High Performance Streaming Analytics Break Out Group

Alison Brizius, Cees de Laat, David Ediger,
Melvin Ramos, Mike Papka, Nick Holliman,
Robert Grossman

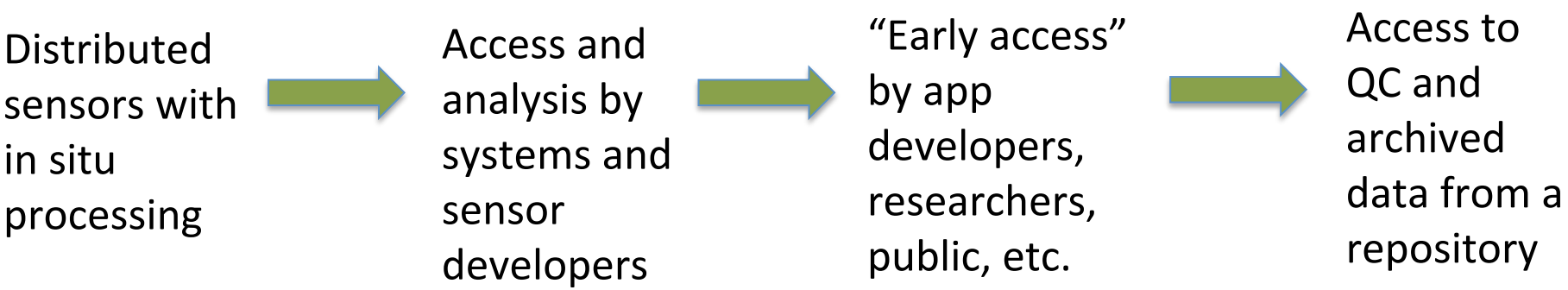
AoT workshop, Chicago, September 4, 2015


streaming
analytics
algorithms
platform
sensors
HPC
databases
visualization
streams
memory
process
problem
cloud
steering
data
couple
hpc
MP
integrate
specialized
pollution/air
fin
Analytic
noise
computing
Topics
failure
algorithmic
azure
reanalysis
prediction
stream
problems
systems
federated
capabilities
interest
Use
unobserved
Questions
distributed
in-situ
given
web
Domains
Discussion
decision
Problems
services
windchill
produce
WRF
continuously
adaptive
Infrastructure
different
link
much
phenomena
policy
identifying
Data
heat
processing
scope/selection
discovery
operations
room
new
fusion
quality
get
emergent
support
etc
emergence
workflow
education
created
collection
BG
app
Tomorrow
scalable
Research
resolutions
map
enough
store
demanded
multi-scale
real-time
batch
integration
today
feedback
Today's
demands
simulation
control
design
software
hardware
Cases
pollution
Huge
city
share

