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."

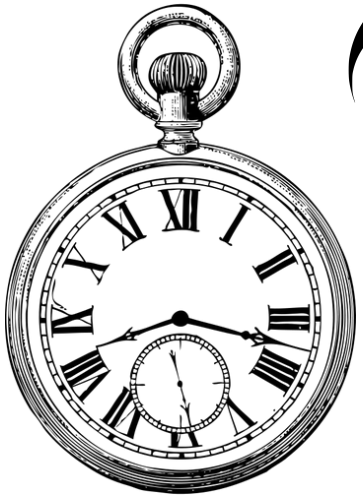
— *Charles Darwin*

“...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

Problem



SEARCH | WHAT'S NEW | METHODS | PARTNERS | PUBLICATIONS

Specimen Details

CalBug

Lasioglossum pacificum

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

Dr Mrs Harry Visscher & Mrs Ella H. Wendell Cr

1905				1905			
July 19	A. J. Van Hoesen			July 1	Rent	Store	$38.75
"	Pol 48.989 Alma			31	"	Office	17.50
"	Plate Glass $9.95	$4.98		Sept 1	"	"	17.50
Aug 15	H & W. Pol 1072.54.71			5	"	Store	3875
"	Royal 38 E M St $27.25	13.63		30	"	Office	17.50
"	Pol 96.29.19 Nat'l			Oct 3	"	Store	38.75
"	38. E M St $16.35	8.18		31	"	Office	17.50
21	L. Dwyer			Nov 1	"	Store	3875
"	38. E Main St	1.80					
Sept 1	Coit & Schoonmaker	7.128					

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, crowdsourcing)? Doesn't solve second problem of wasted effort, and still takes time and coordination effort.

Can modern technology help?

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]

[1]Department 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]
[1]Department 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 procedurally moved into the scanner, some metadata is entered, and the photo is stored in a digital repository
- We 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

Once we have a digital collection, there's plenty to do.

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)

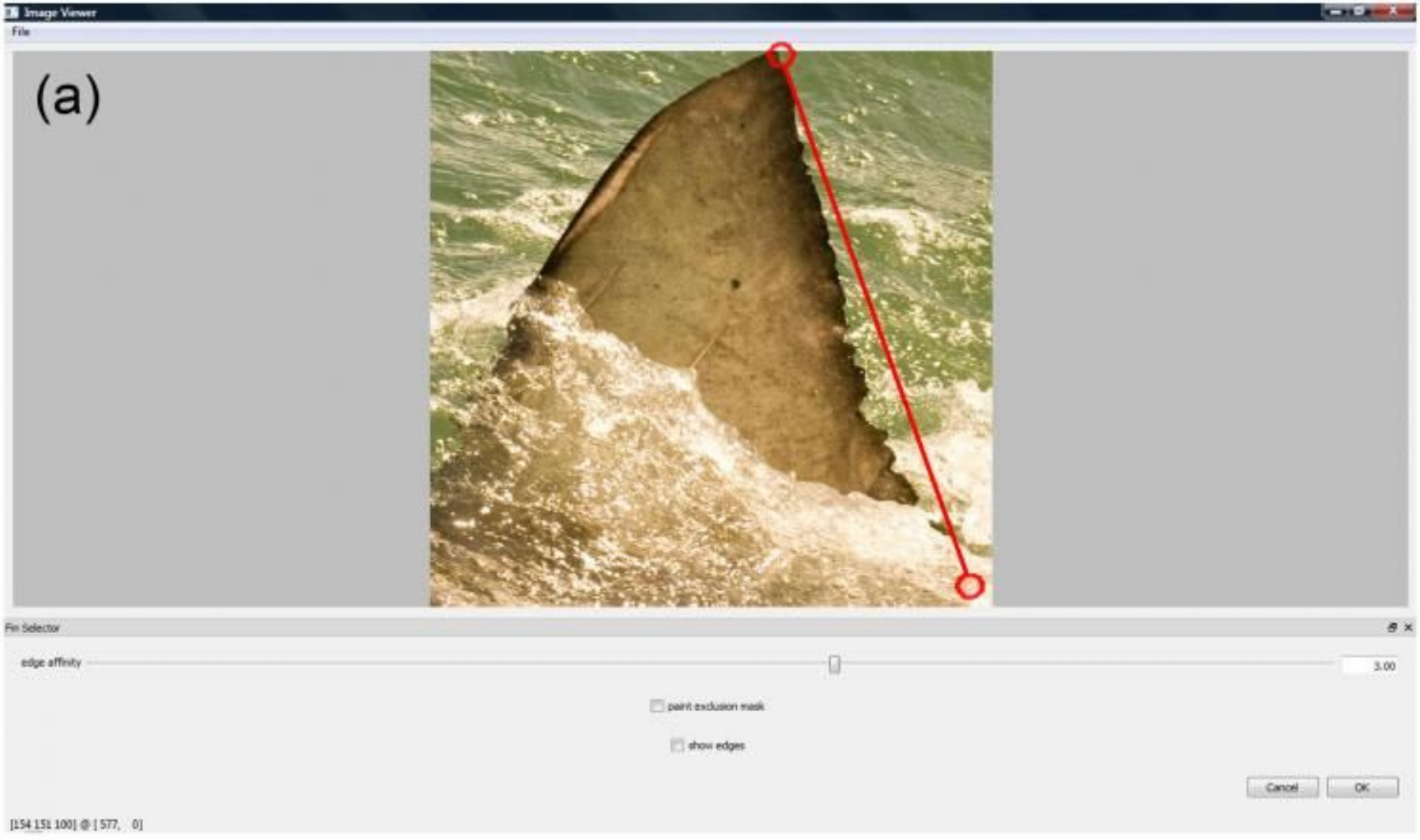


Image Viewer
File
(a)
Pin Selector
edge affinity
5.00
paint exclusion mask
show edges
Cancel
OK
[154 151 100] @ [577, 0]

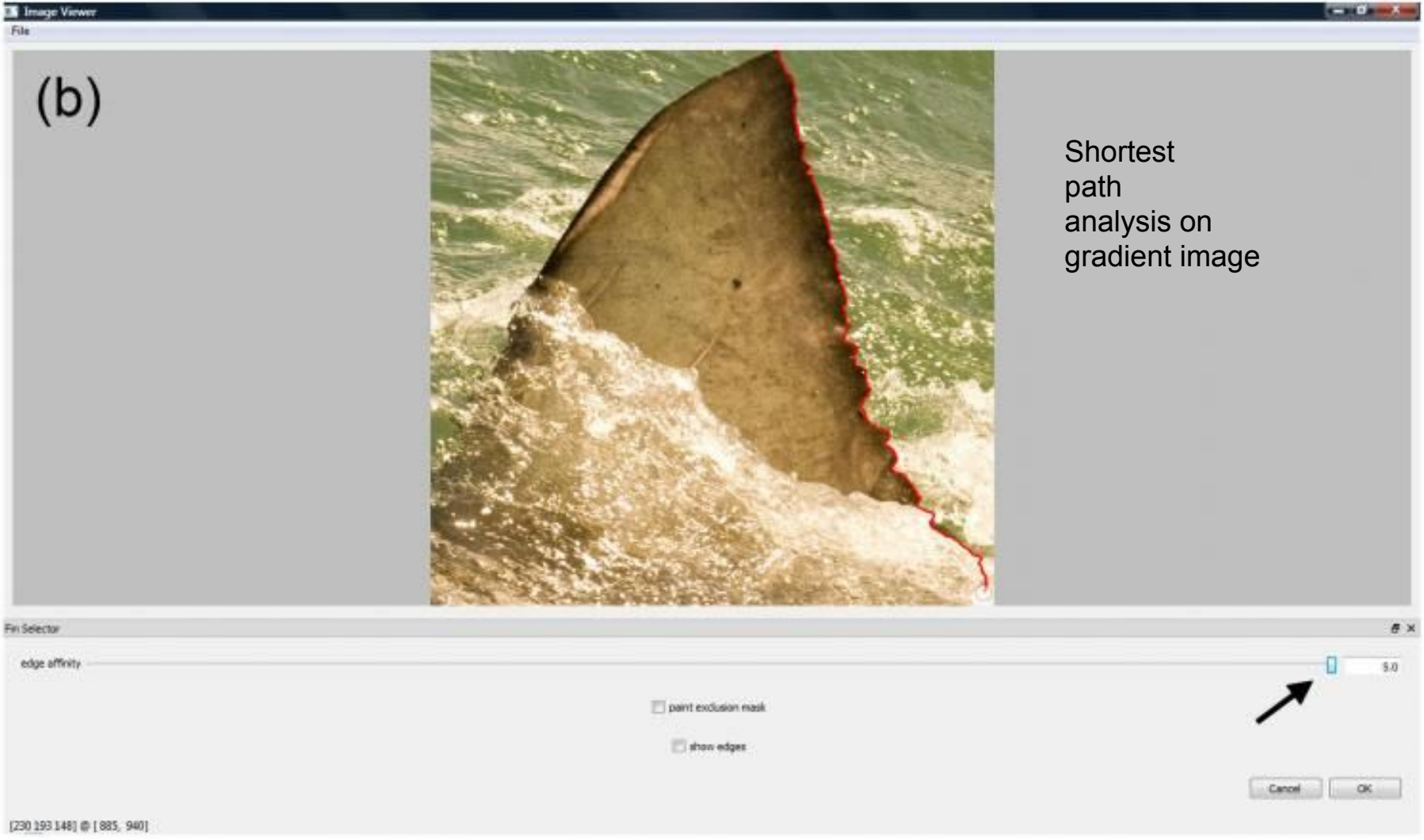
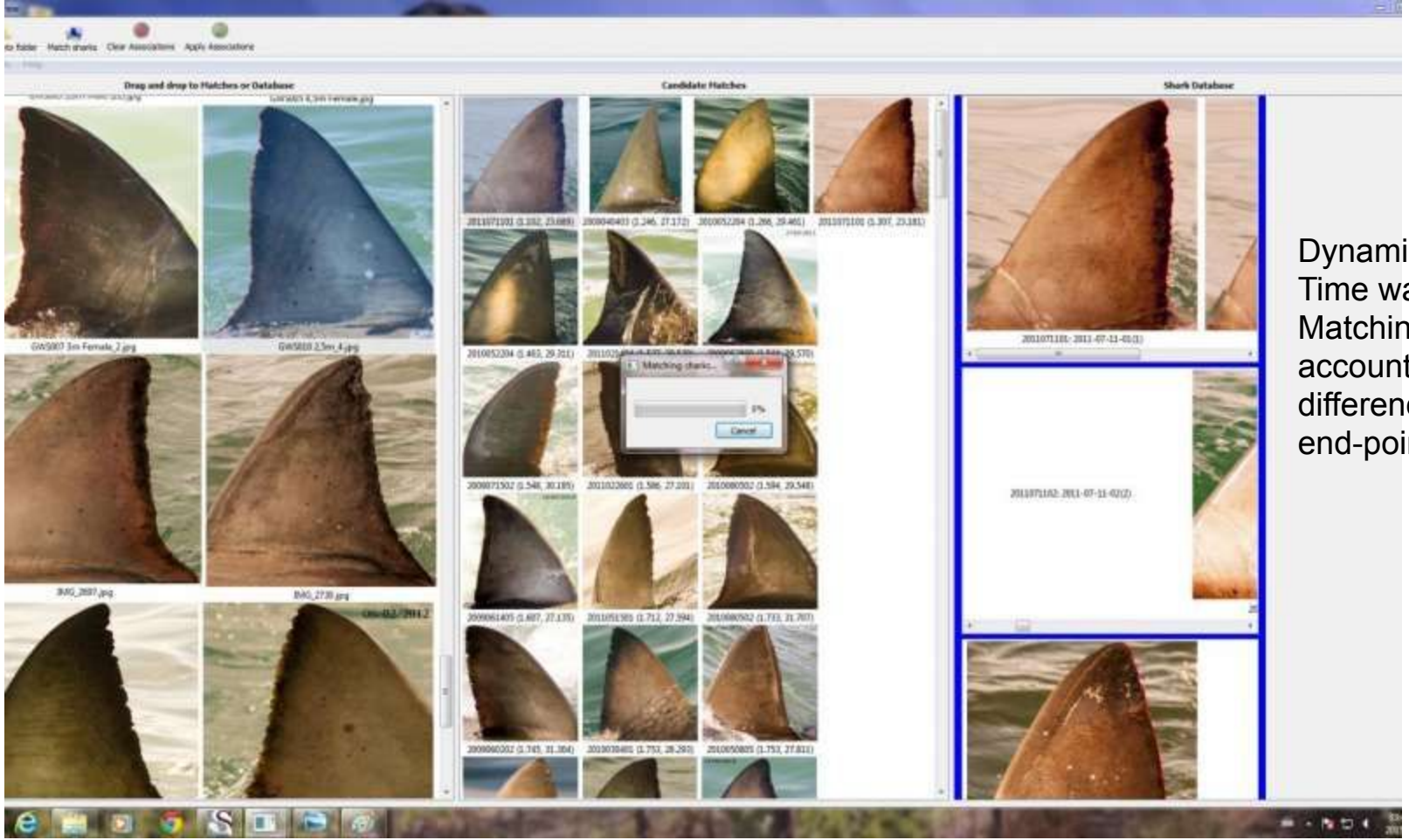


