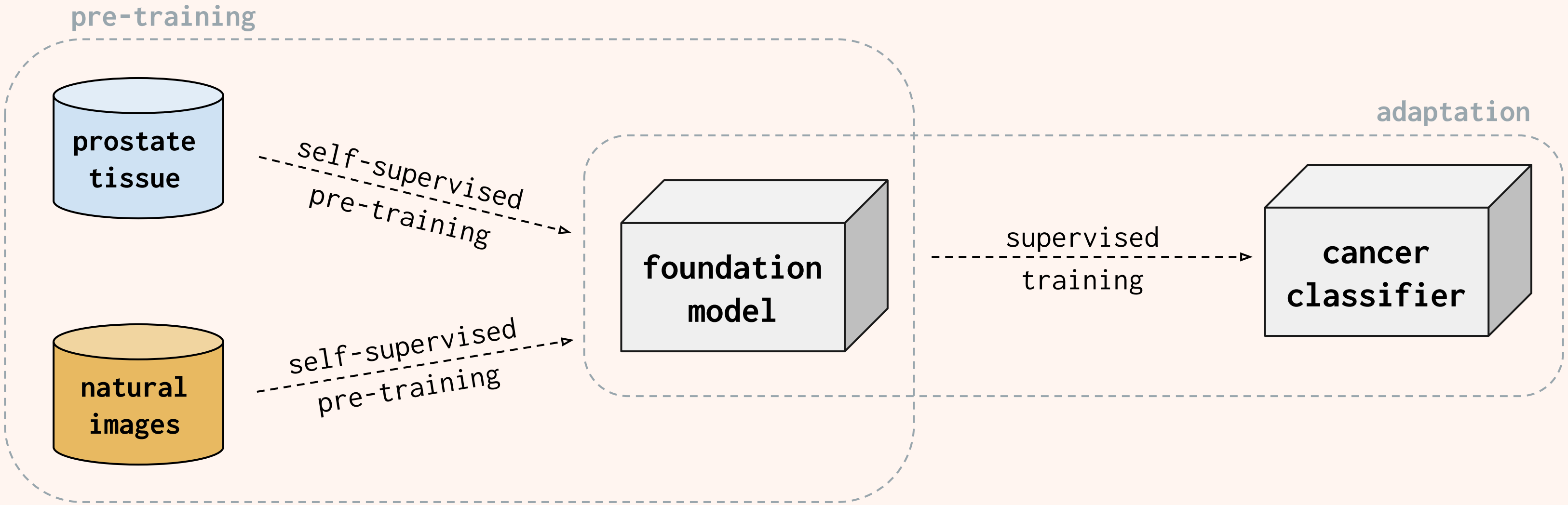
supervised

ResNet50

supervised

downstream task

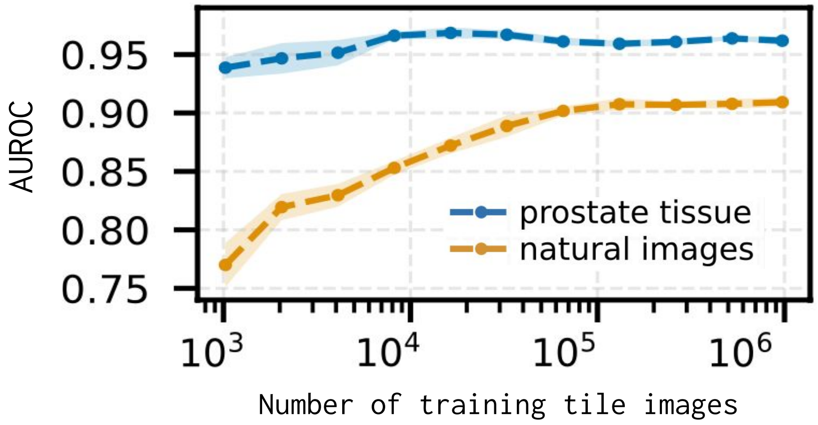
Does domain specific pre-training help?

