PRE-PROGRAM: INTRODUCTION TO EXCEL, STATA & DATA ANALYSIS

EMBA Data Analysis

Professor Timothy Simcoe
Boston University School of Management

Introduction: Management as Measurement

In the 1830's, France imported about 40 million leeches each year for medicinal purposes.*

In 1836, Pierre Louis ran one of the first randomized controlled experiments: treats pneumonia patients with either (a) early aggressive blood-letting, or (b) less aggressive measures.

"I was surprised to see that more than half our ideas failed to move the metrics they were designed to move. Humbling."

- Ronny Kohani, Amazon's director of personalization

"If you cannot measure it, you cannot improve it."

- Lord Kelvin

* Source: "Bad Medicine" David Wootton

Pre-program Objectives

Introduce key concepts and terminology

- Observation, Unit of analysis, Variable types

Introduce software: Excel and Stata

- Cleaning and loading data
- Creating variables
- Summary statistics
- Excel <=> Stata

Get to know each other!

When you ask for data, how does it arrive?

What is the file format?

- Excel (.xls), Access (.mdb), ASCII Text (.txt), Stata (.dta), etc…

How is the data "structured"?

- Tables, spreadsheet, relational DB, flat file, unstructured…
- How many files and/or tables?

Has the information been "cleaned"?

- Documentation available
- Checks for missing data and/or logical consistency

Spreadsheets vs. Databases

Spreadsheets

- Easy to enter and manipulate data by hand
- Lots of flexibility and low setup costs
- User is responsible for tracking relationships
- No automated consistency / quality checks

Databases

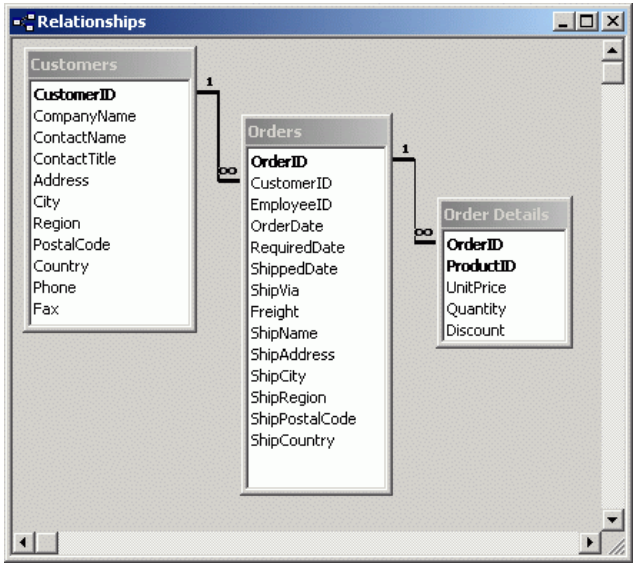
- Up-front cost to design tables and define links
- Specialized knowledge, e.g. Structured Query Language (SQL)
- Greater long-run efficiency in storage & retrieval
- Greater consistency: machine enforces design rules

Spreadsheets have low fixed costs, but high variable costs, compared to relational databases

Relational Databases (in one slide)

Shopping Cart Application

- Multiple Tables: Customer, Order & Item
  - Each table has “records” (rows) and “fields” or “variables” (columns)
  - Each records has a unique “key” or identifier (ID) variable
- Relational Links: ID variables that appear in multiple tables
  - Link types: 1-to-1, many-to-1 or Many-to-many



The "Big Data" Value Chain

Skills	Tools
Collect & Reduce Data	Hadoop / SQL
Mine & Analyze Data	Stata / R / Excel
Visualize & Present Results	Adobe CS / Stata / Office

These activities are complements, not substitutes

EMBA DA will emphasize Analysis & Presentation

Statistical analysis almost always begins with "rectangular" data

Rectangular data is analogous to a single table or worksheet

- Each row is called an <u>observation</u>
  - We will use letter "N" to denote total # of observations
- Each column contains a single <u>variable</u>
  - We will use letter "X" to denote a (generic) variable
  - More on variable types in a few minutes

Creating a rectangular dataset can be a huge amount of work

- Matching, merging, appending, cleaning, etc.
- But not in this class!

Group Exercise: Re-format AnnualSales.xls as rectangular data

What is the unit of analysis in AnnualSales.xls?

