acm

ACM Europe
Summer School on
Data Science,
Athens, Jul. 17, 2018

Learning from our Movements – Mobility Data Analytics

Yannis Theodoridis

Data Science Lab. @ University of Piraeus
ytheod@unipi.gr; www.datastories.org

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ

UNIVERSITY OF PIRAEUS
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

DATA STORIES

Outline

1. **Getting to know your data**
   - Nature of mobility data; sources; applications; similarity measures
2. **Pre-processing your data**
   - Data curation (cleansing, simplification, enrichment, etc,)
   - Data storage (and querying)
3. **Analyzing your data**
   - Cluster analysis (group behaviour) and outlier detection
   - Frequent pattern (path, location) discovery
   - Classification and Prediction
4. **Summary – the Future**
   - A real-world use case; What’s next

Sources of material used

Slides mainly based on:

N. Pelekis & Y. Theodoridis (2014) Mobility Data Management and Exploration. Springer.
URL: infolab.cs.unipi.gr/MDMEbook

Other sources used:

- Slides from EU H2020 DATACRON project
- Slides from EU H2020 DART project
- Slides from EU H2020 Track&Know project

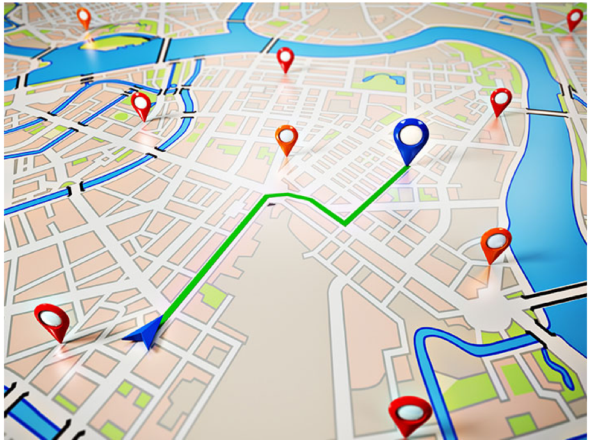
Part I:
Getting to know your data

“Τά πάντα ρεί, μηδέποτε κατά τ’ αυτό μένειν –
Everything changes, nothing remains still.”
Heraclitus

Mobile devices and services

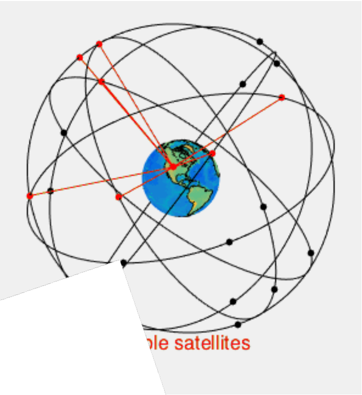
- Large diffusion of mobile devices and related services and apps
  ➔ **mobility-aware data**

- Mobility-aware data are generated by
  - … mobile phones (e.g. cell positions in the GSM network)
  - … GPS devices (e.g. humans' smartphone)
  - … RFIDs, Wi-Fi access points, Bluetooth sensors, etc.

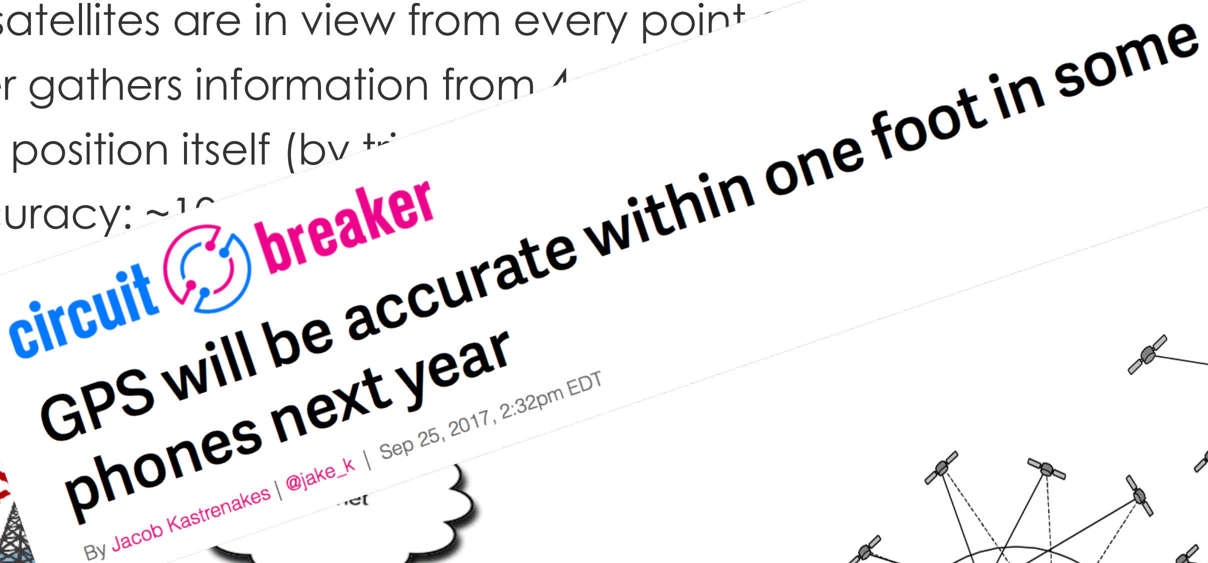


- In this course, we focus on **GPS data**

Geo-positioning



- **GPS (Global Positioning System)**
  - 24-satellite constellation around globe
    - At least 5 satellites are in view from every point
  - GPS receiver gathers information from 4
    - in order to position itself (by tr
  - Position accuracy: ~1

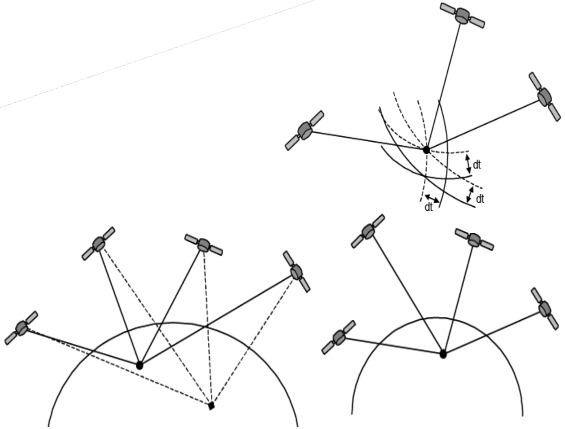


circuit breaker

GPS will be accurate within one foot in some phones next year

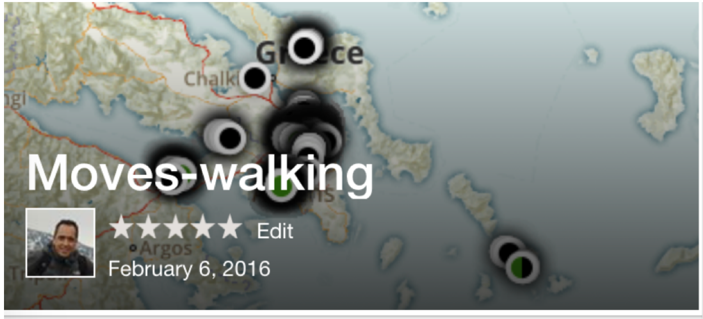
By Jacob Kastrenakes | @jake_k | Sep 25, 2017, 2:32pm EDT

GPS/online phone
GSM/cell network



GPS data – an example

AllTrails

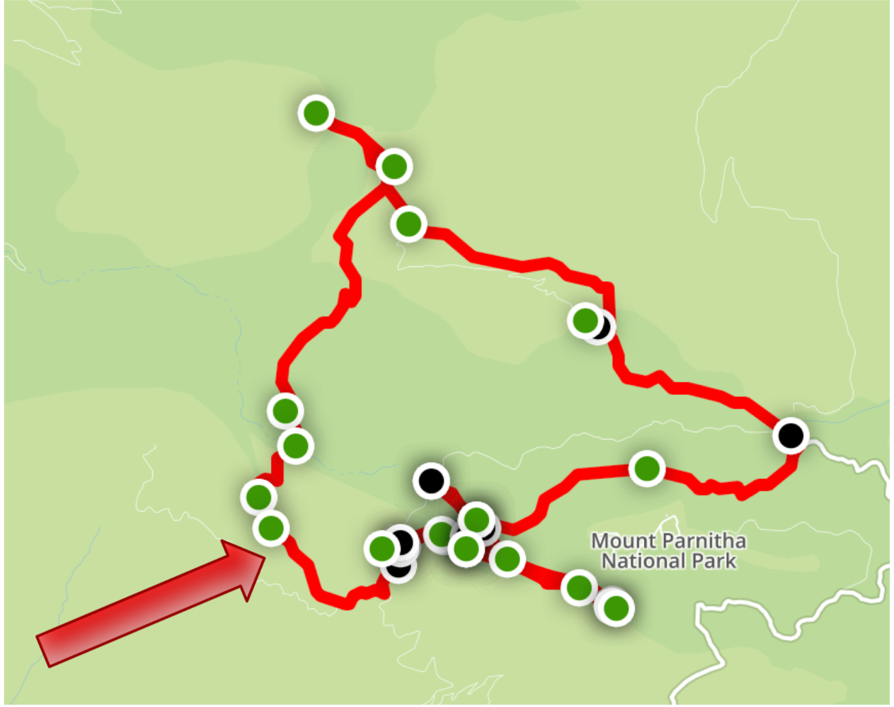


Moves-walking

★★★★★ Edit

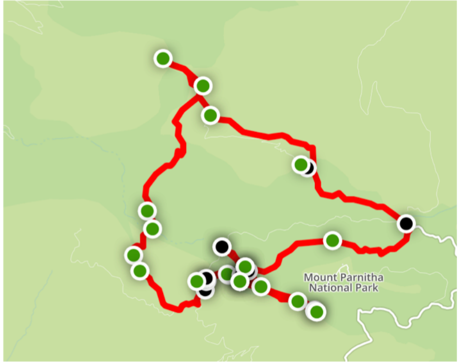
February 6, 2016

```
<trkpt lat="38.16685" lon="23.72597">
        <ele>1132.17</ele>
        <time>2015-10-
02T08:08:29Z</time>
      </trkpt>
```



GPS data – an example (cont.)

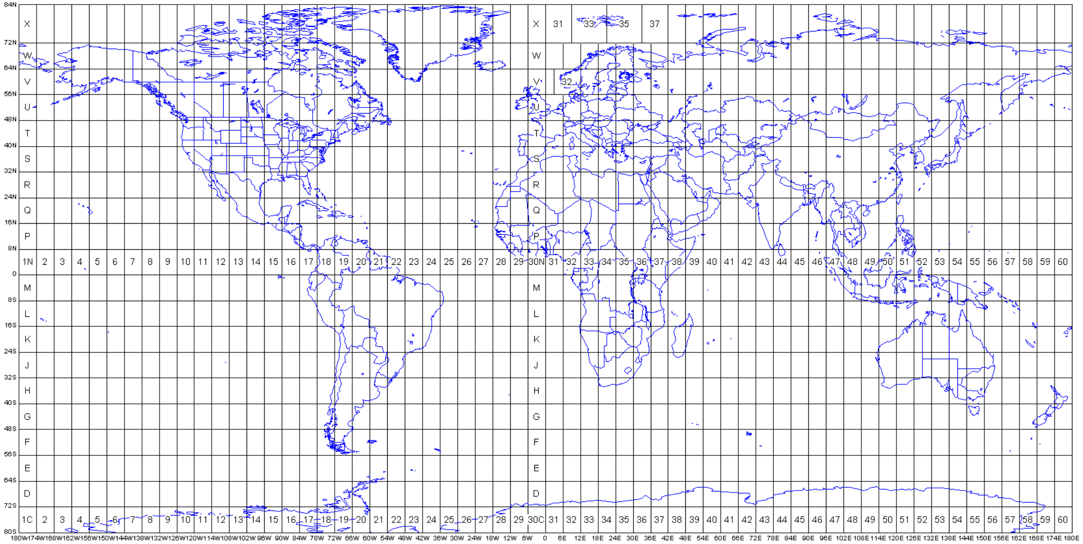
- Raw data:
  .gpx format



```
<trk>
…
<trkpt lat="38.16685" lon="23.72597”>
        <ele>1132.17</ele>
        <time>2010-10-02T08:08:29Z</time>    </trkpt>
<trkpt lat="38.16682" lon="23.72601”>
        <ele>1131.98</ele>
        <time>2010-10-02T08:08:34Z</time>    </trkpt>
<trkpt lat="38.16678" lon="23.7261”>
        <ele>1130.6</ele>
        <time>2010-10-02T08:08:58Z</time>    </trkpt>
…
</trk>
```

From spherical (WGS84) to plane coordinates

- **Universal Transverse Mercator** (UTM): a type of cylindrical projection
  - Internationally standard coordinate system
  - 60 zones (6 degrees of lon, each); 20 cells per zone (9 degrees of lat, each)
- A UTM geo-reference consists of a zone cell, a 6-digits easting and a 7-digits northing
  - Eastings and Northings are in meters
  - e.g. Athens: (34S; 739,545.42; 4,207,529.27)



84N
72N
64N
56N
48N
40N
32N
24N
16N
8N
0
8S
16S
24S
32S
40S
48S
56S
64S
72S
80S
X
W
V
U
T
S
R
Q
P
M
L
K
J
H
G
F
E
D
1N
1C
30N
30C
31
33
35
37
32
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
180W 174W 168W 162W 156W 150W 144W 138W 132W 126W 120W 114W 108W 102W 96W 90W 84W 78W 72W 66W 60W 54W 48W 42W 36W 30W 24W 18W 12W 6W 0 6E 12E 18E 24E 30E 36E 42E 48E 54E 60E 66E 72E 78E 84E 90E 96E 102E 108E 114E 120E 126E 132E 138E 144E 150E 156E 162E 168E 174E 180E

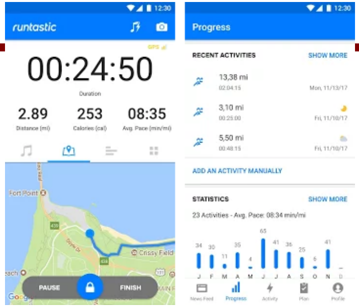
image source: http://www.dmap.co.uk

Location- and mobility-aware apps

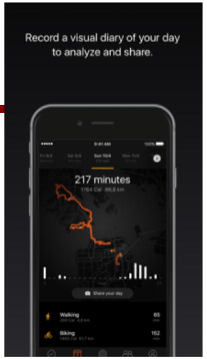
- **Navigation** (vehicle or pedestrian) & **Location-aware information**
  - Routing (walking, driving, eco-friendly, …)
  - Search around for nearby points-of-interest (POI)
- **Resource management** & **Tracking**
  - Fleet (taxis, trucks, vessels, planes, etc.) management
  - Tracing of a stolen car, locating persons in an emergency situation
- **Fitness apps** and **Location-aware social networking**
  - Runtastic, Runkeeper, Human, Moves, etc.
  - Google Maps Location Sharing, Facebook Nearby Friends, Tinder Places, etc.

Commercial examples

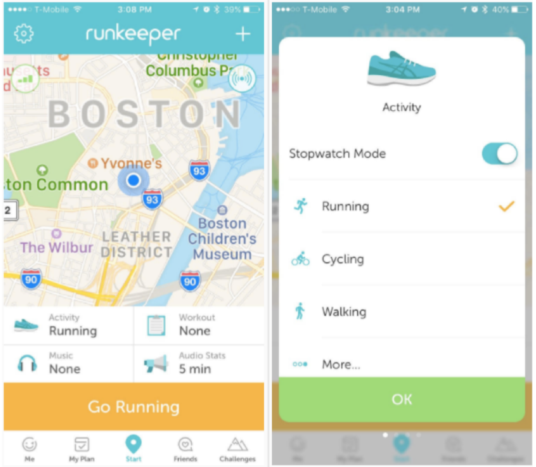
- **Track your activity** (walking, running, cycling, hiking, …)



Runtastic



Human



Activity: Running

Workout: None

Music: None

Audio Stats: 5 min

Go Running

Stopwatch Mode

Running ✓

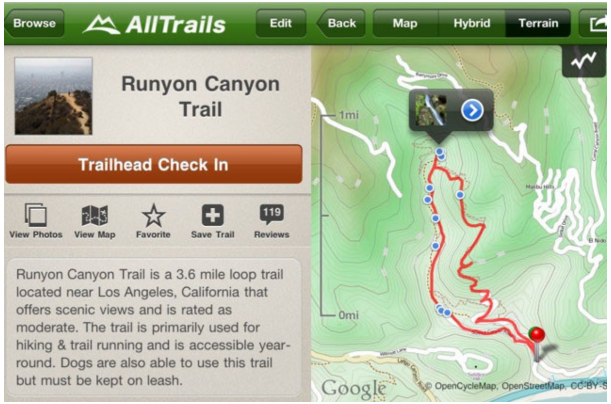
Cycling

Walking

More...

OK

RunKeeper



Runyon Canyon Trail is a 3.6 mile loop trail located near Los Angeles, California that offers scenic views and is rated as moderate. The trail is primarily used for hiking & trail running and is accessible year-round. Dogs are also able to use this trail but must be kept on leash.

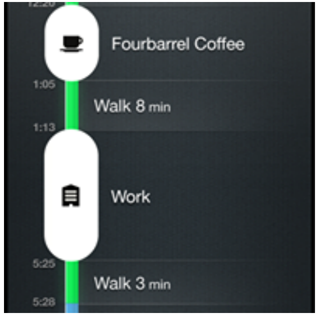
AllTrails

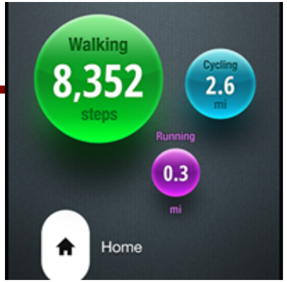
The most advanced application for cyclists on the App Store.
Tracks cycling, running, hikes, and more!
56:54
16.36
22.55
929
Stop
Lap

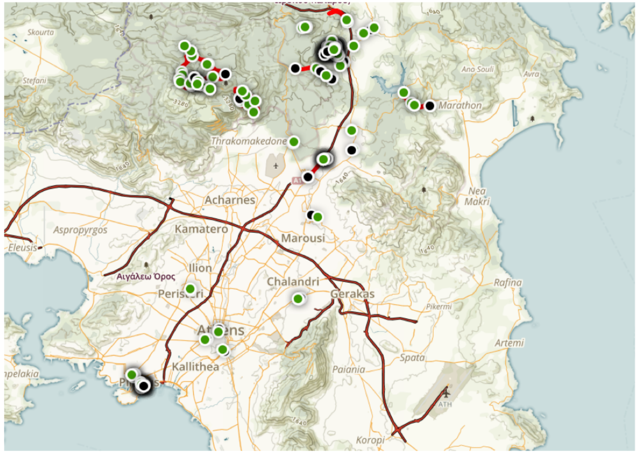
Cyclemeter

Commercial examples (cont.)

- A special case: **Moves** (moves-app.com)
- Activity inference !!
  - movement type,
  - home/work places, etc.







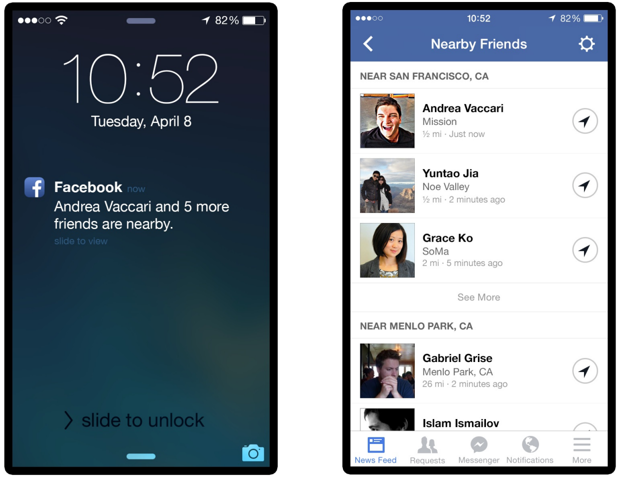
How many hot spots do you see?
(green are Start points; black are End points)

What do you infer about Yannis?

Commercial examples (cont.)

- **Social networking apps** - See in real time where your friends are
  - Google Maps Location Sharing,
  - Facebook Nearby Friends,
  - Tinder Places
  - etc.

Tinder



Facebook now

Andrea Vaccari and 5 more friends are nearby.

slide to view

NEAR SAN FRANCISCO, CA

Andrea Vaccari
Mission
½ mi · Just now

Yuntao Jia
Noe Valley
½ mi · 2 minutes ago

Grace Ko
SoMa
2 mi · 5 minutes ago

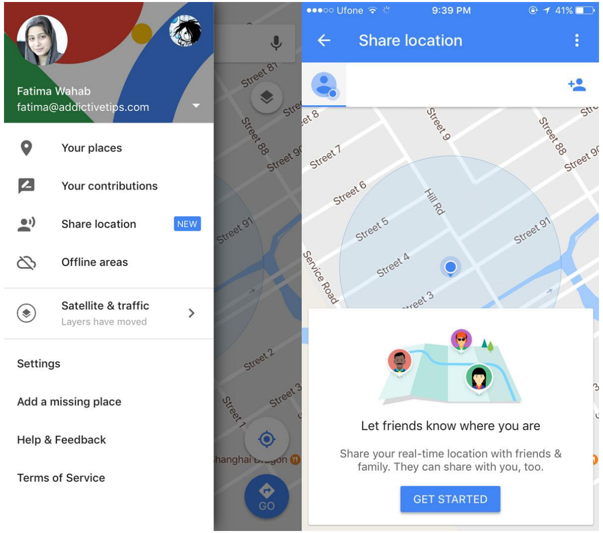
See More

NEAR MENLO PARK, CA

Gabriel Grise
Menlo Park, CA
26 mi · 2 minutes ago

Islam Ismailov

Facebook



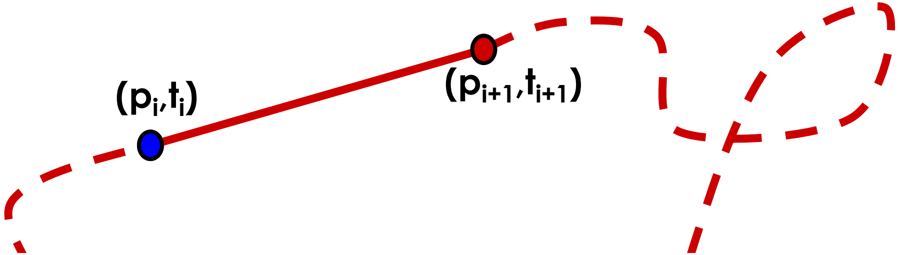
Let friends know where you are

Share your real-time location with friends & family. They can share with you, too.

Google Maps

From location data to trajectories

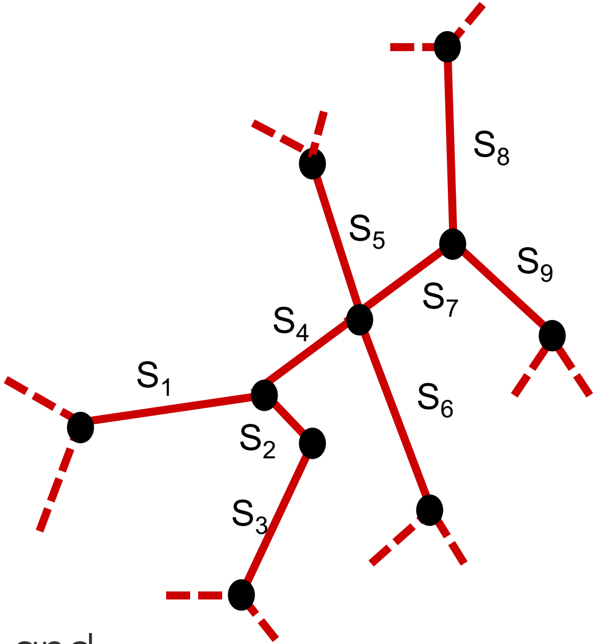
- GPS records samples (p_i, t_i) of our movement – inferring ‘continuous’ movement is not trivial.
- A typical representation of a moving object’s trajectory is a **polyline** (in 4D space; x-, y-, z-, t-) – vertices correspond to (p_i, t_i)
- Usually, **linear interpolation** is assumed between (p_i, t_i) and (p_{i+1}, t_{i+1})
  - to be revisited later (part II)



$$p(t) = \left(x_i + \frac{t - t_i}{t_{i+1} - t_i}(x_{i+1} - x_i), y_i + \frac{t - t_i}{t_{i+1} - t_i}(y_{i+1} - y_i) \right)$$

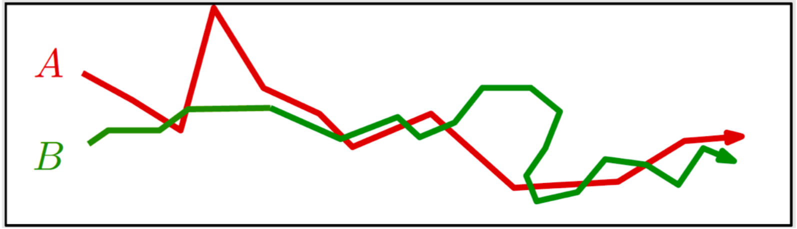
From location data to trajectories (cont.)

- Special case: Network-constrained movement
- Assumes a network / graph G = (V, N)
- Alternative models:
  - **Segment-oriented model**: <S1>, <S2>, etc.
  - **Edge-oriented model**: <S1>, <S2, S3>, etc.
  - **Route-oriented model**: <S1, S4, S7>, <S2, S3>, etc.
- The location of an object is represented by:
  - the entity (segment / edge / route) it is located on, and
  - an offset in [0, 1] denoting the relative location in the entity



Trajectory Similarity

- Key question: How do we measure **similarity** between two trajectories A, B? not so trivial as it sounds

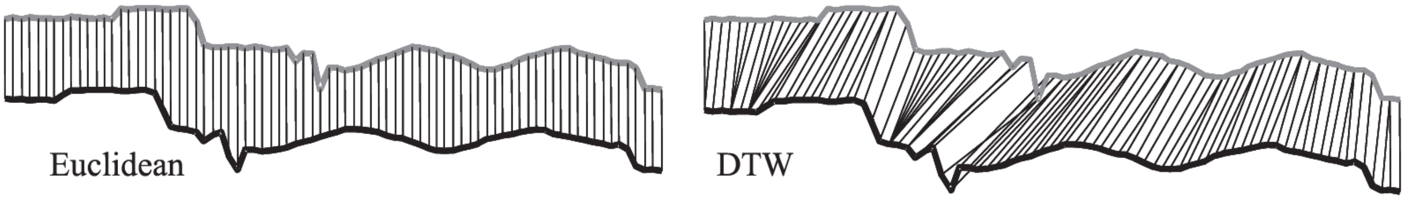


Mount Parnitha National Park
PARNITHOS

- Alternative approaches:
  - Trajectory as a multi-dim. time-series
  - Trajectory as a multi-dim. polyline
  - Trajectory as a movement function

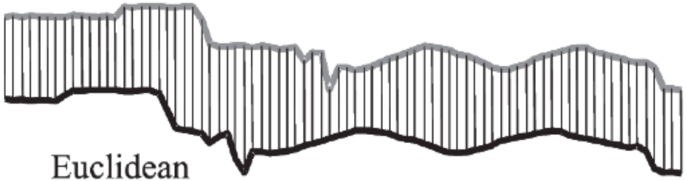
Trajectory as a time-series

- Time-series similarity has been studied extensively (e.g. Vlachos et al. 2002; Chen et al. 2005). Examples:
  - Euclidean distance, Chebyshev distance, Dynamic Time Warping (DTW),
  - Longest Common SubSequence (LCSS),
  - Edit Distance on Real sequences (EDR),
  - Edit distance with Real Penalty (ERP),
  - Swale, etc.



Trajectory as a polyline

- **DISSIM** (Nanni & Pedreschi, 2006; Frentzos et al. 2007)
  - Extension of Euclidean distance:



$$DISSIM(R,S) = \int_{t_1}^{t_n} L_2\big(R(t),S(t)\big)dt$$

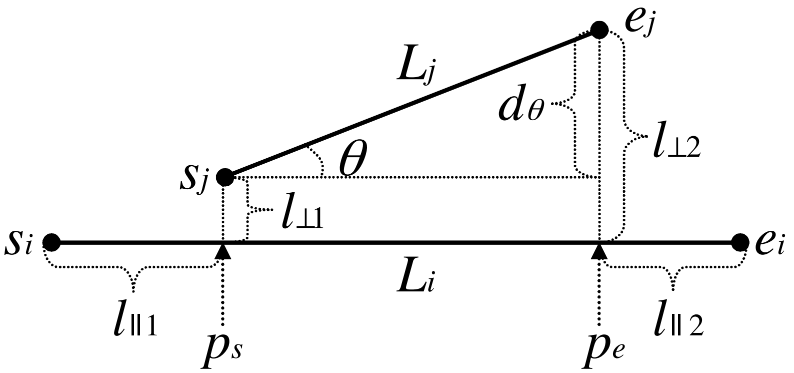
$$DISSIM(R,S) \approx \frac{1}{2}\sum_{k=1}^{n-1}\Big(\Big(L_2\big(R(t_k),S(t_k)\big) + L_2\big(R(t_{k+1}),S(t_{k+1})\big)\Big) \cdot (t_{k+1} - t_k)\Big)$$

- DISSIM function is a metric
  - Conditions: (1) non-negativity; (2) identity of indiscernibles; (3) symmetry; (4) triangle inequality

1. $d(x,y) \geq 0$
2. $d(x,y) = 0 \Leftrightarrow x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)$

Trajectory as a polyline (cont.)

