Introductions

Kellogg research support team:
- Alain Bonacossa
- Andrei Jimyi
- Simone Cavallaro
- Tomek Wisniewksi

And our trusty web site:

www.kellogg.northwestern.edu/researchcomputing/
What do we do?

- Administer school wide research resources
- Help faculty and doctoral students in search of data, taking a look at buggy programs...
- Provide training for faculty and doctoral students
- Assist faculty in the creation of training material for MBA students that uses research data or quantitative tools
- Take on research-related projects from the Dean's Office, e.g. *Kellogg Insight*

Agenda

- Overview of resources
- Data
- Behavioral research
- Software
- Operating systems
  - Brief demo
- Tips on doing empirical research
- Full text databases
- Tips on bibliography management, technical word processing
Getting 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
Key corporate data holdings

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

Getting new data sources

- Cost sharing approach
- Faculty research budgets
- Department and research center funds
- Dean’s Office will match for school wide licenses, but not individual licenses
  - Talk to your faculty adviser, researchers with similar interests
  - Approach Office of Research
  - Why is this dataset interesting (especially if we have a similar resource)? Have a description at hand.
  - Have time horizons in mind: how long is the data going to be needed, are periodic updates needed.
  - Keep in mind time horizons – acquiring new data takes time (funding arrangements, legal review of the contract)
Primary data collection

- Surveys
- Experiments
- Games
- Negotiations
- Unstructured data: field / interview notes, corporate publications, video / audio recordings, transcripts

What package should I use? What OS?

Be flexible.
Invest the time to learn.
Do not handicap yourself.

- 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)
Software categories

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

Operating systems

Microsoft Windows versus UNIX or Linux

- Linux: skew5 (Kellogg server), SSCC (NU Academic Technologies / Kellogg / Weinberg, and Quest (all Northwestern)
- UNIX: wrds (Wharton Research Data Services) – data retrieval
- Windows: sheridan (behavioral research, Kellogg / Weinberg), future application server attached to data warehouse

Why UNIX/Linux?

- Software availability
- Data availability
- Powerful system
- Work from anywhere else without moving your files
- While your programs run, you can work in your PC
### Servers

<table>
<thead>
<tr>
<th>Server</th>
<th><em>Audience</em></th>
<th>Space</th>
<th>Software</th>
<th>Type of Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>server-5</td>
<td>Kellogg faculty and doctoral students</td>
<td>Red Hat Linux, 4 x six core Intel Xeon E5640, 3.07GHz, 256GB of RAM, 10TB of storage</td>
<td>Matlab, Mathematica, Maple, SAS, Stata, R, SPSS, TIF, R, Level Up, Ruffin, Polaris, C++, compilers, COLENS, KINETRO</td>
<td>General-use jobs that demand RAM or &quot;boutique&quot; applications, jobs with specific deadlines (Beowulf, Zeeko, Historical Files)</td>
</tr>
<tr>
<td>SSCC (Social Science Computing Cluster)</td>
<td>Kellogg and Weinberg faculty and doctoral students</td>
<td>Red Hat Linux, 260 AMD Opteron CPU cores, 2.80GHz, 40GB of RAM per CPU, 14TB of storage</td>
<td>Matlab, SAS, Stata, R, SPSS, R, Level Up, Ruffin, Polaris, C++, compilers, KINETRO, Lectra, HLM</td>
<td>Simulation/optimization, multiple jobs, parallel computing (user limited to 20 cores), &quot;boutique&quot; applications</td>
</tr>
<tr>
<td>Quest</td>
<td>Northwestern faculty and doctoral students</td>
<td>Red Hat Linux, 4072 cores Intel Xeon E5520 (2.26GHz, 60GB of RAM per core) and 10144 Intel Westmere E5509 (2.66GHz, 50GB of RAM per core), 100 TB of storage</td>
<td>Matlab, Mathematica, R, KINETRO, OpenMPI, Fortran, C++, Home erect</td>
<td>Simulation/optimization, parallel computing, computationally intensive jobs</td>
</tr>
<tr>
<td>Sheridan Cluster</td>
<td>Kellogg and Weinberg behavioral faculty and doctoral students</td>
<td>We install 64-bit/64-bit Linux, up to 15 simultaneous users, clients get virtual terminal with 2 cores (2.3 GHz) and 60GB of RAM</td>
<td>Atlas, Ames, HLM, Jipes, Jopes, R, Shure, SEDS, Uxnet, Win32, Ruffin, Polaris</td>
<td>Work that requires special applications, not much RAM, &quot;boutique&quot; applications</td>
</tr>
<tr>
<td>app01</td>
<td>Application server +  data warehouse</td>
<td>Kellogg faculty and doctoral students</td>
<td>Application server: Win 7 64-bit, 8G, Applications server: Matlab, SAS, core, 2 Kirkwood/E52, 256GB of DataNFS12 iSQL data warehouse: 32G iSQL data warehouse: 32GB</td>
<td>Data-intensive jobs, jobs that require specific datasets, like CME data, jobs that demand RAM, some types of computationally intensive jobs</td>
</tr>
</tbody>
</table>

