A Functional Model of Sensor Data

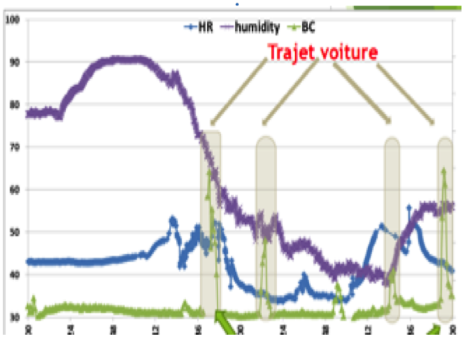
CONCEPTS AND FRAMEWORK

Karine Zeitouni
Joint work with Ahmad Mustapha and Yehia Taher

DAVID Lab.
University of Versailles St-Quentin , Paris-Saclay University

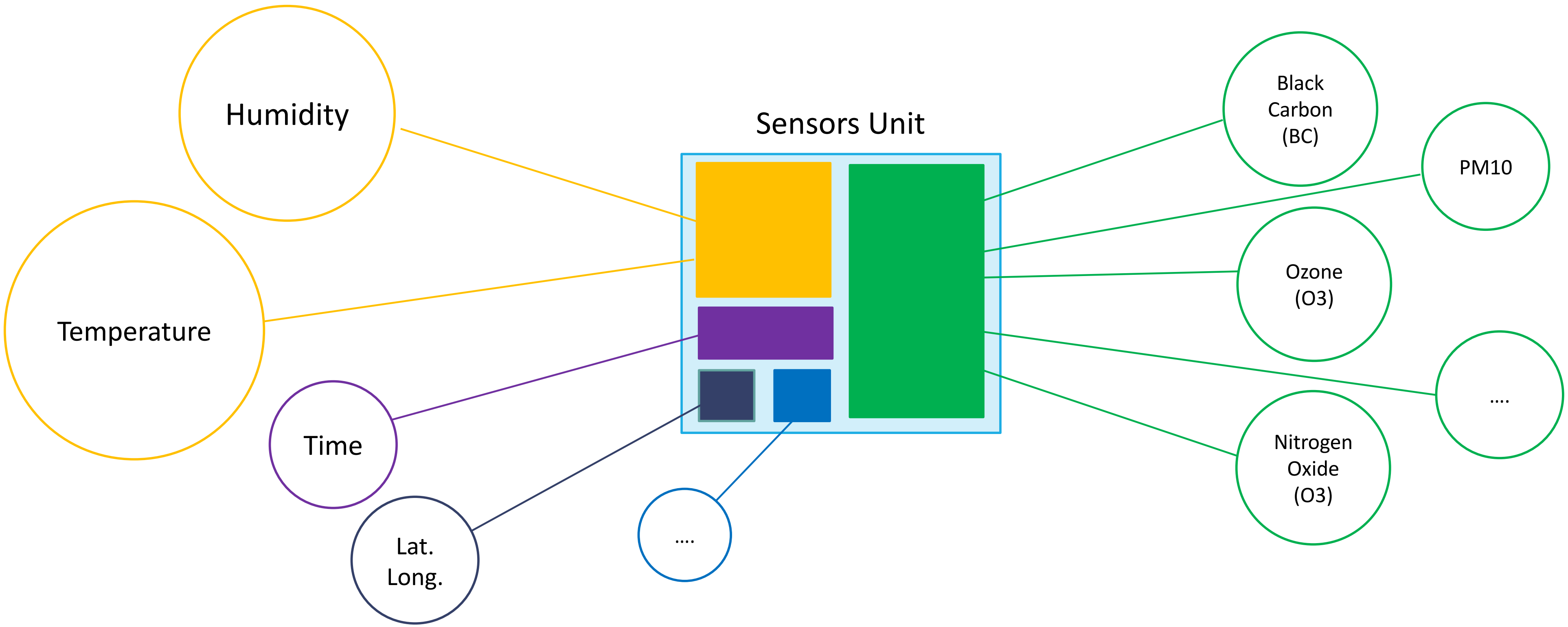
Context & Motivation

POLLUSCOPE



- **Opportunistic Mobile Monitoring** is a new paradigm of data collection during our daily activity and mobility by mean of mobile / wearable sensors.
- Polluscope ANR project leverages this paradigm for measuring individual exposure to air pollution and its health effects.

