ACM Summer School on User Modeling and Personalization in Urban Computing:

Data Fusion

Dr Antonela Tommasel

ISISTAN, CONICET-UNICEN, Argentina

CONICET

I S I S T A N

Who am I?

- Dr Antonela Tommasel
  - PhD in Computer Sciences at UNICEN

- Work at ISISTAN, CONICET-UNICEN.

- Teacher Assistant at UNICEN.

- Research Interests:
  - Recommender systems
  - Text Mining
  - Social Media
  - Social Computing
  - …

CONICET

I S I S T A N

Table of contents

Table of contents

Big Challenges in Big Cities

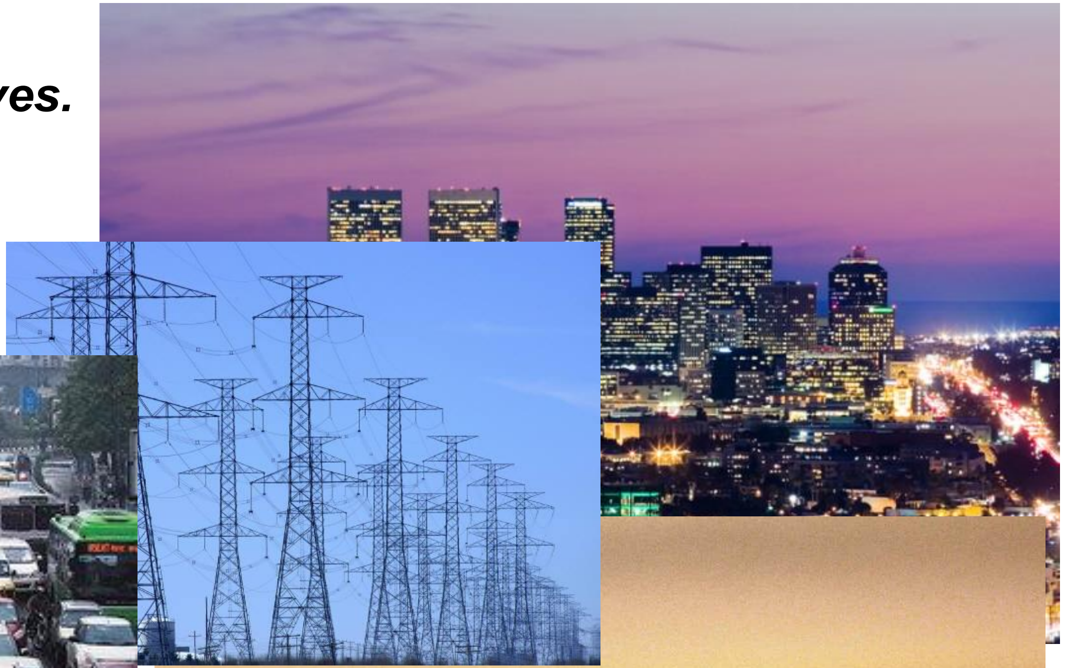
Urbanization's rapid progress has modernized many people's lives.

Big Challenges in Big Cities

Urbanization's rapid progress has modernized many people's lives.

However, it has also generated many big challenges:

- Traffic congestion.
- Environmental pollution.
- Increased energy consumption.
- …

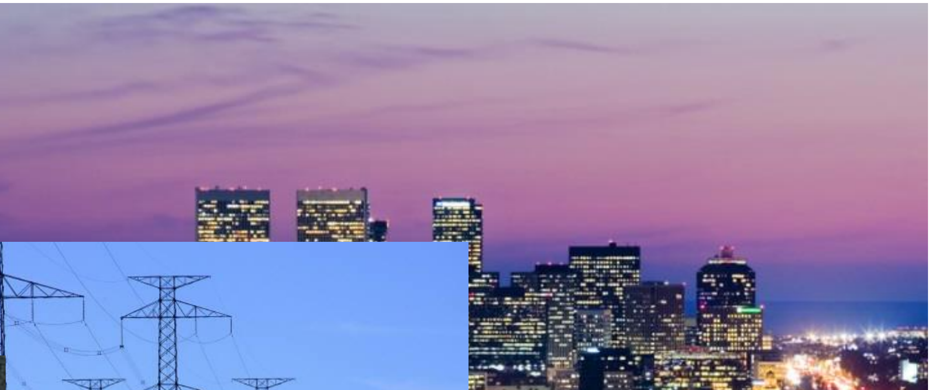


Big Challenges in Big Cities

Urbanization's rapid progress has modernized many people's lives.

However, it has also generated many big challenges:

- Traffic congestion.
- Environmental pollution.
- Increased energy consumption.
- …



Tackling these challenges seemed <u>nearly impossible</u> years ago given the complex and dynamic settings of cities.

Big Challenges in Big Cities

Urbanization's rapid progress has modernized many people's lives.

However, it has also generated many big challenges:

- Traffic congestion.
- Environmental pollution.
- Increased energy consumption.
- …

- Recently, sensing technologies and large-scale computing infrastructures have produced a variety of **big data** in urban spaces:
  - Human mobility.
  - Air quality.
  - Traffic patterns.
  - Geographical data.

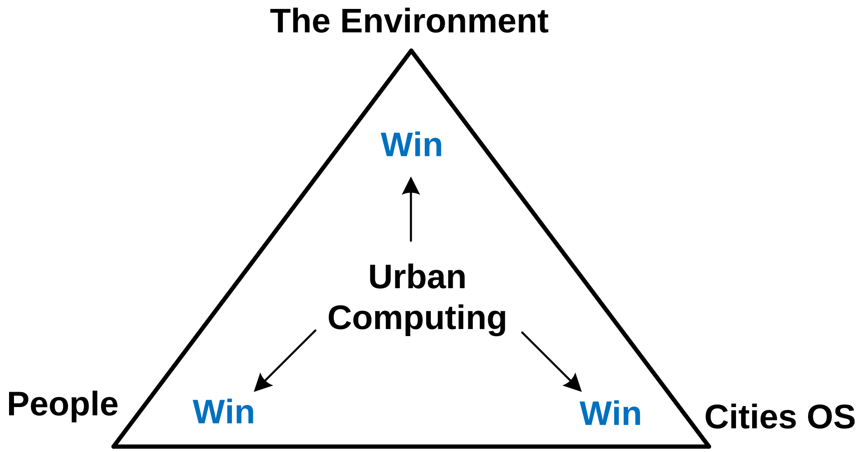
What is “Urban computing”?

“Urban computing is a concept where **every sensor, device, person, vehicle, building, and street** in the urban areas can be used as a component to sense **city dynamics** to enable a **city-wide computing** to tackle the **challenges** in **urban areas** to serve people and cities.

What is “Urban computing”?

“Urban computing is a concept where **every sensor, device, person, vehicle, building, and street** in the urban areas can be used as a component to sense **city dynamics** to enable a **city-wide computing** to tackle the **challenges** in **urban areas** to serve people and cities.

- A process of acquisition, integration, and analysis of big and heterogeneous data generated by a diversity of sources in cities.
- Connects unobtrusive and ubiquitous sensing technologies, advanced data management and analytics models, and novel visualization methods.
- An inter-disciplinary field where computer science meets urban planning, transportation, economy, the environment, sociology, and energy, etc., in the context of urban spaces.



What is “Urban computing”?

Differences and relations

Smart cities

- Current cities → Urban computing → Smart cities
- Unobtrusively sensing (leverage what we already have)

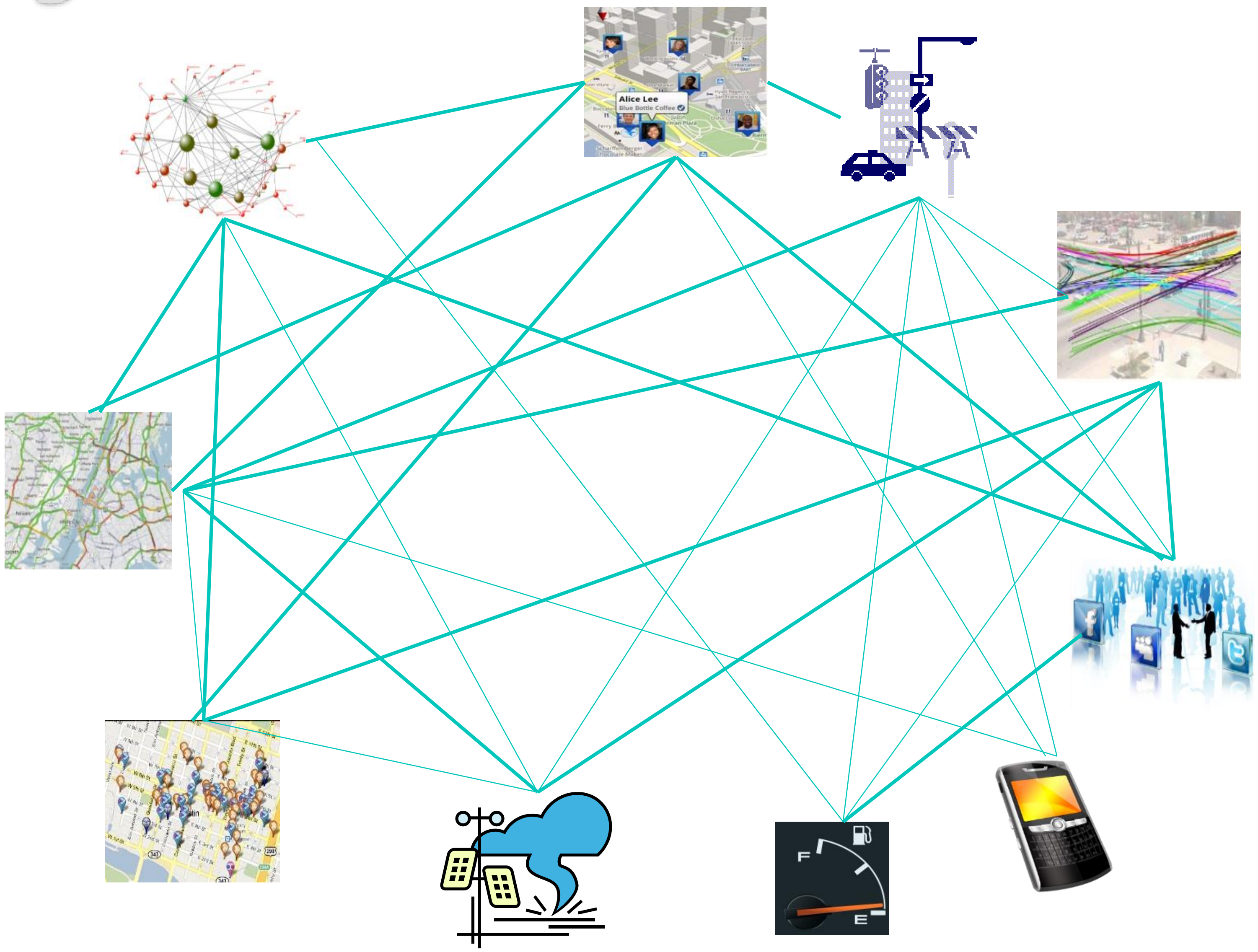
Internet of Things

- Infrastructure connecting objects
- Lack of human and social (only technology)

Cloud computing

- Technology & platform
- Many urban computing scenarios can be built on the Cloud

Big Data in Cities



These big data actually *reflect* the **<u>underlying problems</u>** of a city and <u>can help tackle</u> these problems when used ***<u>correctly</u>***.

Big Data in Cities

Scope

- Traffic flow
- Human mobility
- Energy consumption
- Environment
- Economic
- Populations
- …

Data available

- Mobile phone signal
- GPS traces
- Ticketing data in public transportation
- User-generated content
- Transportation sensors
- Environmental sensors
- …

What kind of data can we get from a city?

Spatially and temporally static

Spatially static and temporally dynamic

Spatially and temporally dynamic

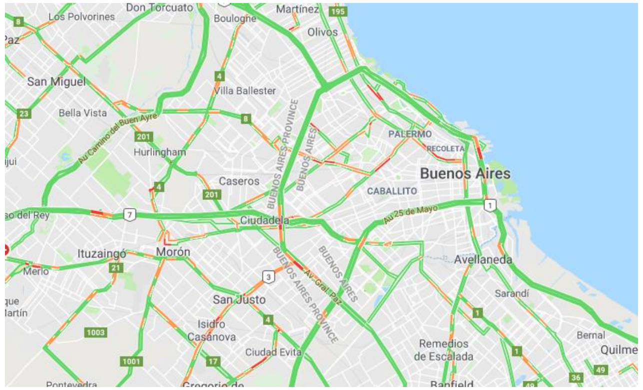
What kind of data can we get from a city?

Spatially and temporarily static

- *Points*
  - Point of interest (POI)
  - Intersections
- *Lines*
  - Route, pipelines
  - Rivers, coasts
- *Graphs*
  - Road Networks
  - Traffic Routes

What kind of data can we get from a city?

Spatially and temporarily static



Geographical Data

- Road network data may be the most frequently used geographical data in urban computing scenarios.
  - Traffic monitoring and prediction
  - Urban planning
  - Routing
  - Energy consumption analysis.

- It is usually represented by a **graph** that is composed of:
  - A set of edges (denoting road segments)
  - A collection of nodes (standing for road intersections).

What kind of data can we get from a city?

Spatially and temporarily static

Point of Interests (POI)

- A POI is usually described by:
  - Name.
  - Address.
  - Category.
  - A set of geospatial coordinates.
- While there are massive POIs in a city, the information of POIs could vary in time.
  - Name or address changes, shut down…

What kind of data can we get from a city?

Spatially and temporarily static

Point of Interests (POI)

- A POI is usually described by:
  - Name.
  - Address.
  - Category.
  - A set of geospatial coordinates.
- While there are massive POIs in a city, the information of POIs could vary in time.
  - Name or address changes, shut down…
- Challenges!
  - How to update the data?
  - How can we verify whether the information of a POI is correct? → Geospatial coordinates may be inaccurate.
  - How can we merge the POI data generated from different sources?

What kind of data can we get from a city?

Spatially static and temporarily dynamic

What kind of data can we get from a city?

Spatially static and temporarily dynamic

Environmental Monitoring Data.

- Meteorological data:
  - Humidity.
  - Temperature.
  - Barometer pressure.
  - Wind speed.
  - Weather conditions.

Sensors.
Public websites.

Air quality monitoring stations
Portable sensors.

- Air quality data
  - Concentration of PM2.5, NO_2, CO and CO_2 …

What kind of data can we get from a city?

Spatially static and temporarily dynamic

Environmental Monitoring Data.

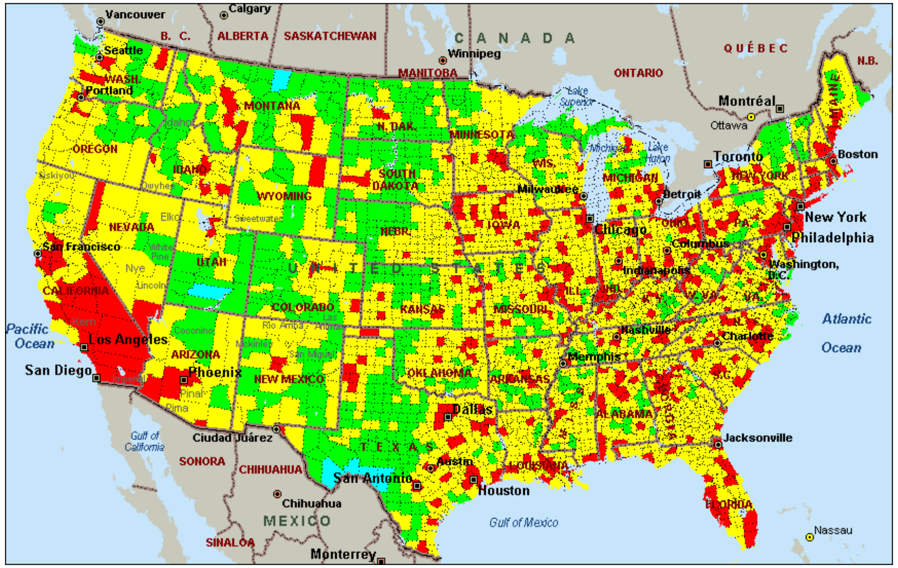
- Meteorological data:
  - Humidity.
  - Temperature.
  - Barometer pressure.
  - Wind speed.
  - Weather conditions.

Sensors.
Public websites.

Air quality monitoring stations
Portable sensors.

- Air quality data
  - Concentration of PM2.5, NO2, CO and CO2 …
- Air Quality Index
  - Influenced by multiple complex factors, such as traffic flow and land uses.
  - Urban air quality varies significantly by location and changes over time.

Good, moderate, and unhealthy



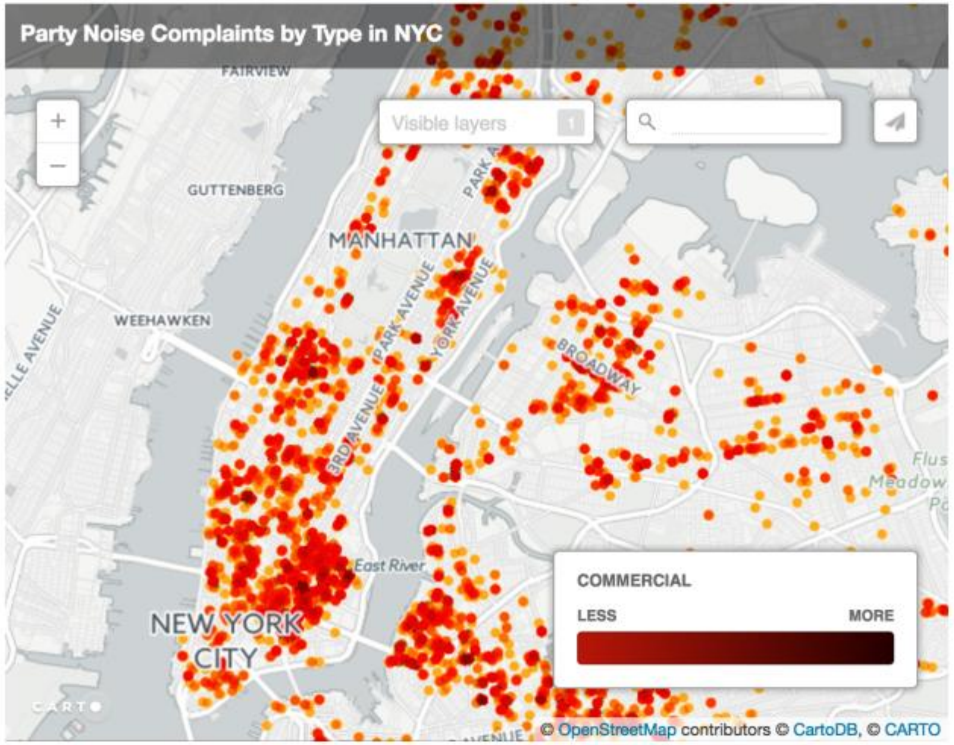
A limited number of monitoring stations cannot reveal the fine-grained air quality throughout a city.

What kind of data can we get from a city?

Spatially static and temporarily dynamic

Noise data

- It has a direct impact on people’s mental and physical health.
- Measuring noise pollution depends on both the intensity of noises and people’s tolerance to noises, which changes over time.
- The data can be used to diagnose a city’s noise pollution.

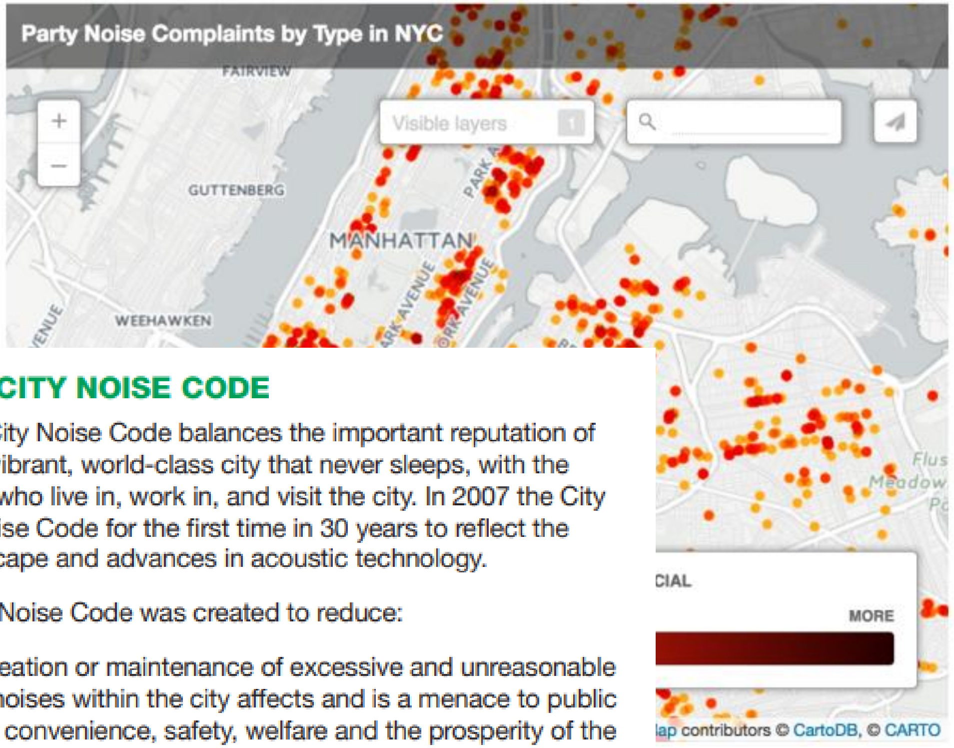


What kind of data can we get from a city?

Spatially static and temporarily dynamic

Noise data

- It has a direct impact on people’s mental and physical health.
- Measuring noise pollution depends on both the intensity of noises and people’s tolerance to noises, which changes over time.
- The data can be used to diagnose a city’s noise pollution.
- There are platforms were people can complain.
    - Each complaint is associated with a timestamp, a location, and a category. Noise is the third largest category in the data.



NEW YORK CITY NOISE CODE

The New York City Noise Code balances the important reputation of New York as a vibrant, world-class city that never sleeps, with the needs of those who live in, work in, and visit the city. In 2007 the City updated the Noise Code for the first time in 30 years to reflect the changing landscape and advances in acoustic technology.

Simply put, the Noise Code was created to reduce:

“The making, creation or maintenance of excessive and unreasonable and prohibited noises within the city affects and is a menace to public health, comfort, convenience, safety, welfare and the prosperity of the people of the city.”

In order to enforce this objective, the New York City Department of Environmental Protection (DEP) and the New York City Police Department (NYPD) share duties based on the type of noise complaint. To report a noise complaint, call 311 and they will direct your grievance to the appropriate agency. For example, the NYPD handles “neighbor to neighbor” noise complaints.

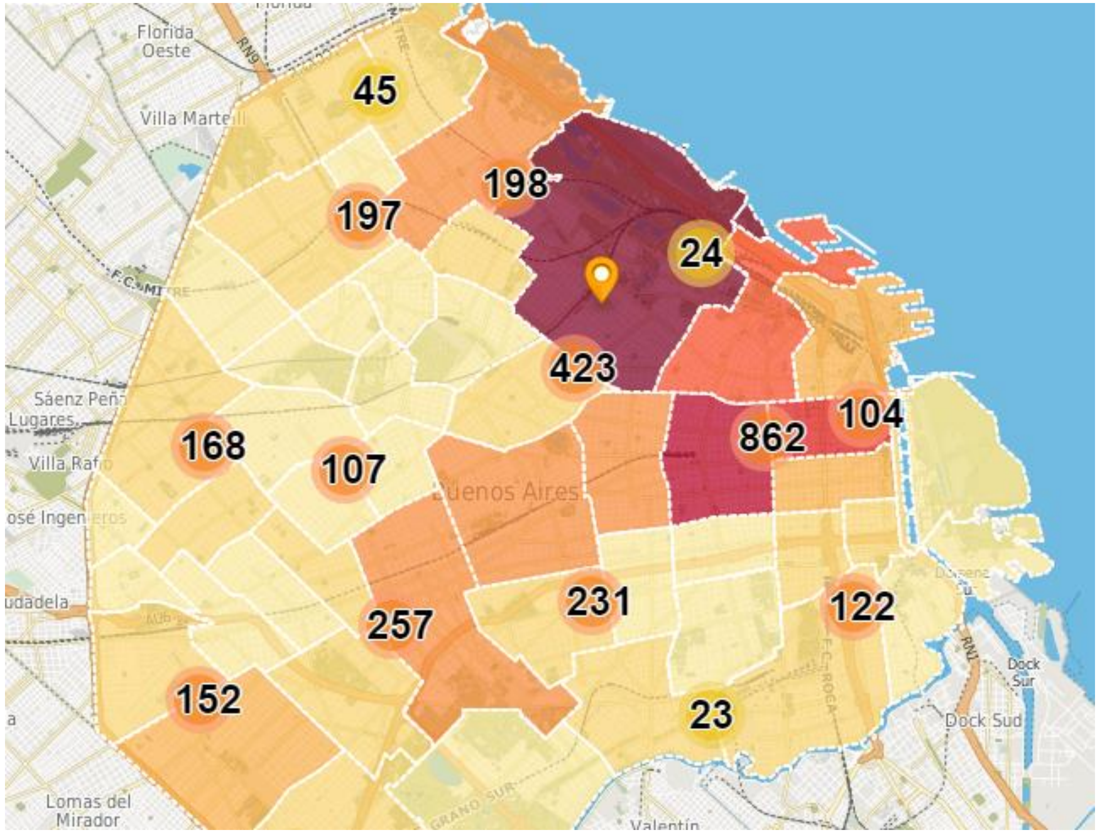
What kind of data can we get from a city?

Spatially static and temporarily dynamic

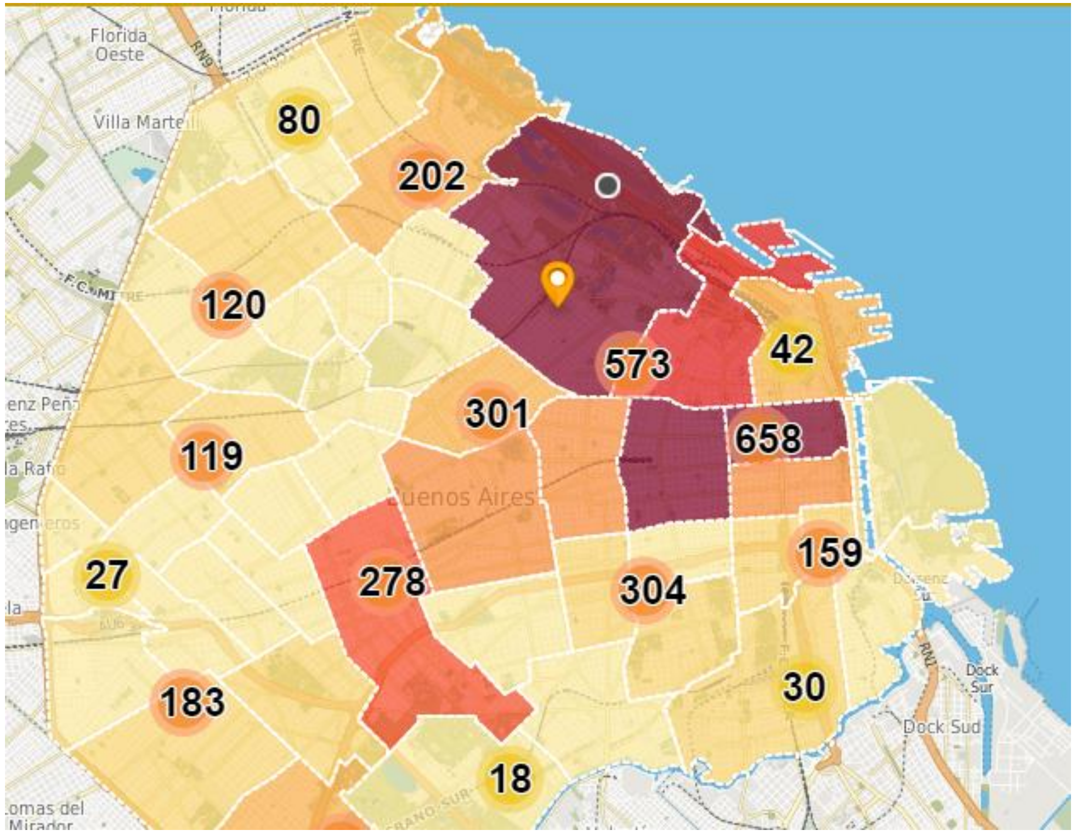
Buenos Aires Ciudad | Mapa del Delito

Crime data

September 2018



December 2018



Spatially static and temporarily dynamic

Crime data

Temporal Pattern Analysis

- Criminal temporal patterns are complicated since temporal resources could be structured in various intervals like weeks, months, seasons, years…
- Generally focuses on learning useful temporal patterns from sequential crime data.

Spatio-Temporal Pattern Analysis

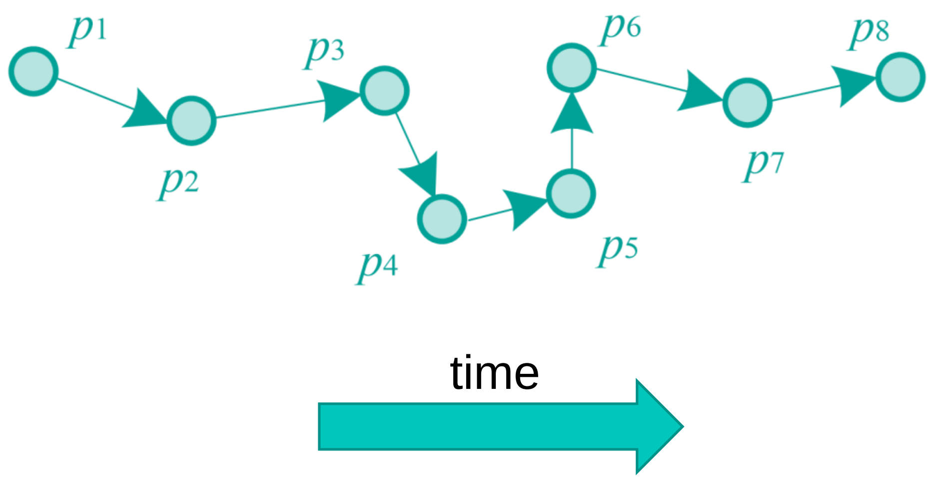
- Aims to obtain understanding from geo- and time-related crime data.
- How to identify patterns from the dynamic interaction among space, time and crime?

Spatial Pattern Analysis

- Crimes are not evenly or randomly distributed in an urban area.
  - Correlated with environmental contexts.
- Aims at learning the aggregation of crime, i.e., hotspots, inside a city.

Spatially and temporarily dynamic

A ***spatial trajectory*** is a trace generated by a moving object in geographical spaces, usually represented by of a series of chronologically ordered points, e. g., $p_1 \rightarrow p_2 \rightarrow \cdots \rightarrow p_n$, where each point consists of a geospatial coordinate set and a timestamp such as $p = (x, y, t)$.



Spatially and temporarily dynamic

- Human mobility
  - Travel logs
  - Sport analysis
  - Check-ins
  - Phone signal
  - Credit card transactions
  - …
- Movement of vehicles
  - Taxis, buses, trucks trajectories
  - Air planes, ferries, …
- Animals migration
  - Birds
  - …

What kind of data can we get from a city?

Spatially and temporarily dynamic

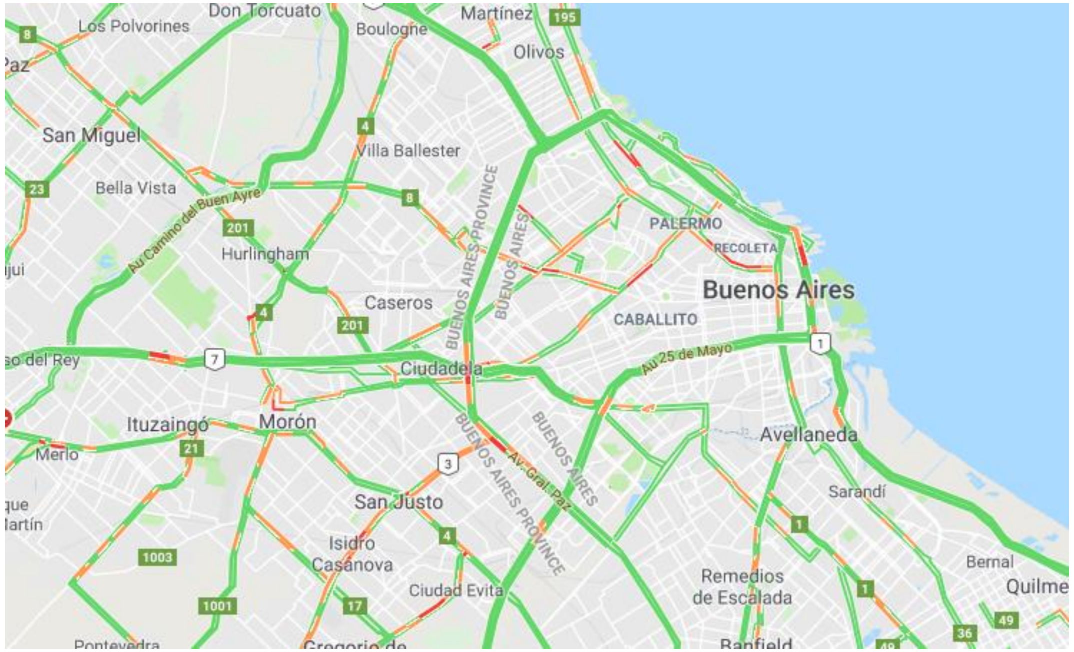
Traffic Data

Different ways to collect traffic data

Loop sensors

- Embedded in pairs in major roads.
- Detect the time interval in which a vehicle travels across two sensors.
- Knowing the distance speed can be calculated.
- The number of vehicles traversing a pair of loop detectors in a time slot determines the traffic volume on a road.

- Expensive! → limited coverage
- Does not tell how a vehicle travels, i.e. waiting times and direction turns cannot be recognised.



What kind of data can we get from a city?

