Research Data Ownership

Personal perspectives from a (data-driven) researcher

Dennis Fok

Erasmus School of Economics

December 2, 2014

Outline

1. Who am I?
2. What do I do?
3. My current data practice
4. Needs and wishes concerning RDM

Who am I?

- Professor of Applied Econometrics at the Erasmus School of Economics
  - Teaching
  - Research
  - Management
    $\rightarrow$ coordinator of Master in Business Analytics & Quantitative Marketing
- Associate Director of the Erasmus Research Institute of Management [ERIM]
  ERIM:
  - Joint venture between Erasmus School of Economics & Rotterdam School of Management
  - Doctoral programme (Research master + PhD programme)
  - Research institute
  - Also active on themes as scientific integrity and professionalism

Research background

I create quantitative/econometric models to

- explain and predict behavior of customers and firms
- help companies optimize their (marketing) strategy
- understand decision-making in markets

Research background

I create quantitative/econometric models to

- explain and predict behavior of customers and firms
- help companies optimize their (marketing) strategy
- understand decision-making in markets

Key features

- Data-analysis
- Modeling
- Non-standard techniques
  $\rightarrow$ developing new techniques and software

Three examples

1. Online product recommendations
   - Use data on previous purchases by individuals
   - Predict the new products that they are likely to buy
   - Main challenge: size of assortment and number of customers

Three examples

1. Online product recommendations
   - Use data on previous purchases by individuals
   - Predict the new products that they are likely to buy
   - Main challenge: size of assortment and number of customers
2. Explain the launch decisions of new products
   - Use data on sales, advertising, quality, price of video games
   - Explain why some games get sequels and some don't
   - Main challenge: launch decision depends on many aspects

Three examples

1. Online product recommendations
   - Use data on previous purchases by individuals
   - Predict the new products that they are likely to buy
   - Main challenge: size of assortment and number of customers
2. Explain the launch decisions of new products
   - Use data on sales, advertising, quality, price of video games
   - Explain why some games get sequels and some don't
   - Main challenge: launch decision depends on many aspects
3. Using purchase histories to identify customer "projects"
   - Use data on customer purchases
   - Use the combination of bought products to predict what project consumers are working on (think about Do-It-Yourself stores)
   - Main challenge: projects are not pre-defined + one product can belong to multiple projects

Data sources

My data comes from many different types of sources

Data sources

My data comes from many different types of sources

Concerning the previous examples:

1. Online product recommendations
   - data made available by an online retailer
     $\rightarrow$ Non-disclosure agreement [NDA]

Data sources

My data comes from many different types of sources

Concerning the previous examples:

1. Online product recommendations
   - data made available by an online retailer
     $\rightarrow$ Non-disclosure agreement [NDA]
2. Explain the launch decisions of new products
   - Sales and price data: bought from company A
   - Advertising data: bought from company B
   - Quality data: manually obtained from websites
   - (Additional) launch data: manually obtained from websites

Data sources

My data comes from many different types of sources

Concerning the previous examples:

1. Online product recommendations
   - data made available by an online retailer
     $\rightarrow$ Non-disclosure agreement [NDA]
2. Explain the launch decisions of new products
   - Sales and price data: bought from company A
   - Advertising data: bought from company B
   - Quality data: manually obtained from websites
   - (Additional) launch data: manually obtained from websites
3. Using purchase histories to identify customer "projects"
   - Cooperation in the Wharton Customer Analytics Initiative
     $\rightarrow$ Non-disclosure agreement

Properties of the data

- Confidential data
- Non-disclosure applies
- Bought-in data
- Multiple sources
- Sometimes large datasets
- Need to apply complex modeling to the data

Current RDM strategies

My current practice

- No formal RDM strategy
  - lack of tools and clarity on ownership/data protection
    $\rightarrow$EUR is currently working on both

Current RDM strategies

My current practice

- No formal RDM strategy
  - lack of tools and clarity on ownership/data protection
    $\rightarrow$EUR is currently working on both
- During the research
  - One of the project members keeps the data (including backups)
  - If necessary data/code/paper is mailed back and forth
  - .. sometimes Dropbox is used
  - .. for some projects we use version-management software that allows for easy cooperation (git on the GitHub platform)

Current RDM strategies

My current practice

- No formal RDM strategy
  - lack of tools and clarity on ownership/data protection
    $\rightarrow$EUR is currently working on both
- During the research
  - One of the project members keeps the data (including backups)
  - If necessary data/code/paper is mailed back and forth
  - .. sometimes Dropbox is used
  - .. for some projects we use version-management software that allows for easy cooperation (git on the GitHub platform)
- After the research
  - Paper is shared through reprint series
  - Data/code is stored on own computer
    - Does not breach non-disclosure (NDA), but
    - sometimes difficult to retrieve data later
    - some NDAs require deletion of data after a certain period

Goals of RDM (in my opinion)

RDM should:

1. simplify the research process
   - make cooperation between researchers easier
   - automatically/easily track important steps/decisions in the process
2. support a professional way of doing research
3. secure long-term storage (after the project)
   - as a service to the researchers
   - necessary in case of suspected misconduct
4. (make sharing data easier)

Short-term needs and wishes

Researchers currently need:

1. IT support
   - make Dropbox obsolete
   - support version management
   - support backups of large data sets
2. support to develop a RDM strategy
   - grant suppliers demand a RDM strategy
3. insight and clarity concerning legal issues
   - who owns the data?
   - how to stay within Non-disclosure agreements?
   - privacy regulation?
   - how to keep data safe?
   - who is responsible for breaches of security?