Sensor Unit



Humidity
Temperature
Time
Lat.
Long.
....
Sensors Unit
Black
Carbon
(BC)
PM10
Ozone
(O3)
....
Nitrogen
Oxide
(O3)

Problem Statement – The Ideal Data Acquisition

Time	Lat.	Long.	NO2 (ppb)	Pressure Atmo. (mV)	Temp. °C	Humidity %	…
02/06/2017 09:17:00	48.8397911	2.0804432	10	4300	34	29	…
02/06/2017 09:18:00	48.8397911	2.0803518	11	4260	34	28	…
02/06/2017 09:19:00	48.8398092	2.0803518	15	4240	34	26	…
02/06/2017 09:20:00	48.8398092	2.0804948	17	4240	34	26	…
02/06/2017 09:21:00	48.8398195	2.0804948	19	4240	34	26	…
02/06/2017 09:22:00	48.8398195	2.0804948	22	4240	34	27	…
02/06/2017 09:23:00	48.8398195	2.0804948	26	4240	34	27	…
02/06/2017 09:24:00	48.8398195	2.0804074	22	4240	34	27	…
02/06/2017 09:25:00	48.8398045	2.0804074	23	4240	34	27	…
02/06/2017 09:26:00	48.8398045	2.0804384	24	4240	33	27	…
02/06/2017 09:27:00	48.8398024	2.0804384	26	4240	33	28	…
…	…	…	…	…	…	…	…

Problem Statement – The Actual Data

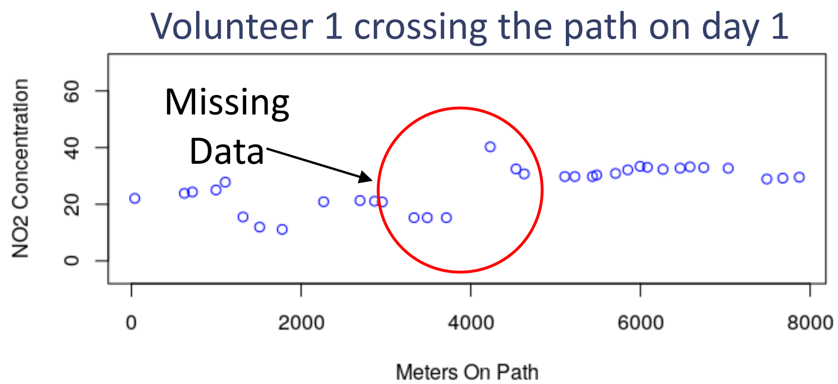
Irregular time intervals inter and intra source, mlissing values, noisy data ☹

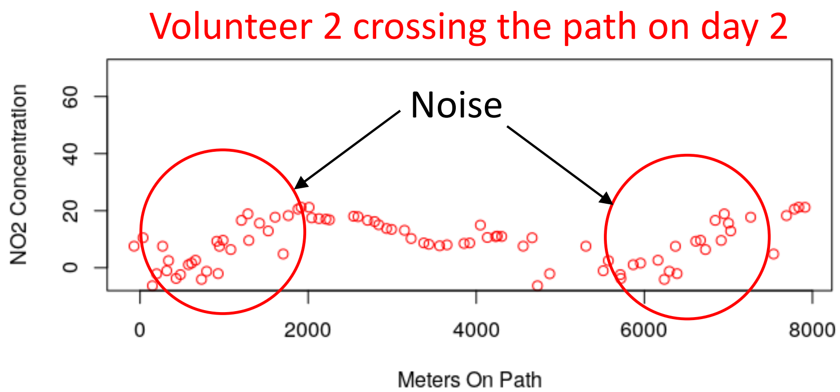
Time	Lat.	Long.	NO2 (ppb)
02/06/2017 09:17:30	48.8397911	2.0804432	3.097
02/06/2017 09:18:30	48.8397911	2.0803518	13.477
02/06/2017 09:20:10	48.8398092	2.0803518	23.103
02/06/2017 09:20:50	48.8398092	2.0804948	23.964
02/06/2017 09:21:32	NA	NA	27.1
02/06/2017 09:23:40	NA	NA	21.681
02/06/2017 09:25:11	NA	NA	24.707
02/06/2017 09:26:46	48.8398195	2.0804074	16.321
02/06/2017 09:27:12	48.8398045	2.0804074	33.231
02/06/2017 09:29:42	NA	NA	32.046
02/06/2017 09:27:00	NA	NA	18.138
...	...	...	...