Spatially and temporarily dynamic

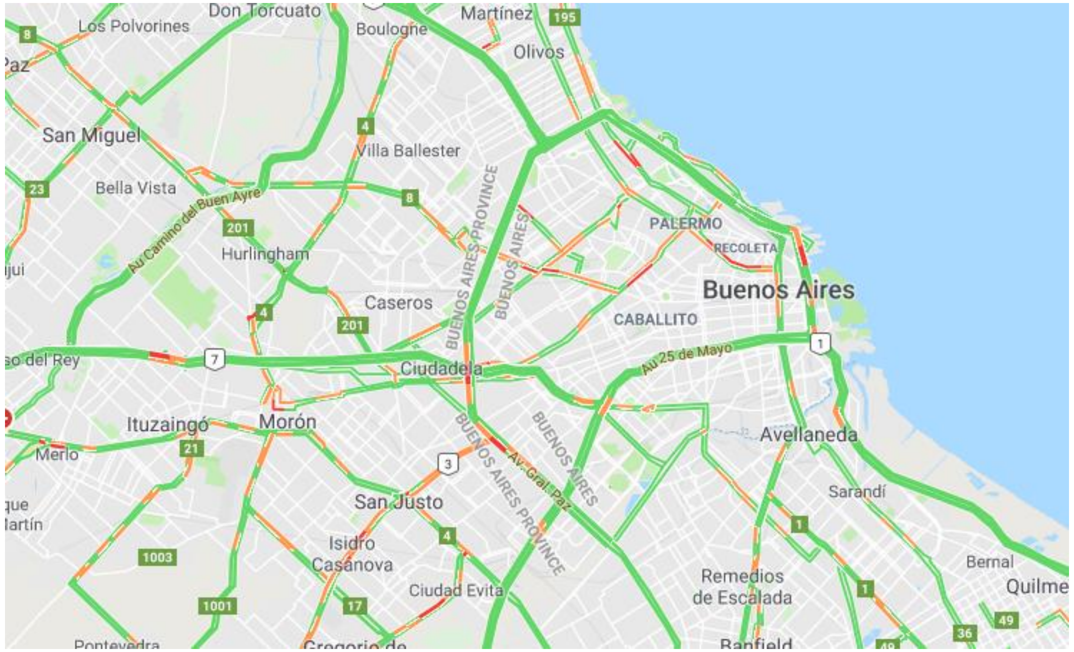
Traffic Data

Different ways to collect traffic data

Surveillance cameras

- Widely deployed in urban areas, generating a huge volume of images and videos reflecting traffic patterns.
- Provides a visual ground truth of traffic conditions to people.

- Automatically turn the images and videos into a specific traffic volume and travel speed is challenging.
- It is difficult to apply a machine-learning model trained for one location to other locations (e.g. differing structure of roads and differing camera settings, angle, and focus).
- Human effort!



What kind of data can we get from a city?

Spatially and temporarily dynamic

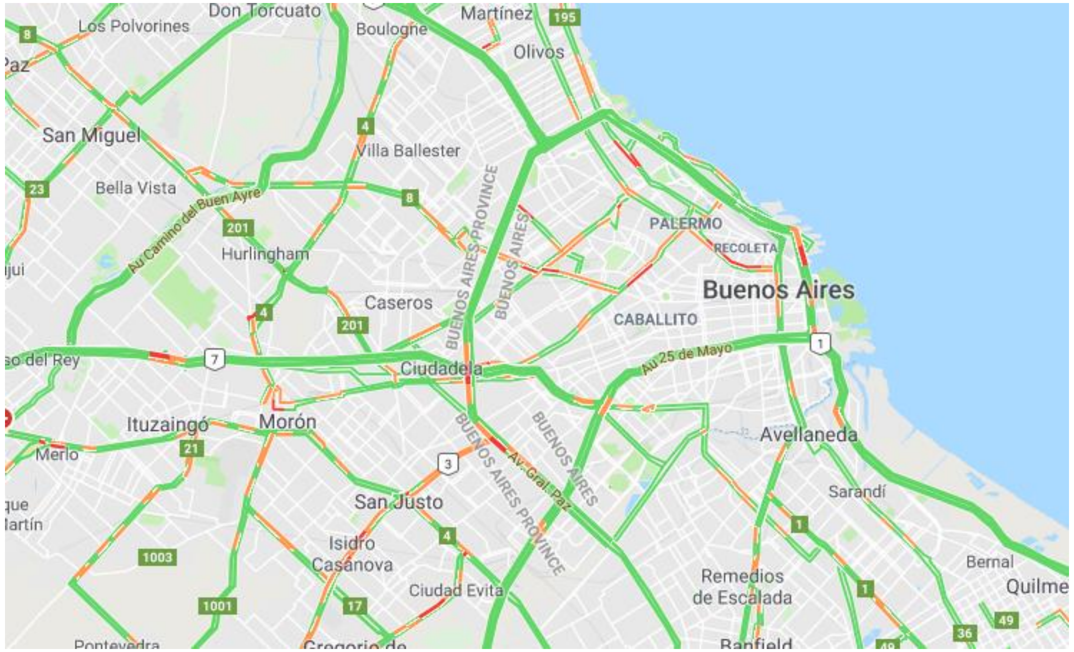
Traffic Data

Different ways to collect traffic data

Floating cars

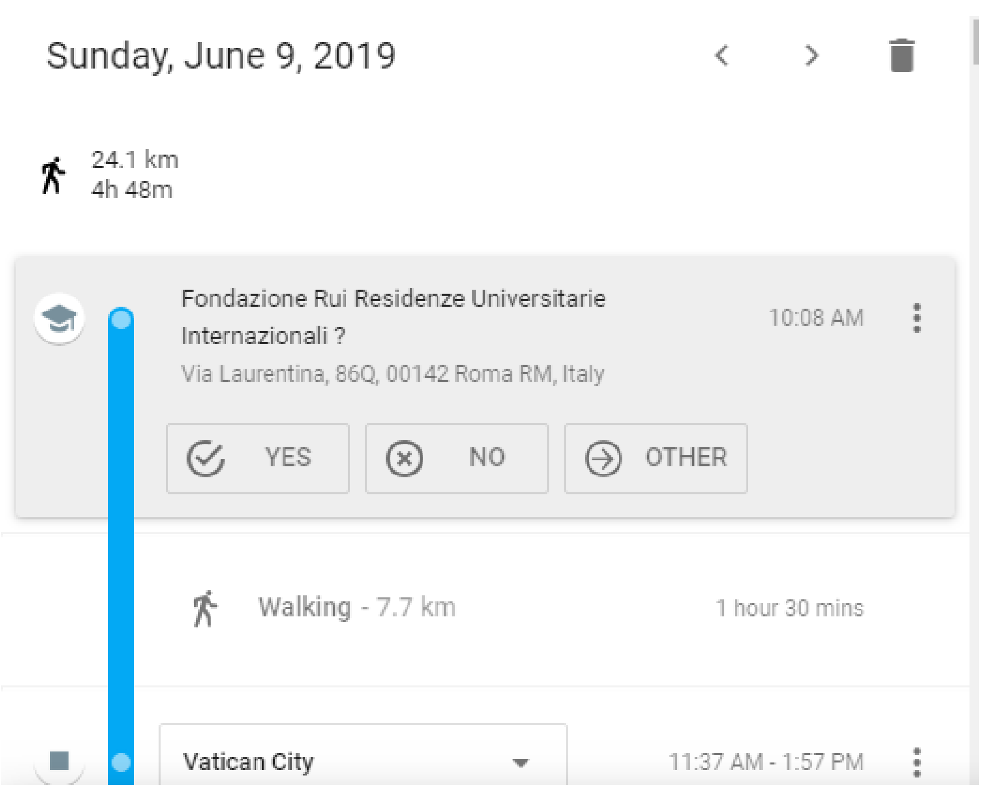
- Generated by vehicles traveling around a city with a GPS sensor.
- The trajectories of these vehicles will be sent to a central system and matched to a road network for deriving speeds on road segments.

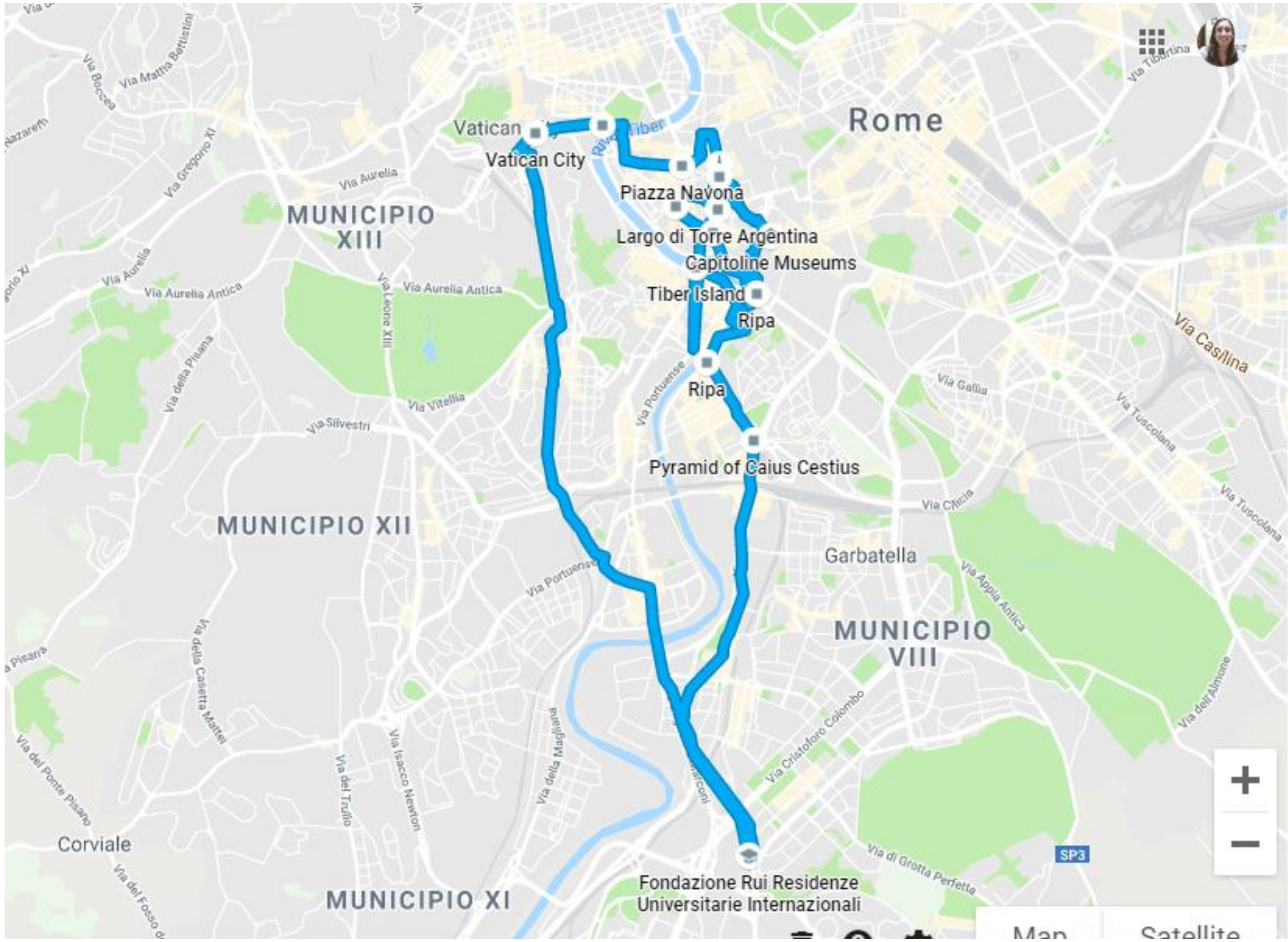
- Coverage depends on the distribution of the probing vehicles.



What kind of data can we get from a city?

Spatially and temporarily dynamic





What kind of data can we get from a city?

Spatially and temporarily dynamic

Commuting Data

- This is another kind of data representing citywide human mobility.

- People traveling in cities generate a huge volume of commuting data.
  - Card swiping data in a subway system or bus line.
  - Ticketing data in parking lots.

- Card swiping data is widely available in a city's public transportation systems.
  - Each transaction record consists of a timestamp of entering/leaving a station and the ID of the station as well as the fare for this trip.

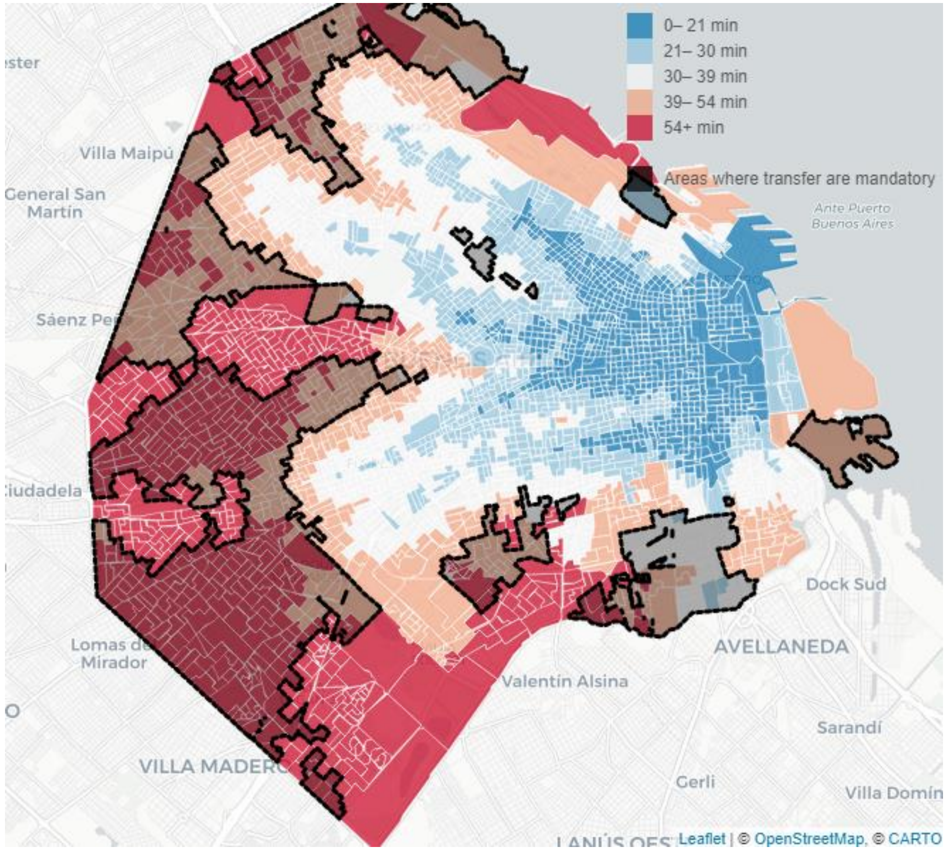
- Street-side parking is usually paid for through a parking meter.
  - The payment information of parking slots may include the time the ticket is issued and the parking fare.
  - The data indicates the traffic of vehicles around a place.

Improve a city's parking infrastructure.
Analyse people's travel patterns.

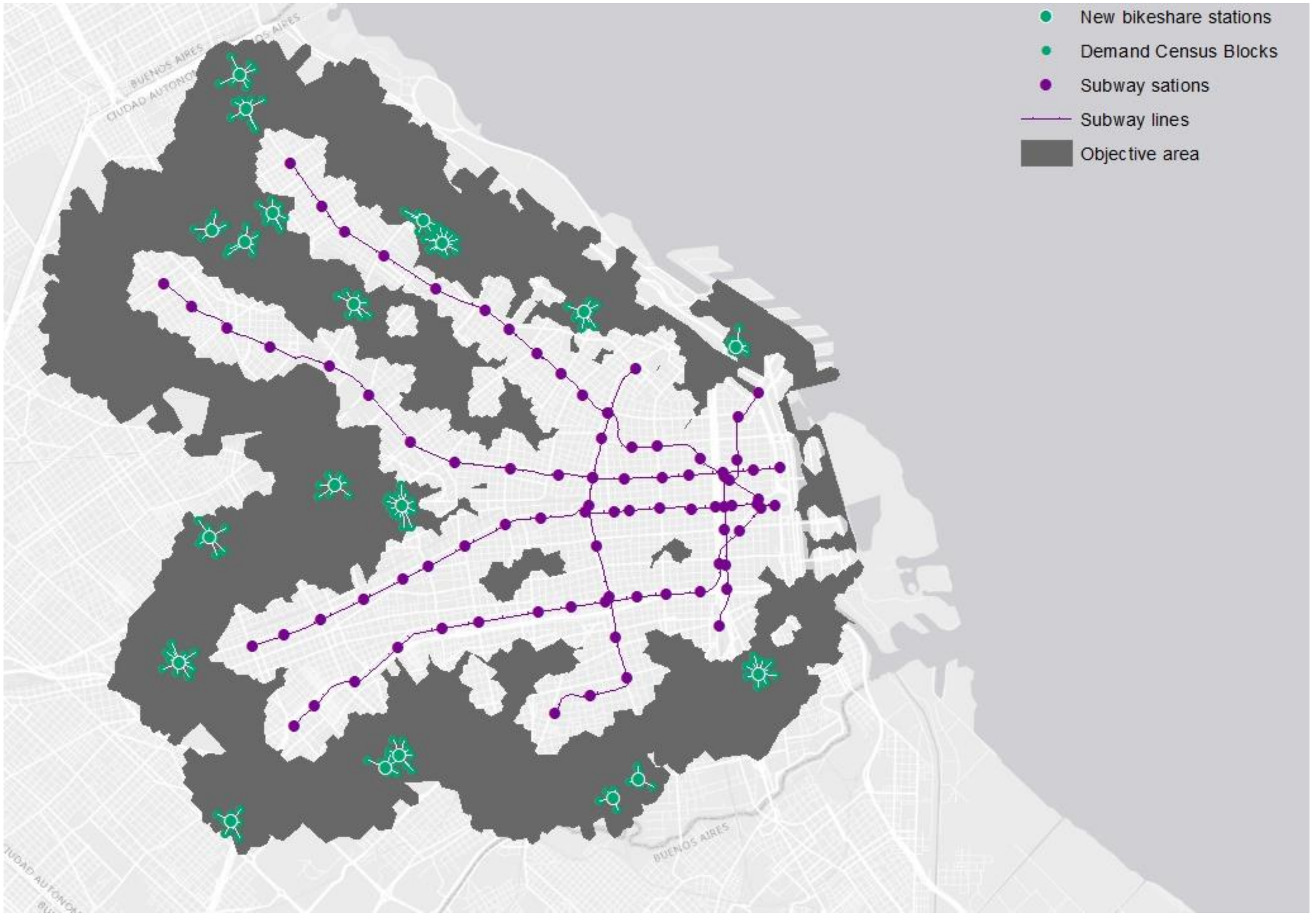
What kind of data can we get from a city?

Spatially and temporarily dynamic

Commuting Data



Using bike share and subway to commute



Spatially and temporarily dynamic

Mobile Phone Signals

- A call detail record is a data record produced by a telephone exchange containing attributes that are specific to a single instance of a phone call:
  - The phone numbers of both the calling and receiving parties.
  - The start time.
  - The duration of that call.

- Study the behaviour of an individual or build a network between different users.

- The similarity between users can also be inferred.

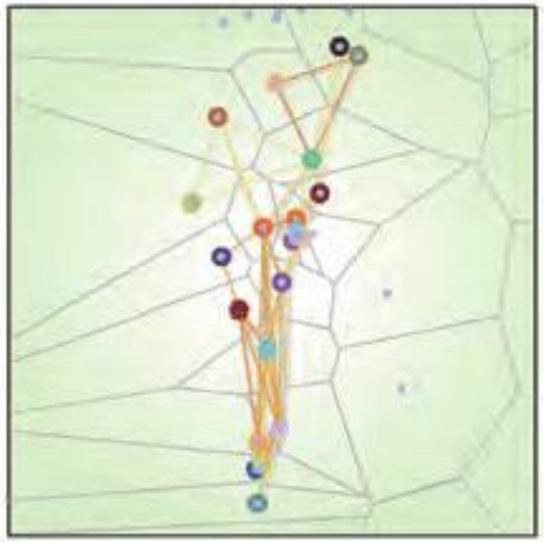
Spatially and temporarily dynamic

Mobile Phone Signals

- Another category of mobile phone signals is more concerned about the **location** of a user **rather** than the **communication** between phones.

- Using a triangle positioning algorithm, a mobile phone’s location can be roughly calculated based on three or more base stations.

- This kind of data denotes **citywide human mobility**, which can be used for detecting urban anomalies or, in the long run, for studying a **city’s functional regions** and urban planning.



Spatially and temporarily dynamic

Crime Data

Crime rate prediction

- Predict the future crime rate of a given urban region.
- Based on previous crime data.
- Based on environmental context data.
  - It is assumed that the distribution of events far from homes raises the opportunities for offense and hence yields higher crime rates.
  - Seasonal crimes.
- Based on social media data.
  - Twitter posts with rich and event-based context is leveraged for predicting criminal incidents

Next-location prediction

- Predict the location where an offender will commit a crime according to the offender's historical trajectories or other information.
- According to crime pattern theory, criminals usually use their most familiar regions as part of their activity space.
- Social network analysis.

Spatially and temporarily dynamic

Crime Data

Criminal network analysis

- Criminal networks are important for crime analysis and prevention.

- A criminal network consists of:
  - Nodes: the individual actors within the criminal network (i.e. the offenders)
  - Ties: the relationships between actors (e.g. co-offenders and crime gang).

What kind of data can we get from a city?

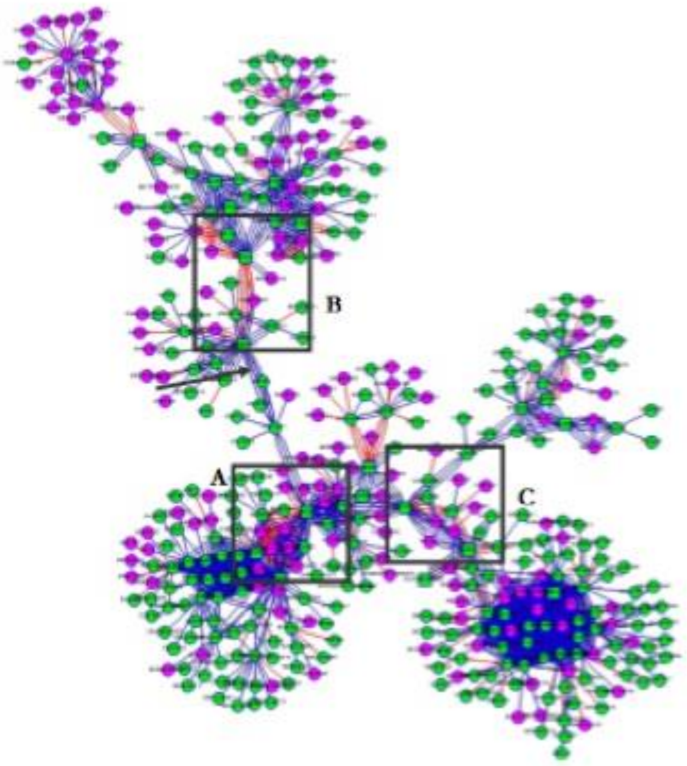
Spatially and temporarily dynamic

Crime Data

Criminal network analysis

- Criminal networks are important for crime analysis and prevention.

- A criminal network consists of:
  - Nodes: the individual actors within the criminal network (i.e. the offenders)
  - Ties: the relationships between actors (e.g. co-offenders and crime gang).



The “tension spots in the network characterized by a high number of overlapping positive and negative ties surroun ding triads” across the two fighting gangs in Richmond.

What kind of data can we get from a city?

Spatially and temporarily dynamic

Crime Data

Criminal network analysis

- Criminal networks are important for crime analysis and prevention.
- A criminal network consists of:
  - Nodes: the individual actors within the criminal network (i.e. the offenders)
  - Ties: the relationships between actors (e.g. co-offenders and crime gang).

Near repeat victimization

- Crime does not happen randomly or evenly across time or space.
- The near-repeat victimization identifies the increased risk of repeat victimization at the same or surrounding regions and within a certain time period.

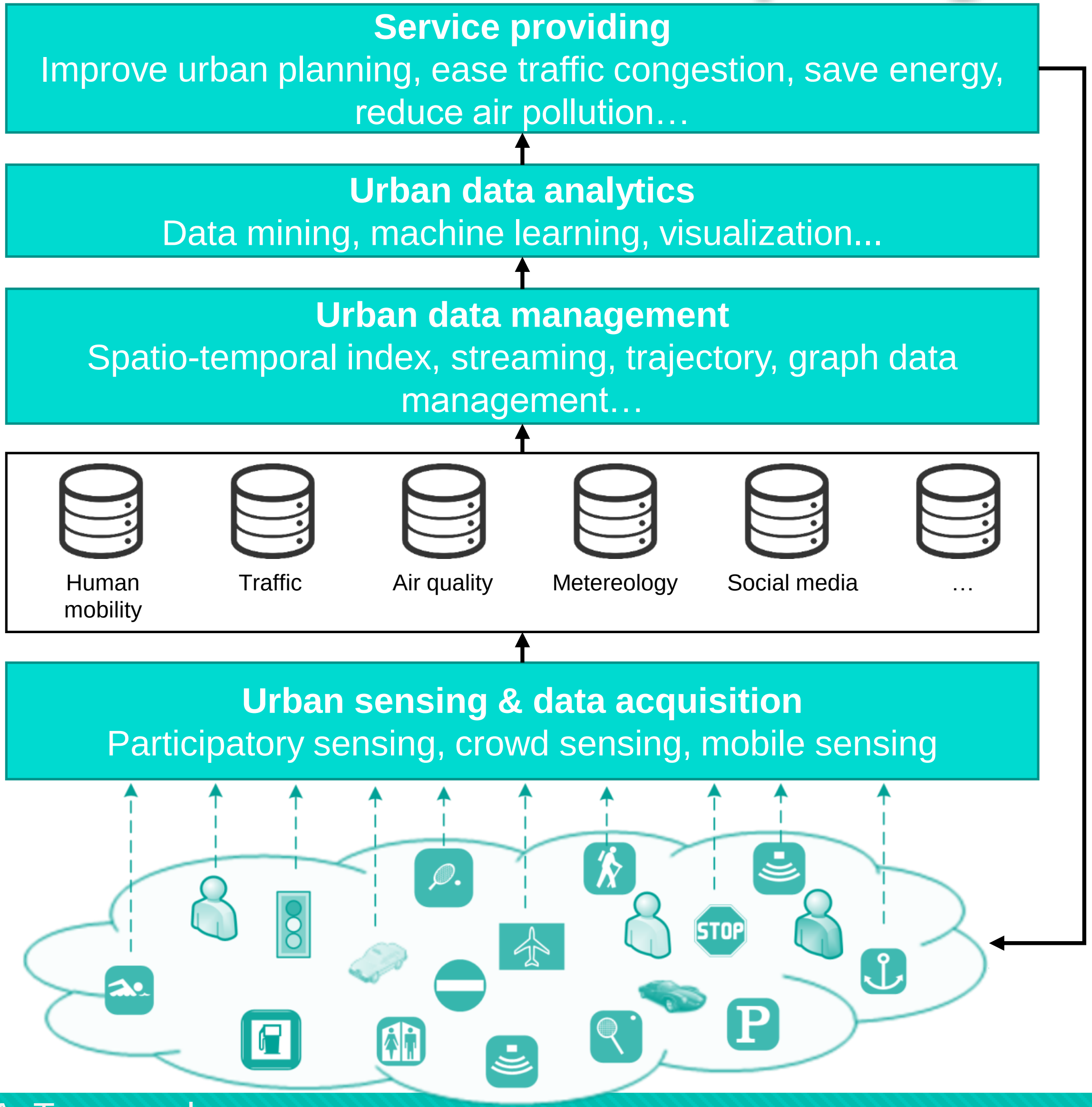
Police patrolling planning

- Helps increase the effectiveness of police patrolling and improve public security simultaneously.
- Patrol area allocation.
- Patrol route planning.

Table of contents

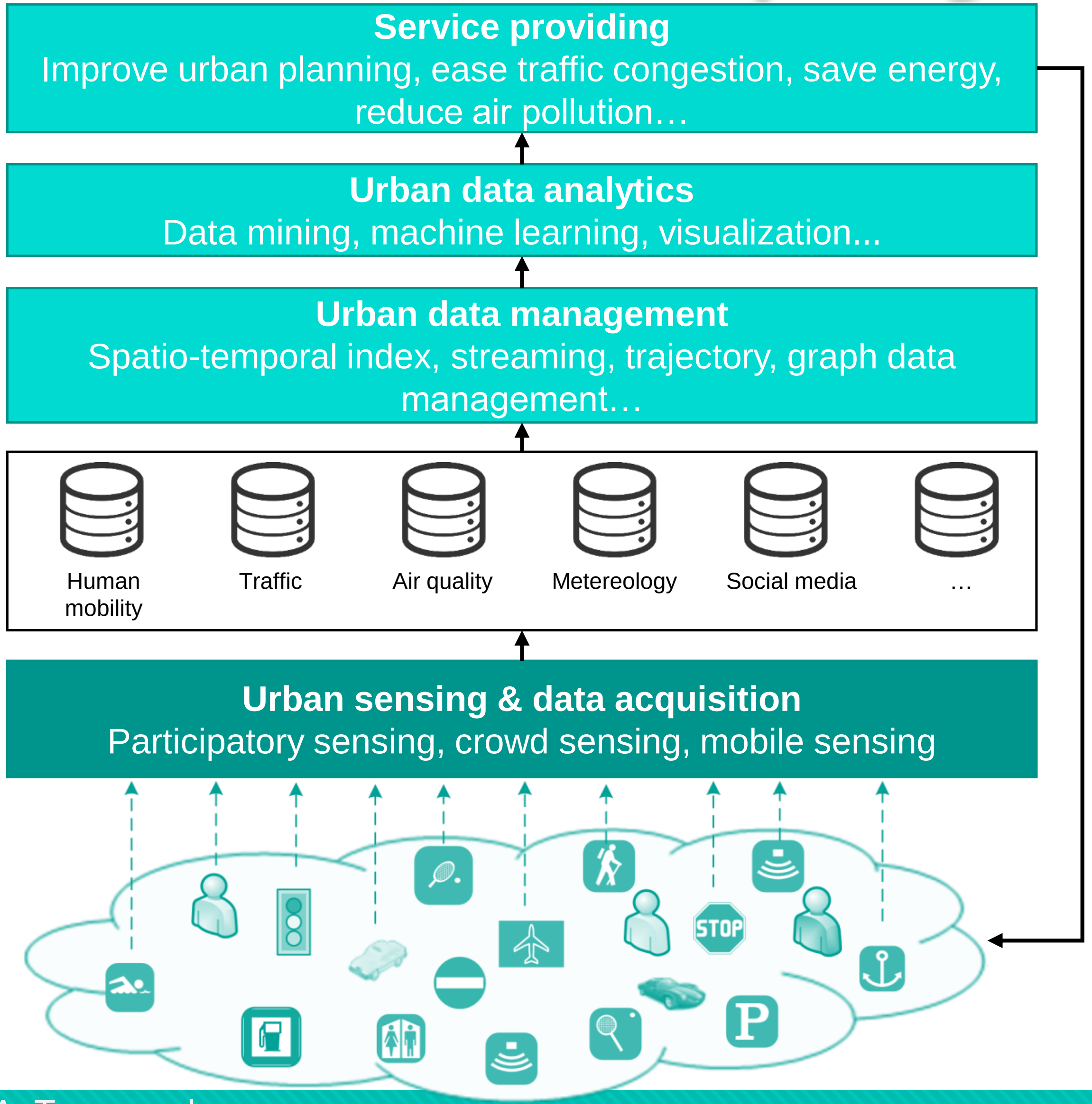
1. Urban computing & Urban data
2. **Urban data (and computing) framework**
3. Data fusion

Urban data (and computing) framework



Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

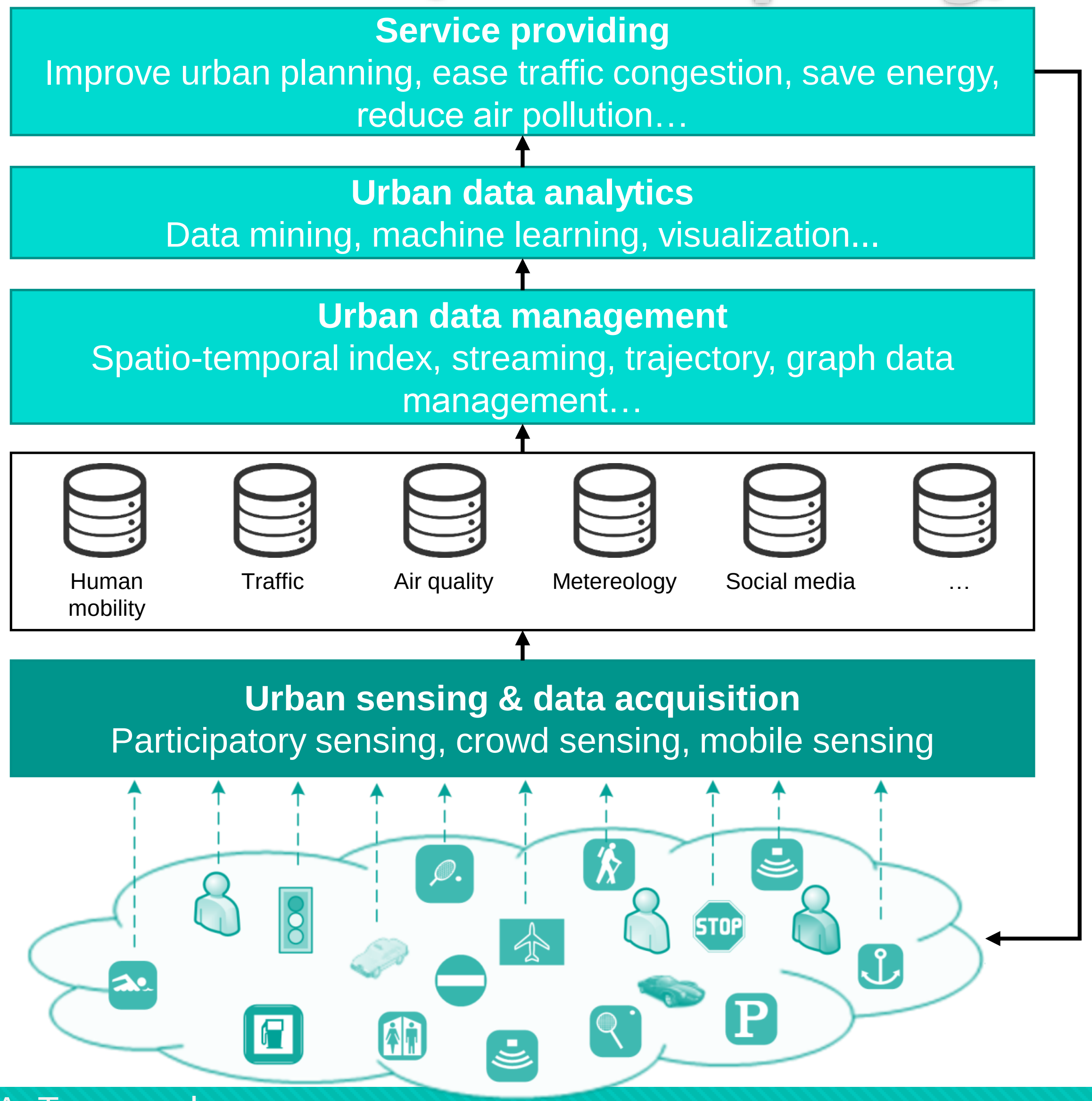
Urban data (and computing) framework



- Collects data from different sources through sensors or humans in a city.
- **Sensor-centric**: deploys a collection of sensors in fixed locations, e.g., at meteorological stations, or with moving objects.
    - Continuously send readings to a backend system without involving people in the loop, once they have been deployed
- **Human-centric**: leverages humans as sensors to probe urban dynamics when they are moving around in cities.
    - The information collected by individuals is then used to solve a problem collectively.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

