Learning Image Representations Without Manual Annotations and Scientific Applications

Piotr Bojanowski, FAIR, Meta

Poznań, Poland

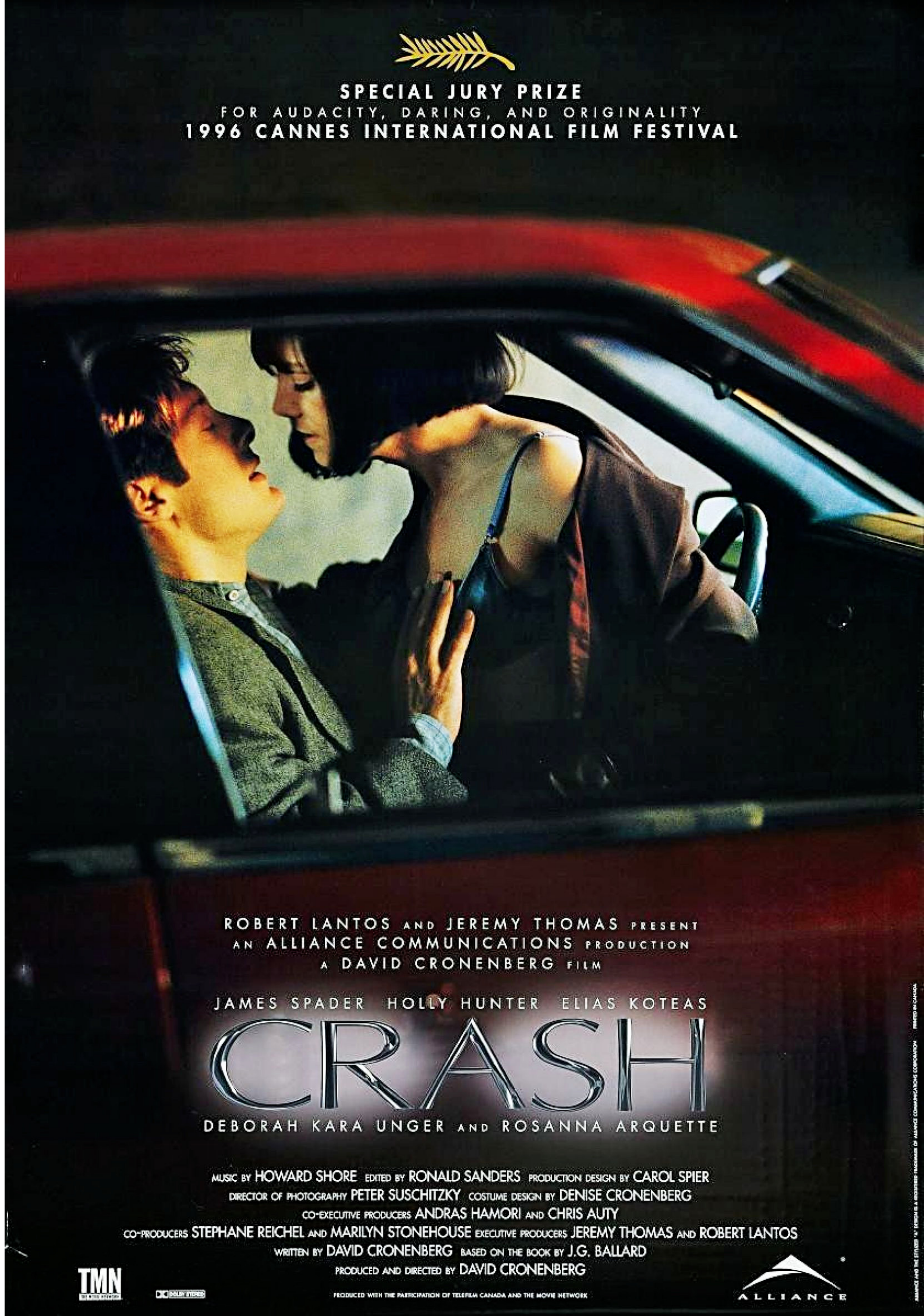
April 5th 2024

Meta AI

Manual annotations...

- Hollywood 2
  - Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. "Actions in context." In *CVPR 2009.*
  - 810 + 884 videos
  - 12 actions
  - 69 Hollywood movies

- Hollywood 3 ?
  - Annotate all movies exhaustively
  - In charge of one of the movies