Each row is an observation

- In AnnualSales.xls we have store-year observations

The unit of analysis is the "object" that you are studying

- For AnnualSales.xls, it could be region, city, store or store-year
- It depends on the question being asked!

Common types of data sets

- Cross-section: A collection of things (observation == unit of analysis)
- Panel Data: Repeated observations (e.g. stores by year)
- Multi-level: Multiple panels (e.g. stores in cities in regions)

Variable Types

Computers recognize two variable types

- Numeric and String variables
- So, 5.7 is not the same as "5.7"
- Don't assume the computer "understands" numbers!

Statisticians recognize (roughly) three variable types

- Metric or continuous, e.g. height, weight, or sales
- Ordered categorical, e.g. Likert (1-5) scale or low-medium-high
- Unordered Categorical, e.g. gender, location

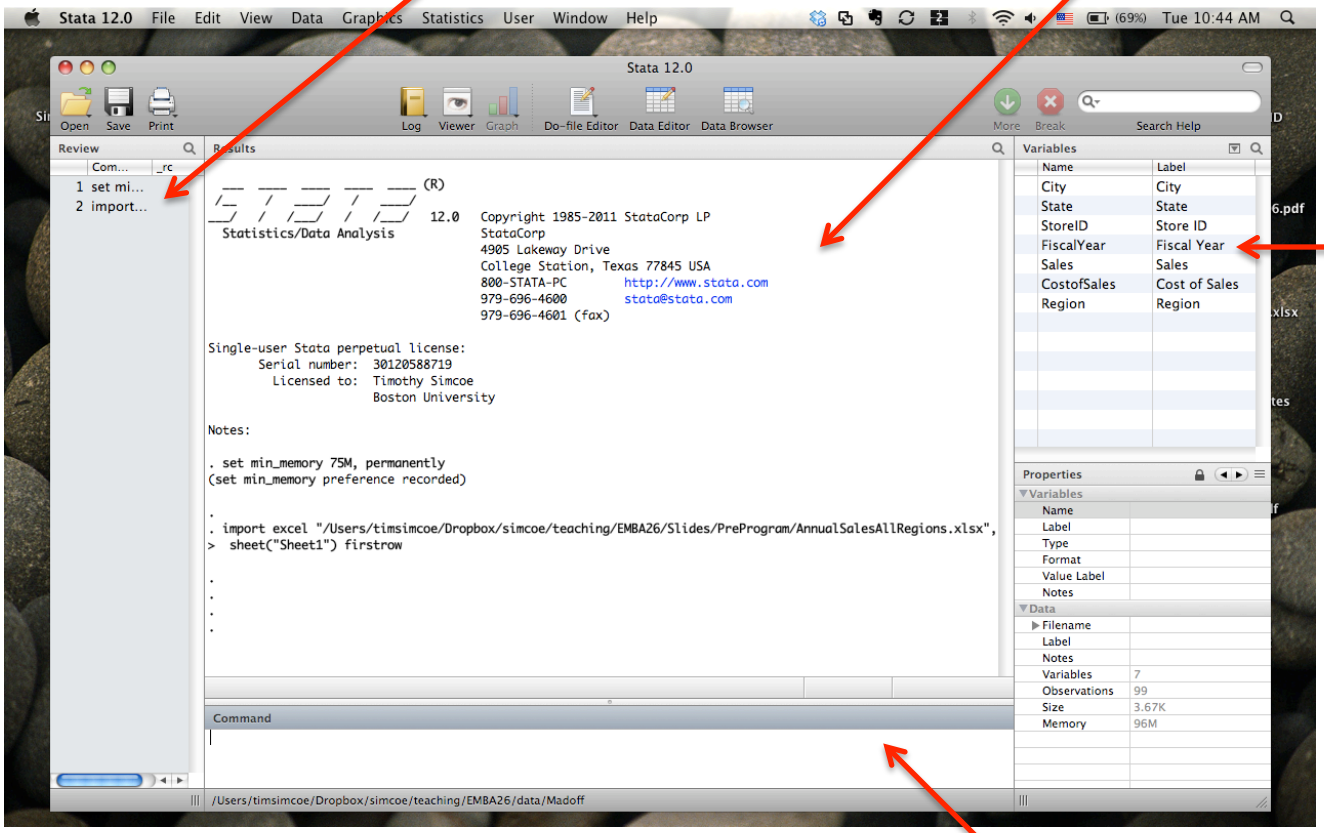
What are the types (computer and statistical) of each variable in AnnualSales.xls?
Examples of metric, ordered and unordered categorical data in your workplace?