### Getting accounts

<table>
<thead>
<tr>
<th>Server</th>
<th>How do I get an account?</th>
<th>How do I access it?</th>
</tr>
</thead>
<tbody>
<tr>
<td>server-5</td>
<td>Account request form <a href="http://server-5.kellogg.northwestern.edu">http://server-5.kellogg.northwestern.edu</a></td>
<td>PaTY, X-Windo2</td>
</tr>
<tr>
<td>SSCC (Social Science Computing Cluster)</td>
<td>Account request form <a href="http://www.k.northwestern.edu/research/ssccform.html">http://www.k.northwestern.edu/research/ssccform.html</a></td>
<td>PaTY, X-Windo2</td>
</tr>
<tr>
<td>Quest</td>
<td>Fill form at <a href="http://www.it.northwestern.edu/research/abv-research/bps/publicquestaccount.html">http://www.it.northwestern.edu/research/abv-research/bps/publicquestaccount.html</a> Requires project proposal, allocation amount (processor-hours), description of computational activities</td>
<td>PaTY, X-Windo2</td>
</tr>
<tr>
<td>Sheridan Cluster</td>
<td>To apply for an account, email <a href="mailto:SheridanCluster@kellogg.northwestern.edu">SheridanCluster@kellogg.northwestern.edu</a> with your name, department affiliation and title</td>
<td>Citrix client</td>
</tr>
<tr>
<td>app01</td>
<td>&quot;Data Warehouse&quot; Email Tomasz Wiesnicki, Application server + data warehouse (in production by summer 2012)</td>
<td>Microsoft Remote Desktop Connection</td>
</tr>
</tbody>
</table>
Servers

“skew” account
- Point your browser to http://skew5.kellogg.northwestern.edu and fill form to apply for account
- Connect via PuTTY (ssh in Mac) or X-Win32 to access Linux account

“wrds” account -- Wharton Research Data Services server
- Point your browser to wrds.wharton.upenn.edu and use web interface
- Connect via PuTTY (ssh in Mac) to access UNIX account
- You can run up to three concurrent jobs
- In addition, there is cluster setup to work with TAQ (the “grid”)

Servers (cont.)

“SSCC” account
- Point your browser to http://sscc.northwestern.edu and fill the form to obtain an account
- Connect via PuTTY (ssh in Mac) or X-Win32 to access this Linux account
- Run up to 48 concurrent jobs
Servers (cont.)

Quest accounts

- Best for numerical problems that can be parallelized
- Point your web browser to [http://www.it.northwestern.edu/research/adv-research/hpc/quest/](http://www.it.northwestern.edu/research/adv-research/hpc/quest/) to request an account and allocation of computing time. Do this a quarter in advance
- Number 236 in the Top 500 (June 2010)

Account types:

- Testing (up to 2,500 CPU/hr): no application required
- Development (up to 25,000 CPU/hr): must be requested 1 month in advance
- Research (25,000+): must be requested a quarter in advance

Windows servers

- `\ksmr01` (for data with Windows interfaces: SDC Platinum, Economatica, EMDB and Datastream; no computation)
- Sheridan cluster, restricted to behavioral researchers in Kellogg (MORS and Marketing) and in Weinberg. For an account and instructions, point to [http://www.wcas.northwestern.edu/sheridan/](http://www.wcas.northwestern.edu/sheridan/)
- `\kap01` ("Data Warehouse") for very large datasets
How do you connect to UNIX and LINUX servers?

Two site licensed software packages (free download for Northwestern members) for Windows:

**PuTTY or SSH**
- 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)

**What about the Mac?** Mac OS X is a UNIX operating system, based on NeXTSTEP. SSH and X11 are part of the operating system. Find the “Terminal” utility and use the ssh command:

```
ssh -Y netid@skew4.kellogg.northwestern.edu
```

On newer Macs, you may need to install Xquartz ([http://xquartz.macosforge.org](http://xquartz.macosforge.org)) as they are shipped without X by default (X allows you to see the GUI in programs)

---

**Behavioral research**

- 3 subject pools
  - Two undergraduate subject pools: one for MORS and one for Marketing and Behavioral Econ/Finance
  - eLab (non Northwestern population; online studies only)
- 2 laboratories for computer based experiments
  - 3rd and 4th floors
- 6 break out rooms (3rd floor) with video cameras
- Audio/video equipment
- Web surveys: Qualtrics (hosted by vendor), Cogix ViewsFlash (hosted by Kellogg)

[www.kellogg.northwestern.edu/kis/websurveys/](http://www.kellogg.northwestern.edu/kis/websurveys/)

- Kellogg Atlantis
- Software: MediaLab, DirectRT, z-tree, Multistage. Custom programs in languages such as C, java, or perl.
Subject pools

- Split in student pools reflects methodological concerns about deception in payment. The Mktg & Behavioral Econ pool does not allow any kind of deception in payment.
- Some overlap between the student pools, since participants cannot be prevented from signing up for both.
- Joint recruiting and split of subjects at the beginning of the school year.
- Use of pools is for Kellogg faculty, post docs and PhD students only – no external colleagues!