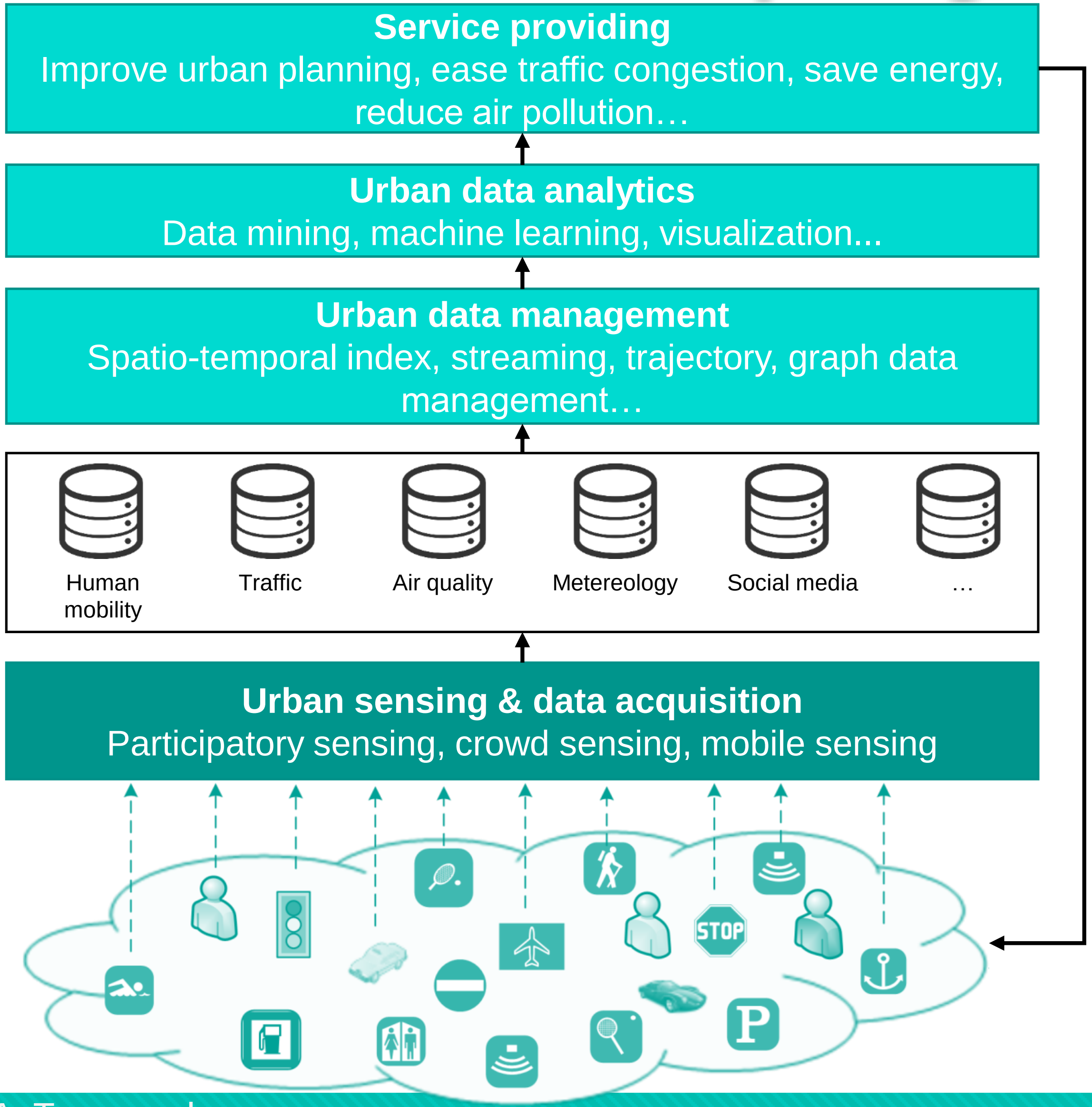


Challenges:

- Skewed sample data
- Data sparsity and missing
- Implicit and noisy data
- Resource deployment

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

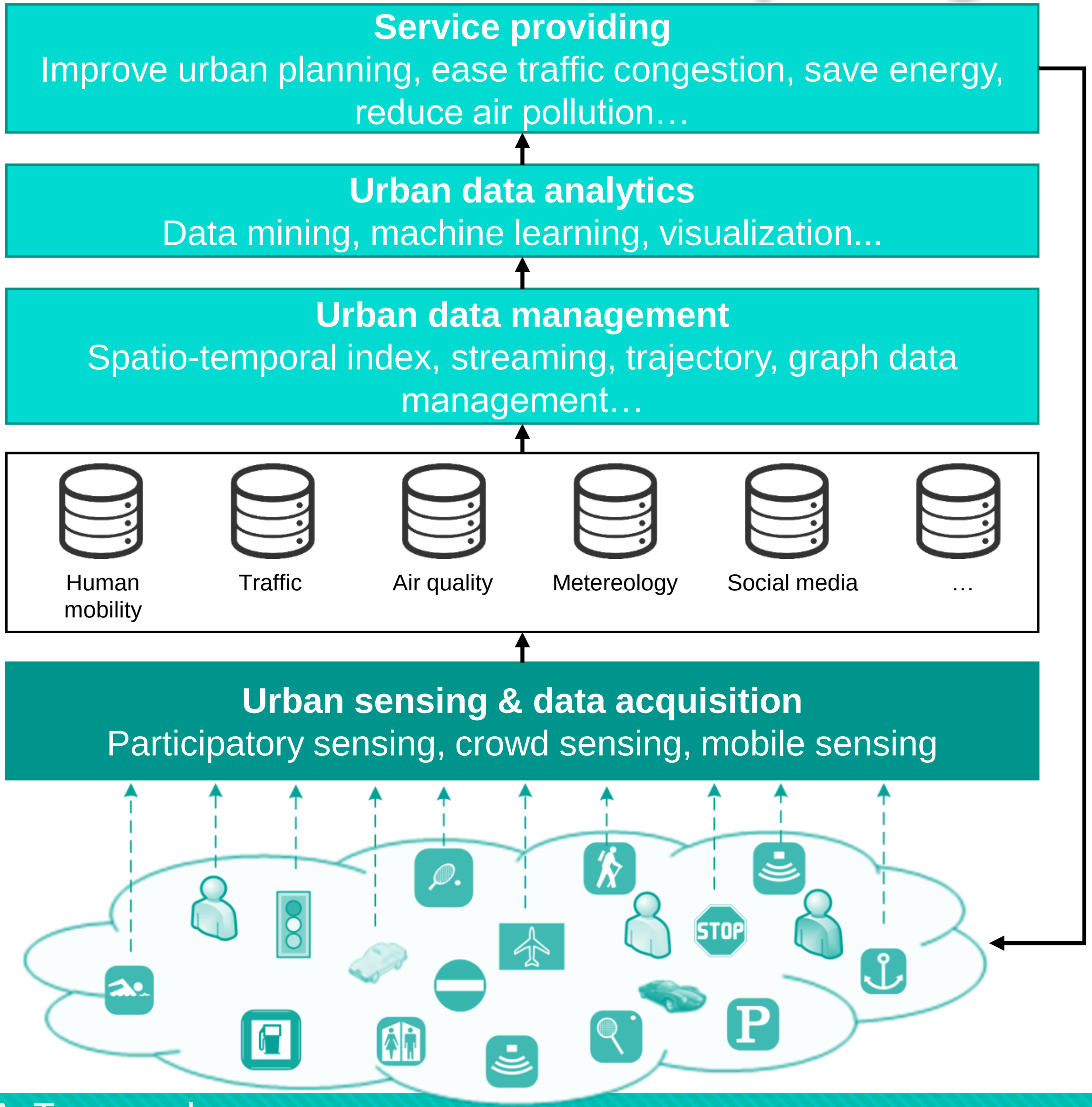


Challenges:

- **Skewed sample data**
- Data sparsity and missing
- Implicit and noisy data
- Resource deployment

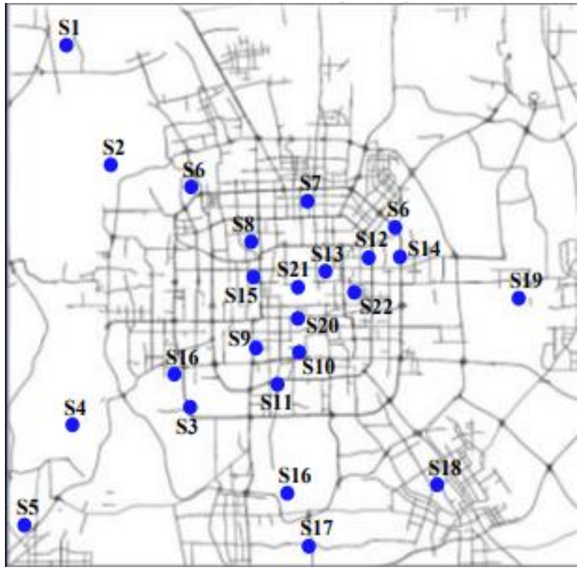
Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

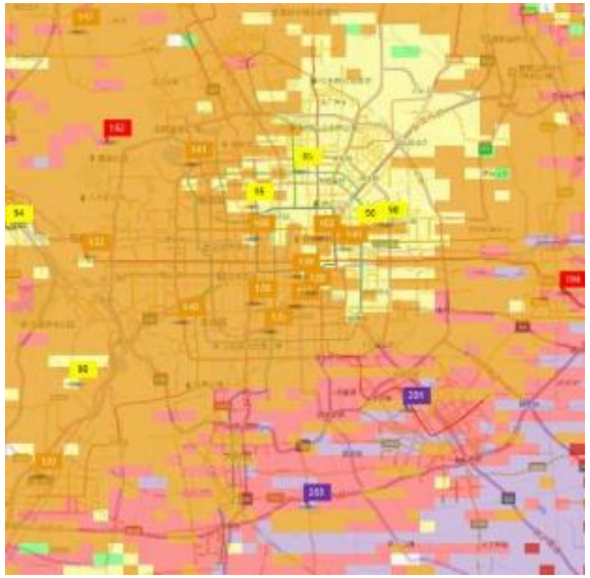


Challenges:

- Skewed sample data
- **Data sparsity and missing**
- Implicit and noisy data



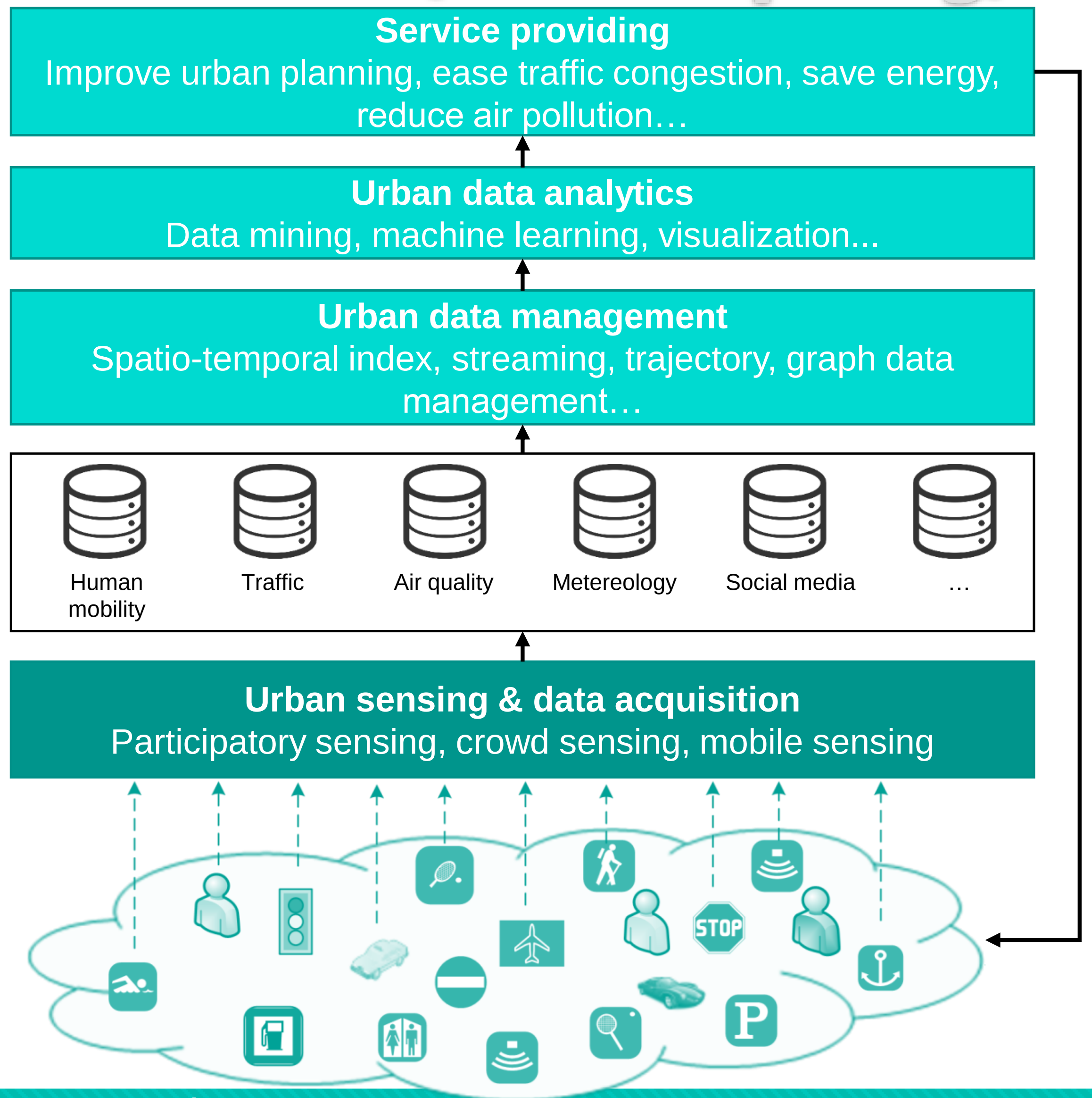
Limited air quality information



Fine-grain air quality information

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework



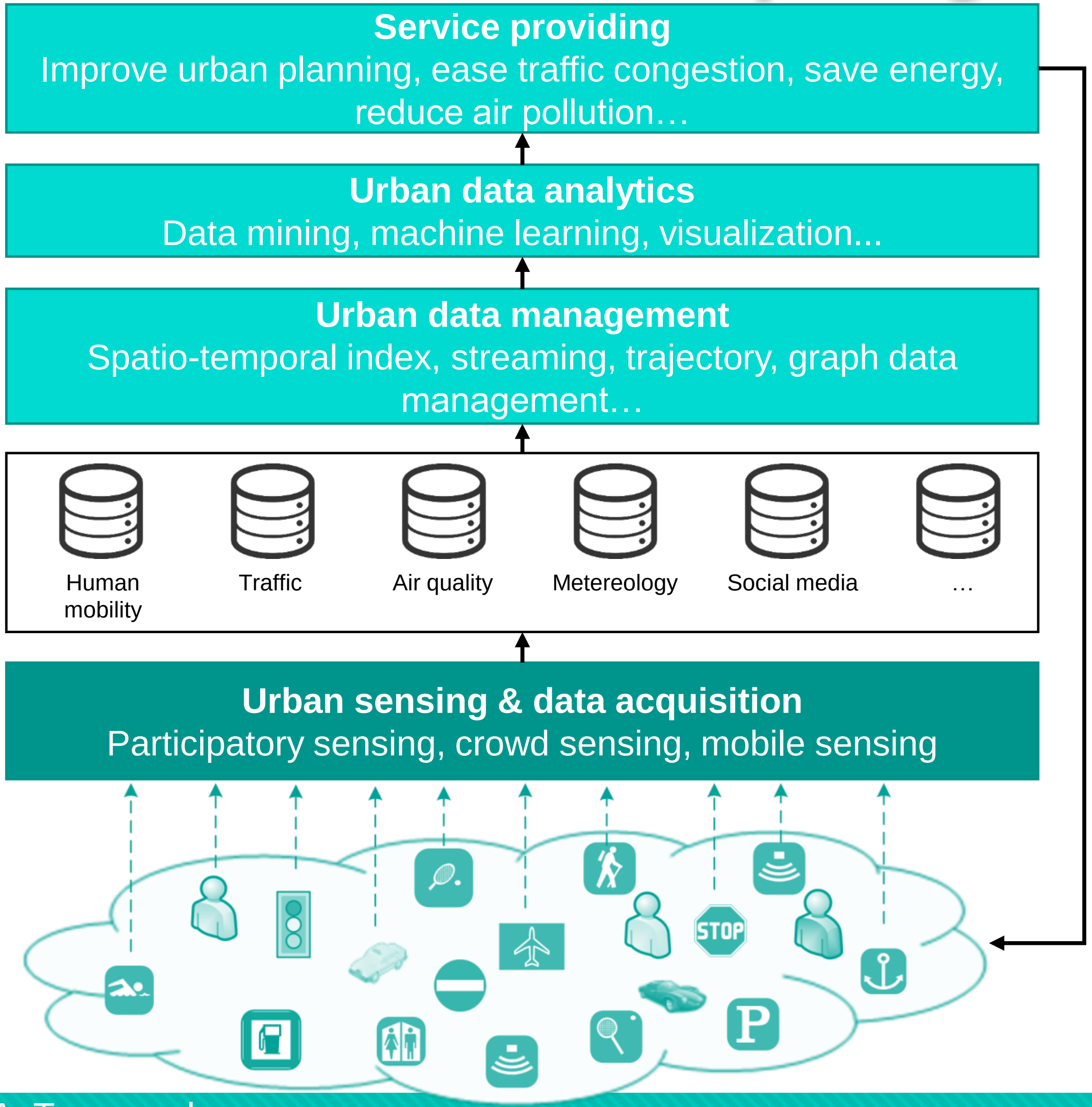
Challenges:

- Skewed sample data
- Data sparsity and missing
- **Implicit and noisy data**
- Resource deployment

- Data lost due to communication or device errors
- Data that is difficult to fill in the blanks:
  - Randomly missing and block missing.
  - Readings changing over time and location non-linearly.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

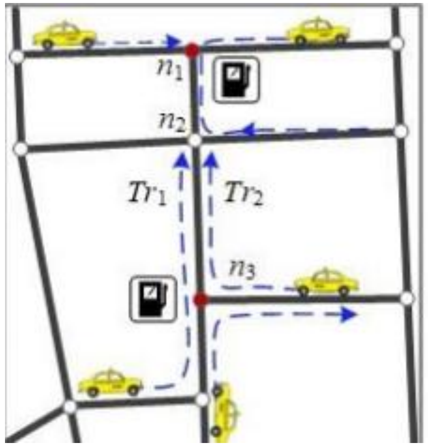
Urban data (and computing) framework



Challenges:

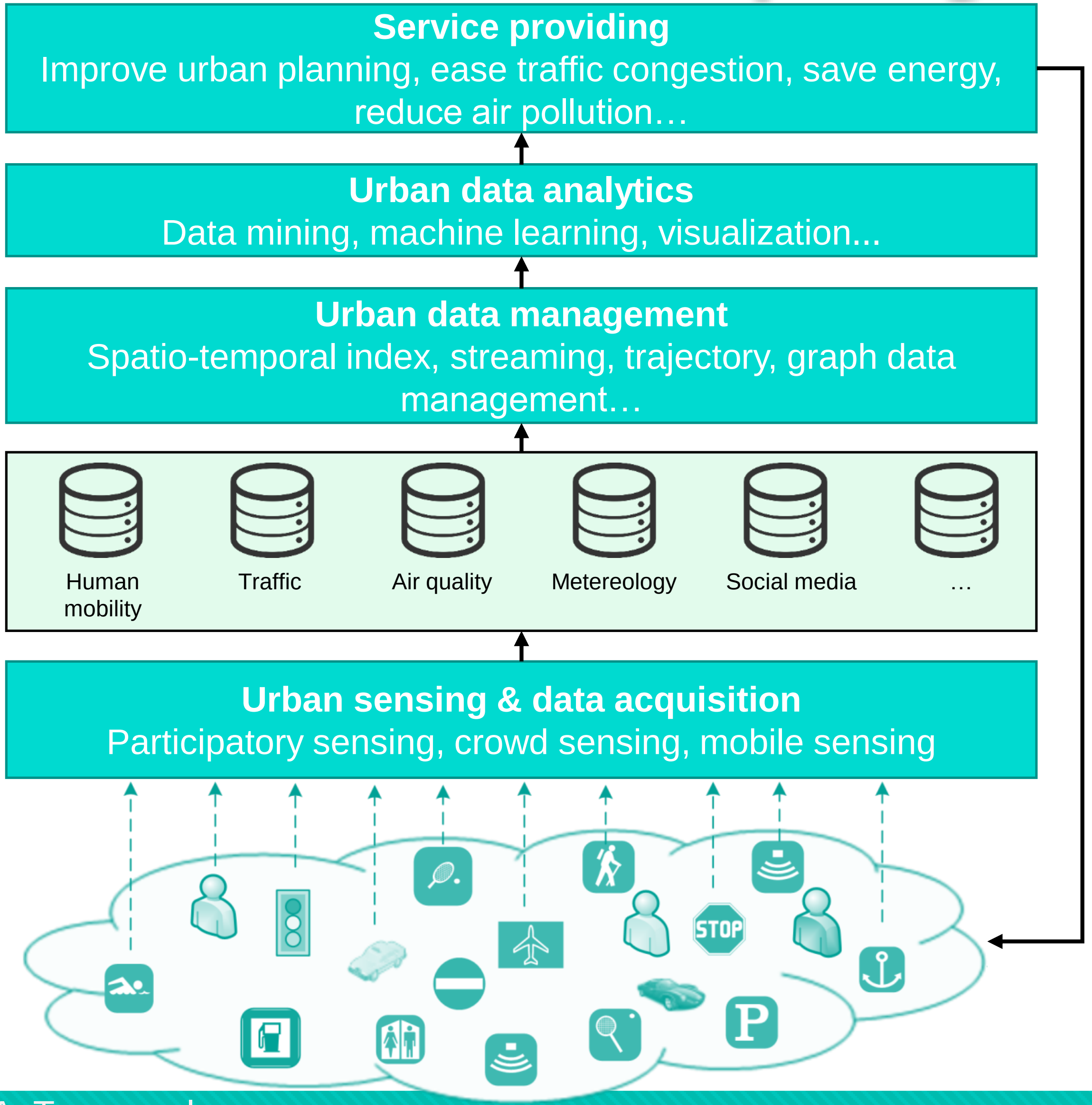
- Skewed sample data
- Data sparsity and missing
- **Implicit and noisy data**
- **Resource deployment**

- Candidate selection could be a NP-hard problem



Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework



- Why unique?
  - Data structures
  - Spatio-temporal dynamics

Spatial	Temporal
• Spatial distance • Spatial correlation • Spatial hierarchy	• Temporal closeness • Period • Trend

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

Service providing
Improve urban planning, ease traffic congestion, save energy, reduce air pollution…

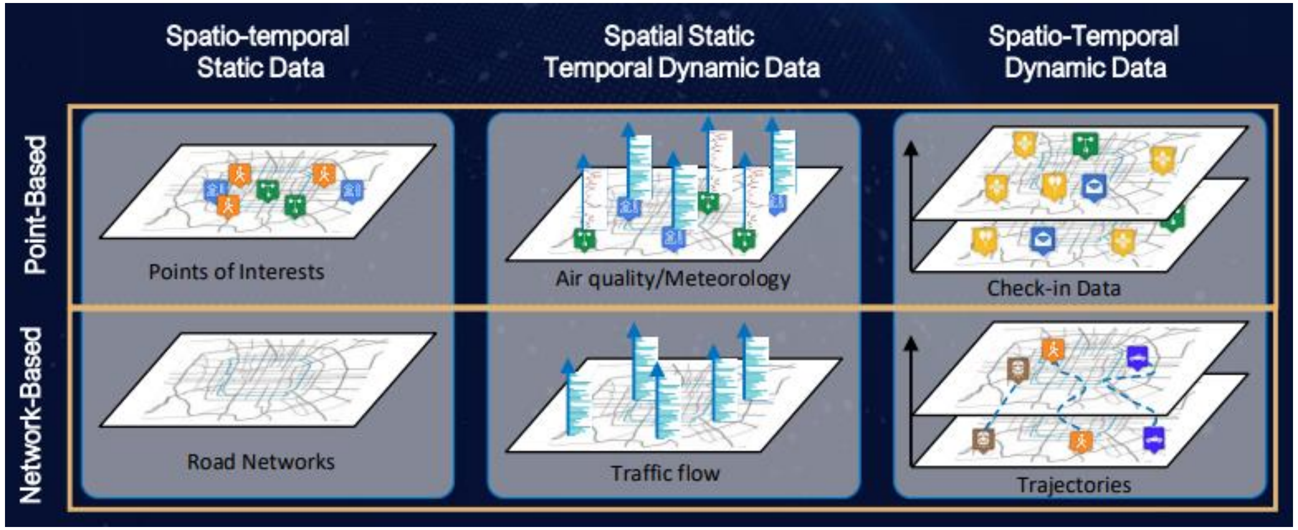
Urban data analytics
Data mining, machine learning, visualization...

Urban data management
Spatio-temporal index, streaming, trajectory, graph data management…

Human mobility | Traffic | Air quality | Metereology | Social media | …

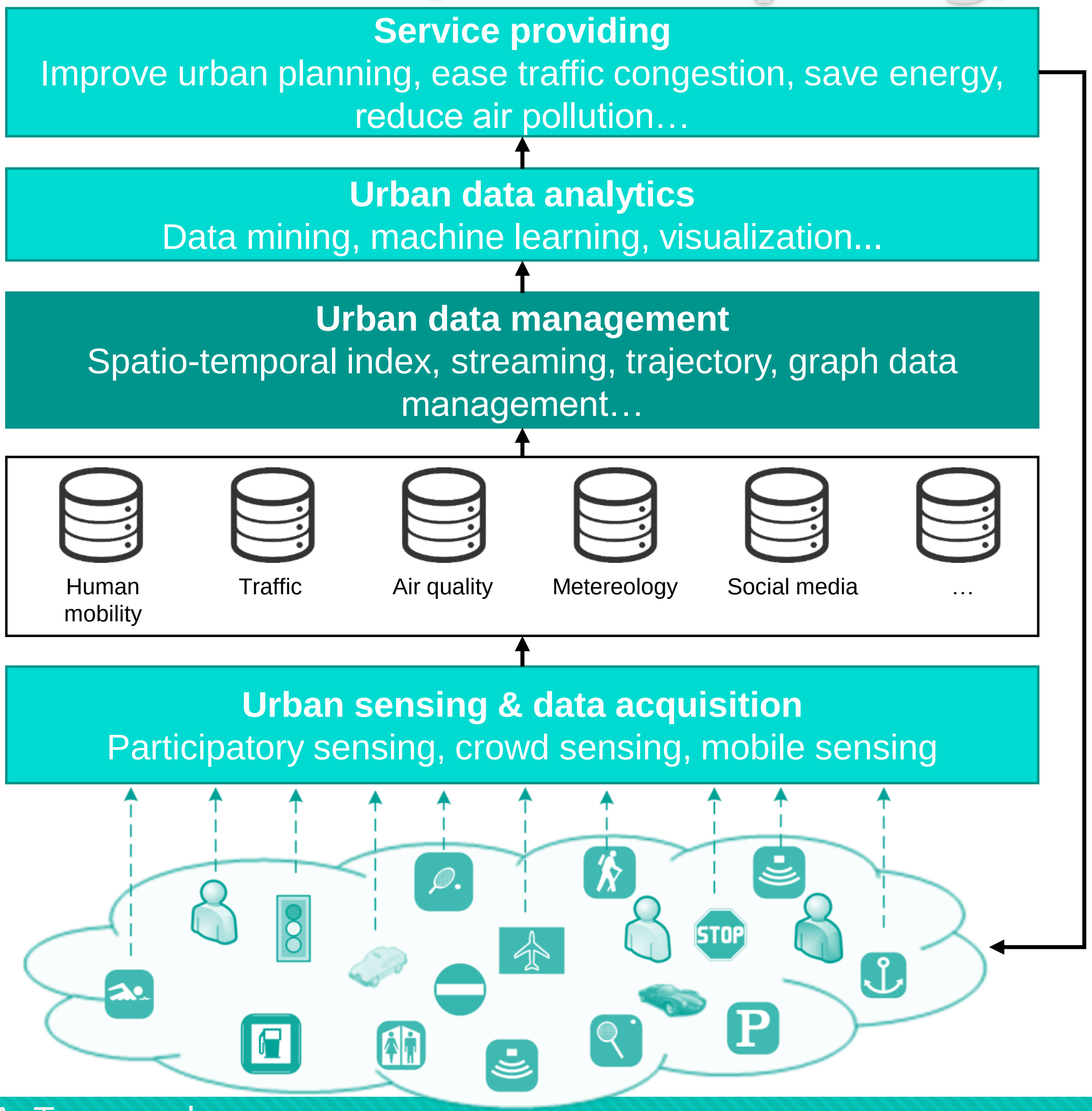
Urban sensing & data acquisition
Participatory sensing, crowd sensing, mobile sensing

- Why unique?
  - Data structures
  - Spatio-temporal dynamics



Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

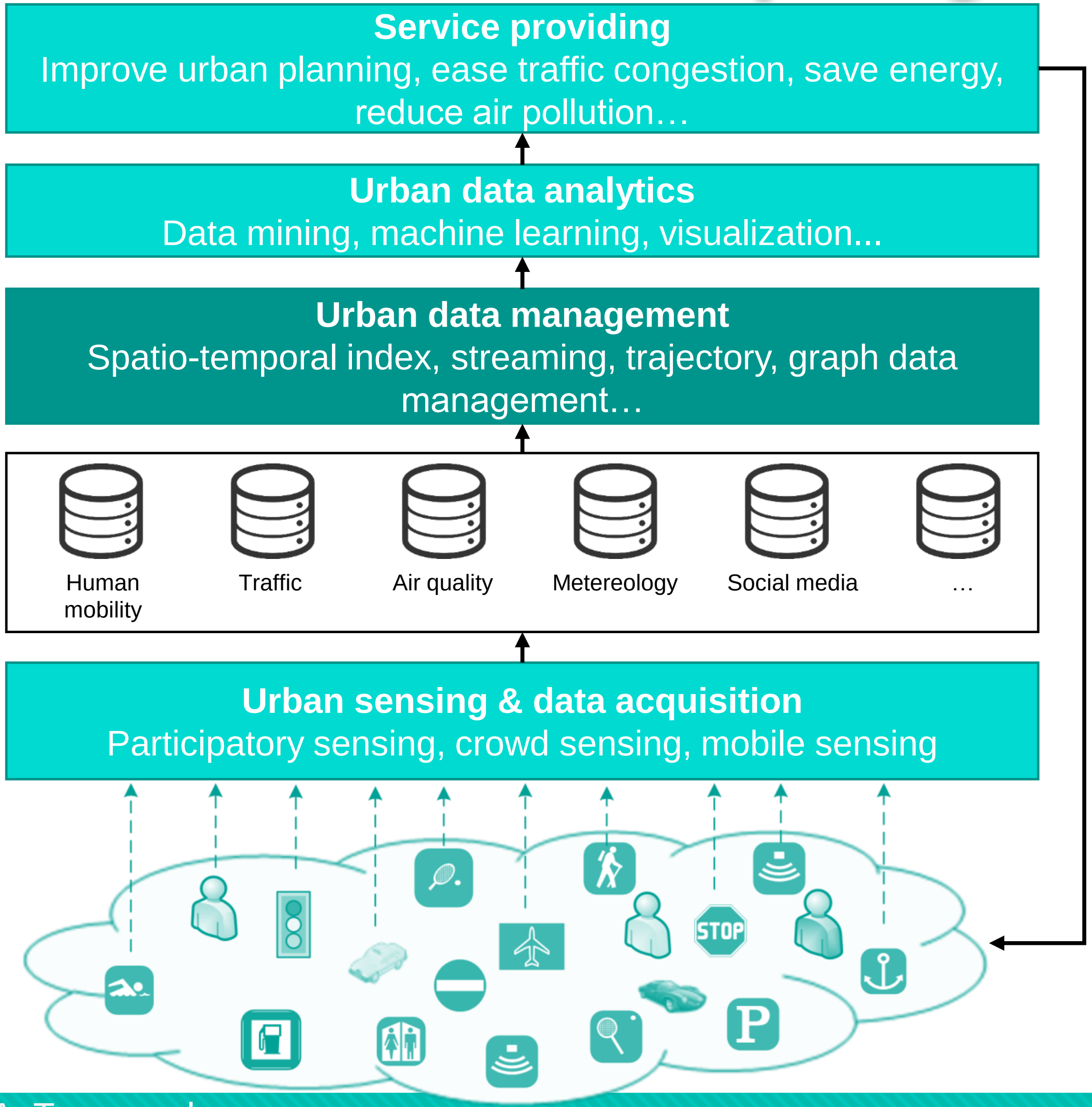
Urban data (and computing) framework



- Manages largescale and dynamic urban data, which is usually associated with a spatial coordinate and a timestamp.
- Devises different storage mechanisms for the different types of urban data.
- The human mobility and social media data are organized and indexed.
  - For multi-modality data.
  - Prepared for retrieval algorithms.
- Simultaneously incorporates spatiotemporal information and texts for supporting efficient data analytics.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

