Bridging European Urban Transformations Workshop, Nov 2016

Data Collections, Storage, and Usage

Peter Triantafillou
Chair of Data Systems
Professor, School of Computing Science
Associate Director, Urban Big Data Centre
University of Glasgow

Outline

- Overview of UBDC
- The Infrastructure:
  - Data and IT Architecture at UBDC
  - Current and Vision
- The iMCD Data collection as a prime example
- Human in the loop
  - People-centered design

Bridging European Urban Transformations Workshop, Nov 2016

Part I

UBDC Overview

Urban Big Data Centre

Bridging European Urban Transformations Workshop, Nov 2016

Mission: Blend novel methods and complex urban data to address social, behavioural and environmental challenges facing cities:

- **Strategic Themes** - dynamic resource management; lifelong learning; economic innovations; citizen engagement / citizen science, planning and policy reform
- **Multiple Urban Sectors**: transport, housing, education, environment, energy – **particularly their connections**
- **Operate a national data service for UK research on cities and urban challenges** - Open data, secure and confidential data, real-time predictive analytics, data capture and linkage, synthetic data generation

Partners

And a network of US, Australian and Chinese institutions
10 Academic Disciplines – Urban Social Science, Data Science and Engineering

Potential Use of UBDC Data

Bridging European Urban Transformations Workshop, Nov 2016

- **Urban operations and management** – e.g., transport operations and traffic flow management, energy management and optimisation, crime detection and prevention
- **Knowledge discovery of patterns and trends** – e.g., understanding emerging issues, behaviours, public mood, critical concerns
- **Citizen engagement/civic participation** – involvement in plan-making, design and idea-generation; crowdsourcing travel and other information, Volunteered Geographic Information
- **Urban planning** – e.g., large-scale: urban land-use planning, mega-infrastructure planning; small-scale: site design, brownfield planning
- **Urban policy analysis and evaluation** – impact of proposed high-speed rail construction, increase in cigarette tax, crime prevention strategies, willingness-to-pay for policy changes

Fundamental to doing these data activities but at different scales:

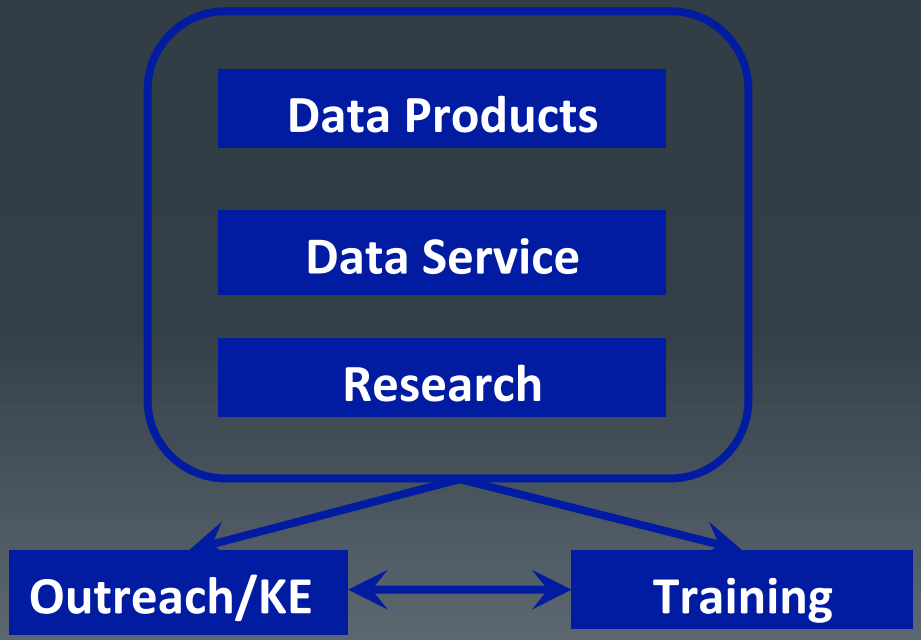
- Detection of changes
- Understanding links, causality and supporting processes
- Forecasting and understanding the future
- Assessing benefits and risks associated with different scenarios
- Evaluation of policies, other actions or potential actions
- Engagement with public and getting feedback

