École de gestion
TELFER
School of Management

LIÉE au LINKED with LEADERSHIP

New Data Opportunities for Linked Environment-Economy-Innovation Research: Linking Web-Scraped Data with Statistics Canada Datasets

Economics and Environmental Policy Research Network Research Symposium

March 1st-2nd, 2018
Canadian Museum of Nature, Ottawa

R. Sandra Schillo, PhD
Assistant Professor
Telfer School of Management
University of Ottawa, Canada
schillo@Telfer.uOttawa.ca

Webscraping and Innovation Research

- Initial projects to explore potential and limitations
  - VC firms and portfolio companies: Identify management team, roles and gender
  - Identify innovation measures
  - Identify other useful KPI (key performance indicators) for small and medium-sized companies
- Key Issue: Validation

Context

- Statistics Canada – Telfer MOU
- SC is showing strong interest in
  - Having their data used
  - Collaborations on indicator development
  - "Big Data"
- We have a few projects using SC data
  - Currently in the journal review stages
- SC is responsible for protecting confidentiality of company responses
  - Cannot give out company names
  - This is why their data are high quality

Issues in Innovation Measurement

Usually survey-based

- Response rates:
  - Typically low when conducted by individual researchers
  - Statistics Canada can require companies to answer
- Survey tools:
  - OSLO Manual: Product/Service, Process, Organizational and Marketing Innovation; also: Innovation Strategy
  - Issues regarding quality of answers: social desirability bias, respondents may not know about all areas of operations
  - Ideally: measures based on observation (Acs & Audretsch 1988)

Types of Web Mining

- Web Structure Mining
  - Using links
- Web Content Mining
  - Using text, images, audio, video
- Web Usage Mining
  - Using data on frequency of views etc.

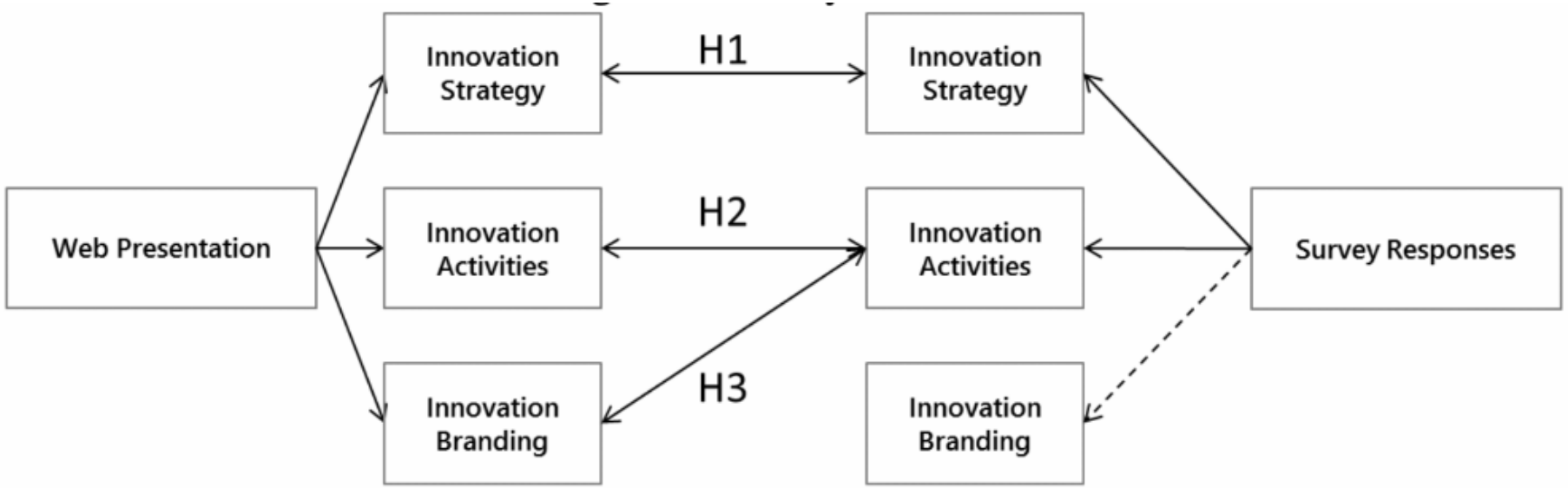
Web Content Mining

- Not common yet: Images, Videos etc.
- Text mining
  - Based on html code
  - Based on Natural Language (NLP)
- Common NLP methods
  - Frequency analysis of keywords
  - Use of dictionaries
  - Statistical analyses of text properties

Unstructured Data!

- Methods are relatively easy to implement
- Obtaining reliable measures is not trivial

Innovation Measurement Model



Literature – Innovation Scraping

	References	Web Content Mining	Web Structure Mining
Activities	Li et al. (2016)	*	
	Katz & Cothey (2006); Kenekayoro et al. (2013); Kenekayoro et al. (2014); Martinez-Torres & Olmedilla (2016); Scharnhorst & Wouters (2006); Rietsch et al. (2016); Martínez-Torres (2014); Martinez-Torres and Olmedilla (2016)		*
	Hyun Kim (2012)	*	*
Strategy	Youtie et al. (2012); Shapira et al. (2016); Arora et al. (2013)	*	
	Ackland et al. (2010); Hyun Kim (2012)	*	*
Branding	*Comments only:* Gök et al. (2015); Shapira et al. (2014)	*	

