Automating Scientific Image Annotation workflow

Searching the "Algorithm Space" to replace engineering jobs with machine learning

Dirk Colbry
Computational Mathematics Science and Engineering

Agenda

- Scientific Image Analysis (SIA)
- Building tools for SIA
- Why existing tools are limiting
- Rescoping the problem
- My approach using Genetic Algorithms

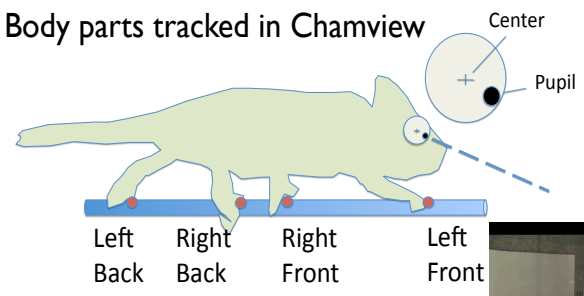
Phenomics

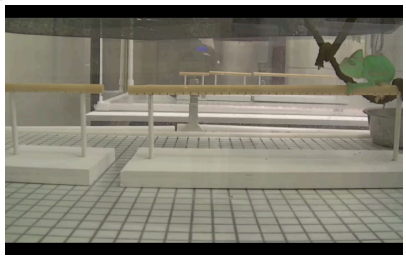
Area of biology concerned with the measurement of phenomes — the physical and biochemical traits of organisms — as they change in response to genetic mutation and environmental influences.

Houle, D.; Govindaraju, D.R.; Omholt, S. (2010), "Phenomics: the next challenge", Nature Reviews Genetics 11 (12): 855–66

Animal Behavior

Dr. Fred Dyer, MSU





Visual Science

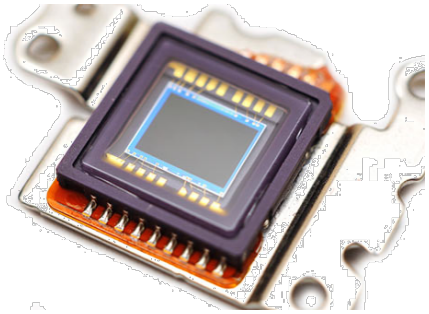
- Long history in Biology
- Traditionally done by hand

Photography

- Changing science
  - Scientists are able to record video without knowing what they will see
  - Cameras may see something the scientists missed
  - Different scientists can view the same data with different scientific questions in mind

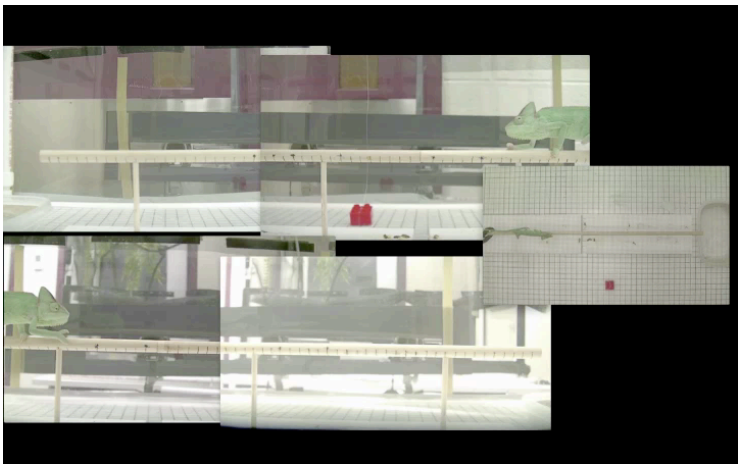
Cameras Everywhere

- Transforming how scientists gather data
- Very affordable
- Data is becoming very cheap to gather, so there is a lot more of it



Charge-Coupled Device (CCD)