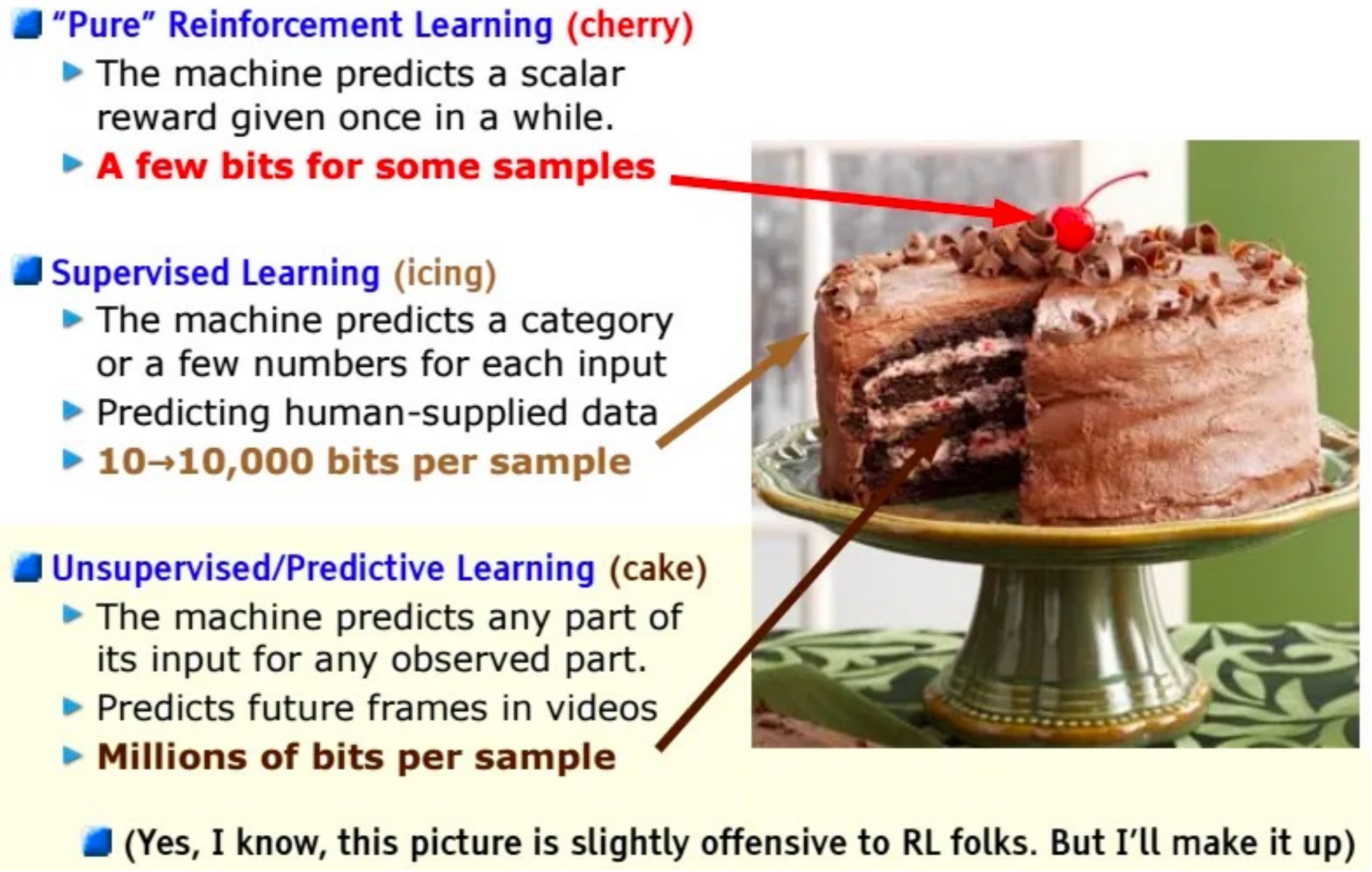


Scarcity of Manual Annotations

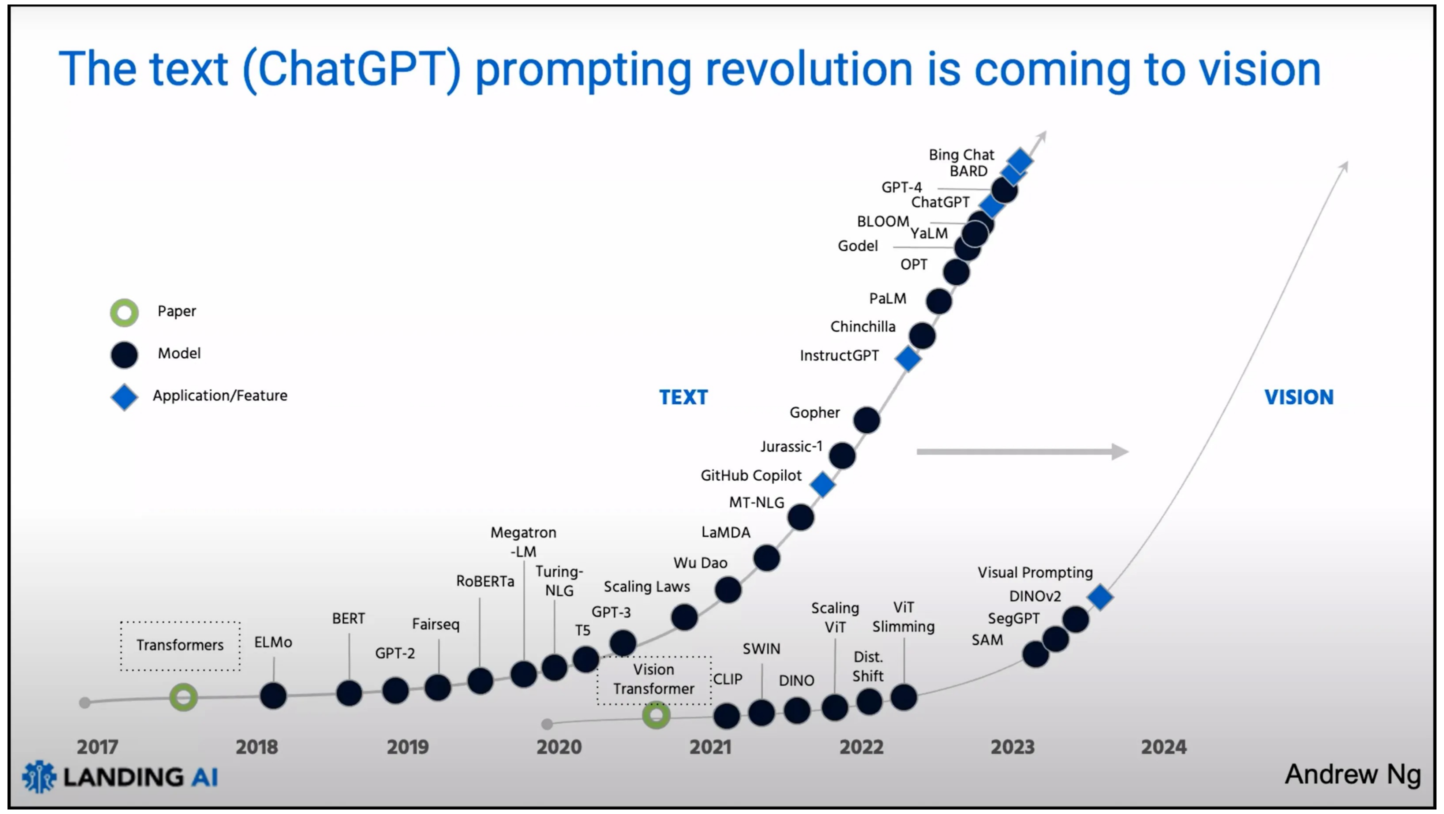
- Annotations are expensive (if high quality)
- Are ambiguous
- Class definitions are not static
- Intractable with increasing complexity of the task

Baking the Cake

- Supervised Learning is needed!
- Unsupervised learning should do the heavy lifting
- Modern success of LLMs follows this exact recipe…
- Is the vision cake ready?



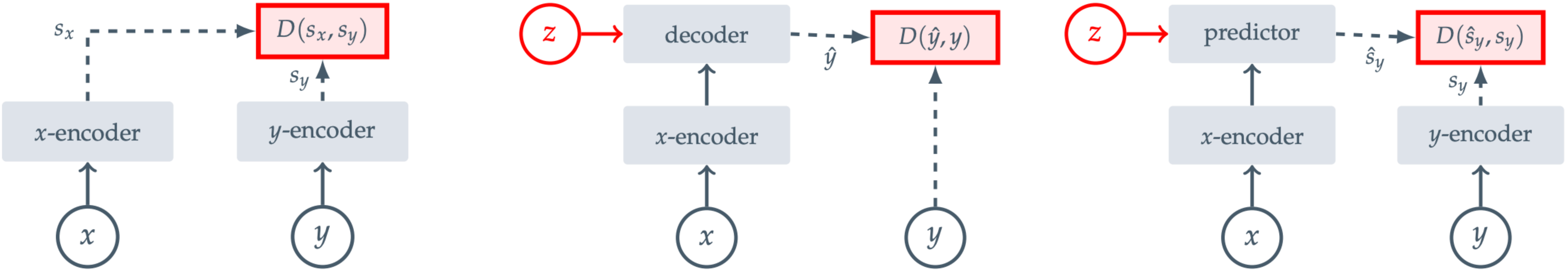
Yann LeCun, ~2016



Andrew Ng, 2023

Meta AI

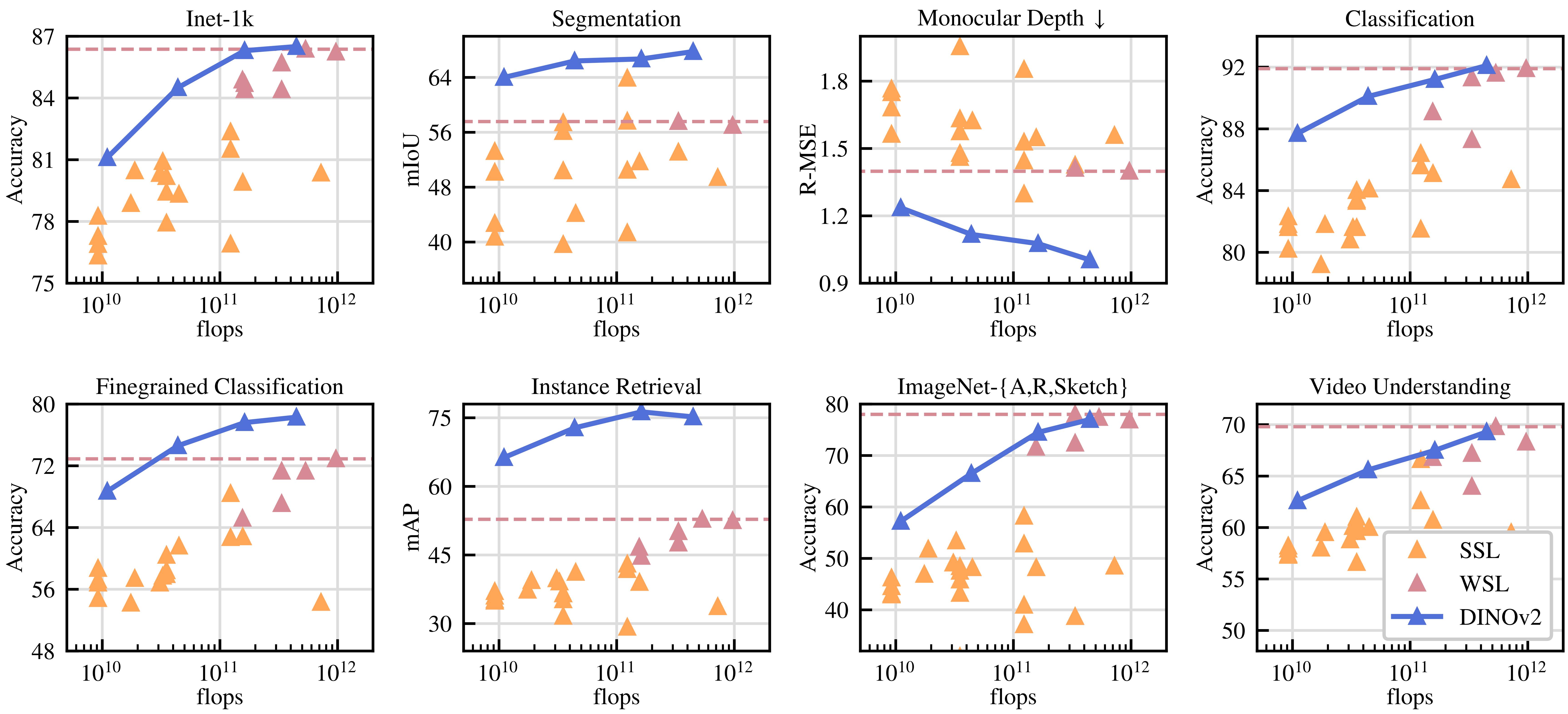
Core principle



(a) **Joint-Embedding Architecture** (b) **Generative Architecture** (c) **Joint-Embedding Predictive Architecture**

Assran, Mahmoud, et al. "Self-supervised learning from images with a joint-embedding predictive architecture." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.

SplitBrainAE, NAT, CPC, DeepCluster, CPCv2, SELA, MoCo, PIRL, SimCLR, SwAV, MoCov2, PCL, BYOL, Barlow Twins, DINO, SimCLRv2, NN-CLR, VicReg, BEiT, MAE...