Time	Pressure Atmo. (mV)
02/06/2017 09:10:00	4300
02/06/2017 09:40:00	4260
02/06/2017 10:10:00	4240
02/06/2017 10:40:00	4240
02/06/2017 11:10:00	4240
02/06/2017 11:40:00	4240
02/06/2017 12:10:00	4240
02/06/2017 12:40:00	4240
02/06/2017 01:10:00	4240
02/06/2017 01:40:00	4240
02/06/2017 02:10:00	4240
...	...

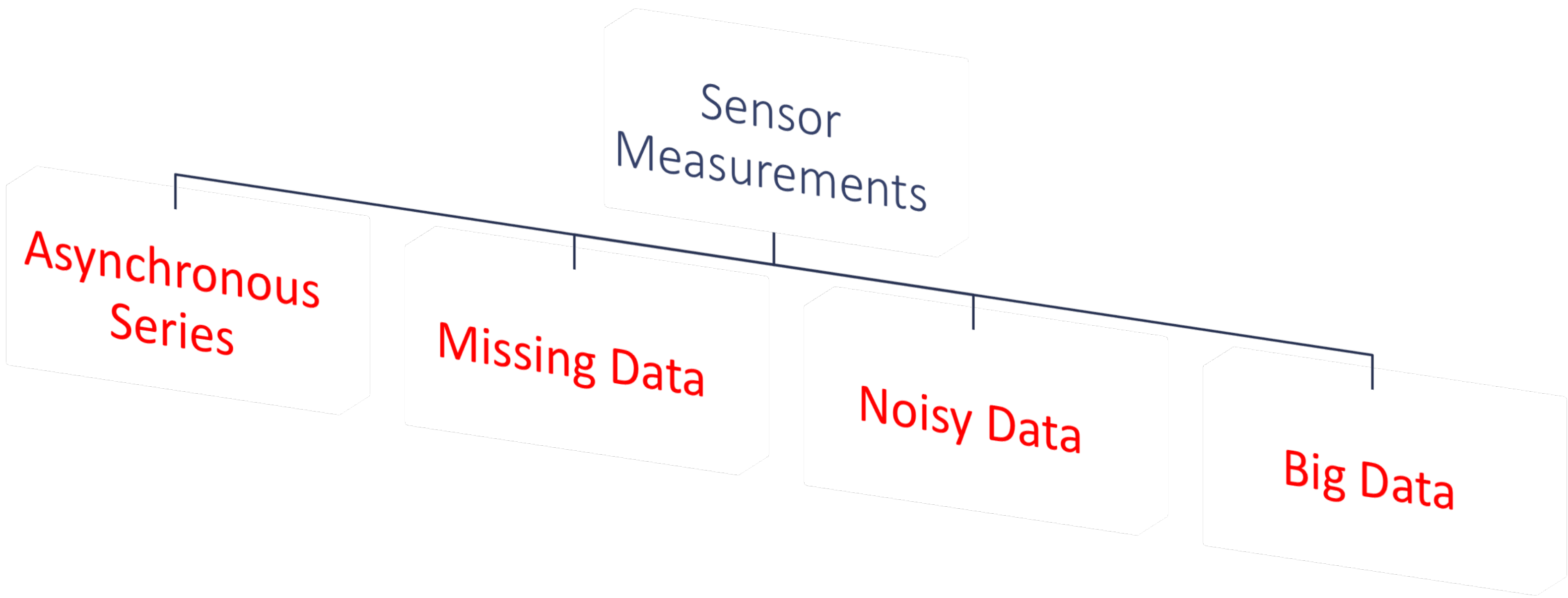
...