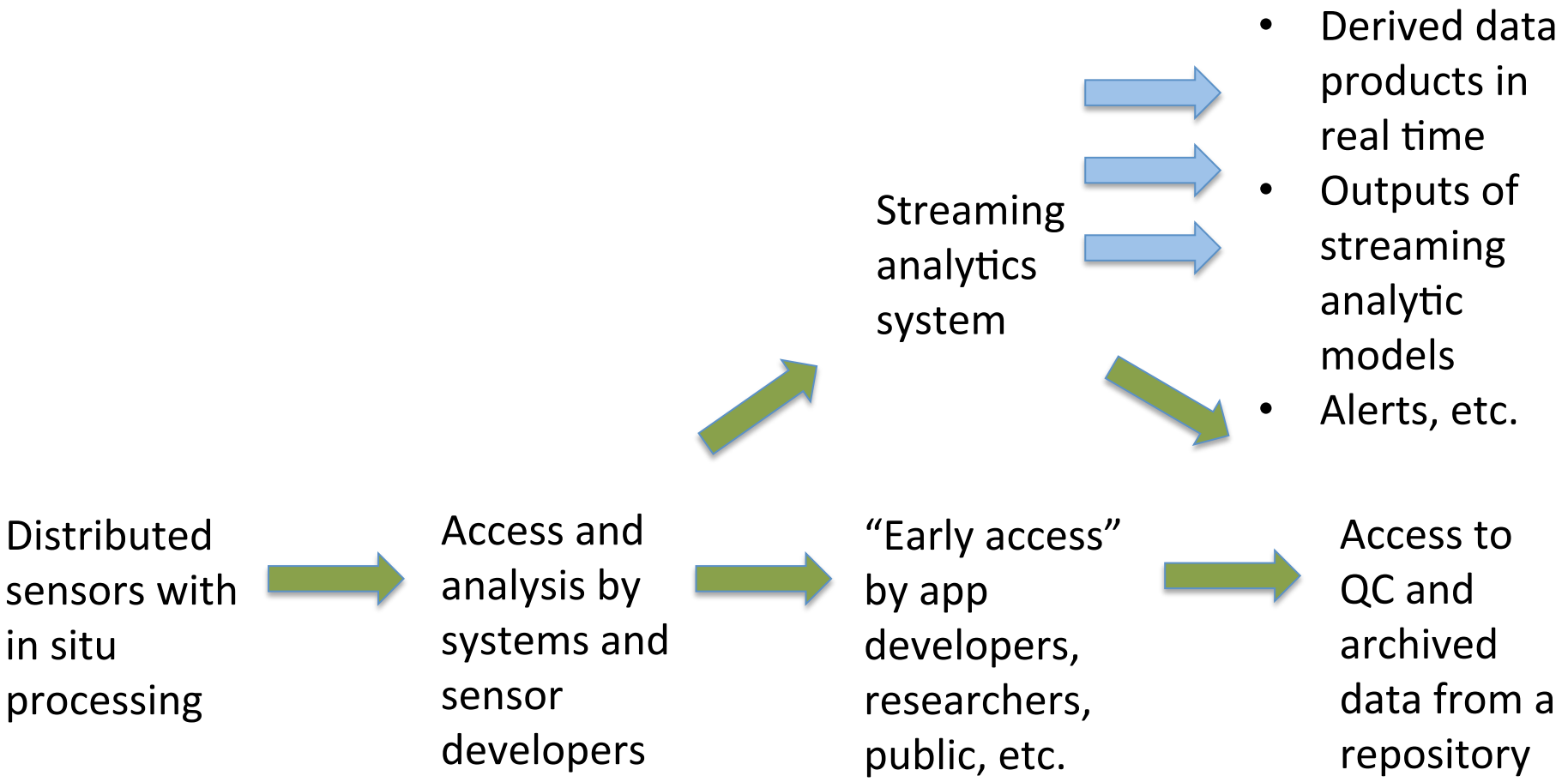
Streaming Analytics

- A data stream is an ordered sequence of data records that can be read only once, with only limited computing and storage capabilities available for processing.
- Think of this as “infinite” data, finite storage, read once.
- Compare to a batch analytics system or HPC system in which data can be read as often as required to complete the analysis.
- There are some research and several commercial systems that can process data streams.
- Over the past several years, there have been some recent open source frameworks (Akka, Storm, etc.) for stream processing that have been used for production deployments.
- There is relatively little work done on streaming ***analytics*** systems and even less on ***high performance*** streaming analytics systems.

