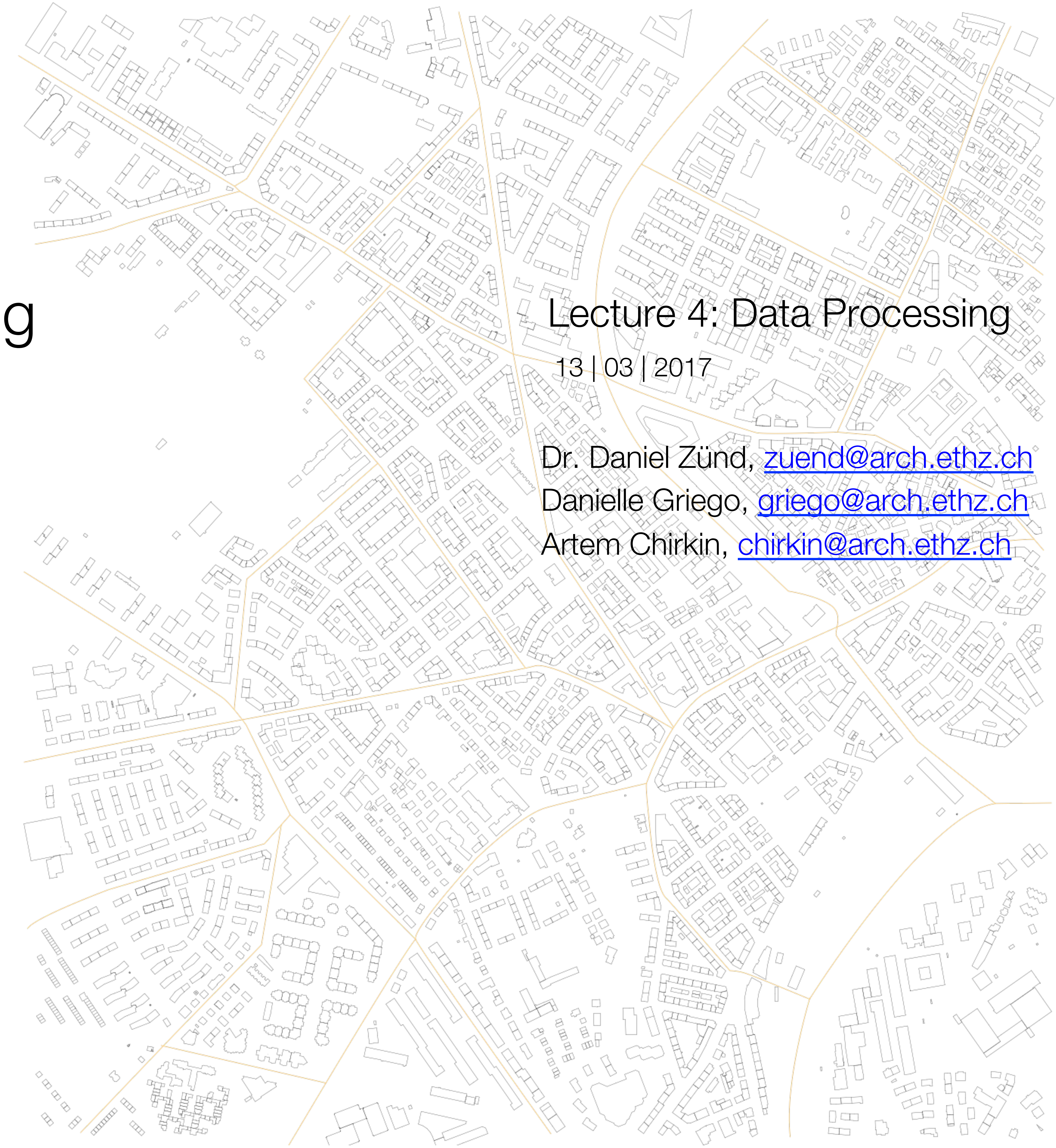
Creative Data Mining

Uncover and evaluate



Lecture 4: Data Processing

13 | 03 | 2017

Dr. Daniel Zünd, zuend@arch.ethz.ch

Danielle Griego, griego@arch.ethz.ch

Artem Chirkin, chirkin@arch.ethz.ch

ETH zürich DARCH iA Chair of Information Architecture

What we'll cover today

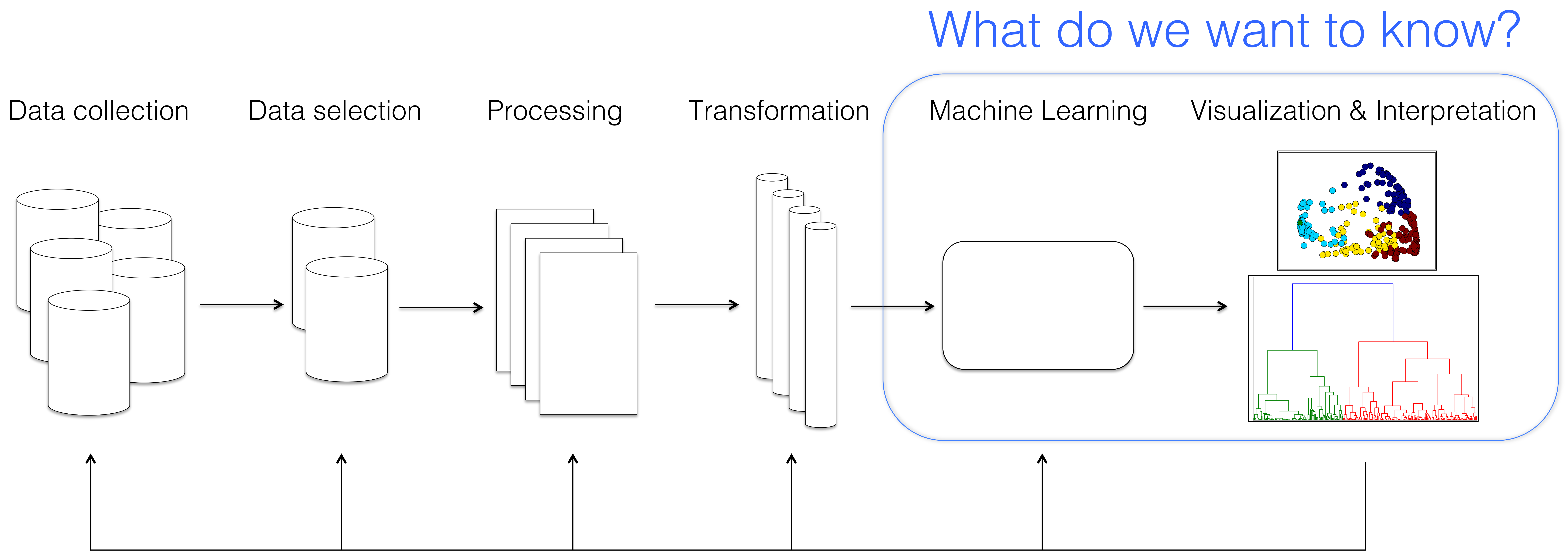
Data Processing

- A bit of theory & background
- Example: ESUM Project
- Start thinking about final project
- 2nd programming assignment

ETH zürich DARCH iA Chair of Information Architecture

Background

How can data mining be creative?

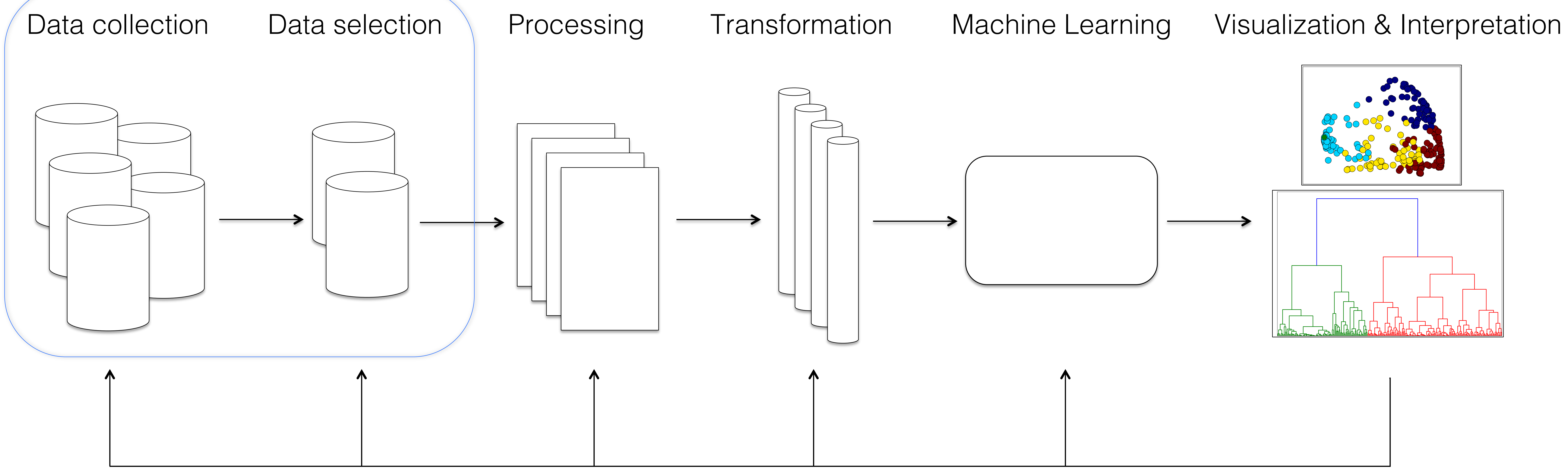


Typical Knowledge Discovery Diagram (KDD)

Background

How can data mining be creative?

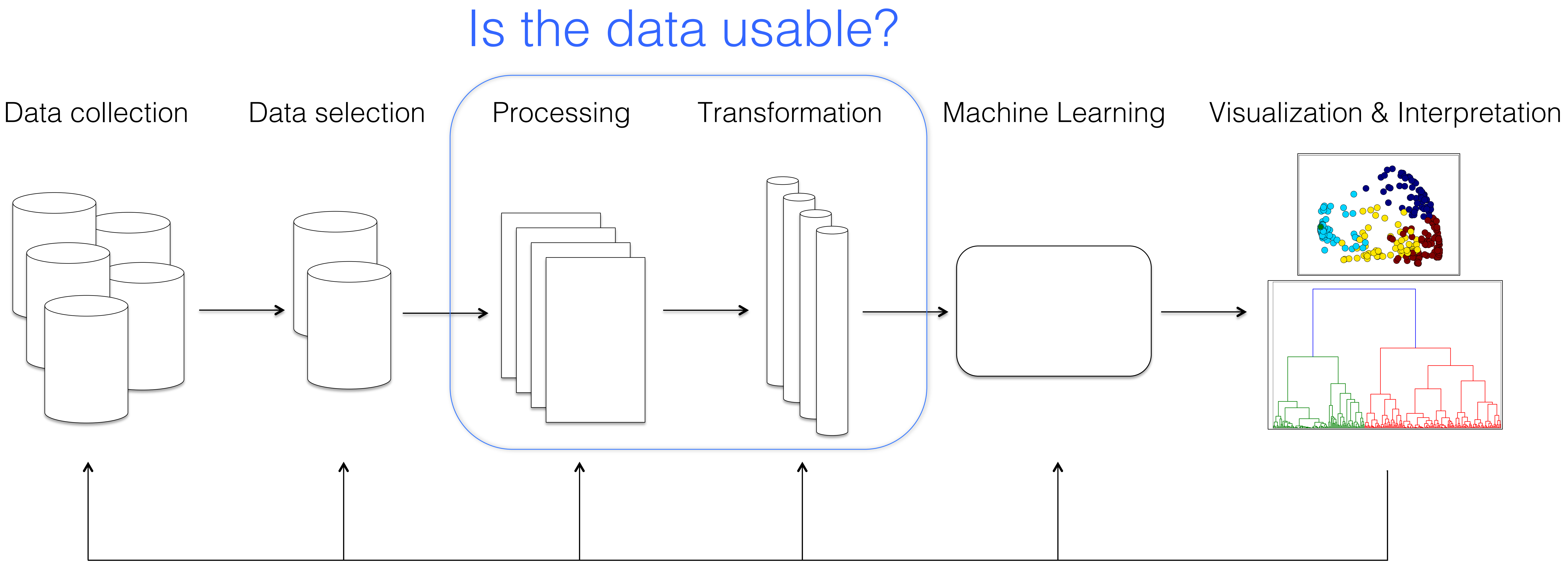
Domain specific data source(s)



Typical Knowledge Discovery Diagram (KDD)

Background

The not-so creative, but essential part of data mining



Typical Knowledge Discovery Diagram (KDD)

Data Quality

"Garbage-in, garbage-out"

http://www.marksgroup.net/blog/zoho-crm-garbage-in-garbage-out-its/

Data Quality

What is dirty data?