What we do

Bridging European Urban Transformations Workshop, Nov 2016

Run a national data service to facilitate urban research

UBDC Portfolio



Interdisciplinary perspectives – Urban Social Science, Data Science and Engineering

- Urban studies, planning & policy
- Statistics
- Economics
- Computer science
- Education
- Geography
- Mathematics
- Civil engineering

Major Achievements to Date

Data Service: Building a national data service focused on urban research:

- Acquired or generated several datasets;
- Made significant progress with data-related engagements (workshops, training, analytics service, impact case preparations) with a large number of stakeholders;
- Active and growing data service user base spanning HE, business, government and third-sector;
- Governance structure to ensure high-quality use of UBDC's data services;
- IT infrastructure and technical processes in place, actively serving analysts.

Research: Major research outputs including a book, conference proceedings from an international workshop, several papers submitted or accepted

Training and capacity-building programme – Well-attended training courses on basic to advanced methods (urban methods and simulations, transport analysis, spatial data management/geographic information science, data management)

International visibility in the area of urban data

Data Service Usage Statistics

Bridging European Urban Transformations Workshop, Nov 2016

- Number of datasets requested or downloaded – 1406 (including open datasets, safeguarded datasets, controlled data)
- Number of internal UBDC data products requested - 343
- Number of external safeguarded datasets requested – 67
- Number of controlled data service users – 3
- Total safeguarded and controlled service users – 109 – vast majority of users from outside the UBDC consortium

- Breakdown of safeguarded and controlled users: HE - 85%; Business – 7%; Government – 6%; 3rd sector – 2%

- Geographic Breakdown of safeguarded and controlled users: UK – 90% (Scotland – 65%; England – 25% - more England than Scotland in latest Call for Expressions of Interest); Europe – 5%; Rest of the world – 5%

Context Driving Our Work

Aspects	Characteristics
Technological	Urban information management: 1) Information generation and capture 2) Management 3) Processing 4) Archiving, curation and storage 5) Dissemination and discovery
Methodological	Data Preparation 1) Information retrieval and extraction 2) Data linkage/information integration 3) Data cleaning, anonymization and quality assessment Urban Analysis 1) Develop and apply methods to analyse various urban challenges 2) Ascertain uncertainty, biases and error propagation in the data