- The **TraClus** approach (Lee et al. 2007)*
- Weighted sum of three components (distances between directed segments):
    - perpendicular $d_{\perp}$
    - parallel $d_{||}$
    - angular $d_{\angle}$



$$d_{\perp} = \frac{l_{\perp 1}^2 + l_{\perp 2}^2}{l_{\perp 1} + l_{\perp 2}}$$

$$d_{\parallel} = \mathrm{MIN}(l_{\parallel 1}, l_{\parallel 2})$$

$$d_\theta = \|L_j\| \times \sin(\theta)$$

* TraClus will be discussed in detail in Part III. Clustering techniques

Trajectory as a movement function

- Trajectory similarity using **Fréchet distance**, e.g. (Buchin et al. 2009)
  - a measure of similarity between curves that takes into account the location and ordering of the points along the curves
  - continuous mapping $\mu :\ \mathrm{A}\ \rightarrow \mathrm{B}$
  - distance $\max_{\alpha \in \mathrm{A}} d\big(\alpha, \mu(\alpha)\big)$

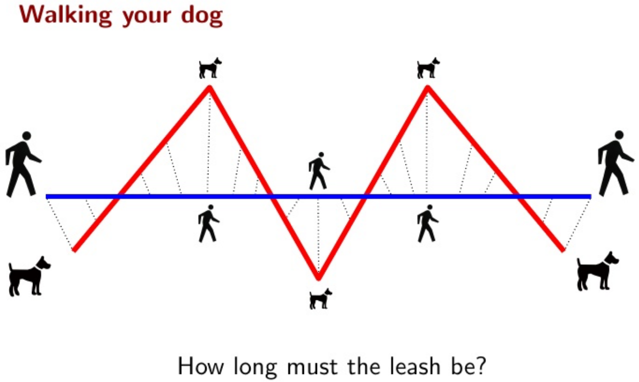


image source: slideshare.net

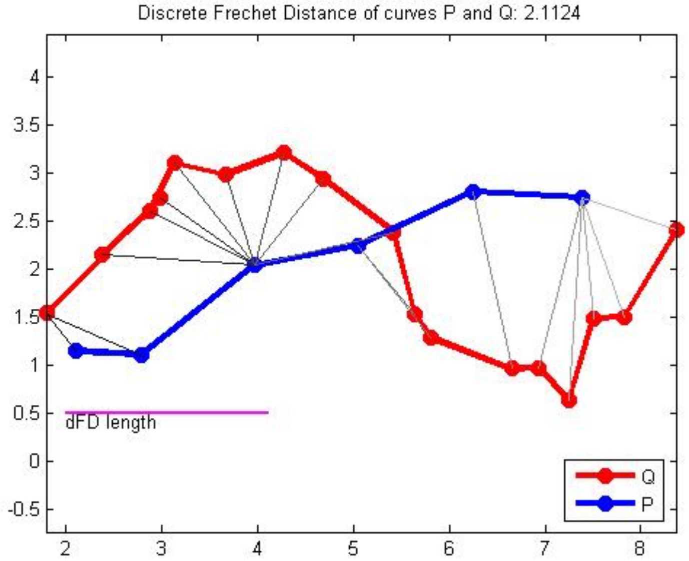


image source: mathworks.com

Examples of datasets @ land (1)

- **GeoLife** (source: Microsoft Research Asia): 182 users under various transportation means; 17,621 trajectories; 68 Km in 2,7 hrs per trajectory, on the average; dense sampling (1 sample every ~5 sec)
- **T-Drive** (source: Microsoft Research Asia): 2,357 taxis in Beijing for 1 week (15 million points, in total); 869 Km per taxi, on the average; sparse sampling (1 sample every ~3 min)

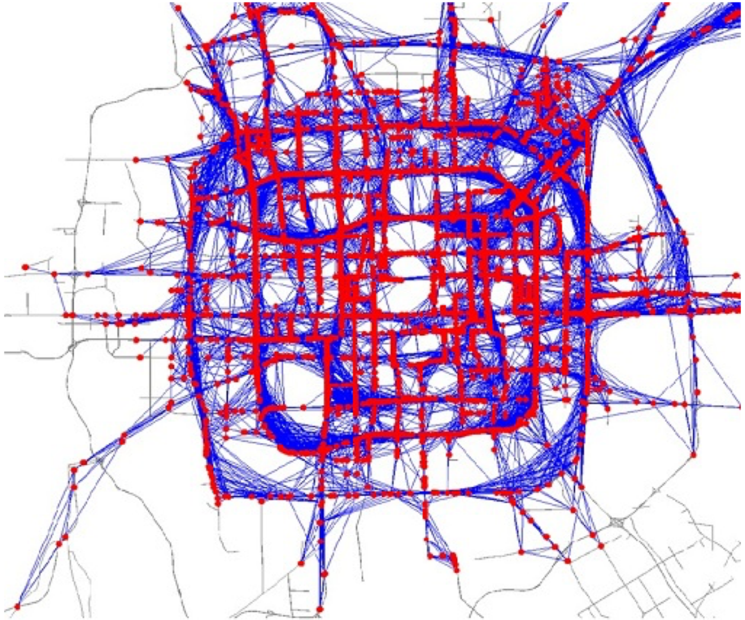
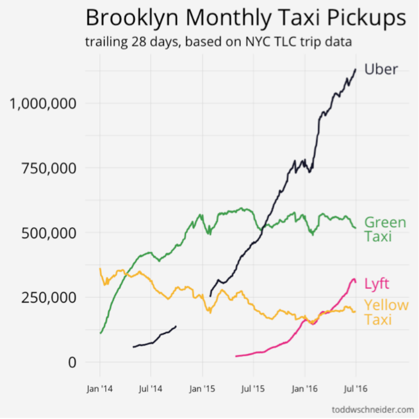
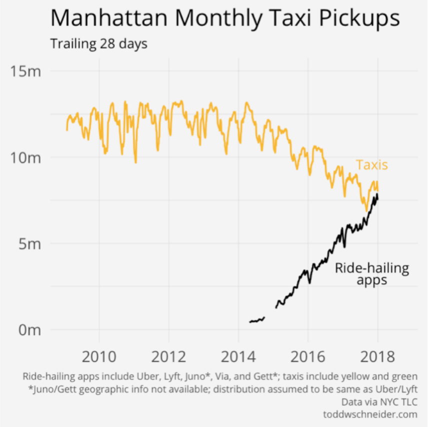


image source: research.microsoft.com

Examples of datasets @ land (2)

- **NYC taxis** (source: NYC Taxi & Limousine Commission): 1.4 billion trips, Jan. 09 – Dec.17.
  - **Ride-hailing apps** data are also provided
  - Attention: pickup – drop-off locations are only available





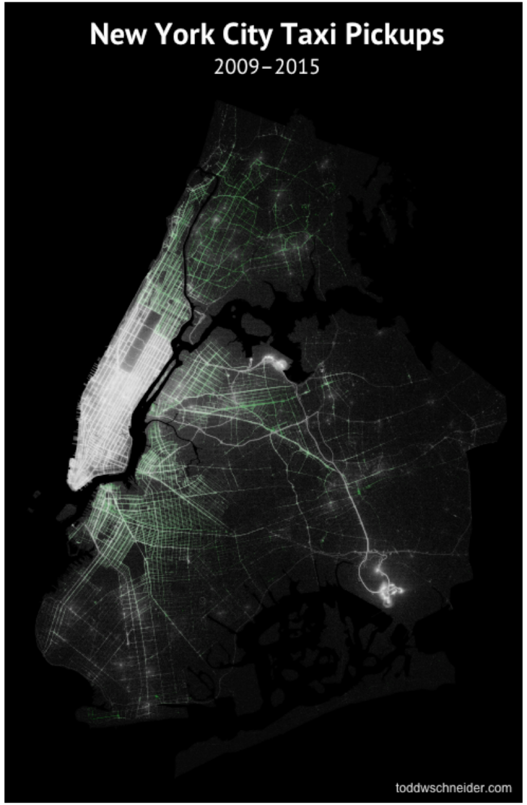
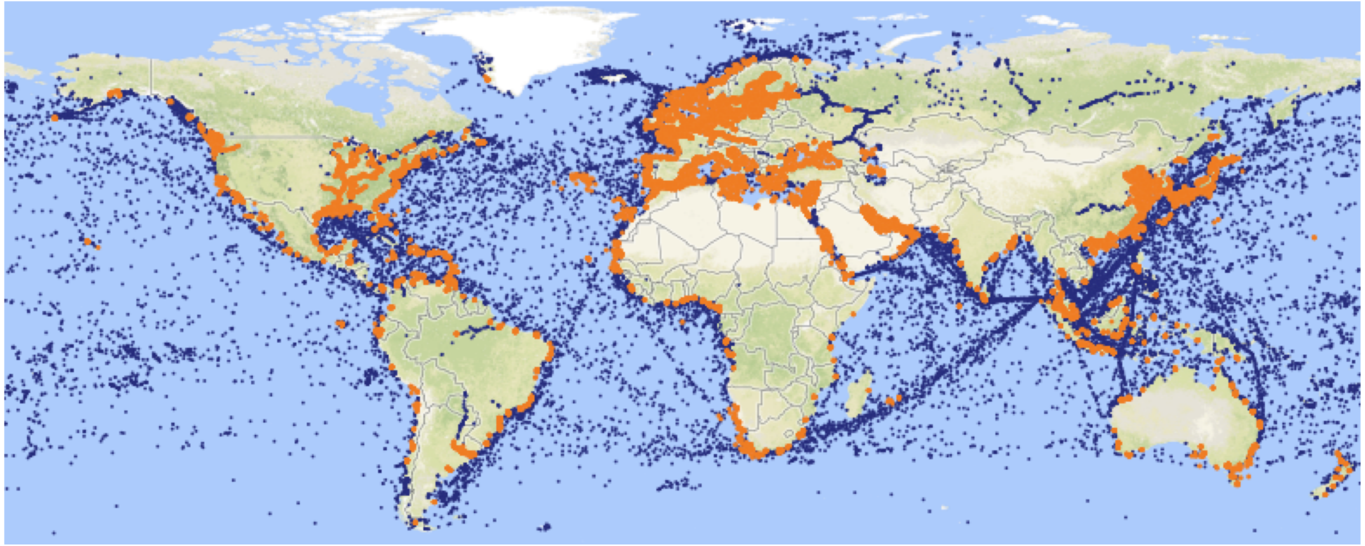


image source: toddwschneider.com

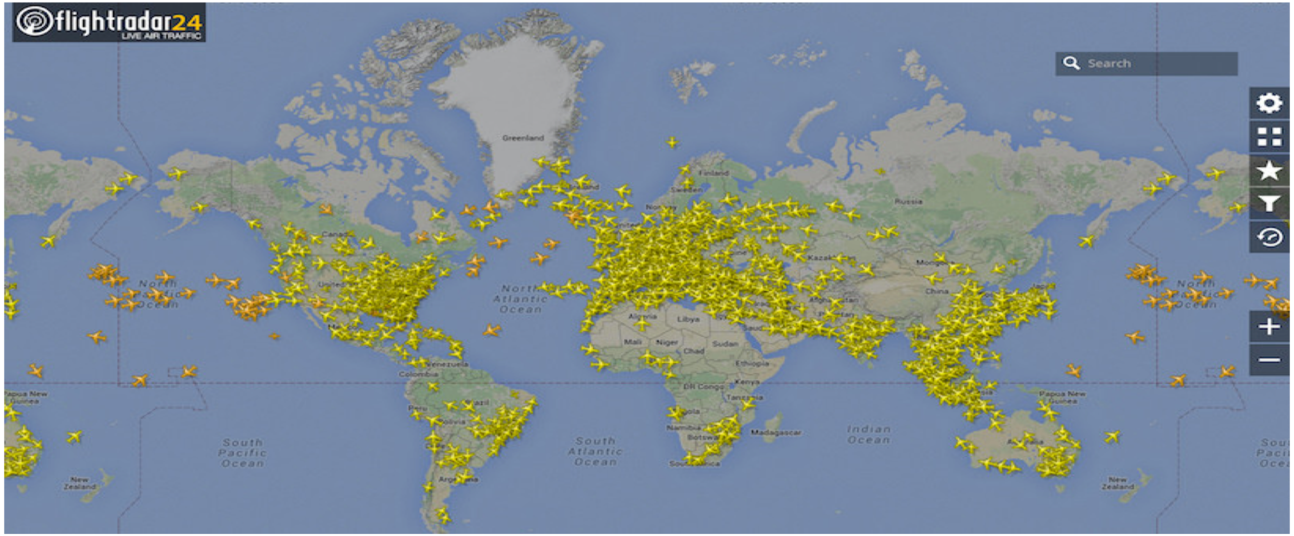
Examples of datasets @ sea

- **AIS** (Automatic Identification System): **tracking system** for identifying & locating vessels at sea
  - 400,000 vessels worldwide (source: vesseltracker.com)



Examples of datasets @ air

- **ADS-B** (Automatic Detection System - Broadcast): **tracking system** for identifying & locating planes on air
  - 50,000 planes flying at the same time worldwide (source: flightradar24.com)

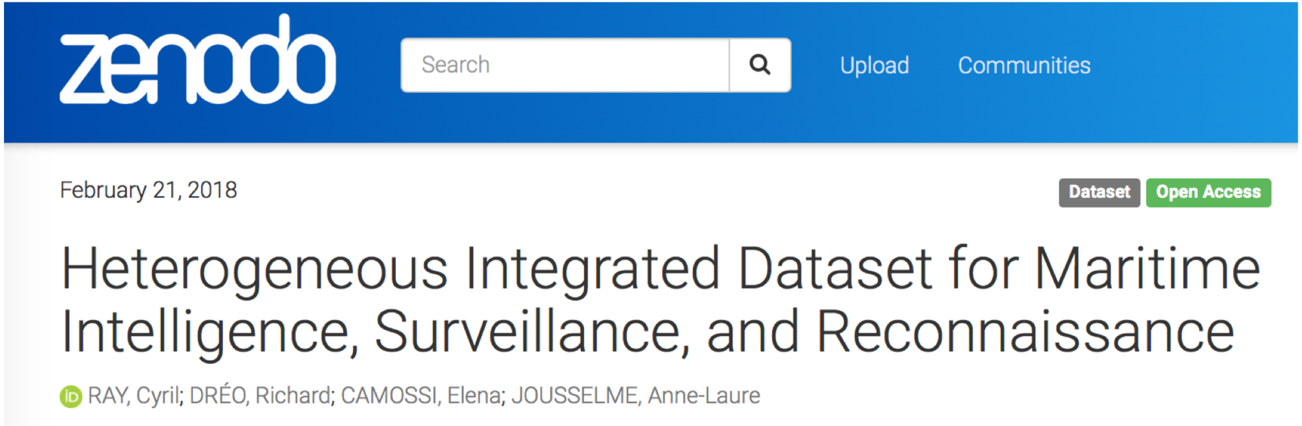


Dataset for hands-on

Southampton
Plymouth
English Channel
Guernsey
Jersey
Celtic Sea
Brest
Rennes
Tours
Nantes
La Rochelle
Limo
Bay of Biscay
Bordeaux

(Ray et al. 2018)

- Collected by Naval Academy, Brest (FR)
- DOI: 10.5281/zenodo.1167595



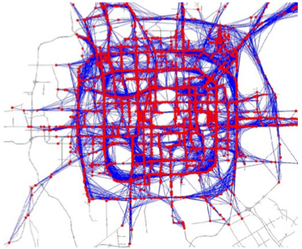
February 21, 2018

Dataset Open Access

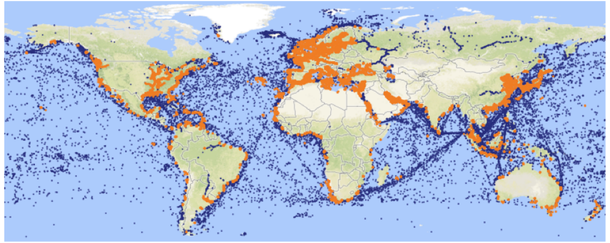
Heterogeneous Integrated Dataset for Maritime Intelligence, Surveillance, and Reconnaissance

RAY, Cyril; DRÉO, Richard; CAMOSSI, Elena; JOUSSELME, Anne-Laure

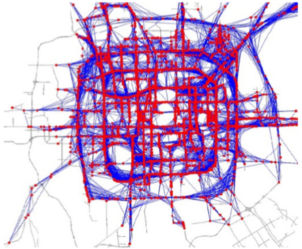
Learning from mobility data



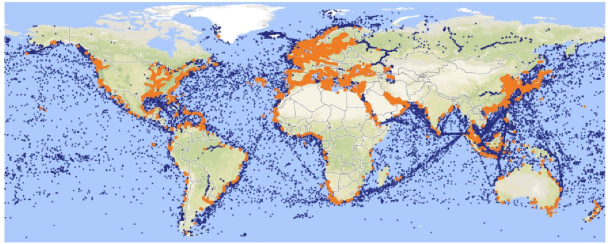
- Analysis at **individual level** (i.e. per moving object):
  - Calculate similarity between an object's actual and expected route
  - Calculate minimum distance between an object's track and a region (e.g. forbidden zone)
  - Calculate maximum number of other objects in an object's vicinity (e.g. 100 m buffer)
  - …



Learning from mobility data (cont.)



- Analysis at **collective level** (i.e. per population of objects)
  - Find objects that move together (for long time)
  - Find the most typical among objects' routes as well as the outliers
  - Find the most crowded places
  - Forecast the near future movement (or even the entire trajectory) of objects
  - etc.



Analytics example -1

- Tanker vessels’ typical movement in Aegean sea, GR
  - Blue lines: typical routes
  - Green rectangles: protected areas
- Further research upon data analytics results
  - e.g. risk analysis

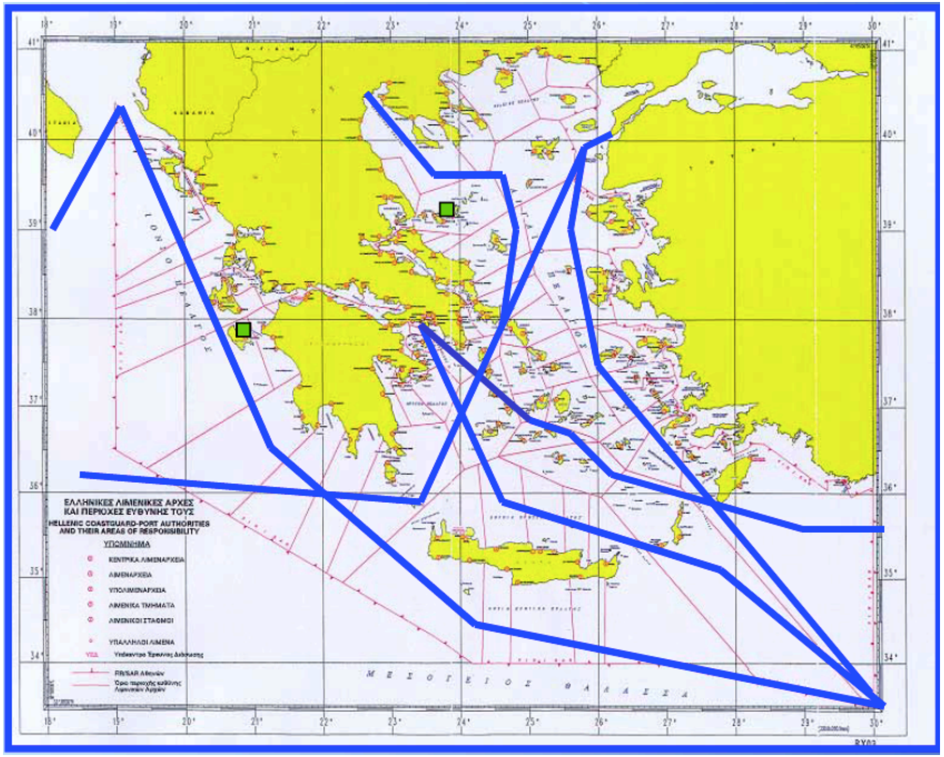
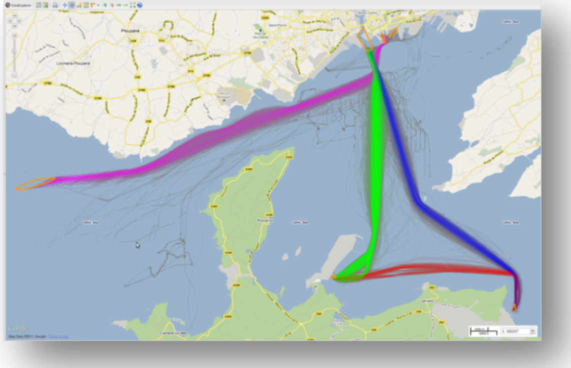

ΕΛΛΗΝΙΚΕΣ ΛΙΜΕΝΙΚΕΣ ΑΡΧΕΣ
ΚΑΙ ΠΕΡΙΟΧΕΣ ΕΥΘΥΝΗΣ ΤΟΥΣ
HELLENIC COASTGUARD-PORT AUTHORITIES
AND THEIR AREAS OF RESPONSIBILITY
ΥΠΟΜΝΗΜΑ
ΚΕΝΤΡΙΚΑ ΛΙΜΕΝΑΡΧΕΙΑ
ΛΙΜΕΝΑΡΧΕΙΑ
ΥΠΟΛΙΜΕΝΑΡΧΕΙΑ
ΛΙΜΕΝΙΚΑ ΤΜΗΜΑΤΑ
ΛΙΜΕΝΙΚΟΙ ΣΤΑΘΜΟΙ
ΥΠΑΛΛΗΛΟΙ ΛΙΜΕΝΑ
FIR/SAR Αθηνών
Μ Ε Σ Ο Γ Ε Ι Ο Σ   Θ Α Λ Α Σ Σ Α

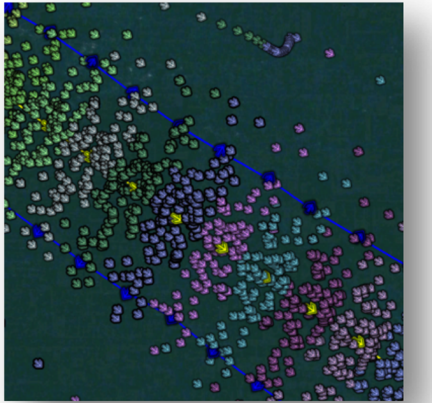
image source: archipelago.gr

Analytics example -2

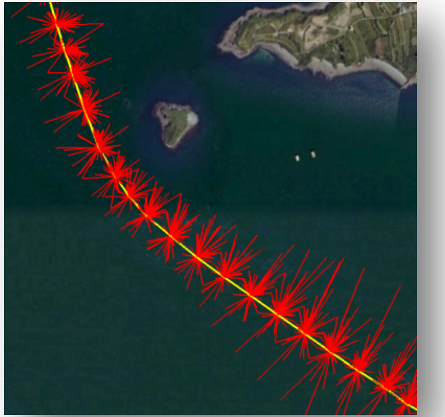
- Vessels' movement in bay of Brest, FR
- Further research:
  - e.g. classification of captains as normal vs. dangerous

frequent patterns





Cloud of locations

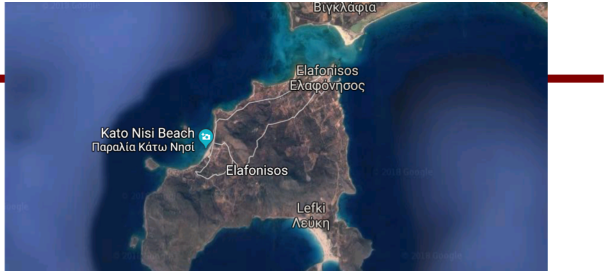


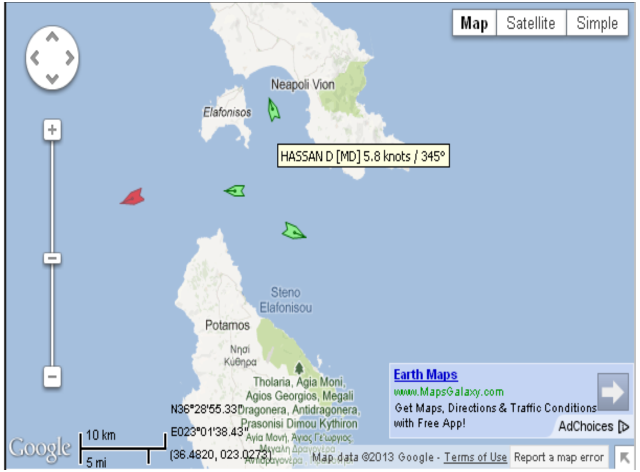
Actual vs. typical locations per route

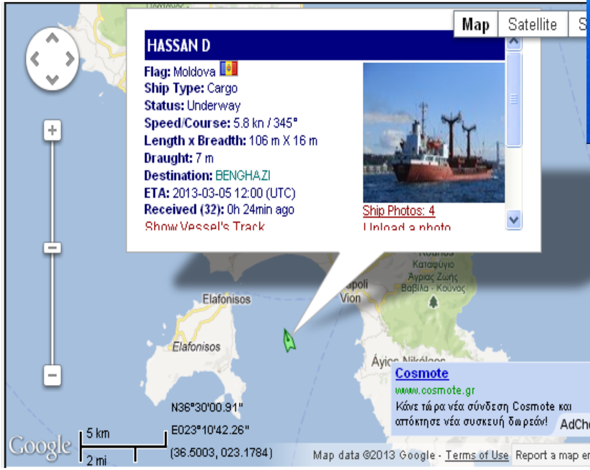
images source: datacron-project.eu

Analytics example -3

- Elafonissos – Peloponnese narrow pass (570 m.)
  - Natura 2000 protected area
  - Searching for suspicious sailing




Map
Satellite
Simple
Neapoli Vion
Elafonisos
HASSAN D [MD] 5.8 knots / 345°
Steno
Elafonisou
Potamos
Νησί
Κύθηρα
Tholaria, Agia Moni, Agios Georgios, Megali Dragonera, Antidragonera, Prasonisi Dimou Kythiron
N36°28'55.33"
E023°01'38.43"
(36.4820, 023.0273)
10 km
5 mi
Google
Earth Maps
www.MapsGalaxy.com
Get Maps, Directions & Traffic Conditions with Free App!
AdChoices
Map data ©2013 Google - Terms of Use
Report a map error



HASSAN D

Flag: Moldova
Ship Type: Cargo
Status: Underway
Speed/Course: 5.8 kn / 345°
Length x Breadth: 106 m X 16 m
Draught: 7 m
Destination: BENGHAZI
ETA: 2013-03-05 12:00 (UTC)
Received (32): 0h 24min ago
Show Vessel's Track

Ship Photos: 4
Upload a photo

image sources: wikipedia.org; marinetraffic.com

Data visualization is a must …

- … in order to “know your data” better
  - Example: major flight routes from Paris to Istanbul
- DataViz is out of scope in this course
- For those interested:
  - e.g. Andrienko et al. (2007; 2008; 2017a; 2017b)

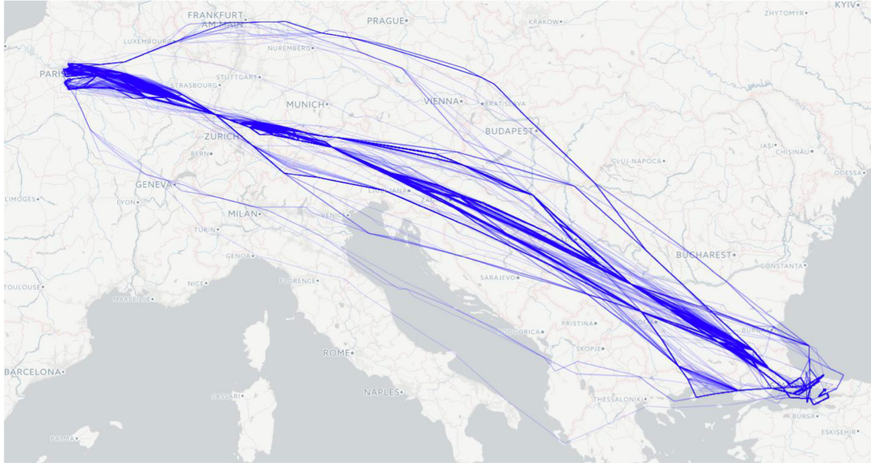
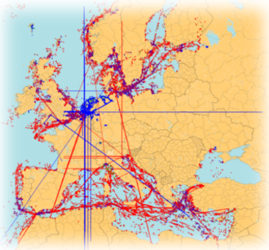


image source: (Andrienko et al. 2017a)

(Big) Mobility Data Analytics Challenges

Volume and Velocity

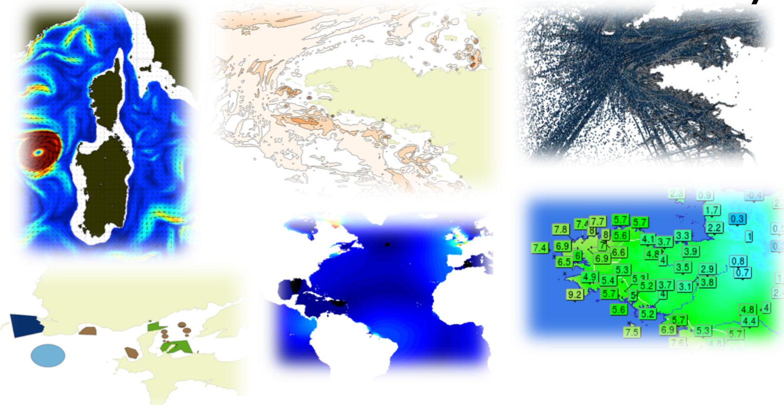
12K distinct ships/day, 200M AIS contacts/month in EU waters



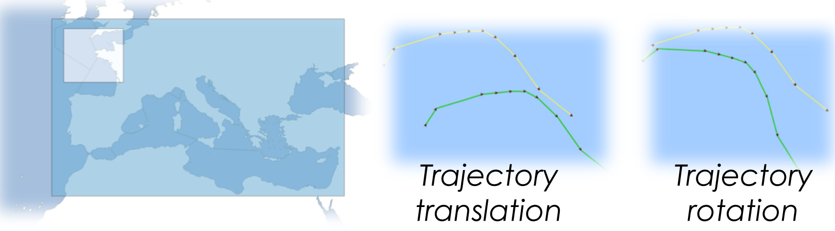
Noisy and error-prone data due to receivers limited coverage, positioning devices switch-off

Veracity Issues

Variety



Historical & aggregated data, geographical & environmental data, contextual data, etc.



Trajectory translation

Trajectory rotation

Multi-scale assessment with pseudo-synthetic labelled data

Image source: (Claramunt et al. 2017)

Summarizing part I ...

- Location- and mobility- aware data is tracked in everyday routine activities
  - Thesaurus of information → challenge for further investigation (= data analytics)
- Issues and challenges
  - How to clean my data?
  - How to store it?
  - How to analyze it?