Inet-1k
Segmentation
Monocular Depth ↓
Classification
Finegrained Classification
Instance Retrieval
ImageNet-{A,R,Sketch}
Video Understanding
Accuracy
mIoU
R-MSE
mAP
flops
SSL
WSL
DINOv2

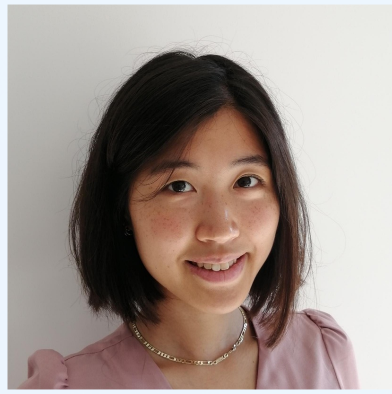
Meta AI

Scientific problems : a perfect application


IMAGINED WITH AI

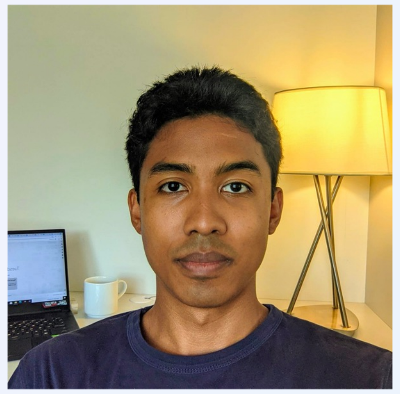








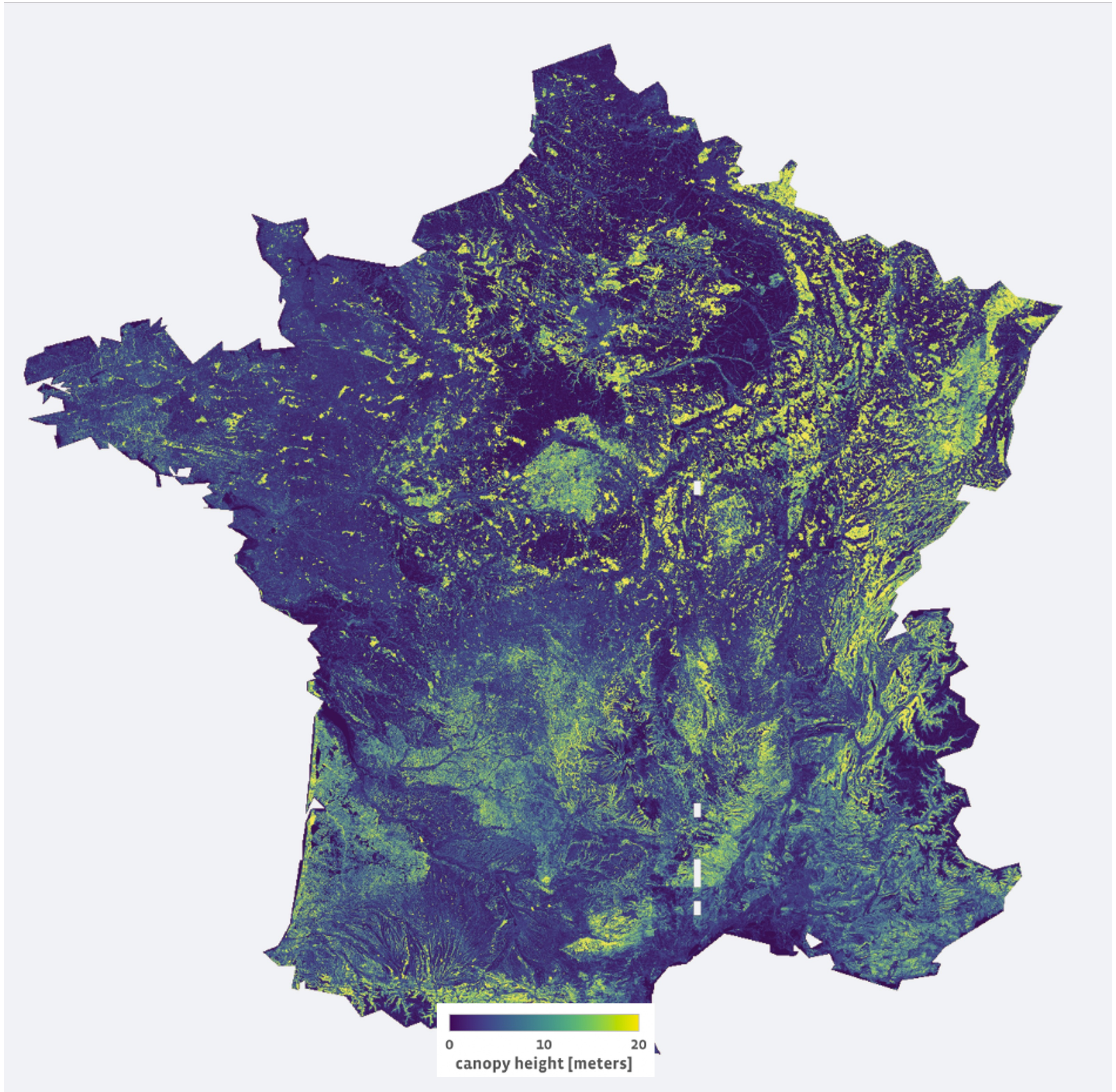


Meta AI

Applications of SSL

Meta AI

High-Resolution Canopy Height Estimation


0
10
20
canopy height [meters]

Physical Modelling @ Meta

Jamie Tolan

Ben Nosarzewski

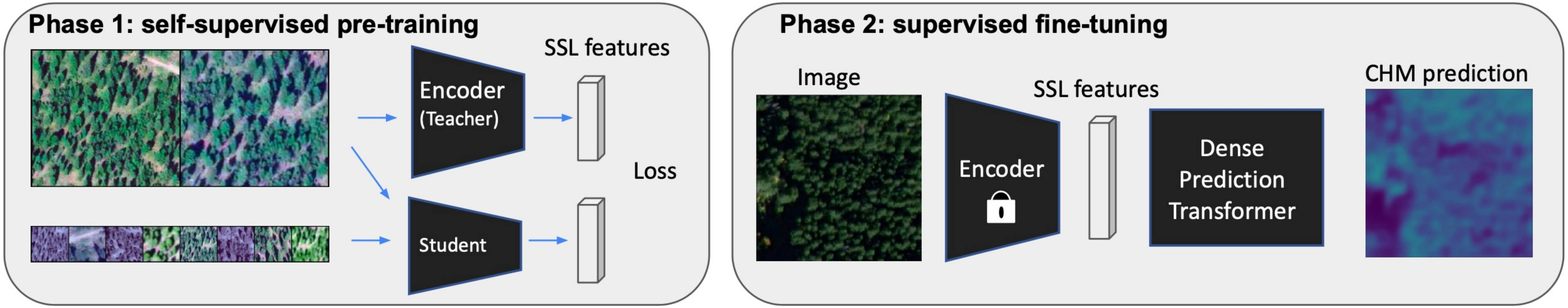
Tobias Tiecke

World Ressource Institute

John Brandt

Justine Spore

	Coverage	Type	Channels	Beam
MAXAR	Global	Satellite	RGB	0.5 m
GEDI	Near-Global	Satellite	RGB + LIDAR	25 m
NEON	Small	Airborne	RGB + LIDAR	1m

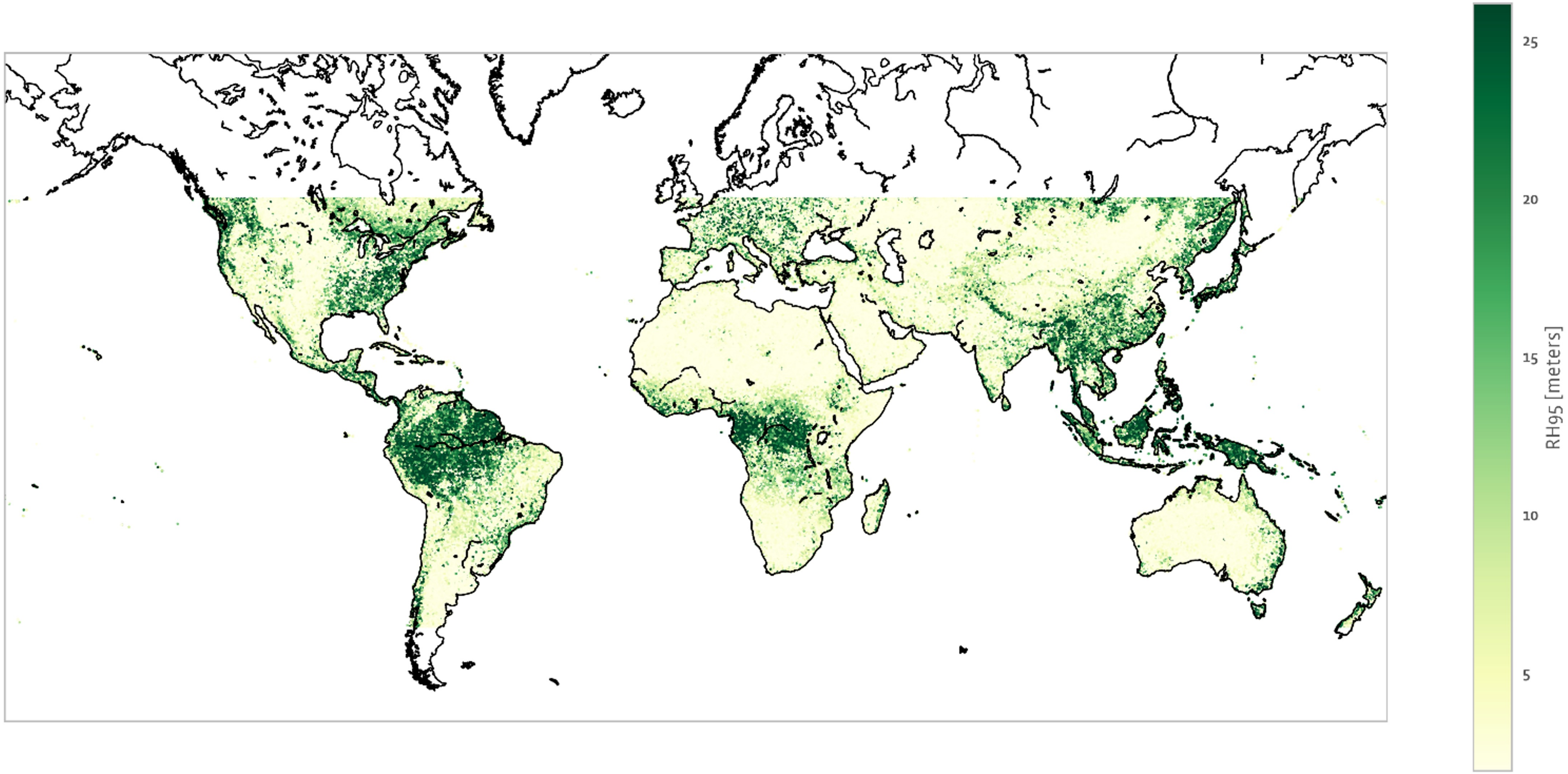


Worldwide Satellite orthorectified map

- 0.5 meter RGB stitched images
- Changes in terrain slope, view angle, sun angle, season, etc.