Image Viewer
File
(b)
Shortest
path
analysis on
gradient image
Fin Selector
edge affinity
5.0
paint exclusion mask
show edges
Cancel
OK
[230 193 148] @ [885, 940]

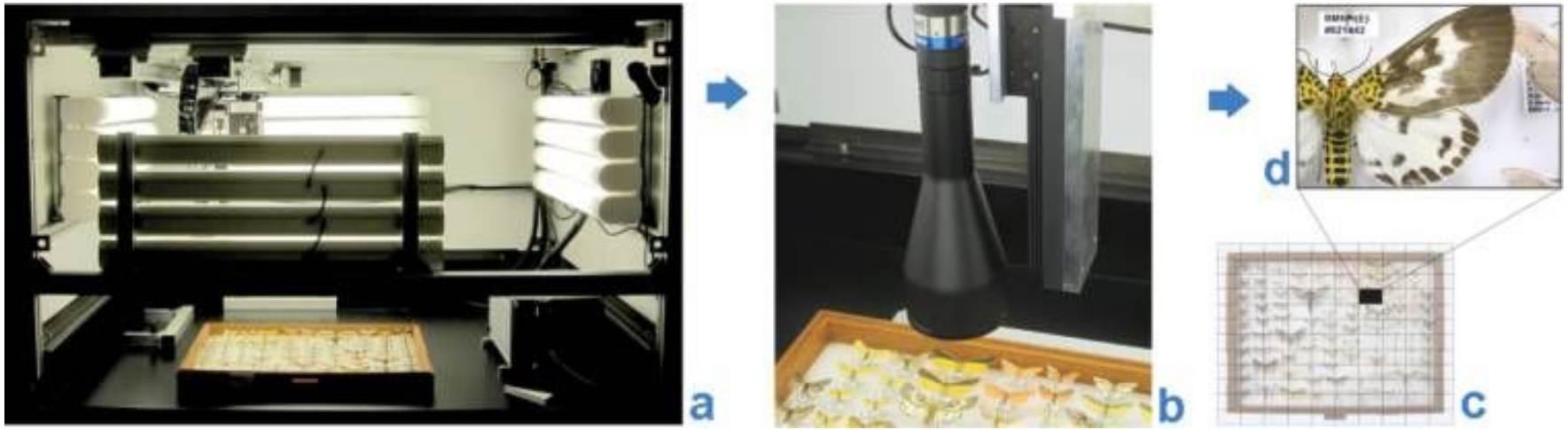


Dynamic Time warping Matching accounting for differences in end-points

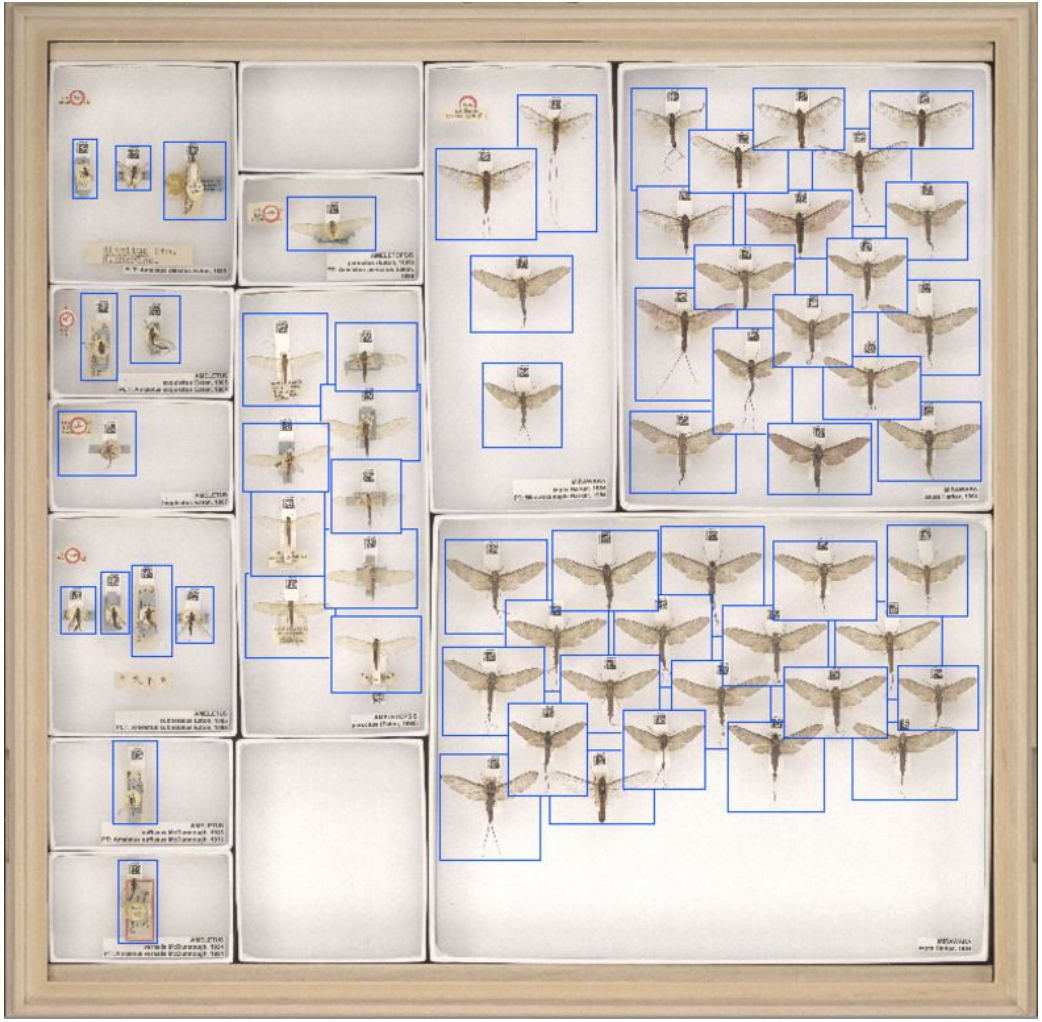
Significant effort went into untangling the pre-existing file system storage scheme!

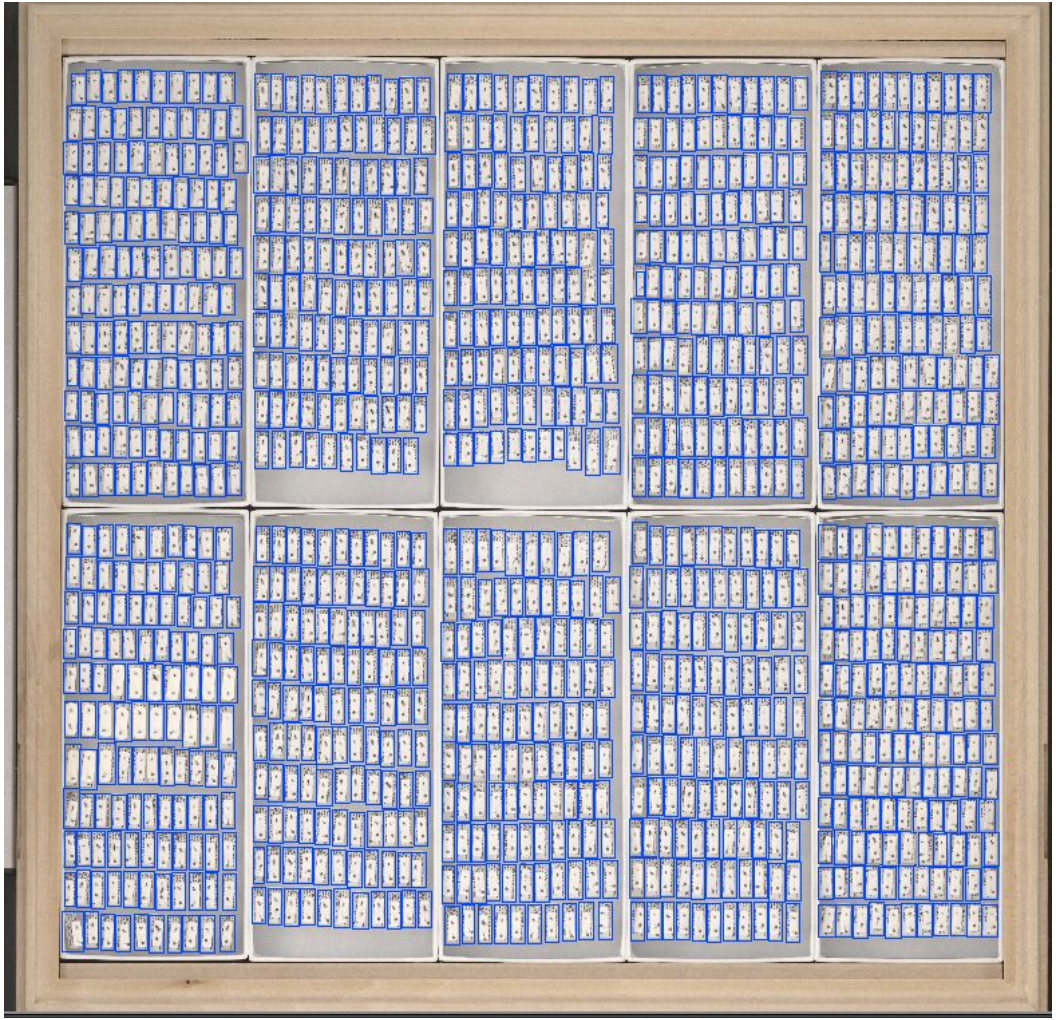
Project 2: **Inselect** insect tray segmentation

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



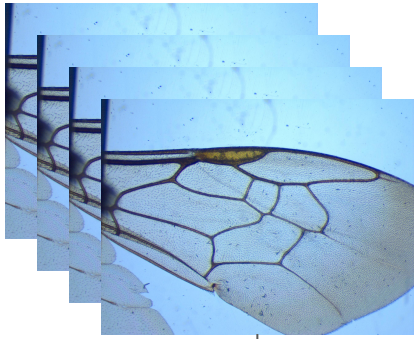
SatScan imaging: **a** SatScan machine **b** specimens being imaged **c** individual frames aligned **d** fragment of a stitched image; final resolution of the stitched image ~11 lines/mm.