Break and Pre-Assessment

Stata's Windows

Review window: A list of your last few commands click on them for a "do over"

Results Window: Where Stata displays output after you tell it to do something



Variables window: Useful info about the variables in current data set

Command window: Where you type in commands (unless you prefer to use the drop down menus)

Getting an Excel File into Stata

1. Make the data rectangular in Excel
2. Make sure numbers are formatted as “numbers”
3. Open Stata*
4. Select “File > Import > Excel Spreadsheet” from dropdown menus
5. Find the file you wish to load into Stata
6. *Choose the correct Excel Worksheet if application*
7. *Check “Import First Row as Variable Names” if applicable*
8. Make sure you have correct # of rows / observations
9. Click “OK”
10. To save in Stata format, select “File > Save as…” drop down

Practice by importing the “StoreManagers” tab from our AnnualSales.xls Spreadsheet and saving it as a Stata file

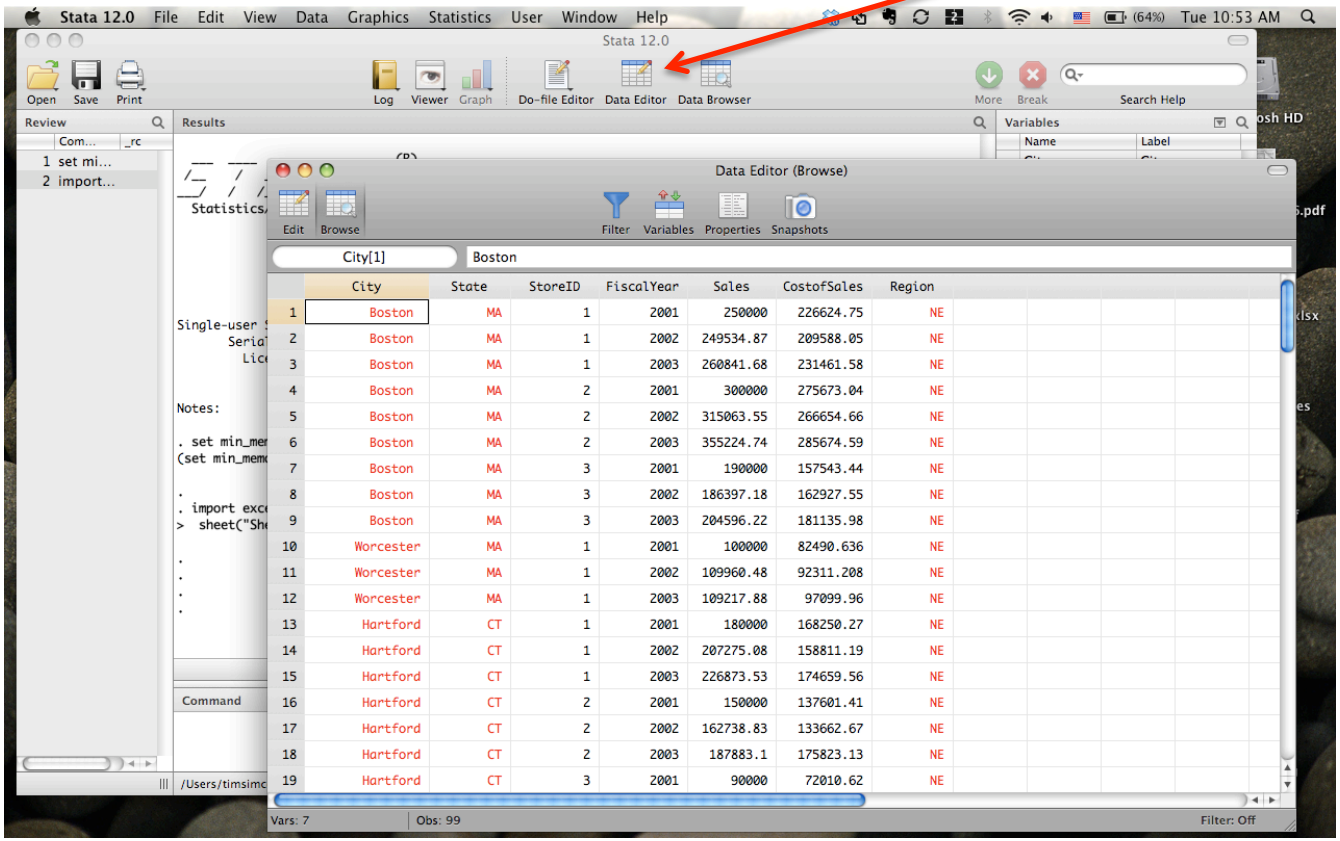
Stata also has a “spreadsheet” window

Data Editor allows editing (not recommended)

