

Introduction to Financial Databases

Sep 18, 2006

By Roger Loh
Department of Finance
Fisher College of Business
Ohio State University

1

Overview

Importance of Data in Financial Research

Describe each Database

How to Access the Data?

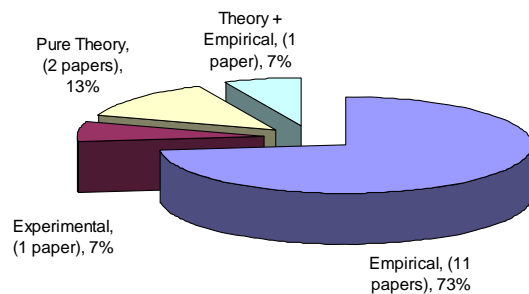
Practice sessions in Fisher 606

2

Importance of Data in Financial Research

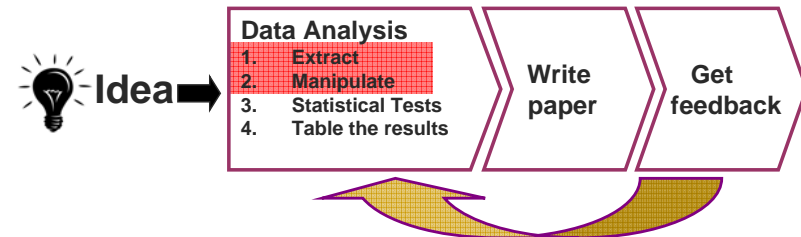
1. Hard to write a paper that doesn't use data.

- Eg., in the June 2006 issue of JF, 12 out of 15 papers (80%) use data.



3

Importance of Data in Financial Research

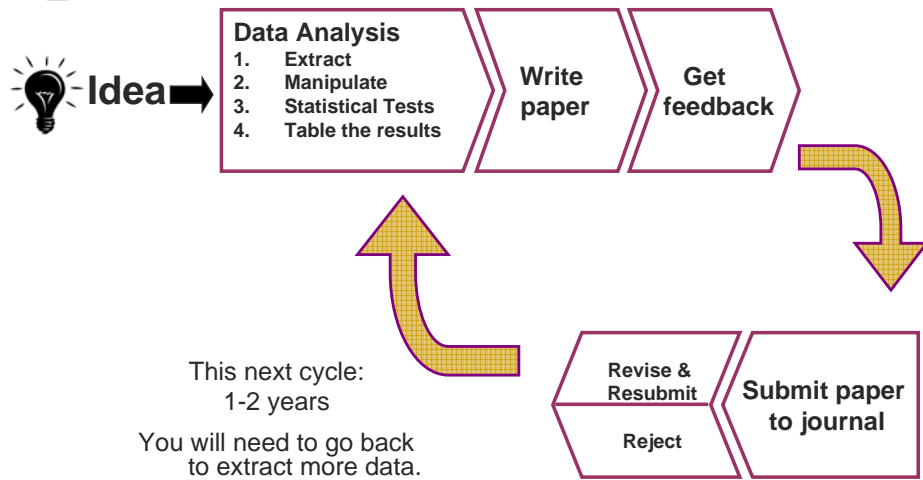


This cycle: 1-3 years

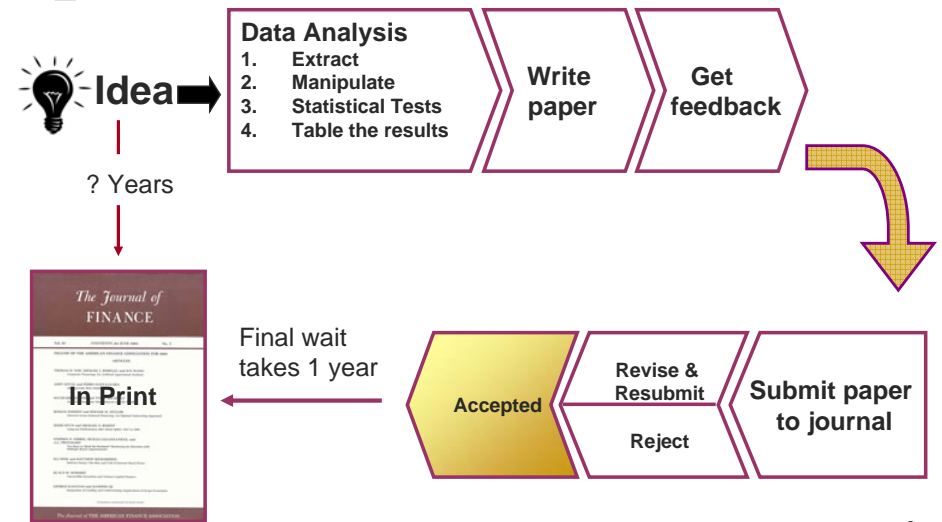
2. If you can access and analyze data quickly, you can shorten the length of this cycle. Writing systematic programs to access data is very important because you need to make repeated amendments to your data requirements.

