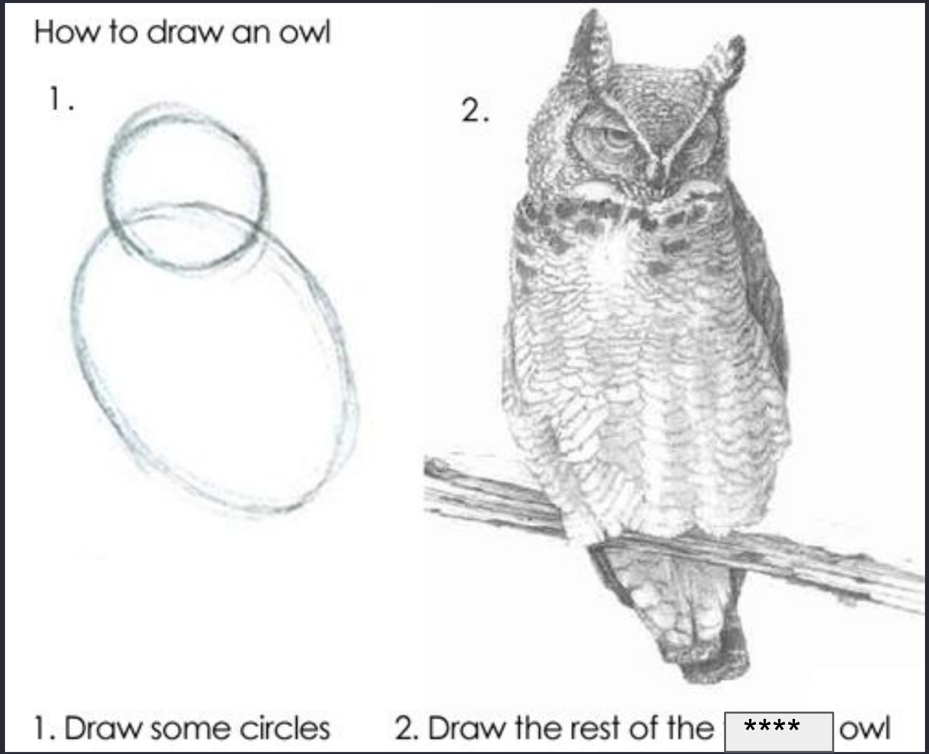
Before I train the model… :
Three adventures in developing ML systems for coastal geomorphology

Evan B. Goldstein (he/him)

Dept. Geography, Environment, and Sustainability
UNC Greensboro

ebgoldst@uncg.edu



Acknowledgements:

Eli Lazarus & Hannah Williams (Southampton)
Dan Buscombe & Sharon Fitzpatrick (USGS contr.)
Jacob Stasiewicz (UNCW)

(Many other colleagues)
My Family

🍞🥖: USGS, NSF, Leverhulme Trust, Microsoft AI 4 Earth, Google TF, NASEM GRP, DoD DARPA, NCDOT

MOTIVATION

- ML is 🔥 right now

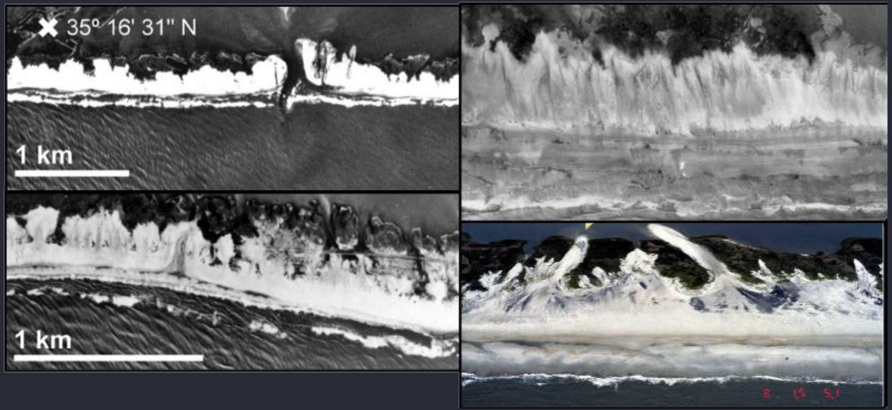
MOTIVATION

- ML is 🔥 right now

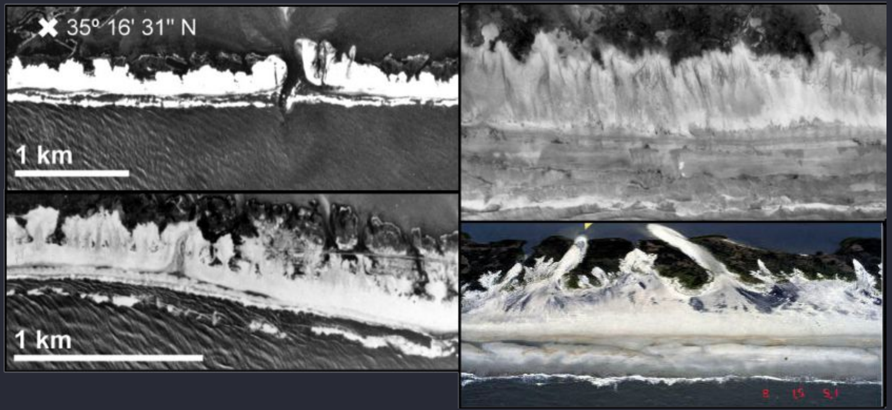
"Everyone wants to do the model work, not the data work":
Data Cascades in High-Stakes AI

