The Internet of Things and Fast Data Streams: Prospects for Geospatial Data Science in Emerging Information Ecosystems

Marc P. Armstrong
The University of Iowa
Iowa City, Iowa
marc-armstrong@uiowa.edu

Shaowen Wang
University of Illinois
Urbana, Illinois
shaowen@illinois.edu

Zhe Zhang
University of Illinois
Urbana, Illinois
zhez@illinois.edu

Purpose

- To address emerging challenges encountered when geospatial methods are applied to streaming geospatial information
- Particular focus on the concept of data velocity (fast data) and its effects on geospatial sampling and analysis
- Current geospatial analysis methods, rooted in the previous century, are unable to cope with emerging data streaming flood
- Need new tools to supplement, not replace, current methods

Velocity: A “V” Component of Big Data

- “Big Data refers to the inability of traditional data architectures to efficiently handle the new datasets. Characteristics of Big Data that force new architectures are volume (i.e., the size of the dataset) and variety (i.e., data from multiple repositories, domains, or types), and the data in motion characteristics of **velocity (i.e., rate of flow)** and variability (i.e., the change in other characteristics).”

(NIST, 2015)

New Data Ecosystem and IoT

- Widespread availability of mobile computing and inexpensive sensors with radios → new data ecosystem, internet of things (IoT)
- Digital components are incorporated into a vast array of "*things*", both large and small that are data stream generators

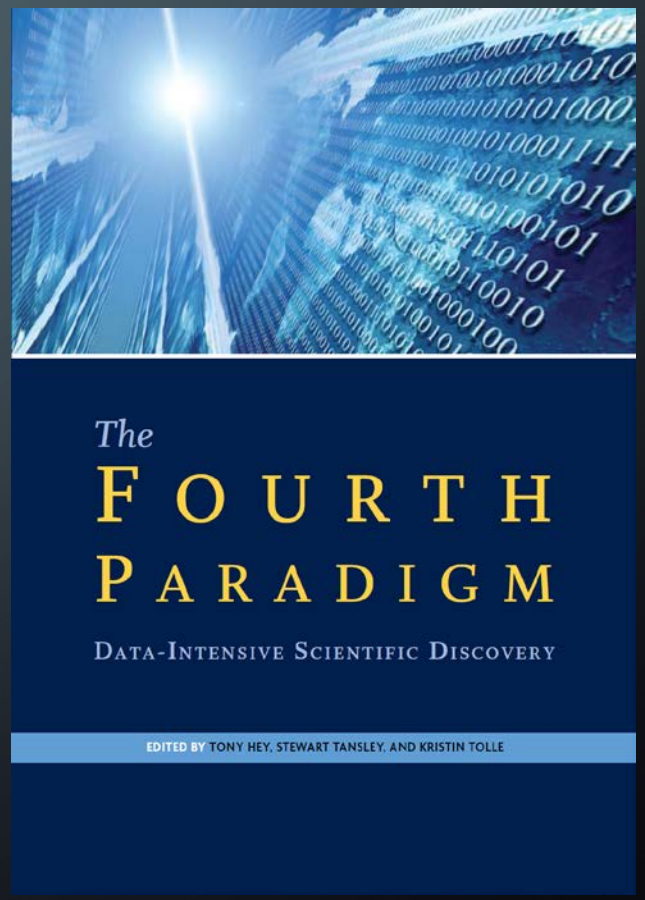
Jarr's Example

- 53,000,000 electric meters stream usage information several times each second to monitor changes in demand
- Feedback provided to "smart" systems about variable charges that can have an environmental impact by holding down demand peaks
- Other examples deal with detection, exceedance, process control, syndromic surveillance, cyber-physical systems and actuators…

Jarr, S. (2015). *Fast Data and the New Enterprise Data Architecture*. Sebastopol, CA: O'Reilly Media, Inc.

