

Accelerating discovery in Natural History through modern computation

Stéfan van der Walt
Berkeley Institute for Data Science
@stefanvdwalt

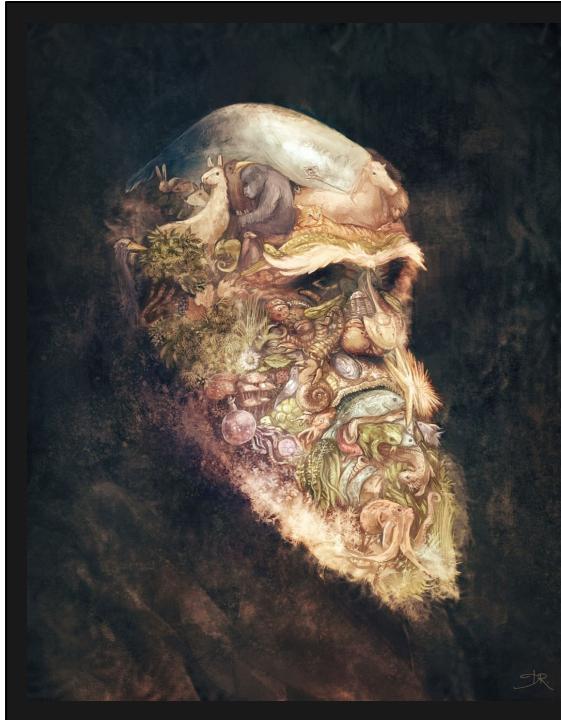






H.M.S. Beagle in the Galapagos Islands. Painting by John Chancellor.





"I love fools' experiments. I am always making them."

— Charles Darwin

"...even a wise experiment when made by a fool generally leads to a false conclusion, but that fools' experiments conducted by a genius often prove to be leaps through the dark into great discoveries."

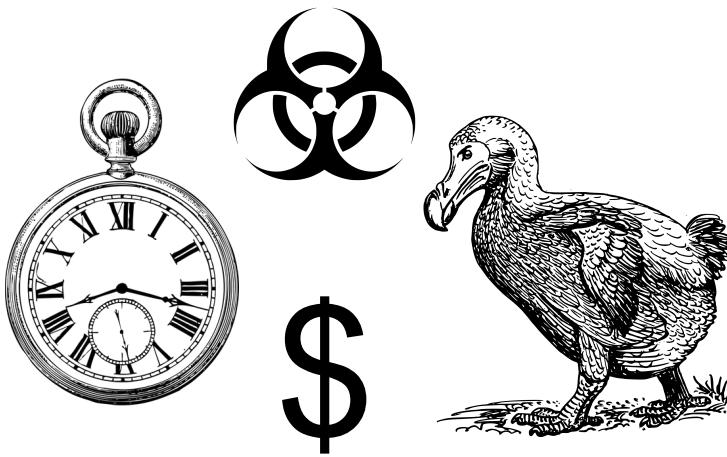
— E. R. Lankester. 'Charles Robert Darwin', 1896.

Artist: David Revoy

License: [CC-BY-NC-ND](https://creativecommons.org/licenses/by-nd/3.0/) David Revoy, www.davidrevoy.com, 16 november 2012.

The idea being that you come up with *many* experiments, some of which are foolish, but some of them provide good insight. The problem we are here to talk about today is how can you maximise the number of experiments you can make on pre-existing data, that was gathered with great care.

Problem



SEARCH WHAT'S NEW METHODS PARTNERS PUBLICATIONS

Specimen Details CalBug

A021 | 640x480 | 2011/12/12 13:47:50

Berkeley, Cal
III-271954

J. W. MacSwain
Collector
Lasiglossum pacificum (Ckll.)
det. McGinley

U.C. Berkeley
EMEC 531,661

Lasiglossum pacificum

Specimen ID:	EMEC531661
Type Status:	
Specimen Preparation:	pin
Institution:	Etsig Museum of Entomology
Other Number:	
Higher Taxonomy:	Animalia, Arthropoda, Insecta, Pierigota, Hymenoptera, Apocrita: Aculeata, Apidae, Halictidae, Halictini,
Scientific Name:	<i>Lasiglossum pacificum</i> (Cockerell, 1898)
Identified By:	McGinley
Location:	Berkeley (Alameda County, California, United States)
LatitudeLongitude:	37.8718 -122.27644 Max error: 4286 m Datum: WGS1984
Elevation:	
Habitat:	
Collection Method:	
Collected By:	J. W. MacSwain
Collection Date:	1954-3-27

- Collections are incredibly valuable (time, money, risk)
 - Often cannot be replicated
- To gather insights from our collections require a large amount of additional effort - These tasks are typically labor intensive, e.g., taking measurements of each gathered species. Some of our collections are now so big, that the task is impossible; i.e., we are forfeiting on scientific discovery.
- Question: can we utilize modern technology to improve our ability to access, analyze, and interpret all this data?

Dr Mrs Harry Vischer & Mrs Ella H. Wendell her

¹⁹⁰⁵ July 19 A.T.Van Atson	¹⁹⁰⁵ July 5 Rent store	\$3875
" Pol 48.989 Aluma	31 " Office	17.50
" Plate Glass \$9.95 #4.98	Sept 1 "	17.50
Aug 15 H&W Pol 1072.5471	5 " Store	3875
" Royal 38 E MSL #2725 1363	30 " Office	17.50
" Pol 96.29.19 Natl	Oct 3 " Store	3875
" 38.8 MSL #1635 8.18	31 " Office	17.50
21 16. Dwyer } 180	Nov 1 " Store	3875
1.117 Post & School Tax 7128		

Let's say 100,000 samples

Each sample takes 10 minutes to fetch and measure

That's ~16,500 hrs, or 8 years of 40 hr workweeks.

Perhaps we can parallelize, so that 8 workers, perfectly coordinated, can do this in ~1 yr. At 62,500 USD per person per annum, that's 500,000 USD.

Complications

What if we had 500,000 samples instead? That's: 8 people, 5 years, 2.5M USD.

What if: we now discovered that a) we forgot to measure some attribute, or b) made a mistake in how we measured. All that work is lost!

- Use cheaper labor (volunteer time, crowd sourcing)? Doesn't solve second problem of wasted effort, and still takes time and coordination effort.

Half-way solution: digitize

"The cost of traditional digitization workflows is vast, both in financial and human terms. Our simple calculations have shown that complete databasing of the ~30 million insect specimens housed in the entomological collection of the Natural History Museum, London, would require 23 years of continuous work from the entire departmental staff to complete (65 people). Depending on the particular collections and curatorial practices used, estimates vary from US\$0.50 to several dollars per specimen to capture full label data ([Heidorn 2011](#)). The cost of traditional imaging and databasing of every natural history object in all European museums was recently estimated as €73.44 per object ([Poole 2010](#)). Thus, the complete digitization of all natural history collections may cost as much as €150,000 million, and take as long as 1,500 years."

— [Vladimir Blagoderov](#),¹ [Ian J. Kitching](#),¹ [Laurence Livermore](#),¹ [Thomas J. Simonsen](#),¹ and [Vincent S. Smith](#)¹

¹Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK

No specimen left behind: industrial scale digitization of natural history collections

Solution: streamlined digitization & automated analysis

- As much as possible of the **procedure must be automated**, except when physical handling of specimens is necessary.
- The approach should, whenever possible, **focus on “wall-to-wall” total digitization of entire collections**, because it is faster to digitize an entire collection than to select individual specimens or drawers of particular interest.
- Complicated **labour-intensive procedures must be divided into** a series of separate, shorter **steps**, each with a distinct outcome. For example, preparation of specimens for imaging should be a separate step from the imaging itself; and unique specimen identifiers can be assigned simultaneously to all specimens in a drawer rather than individually and sequentially. Such a modularised process can then be more easily crowd-sourced among the professional and volunteer communities. Properly organized crowd-sourcing projects would be able to mobilise the efforts of thousands of enthusiasts around the world ([Hill et al. 2012](#)).
- **Collection of metadata must be simplified and standardized.** In most cases, digital representation of the specimen and minimal metadata (uID, specimen location in the collection) is sufficient for collection management purposes. Only minimal information should be collected when initially digitizing an entire collection, but in such a way that it can be amended and expanded upon later.