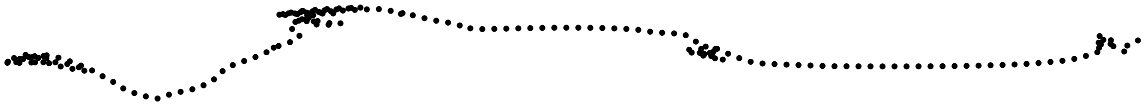
Part II:
Pre-processing your data

“It is a very sad thing that nowadays there is so little useless information." Oscar Wilde

Data pre-processing

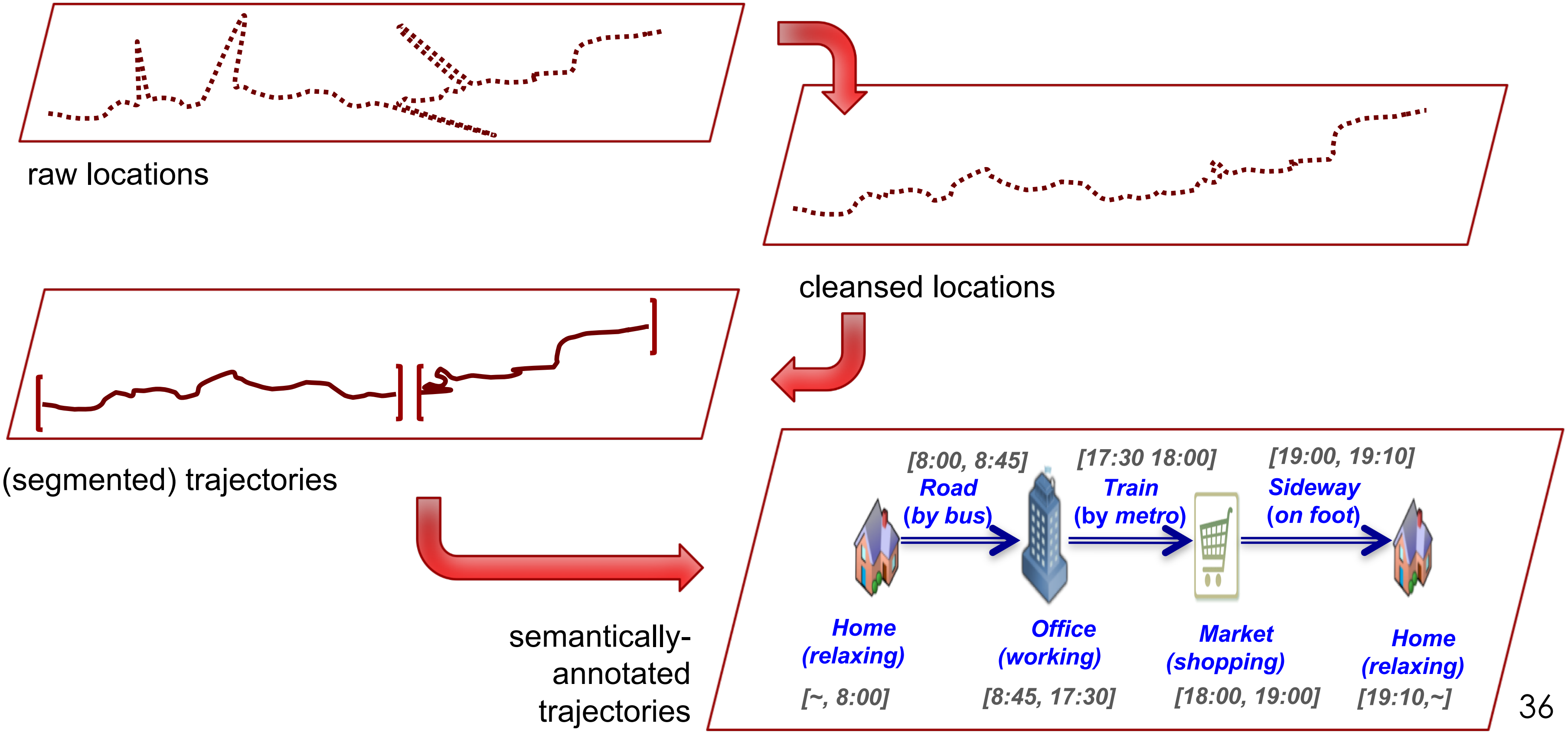
- Definition: **preparing data for analytics purposes**

$T = \{ <p_1, t_1>, <p_2, t_2>, \ldots, <p_n, t_n> \}$



- Data pre-processing tasks:
  - **Cleansing** (noise removal, smoothing, map matching, etc.)
  - **Transformation** (trajectory segmentation, simplification, re-sampling, etc.)
  - **Enrichment** (semantic annotation, data fusion, etc.)
  - **Sampling** (the entire dataset)
- Data storage (and indexing)
  - Moreover, generating 'realistic' synthetic datasets (why?)

Data pre-processing tasks



raw locations

cleansed locations

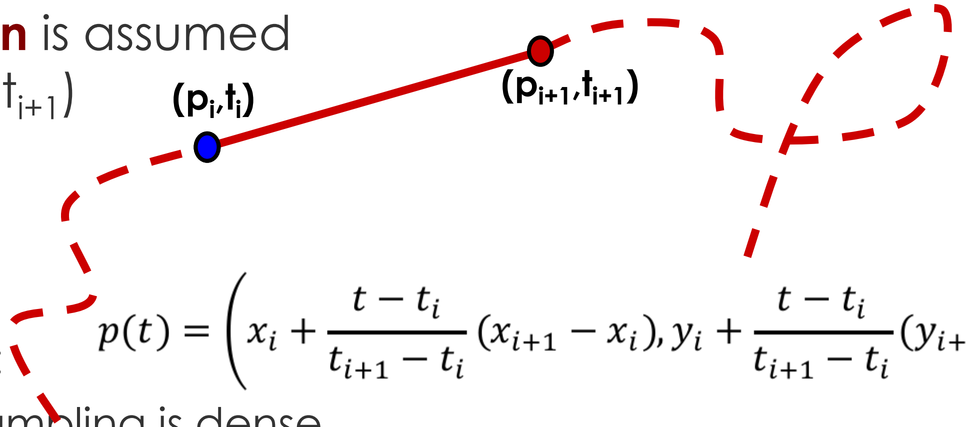
(segmented) trajectories

semantically-annotated trajectories

From GPS data to trajectories

- Recall that … a typical representation of a moving object's trajectory is a **polyline** (in 4D space; x-, y-, z-, t-)
  - vertices correspond to time-stamped locations

- Usually, **linear interpolation** is assumed between (p_i, t_i) and (p_{i+1}, t_{i+1})

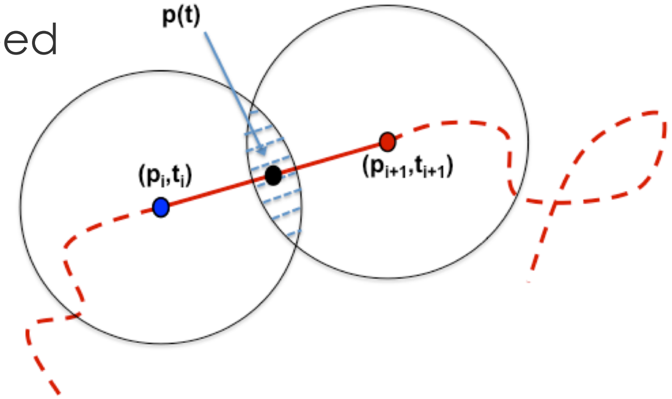


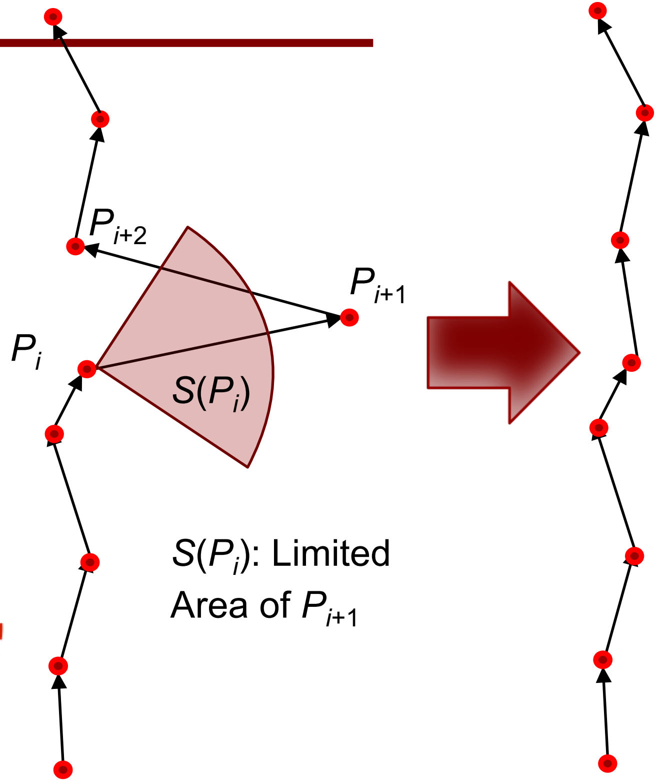
$$p(t) = \left(x_i + \frac{t - t_i}{t_{i+1} - t_i}(x_{i+1} - x_i), y_i + \frac{t - t_i}{t_{i+1} - t_i}(y_{i+1} - y_i) \right)$$

- Notes on linear interpolation:
  1. Makes sense only when sampling is dense
  2. Does not obey the physical rules (why?) … but who cares (why?)

GPS Data Cleansing

- Erroneous recordings: noise vs. random errors
- **Noise** corresponds to values that are 'impossible' to appear
- Can be detected and removed using appropriate filters
  - e.g. maximum speed
- **Potential Area of Activity** (PAA)

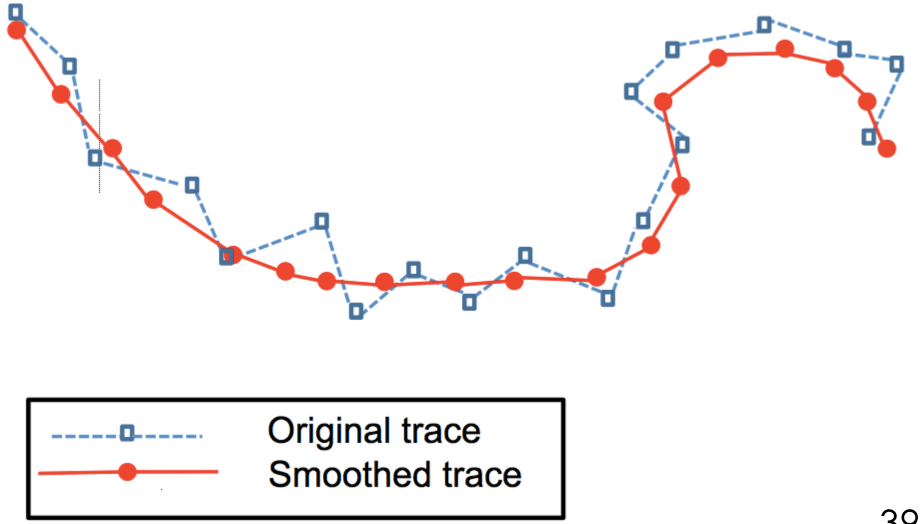




$S(P_i)$: Limited Area of P_{i+1}

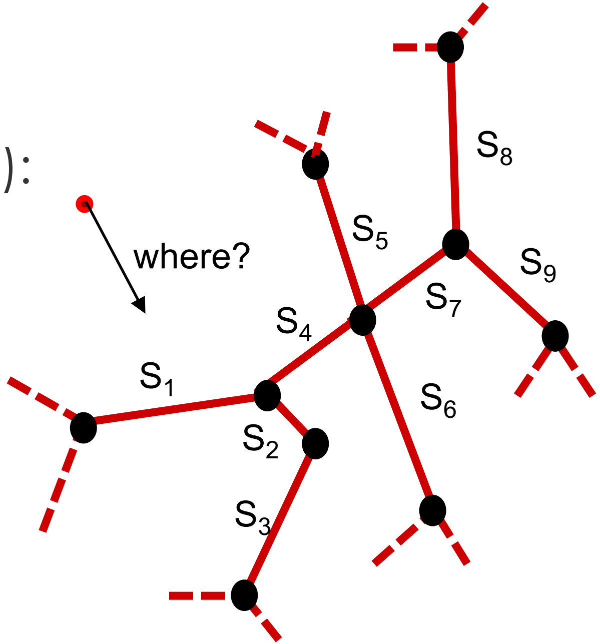
GPS Data Cleansing (cont.)

- Erroneous recordings: noise vs. random errors
- **Random errors** correspond to 'possible' values that appear to be small deviations from actual ones
- Can be smoothed using a plethora of statistical methods
  - e.g. least squares spline approximation (de Boor, 1978)



GPS Data Cleansing (cont.)

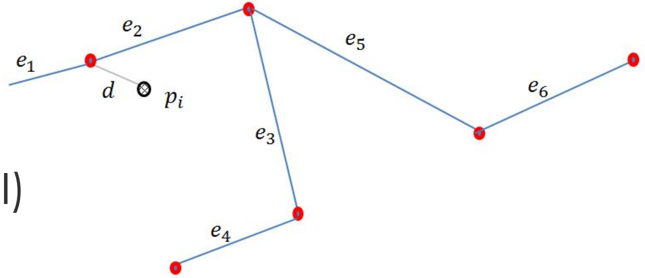
- Special case: network-constrained movement
- Requires an additional step: **map-matching**
- Several techniques (Quddus et al. 2003; 2007):
  - Geometric map-matching
  - Topological map-matching
  - Probabilistic map-matching
  - Hybrid map-matching

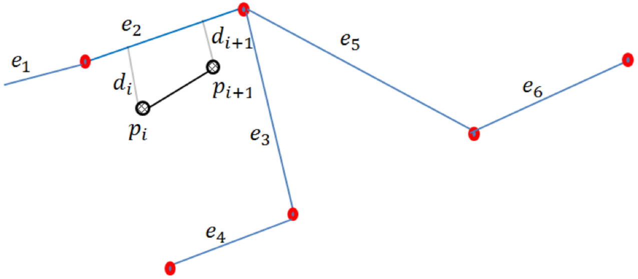


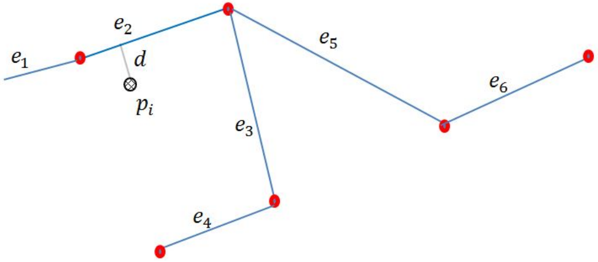
- Examples…

Geometric map-matching

- The basic idea: map a point into its closest position on the network
- Three types:
  - Point-to-point (e.g. Euclidean distance)
  - Point-to-curve (e.g. perpendicular distance)
  - Curve-to-curve (e.g. Fréchet distance; see part III)

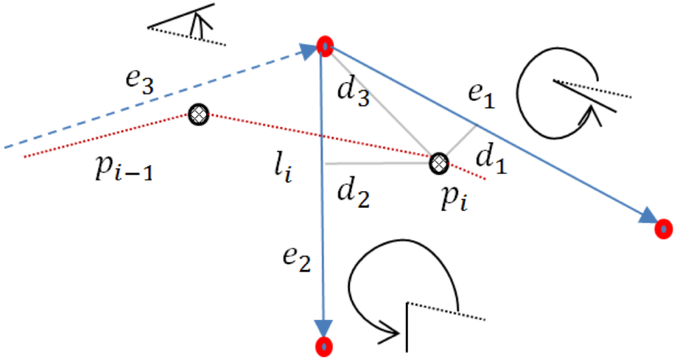






Topological map-matching

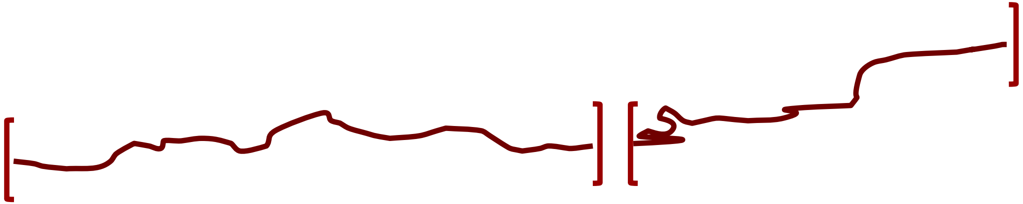
- Utilize both the geometry and the connectivity / adjacency of the graph
- Two steps:
  - Choose the most suitable node(s) of the graph
  - Match the point
- Could be enhanced by a "look-ahead" approach



Trajectory identification (segmentation)

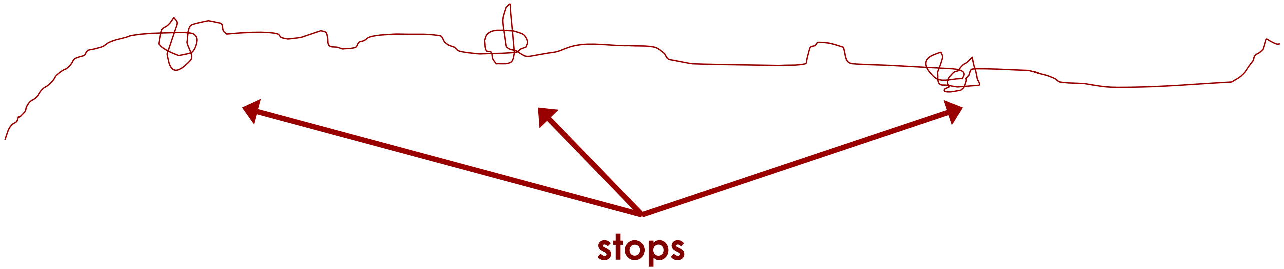
- Goal: **Segment sequences of points** in homogeneous sub-sequences (= trajectories)

- Various approaches:
  - Identification via raw (spatial / temporal) gap
  - Identification via prior knowledge (e.g. office hours, sleeping hours)
  - Correlation-based identification (ideas from time-series segmentation)
  - etc.



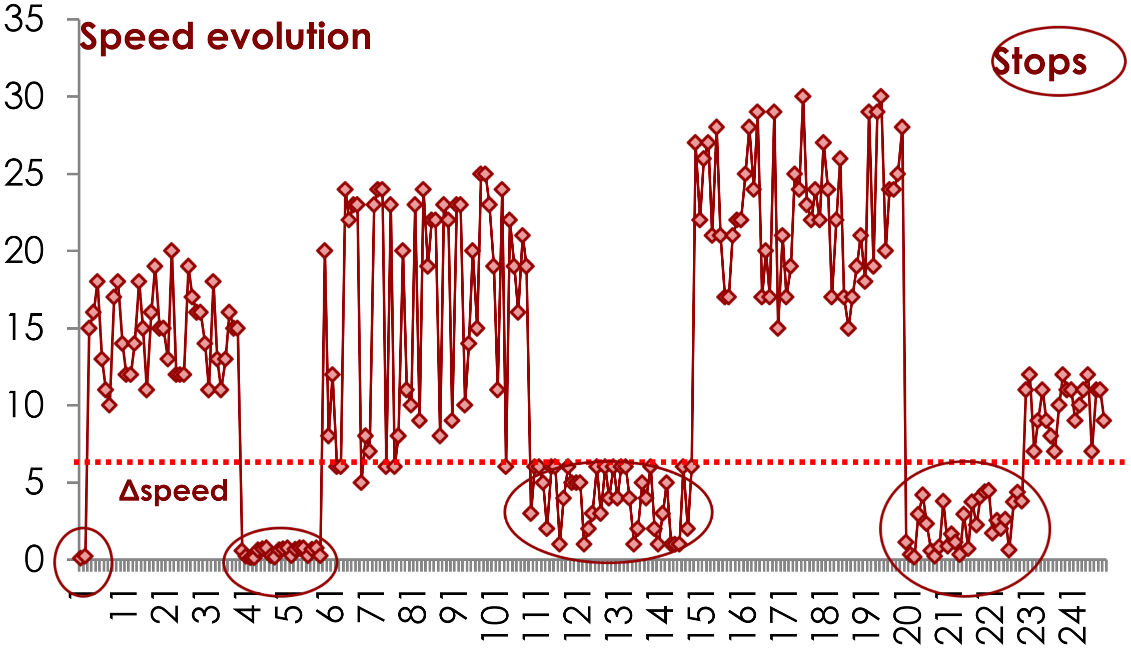
Stop discovery

- How can Stop be detected in a raw trajectory? Solutions:
  - when the trajectory intersects the geometry of a POI and the duration of intersection is above a given temporal duration threshold: **SMoT technique** (2007)
  - when dense areas of the trajectory points are detected, using e.g. a density-based clustering algorithm, and those areas are mapped to a POI: **CB-SMoT technique** (2008)



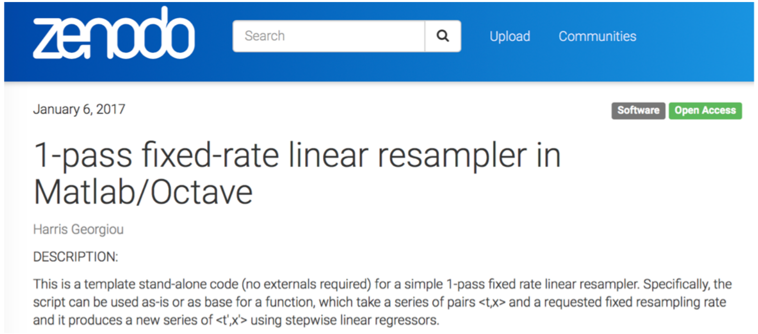
Stop discovery (cont.)

- Alternative: velocity-based stop identification



Trajectory re-sampling

- The need for **fixed re-sampling**: prerequisite by some algorithms ☹
- Possible approach: interpolation over sampled location data
- 1-pass tool (Georgiou, 2017) – linear interpolation



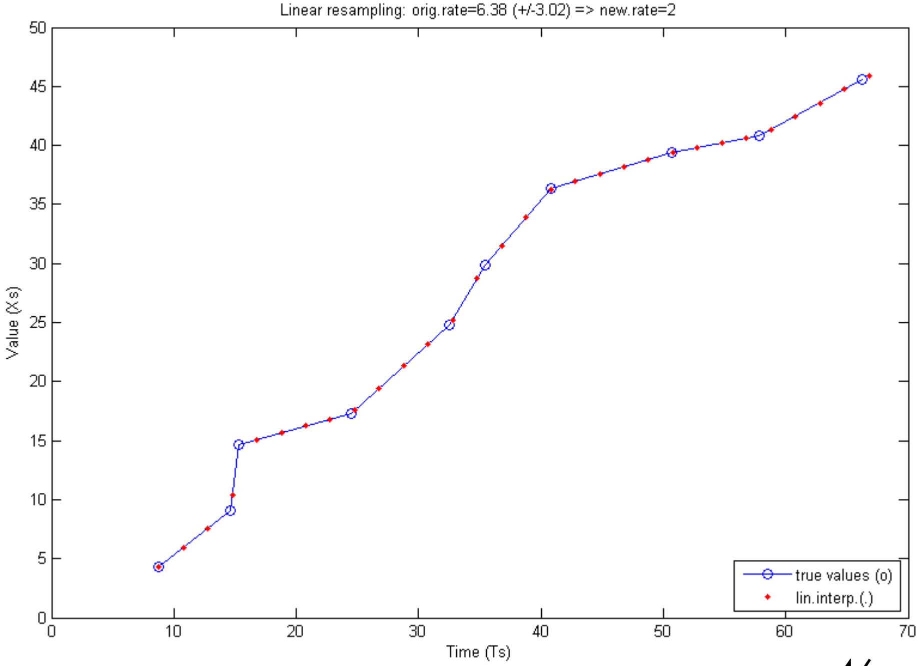
January 6, 2017

1-pass fixed-rate linear resampler in Matlab/Octave

Harris Georgiou

DESCRIPTION:

This is a template stand-alone code (no externals required) for a simple 1-pass fixed rate linear resampler. Specifically, the script can be used as-is or as base for a function, which take a series of pairs <t,x> and a requested fixed resampling rate and it produces a new series of <t',x'> using stepwise linear regressors.



Trajectory simplification

- The need for simplification: efficiency in storage, processing time, etc.
  - Actually, a form of data compression
- Goal: maintain the original signature as much as possible by keeping a set of **critical points** only
- Approaches
  - Offline, i.e. multi-pass, vs.
  - Online, i.e. 1-pass

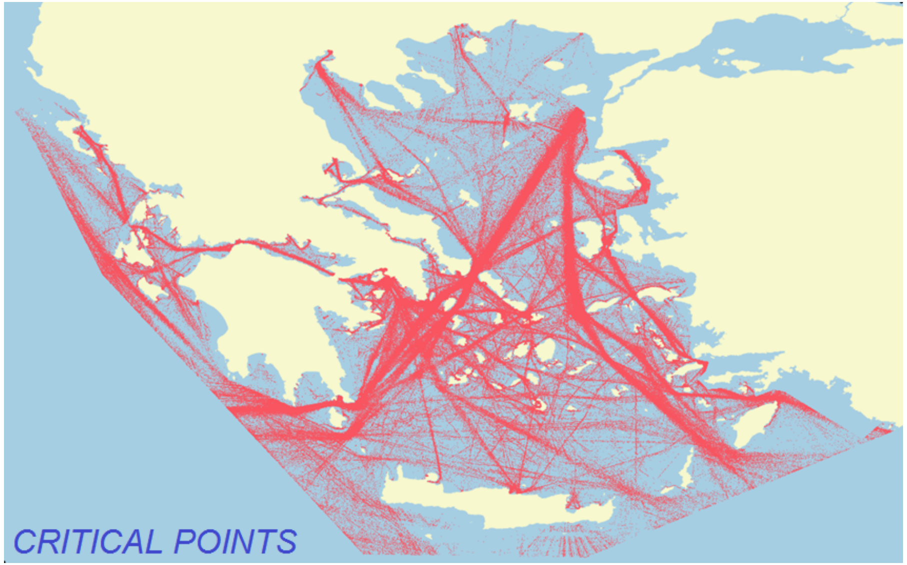
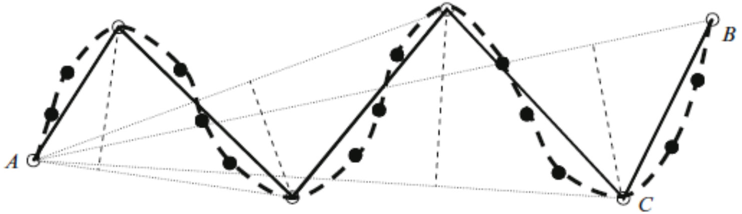
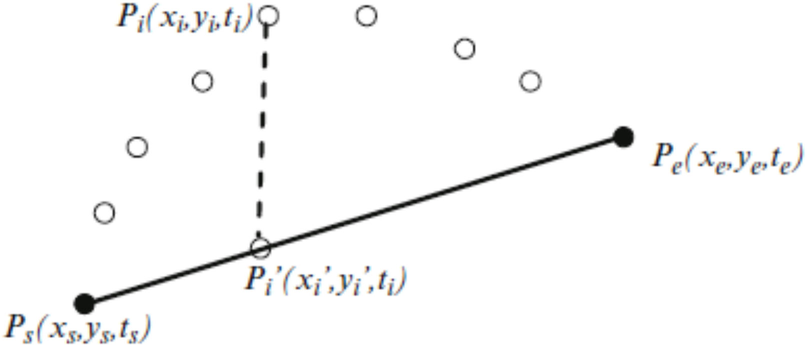


image source: aminess.eu

Trajectory simplification (cont.)

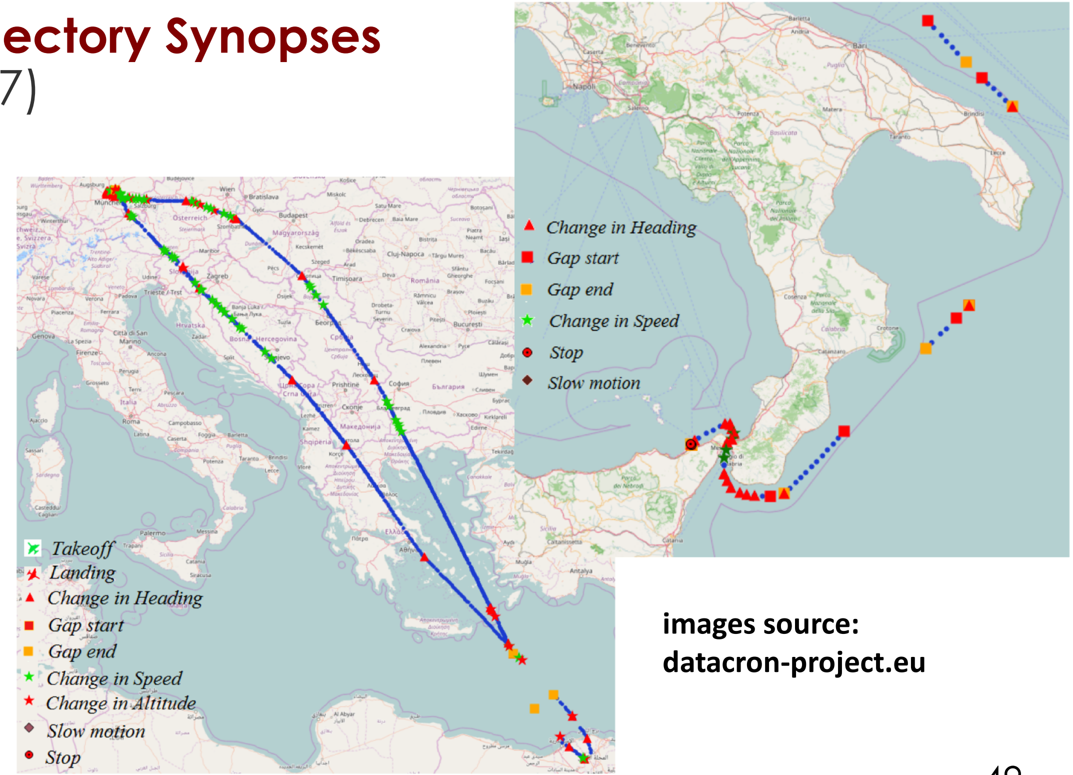
- Offline approaches:
  - top-down vs. bottom-up vs. sliding window vs. opening window
- e.g. **Synchronous Euclidean Distance – SED** (Meratnia & de By, 2004)
  - Customizes polyline simplification (Douglas & Peucker, 1973) to the mobility domain





Trajectory simplification (cont.)