Effect of Data Volume and Velocity on Research Paradigms

- ***Paradigm One: Observational and Experimental***
- ***Paradigm Two: Theory and Models***
- ***Paradigm Three: Simulation and Computation***
- ***Paradigm Four: Data-Intensive Discovery***



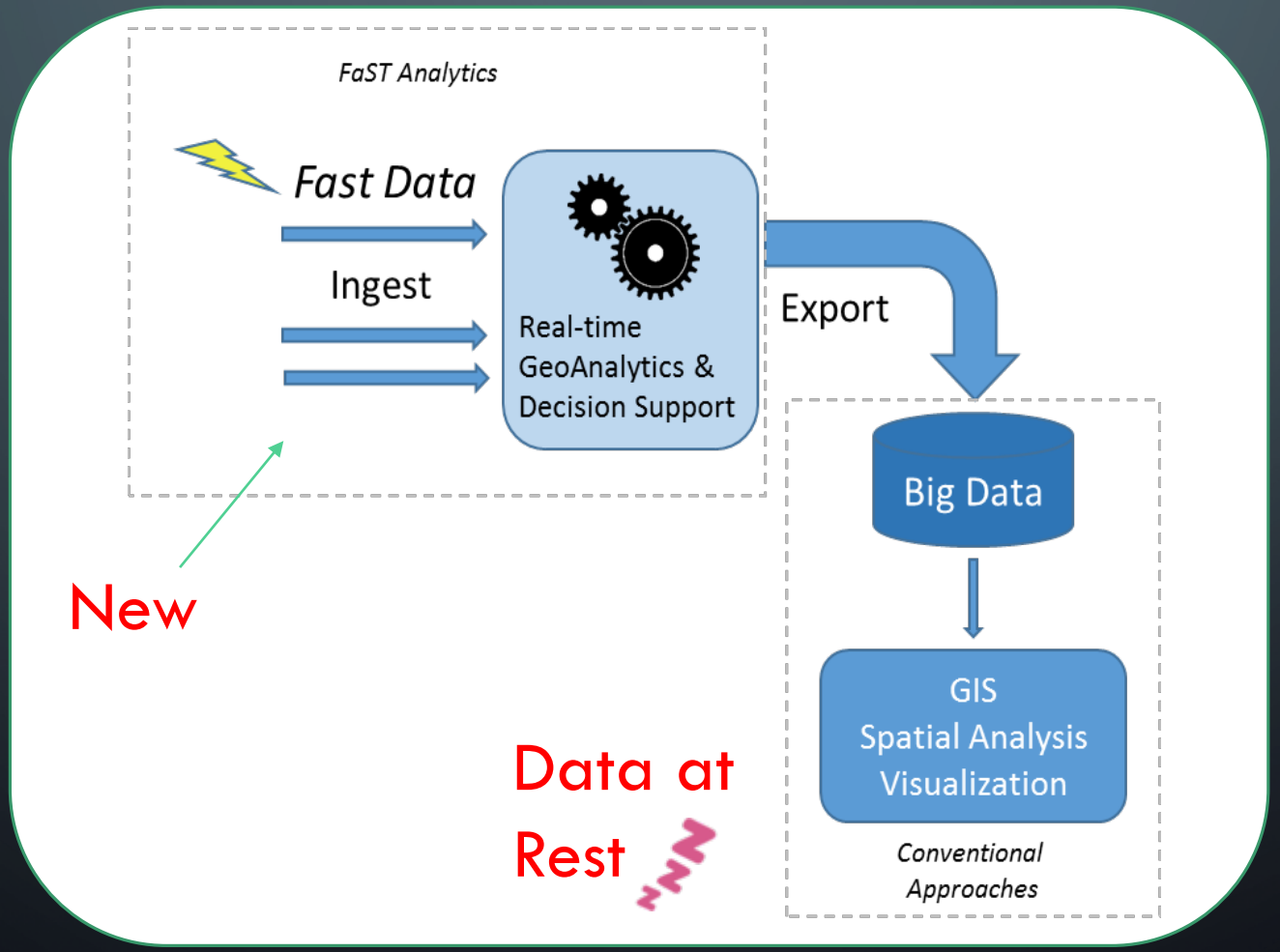
Data-Driven Discovery

- Different way of envisioning the relationship between theory and observation, effectively turning it on its head
- What pattern of observations yields a desired effect?... discover new relationships that can then lead to theoretical insight
- Important not to forget correlation ≠ causation, inductive fallacies, and falsification to help strengthen or demolish relations