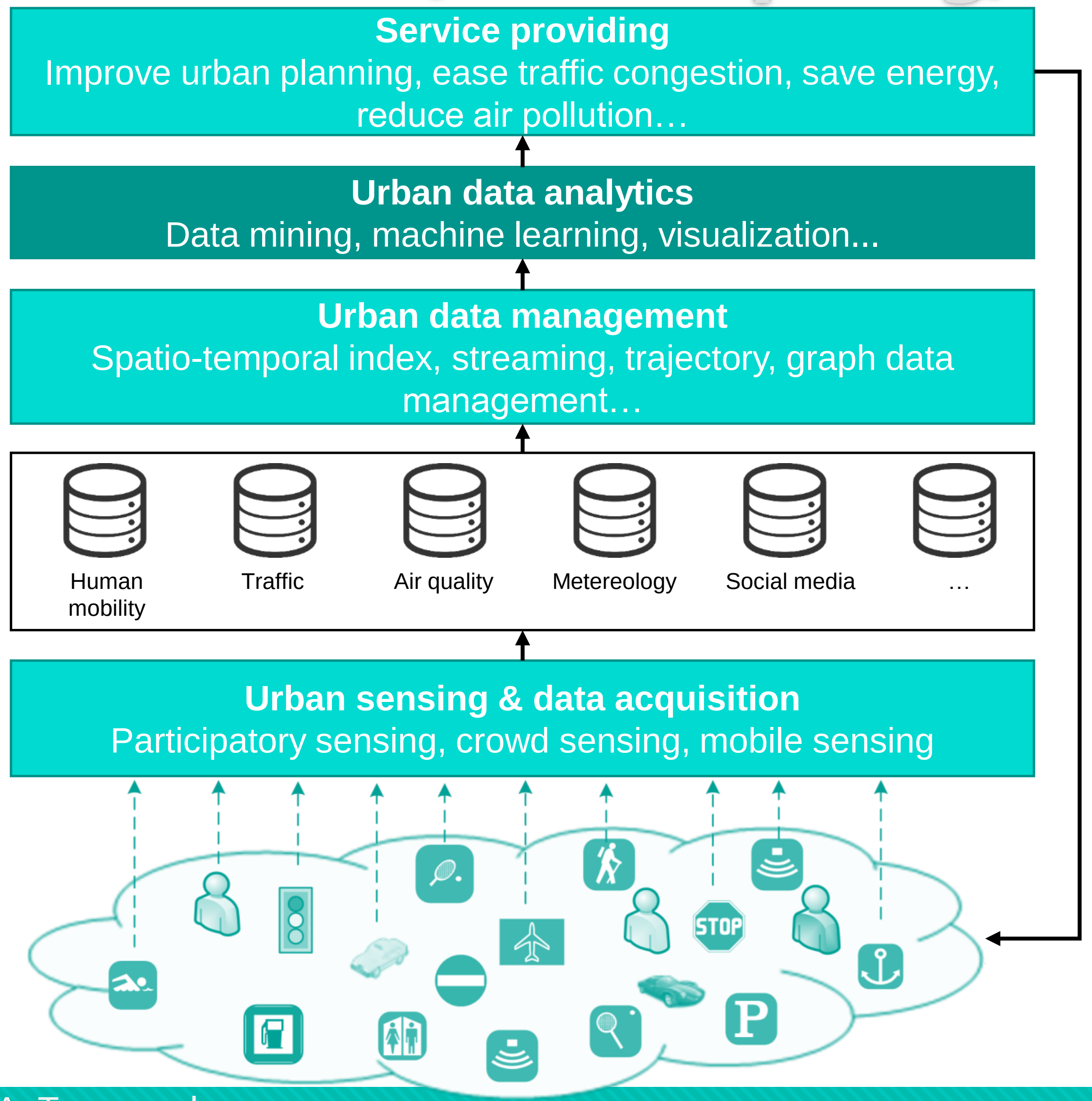


Challenges to supporting spatio-temporal data:

- The data structure of spatio-temporal data is very different from texts and images.
- Queries on spatio-temporal data are different from key words matches.
- Need to organize different datasets organically and in advance.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework



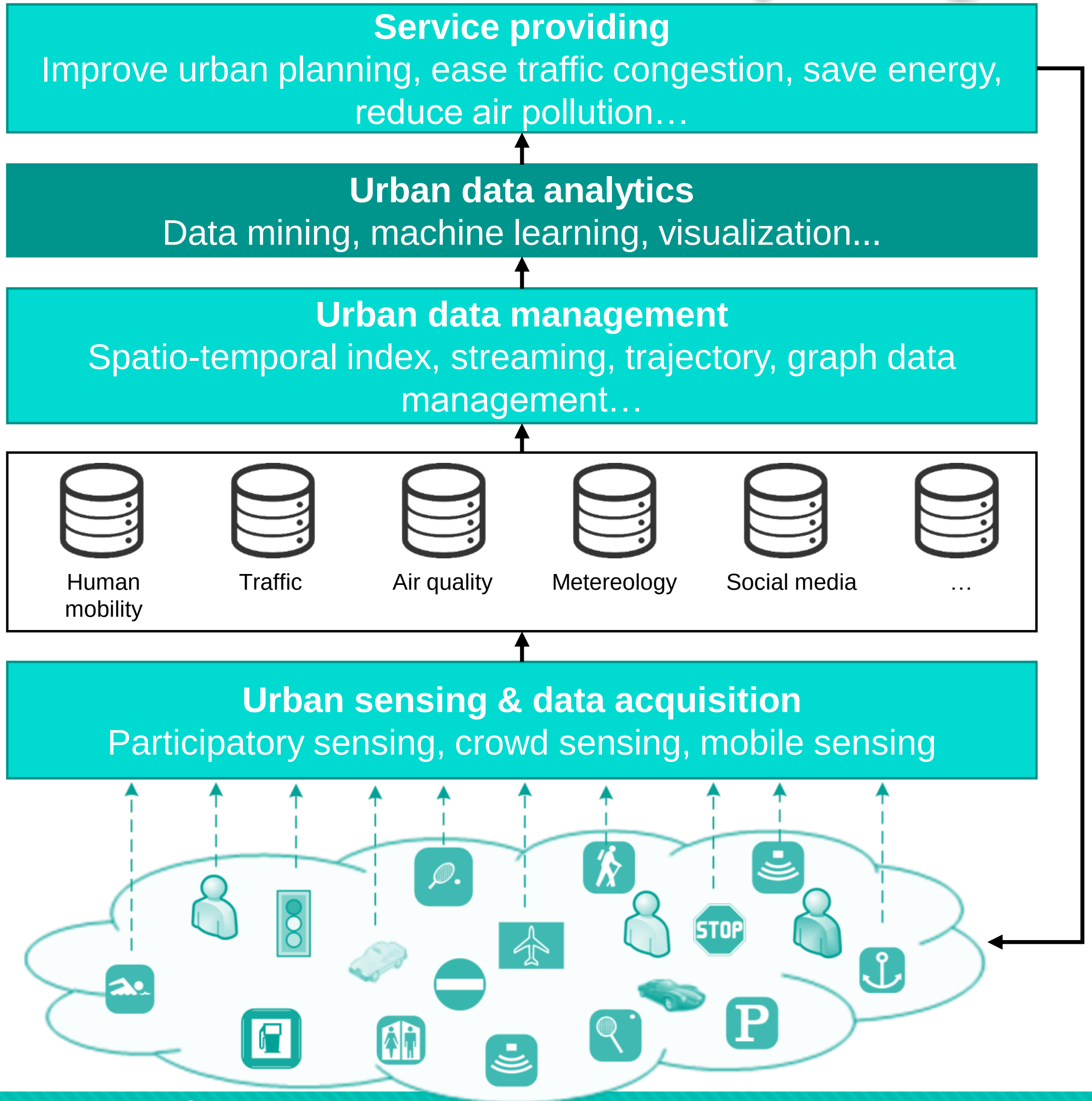
- Applies a diversity of data mining models and machine learning algorithms to the data across different domains.
- Adapts basic data mining and machine learning models, to handling spatio-temporal data.

clustering, classification, regression, anomaly detection algorithms

- Fuses the knowledge from multiple disparate datasets based on cross-domain data fusion methods.
- Database techniques + machine learning algorithms
  - “Real-time” response’

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework



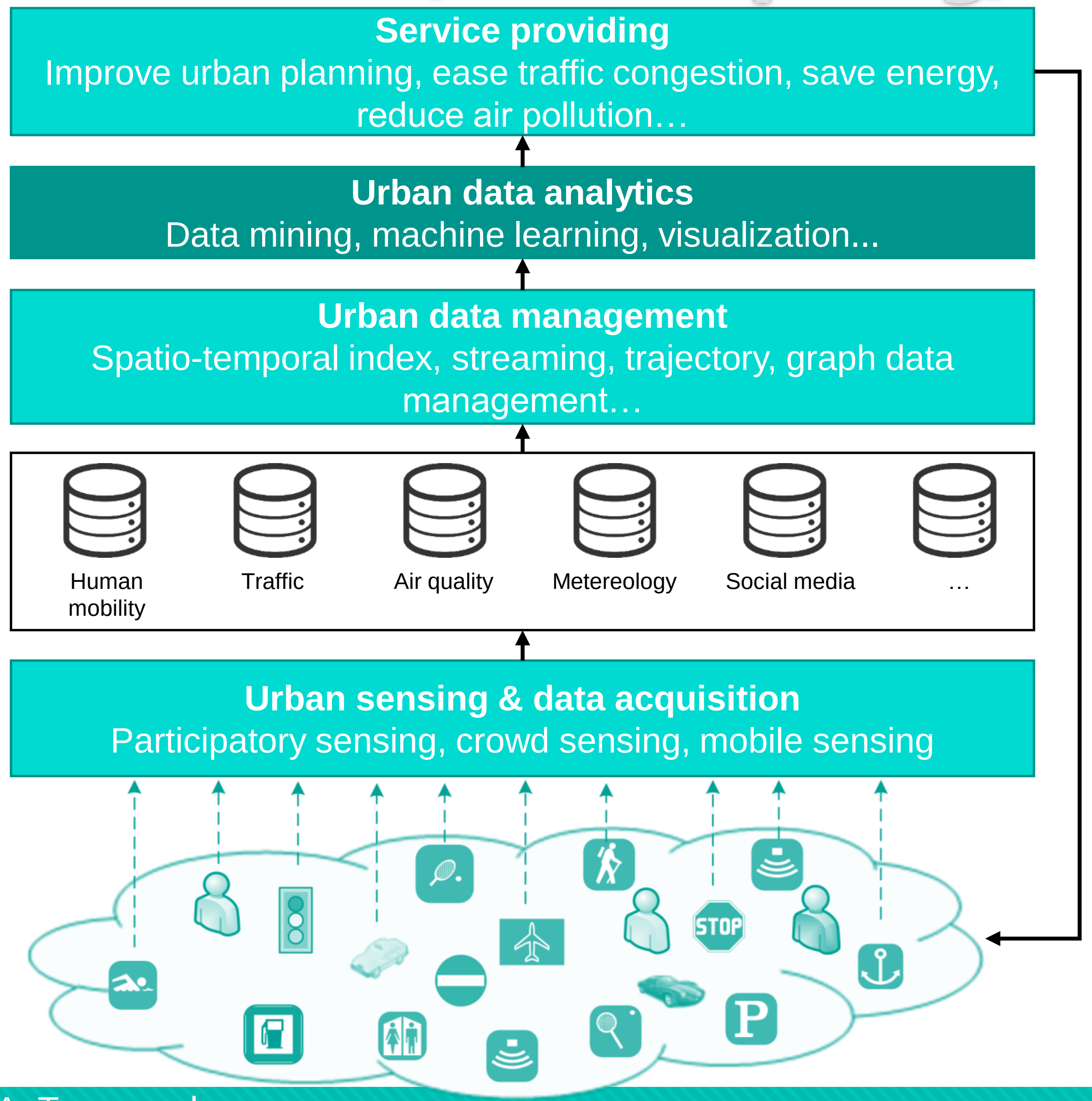
For example:

In the event of an anomaly…

- Identify the locations where people's mobility significantly differs from its origin patterns.
- Describe the anomaly by mining representative terms from the social media that is related to the locations and time span.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

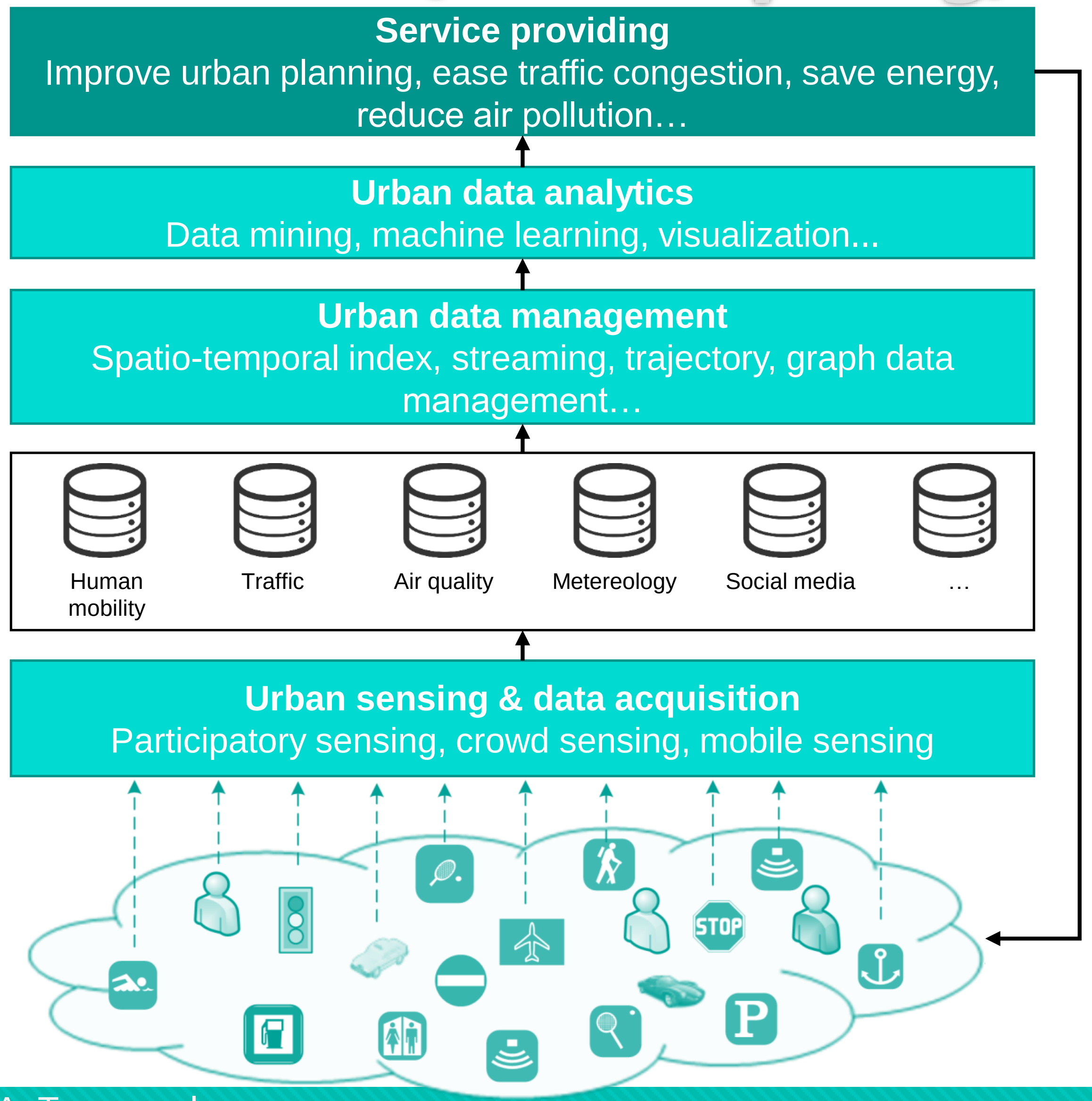


Challenges:

- Text and images != spatial and spatio-temporal data
- Mining a single data source != mining data across different domains
- Pure machine learning != visual and interactive machine learning

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

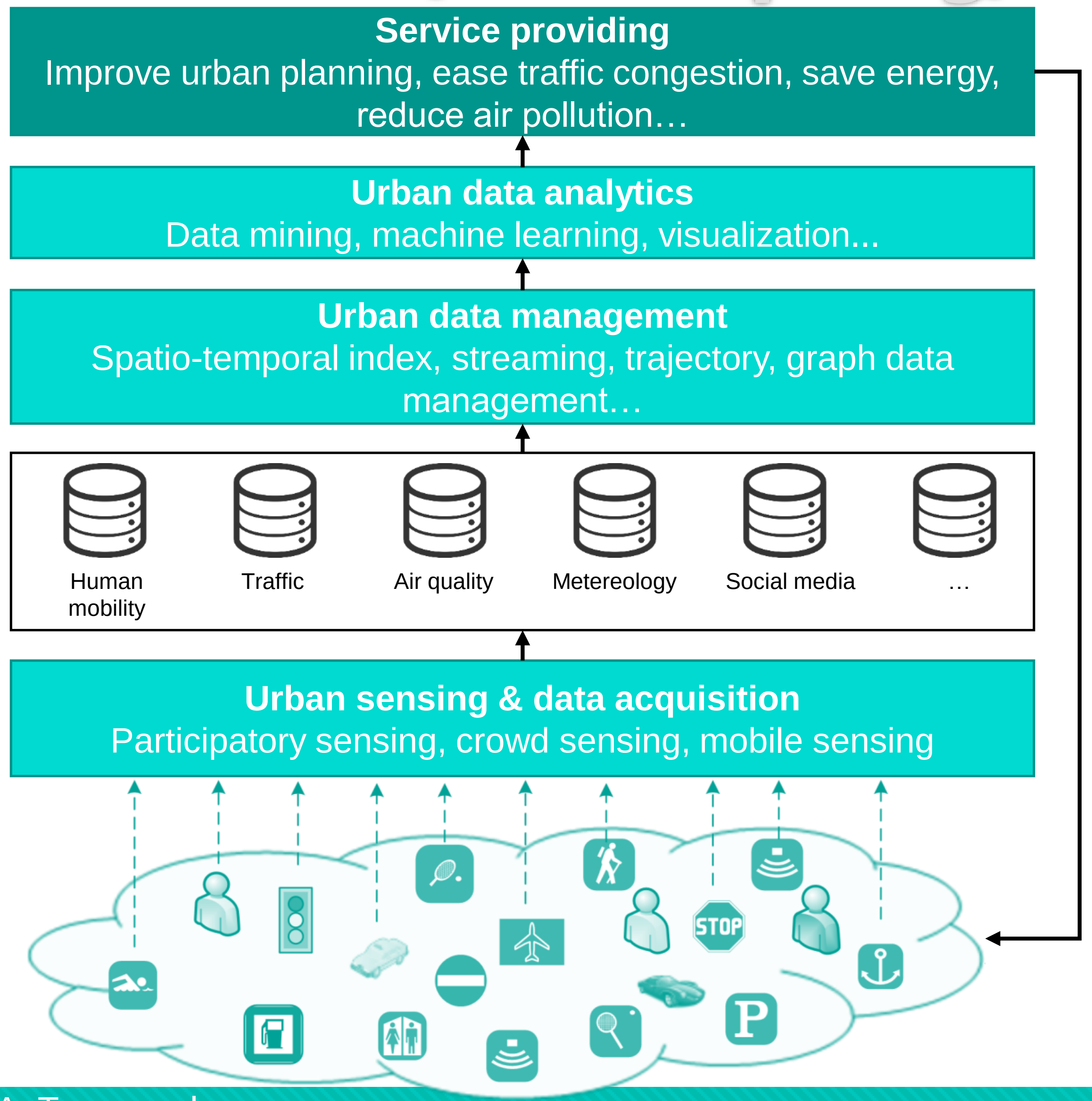


- Offers interface that allows domain systems to call the knowledge from an urban computing application, through cloud computing platforms.
- The knowledge from data must be integrated into existing domain systems to inform their decision making.
- Imperative to enable interactive visual data analytics.

combines human wisdom with machine intelligence by keeping domain experts in a learning loop

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework



- The services provided by this layer range from transportation to environmental protection, to urban planning, energy saving, social and entertainment, and public security.

- Three types of services:
  - Understanding current situation.
  - Predicting the future.
  - Diagnosing history.

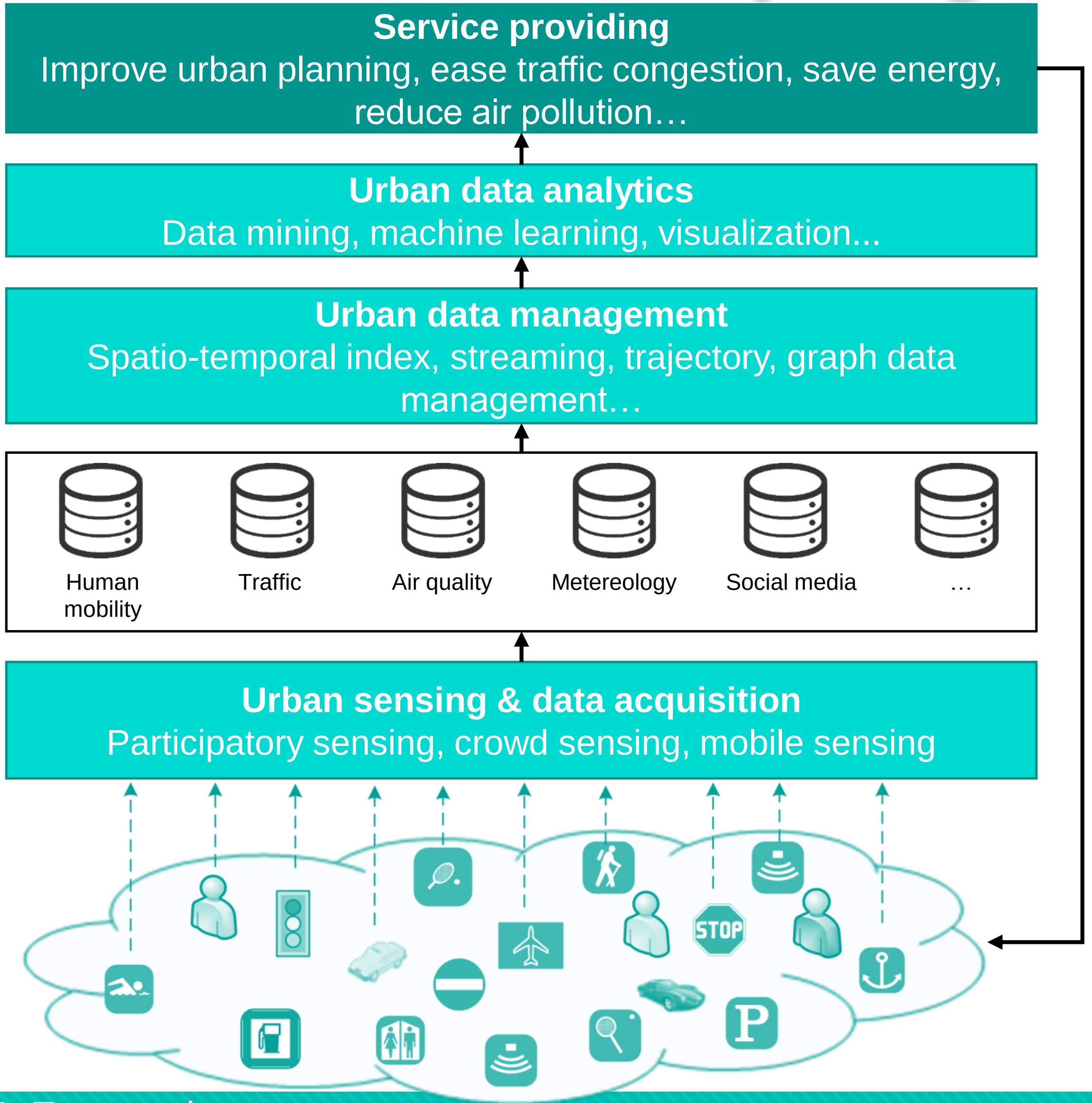
Inferring the real-time and fine-grained air quality throughout a city based on big data.

Forecasting air quality over future time.

Diagnosing the root cause of air pollution based on data accumulated over a long period.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework

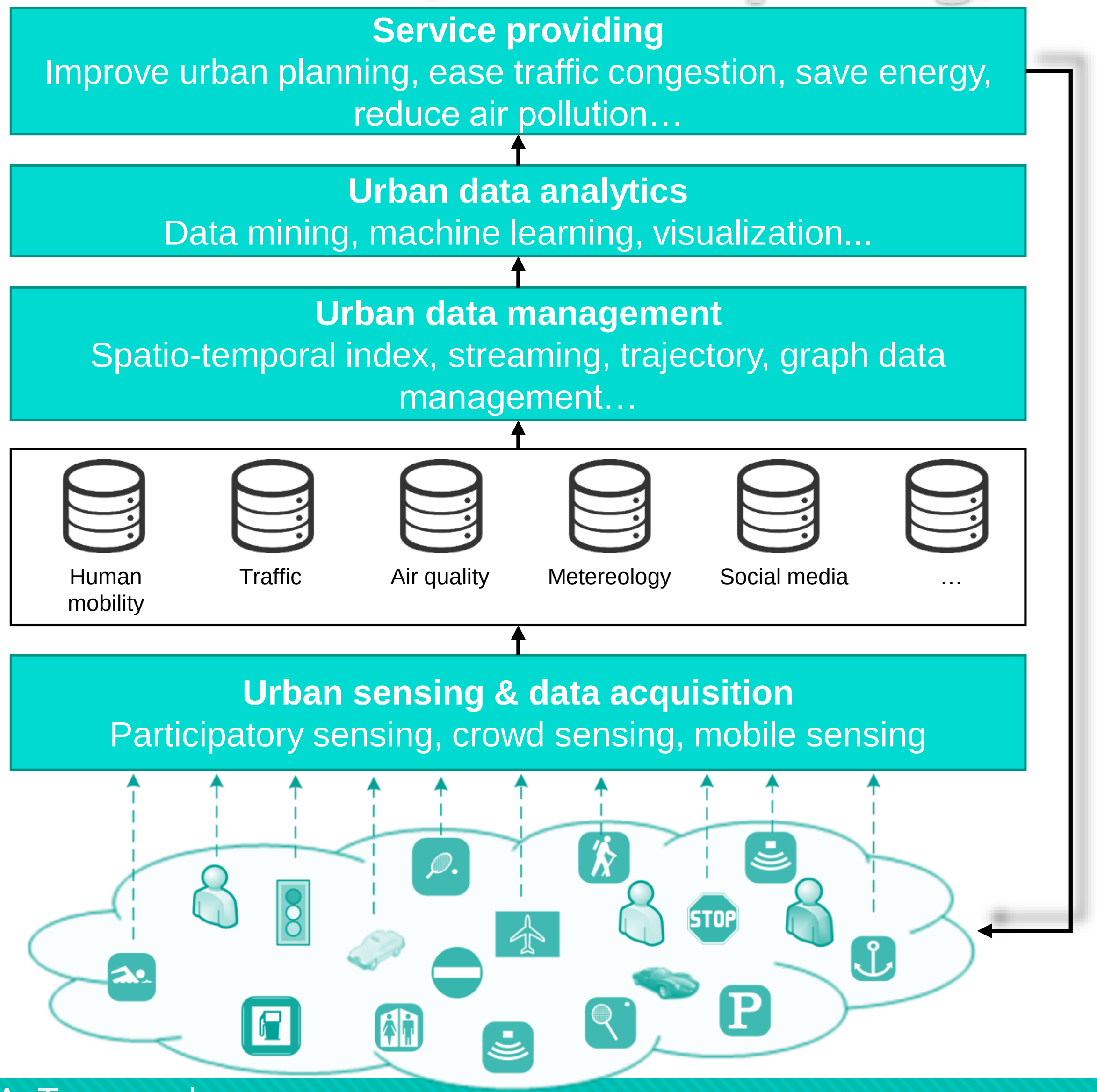


For example, traffic anomaly…

- The locations and description of the anomaly will be sent to the drivers nearby so that they can choose a bypass.
- The information will be delivered to the transportation authority for dispersing traffic and diagnosing the anomaly.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

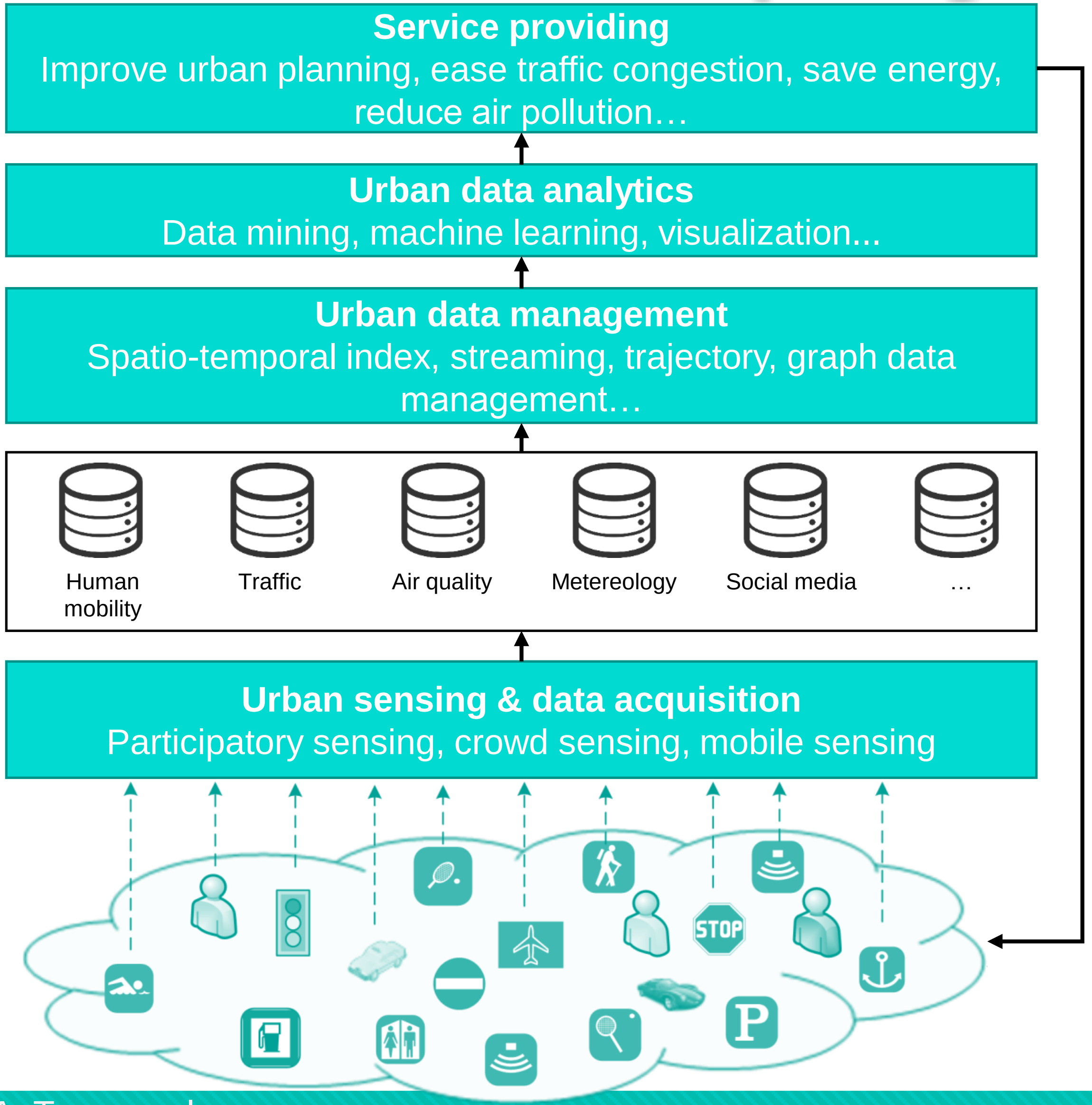
Urban data (and computing) framework



The system continues the loop for an instant and unobtrusive detection of urban anomalies

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

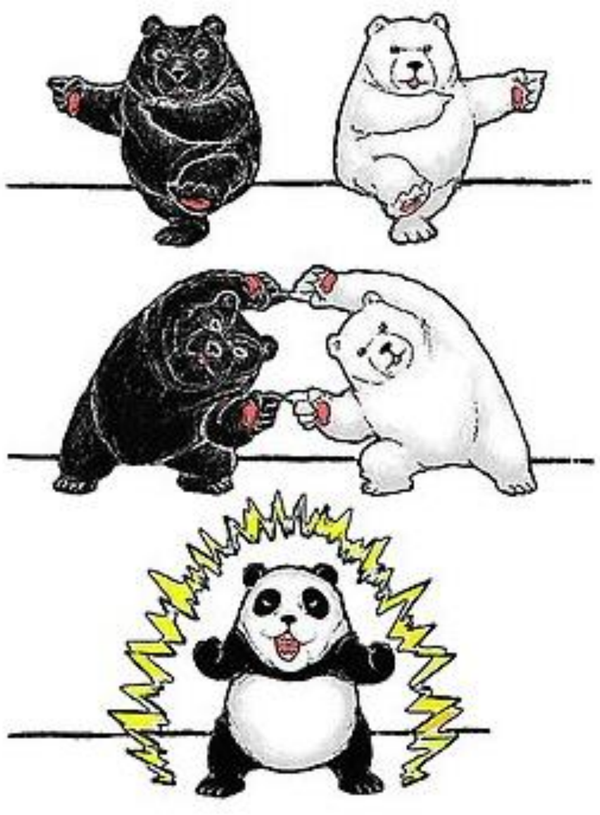
Urban data (and computing) framework



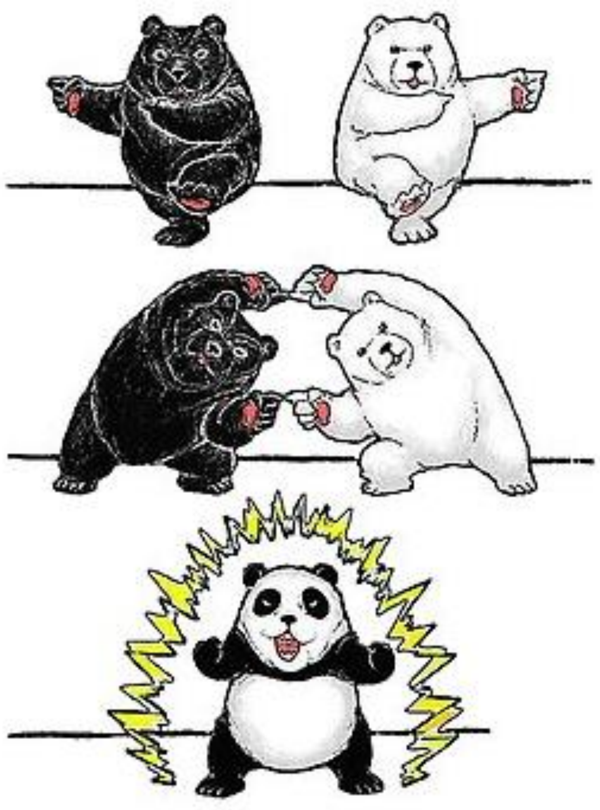
- Compared with other systems that are based on a single (modal)-data/single-task framework, urban computing holds a multi(modal)-data/multitask framework.
- Different tasks can be fulfilled by combining different data sources with different data acquisition, management, and analytics techniques from different layers of the framework.

