Environmental Monitoring
&
The “Big Data” Revolution

QSENSE

Nick Masson, Qsense Inc.
“Making Sense of Sensors”, Sept 27-28, 2017

Premise:

- New AQ sensor technology is becoming accessible and affordable
- Per-unit monitor cost is several orders less expensive FEM/FRM analyzers

   → potential for (1) prolific adoption (2) low/no maintenance (3) high-frequency data

- Solving real problems requires aggregating/centralizing data across the ecosystem
- These new data can be applied in multiple arenas

   → policy/zoning, regulation, community health, personal health, etc…

Challenge:

- Even modest adoption means *huge volumes* of data
- Few monitors are currently commercially available
- Each manufacturer uses it’s own protocols/standards, storage methods, etc…
- Quality is sometimes questionable and variable among manufacturers
- Each manufacturer addresses quality control/correction differently
- Monitors and sensors are in continuous evolution

‘Even modest adoption means *huge volumes* of data’

STATUS QUO

Example:

- SCAQMD has 43 permanent monitoring sites (ref: Table 1, “Annual Air Quality Monitoring Network Plan”, July 1, 2017.)
- Suppose system surfaces hour-averaged data
- Suppose average of 4-variables per site

$43_{sites} * 4_{pt/site} * 730_{hr/mo}$

= 125.5K data points per month

FUTURE

Example:

- Given a monitor:
  (1) each 10 sq-km (130 total in LA)
  (2) to 1-in-50 asthmatics (~6500 total in LA)
- Suppose system surfaces minute-averaged data
- Average of 4-variables per monitor

$6630_{monitors} * 4_{pt/monitor} * 43800_{min/mo}$

= 1,2B data points per month

~10,000x increase

Example

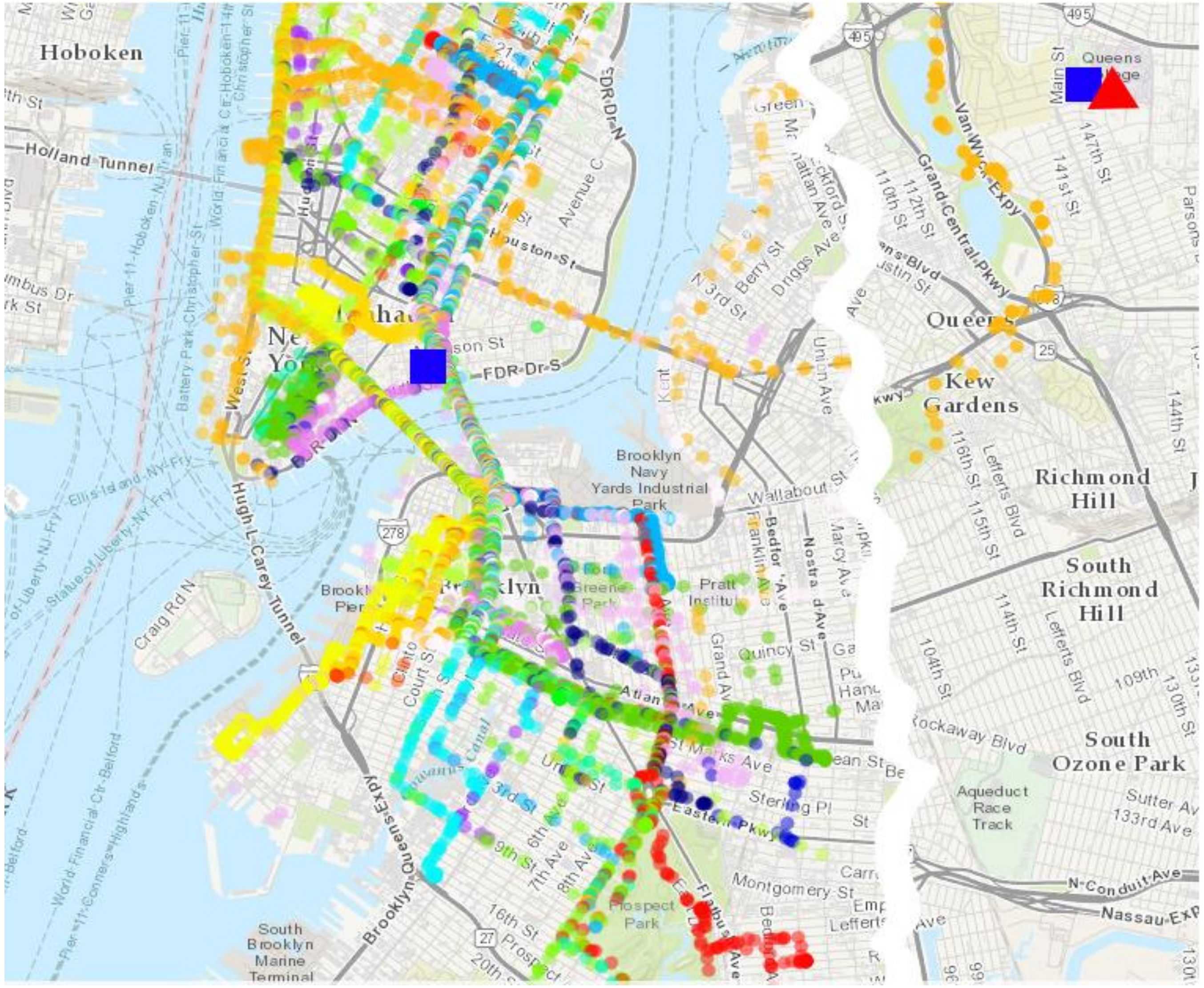
Heterogeneous sensor deployment (multiple types) & ‘Cloud Calibration’ in New York City