— [Vladimir Blagoderov](#),¹ [Ian J. Kitching](#),¹ [Laurence Livermore](#),¹ [Thomas J. Simonsen](#),¹ and [Vincent S. Smith](#)¹

¹Department of Life Sciences, Natural History Museum, Cromwell Road, London, SW7 5BD, UK

No specimen left behind: industrial scale digitization of natural history collections

Vision

- The museum has a device that can rapidly scan a sample
- Samples are moved into the scanner, the shelf nr is entered, and the photo is stored in a digital repository
- Do our analyses and measurements digitally, so that any analysis can be repeated and modified “free of charge” (a bit of compute time).
- For this solution to work, we need to optimize throughput (scanning should be much quicker than the originally intended measurements)
- This (often) requires **specialized hardware and automated analysis software**

Project showcase

1. **Shark identification** — algorithm guided by human input
2. **Inselect: insect tray segmentation** — automated rules, humans verify
3. **Butterfly measurement** — automated rules, tested on human measurements
4. **Bee identification** — automation of rules + learn from human classifications

5. **3D label recovery** — in progress; machine vision, best model unclear

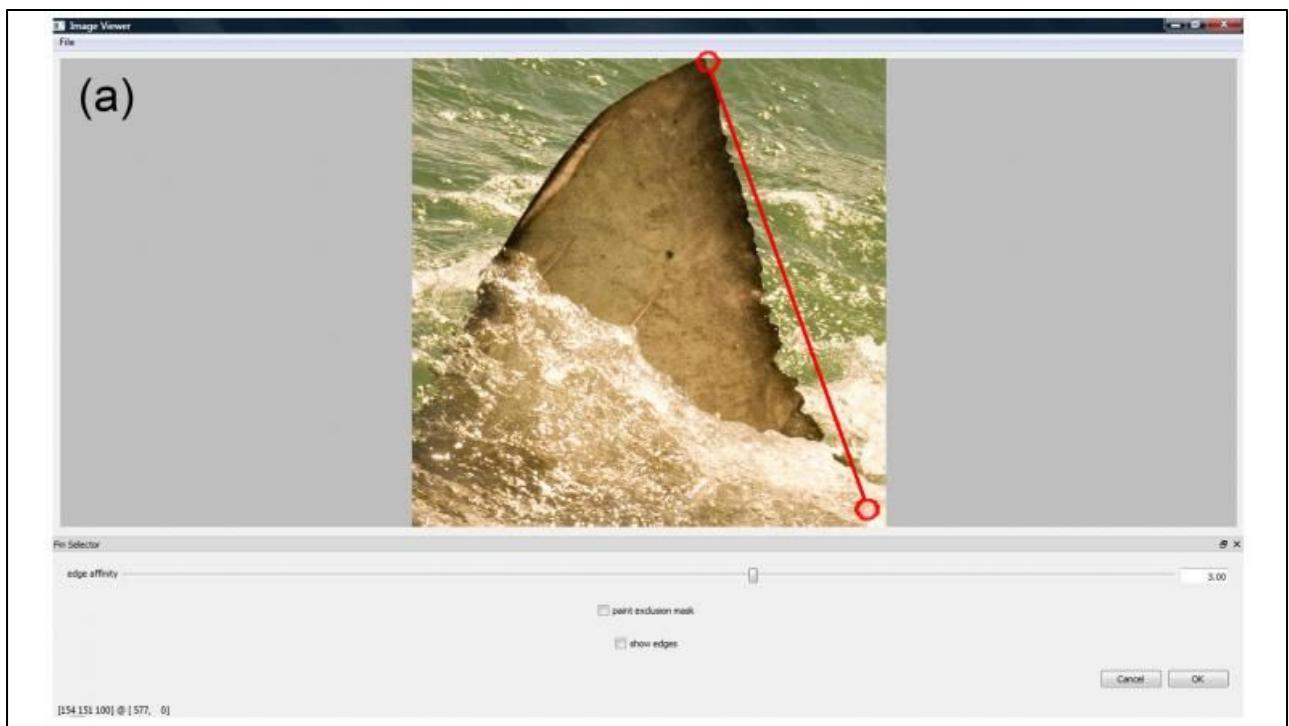
Project 1: **Shark** identification

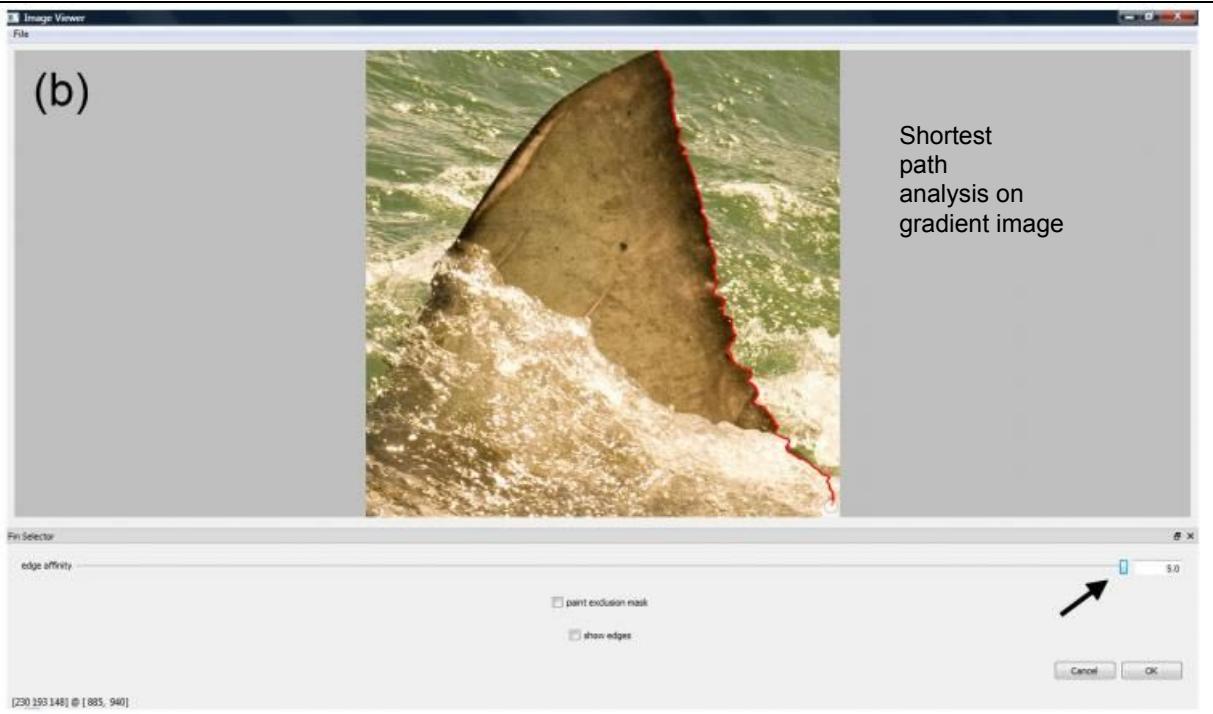
With Sara Andreotti, Pieter Holtzhausen (Stellenbosch University)

