

## Research Computing

---

PhD Student Orientation

Patricia Ledesma Liébana  
September 13, 2006

## Introductions

---

- Patricia Ledesma
- Alain Bonacossa
- Damba Lkhagvasuren

## Agenda

---

- Overview of resources
- Data
- Behavioral research
- Software
- Operating systems
  - Brief demo
- Tips on doing empirical research
- RC "library" resource room
- Full text databases
- Tips on bibliography management, technical word processing

09/13/2006

3

Empirical Dissertations and Datasets Used \*

Department	<u>Empirical theses</u>			<u>Of the empirical dissertations:</u>	
	<u># of theses found</u>	<u>#</u>	<u>%</u>	<u>% using secondary data</u>	<u>Data sources most frequently cited (number of mentions)</u>
Marketing	32	31	96.9	32.3	Anderson & Narus (2), IRI (1), ACNielsen (2)
Management & Organizations	26	26	100.0	46.2	SEC (2), SDC (2), CRSP (1), Compustat (1)
Accounting Information & Management	12	10	83.3	100.0	CRSP (5), Compustat (5), SEC (2), I/B/E/S (1), Chicago Fed bank data (2)
Finance	21	17	81.0	100.0	CRSP (5), Compustat (7), Datastream (4), TORQ (2), DRI (2), I/B/E/S (2), TAQ (1), SDC (2)
Managerial Economics & Decision Sciences	30	12	40.0	83.3	IRI (1), Compustat (1), CRSP (1), SDC (1), ACNielsen (1)
	<b>121</b>	<b>96</b>	<b>79.3</b>	<b>28.1</b>	

\* Includes only theses found in the NU library and in the UMI Dissertations Abstracts database

09/13/2006

4

## Where's the data?

---

### What did other researchers in your area use?

- Kellogg holdings
- NU holdings (ICPSR, Roper)
- Government agencies, international organizations
- Surveys or experiments
- Collect your own data from printed publications
- Datasets from other researchers
- Data collected by companies
- Department/school funding
  - Cost sharing for school wide access
  - Talk to your adviser / department chair first

09/13/2006

5

## Key corporate data holdings

---

- Accounting data: Compustat (Global, North America), Datastream, Global Access, Economatica
- Stock market: CRSP daily/monthly stocks, Datastream, TAQ, ISSM
- Bonds: FISD/Mergent, CRSP US Treasuries, Moody's DRS
- Derivatives: OptionMetrics, R&C Futures, Datastream
- Mutual funds: CRSP Mutual Fund database, Thomson Financial Mutual Fund Holdings, MFlinks
- Analyst forecasts: Zacks, First Call, ValueLine, InvestText
- Corporate governance: IRRC (Governance, Directors, Dilution), Execucomp, Thomson Financial Insiders ("Lancer Analytics"), Hemscott Executive Compensation
- Financial transactions (mergers, acquisitions, initial public offerings, seasoned equity offerings, bankruptcies): SDC Platinum, most dates available in CRSP events files.
- Ownership: Thomson Financial 13F filings, Mutual Fund Holdings, Insiders, IRRC Directors, ExecuComp, Hemscott, Blockholders

09/13/2006

6

## Primary data collection

---

- Surveys
- Experiments
- Games
- Negotiations

09/13/2006

7

## What package do I use?

---

**Be flexible.**

**Invest the time to learn.**

- Learning more than one package is unavoidable
- Willingness to run existing programs/routines in a different package can save you time (do not reinvent the wheel)
- Some datasets will require a specific program (e.g., TAQ)

09/13/2006

8

## Software categories

---

- Statistical/mathematical
- Terminal emulation – to access UNIX and Linux servers
- Programming languages and compilers
- Text editors
- Utilities
- Word processing / typesetting

09/13/2006

9

## Operating systems

---

- Microsoft Windows versus UNIX or Linux
  - UNIX (Sun Solaris)
    - skew3 (Kellogg server)
    - wrds (Wharton Research Data Services) – data retrieval
  - SSCC (Academic Technologies cluster) - Linux
- **Why UNIX/Linux?**
  - Software availability
  - Data availability
  - Powerful system (more RAM than in any PC, more processors)
  - Work from anywhere else without moving your files
  - While your programs run, you can continue to work in your PC

09/13/2006

10

## Beyond the netid

---

- "skew" account
  - Point your browser to <http://skew3.kellogg.northwestern.edu> and fill form to apply for account
  - Connect via SSH or X-Win32 to access UNIX account
- "wrds" account -- Wharton Research Data Services server
  - Point your browser to <http://wrds.wharton.upenn.edu> and use web interface
  - Connect via SSH to access UNIX account

