Data Mining

Programs that Generate Programs

and

Parsing GPS Data and Geographic Information Systems (GIS)

Dr. Thomas B. Kinsman, Ph.D.

Agenda

Administration:

- Future HW is to write your own decision tree.
- Or, you need to generate a program for data visualization in Google Earth.
- In order to do that you will need to know how to write a program which writes a program.
- This is done all of the time.

Engineering Story

Tom Petty (and the Heartbreakers)

- Doing the same thing,
  the same way,
  every time.
- Coffee was always the same
- Brunn coffee makers − always were the same
- Scoops of coffee were measured off exactly.

 3

Programs which Generate Programs:

Two major benefits:

1. You debug the code only once.
   It is more difficult to debug,
   but you only need to do it once.

2. The resulting code might be much more complicated then any code you could write as a human − there are too many choices for you to make, and too many parameters (and thresholds) for you to select.

 4

Two Programming Patterns:

- In the extremes, there are only two programs you can write.
- Only two pure "end member" programs.

Exemplar A:

1. Read in all of the data
2. Process all of the data
3. Write out all of the results

Exemplar B:

1. Read in some of the data
2. Process some of the data
3. Write out some of the results

 5

Factoring Programs

Programs can be factored into 3 parts:

1. The header, or prolog
2. The body
3. The trailer, or epilog

 6

The header might:

- Setup import statements
- Define any global interfaces or variables
- Define any operational parameters

The body might:

- Describe how to handle data as it comes in
- Tell how to visualize any data
- Check for special situations

The trailer might:

- Close off any open definitions
- Emit any special situations found
  - example: stop lights
  - stop signs
  - right turns
  - left turns
- Do the final rendering

 9

PostScript File Example:

- Header sets up and defined the page size.
- Body does all the drawing of vector graphics.
- Trailer says, “showpage()”, which tells the graphics engine to render and print the page and print it.

 10

HTML File Example:

```
<HTML>
<HEAD>
<TITLE>DOCUMENT TITLE</TITLE>
</HEAD>

<BODY>
Body of Document Here…

</BODY>
</HTML>
```

 11

XML File Example:

- Header sets up and defined the protocol.
- Body does all the drawing of vector graphics.
- Trailer closes off the XML file.

 12

kml files

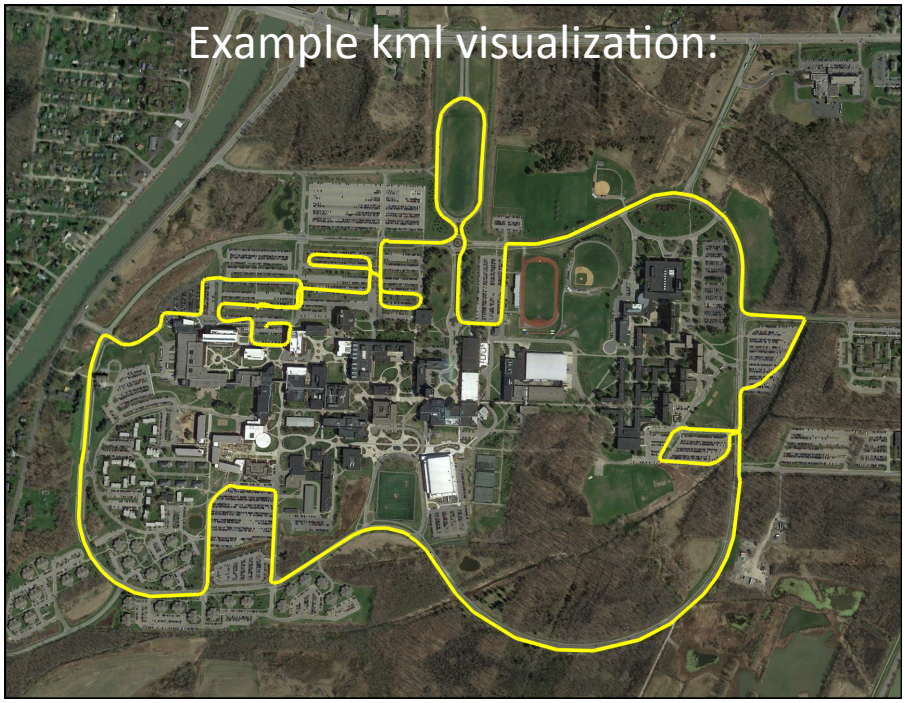
- kml stands for “Keyhole Markup Language”.
- Used by Google Earth to visualize GPS data.