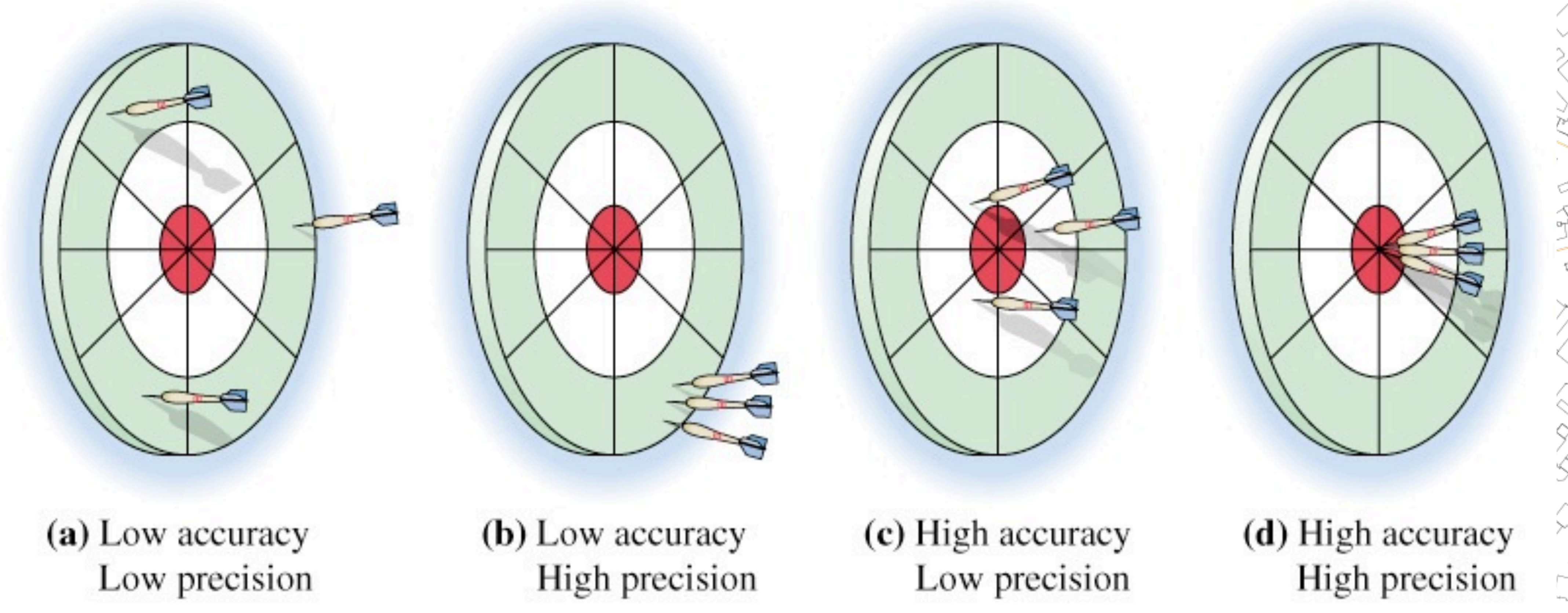
Dirty data is……

inaccurate, imprecise, unreliable, incomplete, inconsistent, irrelevant, invalid, ambiguous, redundant, or faked

Varun Ojha, (2015) A Computational Intelligence Perspective of Modelling and Data Analysis (Knowledge Discovery)

Data Quality

- Accurate
- Precise
- Reliable
- Complete
- Consistent
- Relevant
- Sufficient (size)



http://blog.extensionengine.com/accuracy-precision/

Data Processing

Major tasks in data preprocessing

Data cleaning

- Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Data integration

- Integration of multiple databases, data cubes, or files

Data transformation

- Normalization and aggregation

Data reduction

- Obtains reduced representation in volume but produces the same or similar analytical results

Data discretization

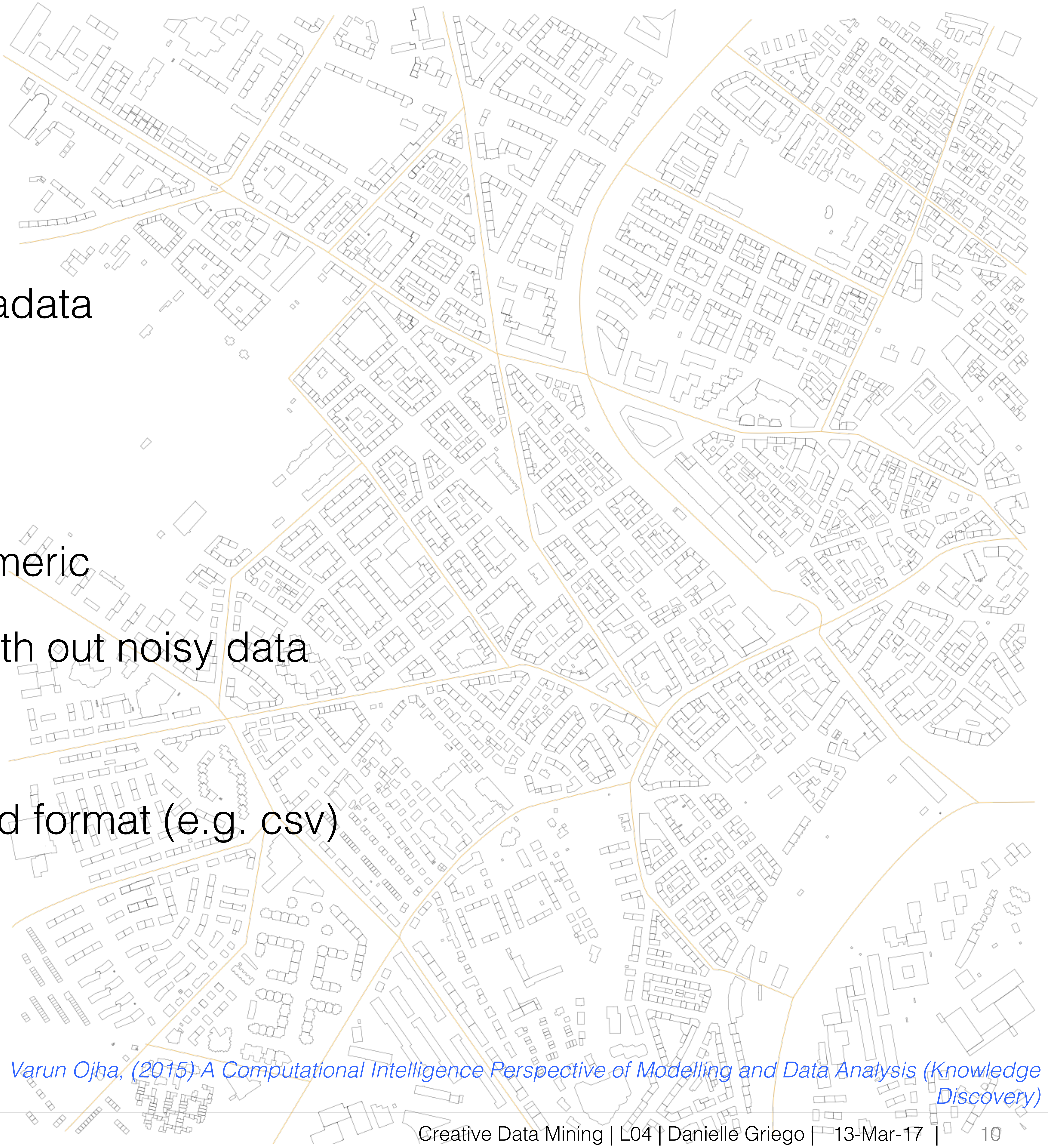
- Part of data reduction but with particular importance, especially for numerical data

https://www.mimuw.edu.pl/~son/datamining/DM/4-preprocess.pdf

Data Processing

Cleaning Tasks

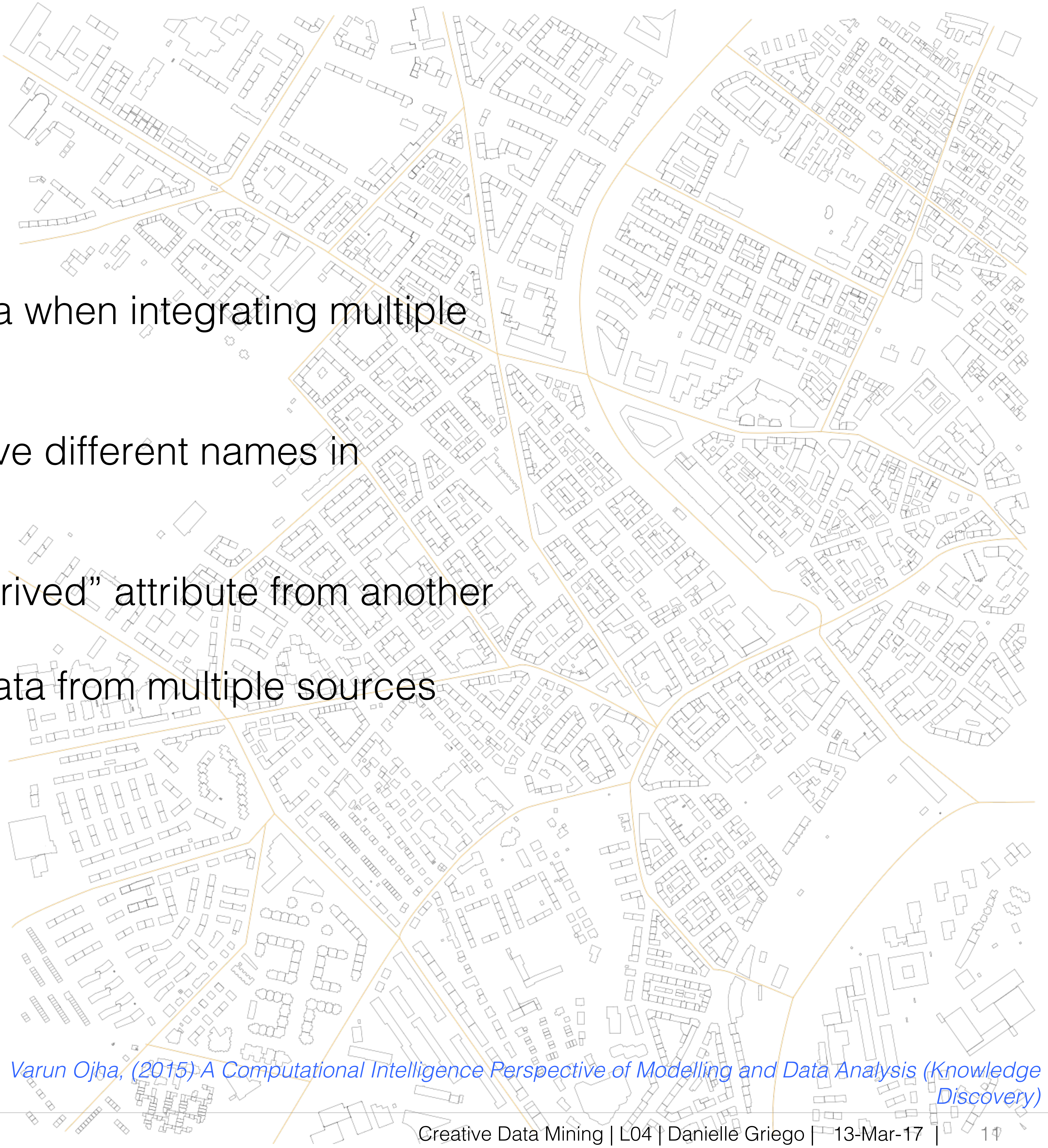
- Data acquisition and metadata
- Fill in missing values
- Unified date format
- Converting nominal to numeric
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Convert data to a standard format (e.g. csv)



Varun Ojha, (2015) A Computational Intelligence Perspective of Modelling and Data Analysis (Knowledge Discovery)

Data Processing

Data Integration



- Look out for redundant data when integrating multiple databases

- The same attribute may have different names in different databases

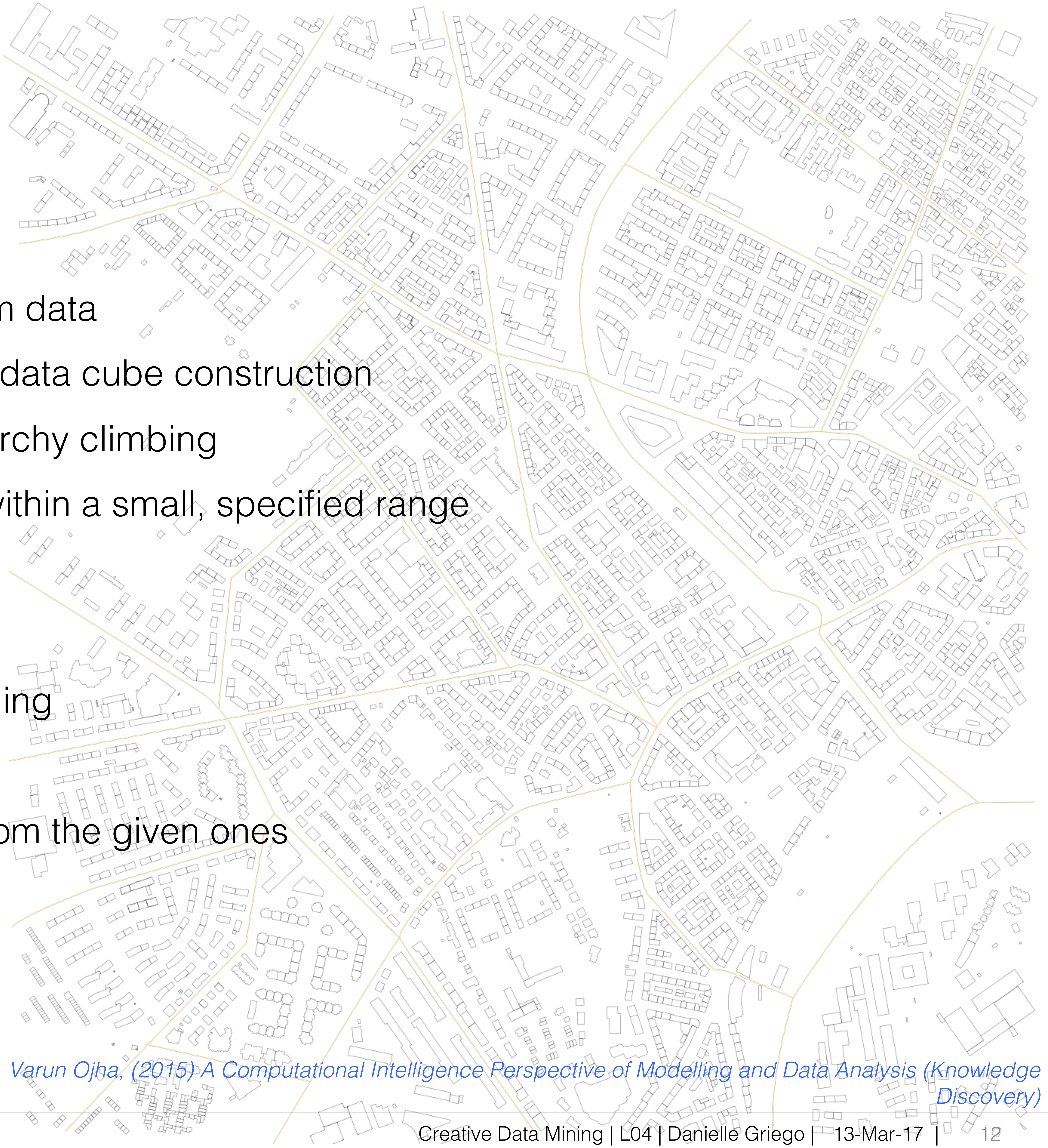
- One attribute may be a "derived" attribute from another

- Careful integration of the data from multiple sources

Varun Ojha, (2015) A Computational Intelligence Perspective of Modelling and Data Analysis (Knowledge Discovery)

