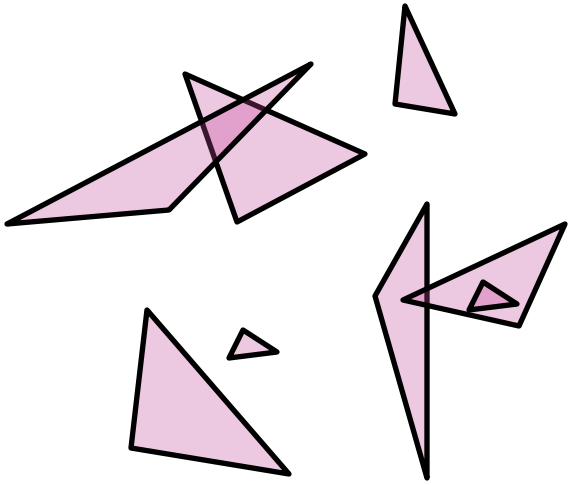
2IL76
Algorithms for Geographic Data

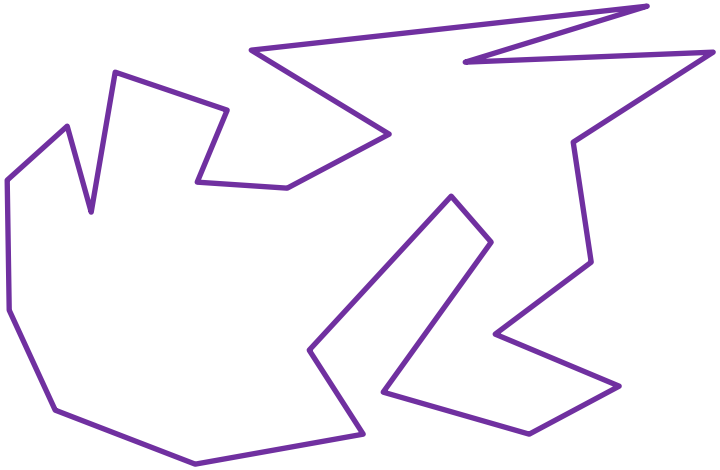
Spring 2015

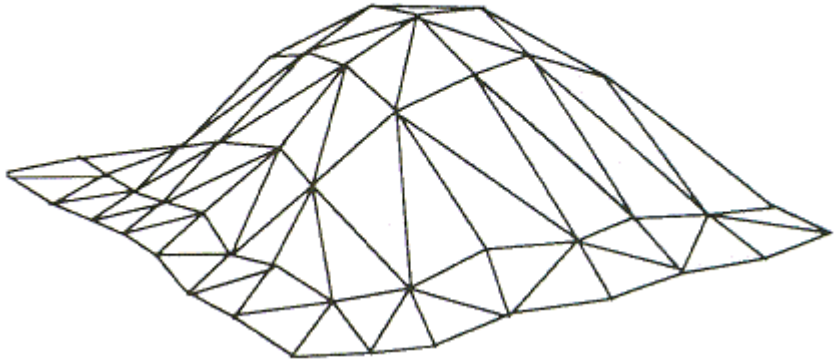
Lecture 1: Introduction

Geometric data

- Sets of points
- Sets of line segments, circles or triangles
- Simple polygons
- Planar subdivisions
- Polyhedral terrains
- Higher-dimensional data



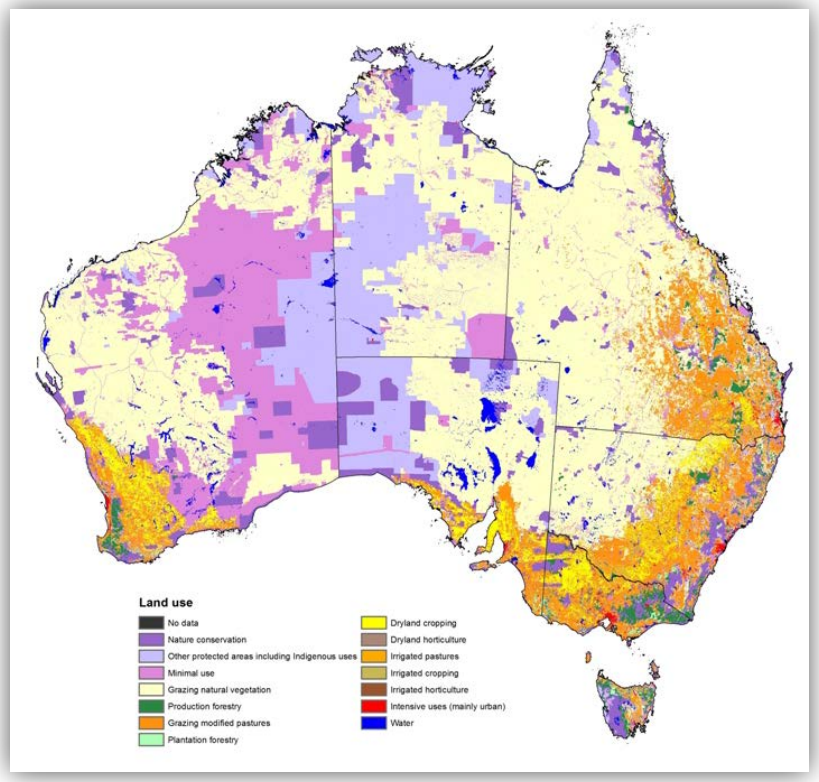




The geographic application

- Geographic Information Science, GIScience
- Data has a concrete meaning

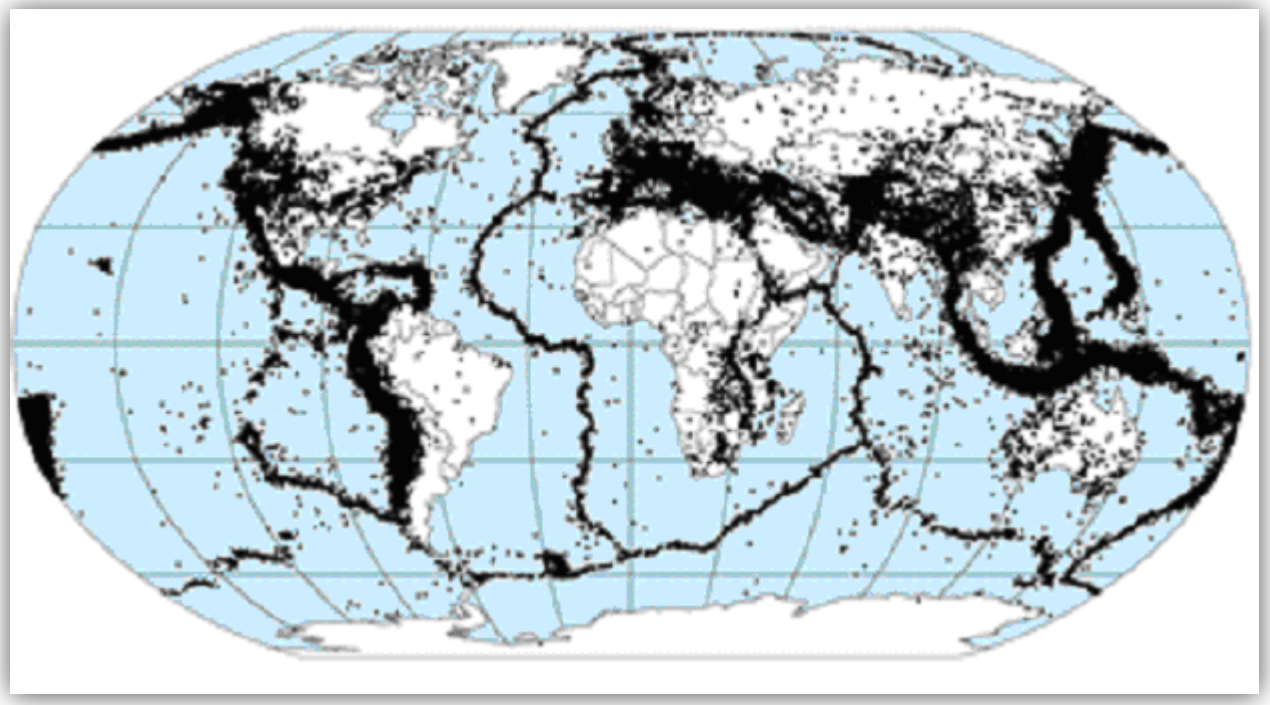
a land-use data set can be modeled by a planar subdivision where each face has an attribute



The geographic application

- Geographic Information Science, GIScience
- Data has a concrete meaning

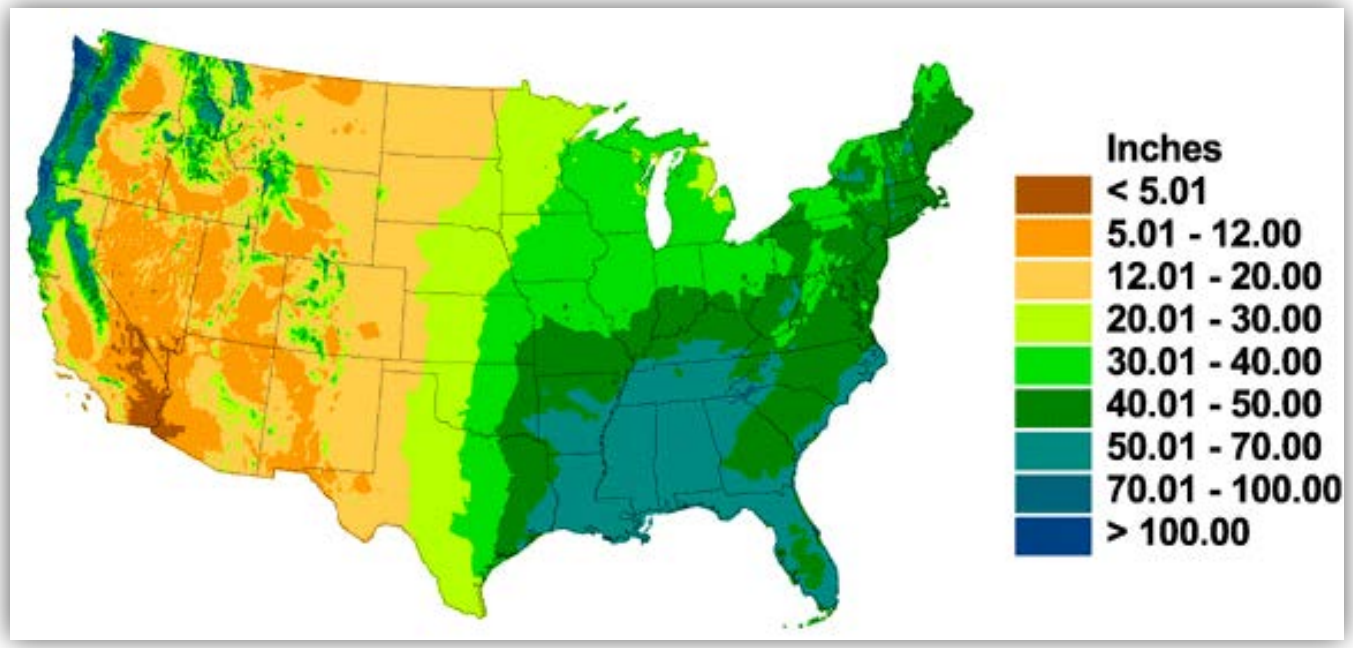
a data set with epicenters of earthquakes can be modeled by a set of points in the plane



The geographic application

- Geographic Information Science, GIScience
- Data has a concrete meaning

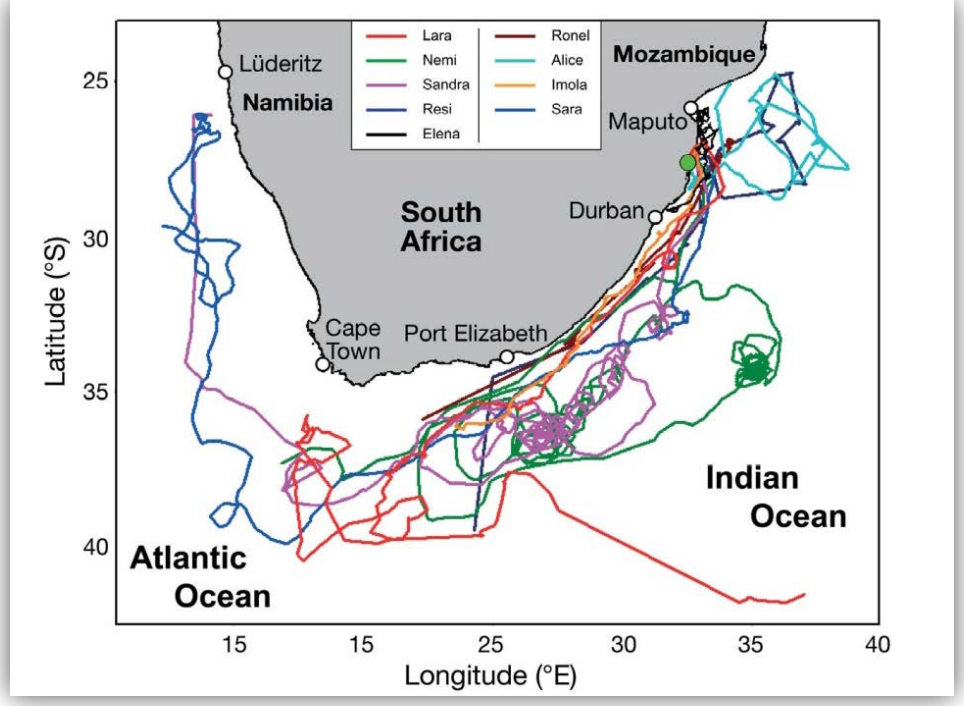
a data set for elevation above sea level, or annual precipitation, can be modeled by a polyhedral terrain



The geographic application

- Geographic Information Science, GIScience
- Data has a concrete meaning

a data set of sea turtle tracks can be modeled by polygonal lines



The course

- Advanced algorithmic tools for geographic data
- Emphasis on algorithms for
  - movement data analysis
  - automated cartography
- But also
  - data and representation
  - modelling

Some administration first

before we really get started ...

Organization

Lecturers: Kevin Buchin, MF 4.101, k.a.buchin@tue.nl
Bettina Speckmann, MF 4.105, b.speckmann@tue.nl

we both travel lots, always email if you want to find us…

Web page:

http://www.win.tue.nl/~kbuchin/teaching/2IL76/index.html

Prerequisites

Basic knowledge of algorithms and mathematics

- O-notation, Ω-notation, Θ-notation; how to analyze algorithms
- Basic calculus
  manipulating summations, solving recurrences,
  working with logarithms, …
- Basic probability theory
  events, probability distributions, random variables, expected values, …
- Basic data structures
  linked lists, stacks, queues
- (Balanced) binary search trees
- Basic sorting algorithms
  MergeSort, InsertionSort, QuickSort, …
- Graph terminology, representations of graphs (adjacency lists and adjacency matrix), basic graph algorithms (BFS, DFS, topological sort, shortest paths)
- Dynamic programming, greedy algorithms

Grading scheme 2IL76

1. 5 homework assignments, the best 4 of which count towards 50% of the final grade.
2. A course project, for which you can get a maximum of 10 points and which counts towards 50% of the final grade. The project is done in groups of three students.