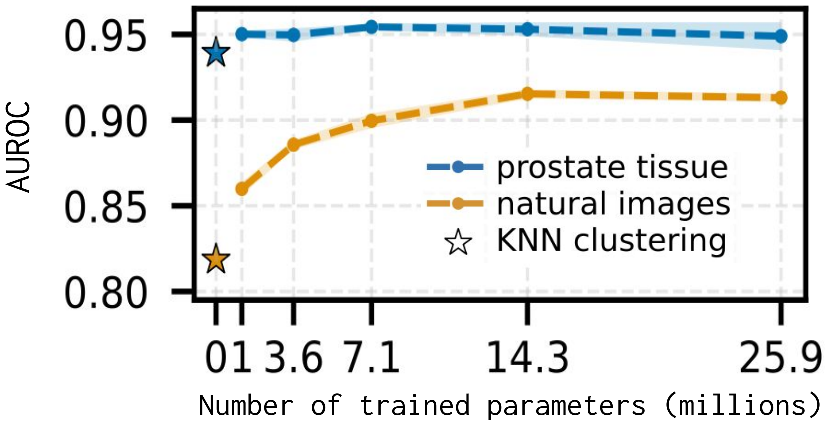


Does domain specific pre-training help?

common model sizes

small MLP	~1 million
ResNet-18	11 million
ResNet-50	24 million
ResNet-101	42 million
ViT-Base	86 million





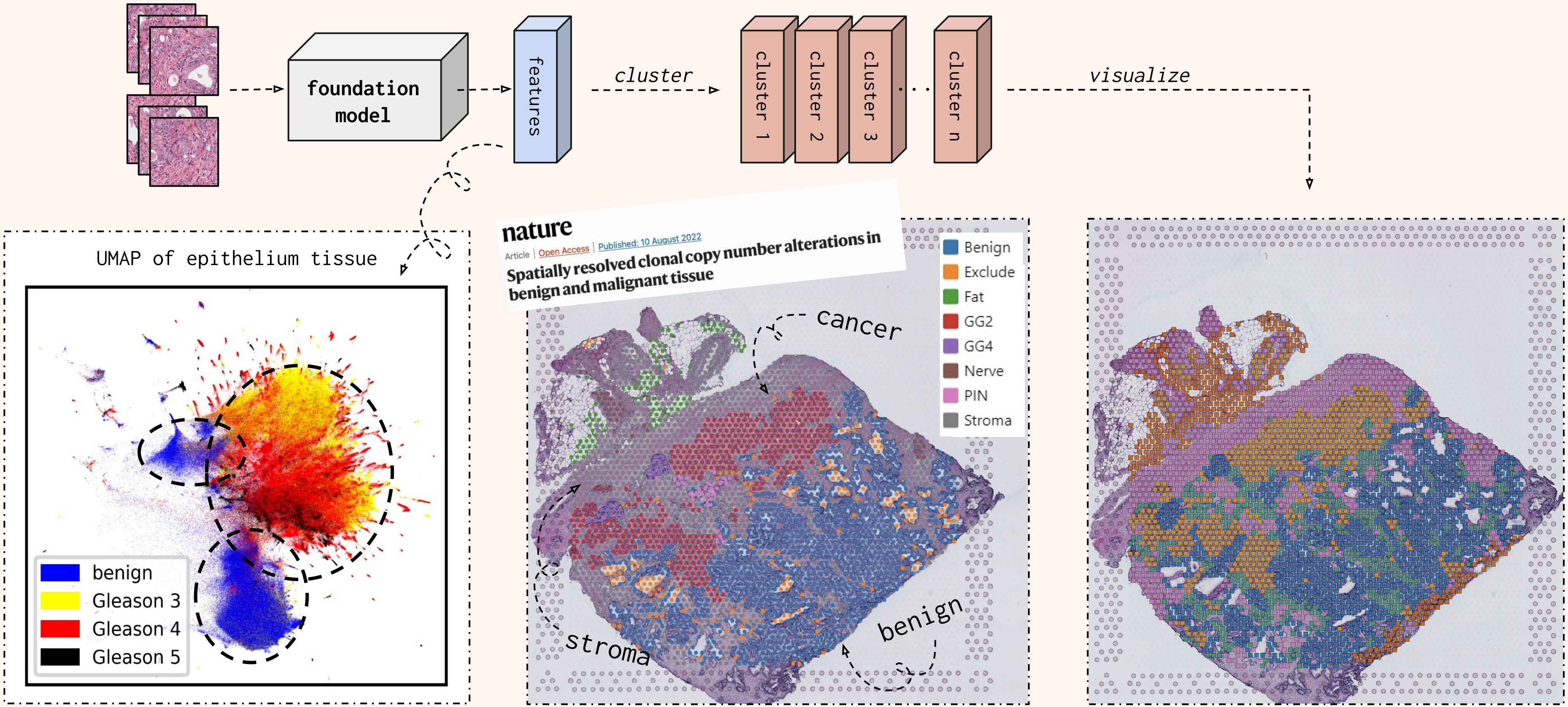
Fine-tuning requires ~1000 times less training data!!