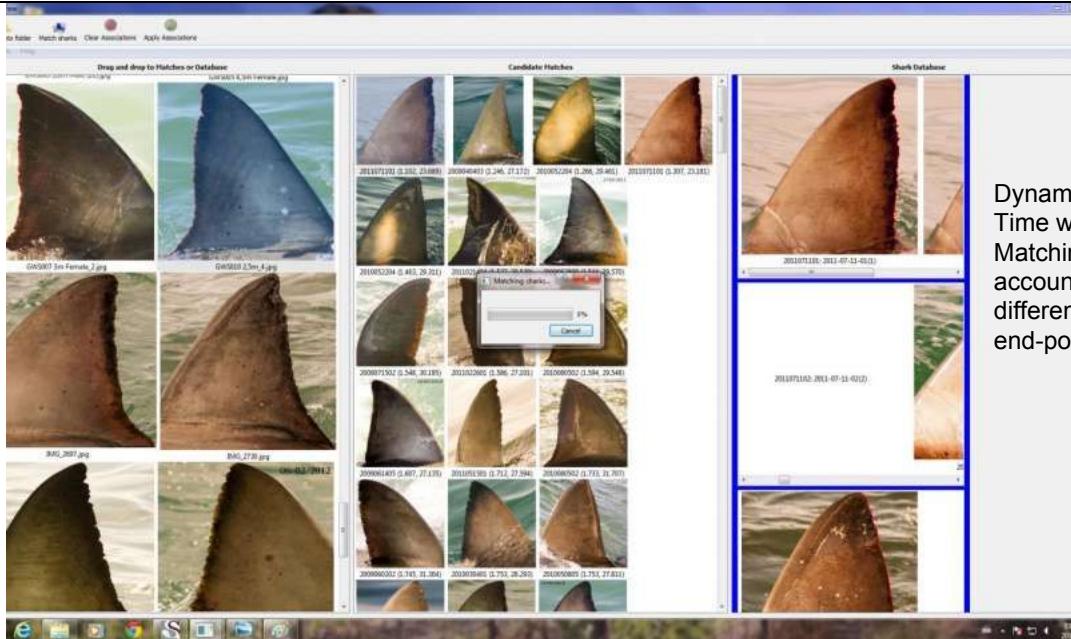


06/05/2009









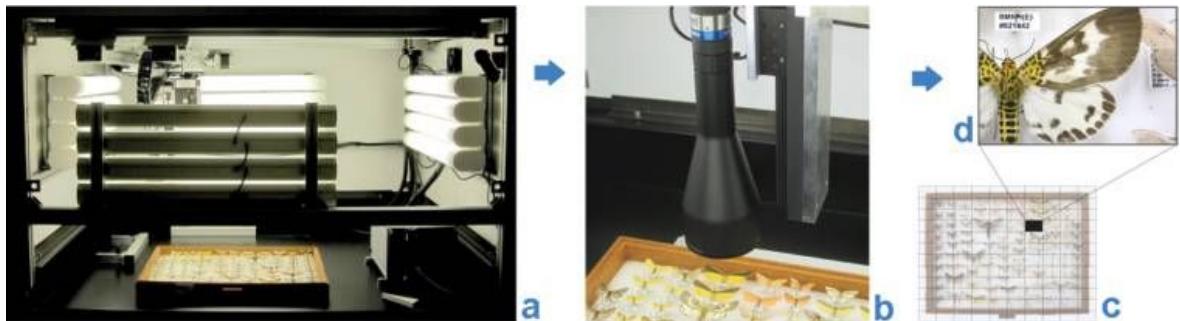
Dynamic
Time warping
Matching
accounting for
differences in
end-points

Significant effort went into untangling the pre-existing file system storage scheme!



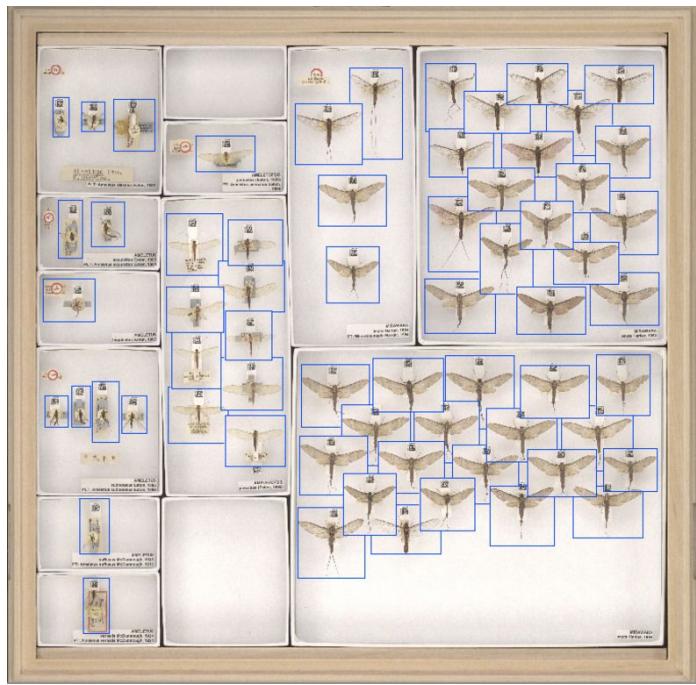
Project 2: Inselect insect tray segmentation

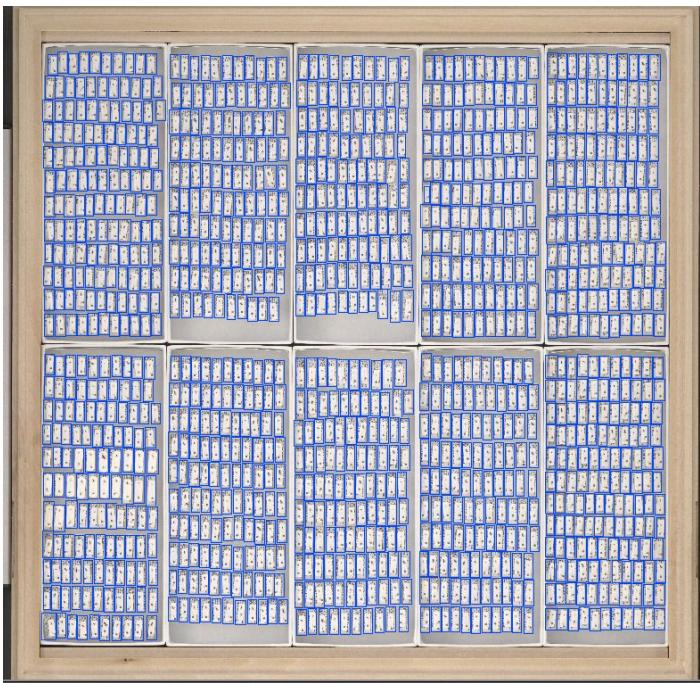
With Ben Price et al. (Natural History Museum, London), Pieter Holtzhausen

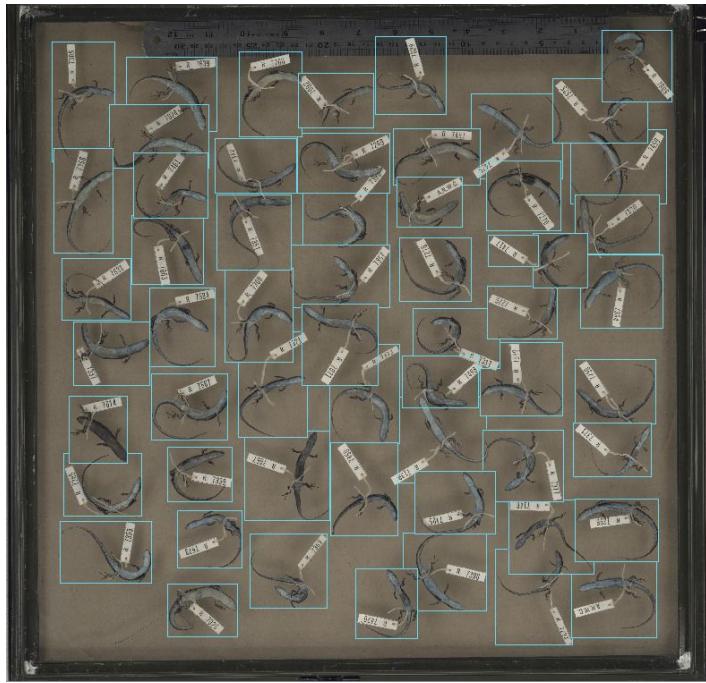


SatScan imaging: **a** SatScan machine **b** specimens being imaged **c** individual frames aligned **d** fragment of a stitched image; final resolution of the stitched image ~11 lines/mm.