- If you reach less than 50% of the possible points on the best four homework assignments or you score less than 5 on the project, then your final grade is FAIL. If you fail the course, you can redo the project and/or one or more sets of homework exercises to replace the set(s) with the lowest score(s).

Academic Dishonesty

Academic Dishonesty

All class work has to be done independently if not stated otherwise. You are of course allowed to discuss the material presented in class, homework assignments, or general solution strategies with me or your classmates, but you have to formulate and write up your solutions by yourself. You must not copy from the internet, your friends, or other textbooks. Problem solving is an important component of this course so it is really in your best interest to try and solve all problems by yourself. If you represent other people's work as your own then that constitutes fraud and will be dealt with accordingly.

Organization

Components:

1.	Lectures	Tuesday 5+6	MF 6
		Thursday 7+8	MF 6
2.	Tutorials	Tuesday 7+8	MF 6

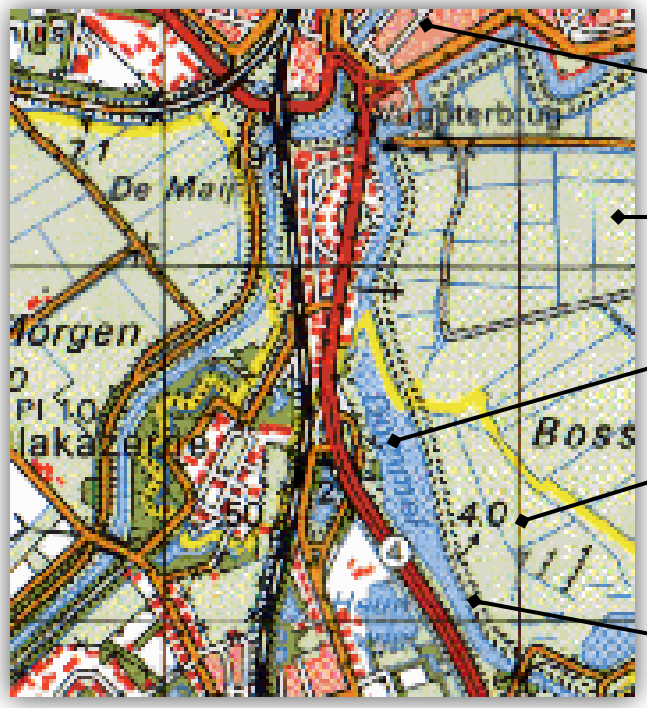
Tutorials are used for many things: assignment of project groups, feed-back to project groups, interim presentations of project groups, feed-back on assignments, presentations of assignment solutions, … integral part of the course

- Check web-page for details

	Lecture					Assignments		
Week	**Date**	**Topic**	**Slides**	**Material**	**Lecturer**	**Due**		
6	03-02-2015	Introduction & similarity			KB + BS	10-02-2015		
	05-02-2015	Similarity of geographic objects			KB			
7	10-02-2015	Simplification			KB	24-02-2015		
	12-02-2015	Schematization			BS			
8	17-02-2015	**Carneval**						
	19-02-2015							
9	24-02-2015	Movement patterns			BS	03-03-2015		
	26-02-2015	Segmentation			KB			
10	03-03-2015	Labeling			BS	10-03-2015		
	05-03-2015	Dynamic Point Labeling			KB			
11	10-03-2015	Cartograms			BS	24-03-2015		
	12-03-2015	Flow & Symbol maps			K.Verbeek			
12	17-03-2015	**Time to wrap up projects**						
	19-03-2015							
13	24-03-2015	Advanced topic/student lectures			(KB)			
	26-03-2015							
14	31-03-2015	student lectures						
	02-04-2015							

Geographic Data

On maps …



urban

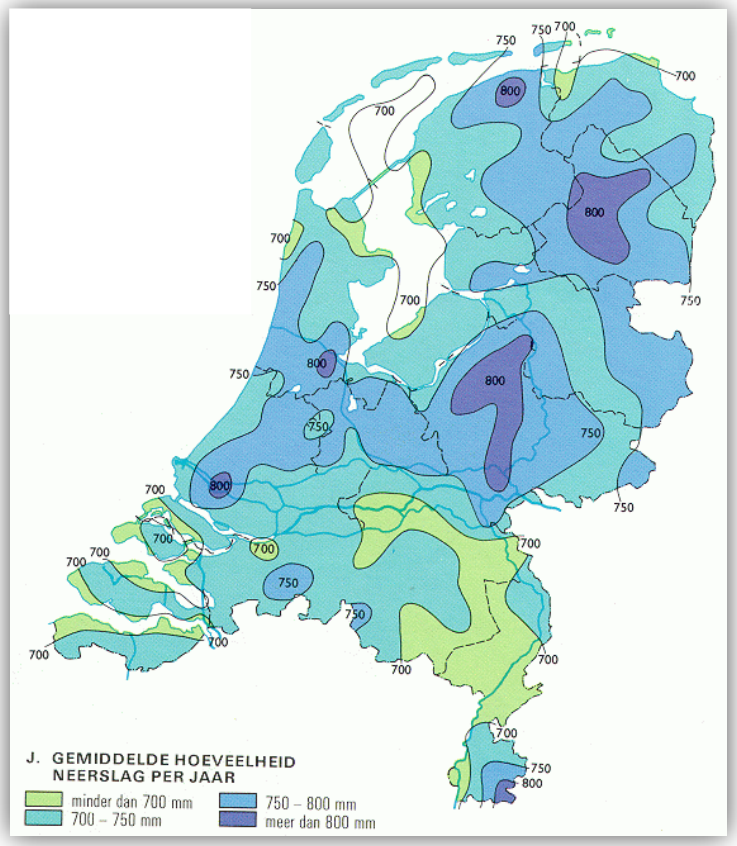
grass

water

text (name, elevation)

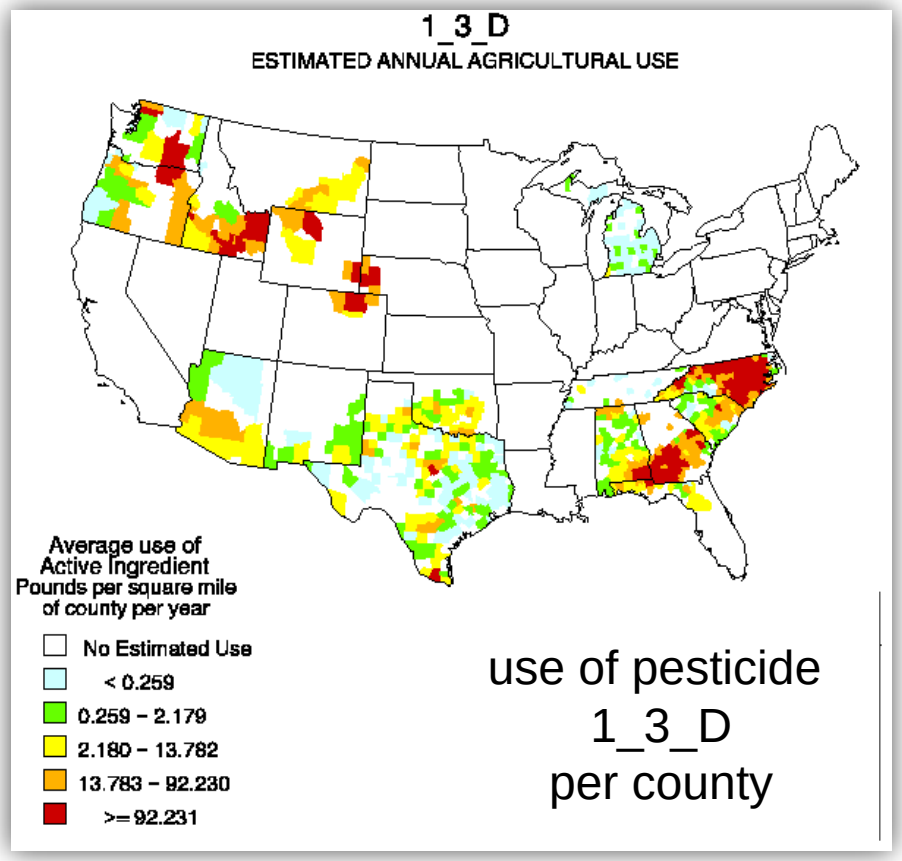
dike

Topographic map



Classified isoline map

On maps …

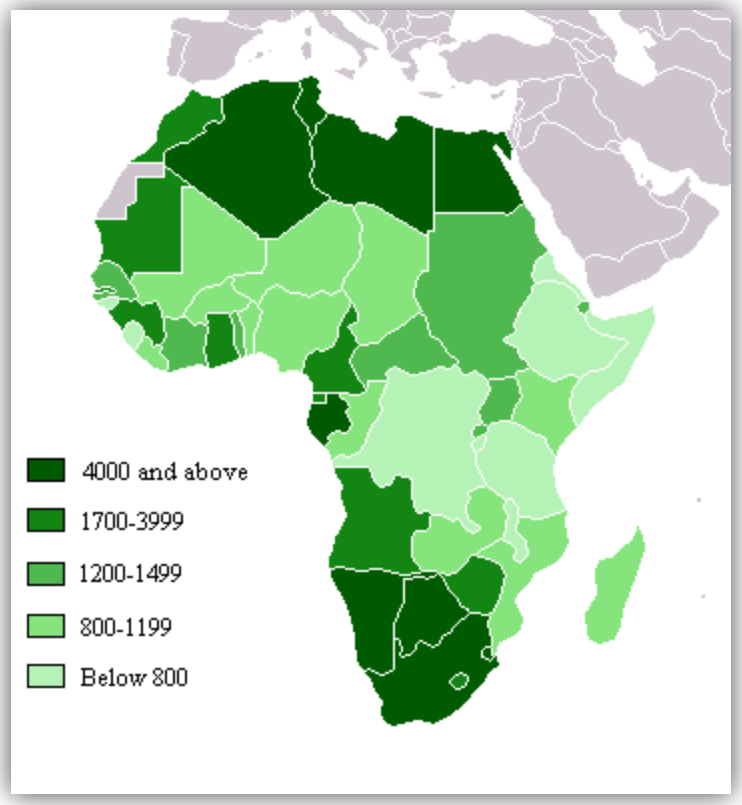


use of pesticide
1_3_D
per county

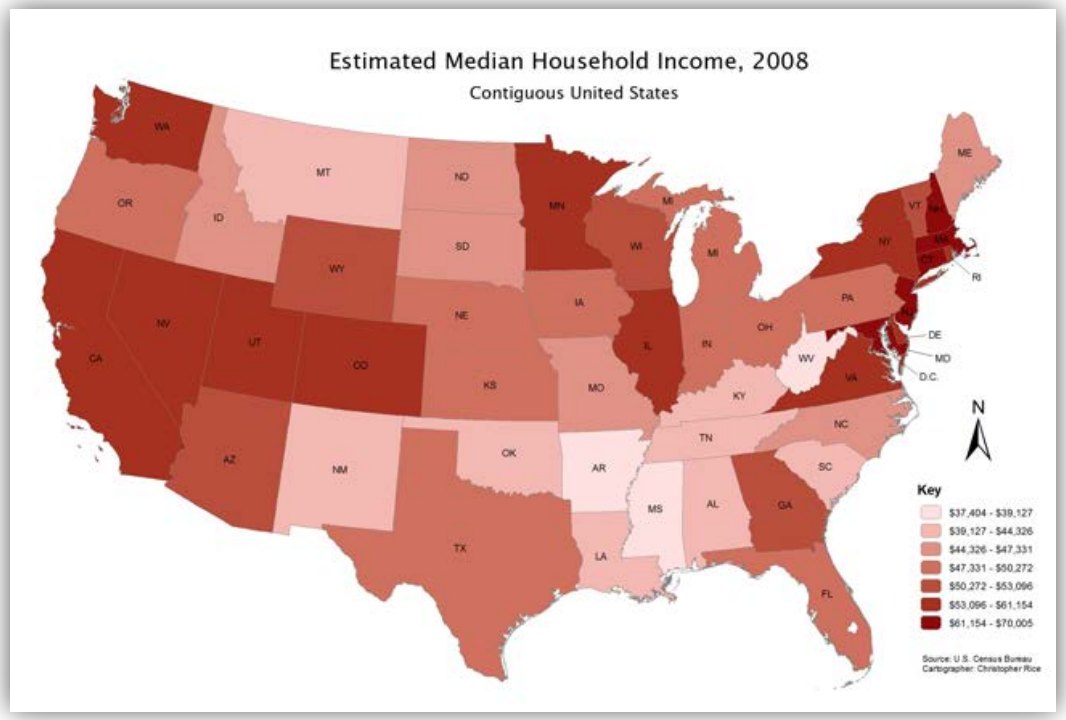
Choropleth map

Map with administrative boundaries which shows per region a value by a color or shade

More choropleth maps



Africa, GDP



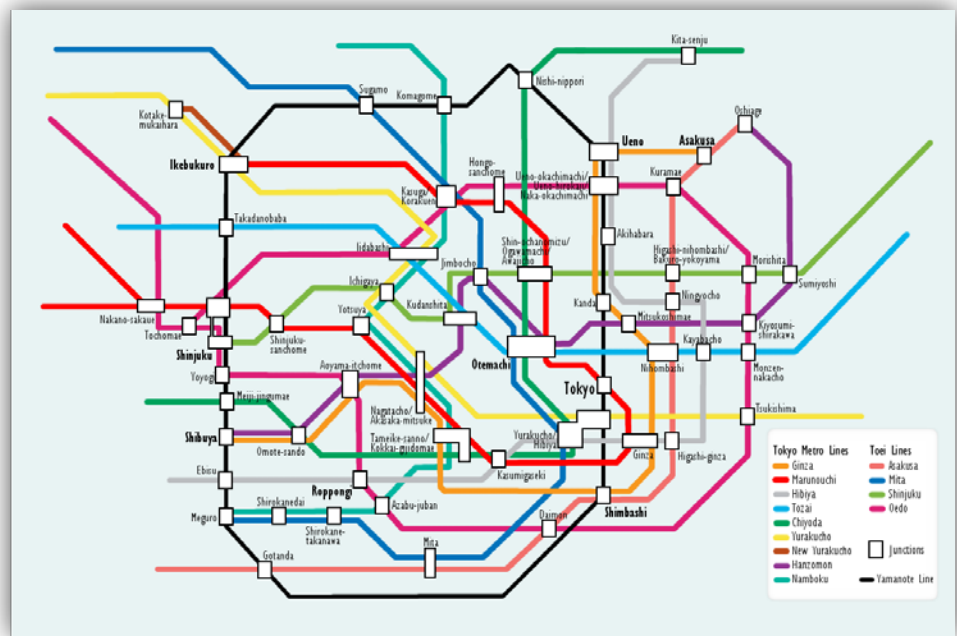
Maps show …

Relation of place (geographic location)
to a value (*here 780 mm precipitation*) or
name (*here is Minnesota*)

Abstraction (model, simplification) of reality

Combination of themes (different kinds of data)

Connections (subway maps)



Tokyo subway map

Scales of measurement

Classification of types of data by statistical properties (Stevens, 1946)

Nominal scale

Ordinal scale

Interval scale

Ratio scale

(Angle/direction, vector, …)

Nominal scale

Administrative maps
names of countries, …

Land use maps
names of land use: urban, grass, forest, water, …

Geological maps
names of soil types: sand, clay, rock, …

- Finite number of classes, each with a name
- Testing is possible for equivalence of name

Ordinal scale

School type

VMBO, HAVO, VWO, …

Wind force on scale of Beaufort

0=no wind, ... 6=heavy wind, …, 9=storm, ...