4

Empirical Research Process



Empirical Research Process



Overview

Importance of Data in Financial Research

Describe each Database

How to Access the Data?

Practice sessions in Fisher 606

Databases by Concept

- Stock returns (CRSP, Datastream)
- Trading data (TAQ)
- Company data (Compustat, IRRIC, Worldscope, SDC)
- Analyst advice (I/B/E/S)
- Institutional holdings data (Thomson 13F)
- Others
 - Economic data (Datastream)
 - Author provided data (e.g. Ken French's website, Robert Shiller's website)
 - Proprietary data (Odean's individual investor database)

List of Databases

1. CRSP
2. Compustat
3. NYSE TAQ
4. I/B/E/S
5. Thomson 13F
6. IRRC
7. Datastream/Worldscop
8. SDC

9

1. CRSP

- Center for Research in Security Prices—most comprehensive US stock returns database.
- Individual stock returns and market returns (daily and monthly).
- Most used items:
 - Price
 - Return
 - Volume
 - Market-cap
 - Shares outstanding
 - SIC Industry code

10

CRSP Data E.g.

Permanent number is the CRSP firm identifier

PERMNO	DATE	COMNAM	PRC	VOL	RET
11081	20050131	DELL INC	41.7600	2676671	-0.00902
11081	20050228	DELL INC	40.0900	3443626	-0.03999
11081	20050331	DELL INC	38.4200	3153971	-0.04166
11081	20050429	DELL INC	34.8300	3768089	-0.09344
11081	20050531	DELL INC	39.9300	3670710	0.14643
11081	20050630	DELL INC	39.4600	3016043	-0.01177
11081	20050729	DELL INC	40.4700	2130263	0.02560
11081	20050831	DELL INC	35.6000	5044521	-0.12034
11081	20050930	DELL INC	34.2000	3611113	-0.03933
11081	20051031	DELL INC	31.8800	4273538	-0.06784
11081	20051130	DELL INC	30.1510	5894065	-0.05423
11081	20051230	DELL INC	29.9500	4104746	-0.00667
12490	20050131	INTERNATIONAL BUSINESS MACHS COR	93.4200	1148066	-0.05234
12490	20050228	INTERNATIONAL BUSINESS MACHS COR	92.5800	802581	-0.00706
12490	20050331	INTERNATIONAL BUSINESS MACHS COR	91.3800	1075088	-0.01296
12490	20050429	INTERNATIONAL BUSINESS MACHS COR	76.3800	2175501	-0.16415
12490	20050531	INTERNATIONAL BUSINESS MACHS COR	75.5500	1389057	-0.00825
12490	20050630	INTERNATIONAL BUSINESS MACHS COR	74.2000	1377932	-0.01787
12490	20050729	INTERNATIONAL BUSINESS MACHS COR	83.4600	1571802	0.12480
12490	20050831	INTERNATIONAL BUSINESS MACHS COR	80.6200	1050638	-0.03163
12490	20050930	INTERNATIONAL BUSINESS MACHS COR	80.2200	1156901	-0.00496
12490	20051031	INTERNATIONAL BUSINESS MACHS COR	81.8800	1413377	0.02069
12490	20051130	INTERNATIONAL BUSINESS MACHS COR	88.9000	1108741	0.08818
12490	20051230	INTERNATIONAL BUSINESS MACHS COR	82.2000	1203340	-0.07537

11

2. Compustat

- US Company Data from financial statements
- Most commonly used Files
 - Industrial Annual
 - Industrial Quarterly
 - CRSP-Compustat Merged File
 - Executive Compensation
- Commonly used items in Annual File (Item #):
 - Long-term Debt (#9)
 - Sales (#12)
 - Earnings (income before extraordinary items, #18)
 - Book value of equity (#60)
 - Total Assets
 - R&D expenditure
 - Cash

12

Compustat Annual Data E.g.