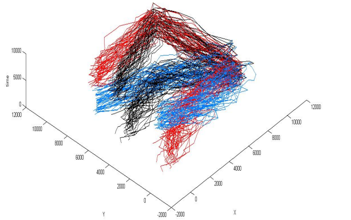
- Online approaches, e.g. **Trajectory Synopses** (Patroumpas et al. 2015; 2017)
- Maintains a **velocity vector** per moving object in order to detect **instantaneous events**
  - stop; change in velocity vector; etc.
- Tradeoff: degree of compression vs. quality of approximation

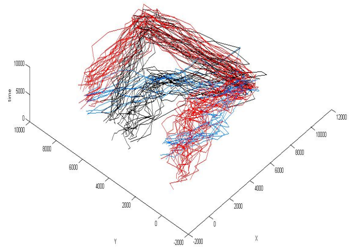


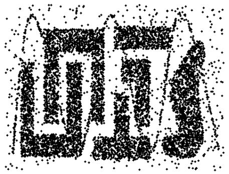
images source: datacron-project.eu

Trajectory dataset sampling

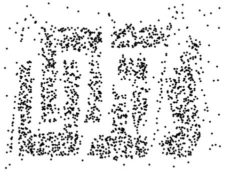
- Motivation: Can we get the gist of a real dataset by working on a sample of it?
- If yes, we can extrapolate our findings on the ‘small’ (sampled) to the ‘large’ (entire) dataset
  - e.g. run a computationally intensive algorithm to discover mobility patterns
- Sampling has been extensively studied in Statistics



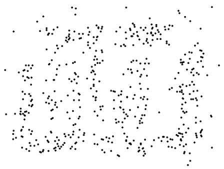




8K points



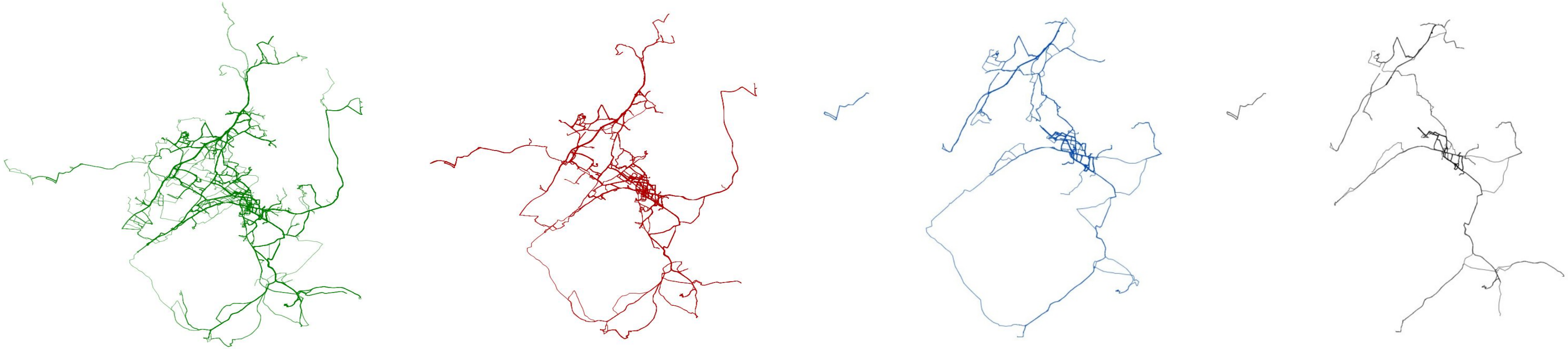
4K points



2K points

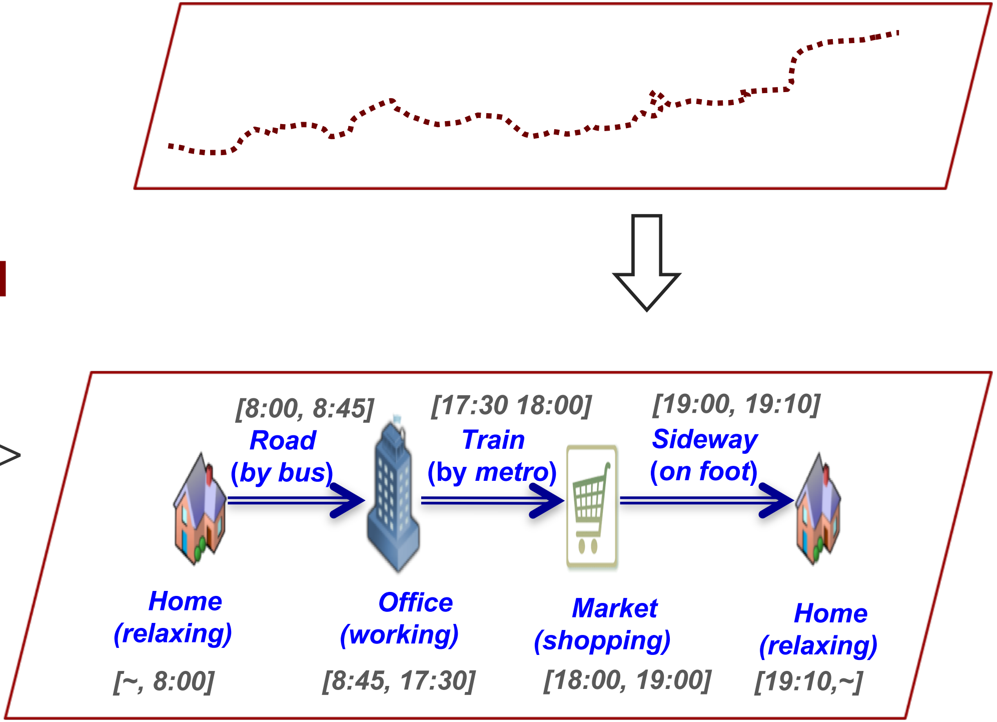
Trajectory dataset sampling (cont.)

- **T-sampling** (Pelekis et al. 2010; Panagiotakis et al. 2012) samples the top-k representative trajectories, following a voting process
  - Trajectory segmentation is neighborhood- rather than geometry-aware
- Example: T-sampling runs (1100 > 200 > 100 > 40 trajectories)



Trajectory enrichment

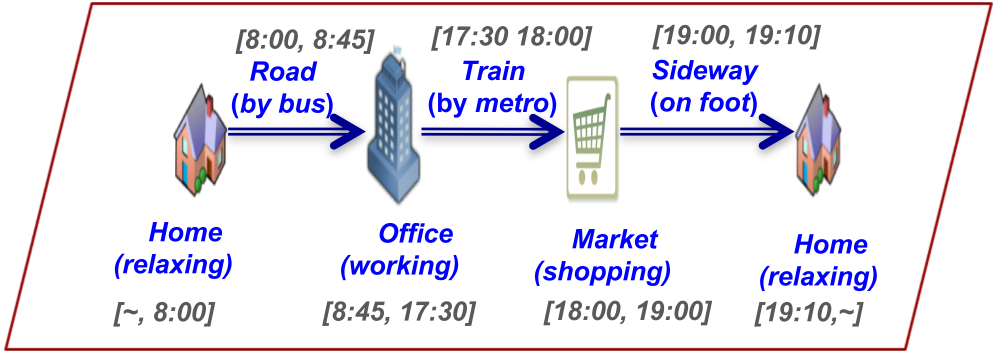
- From “raw” trajectories …
  - sequences of time-stamped locations (p,t)
- … to **semantically-annotated** trajectories
  - meaningful mobility tuples <where, when, what/how/why>
  - Not only a matter of down-scaling the dataset size
  - Mainly, towards enhanced analysis and understanding of movement



Trajectory enrichment (cont.)

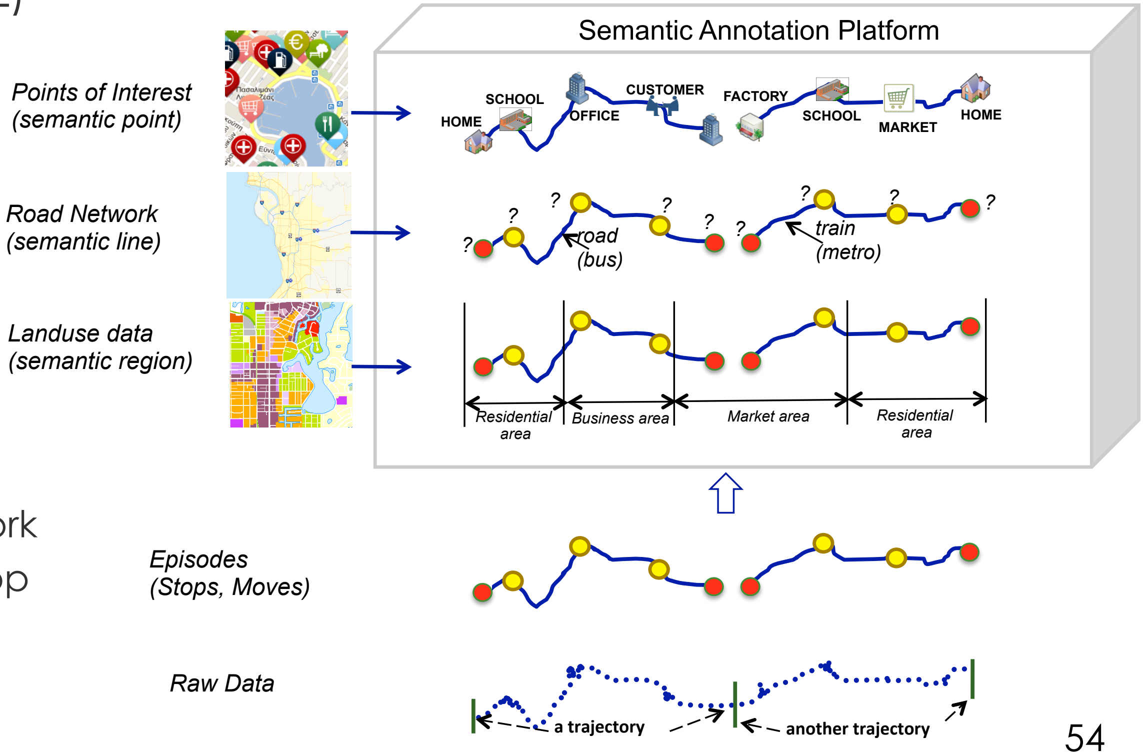
- **Semantic trajectory** (Yan et al. 2011; Parent et al. 2015): an alternative (semantically-annotated) representation of the motion path of a moving object
  - homogenous fractions of movement

- A trajectory is reconstructed as a **sequence of episodes (stops/moves)** along with appropriate **tags**
  - when? where? how? what? why?



Trajectory enrichment (cont.)

- **SeMiTri** (Yan et al. 2011; 2012)
- Preliminary: **segmentation**
  - Detecting stops, changes in movement pattern, etc.
- Core: **semantic annotation**
  - **Semantic regions**: annotate episodes with geographic ROIs (using e.g. OSM)
  - **Semantic lines**: annotate episodes with underlying infrastructure, e.g. road network
  - **Semantic points**: annotate Stop episodes with POI types (using e.g. HMM techniques)



Semantic Annotation Platform

Points of Interest (semantic point)

Road Network (semantic line)

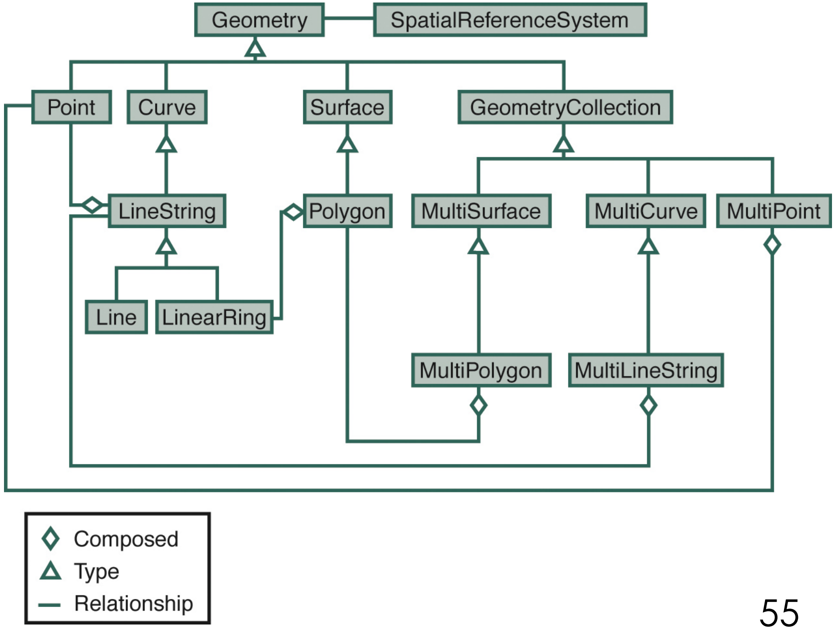
Landuse data (semantic region)

Episodes (Stops, Moves)

Raw Data

DBMS storage options

- Issue: could spatial DBMS efficiently organize mobility information?
  - Objective: both space and time should be considered as **first-class citizens**.
- Current options:
  - **Spatial DBMS** simulated to handle trajectories as polylines, e.g. PostGIS
    - PostGIS supports 2D/3D/4D geometry data types
    - A trajectory can be simulated by a 3D/4D linestring (= sequence of points)
  - vs. dedicated **Moving Object Databases** (MOD)



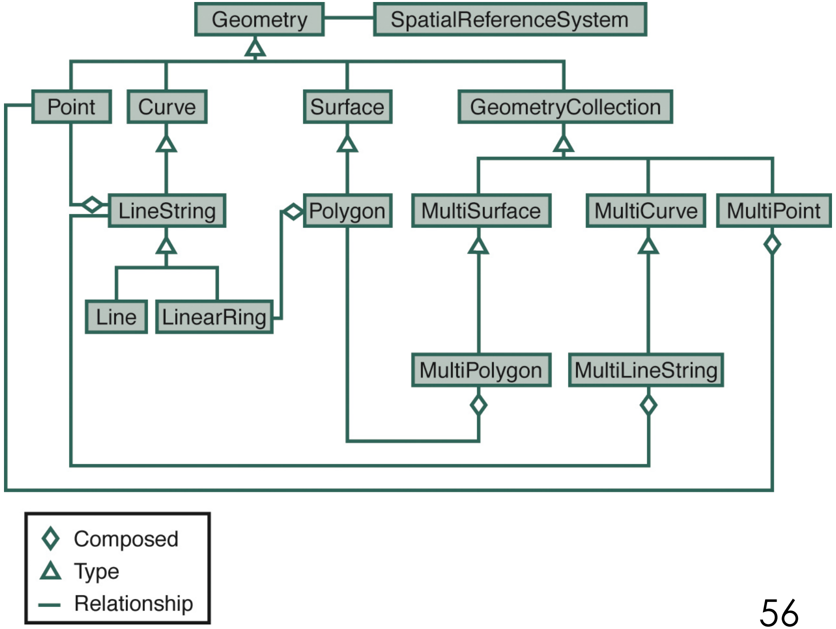
The PostGIS solution

- Create a table of 3D polylines …

```
CREATE TABLE trajectories (
            id integer PRIMARY KEY,
            geom geometry(LINESTRINGZ)
);
```

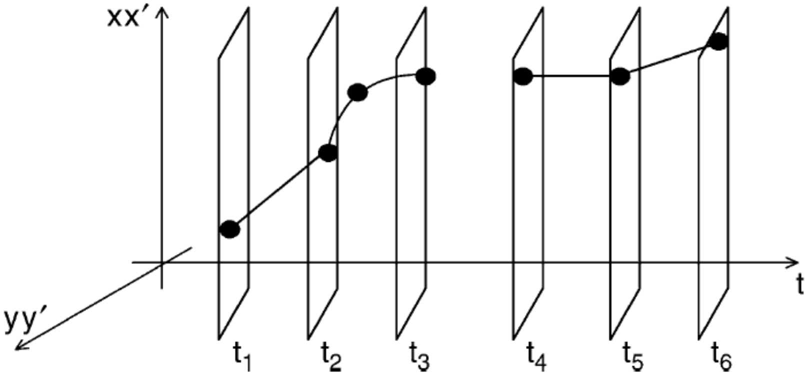
- … then insert WKT converted to geometry

```
INSERT INTO trajectories(id, geom)
             VALUES (1, ST_GeomFromText
             ('LINESTRING(0 0 0, 1 1 1, 2 2 2)')
);
```



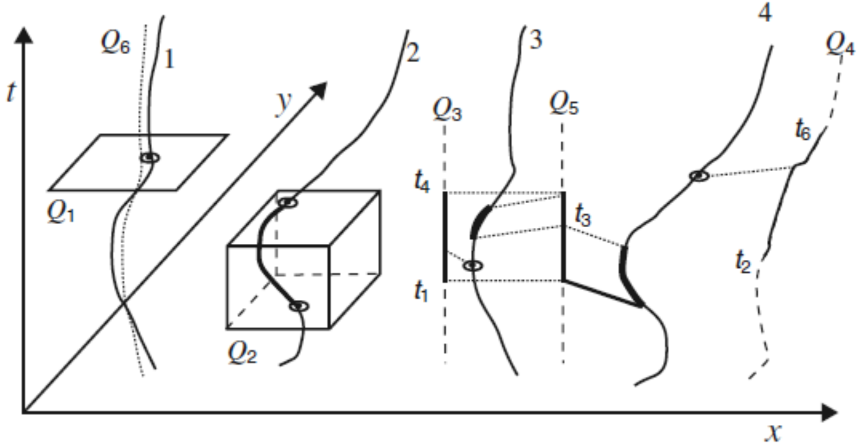
Prototype MOD Engines

- Prototype MOD engines for archival (trajectory) data
  - **SECONDO** (de Almeida et al. 2006) @ Uni. Hagen
  - **HERMES** (Pelekis et al. 2014) @ Uni. Piraeus
- Based on the **'sliced' representation** of trajectories
  - Within each slice, the movement is modeled by a 'simple' function (linear, arc, etc. interpolation)



- Further discussion on MODs is out of scope in this course
  - See e.g. (Pelekis & Theodoridis, 2014), ch.5

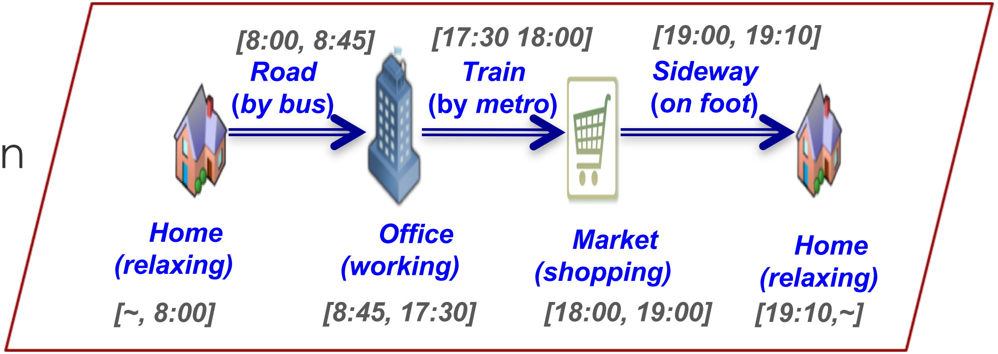
Querying trajectory datasets



- **Time-slice queries**
  - find the locations of trajectories at a given timestamp
- **Spatiotemporal range / NN queries**
  - find objects located inside a given spatial region during a given time interval
  - find objects located nearest to a given (fixed) position / (moving) object during a given time interval
- **Topological queries**
  - find the trajectories that entered (crossed, bypassed, etc.) a given region during a given time interval
- **Trajectory similarity queries**
  - find the trajectories that are similar to a given trajectory

Querying trajectory datasets (cont.)

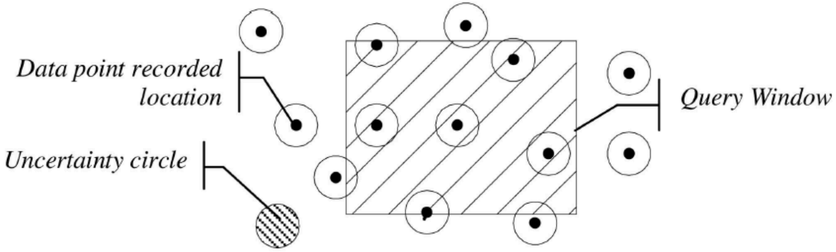
- **Queries on semantically-enriched data**. Examples:
  - Find people who follow the pattern "home – office – home" Mon-Fri
  - Find people who cross the city center from office back to home (by making intermediate stops of at least ½ hour duration there)
  - e.g. (Sideridis et al. 2016)

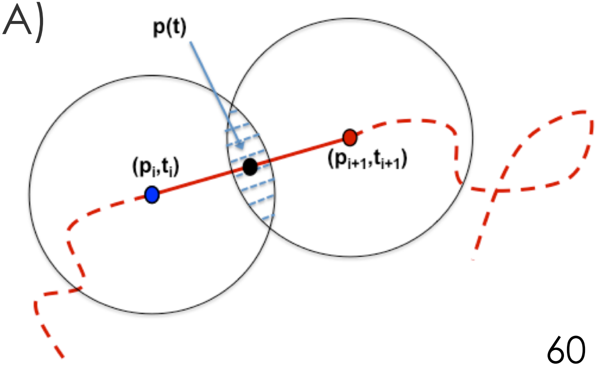


- **Spatio-temporal-textual pattern (ST^{2}P) queries**. Example:
  - Find people who (i) started from home between 8am-9am, (ii) walked for at least 1 hour, and (iii) returned back home between 7pm-8pm.
  - e.g. (Sakr & Guting, 2011; Gryllakis et al. 2017)

Querying under uncertainty

- Our ground truth consists of (i) sampled locations, which (ii) are possibly incorrect !! (due to GPS measurement error)
  - Result: uncertainty in query results (false hits, missed hits, etc.)
  - e.g. find the trajectories that **definitely** / **possibly** entered a given area
- Technically: where could an object have been located at any time t in between two sampled locations at t_i and t_{i+1}?
  - The union of all lenses: **Potential Area of Activity** (PAA)



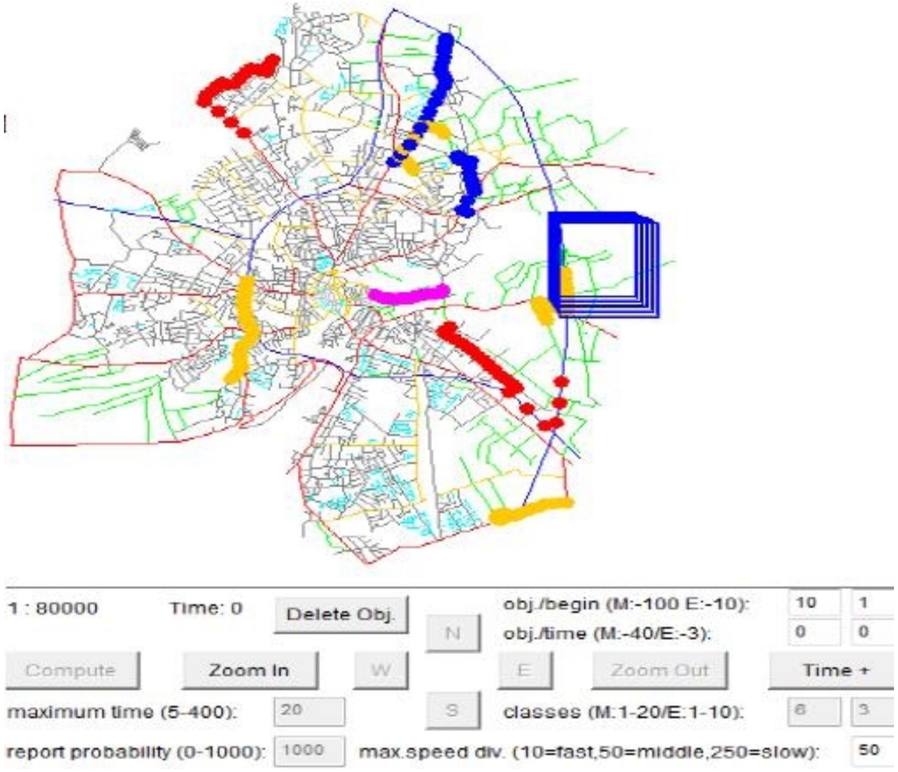


The requirement for synthetic data generators

- Necessary for performance evaluation purposes
- **Micro-** (i.e., dealing with single moving objects) vs. **Macro-scopic** (i.e., dealing with the traffic flow rather than single moving objects)
- Microscopic generator example:
  - Free movement on the plane: **GSTD** (Theodoridis & Nascimento, 2000)
- Macroscopic generator examples:
  - Movement under network-constraints: **Brinkhoff** (Brinkhoff, 2002), **BerlinMOD** (Düntgen et al. 2008)
  - Semantically-annotated movement following predefined patterns under network-constraints: **Hermoupolis** (Pelekis et al. 2013; 2016)

Brinkhoff’s generator

- Methodology:
  - generate starting points
  - generate length of route (depending on object class)
  - generate destination for each object
  - compute the route
  - compute the trajectory by generating a random speed every time unit
    - based on capacity, weather, edge class, etc.



1 : 80000

Time: 0

obj./begin (M:-100 E:-10): 10 1

obj./time (M:-40/E:-3): 0 0

maximum time (5-400): 20

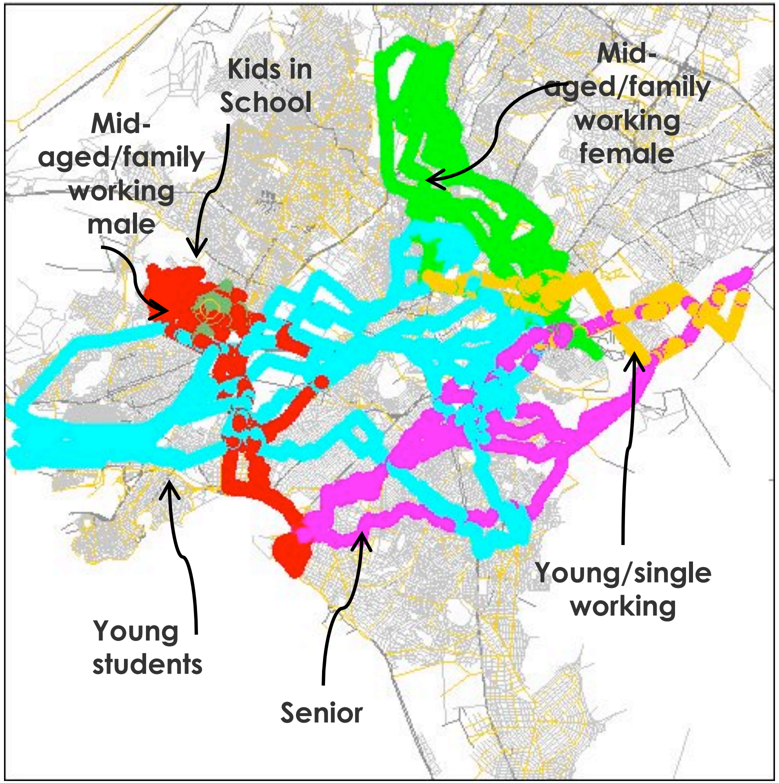
classes (M:1-20/E:1-10): 6 3

report probability (0-1000): 1000

max.speed div. (10=fast,50=middle,250=slow): 50

Hermoupolis generator

- Generate objects moving in an urban (network-constrained) area
  - … according to different population profiles of given distribution, e.g.
    - Kids in school: 20%
    - Young students: 10%
    - etc.
- Dual output: synchronized raw (GPS-like) + semantic trajectories
- Towards the “by-example” paradigm



Summarizing part II …

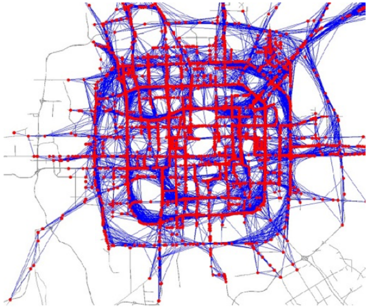
- Building (and maintaining) meaningful trajectory datasets from raw GPS data involves:
  - Data cleansing (noise removal, random errors smoothening)
  - Trajectory identification – segmentation – simplification – enrichment, etc.
  - Efficient data storage and querying (past vs. current locations)
- Trends in this area include:
  - Spatio-temporal-textual query processing (the era of Semantic trajectories)
  - Predictive query processing
  - Building synthetic data generators "by-example"

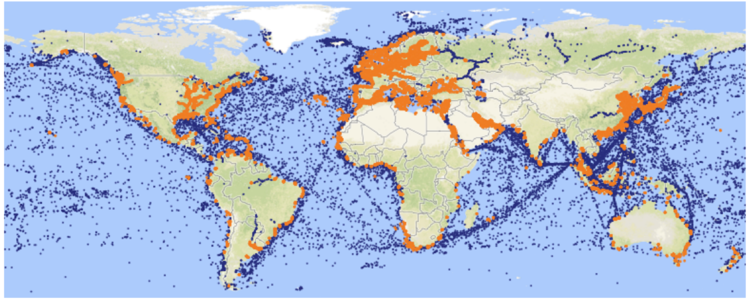
Part III:
Analyzing your data

"The only source of knowledge is experience."
Albert Einstein

Types of mobility data analytics

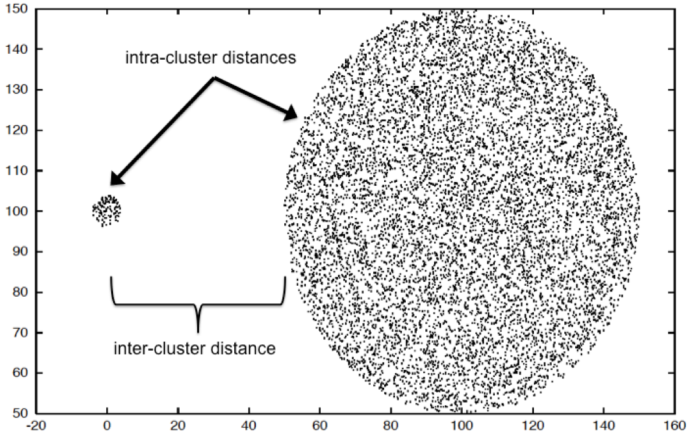
- Discovering **groups** and **outliers**
- Discovering **frequent routes** (hot paths) and **frequent locations** (hot spots)
- **Classification** and **prediction** tasks
- etc.





Cluster analysis principles

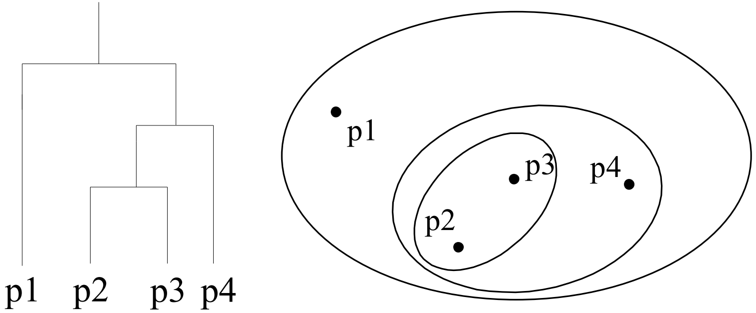
- Objective: find groups of objects, such that:
  - the objects assigned to the same group are expected to be quite similar to each other, whereas
  - the objects assigned to different groups are expected to be quite dissimilar to each other



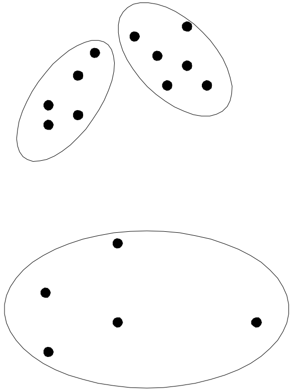
- Goal:
  - **intra- (inter-) cluster distance** should be minimized (maximized, resp.)
- Issue: appropriate "similarity" measures (recall Part I. Similarity)

Clustering techniques

- **Hierarchical clustering**: a set of nested clusters, organized as a hierarchical tree
  - Hierarch is built upon objects' similarity

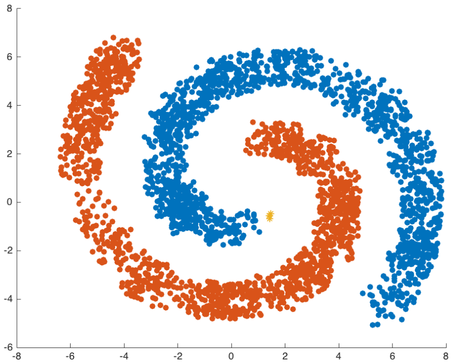


- **Partitional clustering**: a partitioning of objects into non-overlapping subsets (clusters), according to their similarity
  - Spherical-oriented methods: K-means, etc.
  - Density-based methods: DBSCAN, OPTICS, etc.