Big Problems with Streaming Analysis

- Statistical analyses assume *data at rest*
- Stream sample size is unknown and unknowable (a known unknown?)
- Stationarity? Neighborhood instability?
- Can’t keep up, fall further and further behind the current

Staged Architecture for Fast Data Discovery



A Few New Approaches for Fast Data

- Reservoir sampling
- Approximate computing
- Sublinear time algorithms
- Edge computing

Reservoir Sampling

- Discarding observations, load shedding to reduce processing, will introduce bias
- Reservoir sampling reduces this bias
- First *n* stream elements = original reservoir & subsequent elements are candidates based on stream index
- Some increase retention likelihood of recently added observations; new data has greater salience than older

Approximate Computing

- Increasingly important to improve energy efficiency: shortcuts reduce energy-consuming cycles (*e.g.*, loop perforation)
- Premise: some applications do not require absolute correctness
- Not a new idea: “lossy” algorithms used for pictures and music
- According to Moreau, Sampson and Ceze (2015: 12) approximate computing is particularly relevant in mobile environments that involve sensor data collection and summarization

Mittal, S. *ACM Computing Surveys*, Vol. 48, No. 4, Article 62, March 2016

Sublinear Time Algorithms

- Use data assumptions to yield imprecise answers
- Rubinfeld and Shapira (2011:1562): “there are many situations in which a fast approximate solution is more useful than a slower exact solution.”
- Geographic theory about expected values and locations… Tobler’s First Law?

Edge Computing

- Data streams from spatially distributed sensors are pre-processed locally before being transferred to the cloud
- Staged or hierarchical architecture
- Decreases in latency and power consumption are reported

Big Data Analytics Needed for Fast Data

- Distributed file systems & parallelism
- Next milestone of 10^{18} operations (exascale) approaching
- Remaining problem areas include:
  - Inter-processor communication latency & memory movement
  - Scalable programming model: failure tolerant and locality aware
  - With billion-way concurrency, load balancing is critical

The CyberGIS Option

- High performance comes only from concurrent processes that do not suffer from data starvation and communication latency
- Data locality reduces communication needs
- CI middleware maps between memory hierarchies and geographic relationships, and also schedules load-balanced resources for large geographical problems

Simple Conceptual Illustration: Streaming Radiation Detectors

- CyberGIS workflow designed to detect anomalous radioactive sources in streaming data anonymized from Safecast (https://blog.safecast.org), a global volunteer-centered citizen science project
- Radiation data recorded in Japan using radiation sensor network
- Each stream observation: ID, time stamp, location (latitude, longitude), and measurement strength in counts per second

Approach

- CyberGIS enabled spatiotemporal data algorithm detects anomalous radiation sources and alarms if found
- Two radiation source types are considered:
  - Naturally occurring background radiation level (computed-on-the-fly using *k*-NN)
  - Anomalous sources that may come from nuclear weapons, dirty bombs, radioactive waste, or the precursors to such threats