Project 3: **Bee** identification from **wing** venation

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



- Agapostemon sericeus
- Bombus bimaculata
- Ceratina calcarata
- Lasioglossum acuminatum

Training

Feature extraction

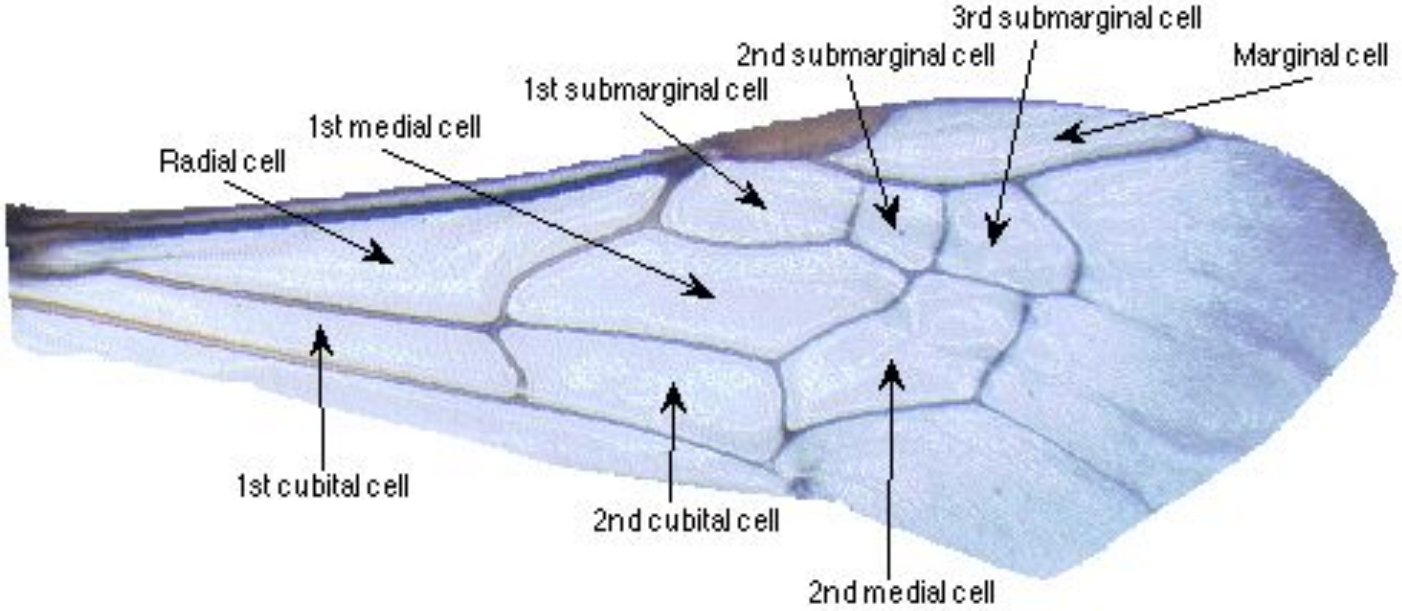
Classifier

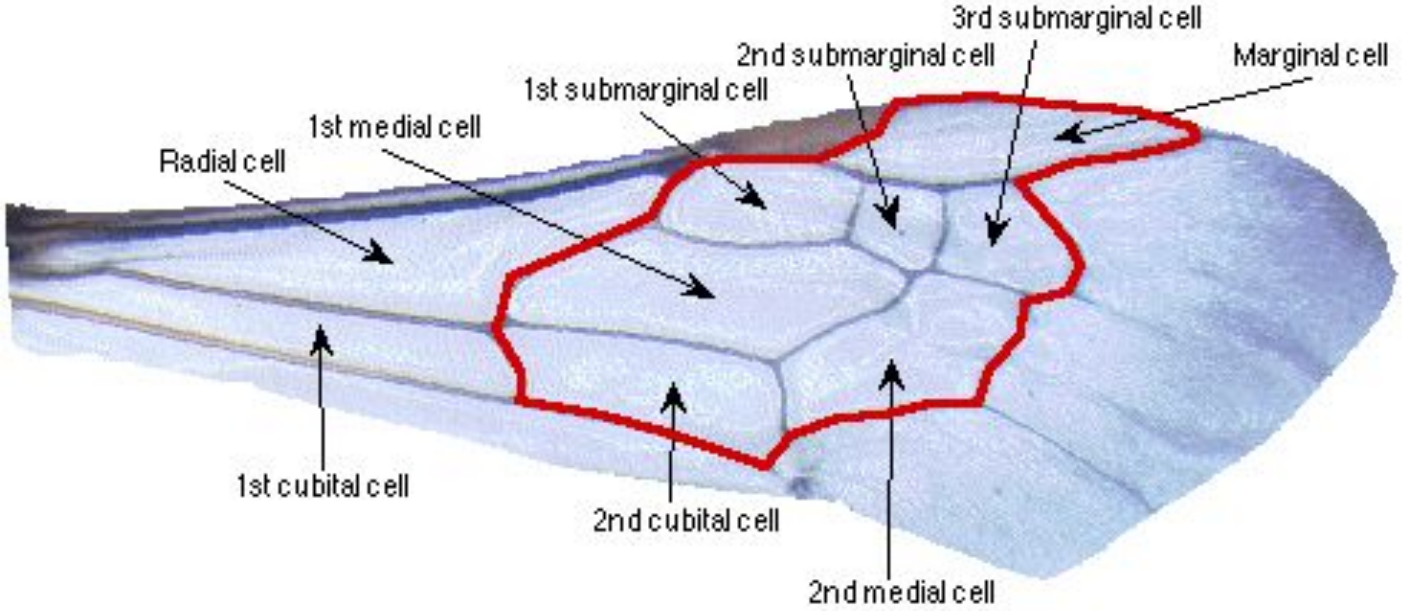
- Agapostemon sericeus
- Agapostemon texanus
- Agapostemon virescens
- Bombus bimaculata
- Bombus griseocolis
- Bombus impatiens
- Ceratina calcarata
- Lasioglossum acuminatum
- Lasioglossum coriaceum
- Lasioglossum leucozonium
- Lasioglossum mawispb
- Lasioglossum nymphaerum
- Lasioglossum pilosum
- Lasioglossum rohweri
- Lasioglossum zephyrum
- Lasioglossum zonulum
- Osmia coloradensis
- Osmia lignaria
- Osmia pusilla
- Osmia ribifloris
- Osmia texana

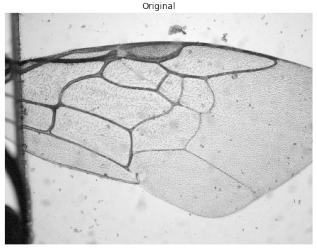
Feature extraction pipeline

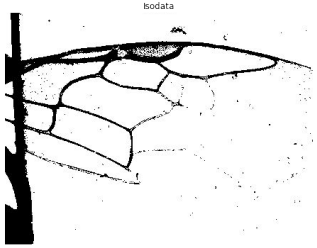
To use with classic machine learning models, we need features from images.

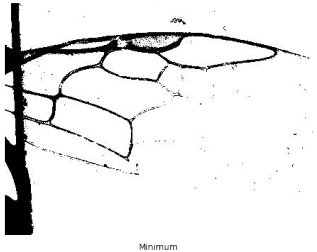
[0.5, 12.3, 7.8, …, 10.3]

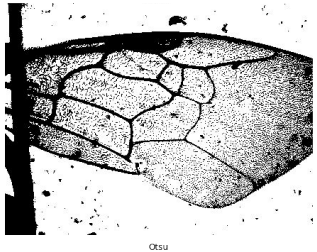

3rd submarginal cell
2nd submarginal cell
Marginal cell
1st submarginal cell
1st medial cell
Radial cell
1st cubital cell
2nd cubital cell
2nd medial cell

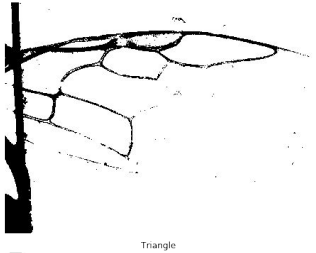

3rd submarginal cell
2nd submarginal cell
Marginal cell
1st submarginal cell
1st medial cell
Radial cell
1st cubital cell
2nd cubital cell
2nd medial cell