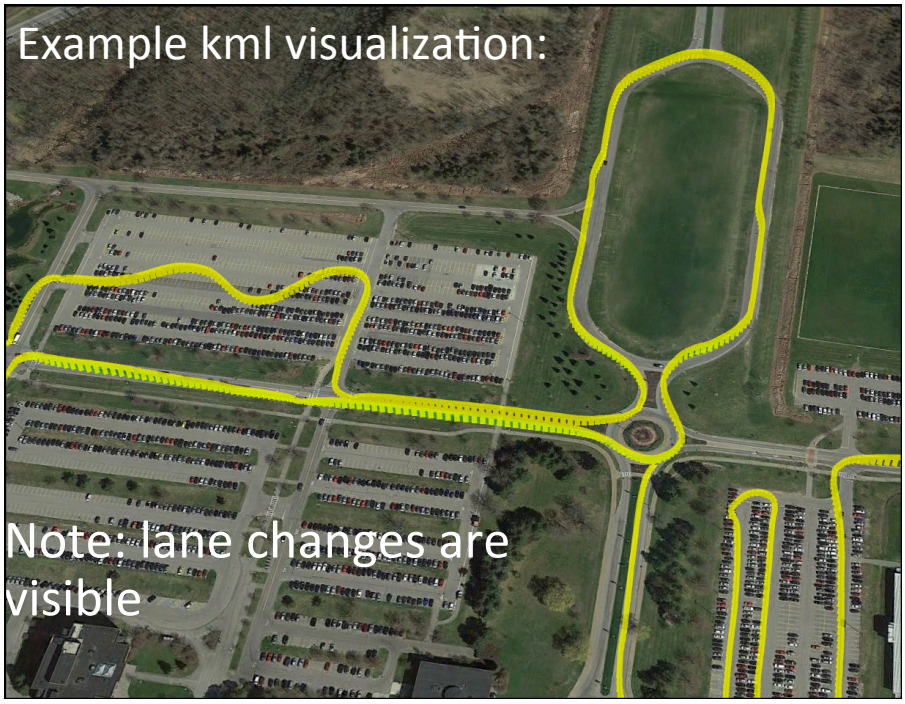
- More information here: https://developers.google.com/kml/documentation/kml_tut

- Or here: https://developers.google.com/kml/documentation

- Or, of course any other source you can find on the internet.

 13





Modified to See Speed Profile:

The car was travelling from left to right.