Two Main Challenges

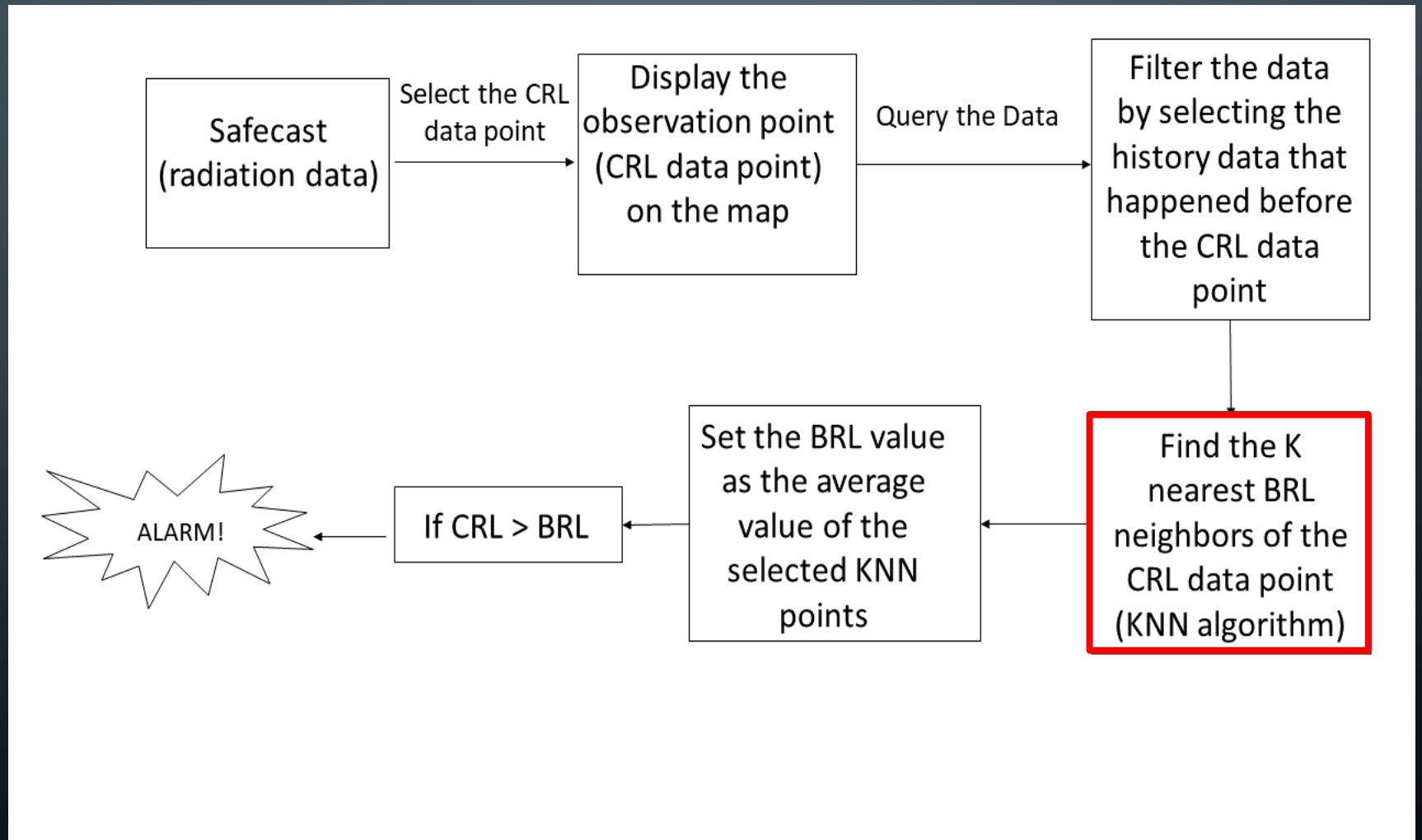
1. False positive alarms: ionizing radiation occurs from cosmic rays, & is emitted by rocks, soil and building materials
2. Safecast data has high volume & dimension with fast streaming speeds; causes data management and latency problems

- Need to detect a source in real-time with a low signal-to-noise ratio, where the source is the signal, and the background radiation is ambient noise, in the presence of confounding factors (GPS accuracy, detector motion, shielding and weather conditions)

