COSCS494/594 Fundamentals of Digital Archeology
Preliminaries

Audris Mockus
University of Tennessee
audris@utk.edu

Preliminaries: Outline

- Preliminaries~[1]
  - Historic context
  - Digital Archeology as Data Science
  - Illustration
  - Why promising?

Why measure?

"... the art of measurement would do away with the effect of appearances, and, showing the truth, would fain teach the soul at last to find rest in the truth, and would thus save our life."

Protagoras, Plato

The absence of romance in my history will, I fear, detract somewhat from its interest; but if it be judged useful by those inquirers who desire an exact knowledge of the past as an aid to the interpretation of the future, which in the course of human things must resemble if it does not reflect it, I shall be content.

The History of the Peloponnesian War, Thucydides

Science(s) of human and collective nature

- A is the study of `past human events and activities`
- B is the study of human cultures through the `recovery, documentation and analysis` of material `remains`
- C is the study of developer culture and behavior through the `recovery, documentation and analysis of` digital `remains`

Digital Archeology

The study of peoples culture and behavior through the `recovery, documentation and analysis of *digital* remains`
Primary method

- `Organizational Tomography` is the reconstruction of people's behavior from the observed projections in digital remains
  - By linking traces from unrelated tools, e.g., by using
    - Chronology of events
    - Patterns of individual and social behavior
    - Nature of roles and tasks
  - Laws of data (see below)

Digital Archeology: Commitment

Understand human endeavor with increasing `precision` and `scope`

Premises

Definition (Knowledge)

A useful model, i.e., simplification of reality

Definition (Big Data)

Data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a reasonable time

Definition (Data Science)

The study of the generalizable extraction of knowledge from data

Why not Science?

`Science extracts knowledge from experiment data`

Definition (Operational Data (OD))

Digital traces produced in the regular course of work or play (i.e., data generated or managed by operational support (OS) tools)

- no carefully designed measurement system

Science: Temperature Experiment Data

Meteorology

- Weather stations
  - Known locations everywhere

Science: Temperature Experiment Data

Meteorology

- Weather stations
  - Known locations everywhere
  - Calibrated sensor, 5 ± 1 ft above the ground, shielded from sun, freely ventilated by air flow ...

Science: Temperature Experiment Data

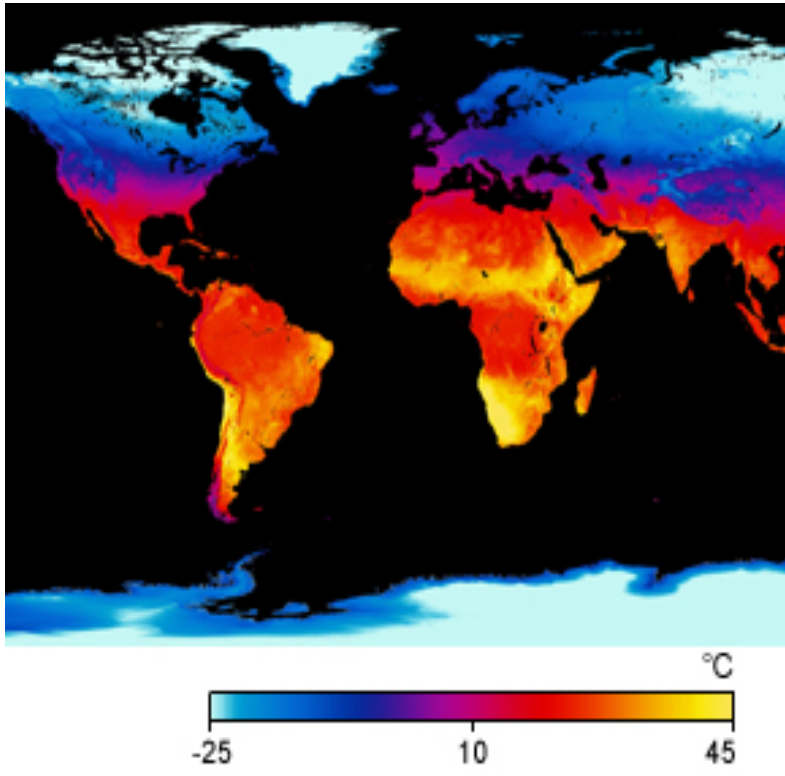
Meteorology

- Weather stations
  - Known locations everywhere
  - Calibrated sensor, 5 ± 1 ft above the ground, shielded from sun, freely ventilated by air flow ...
  - Measures collected at defined times

