Experimental Research in GMT

Scientific Perspectives on GMT 2019/2020

Marc van Kreveld

Experimental research

- A scientific experiment
  - (usually) tests a hypothesis
  - is repeatable
    - fully described (method and parameters) / executable available
    - data is easily available or can be generated

According to some philosophies of science, an experiment can never "prove" a hypothesis, it can only add support.

Experimental research

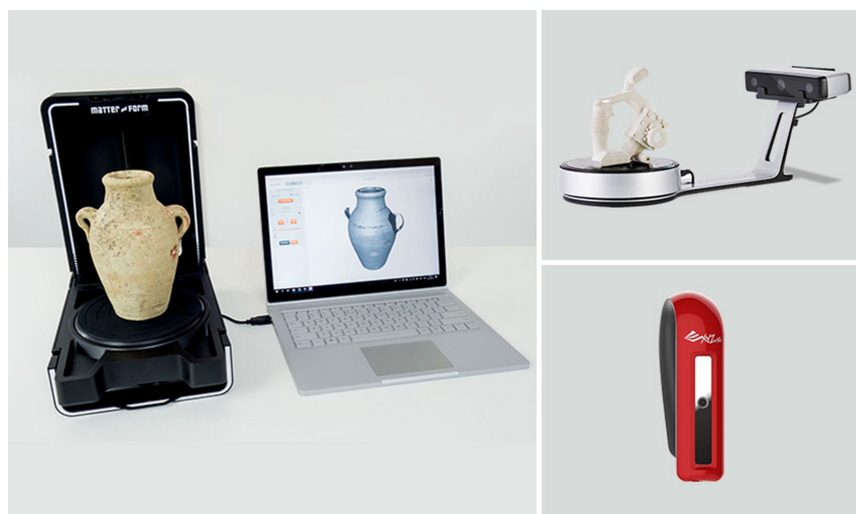
- A scientific experiment
  - (usually) tests a hypothesis
  - is repeatable
    - fully described (method and parameters) / executable available
    - data is easily available or can be generated

According to some philosophies of science, an experiment can never "prove" a hypothesis, it can only add support.

On the other hand, an experiment that provides a counterexample can disprove a theory or hypothesis, but a theory can always be salvaged by appropriate ad hoc modifications at the expense of simplicity.

Experimental research

- Example: method *XYZ* requires O(n^2) time to reconstruct a shape from a 3D point cloud in theory, but in practice an efficiency closer to O(n log n) time is observed
- A paper might demonstrate ("prove") this on data from miniature sculptures with a 3D scanner



Experimental research

- Example: method *XYZ* requires O(n^2) time to reconstruct a shape from a 3D point cloud in theory, but in practice an efficiency closer to O($n \log n$) time is observed
- A paper might demonstrate (“prove”) this on data from miniature sculptures with a 3D scanner
- Another researcher may discover that the practical performance of *XYZ* is much worse on 3D point clouds from aerial laser scanning

Experimental research

- Example: method *XYZ* requires O(n^2) time to reconstruct a shape from a 3D point cloud in theory, but in practice an efficiency closer to O($n \log n$) time is observed
- A paper might demonstrate ("prove") this on data from miniature sculptures with a 3D scanner
- Another researcher may discover that the practical performance of *XYZ* is much worse on 3D point clouds from aerial laser scanning
- The practical performance "theory" on *XYZ* must be reduced in breadth of validity

Experimental research

- Example: method *XYZ* requires $O(n^2)$ time to reconstruct a shape from a 3D point cloud in theory, but in practice an efficiency closer to $O(n \log n)$ time is observed
- If another researcher discovers that the practical performance of *XYZ* is similar on 3D point clouds from aerial laser scanning, the theory is supported but not proved

Experimental Research in GMT: why?

- Algorithmic efficiency (speed in practice)
- Quality of result (optimization)
- Quality of model (sanity check)
- Dependency analysis (of two variables on each other)
- Sensitivity analysis (of one variable or parameter and an artifact)

Algorithmic efficiency experiments

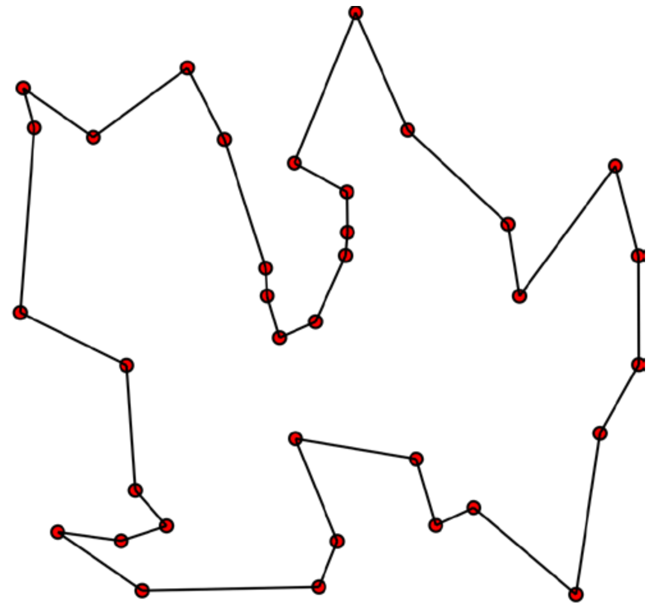
- In algorithms research, worst-case asymptotic efficiency can be proved mathematically
  - efficiency = scaling behavior
  - efficiency ≠ running time

    → suppose we need to know running time …

- Is "worst-case" not too pessimistic?
  - efficiency ≠ scaling behavior in practice

Optimization experiments

- For computational problems that are too hard to optimize:
  - Maximum independent set in a graph
  - Traveling Salesperson Tour of points in the plane
  - ….
- How well do heuristics perform in practice?
- Some approximation algorithms guarantee to be at most a factor 2 off. How far off are they in practice?