[Copyright](#) Vladimir Blagoderov, Ian J. Kitching, Laurence Livermore, Thomas J. Simonsen, Vincent S. Smith. CC-BY 3.0.



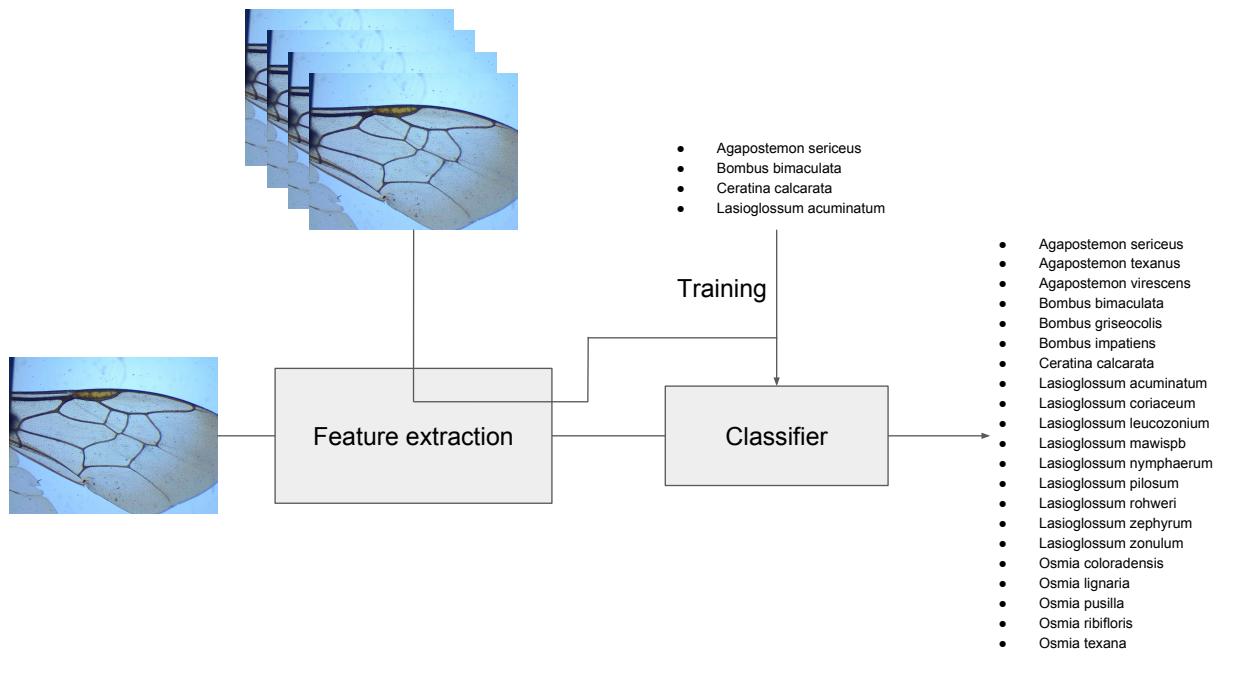




Project 3: Bee identification from wing venation

With Lauren Ponisio (UC Riverside), Théo Bodrito (Mines ParisTech), BIDS
Machine Shop