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Active</th>
</tr>
</thead>
<tbody>
<tr>
<td>MORS</td>
<td>4,000</td>
</tr>
<tr>
<td>Marketing and Behavioral Economics &amp; Finance</td>
<td>3,500</td>
</tr>
<tr>
<td>eLab (non NU, online studies only)</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Behavioral labs
• Collection of ready-to-use (MediaLab and Qualtrics) measures in personality/social psychology and marketing for online studies.

• [http://www.kellogg.northwestern.edu/rc/sec/faculty_and_phd-only/labs/atlantis.htm](http://www.kellogg.northwestern.edu/rc/sec/faculty_and_phd-only/labs/atlantis.htm)

<table>
<thead>
<tr>
<th>Scale</th>
<th>Citation</th>
<th>Description</th>
<th>MediaLab</th>
<th>Qualtrics*</th>
</tr>
</thead>
</table>

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

• IRB considerations when using Kellogg subject pools:
  • Do not collect **any identifiable information** from participants (e.g., emails, names etc.)
  • Use SONA anonymous ID codes to select raffle winners and for credit granting
  • Do not contact participants directly through SONA - messages should go through the lab manager
Plan, baby, plan!

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
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
    - For example, why did you drop certain observations?
  - 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.

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), and the operating system.
  - **Example:** The Matlab code of a current student could run 200 times faster by (a) re-writing the code eliminate nested loops and “vectorizing” (using Matlab matrix and vector operations); (b) outputting some data and running some Fortran code that returned results to Matlab.
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)

Bibliography management and technical word processing

- You will become a technical writer… learn how to use software your main tool:
  - Learn LaTeX (MikTeX or Scientific Word/WorkPlace), learn to use BibTeX
  - In MS Word, use and modify styles to be able to create tables of contents and work with outlines. Word 2007 also has a bibliography manager.
- Invest in a good reference book.
- EndNote is available for free through a library license (http://libguides.northwestern.edu/content.php?pid=26823&sid=194002).
LaTeX file

\begin{document}
\title{The Title of a Standard LaTeX Article}
\author{A. U. Thor}
\thanks{The University of Stewart Island}
\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 addedChanges to the typeset format of this shell are.
\end{document}

BibTeX file

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

\@book{akyuz93,  
Editor = {Akyuz, Yilmez and Held},  
Title = {Finance and the Real Econ Countries.},  
Publisher = {United Nations Univ. Address = {Santiago}. Year = {1993}}}

Chicago style \TeX citations

(Author, 1990) \cite{label}
(Author 1990, chapter 1) \cite{Chapter-1}{label}
Author1, Author2 (1990) \citeN{label}
Author1 et al, 1990) \shortcite{label}
Author1 et al (1990) \shortciteN{label}
(1990) \citeyear{label}
1990 \citeyearNP{label}

More information about \LaTeX and Scientific WorkPlace/Word:
   www.kellogg.northwestern.edu/researchcomputing/tex.htm
   www.kellogg.northwestern.edu/researchcomputing/sciword.htm
For Windows, many faculty use MikTeX with WinEDT or emacs; for the Mac, faculty recommend MacTeX with TeXShop
More tips

- As a PhD student, you will have more spare time than as a professor.
- **Attend all the seminars!**
- **Work as a research assistant**
- Take courses out of your comfort zone…
  - In economics, labor and industrial organization are ahead of the curve in empirical analysis.
  - In sociology, there are event history analysis courses.
  - Within Kellogg, courses on politics of the corporation and Empirical Corporate Finance are great resources.

Miscellaneous 1

- Accent modification clinic at the Evanston Speech and Language Clinic: http://www.communication.northwestern.edu/clinics/speech_language/services.php
- Intellectual integrity – refer to Northwestern’s University Senate document “How to avoid plagiarism”: www.northwestern.edu/uacc/plagiar.html
- Access to university resources is governed by license agreements and copyright laws:
  - Do not share software, do not download pirated material (use of the network is monitored for illegal activity)
  - Do not share proprietary data
  - Some references:
    - Appropriate use of electronic resources: http://www.it.northwestern.edu/policies/electronic-resources.html
Miscellaneous 2: NU Library staff

All emails in this slide are @northwestern.edu

- Jami Xu, Reference Librarian & Liaison to the Kellogg School of Management (email: jamixu)
- Carol Doyle, Reference Librarian and Business librarian (email: c-doyle4)
- Anne Zald, Social Science Data Services Librarian (email: anne.zald)
- Bibliographer for Management and Psychology (NU is conducting a search)
- Other bibliographers: John Hernandez (econ, soc), Scott Garton (anthrop), Lucy Lyons (poli sci). Other areas: www.library.northwestern.edu/collections/selectors.html

- There is a lot of information online: www.kellogg.northwestern.edu/researchcomputing
- Develop good habits now, learn as much as possible.
- Where are we?
  - Alain: room 4227, a-bonacossa@, 7-1854
  - Andrei: room 4222, a-jimyi@, 7-4070
  - Simone: room 370A, s-cavallaro@, 7-4625
  - Patricia: room 4219, pledesma@, 7-7658
  - Tomek: room 4226, t-wisnewski@, 7-4334