Quality-of-model experiments

- How plausible or reasonable is the output?
- Are there unexpected (undesirable) artifacts in the output?
- Analysis method: quantitative and visual inspection
- “Model” could be “measure”
- Example (Thursday): Grid maps

Dependency analysis

- Changing one parameter of the method and analyzing how this influences the output
- Example (Thursday): removing local minima with higher-order Delaunay triangulations

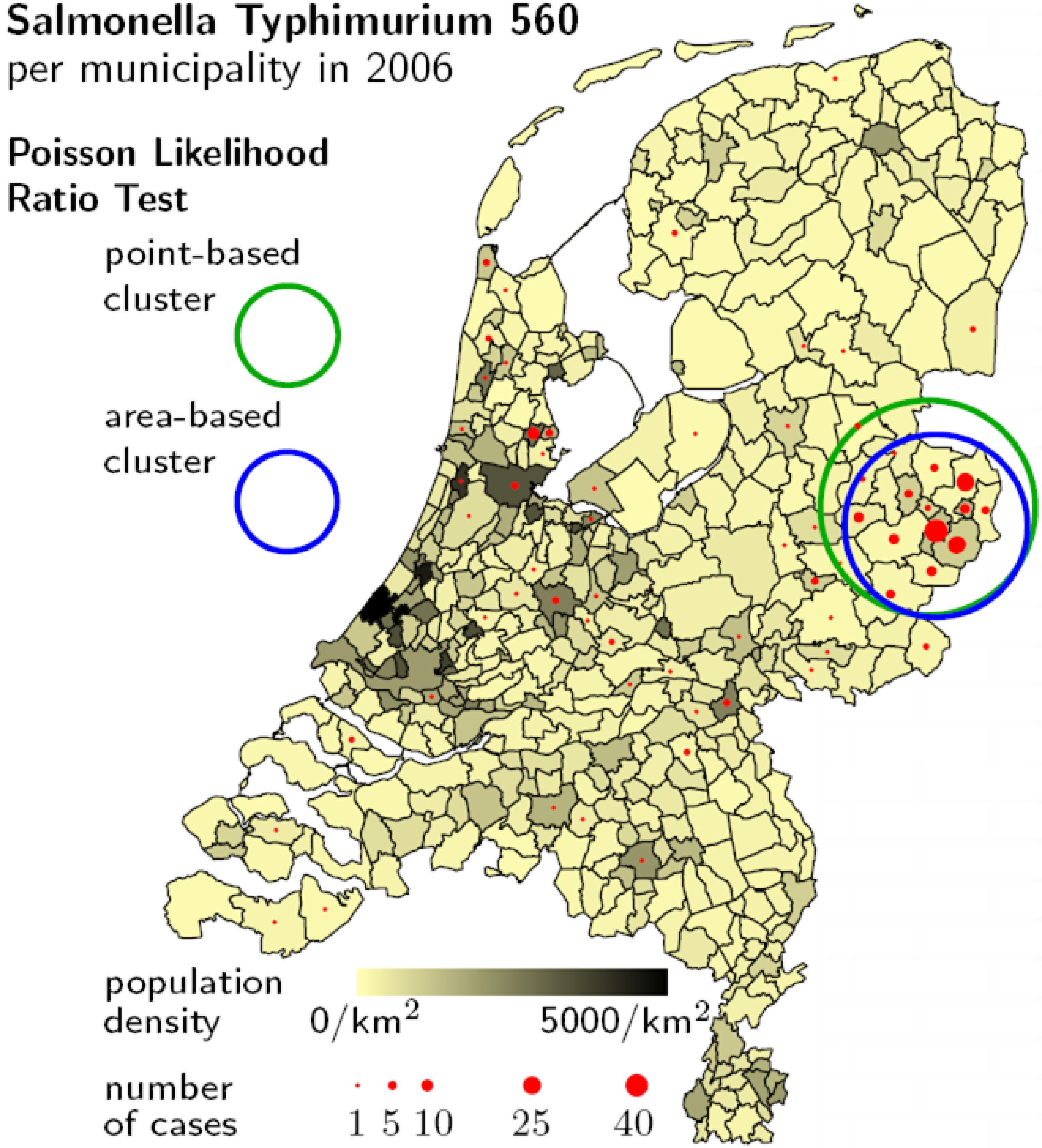
Sensitivity analysis

- Similar objective as dependency analysis, but now a parameter of the input data and not of the method
- Example of dependency: the assignment;
  RANSAC under (synthetic) measurement imprecision

Synthetic data and real-World data

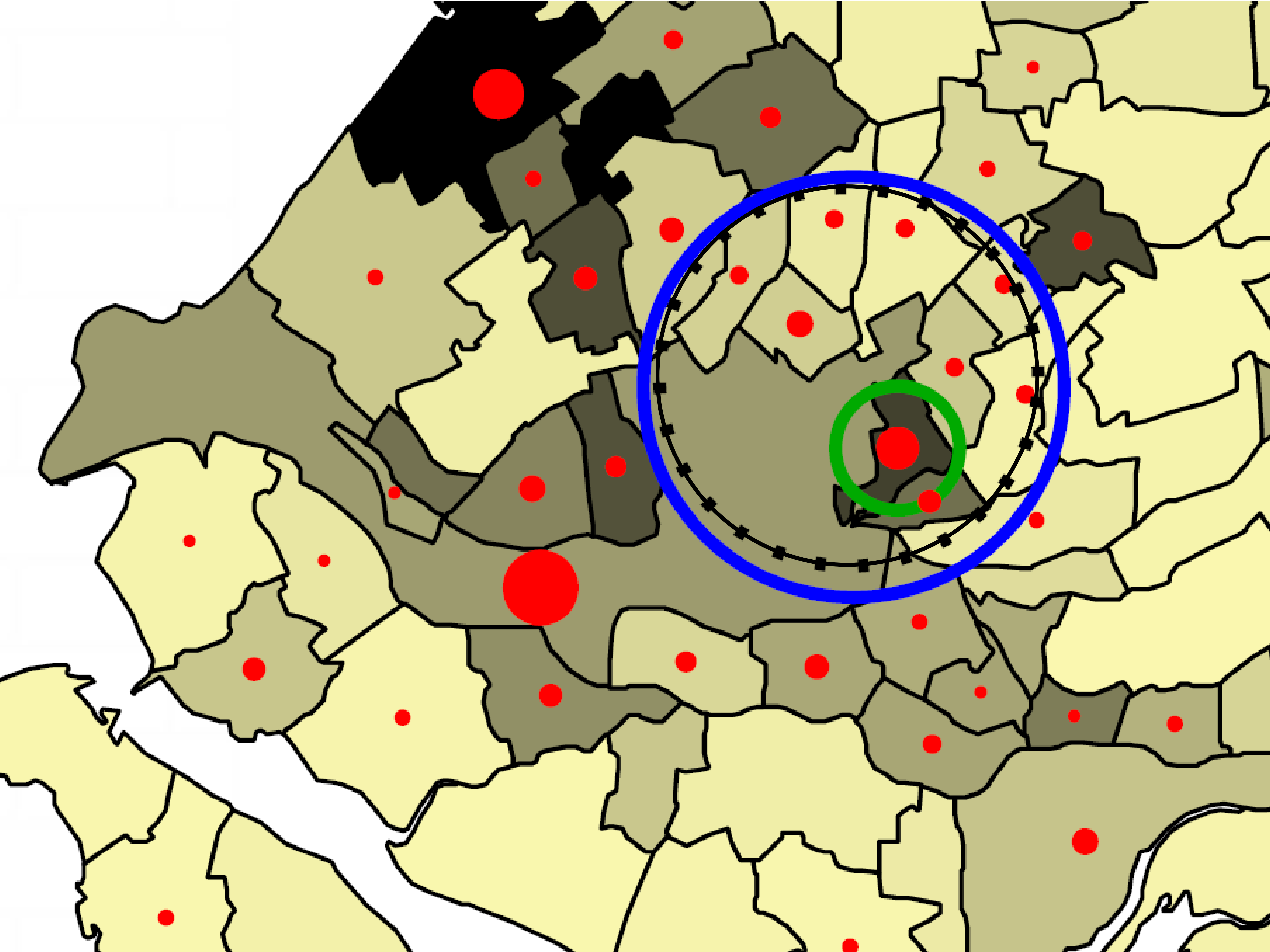
- Synthetic data
  - Comes with a ground truth (real/fundamental truth)
  - Can be generated in large quantities
  - Can have any input size
  - … but might not show what happens on real data
- Real-World data
  - Does not have a ground truth
  - Cannot be controlled well
  - Only few data sets may be available