Science: Temperature Experiment Data

Meteorology

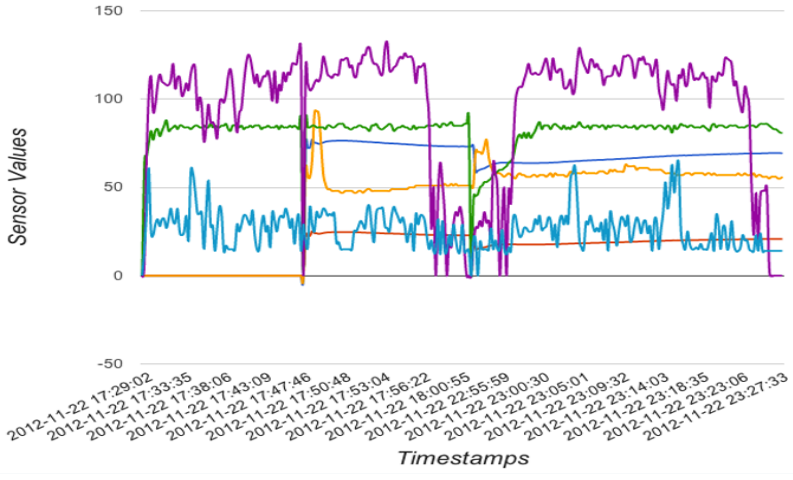
- Weather stations
  - Known locations everywhere
  - Calibrated sensor, 5 ± 1 ft above the ground, shielded from sun, freely ventilated by air flow ...
  - Measures collected at defined times
- Use measures directly in models



Data Science: Operational Data

Mobile Phones

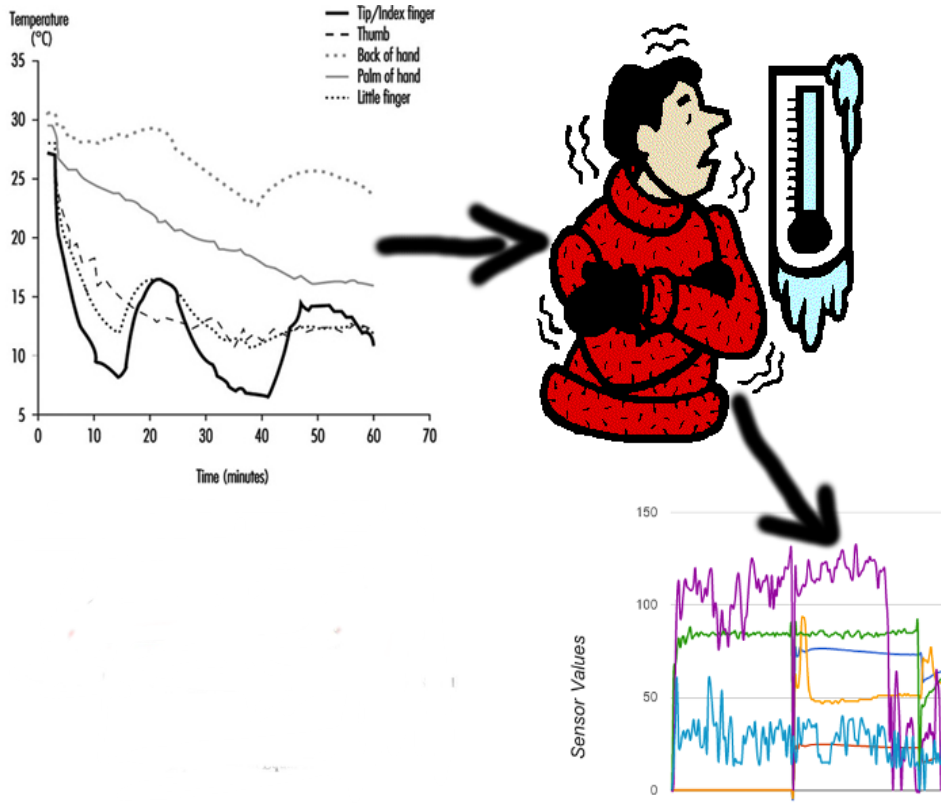
- Location, accelerometer, no temperature
  - No context: indoors/outside
  - Locations/times missing
  - Incorrect values



