Indexing Uncertainties in Data

Justin Meyer
Mentor: Nick Larusso
Faculty Advisor: Dr. Ambuj Singh
Santa Barbara City College
Major: Electrical Engineering
Funding: N.S.F.

http://www.crustal.ucsb.edu/images/banner-logo-ucsb.png

http://www.ccmr.cornell.edu/images/logos/logo-NSF-CMYK.GIF

http://www.sbcc.edu/marketing/ind
p?sec=1331

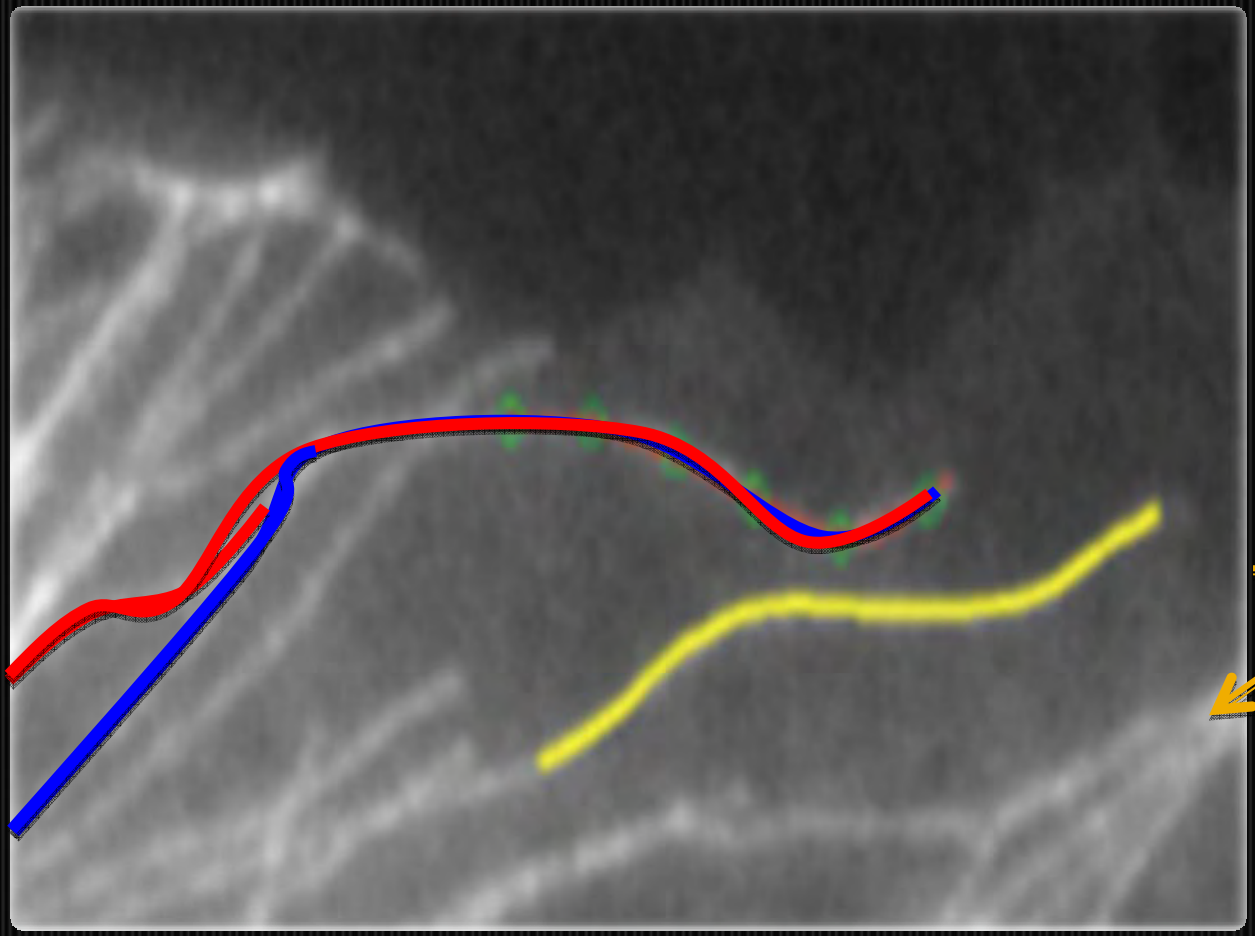
Motivation

- Applications of querying uncertain data
  - Alzheimer's microtubule length measurements
  - Bio-imaging
- Faster access to data
  - Larger dataset better representation of reality
  - Selection and range queries