Datasets (Ground truth 1)

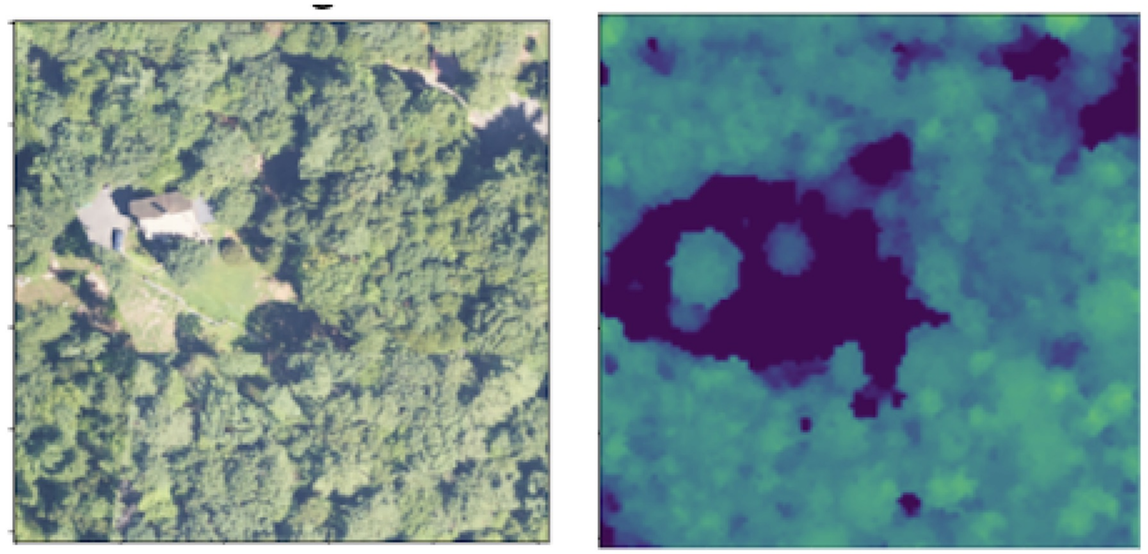
GEDI: Space based lidar: Global (4% area coverage), 25 meter beamwidth, 1064nm, 1.28B data points:



Neon dataset (Ground truth 2)

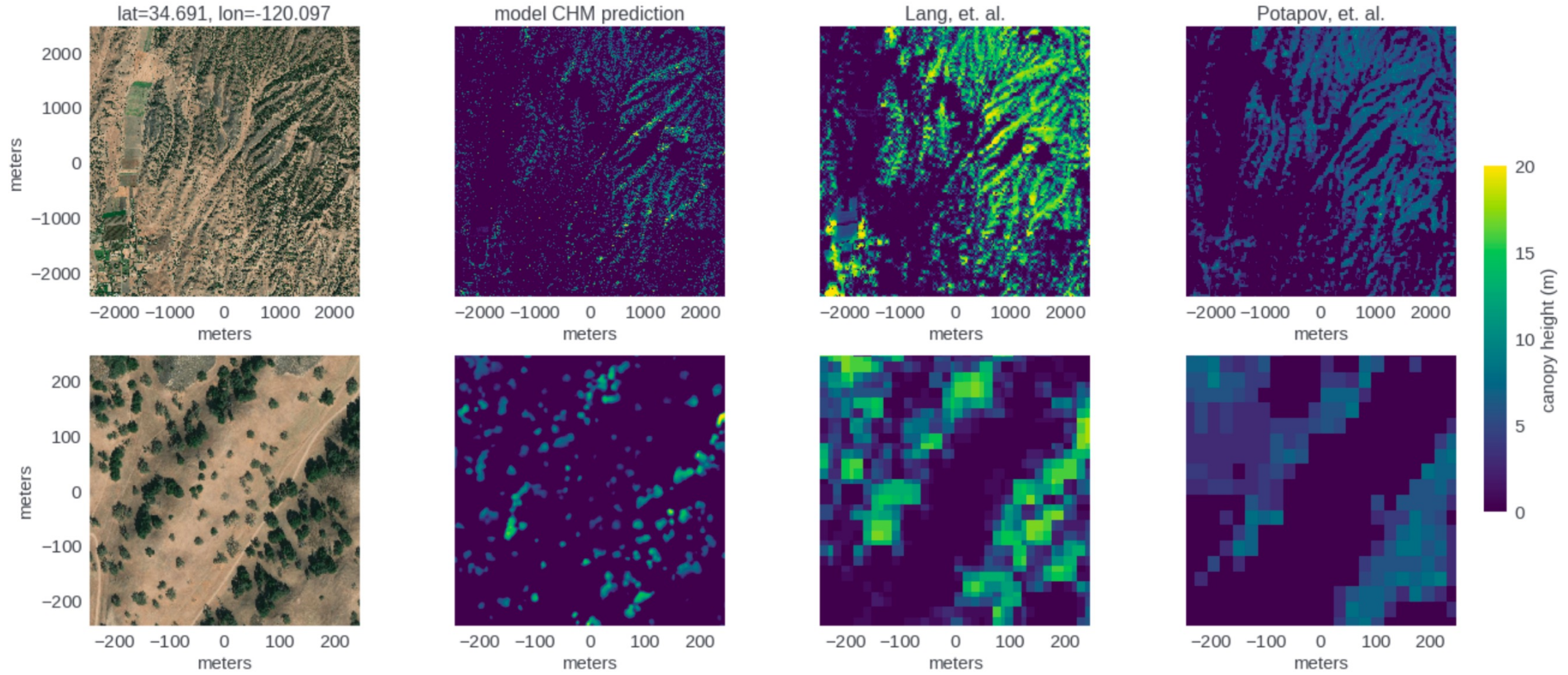
- Aerial Lidar canopy height maps
- 38 sites across North America
- About 5000 training image / lidar pairs of resolution 2200 pixels
- Dataset setup:
  - 80%-10%-10% train/val/test split
  - 5 Test sites completely independent of training sites

2



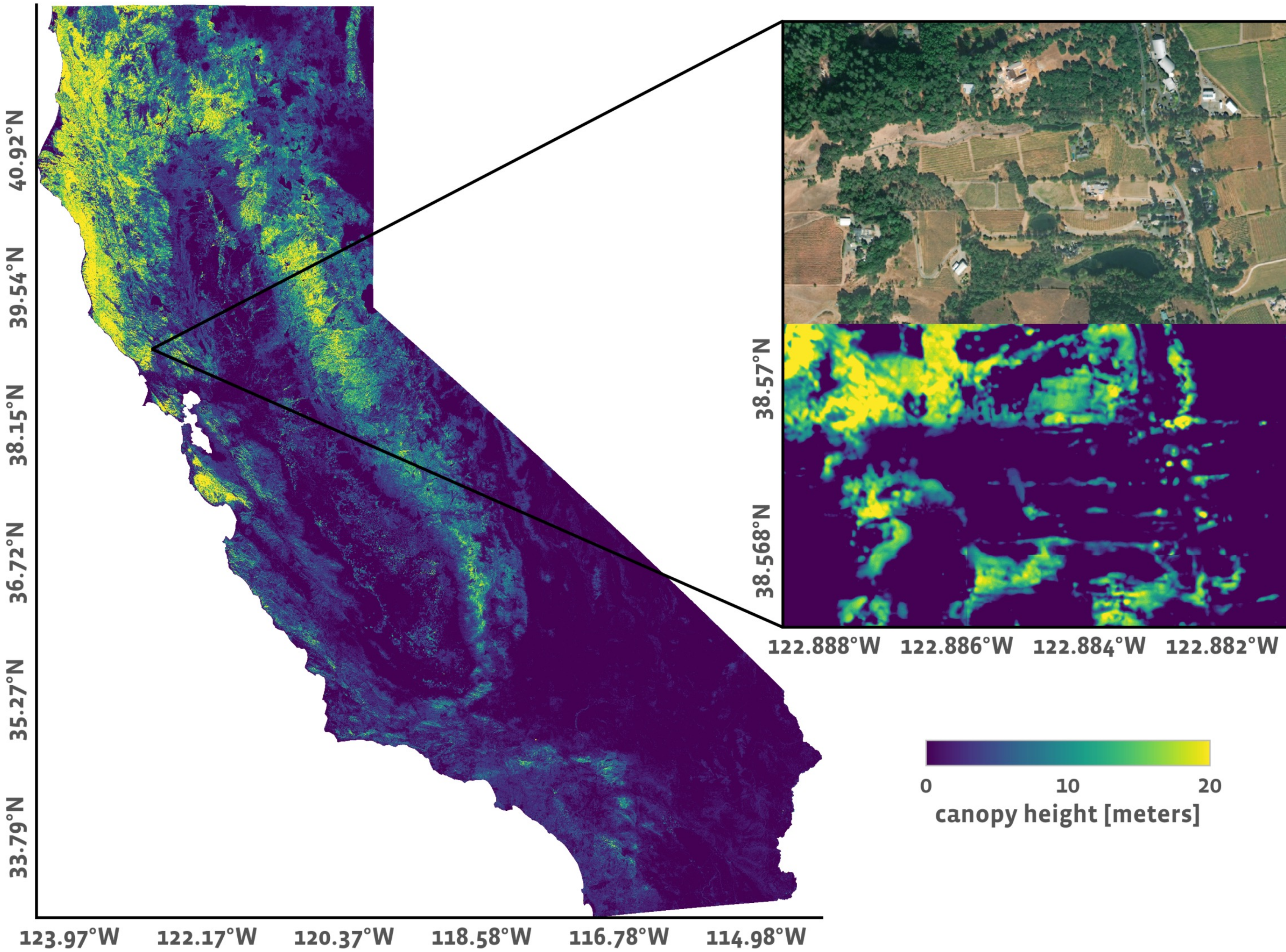
In addition to the Neon dataset, we used some lidar data in Sao Paulo, California and France for evaluation.