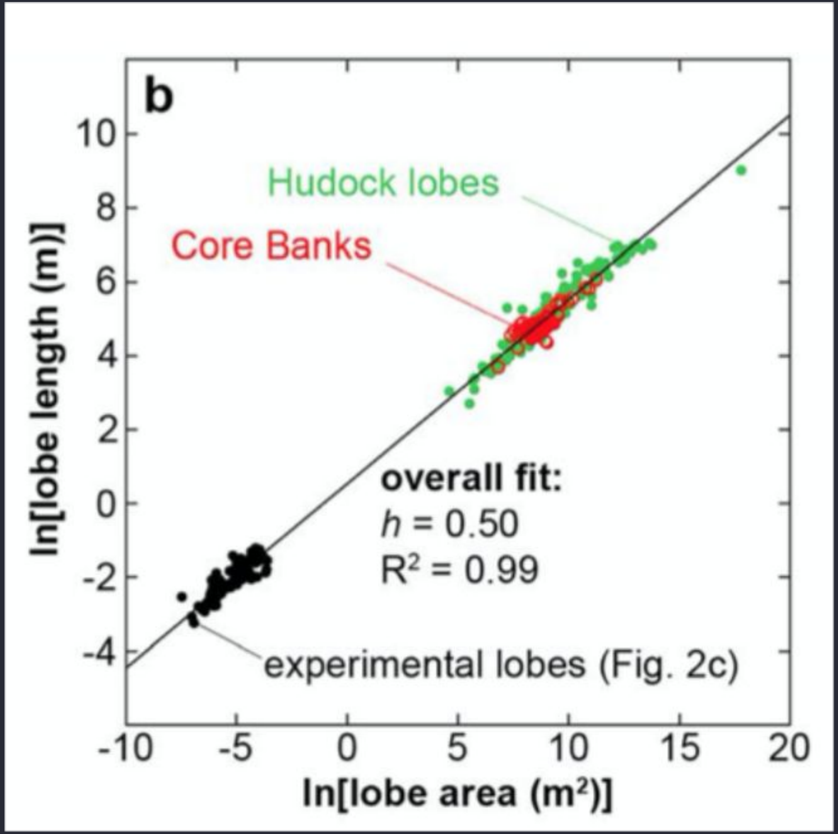
Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo
[nithyasamba,kapania,hhighfill,dakrong,pkp,loraa]@google.com
Google Research
Mountain View, CA

Ex. 1: Labeling images to study barrier island overwash (process) & washover (deposit)



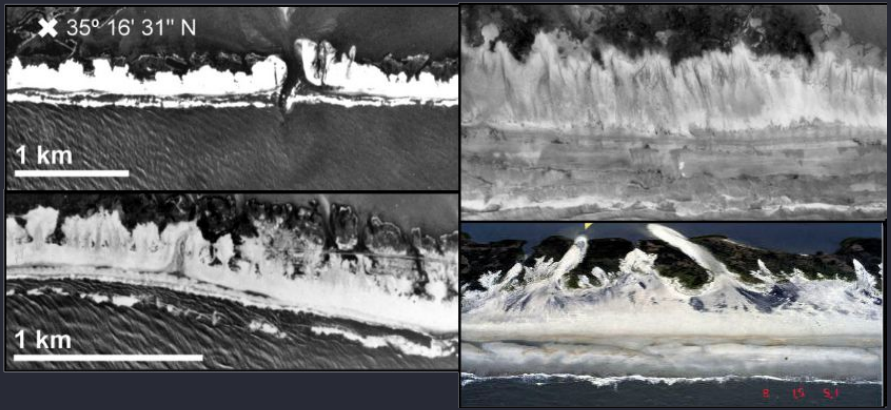
Ex. 1: Labeling images to study barrier island overwash (process) & washover (deposit)