Data Browser only permits viewing data

Click here to view data in “spread-sheet” format

Note that **string** variables are displayed in **RED**, & **numeric** variables are displayed in **BLACK**



	City	State	StoreID	FiscalYear	Sales	CostofSales	Region
1	Boston	MA	1	2001	250000	226624.75	NE
2	Boston	MA	1	2002	249534.87	209588.05	NE
3	Boston	MA	1	2003	260841.68	231461.58	NE
4	Boston	MA	2	2001	300000	275673.04	NE
5	Boston	MA	2	2002	315063.55	266654.66	NE
6	Boston	MA	2	2003	355224.74	285674.59	NE
7	Boston	MA	3	2001	190000	157543.44	NE
8	Boston	MA	3	2002	186397.18	162927.55	NE
9	Boston	MA	3	2003	204596.22	181135.98	NE
10	Worcester	MA	1	2001	100000	82490.636	NE
11	Worcester	MA	1	2002	109960.48	92311.208	NE
12	Worcester	MA	1	2003	109217.88	97099.96	NE
13	Hartford	CT	1	2001	180000	168250.27	NE
14	Hartford	CT	1	2002	207275.08	158811.19	NE
15	Hartford	CT	1	2003	226873.53	174659.56	NE
16	Hartford	CT	2	2001	150000	137601.41	NE
17	Hartford	CT	2	2002	162738.83	133662.67	NE
18	Hartford	CT	2	2003	187883.1	175823.13	NE
19	Hartford	CT	3	2001	90000	72010.62	NE

Creating new variables

Problem: Calculate Gross Income and Gross Margin by store-year

Using Formulas in Excel

- Create a new column for Gross Income
- Enter the formula (Gross Income = Sales – Cost of Sales)
- Useful Trick: Absolute references ($) and cut & paste

Using Stata's "generate" command

- Type "generate GrossIncome = Sales – CostofSales" in command window
- Useful trick: Click next to variable name to paste in command window
- ***No spaces allowed in variable names***
- Can change variable names, e.g. "rename CostofSales COGS"

Summarizing a variable

What is the average Gross Margin for all store-years in our data?

Using Functions in Excel

- Create a new column containing GrossMargin = GrossIncome / Sales
- Use the function "= average([data range])" to compute an average
- Other functions: count([data]), stdev([data]), min([data]), max([data])

Using Stata's "summarize" command

- Type "summarize [variable name]" in command window
- For more information use "summarize [variable name], detail"

Making tables (for categorical data)

Problem: What is the average Gross Margin for each region by year?