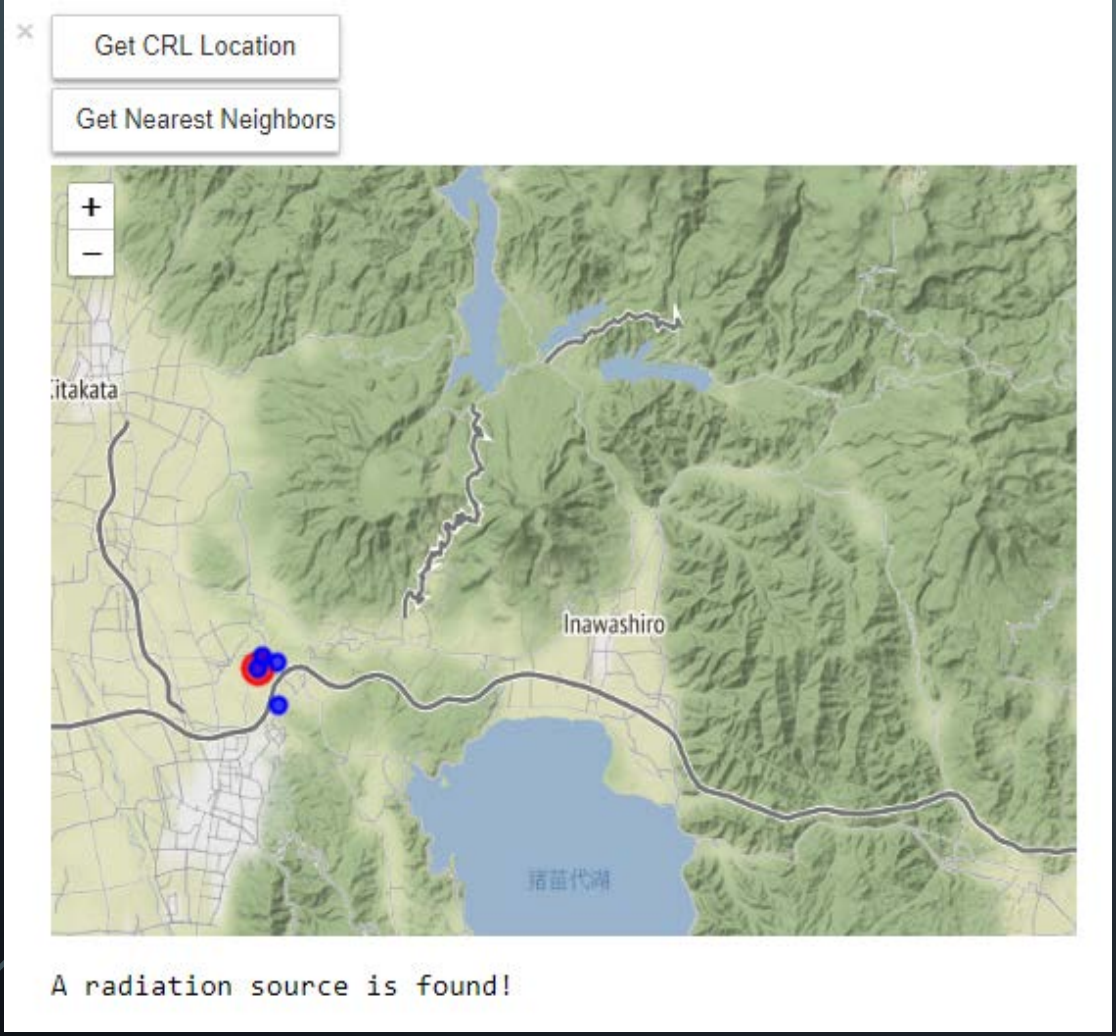
CyberGIS Workflow for Radiation Detection

CRL= Current Radiation Level

BRL= Background Radiation Level



Implementation



- CyberGIS-Jupyter Environment
  - Jupyter notebook
  - Docker containers
  - Cloud-based infrastructure provisioning
  - High-performance computing resources