→ experiments on these types gives complementary results


Salmonella Typhimurium 560
per municipality in 2006
Poisson Likelihood
Ratio Test
point-based
cluster
area-based
cluster
population
density
0/km²
5000/km²
number
of cases
1 5 10 25 40

Mixed real and synthetic test data

- Take real municipality boundaries and population
- Choose a random location for an infected area
  → circle with radius 10 km or 20 km
- All people outside the circle have relative risk 1
- All people inside the circle have relative risk 2.5, 4, or 10
- In this way we generate case counts for all municipalities



Some results

- 10,000 tests with cases, average distance to source

	relative risk 2.5	relative risk 4	relative risk 10
centroid, discrete	4.89	4.59	4.52
centroid, continuous	3.02	2.54	2.42
area, homogeneous	2.64	2.34	2.28
area, non-homogeneous	2.51	1.82	1.76

(in km; cluster radius is 20 km)

Some results

- 10,000 tests with cases, average radius of circle found

	relative risk 2.5	relative risk 4	relative risk 10
centroid, discrete	20.81	21.09	21.81
centroid, continuous	20.37	20.67	21.32
area, homogeneous	21.29	21.66	22.49
area, non-homogeneous	19.94	20.40	21.21

(in km; cluster radius is 20 km)

Other experiments?

- We could have tested other relative risks
- We could have tested other circle radii
- We could have tested other maps and populations

… but, what other things would have been possible?

… and, how should relative risks and circle radii be chosen?

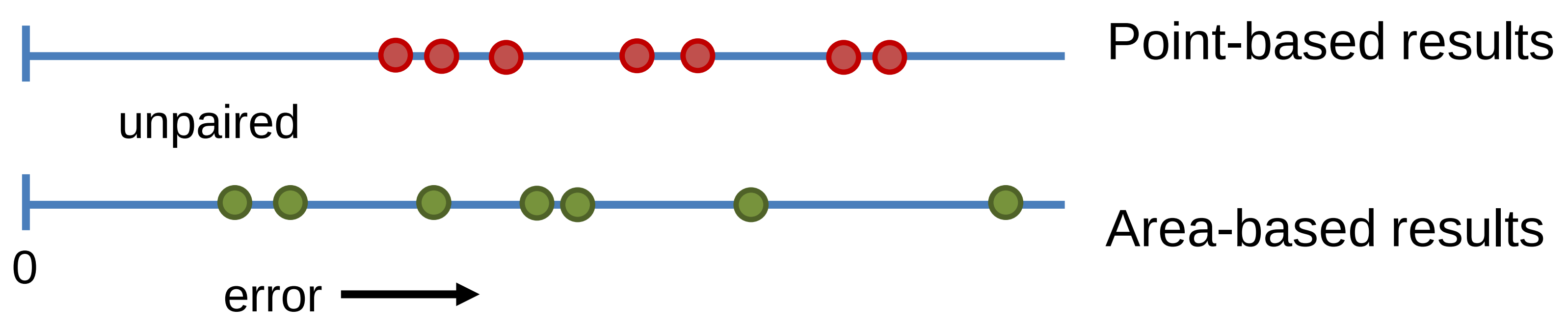
Discuss with your neighbors…

Other experiments?

- We could have done a t-test on the hypothesis that area-based clustering predicts the center of outbreak better than point-based clustering

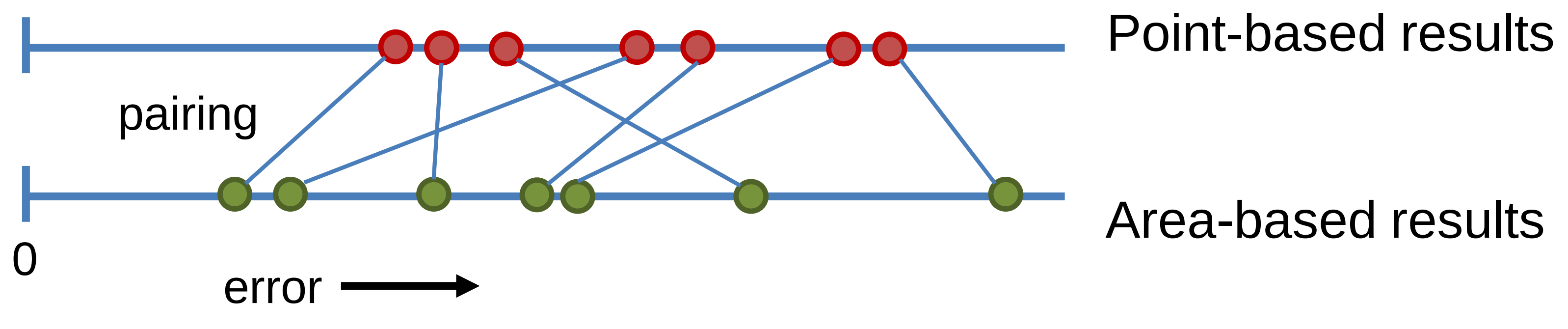
Other experiments?