Illustration in Mobile Crowd Sensing Sensor Data are Imperfect Snapshots





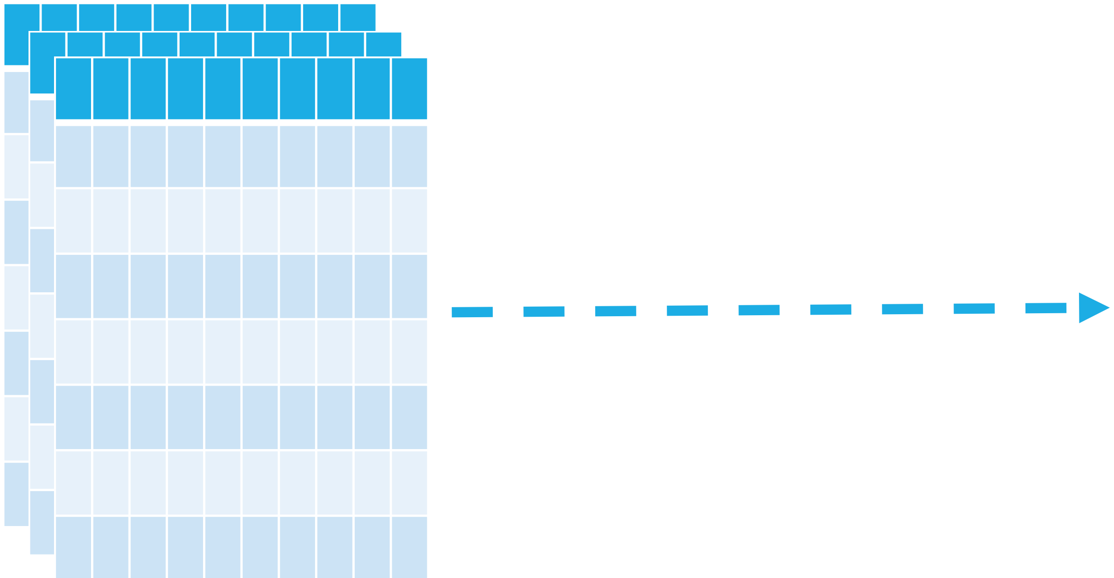
Problem Statement –
How to Deal with the Sensor Data Problem ?