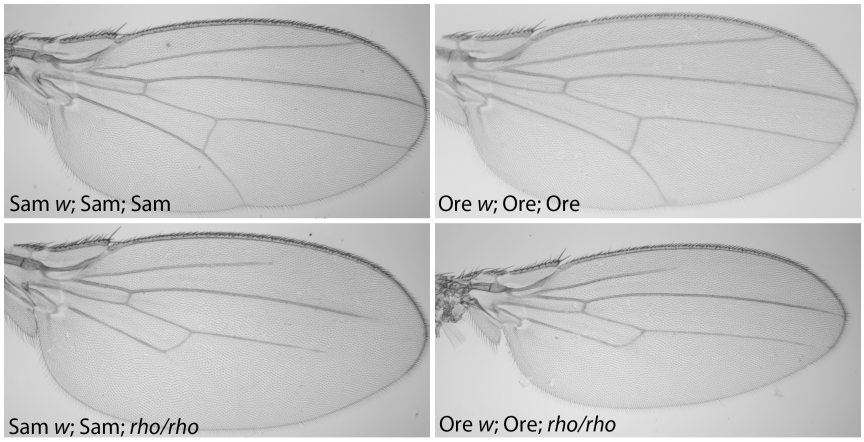
Even small projects grow fast



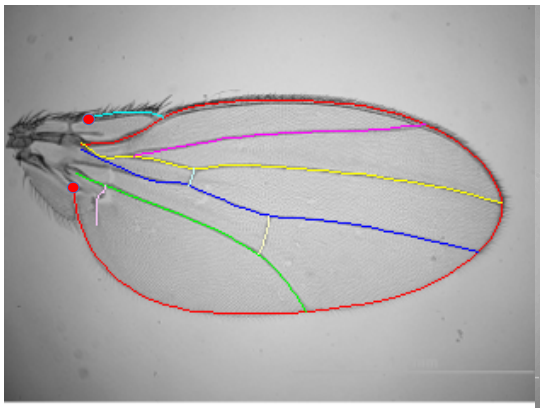
5 cameras x 5 minutes x 30frames/sec – 9000 Images

Wing Images

Dr. Ian Dworkin, Formerly MSU



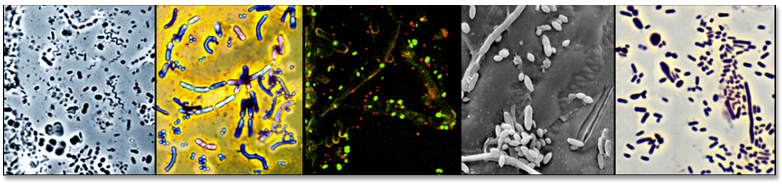
Wingmachine



Wingmachine software developed by the Houle lab, FSU
Houle *et al.* 2003 BMC Evo. Biol. 3:25

CMEIAS – Center for Microbial Ecology Image Analysis System

Dr. Frank Dazzo, MSU



Comprehensive suite of bioimage informatics analysis software applications designed to strengthen quantitative, microscopy-based approaches for understanding microbial ecology, at spatial scales relevant to the individual microbes and their ecological niches.

http://cme.msu.edu/cmeias/

ImageJ
Image Processing and Analysis in Java

- Free, open source
- Many thousand users
- More than 500 plugins listed on their website
- Many, many features

- Python and MATLAB are good at images too…

Existing automated approaches

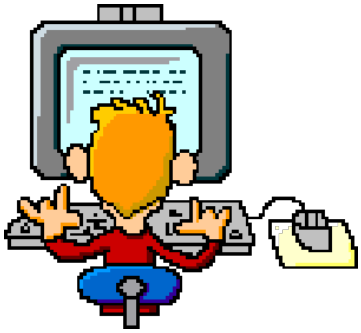
Task Specific

- Program solves a specific problem for a specific type of input
- Domain specific assumptions make it easy to automate image analysis
- Examples:
  CMEIAS, wingmachine

General

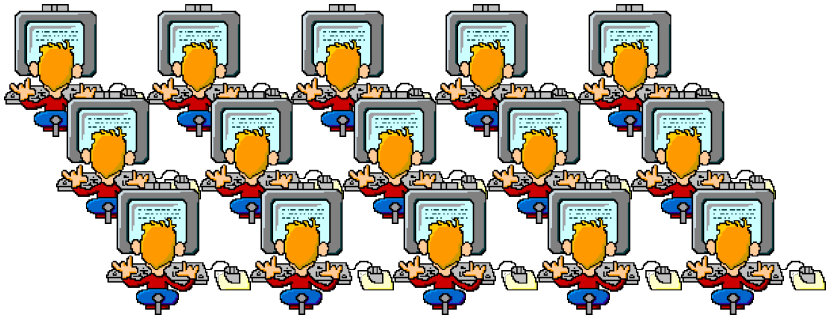
- Tools make it easy to do global manipulation of images
- Difficult to do anything specific to a problem
- Examples:
  Photoshop, ImageJ, Python, MATLAB