Using Pivot Tables in Excel

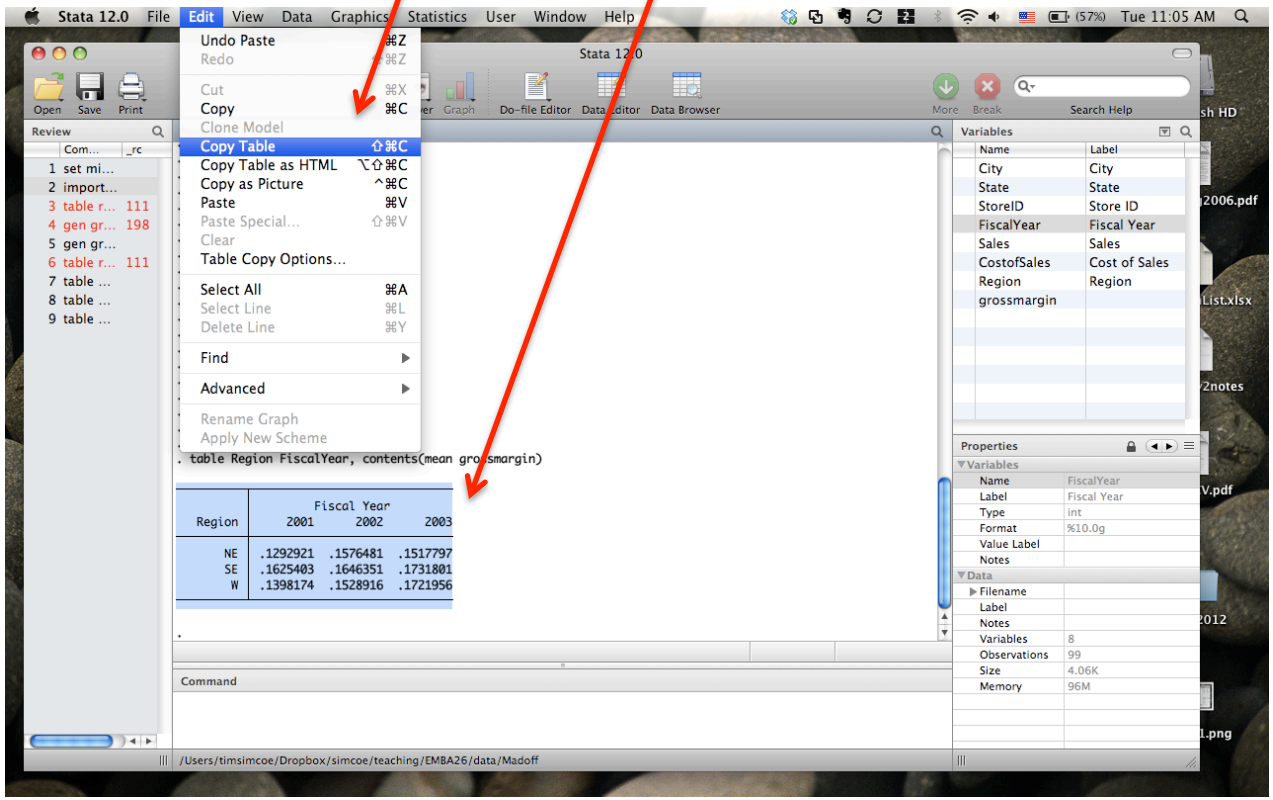
- ***Not required material, but you might find it useful***
- Select "Data > Pivot Table…" from drop down menus
- Use the Wizard Tool to build your table

Using Stata's "table" command

- Type "table Region FiscalYear, contents(mean GrossMargin)"
- OR use Stata's drop-down menus
  - "Statistics > Summaries, tables and tests > Tables > Table of summary statistics"

Copying Stata output into Excel

1. Select a Table in the Results Window
2. Go to “Edit > Copy Table”
3. Paste into Excel



Merging datasets

Forget about doing this in Excel

In Stata

1. Make sure two files correctly formatted (.dta)
2. Place files in same directory (for simplicity)
3. Make sure merge variable names are identical
4. Move to correct directory ("File > Change Working Directory…")
5. Type the following command

```
merge m:1 City State StoreID Region using managers
```

Merge Type
- m:1 = many-to-1
Could be 1:m or m:m

Merge Variable Names

Name of data set to merge

Stata vs. Excel: Frequently Asked Questions

Will this course teach me to use Excel?

- No. I assume you are have some basic Excel skills (e.g. arithmetic formulas, copy & paste) and will introduce advanced material in class as needed.

Why do we use STATA?

- Instead of Excel
  - Excel lacks tools for serious statistical analysis
  - Stata is <u>much easier to use</u> for any data analysis that goes beyond the basics
  - STATA will familiarize you with quantitative analysts' "tools of the trade"

- Instead of SPSS or SAS
  - My goal: "democratize" evidence-based decision-making; STATA is the perfect tool for this
  - PC-based (not designed for mainframe)
  - Extremely powerful (equal to SPSS or SAS)
  - Intuitive menu-driven interface
  - You can TAKE IT WITH YOU wherever you work and use it independent of firm IT infrastructure

Additional Stata Resources

The Instructor

- I am available to answer Stata questions any time
- I prefer to take these questions by email

Help Manuals

- Type “help” or “help [command name]” in Stata to see the manuals

Internet

- Google “Stata Help” for a wealth of information

SMG Tools

- I’ve posted a Tutorial, Cheat Sheet and Tips to a “Stata Help” folder

Reference Books

- “A Gentle Introduction to Stata”

Data Analysis: where are we going?