Where Uncertainty Comes From

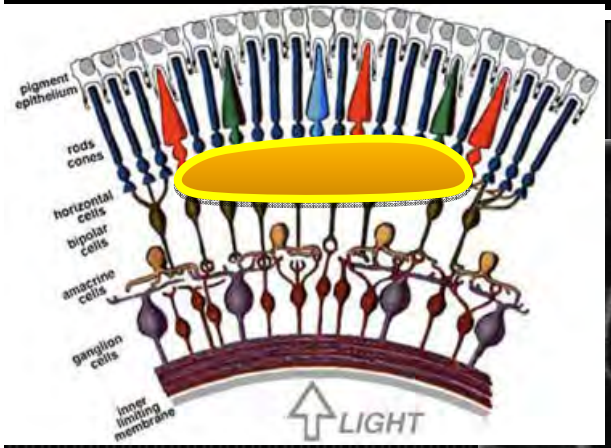
- Bio-imaging
  - Dyeing Techniques are imperfect
    - Dyeing multiple images
    - Hand Dyeing is intensive
  - 3D compression into a 2D image
  - Confocal Microscope
- Sensory data (tracking systems)
  - Update times

Microtubule Images



http://www.bioimage.ucsb.edu/component/content/article/53-frontpage-highlights/116-bisquewebapps