Cloud QA/QC calibration process

- Automated flagging and rejection of bad data
- Automated update of calibrations on a per-device basis

Summary

- Twelve (12) month dataset
- RMSE of raw data ~ 9ug/m^3
- RMSE of cloud calibrated data ~4 ug/m^3
- ~12% of data omitted in QA/QC flagging



LEGEND

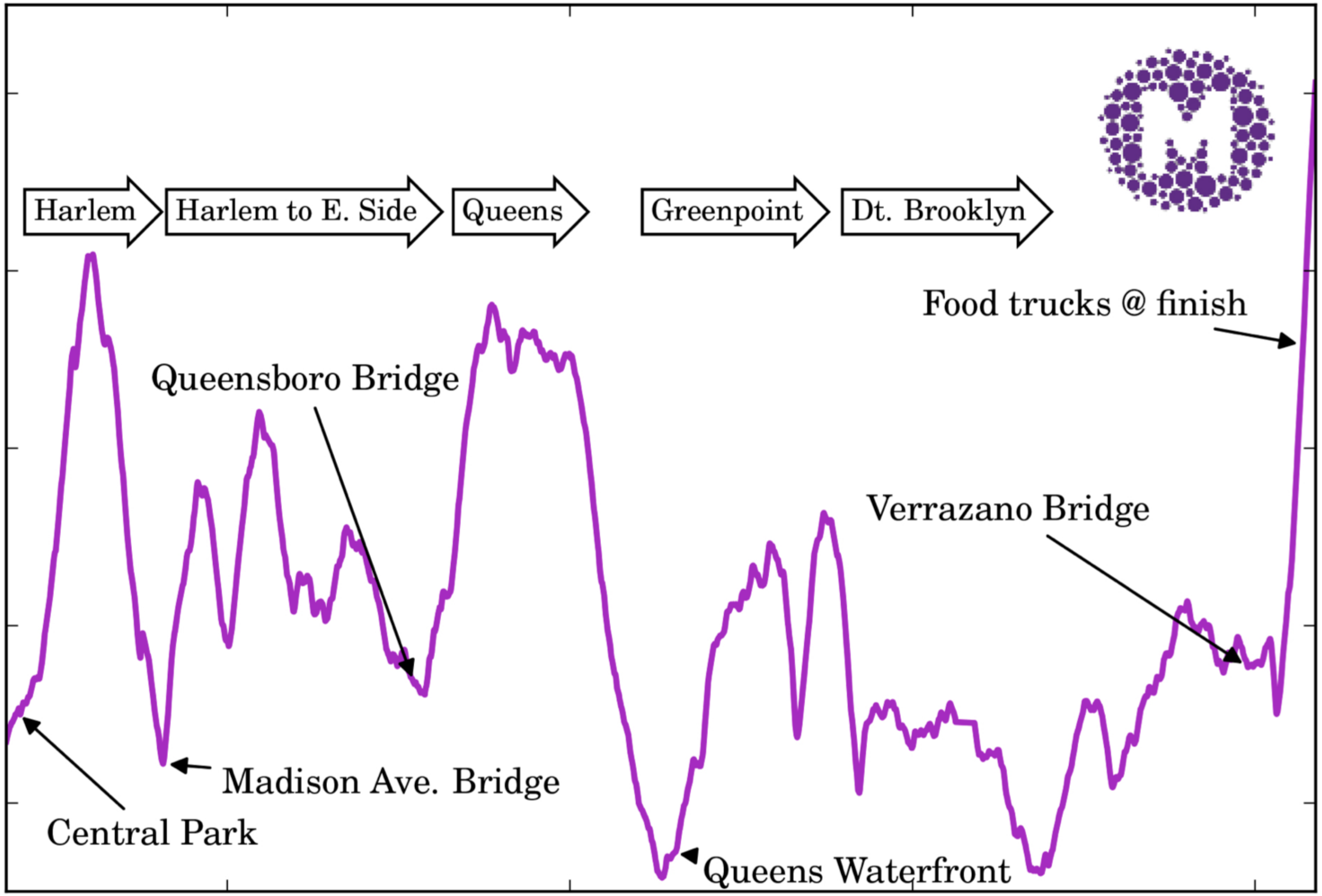
stationary PM2.5 (ZooBox)