Zheng, Y., et al. Urban Computing: concepts, methodologies, and applications. *ACM transactions on Intelligent Systems and Technology*.

Urban data (and computing) framework



Urban data (and computing) framework



Data fusion!

Table of contents

Data fusion

Traditional data mining usually deals with data from a single domain.

In the big data era, we face a diversity of datasets from different sources in **different domains**, consisting of **multiple modalities**:

- Representation.
- Distribution.
- Scale.
- Density.

How to unlock the power of knowledge from multiple disparate (but potentially connected) datasets?

Treating different datasets equally?

Concatenating the features from disparate datasets?

Data fusion

Traditional data mining usually deals with data from a single domain.

In the big data era, we face a diversity of datasets from different sources in **different domains**, consisting of **multiple modalities**:

- Representation.
- Distribution.
- Scale.
- Density.

How to unlock the power of knowledge from multiple disparate (but potentially connected) datasets?

Treating differe tasets equally?

Concatenating the featur om disparate datasets?

NO!

Data fusion

Traditional data mining usually deals with data from a single domain.

In the big data era, we face a diversity of datasets from different sources in **different domains**, consisting of **multiple modalities**:

- Representation.
- Distribution.
- Scale.
- Density.

How to unlock the power of knowledge from multiple disparate (but potentially connected) datasets?

Treating differe[illegible]tasets equally? Concatenating the featur[illegible]om disparate datasets?

NO!

Use advanced data fusion techniques that can fuse **knowledge** from various datasets organically in a machine learning and data mining task

Data fusion

Data fusion defines a **combination of multiple sources** to obtain **improved information**

less expensive, **higher quality**, or **more relevant information**.

Data fusion = Information fusion

Data fusion

Data fusion defines a **combination of multiple sources** to obtain **improved information**

less expensive, **higher quality**, or **more relevant information**.

Data fusion	≠	Information fusion
Defines raw data (obtained directly from the sensors)		Defines already processed data

Data fusion

Data fusion defines a **combination of multiple sources** to obtain **improved information**

less expensive, **higher quality**, or **more relevant information**.

Data fusion < Information fusion

Higher semantic level

Data fusion

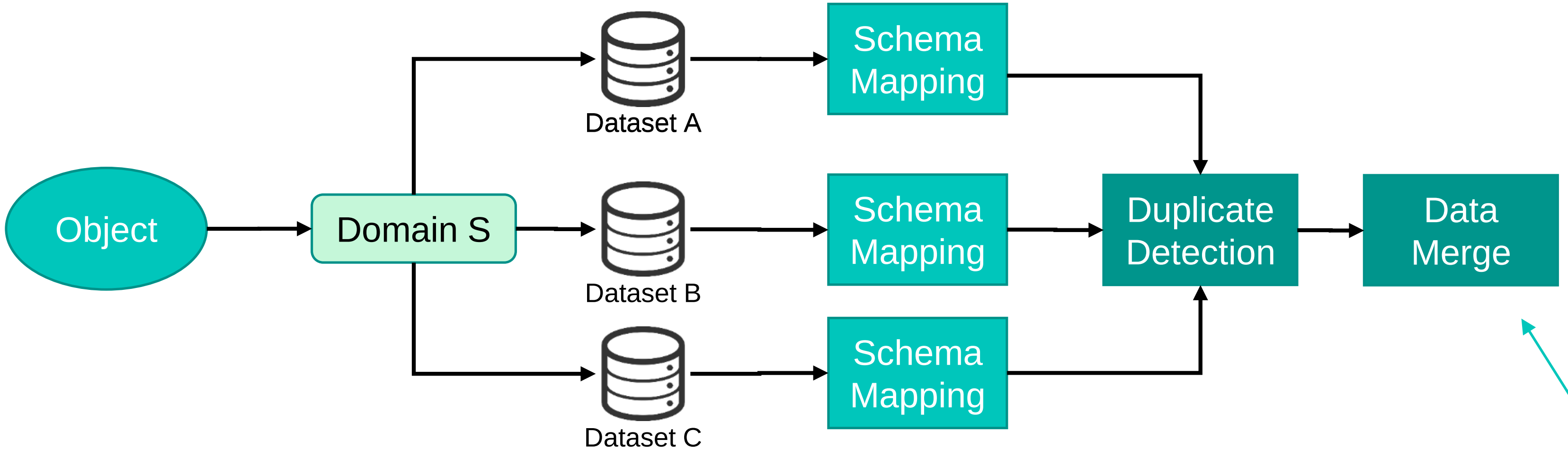
Data fusion defines a **combination of multiple sources** to obtain **improved information**

less expensive, **higher quality**, or **more relevant information**.

Decision fusion	Data combination	Data aggregation	Multi-sensor data fusion	Sensor fusion

Data fusion vs Traditional data integration

- Conventional data integration is a process of integration of multiple data representing the **same real-world object** into a consistent, accurate, and useful representation.



same domain!

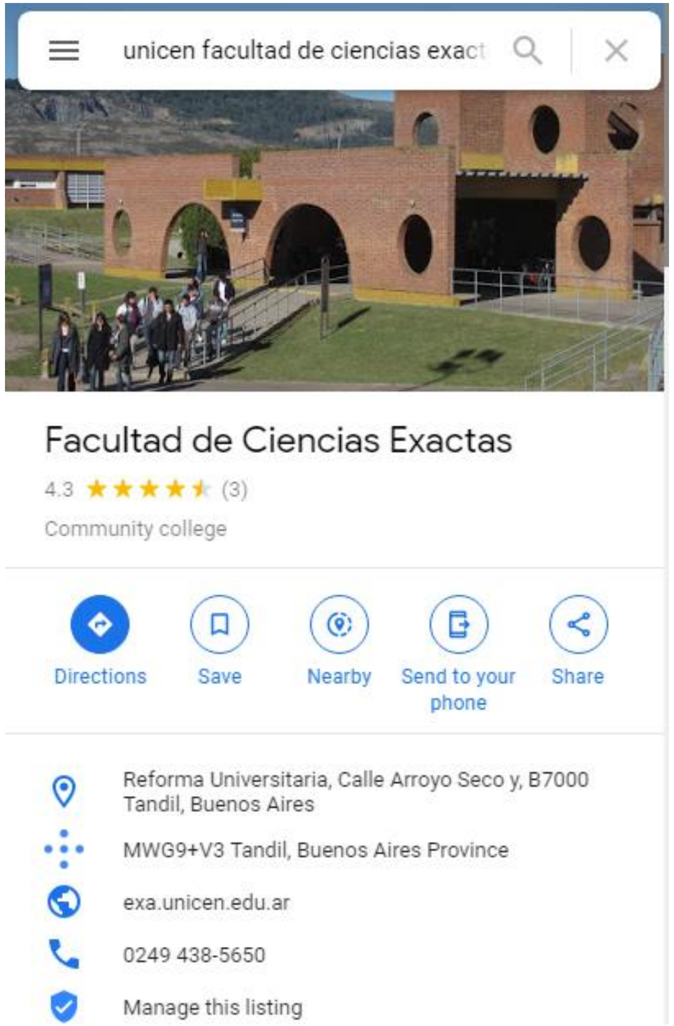
For example, there are two POI datasets generated from diverse sources (e.g. Facebook, Foursquare, Google Maps).

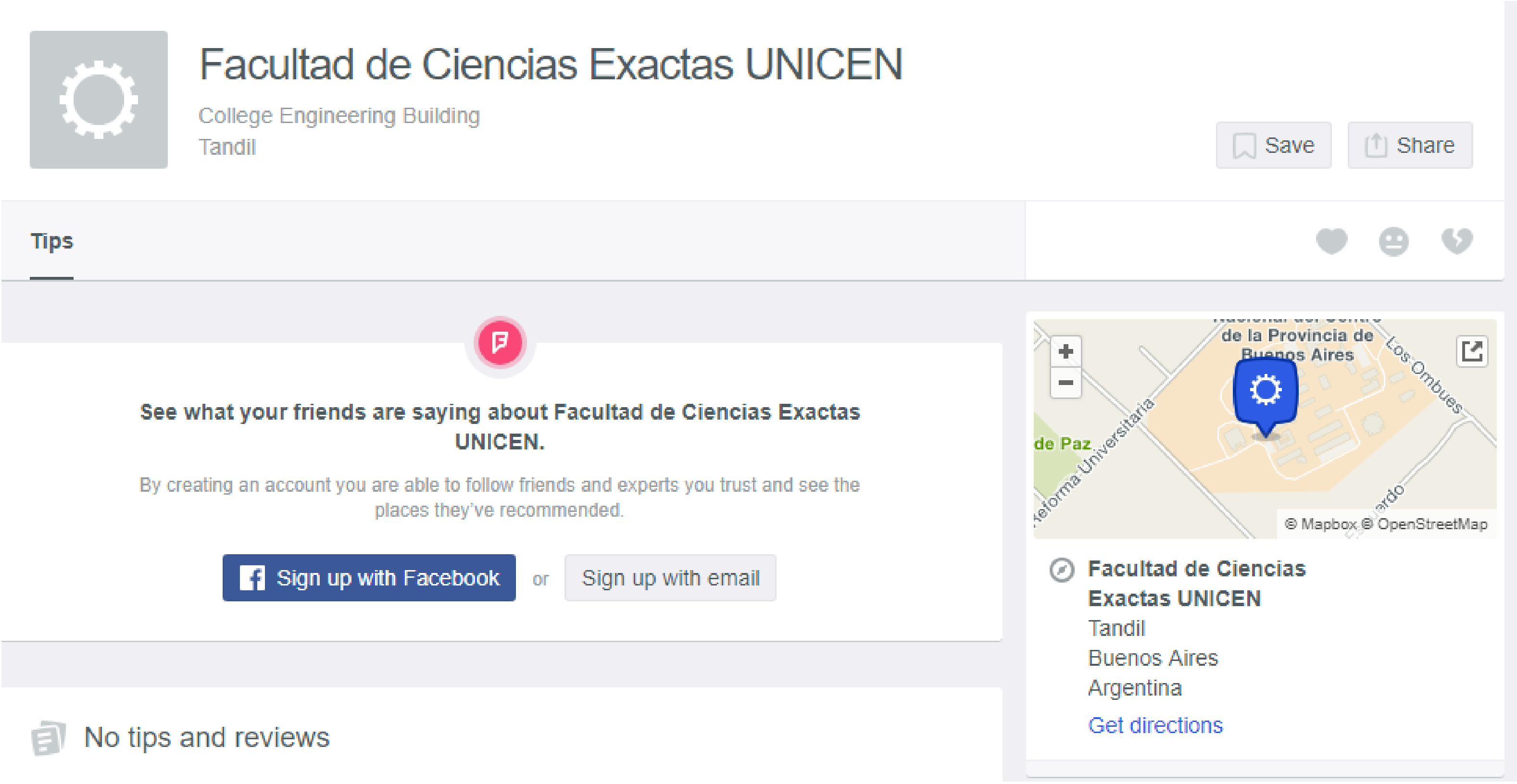
The records describing the same POI will be merged and each POI will be represented by a single POI.

a dataset with a consistent data schema

Let's forget about sensors for a moment

Data fusion vs Traditional data integration





Data fusion vs Traditional data integration

- In the era of big data there are multiple datasets generated in different domains.

Data fusion vs Traditional data integration

- In the era of big data there are multiple datasets generated in different domains.

Implicitly connected by a latent object

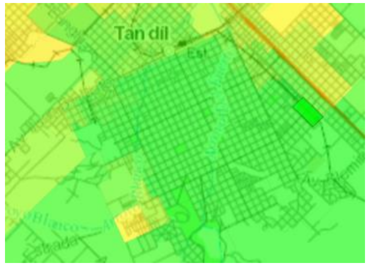
Data fusion vs Traditional data integration

- In the era of big data there are multiple datasets generated in different domains.

Implicitly connected by a latent object

traffic conditions

POIs



demography

city

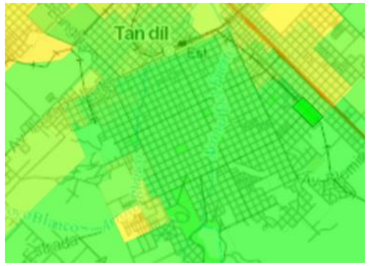
Data fusion vs Traditional data integration

- In the era of big data there are multiple datasets generated in different domains.

Implicitly connected by a latent object

traffic conditions

POIs



demography

city

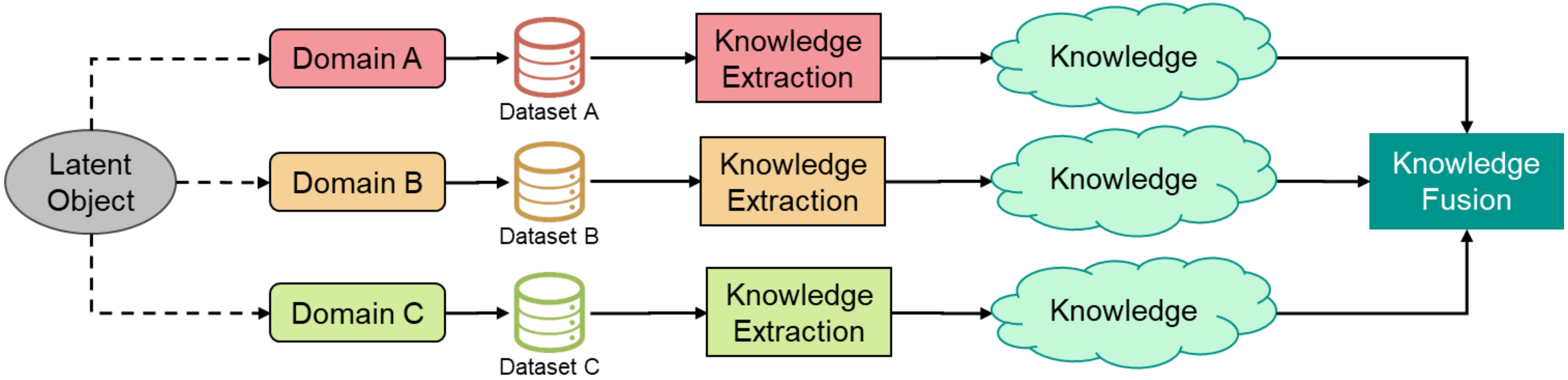
- Different domains.
- Describe different objects.

Data fusion vs Traditional data integration

- In the era of big data there are multiple datasets generated in different domains.

Implicitly connected by a latent object

- **Cannot** merge them straightforwardly by a schema mapping and duplication detection.
- Instead, **extract knowledge** from each dataset by different methods, **fusing the knowledge** from them organically to understand a region's **function collectively**.
- **Knowledge fusion** instead of schema mapping.



Data fusion vs Heterogeneous information network

- An information network represents an abstraction of the real world, focusing on **objects and interactions** between objects.

- Departing from many existing network models that view interconnected data as homogeneous graphs or networks, a heterogeneous information network consists of **nodes and relations of different types**.
  - Can be constructed in any domain.

- This level of abstraction has great power in:
  - **Representing and storing** essential information about the real-world.
  - **Providing** a useful tool to mine knowledge from it.

Data fusion vs Heterogeneous information network

- An information network represents an abstraction of the real world, focusing on **objects and interactions** between objects.

- Departing from many existing network models that view interconnected data as homogeneous graphs or networks, a heterogeneous information network consists of **nodes and relations of different types**.
  - Can be constructed in any domain.

- This level of abstraction has great power in:
  - **Representing and storing** essential information about the real-world.
  - **Providing** a useful tool to mine knowledge from it.

- **However, it only links the object in a single domain rather than data across different domains.**

In a bibliographic information network, people, papers, and conferences are all from a bibliographic domain.

In a Twitter information network, users, images, tags, and comments are all from a social media domain.

Data fusion vs Heterogeneous information network

- An information network represents an abstraction of the real world, focusing on **objects and interactions** between objects.

- Departing from many existing network models that view interconnected data as homogeneous graphs or networks, a heterogeneous information network consists of **nodes and relations of different types**.
  - Can be constructed in any domain.

- This level of abstraction has great power in:
  - **Representing and storing** essential information about the real-world.
  - **Providing** a useful tool to mine knowledge from it.

- **However, it only links the object in a single domain rather than data across different domains.**

- To fuse data across totally different domains such heterogeneous networks may not be able to find explicit links with semantic meanings between objects of different domains.

Data fusion vs Heterogeneous information network

- An information network represents an abstraction of the real world, focusing on **objects and interactions** between objects.
- Departing from many existing network models that view interconnected data as homogeneous graphs or networks, a heterogeneous information network consists of **nodes and relations of different types**.
  - Can be constructed in any domain.
- This level of abstraction has great power in:
  - **Representing and storing** essential information about the real-world.
  - **Providing** a useful tool to mine knowledge from it.
- **However, it only links the object in a single domain rather than data across different domains.**
- To fuse data across totally different domains such heterogeneous networks may not be able to find explicit links with semantic meanings between objects of different domains.

How do social media and traffic
or air quality data relate?

Data fusion vs Heterogeneous information network

- An information network represents an abstraction of the real world, focusing on **objects and interactions** between objects.

- Departing from many existing network models that view interconnected data as homogeneous graphs or networks, a heterogeneous information network consists of **nodes and relations of different types**.
  - Can be constructed in any domain.

- This level of abstraction has great power in:
  - **Representing and storing** essential information about the real-world.
  - **Providing** a useful tool to mine knowledge from it.

- **However, it only links the object in a single domain rather than data across different domains.**

- To fuse data across totally different domains such heterogeneous networks may not be able to find explicit links with semantic meanings between objects of different domains.

Consequently, algorithms proposed for mining heterogeneous information networks **cannot** be applied to **cross-domain data fusion** directly.

Classification of techniques

Relations between the data sources

Abstraction levels

Type of architecture

(Amongst the most common ones)

Relationship between the data sources

Complementary

- The information provided by the input sources represents different parts of the scene and could thus be used to obtain more complete global information.
- E.g. Information provided by two cameras with different fields of view.

Redundant

- Two or more input sources provide information about the same target and could thus be fused to increment the confidence.
- E.g. Data coming from overlapped areas in visual sensor networks.

Cooperative

Information that is typically more complex than the original information.

E.g. Multi-modal (audio and video) data fusion.

Abstraction levels

Signal level

- Directly addresses the signals that are acquired from the sensors.

Pixel level

- Operates at the image level.
- Could be used to improve image processing tasks.

- E.g. NN kernels applied to extract specific characteristics of pixels in images.

Characteristic

- Employs features that are extracted from the images or signals.

- E.g. shape, velocity.

Symbol

- Information is represented as symbols.
- Also known as the decision level.

Abstraction levels II

Low level

- Raw data is directly provided as an input to the data fusion process.
- The fusion process provides more accurate data than the individual sources.
  - Lower noise.

Medium level

- The characteristics or features are fused to obtain features that could be employed for other tasks.
  - E.g. shape, texture, and position.
- Also known as the feature or characteristic level.

High level

- Takes symbolic representations as sources and combines them to make decisions.
- Creates "knowledge".
- Involves learning.
- Also known as decision fusion.

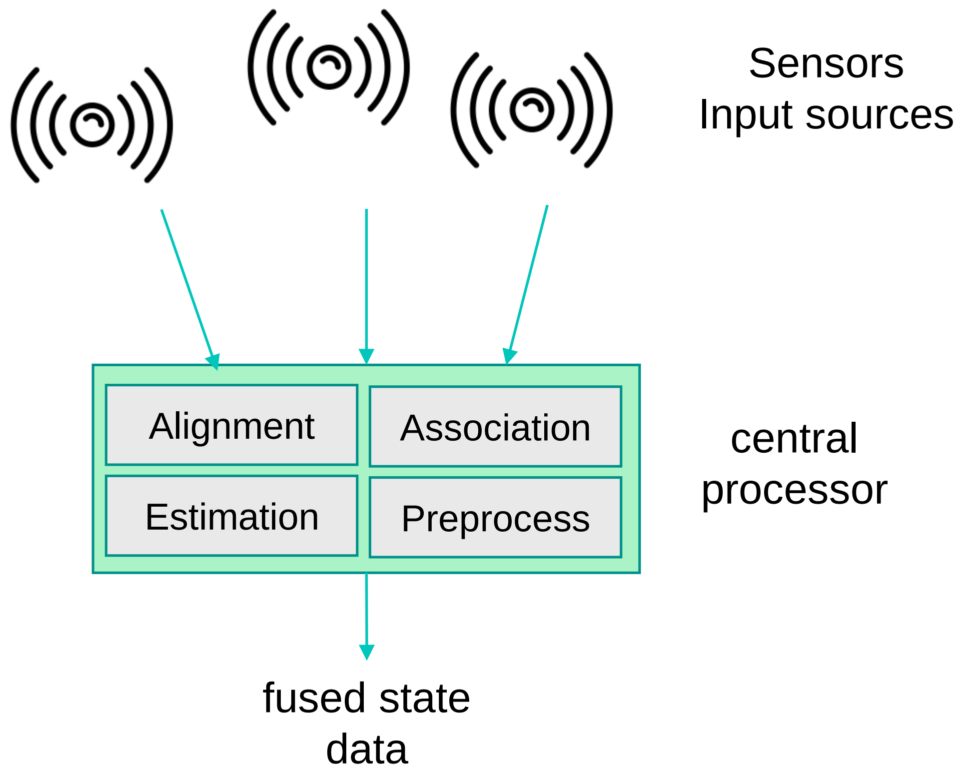
• Mutiple level

- Addresses data provided from different levels of abstraction.
- E.g. a measurement is combined with a feature to make a decision.

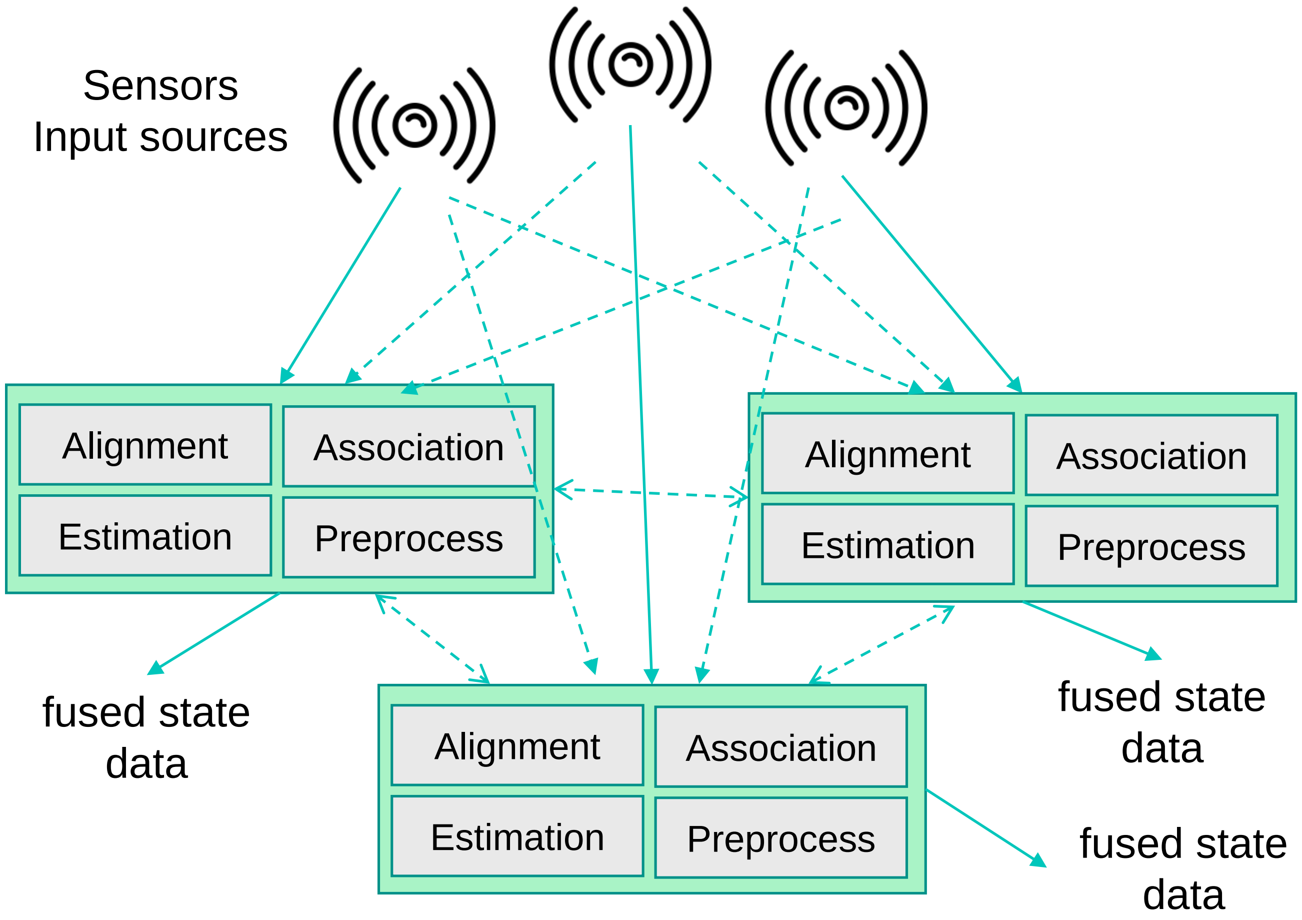
Type of architecture

Centralised architecture

- The fusion node resides in the central processor that receives the information from all of the input sources.
- The sources obtain only the observation as measurements and transmit them to a central processor, where the data fusion process is performed.
- May be theoretical optimal. But, needs to assume that data associations are correct and transfer times are insignificant.
  - Spoiler alert, they are significant.
- Requires large amount of bandwidth.
  - Bottleneck when used for fusing visual sensor networks.



Type of architecture



Decentralised architecture

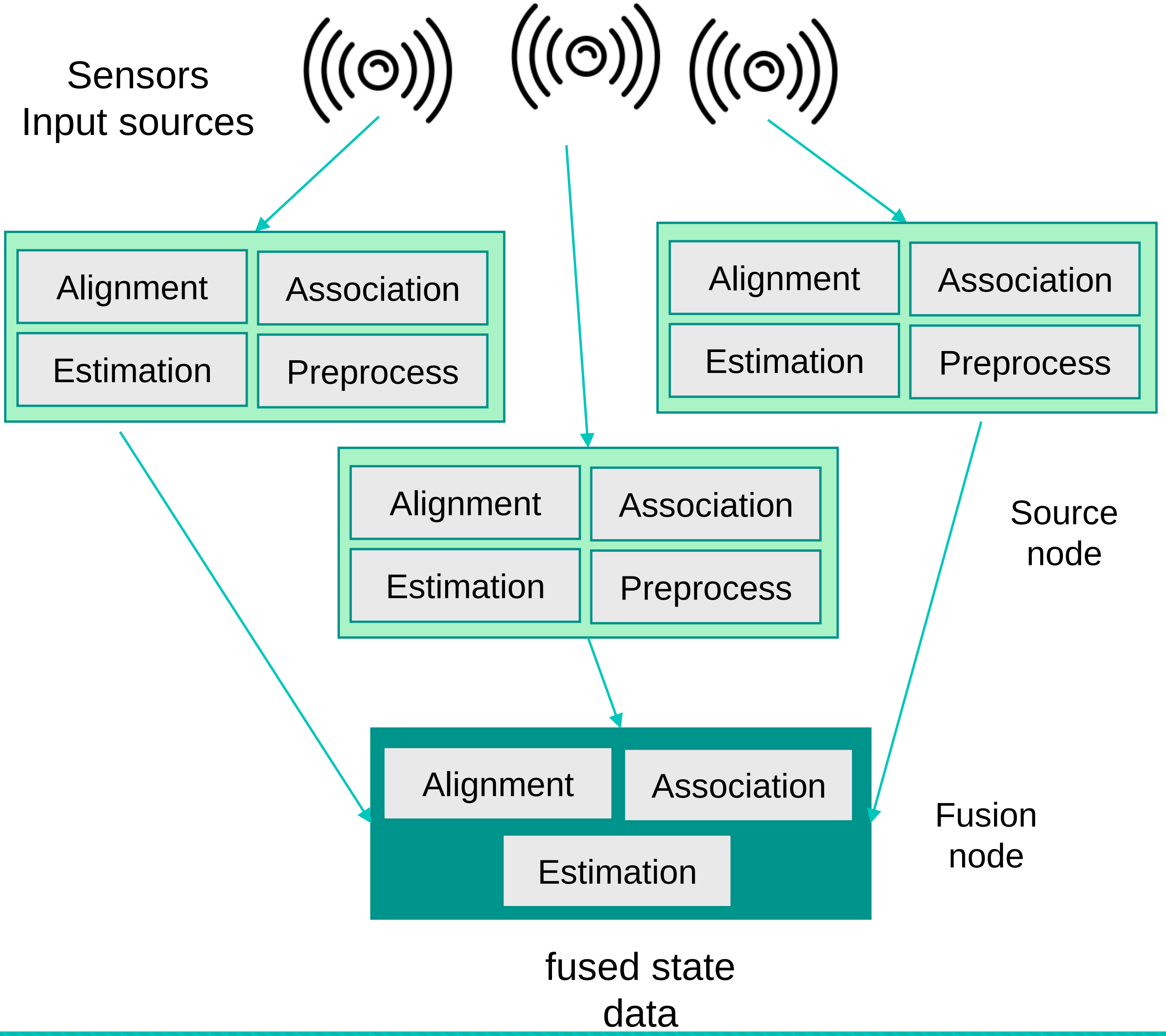
- It is composed of a network of nodes in which each node has its own processing capabilities and there is no single point of data fusion.
- Each node fuses its local information with the information that is received from its peers.
  - Autonomous nodes.
- High communication cost.
  - Linear in the number of nodes.
  - Scalability problems!

