# Indexing Uncertainties in Data

Justin Meyer
Santa Barbara City College
Electrical Engineering

Mentor: Nick Larusso
Advisor: Dr. Ambuj Singh

## Motivation

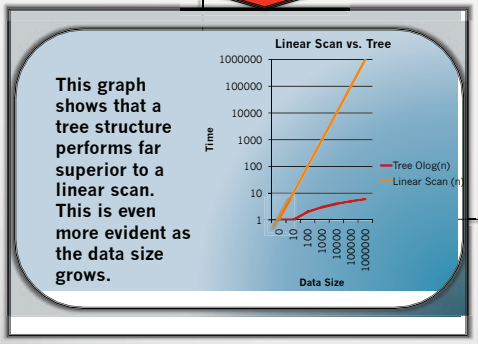
- ❑Faster Access to data
  - ❑Faster query results
  - ❑Better representation of reality
- ❑Applications of querying uncertain data
  - ❑Bio-image analysis
    - ❑Alzheimer's microtubule analysis
  - ❑Tracking systems

## Goals

- ❑Faster query results
  - ❑Faster results from range and selection queries
- ❑Dynamic structure
  - ❑Allowing for larger dataset
  - ❑Can be applied to all areas of uncertainty
- ❑Efficiently handle large datasets

## Results

- ❑Confirm with hypothesized results
- ❑Compare to other existing indexing models




This graph shows that a tree structure performs far superior to a linear scan. This is even more evident as the data size grows.

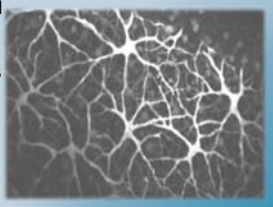
## Studying Bio-images

Bio-images are used to study biology on a cellular level. Cellular interactions are the foundations of all life. Through the techniques used to generate these images uncertainty is created.

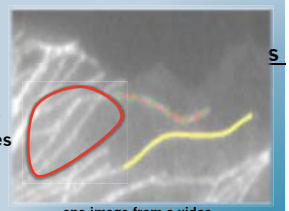
## Example Applications

### Retinal Images

This image is a horizontal cell which is a 3D cell that is then compressed into a 2D image. This is done as a simplification measure. These images are used to study the morphology of the cell.

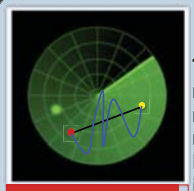


This image is used to measure the lengths of the microtubules. Microtubules are used to transport proteins in the brain. Uncertainty in this image comes from the clarity of the image and the microtubules intersecting each other.



one image from a video

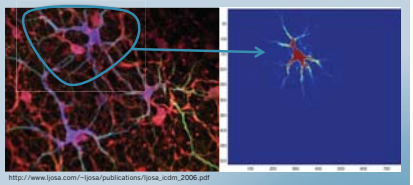
## Tracking System Uncertainty



The blue path would cause great uncertainty if you assume a straight line.

The yellow dot represents the first time reading and the red represents the second. If this is tracking a ship it might be safe to assume that the "dot" traveled in a straight line. If you are tracking a particle is this still a safe assumption?

## Data Generation



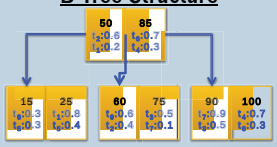
http://www.ljosa.com/~ljosa/publications/ljosa_icdm_2006.pdf

Data from bio-images is most often generated by using a computer program to attempt to study the morphology of the cell. This program indexes the probability of the cell through color. The darker the red the higher the probability.

## Methods: B Tree

- ❑Used B Tree structure
- ❑Can quickly return results from range and selection queries
- ❑Tuples are indexed in an inverted fashion

### B Tree Structure




| 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) |

## Methods: Inverted Index

- ❑Tuples are split up according to values
- ❑Indexed in order with largest probability first
- ❑This allows for fast results for queries
  - ❑Queries don't look through every value

## Where Uncertain Data Comes From

- ❑Bio-Images
  - ❑Dyeing techniques
    - ❑Multiple cells dyed in the same image
    - ❑Dyeing a single cell is very intensive
  - ❑Compression of a 3D image into a 2D
- ❑Update times in sensory networks
  - ❑Tracking systems
  - ❑Temperature systems

## Acknowledgements

Dr. Ambuj Singh
Nick Larusso
INSET, UCSB, CNSI

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