Data Processing

Data transformation

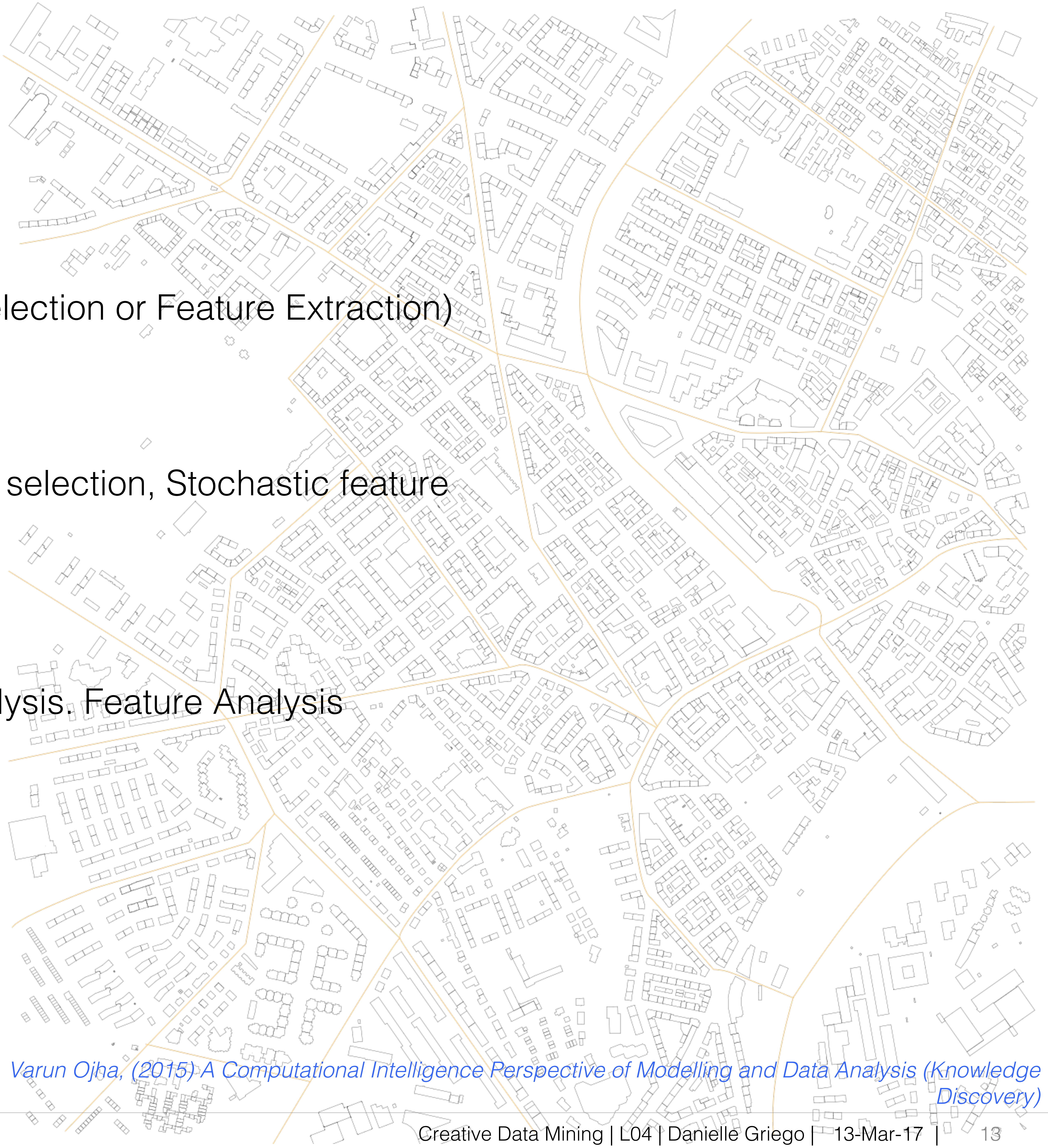


- Smoothing: remove noise from data
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
- min-max normalization
- z-score normalization
- normalization by decimal scaling
- Attribute/feature construction
- New attributes constructed from the given ones

Varun Ojha, (2015) A Computational Intelligence Perspective of Modelling and Data Analysis (Knowledge Discovery)

Data Processing

Data reduction



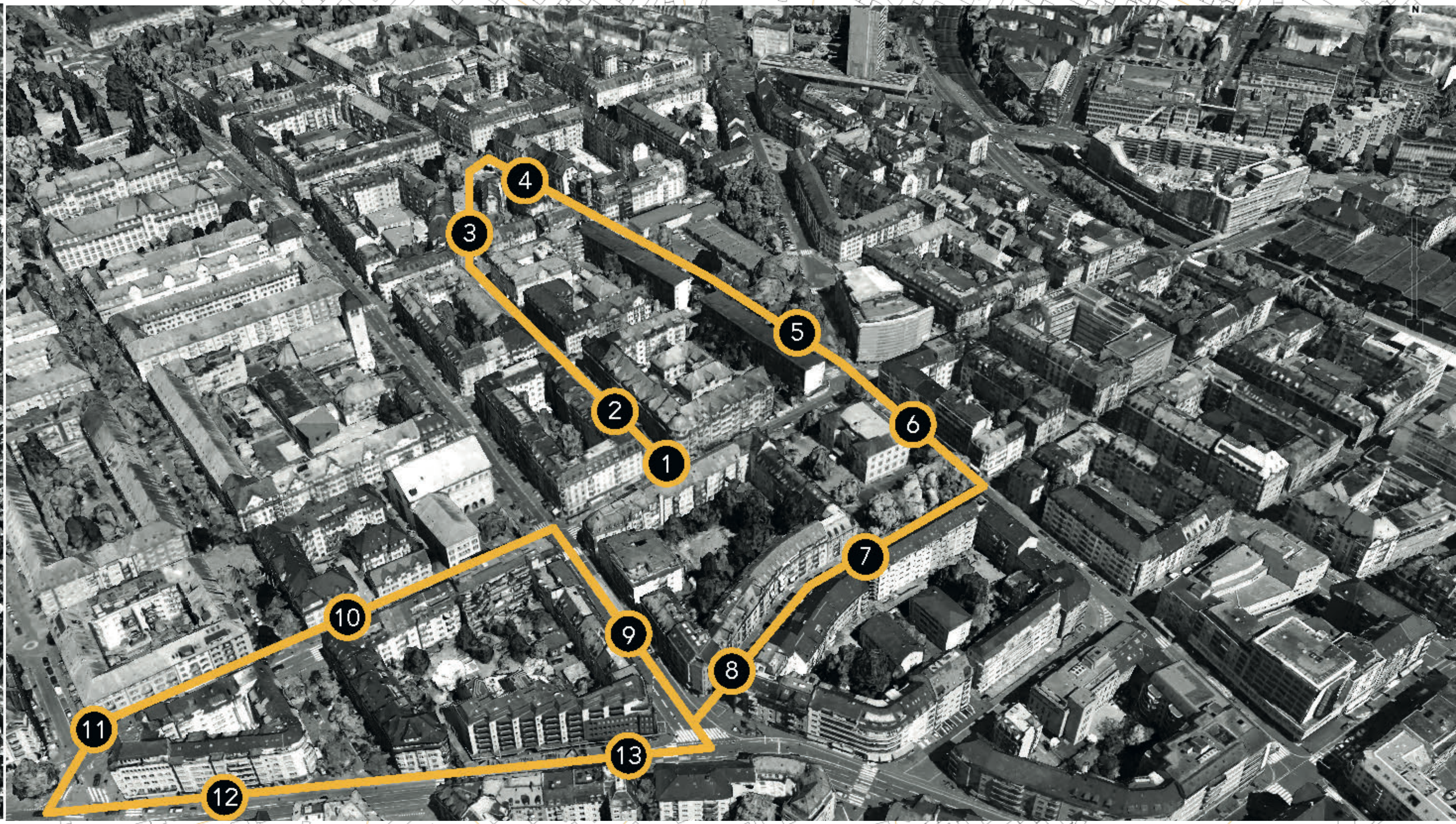
- Feature Reduction (Feature Selection or Feature Extraction)
- Feature Selection
- E.g. Correlation-based feature selection, Stochastic feature Selection.
- Feature Extraction
- E.g. Principle Component Analysis. Feature Analysis

Varun Ojha, (2015) A Computational Intelligence Perspective of Modelling and Data Analysis (Knowledge Discovery)

Example

ESUM- Analyzing trade-offs between Energy and Social performance of Urban Morphologies

Location Wiedikon Zürich



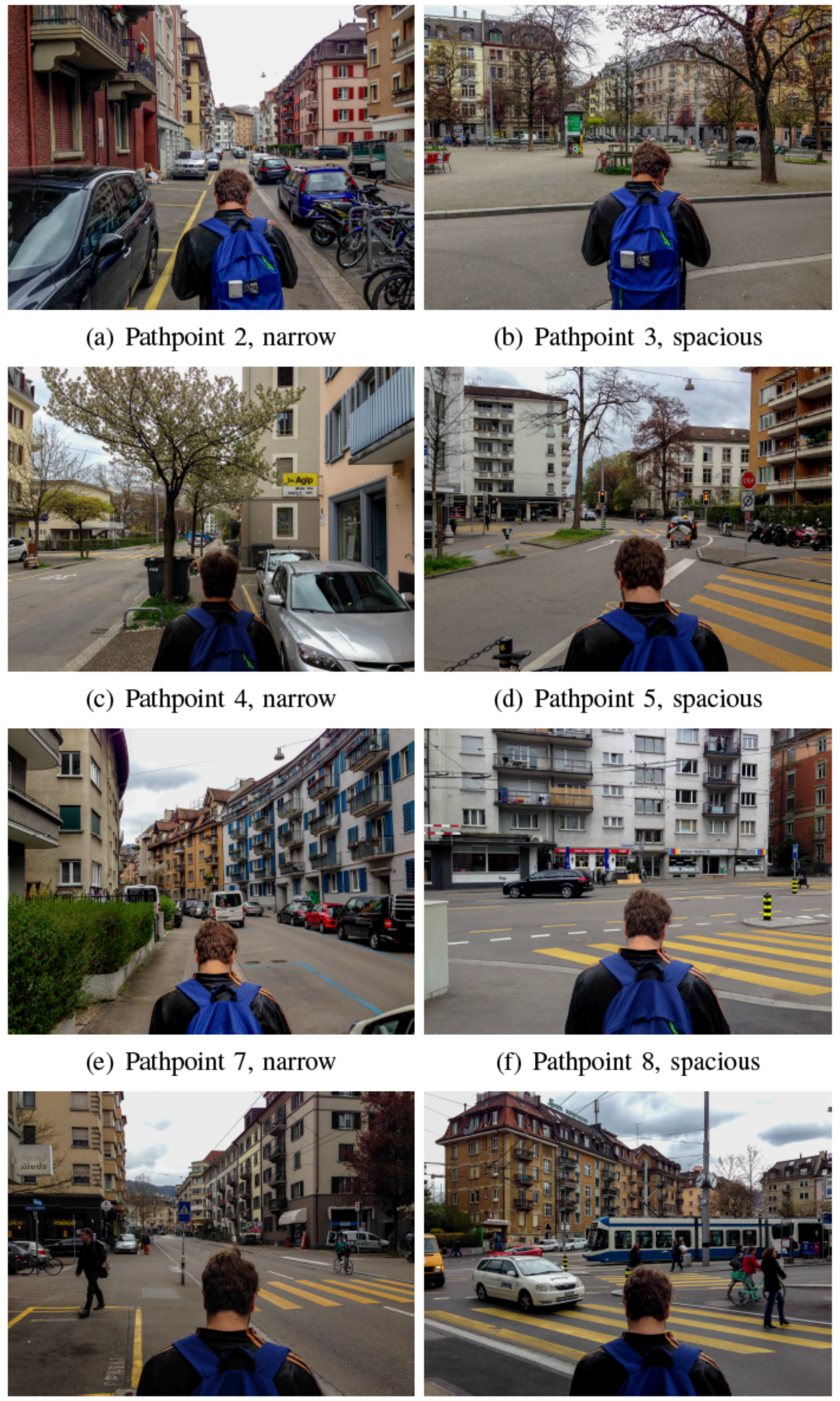
14 survey checkpoints along experimental path

Example

ESUM- Analyzing trade-offs between Energy and Social performance of Urban Morphologies

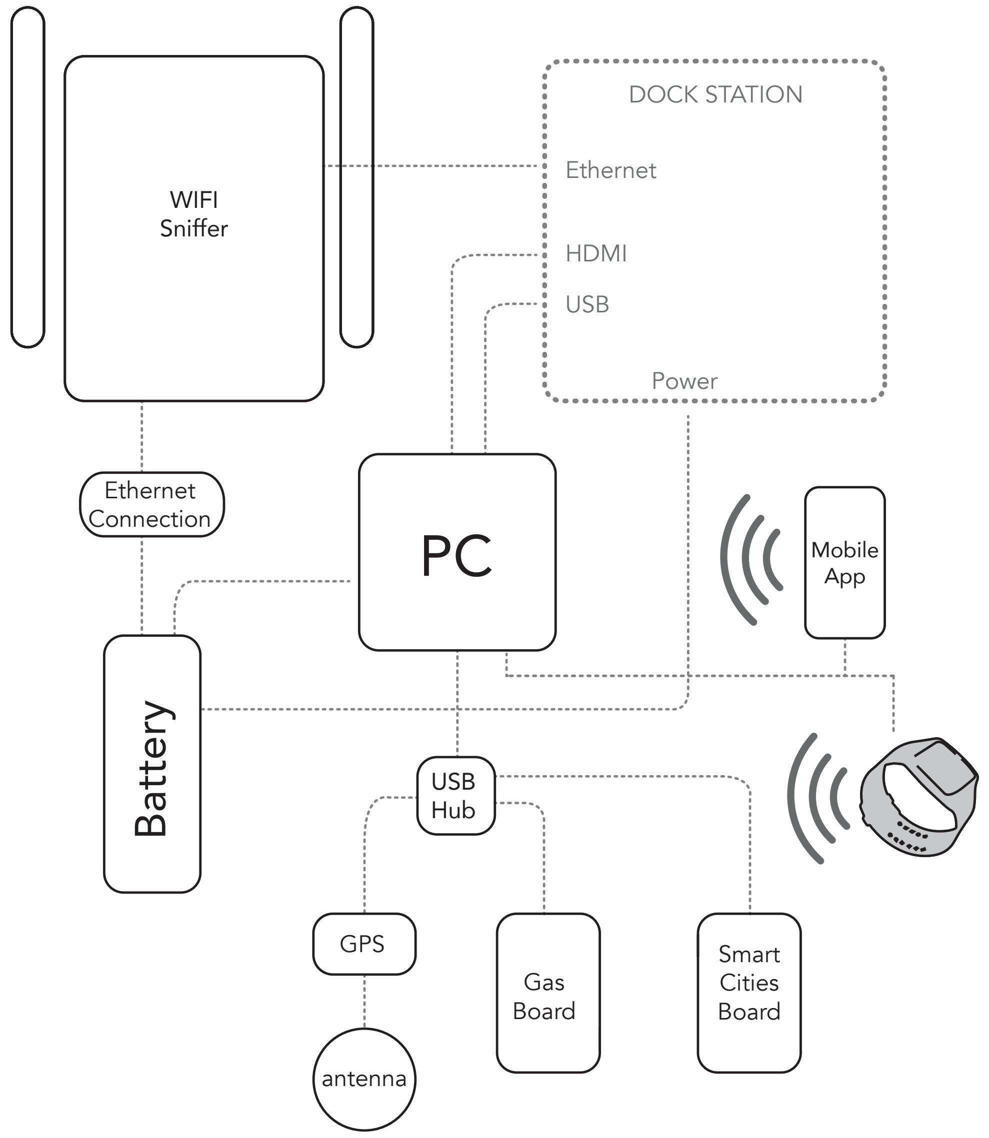
Data from 37 participants in Zurich to:

- Investigate impact of static (urban morphology) and dynamic features (environmental sensors) of the built environment on perception (using surveys and biofeedback data)



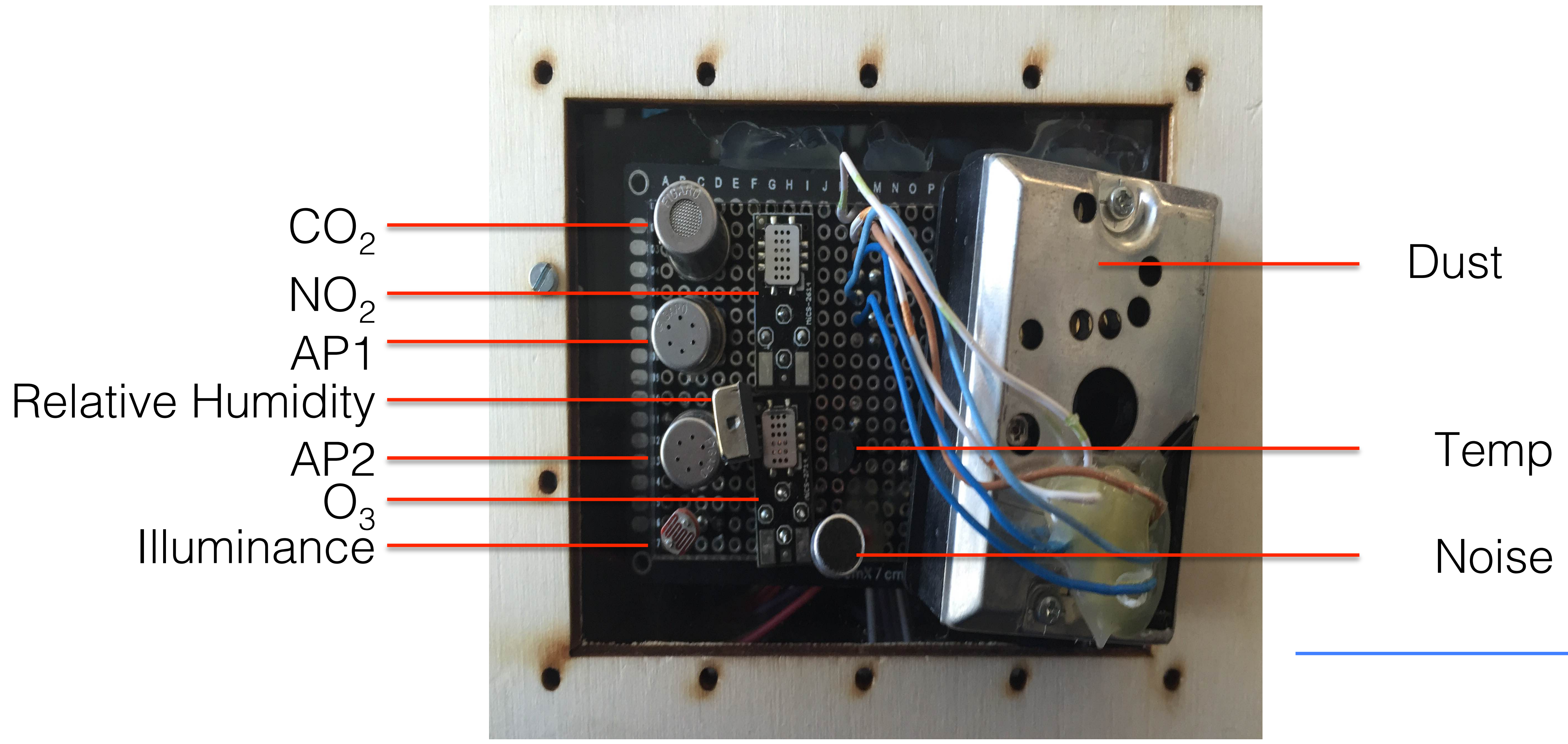
(a) Pathpoint 2, narrow (b) Pathpoint 3, spacious (c) Pathpoint 4, narrow (d) Pathpoint 5, spacious (e) Pathpoint 7, narrow (f) Pathpoint 8, spacious (g) Pathpoint 10, narrow (h) Pathpoint 11, spacious

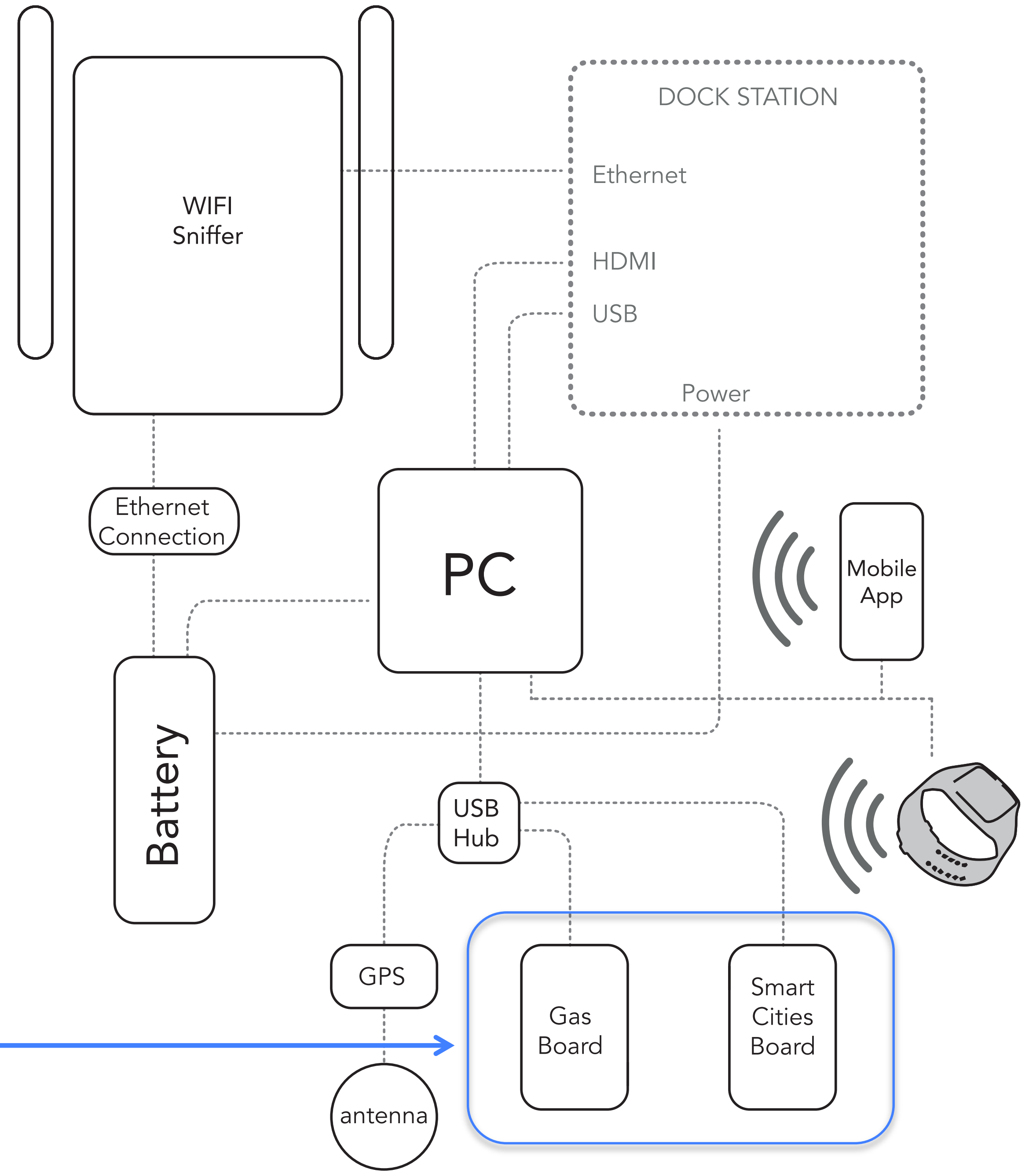
Fig. 2. Four instances of narrow-spacious spatial configurations and their corresponding pathpoints along the select path.



Mobile sensor equipment

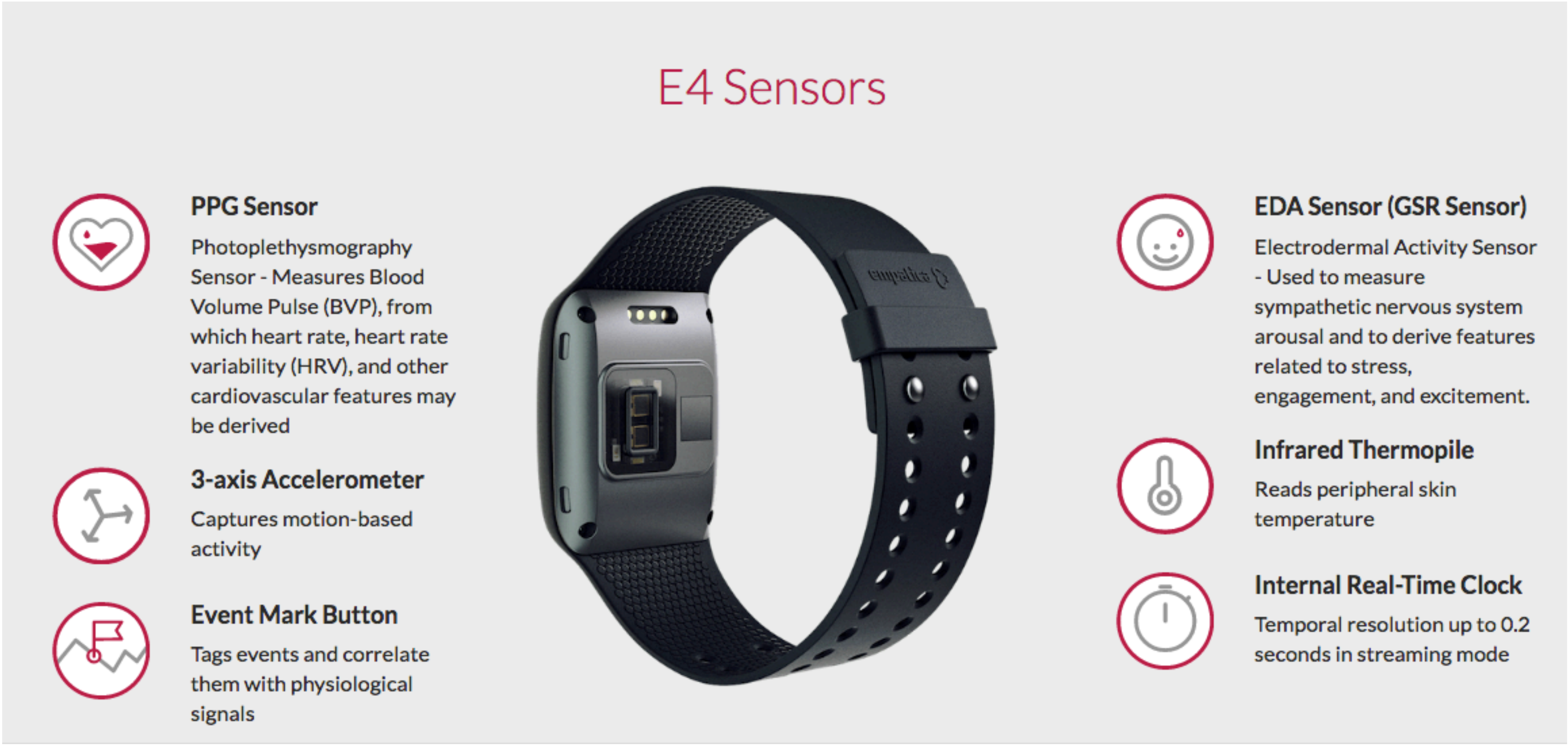
Sensorbackpack with environmental and position sensors