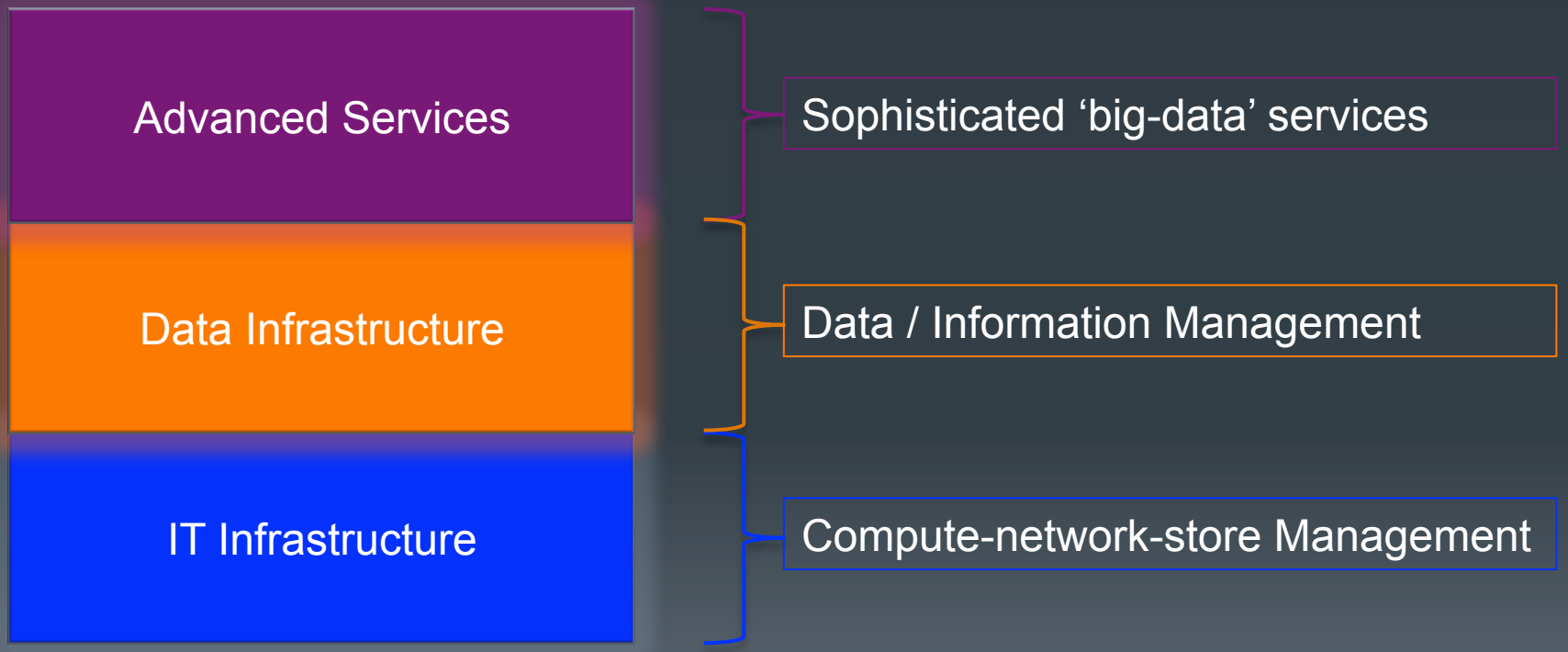
Bridging European Urban Transformations Workshop, Nov 2016

Part II

UBDC
IT & Data Infrastructures & Services

A Layered View

Bridging European Urban Transformations Workshop, Nov 2016



UBDC IT/Data Infrastructure

IT Infrastructure

Bridging European Urban Transformations Workshop, Nov 2016

- Hardware:Servers:
  - 2 Dell 920s:
    - each with 256GB RAM, 11TB diskspace, 96 cores
    - Hold VMs (~10 VMs: iMCD / CS research projects)
  - 2 Dell 720s,
    - each with 64GB RAM, 24.5TB diskspace, 32 cores
    - DB servers (tables extracted from CKAN)
  - 2 Dell 620s, each with 64GB RAM, 4TB disks, 8 cores
  - 3 mac mini servers
    - CKAN server + files, minecraft server, …
  - 2 DDN Storage platforms, each with 240TB

IT Infrastructure

Bridging European Urban Transformations Workshop, Nov 2016

- Software:
  - OSs / File Systems:
    - LINUX installations (mostly CS researchers) and
    - Windows 7 (social scientists)
  - VMs, Hypervisors
    - Offer independent insulated ‘servers’ to users
  - Bespoke software:
    - ArcGIS server, QGIS,
    - R, SPSS, …

IT Infrastructure

- Advanced IT Services
  - Backup and versioning on work performed on VMs
  - Resource Management
    - VM configuration/dimensioning
    - Monitoring
    - Load Balancing
    - Resource utilisation

In the process of establishing user requirements:
Hold directed consultations with user groups

Data Infrastructure

- A variety of **Data Formats**
  - Unstructured (e.g., text files, web pages, images, news feeds, twitter streams…)
  - Structured (tabular forms: from relational tables to .csv files and to spreadsheets)
  - "Specialised" (maps, etc).

- ➔ A Variety of **Data Systems**
  - NoSQL DBs:
    - Document DBs (e.g., MongoDB)
    - Graph DBs (e.g., Neo4J)
  - SQL DBs (Postgres Servers)

Does one system type fit all (e.g., SQL DB) ?