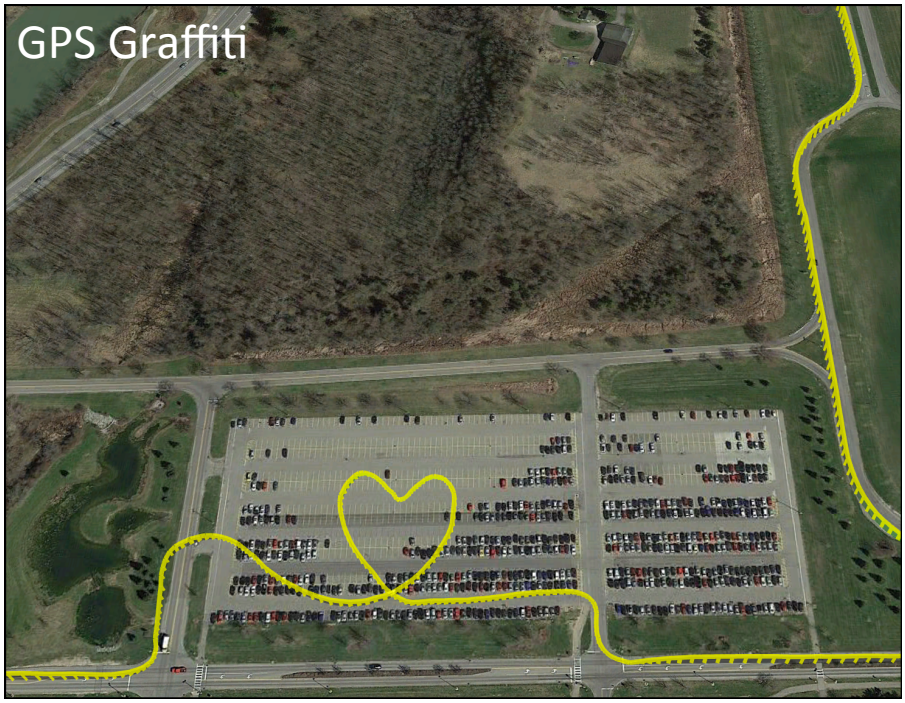
Using temporal sequential processing, we can detect the stops,
ignore stops which are soon followed by other stops,
and then locate the final position of the stop sign.

16

Possible Temporal Sequential Analysis:

1. Normal forward motion.
2. if (stopped for more then $Thresh_1$, and speed below $Thresh_2$)
   A. compare current position to last stop position
   B. If (current position ≤ $Thresh_3$ miles from last) pop last position off list
   C. Add current position to list of stop signs.
3. Go to state 1.

 17



KML File Example Header:

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
<Document>
<Style id="yellowPoly">
  <LineStyle>
    <color>Af00ffff</color>
    <width>6</width>
  </LineStyle>
  <PolyStyle>
    <color>7f00ff00</color>
  </PolyStyle>
</Style>
<Placemark><styleUrl>#yellowPoly</styleUrl>
<LineString>
<Description>Speed in MPH, not altitude.</Description>
  <extrude>1</extrude>
  <tesselate>1</tesselate>
  <altitudeMode>absolute</altitudeMode>
  <coordinates>
```

 19

KML File Example Body:

```
-77.607077,43.152722,1.12
-77.607077,43.152713,1.13
-77.607077,43.152705,1.14
-77.607073,43.152697,1.14
-77.607072,43.152688,1.14
-77.607070,43.152678,1.14
-77.607070,43.152668,1.14
-77.607072,43.152658,1.13
-77.607073,43.152647,1.11
-77.607073,43.152637,1.10
```

 20

KML File Example Trailer:

```
    </coordinates>
   </LineString>
  </Placemark>
 </Document>
</kml>
```

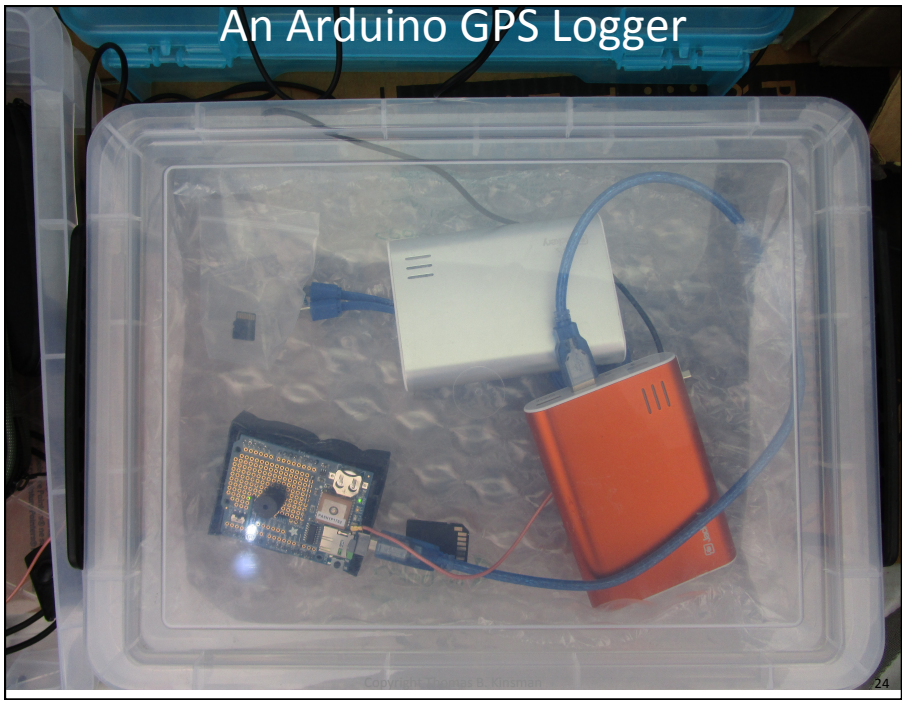

Parsing GIS / GPS Data

Questions regarding GPS Data

- Where Does GPS Data come from?
- Where does time originate?
- Where else have we had this time problem?
- What does the origin of time give up?
- How accurate is GPS data if all bits are exact?

 23

An Arduino GPS Logger



 24

GPS Data: RMC Fields

```
$GPRMC,201430.800,A,4305.1517,N,
07740.8502,W,1.21,61.08,200918,,,A*44
```

1. Field 1 is type of data
2. Time in GMT, or UTC
3. A
4. 4305.1517 → 43 degree, and 5.1517 minutes
5. N is North (+ degrees) or S (- degrees)
6. 07740.8502 →77 degrees and 40.85 minutes
7. E is East (+ degrees) or W is West (- degrees)
8. Speed in knots. In this case 1.21 knots.
9. Tracking angle, 61.08 degrees
10. 200918 → 20th of Sept. 2018.

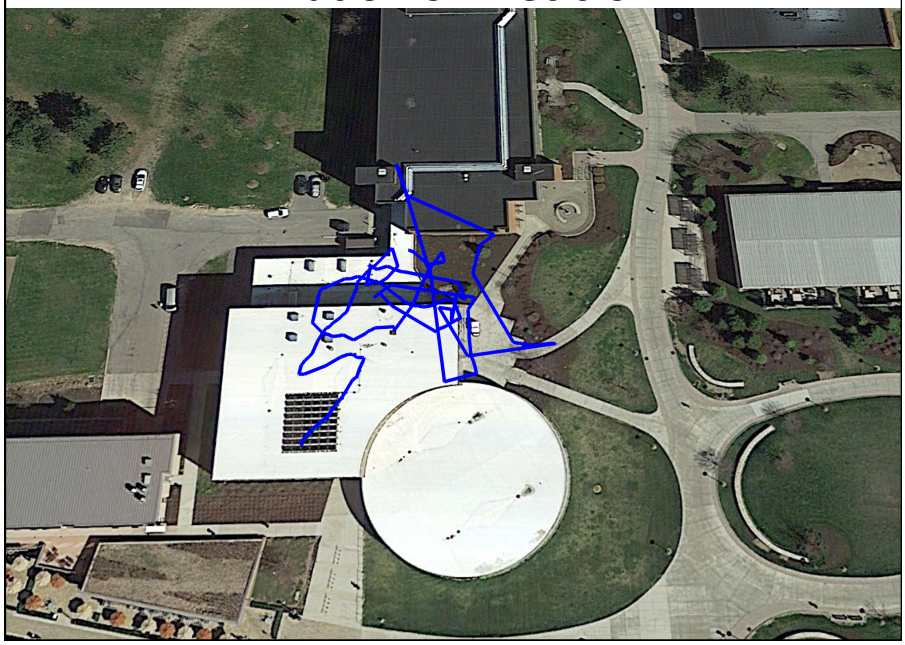
GPS Data: GGA Fields