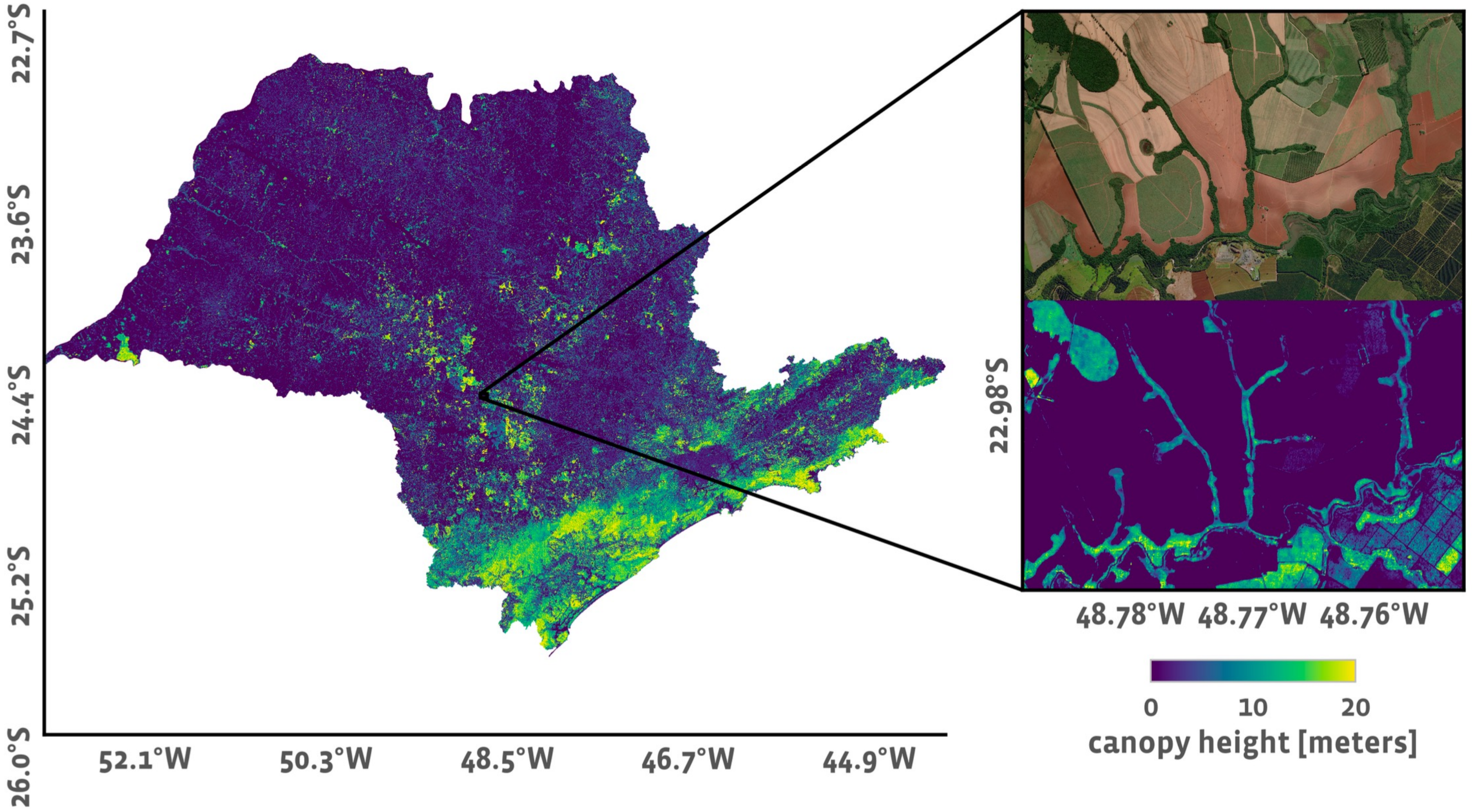
Canopy Height Estimation

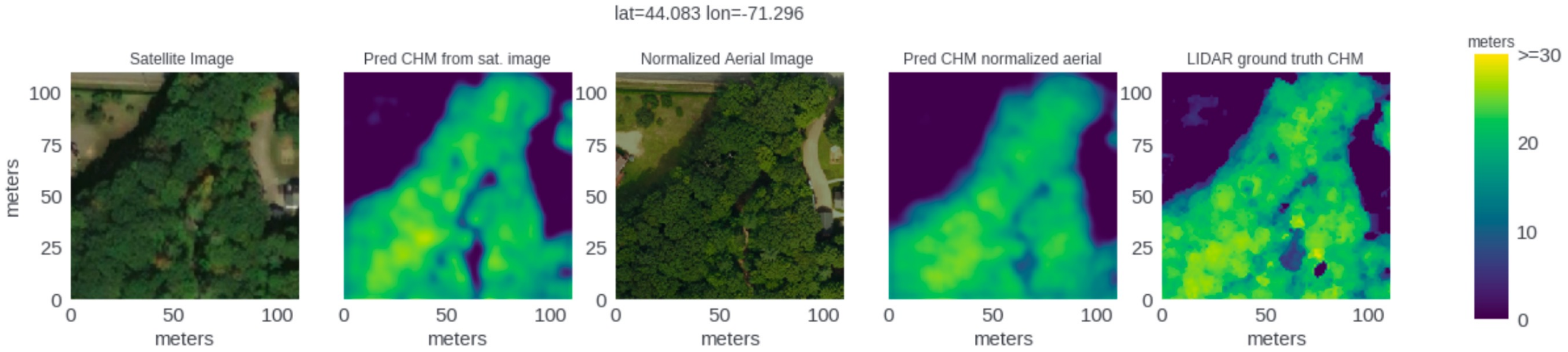


Meta AI

Canopy Height Estimation






lat=44.083 lon=-71.296
Satellite Image
Pred CHM from sat. image
Normalized Aerial Image
Pred CHM normalized aerial
LIDAR ground truth CHM
meters
100
75
50
25
0
0
50
100
meters
>=30
20
10
0