- We could have done a t-test on the hypothesis that area-based clustering predicts the center of outbreak better than point-based clustering
- Null hypothesis: the two clustering methods predict equally well



Other experiments?

- We could have done a t-test on the hypothesis that area-based clustering predicts the center of outbreak better than point-based clustering
- Null hypothesis: the two clustering methods predict equally well



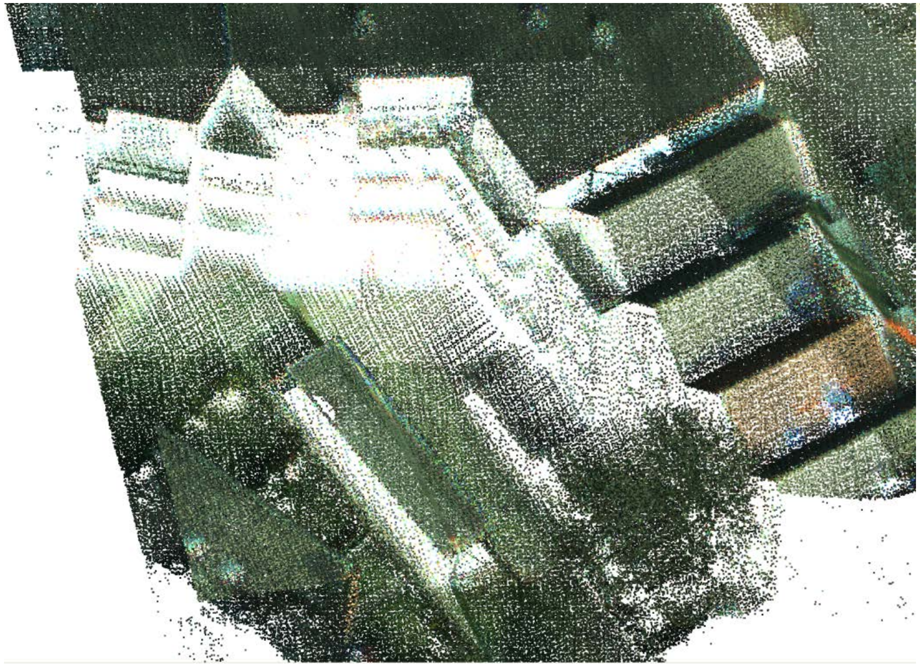
Benchmark data

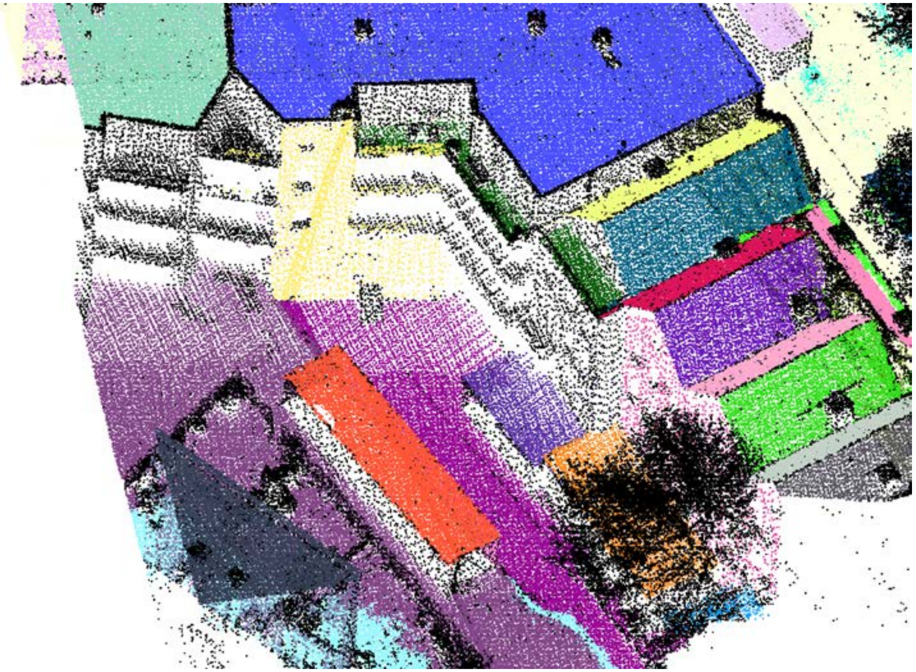
- To allow comparison with other approaches without having to run or implement them
- To have good data readily available
- Examples:
  - Stanford Large Network Dataset Collection
    - Social networks
    - Web networks
    - …
  - Stanford GraphBase
  - Princeton Shape Benchmark: 3D polygonal models
  - Movebank: Animal tracking data