Summary

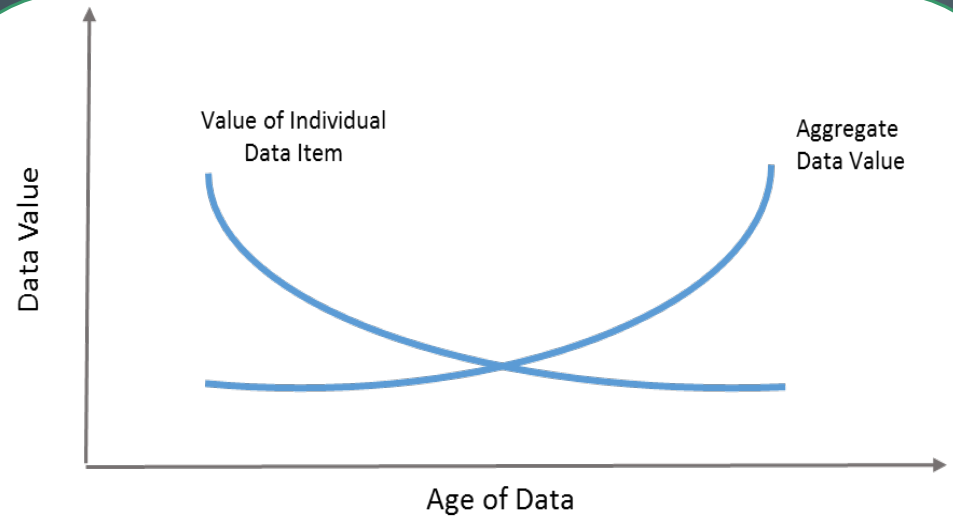
- Fast data sources are growing
- Need new methods of sampling & analysis to support data-driven discovery from fast data
- High performance computing and networking is required
- CyberGIS approach shows promise
- Real-time exceedance detection is but one motivating example
- **Many fruitful areas for future research**

Thank you!