Retinal Images



Horizontal Cells

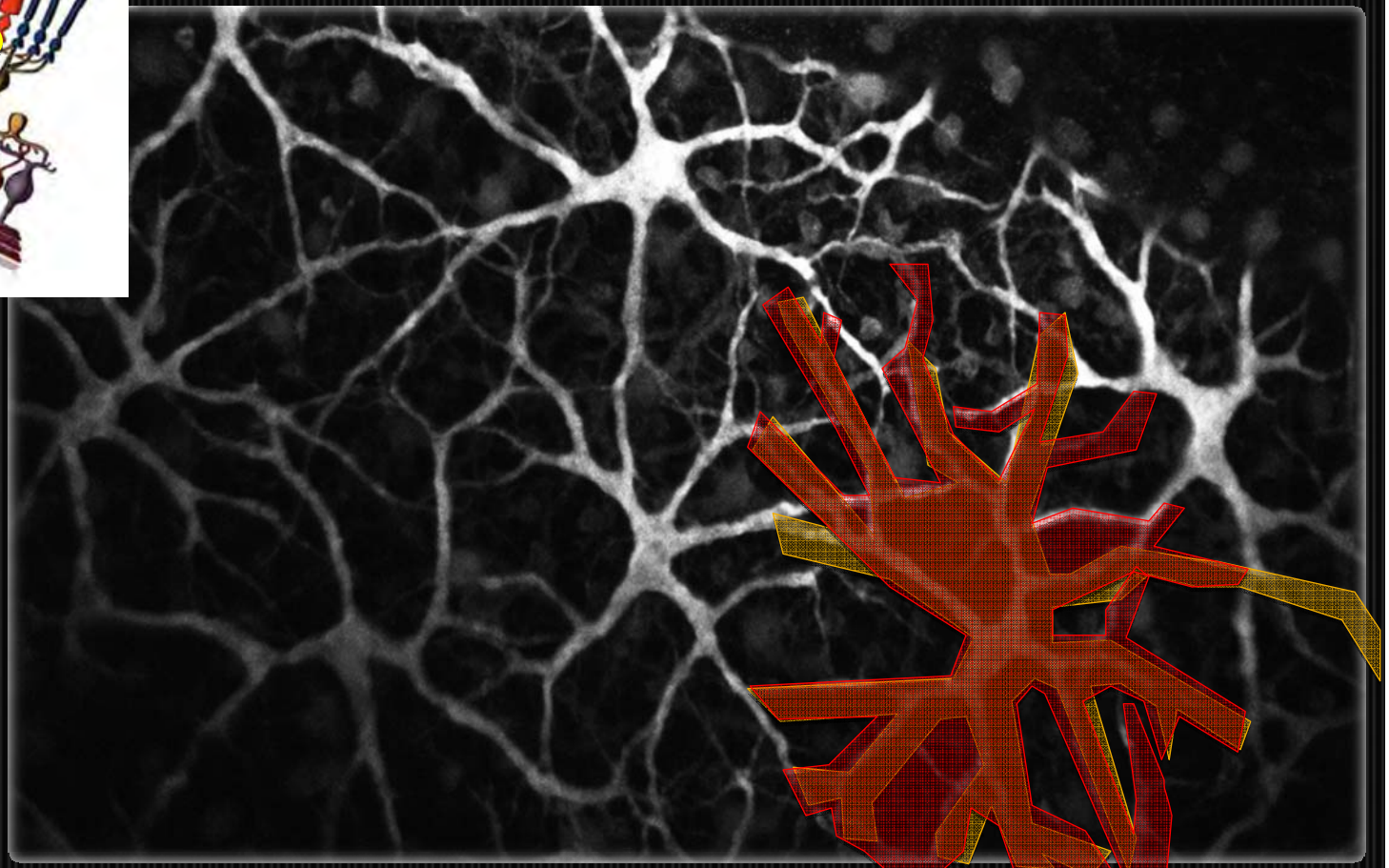
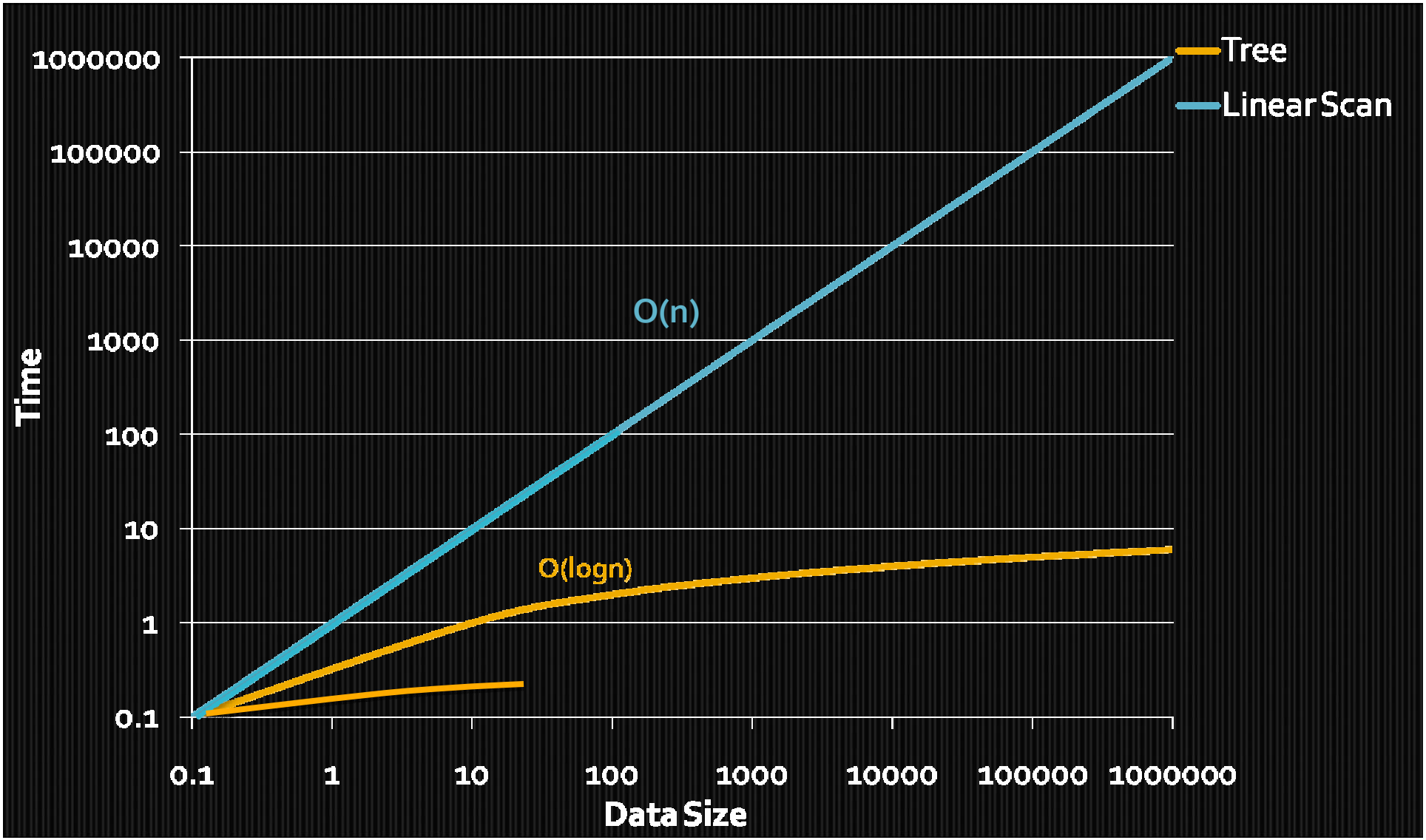