Problem: Should this chain expand in the Southeast?

1. Reformulate the question: How confident are we that higher margins in the Southeast are not random?
2. Build a "model" to test that question
   - Gross Margins = Year Effects + Region Effects + Store-level "noise"
3. Use the model to conduct a statistical test
   - "regress grossmargin i.fiscalyear i.regionCode"
4. Use results to make **informed business decisions**

Iterative process: each step informs the prior step
Statistical software speeds up steps #2 and #3 dramatically

A Quick Experiment

Naïve measurement can (and does) lead us astray

- Sampling on the outcome
  - "Airport book" consulting
  - "Survivor bias" in hedge/mutual fund returns
- Correlation versus causation
  - Study: nightlights produce myopia
  - You might like this 30" monitor because you purchased a display cable

We will develop tools and frameworks for thinking about *causality* when we observe relationships in the data

Another Tale of Caution

Dear Dr. Dally:

We appreciate the opportunity to provide comment regarding the center-specific statistics which you have provided to us for our heart/lung and lung transplant programs. We are a fledgling program, and you have statistics on one heart/lung and one single lung transplant patient, both of whom are alive and well after 12 months. Since our observed survival rate is 100%, we are fascinated by the UNOS calculated 12 month risk-adjusted survival rate of 55% for our heart/lung patient and 47% for our lung transplant recipient. This, of course, betrays the difficulties of providing meaningful statistics, particularly with low volume numbers. Both patients, we can assure you, are 100% alive and not 50% alive and well as the risk-adjusted survival statistics might imply to the consuming public. We would believe this provides an example of one of the weaknesses of extrapolated statistics.

Sincerely yours,

Barry L. Fields, M.D. Robert M. Mentzer, Jr., M.D.

Co-Directors of Cardiopulmonary Transplantation

We will develop tools and frameworks for thinking about *inference* (statements about precision of measured relationships)

BIG PICTURE: Information Technology has given firms the opportunity know much more…

About customers

- Credit card transactions
- Loyalty programs
- Payment history
- Behaviors & attitudes

About Employees

- Retention
- Performance
- Knowledge mgmt
- Comp & incentives

About operations

- SCM systems / RFID
- Mfg. process control

About Suppliers

- Bidding/pricing history
- On-time performance

How can firms use this information trail to create better products & services, make better decisions, and compete more effectively?

Problem: It's hard to make data-driven decisions

DIFFICULTIES WITH EVIDENCE-BASED MANAGEMENT

- Many organizations are not used to thinking this way
  - Managers don't prioritize data acquisition
  - Powerful constituencies feel threatened
  - Few incentives to gather / act-on data
- Data are costly to gather
  - New IT systems are expensive
  - Existing databases ill-suited and difficult to combine
- Requires specialized skills
  - Deep analysis can require sophisticated modeling
  - Managers are bad at analytics, statisticians can't manage…

In this course you learn to overcome these problems

Step 1: To become a sophisticated *proponent & consumer* of quantitative analyses

- To develop a solid conceptual grasp of statistical methods
  - e.g. random sampling, hypothesis testing, multiple regression
- To understand and articulate the power of controlled randomized experiments
  - correlation vs. causation, description vs. prediction
- To understand when analytical methods are appropriate and when they fail

Step 2: To apply *workhorse analytical methods* on real-world data sets

- To become familiar with the critical skills of a quantitative analyst
  - building, testing and refining an empirical model
  - managing large(ish) data sets
  - communicating key assumptions and results

- To develop an appreciation for the broad scope of simple statistical methods
  - applications to operations, marketing and finance

Step 3: To identify and address *organizational barriers* to evidence-based management

- To learn to identify opportunities for data analysis and/or systematic experimentation
- To effectively promote the evidence-based approach within a complex organization
- To become sensitive to ethical and security issues that arise in working with large databases

What you will have learned in these three steps enables you to implement "evidence-based management" in practice

Short Bio

Experience

- Five years in consulting, 6 months at CEA
- PhD from Berkeley (Business & Public Policy)
- Previously taught at U of Toronto (5 years)

Research interests

- Standardization (Technology & Quality)
- Innovation, particularly in ICT industries
- Intellectual property
- Applied Econometrics

Personal

- Married with 3 kids: Kate (10), Anne (8) & Teddy (3)
- Enjoy skiing, golf, running and Red Sox

Questions?