Build competitive classifiers without any training!

Task 1: Limit the amount of training data

Task 2: Limit the number of trained parameters

Automatically annotate datasets by clustering...?

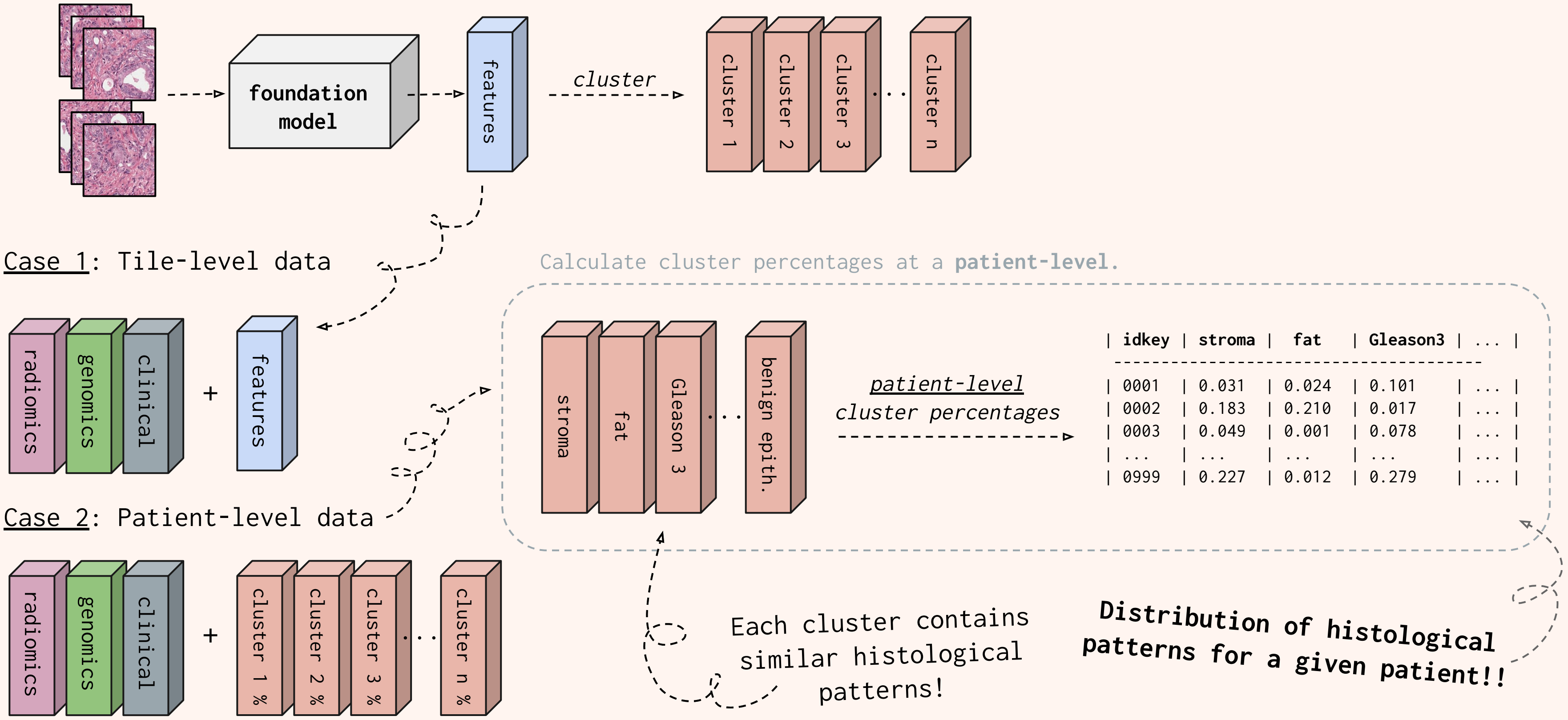


nature

Article | Open Access | Published: 10 August 2022

Spatially resolved clonal copy number alterations in benign and malignant tissue

Building multi-modal models...?



<u>Case 1</u>: Tile-level data

<u>Case 2</u>: Patient-level data

Calculate cluster percentages at a **patient-level**.

idkey	stroma	fat	Gleason3	...
0001	0.031	0.024	0.101	...
0002	0.183	0.210	0.017	...
0003	0.049	0.001	0.078	...
...	...	...	...	...
0999	0.227	0.012	0.279	...

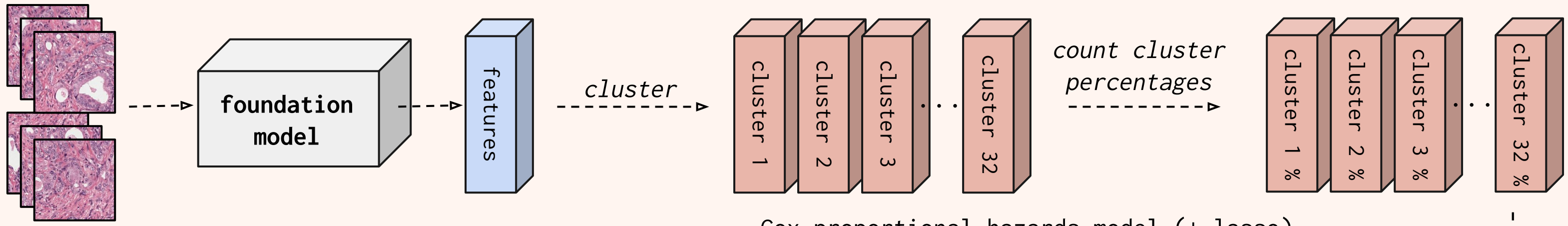
Each cluster contains similar histological patterns!

Distribution of histological patterns for a given patient!!

Prostate cancer specific death

432 patients with 19 year median follow-up

32 clusters!


foundation model