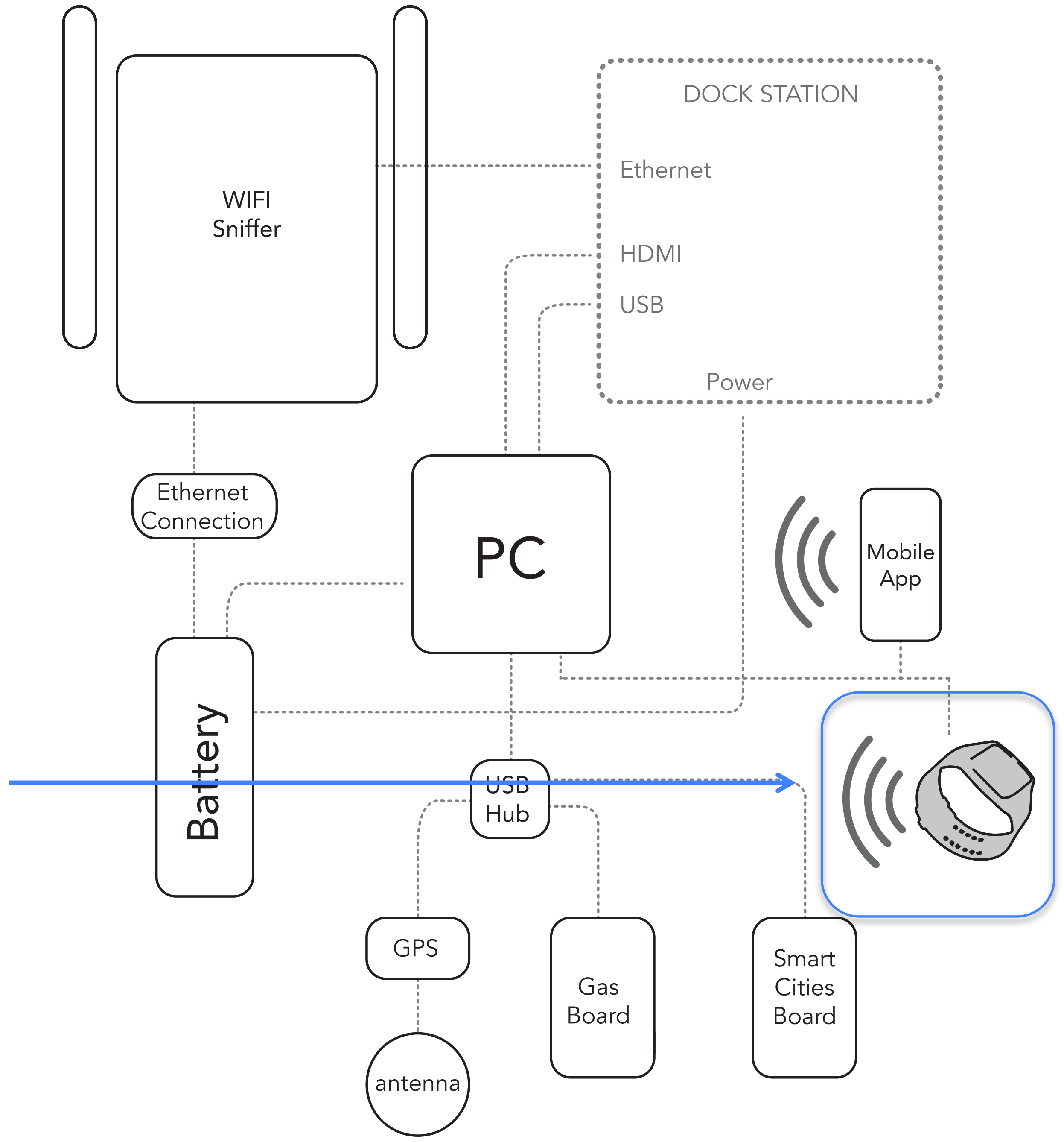




Mobile Sensor equipment

Biofeedback wristband

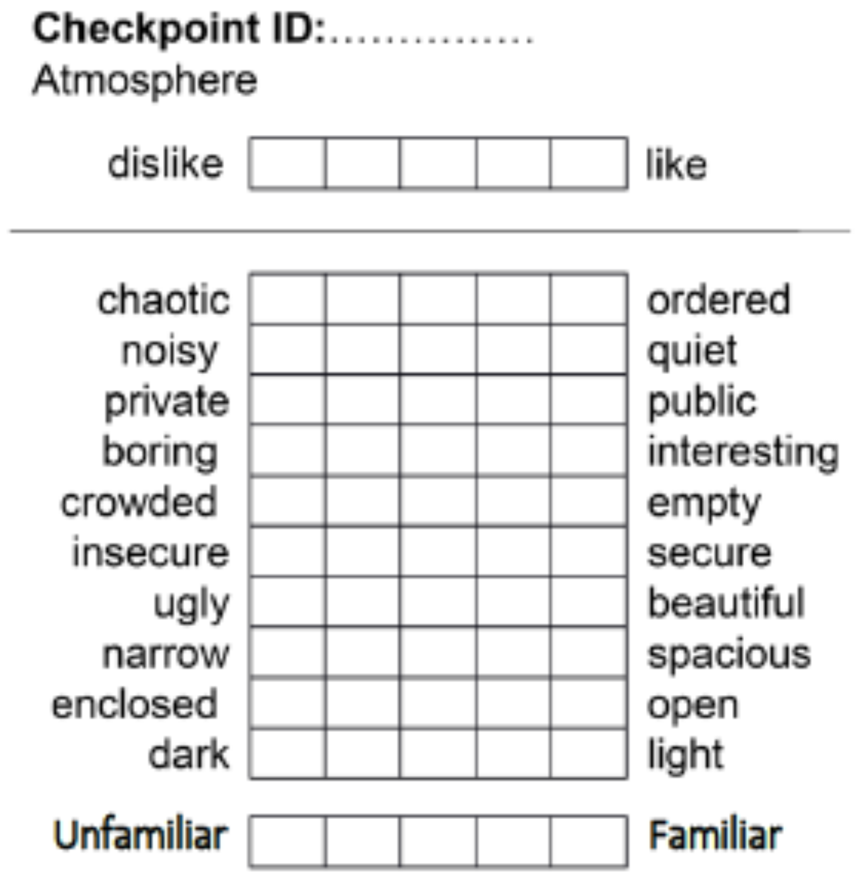


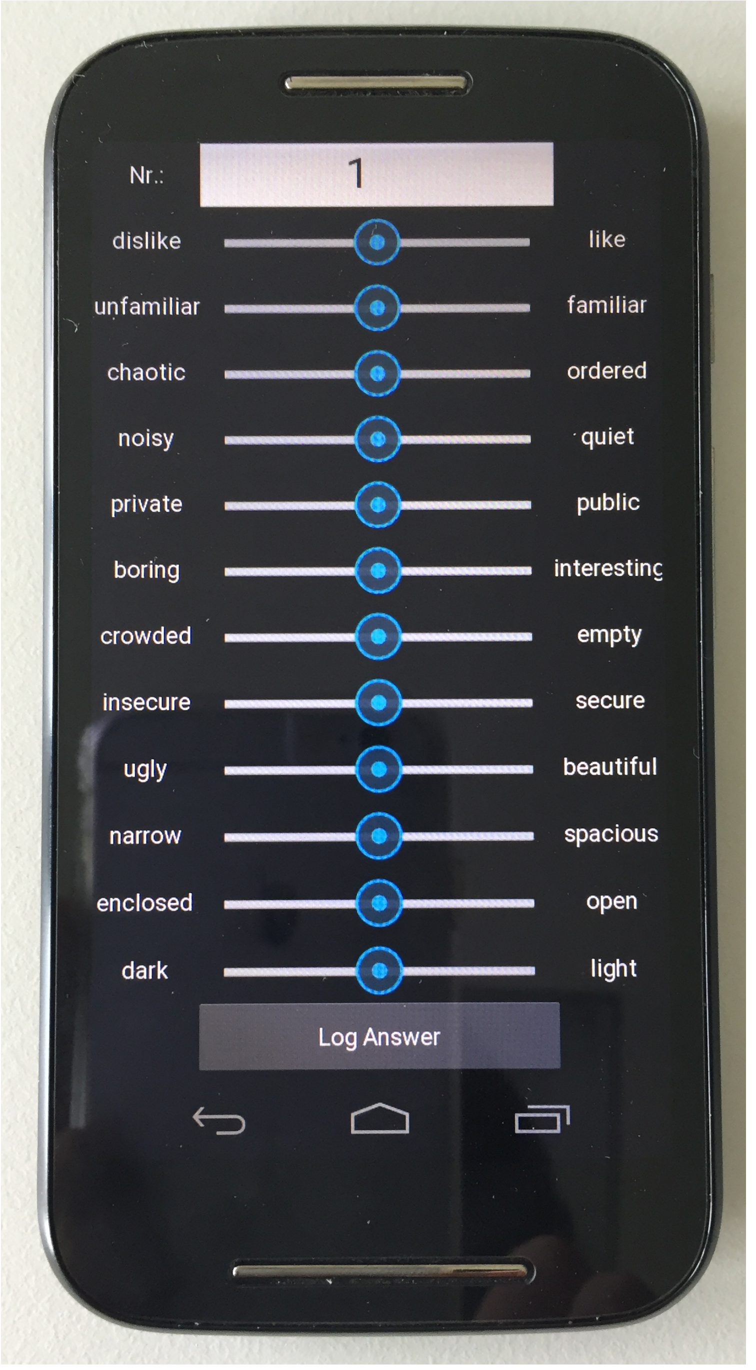


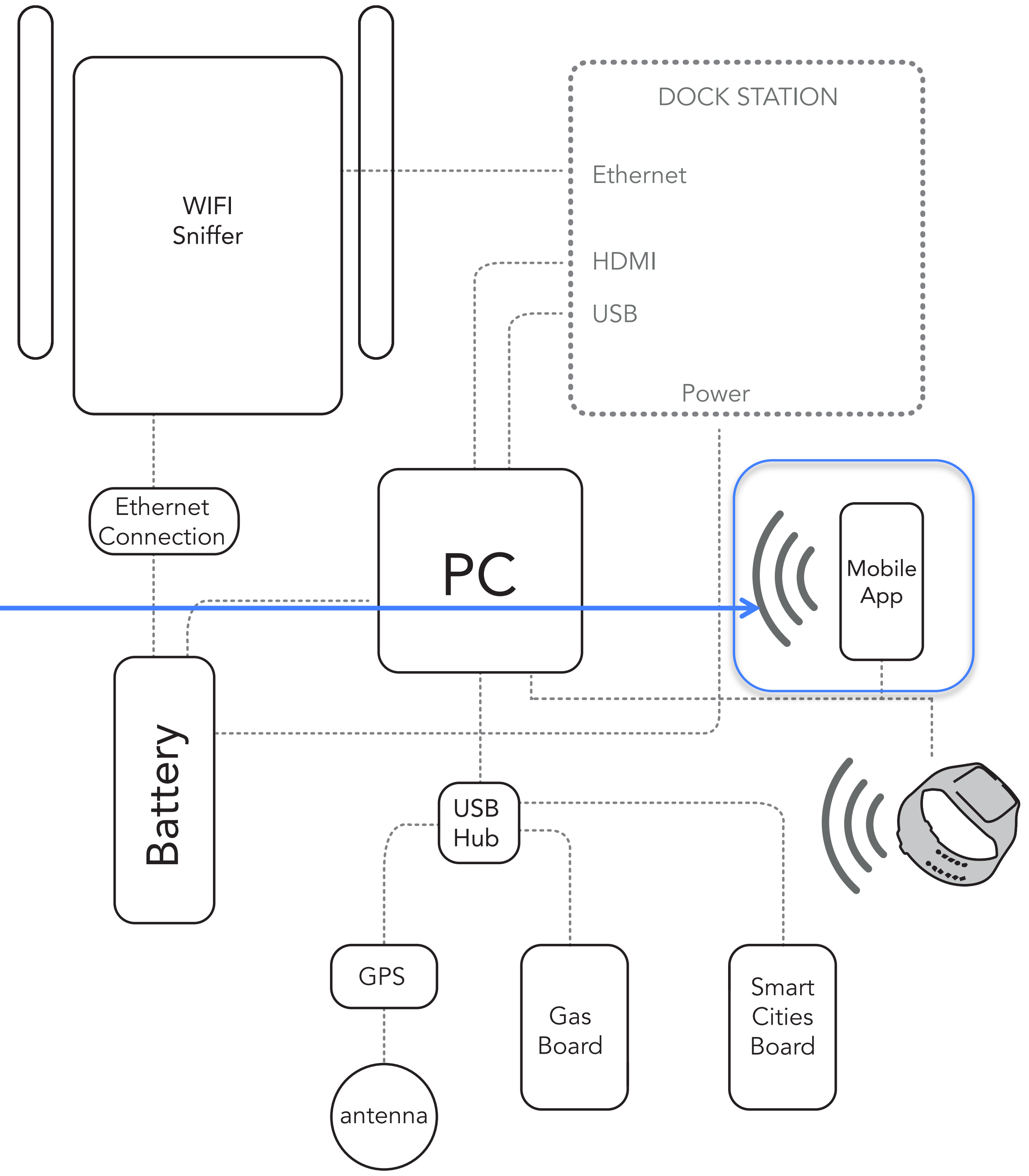
https://www.empatica.com/e4-wristband

Mobile Sensor equipment

Biofeedback wristband







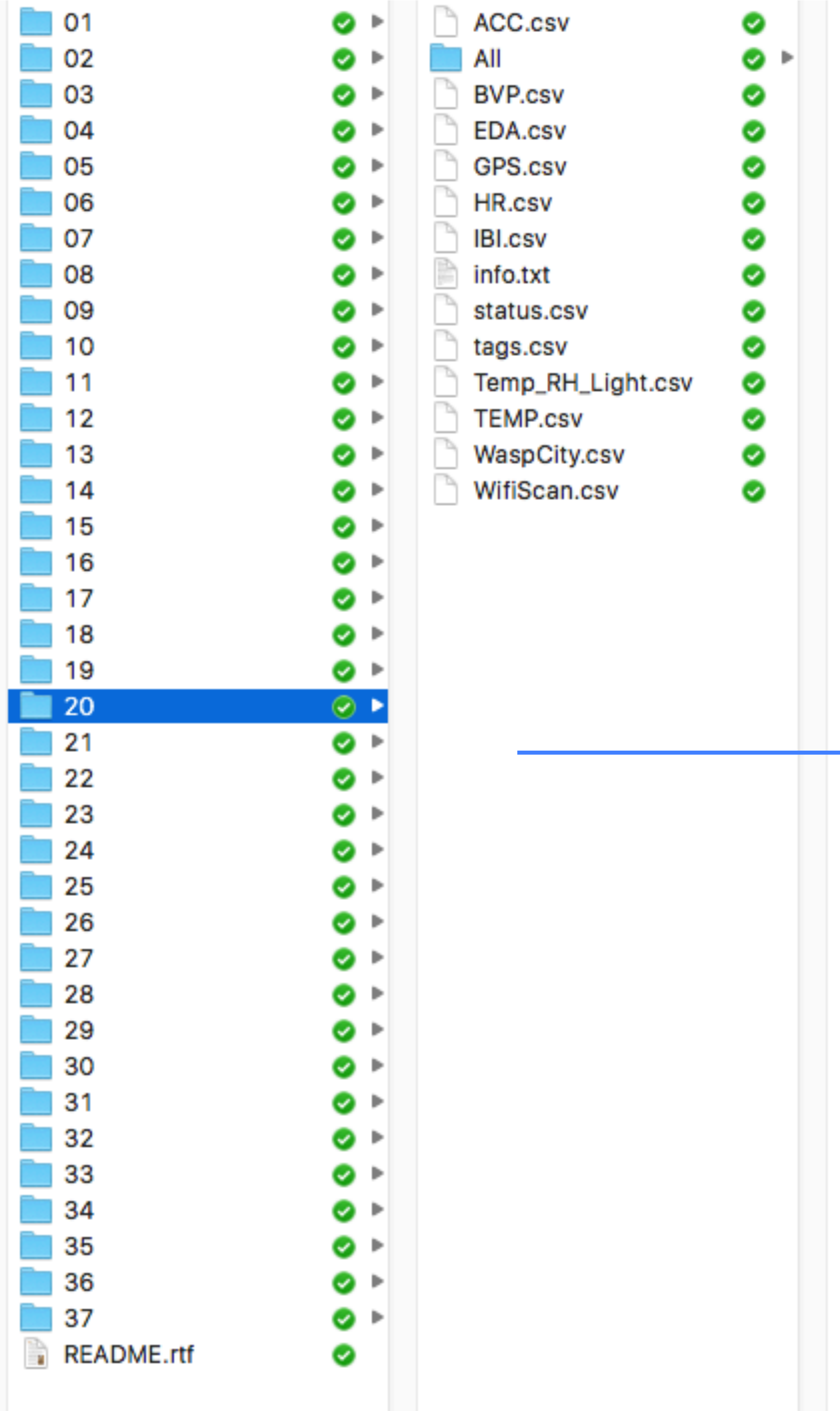
Experimental data-set

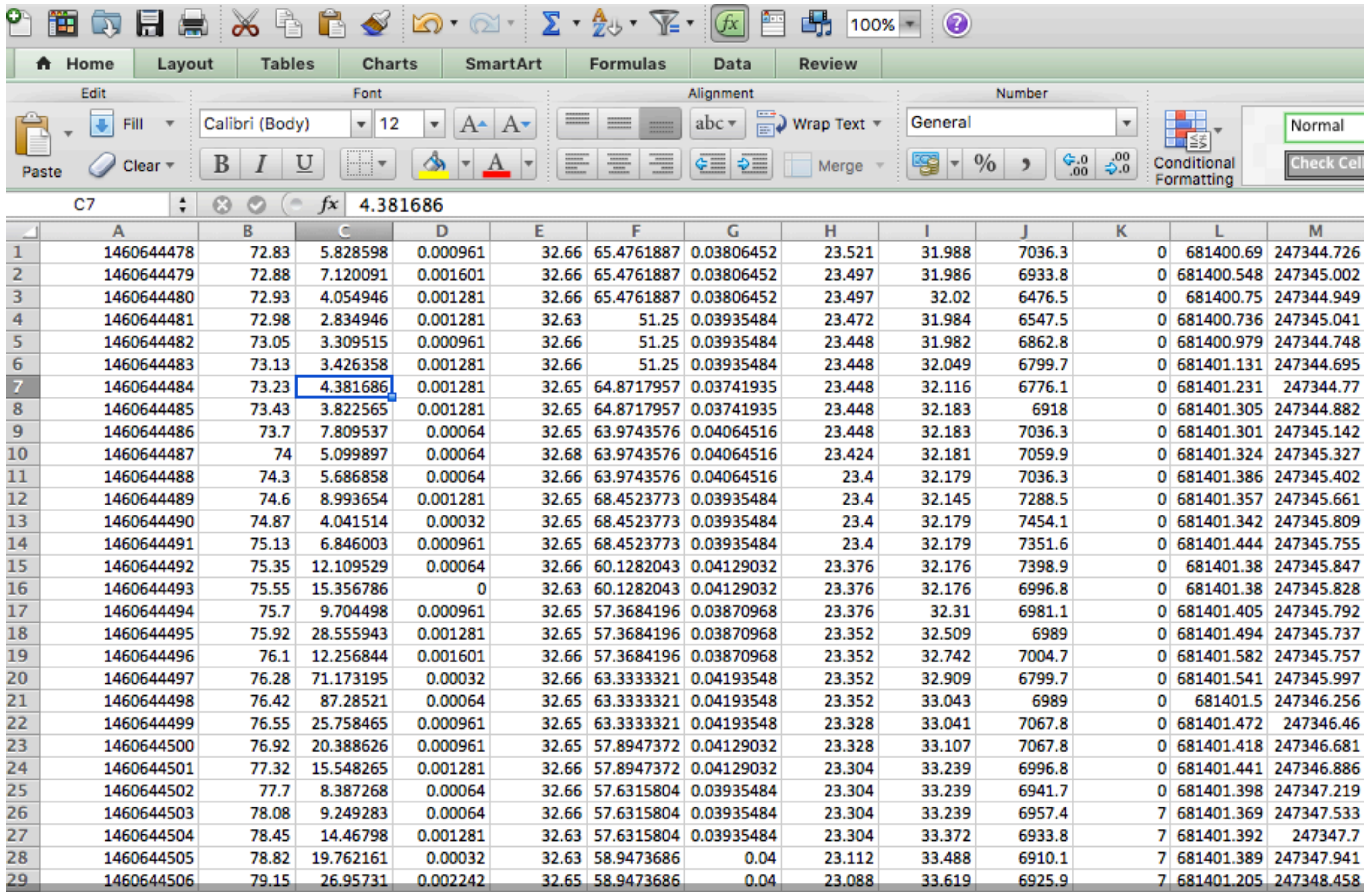
ESUM- Analyzing trade-offs between Energy and Social performance of Urban Morphologies

Device	*Sensor/ Measurement*	*units*	*Measurement range*	*Measurement frequency*	*Accuracy*	*Response time*
WaspCity	Sound Pressure	dB	50-100 dB	0.4 Hz	±2.5 dB	Not Given
	Luminosity	%	0-100% (400- 700 nm)	0.4 Hz	Resistive sensor 20MOhm (Darkness) 5-20 kOhm(Light)	Not Given
	Dust	mg/m3	Typical 0.5V/(0.1mg/m3)	0.4 Hz	Operating supply voltage 5±0.5V	10±1ms
WaspGas	Temperature	C	-40 ~ 125 C	0.25 Hz	±2 C(0-70 C), ±4 C(<0 C, >70C)	1.65 seconds
	Atmospheric Pressure	kPa	15 - 115 kPa	0.25 Hz	<±1.5% V	20 ms
	Humidity	%RH	0-100% RH	0.25 Hz	<±4% RH (a 25C, range 30-80%), ±6 %RH(range 0-100)	<15 seconds
Meshlium Scanner AP	Wifi Scanner	MAC address	Wifi Scanner (50-200m) Bluetooth Scanner (20-30m)	push values @ 0.016 Hz	Measurement range depends on he antenna and line of sight to the device	60 seconds
	Wifi Scanner	AP		push values @ 0.016 Hz		
	Wifi Scanner	RSSI (Received Signal Strenght Indicator)	-40 dBm (nearest node) to -90 dBm (marthes nodes)	push values @ 0.016 Hz	distance of 10m ~=(50dBm), 50m ~=(75dBm)	
Mobile Device	GPS	Lat/Long	outdoor only	variable, dependent on device satellite connection		
	Survey	12 questions, scale -2 to 2	NA	At checkpoint		
GPS	GPS	Lat/Long	outdoor only	1 Hz		
Biofeedback Wristband	PPG (Photoplethysmography)	Sensor output: Blood Volume Pulse (BPV)		64 Hz	0.9 nW/Digit	
	EDA (Electrodermal Activity)		0.01 mSiements -100 mSiemens	4 Hz		
	Skin Temperature Infrared thermopile	C	-40-115 C	4 Hz	±0.2 C within 36-39 C	
	3 Axis accelerometer	x, y, z		32 Hz		

Data Processing: ESUM Experiment

Data cleaning: unified date/time, convert WGS84 spherical coordinates to CH1903 planar coordinates





	A	B	C	D	E	F	G	H	I	J	K	L	M
1	1460644478	72.83	5.828598	0.000961	32.66	65.4761887	0.03806452	23.521	31.988	7036.3	0	681400.69	247344.726
2	1460644479	72.88	7.120091	0.001601	32.66	65.4761887	0.03806452	23.497	31.986	6933.8	0	681400.548	247345.002
3	1460644480	72.93	4.054946	0.001281	32.66	65.4761887	0.03806452	23.497	32.02	6476.5	0	681400.75	247344.949
4	1460644481	72.98	2.834946	0.001281	32.63	51.25	0.03935484	23.472	31.984	6547.5	0	681400.736	247345.041
5	1460644482	73.05	3.309515	0.000961	32.66	51.25	0.03935484	23.448	31.982	6862.8	0	681400.979	247344.748
6	1460644483	73.13	3.426358	0.001281	32.66	51.25	0.03935484	23.448	32.049	6799.7	0	681401.131	247344.695