Type of architecture

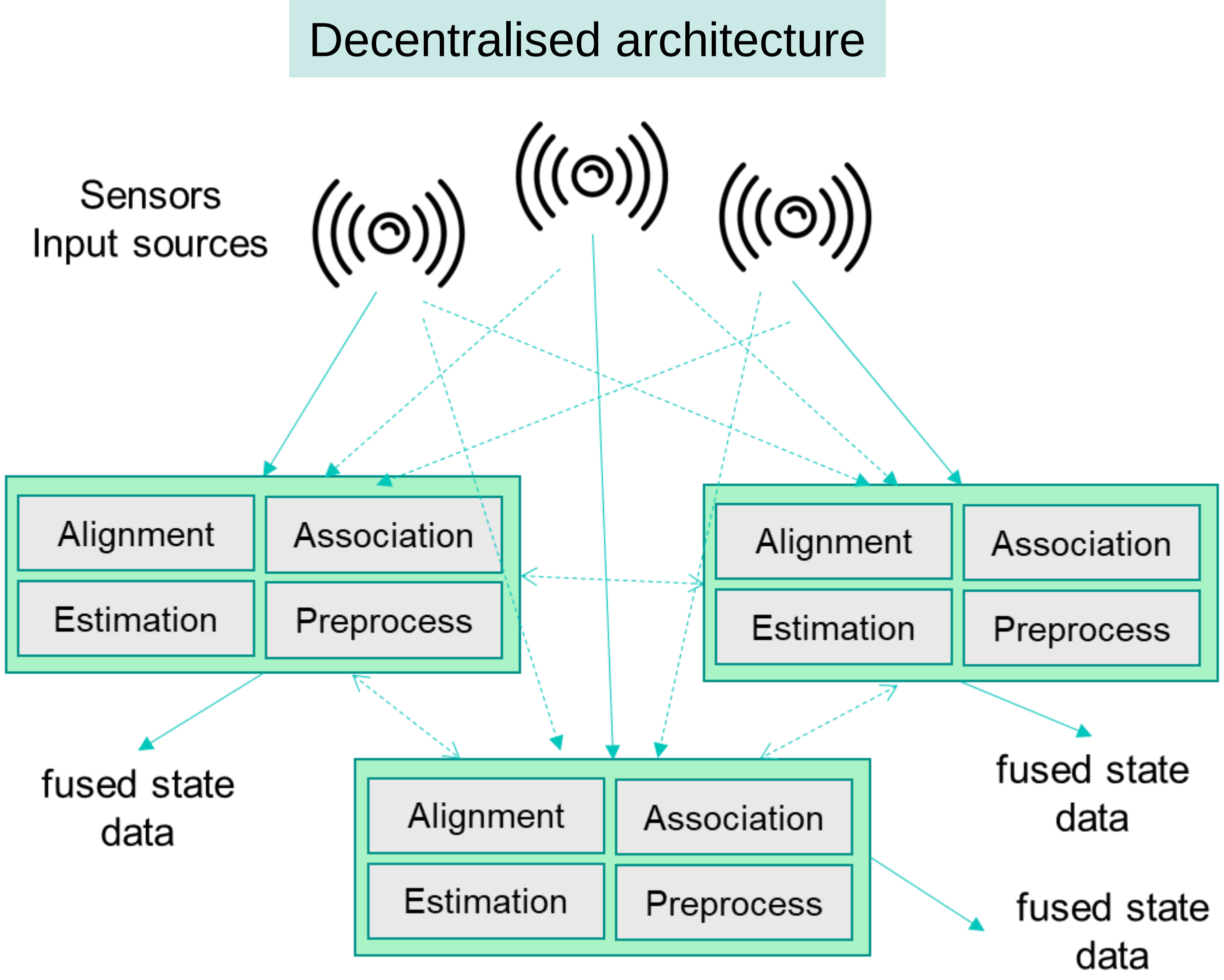
Distributed architecture

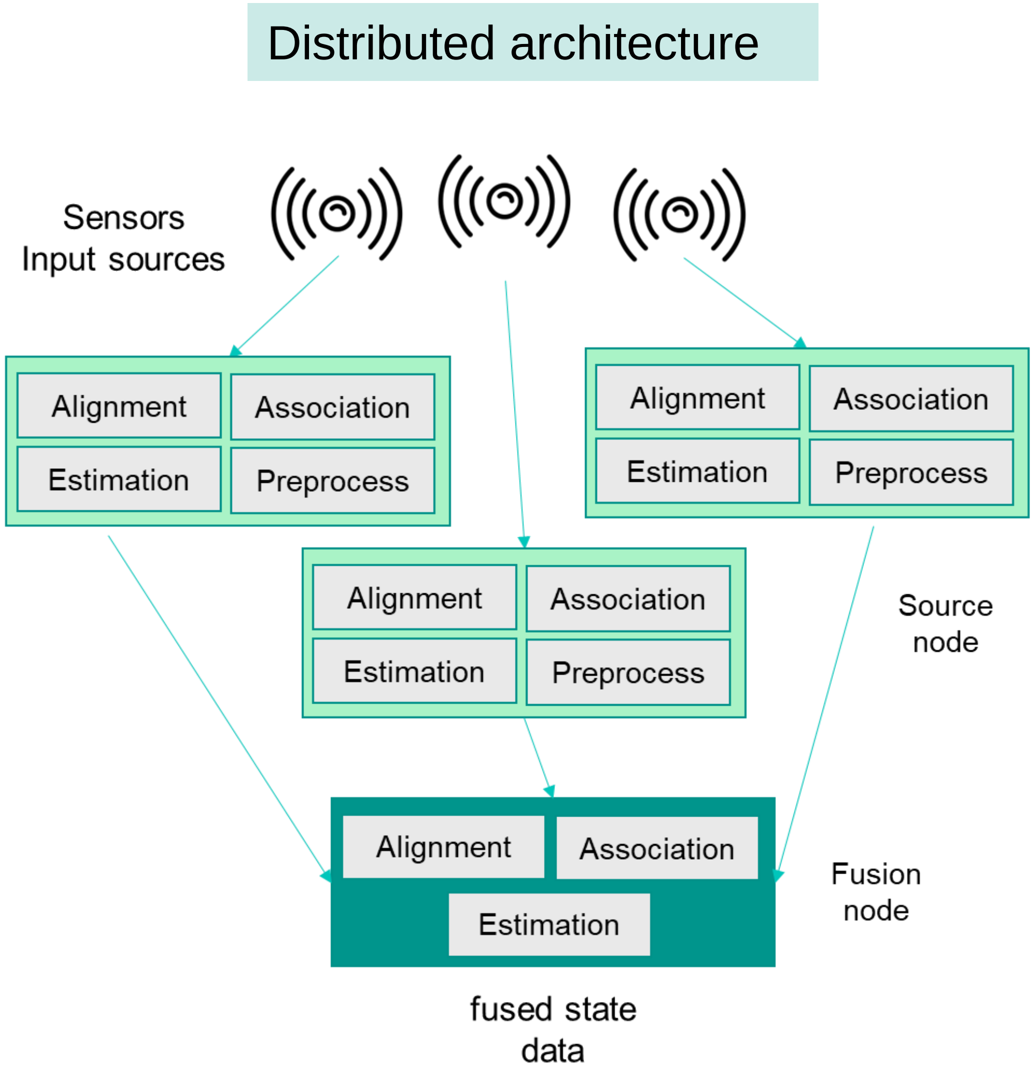
- Measurements from each source node are processed independently before the information is sent to the fusion node.

- Source node: data association and state estimation.
- Fusion node: fuses the estimations from the source nodes.
    - Global view.



Type of architecture





Type of architecture

Decentralised architecture	Distributed architecture
• The complete data fusion process is conducted in each node, and each node provides a globally fused result.	• A pre-processing of the obtained states is performed, and data fusion is performed after.
• Not necessarily communicate a shared notion of state, may communicate raw data.	• Share a common notion of state (e.g. position, velocity, and identity) which is used to perform the fusion process.
• Process is additive. Can separate old information/knowledge.	• The fusion of states is not additive.
• More communication costs.	• Reduce the necessary communication and computational costs. • Some tasks are computed in the distributed nodes before data fusion is performed in the fusion node.

Classification of techniques

Relations between the data sources

Abstraction levels

Type of architecture

Classification of techniques

- **Stage-based:** Use different datasets at different stages of a data mining task.
- **Feature level-based:** Learns a new representation of the original features extracted from different datasets.
- **Semantic-based:** Blends data based on their semantic meanings.

Abstraction levels

Yu Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data

Classification of techniques

- **Stage-based:** Use different datasets at different stages of a data mining task.

Abstraction levels

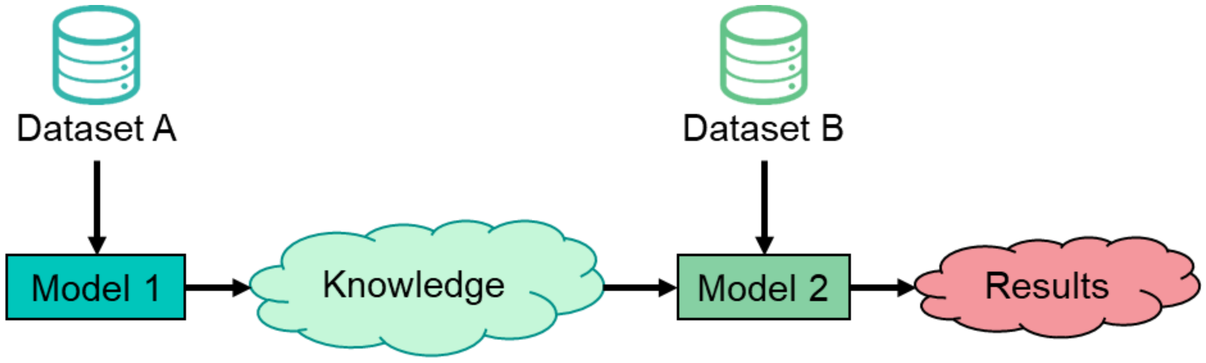
Yu Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data

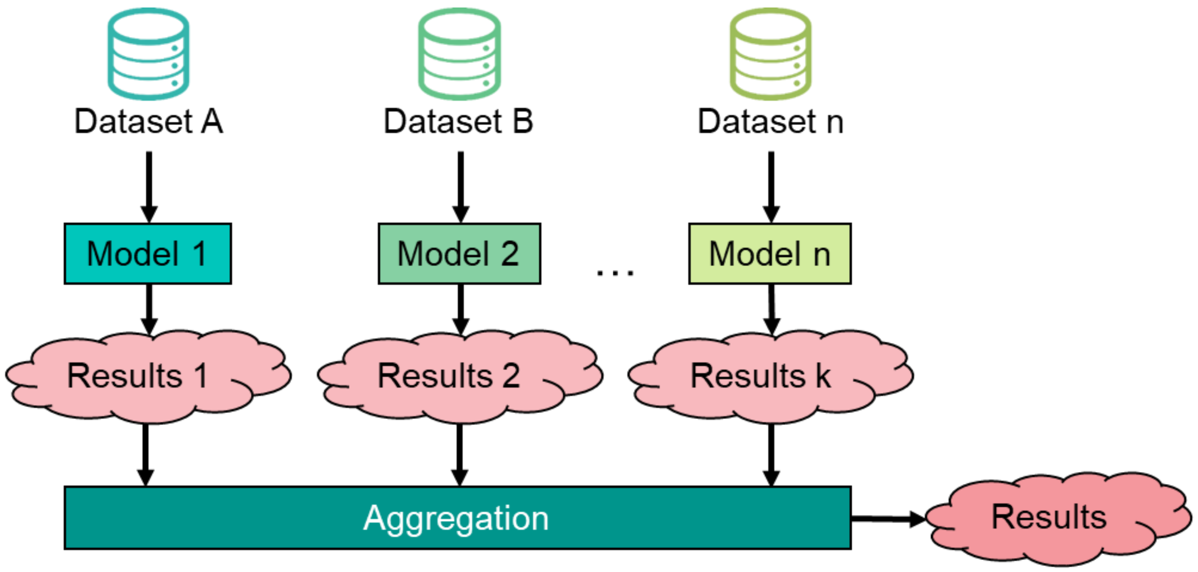
Stage-based data fusion methods

- Different datasets at **different stages** of a data mining task.
- Datasets are **loosely coupled**, without **any consistency requirement.**
- Can be a meta-approach used together with other data fusion methods.

Stage-based data fusion methods

- Different datasets at **different stages** of a data mining task.
- Datasets are **loosely coupled**, without **any consistency requirement.**
- Can be a meta-approach used together with other data fusion methods.





Stage-based data fusion methods

- Different datasets at **different stages** of a data mining task.

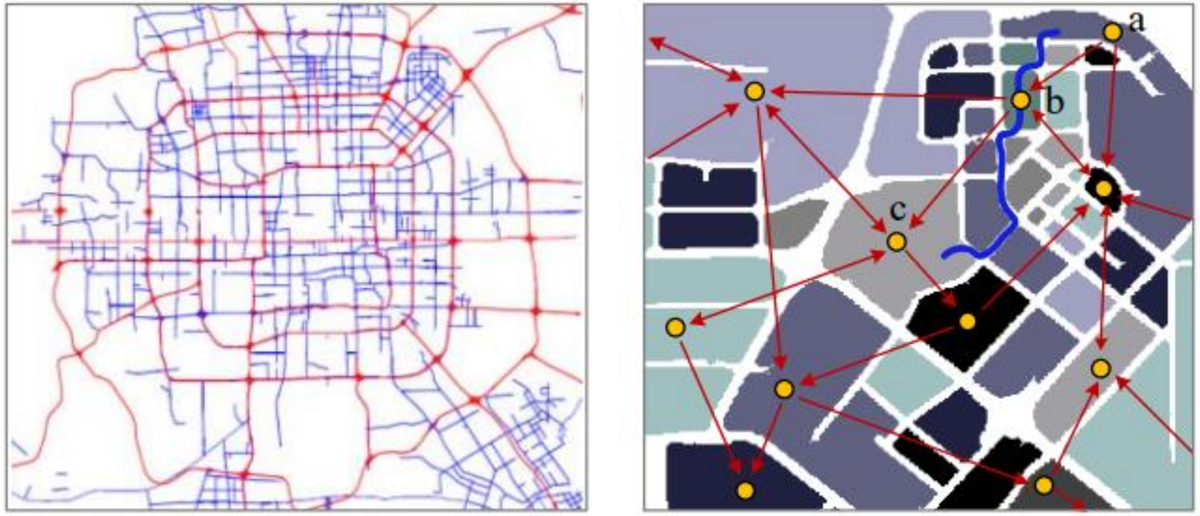
- Datasets are **loosely coupled**, without **any consistency requirement.**

- Can be a meta-approach used together with other data fusion methods.

Example:

- Partition of a city into regions by major roads using a map segmentation method.
- Then, GPS trajectories of taxicabs are then mapped onto the regions to formulate a region graph.
  - Nodes are a region and an edge denotes the aggregation of commutes (by taxis in this case) between two regions.

- Region graph blends knowledge from the road network and taxi trajectories.
  - Identify improper design of a road network.
  - Detect and diagnose traffic anomalies
  - Find urban functional regions.



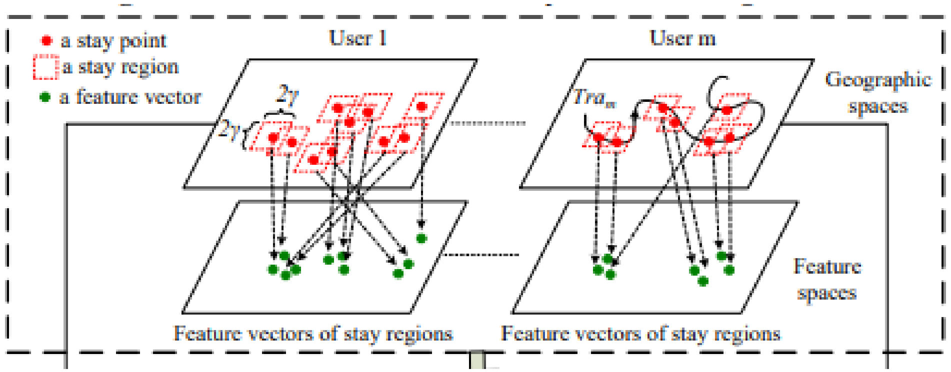
Y. Zheng. Trajectory Data Mining: An Overview. ACM Transactions on Intelligent Systems and Technology, vol. 6, issue 3,pp.1-29, 2015.

Stage-based data fusion methods

Example Social Tie Inference:

- Detect the stay points from an individual's location history.

- Each stay point is then converted into a feature vector based on its surrounding POIs.
    - The distance between these feature vectors denotes the similarity between the places people have visited.



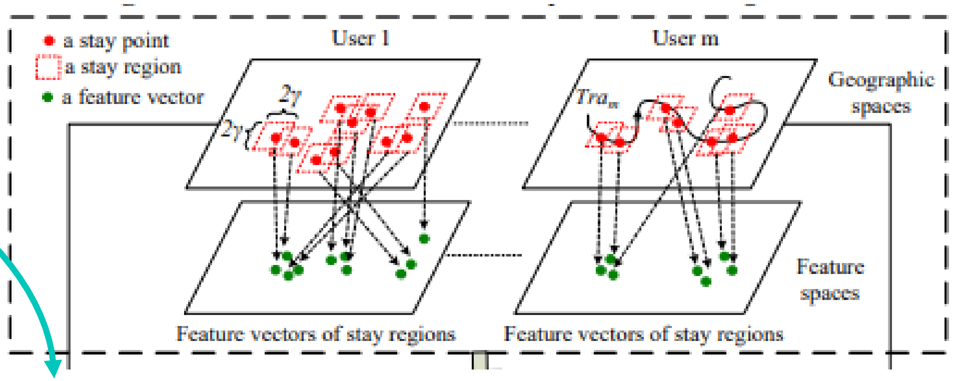
X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring Social Ties between Users with Human Location History," *J. of Ambient Intelligence and Hu-manized Computing*, vol. 5, no. 1, pp. 3-19, 2014.

Stage-based data fusion methods

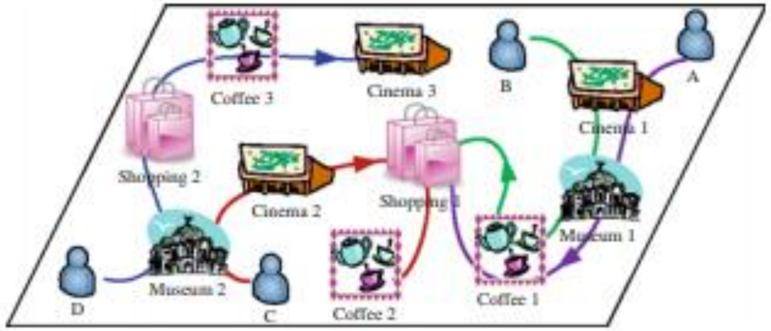
Example Social Tie Inference:

- Detect the stay points from an individual's location history.

- Each stay point is then converted into a feature vector based on its surrounding POIs.
    - The distance between these feature vectors denotes the similarity between the places people have visited.



in the form of spatial trajectories

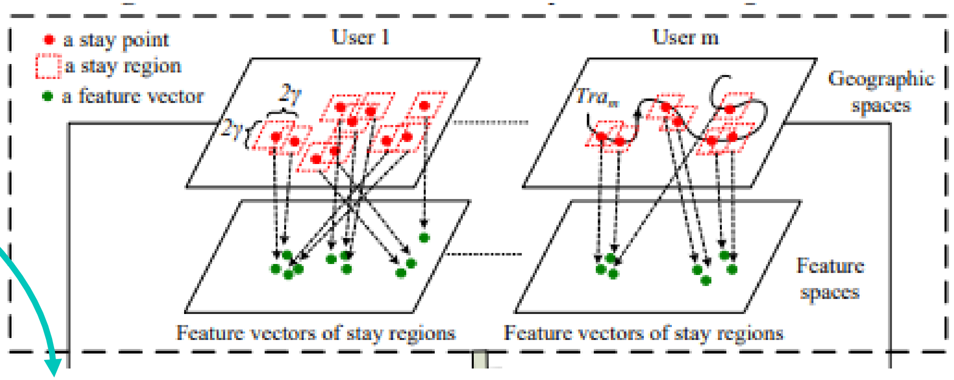


X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring Social Ties between Users with Human Location History," *J. of Ambient Intelligence and Hu-manized Computing*, vol. 5, no. 1, pp. 3-19, 2014.

Stage-based data fusion methods

Example Social Tie Inference:

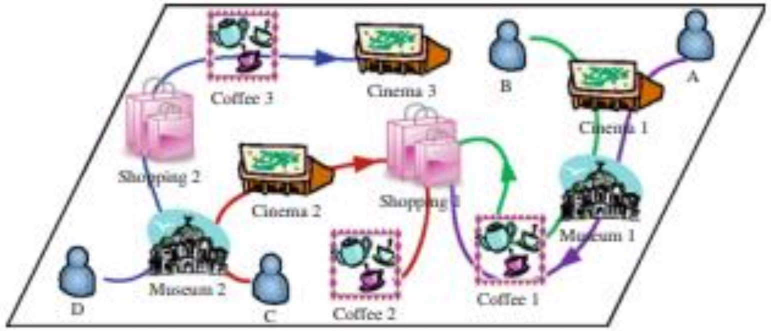
- Detect the stay points from an individual's location history.
- Each stay point is then converted into a feature vector based on its surrounding POIs.
  - The distance between these feature vectors denotes the similarity between the places people have visited.



in the form of spatial trajectories

avoid disjoint location histories

E.g. 5 restaurants, 1 shopping mall and 1 gas station around a stay point.



X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring Social Ties between Users with Human Location History," *J. of Ambient Intelligence and Hu-manized Computing*, vol. 5, no. 1, pp. 3-19, 2014.

Stage-based data fusion methods

Example Social Tie Inference:

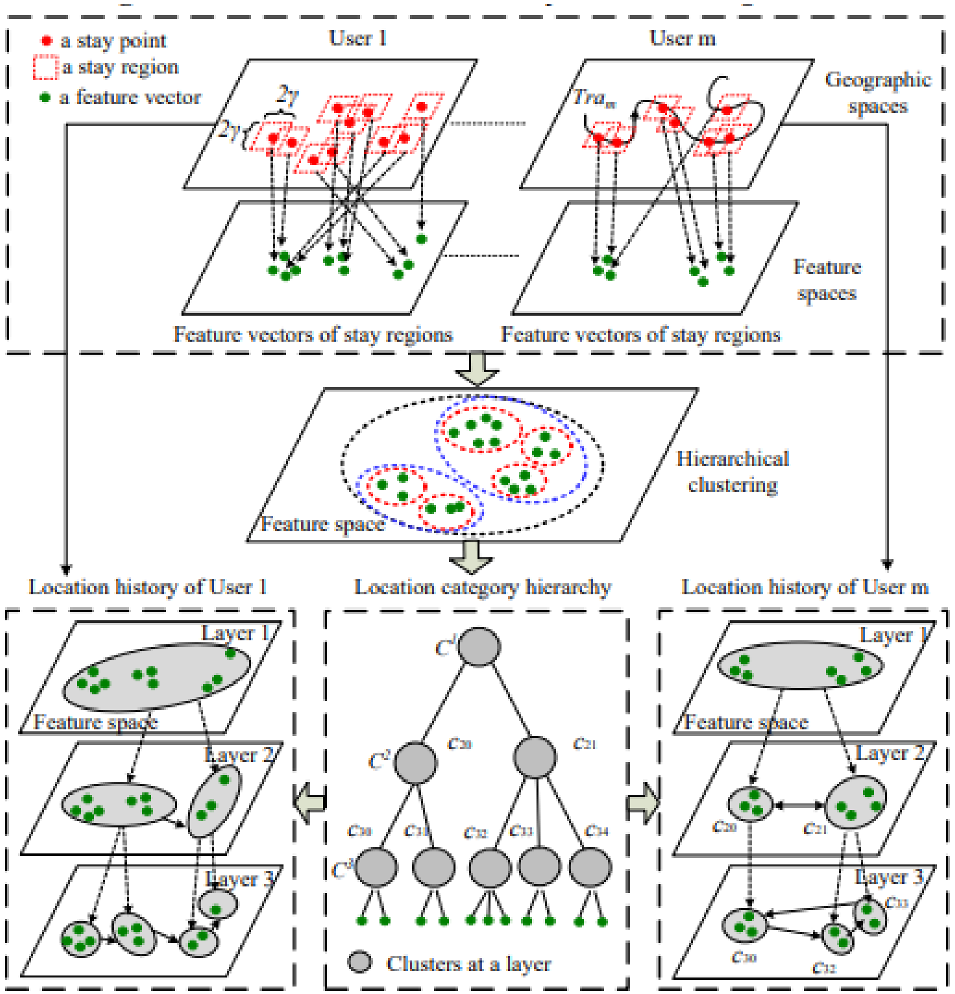
- Detect the stay points from an individual's location history.

- Each stay point is then converted into a feature vector based on its surrounding POIs.
  - The distance between these feature vectors denotes the similarity between the places people have visited.

- Stay points are hierarchically clustered into groups according to their feature vectors of POIs.

- Users' location history can be represented with a partial tree.
  - Also, hierarchical graphs are built representing the original trajectories.

- As users' hierarchical graphs are built based on the same tree structure, their location histories become comparable.
  - Similarity between users can be computed.



X. Xiao, Y. Zheng, Q. Luo, and X. Xie, "Inferring Social Ties between Users with Human Location History," *J. of Ambient Intelligence and Hu-manized Computing*, vol. 5, no. 1, pp. 3-19, 2014.

Classification of techniques

Abstraction levels

- **Feature level-based:** Learns a new representation of the original features extracted from different datasets.

Yu Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data

Feature level-based data fusion methods

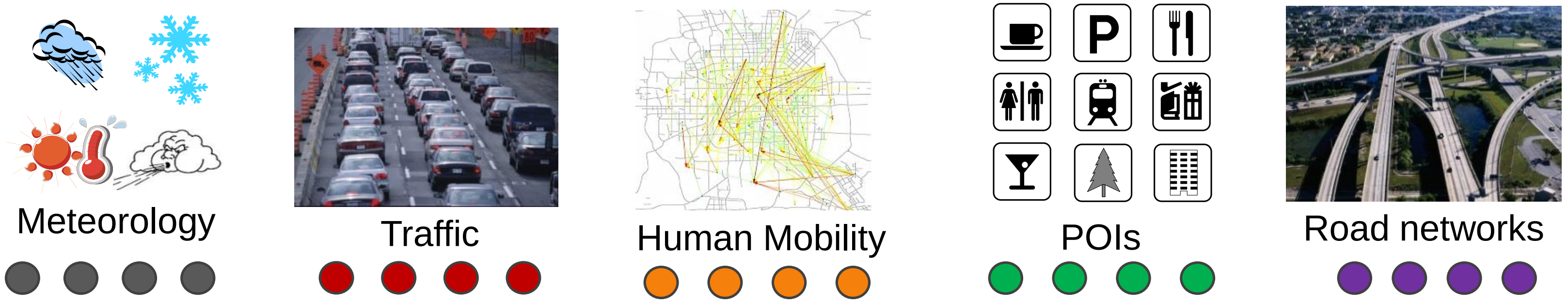
Direct concatenation

- Treat features extracted from different datasets equally, concatenating them sequentially into a feature vector.

Feature level-based data fusion methods

Direct concatenation

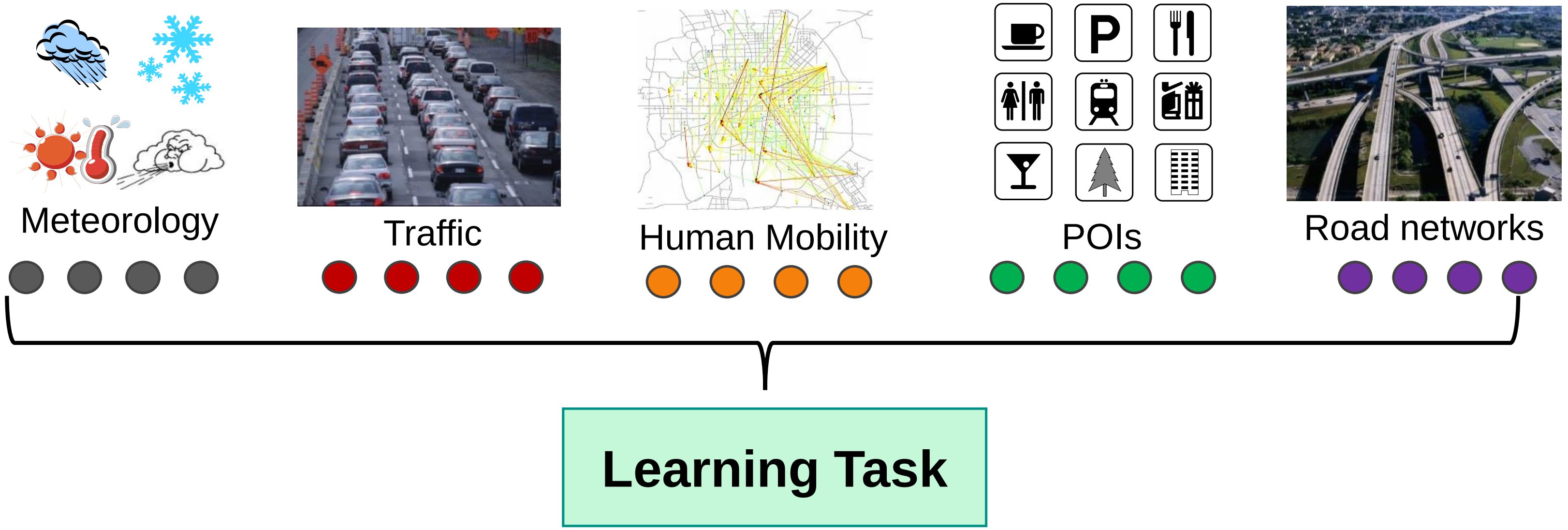
- Treat features extracted from different datasets equally, concatenating them sequentially into a feature vector.



Feature level-based data fusion methods

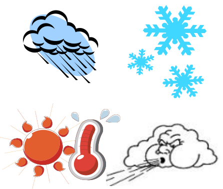
Direct concatenation

- Treat features extracted from different datasets equally, concatenating them sequentially into a feature vector.
- The feature vector is then used in clustering and classification tasks.



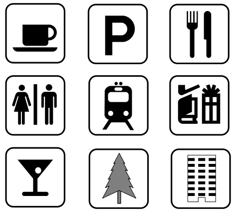
Direct concatenation

- Treat features extracted from different datasets equally, concatenating them sequentially into a feature vector.
- The feature vector is then used in clustering and classification tasks.
- **The representation, distribution and scale of different datasets may be very different!**


Meteorology


Traffic


Human Mobility



POIs

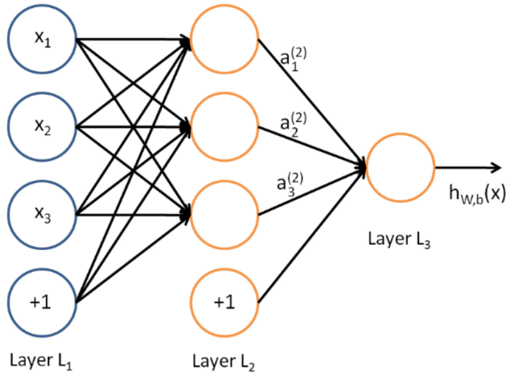

Road networks

Feature level-based data fusion methods

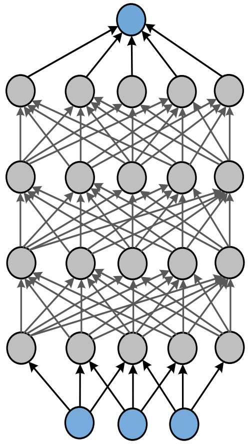
Direct concatenation

- Treat features extracted from different datasets equally, concatenating them sequentially into a feature vector.
- The feature vector is then used in clustering and classification tasks.
- **The representation, distribution and scale of different datasets may be very different!**
  - **Over-fitting** in the case of a small size training sample, and the specific statistical property of each view is ignored.
  - Difficult to discover **highly non-linear relationships** that exist between low-level features across different modalities.
  - **Redundancies and dependencies** between features extracted from different datasets which may be correlated.

Neural Network

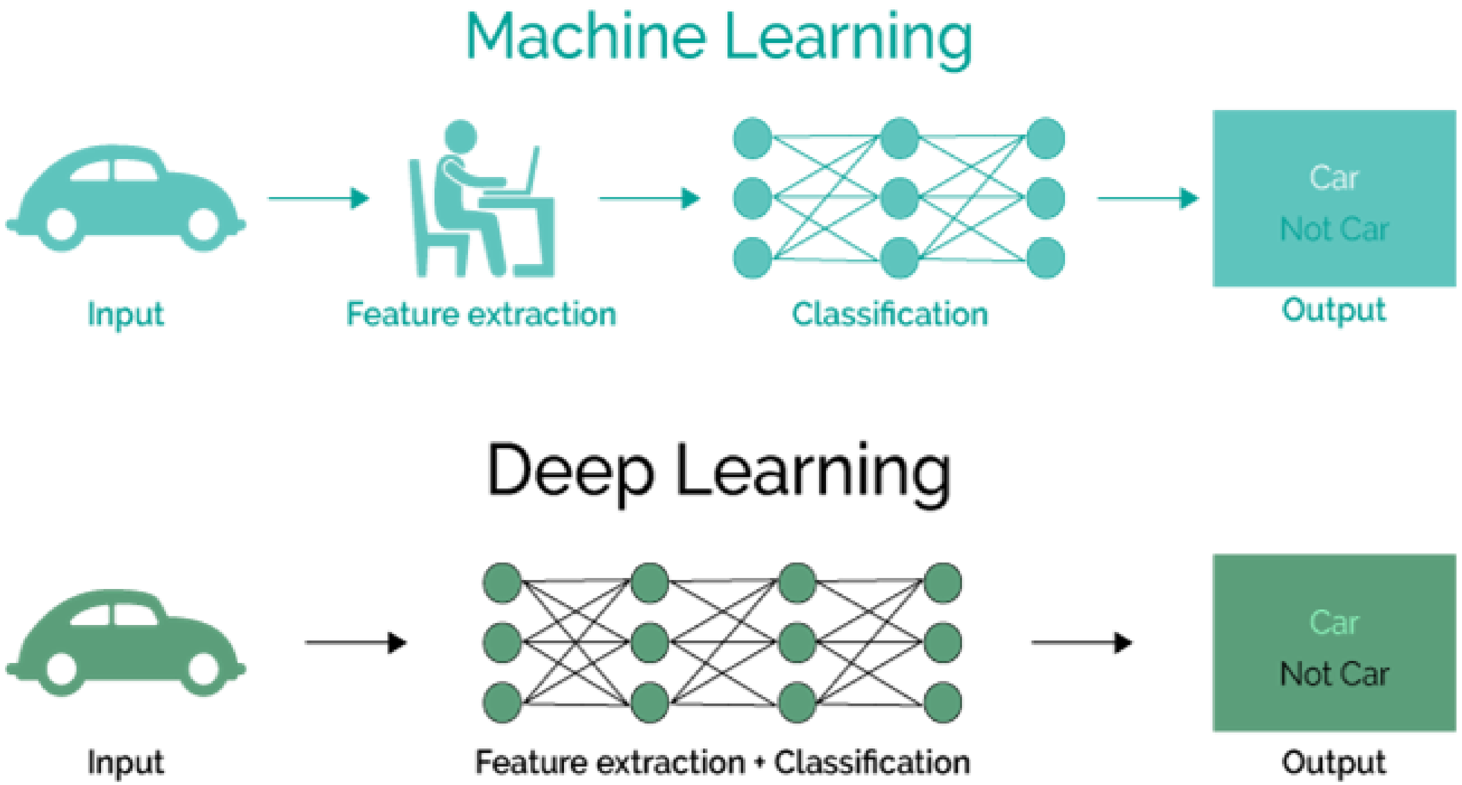


- Using supervised, unsupervised and semi-supervised approaches, learns multiple levels of representation and abstraction
- Produce unified feature representation from disparate dataset
  - Majority of NNs are applied to handle data with a single modality.
  - Lately, they can learn feature presentations from data with different modalities.
  - Useful for classification and information retrieval tasks.
- Learn a **middle-level** feature representation and abstraction.
- Work well for image, text and sound when having sufficient data.
- The learned representations can be fed into other classifiers and predictors.



Feature level-based data fusion methods

Neural Network



Neural Network

What defines a good multi-modality learning?

1. The learned shared feature representation preserves the similarity of “concepts”.

2. The joint feature representation is easy to obtain in the absence of some modalities, and thus fills in missing modalities.

3. The new feature representation facilitates retrieval of one modality when querying from the other.

Neural Network

What defines a good multi-modality learning?

1. The learned shared feature representation preserves the similarity of “concepts”.
2. The joint feature representation is easy to obtain in the absence of some modalities, and thus fills in missing modalities.
3. The new feature representation facilitates retrieval of one modality when querying from the other.

E.g. given text and images, a probabilistic model can correlate the occurrence of the words “magic wand” with the visual properties of an image of a wand, and represent them jointly.

Neural Network

What defines a good multi-modality learning?

1. The learned shared feature representation preserves the similarity of "concepts".

E.g. given text and images, a probabilistic model can correlate the occurrence of the words "magic wand" with the visual properties of an image of a wand, and represent them jointly.

2. The joint feature representation is easy to obtain in the absence of some modalities, and thus fills in missing modalities.

Can we use a model that was trained on images and text, when we only have images at test time? Can this model do better than one that was trained on images alone?

3. The new feature representation facilitates retrieval of one modality when querying from the other.

Neural Network

What defines a good multi-modality learning?

1. The learned shared feature representation preserves the similarity of “concepts”.

E.g. given text and images, a probabilistic model can correlate the occurrence of the words “magic wand” with the visual properties of an image of a wand, and represent them jointly.