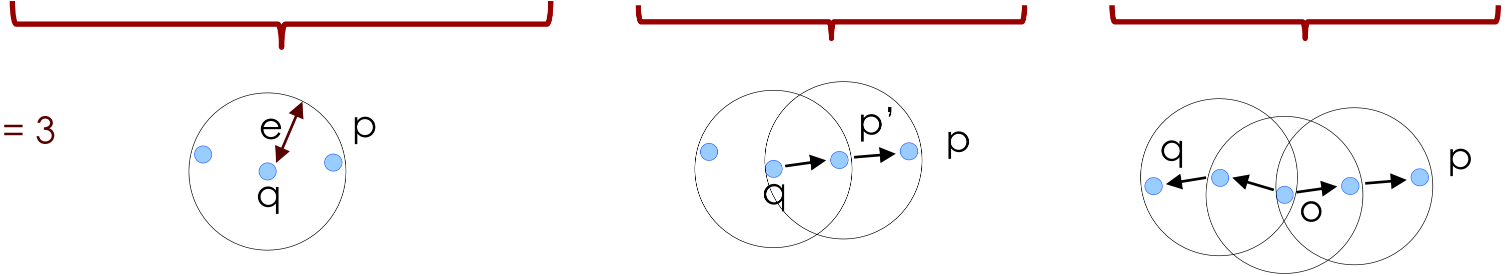
Clustering techniques (cont.)

- **DBSCAN** (Ester et al. 1996): density-based clustering
  - 'density' corresponds to the population within an object's neighborhood
  - Method parameters: radius of the neighborhood (e); minimum population within the neighborhood (m)



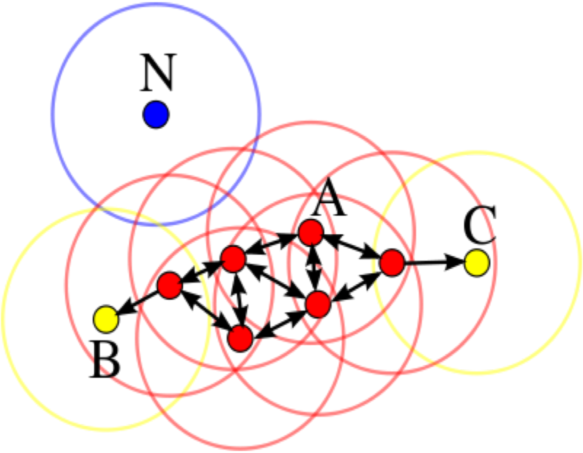
- The notion of **density reachability**
  - Directly Density-Reachable vs. Density-Reachable vs. Density Connected

m = 3



Clustering techniques (cont.)

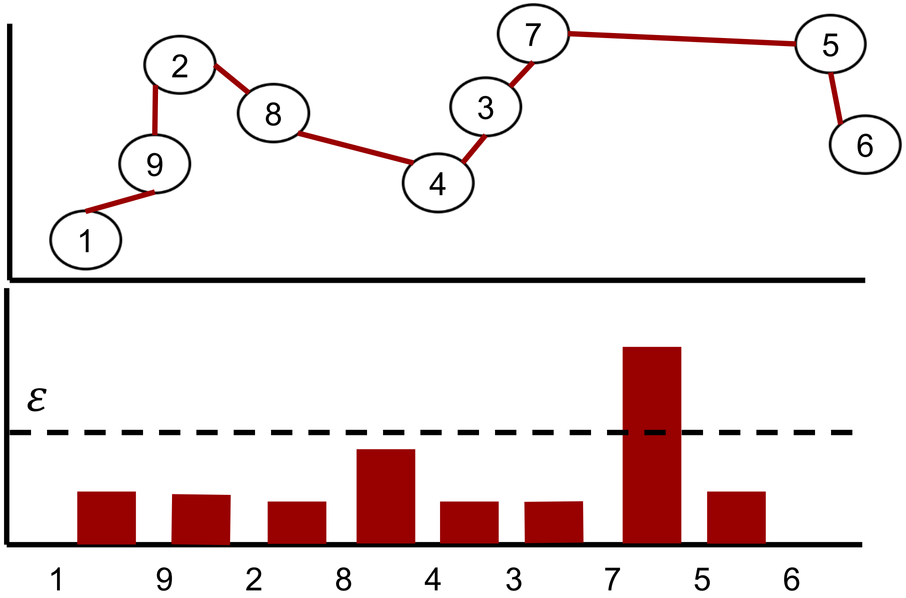
- **DBSCAN** (cont.) - A point is characterized as:
  - **core**, if it has at least m points within its e- neighborhood
  - **border**, if itself is not a core point, but it lies in the neighborhood of a core point
  - **noise**, otherwise



- Core vs. Border vs. Noise points
  - Core points build clusters
  - Border points are assigned to the clusters built by their cores
  - Noise points are marked as outliers

Clustering techniques (cont.)

- **OPTICS** (Ankerst et al, 1999)
  - The concept of 'core' objects (again…)
  - Objects are visited according to their 'reachability'
  - Parameter: reachability threshold
- The **reachability plot** produces "valleys" and "hills"
  - Valleys → clusters
  - Hills → outliers (noise)
- Example:
  - $C_1 = \{1, 9, 2, 8, 4, 3, 7\}$; $C_2 = \{5, 6\}$



Trajectory clustering

- Challenging !! Objectives:
  - Cluster trajectories w.r.t. similarity
  - Eventually, detect outliers
- Issues:
  - Which similarity function?
  - Upon the entire trajectories or portions (sub-trajectories?

Could you detect clusters? outliers?

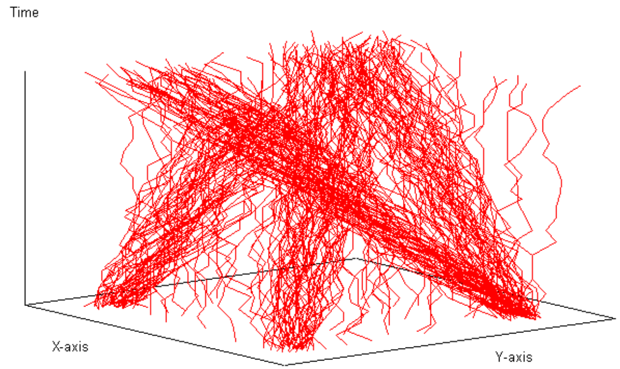
- State-of-the-art:
  - Clustering on entire trajectories: **T-OPTICS** (Nanni & Pedreschi, 2006)
  - Clustering on sub-trajectories: **TraClus** (Lee et al. 2007); **S^{2}T-Clustering** (Pelekis et al. 2017a; 2017b)

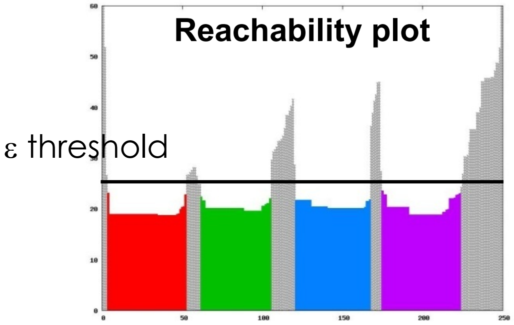
T-OPTICS (Trajectory OPTICS)

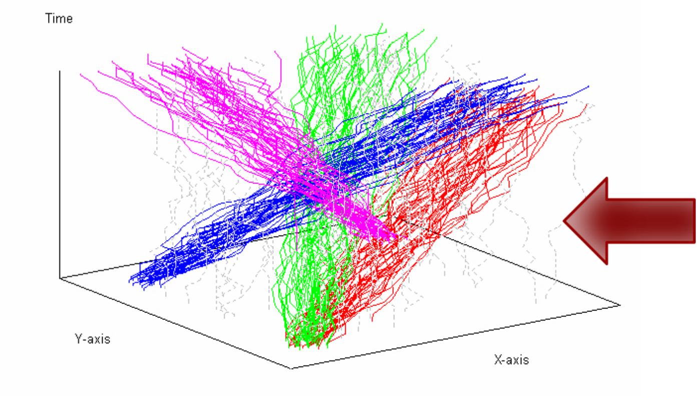
- Builds upon OPTICS method and DISSIM distance function

$$DISSIM(R,S) = \int_{t_1}^{t_n} L_2(R(t),S(t))dt$$

  - Recall that DISSIM is a metric → indexing is allowed

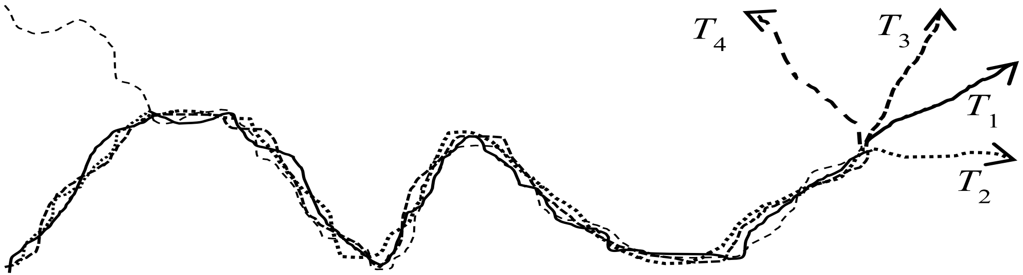


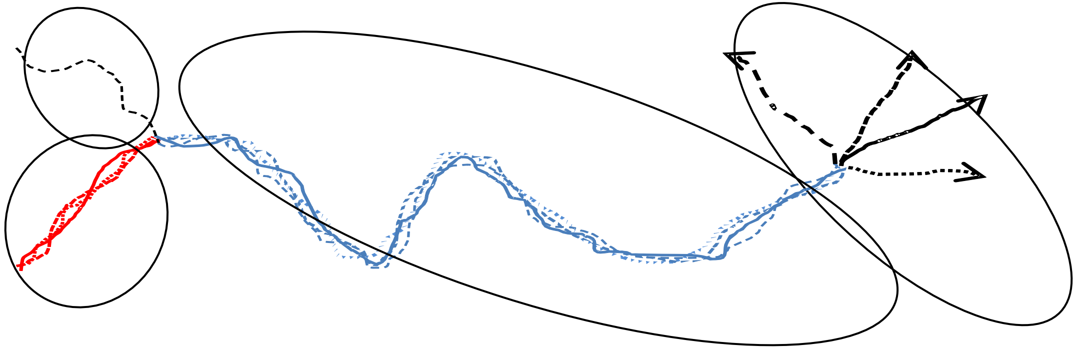




Sub-trajectory clustering

- Motivation: how many clusters (and outliers) are formed by trajectories T_1 … T_4?
  - one (zero)? zero (four)?
- What if we work at sub-trajectory level?
- Challenge: how do we detect the appropriate sub-trajectories?

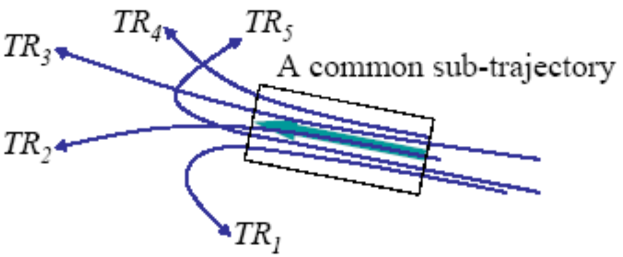


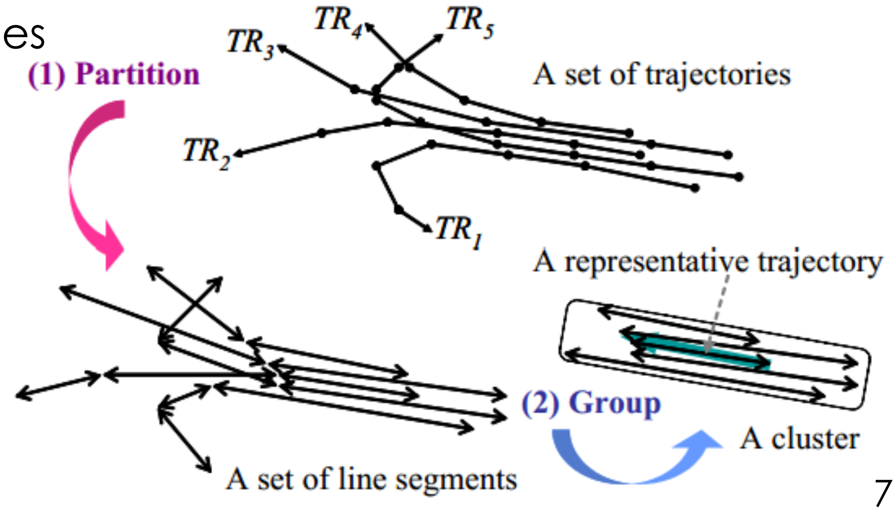


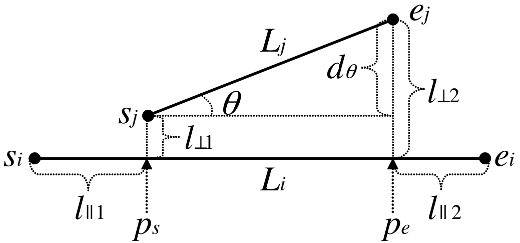
TraClus (Trajectory Clustering)

- Discovers portions (sub-trajectories) of a trajectory wrt. **homogeneity in movement**

- TraClus works in two phases:
  - **Partition** trajectories in sub-trajectories
  - **Group** sub-trajectories together
    - Recall TraClus distance function (discussed in Part I.Similarity)







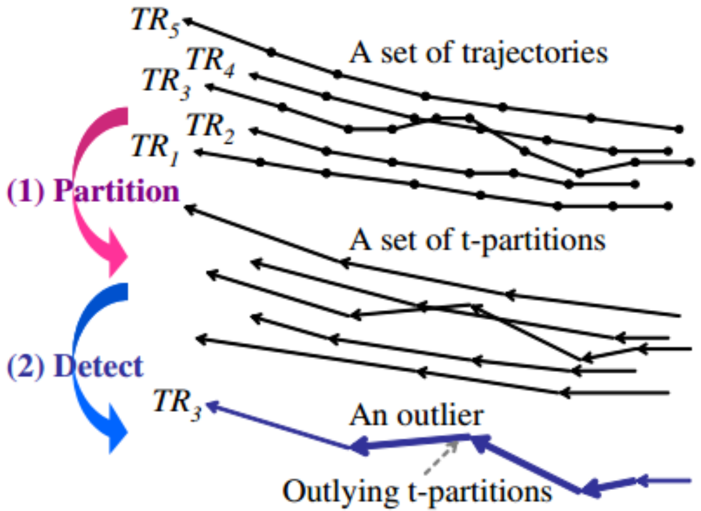
$$d_\perp = \frac{l_{\perp1}^2 + l_{\perp2}^2}{l_{\perp1} + l_{\perp2}}$$

$$d_\parallel = \mathrm{MIN}(l_{\parallel1}, l_{\parallel2})$$

$$d_\theta = \|L_j\| \times \sin(\theta)$$

TraOD (Trajectory Outlier Detection)

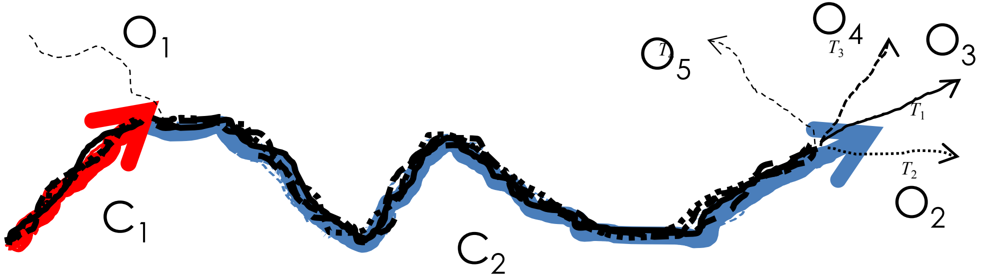
- TraCLus methodology can be exploited for **outlier detection**
- **TraOD** (Lee et al. 2008) works in two phases:
  - **Partition**: trajectories are segmented into t-partitions (sub-trajectories); recall TraClus
  - **Detect**: a trajectory is considered outlier if it contains a sufficient number of outlying t-partitions

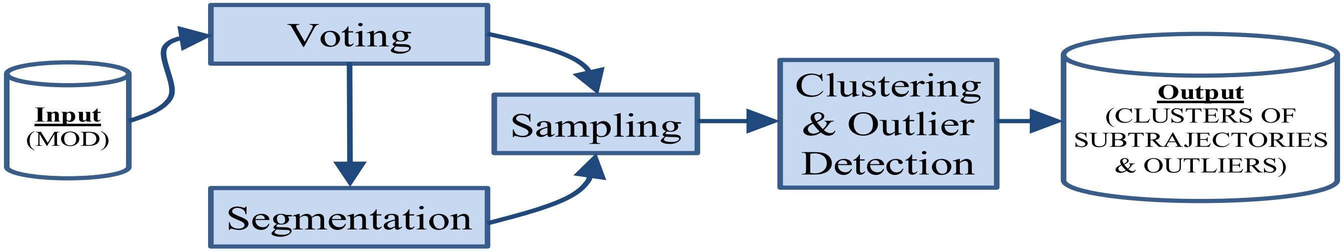


S^{2}T-Clustering (Sampling-based Sub-Trajectory Clustering)

A three-step process:

- neighborhood-aware trajectory **segmentation** (via a **voting** process)
- sub-trajectory **sampling**
- sub-trajectory **clustering** and **outlier detection**





Input
(MOD)

Voting

Segmentation

Sampling

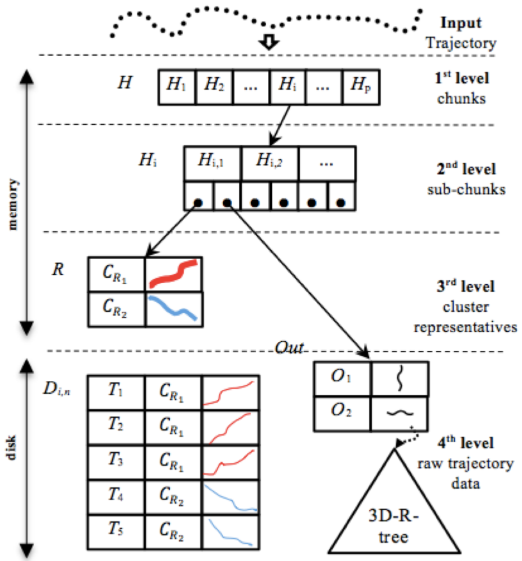
Clustering & Outlier Detection

Output
(CLUSTERS OF SUBTRAJECTORIES & OUTLIERS)

image sources: (Pelekis et al. 2017a)

S^{2}T-Clustering (cont.)

- **ReTra index** (Pelekis et al. 2017b)
  - Maintaining an index of clusters over sub-trajectories (and outliers)



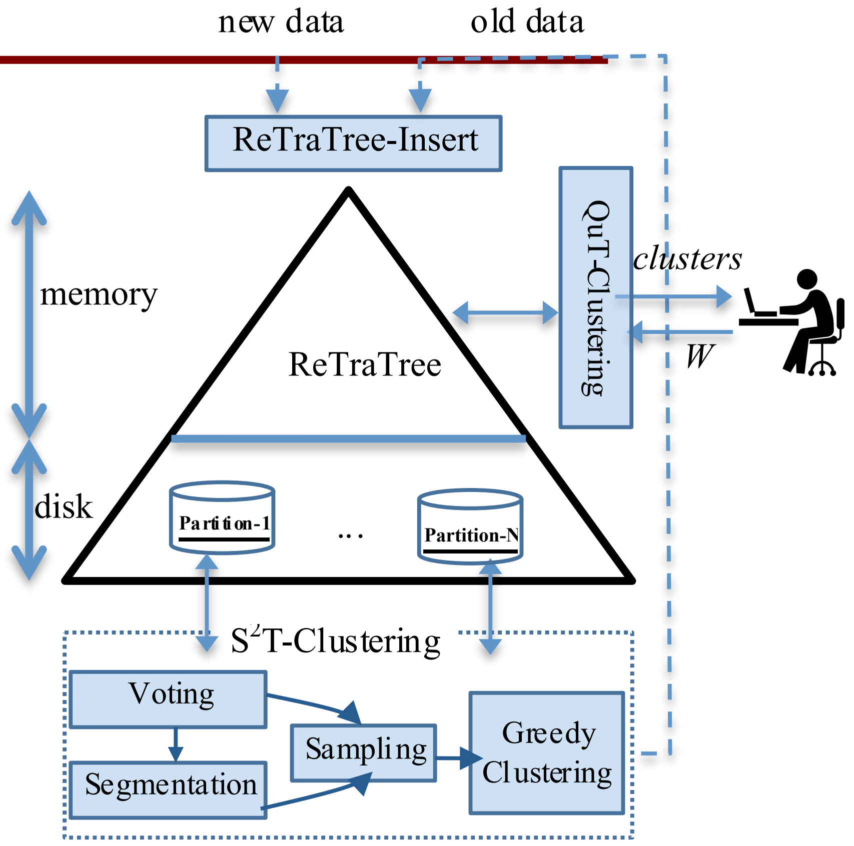
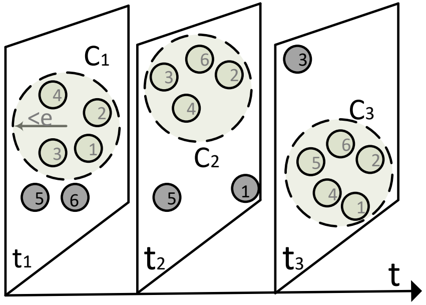
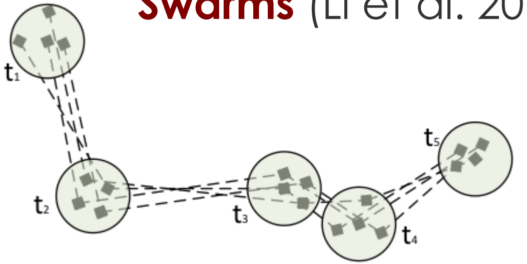


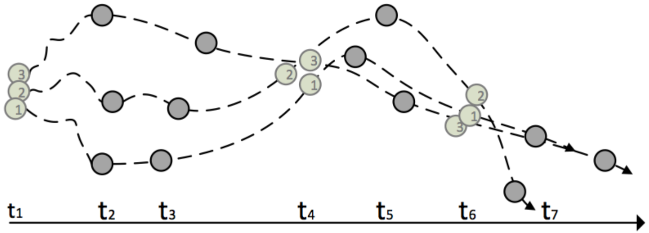
image sources: (Pelekis et al. 2017b)

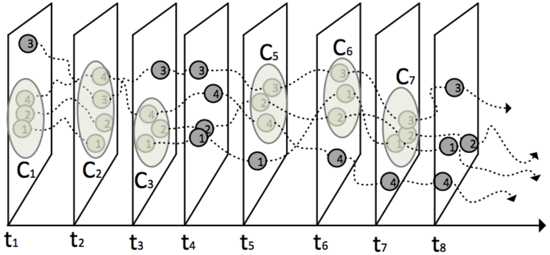
Discovering group patterns

- Several variants
  - Spherical-like clustering: **Flocks** (Laube et al. 2005; Gudmundsson & van Kreveld, 2006)
  - Density-based clustering: **Convoys** (Jeung et al. 2008); **Swarms** (Li et al. 2010), etc.



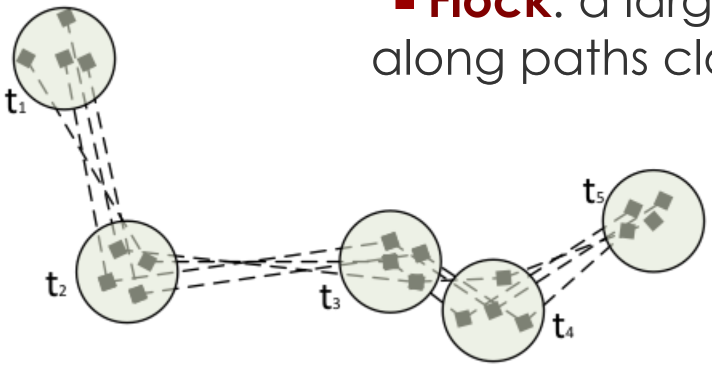






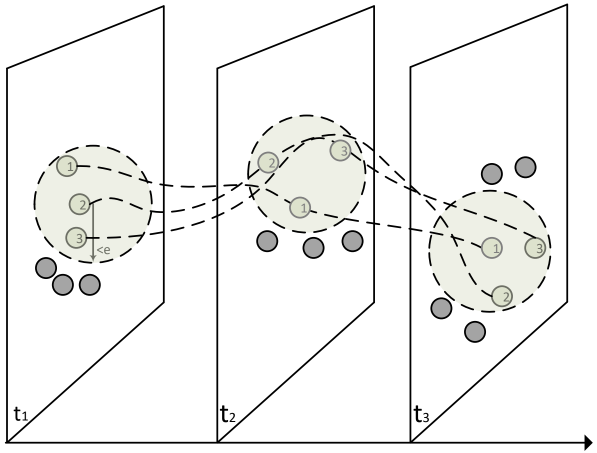
- Note: they work on time-aligned location sequences
  - cf. fixed re-sampling preprocessing task (part II)

Flocks and variants



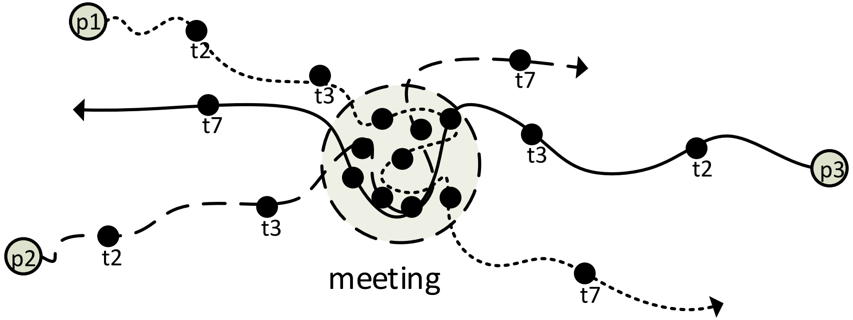
- **Flock**: a large enough subset of objects moving along paths close to each other for a certain time
  - Circular cluster

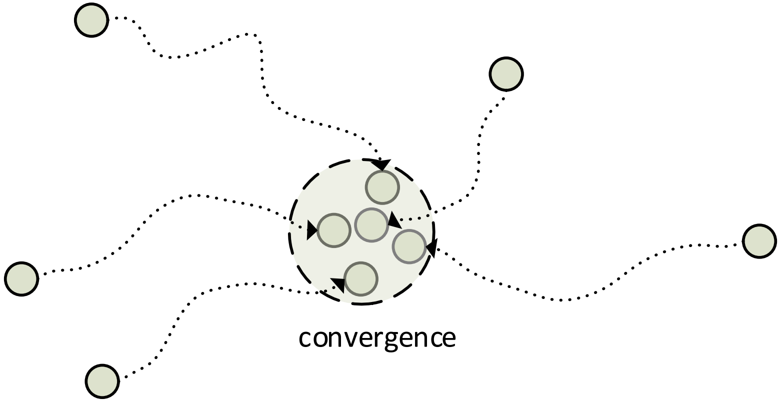
- Side-effect: the **lossy-flock problem**



Flocks and variants (cont.)

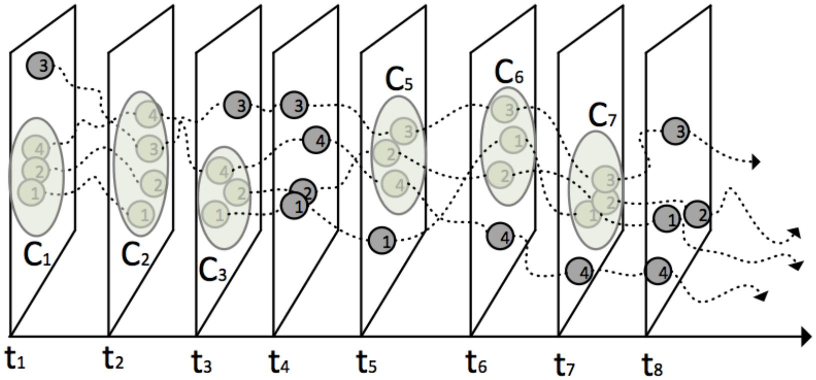
- Interesting problems arise over the flock concept:
  - Identify long flock patterns (**top-k longest flock pattern discovery**)
  - Discover **meetings** (fixed- vs. varying- versions)
  - Discover **convergences**
  - Discover **leaders** and **followers**



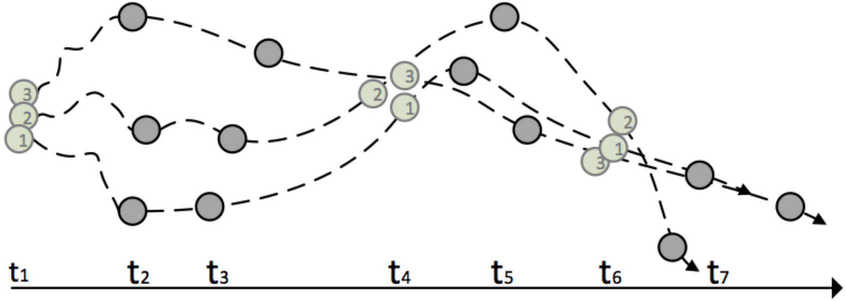


Convoys vs. Swarms

- **Convoy**: a group of objects with cardinality at least m, which are density-connected with respect to a distance threshold e, during at least k timepoints
  - Timestamps are required to be consecutive

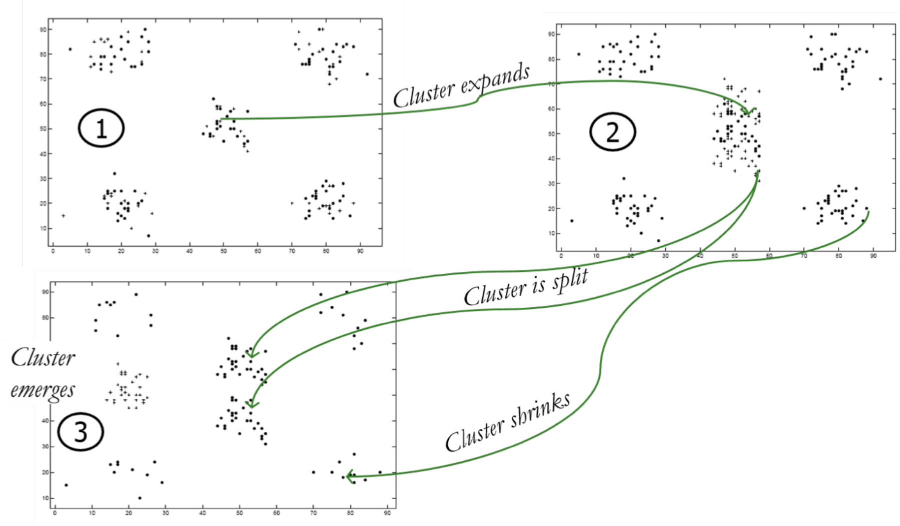


- **Swarm**: a group of objects with cardinality at least m, that are part of the same cluster, during at least k timepoints
  - Timestamps are not required to be consecutive



Cluster evolution

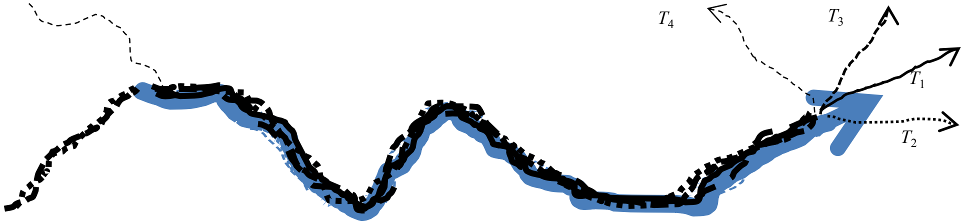
- Clusters may evolve with time (Spiliopoulou et al. 2006)*
  - A cluster may **expand** or **shrink**
  - A cluster may be **split** in two or more
  - A cluster may be **absorbed** by another cluster
  - Two or more clusters may be **merged** to a new cluster,
  - etc.