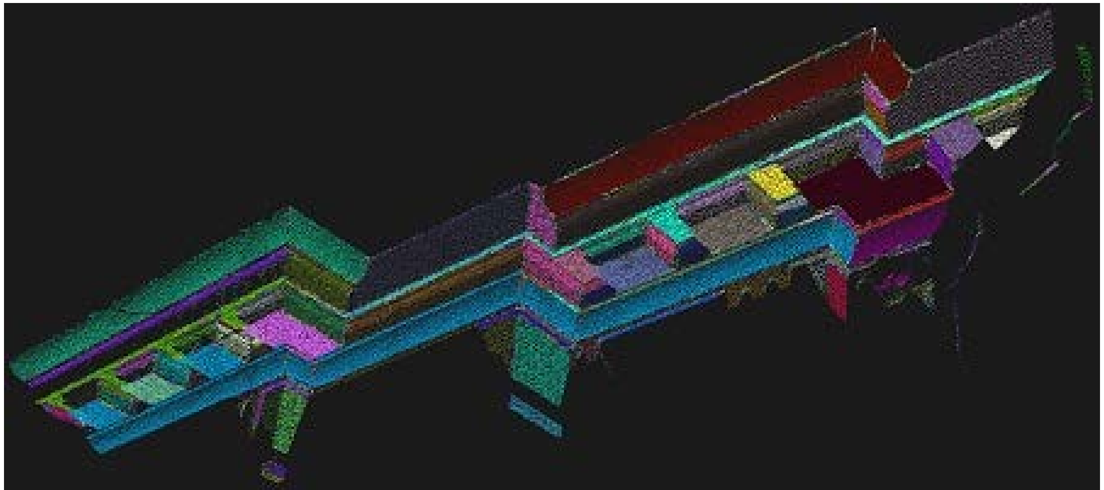
Matching parts of a point set to a model

- Two popular methods: RANSAC and the Hough transform
- Used for 3D reconstruction, in particular, to find many points close to a plane in 3D, yielding facets of buildings

Points clustered by planes

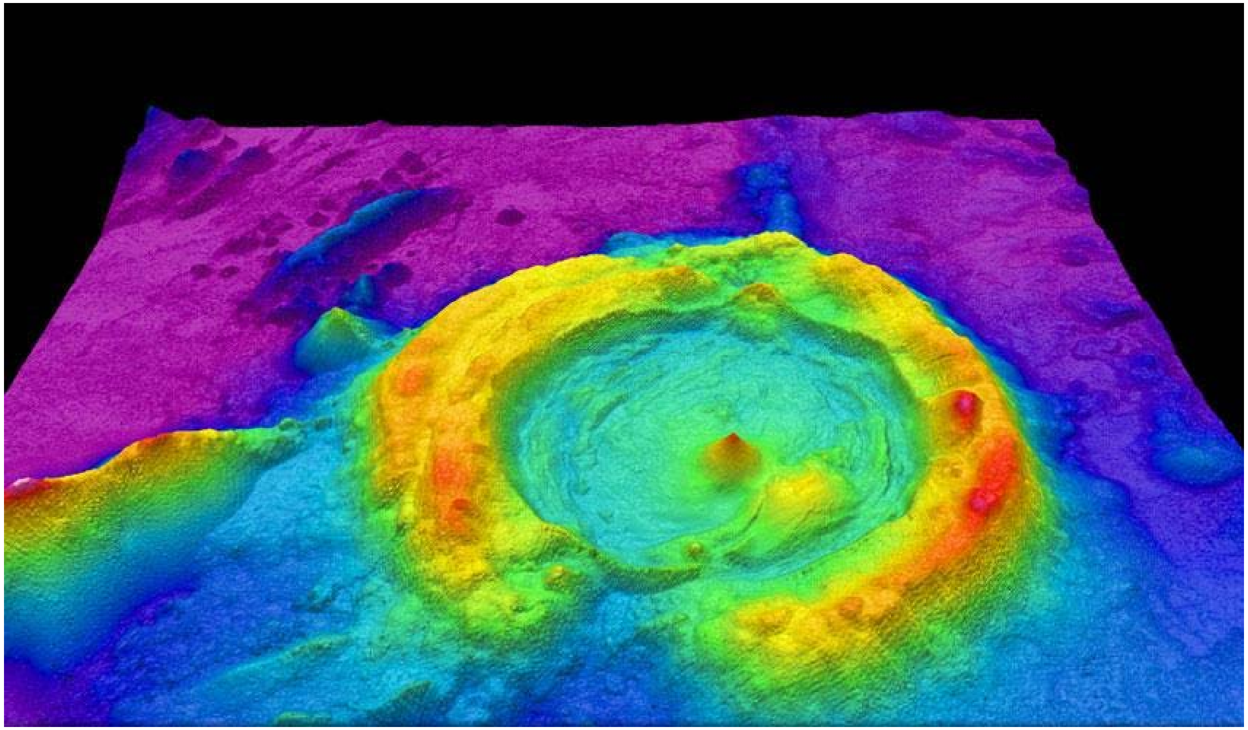






LiDAR point cloud data

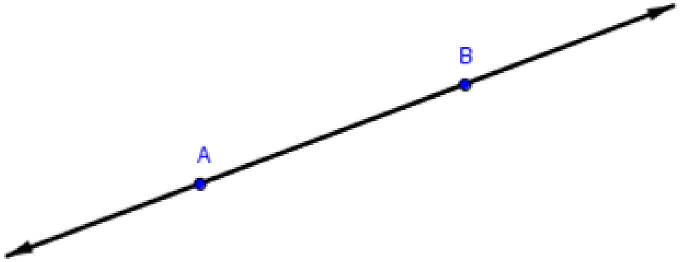
- **Li**ght **D**etection **A**nd **R**anging
- Yields 3D point cloud using time to return of a laser ray

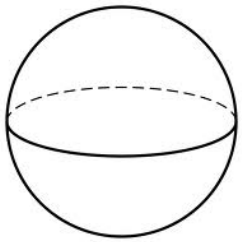


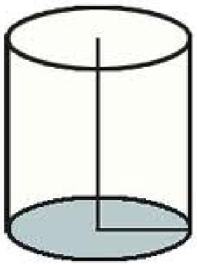
LiDAR data where points are colored by their third coordinate

RANSAC

- RANdom SAmple Consensus: method that can be used to detect planes (and other shapes) in point sets
  - randomized
  - assumes a model defined by few points
    - line (2D): defined by 2 points
    - plane: defined by 3 points
    - sphere: defined by 4 points
    - vertical cylinder: defined by 3 points