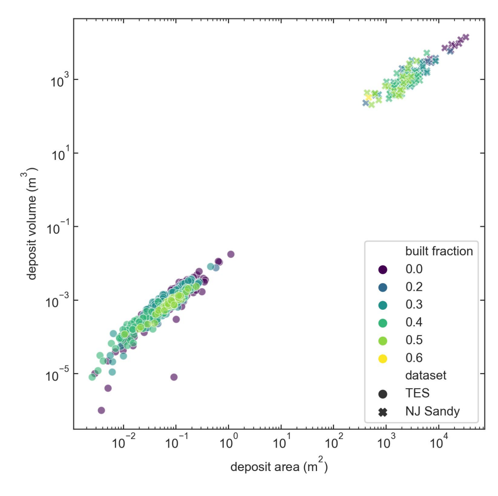


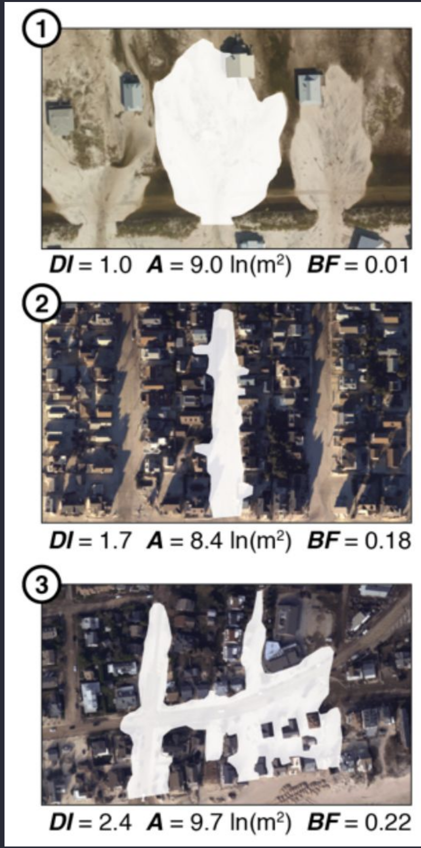


Lazarus, 2016

Ex. 1: Labeling images to study barrier island overwash (process) & washover (deposit)

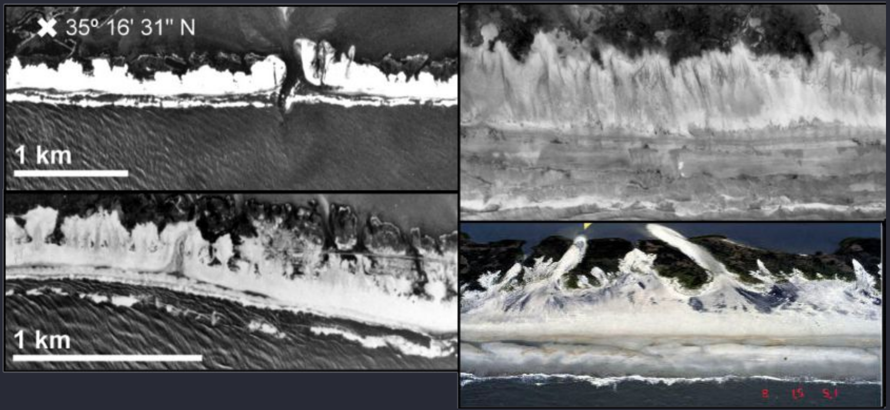


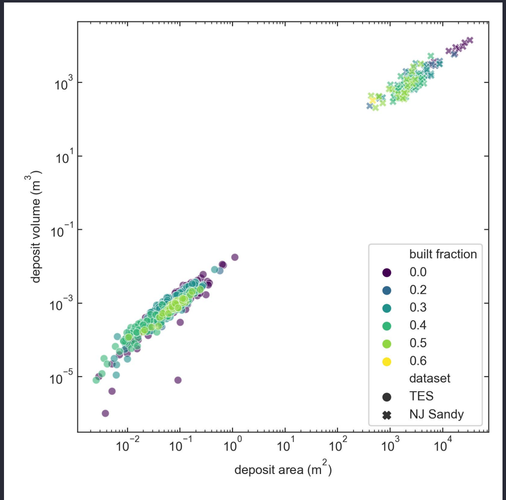


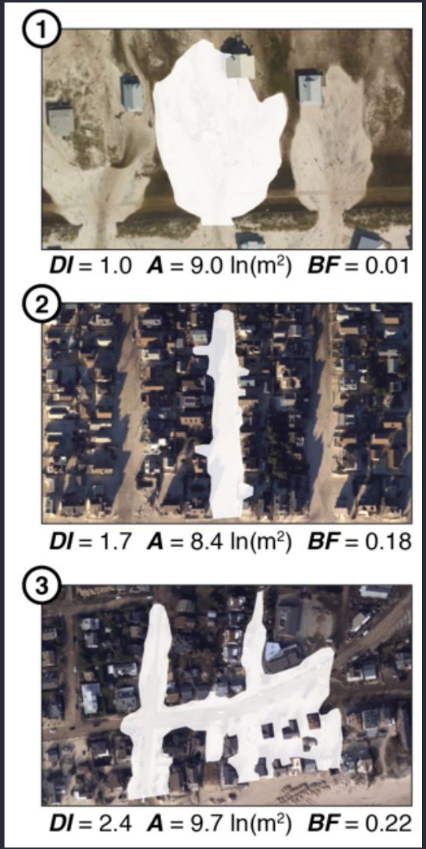


Lazarus et al (2020, 2022)

Ex. 1: Labeling images to study barrier island overwash (process) & washover (deposit)







Hurricane Florence 2018:
29k images, 7.7Mb each (!!!)