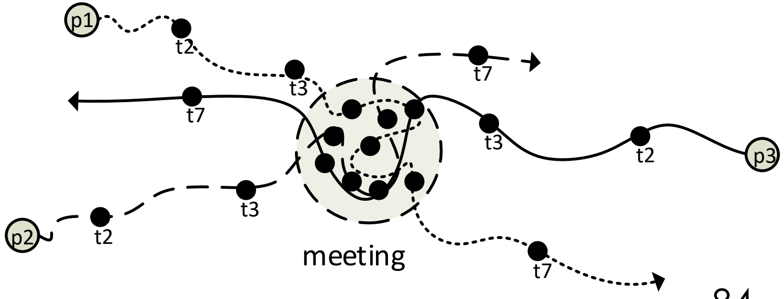
* Applicable to mobility data, though originally proposed for use in other domain (document clustering)

Frequent pattern mining

- Technical objective: identify ‘frequent’ or ‘popular’ patterns
  - Patterns could be routes (hot paths, etc.) or places (hot spots, etc.)



- Approaches:
  - techniques that identify regularities in the behavior of a single user, e.g. **Periodic patterns** (Cao et al. 2007)
  - techniques that reveal collective sequential behavior of a set of users, e.g. **T-Patterns** (Giannotti et al. 2007)

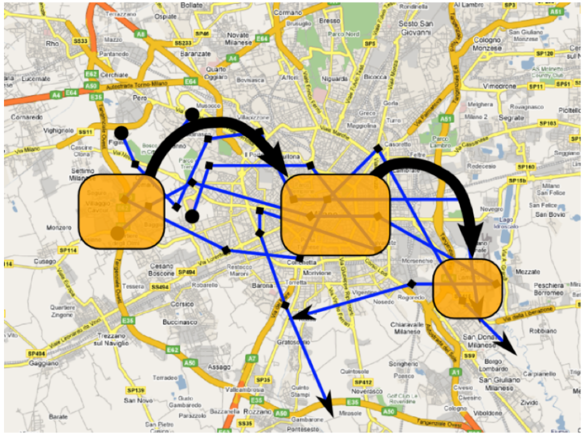


T-Pattern (Trajectory pattern)

- A **T-Pattern** is a pair (**s**, **α**):
  - **s** = <(x_i,y_i)> is a sequence of locations
  - **α** = <α_i> are the respective transition times (annotations)

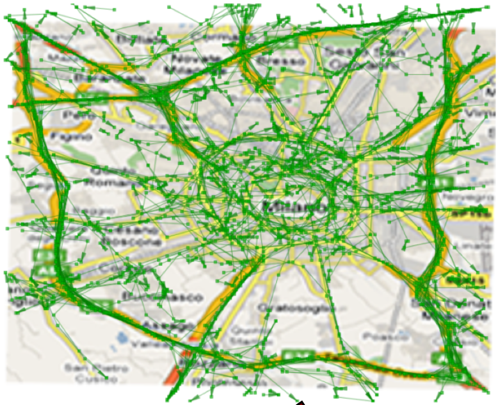
  also written as:

$$(x_0, y_0) \xrightarrow{\alpha_1} (x_1, y_1) \xrightarrow{\alpha_2} \cdots \xrightarrow{\alpha_k} (x_k, y_k)$$



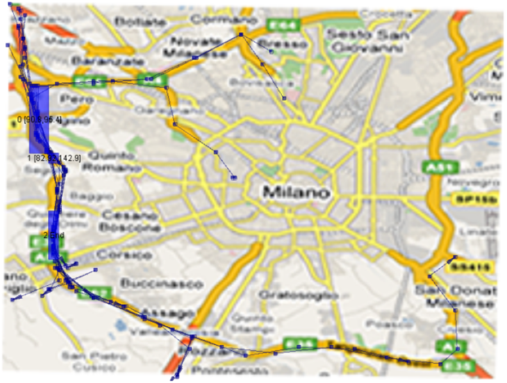
- A T-pattern T_p **occurs** in a trajectory T if T contains a sub-sequence S, such that:
  - each point in T_p is **close** to a point in S (spatial closeness)
  - transition times in T_p are **similar** to those in S (temporal closeness)

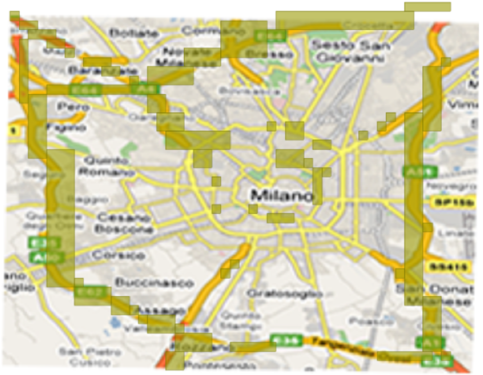
T-Pattern discovery



Input:
Trajectory
Dataset

Output: T-Patterns

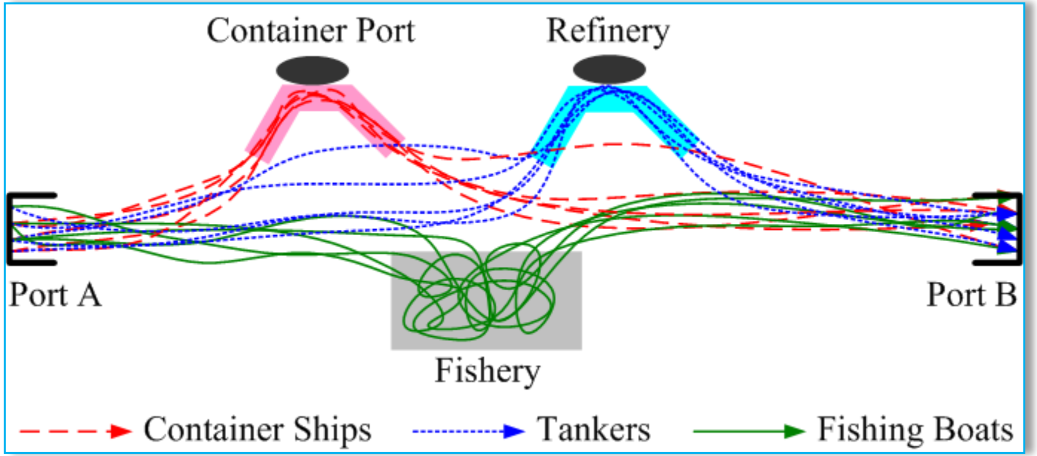




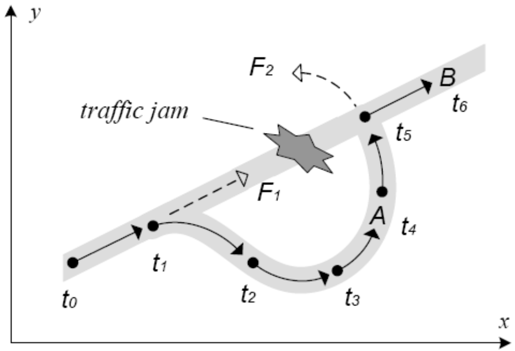
Intermediate result:
Regions of Interest

Classification

- **Classification** aims to predict the class label of a moving object based on its features. State-of-the-art: TraClass (Lee et al. 2008b)
- **TraClass** (Trajectory Classification) works in three phases:
  1. Partitions trajectories based on their shapes (using a TraClus variant)
  2. Discovers regions that contain sub-trajectories mostly from the same class (region-based clustering)
  3. Discovers common movement patterns for each class of sub-trajectories (trajectory-based clustering)

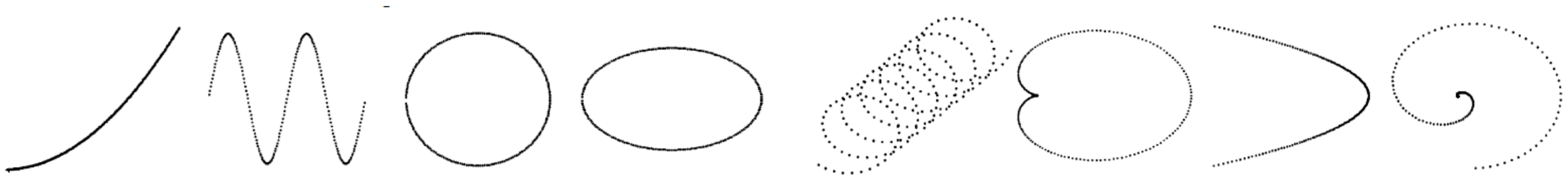


Prediction



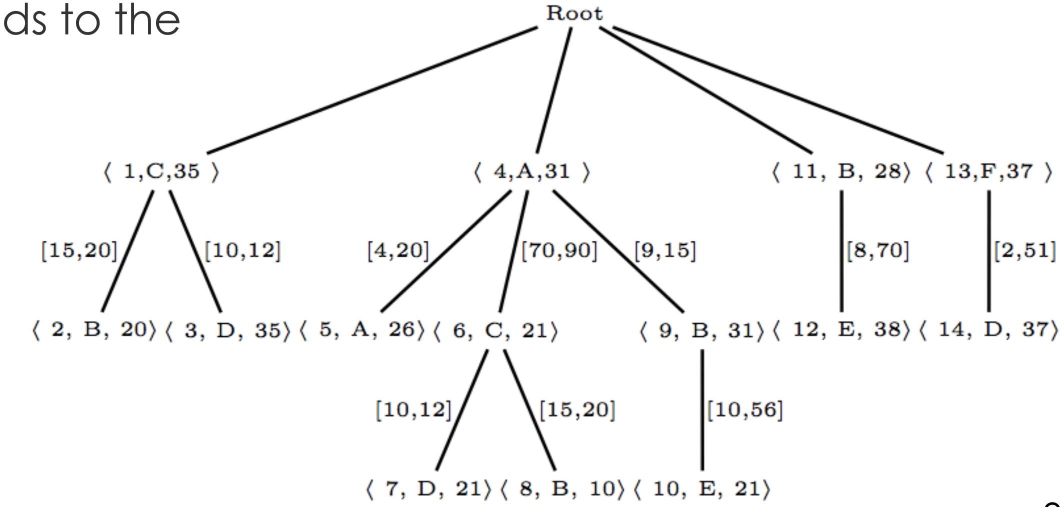
- **Prediction** aims to predict the future location(s) of (or even the entire trajectory to be followed by) a moving object.

- Two main approaches: **Formula-** vs. **Pattern-based** prediction
  - Motion function models, e.g. RMF (Tao et al. 2004)
  - vs. patterns built upon the history, e.g. Sequential patterns (Monreale et al. 2009), Personal profiles (Trasarti et al. 2017)
  - Recent survey of 50+ methods: (Georgiou et al. 2018)



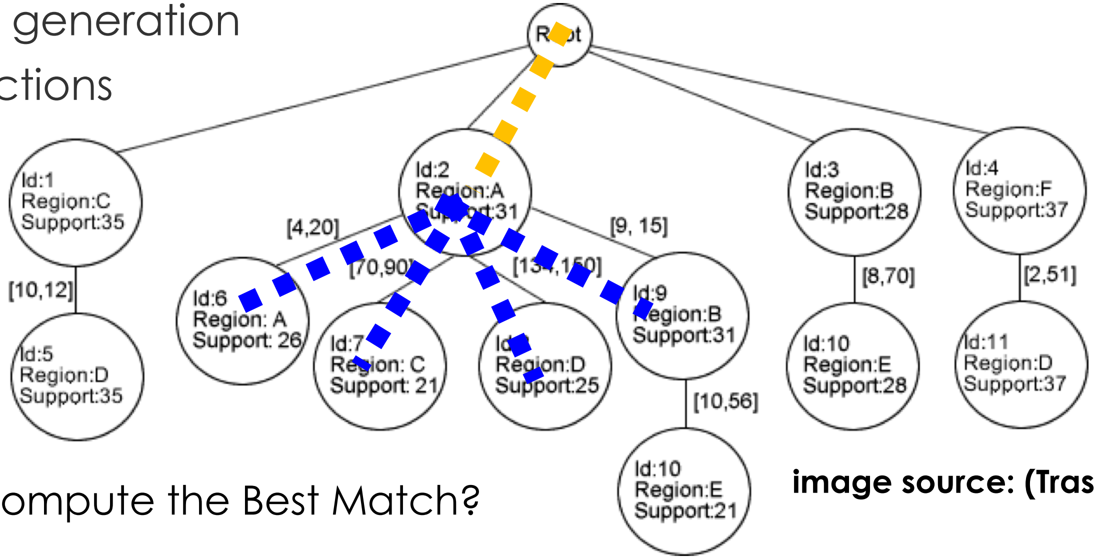
Prediction (cont.)

- **WhereNext** (Monreale et al. 2009) builds upon the T-pattern concept: extracts a set of T-patterns and builds a T-pattern tree
  - the best path is found for a given trajectory
  - the predicted future location of the trajectory is the region that corresponds to the final node of the best path
- Example …



Prediction (cont.)

- **WhereNext** (cont.)
- Having a new trajectory, the method follows 3 steps:
  - Search for best match
  - Candidates generation
  - Make predictions



Best Match

Candidates

Predictions

Issue: how to compute the Best Match?

image source: (Trasarti et al. 2009)

Prediction (cont.)

- **MyWay** (Trasarti et al. 2017) maintains a Personal Mobility Data Store (PMDS) per participating person
  - How a person is moving?
    - According to his/her past movement patterns
  - What if the personal datastore is not adequate?
    - Look into the collective knowledge base
- 3 predictors: personal (red), collective (blue), hybrid (green)

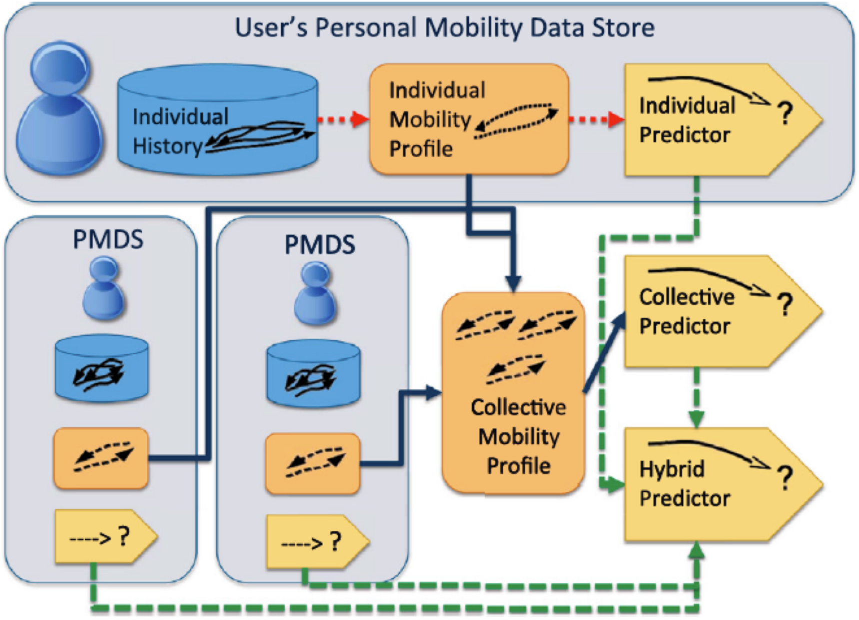
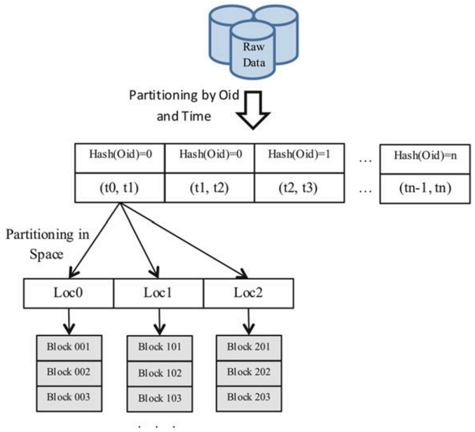
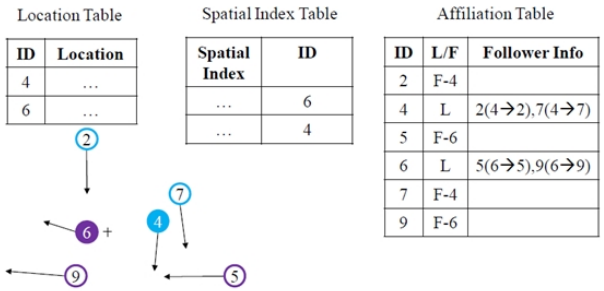


image source: kdd.isti.cnr.it

What's new in big MDA?

- Mobility data applications: historical vs. real-time
  - Offline management of archived past data
  - Online management of streaming current (and recent past) data
- Queries and operations of interest: spatiotemporal range, NN, etc.
- **Offline** vs. **Online MDA**. Examples (resp.):
  - **CloST** (Tan et al. 2012)
  - **MOIST** (Jiang et al. 2012)





Location Table

ID	Location
4	…
6	…

Spatial Index Table

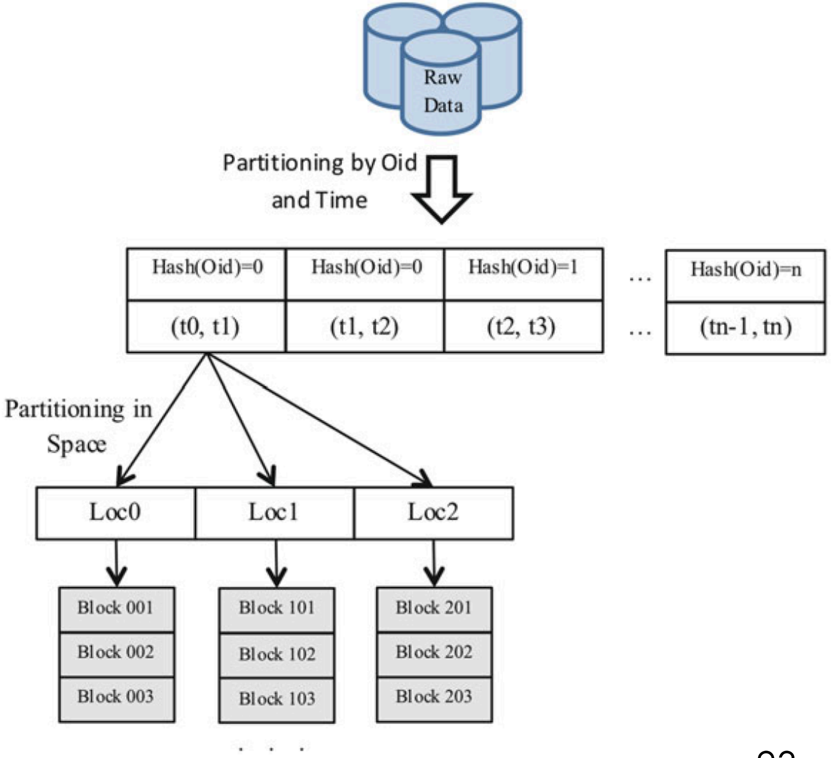
Spatial Index	ID
…	6
…	4

Affiliation Table

ID	L/F	Follower Info
2	F-4	
4	L	2(4→2),7(4→7)
5	F-6	
6	L	5(6→5),9(6→9)
7	F-4	
9	F-6	

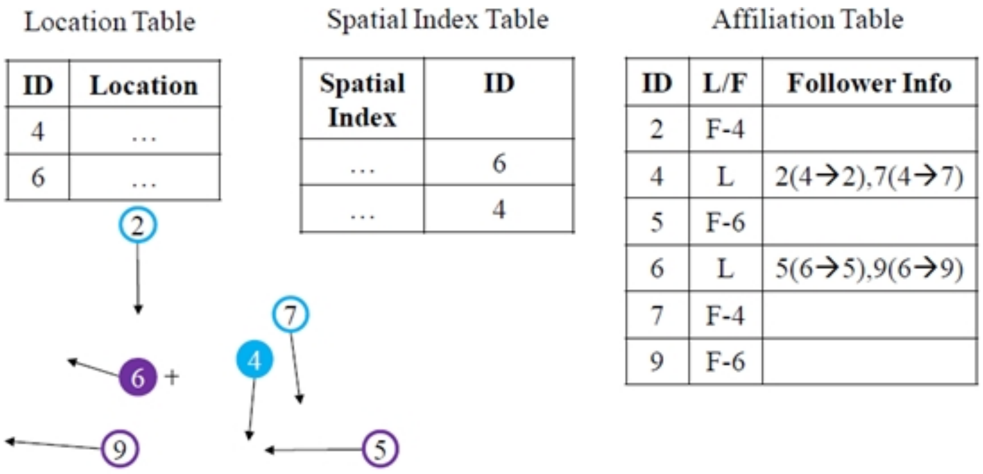
Offline data analytics: CloST

- **CloST**: a scalable spatiotemporal data storage system that supports data analytics using Hadoop
- Two types of queries are ssupported
  - single-object spatiotemporal range queries
  - all-objects spatiotemporal range queries
- Three-level hierarchical partitioning:
  1. partitions according to hash values of the object ids and coarse ranges of time
  2. partitions according to a spatial index on the location attribute
  3. actual data



Online data analytics: MOIST

- **MOIST** (Moving Object Indexer with School Tracking)
- Methodology
  - The space is divided into cells of different resolutions and a space filling curve is constructed
  - Nearby and of similar moving behavior objects are grouped into one school
  - The leader object is tracked, distances between the followers and the leader are recorded
  - Aged data are flushed onto disk so that the history of objects be analyzed



Location Table

ID	Location
4	…
6	…

Spatial Index Table

Spatial Index	ID
…	6
…	4

Affiliation Table

ID	L/F	Follower Info
2	F-4	
4	L	2(4→2),7(4→7)
5	F-6	
6	L	5(6→5),9(6→9)
7	F-4	
9	F-6	

Summarizing part III …

- Typical lines of research in MDA include:
  - (Sub-) trajectory clustering and outlier detection
  - Detecting collective / group behavior
  - Discovering frequent patterns (routes, places, etc.)
  - Predicting the anticipated movement (or other features)
- Trends in this area include:
  - Semantic- (i.e. context-) aware MDA (clustering, frequent pattern mining, prediction, etc.)
  - MDA under the Big Data prism
  - Incremental (online) MDA

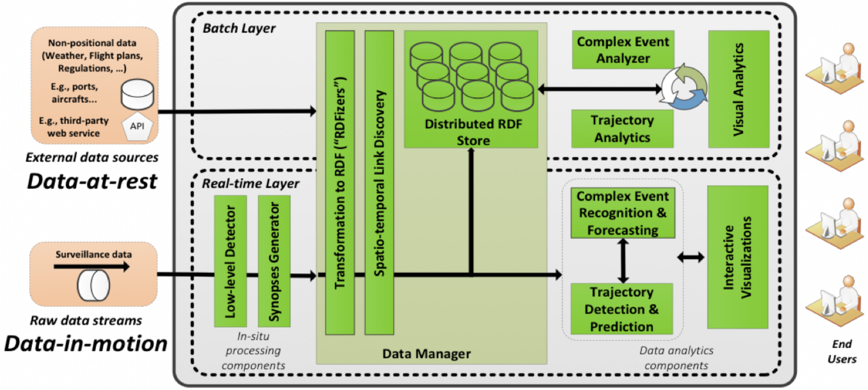
Part IV:
Summary – the Future

"As you set out for Ithaca, hope the voyage is a long one, full of adventure, full of discovery…"
Constantine Cavafy

An real-world MDA example

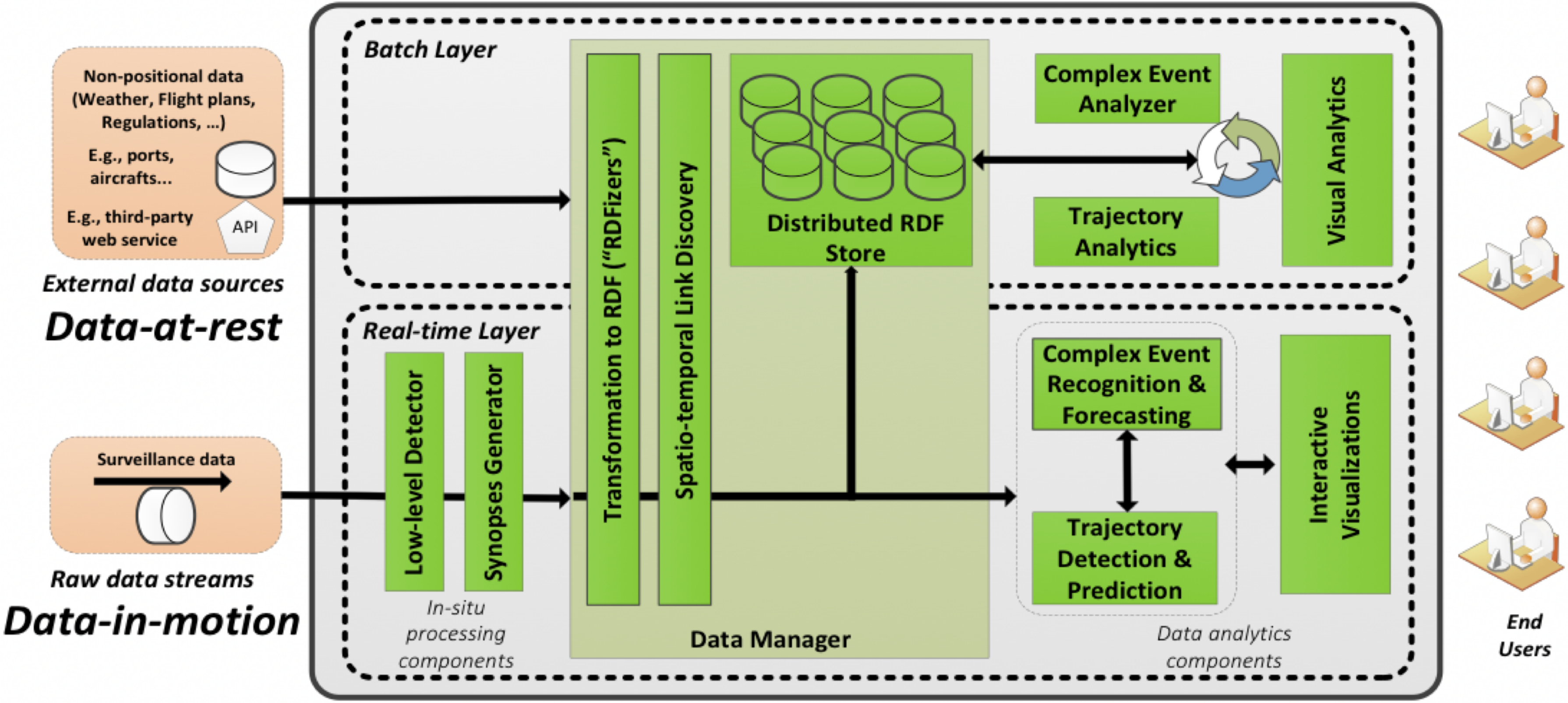
- The problem: **data-driven aircraft trajectory prediction** *
  - … instead of model-based prediction
  - Data sources available include aircraft surveillance data (from multiple sources), flight plans, air space zones, weather info, etc.
- **datAcron** system architecture (Claramunt et al. 2017; Vouros et al. 2018; Santipantakis et al. 2018)

* For the following slides, credits to all datAcron partners, especially BRTE and CRIDA (aviation use case)

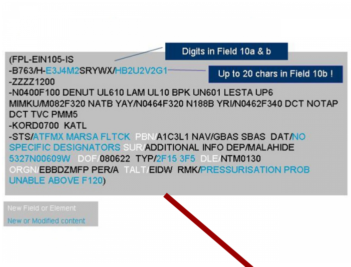


Non-positional data (Weather, Flight plans, Regulations, …)
E.g., ports, aircrafts...
E.g., third-party web service
API
External data sources
Data-at-rest
Surveillance data
Raw data streams
Data-in-motion
Batch Layer
Real-time Layer
Low-level Detector
Synopses Generator
In-situ processing components
Transformation to RDF ("RDFizers")
Spatio-temporal Link Discovery
Distributed RDF Store
Data Manager
Complex Event Analyzer
Trajectory Analytics
Visual Analytics
Complex Event Recognition & Forecasting
Trajectory Detection & Prediction
Interactive Visualizations
Data analytics components
End Users

datAcron system architecture


Non-positional data (Weather, Flight plans, Regulations, …)
E.g., ports, aircrafts…
E.g., third-party web service
API
External data sources
Data-at-rest
Surveillance data
Raw data streams
Data-in-motion
Batch Layer
Real-time Layer
Transformation to RDF ("RDFizers")
Spatio-temporal Link Discovery
Distributed RDF Store
Data Manager
Complex Event Analyzer
Trajectory Analytics
Visual Analytics
Low-level Detector
Synopses Generator
In-situ processing components
Complex Event Recognition & Forecasting
Trajectory Detection & Prediction
Interactive Visualizations
Data analytics components
End Users

Trajectory prediction (model-based)

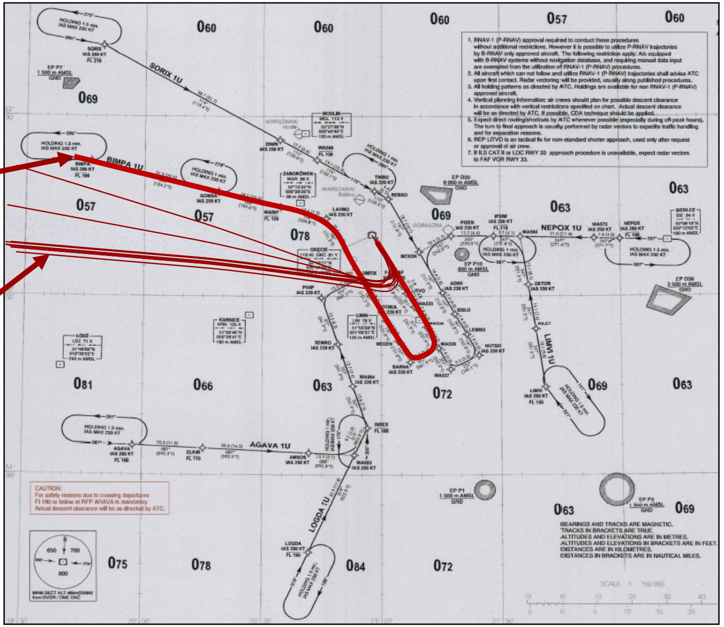


```
(FPL-EIN105-IS
-B763/H-E3J4M2SRYWX/HB2U2V2G1
-ZZZZ1200
-N0400F100 DENUT UL610 LAM UL10 BPK UN601 LESTA UP6
MIMKU/M082F320 NATB YAY/N0464F320 N188B YRI/N0462F340 DCT NOTAP
DCT TVC PMM5
-KORD0700  KATL
-STS/ATFMX MARSA FLTCK  PBN/A1C3L1 NAV/GBAS SBAS  DAT/NO
SPECIFIC DESIGNATORS SUR/ADDITIONAL INFO DEP/MALAHIDE
5327N00609W   DOF/080622  TYP/2F15 3F5  DLE/NTM0130
ORGN/EBBDZMFP PER/A  TALT/EIDW  RMK/PRESSURISATION PROB
UNABLE ABOVE F120)
```