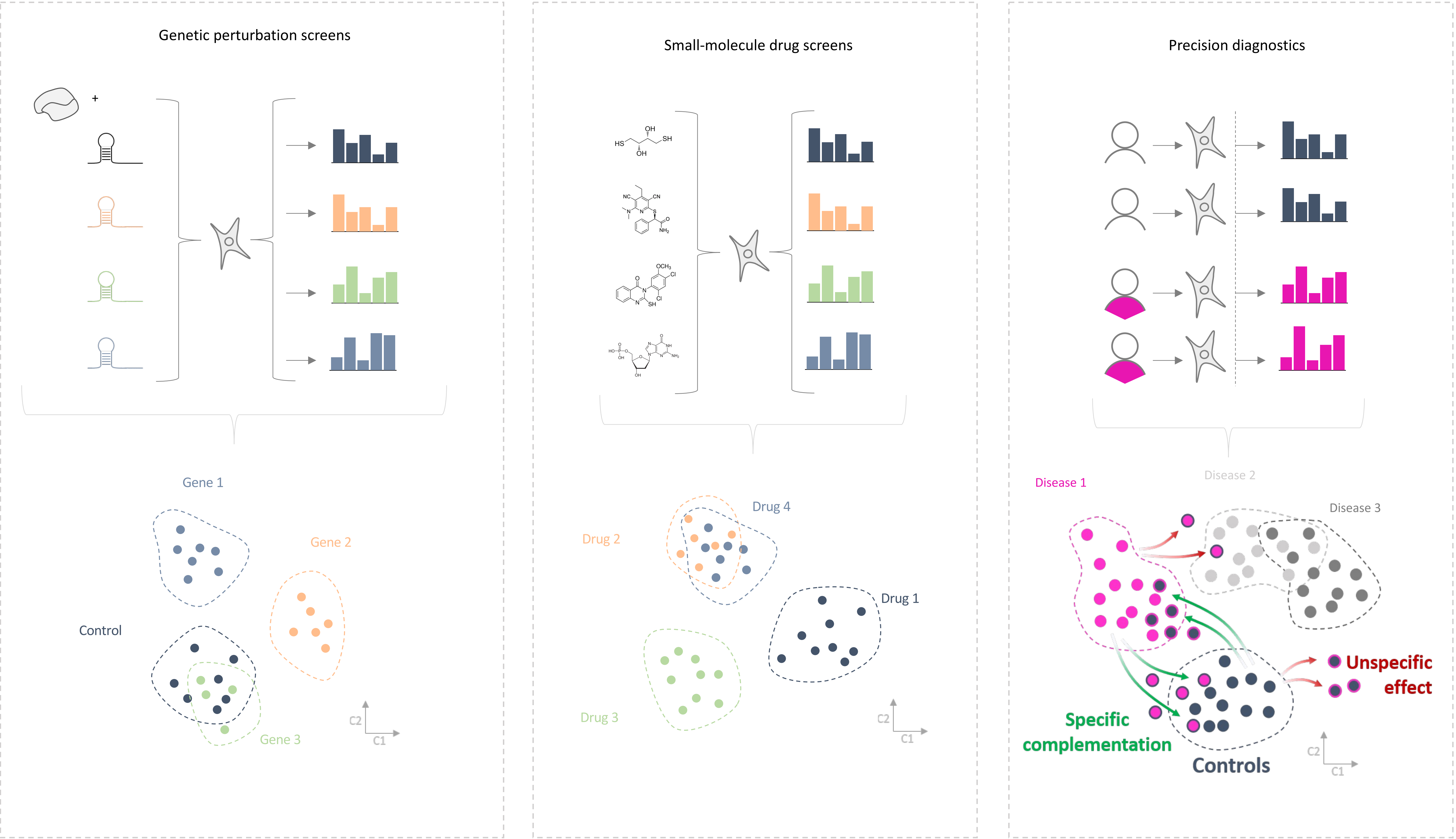
Single-Cell Microscopy

Juan C. Caicedo
University of Wisconsin-Madison /
Broad Institute of MIT

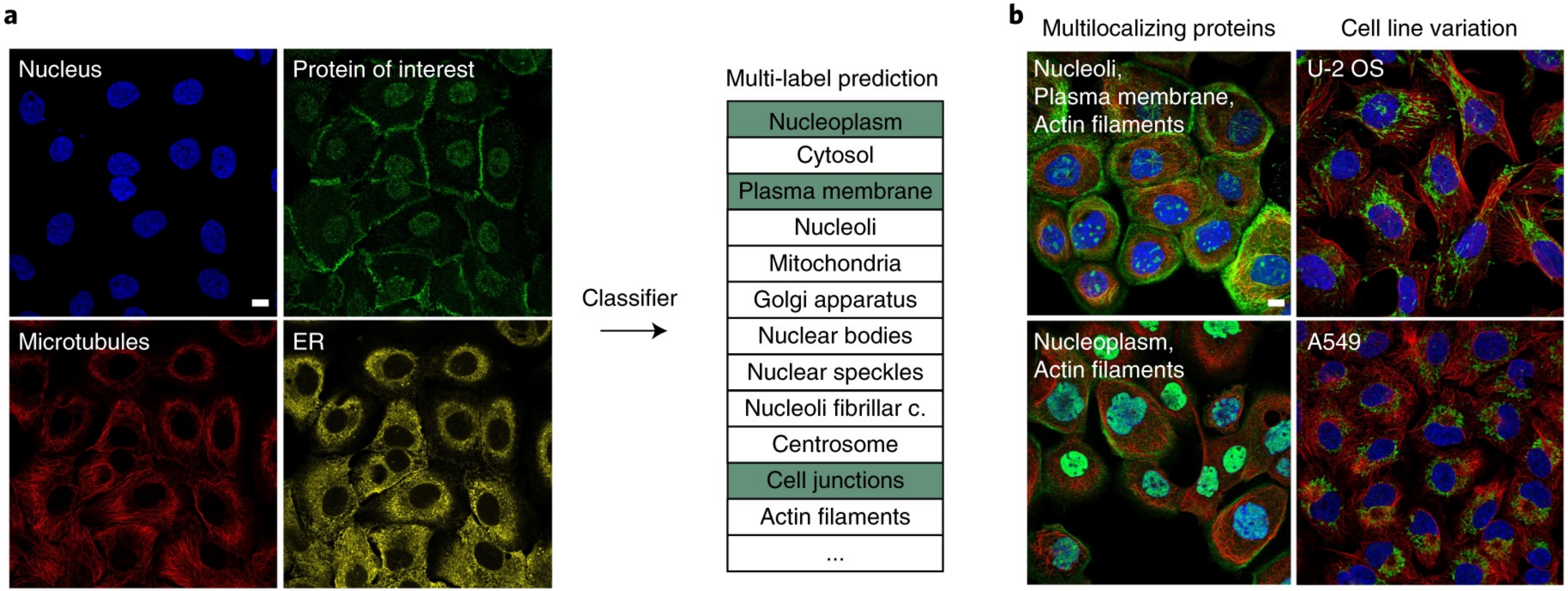
Wolfgang Pernice
Columbia University Irving Medical Center

