Data Science for Official Statistics

Key ideas

- New information =>decreased role for national statistical agencies and growth of independent data institutes
- New insights =>dynamic linkages; new units of analysis
- New challenges =>new types of human capital; privacy and confidentiality

=>New approaches needed

New types of information

Big Data

- **Observational Science**
  - Scientist gathers data by direct observation
  - Scientist analyzes data
- **Analytical Science**
  - Scientist builds analytical model
  - Makes predictions.
- **Computational Science**
  - Simulate analytical model
  - Validate model and makes predictions
- **Data Exploration Science**
  - **Data-driven science**
    Data captured by instruments or from the web, or data generated by simulation
  - Information extraction
  - Processed by software
  - Placed in a database / files
  - Scientist(s)/scholar(s) analyze(s) database / files
  - *Access crucial*

Source: Lee Giles

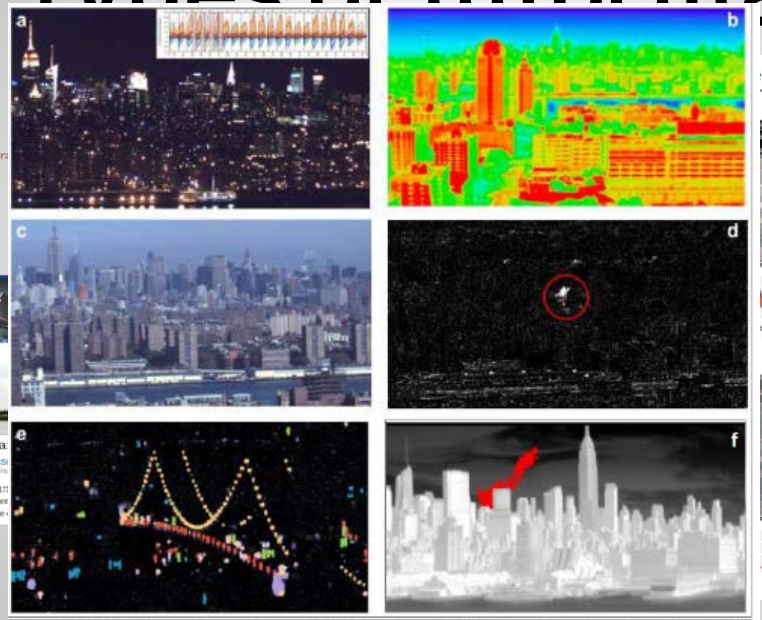
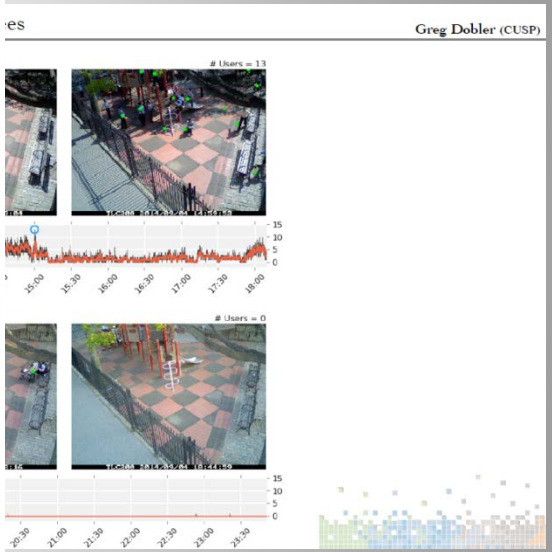
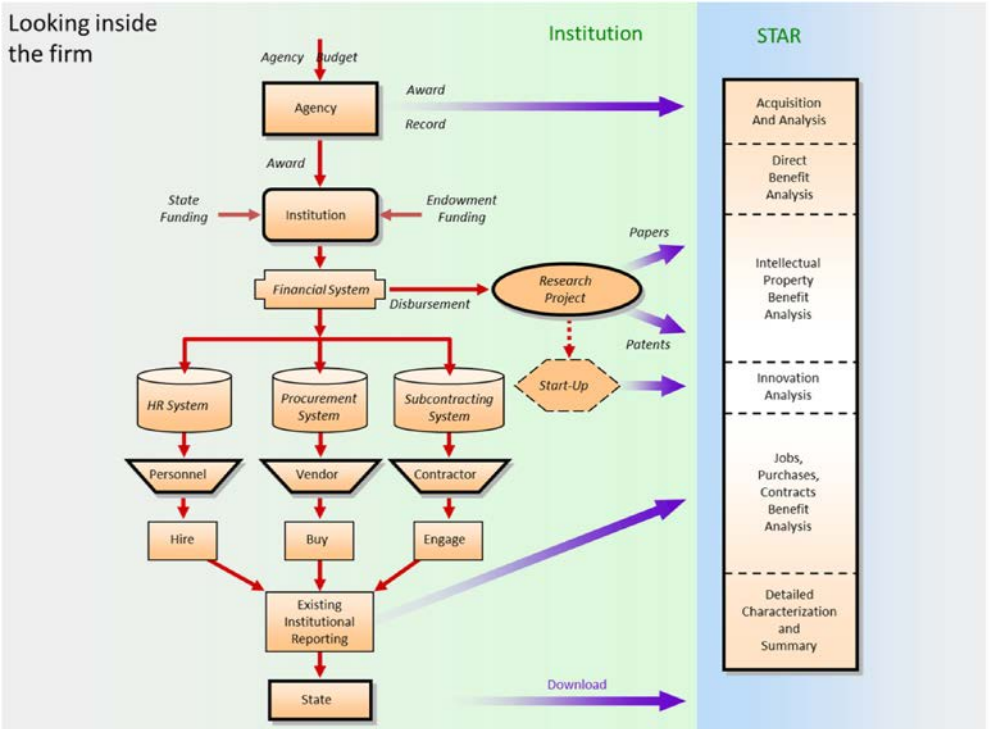