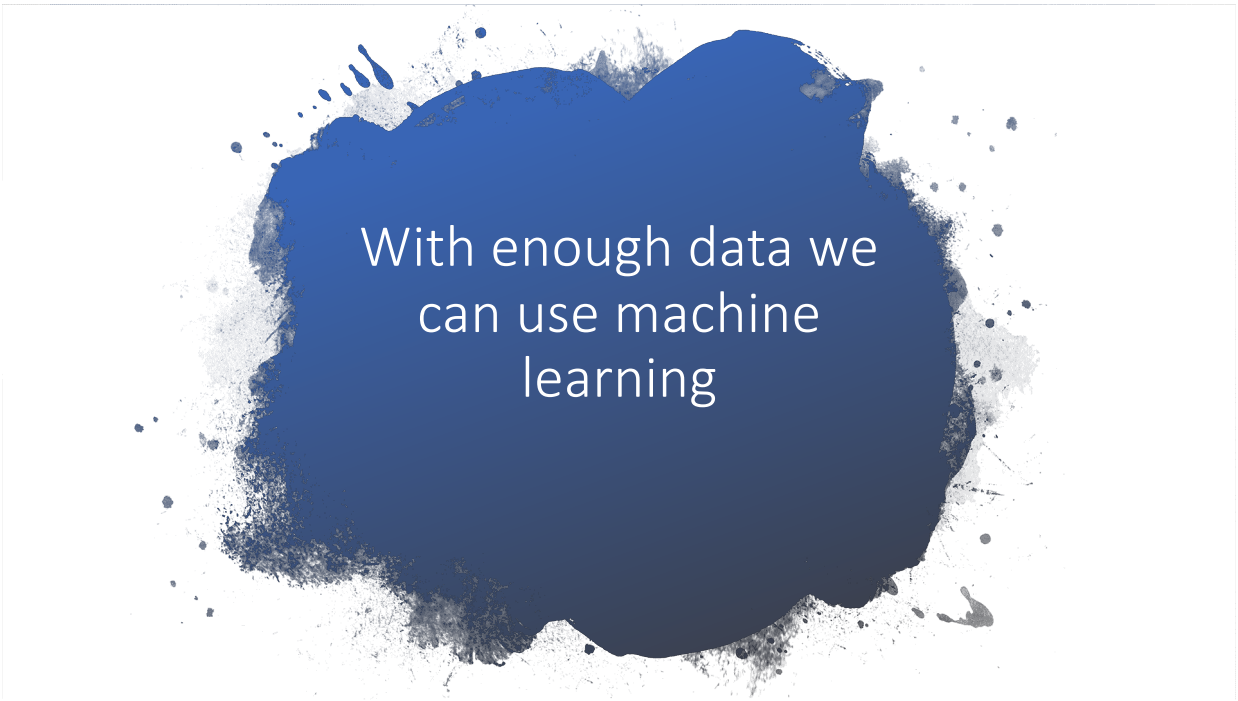
How are digital images analyzed?



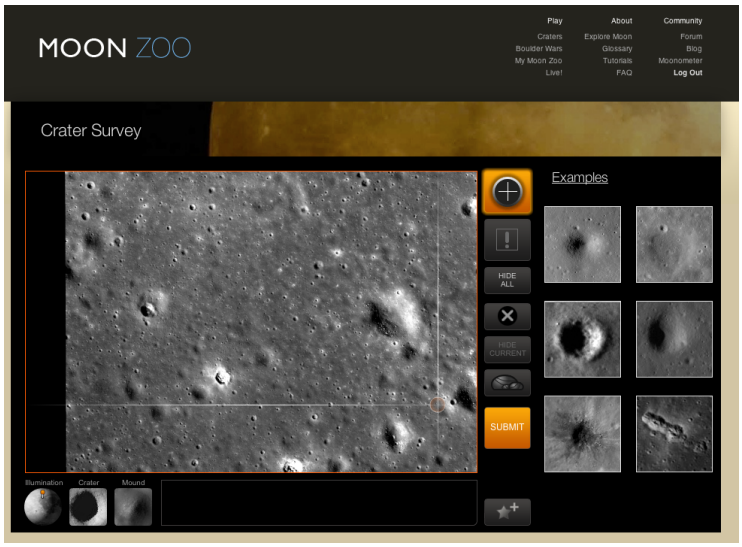
Graduate students are cheap…
Undergraduates are even cheaper!