Questionnaire-answers

disagree, partly disagree, neutral, partly agree, agree

- Finite number of classes, each with a name
- Testing for equivalence of name and for order

Interval scale

Temperature

in degrees Celsius or Fahrenheit

Time/year

on Christian calendar

- Unbounded number of classes, each with a value
- Testing for equivalence, for order and for difference (a unit distance exists)

Ratio scale

Measurements

concentration of lead in soil, …

Counts

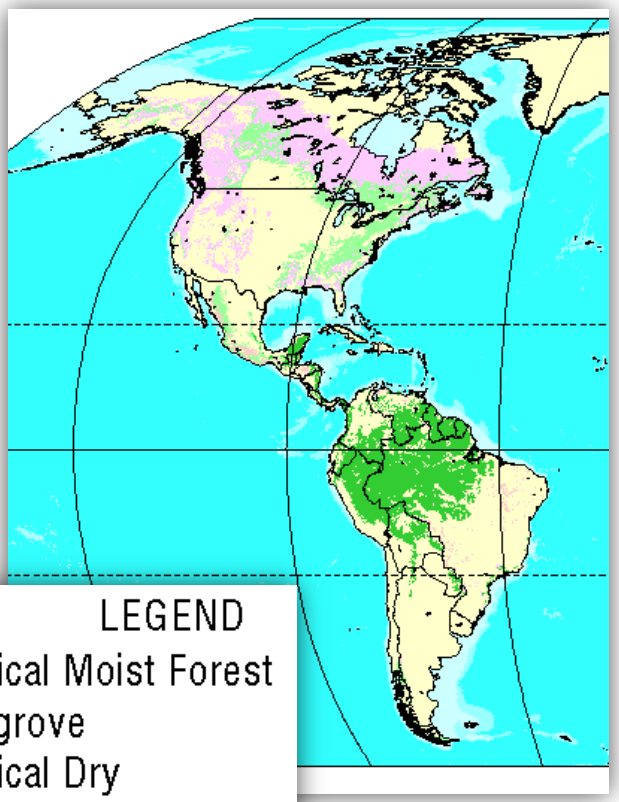
population, number of airports, …

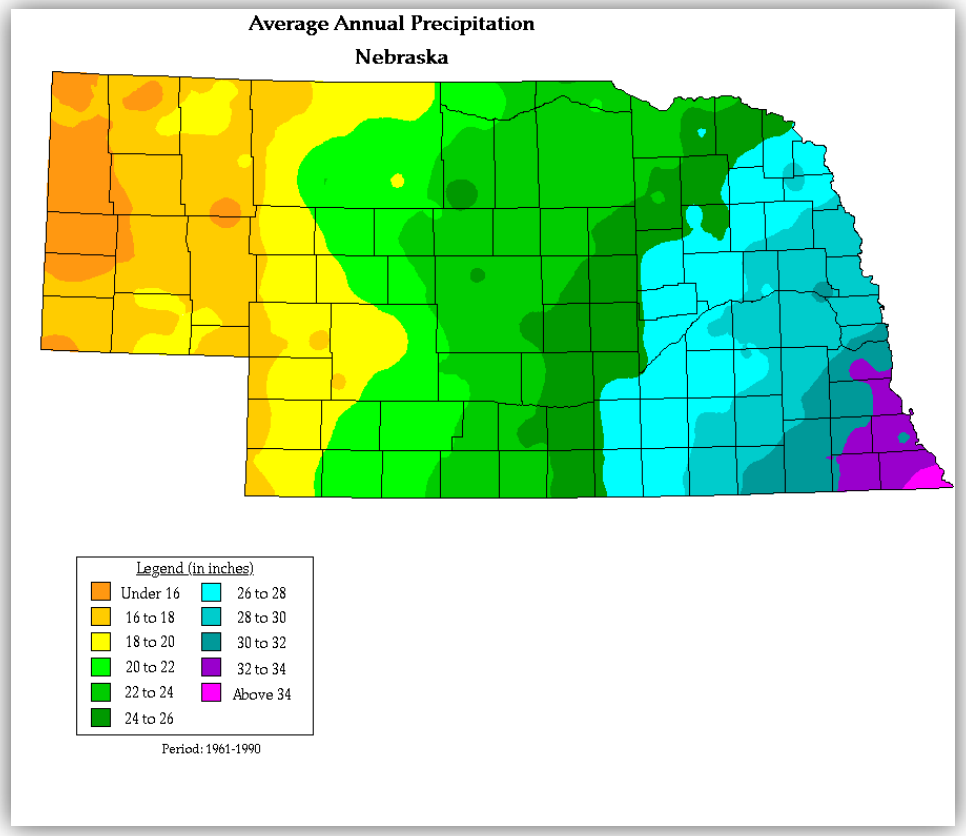
Percentages

unemployment percentage, percent of landuse type forest, …

- Unbounded number of classes, each with a value
- Testing for equivalence, for order, for difference and for ratio (a natural zero exists)

Examples


LEGEND
Tropical Moist Forest
Mangrove
Tropical Dry
Temperate Broadleaf/
Temperate Mixed
Temperate Needleleaf


Average Annual Precipitation
Nebraska
Legend (in inches)
Under 16
16 to 18
18 to 20
20 to 22
22 to 24
24 to 26
26 to 28
28 to 30
30 to 32
32 to 34
Above 34
Period: 1961-1990

Overview

		two items	collection
nominal	categories	equivalence	number of occurrences, mode
ordinal	categories	… and order	… and median
interval	unbounded	… and difference	… and average
ratio	unbounded	… and ratio	… and normalize

Other scales

Angle

wind direction, direction of spreading, …

Vector

angle and value (primary wind direction and speed)

Categorical scales with partial membership

fuzzy sets, points on indeterminate boundary between “plains” and “mountains”, location of coast line: tide, …

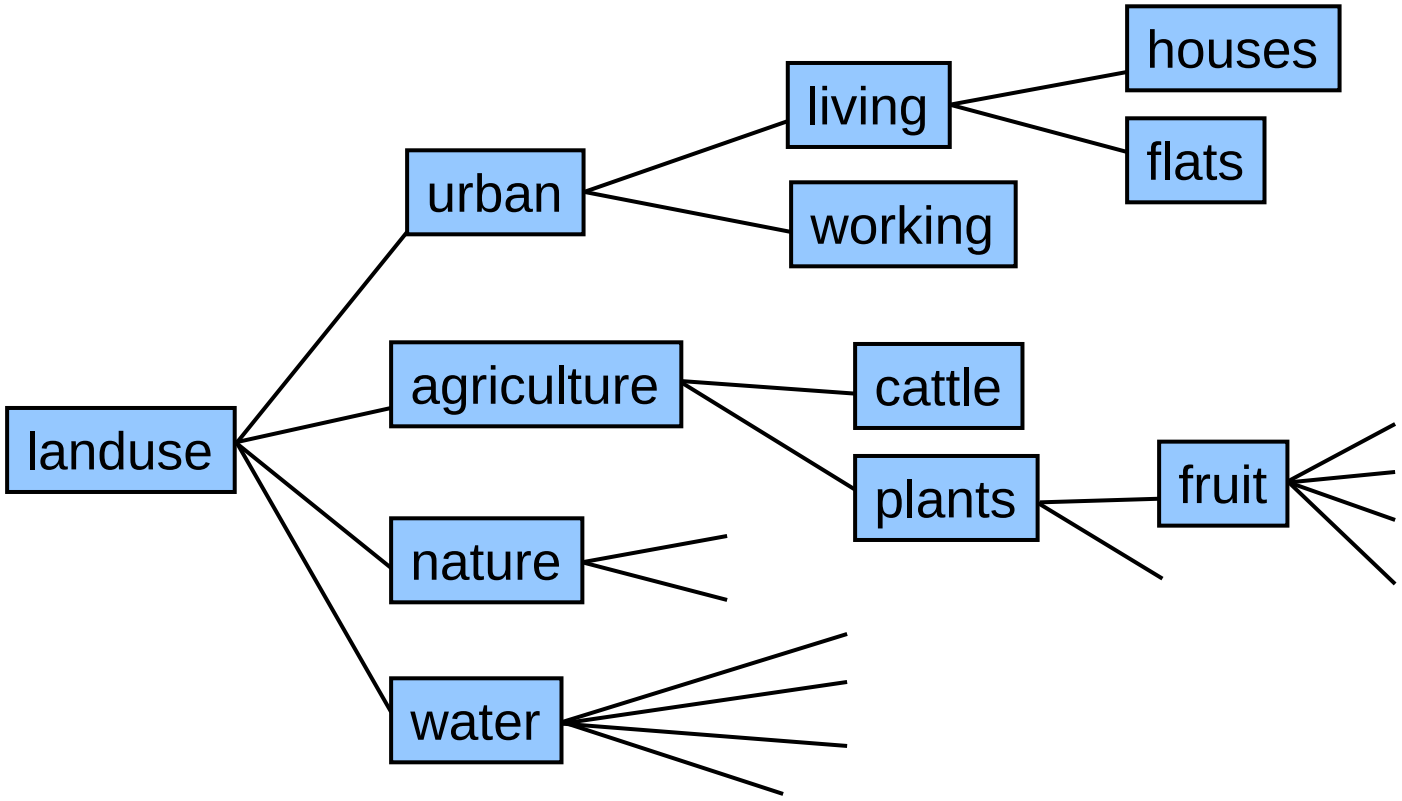
Example


Miami
Nassau
Le Havane
Cancún
Belize City
Kingston
Port-au-Prince
San Juan
San Salvador
Bridgetown
San José
Caracas
Panama
1010
The Weather Channel
weather.com
24 Apr 2007 07:02 GMT

Classification schemes

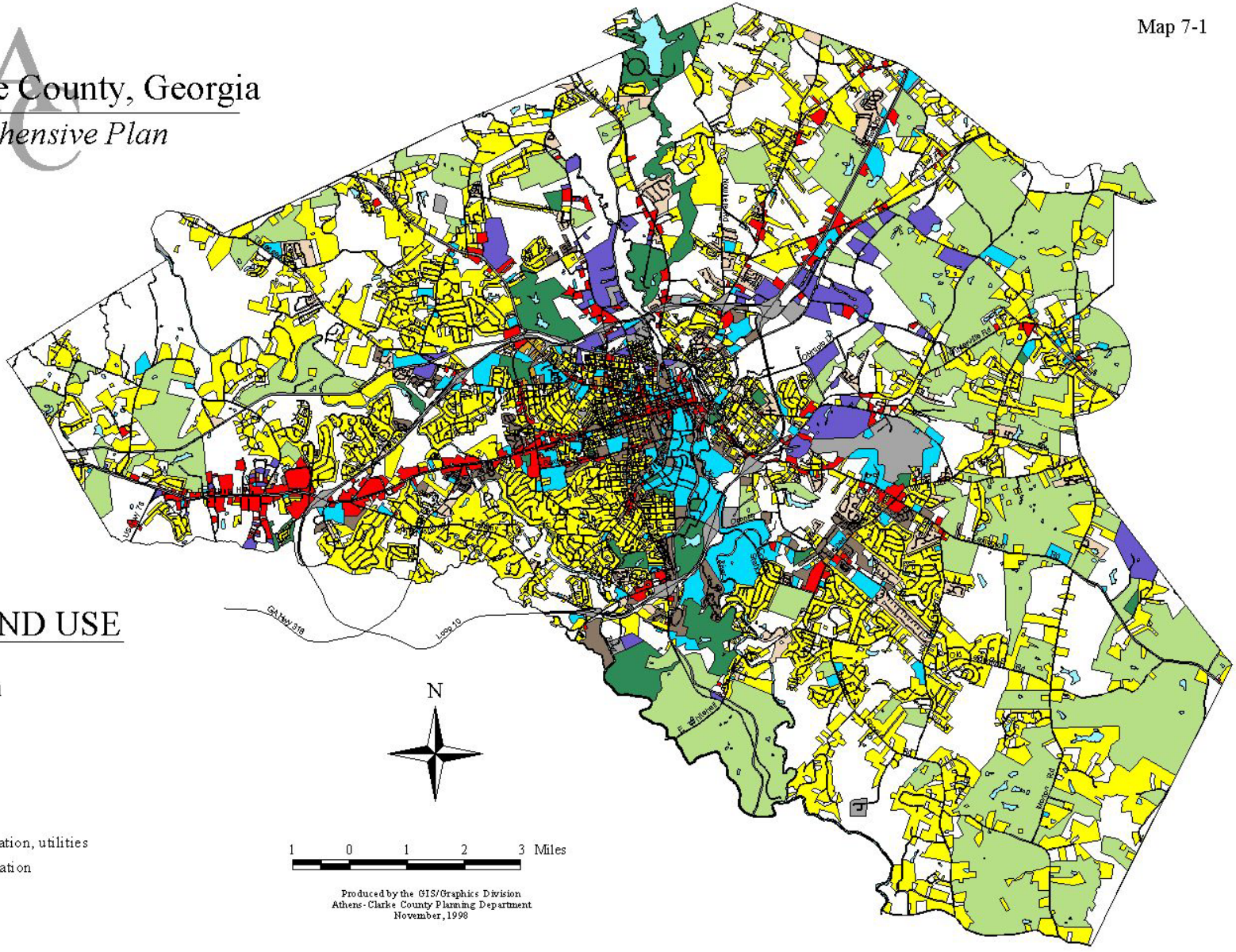
Data on nominal scale

hierarchical classification schemes



Athens-Clarke County, Georgia

Comprehensive Plan



EXISTING LAND USE

- Low density residential
- Medium density residential
- High density residential
- Office
- Commercial
- Industrial
- Public, institutional
- Transportation, communication, utilities
- Park, recreational, conservation
- Agriculture, forestry
- Open water
- Undeveloped

N

1 0 1 2 3 Miles

Produced by the GIS/Graphics Division
Athens-Clarke County Planning Department
November, 1998

Classification schemes

Data on interval and ratio scales	4, 5, 5, 8, 12, 14, 17, 23, 27
Fixed intervals	[1-10], [11-20], [21-30]
Fixed intervals based on spread	[4-11], [12-19], [20-27]
Quantiles equal representatives	[4-5], [8-14], [17-27]
“Natural” boundaries	[4-5], [8-17], [23-27]

Statistical boundaries

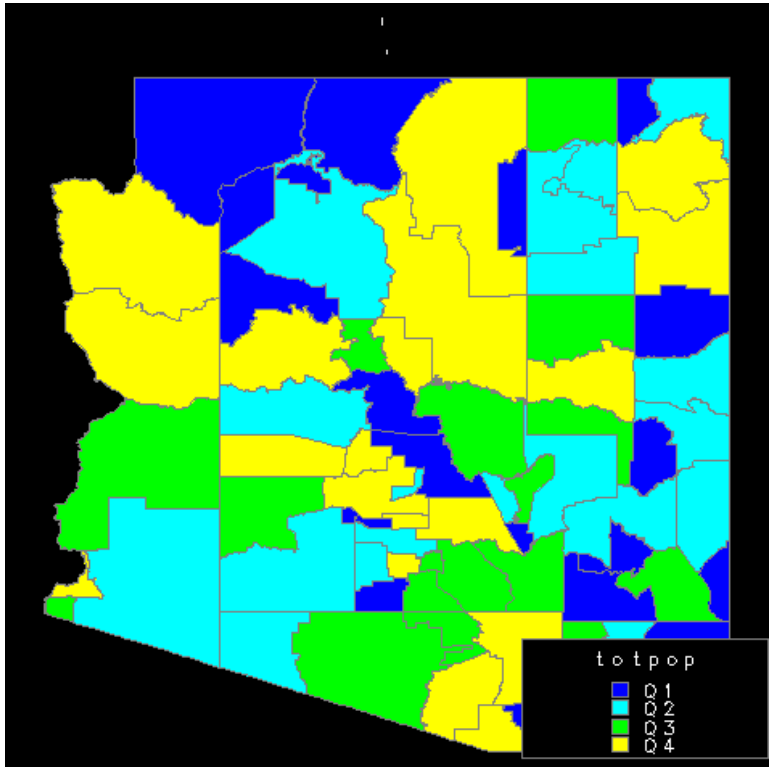
average μ, standard deviation σ, for example boundaries