Validation

- This is the key issue for research on innovation (environmental innovation, as well)
  - Requires 'gold standard'
- Simple keyword frequency counts are unlikely to be good measures
  - Statistical analysis of keywords
  - Machine learning
- Statistics Canada can be 'gold standard'
  - Develop method to access to data

Our Project

1. Identify company URLs (UO/ISED)
2. Download web site text (UO)
3. Extract identifying info from web (UO)
4. Extract R&D / innovation measures (UO)
5. Match to Business Numbers (SC)

---

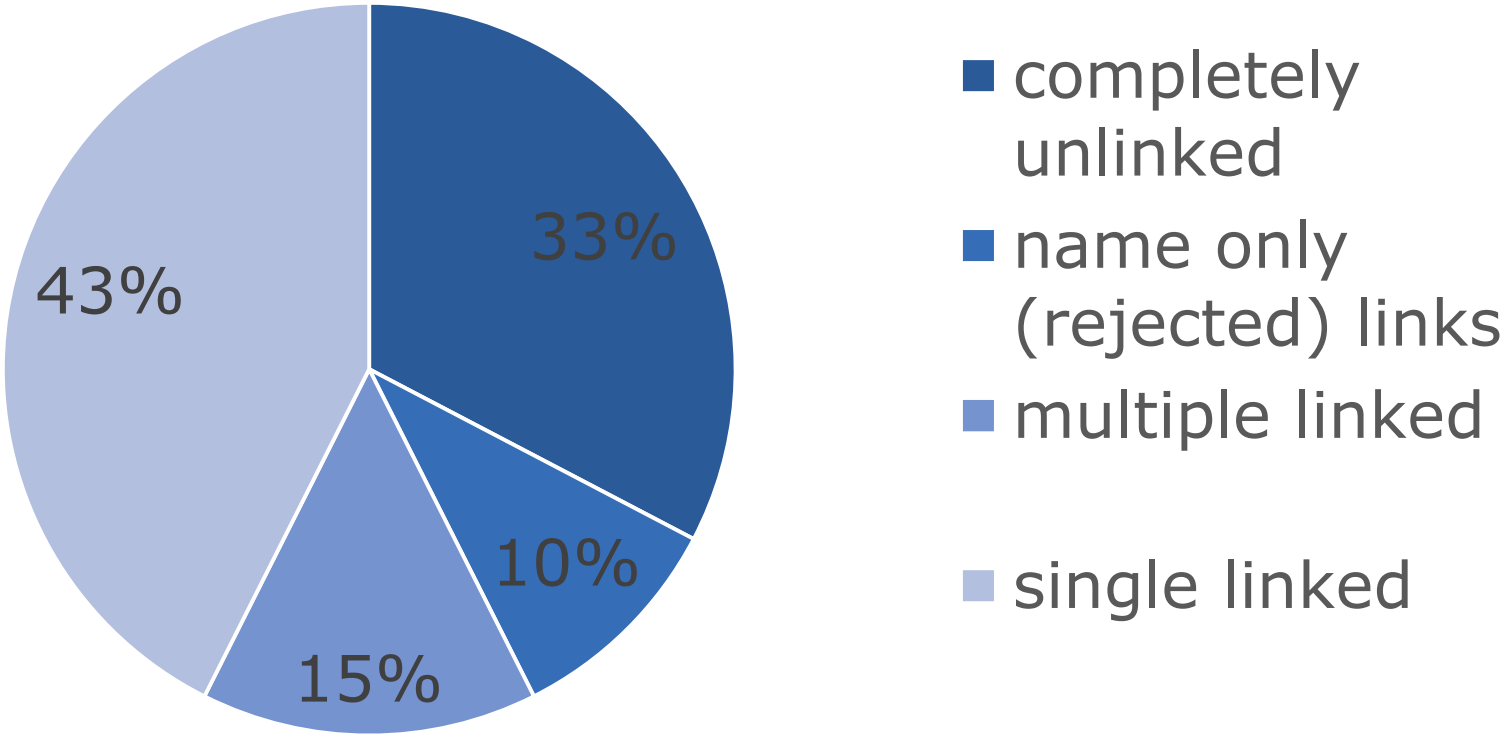
6. Retrieve innovation data (SC)
7. Compare R&D / innovation measures from web to SC data (UO@SC)
8. Use validated indicators for future research

Data Set

- 7944 businesses
- Manufacturing (NAICS 31-33)
- URLs provided
- Text analysis to extract
  - Addresses
  - Postal codes
  - Phone numbers
- Generally not available from web sites: Business Numbers

Noisy!!

Matching of Business Numbers



Overall Matching Rate: 57.5%

Discussion

- First project to attempt this
- 60% matched
  - Better than feared
  - Not good enough to replace SC surveys (that would be an unrealistic expectation)
  - Resulting data set 4500 companies
- Identified some room for improvement
- Potential to assist with frame-building, validation (of survey questions, as well)

Future Research

- Finish this project
  (link to innovation & research data)
- Explore opportunities for improvement
- Use not only web sites, but also social media
- Can we measure dimensions of inclusiveness of innovation (Schillo & Robinson 2017)?
  - Participation in innovation process, governance
  - Social, environmental impacts

École de gestion
TELFER
School of Management

LIÉE au LINKED with LEADERSHIP

schillo@telfer.uottawa.ca

École de gestion Telfer
Université d'Ottawa
55, avenue Laurier Est
Ottawa ON K1N 6N5

Telfer School of Management
University of Ottawa
55 Laurier Avenue East
Ottawa ON K1N 6N5

www.telfer.uOttawa.ca