2. The joint feature representation is easy to obtain in the absence of some modalities, and thus fills in missing modalities.

Can we use a model that was trained on images and text, when we only have images at test time? Can this model do better than one that was trained on images alone?

3. The new feature representation facilitates retrieval of one modality when querying from the other.

Each data point in the database can be mapped to the created latent space. Queries can also be mapped to this space and an appropriate distance metric can be used to retrieve results that are close to the query.

Neural Network

Step 50	Step 100	Step 150	Step 200	Step 250
travel	beach	sea	water	italy
trip	ocean	beach	canada	water
vacation	waves	island	bc	sea
africa	sea	vacation	britishcolumbia	boat
earthasia	sand	travel	reflection	italia
asia	nikon	ocean	alberta	mare
men	surf	caribbean	lake	venizia
2007	rocks	tropical	quebec	acqua
india	coast	resort	ontario	ocean
tourism	shore	trip	ice	venice

Input tags	Step 50	Step 100	Step 150	Step 200	Step 250
purple, flowers					
car, automobile					

N. Srivastava, R. Salakhutdinov,“Multimodal Learning with Deep Boltzmann Machines,” *Proc. Neural Information and Processing Systems*, 2012.

Neural Network

- A deep auto-encoder architecture to capture the “middle-level” feature representation between two modalities (e.g., audio and video).

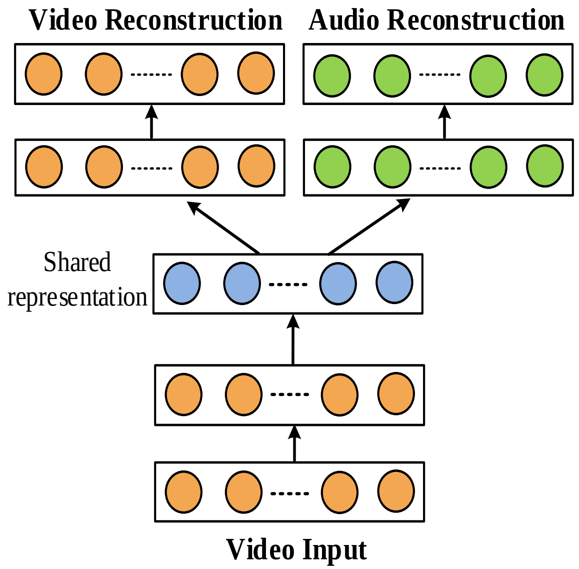
- Two alternatives:
  - A single modality (e.g. video or audio) is used as the input to reconstruct a better feature representation for video and audio respectively.
  - The shared representation learning and multi-modal fusion.

J.Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multi-modal deep learning, ”*Proc. the 28th International Conference on Machine Learning*, pp. 689-696, 2011

Neural Network

- A deep auto-encoder architecture to capture the “middle-level” feature representation between two modalities (e.g., audio and video).

- Two alternatives:
  - **A single modality (e.g. video or audio) is used as the input to reconstruct a better feature representation for video and audio respectively.**
  - The shared representation learning and multi-modal fusion.

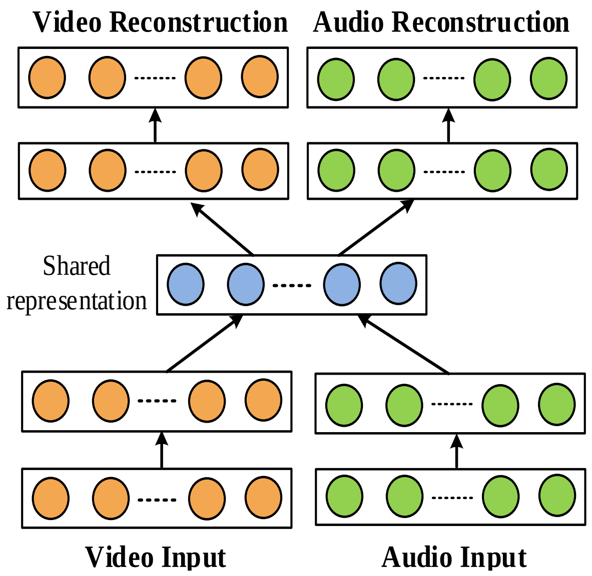


J.Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multi-modal deep learning, ”*Proc. the 28th International Conference on Machine Learning*, pp. 689-696, 2011

Neural Network

- A deep auto-encoder architecture to capture the “middle-level” feature representation between two modalities (e.g., audio and video).

- Two alternatives:
  - A single modality (e.g. video or audio) is used as the input to reconstruct a better feature representation for video and audio respectively.
  - **The shared representation learning and multi-modal fusion**.



J.Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multi-modal deep learning, ”*Proc. the 28th International Conference on Machine Learning*, pp. 689-696, 2011

Neural Network

- A deep auto-encoder architecture to capture the “middle-level” feature representation between two modalities (e.g., audio and video).

- Two alternatives:
  - A single modality (e.g. video or audio) is used as the input to reconstruct a better feature representation for video and audio respectively.
  - The shared representation learning and multi-modal fusion.

- According to the authors:
  - NN achieve better single modality representation with the help of other modalities.
  - The shared representations can capture the correlations across multiple modalities.

J.Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multi-modal deep learning, ”*Proc. the 28th International Conference on Machine Learning*, pp. 689-696, 2011

Feature level-based data fusion methods

Neural Network

- Performance heavily depend on how well parameters can be tuned.
- Finding optimal parameters is a labour intensive and time-consuming process given a large number of parameters.
  - Heavily relies on human experience!
- Hard to explain what the middle-level feature representation stands for.
- It is still unknown how NN makes raw features a better representation.

Classification of techniques

Abstraction levels

- **Semantic-based:** Blends data based on their semantic meanings.

Yu Zheng. Methodologies for Cross-Domain Data Fusion: An Overview. IEEE Transactions on Big Data

Semantic-based data fusion methods

Feature-based methods
do not care about the meaning of each feature, regarding a feature solely as a **real-valued number** or a **categorical value**.

Semantic-based methods
understand the insight of each dataset and relations between features across different datasets.

Semantic-based data fusion methods

Feature-based methods
do not care about the meaning of each feature, regarding a feature solely as a **real-valued number** or a **categorical value**.

Semantic-based methods
understand the insight of each dataset and relations between features across different datasets.

Carry a **semantic** meaning derived from the ways that **people think** of a problem with the help of multiple datasets.

Know what each dataset stands for, why different datasets can be fused, how they complement each other.

Semantic-based data fusion methods

Feature-based methods
do not care about the meaning of each feature, regarding a feature solely as a **real-valued number** or a **categorical value**.

Semantic-based methods
understand the insight of each dataset and relations between features across different datasets.

Carry a **semantic** meaning derived from the ways that **people think** of a problem with the help of multiple datasets.

interpretable

meaningful

Know what each dataset stands for, why different datasets can be fused, how they complement each other.

Semantic-based data fusion methods

Feature-based methods
do not care about the meaning of each feature, regarding a feature solely as a **real-valued number** or a **categorical value**.

Semantic-based methods
understand the insight of each dataset and relations between features across different datasets.

Carry a **semantic** meaning derived from the ways that **people think** of a problem with the help of multiple datasets.

interpretable

meaningful

Know what each dataset stands for, why different datasets can be fused, how they complement each other.

Multi-view	Similarity	Probabilistic dependency	Transfer learning

Multi-view based

- Different datasets or different feature subsets about an object can be regarded as different views on the object.

A **person** can be identified by face, fingerprint, or signature
An **image** can be represented by colour or texture features

- As these datasets describe the same object, there is a **latent consensus** among them.
  - Datasets are **complementary** to each other, containing knowledge that other views do not have.
  - **Combining** multiple views can describe an object **comprehensively** and **accurately**.

Co-training

Multi kernel learning

Subspace learning

Multi-view based

- Different datasets or different feature subsets about an object can be regarded as different views on the object.

A **person** can be identified by face, fingerprint, or signature
An **image** can be represented by colour or texture features

- As these datasets describe the same object, there is a **latent consensus** among them.
  - Datasets are **complementary** to each other, containing knowledge that other views do not have.
  - **Combining** multiple views can describe an object **comprehensively** and **accurately**.

train alternately to maximize the mutual agreement on two distinct views of the data

Co-training

Multi kernel learning

Subspace learning

Multi-view based

- Different datasets or different feature subsets about an object can be regarded as different views on the object.

A **person** can be identified by face, fingerprint, or signature
An **image** can be represented by colour or texture features

- As these datasets describe the same object, there is a **latent consensus** among them.
  - Datasets are **complementary** to each other, containing knowledge that other views do not have.
  - **Combining** multiple views can describe an object **comprehensively** and **accurately**.

exploit kernels that naturally correspond to different views and combine kernels either linearly or non-linearly to improve learning

Co-training | Multi kernel learning | Subspace learning

Multi-view based

- Different datasets or different feature subsets about an object can be regarded as different views on the object.

A **person** can be identified by face, fingerprint, or signature
An **image** can be represented by colour or texture features

- As these datasets describe the same object, there is a **latent consensus** among them.
  - Datasets are **complementary** to each other, containing knowledge that other views do not have.
  - **Combining** multiple views can describe an object **comprehensively** and **accurately**.

obtain a latent subspace shared by multiple views, assuming that the input views are generated from this latent subspace

Co-training	Multi kernel learning	Subspace learning

Multi-view based: Co-training

- One of the earliest schemes for multi-view learning.

- Co-training considers a setting in which each example can be partitioned into two distinct views.

- Three main assumptions:

    - **Sufficiency.** Each view is sufficient for classification on its own.

    - **Compatibility.** The target functions in both views predict the same labels for co-occurring features with high probability.

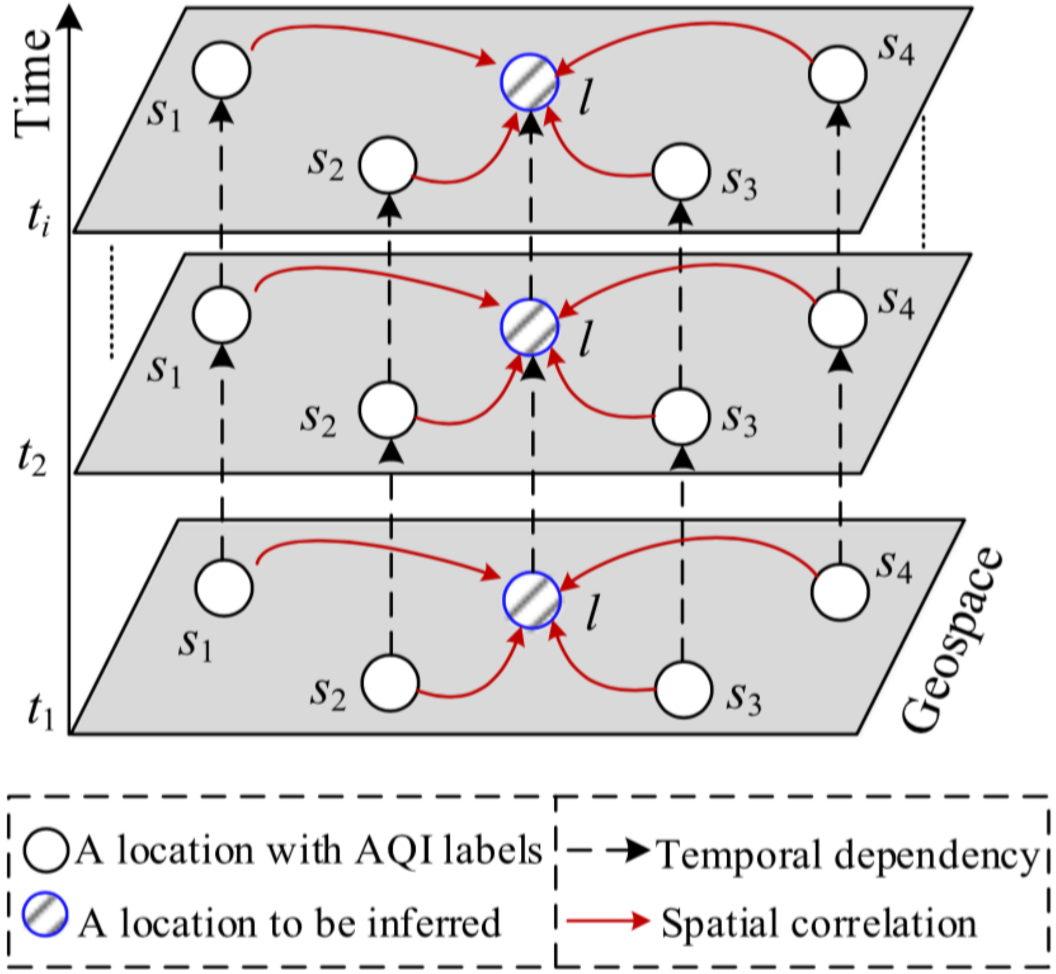
    - **Conditional independence.** The views are conditionally independent given the class label.
        - Too strong in practice.
        - If the independence assumption is violated, on average the added examples will be less informative.
        - It may not be that successful.

Multi-view based: Co-training

- Example. Co-training model to infer the fine-grained air quality.
- Five datasets:
  - Air quality.
  - Meteorological data.
  - Traffic.
  - POIs.
  - Read networks.

- Air quality has a temporal dependency in an individual location and the spatial correlation amongst different locations.
  - The current air quality depends on the last hours.
  - The air quality of a place could be bad if the air quality of its surrounding locations is bad.

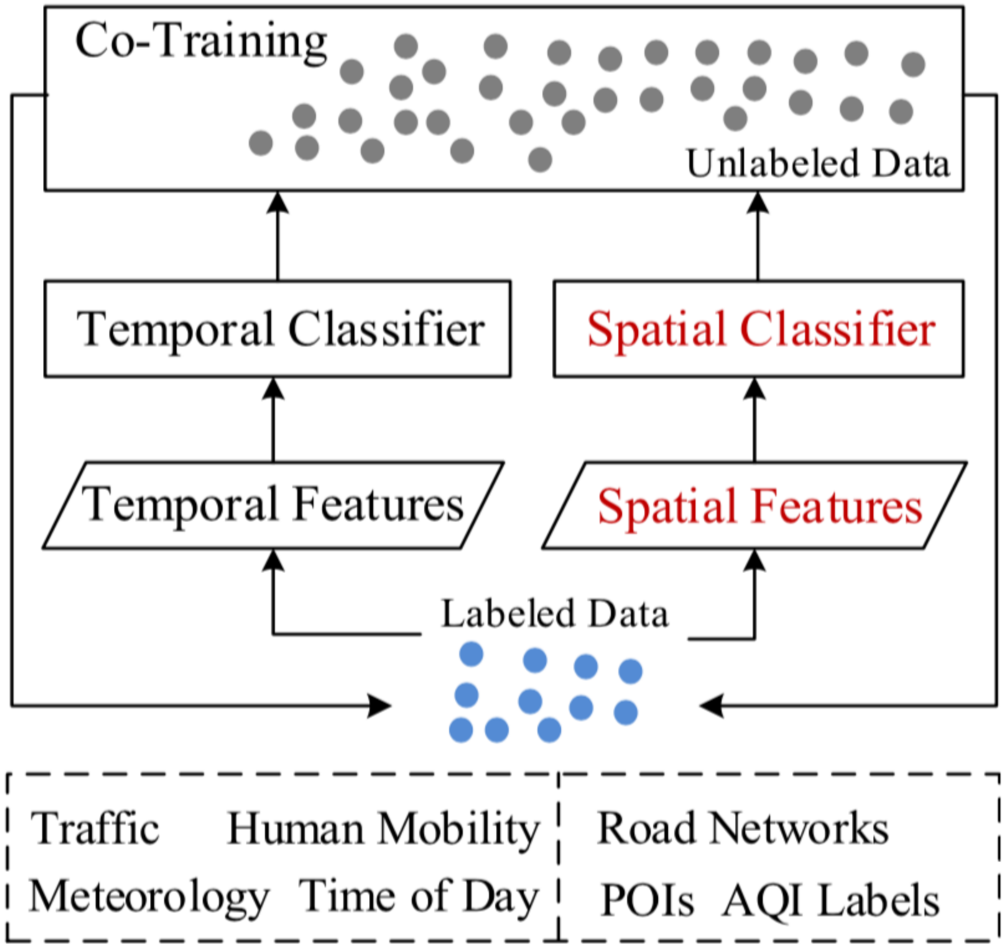
- The problem can be formulated based on two distinct views: temporal dependency + spatial correlation.



N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng andH. Xiong,“Discov-ering Urban Functional Zones Using Latent Activity Trajectories,”*IEEE Transactions on Knowledge and Data Engineering*, vol. 27,no.3, pp. 1041-4347, 2015

Multi-view based: Co-training

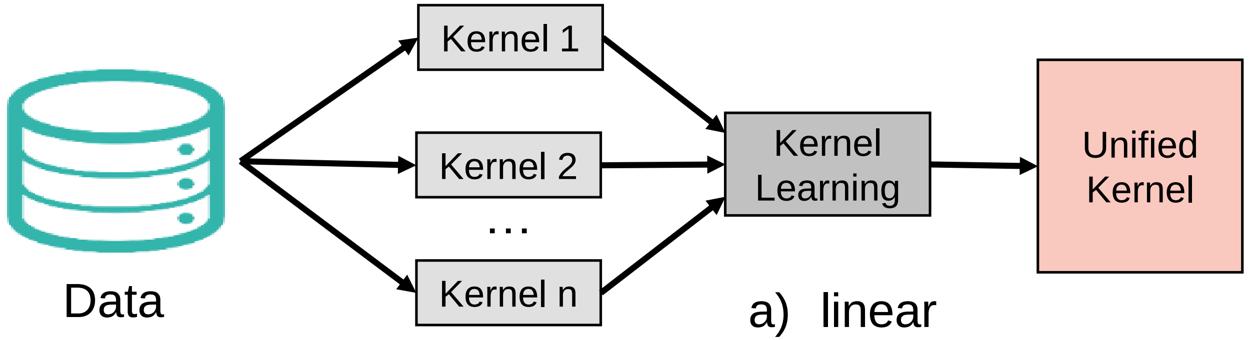
- Two classifiers.
  - <u>Spatial classifier</u> based on NN that takes spatially-related features as input to model the spatial-correlation between air qualities of different lo-cations.
  - <u>Temporal classifier</u> based on conditional random fields that includes temporally-related features to model the temporal dependency of air quality in a location.
- Classifiers are first trained based on limited labelled data using non-overlapped features.
- The instances that are confidently inferred by a classifier in each round are included in the next round of training.
- For classifying an instance, the different classifiers receive different features → two sets of probabilities across different labels.



N. J. Yuan, Y. Zheng, X. Xie, Y. Wang, K. Zheng andH. Xiong,"Discov-ering Urban Functional Zones Using Latent Activity Trajectories,"*IEEE Transactions on Knowledge and Data Engineering*, vol. 27,no.3, pp. 1041-4347, 2015

Multi-view based: Multi kernel learning

- Uses a predefined set of kernels and learns an optimal linear or non-linear combination of kernels as part of the algorithm.
- A kernel is a hypothesis on the data, which could be a similarity notion, or a classifier, or a regressor.
- E.g. Ensemble and boosting methods, such as Random Forest.



a) a sample of data
b) a view on entire data

a) linear
b) non-linear
c) data-dependent

Multi-view based: Multi kernel learning

Different kernels correspond to different notions of similarity.

- A learning method picks the best kernel, or uses a combination of these kernels.
- A sample of data used to train a kernel based on all features.
- Using a specific kernel may be a source of bias, allowing a learner to choose among a set of kernels can result in a better solution.
- A learning method picks the best kernel, or uses a combination of these kernels

Multi-view based: Multi kernel learning

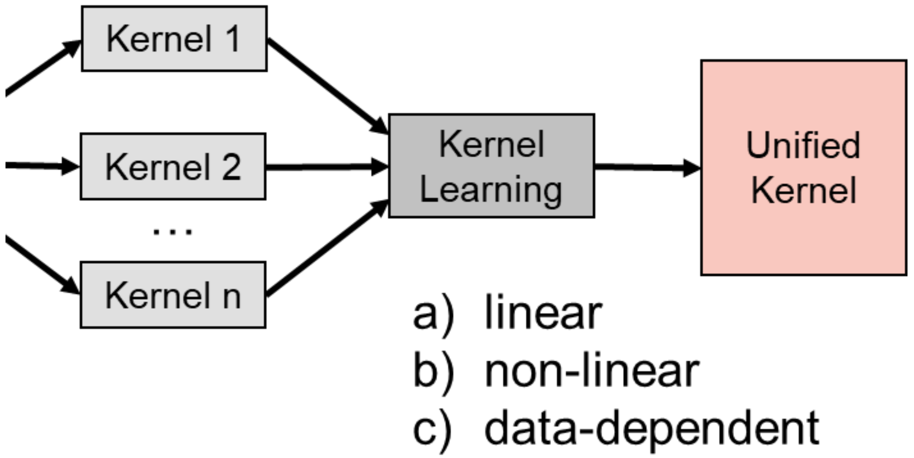
Different kernels correspond to different notions of similarity.

- A learning method picks the best kernel, or uses a combination of these kernels.
- A sample of data used to train a kernel based on all features.
- Using a specific kernel may be a source of bias, allowing a learner to choose among a set of kernels can result in a better solution.
- A learning method picks the best kernel, or uses a combination of these kernels

An alternative… train different kernels using inputs coming from different sources or modalities.

- Since these are different representations, they have different measures of similarity corresponding to different kernels.
- Combining kernels is one possible way to combine multiple information sources.
- The reasoning is similar to combining different classifiers

Multi-view based: Multi kernel learning



Kernel combination

Linear combination

- Unweighted and weighted sum of all results.

Nonlinear combination

- Nonlinear functions of kernels.
  - E.g. multiplication, power and exponentiation.

Data-dependent combination

- Assign specific kernel weights for each data instance.
- Identify local distributions in the data.
- Can learn proper kernel combination rules for each region.

Multi-view based: Multi kernel learning

Training methodology

One-step methods

- Calculate the parameters of the combination function and base learners in a single pass.

- Sequential approach. The combination function parameters are determined first, and then a kernel-based learner is trained using the combined kernel.

- Simultaneous approach. Both set of parameters are learned together.

Two-step methods

- Iterative approach.
- In each iteration.
  - Update the parameters of the combination function while fixing that of the base learner.
  - Update the parameters of base learners while fixing the parameters of the combination function.
- Repeated until convergence.

Multi-view based: Multi kernel learning

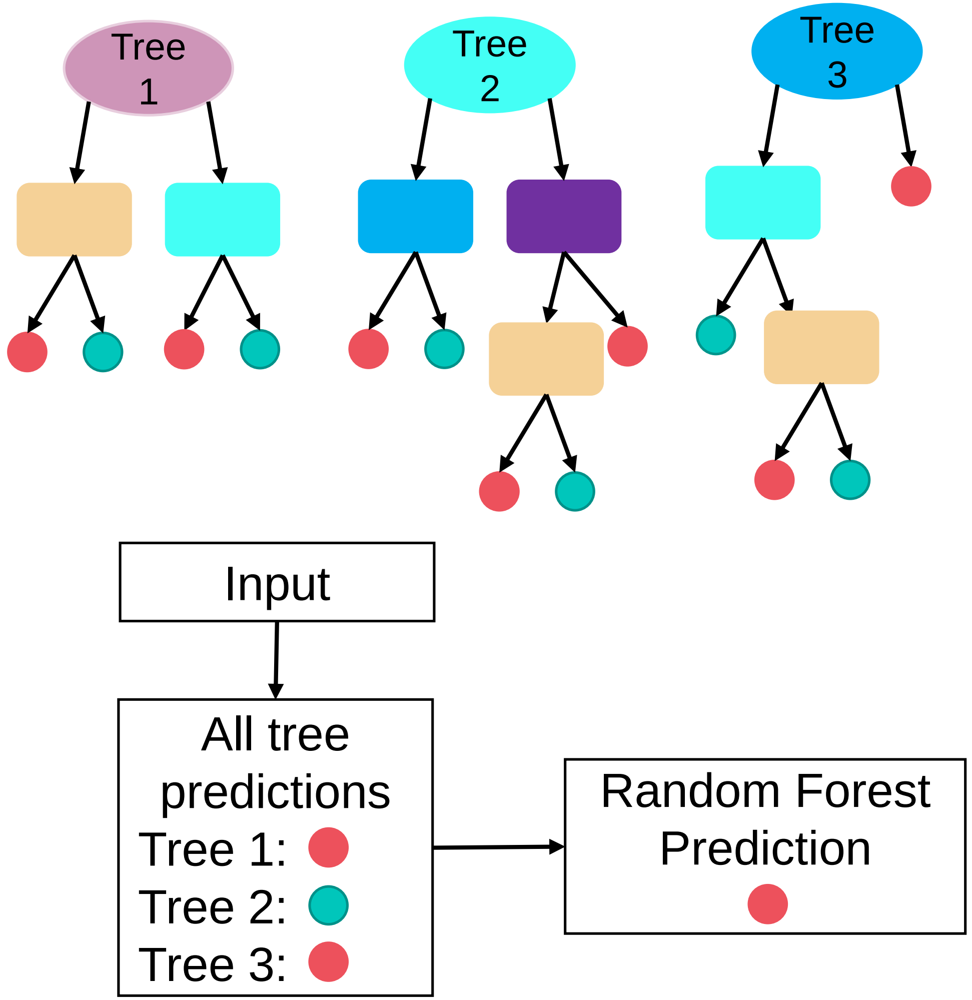
Example. Random Forest.

- Combines the idea of Bootstrap Aggregating and the random selection of features.
- Builds a collection of decision trees with controlled variance.

- It trains multiple Decision Trees by selecting a portion of training data each time based on Bagging and a portion of features.

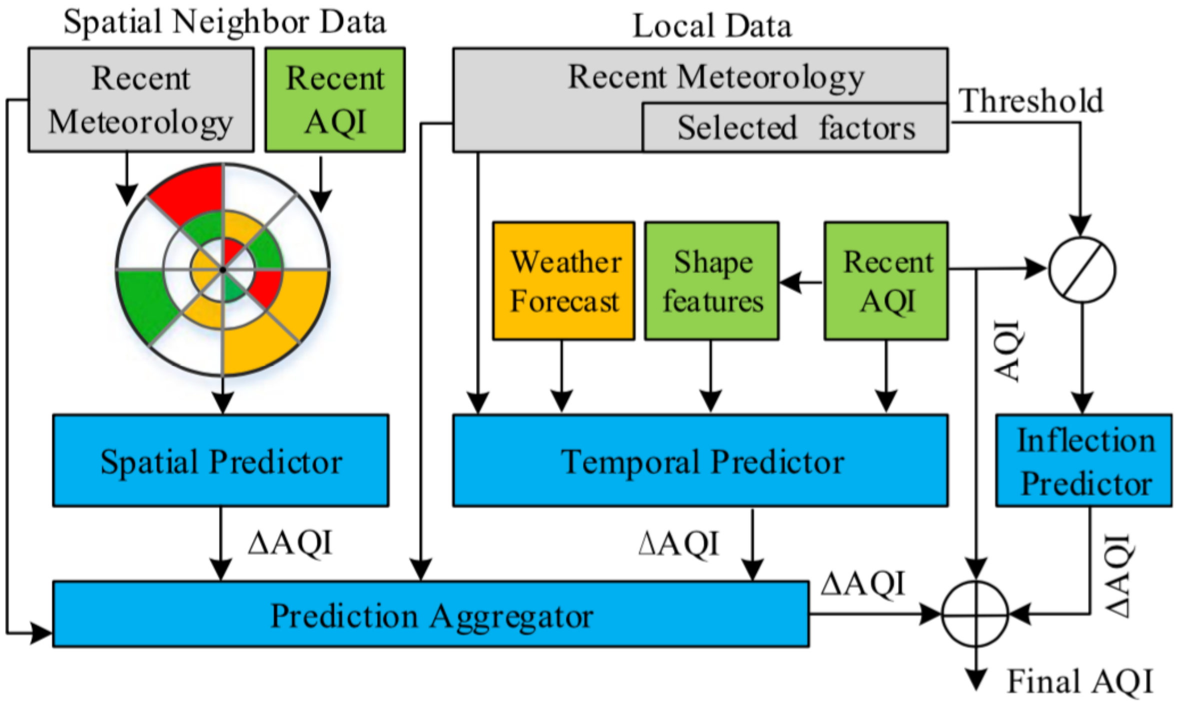
- To classify an instance, different selections of the case's features are sent to corresponding Decision Trees (i.e. kernels) simultaneously.

- Each kernel generates a prediction, which is then linearly aggregated.



Multi-view based: Multi kernel learning

- Example. Model to infer the fine-grained air quality.
- This model works better than the one we already saw.
- **Feature space:**
  - The features used by the spatial and temporal predictors do not have any overlaps, providing different views on a station's air quality.
- **Model:**
  - The spatial and temporal predictors model the local factors and global factors which have significantly different properties.
- **Parameter learning:**
  - Decomposing a big model into 3 coupled small ones scales down the parameter spaces tremendously.
  - More accurate learning and predictions.



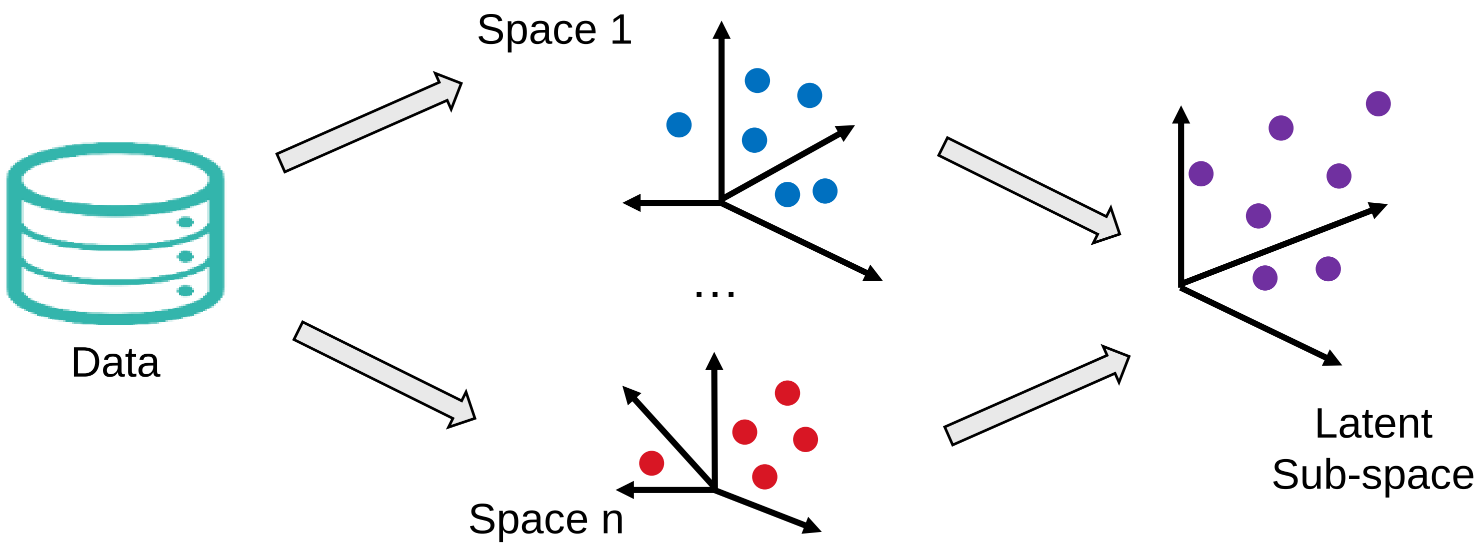
Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, T. Li, Forecasting Fine-Grained Air Quality Based on Big Data. *Proc.ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD'15)*, 2015.

Multi-view based: Subspace learning

- Aim to obtain a latent subspace shared by multiple views by assuming that input views are generated from this latent subspace.

- The subspace can be used in learning tasks (e.g. classification and clustering).

- The subspace usually has a lower dimensionality than the inputs.
    - Reduces the “curse of dimensionality.



Multi-view based: Subspace learning

- Example.
  - Single view: Principal Component Analysis
  - Multi view: several alternatives
    - Linear: Canonical Correlation Analysis (CCA):
      - Maximises the correlation between two views in the subspace and outputs one optimal projection on each view.
      - Subspace is linear → cannot be directly applied to non-linearly embedded datasets.
    - Non-linear: Kernel variant of CCA.
      - Map each non-linear point to a higher space in which linear CCA operates.
    - Gaussian processes.

Multi-view based: Subspace learning

- Example.
  - Single view: Principal Component Analysis

“A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance”

  - Multi view: several alternatives
    - Linear: Canonical Correlation Analysis (CCA):
      - Maximises the correlation between two views in the subspace and outputs one optimal projection on each view.
      - Subspace is linear → cannot be directly applied to non-linearly embedded datasets.
    - Non-linear: Kernel variant of CCA.
      - Map each non-linear point to a higher space in which linear CCA operates.
    - Gaussian processes.

Multi-view based: Subspace learning

- Example.
  - Single view: Principal Component Analysis

“A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance”

  - Multi view: several alternatives
    - Linear: Canonical Correlation Analysis (CCA):
      - Maximises the correlation between two views in the subspace and outputs one optimal projection on each view.
      - Subspace is linear → cannot be directly applied to non-linearly embedded datasets.
    - Non-linear: Kernel variant of CCA.
      - Map each non-linear point to a higher space in which linear CCA operates.
    - Gaussian processes.

“A stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed.”

Similarity based

- Similarity lies between different objects.
- If elements X and Y are similar → The information of one element can be leveraged by the other when there is no data available.
- When X and Y have multiple datasets → Multiple similarities between the two objects can be learned and computed.
- These similarities can mutually **reinforce** each other, consolidating the correlation between two objects collectively.
  - For example, the similarity learned from a dense dataset can reinforce those derived from other sparse datasets, thus helping fill in the missing values of the latter.
- More likely to accurately estimate the similarity between two objects by combining multiple datasets.
- Different datasets can be **blended together** based on similarities.

Semantic based data fusion methods

Similarity based

Matrix Factorisation

Manifold Alignment

Similarity based: Matrix Factorisation

- Decomposes a (sparse) matrix X into the production of low-rank matrices.