Also, easy to run in parallel



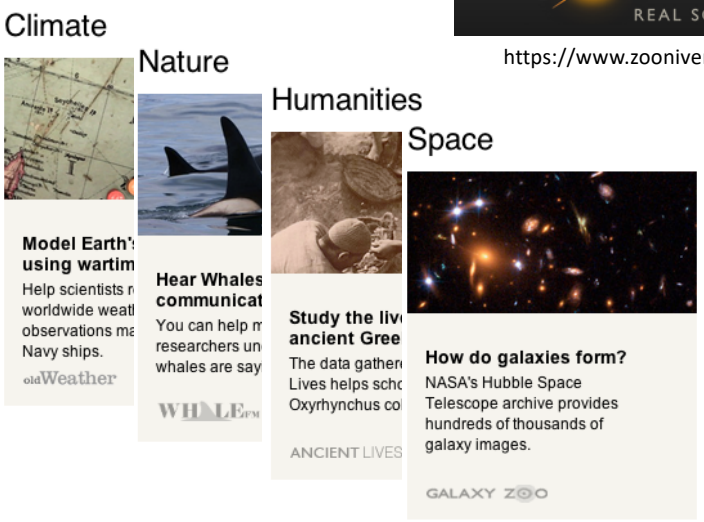


Leveraging “Citizen Scientists”



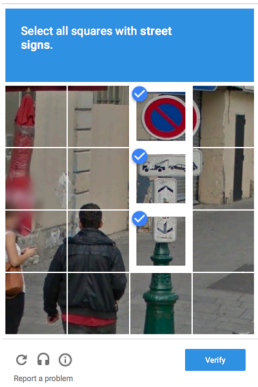
Massively Parallel Image Analysis

https://www.zooniverse.org/

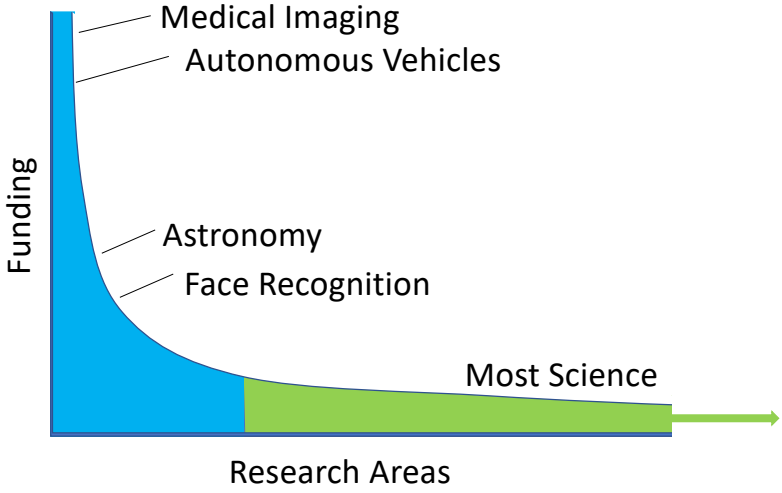


reCAPTCHA™ Digitizing books one word at a time

- Idea invented by Luis von Ahn, Ph.D. Carnegie Mellon
- Crowdsourcing – Distributing tasks to large numbers of people
- Winner of the MacArthur Award
- Security pass-phrase



Serving the Long Tail of Science Imaging



Why is the long tail of science hard?

- In exploratory Science, features change with every problem
- Projects can't afford an engineer for every new idea
- Not everyone can be an expert in image analysis, so training every scientist doesn't always work
- Annotating images is time consuming
- By the time you are done annotating a training set you may be done with the research!