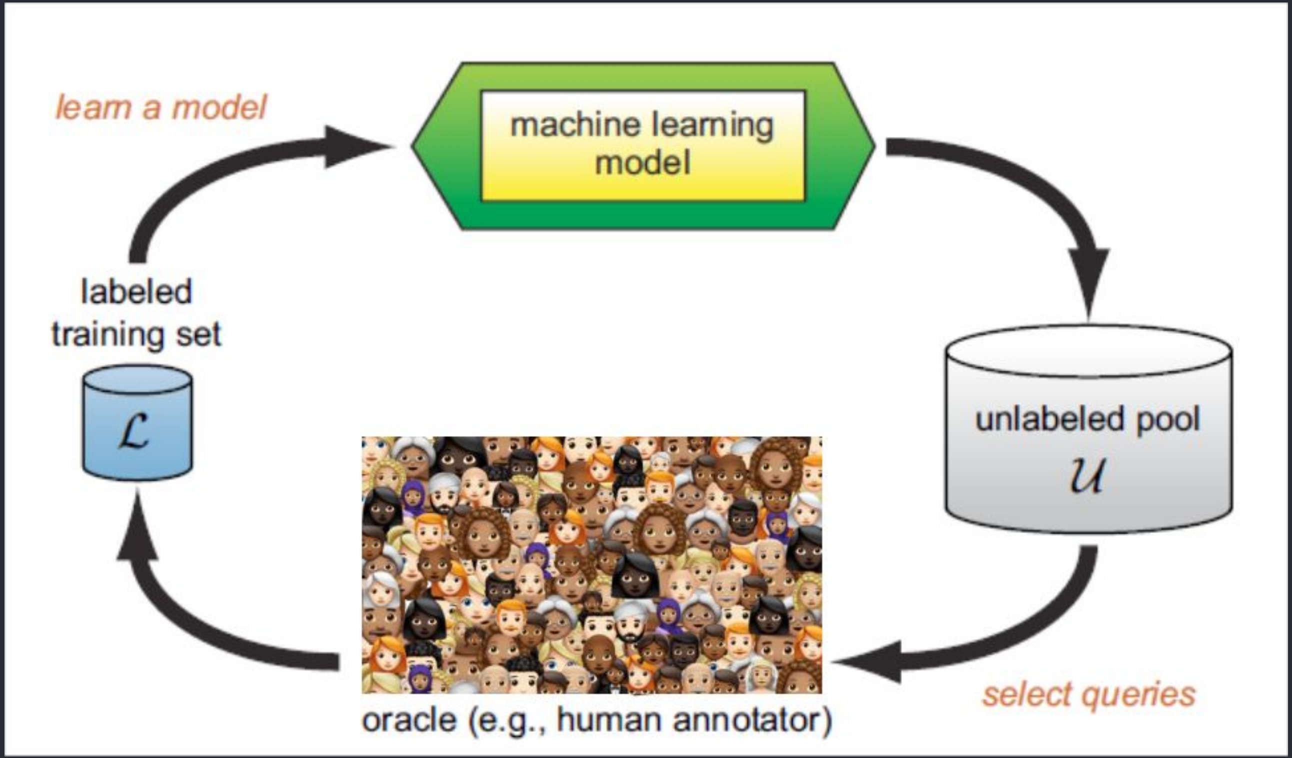
>140k images over 12 storms

Lazarus et al (2020, 2022)