Michael Doron
Q.Ai / Broad Institute of MIT

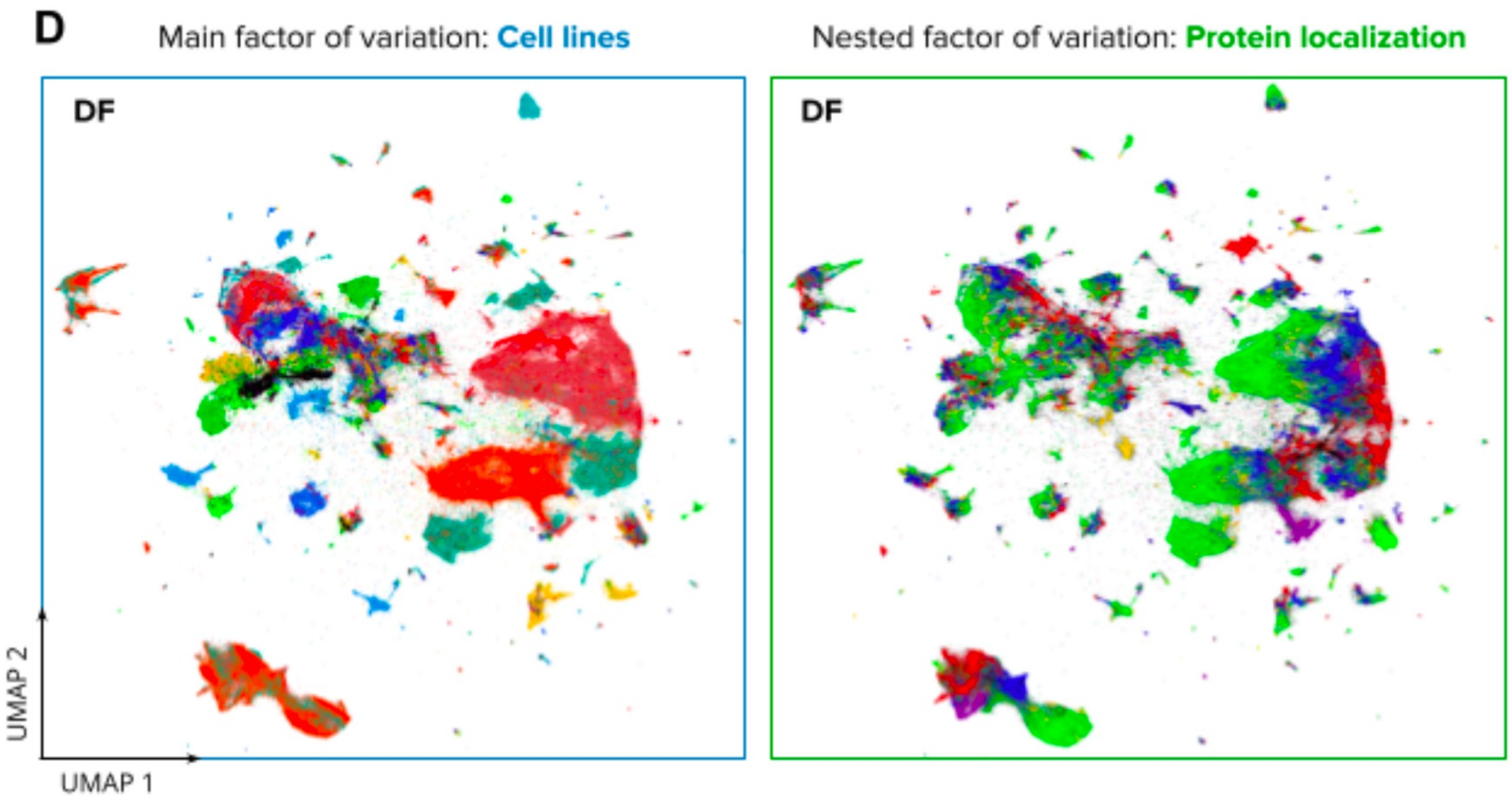
Single-Cell Microscopy

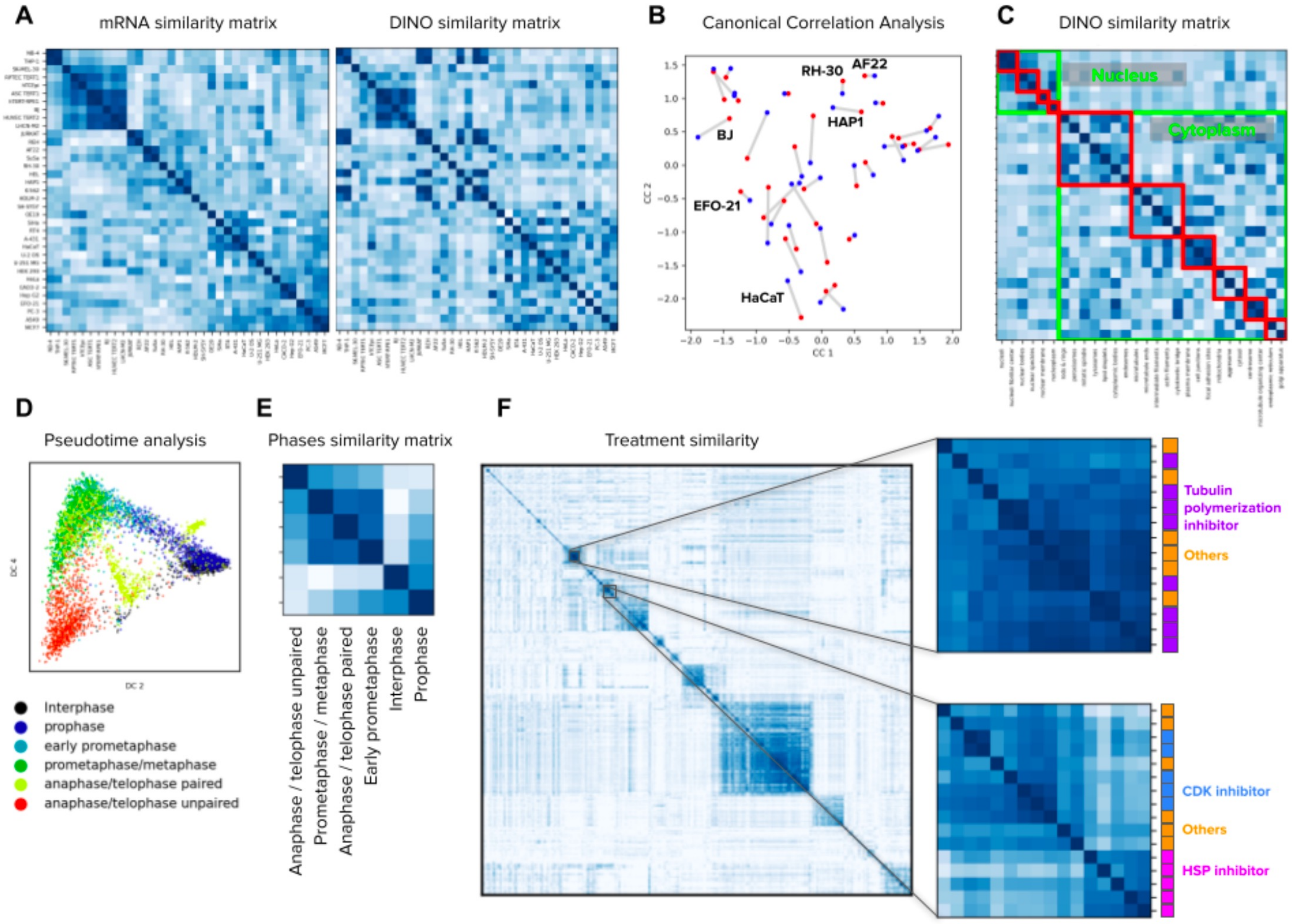


Meta AI



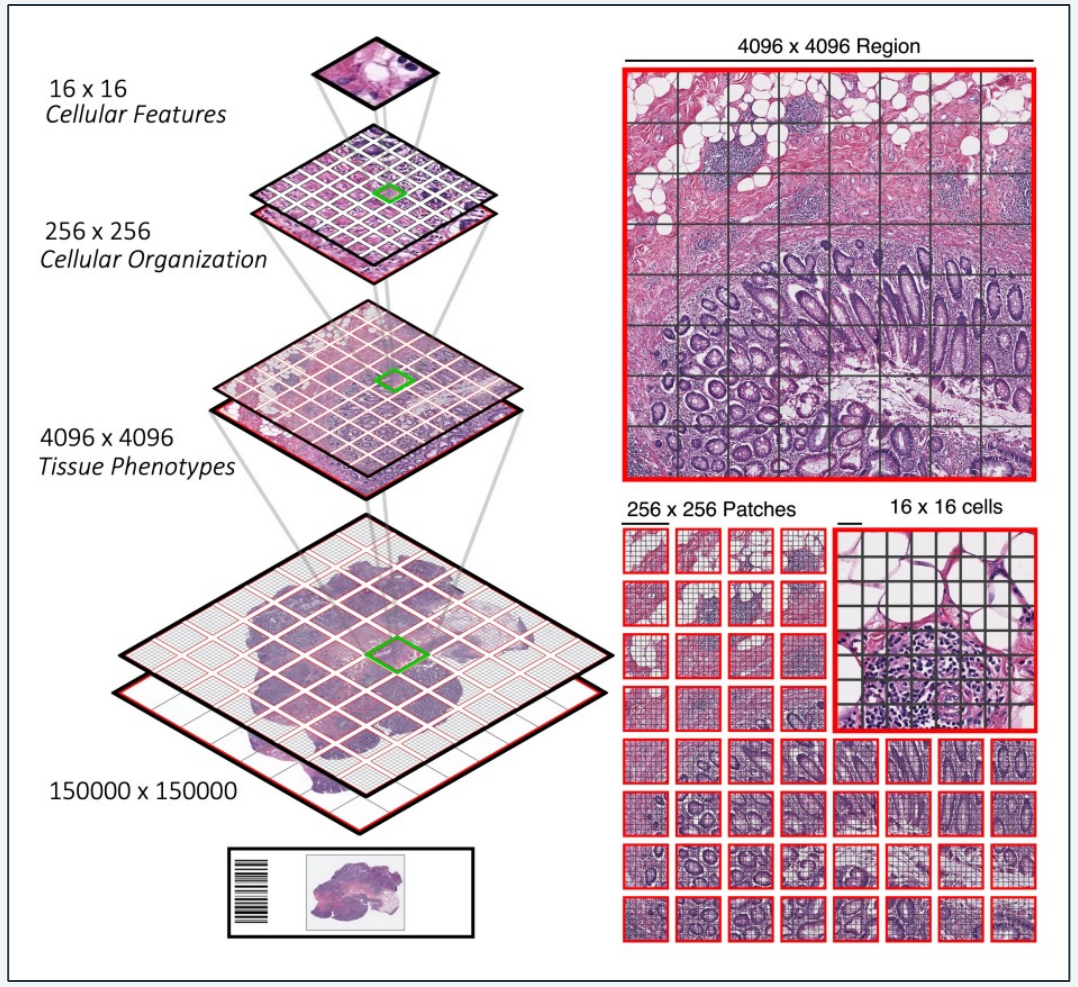
Uhlén, Mathias, et al. "Tissue-based map of the human proteome." *Science* 347.6220 (2015): 1260419.


D
Main factor of variation: Cell lines
Nested factor of variation: Protein localization
DF
DF
UMAP 2
UMAP 1