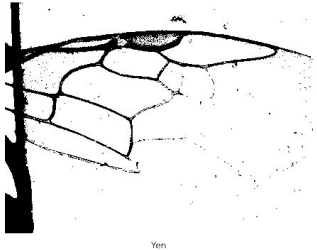

Original

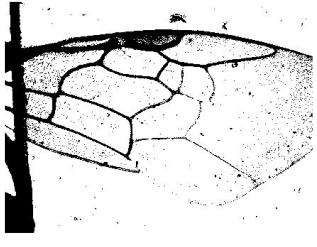

Isodata

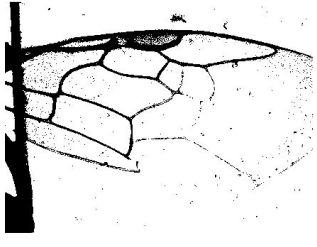

Li

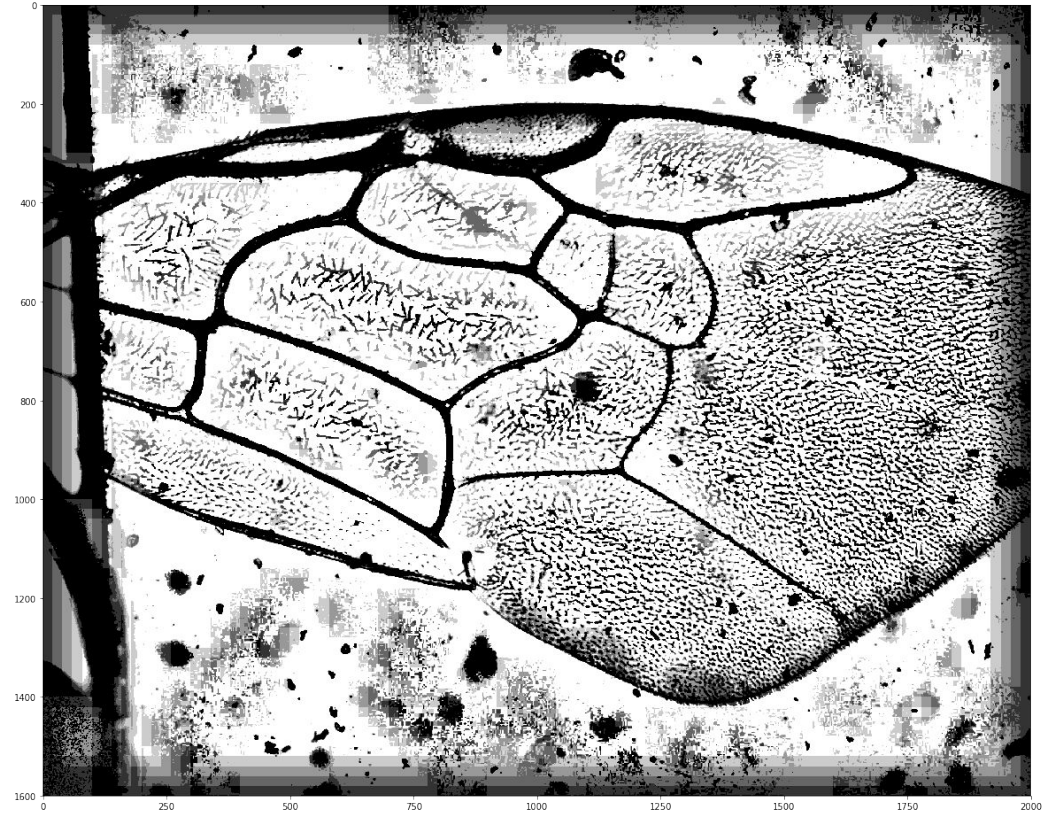

Mean


Minimum


Otsu

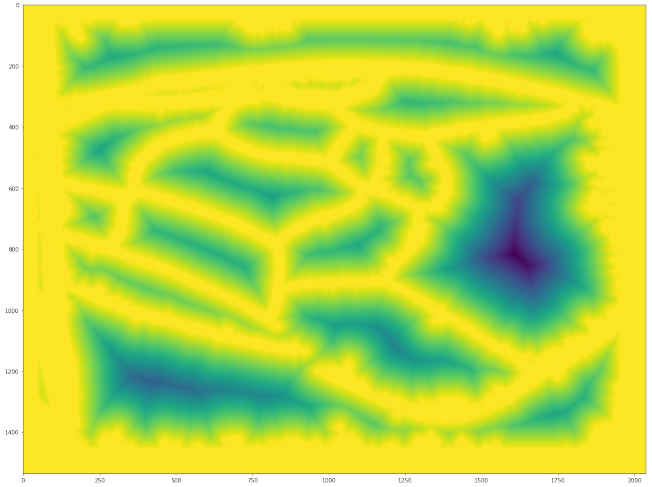

Triangle

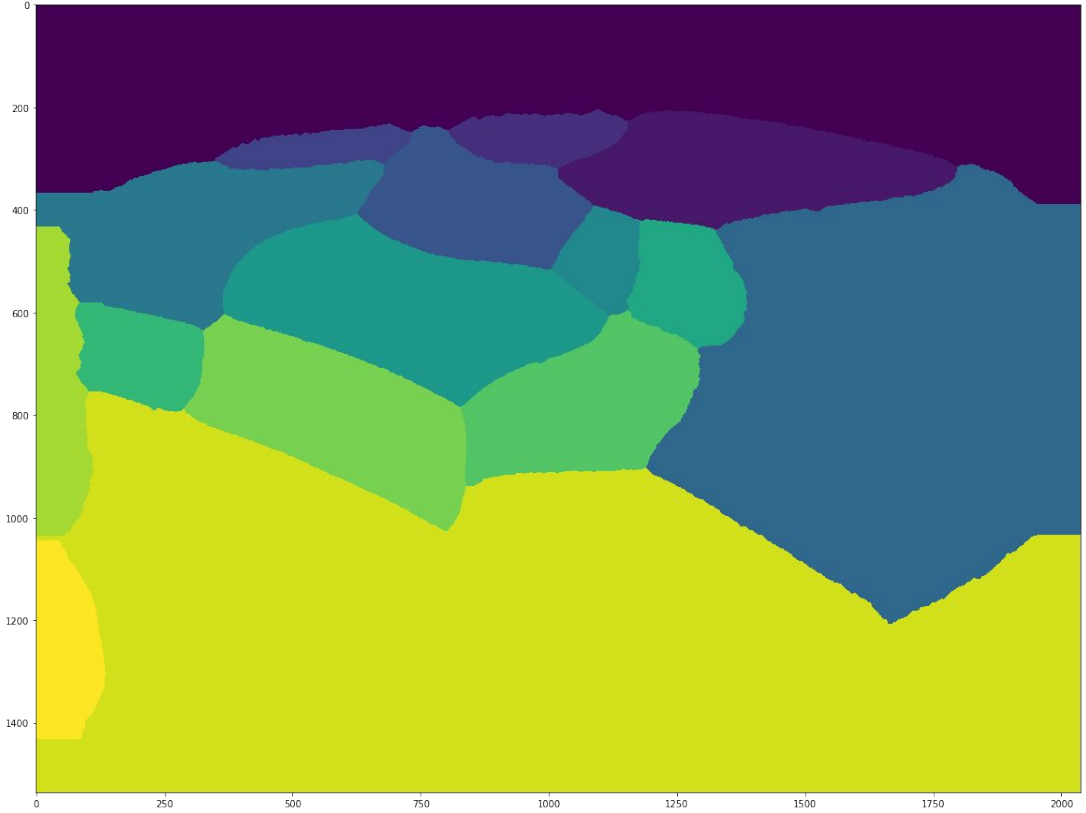

Yen

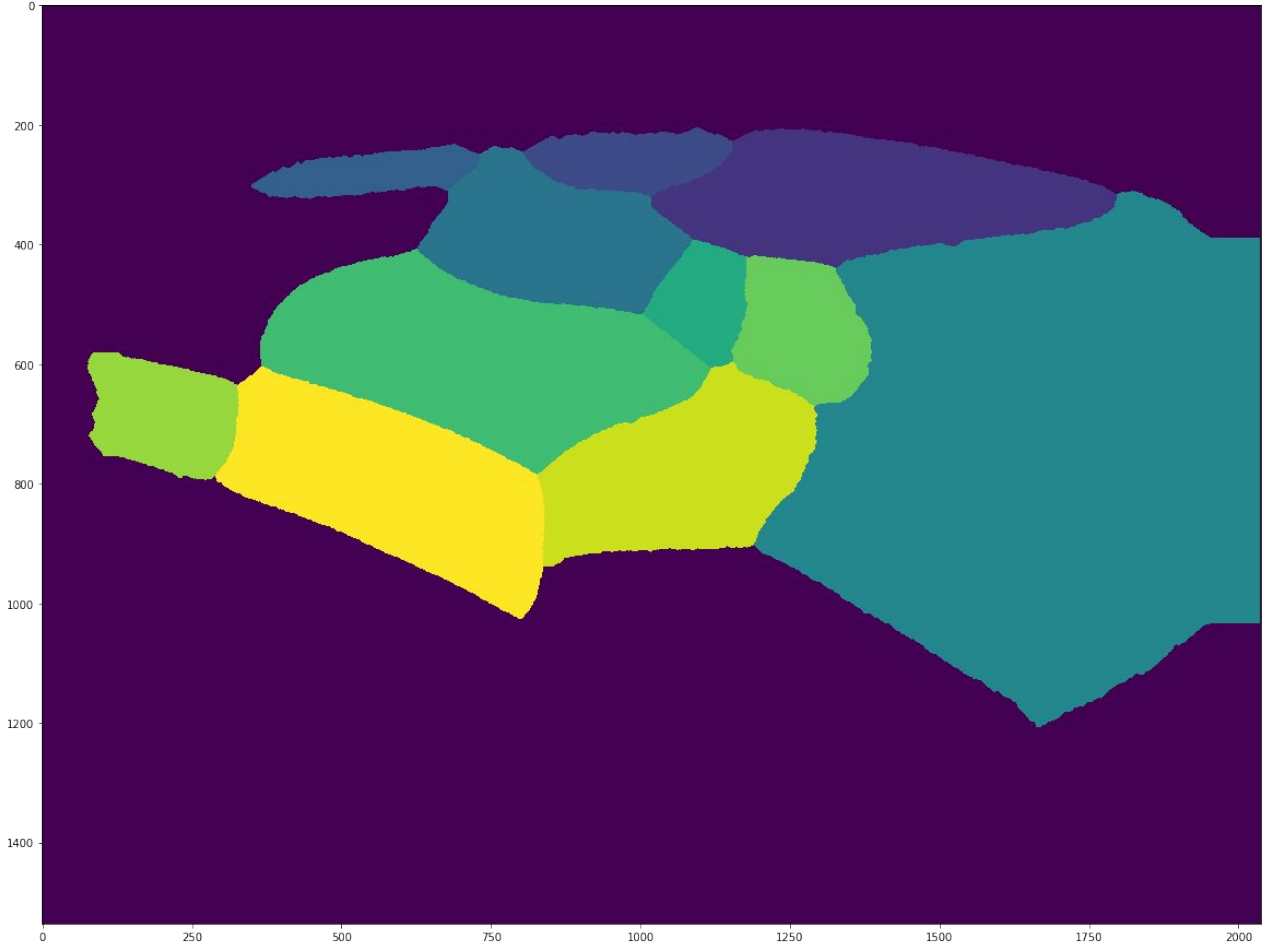

0
200
400
600
800
1000
1200
1400
1600
0
250
500
750
1000
1250
1500
1750
2000

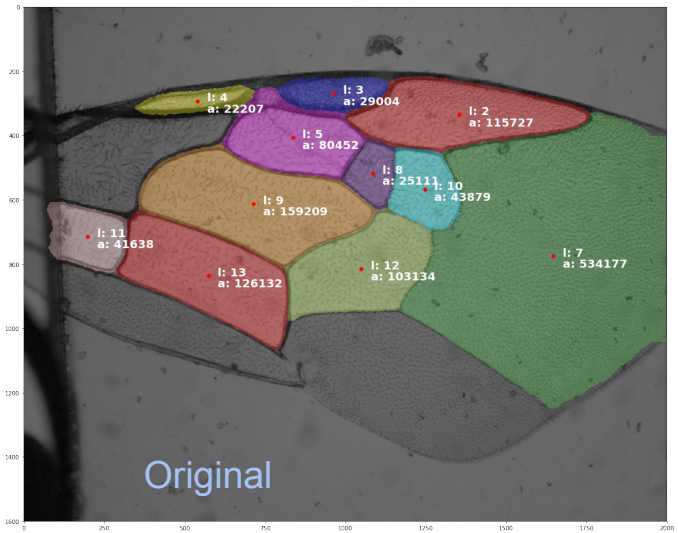

0
200
400
600
800
1000
1200
1400
0
250
500
750
1000
1250
1500
1750
2000

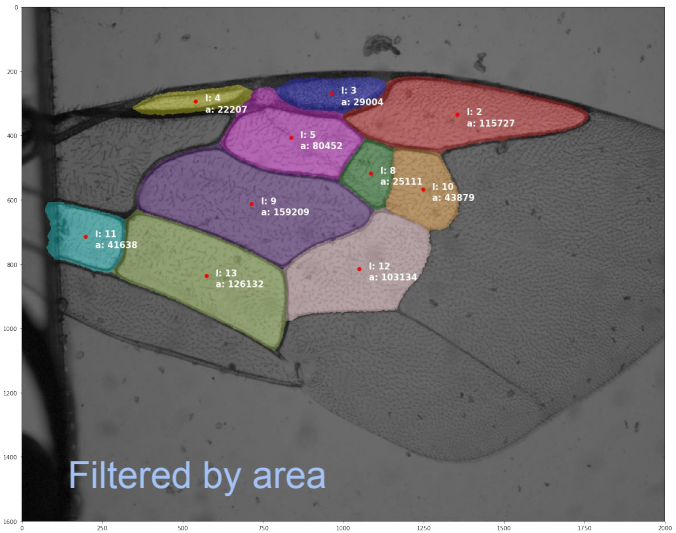

0
200
400
600
800
1000
1200
1400
0
250
500
750
1000
1250
1500
1750
2000

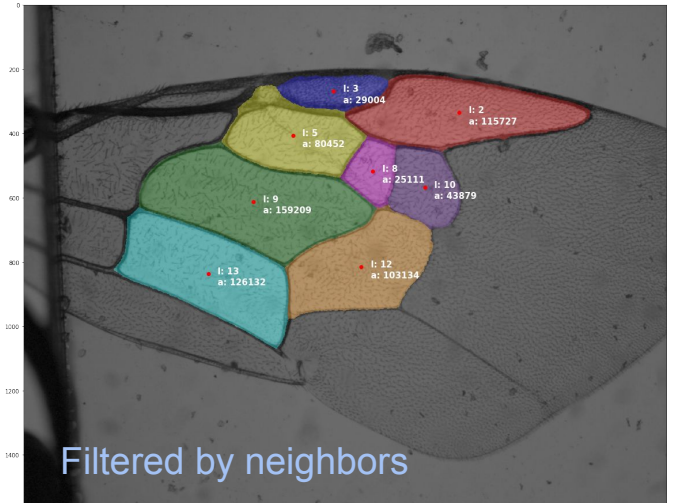

0
200
400
600
800
1000
1200
1400
0
250
500
750
1000
1250
1500
1750
2000

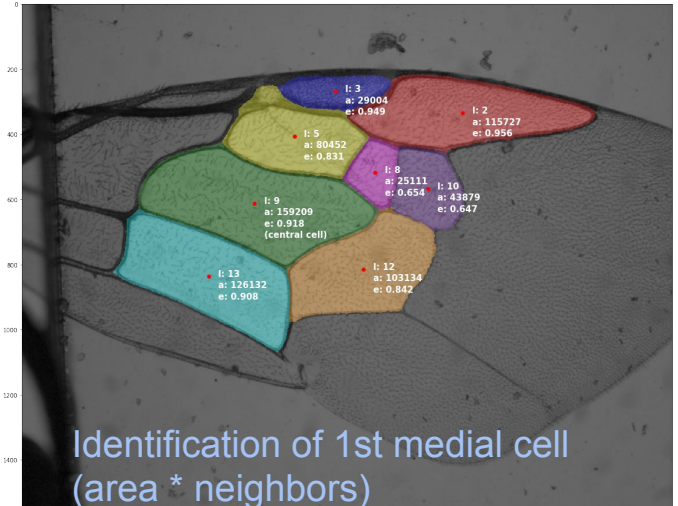

0
200
400
600
800
1000
1200
1400
0
250
500
750
1000
1250
1500
1750
2000

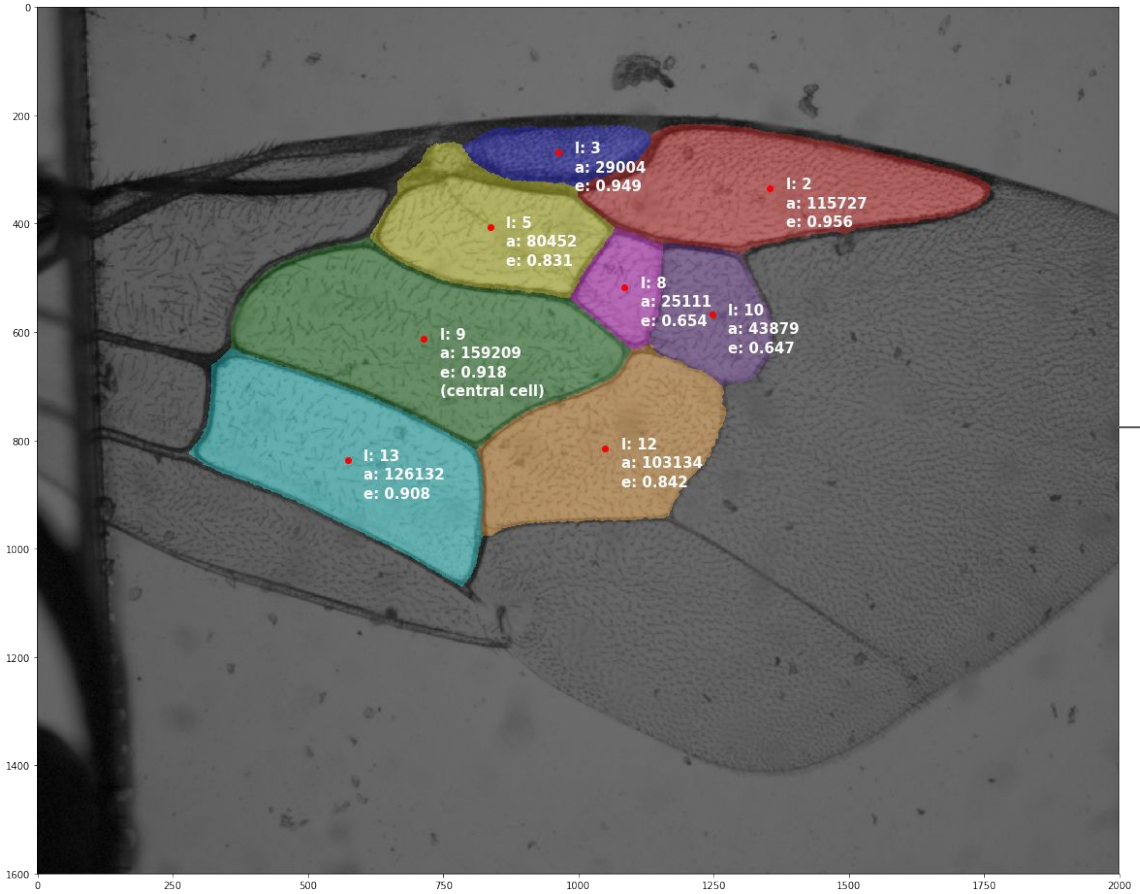

0
200
400
600
800
1000
1200
1400
0
250
500
750
1000
1250
1500
1750
2000


l: 4
a: 22207
l: 3
a: 29004
l: 2
a: 115727
l: 5
a: 80452