$\mu - 2\sigma$, $\mu - \sigma$, μ, $\mu + \sigma$, $\mu + 2\sigma$

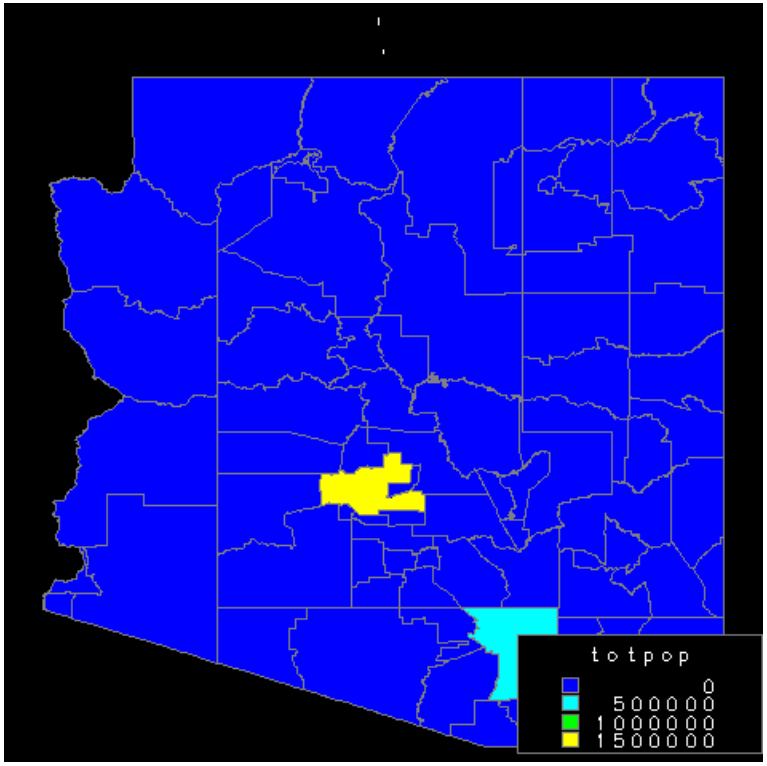
Arbitrary

Two classifications

Counties of Arizona, total population



Quartiles



Four equal intervals

Why is choice of classification important?

- Visualization often needs classification
- Choice of class intervals influences interpretation

Imagine a report that addresses air pollution due to a factory compiled by the board of the factory or by an environmental organization …

Data: object and field view

Object view

discrete objects in the real world

- road
- telephone pole
- lake

Field view

geographic variable has a “value” at every location in the real world

- elevation
- temperature
- soil type
- land cover

Spatial objects

Points
0-dimensional, *measurement point*

(Polygonal) line
1-dimensional, *border between Bolivia and Peru*

Polygons
2-dimensional, *Switzerland*

Sets of points
locations of accidents, …

Systems of lines (trees, graphs),
street network, …

Sets of polygons, subdivisions
island group, provinces of the Netherlands, …

Dependency of dimension

Dimension of an object can be scale dependent

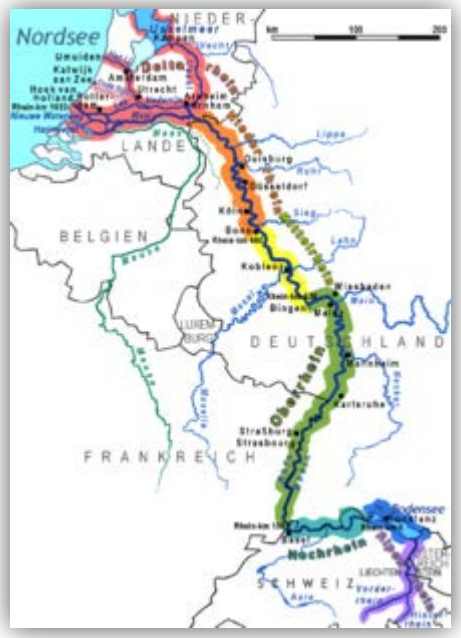
Rhine at scale 1 on 25.000 is 2-dimensional

Rhine at scale 1 on 1.000.000 is 1-dimensional

Dimension of an object can be application dependent

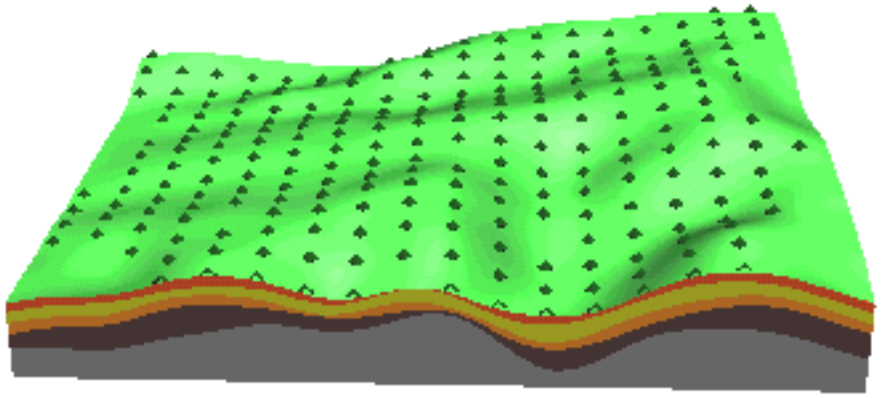
Rhine as transport route is 1-dimensional
length is relevant; not the surface area

Rhine as land cover in the Netherlands is 2-dimensional



The third dimension

- Elevation can be considered an attribute on the ratio (!?) scale at (x,y)-coordinates
- For civil engineering: crossing of street and railroad can be at the same level, or one above the other
- Data on subsurface layers and their thickness



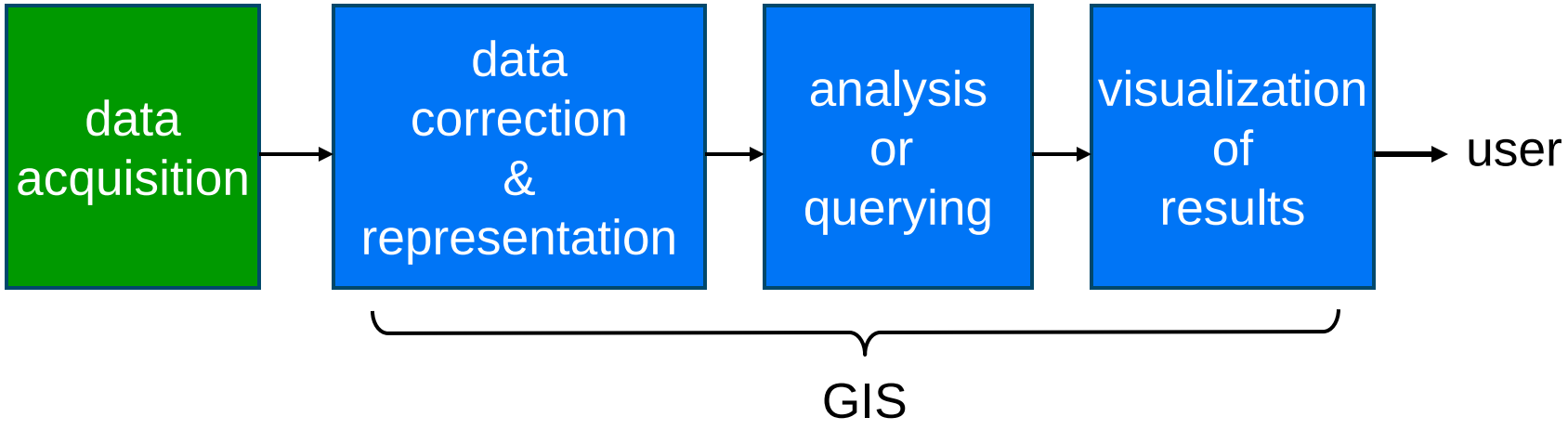
GIS

Geographic Information Systems / Science

Geographic Information System

Geographic Information System (GIS)

stores, manipulates, analyzes, and visualizes geographic data



Automated cartography

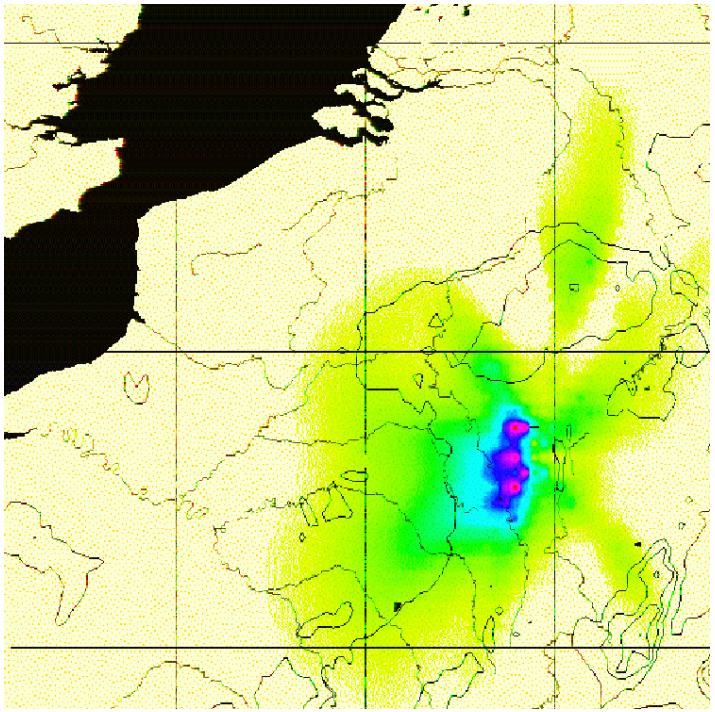
automated methods for the visualization of geographic data

Automated cartography

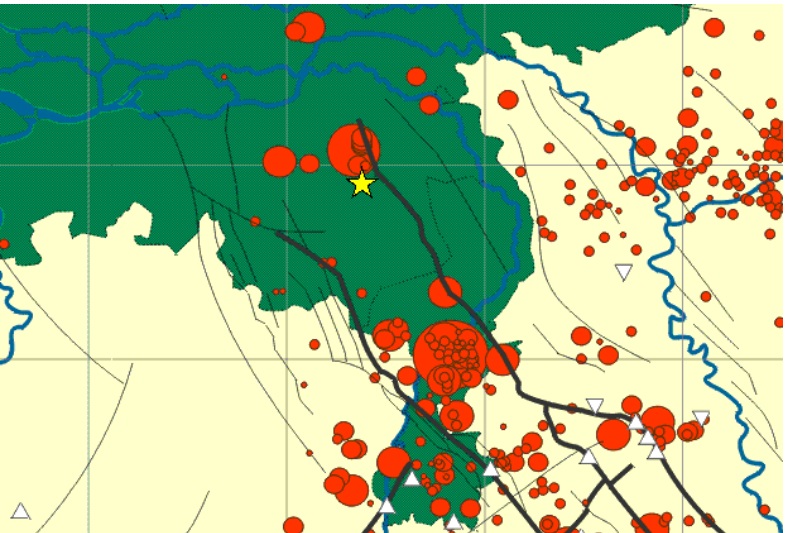
Automated cartography

... *strives to automate all task that are traditionally performed manually by a cartographer …*

- Visualize the results of an analysis
- Make maps
  - general / special purpose
  - cartographic generalization
  - label placement



air pollution analysis



earthquakes and break lines

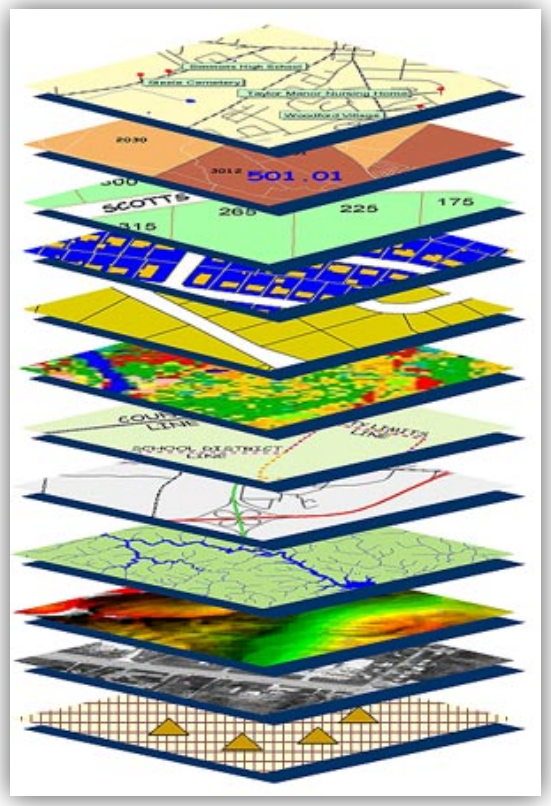
Thematic map layers

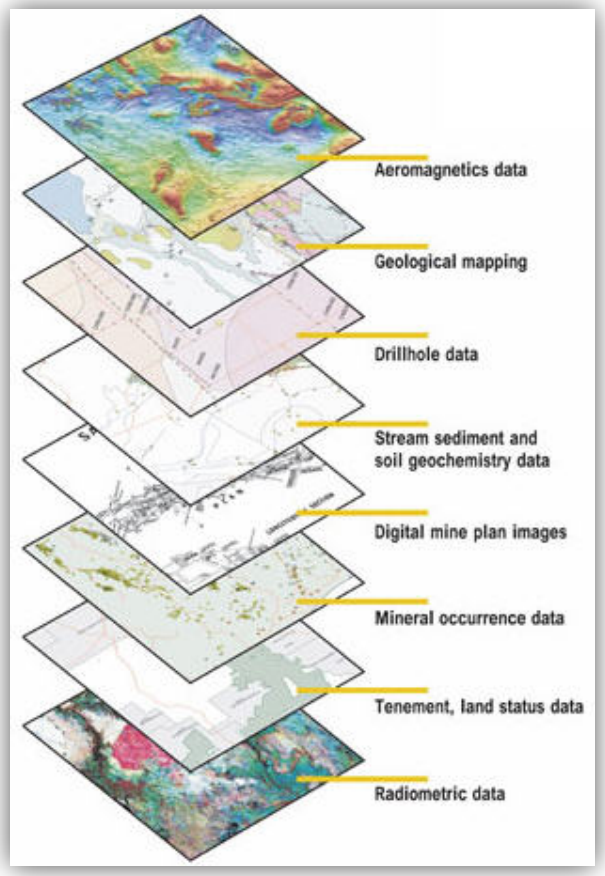
Map layers (or data layers)

separate storage of data according to theme

GIS typically use tens to hundreds of map layers

municipality borders,
land use,
cadastral boundaries,
water pipes,
churches, …





Geometry, topology and attributes

Geometry

coordinates

Topology

adjacency relations of objects

Attributes

properties, values

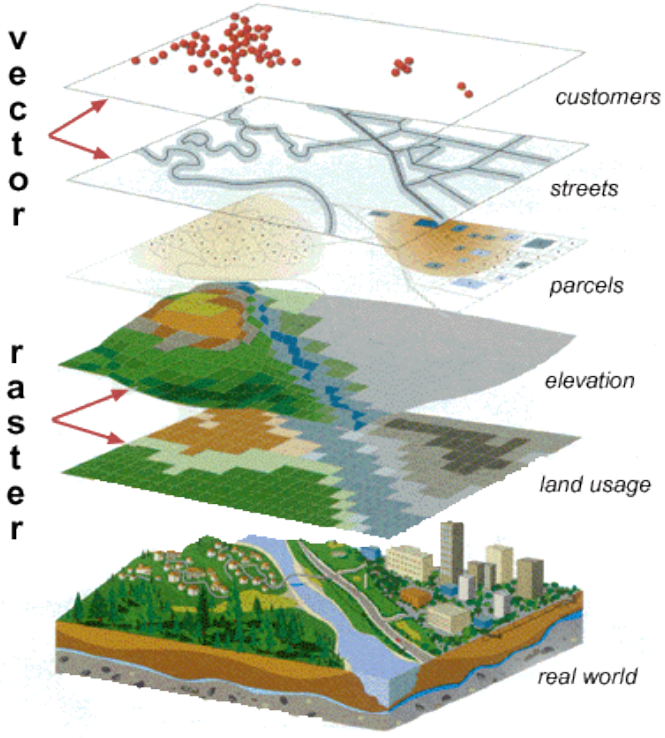
Example: Country map of South America

- geometry
  - coordinates of the borders
- topology
  - which countries border which
- attributes
  - names of countries, population, …