Data Infrastructure

Bridging European Urban Transformations Workshop, Nov 2016

- Data and metadata quality: To enable
  - Searchability
  - Discovery
  - Linkability
- Cataloguing
  - Searchability of data resources
- Access to sensitive and open data
  - Setup access paths to appropriate data/IT infrastructure, both to UG and Edinburgh sites

Principle :
"If we build it they will come …"

Account for / accommodate all user types / data formats / data uses, etc…

En route to a truly rich and useful resource

Advanced Services

Bridging European Urban Transformations Workshop, Nov 2016

- Information Retrieval Type searchability
  - Keyword queries
    - Not just on metadata descriptors/terms
  - Lucene ? ElasticSearch ? Solr ?
- Semantic alignment of data resources
  - Deriving “latent” linkability
  - Enable both:
    - Users who know exactly what they want to do with which data resources
    - Users who inquire about the usefulness of resources to their tasks !
- Discovery of services
  - Cataloguing and enabling searching for available services derived either by us or by SS researchers who have previously analysed our data resources and their results

Vision: Added Value

Bridging European Urban Transformations Workshop, Nov 2016

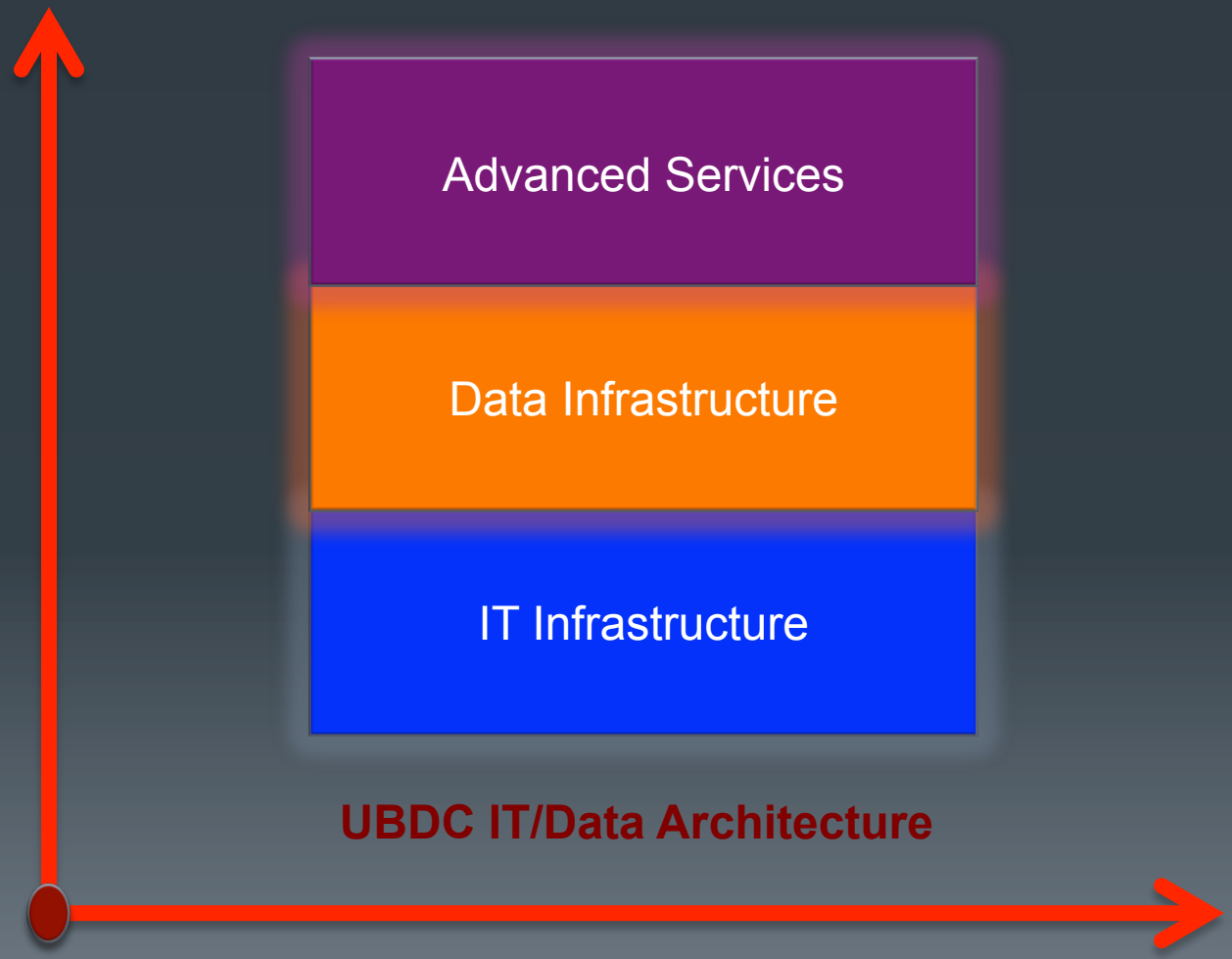
Principle is:
So we built it, and they came,