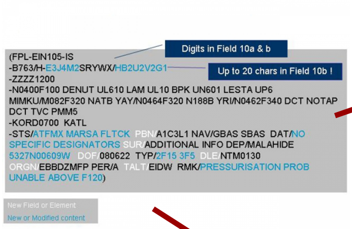
Model-based Trajectory prediction

≠

Surveillance Data



Trajectory prediction (data-driven)



Digits in Field 10a & b

Up to 20 chars in Field 10b !

```
(FPL-EIN105-IS
-B763/H-E3J4M2SRYWX/HB2U2V2G1
-ZZZZ1200
-N0400F100 DENUT UL610 LAM UL10 BPK UN601 LESTA UP6
MIMKU/M082F320 NATB YAY/N0464F320 N188B YRI/N0462F340 DCT NOTAP
DCT TVC PMM5
-KORD0700  KATL
-STS/ATFMX MARSA FLTCK  PBN/A1C3L1 NAV/GBAS SBAS  DAT/NO
SPECIFIC DESIGNATORS SUR/ADDITIONAL INFO DEP/MALAHIDE
5327N00609W   DOF/080622  TYP/2F15 3F5  DLE/NTM0130
ORGN/EBBDZMFP PER/A  TALT/EIDW  RMK/PRESSURISATION PROB
UNABLE ABOVE F120)
```

New Field or Element

New or Modified content

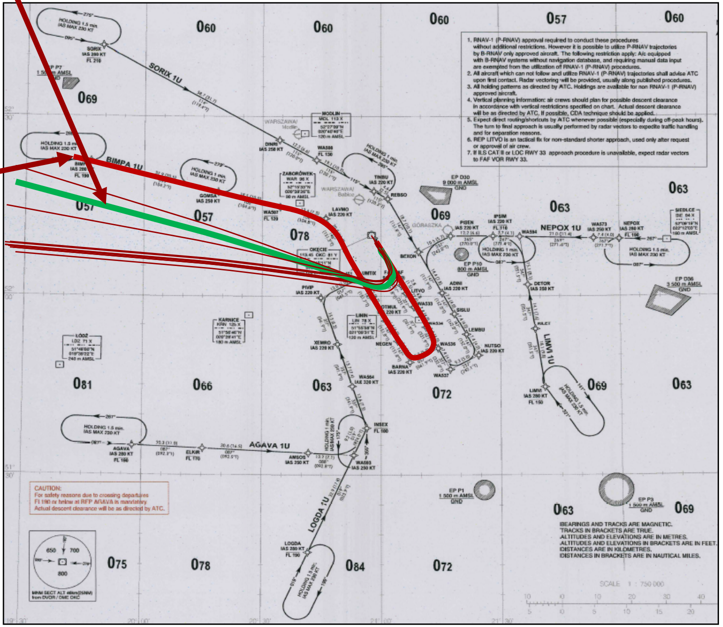
DatACRON Trajectory prediction

Historical data + context data

Model-based Trajectory prediction

≠

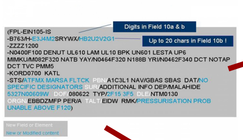
Surveillance Data



Trajectory prediction (data-driven)

- Formally:
  Given a Flight Plan, predict the **actual trajectory** of the corresponding flight, w.r.t. information that really matters
  - Current and forecasted weather info,
  - Predicted air-space traffic,
  - Aircraft type, etc.

sequence of (lat-, lon-, alt-, t-) tuples



Digits in Field 10a & b

Up to 20 chars in Field 10b !

```
(FPL-EIN105-IS
-B763/H-E3J4M2SRYWX/HB2U2V2G1
-ZZZZ1200
-N0400F100 DENUT UL610 LAM UL10 BPK UN601 LESTA UP6
MIMKU/M082F320 NATB YAY/N0464F320 N188B YRI/N0462F340 DCT NOTAP
DCT TVC PMM5
-KORD0700  KATL
-STS/ATFMX MARSA FLTCK  PBN/A1C3L1 NAV/GBAS SBAS  DAT/NO
SPECIFIC DESIGNATORS SUR/ADDITIONAL INFO DEP/MALAHIDE
5327N00609W   DOF/080622  TYP/2F15 3F5  DLE/NTM0130
ORGN/EBBDZMFP PER/A  TALT/EIDW  RMK/PRESSURISATION PROB
UNABLE ABOVE F120)
```

New Field or Element

New or Modified content

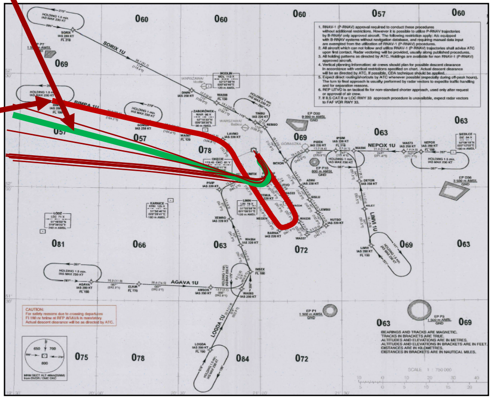
DatACRON Trajectory prediction

Historical data + context data

Model-based Trajectory prediction

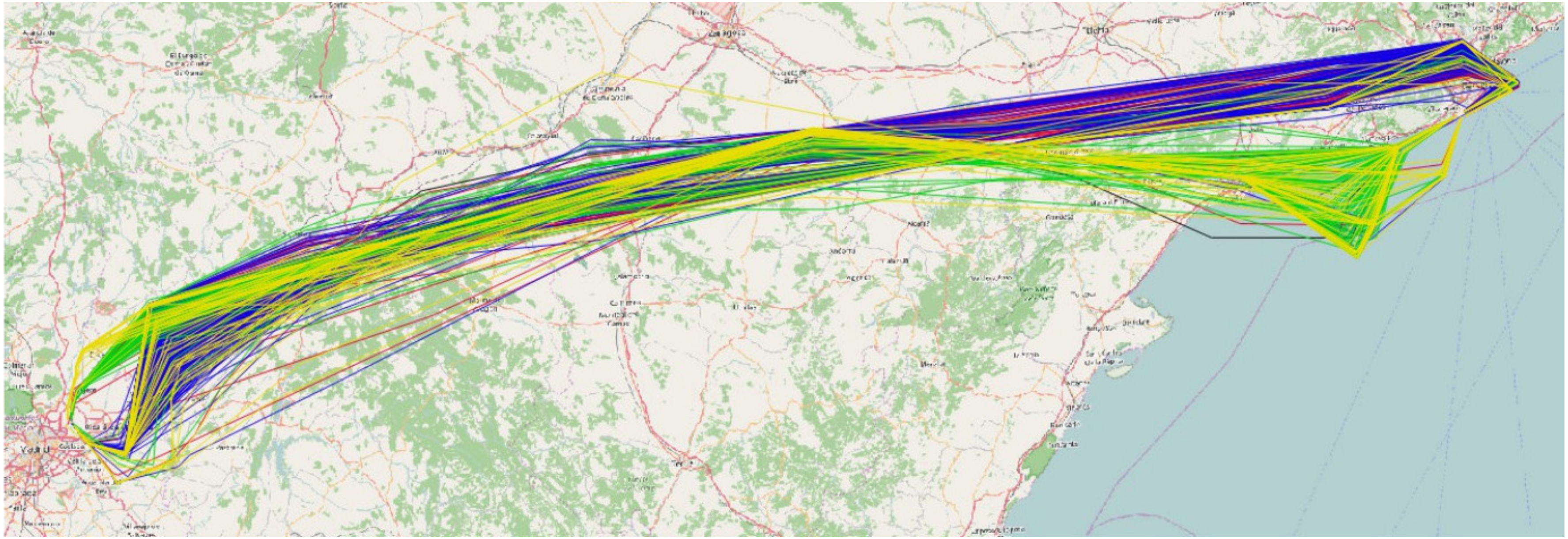
≠

Surveillance Data



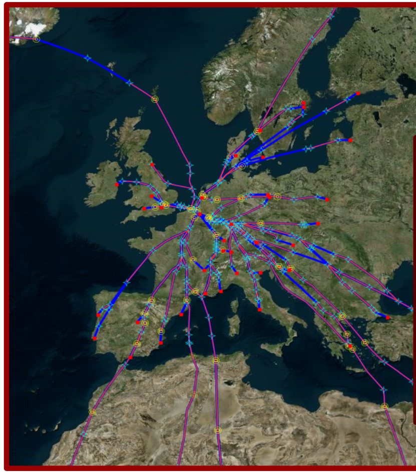
Experimental dataset

- Spain (Madrid-Barcelona flights), April 2016



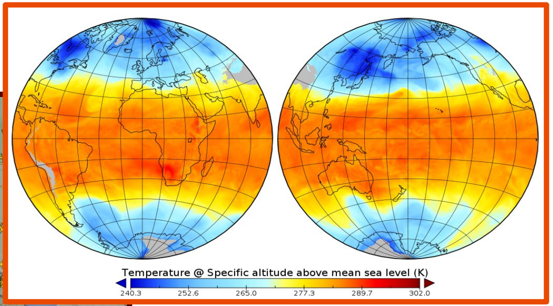
Data sources

- DataSets:
  - Initial Flight Plans
  - Actual trajectories from Surveillance
  - Weather live data and forecasts
  - Other context data



EBBR Outbound Fligh Plans for a 2 hour timeslot

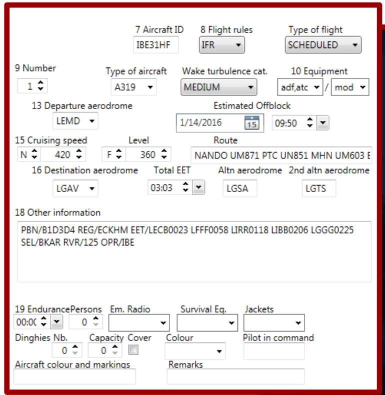
ADS-B Surveillance traffic



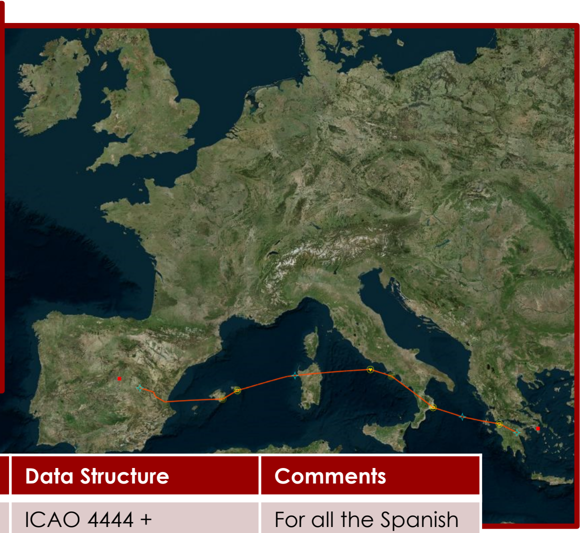
European Sector static information

Data sources – Flight plan

- Specified information provided to air traffic services units, relative to an intended flight or portion of a flight of an aircraft.
- Standards and data format
  - ICAO 4444 + amendments
  - NM 19.0.0 - NOP/B2B Reference
  - Manuals – FlightServices
  - FIXM



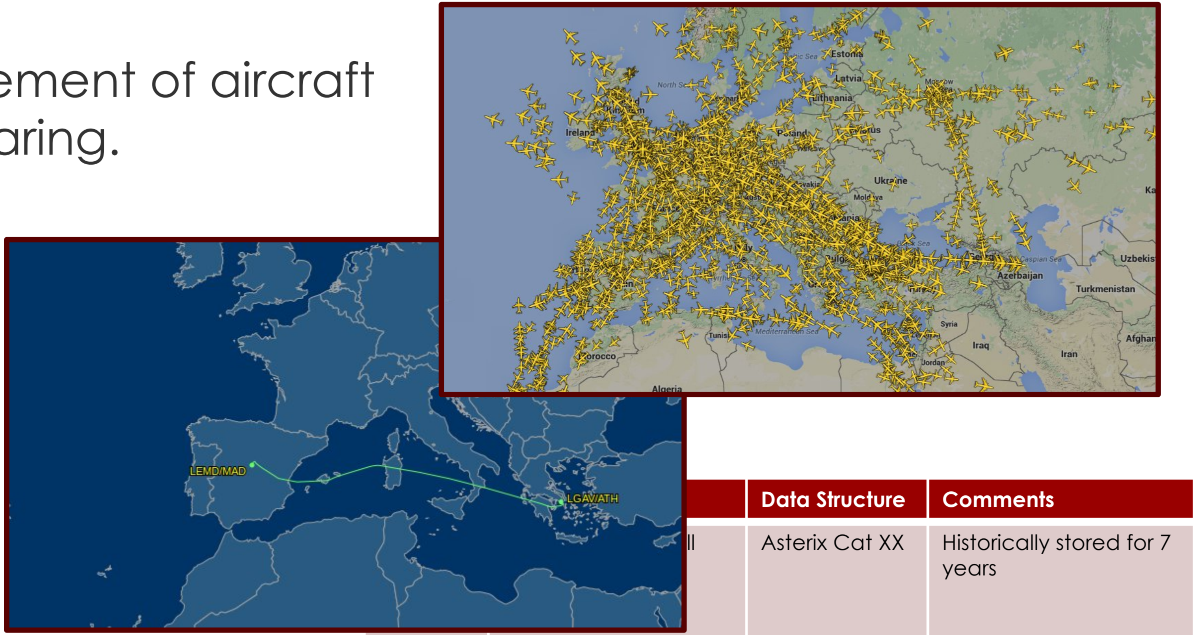
7 Aircraft ID: IBE31HF
8 Flight rules: IFR
Type of flight: SCHEDULED
9 Number: 1
Type of aircraft: A319
Wake turbulence cat.: MEDIUM
10 Equipment: adf,atc / mod
13 Departure aerodrome: LEMD
Estimated Offblock: 1/14/2016 09:50
15 Cruising speed: N 420
Level: F 360
Route: NANDO UM871 PTC UN851 MHN UM603 E
16 Destination aerodrome: LGAV
Total EET: 03:03
Altn aerodrome: LGSA
2nd altn aerodrome: LGTS
18 Other information:
PBN/B1D3D4 REG/ECKHM EET/LECB0023 LFFF0058 LIRR0118 LIBB0206 LGGG0225
SEL/BKAR RVR/125 OPR/IBE
19 Endurance: 00:0(
Persons: 0
Em. Radio:
Survival Eq.:
Jackets:
Dinghies Nb.: 0
Capacity: 0
Cover: □
Colour:
Pilot in command:
Aircraft colour and markings:
Remarks:



Sources	Description	Data Structure	Comments
Spanish ATC Platform Flight Plan Data	Relevant flight messages for all the flights in Spanish airspace (Flight plan creation, deletion and major updates, sector entry, sector leave,…)	ICAO 4444 + Amendments (FPL 2012)	For all the Spanish airspace, 1 Gb/day. Historically stored for 7 years. Streaming can be emulated
Network Manager Flight Information	Flight history for inbound and outbound flights in European Airspace	NM 19.0.0 - NOP/B2B Reference Manuals – FlightServices	

Data sources – Surveillance

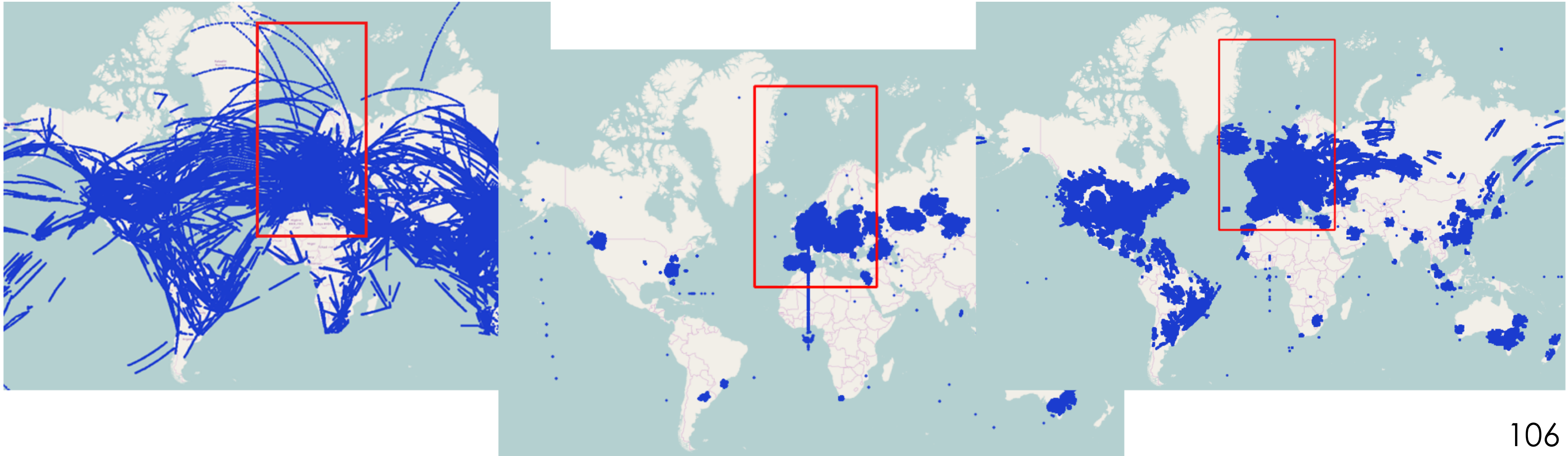
- Detection and measurement of aircraft position, range and bearing.
- Standards and data format
  - ASTERIX CATXX
  - ASDI
  - Plain ADS-B (RTCA DO-260)



		Data Structure	Comments
	ll	Asterix Cat XX	Historically stored for 7 years
ADSB	Global network of 70 ADS-B stations (53 in Europe)	DO-260 and decoded CSV text	Hundreds of flights 3D position, velocity… etc (all ADS-B message fields) each 0.5 seconds

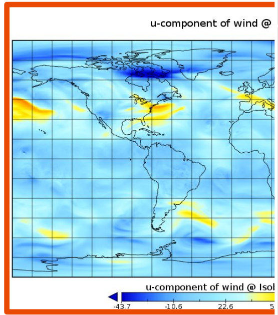
Data sources – Surveillance (cont.)

ADS-B positions provided by FlightAware (left), ADSBHun (middle), ADSBExchange (right)



Data sources – Weather

- Involving predictions and observations
- Standards and data format
  - GRIB / GRIB-2
  - netCDF
  - TAF
  - METAR

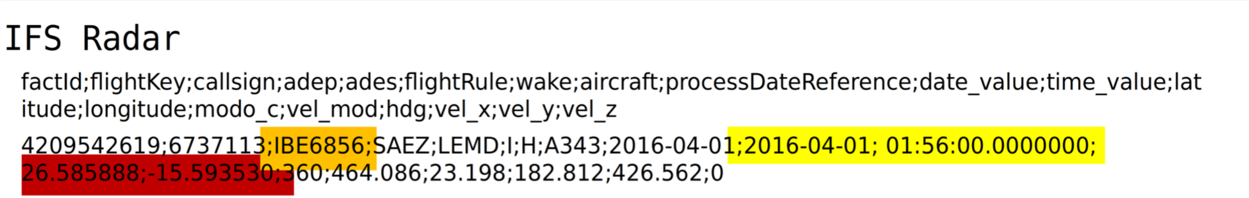


Sources	Description	Data Structure	Comments
ECMWF	Re-analyses from 1979 to date. Useful for climatological purposes	Original data: 6-hourly Analyses from 1979 to date. 0.72 degree horizontal resolution, over Surface and 37 vertical pressure levels. Climatological data: means, medians and standard deviations for all relevant variables at surface	Limited by ECMWF data Policy The Statistical variable might be daily, monthly or number of occurrences per month or… depending upon the variable type. On demand other statistical indicators can be calculated.
ECMWF	Simultaneous forecast of the same model run with slightly different initial conditions High Resolution	15 days forecasts with 3 hourly time step of 51 parallel forecasts (ensemble members). 0.25 degrees horizontal resolution, several vertical pressure levels . Two drops a day (00,12Z) Up to 10 days forecast time range and 3 hourly /hourly time step. 0.125 degrees horizontal resolution, several vertical levels both pressure and hybrid. Two drops a day (00,12Z)	Derived quantities like Ensemble means, STD, probabilities can be made available over the period and area requested. Need to decide which variable and which level make available.
	s Resolution Global	15 days forecasts with 3 hourly time step of 20 parallel forecasts (ensemble members). 0.50 degrees horizontal resolution, several vertical pressure levels . Four drops a day (00,06,12,18Z) Up to 10 days forecast time range and 3 hourly /hourly time step. 0.25 degrees horizontal resolution, several vertical levels both pressure	Derived quantities like Ensemble means, STD, probabilities can be made available over the period and area requested. Need to decide which variable and which level make available.

Trajectory Reconstruction

- Recall part II tasks:
  - Trajectory reconstruction (cleansing, summarization, etc.)
  - Fusion from different sources and trajectory enrichment

… to be performed **online**



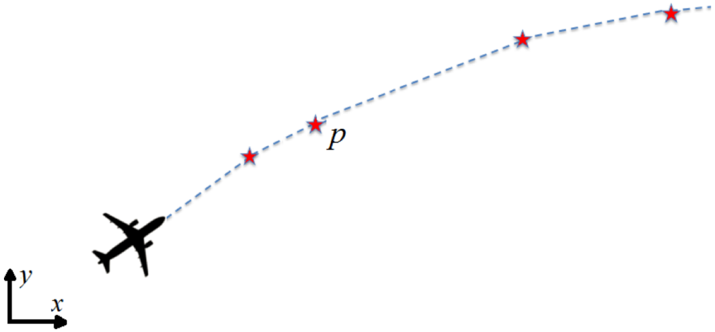
IFS Radar

```
factId;flightKey;callsign;adep;ades;flightRule;wake;aircraft;processDateReference;date_value;time_value;lat
itude;longitude;modo_c;vel_mod;hdg;vel_x;vel_y;vel_z
4209542619;6737113;IBE6856;SAEZ;LEMD;I;H;A343;2016-04-01;2016-04-01; 01:56:00.0000000;
26.585888;-15.593530;360;464.086;23.198;182.812;426.562;0
```

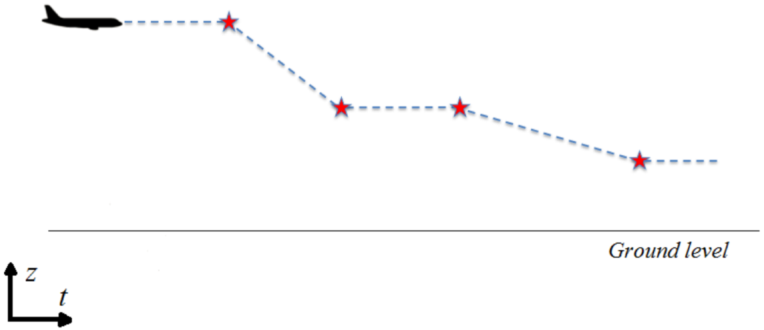
Challenges …

Trajectory Reconstruction (cont.)

Challenge 1: Identifying critical points



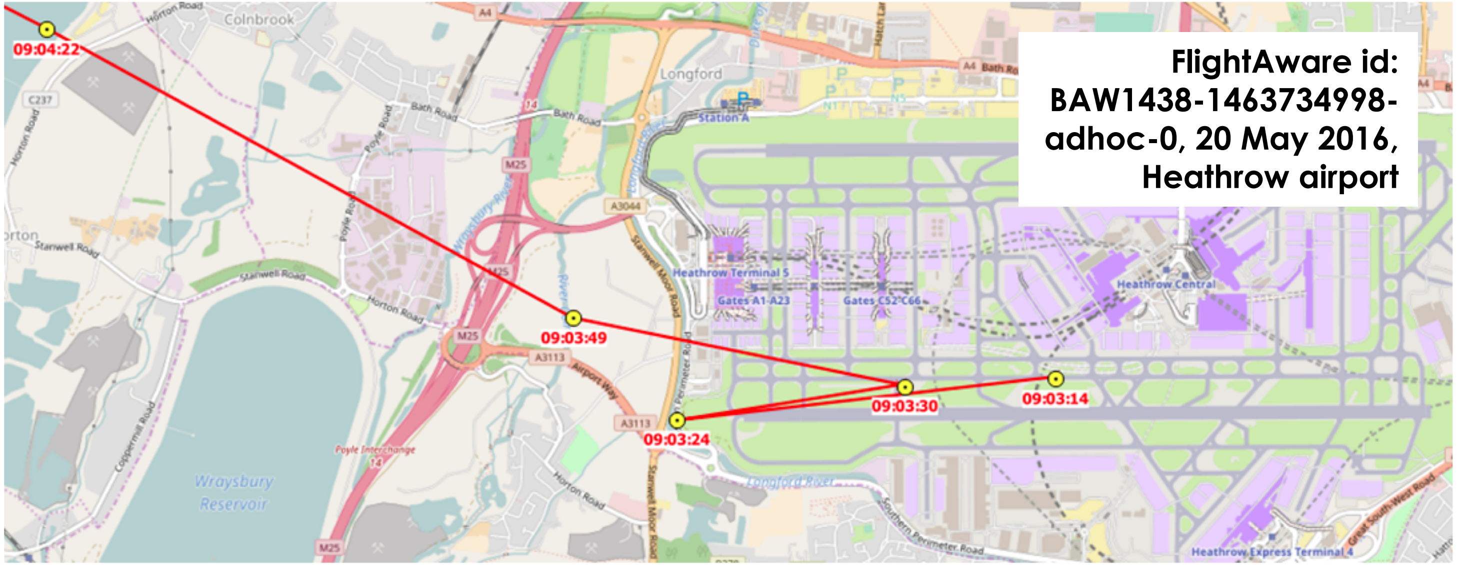
vertical view



lateral view

Trajectory Reconstruction (cont.)

Challenge 2: detect and eliminate noise



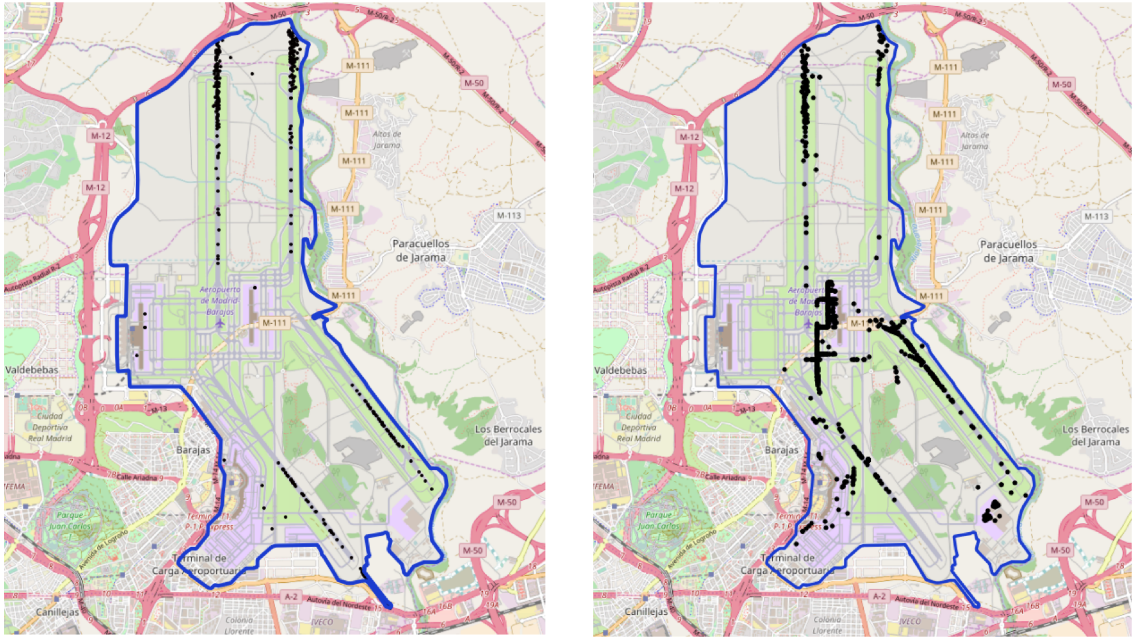
FlightAware id: BAW1438-1463734998-adhoc-0, 20 May 2016, Heathrow airport

Noise in ADS-B Flight Aware positions during takeoff of an aircraft (see timestamps)

Trajectory Reconstruction (cont.)

Challenge 3: fuse information from different sources

Samples of ADS-B positions at Madrid airport - FlightAware (left) vs. ADSBHub (right)

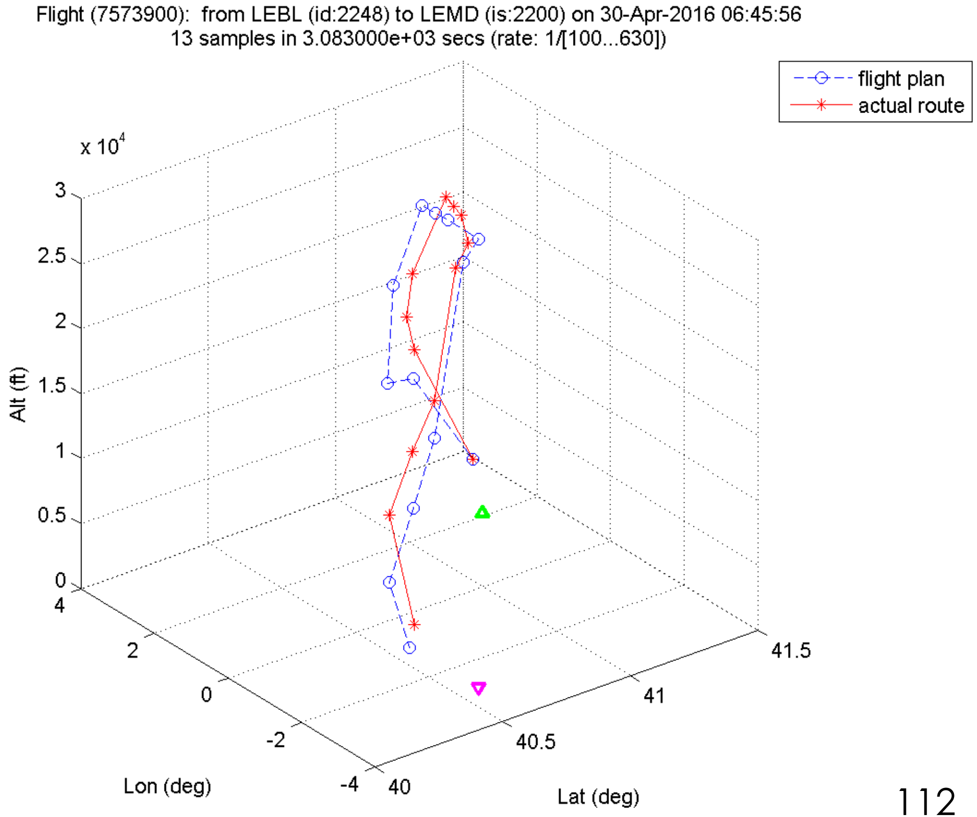


Data-driven trajectory prediction

Method sketch:

Input: Flight plans, actual routes, local weather, aircraft type, etc.