7	1460644484	73.23	4.381686	0.001281	32.65	64.8717957	0.03741935	23.448	32.116	6776.1	0	681401.231	247344.77
8	1460644485	73.43	3.822565	0.001281	32.65	64.8717957	0.03741935	23.448	32.183	6918	0	681401.305	247344.882
9	1460644486	73.7	7.809537	0.00064	32.65	63.9743576	0.04064516	23.448	32.183	7036.3	0	681401.301	247345.142
10	1460644487	74	5.099897	0.00064	32.68	63.9743576	0.04064516	23.424	32.181	7059.9	0	681401.324	247345.327
11	1460644488	74.3	5.686858	0.00064	32.66	63.9743576	0.04064516	23.4	32.179	7036.3	0	681401.386	247345.402
12	1460644489	74.6	8.993654	0.001281	32.65	68.4523773	0.03935484	23.4	32.145	7288.5	0	681401.357	247345.661
13	1460644490	74.87	4.041514	0.00032	32.65	68.4523773	0.03935484	23.4	32.179	7454.1	0	681401.342	247345.809
14	1460644491	75.13	6.846003	0.000961	32.65	68.4523773	0.03935484	23.4	32.179	7351.6	0	681401.444	247345.755
15	1460644492	75.35	12.109529	0.00064	32.66	60.1282043	0.04129032	23.376	32.176	7398.9	0	681401.38	247345.847
16	1460644493	75.55	15.356786	0	32.63	60.1282043	0.04129032	23.376	32.176	6996.8	0	681401.38	247345.828
17	1460644494	75.7	9.704498	0.000961	32.65	57.3684196	0.03870968	23.376	32.31	6981.1	0	681401.405	247345.792
18	1460644495	75.92	28.555943	0.001281	32.65	57.3684196	0.03870968	23.352	32.509	6989	0	681401.494	247345.737
19	1460644496	76.1	12.256844	0.001601	32.66	57.3684196	0.03870968	23.352	32.742	7004.7	0	681401.582	247345.757
20	1460644497	76.28	71.173195	0.00032	32.66	63.3333321	0.04193548	23.352	32.909	6799.7	0	681401.541	247345.997
21	1460644498	76.42	87.28521	0.00064	32.65	63.3333321	0.04193548	23.352	33.043	6989	0	681401.5	247346.256
22	1460644499	76.55	25.758465	0.000961	32.65	63.3333321	0.04193548	23.328	33.041	7067.8	0	681401.472	247346.46
23	1460644500	76.92	20.388626	0.000961	32.65	57.8947372	0.04129032	23.328	33.107	7067.8	0	681401.418	247346.681
24	1460644501	77.32	15.548265	0.001281	32.66	57.8947372	0.04129032	23.304	33.239	6996.8	0	681401.441	247346.886
25	1460644502	77.7	8.387268	0.00064	32.66	57.6315804	0.03935484	23.304	33.239	6941.7	0	681401.398	247347.219
26	1460644503	78.08	9.249283	0.00064	32.66	57.6315804	0.03935484	23.304	33.239	6957.4	7	681401.369	247347.533
27	1460644504	78.45	14.46798	0.001281	32.63	57.6315804	0.03935484	23.304	33.372	6933.8	7	681401.392	247347.7
28	1460644505	78.82	19.762161	0.00032	32.63	58.9473686	0.04	23.112	33.488	6910.1	7	681401.389	247347.941
29	1460644506	79.15	26.95731	0.002242	32.65	58.9473686	0.04	23.088	33.619	6925.9	7	681401.205	247348.458

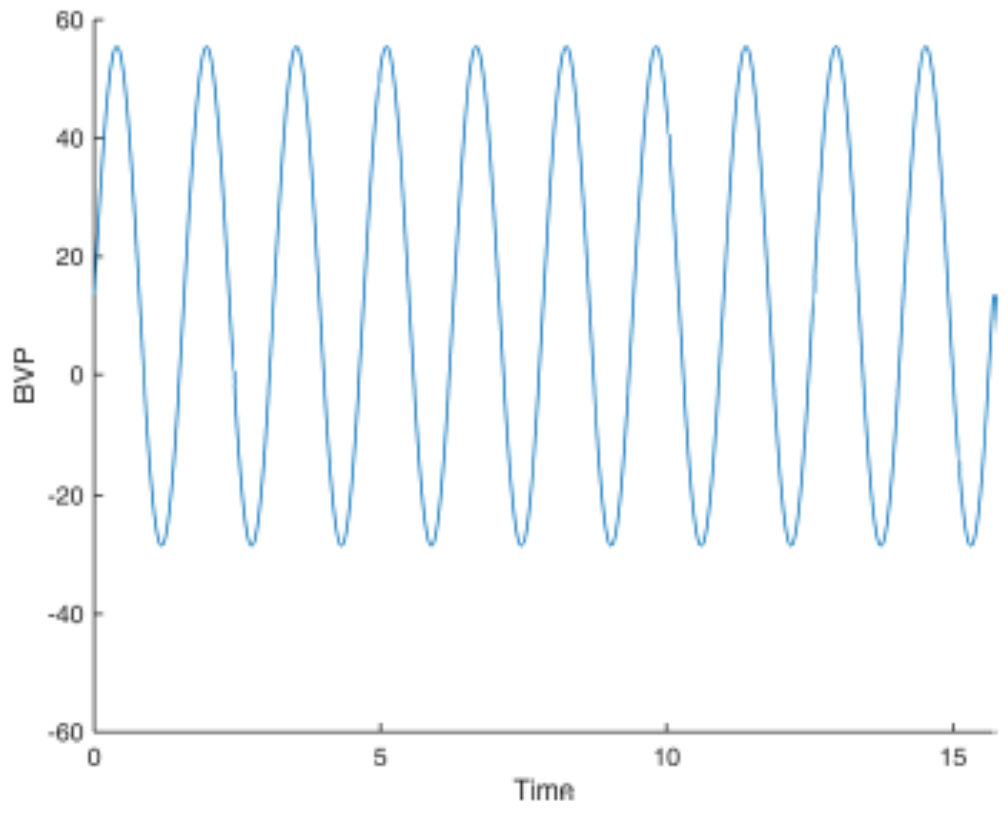
D. Griego, V. Buff, E. Hayos, I. Moise, E. Pournaras (2017), Sensing and mining urban qualities in smart cities, proceedings in AINA IEEE 31st Conference

Data Processing: ESUM Experiment

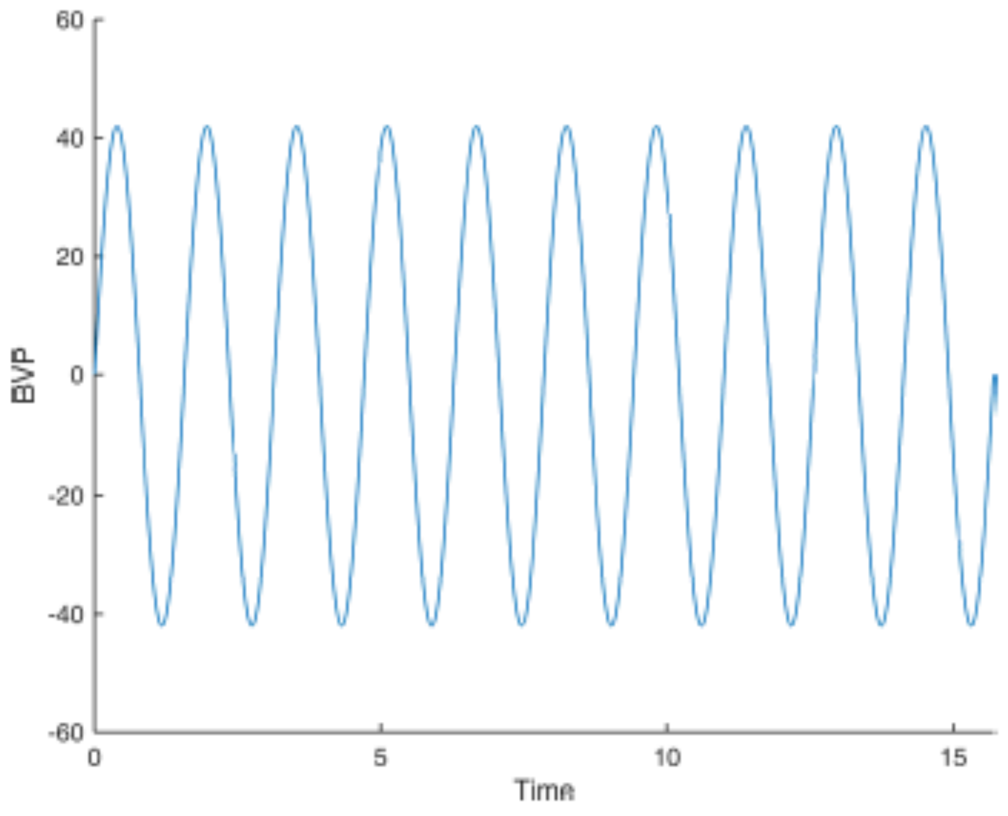
Data integration and reduction: frequency reduction and and integration from multiple data sources

TABLE III. DATA COLLECTION FREQUENCIES.