Current model under discussion



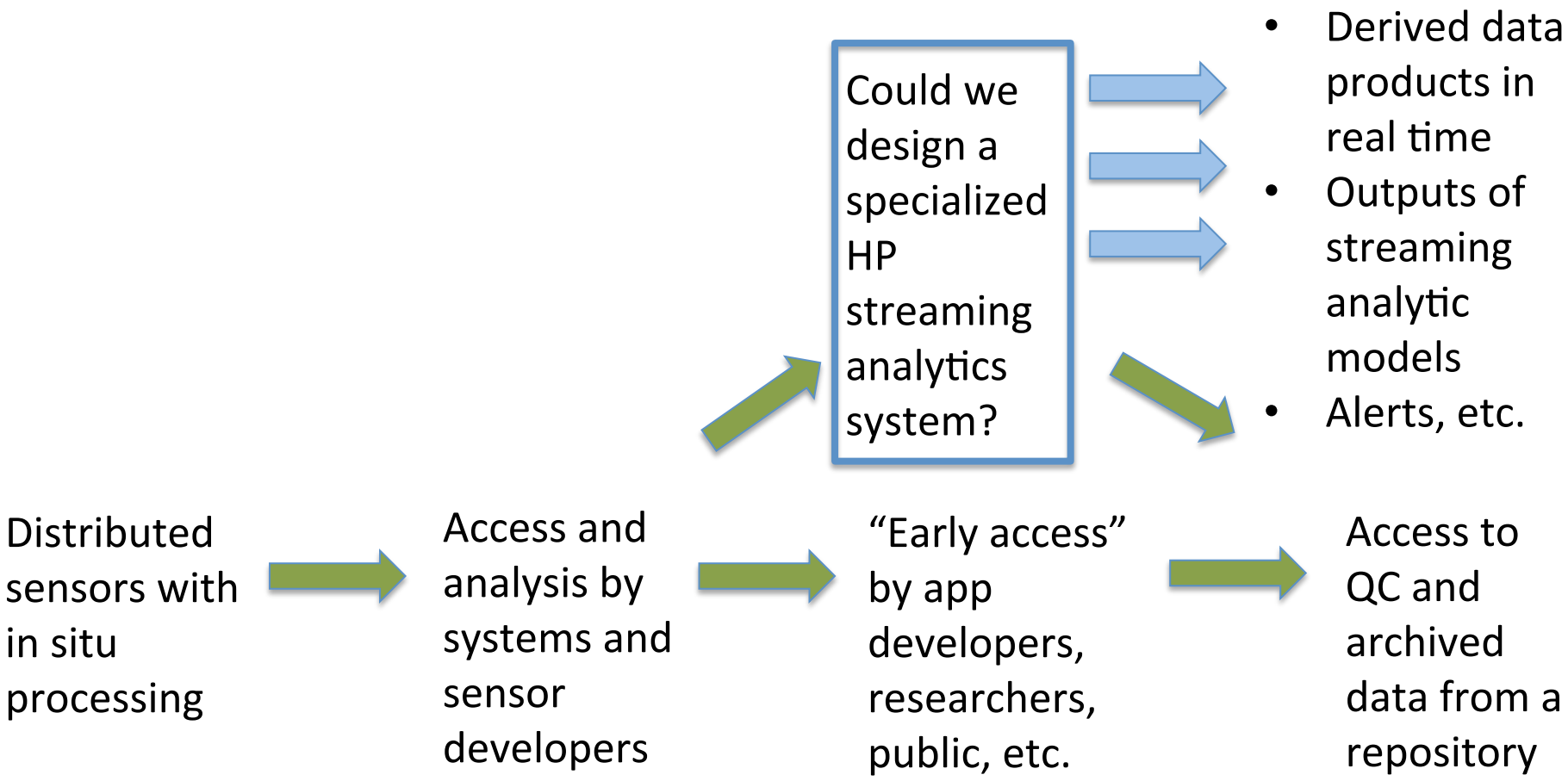
High performance streaming analytics systems