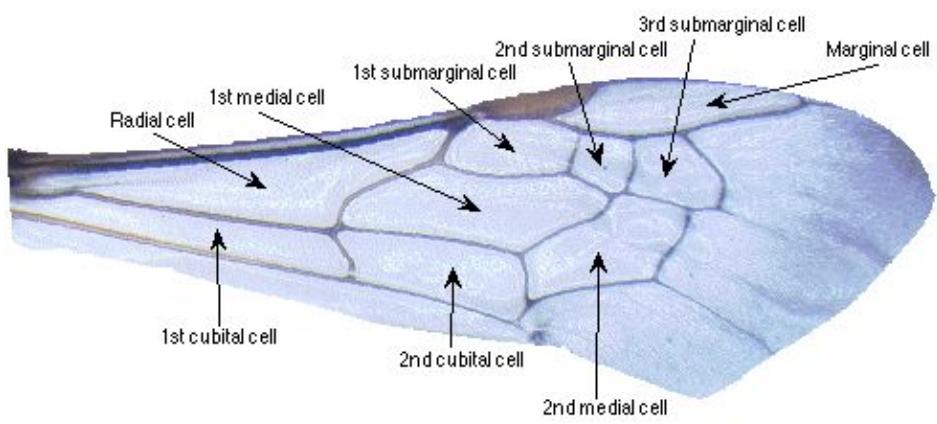


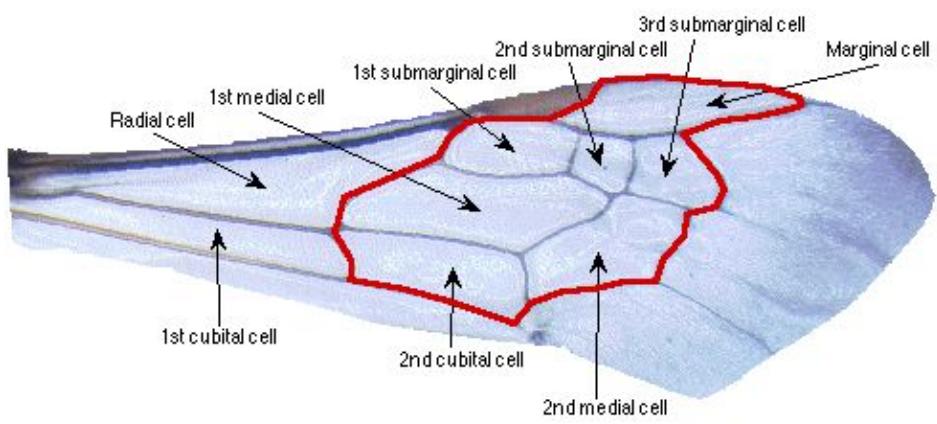
Feature extraction pipeline

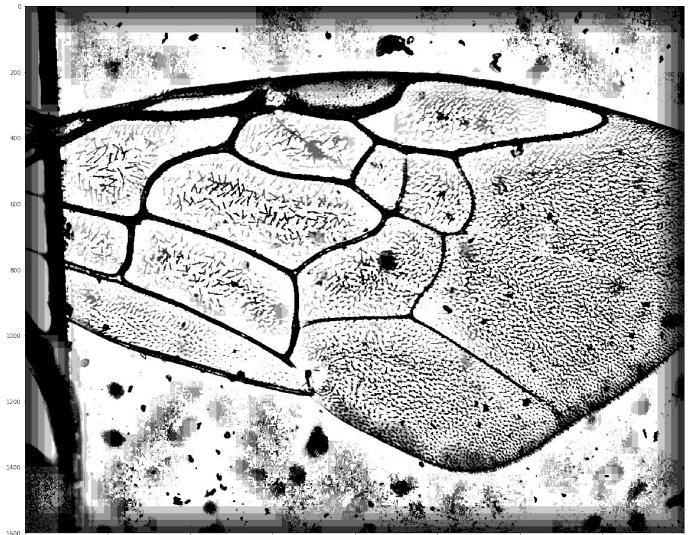
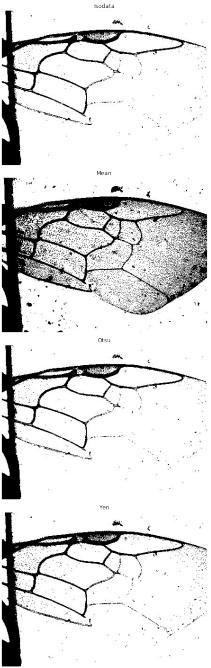
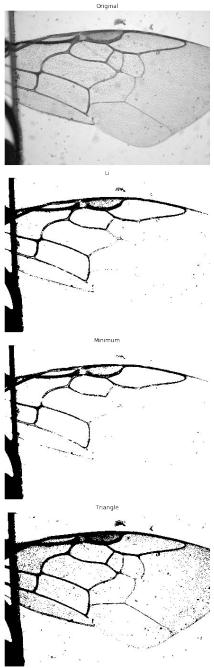
To use with classic machine learning models, we need features from images.



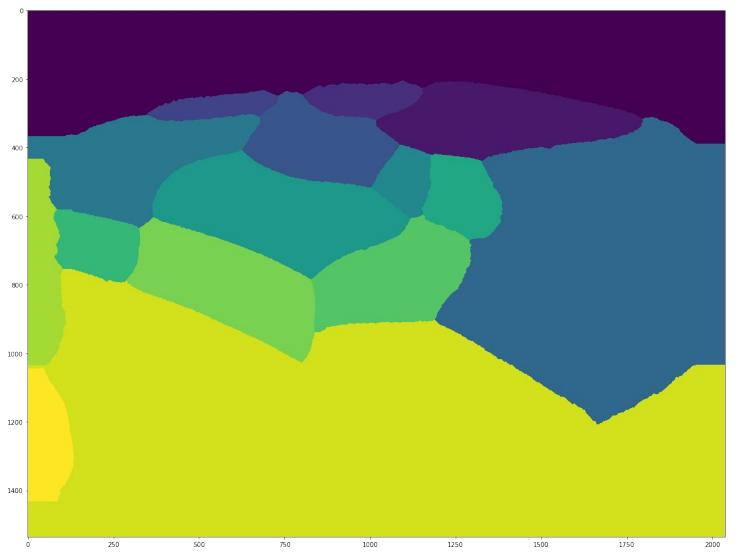
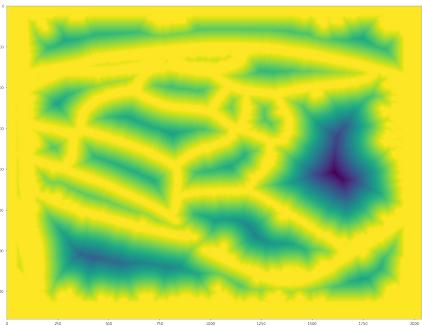
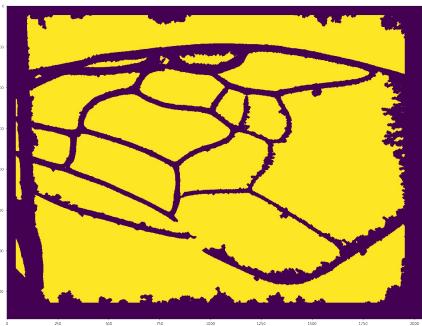
[0.5, 12.3, 7.8, ..., 10.3]



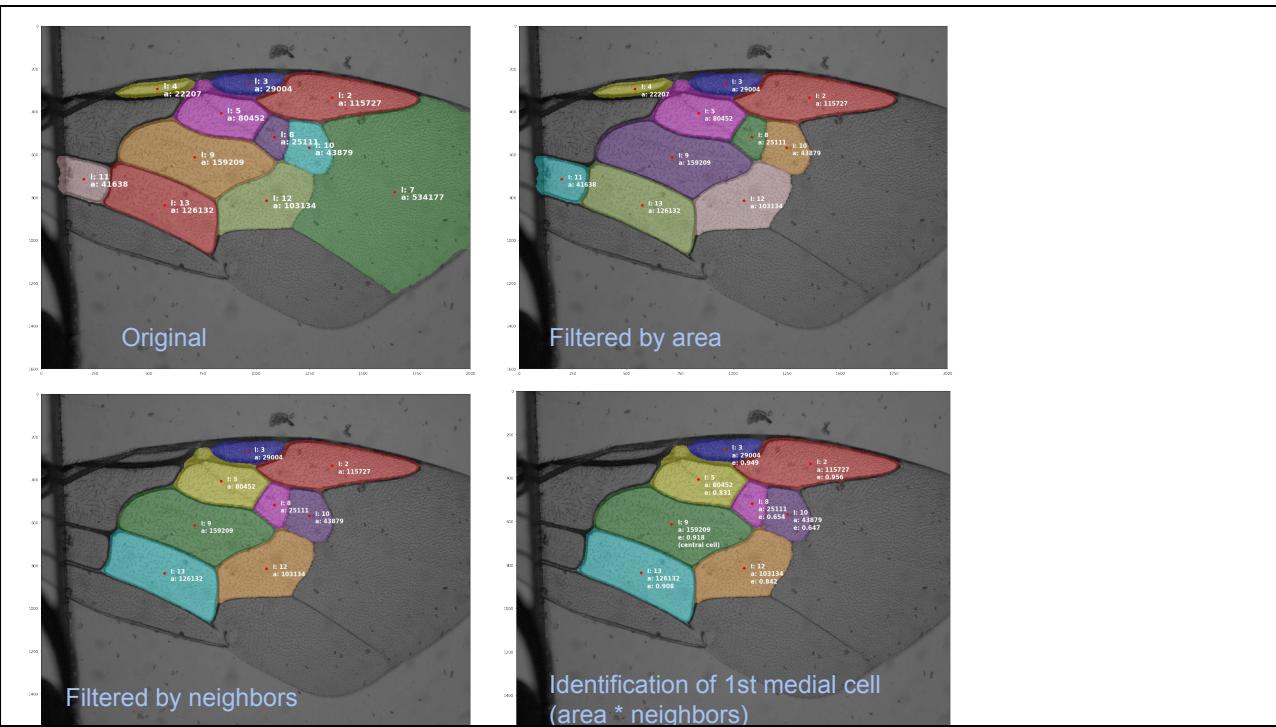


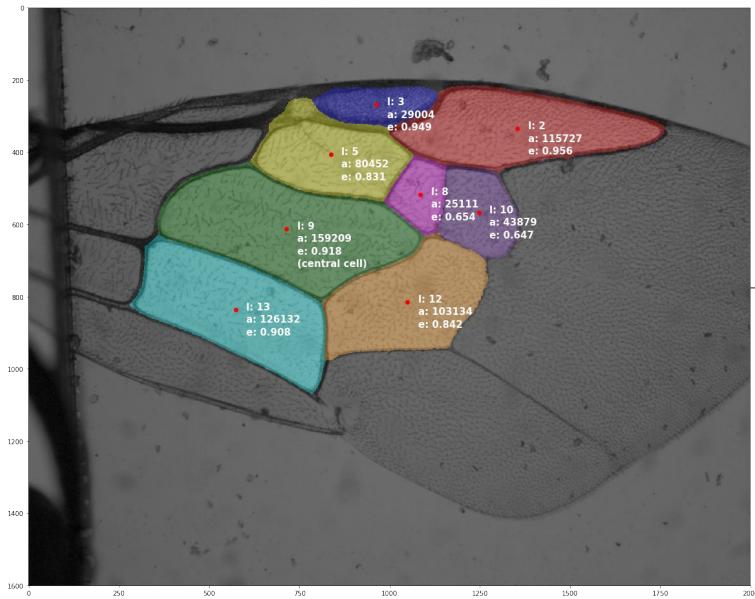












Calculate properties for each cell;

Area, circumference,
boundary coefficients
(we store a compact representation of
the *shape* of the cell), etc.