We need labeled data

- Experts? Expensive

- Active learning



System described in Goldstein et al., AI for Earth NeurIPS workshop, 2020

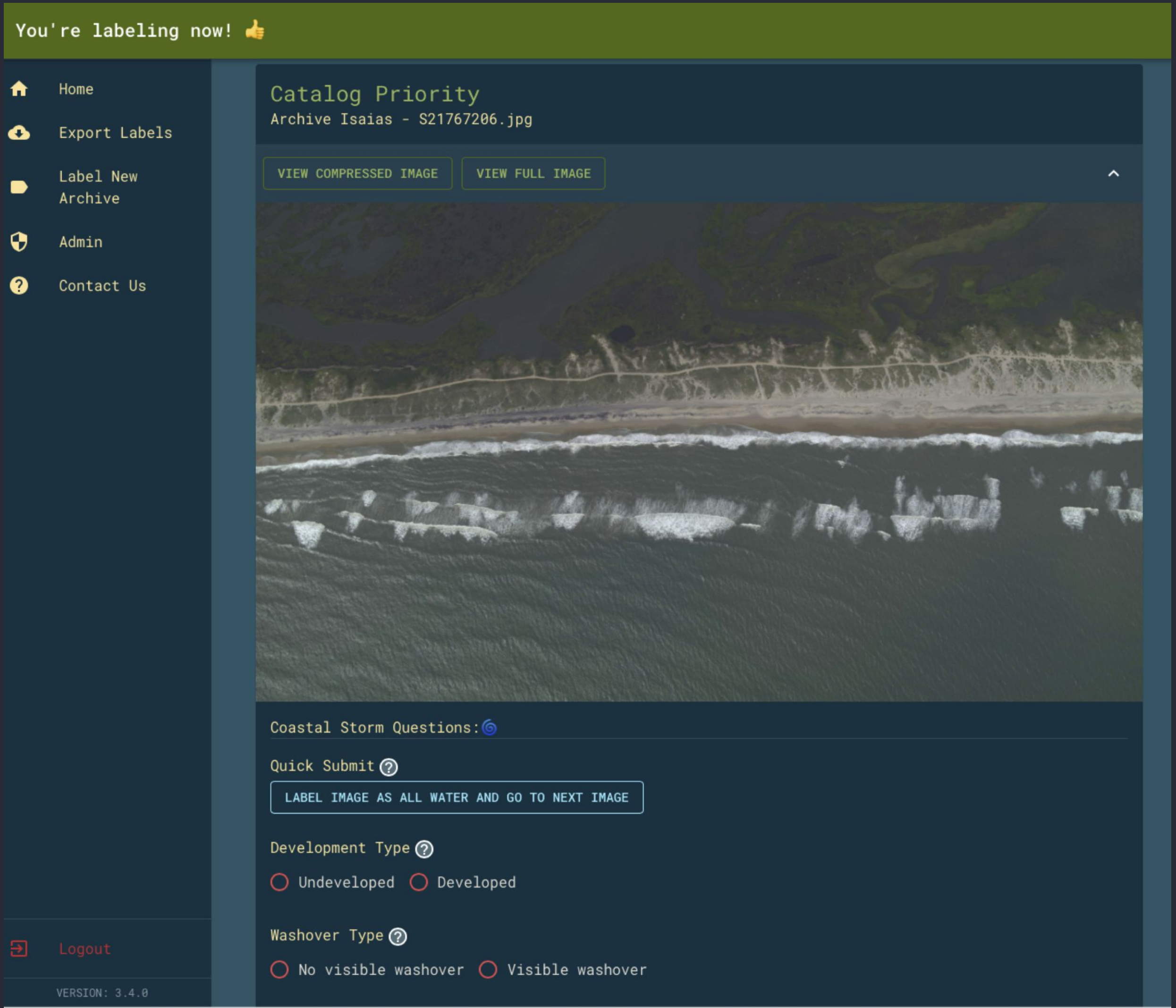
Coastal Image Labeler:

Public-facing, deployed via Azure VM from 2020-2023

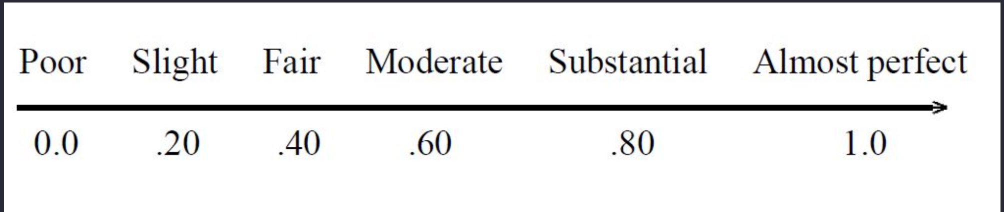
31 people

10.2k labels for 4250 images

https://doi.org/10.5281/zenodo.4272063

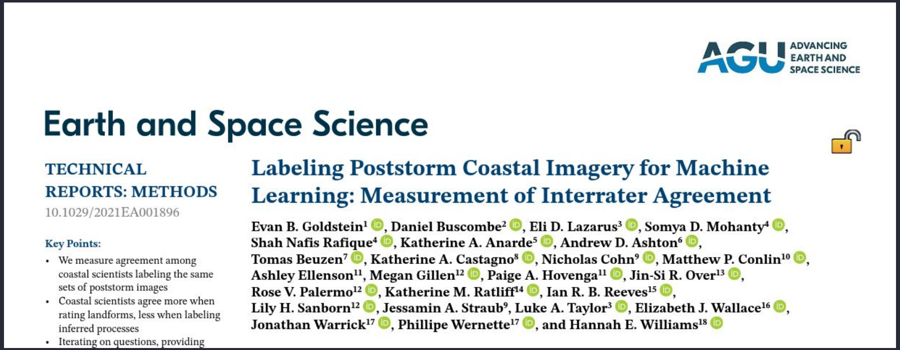


Inter-rater agreement



Creating labeled data is challenging - & the ‘labeled’ data is not even clean, unbiased, or error free.

Some objects/definitions are not as clear as we think.