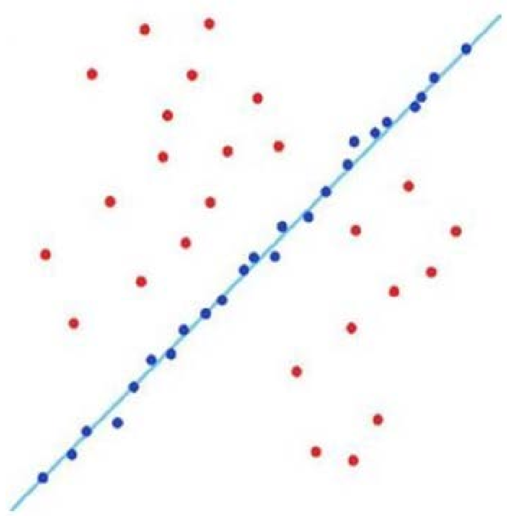






RANSAC

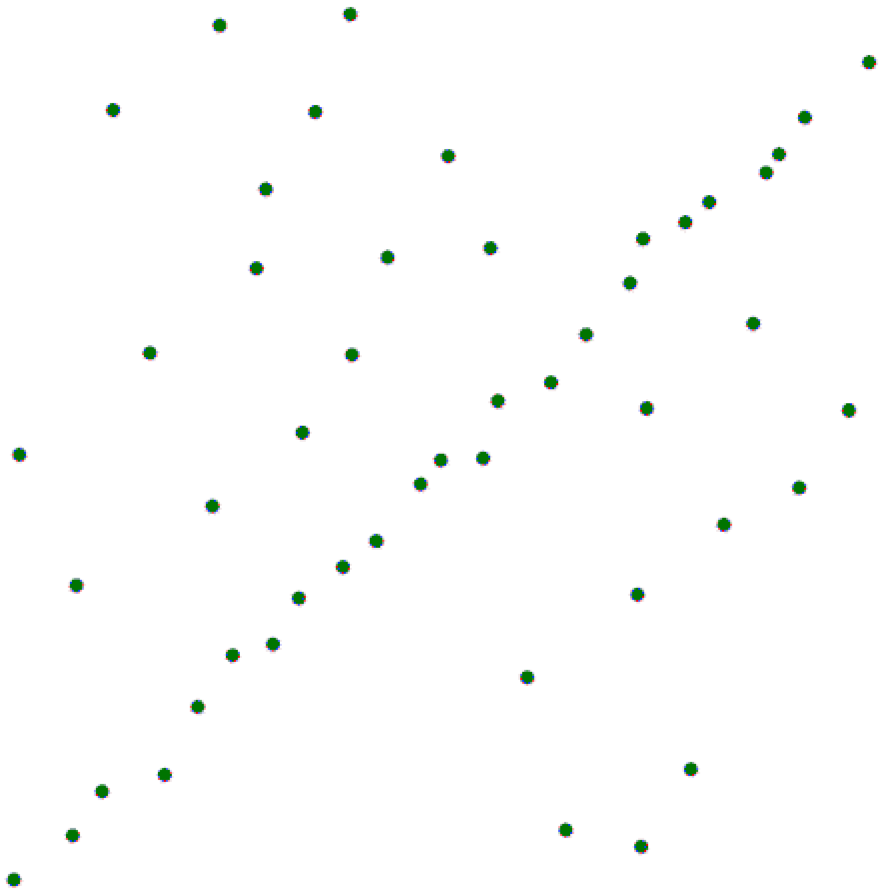
- Simplest case: 2D point set, we want to find a line with most points on (or near to) it
  - points on/near this line are called inliers, they support the line
  - other points are outliers, they do not support the line



RANSAC

1. Choose a threshold distance *d*
2. For #iterations do
   - Choose 2 points, make line *L*
   - For each point *q* in *P*, test if *q* lies within distance *d* from *L*
     If yes, increase the support of *L* by 1
   - If *L* has higher support than the highest-support line found so far, remember *L* and its support
3. Return *L* as the line with most points near it

RANSAC



RANSAC

- For testing whether a point q supports a line L, we do not actually compute the distance from q to L
- Instead, we generate two lines at distance d from L
- Then we test for each point whether it lies below the upper and above the lower line

RANSAC

- How large should the threshold distance *d* be?
- How many iterations should we do to have a high probability of finding the line with highest support?

→ the threshold distance is related to the measurement error (~5 cm) and the flatness of the surface

→ the number of iterations depends on the inlier-outlier ratio and with how much probability we want to find the best line

RANSAC, iterations

- Suppose we want to have 95% probability, p=0.95, of finding the best line
- Suppose there are k points on the line (inliers) and n points in total
- Then the probability of choosing 2 points on the line is $(k/n)^2$
- The probability of never selecting 2 points on the line in r iterations is $(1 – (k/n)^2)^r$
- The probability of finding the line in r iterations is $1 – (1 – (k/n)^2)^r$

RANSAC, iterations

- So we want $1 - (1 - (k/n)^2)^r > p$

$$
\begin{aligned}
(1 - (k/n)^2)^r &< 1 - p \\
\log (1 - (k/n)^2)^r &< \log (1 - p) \\
r \log (1 - (k/n)^2) &< \log (1 - p)
\end{aligned}
$$

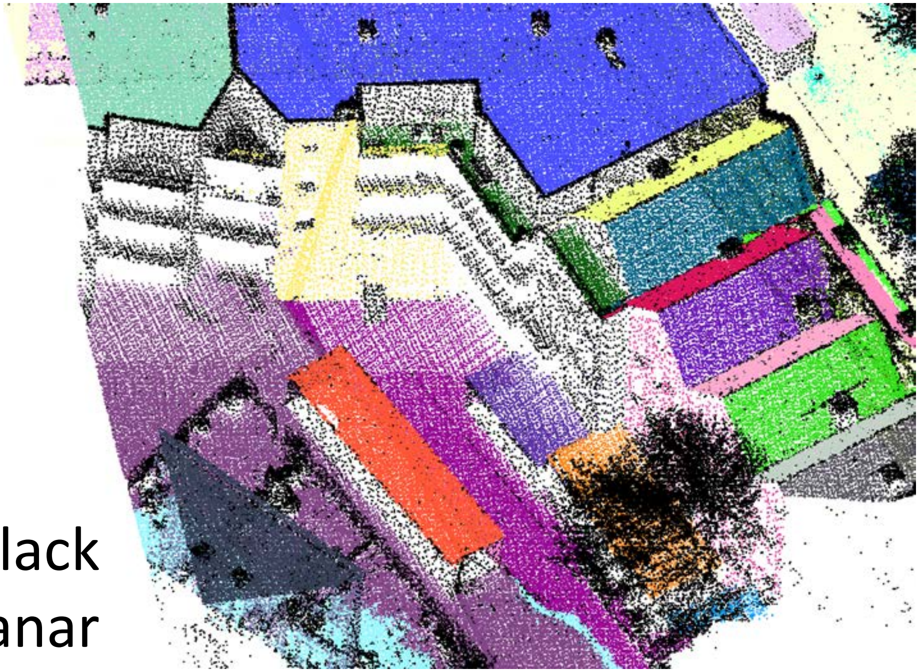
$$r > \log (1 - p) / \log (1 - (k/n)^2)$$