Representation of geometry

Raster and vector

- Can be mixed in a GIS, any map layer
- Conversion raster-vector and vice versa
- Representation depends on
  type of data,
  way of acquisition,
  desired operations, …

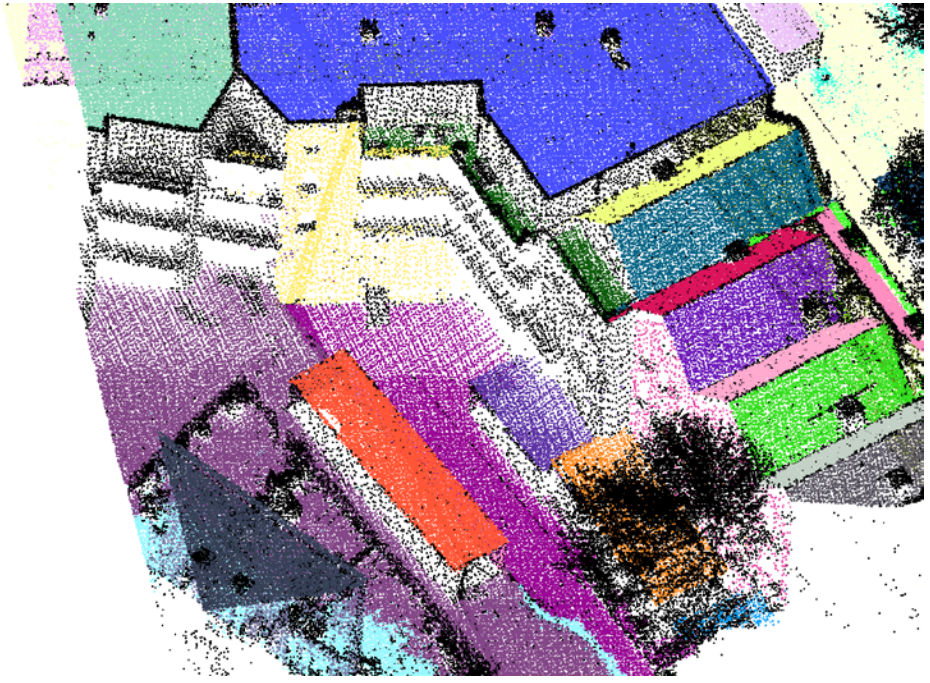


What is typical about GIS?

- Applications are spatial
- Problems are not well-defined: multiple criteria must be met simultaneously, concepts often not well-defined (e.g. Groene Hart, surroundings of Arnhem)
- Data is unreliable, expensive to get, sometimes outdated, sometimes only partially relevant, difficult to integrate, ...

GIS problems

- ... do not have an algorithmic problem statement yet
  - placing a river label nicely curved along it on a map
  - making a 3D city model from a LIDAR point cloud
  - segmenting a trajectory into pieces of similar behavior
  - selecting roads to keep for map generalization



GIS problems

Input: Usually not hard to specify

Output: ???

Formalizing a GIS problem

river labeling

Formalizing a GIS problem: river labeling

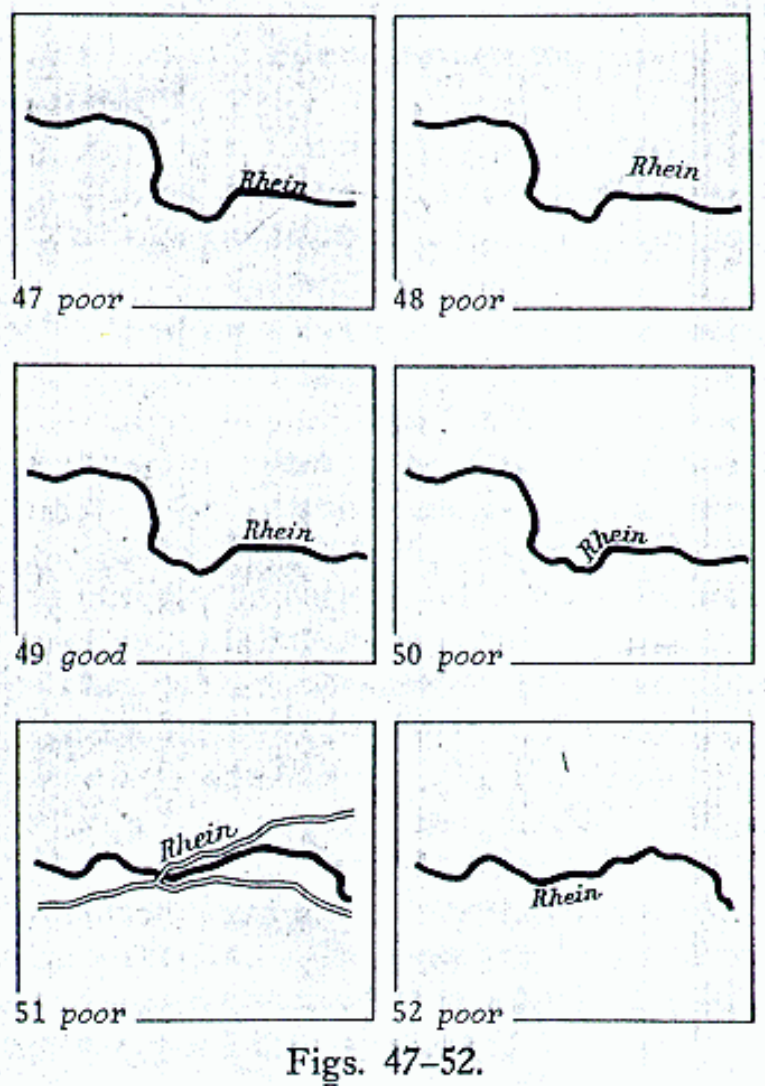
- A river is represented as a polygonal line, spline, or other curve representation
- A text is given and a font size

Input: A polygonal line with n vertices, and a label text and font size

Output: ???

Formalizing a GIS problem: river labeling

- Find out what experts think
  - read papers, books
  - talk to experts
- If there are multiple criteria, study them separately at first

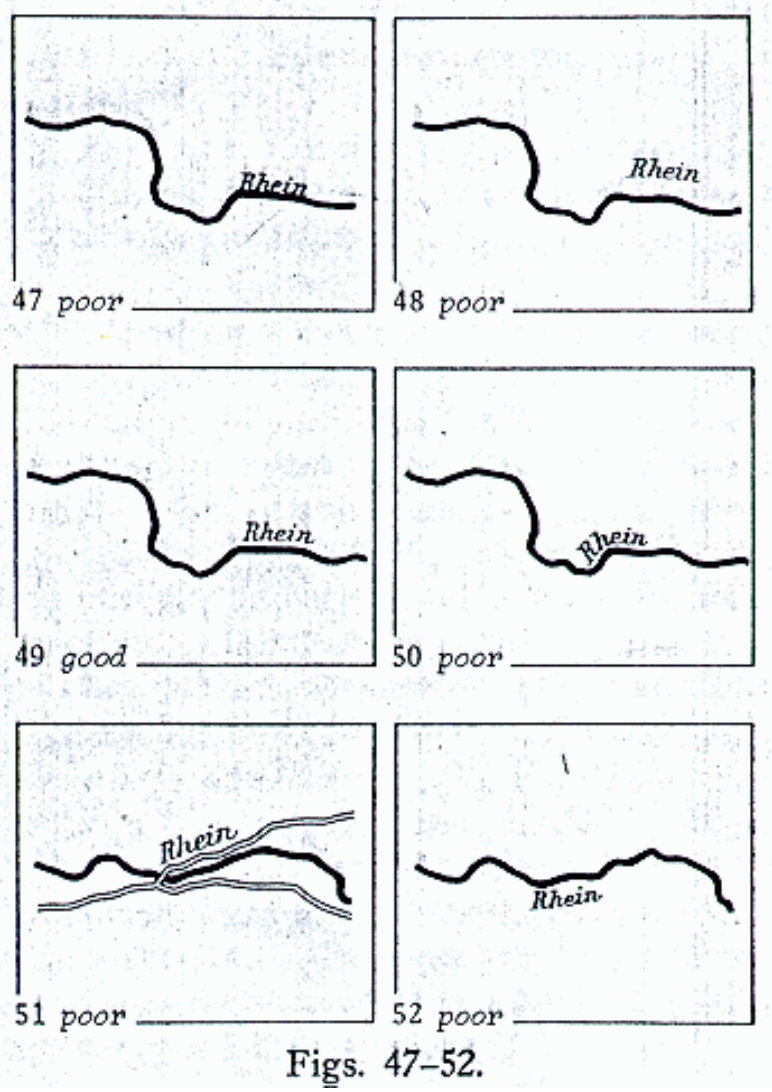


Figs. 47–52.

River label placement

- River label placement rules according to Imhof (1962):
  - Not too close
  - Not too far
  - Not too curved
  - No other line feature in between
  - Not below

When Imhof says “Not too far”, what does he mean, and how should we measure distance?



Figs. 47–52.

River label placement

- A river representation is a geometric object
- A label, when placed, can also be seen as a geometric object

… so we use a geometric distance measure for “not too far”

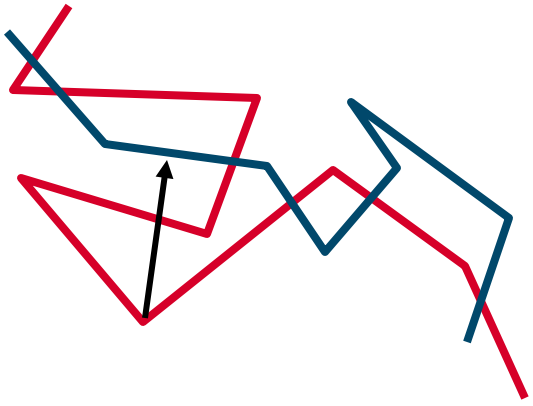
How about simply using shortest distance?

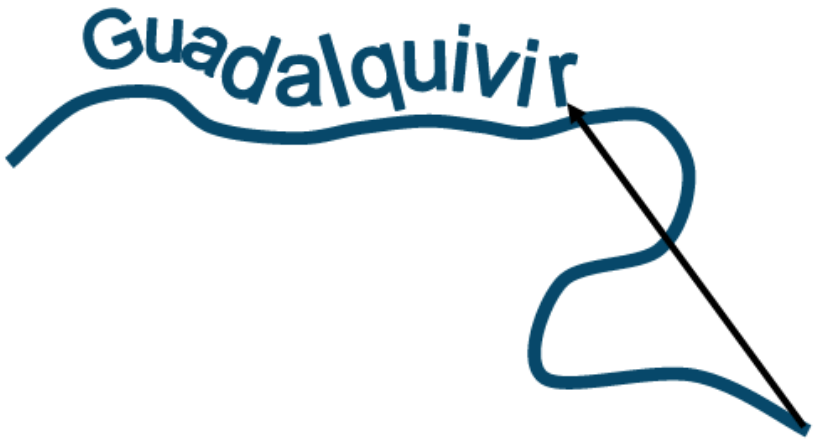
River label placement

- A river representation is a geometric object
- A label, when placed, can also be seen as a geometric object

… so we use a geometric distance measure for “not too far”

How about using the Hausdorff distance?





River label placement

- A river representation is a geometric object
- A label, when placed, can also be seen as a geometric object

… so we use a geometric distance measure for “not too far”

How about using the directed Hausdorff distance, from the baseline of the label to the river?