Photo Courtesy Dr. Ambuj Singh

Project Goals

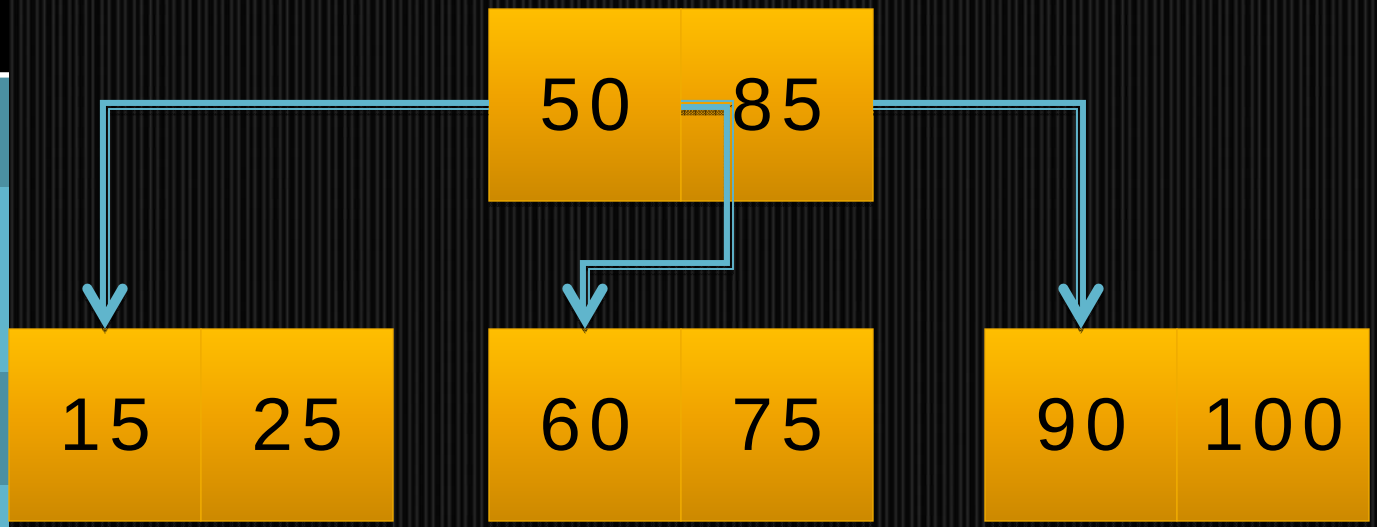
- Efficiently query with use of an index structure
  - Without indexing, linear scan required
  - Indexing is more scalable as datasets grow
- Problem with indexing due to uncertainties within data
  - Produce results for range and selection queries faster