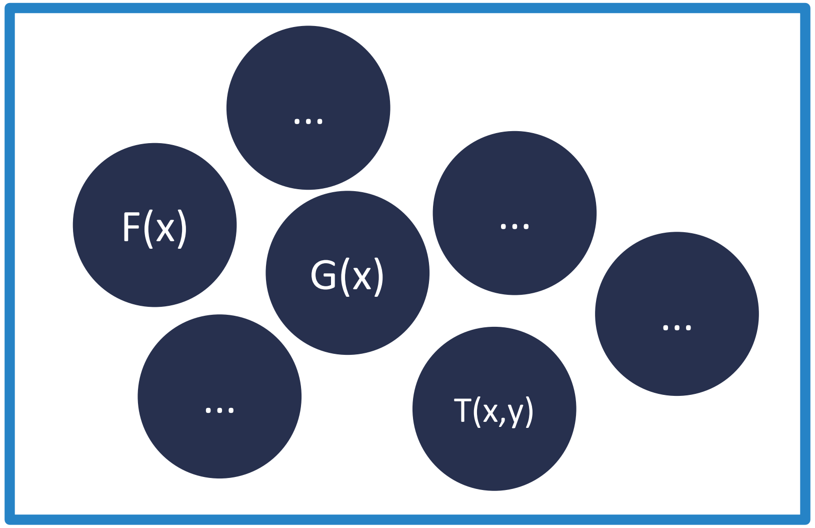
Outline

The Data Model

The Intuition –
Continuous Views as a Solution

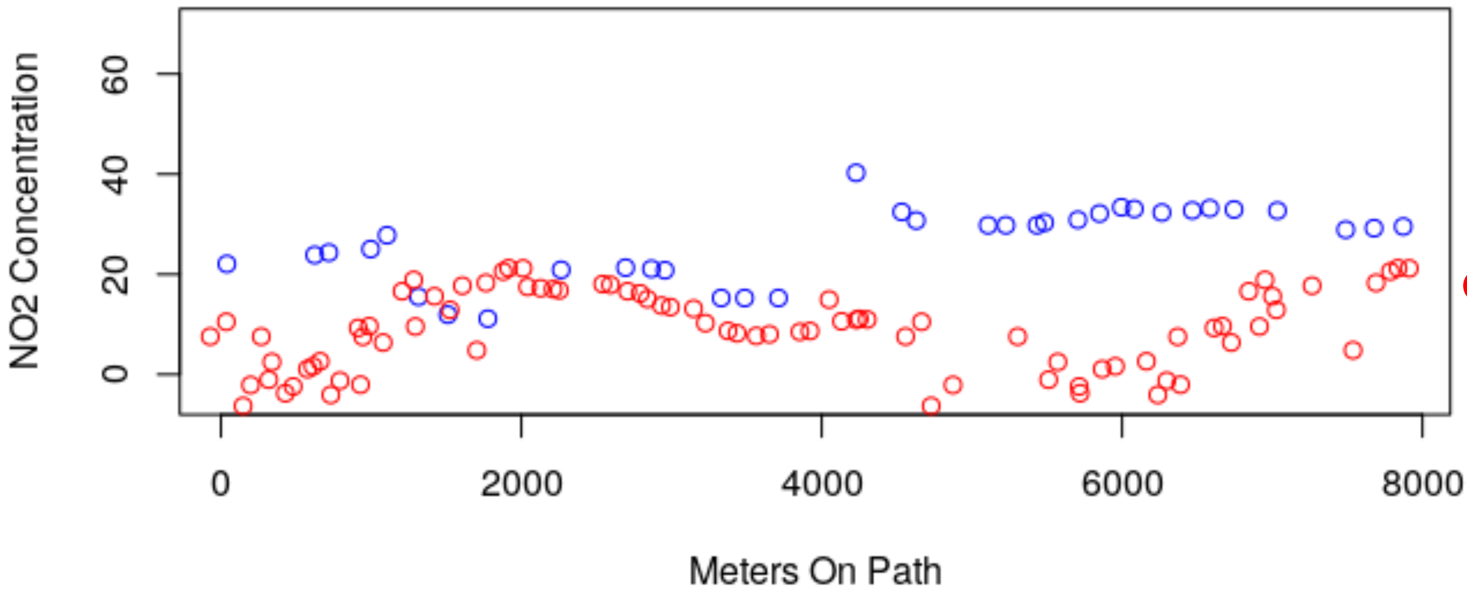


Raw Multi-dimensional
Spatio-temporal Sensors Data



Sensor Data Representation
as Continuous Functions of Space and/or Time

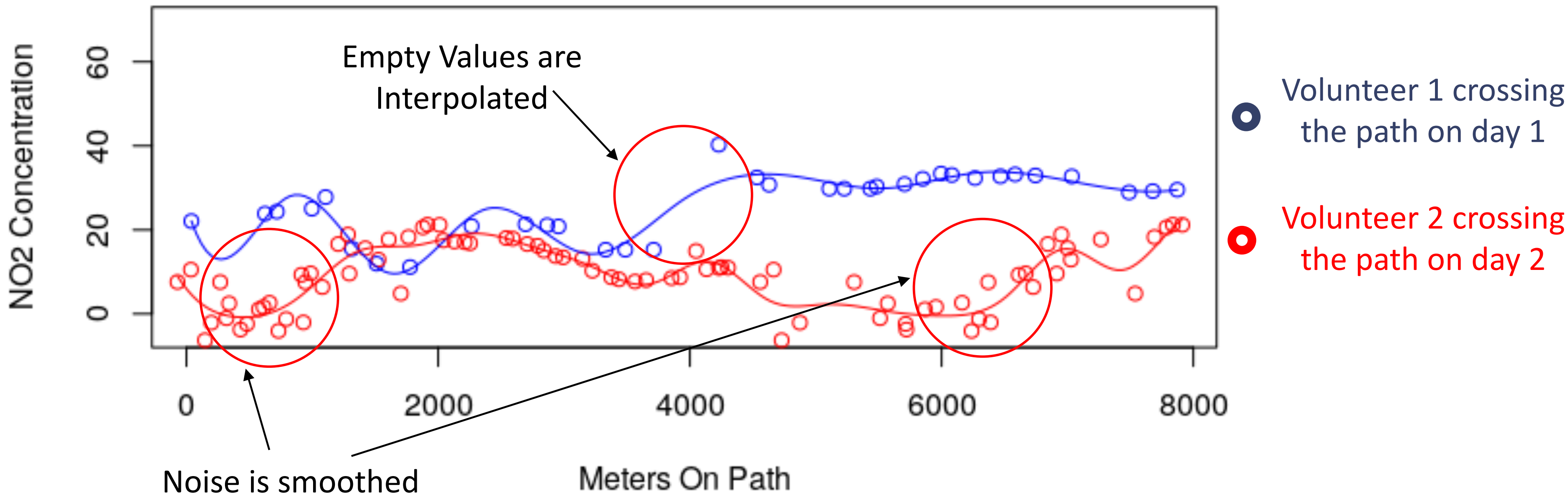
Continuous Views as a solution – Original Data