l: 8
a: 25111
l: 10
a: 43879
l: 9
a: 159209
l: 11
a: 41638
l: 13
a: 126132
l: 12
a: 103134
l: 7
a: 534177
Original


l: 4
a: 22207
l: 3
a: 29004
l: 2
a: 115727
l: 5
a: 80452
l: 8
a: 25111
l: 10
a: 43879
l: 9
a: 159209
l: 11
a: 41638
l: 13
a: 126132
l: 12
a: 103134
Filtered by area


l: 3
a: 29004
l: 2
a: 115727
l: 5
a: 80452
l: 8
a: 25111
l: 10
a: 43879
l: 9
a: 159209
l: 13
a: 126132
l: 12
a: 103134
Filtered by neighbors


l: 3
a: 29004
e: 0.949
l: 2
a: 115727
e: 0.956
l: 5
a: 80452
e: 0.831
l: 8
a: 25111
e: 0.654
l: 10
a: 43879
e: 0.647
l: 9
a: 159209
e: 0.918
(central cell)
l: 13
a: 126132
e: 0.908
l: 12
a: 103134
e: 0.842
Identification of 1st medial cell
(area * neighbors)



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.

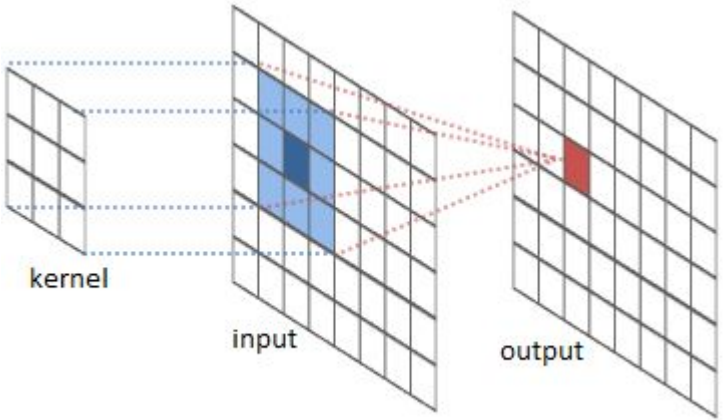
Or...

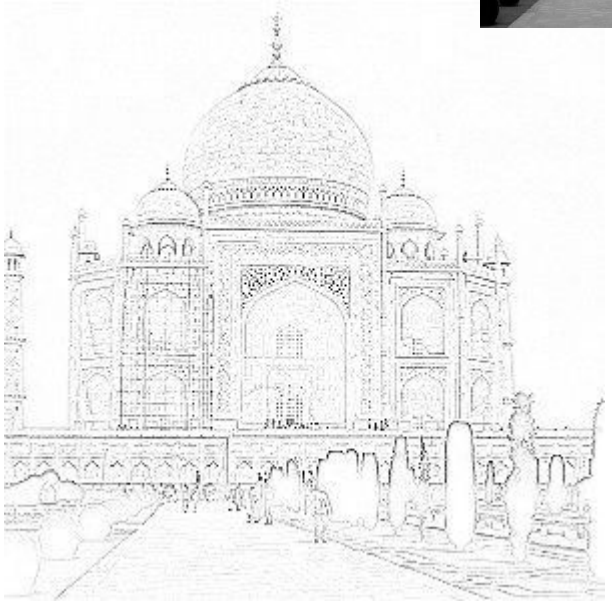
...forget *all* that

and use a convolutional neural network.

¯_(ツ)_/¯

	0	1	0	
	1	-4	1	
	0	1	0	

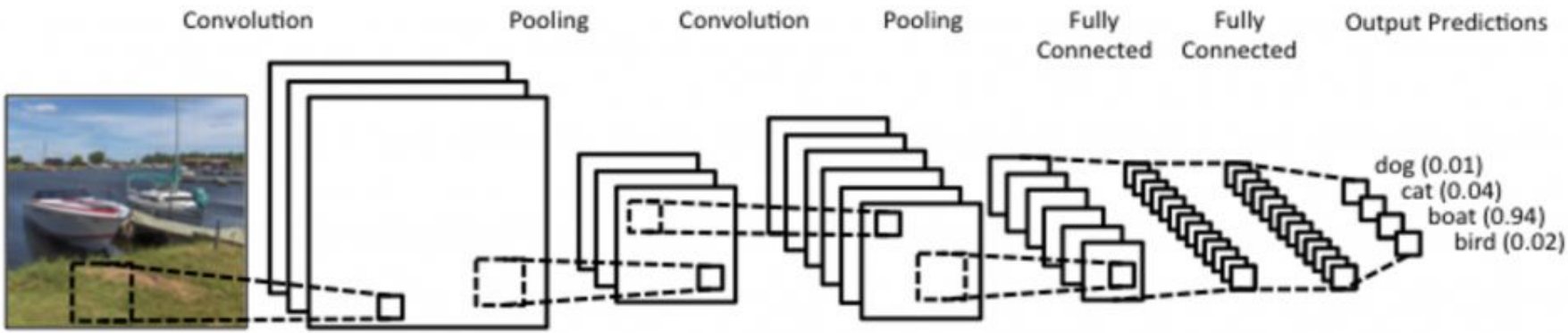




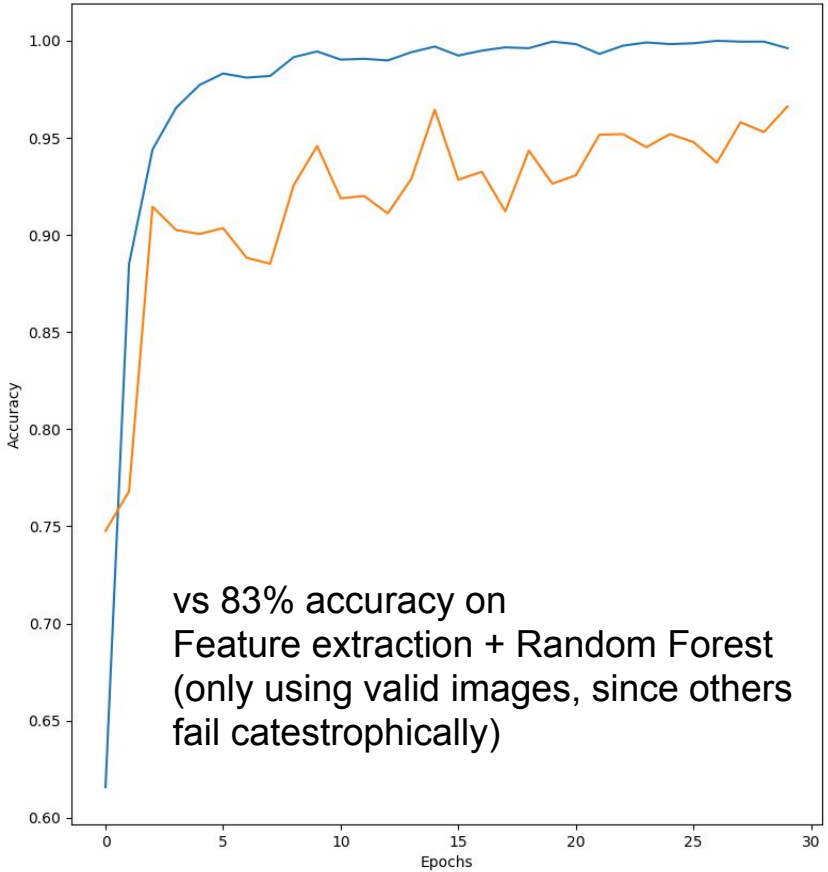
http://intellabs.github.io/RiverTrail/tutorial/