Figure 1. (a) Nighttime panchromatic image of Manhattan from the CUSP Urban Observatory in Brooklyn, NYC. The inset shows aggregate on/off transitions of some 4,000 lights in the scene's 1,000 buildings sampled at 0.1 Hz over three weeks – an unprecedented level of spatial and temporal granularity. (b) is a daytime thermal infrared image from the same site. The observed building temperatures result from a complex interplay of internal heating, insulation, insolation, surface material, and inter-building radiative coupling. The middle panels illustrate the value of detailed temporal observations: (c) is a daytime view of the field in (a), while (d) is the difference between (c) and an image taken 10 seconds later; a transient soot plume emitted during boiler start-up is evident. The bottom two panels demonstrate the potential of hyperspectral sensors: (e) is a view of the Manhattan Bridge and surroundings in which Visible/Near-Infrared hyperspectral imaging has been used to color-code the various lighting technologies identified, while (f) shows a transient plume of chlorodifluorosomethane identified during persistent Long-Wave Infrared (8-13 micron) hyperspectral imaging of Manhattan's west side. That ozone-destroying



time series of attendees

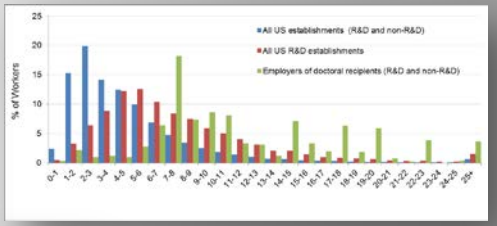
Greg Dobler (CUSP)



two panels demonstrate the potential of hyperspectral sensors: (e) is a view of the Manhattan Bridge and surroundings in which Visible/Near-Infrared hyperspectral imaging has been used to color-code the various lighting technologies identified, while (f) shows a transient plume of chlorodifluoromethane identified during persistent Long-Wave Infrared (8-13 micron) hyperspectral imaging of Manhattan's west side. That ozone-destroying chemical, used for refrigeration, is being phased out by 2020.