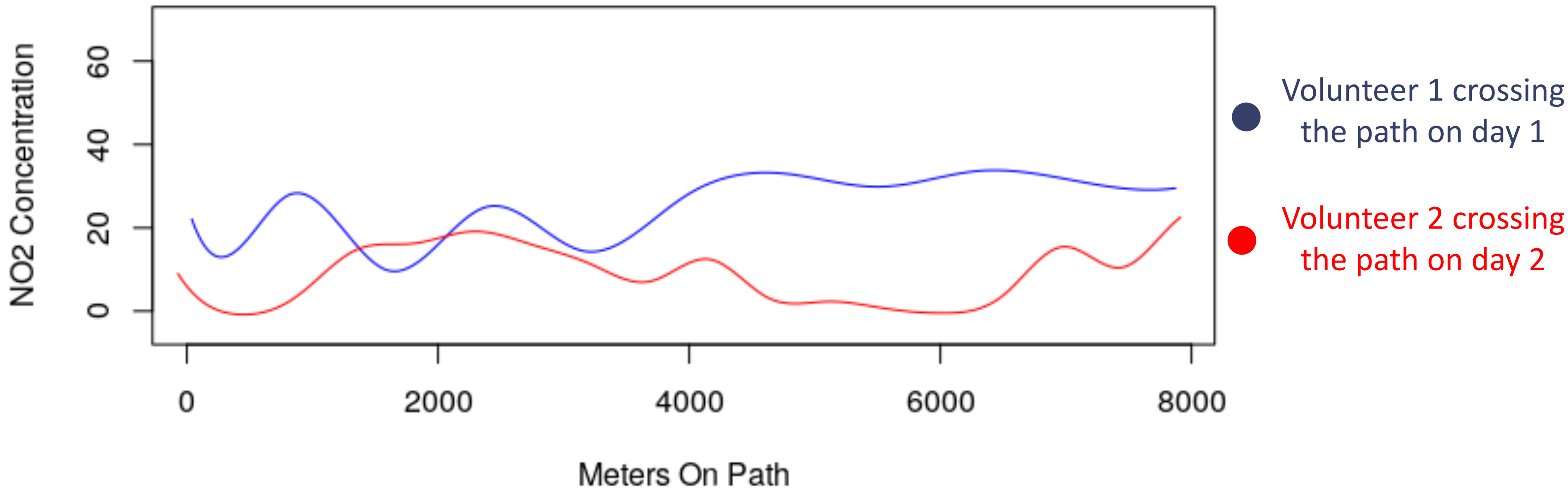
Volunteer 1 crossing the path on day 1

Volunteer 2 crossing the path on day 2

Continuous Views as a solution – Model Fitting



Continuous Views as a solution – Final Data Representation



The Framework

The Framework

The Framework

- Data Approximation
- Data Storage
- SQL-Like Querying
- Data Analysis

Data Approximation

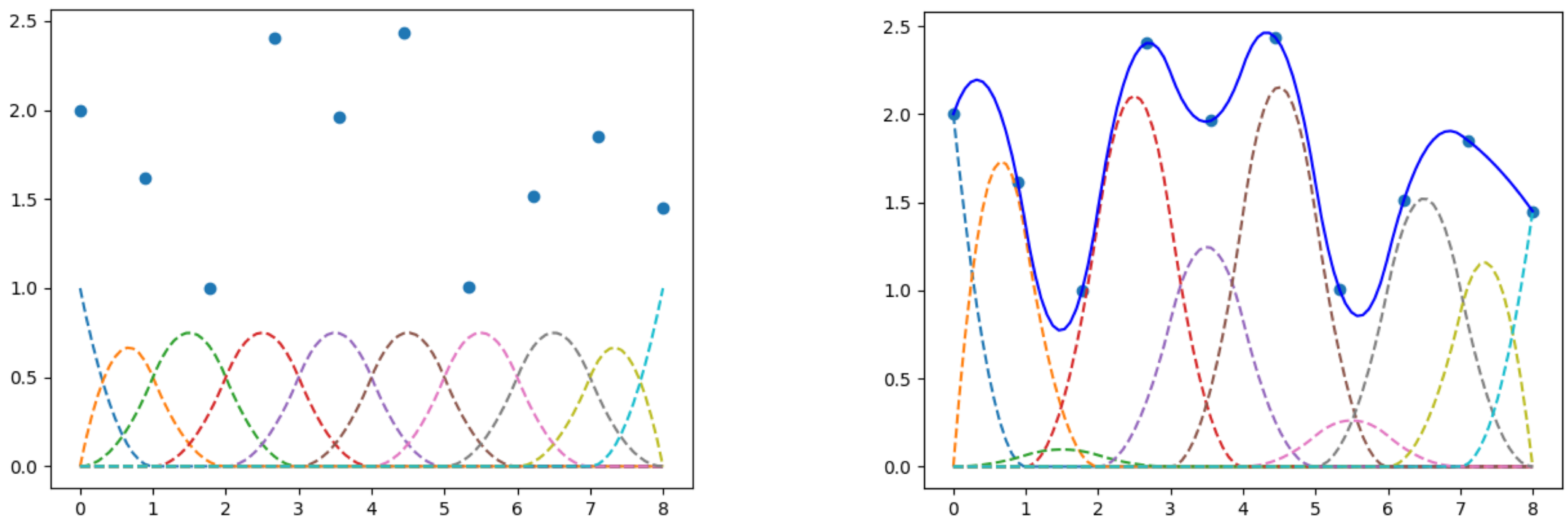
- We adopted "Basis Function Expansion" technique
- Given a list of observations (**x**, **y**) = $\begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ x_4 & y_4 \\ . & . \\ x_n & y_n \end{pmatrix}$
- We aim to have F(x) such that: $y = F(x) - e$
- F(x) will be represented by a linear aggregation of basis function B_i (x).