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

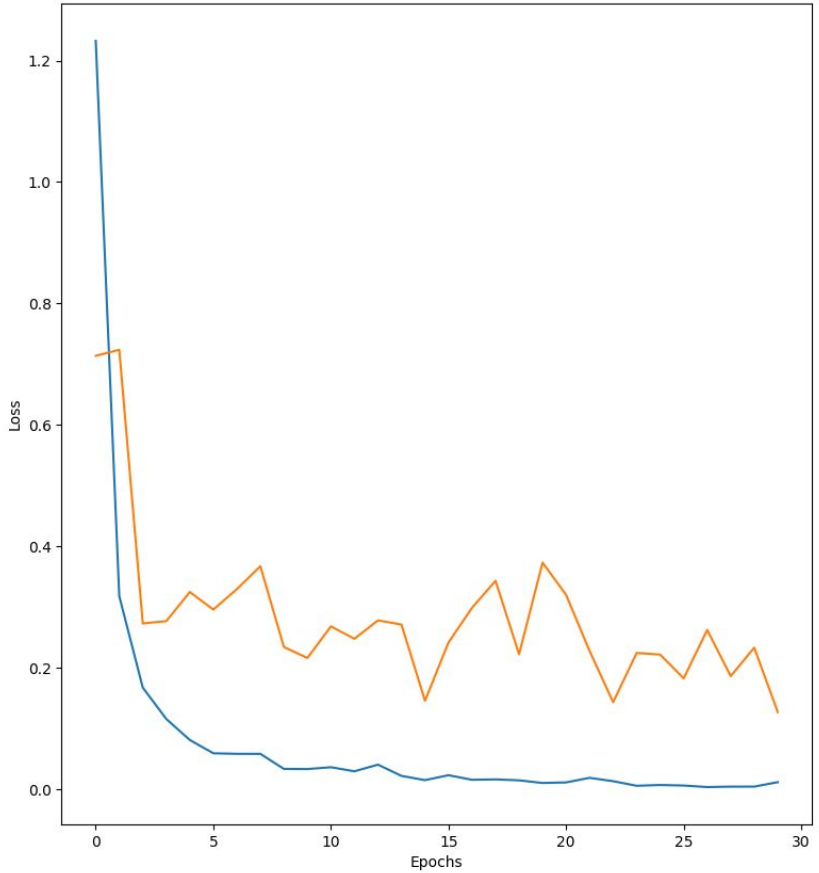
Example architecture (for illustration, we used DenseNet instead)



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



vs 83% accuracy on
Feature extraction + Random Forest
(only using valid images, since others fail catestrophically)



Advantages:

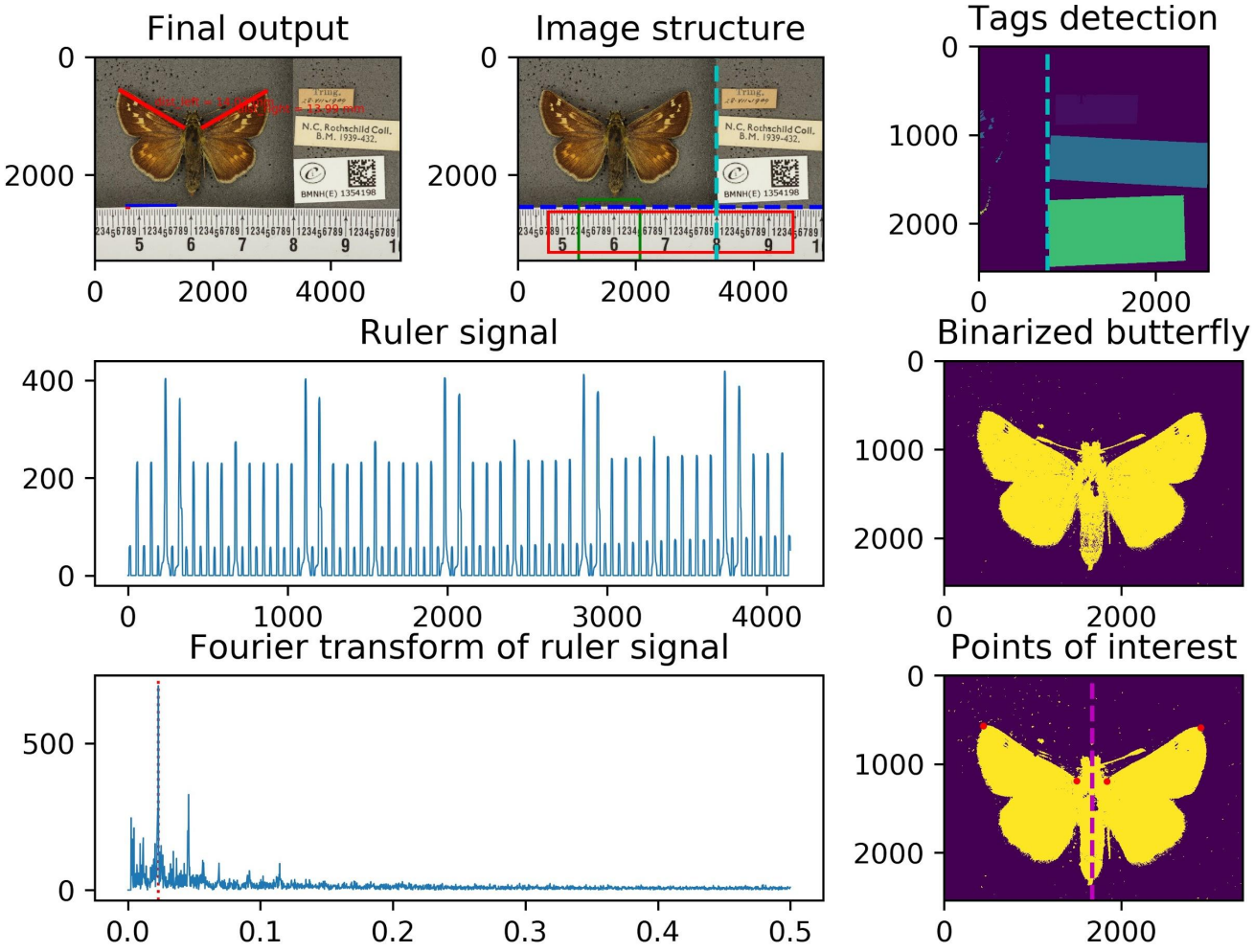
- Raw data as input (perhaps augmented); no feature engineering
- Network can be engineered; little domain knowledge needed

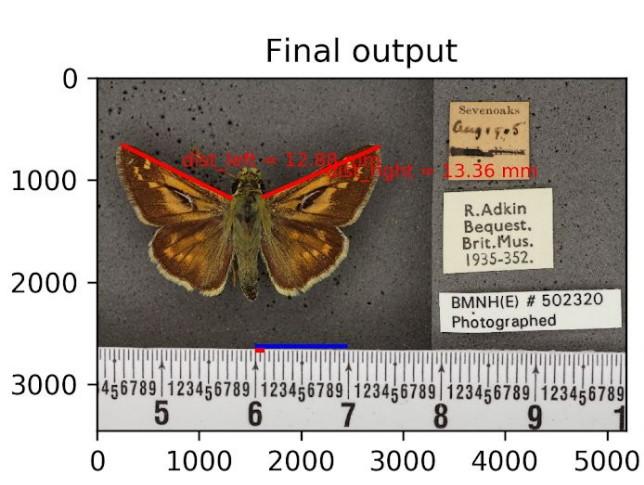
Disadvantages:

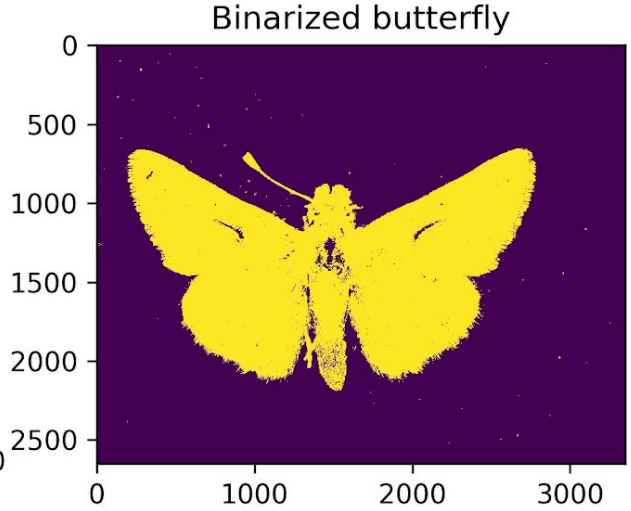
- Unknown failure modes; operates well only within reaches of seen data
- Working is not easily interpretable
- Often requires a significant amount of training data