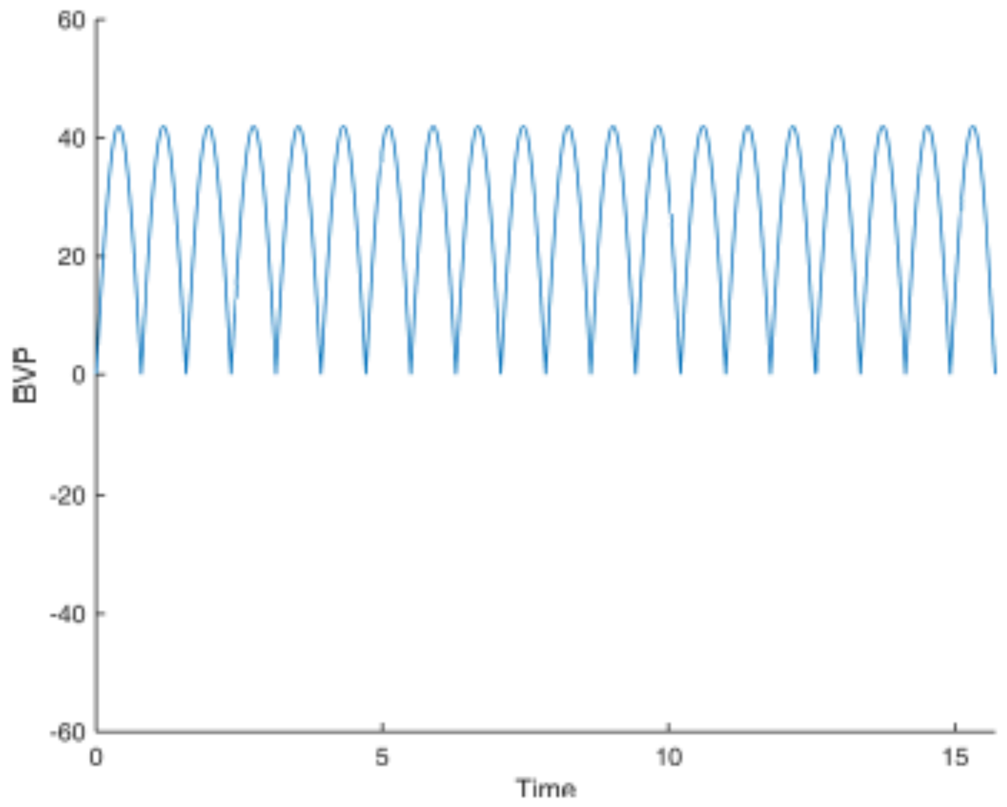
Sensor description	Frequency [Hz]
Heart rate (HR)	1
Blood volume pressure (BVP)	64
Electrodermal activity (EDA)	4
Biofeedback temperature (T-BF)	1
Sound level (S)	0.3
Dust (D)	0.3
Environment temperature (T-EN)	1
Relative humidity (RH)	1
Illuminance (IL)	1
People density (PD)	1 (if many), 0.024 (if few)
Longitude (LON)	1
Latitude (LAT)	1
Survey answers	In each of the 14 checkpoints



(a) Raw signal.



(b) Signal after adding average offset.



(c) Absolute positive signal.

Fig. 4. Frequency reduction applied to the bloom volume pressure measurements.

D. Griego, V. Buff, E. Hayos, I. Moise, E. Pournaras (2017), Sensing and mining urban qualities in smart cities, proceedings in AINA IEEE 31st Conference

Data processing: ESUM Experiment

Data integration and reduction: Geo-referencing data to specific locations

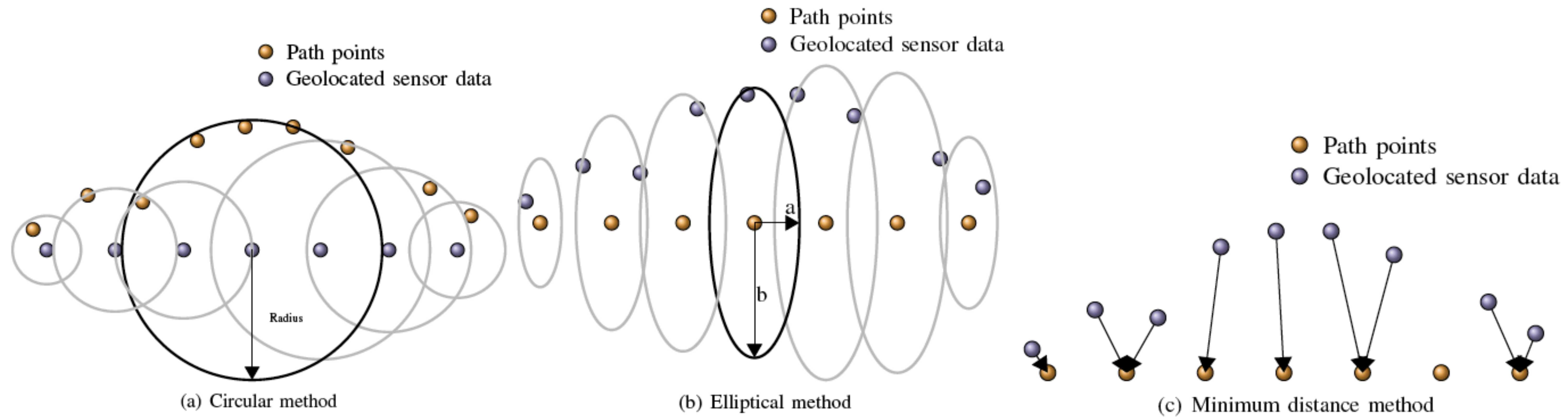
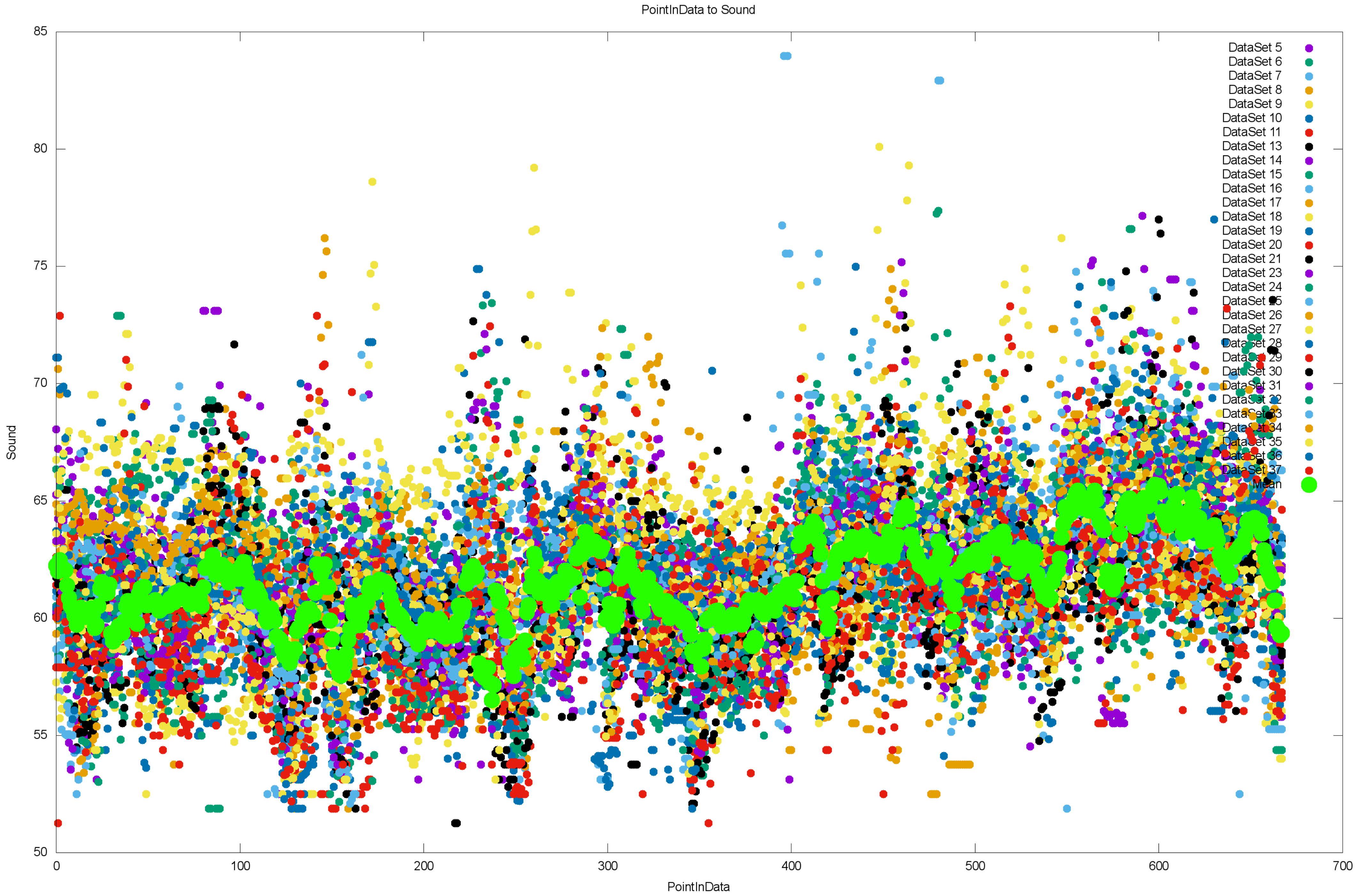


Fig. 5. Three data localization methods.

D. Griego, V. Buff, E. Hayos, I. Moise, E. Pournaras (2017), Sensing and mining urban qualities in smart cities, proceedings in AINA IEEE 31st Conference

Data processing: ESUM Experiment

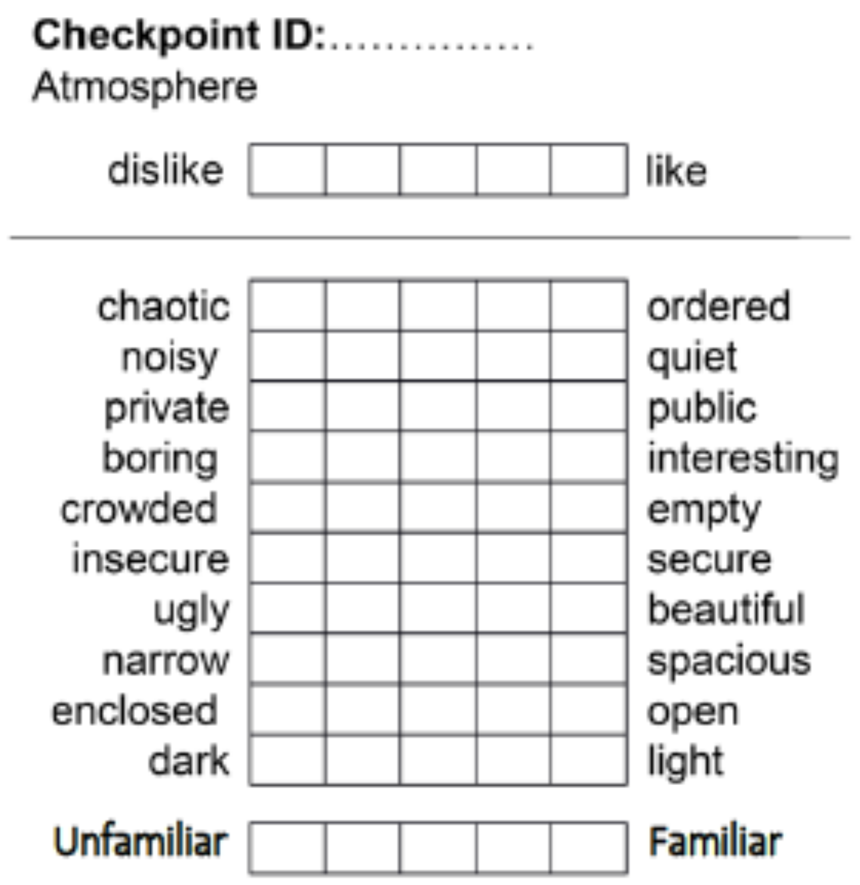
Data integration and reduction: Geo-referencing data to specific locations



D. Griego, V. Buff, E. Hayos, I. Moise, E. Pournaras (2017), Sensing and mining urban qualities in smart cities, proceedings in AINA IEEE 31st Conference