→ [0.5, 12.3, 7.8, ...]

Varying numbers of medial cells?



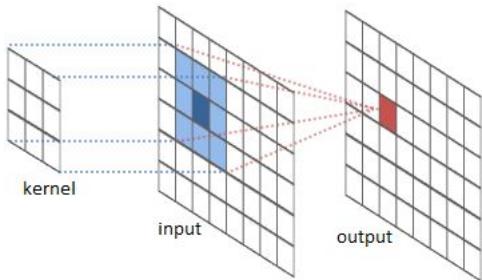
Split classifier.

Or...

...forget all that

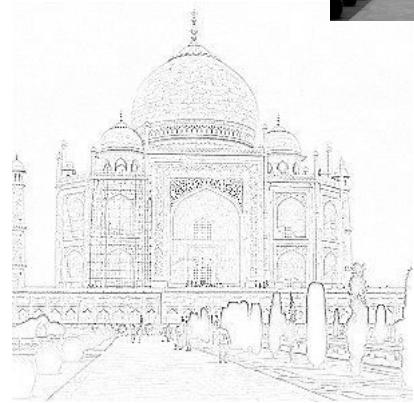
and use a convolutional neural network.

↖_(ツ)_↗



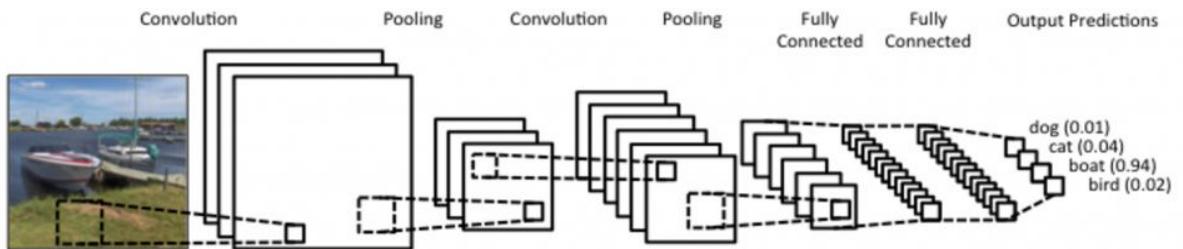
<http://intellabs.github.io/RiverTrail/tutorial/>

0	1	0
1	-4	1
0	1	0

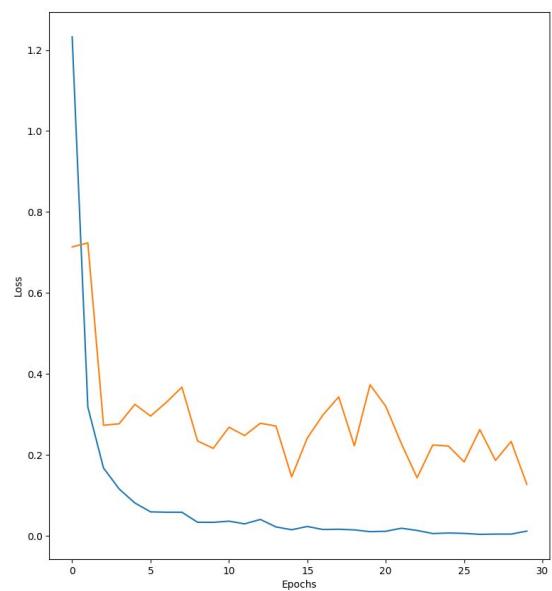
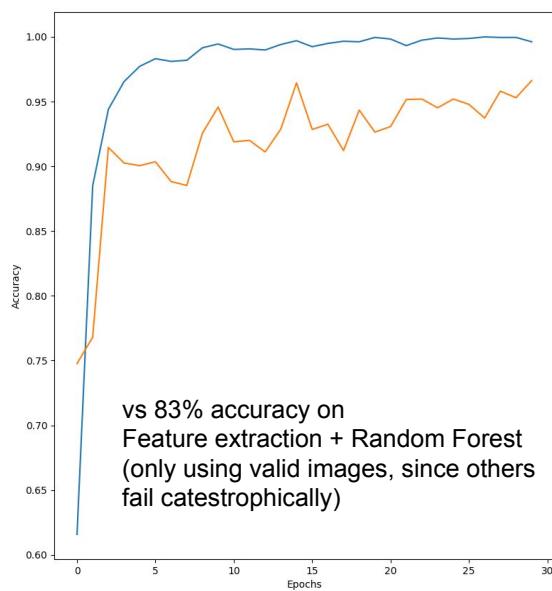


From Gimp 2.8 manual at
<https://docs.gimp.org/2.8/en/plug-in-convmatrix.html>

Example architecture (for illustration, we used DenseNet instead)



From <http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/>



Advantages:

- Raw data as input (perhaps augmented); no feature engineering
- Network can be engineered; little domain knowledge needed

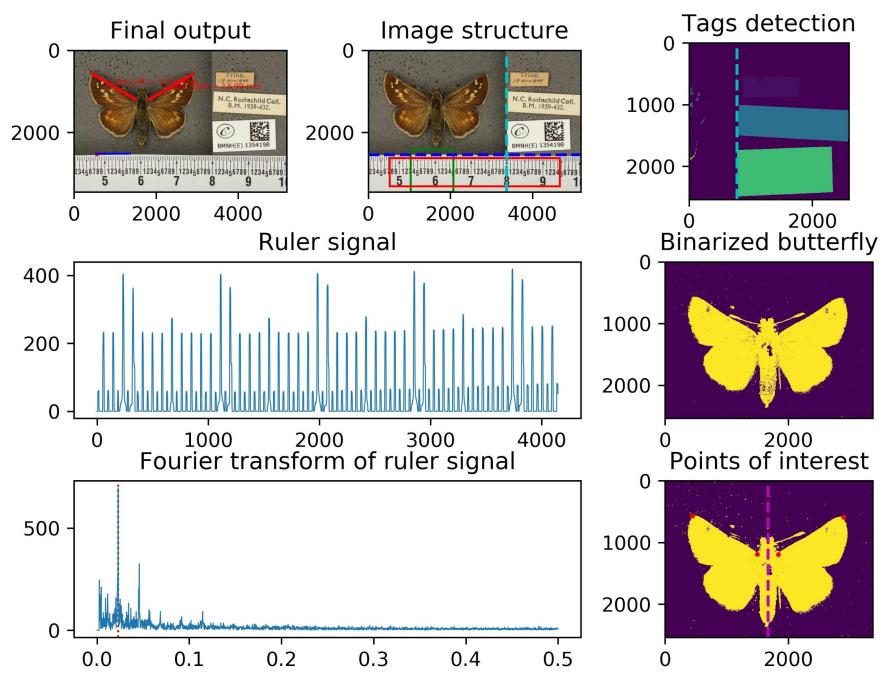
Disadvantages:

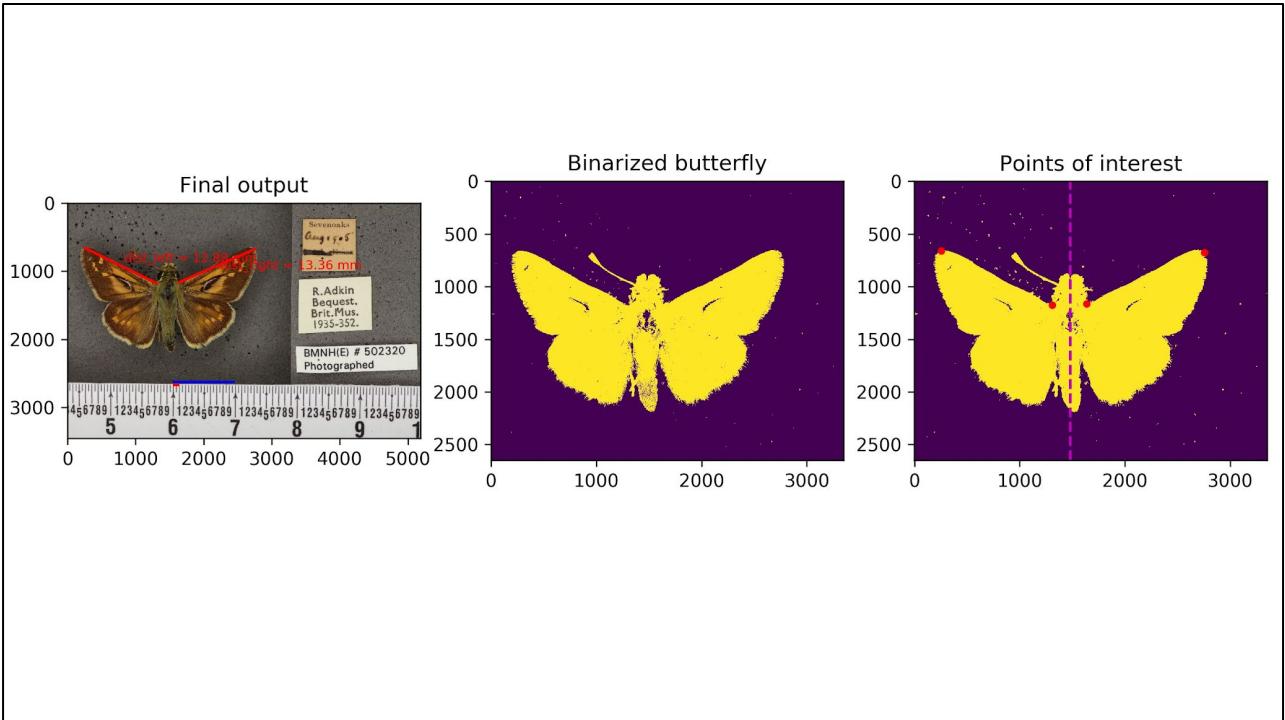
- Unknown failure modes; operates well only within reaches of seen data
- Working is not easily interpretable
- Often requires a significant amount of training data

Project 4: Butterfly measurement

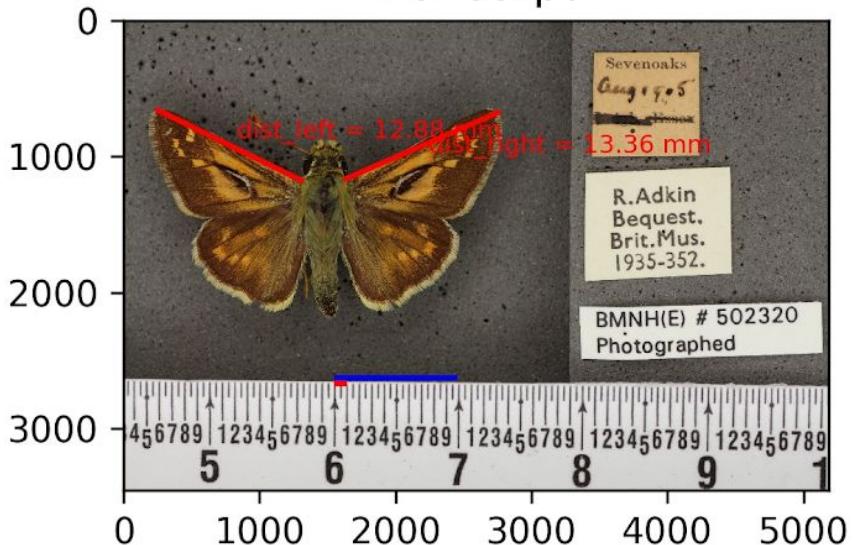
With Ben Price (NHM), Phil Fenberg (Southampton/NHM), Théo Bodrito (Mines ParisTech), Dennis Feng (UC Berkeley),
Rebecca Wilson (Southampton/NHM), Hannah O'Sullivan (NHM), Steve Brooks (NHM), BIDS Machine Shop (UC Berkeley)