09/13/2006

11

## Beyond the netid (cont.)

---

- "SSCC" account
  - Point your browser to <http://sscc.northwestern.edu> and fill the form to obtain an account
  - Connect via SSH or X-Win32 to access this Linux account

09/13/2006

12

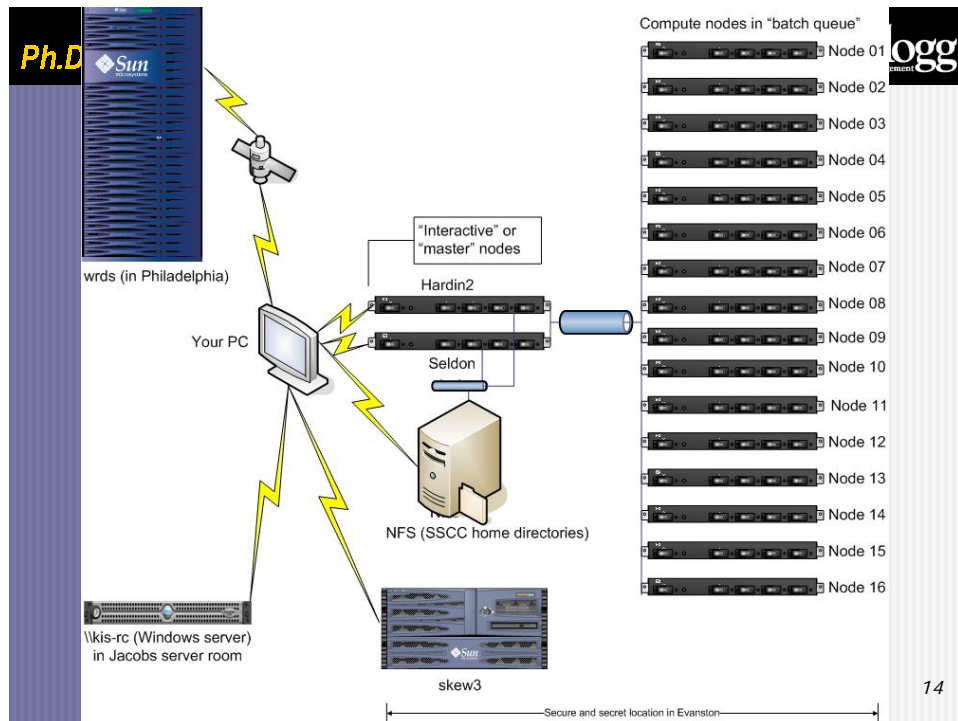
## How do you connect?

- Two site licensed software packages (free download for Northwestern members):
- **SSH Secure Shell**
  - Encrypted communication
  - Very efficient (does not use a lot of resources)
  - Text-mode
- **Starnet X-Win32**
  - Allows graphical interface emulation
  - Requires VPN from off-campus
  - Can be slow due to rendering of graphics

[www.kellogg.northwestern.edu/researchcomputing/terminal-emulators.htm](http://www.kellogg.northwestern.edu/researchcomputing/terminal-emulators.htm)

09/13/2006

13



14

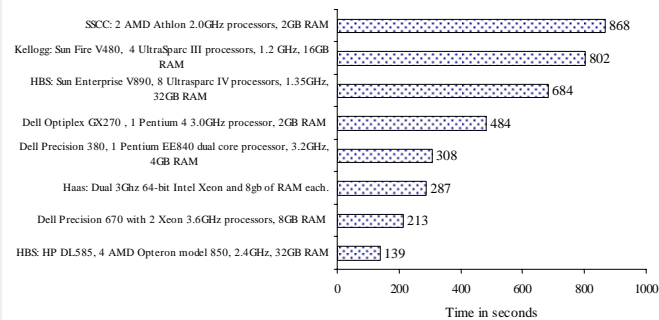
## skew3, SSCC or wrds?

- Both the WRDS server and skew3 are shared resource environment
  - Use WRDS for data retrieval from datasets stored there. 32 CPUs, but users from 150 universities!
  - Use skew for your actual computing, especially large jobs (with RAM requirements beyond the capacity of a PC or "I/O intensive" work)
- The SSCC is setup more for jobs that will run for a long time, for multiple specifications or parameter values.