AGU ADVANCING EARTH AND SPACE SCIENCE

Earth and Space Science

TECHNICAL REPORTS: METHODS

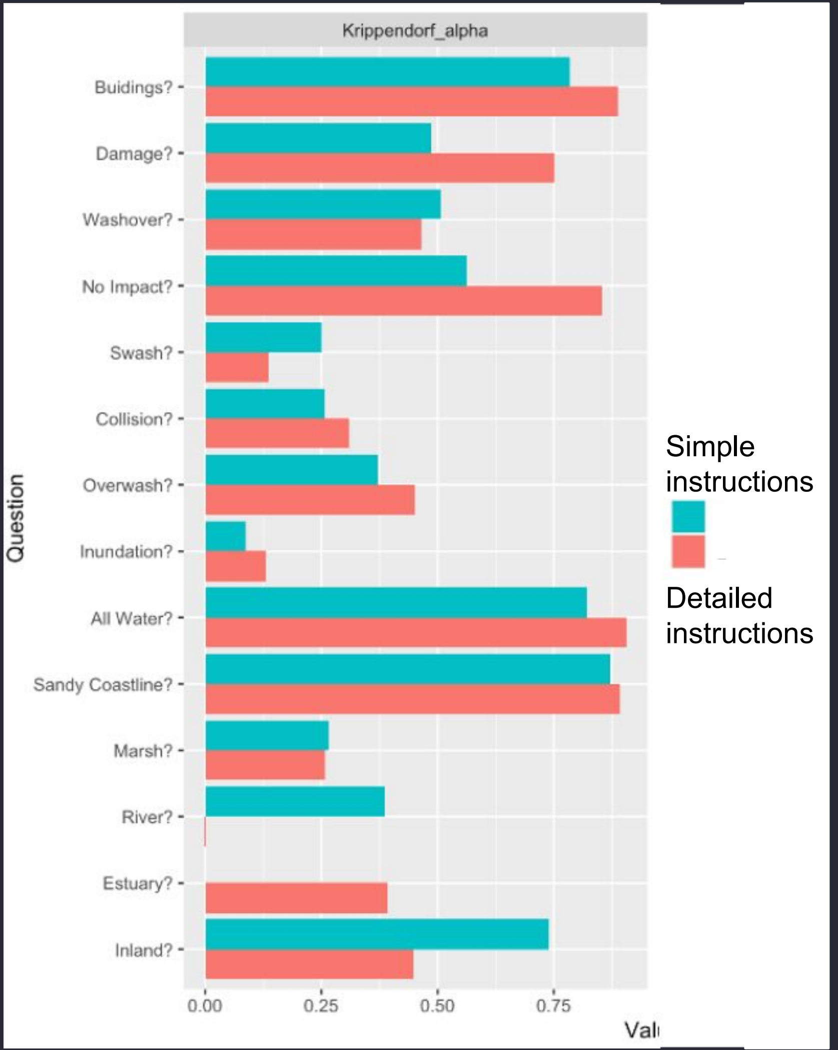
10.1029/2021EA001896

Key Points:

- We measure agreement among coastal scientists labeling the same sets of poststorm images
- Coastal scientists agree more when rating landforms, less when labeling inferred processes
- Iterating on questions, providing

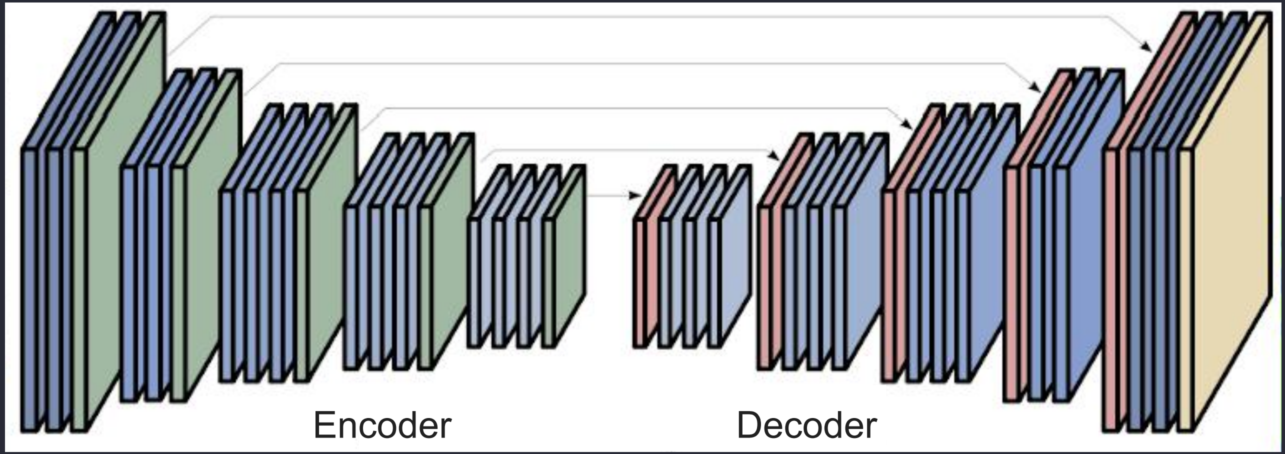
Labeling Poststorm Coastal Imagery for Machine Learning: Measurement of Interrater Agreement