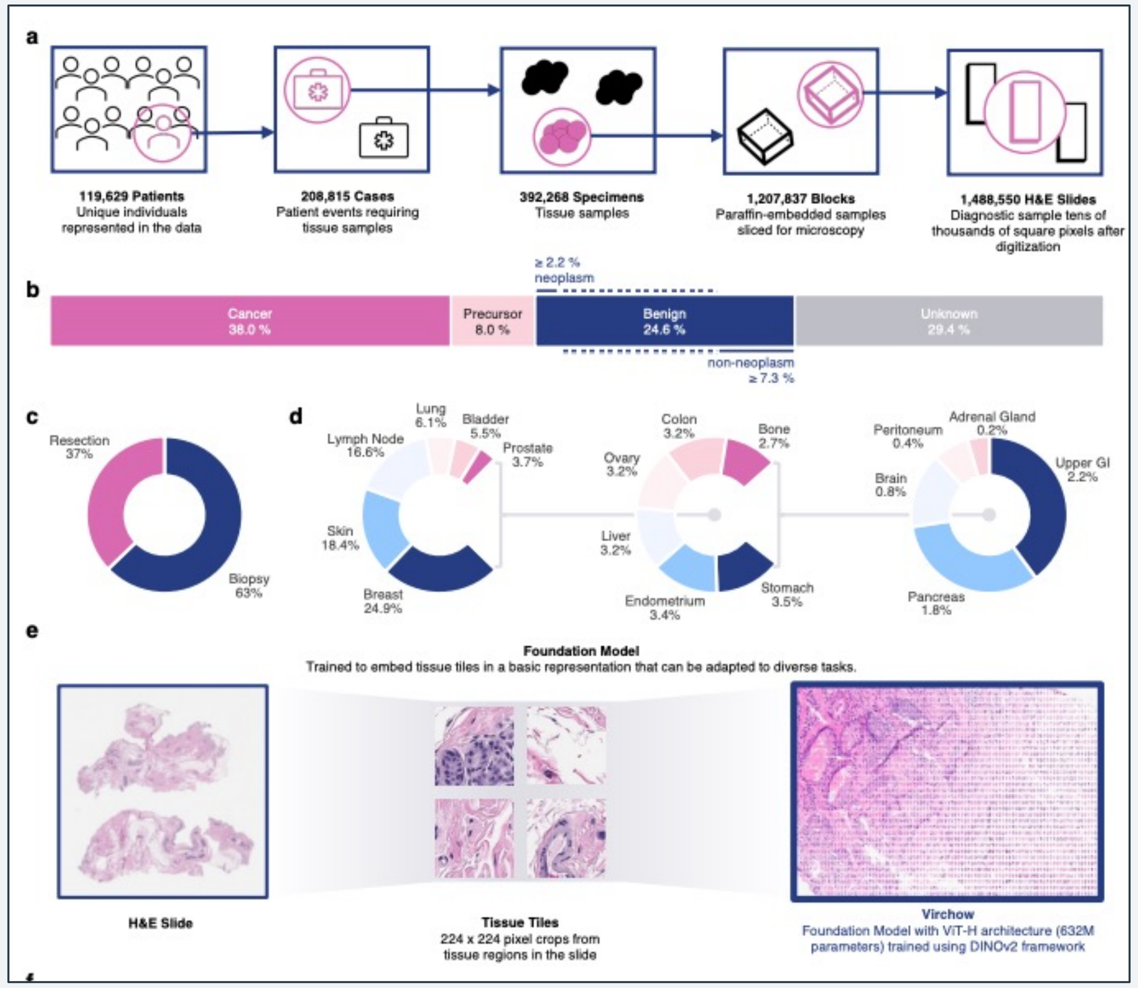

A
mRNA similarity matrix
DINO similarity matrix
B
Canonical Correlation Analysis
RH-30
AF22
HAP1
BJ
EFO-21
HaCaT
CC 1
CC 2
C
DINO similarity matrix
Nucleus
Cytoplasm
D
Pseudotime analysis
DC 2
DC 4
Interphase
prophase
early prometaphase
prometaphase/metaphase
anaphase/telophase paired
anaphase/telophase unpaired
E
Phases similarity matrix
Anaphase / telophase unpaired
Prometaphase / metaphase
Anaphase / telophase paired
Early prometaphase
Interphase
Prophase
F
Treatment similarity
Tubulin polymerization inhibitor
Others
CDK inhibitor
Others
HSP inhibitor

Conclusion and Future Work

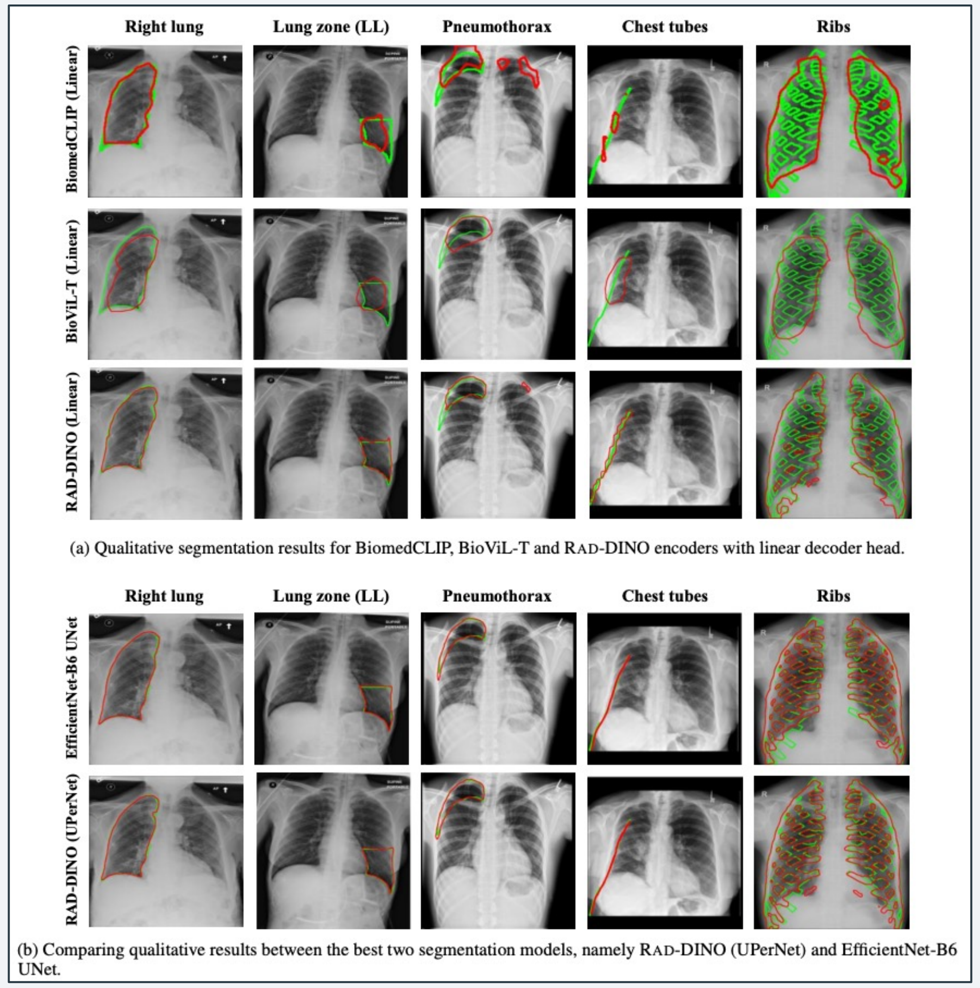
DINOv2 : a foundation model factory



Chen, Richard J., et al. "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.



Vorontsov, Eugene, et al. "Virchow: A million-slide digital pathology foundation model." arXiv preprint arXiv:2309.07778 (2023).

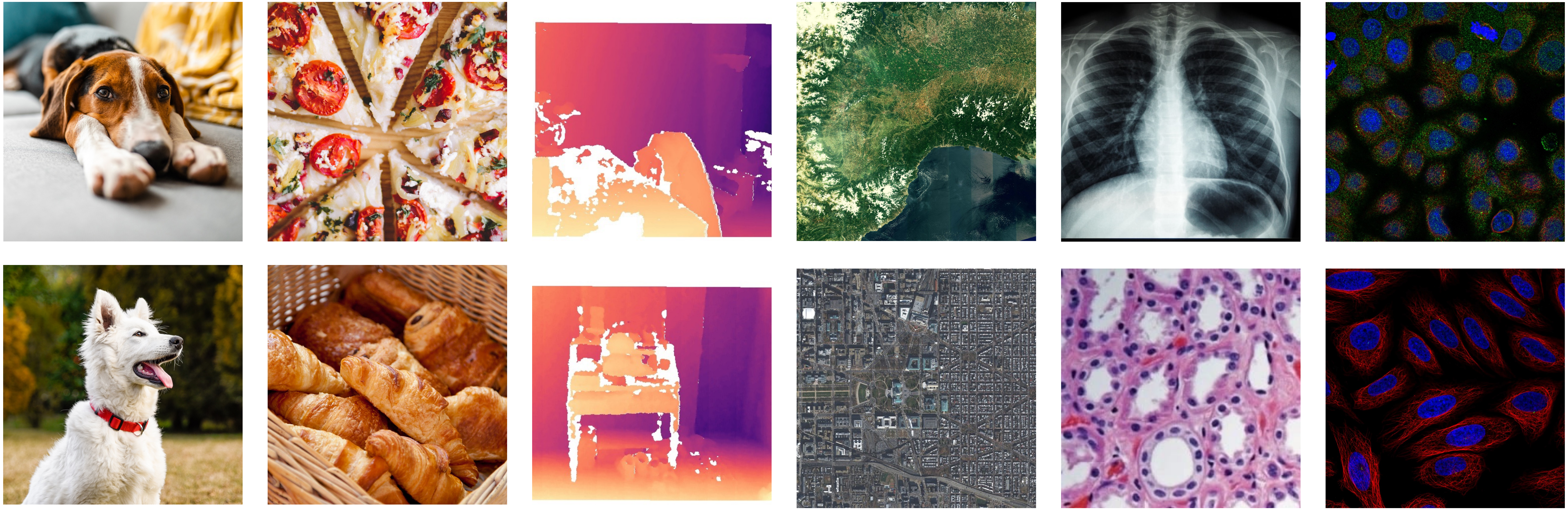


(a) Qualitative segmentation results for BiomedCLIP, BioViL-T and RAD-DINO encoders with linear decoder head.

(b) Comparing qualitative results between the best two segmentation models, namely RAD-DINO (UPerNet) and EfficientNet-B6 UNet.

Pérez-García, Fernando, et al. "RAD-DINO: Exploring Scalable Medical Image Encoders Beyond Text Supervision." arXiv preprint arXiv:2401.10815 (2024).

Learning Universal Visual Representations



Meta AI