mobile PM2.5 (AirBeam)

regulatory PM2.5 (TEOM)

QSENSE

*Example spatial granularity resolved from a **single** mobile $PM_{2.5}$ monitor*



*worn on a cyclist in New York City

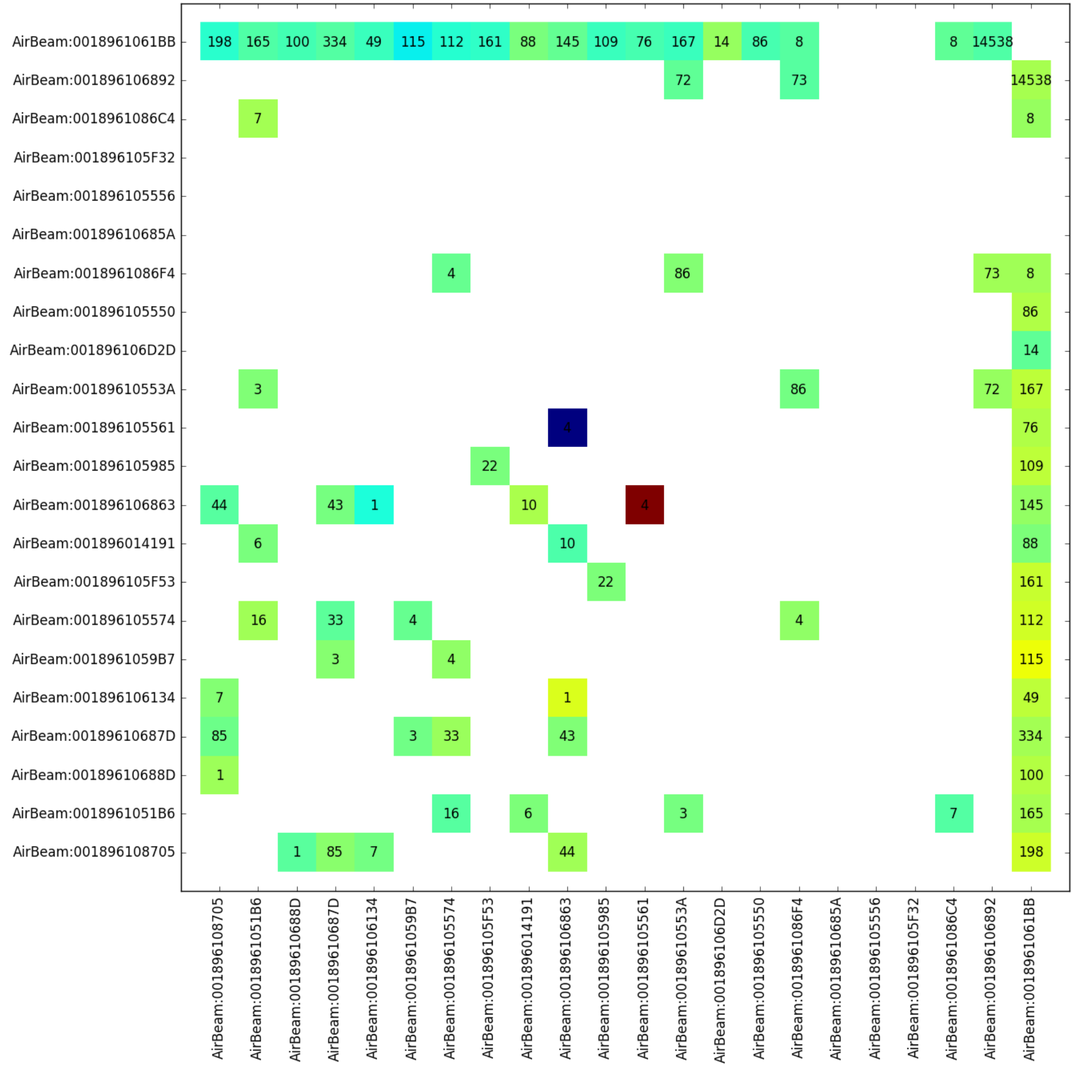
“Big Data” applies to analysis software as much as it does to data management software.