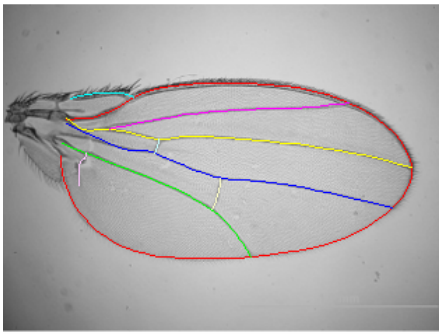
Our Goal

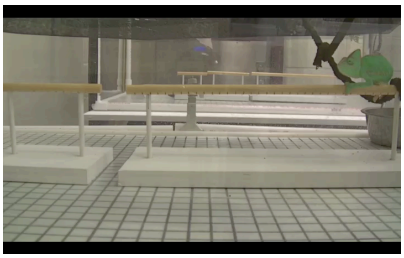
Focus on the workflow of making tools, not on the individual tools themselves

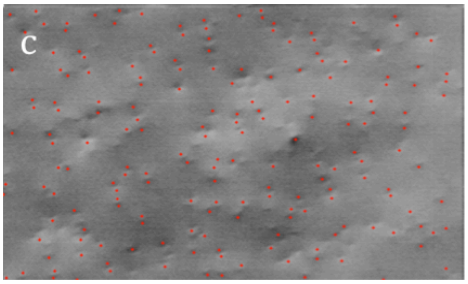
Make it easer to make new tools

How?