RANSAC, iterations

- Examples:
  - if 10% of the points lie on the line and we want to find it with 95% certainty, we need nearly 300 iterations
  - if 5% of the points lie on the line and we want to find it with 95% certainty, we need nearly 1200 iterations
  - if 10% of the points lie on the line and we want to find it with 90% certainty, we need nearly 230 iterations

Iterated RANSAC

- After finding the plane with the most points, remove the points from the set and remember them as a cluster
- Then continue and find more planes, until no plane seems to have sufficient support



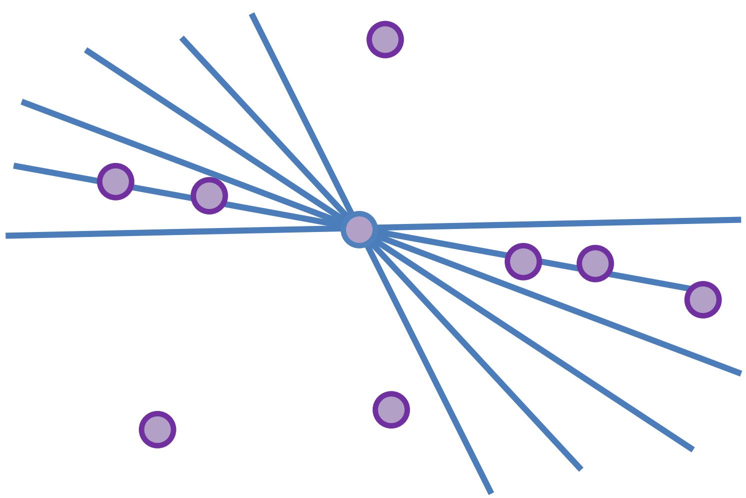
Points not in clusters are black
Why are the outlines of planar regions black?

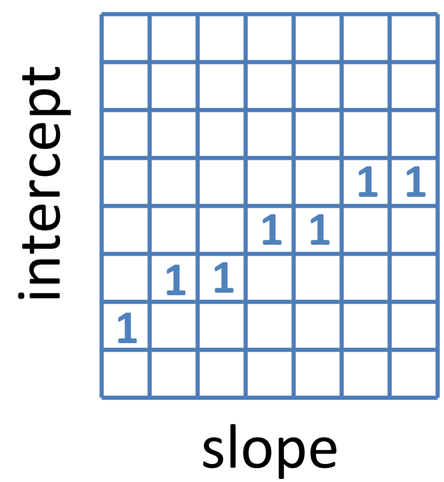
Iterated RANSAC

- The remaining points are
  - vegetation
  - curved surfaces (cars, domes)
  - traffic signs, lamp posts, mailboxes, garbage bins, bicycles, drainage pipes, …
  - points on planes whose normal was incorrect (possibly, close to corners)
  - points on very small or largely occluded planes
  - points inside buildings measured through windows
- These points may still help for reconstruction

Hough transform

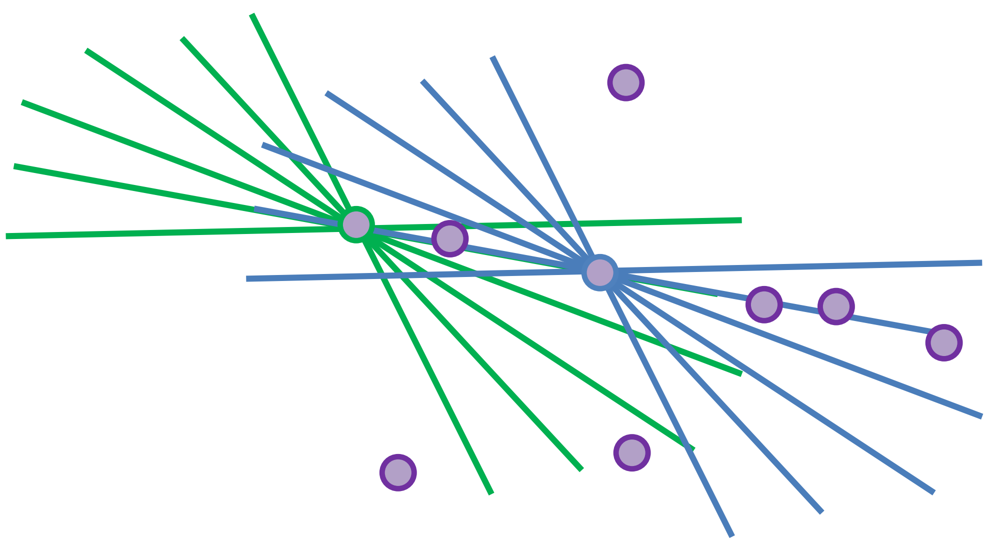
- The Hough transform is an alternative to RANSAC and can also give a plane close to many points
- It discretizes the set of all lines by a grid; points give a count to all grid cells whose lines come near that point