Now: Gather ‘knowledge’ and produced resources and make them available to community

Feedback Loop

Vision: Added Value

Bridging European Urban Transformations Workshop, Nov 2016



UBDC IT/Data Architecture

Bridging European Urban Transformations Workshop, Nov 2016

Part III

The iMCD (Integrated Multimedia City Data) Data Collection/Project

iMCD

Objective – Create a multi-strand data platform for understanding Glasgow

- **Primary survey** of 1500+ households in Greater Glasgow, UK, and household members (about 2600 persons)
  - *Questionnaire-based survey* – transport/travel and activity diary, education and literacy/skills tests, energy use, ICT/technology, cultural/civic engagement, attitude and preferences, caregiving, volunteering activities
  - *Sensing survey* (GPS and lifelogging use by participants)
- Significant **Information Retrieval** for a year (data from various text-based and multimedia data from the Internet, eg Twitter, online news)
- **Remote Sensing**: Very High Resolution satellite data and LiDAR data to construct dynamic Digital Surface Model of Glasgow
- **Sensor networks**: transportation, emissions, weather, lighting systems
- **Multiple private sector datasets**
- Glasgow City's **Open Data Portal &** other administrative data

Different strands of data collected, to the extent possible, for same time periods and same study area – however, important exceptions based on availability

iMCD Motivations

- To study biases in various types of Big Data through a combination of a large-scale structured household survey collected using statistically valid approaches and a large number of unstructured “Big Data” sources – establishes ground-truth and benchmark data;

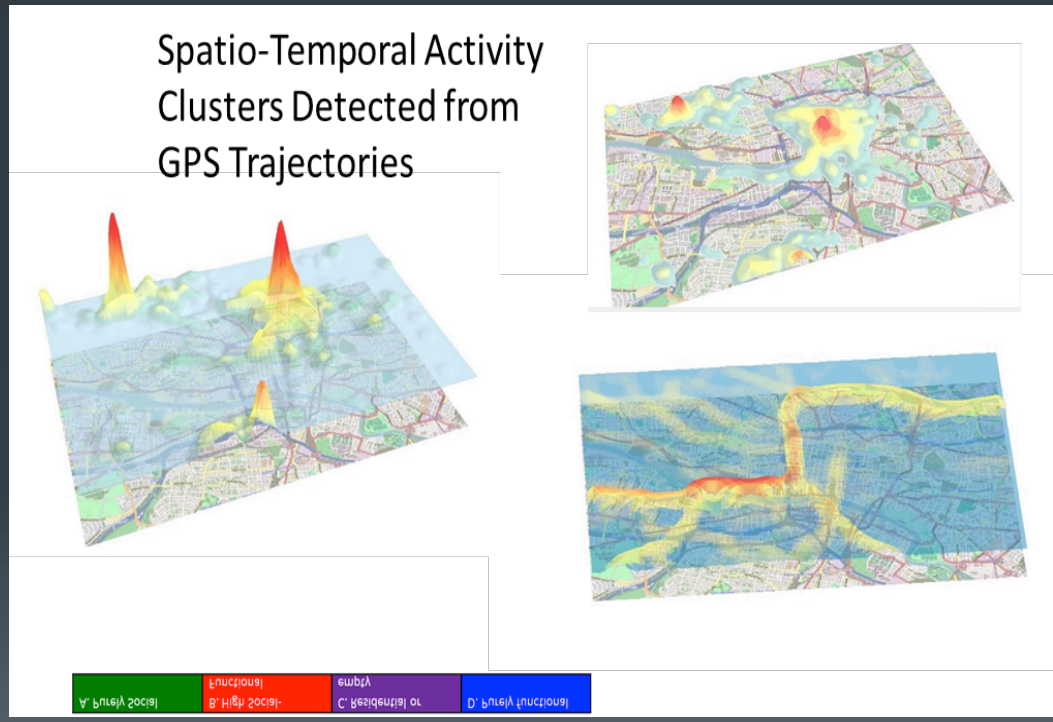
- To generate a data source that helps in the formation of targeted (sustainability, learning, healthy etc) behavioural interventions and to generate data-driven hypothesis that can be analyzed and evaluated using urban models and simulations;