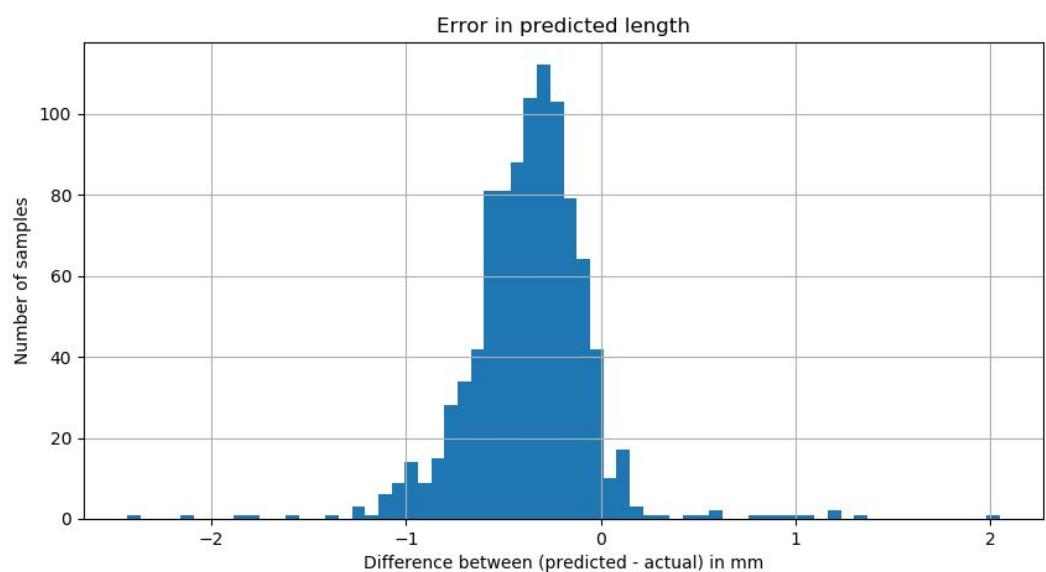


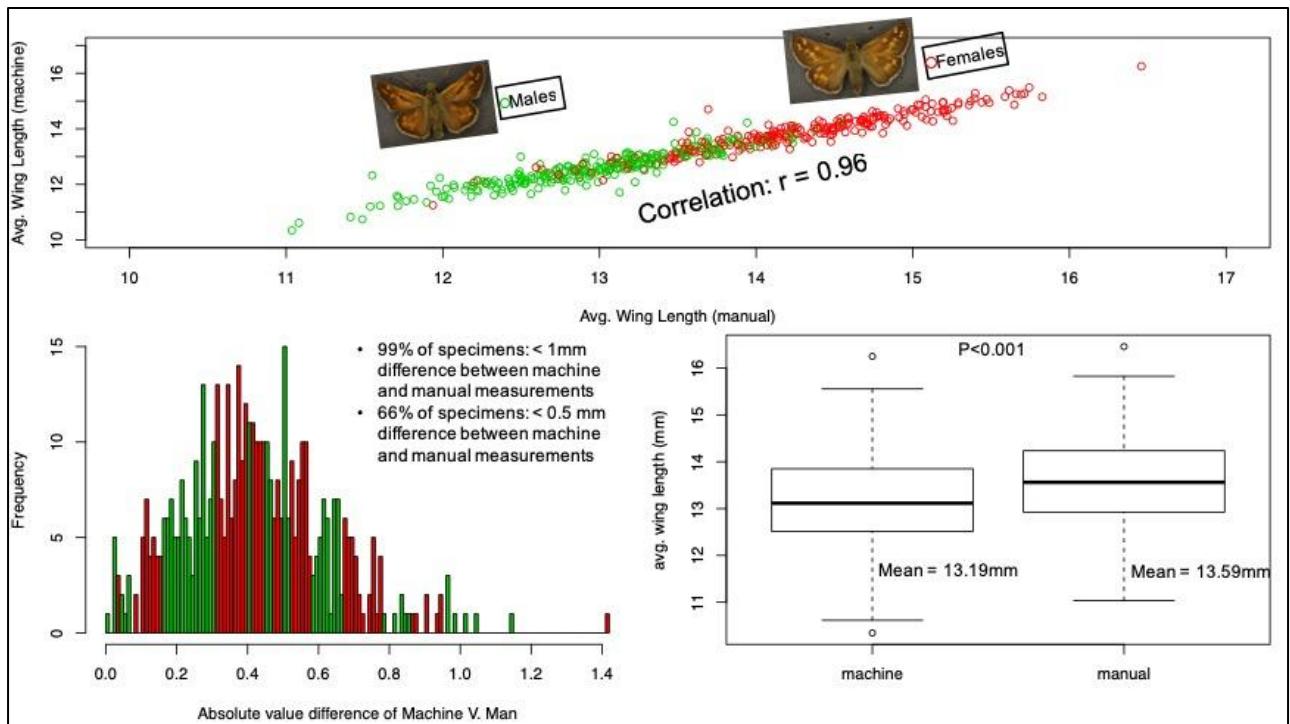




Final output







ALICE: Angled Label Image Capture and Extraction for high throughput insect specimen digitisation



AUTHORS

Benjamin Price, Steen Dupont, Elizabeth Allan, Vladimir Blagoderov, Alice Butcher, James Durrant, Pieter Holtzhausen, Phaedra Kokkini, Laurence Livermore, Helen Hardy, Vincent Smith

CREATED ON

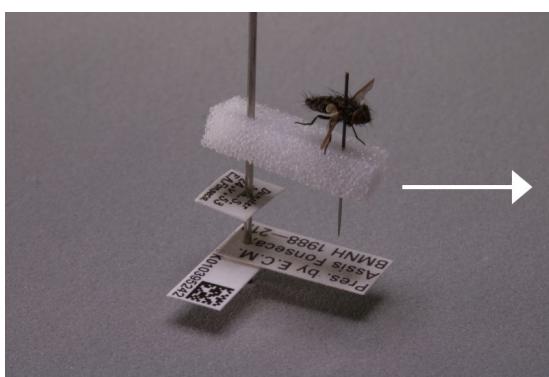
November 05, 2018

LAST EDITED

November 05, 2018

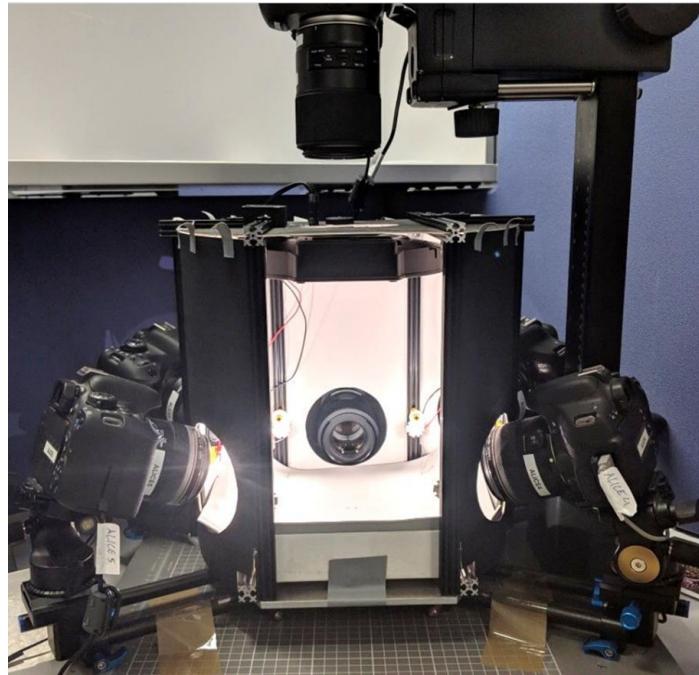
SUPPLEMENTAL MATERIALS

osf.io/9p4f6/



Future work

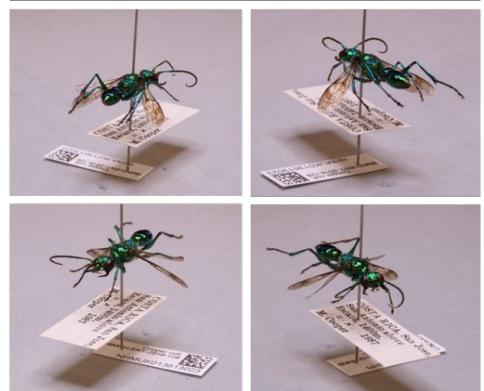
<https://github.com/NaturalHistoryMuseum/ALICE>

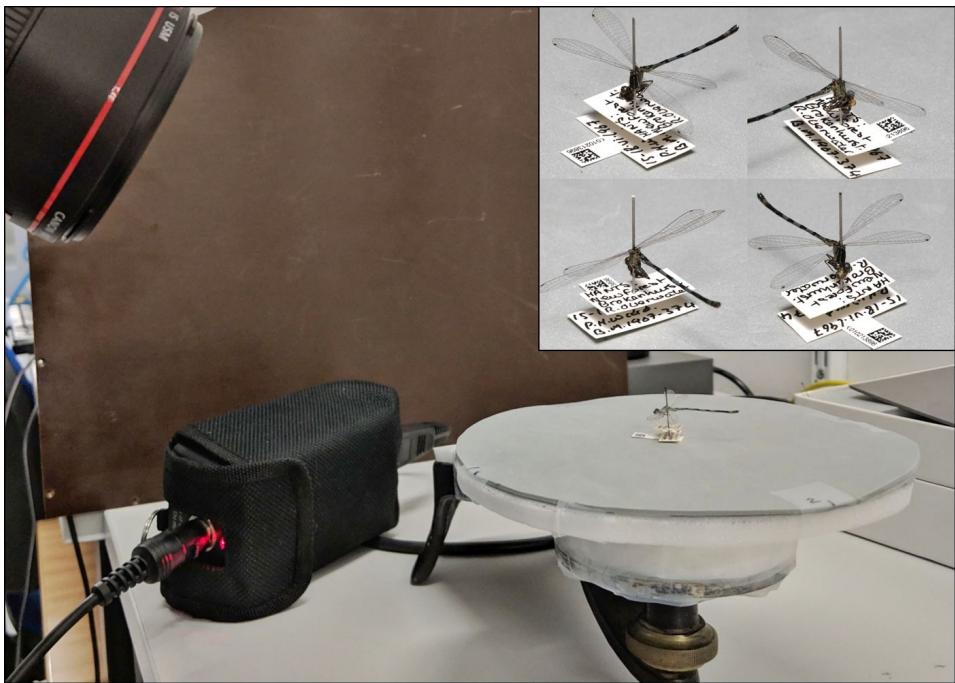


Specimen Images



Label Images





Conclusion

We have very **expensive collections**

These collections contain **great riches**

But, we must **improve our ability to analyze them**

We can do that, with the help of **modern imaging and computation**

And **we are doing just that**, here at Berkeley.