Given:
500 *sensors*
1 *months of data*
minute-frequency sample rate

Objective:
Given a cohort of sensors, apply an algorithm to every A<>B pair of sensors in the cohort, for every timestamp of data.

Requires:
$(500_{sensors} * 1_{pt/min} * 43800_{min/mo} * 1_{mo})^2$

$= 4.8*10^{14}$ (about half a *quadrillion)* operations



‘Each manufacturer uses it’s own protocols/standards, storage methods, etc…’

Will solutions be engineered from a single hardware manufacturer’s product? **NO.**

- Pipelining and unifying data with different formats and protocols is an expensive task with limited value-add.

As such:

- Hardware manufacturers should organize and expose data using industry-standard methods (e.g. RESTful APIs with appropriate query fields)
- Hardware manufacturers and software developers should work together to establish criteria and minimum standards for formatting data.

e.g. https://github.com/qsenseinc/protocol-standards

(mirror https://github.com/FullCircleEngineering/qsense-protocol-standards **)**

‘Quality is sometimes questionable and variable among manufacturers’

‘Each manufacturer addresses quality control/correction differently’

- Monitoring hardware needs to be verified and characterized. e.g. http://www.aqmd.gov/aq-spec
- Ultimately, data should be post-processed and “calibrated” on the Cloud, *as close to the decision-maker as possible.*
- Data “calibration” methods need to be both generalizable across all devices, yet also accommodate device-specific characteristics.
- Manufacturers should avoid processing of data ‘on-device’ *if such processing creates variability in the characterization of that device*. Doing so makes top-level synthesis and analysis difficult, if not impossible.
- Extracting intelligence from these data requires a completely new interpretive framework.

  → This framework must evolve from intimate collaboration between regulators and developers of data software/analytics.

The Good News!

These are all solvable problems (indeed, we're already solving them!). The tech industry already knows how to solve Big Data problems. Scalable compute and storage technologies can be purposed to address the aforementioned challenges. With the craze for data-driven-intelligence, almost every tech company now has engineers and software that could easily accommodate the data-management needs given even the most prolific dissemination of these sensors.

The Not-so-good News…

Building and running the "right" data system is *capital intensive.* Conventional data management and data analysis methods are not suitable (or will only work in narrowly constrained applications). AQ data systems need to integrate innovative and complex methods in order to *extracte value* from evolving ecosystems of environmental sensors.

Questions?

QSENSE

Nick Masson
Qsense Inc.
nicholas.b.masson@gmail.com