Gvkey is the Compustat firm identifier

GVKEY	yeara	SMBL	DATA9	DATA12	DATA18	DATA60
6066	1995	IBM	10060	71940	4178	22170.00
6066	1996	IBM	9872	75947	5429	21375.00
6066	1997	IBM	13696	78508	6093	19564.00
6066	1998	IBM	15508	81667	6328	19186.00
6066	1999	IBM	14124	87548	7712	20264.00
6066	2000	IBM	18371	88396	8093	20377.00
6066	2001	IBM	15963	85866	7723	23614.00
6066	2002	IBM	19986	81186	5334	22782.00
6066	2003	IBM	16986	89131	7613	27864.00
6066	2004	IBM	14828	96293	8448	29747.00
6066	2005	IBM	15425	91134	7994	33098.00
14489	1995	DELL	113	5296	272	973.00
14489	1996	DELL	18	7759	531	806.00
14489	1997	DELL	17	12327	944	1293.00
14489	1998	DELL	512	18243	1460	2321.00
14489	1999	DELL	508	25265	1666	5308.00
14489	2000	DELL	509	31888	2236	5622.00
14489	2001	DELL	520	31168	1246	4694.00
14489	2002	DELL	506	35404	2122	4873.00
14489	2003	DELL	505	41444	2645	6280.00
14489	2004	DELL	634	49205	3043	6485.00
14489	2005	DELL	559	55908	3572	4129.00

13

3. I/B/E/S

- Institutional Brokers Estimate System.
- Contains sell-side security analysts' earnings forecasts and stock recommendations for US firms and international firms.
- Earnings Forecasts
 - Summary File (monthly consensus forecasts)
 - Detail File (Individual analyst forecasts)
- Recommendations
 - Summary File (monthly consensus recommendations)
 - Detail File (Individual analyst recommendations)

14

I/B/E/S Detail Recommendations E.g.

OFFIC	TICKER	RECDATS	BROKER	AMASKCD	ANALYST	ITEXT
DELL	DELL	20060110	JYSKE	00073425	JACOBSEN	P BUY
DELL	DELL	20060110	FRIEDMAN	00080211	SUMNER	C HOLD
DELL	DELL	20060120	BAIRD	00053931	RENOUARD	D STRONG BUY
DELL	DELL	20060206	CARISCO	00001032	STAHLMAN	M HOLD
DELL	DELL	20060208	BERN	00058377	SACCONAGHI, JR T	STRONG BUY
DELL	DELL	20060217	MONTSEC	00070706	BACHMAN	K HOLD
DELL	DELL	20060219	NUTMEG	00000678	LABE	P HOLD
DELL	DELL	20060307	GRUMMAN	00065515	CHU	R STRONG BUY
DELL	DELL	20060421	SMITH	00047225	GARDNER	R SELL
DELL	DELL	20060430	NUTMEG	00000678	LABE	P BUY
DELL	DELL	20060505	GOLDMAN	00001176	CONIGLIARO	L HOLD
DELL	DELL	20060509	BAIRD	00053931	RENOUARD	D HOLD
DELL	DELL	20060509	DAKIN	00075117	SHAW	C SELL
DELL	DELL	20060519	FBOSTON	00085160	SEMPLE	R HOLD
DELL	DELL	20060519	NEEHDAM	00001047	WOLF	C HOLD
DELL	DELL	20060519	LEHMAN	00001924	BLOUNT	H HOLD
DELL	DELL	20060519	SMITH	00047225	GARDNER	R HOLD
DELL	DELL	20060519	FRIEDMAN	00080211	SUMNER	C STRONG BUY
DELL	DELL	20060522	ARGUS	00000938	ABRAMOWITZ	W HOLD
DELL	DELL	20060526	FGS	00111945	DALAL	N UNDERPERFORM
DELL	DELL	20060530	FIRSTALB	00114594	MAI	H HOLD
IBM	IBM	20060109	JFMORGAN	00072446	SHOPE	B HOLD
IBM	IBM	20060421	FGS	00111945	DALAL	N HOLD
IBM	IBM	20060505	GOLDMAN	00001176	CONIGLIARO	L BUY
IBM	IBM	20060601	CANACCOR	00072029	MISEK	P HOLD