New Insights

- Employment dynamics
- Link between R&D and Innovation
- Credit applications



New Insights

Data

- Lending Club
  - provides data on loans underwritten from 2007-2015 including:
    - Hard data
      - Loan purpose
      - Income, debt-to-income
      - Delinquency behavior prior to application
      - FICO scores, Lending Club assigned loan grade (A-G)
      - Debt history (#lines of credit, utilization rates, etc.)
      - Employment history
      - Loan performance (current, paid off, charged off/default)
    - Soft data
      - Loan description (provided by borrower)

7

Credit scoring: Soft Data Examples

[1] If funded, I would use this loan consolidate two loans with interest rates of 15 and 16 percent respectively. I have no mortgage. One car is paid for and the other I bought from my sister. I pay her $200 / month. I owe her about $1000. The biggest monthly expense we have is tuition for two kids going to Catholic School, ($600 / month). I have been on the same job since 1990, with a salary of $54,000. My husband has been on the same job since 1995, with a salary of $30,000. My monthy expenses run about $2750. Borrower added on 03/11/10 > We have **really worked hard to clean up our credit** during the past five years. We are really wanting to use this loan to continue that by paying off higher interest loans with this loan. [debt-consolidation; paid-in full]

[2] Due to a lack of personal finance education and exposure to poor financing skills growing up, I was easy prey for credit predators. I am **devoted to becoming debt-free** and can assure my lenders that I will pay on-time every time. I have never missed a payment during the last 16 years that I have had credit. [debt-consolidation; paid-in full]

14

Credit Score: Soft Data Examples

[3] Purpose of loan: This loan will be used to payoff an existing 401K savings plan loan so that I can start saving money to buy a home. My financial situation: I've been employed with the same company for almost 7 years now. I've always **paid my bills on time and have paid off previous loan obligations** which included 2 car loans and 2 school loans. Monthly net income: $1,000 [debt consolidation; paid in full]

[4] This loan is for the purpose of gaining legal custody and visitation rights for my daughter. I want to hire an attorney of quality that will help me obtain my goal. My daughter lives in another state and I just **want what is my mine** legally and naturally as her father. I love her very much, she is 5 years old and also loves me very much. [debt-consolidation; *charged off*]

[5] I need this loan for a few different reasons: I need to fix the transmission on my car and the mechanic quoted me $700 I want to pay off the remainder of my credit card debt which stands at $450 I **want to have money** to apply to grad schools. I recently graduated with my B.A. and I want to pursue my M.F.A. Each grad school application is $150 and the GRE (Which I need to take) costs $100 [debt consolidation; charged off]

15

Source: Dennis Glennon

New analytical challenges

- Frames created by linkage
  - ⇒New error framework
  - ⇒Replication essential
  - ⇒Privacy and confidentiality addressed
- Database management skills
- Technical skills

This is really hard work

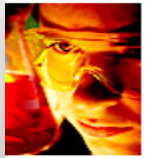
⇒Community effort essential and access is essential