River label placement

- “Not too far” can be formalized best by the directed Hausdorff distance from the baseline of the label to the river
- “Not too close” can be formalized by requiring a minimum distance between the label and the river

- Not too curved: need to define curvature of a text
- No line feature in between: need to define “where” is in between
- Not below

River label placement

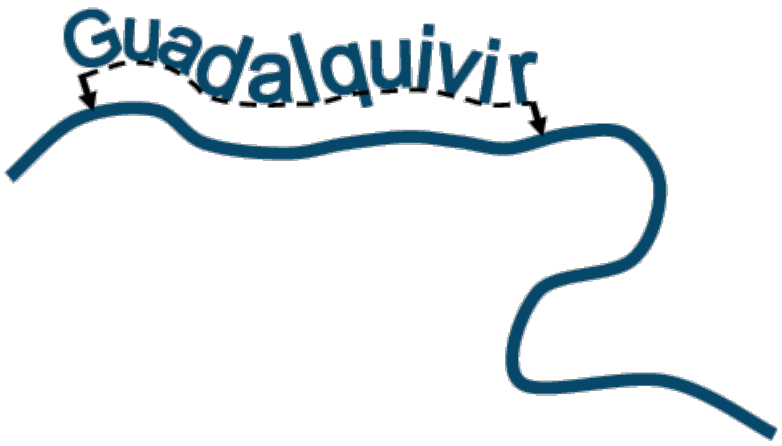
- “Not too far” can be formalized best by the directed Hausdorff distance from the baseline of the label to the river
- “Not too close” can be formalized by requiring a minimum distance between the label and the river

- curvature c: not more bent than a circle with radius 1/c

- Not too curved: need to define curvature of a text
- No line feature in between: need to define “where” is in between
- Not below

River label placement

- “Not too far” can be formalized best by the directed Hausdorff distance from the baseline of the label to the river
- “Not too close” can be formalized by requiring a minimum distance between the label and the river



- Not too curved: need to define curvature of a text
- No line feature in between: need to define “where” is in between
- Not below

River label placement

Input: A polygonal line with n vertices, and a label text and font size, a minimum distance d, and a maximum curvature c

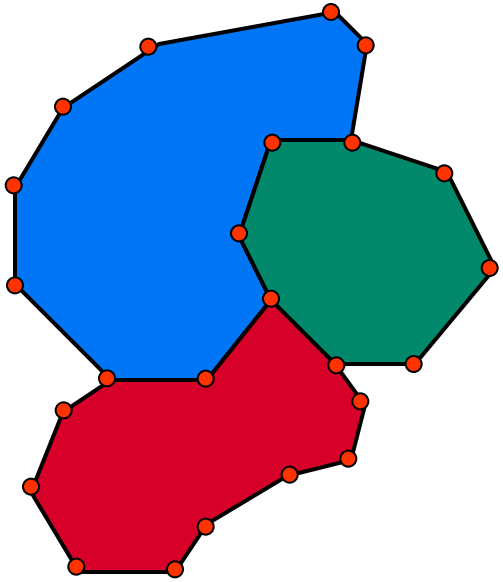
Output: A shape and position of the baseline of the label so that:

- the label is at least d from the river
- the baseline has curvature at most c
- the directed Hausdorff distance from the baseline of the label to the river is minimized
- the bounded region between the ends of the baseline and their closest points on the river does not intersect a line feature
- “not below”

Useful geometric tools

Geometric algorithms

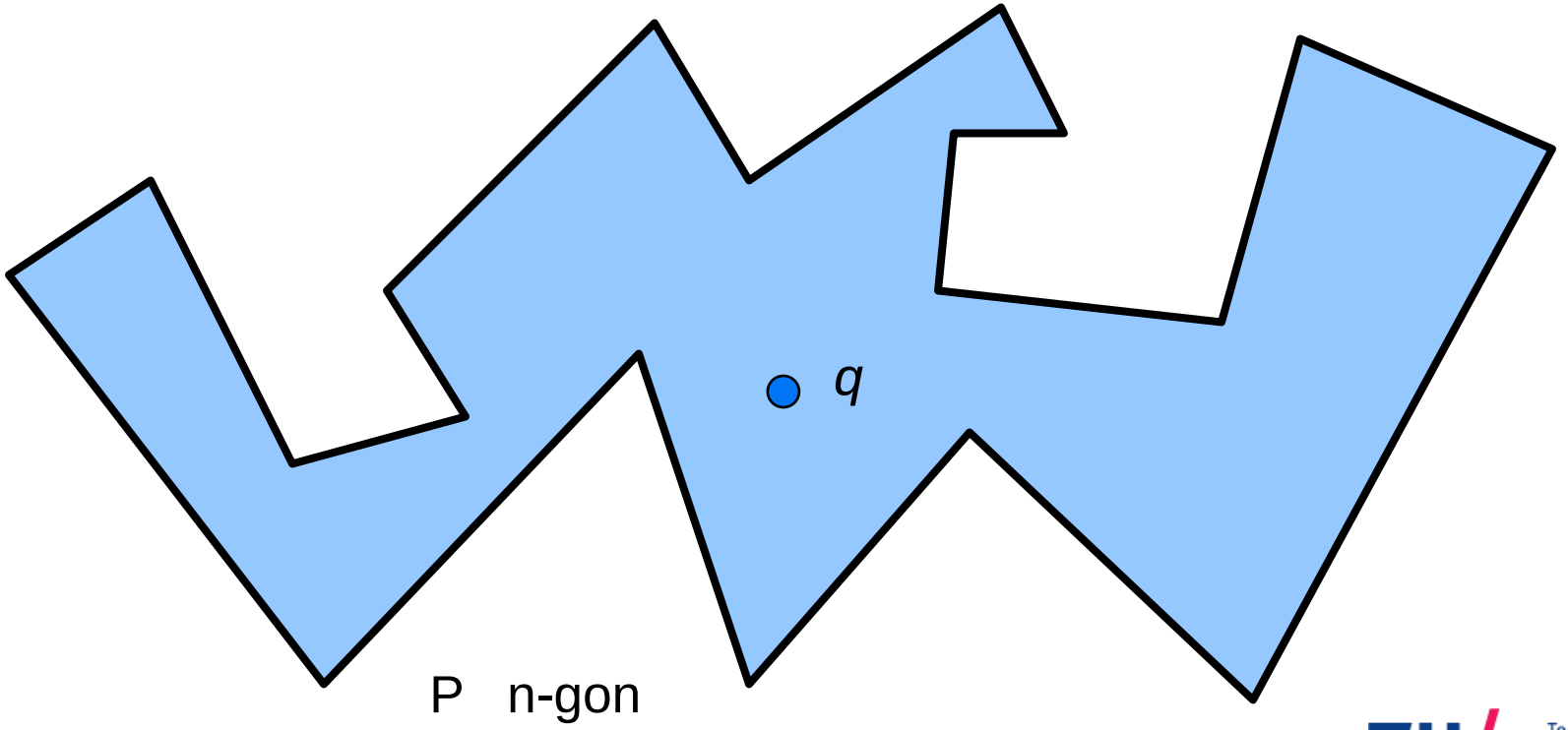
- Geometric data structures useful for geographic data
- This course does not focus on geometry, and geometric algorithms are not a prerequisite
- We will sometimes use geometric algorithms/data structures as a blackbox



Point Location

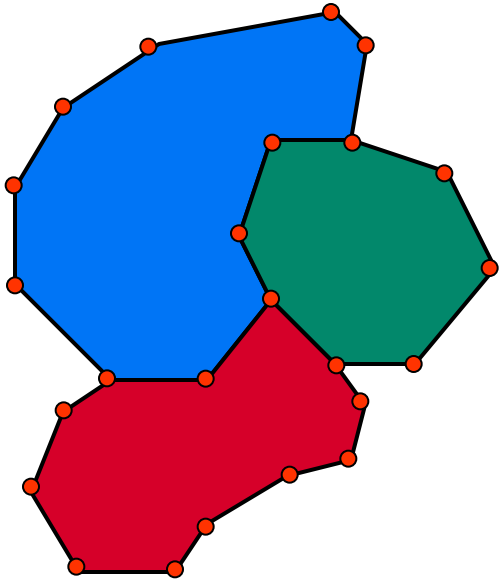
- Is point q inside simple polygon P?

- Naïve: O(n) per test
- CG: O(log n) after preprocessing



Subdivisions: topological structure

- Nodes are objects with coordinates
- Edges are connections of nodes
- Sequences of edges along polygon boundaries are connected
- Polygons are objects of which the boundary is stored

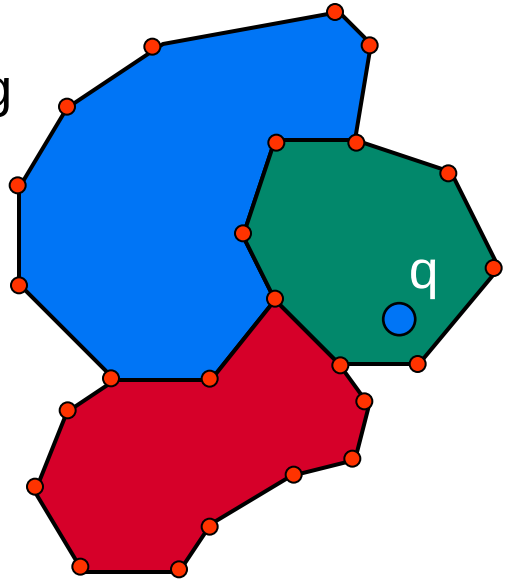


DCEL
Doubly-connected edge list

Point Location

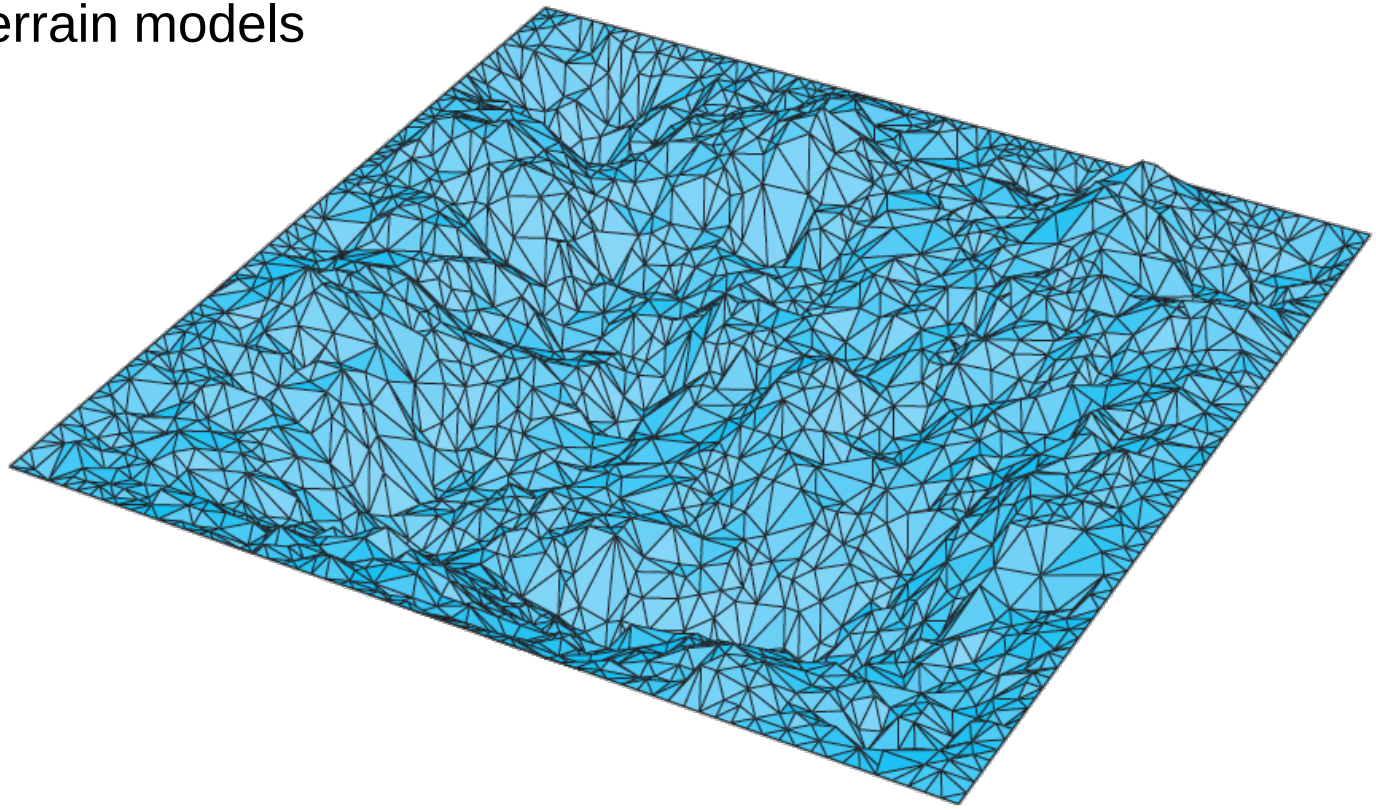
- Which cell of a planar subdivision contains q?

- Naïve: O(n) per test
- CG: O(log n) after O(n log n) preprocessing



Triangulations

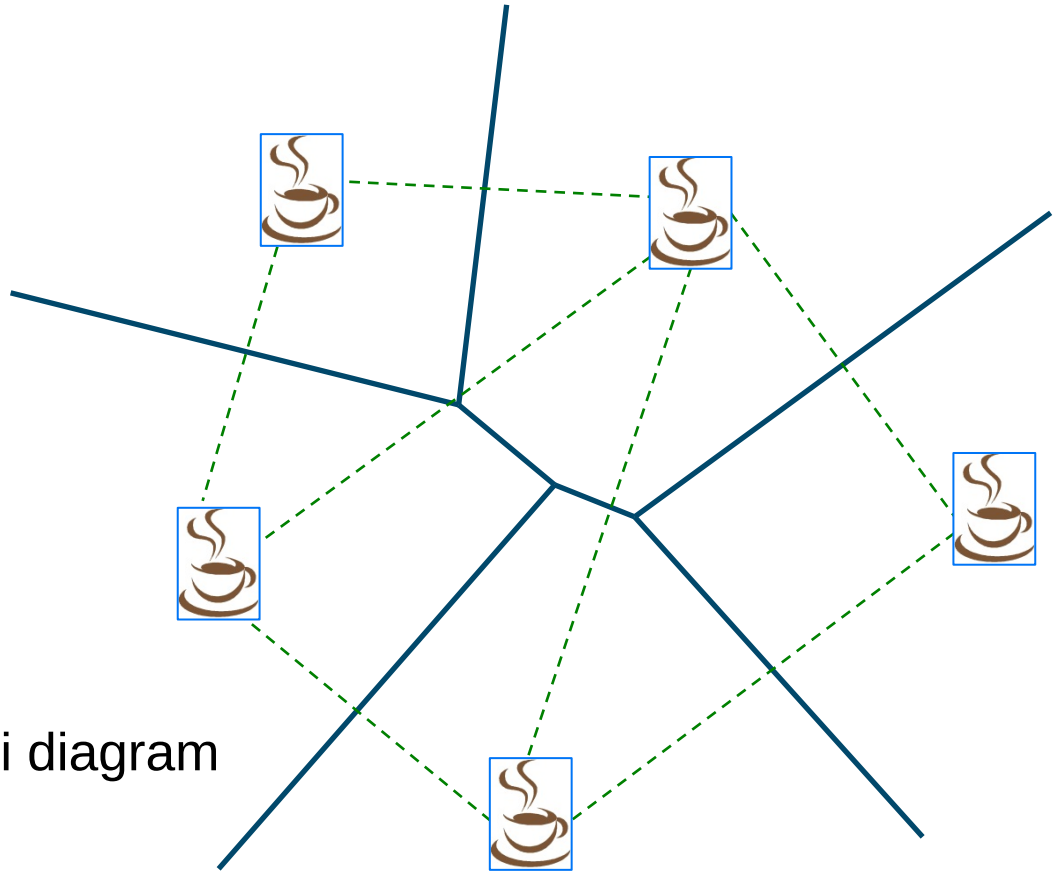
- For terrain models



- A suitable triangulation of a planar point set can be computed in O(n log n) time

Proximity

- Closest coffee machine?