```
$GPGGA,201431.400,4305.1514,N,07740.8503,W,
1,06,1.18,145.9,M,-34.4,M,,*5F
```

1. Field 1 is type of data
2. Time in GMT, or UTC
3. 4305.1514 → 43 degree, and 5.1514 minutes
4. N is North
5. 07740.8503 →77 degrees and 40.8503 minutes
6. W is West
7. 1 is if you have a fix or not
8. Number of satellites you are tracking
9. Dilution of precision
10. Altitude

 27

Dilution of Precision



Just a bit more on KML Files:

More on KML Files:

Order swap:

- GPS Location comes in as:
  (lat, N/S, long, E/W).
- KML wants to see
  (+/-long, +/-lat)

Format Change:

- Comes in as:
  Degrees*100 + Minutes. (N,S,W,E)
- Need to change to:
  +/- Degrees.fractional_degrees

Creating Decision Trees

Header for a Decision Tree:

```
define the input file
open the input file

example: open up the CSV file
```

Body for a Decision Tree:

```
read in one record
set default class
class_to_print = default;
if ( attribA >= thresholdA )
  if ( attribB >= thresholdB )
    class_to_print = …;
  else
    class_to_print = …;
else
etc…
print out the class_to_print
```

 33

Trailer for a Decision Tree:

```
Close the program, if needed
```

 34

Variation - 1

The "emit header" writes out:

1. Header
   Open the input file…
   Main program …
   While there is more data to read in:
   A. Read in a record

the "emit body" writes out:

   B. The classification code for this record.

the "emit trailer" writes out:

   C. Write out the result of the classification
   Close the function…

 35

Variation - 2

1. Header
2. Forward declaration for a classification function. (But do not define the function yet.)
3. Main program
   While there is more data to read in:
   A. Read in a record
   B. Call the classification function
   C. Write out the results

4. Now actually define the classification function

 36

Variation - 3

The "emit header" writes out:

1. Header.

The "emit body" writes out:

2. Actually define the classification function.

Then the "emit trailer" writes out:

3. Main program:
   While there is more data to read in:
   A. Read in a record
   B. Call the classification function
   C. Write out the results

 37

Summary

- It is a good engineering practice to do the same thing, the same way, every time.
- The goal of the decision tree homework will be to write a program, which writes a program.
- The goal of the GPS visualization task is to write a program, which outputs a KML file for visualization using Google Earth.
- Both of these tasks involve processing data, and making decisions.
- In both cases, your program writes a program which writes out: a header, a body, and a trailer.

 38

END

 39

Questions From Lecture:

- Given data containing a mixture of classes, what is an end member?
- What does the story of Tom Petty making coffee have to do with any of this?
- What are the two general programming patterns?
- What are the three main parts of a program?
- Give two examples where programs use these patterns?
- Can you give an example end member of shoppers?
- What are two "end member" patterns for general programming?
- What three attributes of GPS does kml usually visualize?
- How did Dr. K change his kml files?
- If GPS data is accurate to a millionth of a degree, what is the approximate accuracy, in inches? (Circumference of Earth is 25,000 miles. 5,280 ft per mile. 12 inches per foot.)
- How can GPS data be cleaned?
  What could possibly go wrong?
- How can GPS data be reduced?
- What does GPS stand for?
- What does kml stand for?
- What does RMC stand for?
- Why can't you simply find a stop sign by one stopping point?
- What is GPS graffiti?
- What file format is kml a subset of?
- What are three issues when converting GPS sentences into kml?
- In which cases have we seem UTC (GMT) time?
- What problem does UTC time solve?
- What is the dilution of precision?
- When is the dilution of precision an issue?
- What is your longitude and latitude at RIT? (-77.6, 43.1)

 40

Other Things Mentioned

- 0° longitude is w.r.t. Greenwhich
  - West is negative
  - East is positive
- 0° latitude is w.r.t. equator
  - North is positive
  - South is negative
- Dilution of Precision
  Can you explain…

41