Evan B. Goldstein[1], Daniel Buscombe[2], Eli D. Lazarus[3], Somya D. Mohanty[4], Shah Nafis Rafique[4], Katherine A. Anarde[5], Andrew D. Ashton[6], Tomas Beuzen[7], Katherine A. Castagno[8], Nicholas Cohn[9], Matthew P. Conlin[10], Ashley Ellenson[11], Megan Gillen[12], Paige A. Hovenga[11], Jin-Si R. Over[13], Rose V. Palermo[12], Katherine M. Ratliff[14], Ian R. B. Reeves[15], Lily H. Sanborn[12], Jessamin A. Straub[9], Luke A. Taylor[3], Elizabeth J. Wallace[16], Jonathan Warrick[17], Phillipe Wernette[17], and Hannah E. Williams[18]



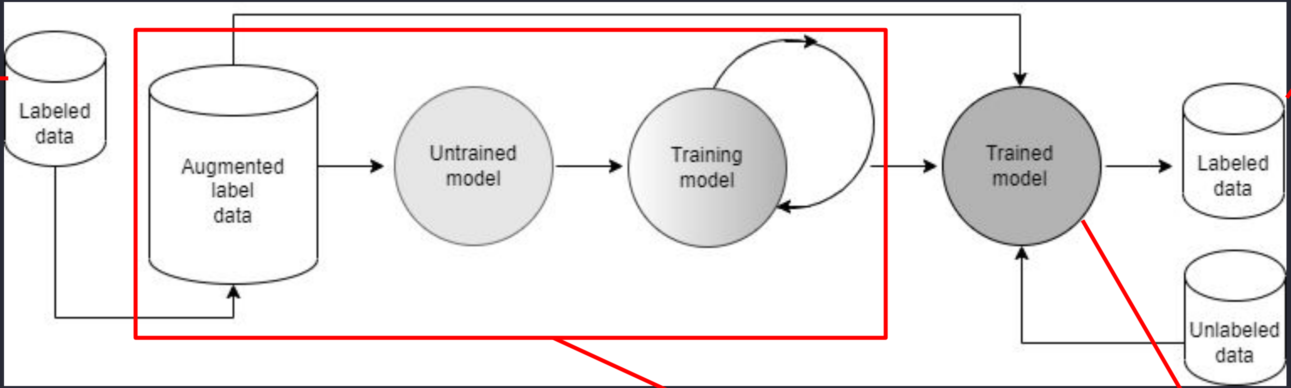
Ex. 2: Developing a generic geoscience segmentation tool

Dan Buscombe: end-to-end solution for labeling & modeling



Doodleverse

Coast Train



Buscombe et al. 2023 *"A 1.2 Billion Pixel Human-Labeled Dataset for Data-Driven Classification of Coastal Environments."* Scientific Data

Doodler

Buscombe et al. (2021) *"Human-in-the-Loop Segmentation of Earth Surface Imagery."* Earth and Space Science

gym

Buscombe and Goldstein (2022) *"A Reproducible and Reusable Pipeline for Segmentation of Geoscientific Imagery."* Earth and Space Science

zoo

Library of trained models..
In progress!

DOODLEVERSE SEGMENTATION OF GEOSCIENTIFIC IMAGERY

https://github.com/Doodleverse

There will always be a better model…

Sept 2022

Earth and Space Science

Method | Open Access

A Reproducible and Reusable Pipeline for Segmentation of Geoscientific Imagery

D. Buscombe, E. B. Goldstein

January 2023

Keras

Search Keras documentation...

About Keras

Getting started

Developer guides

Keras API reference

Keras Core: Keras for TensorFlow, JAX, and PyTorch

» Code examples / Computer Vision / Semantic segmentation with SegFormer and Hugging Face Transformers

Semantic segmentation with SegFormer and Hugging Face Transformers

Author: Sayak Paul

Date created: 2023/01/25

Quickly incorporated because of open datasets.

We need more (cleaned, labeled) data and to really value people who share (FAIR-ly)

Ex. 3: EdgeML for instant grain size

grain size variability is **wild**

- Small (cm) & fast (sec.) scales.
- Spatio-temporal data sparsity
- Measuring is onerous
- Models can use distributions, but most people use D50

Decades of work on instant grain size measurements

ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Sedimentary Geology 202 (2007) 402 – 408
Sedimentary Geology
www.elsevier.com/locate/sedgeo

Underwater microscope for measuring spatial and temporal changes in bed-sediment grain size

David M. Rubin [a,*], Henry Chezar [b], Jodi N. Harney [a,1], David J. Topping [c], Theodore S. Melis [c], Christopher R. Sherwood [d]
[a] U.S. Geological Survey, 400 Natural Bridges Dr., Santa Cruz, CA 95060, USA
[b] U.S. Geological Survey, 345 Middlefield Rd., Menlo Park, CA 94025, USA
[c] U.S. Geological Survey, 2255 N. Gemini Dr., Flagstaff, AZ 86001, USA
[d] U.S. Geological Survey, 384 Woods Hole Road, Woods Hole, MA 02543, USA