Linear Scan vs Indexing Structure



Indexing Certain Data

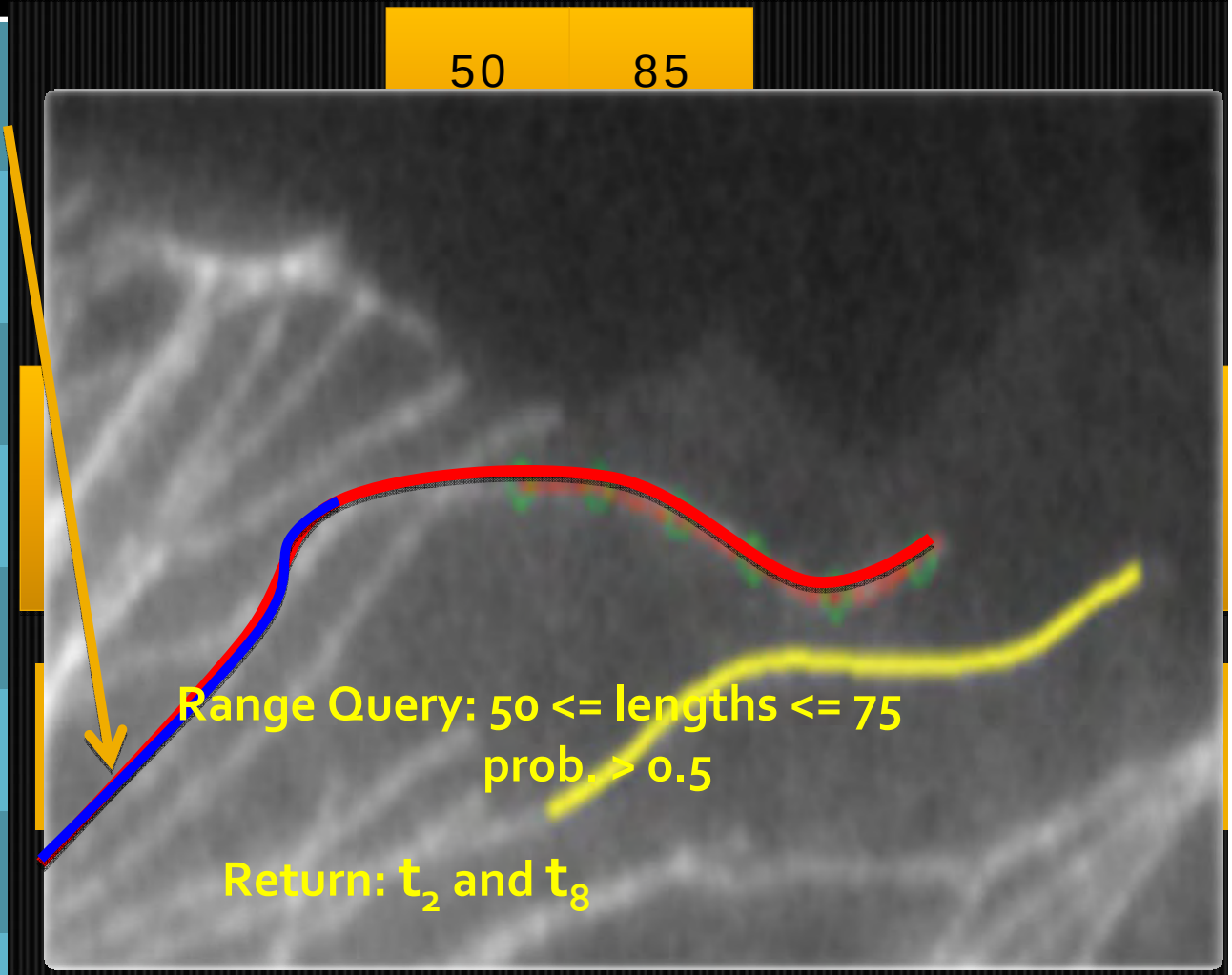
Microtubule	Length
t_1	25
t_2	50
t_3	75
t_4	100
t_5	15
t_6	85
t_7	90
t_8	60