Voronoi diagram

- Computing the Voronoi diagram takes O(n log n) time

Movement Data

Trajectories

- Model for the movement of a (point) object;

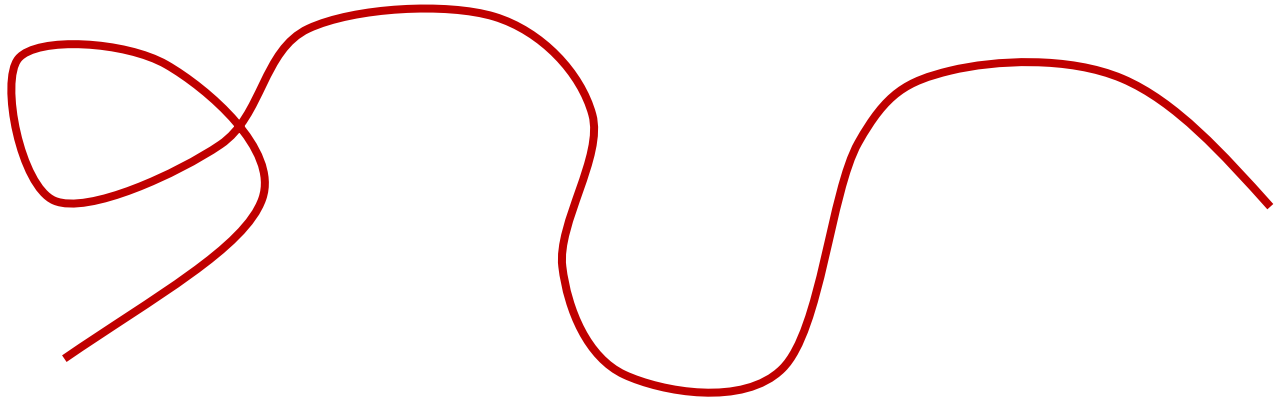
  f : [time interval] → 2D or 3D

Trajectories

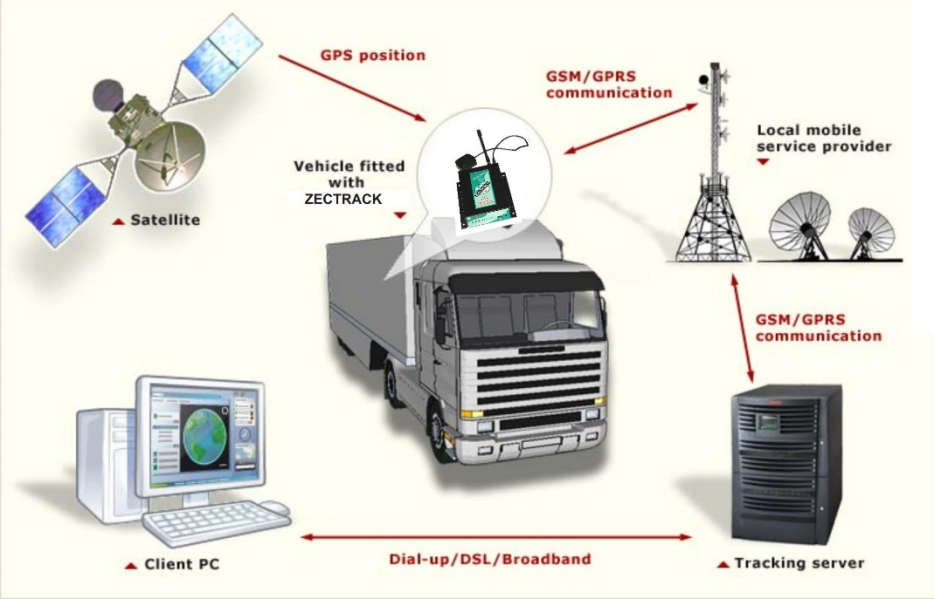
- Model for the movement of a (point) object;