09/13/2006

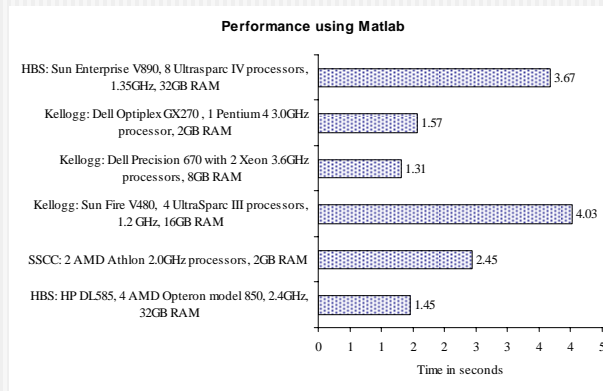
15

Performance using Stata



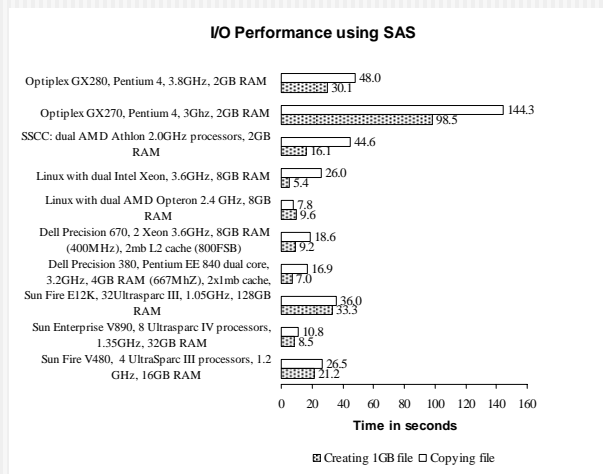
09/13/2006

16



09/13/2006

17



09/13/2006

18

## Behavioral research facilities

---

- Subject pool
- 2 laboratories for computer based experiments (3<sup>rd</sup> and 4<sup>th</sup> floors)
- 6 break out rooms (3<sup>rd</sup> floor) with video cameras
- Audio/video equipment
- Web surveys: Cogix ViewsFlash (hosted by Kellogg), Qualtrics (hosted by vendor)  
[www.kellogg.northwestern.edu/kis/websurveys/](http://www.kellogg.northwestern.edu/kis/websurveys/)
- Software: MediaLab, DirectRT, z-tree, Multistage. Custom programs in languages such as C, java, or perl.

09/13/2006

19

## Behavioral labs

---



09/13/2006

20

## Behavioral research

---

- Institutional Review Board (IRB):  
[www.research.northwestern.edu/research/OPRS/irb/](http://www.research.northwestern.edu/research/OPRS/irb/)
- Get IRB certification by taking the CITI BASIC course: [www.citiprogram.org/](http://www.citiprogram.org/)
- When submitting a research proposal to the IRB...
  - Be certified!
  - Fill all the fields required
  - Check whether you qualify for an expedited review
  - Make sure you have a faculty as Principal Investigator

09/13/2006

21

## Good empirical research practices

---

An empirical research paper is a 2-4 year project, from first draft to publication, with countless revisions (committee suggestions, brown bags, seminars, conferences, referee reports, ...). Hence:

- Become intimately familiar with your data
- Avoid shortcuts:
  - learn to code properly – don't modify your data by hand, don't hard code numeric results
  - learn to do the data extraction (avoid WRDS web interface)
- Create as few intermediate datasets and programs as possible

09/13/2006

22

## Good empirical research practices (cont.)

---

- Test data modifications on a “representative” sample (all data scenarios), not your main data set
- Document, document, document:
  - Source of each series, how it was modified
  - Include detailed comments in your programs
  - Create a document in which you have a flow chart (what program creates what files, what inputs are needed), a summary description of your files, a list of things to do
- After each revision, do some house-keeping to eliminate unnecessary files and update your documents.

09/13/2006

23

## If you are doing simulations:

---

- Test and benchmark every portion of your code. If you try it on a different operating system (Windows/Linux/Unix), test again.
- Learn to use the appropriate tools.
  - *Example:* A 2003 MECS graduate found that using Fortran 90 with some IMSL library routines increased speed in an optimization by nearly **100 times** (8 minutes versus more than 12 hours) versus using Matlab.
  - Differences in the version of a package can be significant (Matlab 6 versus 7).

09/13/2006

24

## Resource ("library") room – Jacobs 4219

---

- Documentation (user guides, related books) about the statistical / mathematical software and datasets available at the school
- One week borrowing – renewable unless someone is waiting for the title.
- Please remember to sign out and return titles – we rely on good citizenship.
- Suggestions for titles are welcome.