$$F(x) = \sum_{1}^{m} c_i B_i(x) = c_1 B_1(x) + c_2 B_2(x) + ... + c_n B_m(x)$$

- so that $D = (y - F(x))^2$ is minimized.
- Solving $\frac{dD}{dc} = 0$ will do the job.

Data Approximation

- The basis functions are a set of functions with certain characteristics
- E.g. Splines, Polynomials, Fourier, …



The Framework

The Framework

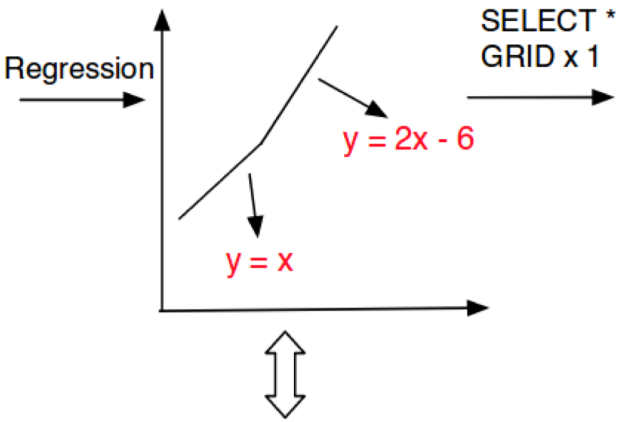
- Data Approximation
- Data Storage
- SQL-Like Querying
- Data Analysis

Data Storage

Raw Data

x	y
1	1.1
2	1.9
3	3.1
4	4
5	4.8
6	6.1
7	7.9
8	9.8
9	11.9
11	16.1
12	18
14	22.2

Regression Curve



Function Table

Start x	End x	Expr.
1	6	x
6	14	2x-6

Materialized Grid

x	y
1	1.0
2	2.0
3	3.0
4	4.0
5	5.0
6	6.0
7	8.0
8	10.0
9	12.0
10	14.0
11	16.0
12	18.0
13	20.0
14	22.0