The End

Change in Value Chain for Fast Data Streams

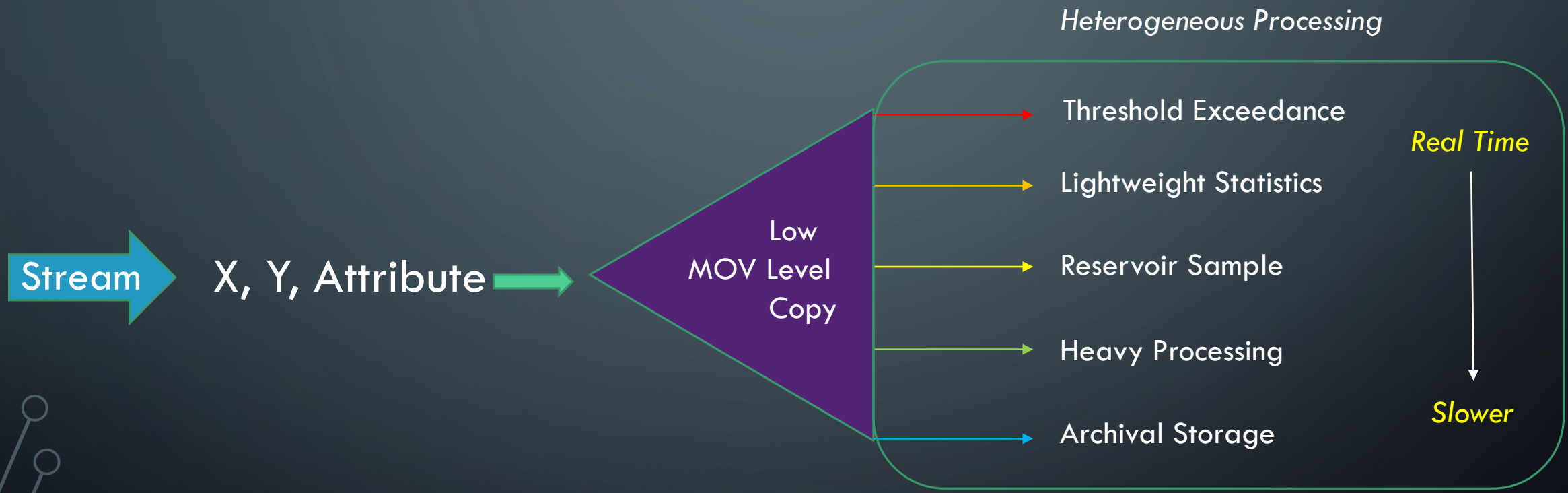
Generate → Store → Analyze → Use



Interactive	Real-time Analytics	Record Access	GIS Analysis	Exploratory Analytics
Milliseconds	Hundredths of seconds	Seconds	Minutes	Hours
Exceedance	• Aggregate • Sum	Retrieve attribute	Buffer	• GWR • Optimization

Fast Slow

A Different Kind of Parallel Perspective



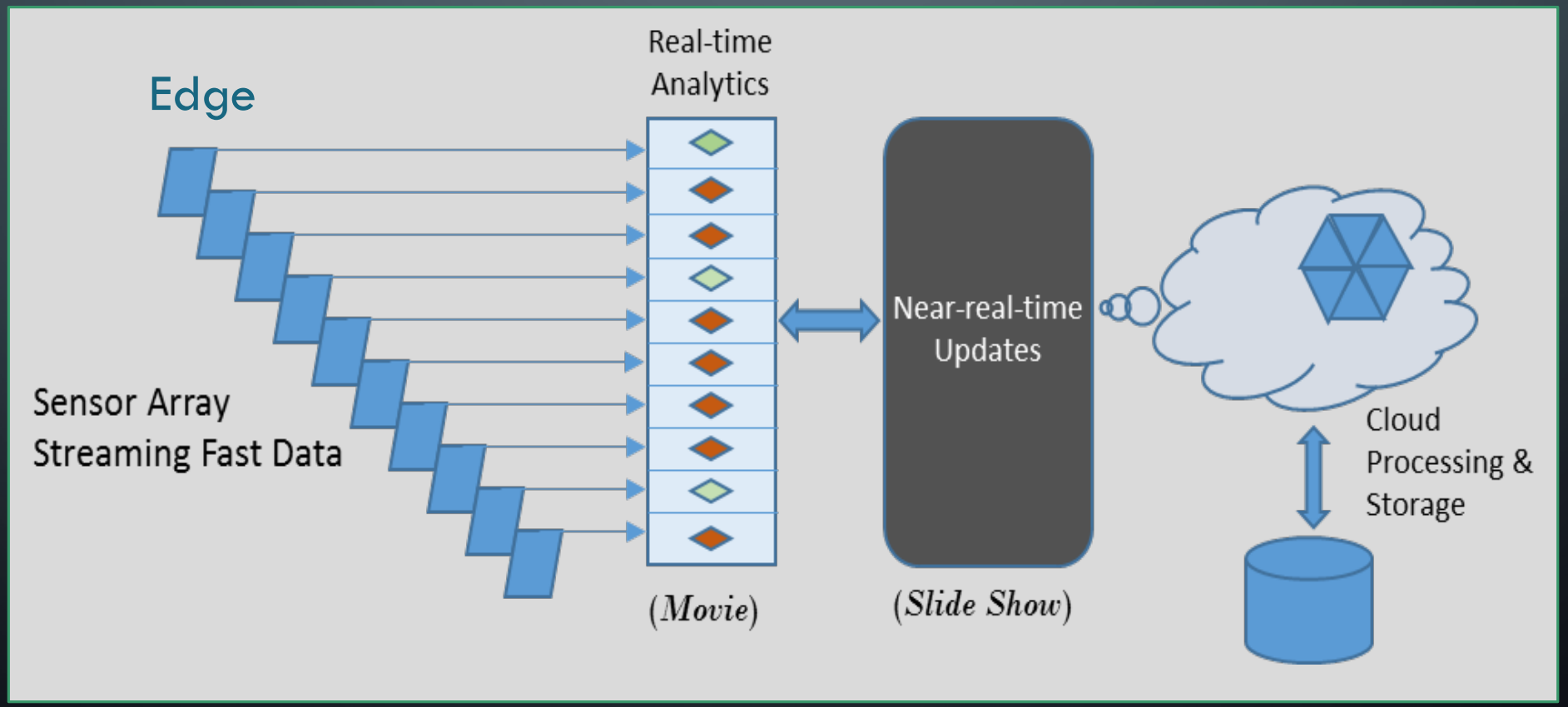
- inductive inference is not foolproof; it is possible to fall into logical traps
- An illustrative syllogism often involves {birds, feathers, flight and penguins}
- Robustness of an induced relation can be tested by evaluating predictions made about future system states
- If a prediction proves to be incorrect, that will require a revision to the theory or model that gave rise to the prediction