1. Past enriched trajectories are **Clustered**; medoids of clusters ('representatives') are also produced
2. A **Predictive Model** (PM) is built for each cluster
3. For each new flight plan FP, the **k-closest matches** (PMs) are found
4. Output: top-*k* PMs w.r.t. query FP

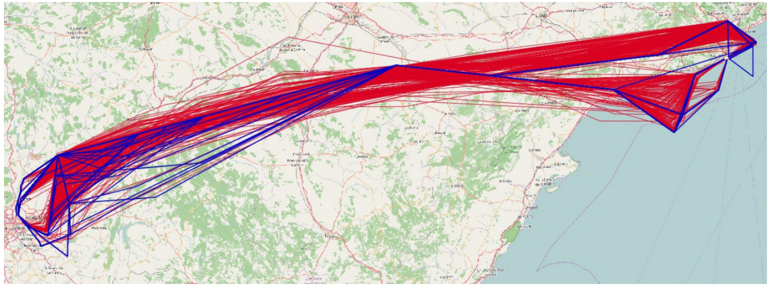


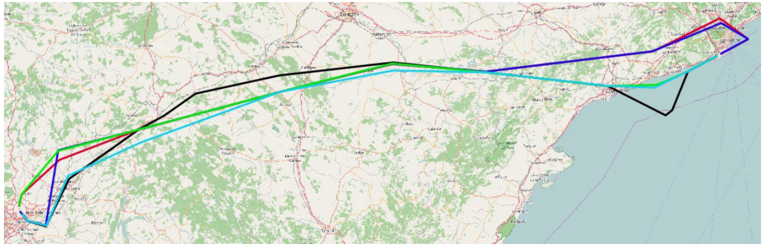
Data-driven trajectory prediction (cont.)

Method sketch:

Input: Flight plans, actual routes, local weather, aircraft type, etc.

1. Past enriched trajectories are **Clustered**; medoids of clusters ('representatives') are also produced
2. A **Predictive Model** (PM) is built for each cluster
3. For each new flight plan FP, the **k-closest matches** (PMs) are found
4. Output: top-*k* PMs w.r.t. query FP



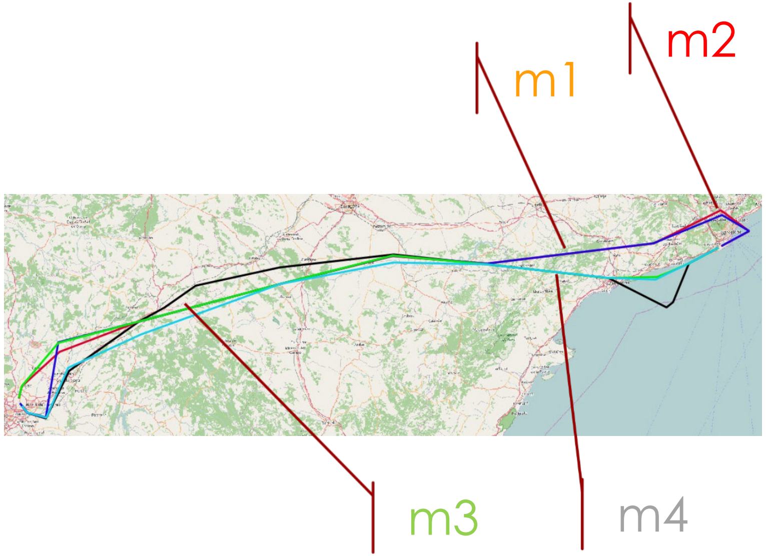


Data-driven trajectory prediction (cont.)

Method sketch:

Input: Flight plans, actual routes, local weather, aircraft type, etc.

1. Past enriched trajectories are **Clustered**; medoids of clusters ('representatives') are also produced
2. A **Predictive Model** (PM) is built for each cluster
3. For each new flight plan FP, the **k-closest matches** (PMs) are found
4. Output: top-*k* PMs w.r.t. query FP

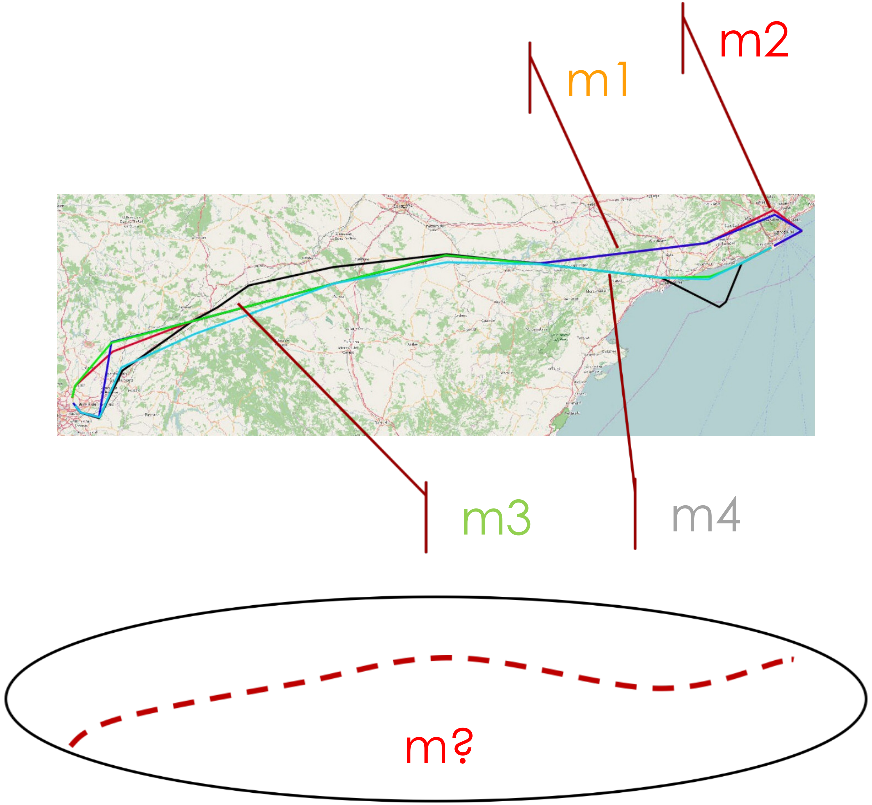


Data-driven trajectory prediction (cont.)

Method sketch:

Input: Flight plans, actual routes, local weather, aircraft type, etc.

1. Past enriched trajectories are **Clustered**; medoids of clusters ('representatives') are also produced
2. A **Predictive Model** (PM) is built for each cluster
3. For each new flight plan FP, the **k-closest matches** (PMs) are found
4. Output: top-*k* PMs w.r.t. query FP

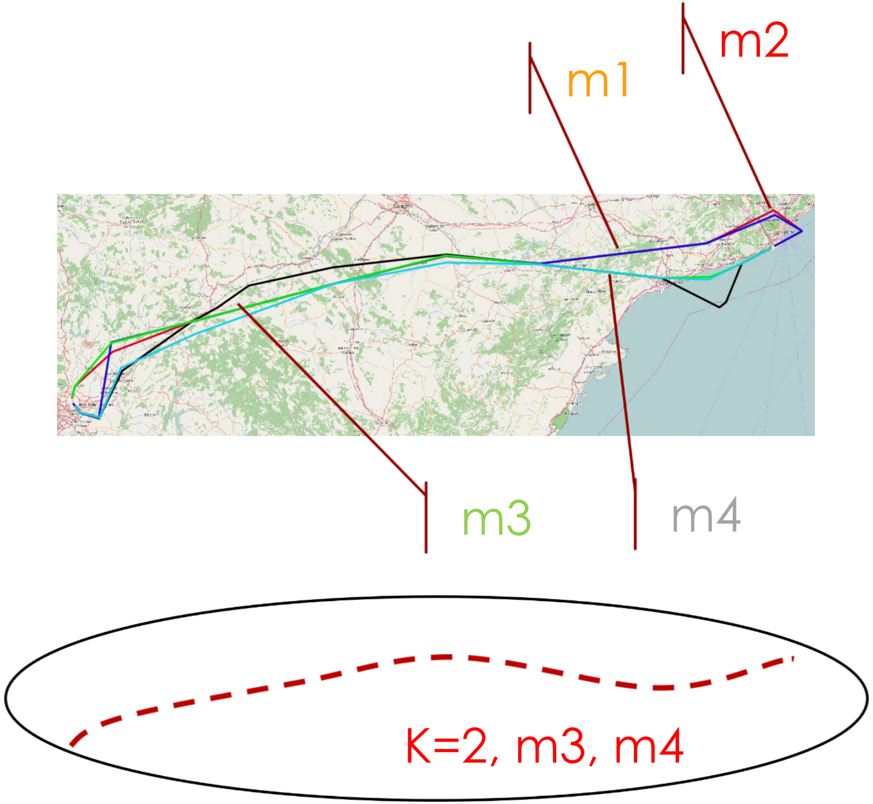


Data-driven trajectory prediction (cont.)

Method sketch:

Input: Flight plans, actual routes, local weather, aircraft type, etc.

1. Past enriched trajectories are **Clustered**; medoids of clusters ('representatives') are also produced
2. A **Predictive Model** (PM) is built for each cluster
3. For each new flight plan FP, the **k-closest matches** (PMs) are found
4. Output: top-*k* PMs w.r.t. query FP



Data-driven trajectory prediction (cont.)

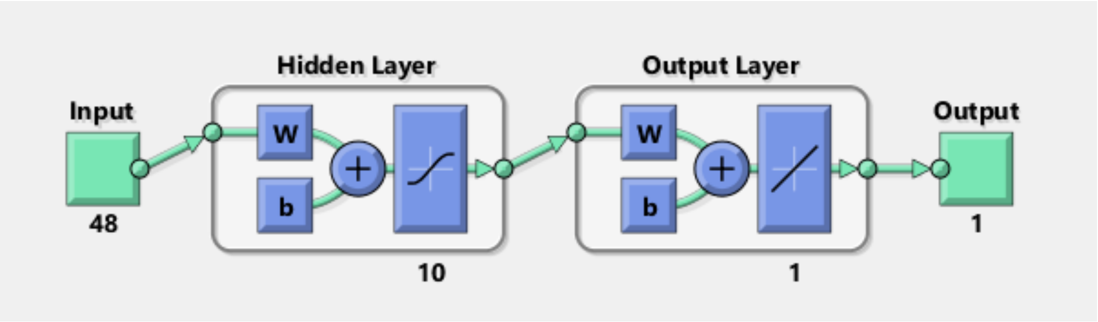
Method sketch:

Input: Flight plans, actual routes, local weather, aircraft type, etc.

1. Past enriched trajectories are **Clustered**; medoids of clusters ('representatives') are also produced
2. A **Predictive Model** (PM) is built for each cluster
3. For each new flight plan F **k-closest matches** (PMs) a found
4. Output: top-*k* PMs w.r.t. qu

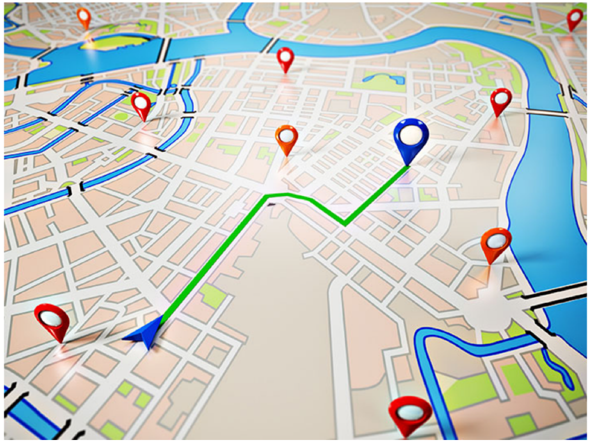
- Hidden Markov Model (HMM)
- Linear Regressor (LR)
- Decision Tree (CART)
- Neural Network (NN-MLP), etc.

Example (below) of Non-linear Regressor: NN-MLP
input (48): Flight Plan waypoints
output (1): deviation of prediction from a waypoint



Summary

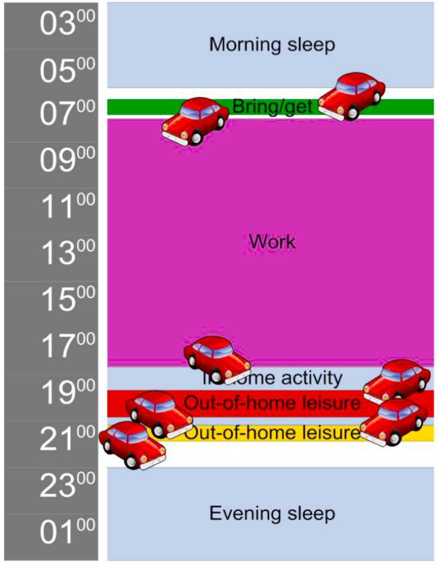
- The field of **Mobility Data Management and Exploration*** has many success stories to narrate on:
  - **Data management** - access methods, query processing techniques, DBMS extensions (the so-called, Moving Object Databases)
  - **Data exploration** – data mining techniques (clusters, flocks, convoys, T-patterns, hot spots, etc.)
  - … mostly based on the sampled spatio-temporal coordinates (x-, y-, z-, t-) of moving objects



* Pelekis N, Theodoridis Y (2014) Mobility data management and exploration. Springer.

Summary (cont.)

- The new era that emerges is around two keywords:
  - **Semantically-annotated trajectories*** – information about when, where, what, how, why
  - **Big mobility data**** – voluminous, streaming, disperse information about movement of objects (at land, sea, air)



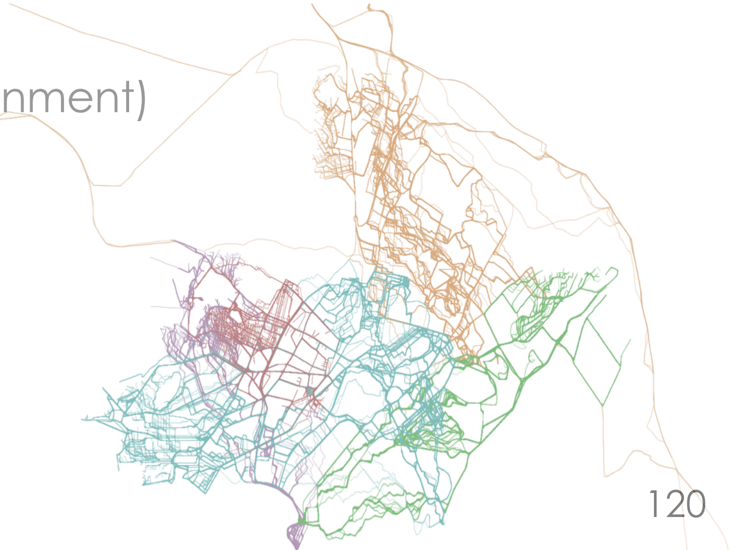
* Parent C, et al. (2013): Semantic trajectories modeling and analysis. ACM Computing Surveys, 45(4).

** Vouros GA, et al. (2018) Big data analytics for time critical mobility forecasting: recent progress and research challenges. In Proceedings of EDBT.

A tentative research agenda

… for the next 5 years:

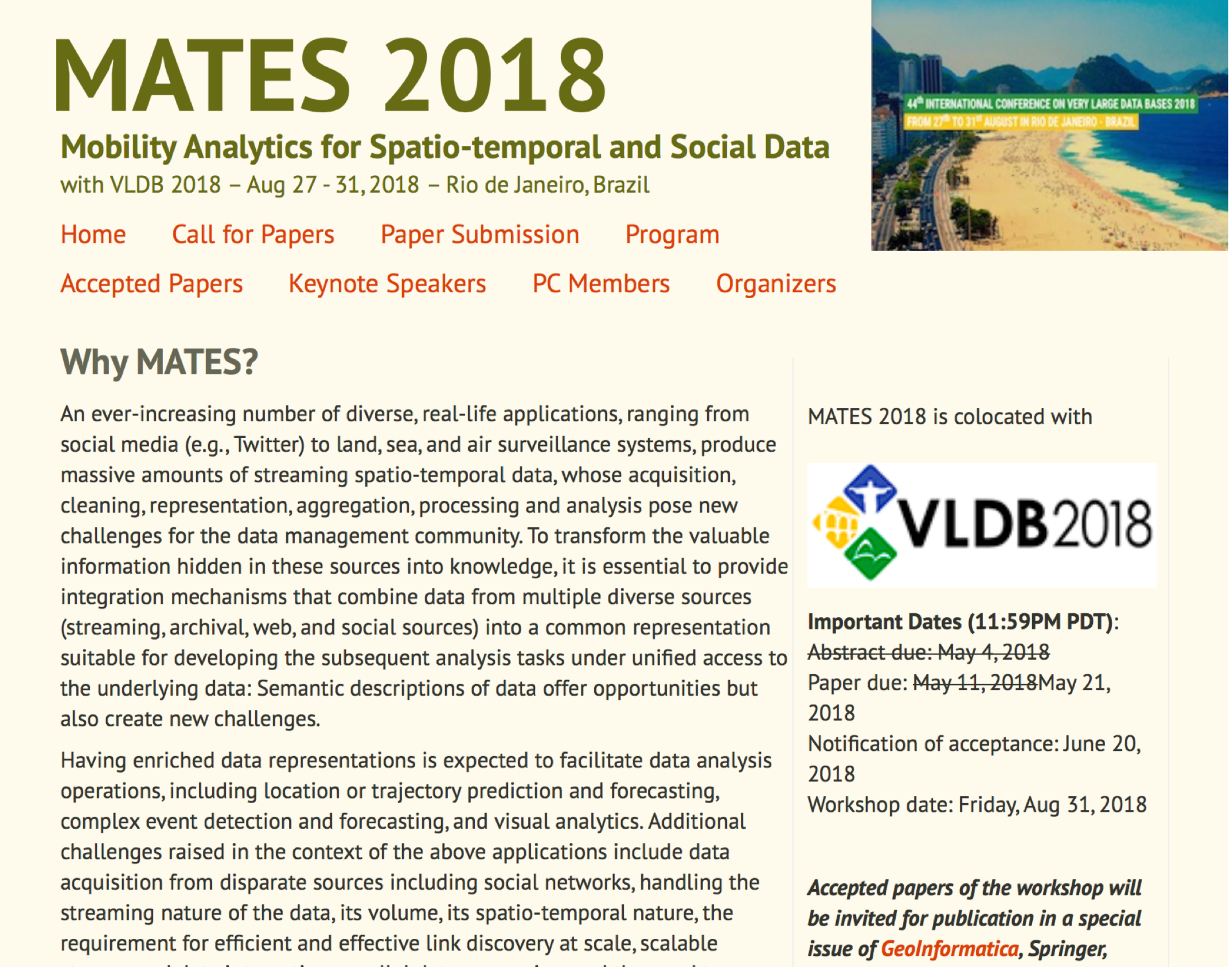
1. Reconstructing semantic trajectories online
2. Generating synthetic mobility data by-example
3. Spatio-Temporal-Textual data analytics
4. Predictive query processing (in big data environment)
5. Analyzing data-intensive mobility apps
6. Data-at-rest vs. data-in-motion: Who wins?



Acknowledgments

- Grateful to Data Science Lab people
  - Nikos Pelekis, and other colleagues and students

- Ack EU support through a series of grants:
  - **Track & Know** – Big Data for Mobility Tracking Knowledge Extraction in Urban Areas. 2018-20 [trackandknowproject.eu]
  - **MASTER** – Multiple Aspect Trajectory Management and Analysis, 2018-22 [http://www.master-project-h2020.eu]
  - **datAcron** – Big Data Analytics for Time Critical Mobility Forecasting, 2016-18 [datacron-project.eu]
  - **DART** – Data-Driven Aircraft Trajectory Prediction Research. 2016-18 [dart-research.eu]


44th INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES 2018
FROM 27th TO 31st AUGUST IN RIO DE JANEIRO - BRAZIL
VLDB2018

MATES 2018

Mobility Analytics for Spatio-temporal and Social Data

with VLDB 2018 – Aug 27 - 31, 2018 – Rio de Janeiro, Brazil

Home · Call for Papers · Paper Submission · Program · Accepted Papers · Keynote Speakers · PC Members · Organizers

Why MATES?

An ever-increasing number of diverse, real-life applications, ranging from social media (e.g., Twitter) to land, sea, and air surveillance systems, produce massive amounts of streaming spatio-temporal data, whose acquisition, cleaning, representation, aggregation, processing and analysis pose new challenges for the data management community. To transform the valuable information hidden in these sources into knowledge, it is essential to provide integration mechanisms that combine data from multiple diverse sources (streaming, archival, web, and social sources) into a common representation suitable for developing the subsequent analysis tasks under unified access to the underlying data: Semantic descriptions of data offer opportunities but also create new challenges.

Having enriched data representations is expected to facilitate data analysis operations, including location or trajectory prediction and forecasting, complex event detection and forecasting, and visual analytics. Additional challenges raised in the context of the above applications include data acquisition from disparate sources including social networks, handling the streaming nature of the data, its volume, its spatio-temporal nature, the requirement for efficient and effective link discovery at scale, scalable

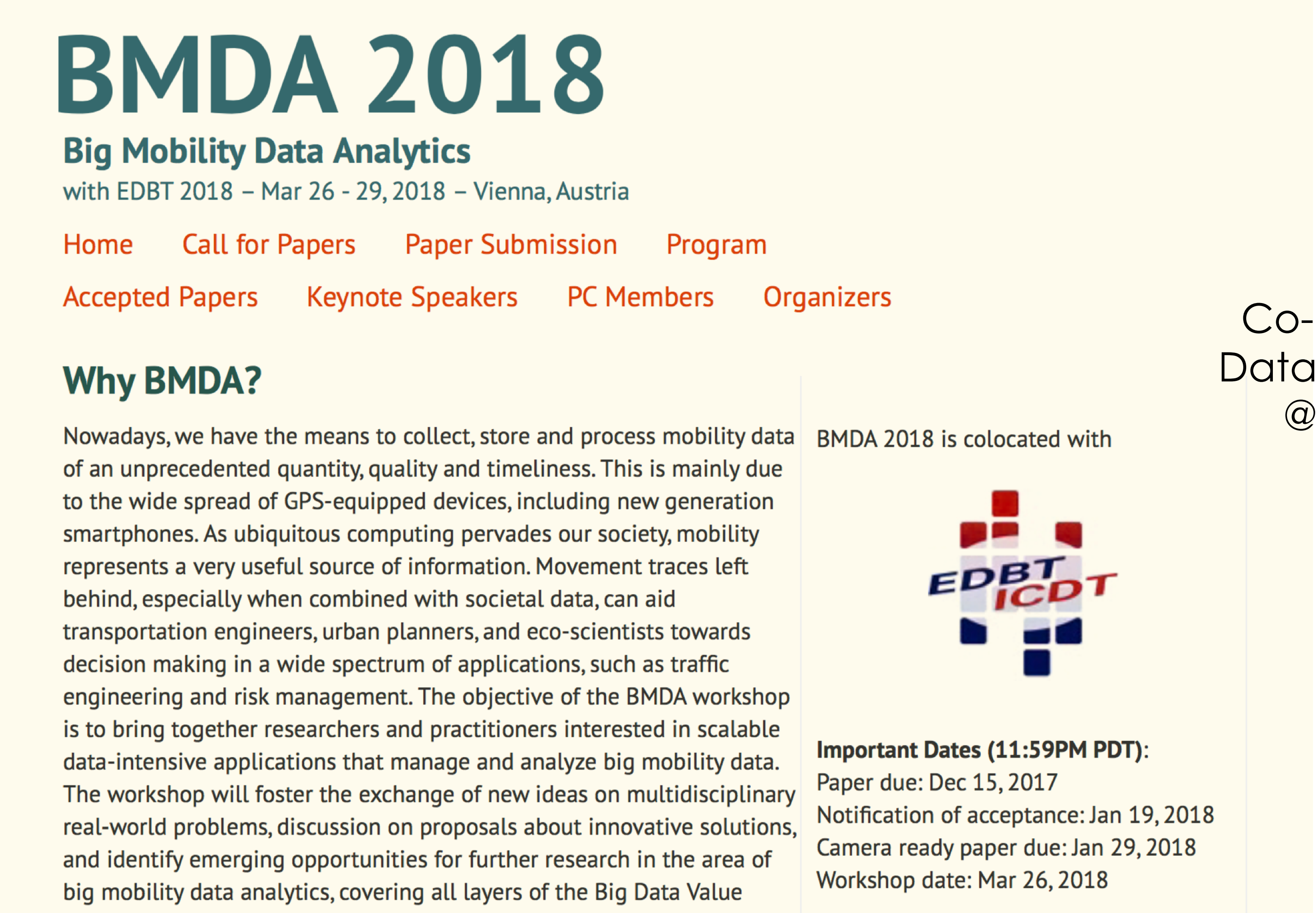
MATES 2018 is colocated with

Important Dates (11:59PM PDT):
~~Abstract due: May 4, 2018~~
Paper due: ~~May 11, 2018~~May 21, 2018
Notification of acceptance: June 20, 2018
Workshop date: Friday, Aug 31, 2018

Accepted papers of the workshop will be invited for publication in a special issue of GeoInformatica, Springer,

Co-organized by Data Science Lab @ Univ. Piraeus people

DATA STORIES



BMDA 2018

Big Mobility Data Analytics

with EDBT 2018 – Mar 26 - 29, 2018 – Vienna, Austria

Home Call for Papers Paper Submission Program

Accepted Papers Keynote Speakers PC Members Organizers

Why BMDA?

Nowadays, we have the means to collect, store and process mobility data of an unprecedented quantity, quality and timeliness. This is mainly due to the wide spread of GPS-equipped devices, including new generation smartphones. As ubiquitous computing pervades our society, mobility represents a very useful source of information. Movement traces left behind, especially when combined with societal data, can aid transportation engineers, urban planners, and eco-scientists towards decision making in a wide spectrum of applications, such as traffic engineering and risk management. The objective of the BMDA workshop is to bring together researchers and practitioners interested in scalable data-intensive applications that manage and analyze big mobility data. The workshop will foster the exchange of new ideas on multidisciplinary real-world problems, discussion on proposals about innovative solutions, and identify emerging opportunities for further research in the area of big mobility data analytics, covering all layers of the Big Data Value

BMDA 2018 is colocated with

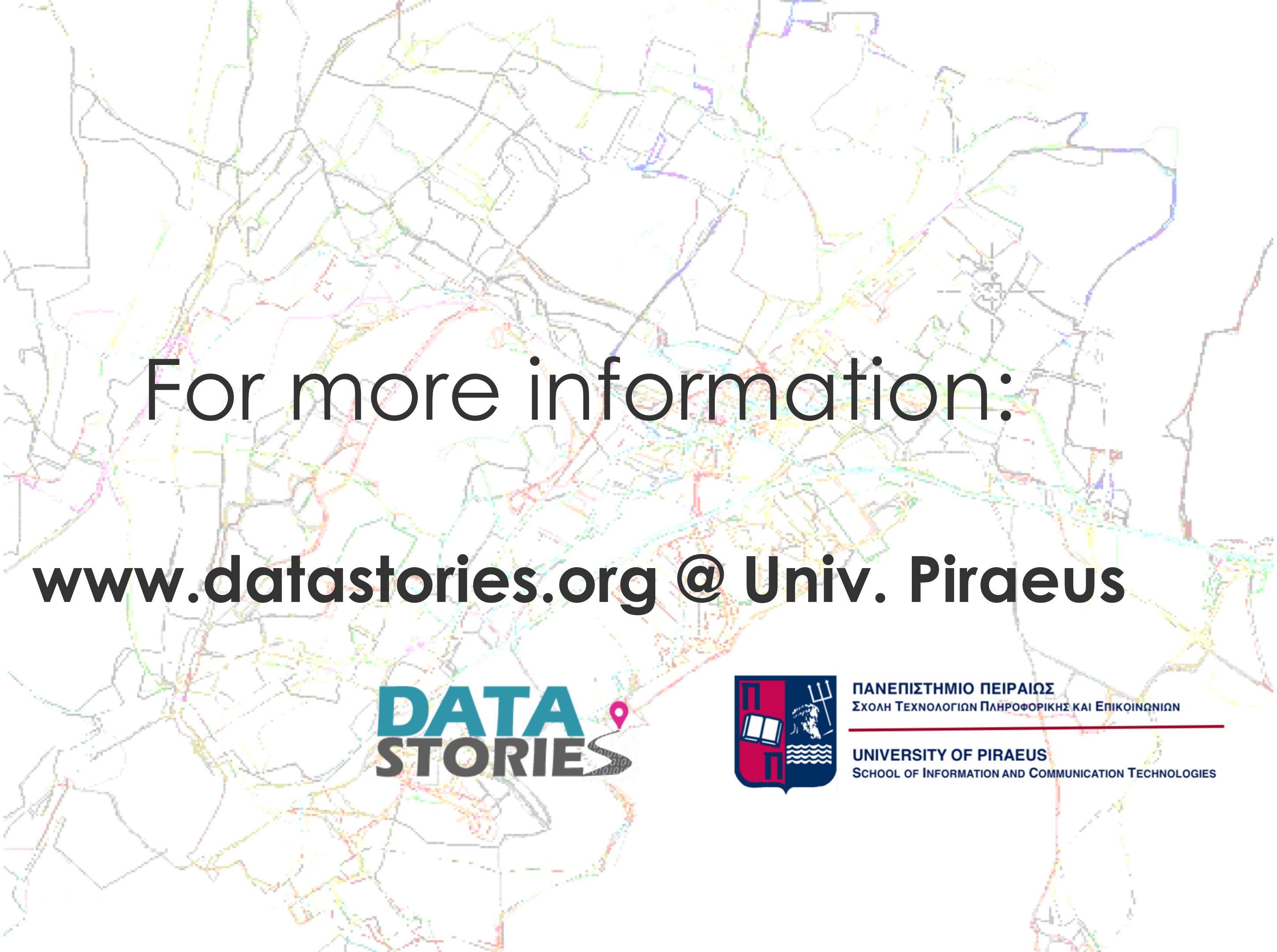
EDBT ICDT

Important Dates (11:59PM PDT):
Paper due: Dec 15, 2017
Notification of acceptance: Jan 19, 2018
Camera ready paper due: Jan 29, 2018
Workshop date: Mar 26, 2018

Co-organized by
Data Science Lab
@ Univ. Piraeus
people

DATA STORIES

Thank you
for your attention !!



For more information:

www.datastories.org @ Univ. Piraeus

DATA STORIES

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Σχολη Τεχνολογιων Πληροφορικης και Επικοινωνιων

UNIVERSITY OF PIRAEUS
School of Information and Communication Technologies