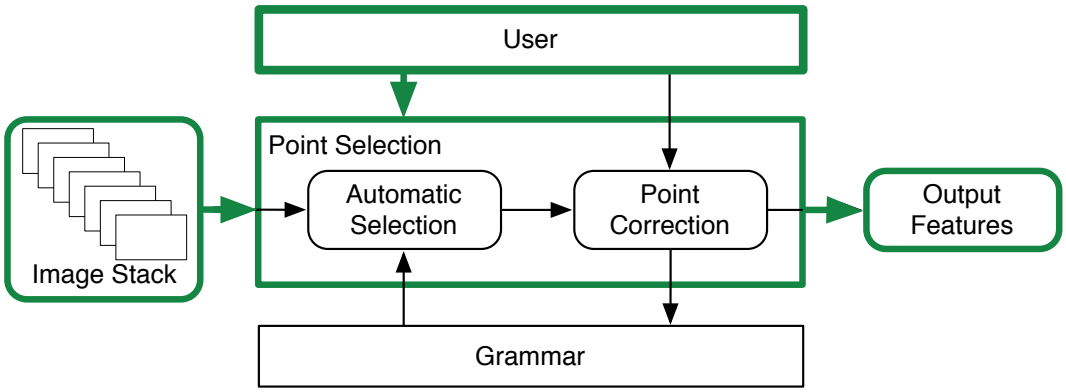
Example #1 - Point Selection Workflow





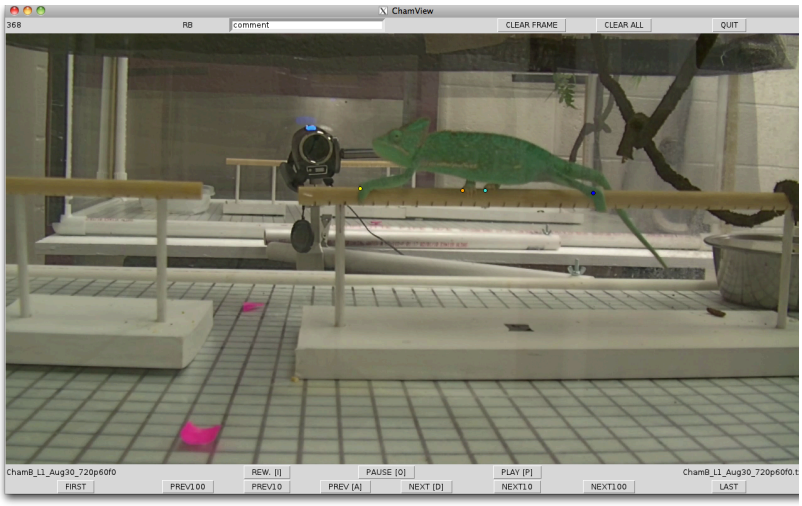


Keeping the researcher in the loop

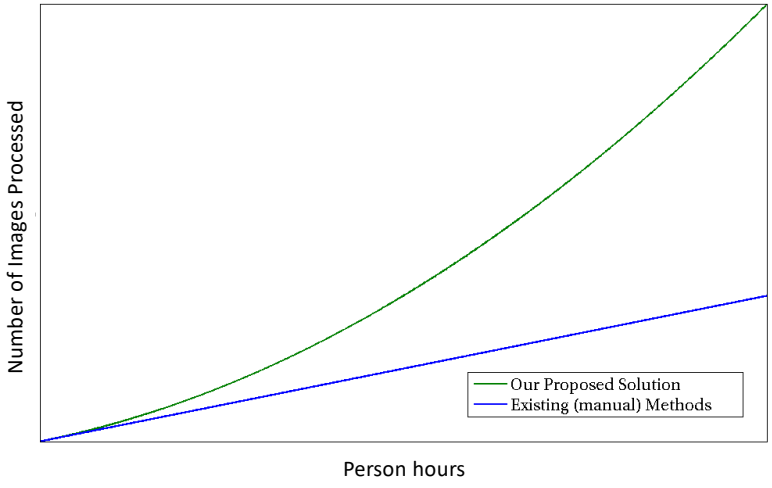


Terms: Incremental vs Batch Learning

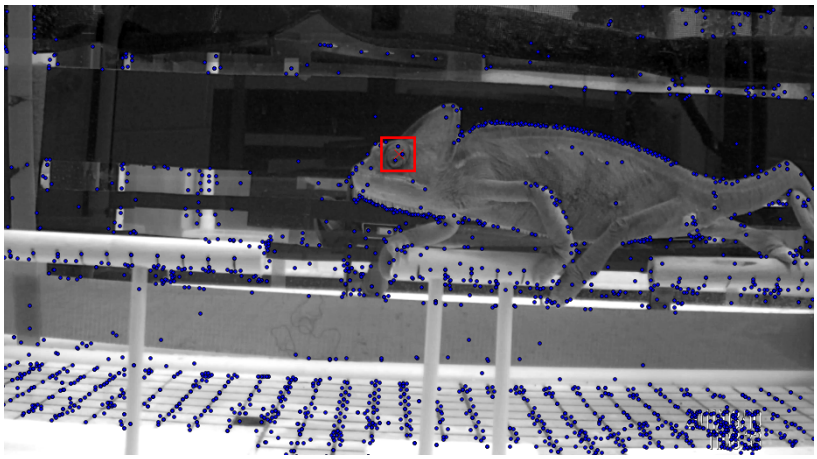
ChamView



How do we measure success?