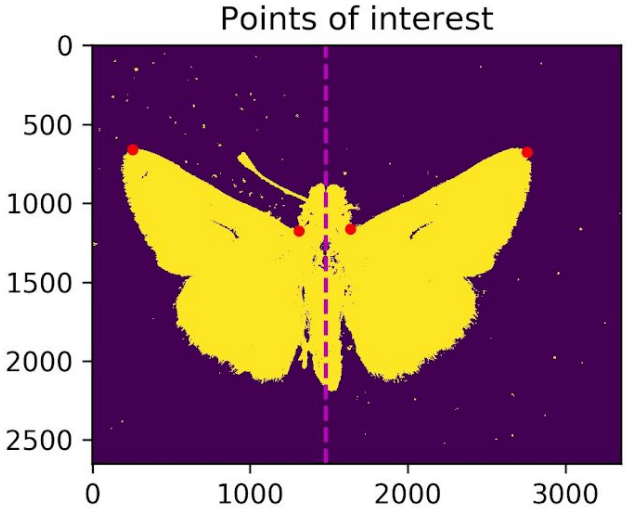
Project 4: Butterfly measurement

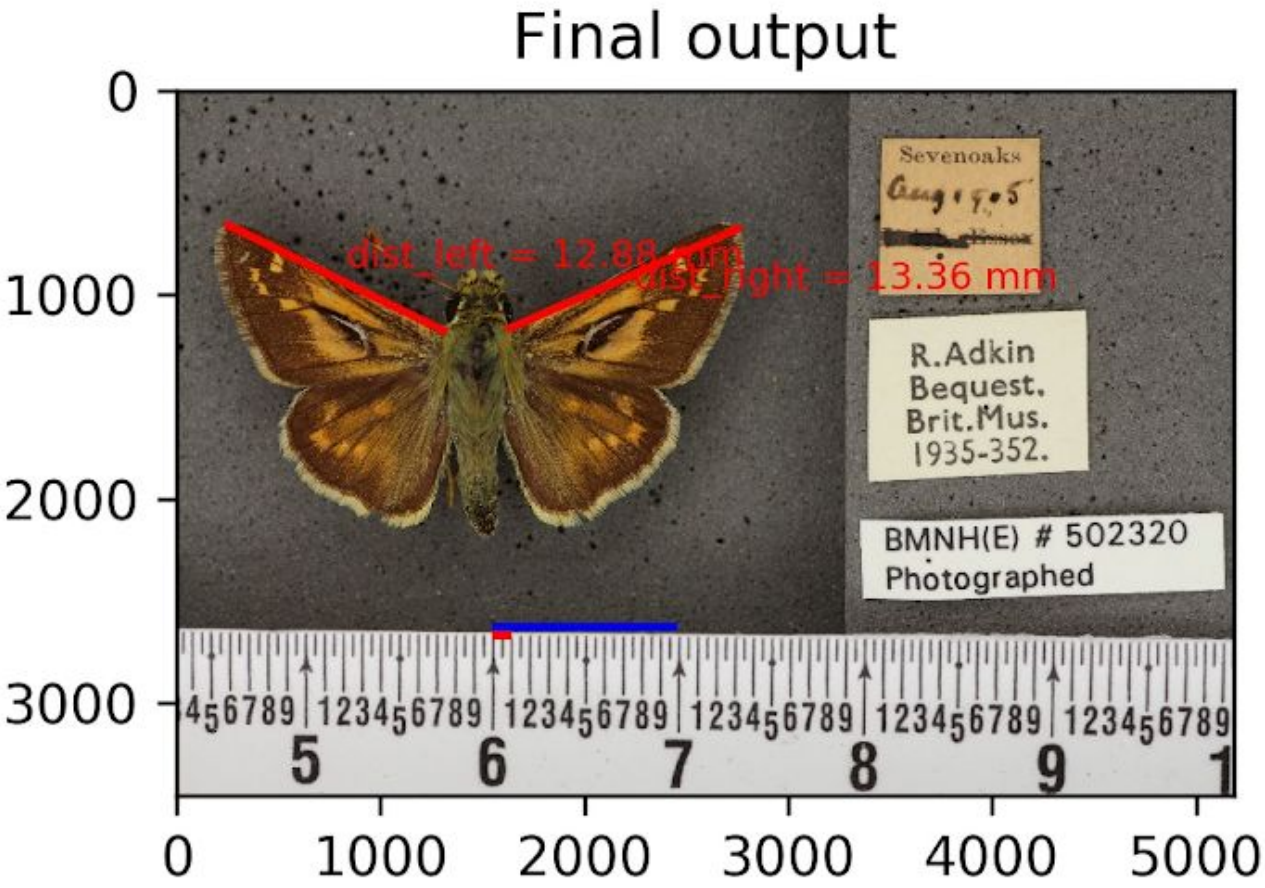
With Ben Price (NHM), Phil Fenberg (Southampton/NHM), Théo Bodrito (Mines ParisTech), Dennis Feng (UC Berkeley), Rebecca Wilson (Southampton/NHM), Hannah O'Sullivan (NHM), Steve Brooks (NHM), BIDS Machine Shop (UC Berkeley)

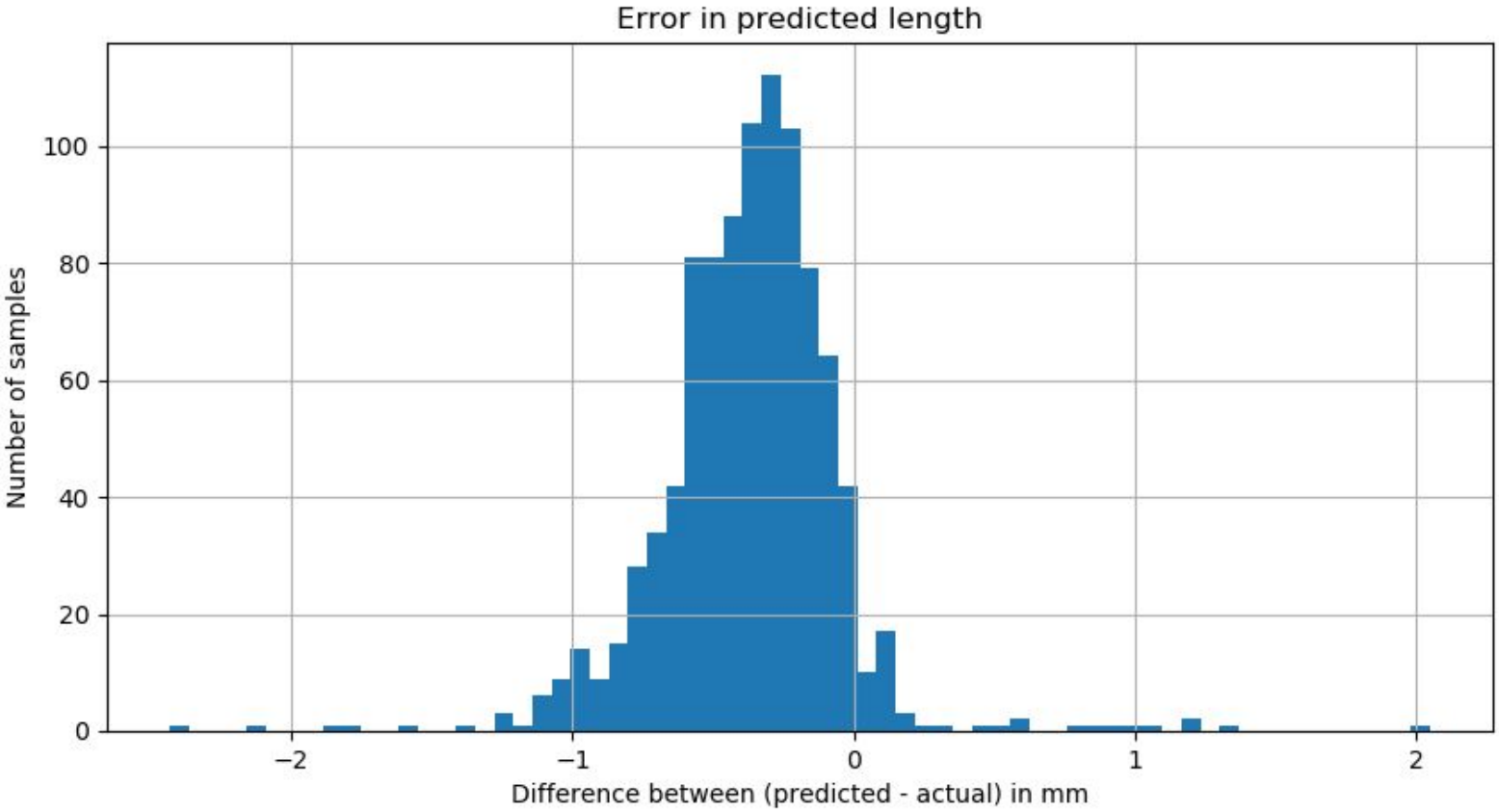

Final output
Image structure
Tags detection
Ruler signal
Binarized butterfly
Fourier transform of ruler signal
Points of interest
N.C. Rothschild Coll. B.M. 1939-432.
BMNH(E) 1354198
0
2000
4000
1000
400
200
500
0.0
0.1
0.2
0.3
0.4
0.5

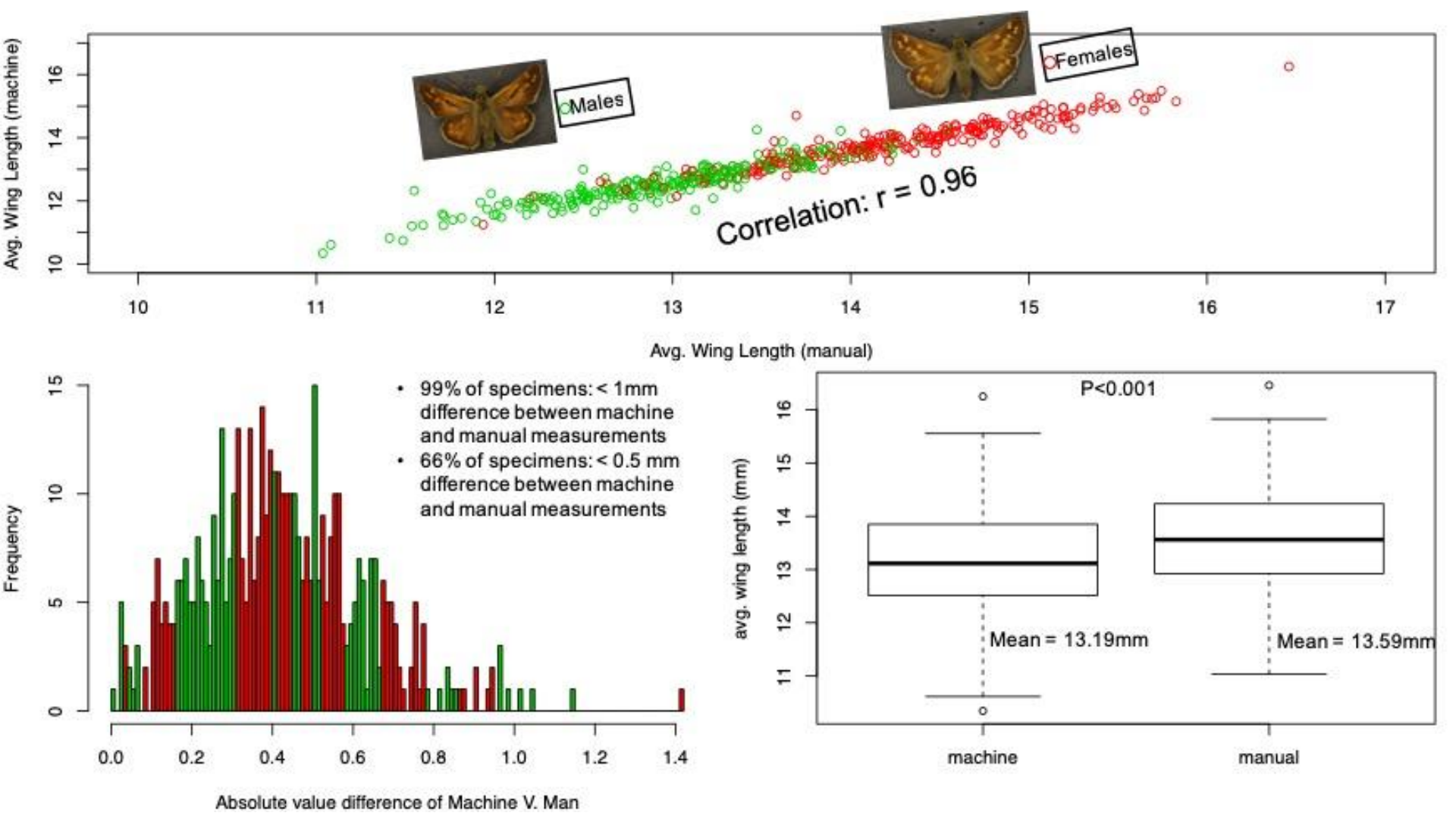

Final output
dist_left = 12.88 mm
dist_right = 13.36 mm
Sevenoaks
R.Adkin
Bequest.
Brit.Mus.
1935-352.
BMNH(E) # 502320
Photographed
0
1000
2000
3000
0
1000
2000
3000
4000
5000