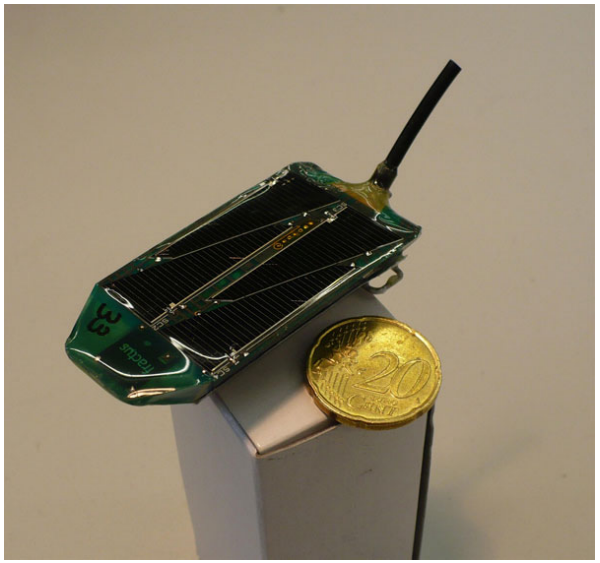
Streaming Analytics Use Cases

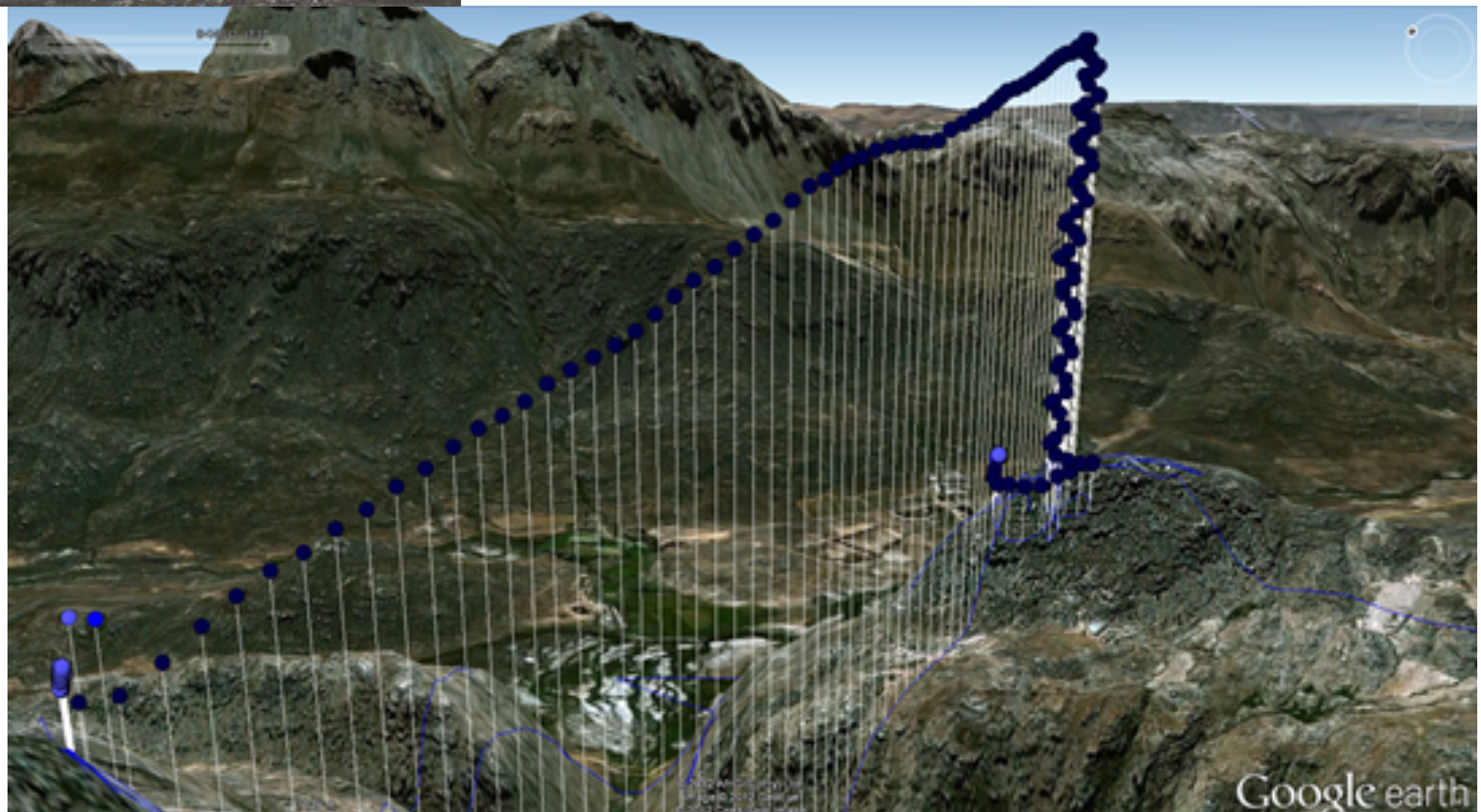
- Assume 10,000+ sensors distributed in a city
- What we can do with today's technology
  - Continuously updated heat map
  - Continuously pollution/air quality map
  - Continuously updated wind chill, noise pollution, etc.
  - Real time anomaly detection
- What we would like to do, but cannot do today
  - Complex data fusion in real time
  - Integrating the results of ad-hoc simulation into data products derived from streaming analytics
  - Streaming "re-analysis" of complex data streams
  - Smart trade off between in situ processing at sensors, centralized streaming processing and batch processing
    - Variant is regional streaming processing and roll up

High performance streaming analytics model



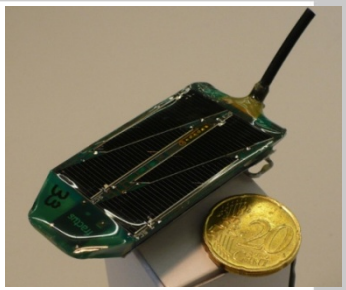
- Many of us have designed high performance computing systems.
- Would it make sense to consider how to design a specialized system for high performance streaming analytics that could be used to produce real time streaming analytics over a smart city?