- Eliminate on the order of half the possibilities with each decision

Indexing Uncertain Data

Microtubule	Length
t_1	(25 , 0.8) (50, 0.2)
t_2	(50, 0.6) (60, 0.4)
t_3	(75, 0.5) (90, 0.5)
t_4	(100, 0.7) (85, 0.3)
t_5	(15, 0.6) (25, 0.4)
t_6	(85, 0.7) (100, 0.3)
t_7	(90, 0.9) (75, 0.1)
t_8	(60, 0.7) (15, 0.3)



Results

- The actual results confirming the hypothesized results
  - The structures cost is better than the linear scan
    - Especially for large datasets
- Future work
  - Appling to all areas of uncertain data
    - Sensory data
  - Plan to compare with other uncertain indexing techniques

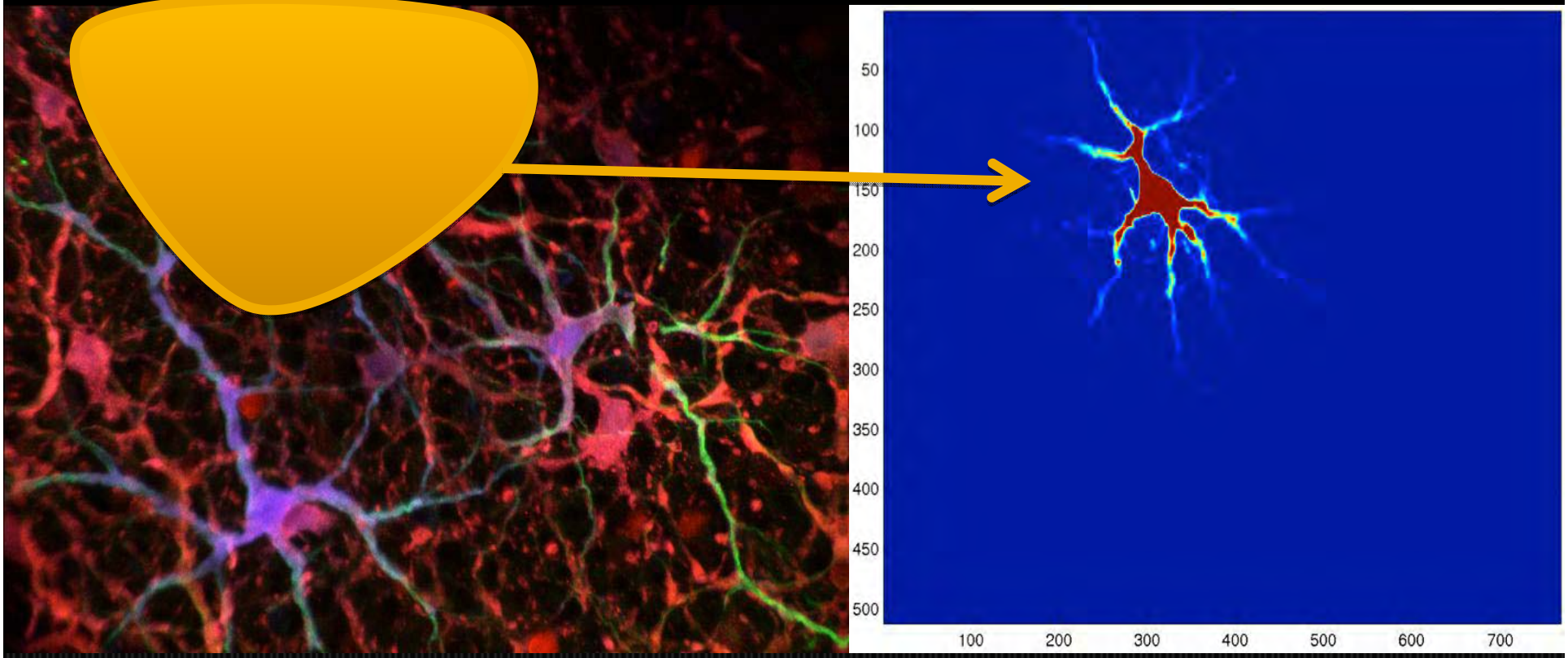
Questions ?

?