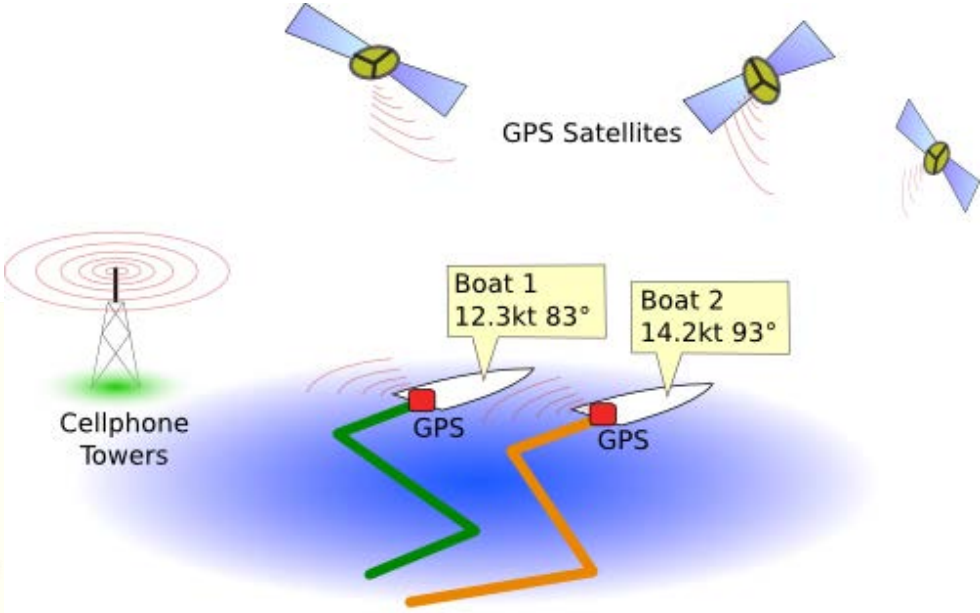
  f : [time interval] → 2D or 3D

- The path of a trajectory is just any curve



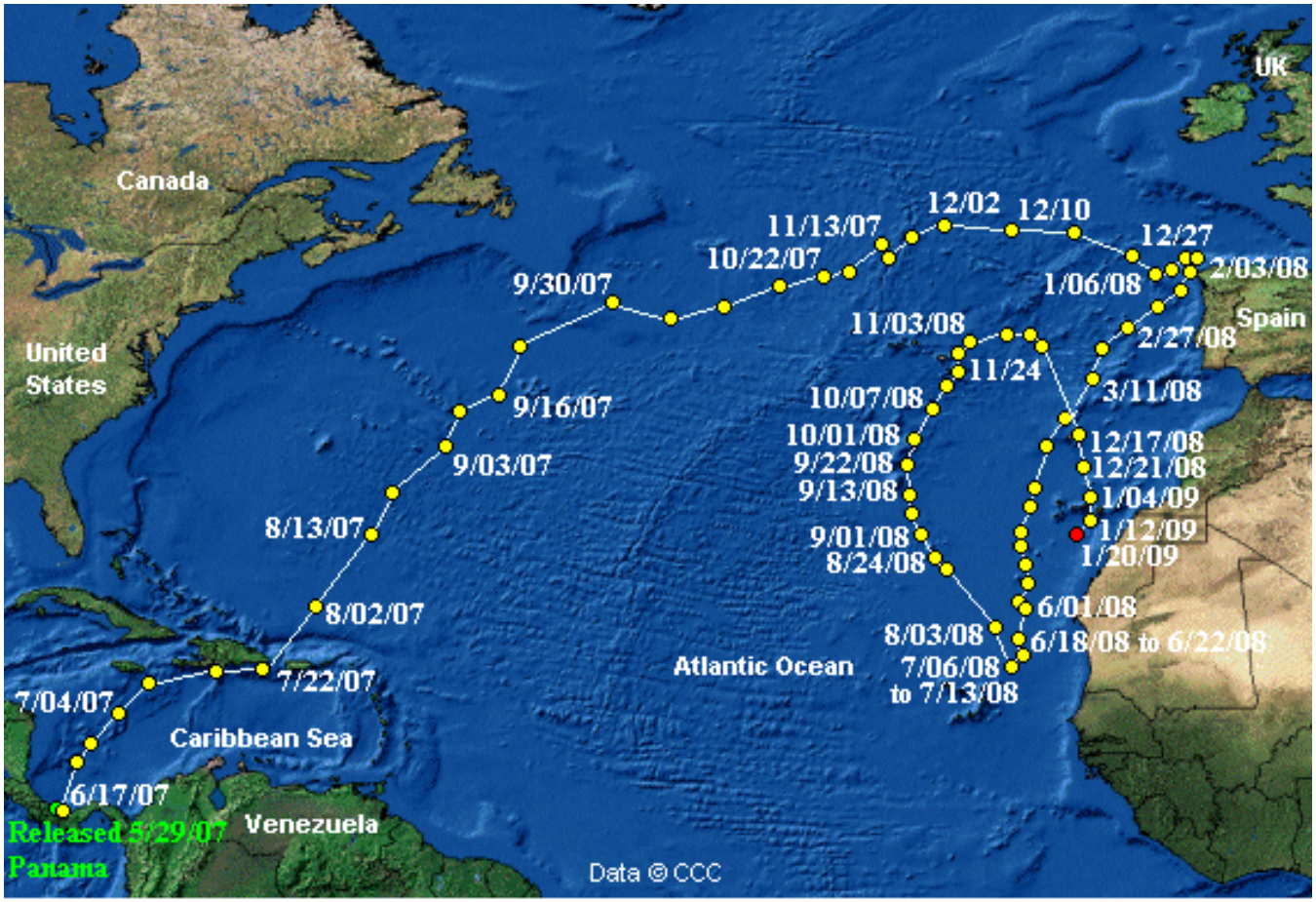
Tracking vehicles



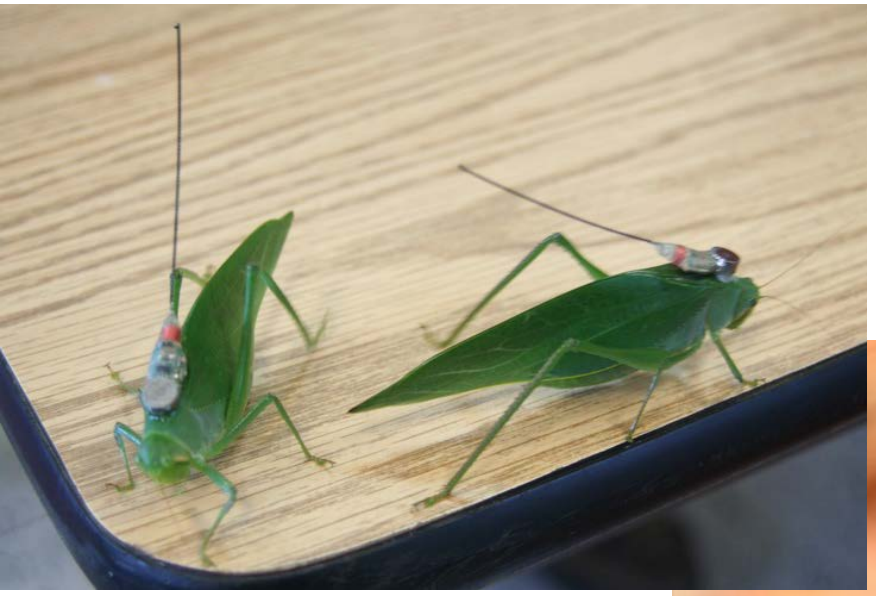


Tracking animals

Tracked turtle

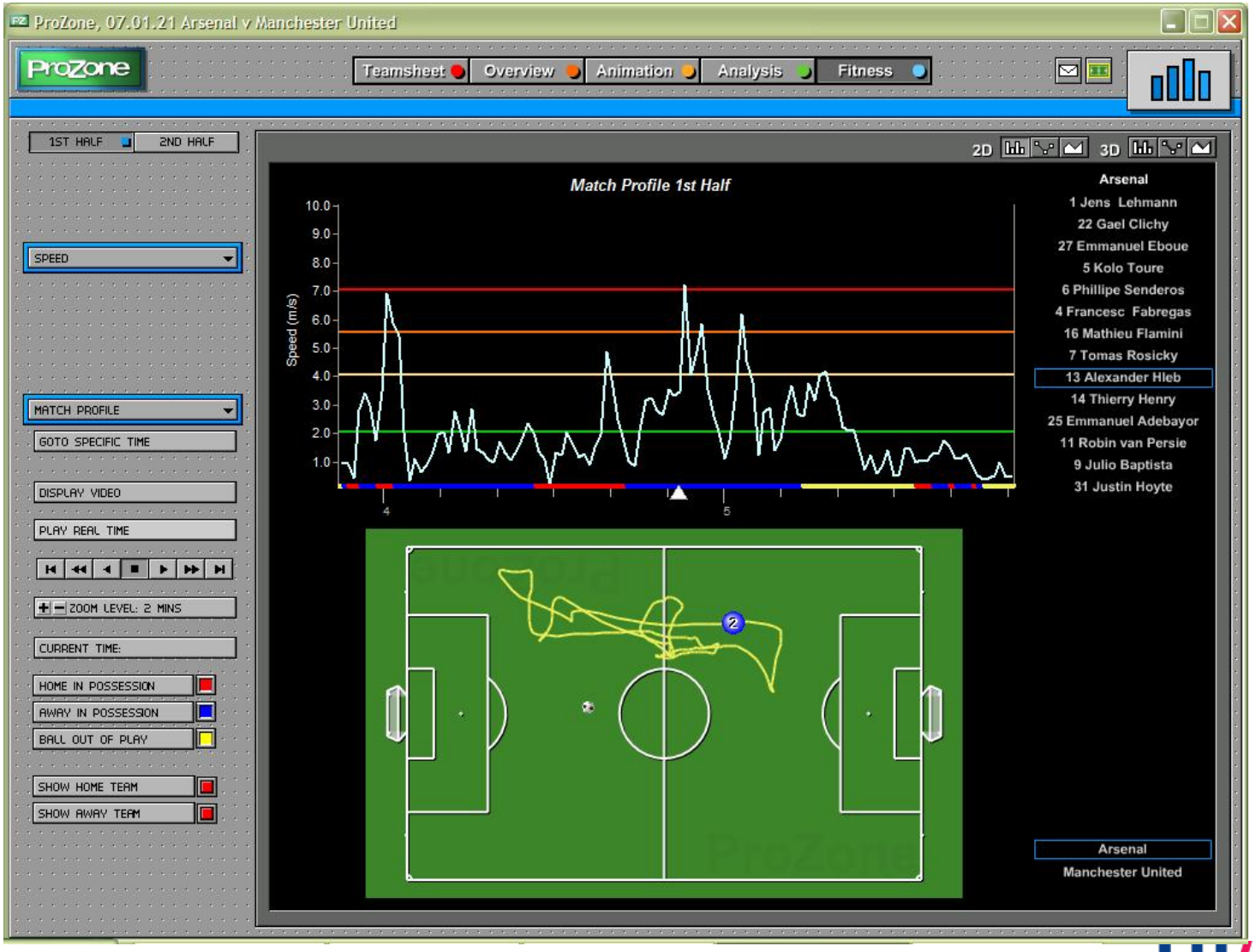


Tracking insects

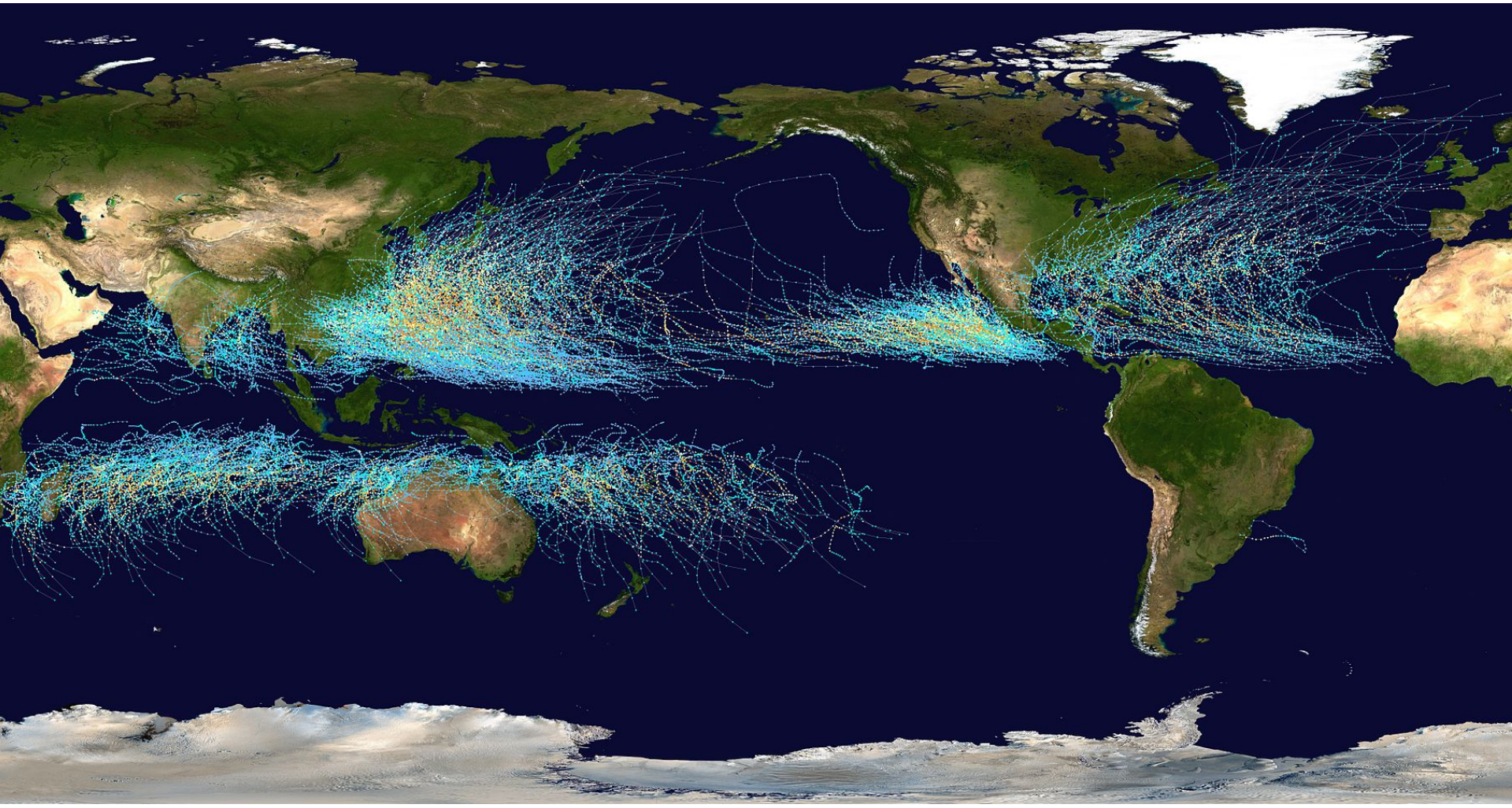


Tracking people

Tracking sports players



Tracking hurricanes



Tracking technology

GPS, RFID, video analysis, …

- Range
- Precision
- Sampling rate

Tracking technology

GPS

- Range
- Precision
- Sampling rate

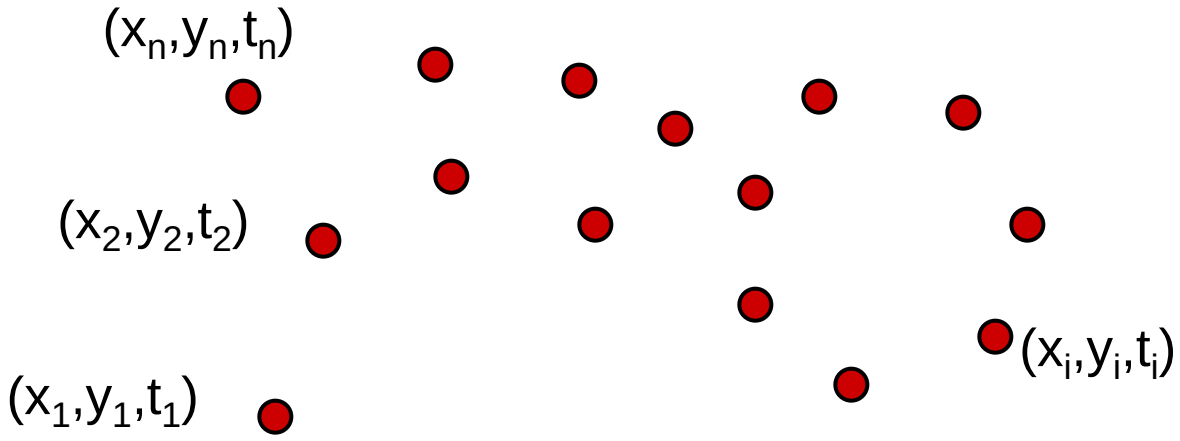
Tracking technology

GPS

- Range: whole world
- Precision: 2-10 meters in *lat-lon*, worse in elevation
- Sampling rate: depends on device, energy source, need not be regular

Trajectory data

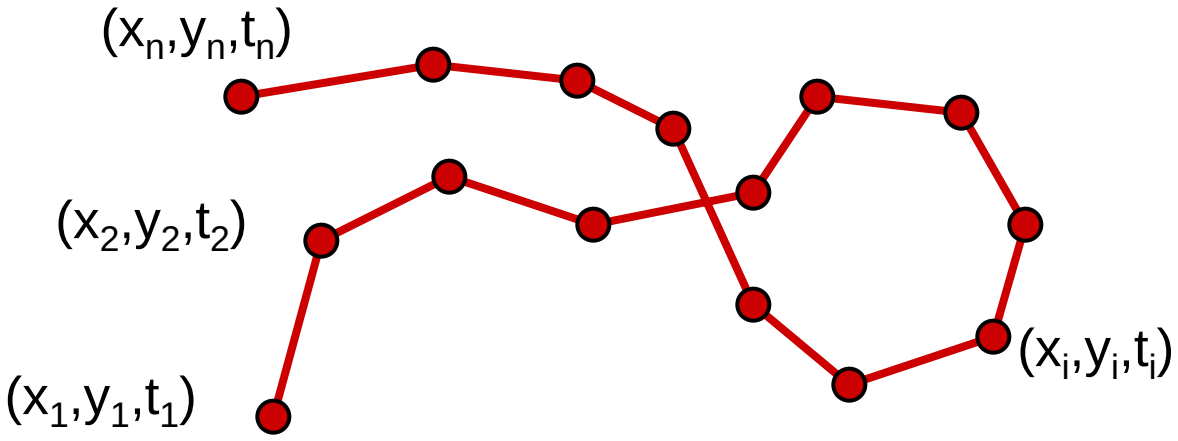
- The data as it is acquired by GPS:
  sequence of triples (spatial plus time-stamp);
  quadruples for trajectories in 3D



Trajectory data

- Typical assumption for sufficiently densely sampled data: constant velocity between consecutive samples

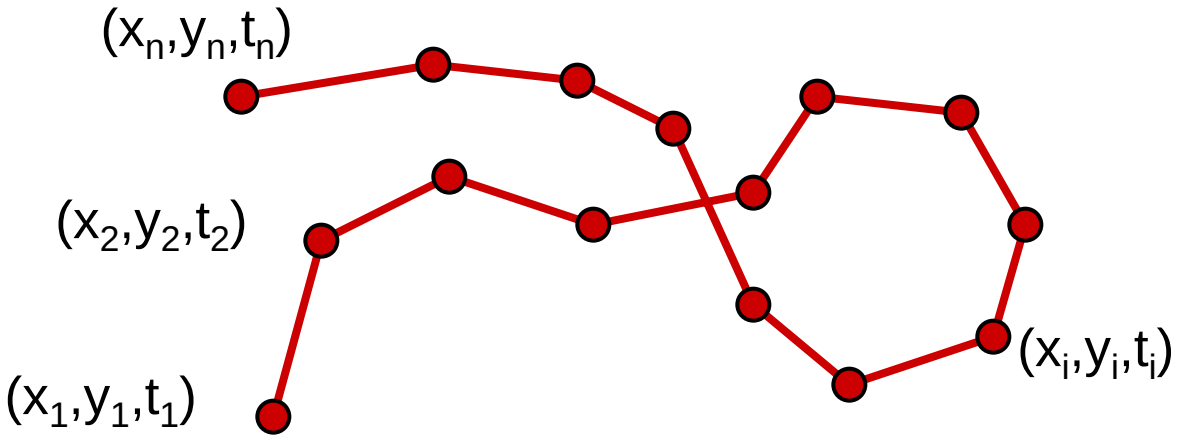
  ➡ velocity/speed is a piecewise constant function



Trajectory data

- Typical assumption for sufficiently densely sampled data: constant velocity between consecutive samples

  What if data is less dense?



Trajectory data

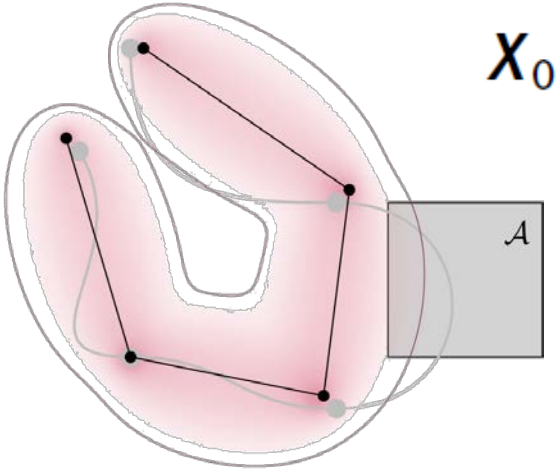
- Typical assumption for sufficiently densely sampled data: constant velocity between consecutive samples

  What if data is less dense?

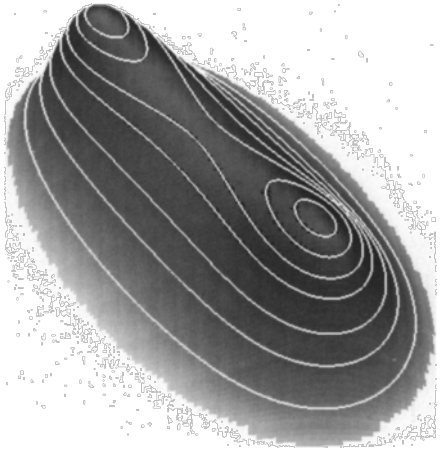
- Other models: Brownian bridges, …

Brownian motion conditioned under starting and ending position

$$\boldsymbol{X}_0 \sim \mathcal{N}\left(\boldsymbol{a}, \delta_a^2\right), \boldsymbol{X}_T \sim \mathcal{N}\left(\boldsymbol{b}, \delta_b^2\right)$$



Probability that object visits region A



Trajectory data analysis

Questions

“How much time does a gull typically spend foraging on a trip from the colony and back?”

Trajectory data analysis

Questions

"How much time does a gull typically spend foraging on a trip from the colony and back?"

"If customers look at book display X, do they more often than average also go to and look at bookshelf Y?"

Trajectory data analysis

Questions

"How much time does a gull typically spend foraging on a trip from the colony and back?"

"If customers look at book display X, do they more often than average also go to and look at bookshelf Y?"

"Is this Alzheimer patient in need of assistance?"

Abstract / general purpose questions

Single trajectory

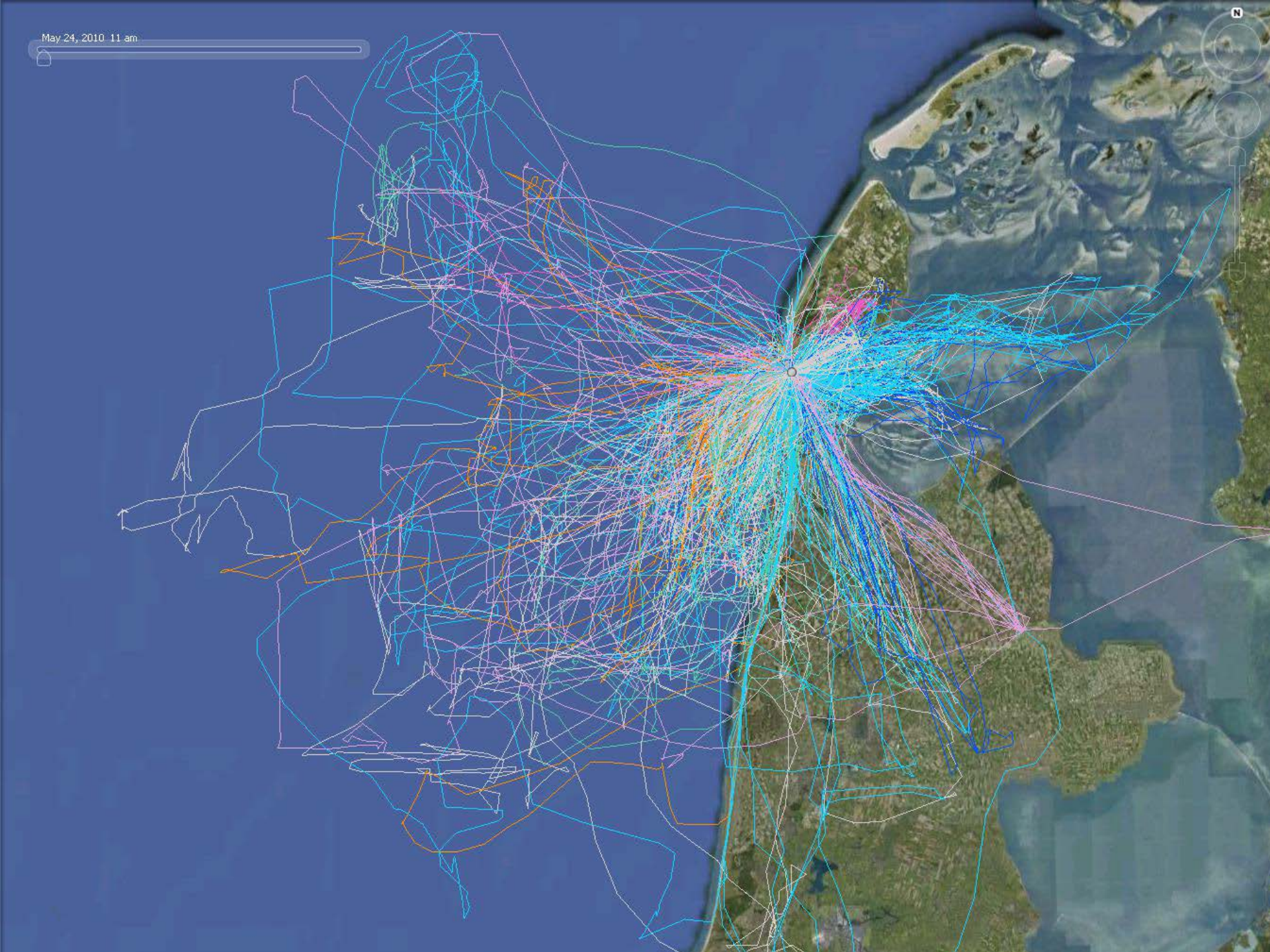
- simplification, cleaning
- segmentation into semantically meaningful parts
- finding recurring patterns (repeated subtrajectories)

Two trajectories

- similarity computation
- subtrajectory similarity

Multiple trajectories

- clustering, outliers
- flocking/grouping pattern detection
- finding a typical trajectory or computing a mean/median trajectory
- visualization


May 24, 2010 11 am

Algorithmic trajectory analysis

Algorithmic research for trajectory analysis

involves (at least) two stages

- proper formalization of the problem
- development of useful, efficient algorithms

Applied algorithmic research for trajectory analysis

also involves

- implementation
- verification
- feedback from domain experts
- revision
- …

usually concerns specific research questions …