Privac dentiality

both

my name

New Approach



NIH Human Embryonic Stem Cell Registry

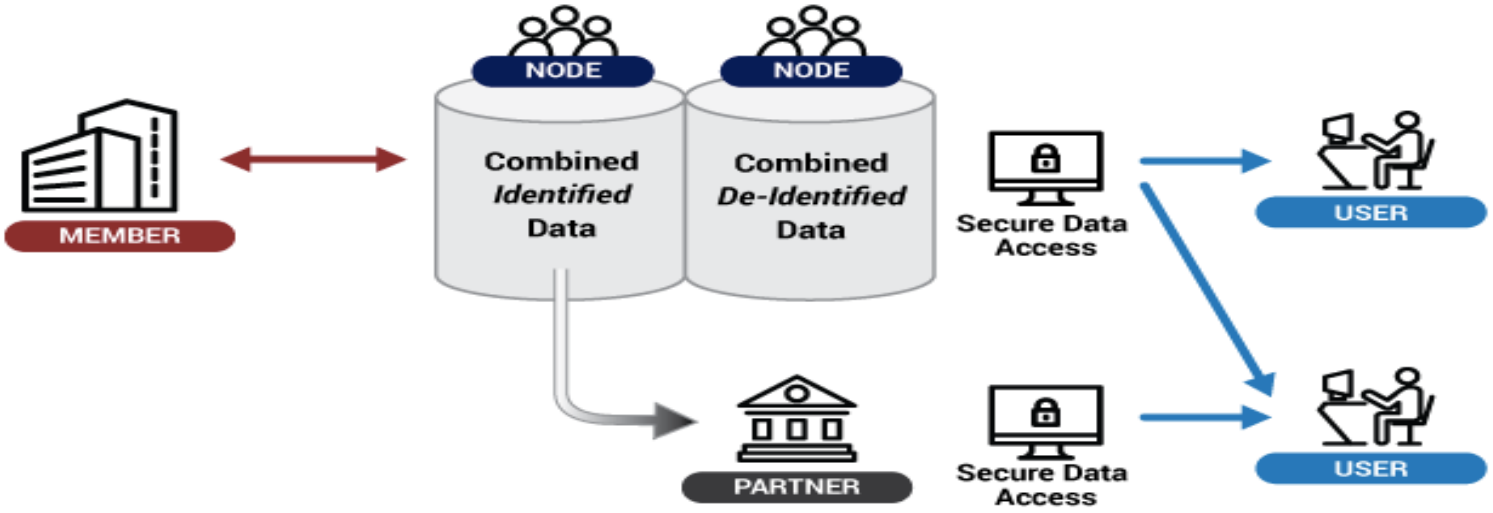
The Registry lists human embryonic stem cell lines that are eligible for use in NIH-funded research.

Review the Registry

MEMBERS: Universities contribute data, support infrastructure and receive campus-specific and aggregate reports

NODES: Approved nodes materially improve data, develop products, and expand user communities

USERS: Approved users securely access de-identified aggregate datasets



PARTNERS: Approved partners receive data from IRIS which they improve and make accessible throught their own secure systems

GOAL 1

GOAL 2

Data Curation and Management

- Data acquisition (Wagner, ENGR, GovLab)
- New data collection
- Curation (Libraries)
- Documentation
- Provenance
- Version Control

Data Access and Discovery

- Access and security controls (NYU ITS)
- Interrogation (Urban Profiler)
- Integration

Data Analysis, Collaboration and Reproducibility

- Collaborative space creation
- Visualization
- Workflow trace tools (Vistrails)
- Privacy and security

User Training and Engagement

- Standard Briefings on privacy and confidentiality
- Data hygiene
- Conceptual approach and new tools (big data class)

Dissemination

- Statistical Disclosure control
- Census RDC network
- API development (libraries)
- Agency engagement
- Citizen input and engagement (libraries)
- Workshops

Key ideas

- New information =>decreased role for national statistical agencies and growth of independent data institutes
- New insights =>dynamic linkages; new units of analysis
- New analytical challenges =>new types of human capital

=>New approaches needed

Thanks!

Questions?

Julia Lane
Julia.lane@nyu.edu

Credits

Multiple coauthors
Jason Owen Smith, Bruce W einberg, Rebecca Rosen, Barbara Allen, Ron Jarmin, Christina Jones, Ahmad Emad, Evgeny Klochikhin, Nathan Goldschlag, Nik Zolas, Kaye Husbands Fealing, Paula Stephan, Jacques Mairesse, Michele Pezzoni, Stefano Bianchini, Patrick Llerena, Joseba Salvo, Frauke Kreuter, Ian Foster, Fayid Ghani, Stefan Bender
Presentation template by SlidesCarnival