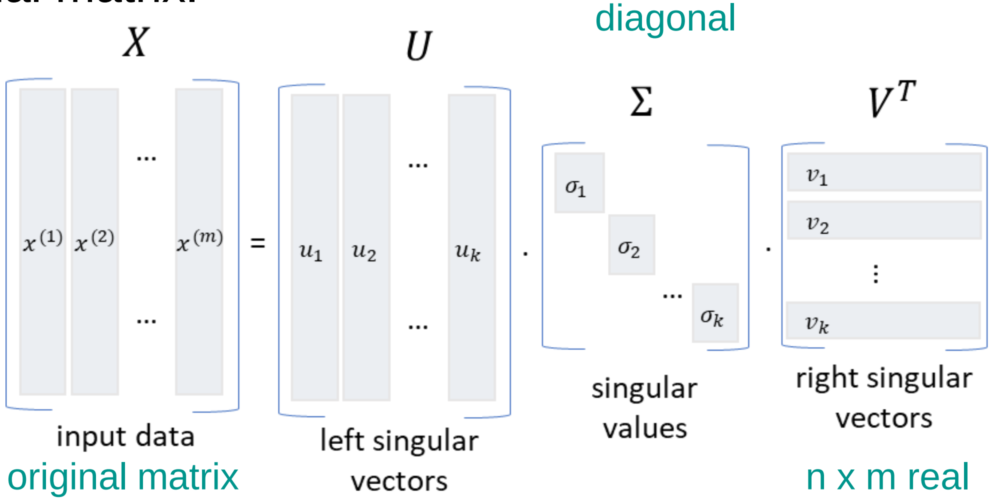
- The low-rank matrices can be used for approximating the original matrix.
  - Helps to fill the missing values.

- Two widely used methods:
  - Singular Value Decomposition.
  - Non-negative Matrix Factorisation.

Similarity based: Matrix Factorisation

- Decomposes a (sparse) matrix X into the production of low-rank matrices.
- The low-rank matrices can be used for approximating the original matrix.
  - Helps to fill the missing values.
- Two widely used methods:
  - **Singular Value Decomposition.**
  - Non-negative Matrix Factorisation.

m x n non-negative real numbers on the diagonal



original matrix

m x m real unitarian matrix

n x m real unitarian matrix

orthogonal

More computationally expensive and harder to parallelize than NFM!!

Similarity based: Matrix Factorisation

- Decomposes a (sparse) matrix X into the production of low-rank matrices.
- The low-rank matrices can be used for approximating the original matrix.
  - Helps to fill the missing values.
- Two widely used methods:
  - Singular Value Decomposition.
  - **Non-negative Matrix Factorisation.**

Factorises an m x n matrix R into a production of an m x K matrix P and a K x n matrix Q

R = P x Q
All elements are non-negative

easier to inspect!

Similarity based: Matrix Factorisation

- Multiple datasets with different properties from different sources cannot be mapped to a unique matrix.
- A single matrix would lead to an inaccurate complementation of missing values in the matrix.

- Advanced methods use coupled factorisations to accommodate different datasets with different matrices **sharing** common dimensions.
- "Collaborative" decomposition.
  - Transfer the similarity between the different elements in a dataset to another one.
  - Complementing more accurately the missing values.

Similarity based: Manifold Alignment

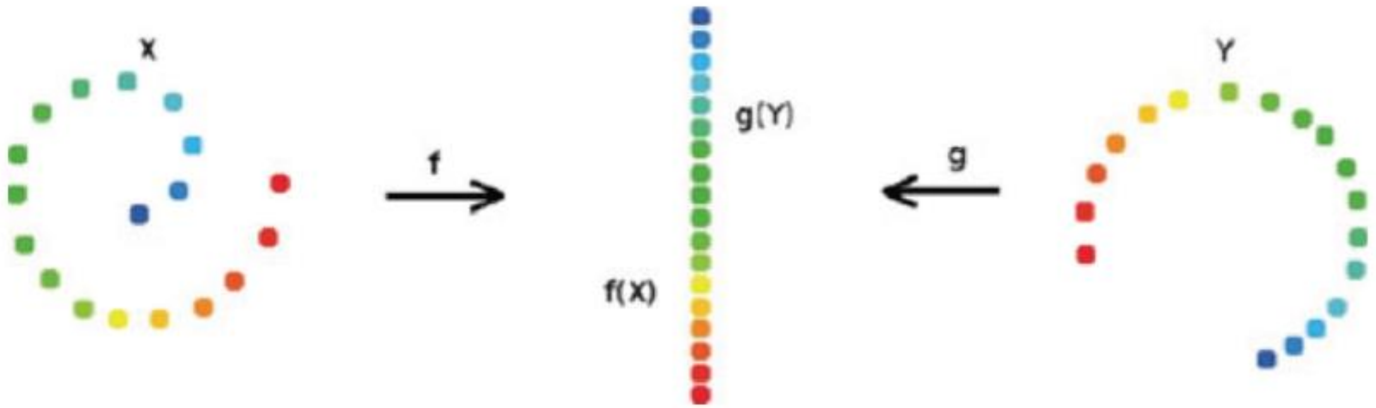
- Uses the relationships of instances **within** each dataset to strengthen the knowledge of the relationships **between** the datasets.
- Joins disparate datasets into a joint latent space.
- Closely related to other manifold learning techniques for dimensionality reduction.
- Given a dataset, it attempts to identify the low dimensional structure of that dataset and preserve that structure in a low dimensional embedding of the dataset.

Similarity based: Manifold Alignment

- Two principles:
  1. It preserves the correspondences across datasets and the individual structures within each dataset by mapping similar instances in each dataset to similar locations in the Euclidean space.

Similarity based: Manifold Alignment

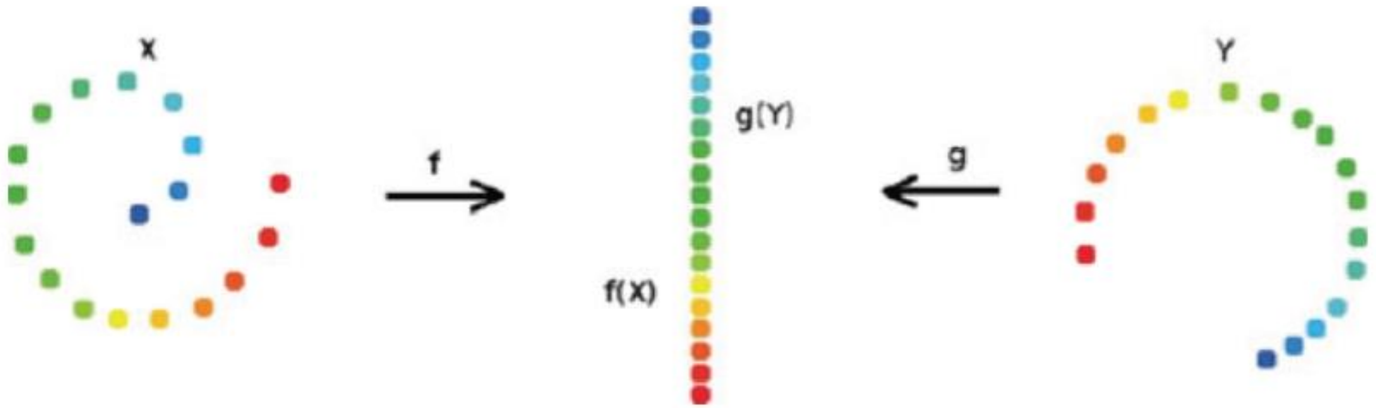
- Two principles:
  1. It preserves the correspondences across datasets and the individual structures within each dataset by mapping similar instances in each dataset to similar locations in the Euclidean space.



Mapping X and Y to a new joint space (f(X),g(Y)), where locally similar instances within each dataset and corresponding instances across datasets are close or identical in that space

Similarity based: Manifold Alignment

- Two principles:
  1. It preserves the correspondences across datasets and the individual structures within each dataset by mapping similar instances in each dataset to similar locations in the Euclidean space.



Mapping X and Y to a new joint space (f(X),g(Y)), where locally similar instances within each dataset and corresponding instances across datasets are close or identical in that space

  2. It assumes that the datasets to be aligned have the same underlying structure.

Probabilistic dependency based

- Bridges the gap between different datasets by the **probabilistic dependency.**

- Emphasizes more about the **interaction** rather than the **similarity** between two elements.

- Generally, uses a graph based representation as the foundation for encoding the complete distribution over a multi-dimensional space.
  - Graphs can be seen as a compact or factorised representations of a set of independences that hold in the specific distribution.

- Two commonly used graph representations.
  - Bayesian Networks.
  - Markov Networks.

Probabilistic dependency based

- Bridges the gap between different datasets by the **probabilistic dependency.**

- Emphasizes more about the **interaction** rather than the **similarity** between two elements.

- Generally, uses a graph based representation as the foundation for encoding the complete distribution over a multi-dimensional space.
  - Graphs can be seen as a compact or factorised representations of a set of independences that hold in the specific distribution.

- Two commonly used graph representations.
  - Bayesian Networks.
  - Markov Networks.

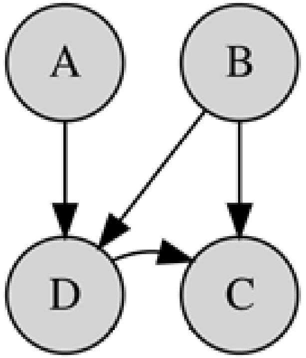
Both families encompass the properties of factorization and independences, but they differ in the set of independences they can encode and the factorization of the distribution that they induce

Probabilistic dependency based

Both families encompass the properties of factorization and independences, but they differ in the set of independences they can encode and the factorization of the distribution that they induce

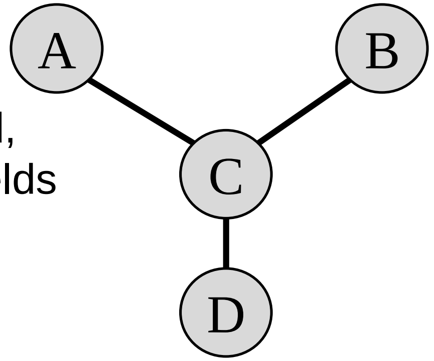
Bayesian Networks

- Directed graph.
- Acyclic.
- Factorises a joint distribution into conditional probabilities.
- E.g. Hidden Markov Model, LDA, Neural Networks.



Markov Networks

- Undirected graph.
- May have cycles.
- Joint probability.
- E.g. Conditional Random Field, Gaussian Markov Random Fields



Transfer-learning based

- An assumption in many machine learning algorithms is that the training and test data must be in the **same feature space** and have the **same distribution**.

- Transfer learning, allows the domains, tasks, and distributions used in training and testing to be **different**.

- Example:
  - A user’s transaction records in Amazon → learn interests → application of travel recommendation.
  - The knowledge learned from one city’s traffic data → another city.

Transfer-learning based

Learning settings		Source and target domains	Source and target tasks
Traditional ML		Same	Same
Transfer Learning	Inductive learning/ Unsupervised	Same	Different but related
		Different but related	Different but related
	Transductive learning	Different but related	The same

- Tasks are different in source and target domains.
- Storing knowledge gained while solving one problem and applying it to a different but related problem.

Transfer-learning based

Learning settings		Source and target domains	Source and target tasks
Traditional ML		Same	Same
Transfer Learning	Inductive learning/ Unsupervised	Same	Different but related
		Different but related	Different but related
	Transductive learning	Different but related	The same

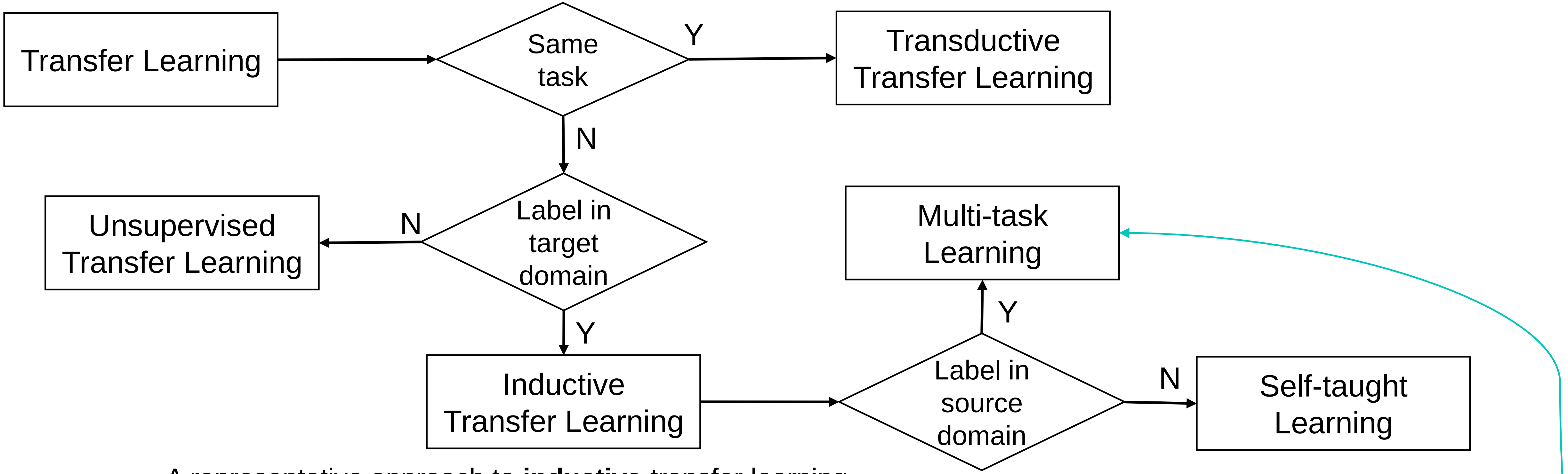
- Two categories.
  1. Same feature spaces between domains → Different marginal probability distribution.
     - Most of transfer approaches.
  2. Different Feature spaces between domains.

Transfer the traffic data from a city to another one where training data are limited

Clustering web pages. One dataset in Spanish, the other in English.

Transfer-learning based

Taxonomy according to whether label data are available in source and target domains.



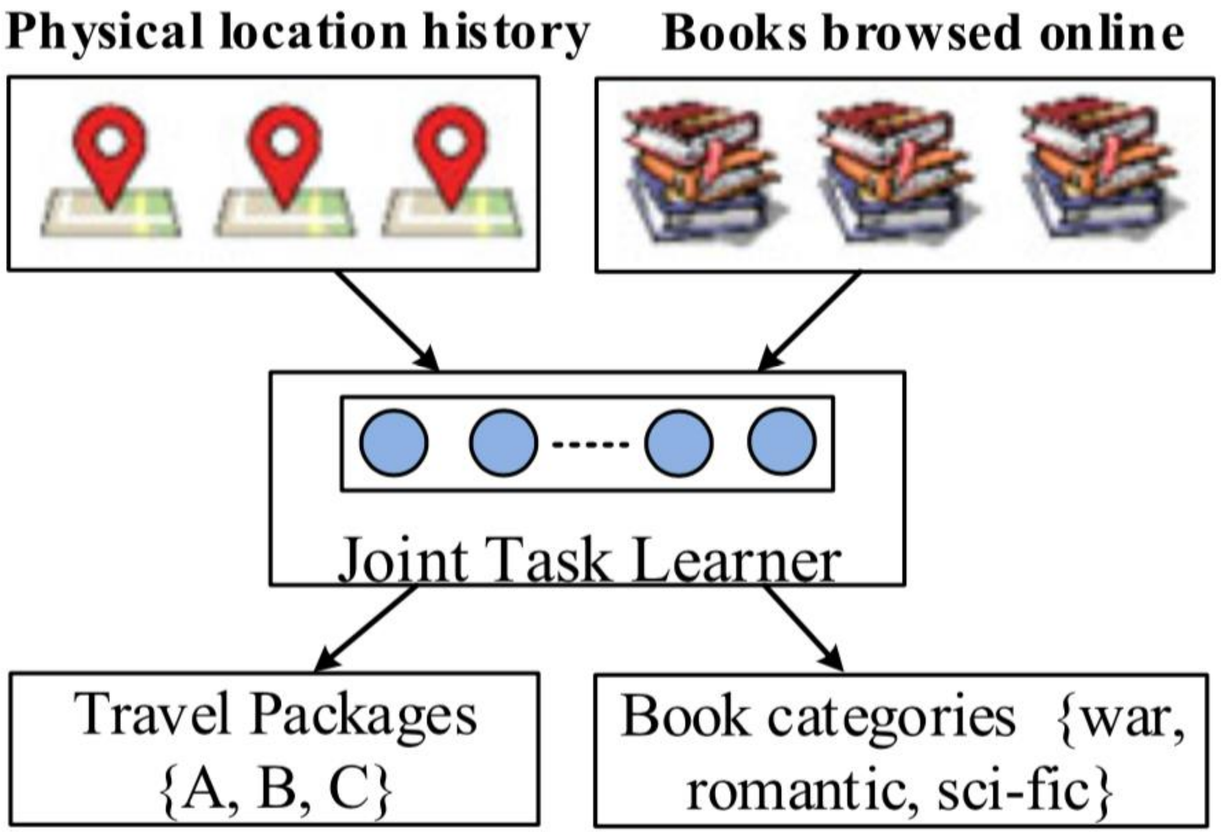
- A representative approach to **inductive** transfer learning.
- Learns a problem together with other related problems at the same time, using a shared representation.
- Often leads to a better model for the main task, because it allows the learner to use the commonality among the tasks.
- Useful in sparse settings.

Transfer-learning based

- Transfer learning between two classification tasks.
  - Infer interests in travel packages in terms of the searched books.
  - Infer interests in book styles in terms of location history.

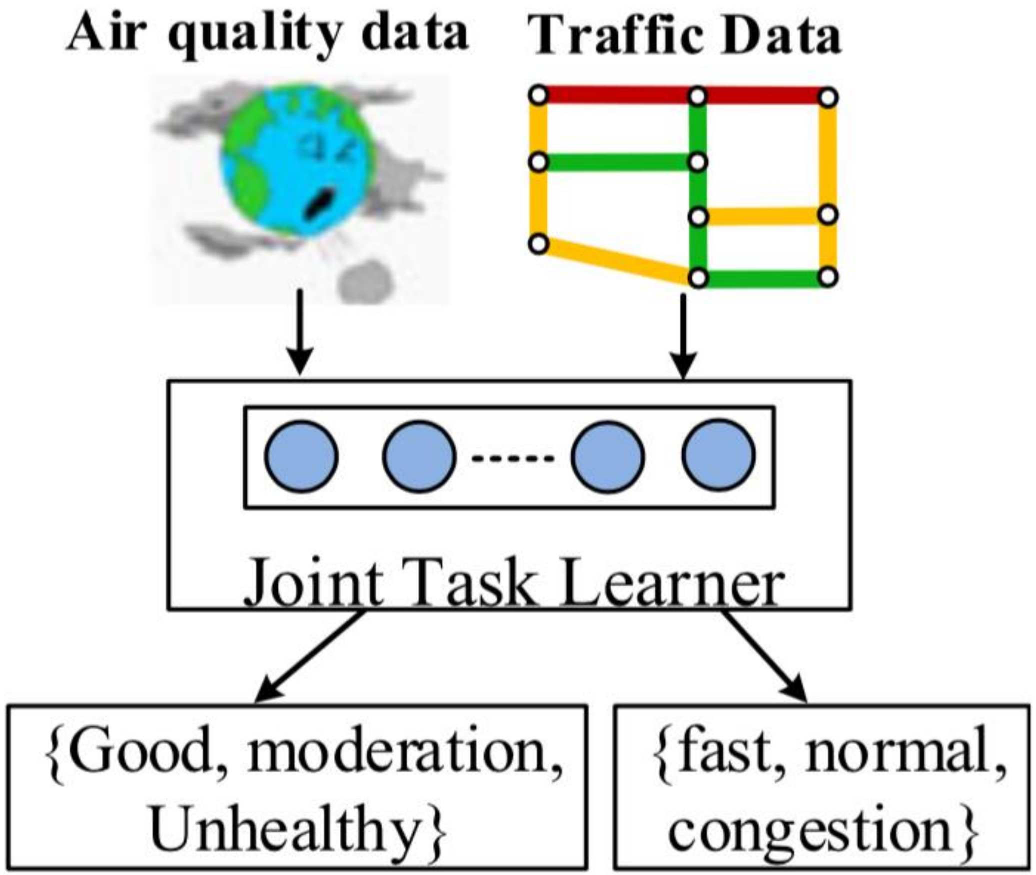
- With the two datasets from the same users, both tasks can be associated into a MTL framework.
  - Learn the shared representations of users' general interests.

- As the books a user has browsed may imply her general interests and characters, which can be transferred into the travel package recommendation.
- The knowledge from a user's physical location can also help estimate a user's interests in different book styles.



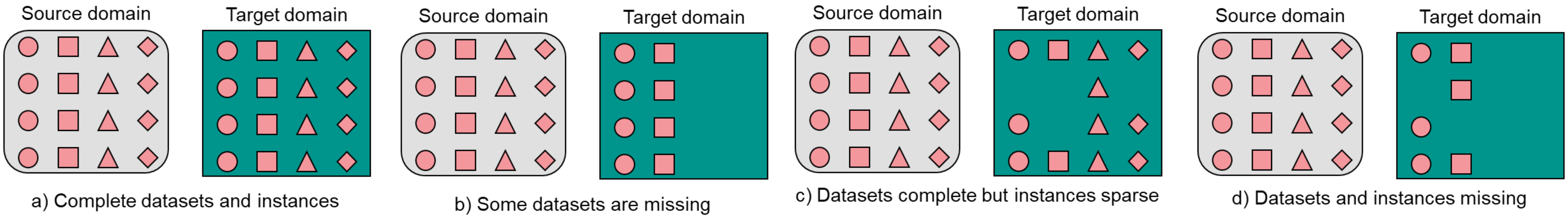
Transfer-learning based

- Co-prediction of the air quality and traffic conditions at near future simultaneously.
- The general insight is that different traffic conditions will generate different volumes of air pollutants, therefore impacting the air quality differently.
- Likewise, people tend to go hiking or picnic in a day with good air quality, while preferring to minimize travel in a day with hazardous air quality.
- As a result, the traffic conditions are also affected by air quality.



Transfer-learning based: Multiple datasets

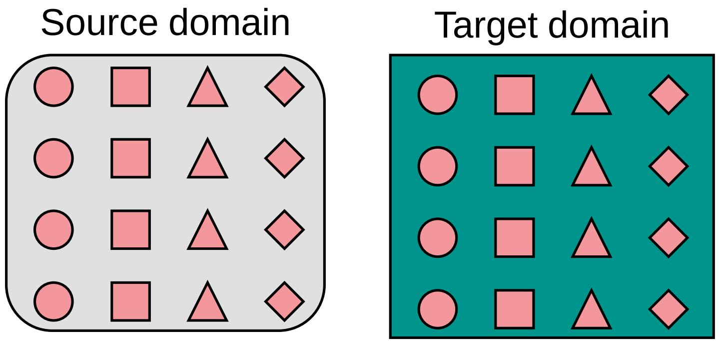
- In the big data era, many machine learning tasks have to harness a diversity of data in a domain so as to achieve a better performance.
- New techniques that can transfer the knowledge of multiple datasets from a source to a target domain!
- Four situations of transfer learning for multiple datasets:



Transfer-learning based: Multiple datasets

Datasets and instances complete

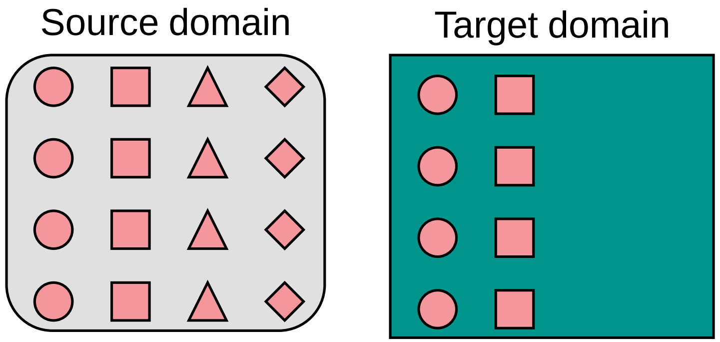
- Target domain has all kinds of source domain.
- Each target dataset has sufficient observations.
- The target domain has the same (and sufficient) feature spaces as the source domain.
- Multi-task learning.



a) Complete datasets and instances

Some datasets are missing

- Target domain is incomplete.
- Other datasets have sufficient observations.
- New techniques for dealing with structurally missing datasets.



b) Some datasets are missing

Semantic based data fusion methods

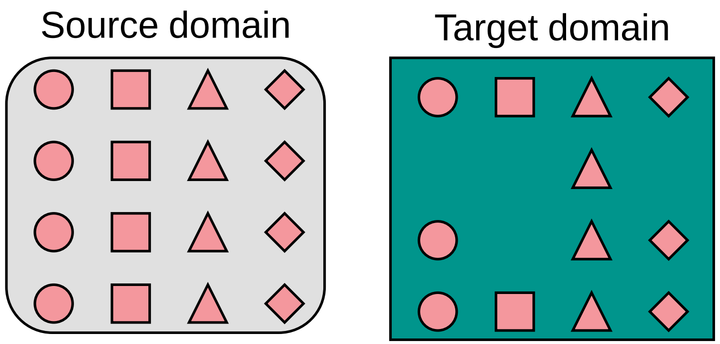
Transfer-learning based: Multiple datasets

Datasets complete but instances sparse

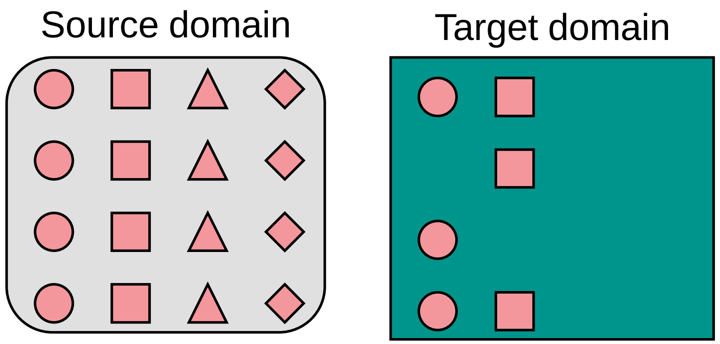
Datasets incomplete and instances sparse

No solution for these cases!

Open research problema!!



c) Datasets complete but instances sparse



d) Datasets and instances missing

A comparison of data fusion techniques

indicates if a method can incorporate other approaches as a meta method. For instance, a semantic meaning-based data fusion methods can be employed in a stage-based fusion method.

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		**Y**	NA	NA	NA	NA	NA
Feature	Concat.+Reg.	**N**	L	Flex	F, P, C, O, R	S	Y
	DNN	**N**	L	Flex	F, P, A, O, R	U/S	Y
Semantic	Multiview	**Y**	S	Fix	F, P, O	S, SS	Y
	Probabil.	**N**	S	Fix	F, P, C, O, A	S/U	N
	Similarity	**N**	S	Flex	F, A, O	U	Y
	Transfer	**Y**	S	Flex	F, P, A, R	S/U	Y

A comparison of data fusion techniques

feature-based methods need a large (L) amount of labelled instances as training data, while the semantic meaning-based methods can be applied to a dataset with a small (S) number of labelled instances.

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		Y	**NA**	NA	NA	NA	NA
Feature	Concat.+Reg.	N	**L**	Flex	F, P, C, O, R	S	Y
	DNN	N	**L**	Flex	F, P, A, O, R	U/S	Y
Semantic	Multiview	Y	**S**	Fix	F, P, O	S, SS	Y
	Probabil.	N	**S**	Fix	F, P, C, O, A	S/U	N
	Similarity	N	**S**	Flex	F, A, O	U	Y
	Transfer	Y	**S**	Flex	F, P, A, R	S/U	Y

A comparison of data fusion techniques

directly combining features extracted from static data and those from dynamic data would cause static features ignored by a machine learning model

E.g. sensors constantly providing air quality measurements vs noise complains
Not all regions have sensors, not all regions received complains, some regions may receive complains but not have sensors

Methods		Meta	Labels		Goals	Train	Scale
			Volume	**Position**			
Stage-based		Y	NA	**NA**	NA	NA	NA
Feature	Concat.+Reg.	N	L	**Flex**	F, P, C, O, R	S	Y
	DNN	N	L	**Flex**	F, P, A, O, R	U/S	Y
Semantic	Multiview	Y	S	**Fix**	F, P, O	S, SS	Y
	Probabil.	N	S	**Fix**	F, P, C, O, A	S/U	N
	Similarity	N	S	**Flex**	F, A, O	U	Y
	Transfer	Y	S	**Flex**	F, P, A, R	S/U	Y

A comparison of data fusion techniques

Bayesian Networks and the straightforward feature-based fusion methods (e.g. when using a linear regression model) are usually good at dealing with causality inference problems (C).

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		Y	NA	NA	NA	NA	NA
Feature	Concat.+Reg.	N	L	Flex	F, P, C, O, R	S	Y
	DNN	N	L	Flex	F, P, A, O, R	U/S	Y
Semantic	Multiview	Y	S	Fix	F, P, O	S, SS	Y
	Probabil.	N	S	Fix	F, P, C, O, A	S/U	N
	Similarity	N	S	Flex	F, A, O	U	Y
	Transfer	Y	S	Flex	F, P, A, R	S/U	Y

Filling Missing Values, Predict Future, Causality Inference, Object Profiling, Ranking, and Anomaly Detection

A comparison of data fusion techniques

not every technique supports unsupervised learning

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		Y	NA	NA	NA	**NA**	NA
Feature	Concat.+Reg.	N	L	Flex	F, P, C, O, R	**S**	Y
	DNN	N	L	Flex	F, P, A, O, R	**U/S**	Y
Semantic	Multiview	Y	S	Fix	F, P, O	**S, SS**	Y
	Probabil.	N	S	Fix	F, P, C, O, A	**S/U**	N
	Similarity	N	S	Flex	F, A, O	**U**	Y
	Transfer	Y	S	Flex	F, P, A, R	**S/U**	Y

supervised (S), unsupervised (U) and semi-supervised (SS) learning

A comparison of data fusion techniques

it is not easy for probabilistic dependency-based approaches to scale efficiency and up (N). A graphical model with a complex structure, e.g. many (hidden) nodes and layers, may become intractable

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		Y	NA	NA	NA	NA	**NA**
Feature	Concat.+Reg.	N	L	Flex	F, P, C, O, R	S	**Y**
	DNN	N	L	Flex	F, P, A, O, R	U/S	**Y**
Semantic	Multiview	Y	S	Fix	F, P, O	S, SS	**Y**
	Probabil.	N	S	Fix	F, P, C, O, A	S/U	**N**
	Similarity	N	S	Flex	F, A, O	U	**Y**
	Transfer	Y	S	Flex	F, P, A, R	S/U	**Y**

A comparison of data fusion techniques

when a matrix becomes very large, NMF, which can be operated in parallel, can be employed to expedite decomposition (Y).

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		Y	NA	NA	NA	NA	**NA**
Feature	Concat.+Reg.	N	L	Flex	F, P, C, O, R	S	**Y**
	DNN	N	L	Flex	F, P, A, O, R	U/S	**Y**
Semantic	Multiview	Y	S	Fix	F, P, O	S, SS	**Y**
	Probabil.	N	S	Fix	F, P, C, O, A	S/U	**N**
	Similarity	N	S	Flex	F, A, O	U	**Y**
	Transfer	Y	S	Flex	F, P, A, R	S/U	**Y**

A comparison of data fusion techniques

given the same amount of training data feature-based methods are not as good as semantic meaning-based approaches, as there are dependencies and correlations between features.
Adding a sparsity regularization can alleviate the problem to some extents, but cannot solve it fundamentally.

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		Y	NA	NA	NA	NA	**NA**
Feature	Concat.+Reg.	N	L	Flex	F, P, C, O, R	S	**Y**
	DNN	N	L	Flex	F, P, A, O, R	U/S	**Y**
Semantic	Multiview	Y	S	Fix	F, P, O	S, SS	**Y**
	Probabil.	N	S	Fix	F, P, C, O, A	S/U	**N**
	Similarity	N	S	Flex	F, A, O	U	**Y**
	Transfer	Y	S	Flex	F, P, A, R	S/U	**Y**

A comparison of data fusion techniques

with a large amount of labelled data DNNs can perform well. However, the performance of the model heavily relies on tuning parameters. Given a huge model with many parameters to learn, this is usually a time-consuming process that needs the involvement of human experiences

Methods		Meta	Labels		Goals	Train	Scale
			Volume	Position			
Stage-based		Y	NA	NA	NA	NA	**NA**
Feature	Concat.+Reg.	N	L	Flex	F, P, C, O, R	S	**Y**
	DNN	N	L	Flex	F, P, A, O, R	U/S	**Y**
Semantic	Multiview	Y	S	Fix	F, P, O	S, SS	**Y**
	Probabil.	N	S	Fix	F, P, C, O, A	S/U	**N**
	Similarity	N	S	Flex	F, A, O	U	**Y**
	Transfer	Y	S	Flex	F, P, A, R	S/U	**Y**

Summary

"Urban computing is a concept where **every sensor, device, person, vehicle, building, and stree**t in the urban areas can be used as a component to sense **city dynamics** to enable a **city-wide computing** to tackle the **challenges** in **urban areas** as so to serve people and cities.

- Data from cities:

Spatially and temporally static	Spatially static and temporally dynamic	Spatially and temporally dynamic

Summary

- The proliferation of big data calls for advanced data fusion methods that can discover knowledge from multiple datasets with underlying connections

Data fusion defines a **combination of multiple sources** to obtain **improved information**; in this context, improved information means **less expensive**, **higher quality**, or **more relevant information**.

- **Stage-based:** Use different datasets at different stages of a data mining task.
- **Feature level-based:** Learns a new representation of the original features extracted from different datasets.
- **Semantic-based:** Blends data based on their semantic meanings.

References

- Zheng, Y., et al. **Urban Computing: concepts, methodologies, and applications**. *ACM transactions on Intelligent Systems and Technology*

- Torra, V., 2003. Trends in information fusion in data mining. In *Information fusion in data mining* (pp. 1-6). Springer, Berlin, Heidelberg.

- Zheng, Y., 2015. Methodologies for cross-domain data fusion: An overview. *IEEE transactions on big data*, *1*(1), pp.16-34.

- Castanedo, F., 2013. A review of data fusion techniques. *The Scientific World Journal*, *2013*.

Thanks!

Questions?

ACM Summer School on User Modeling and Personalization in Urban Computing:

Data Fusion

Dr Antonela Tommasel

ISISTAN, CONICET-UNICEN, Argentina

CONICET

I S I S T A N