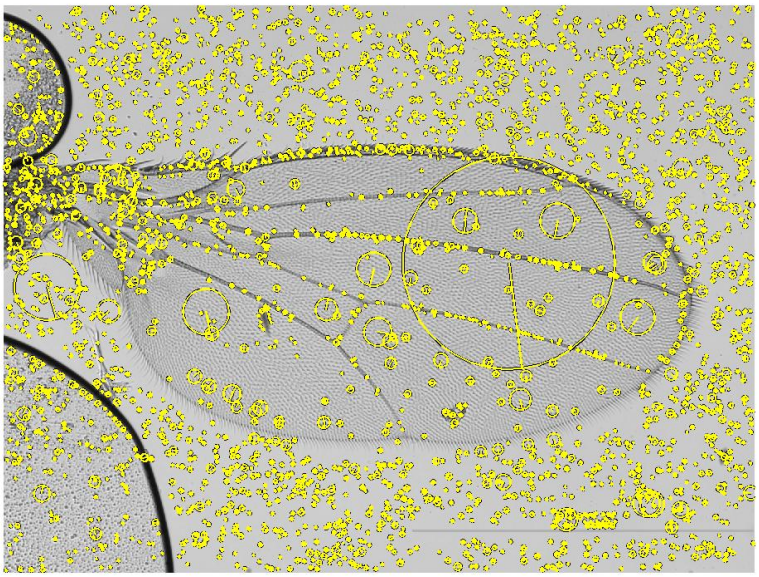
Point tracking on Chameleons



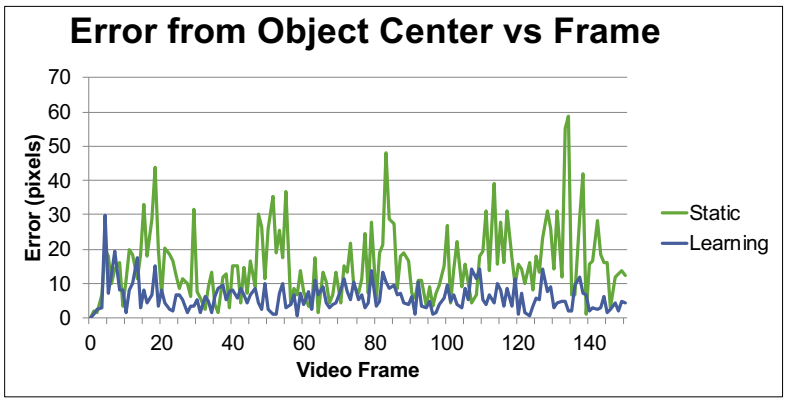
Patrick Korth, CSE University of Michigan