- To provide concrete exemplars regarding technological, methodological, epistemological and political economy challenges

Urban Informatics

Bridging European Urban Transformations Workshop, Nov 2016

- Identifying changing **land use** patterns;
- Monitoring and predictive analytics of "**urban metabolism**" – traffic congestion, waste generation, energy consumption;
- Learning, engagement patterns and **education policy**;
- Connections to **digital literacy**, social exclusion and **public health**;
- **Housing policy** and private rental;
- Agent-based Modeling to simulate effects of **complex urban policies**



Bridging European Urban Transformations Workshop, Nov 2016

Part IV

The Human In the Loop

Vision: People-Centered Design

Bridging European Urban Transformations Workshop, Nov 2016

1. We gather data and organize it along with IT infrastructure
2. We make it available to the community
   - Enable smart citizens
   - Participate in data collection / production and consumption
3. Smart citizens generate new
   - information and knowledge and
   - New data services
     - Analytics tools, reseach methods, reports/conclusions, ..
   - Which UBDC
     - stores / curates / manages and
     - Makes available to the community
4. Repeat ad infinitum

People-Centered Design

Bridging European Urban Transformations Workshop, Nov 2016

Major Obstacles:

- Getting to the Centre
  - Humans: Digital divide and related social exclusions remain
  - Data: acquisitions
- Once there:
  - Sharability of Obtained Data / Information / Services

People-Centered Design

Acquiring data can be

- costly and time-consuming !
- Example: Zoopla
  - Purchased a data pipeline
    - ~3,000 calls to data access APIs per hour
    - physically acquiring the whole historical DB
      - can take a long time
      - requires dedicated human resources

People-Centered Design

Bridging European Urban Transformations Workshop, Nov 2016

- Sharing data is not easy!!!
  - **Licensing** restrictions
    - Who can use it and how much of it
- Legal expertise needed – cost: £ and time
- UBDC is a broker: need one license
  - Between UBDC and data owner and
  - Between UBDC and end-user
    - Too many possible end users
    - Hard to come up with a single EULA
- Liability risks:
  - Pass them on to end-users ?
  - What if they cannot afford these ? (e.g., private citizens)
  - How can we know of organisation or citizen can afford these?

People-Centered Design

Bridging European Urban Transformations Workshop, Nov 2016

- **Privacy** restrictions
  - UBDC has outsourced sensitive data access to a safe haven
  - But it stll must deal with understanding which dataset and which data access request falls in which category:
    - Open
    - Safeguarded (EULA)
    - Controlled (safe haven)
- Cost: £ and time
  - Information compliance services and Data Protection Acts specialists
  - Resource demands
  - Risks
- **Curb user expectations ?**

Bridging European Urban Transformations Workshop, Nov 2016

THANKS !