Binarized butterfly
0
500
1000
1500
2000
2500
0
1000
2000
3000


Points of interest
0
500
1000
1500
2000
2500
0
1000
2000
3000


Final output
0
1000
2000
3000
0
1000
2000
3000
4000
5000
dist_left = 12.88 mm
dist_right = 13.36 mm
Sevenoaks
R.Adkin
Bequest.
Brit.Mus.
1935-352.
BMNH(E) # 502320
Photographed


Error in predicted length
Number of samples
0
20
40
60
80
100
−2
−1
0
1
2
Difference between (predicted - actual) in mm


Avg. Wing Length (machine)
Avg. Wing Length (manual)
Males
Females
Correlation: r = 0.96
Frequency
Absolute value difference of Machine V. Man
• 99% of specimens: < 1mm difference between machine and manual measurements
• 66% of specimens: < 0.5 mm difference between machine and manual measurements
avg. wing length (mm)
P<0.001
Mean = 13.19mm
Mean = 13.59mm
machine
manual

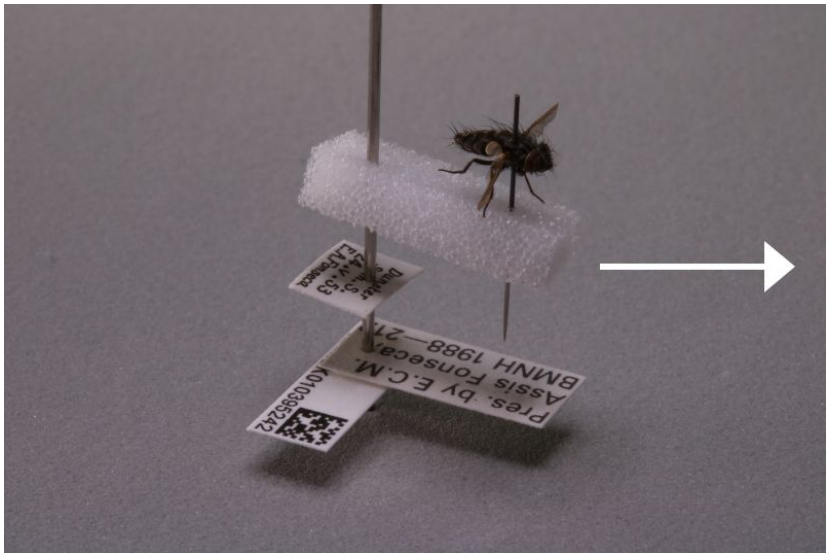
ALICE: Angled Label Image Capture and Extraction for high throughput insect specimen digitisation

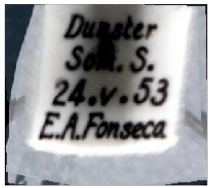
AUTHORS
Benjamin Price, Steen Dupont, Elizabeth Allan, Vladimir Blagoderov, Alice Butcher, James Durrant , Pieter Holtzhausen , Phaedra Kokkini, Laurence Livermore, Helen Hardy, Vincent Smith

CREATED ON
November 05, 2018

LAST EDITED
November 05, 2018

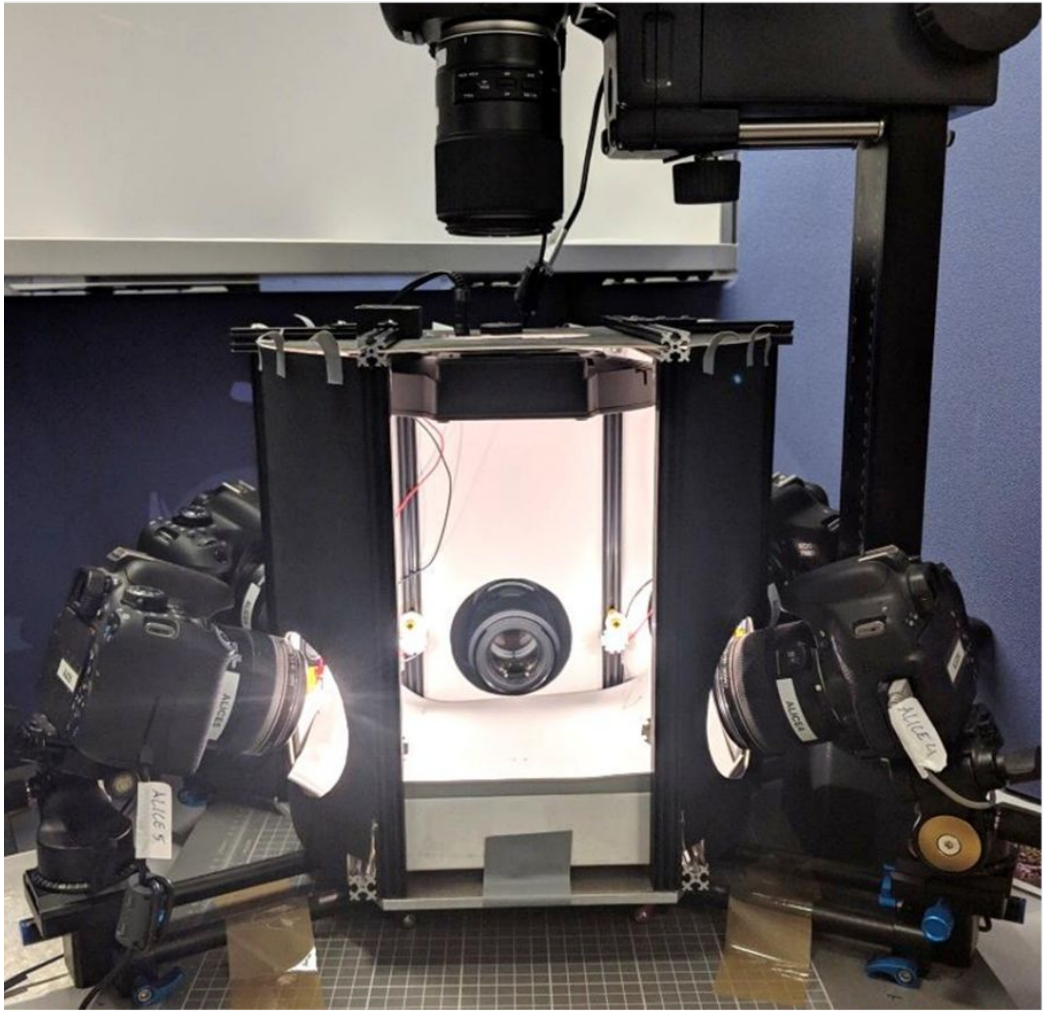
SUPPLEMENTAL MATERIALS
osf.io/9p4f6/



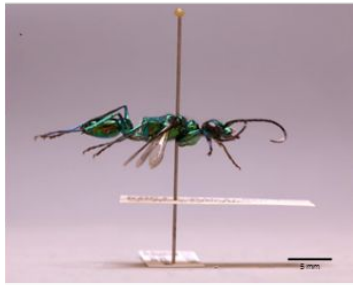


Future work

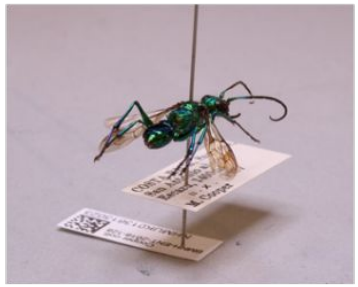
https://github.com/NaturalHistoryMuseum/ALICE

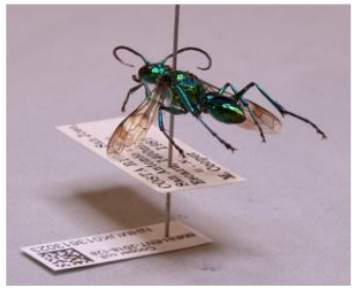


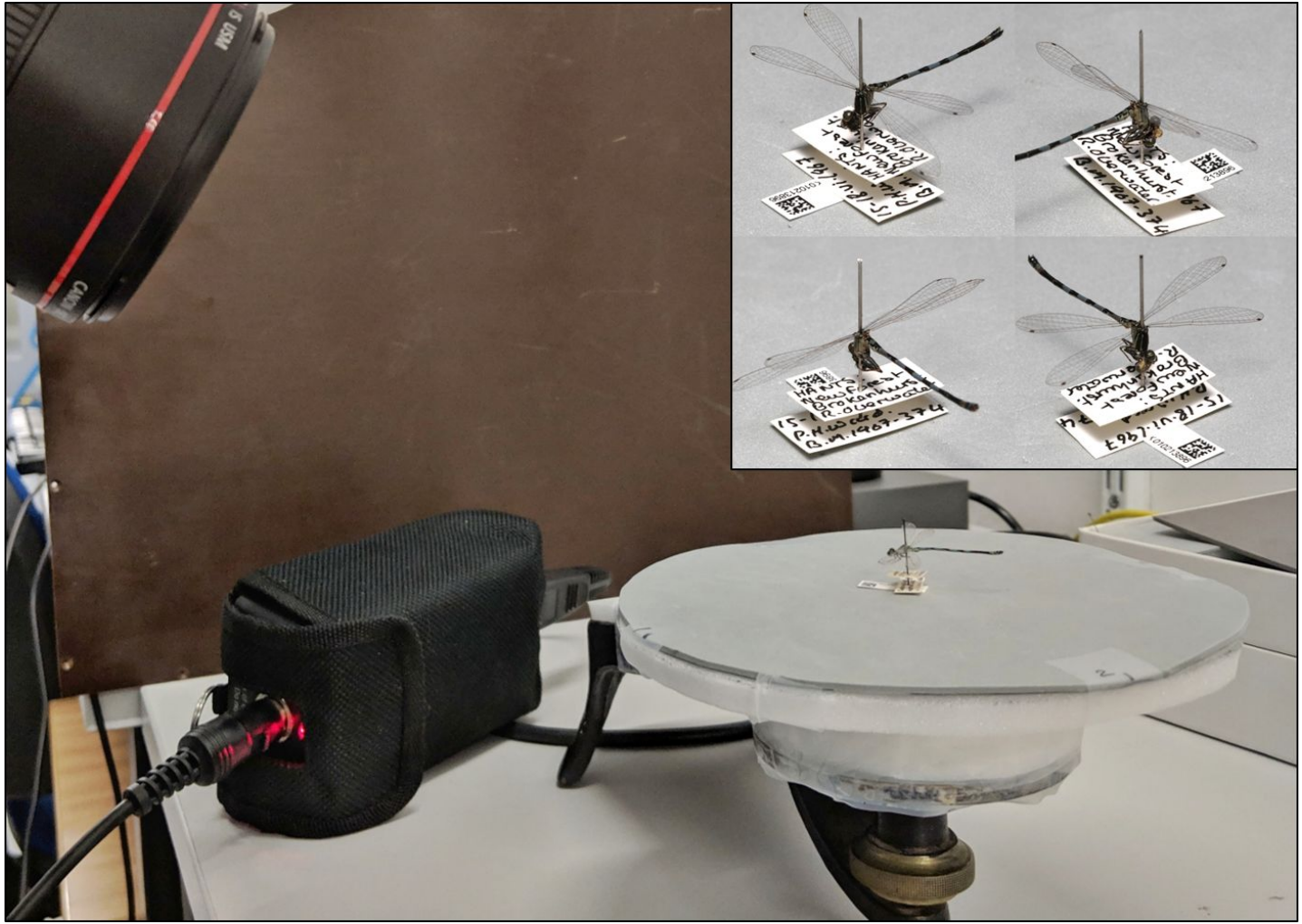
Specimen Images



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.