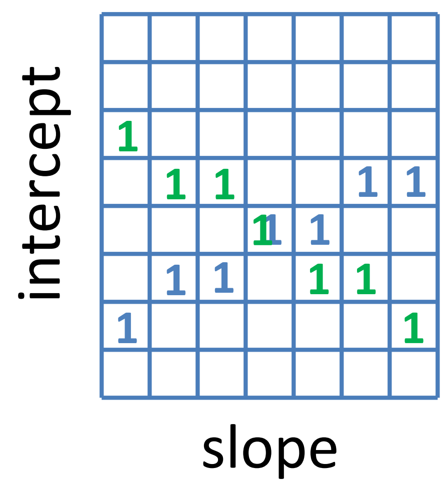




Hough transform

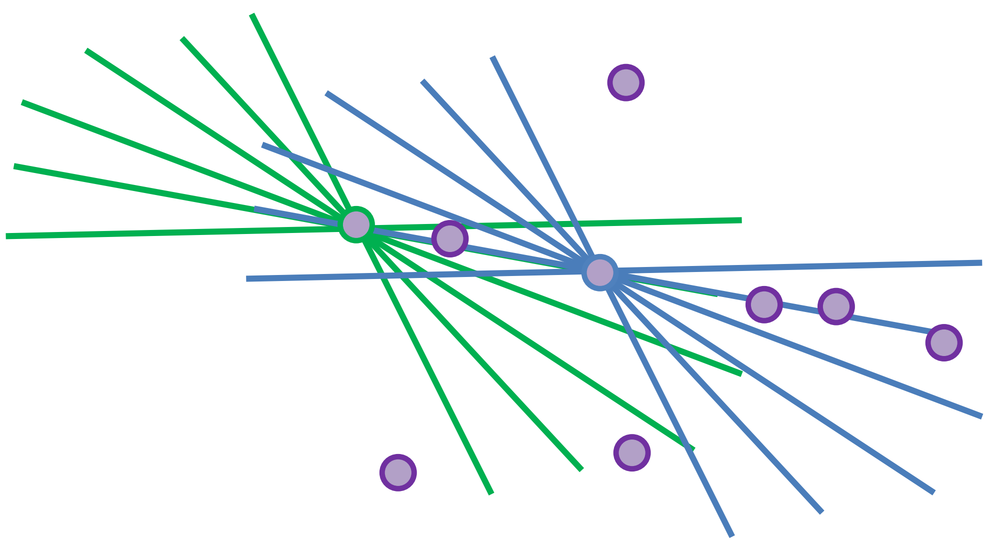
- The Hough transform is an alternative to RANSAC and can also give a plane close to many points
- It discretizes the set of all lines by a grid; points give a count to all grid cells whose lines come near that point

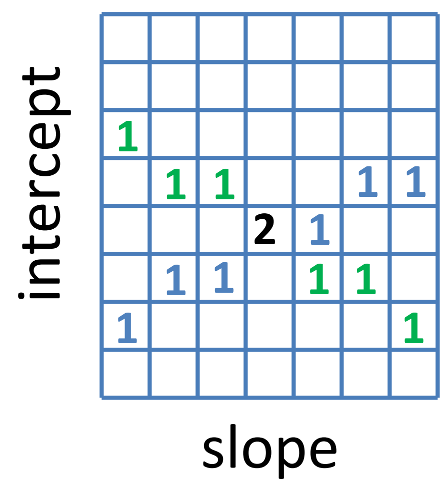




Hough transform

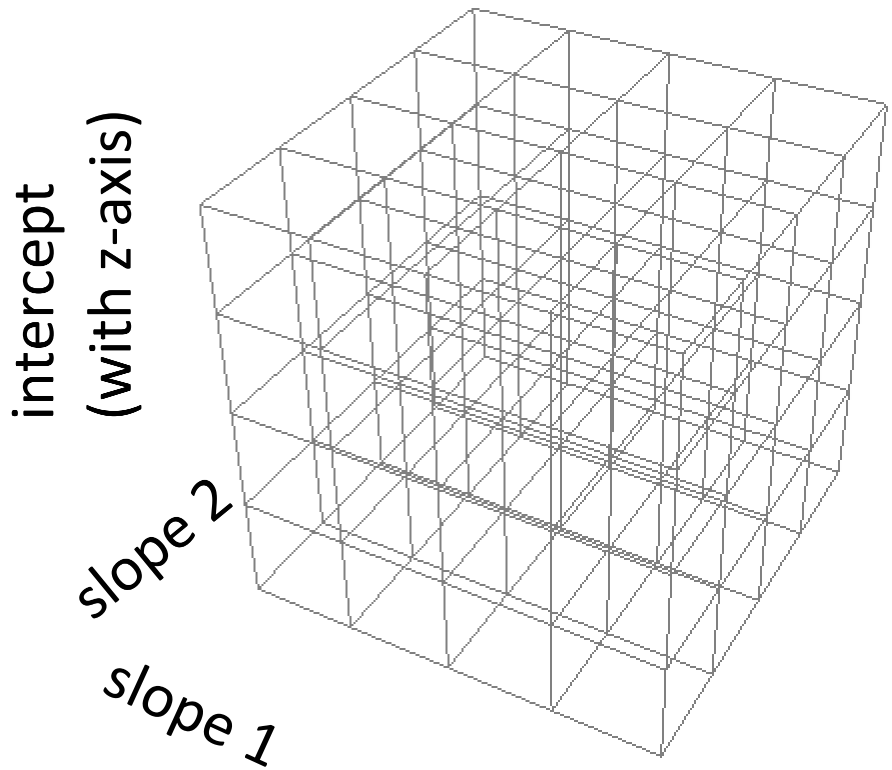
- The middle of the cell with the highest number gives slope and intercept of a line with many points close by





Hough transform

- In 3D, the grid for representing all planes is 3-dimensional (and may take up a lot of storage space)



Scientific misconduct

- Very easy in experimental research
- Possibly accidental!
- Read the Wikipedia page on the topic

Scientific misconduct

- For example: you propose (and want to publish) a new method *M* for task *X*
  - Leaving out datasets where *M* does not perform well
  - Not extensively searching for other methods that solve task *X*
  - Knowingly fail to mention method *N* if that method works better for task *X*
  - Using different environments when running *M* and *N* when comparing them
  - Postprocessing the output to make method *M* look better
  - Not describing *M* fully correctly in the paper
  - …

Scientific misconduct?

- You propose (and want to publish) a new method *M* for task *X*
- Your method's performance depends on a parameter *q*
- You compare existing method *N* with your method *M* using 10 different parameter settings for *q*
- You conclude that your method is better because in 93% of the cases method *M* is better than method *N* for some parameter setting

Summary

- Experimental research is done to learn about models or algorithmic solutions in practice
- It is possibly the most important type of research in GMT
- It is important to think about honesty and avoid scientific misconduct