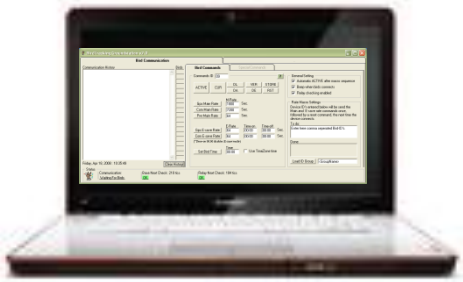

Google earth

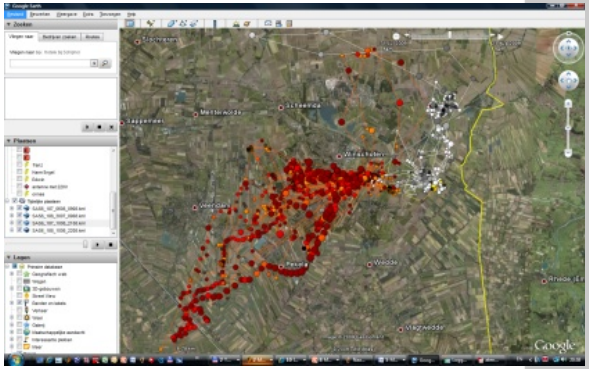
UvA- GPS Bird tracking

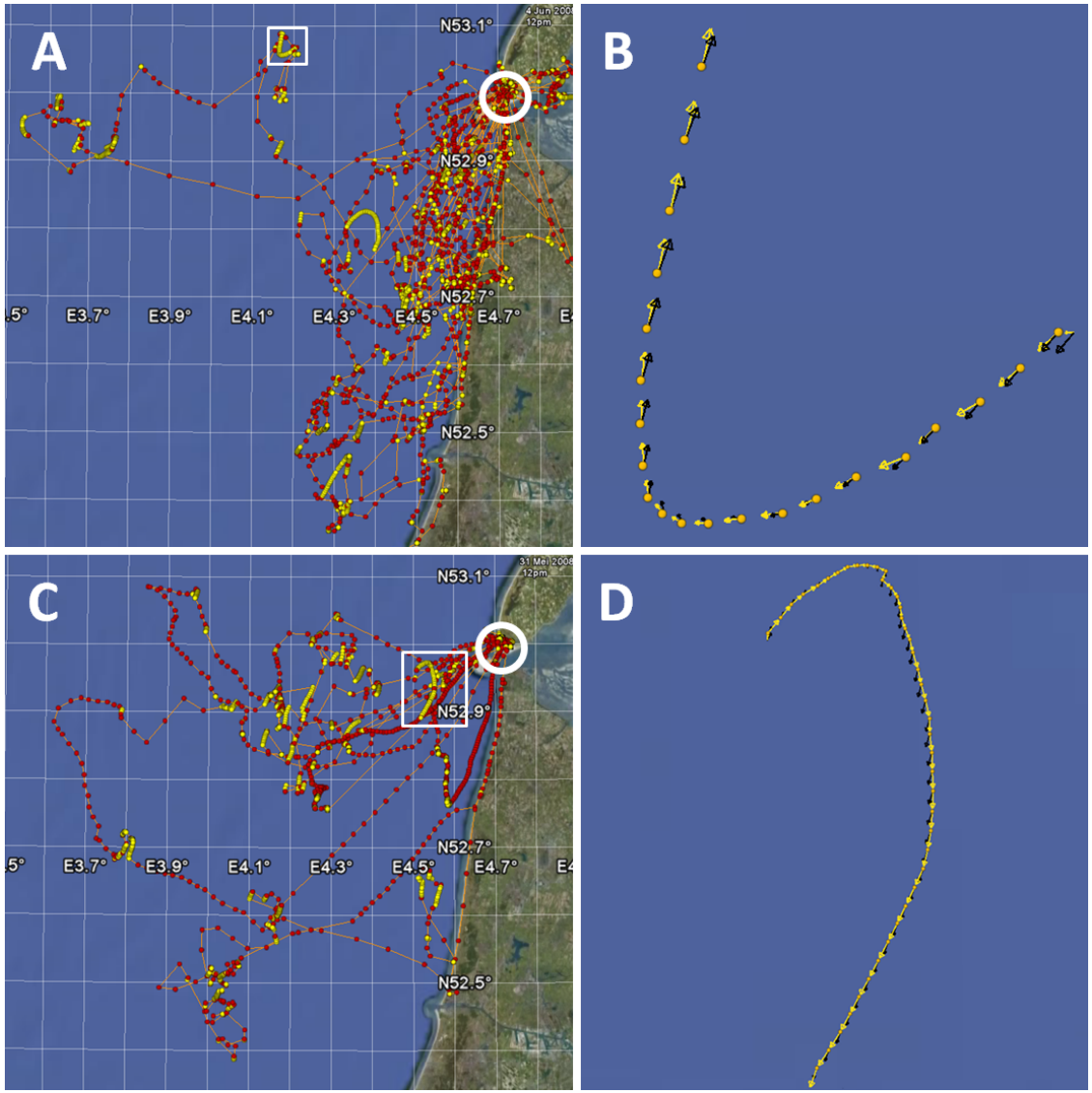
- Solar powered GPS logger 12 – 18 g
- 50 - 10.000 measurements per day
- Bi-directional wireless data transmission
- Upload new measurement schemes
- Base station & (mobile) relay network



Remote control







Shamoun-Baranes et al 2011