- FunctionDB is a validated continuous functions storage and querying database (Thiagarajan & Madden, SIGMOD'08)
- For more genericity, we will represent functions by symbolic expressions rather than coefficients.
  - e.g. 2x^2 + 1.3x^1 + 4x
  - FunctionDB is resticted to linear regression

The Framework

The Framework

Data Approximation

Data Storage

SQL-Like Querying

Data Analysis

Data Query

- SQL - Like queries that makes use of function views
- The user can create, query, and aggregate continuous functions

```
CREATE VIEW NO2Time
AS FIT NO2 OVER Time
USING BASIS Fourier(...)
USING ALGO LSSE(...)
USING PARTITION SplitEqually(...)
TRAINING DATA SELECT * FROM Somedata
```

Data Query

- The Query Language supports Functions Selection and Aggregation

```
SELECT NO2 FROM NO2Pos
WHERE Pos > 300m and Pos< 4000m
GRID Time 1m   /* Discretized output */
```

Position	NO2
300	...
301	...
...	...
4000	....

```
SELECT SUM(NO2) FROM NO2Pos
WHERE NO2 < 10  /* Functional output */
```

Exprs.
expr1
...
exprn

- Note that the queries are executed symbolically (algebraically) whenever possible

The Framework

The Framework

- Data Approximation
- Data Storage
- SQL-Like Querying
- Data Analysis

Data Analysis

- Contemporary Machine Learning and data analysis techniques focuses on vectorized data
- Our framework supports this type of analysis as functions can be discretized using the GRID statement we see previously
- However, we thought about complementing discrete data analysis with analysis techniques that focuses on continuous data or functions.
- We will integrate **Functional Data Analysis** (FDA) in our framework

Data Analysis - FDA

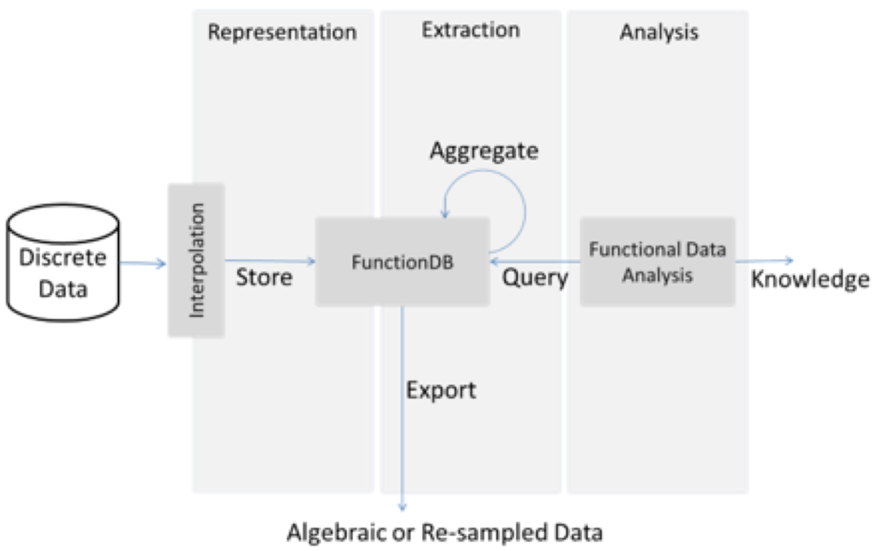
- Functional Data Analysis (FDA) is a statistical field that analysis function data, i.e. data are assumed to be smooth. In FDA functions are the atomic data structure and the analysis target.

Basic Operations	Advanced
• Mean • Correlation • Variance • Derivation • …	• Functional PCA • Clustering • Regression • Classification • …

The Proposal Intended Benefits

- Semi-automation of data preprocessing.
- Compression
- Provide a familiar query language.
- Raw data abstraction.

Conclusion



- We presented a blueprint to ease the acquisition, storage, processing, and analysis of sensors data.
- The main idea is to approximate discrete data with continuous functions.
- An implementation that uses Apache Spark is being under development
- Can/How this applies to TransiXplore?