09/13/2006

25

## Full text data sources

---

- Many journals online
- Search NUCat for links
- Some resources omitted from library catalog
- Databases you should look into:
  - Abstracts/citations: EconLit, PsychInfo, Sociological Abstracts, Social Science Citation Index (Web of Science)
  - Working papers: subscribe to SSRN (<http://www.kellogg.northwestern.edu/researchcomputing/ssrn.htm>)

09/13/2006

26

## Bibliography management and technical word processing

---

- Learn how to use software...
  - In MS Word, use and modify styles to be able to create tables of contents and work with outlines
  - In Scientific Word/WorkPlace or MikTeX, learn to use BibTeX
- Invest in a good reference book
- EndNote is available for free through a library license

09/13/2006

27

## What is LaTeX?

---

### LaTeX file

```

\begin{document}
\title{The Title of a Standard LaTeX Article}
\author{A. U. Thor \\\%EndName
The University of Stewart Island}
\maketitle
\begin{abstract}
We study the effects of warm water on the local
major finding is that it is extremely difficult
warm water. The success factor is approximately
\end{abstract}
\section{The Chicago Bibliography System}
\subsection{About This Shell}
This shell document provides a sample layout of
The Typeset Bibliography Choice has been set to
package has been added.
Changes to the typeset format of this shell and
formatting files (\TeX\files\article.cls, \TeX\files

```

### BibTeX file

```

@techreport{Aiyer97,
Author = {Aiyer, Sri-Ram},
Title = {Pension Reform in Latin
Institution = {1865},
Year = {1997} }

@book{akyuz93,
Editor = {Akyüz, Yilmaz and Held,
Title = {Finance and the Real Eco:
Countries.},
Publisher = {United Nations Unive:
Address = {Santiago},
Year = {1993} }

```

09/13/2006

28

## Chicago style TeX citations

---

(Author, 1990)	<code>\cite{label}</code>
(Author 1990, chapter 1)	<code>\cite[Chapter~1]{label}</code>
Author1, Author2 (1990)	<code>\citeN{label}</code>
(Author1 et al, 1990)	<code>\shortcite{label}</code>
Author1 et al (1990)	<code>\shortciteN{label}</code>
(1990)	<code>\citeyear{label}</code>
1990	<code>\citeyearNP{label}</code>

More information about LaTeX and Scientific  
WorkPlace/Word:

[www.kellogg.northwestern.edu/researchcomputing/tex.htm](http://www.kellogg.northwestern.edu/researchcomputing/tex.htm)

[www.kellogg.northwestern.edu/researchcomputing/sciword.htm](http://www.kellogg.northwestern.edu/researchcomputing/sciword.htm)

09/13/2006

29

## Miscellaneous 1

---

- Accent modification clinic at the Evanston Speech and Language Clinic:  
[www.communication.northwestern.edu/csd/clinics/eslc/](http://www.communication.northwestern.edu/csd/clinics/eslc/)
- Intellectual integrity – refer to Northwestern's University Senate document "How to avoid plagiarism":  
[www.northwestern.edu/uacc/plagiar.html](http://www.northwestern.edu/uacc/plagiar.html)

09/13/2006

30

## Miscellaneous 2: NU Library staff (email @northwestern.edu)

---

- Leslie Bjorncrantz, Bibliographer for Management and Psychology (email: l-bjorncrantz)
- Jeannette Moss, Reference Librarian / Instruction Librarian (email: j-moss)
- Jami Xu, Reference Librarian & Liaison to the Kellogg School of Management (email: jamixu)
- Kathleen Murphy, Social Science Data Services Librarian (email: kemurphy)
- Other bibliographers: Harriet Lightman (econ, soc), Scott Garton (anthrop), Lucy Lyons (poli sci). Other areas: [www.library.northwestern.edu/collections/selectors.html](http://www.library.northwestern.edu/collections/selectors.html)
- Schaffner reference librarians: Julie Borden (Head, j-borden, 3-0720), Carol Doyle (c-doyle, 3-6618), Qiana Johnson (q-johnson, 3-6617)

09/13/2006

31

## Takeaways

---

- There is a lot of information online: [www.kellogg.northwestern.edu/researchcomputing](http://www.kellogg.northwestern.edu/researchcomputing)
- Pick good habits now.
- Where are we?  
 Alain: room 4242, a-bonacossa@, 7-1854  
 Damba: room 4222, d-lkhagva@, 7-4070  
 Patricia: room 4219, pledesma@, 7-7658

09/13/2006

32