Acknowledgements:
INSET
CNSI
Dr. Ambuj Singh
Nick Larusso
NSF

Justin Meyer
Email: jmeyeroct22@gmail.com

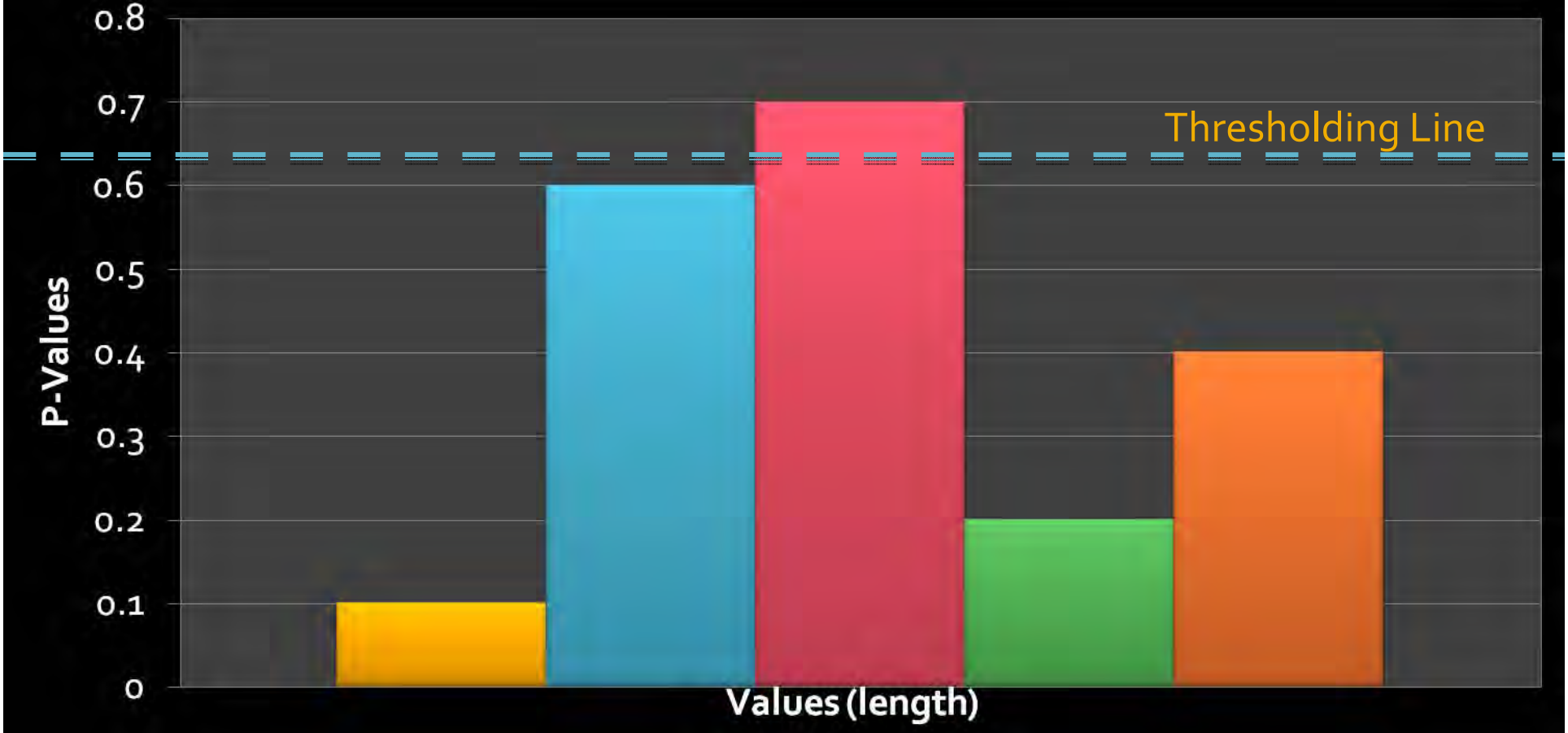
Software Generated Uncertainty



- The probability is represented as the color intensity of the red
  - Red is cell
  - Blue is background
- Darker the color the higher the probability

http://www.ljosa.com/~ljosa/publications/ljosa_icdm_2006.pdf

Thresholding



- Only the values above the line a represented as a "average" value
- This is how most databases handle uncertainty