features

cluster

cluster 1

cluster 2

cluster 3

. . .

cluster 32

count cluster percentages

cluster 1 %

cluster 2 %

cluster 3 %

. . .

cluster 32 %

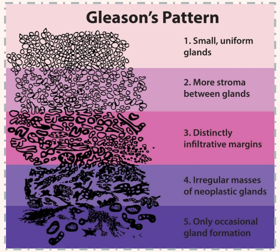
The CAPRA-S score: a straightforward tool for improved prediction of outcomes after radical prostatectomy

Matthew R. Cooperberg, MD, MPH,[1]† Joan F. Hilton, MPH, ScD,[1][2] and Peter R. Carroll, MD, MPH[1]

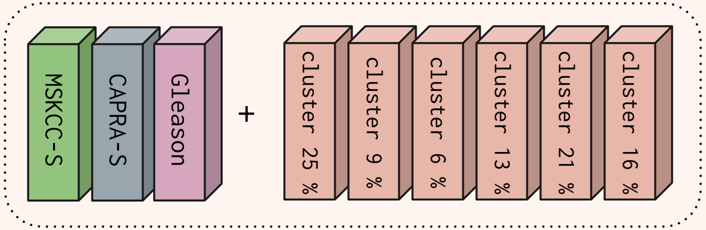
▸ Author information ▸ Copyright and License information Disclaimer

Risk of Dying of Prostate Cancer in Men With a Rising PSA After Radical Prostatectomy

This nomogram can be used by patients to estimate the risk of dying of prostate cancer if their cancer recurs, signaled by a rising PSA, after radical prostatectomy. The nomogram predicts the likelihood, in a man initially treated with surgery, that he will die of prostate cancer five, ten, and 15 years from the time his PSA begins to rise.