15

4. NYSE Trade and Quote (TAQ)

- Consolidated Trades, Consolidated Quotes data.
- Used when intra-day price and quotes are needed, especially in market microstructure/liquidity research.
- Commonly used variables
 - Trade price
 - Trade size
 - Bid/offer price
 - Bid/offer size
- Managing the large size of the dataset is the challenge of using TAQ data.

16

TAQ Quotes Data E.g.

SYMBOL	DATE	TIME	BID	OFB	BIDSIZ	OFBSIZ
IBM	20050606	9:30:01	0.01	1000.00	1	1
IBM	20050607	9:30:01	0.01	1000.00	1	1
IBM	20050607	9:30:04	71.56	78.95	1	1
IBM	20050607	9:30:04	71.56	78.95	1	1
IBM	20050607	9:30:06	71.56	78.98	1	1
IBM	20050607	9:30:06	71.56	78.98	1	1
IBM	20050607	9:30:08	71.56	79.00	1	1
IBM	20050607	9:30:08	71.56	79.00	1	1
IBM	20050607	9:30:19	75.11	75.48	38	5
IBM	20050607	9:30:19	75.00	75.06	237	21
IBM	20050607	9:30:19	0.00	0.00	0	0
IBM	20050607	9:30:19	75.11	75.48	5	5
IBM	20050607	9:30:19	75.00	75.06	232	21
IBM	20050607	9:30:19	7.50	142.50	1	1
IBM	20050607	9:30:19	74.70	0.00	1	0
IBM	20050607	9:30:19	74.70	75.36	1	1
IBM	20050607	9:30:19	75.11	75.28	5	10
IBM	20050607	9:30:19	75.01	75.36	1	1
IBM	20050607	9:30:19	75.01	75.16	1	1
IBM	20050607	9:30:19	75.11	75.19	5	2
IBM	20050607	9:30:20	0.00	0.00	0	0
IBM	20050607	9:30:20	0.00	0.00	0	0
IBM	20050607	9:30:20	0.00	0.00	0	0
IBM	20050607	9:30:20	74.40	76.01	10	10
IBM	20050607	9:30:20	74.40	76.01	10	10
IBM	20050607	9:30:20	74.81	75.11	13	1
IBM	20050607	9:30:20	74.90	75.16	1	1

17

5. Thomson 13F

- Institutions are required to report their ownership of equities in quarterly 13F filings to the SEC.
- Aggregate holdings for the institution, regardless of the number of individual fund portfolios.
- Shows how many shares of a firm are held by each institution.
- Usage of this database among published papers increased recently.

18

13F Holdings Data E.g.