Data Science: Operational Data

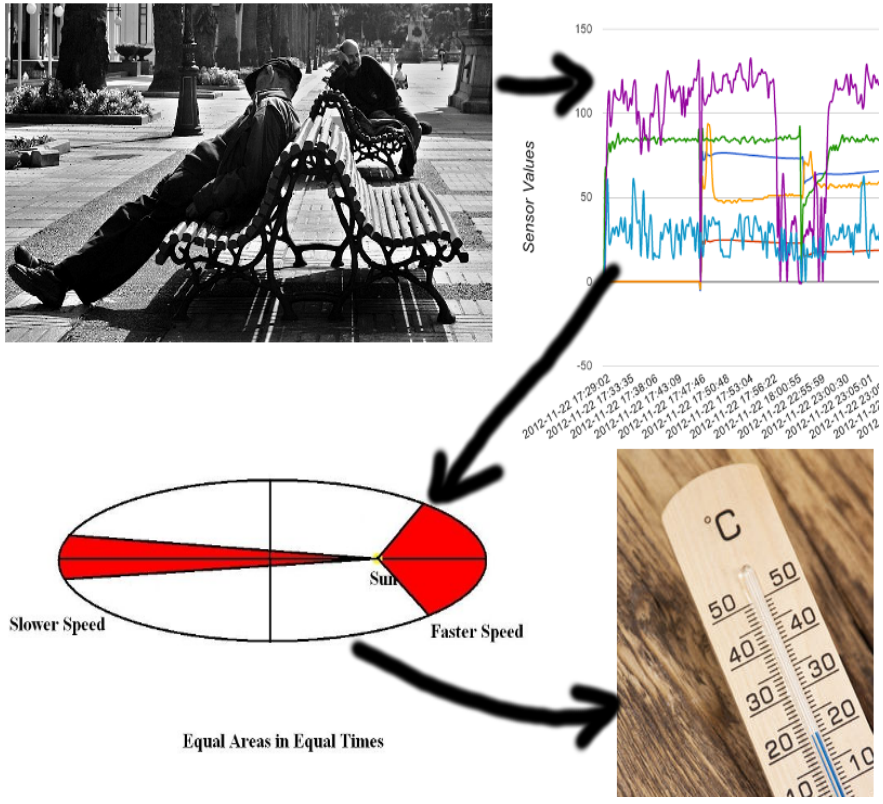
Mobile Phones

- Data Laws, e.g,
  - Temperature $\rightarrow$ sensor?
  - When outside?



Data Science: Operational Data

Mobile Phones



- Use Data Laws
  - Recover context, correct, impute missing
  - Map sensor output into temperature

Example Tools Producing OD

- Communication
  - Twitter, IM, Forums
- Documentation
  - StackOverflow, Wikies
- Issue tracking and customer relationship mgmt
  - Bugzilla, JIRA, ClearQuest, Siebel
- Text editors
  - Emacs, Eclipse, Sublime
- Version control systems (VCS)
  - SCCS, CVS, ClearCase, SVN, Bzr, Hg, Git

Why OD is a Promising Area?

- Prevalent
  - Massive data from software development
  - Increasingly used in practice
  - Many activities transitioning to a digital domain
- Treacherous - unlike experimental data
  - Multiple contexts
  - Missing events
  - Incorrect, filtered, or tampered with
- Continuously changing
  - OS systems and practices are evolving
  - New OS tools are being introduced in SE and beyond
  - Other domains are introducing similar tools

Engineering OD Solutions: Goals

Premise

- OD Solutions (ODS) are software systems
  - Complex/large data, imputation/cleaning/correction
- ODS feeds on (and feeds) OS tools

Goal

- Approaches and tools for engineering ODS
  - To ensure the integrity of ODS
- To simplify building and maintenance of ODS

Method

- Discover by studying existing ODS
  - Integrity issues tend to be ignored
  - Cleaning/processing scripts offered
- Borrow suitable techniques from other domains
  - software engineering, databases, statistics, HCI, ...
- New approaches for unique features of ODS

OD: Multi-context, Missing, and Wrong

- With each source of data we will try to identify
  - What events/records may represent different types of phenomena
  - What relevant observation we know are missing
  - What relevant observations may be missing
  - What values may be wrong, how, and why

Summary: defining features

- Digital Archeology
  - Is badly needed and challenging
  - Should be a fruitful area for decades to come
- Defining features of OD (digital traces)
  - No two events have the same context
    - Observables represent a mix of platonic concepts
  - Not everything is observed
  - Data may be incorrect

Summary: how to cope?

- Understand practices of how operational systems are used
- Establish Data Laws
  - Use other sources, experiment, ...
- Use Data Laws to
  - Recover the context
  - Correct data
  - Impute missing information
- Bundle with existing operational support systems

References


Audris Mockus.
Engineering big data solutions.
In *ICSE'14 FOSE*, 2014.