Data processing: ESUM Experiment

Data reduction: feature extraction



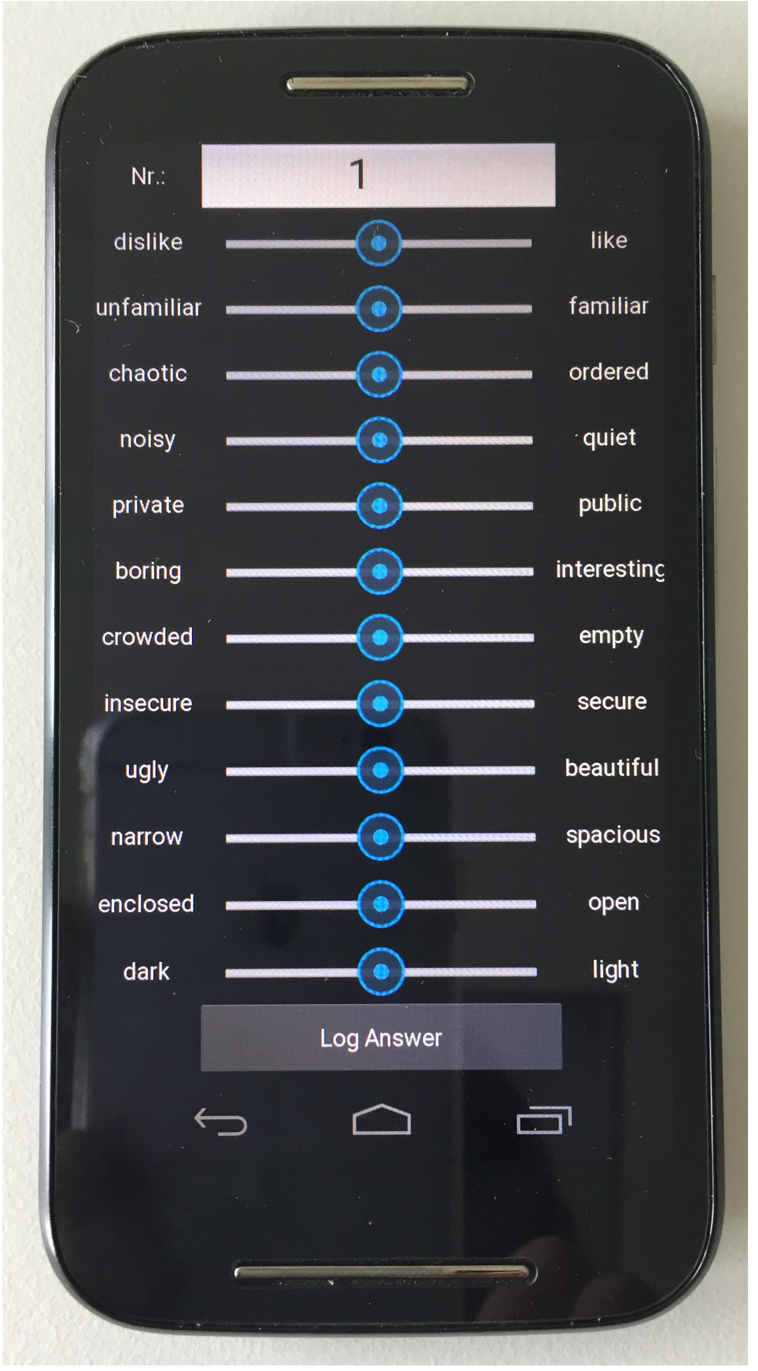


TABLE I. QUESTIONS AND RANGES OF RATINGS FROM WHICH CITIZENS MAKE SELECTIONS. THE VIRTUAL SENSOR USING EACH QUESTIONS IS INDICATED.

Question #	1	2	3	4	5	6
Question Range	Beautiful Ugly	Empty - Crowded	Familiar - Unfamiliar	Interesting Boring	Light - Dark	Like - Dislike
Used in virtual sensor:	APP	-	APP	ACT	ACT	APP
Question #	7	8	9	10	11	12
Question Range	Open - Enclosed	Ordered Chaotic	Public Private	Quiet Noisy	Secure - Insecure	Spacious Narrow
Used in virtual sensor:	ACT	APP	APP	EXP	EXP	EXP

Course Schedule

Start thinking about your final projects

Creative Data Mining
Uncover and Evaluate

The participants of this course learn how to collect, process, analyze and interpret real spatial and temporal data in order to work with quantifiable qualities in urban planning. This is achieved by using actual data from a recent study and analysing it with different data processing and machine learning techniques.
The goal of the course is to explore a specific research question about the urban environment and test the stated hypothesis using different techniques presented in the course, thus preparing students with a skill-set to further support their design and decision making processes.
The course focuses on creating deeper insights to critically evaluate design alternatives for urban planning projects. Students will work with time-series and geo-referenced data including temperature, relative humidity, illuminance, noise, people density, and dust particulate matter. Subjective impression survey data will also be integrated into the student projects to further explore influencing factors of the urban environment on our perceptual experiences. Non-architectural skills the participants can develop during this course are 1) an introduction to programming 2) how clustering methods like PCA or K-Means could be applied in an architectural context.

Where
HIT H 34.1 (Video Wall)

Supervision

Danielle Griego	griego@arch.ethz.ch
Daniel Zünd	zuend@arch.ethz.ch
Artem Chirkin	chirkin@arch.ethz.ch

Prof. Dr. Gerhard Schmitt
Chair of Information Architecture
Information Science Lab
Wolfgang-Pauli-Strasse 27, 8093 Zurich
www.ia.arch.ethz.ch

20.02.2017	**Course Introduction**
27.02.2017	**Introduction to Python & Programming I**
06.03.2017	**Introduction to Python & Programming II**
13.03.2017	**Data Processing**
20.03.2017	**Seminar week (No lecture)**
27.03.2017	**Intro to time-series data analysis**
03.04.2017	**Time series data analysis ctd. & Machine learning**
10.04.2017	**Machine learning ctd.**
17.04.2017	**Holiday (No lecture)**
24.04.2017	**Programming tutorial applications**
01.05.2017	**Holiday (No lecture)**
08.05.2017	**Q&A Feedback Workshop I**
15.05.2017	**Final iA critique** Combined critique with the other iA courses (13:00 - 18:00)

Requirement Former knowledge of any digital tool or coding language is most welcome but NOT required. You only need to provide a reasonable amount of motivation and of course a notebook.

* Total 60 h = 2 ECTS
Ungraded Semester Performance

The most recent outline will be found on www.ia.arch.ethz.ch

iA Chair for Information Architecture

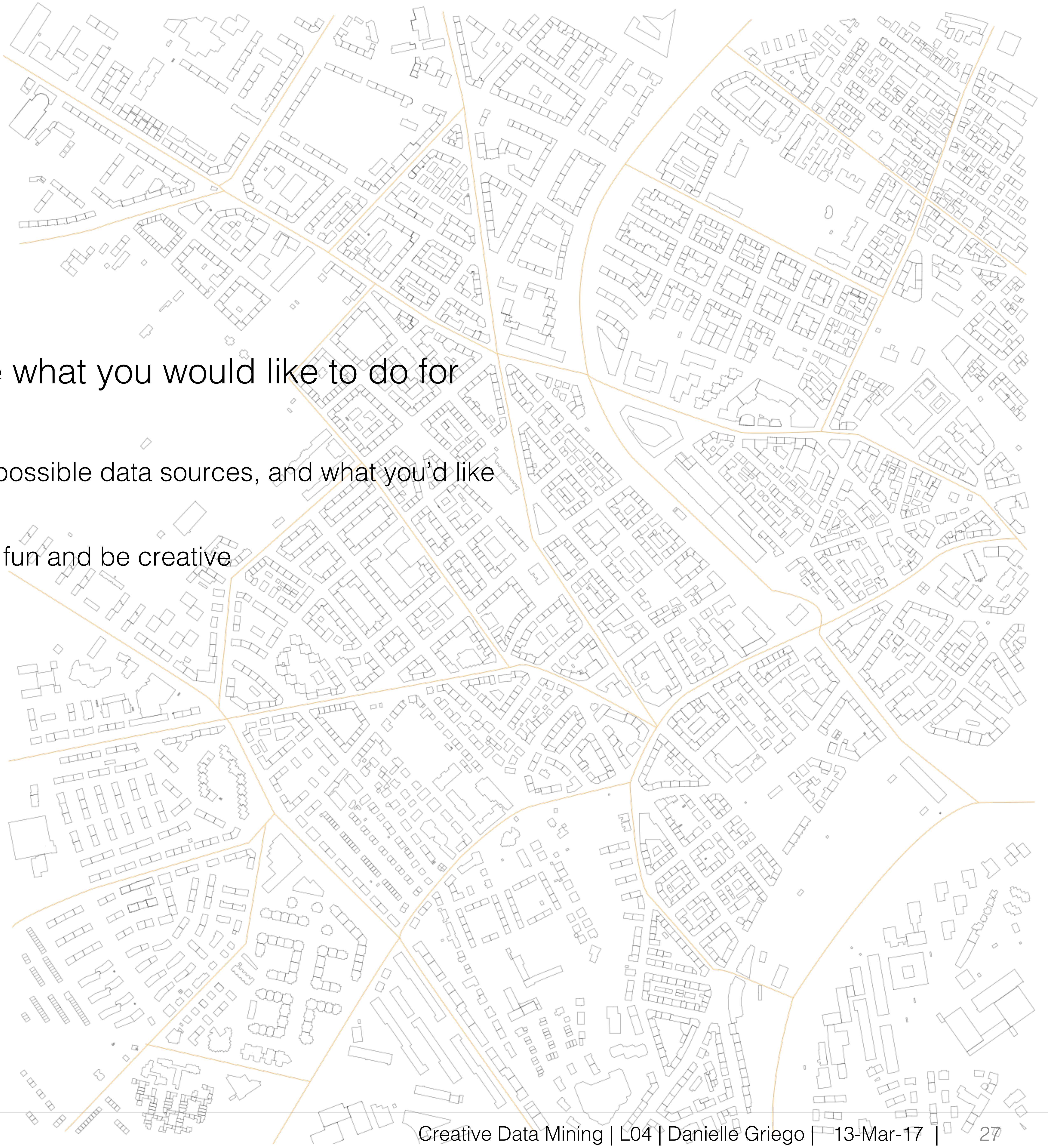
ETH zürich DARCH iA Chair of Information Architecture

Semester Project

1. Formulate 1-2 specific question(s) of interest to you
2. State your hypothesis/expected outcome based on supporting literature (minimum one source) your expertise, and intuition
3. Answer that question through your analysis, for this:
    - Select the best available data sources for your question (min. of 2 data sources)
    - Include a time series and/or clustering analysis
4. Summarize your results
    - Show a clear conclusion, does your analysis answer your question(s)?
5. Conclusions & lessons learned
6. Include motivation and references

2nd Conceptual Homework

Project Ideas



For next class:

1. Prepare 1-3 slides to describe what you would like to do for your final project
    - This includes a specific hypothesis, possible data sources, and what you'd like to know (research question)
    - Come up with questions for us, have fun and be creative

2nd Programming assignment

Data Processing: data transformation

Transform survey response data from each ESUM participant by survey question:

→(12 questions, 14 check points) per participant to a single matrix for each question (32 participants, 14 check points)

→Headers from participant data files: ["Date-Time", "path-point", "like-dislike", "familiar-unfamiliar", "ordered-chaotic", "quiet-noisy", "public-private", "interesting-boring", "empty-crowded", "secure-insecure", "beautiful-ugly", "spacious-narrow", "open-enclosed", "light-dark"]

ETH zürich DARCH iA Chair of Information Architecture

Questions?

Lecture 4: Data Processing

13| 03 | 2017

Dr. Daniel Zünd, zuend@arch.ethz.ch
Danielle Griego, griego@arch.ethz.ch
Artem Chirkin, chirkin@arch.ethz.ch

ETH zürich DARCH iA Chair of Information Architecture

One example from last year

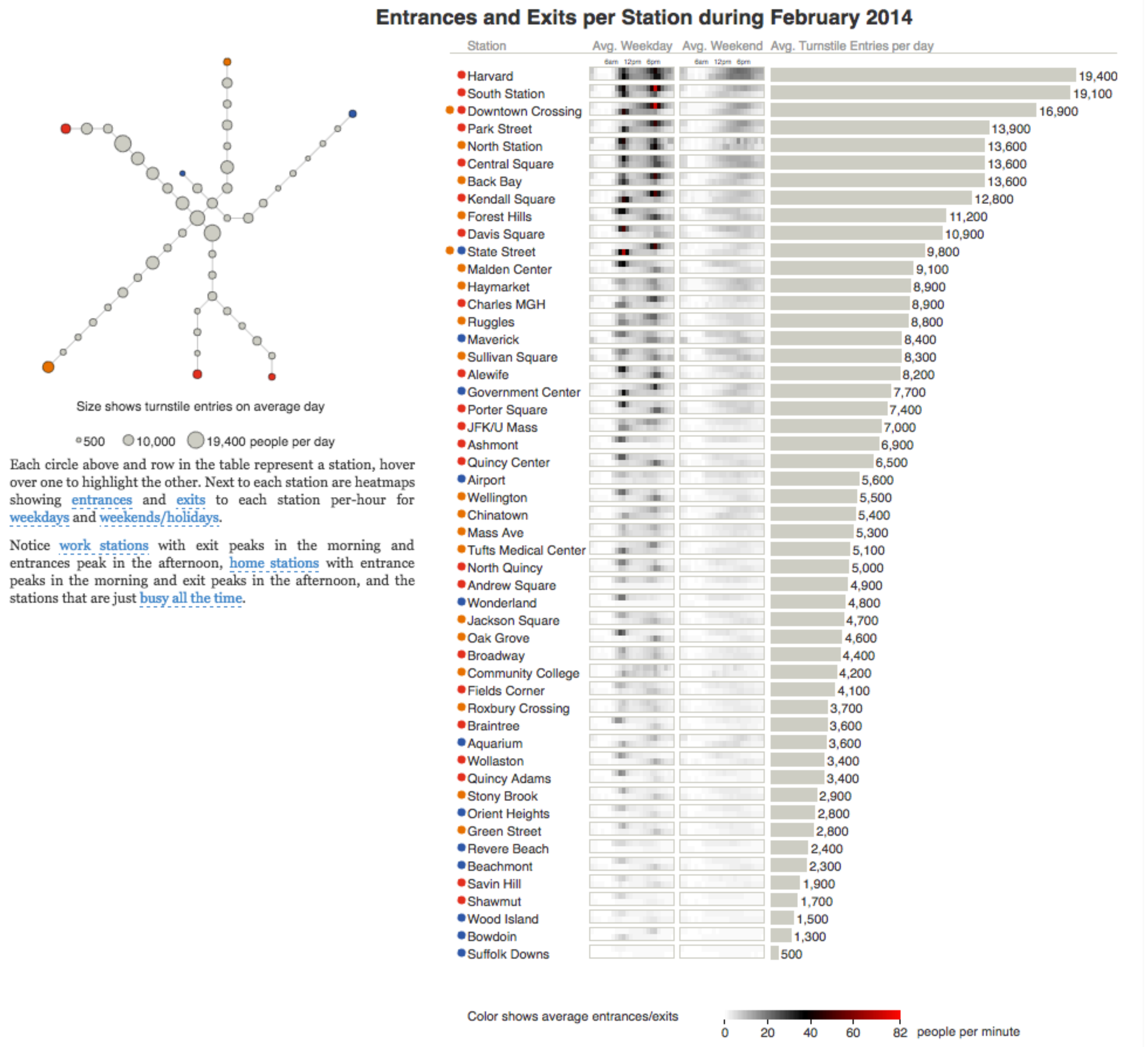
DARCH

iA Chair of Information Architecture

Final project from Jochen Aarts and Stéphane de Weck Creative Data Mining FS2016

Another example

Interactive data visualization from Boston metro



http://mbtaviz.github.io/