- In the Introduction to the Second Edition of *Theoretical Geography*, Bunge (1966: XVI) states that it is “time for geography to mature as a predictive science … seeking to predict locations where before there was contentment with simply describing and classifying them.”
- Tobler and Wineburg (1971) explicitly took up Bunge’s forecasting challenge by trying to identify the location of Bronze Age settlements in Anatolia using transformed place name dyads, a gravity model and trilateration.

Induction

- Applied to fast data streams to gain insights into processes.
- Holland *et al*. (1986:1) induction encompasses all inferential processes that expand knowledge in the face of uncertainty
- In an inductive analysis of a data stream, evidence in the form of new observations is accumulated in order to arrive at a conclusion
- Induction may be thought of as truth estimation techniques designed to yield best answer given information available

Staged View


Edge
Real-time
Analytics
Sensor Array
Streaming Fast Data
(Movie)
Near-real-time
Updates
(Slide Show)
Cloud
Processing &
Storage

SKETCHING

create synopses and reduce the dimensionality of streaming data.

For each new streamed element,

determine set membership (has this element appeared before or is it a member of a predefined set?),

cardinality (how many different types have appeared in the stream?) and frequency.

Put another way, if we have a set ***S*** it would be useful to be able to add additional elements to it, to test whether a new element is already a member of the set, and if true, increment a counter. Ellis (2014: 331) states that sketch algorithms have three desirable features:

1. Data updates are performed in constant time;
2. Storage space is independent of stream size; and
3. Queries are performed in linear time for the worst case.

A Bloom filter is one widely adopted approach to sketching (Bloom, 1970) that enables the efficient determination of “heavy hitters” or most frequent values in the stream

Compositing

- Sample taken from each individual to permit two tests
- Rather than test all *n* samples, composite into groups of, say, ten
- Each grouped sample is tested for a signal and if there is none, then all members of that group are negative
- If a signal is detected, each individual within the group is tested to determine the origin of the positive signal(s).

Sensor Networks

- Urban *in situ* sensors monitor particulates, gases, temperature... interest by technology companies, including AT&T, and Microsoft
- Others are mobile, transported by vehicles or worn by people
- Satellite broadband support is coming (OneWeb, LeoSat, SpaceX) using constellations (100s) of LEO satellites
- Concerned primarily with monitoring, though some close the loop with feedback & actuator control over a cyber-physical system

Environmental Remote Sensing

- LiDAR systems collect massive point clouds, with some systems capable of generating > 200,000 3D coordinates *each second*
- Cubesats standard form factor: 10 x 10 x 10 cm
- Original Skybox uses a mini-fridge form factor to acquire sub-meter imagery and video; generates > terabyte of data each day
- Questions about constellation density and collision cascades that could turn orbits into junkyards (“debris sats”)