ELSEVIER
Available online at www.sciencedirect.com
ScienceDirect
Sedimentary Geology 201 (2007) 180 – 195
Sedimentary Geology
www.elsevier.com/locate/sedgeo

Field test comparison of an autocorrelation technique for determining grain size using a digital 'beachball' camera versus traditional methods

Patrick L. Barnard [a,*], David M. Rubin [a], Jodi Harney [a], Neomi Mustain [b]
[a] United States Geological Survey, Coastal and Marine Geology Team, Pacific Science Center, 400 Natural Bridges Drive, Santa Cruz, CA 95060, United States
[b] Department of Earth and Planetary Sciences, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95060, United States

Received 27 June 2006; received in revised form 18 May 2007; accepted 22 May 2007

A SIMPLE AUTOCORRELATION ALGORITHM FOR DETERMINING GRAIN SIZE FROM DIGITAL IMAGES OF SEDIMENT

DAVID M. RUBIN
U.S. Geological Survey, Santa Cruz, California 95060, U.S.A.
e-mail: drubin@usgs.gov

EARTH SURFACE PROCESSES AND LANDFORMS
Earth Surf. Process. Landforms **34**, 1811–1821 (2009)

Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/esp.1877

Cobble cam: grain-size measurements of sand to boulder from digital photographs and autocorrelation analyses†

Jonathan A. Warrick,[1] David M. Rubin,[1] Peter Ruggiero,[2] Jodi N. Harney,[3] Amy E. Draut[1] and Daniel Buscombe[1]
[1] US Geological Survey, Coastal and Marine Geology, Santa Cruz, CA, USA
[2] Oregon State University, Geosciences Department, Corvallis, OR, USA
[3] Coastal & Ocean Resources Inc., Sidney, BC, Canada

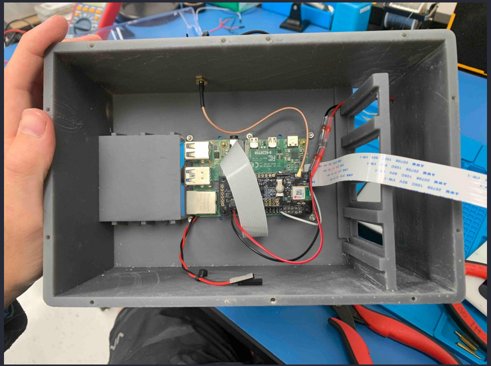
LIMNOLOGY and OCEANOGRAPHY: METHODS
Limnol. Oceanogr.: Methods 12, 2014, 390–406

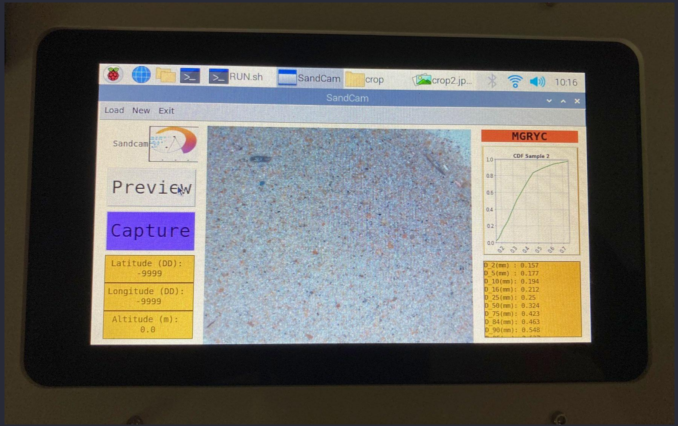
Autonomous bed-sediment imaging-systems for revealing temporal variability of grain size

Daniel Buscombe[1], David M. Rubin[2], Jessica R. Lacy[2], Curt D. Storlazzi[2], Gerald Hatcher[2], Henry Chezar[2], Robert Wyland[2], and Christopher R. Sherwood[3]*
[1]United States Geological Survey, Flagstaff, Arizona, USA
[2]United States Geological Survey, Santa Cruz, California, USA
[3]United States Geological Survey, Woods Hole, Massachusetts, USA

- Cameras and methods exist, but requires calibration and hand-off (i.e., off-device processing)

Instagrain





- Image regression on RPi using TF lite
- Multi-output: D2 to D98
- 2 sec. / obs. (mostly I/O)
- 15% MAPE on D50

**Only worked as co-designed system

https://github.com/UNCG-DAISY/Instagrain

Wrap Up

- ‘Labeled’ data is not as clean, unbiased, or error-free as we hope. Objects/definitions are not as clear as we think.
- There will always be a new, potentially better model. (Open data helps you quickly test new models.)
- On the edge, solutions likely require hardware + software + ML co-design.

Thank you for the invitation to speak and thanks for listening!