Fly Wings

Ian Dworkin

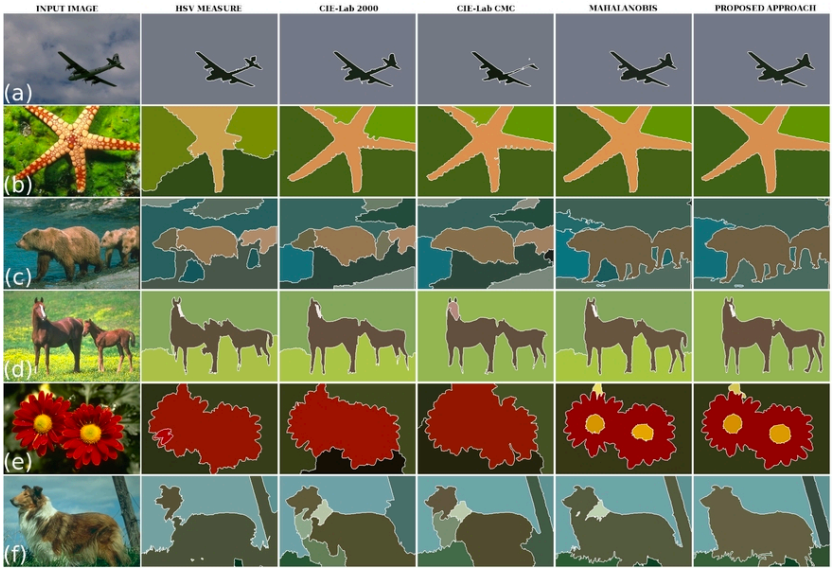


Preliminary SIFT Results



Patrick Korth and Dirk Colbry,

Example #2: Image Segmentation



Python Skimage.Segmentaiton Library

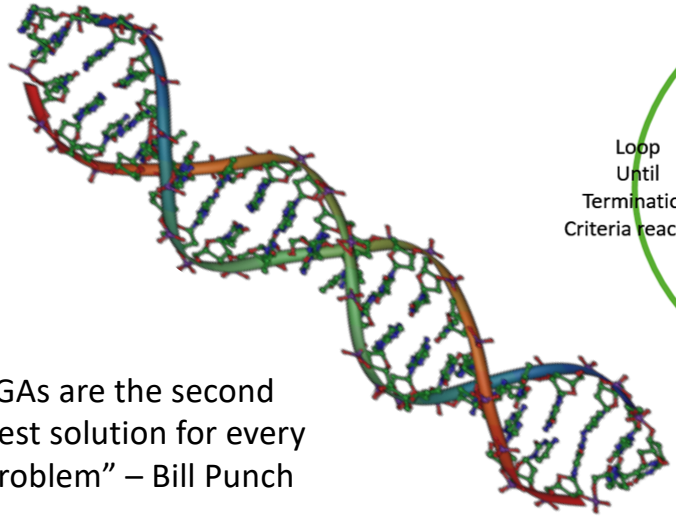
1. thresholding(image[, …]) - Basic image thresholding (not part of skimage)
2. random_walker(data, labels) - Random walker algorithm for segmentation from markers.
3. active_contour(image, snake)- Active contour model.
4. felzenszwalb(image[, …]) - Computes Felsenszwalb's efficient graph based image segmentation.
5. slic(image[, …]) - Segments image using k-means clustering in Color-(x,y,z) space.
6. quickshift(image[, …]) - Segments image using quickshift clustering in Color-(x,y) space.
7. watershed(image[, …]) - Find watershed basins in image flooded from given markers.
8. chan_vese(image[, mu, …]) - Chan-Vese segmentation algorithm.
9. morphological_geodesic_active_contour(…) - Morphological Geodesic Active Contours (MorphGAC).
10. morphological_chan_vese(…) - Morphological Active Contours without Edges (MorphACWE)
11. inverse_gaussian_gradient(image) - Inverse of gradient magnitude.
12. circle_level_set(…[, …]) - Create a circle level set with binary values.
13. checkerboard_level_set(…) - Create a checkerboard level set with binary values.

Which one is Best?

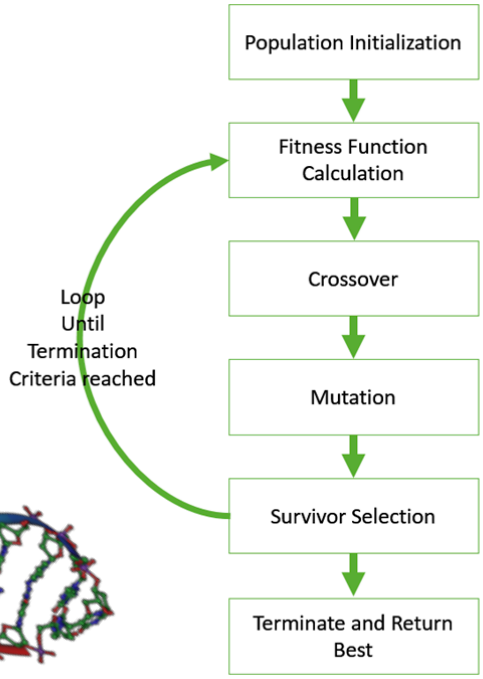
Machine Learning

- How do we search the Algorithm/Parameter space to find the best algorithm for a specific job?
  - Don't always have a lot of training data
  - Every problem is different
  - There is no best algorithm

Basic Genetic Algorithms



"GAs are the second best solution for every problem" – Bill Punch

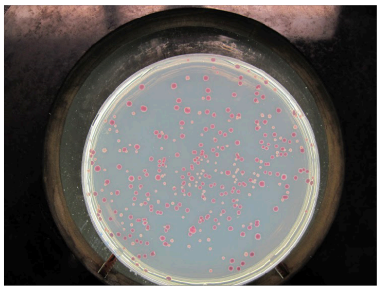


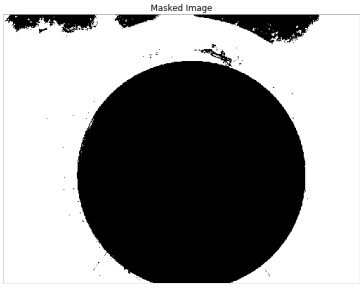
Part 1: Define Your Population Space

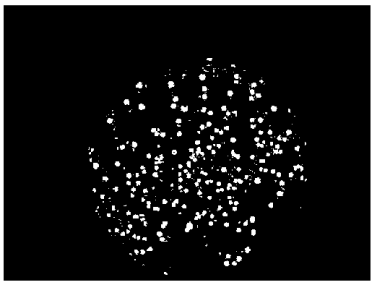
[Algorithm (1-13), option 1, option2, option3, …, optionN]

[5, 244, 0.44, 72, …, -28] … [2, 10, 0.1, 1, …, 1035]

Part 2: Fitness Function







47% Correct

62% Correct

Why Use GAs?

- Highly Heterogenous search space
- Easy to seed search space with known engineered solutions
- Can scale easily (task scaling)
- Output is human readable

Questions