Gleason's Pattern
1. Small, uniform glands
2. More stroma between glands
3. Distinctly infiltrative margins
4. Irregular masses of neoplastic glands
5. Only occasional gland formation

Cox proportional hazards model (+ lasso)


MSKCC-S

CAPRA-S

Gleason

+

cluster 25 %

cluster 9 %

cluster 6 %

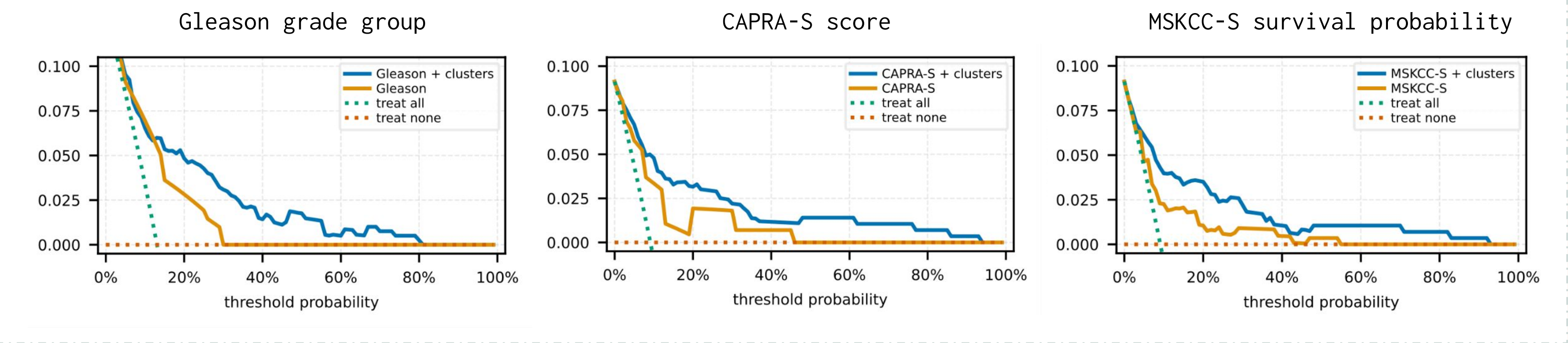
cluster 13 %

cluster 21 %

cluster 16 %

drop clusters based on parameter importance

Significantly higher net benefit!


Gleason grade group

CAPRA-S score

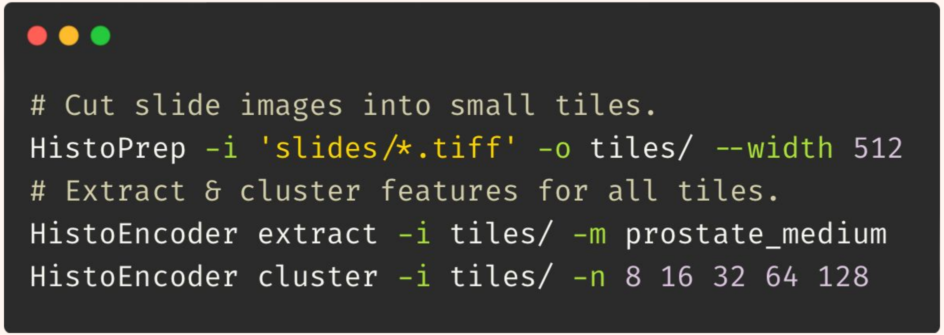
MSKCC-S survival probability

Summary

Foundation model fine-tuning requires no training or 1000x less training data.

Images with similar histological patterns produce similar embedded features.

Pre-annotate whole datasets!!



```
# Cut slide images into small tiles.
HistoPrep -i 'slides/*.tiff' -o tiles/ --width 512
# Extract & cluster features for all tiles.
HistoEncoder extract -i tiles/ -m prostate_medium
HistoEncoder cluster -i tiles/ -n 8 16 32 64 128
```

Combine histomics with other data modalities!!!

https://github.com/jopo666/HistoEncoder