mgrno	rdate	mgrname	cusip	ticker	shares	shrouttr
110	30JUN2005	A R ASSET MANAGEMENT, INC.	24702R10	DELL	20250	2421
110	30SEP2005	A R ASSET MANAGEMENT, INC.	24702R10	DELL	20250	2397
110	31DEC2005	A R ASSET MANAGEMENT, INC.	24702R10	DELL	20250	2354
120	31MAR2005	AAL CAPITAL MANAGEMENT CORP.	24702R10	DELL	1442976	2459
120	30JUN2005	AAL CAPITAL MANAGEMENT CORP.	24702R10	DELL	1497626	2421
120	30SEP2005	AAL CAPITAL MANAGEMENT CORP.	24702R10	DELL	1383126	2397
120	31DEC2005	AAL CAPITAL MANAGEMENT CORP.	24702R10	DELL	1103726	2354
185	31MAR2005	ASB CAPITAL MANAGEMENT, INC.	24702R10	DELL	1972001	2459
185	30JUN2005	ASB CAPITAL MANAGEMENT, INC.	24702R10	DELL	1960261	2421
185	30SEP2005	ASB CAPITAL MANAGEMENT, INC.	24702R10	DELL	2191936	2397
185	31DEC2005	ASB CAPITAL MANAGEMENT, INC.	24702R10	DELL	2088638	2354
195	31MAR2005	ABERDEEN ASSET MANAGERS LTD.	24702R10	DELL	377300	2459
195	30JUN2005	ABERDEEN ASSET MANAGERS LTD.	24702R10	DELL	368900	2421
195	30SEP2005	ABERDEEN ASSET MANAGERS LTD.	24702R10	DELL	323400	2397
195	31DEC2005	ABERDEEN ASSET MANAGERS LTD.	24702R10	DELL	323400	2354
205	31MAR2005	ABNER HERRMAN&BROCK ASSET MGMT	24702R10	DELL	103175	2459
205	30JUN2005	ABNER HERRMAN&BROCK ASSET MGMT	24702R10	DELL	94907	2421
220	30JUN2005	ACADIAN ASSET MANAGEMENT, INC.	24702R10	DELL	1200	2421
260	31MAR2005	ADAMS EXPRESS COMPANY	24702R10	DELL	400000	2459
260	30JUN2005	ADAMS EXPRESS COMPANY	24702R10	DELL	400000	2421
260	31DEC2005	ADAMS EXPRESS COMPANY	24702R10	DELL	400000	2354
350	31MAR2005	ADELL HARRIMAN& CARPENTER INC.	24702R10	DELL	86389	2459
350	30JUN2005	ADELL HARRIMAN& CARPENTER INC.	24702R10	DELL	84064	2421
350	30SEP2005	ADELL HARRIMAN& CARPENTER INC.	24702R10	DELL	85589	2397
350	31DEC2005	ADELL HARRIMAN& CARPENTER INC.	24702R10	DELL	93819	2354
440	31MAR2005	ADVANCE CAPITAL MGMT, INC.	24702R10	DELL	19200	2459

19

IRRC

- Investor Responsibility Research Center
- Corporate Governance data
 - Gompers, Ishii, Metrick (2003 <http://papers.ssrn.com/id=278920>) corporate governance index. Lists the anti-takeover provisions that a firm has. More provisions=more entrenched management=poor governance. Commonly used for recent corporate governance studies.
- Directors data
 - Information on the directors of a firm—board size, age, whether they are independent, whether they hold shares in the firm.

20

IRRC Directors Data E.g.

TICKER	year	DID	CHAIRMAN	CEO	FNAME	LNAME	DIRTYPE	STKHOLDING
DELL	2000	31702	0	0	Michael H.	Jordan	I	0.0
DELL	2000	33232	1	1	Michael S.	Dell	E	12.4
DELL	2000	33233	0	0	Thomas W.	Luce III	L	0.0
DELL	2000	34718	1	1	Alex J.	Mandl	I	0.0
DELL	2000	34778	0	0	Michael A.	Miles	I	0.0
DELL	2000	38267	1	1	Mary Alice	Taylor	I	0.0
DELL	2000	38942	0	0	Klaus S.	Luft	I	0.0
DELL	2000	38943	1	1	Donald J.	Carty	I	0.0
DELL	2000	40255	0	0	Sam	Nunn	I	0.0
DELL	2000	38929	0	0	Morton L.	Topfer	E	0.0
DELL	2001	31702	0	0	Michael H.	Jordan	I	0.0
DELL	2001	32244	0	1	William H.	Gray III	I	0.0
DELL	2001	33232	1	1	Michael S.	Dell	E	12.0
DELL	2001	33233	0	0	Thomas W.	Luce III	L	0.0
DELL	2001	34718	0	0	Alex J.	Mandl	I	0.0
DELL	2001	34778	0	0	Michael A.	Miles	I	0.0
DELL	2001	37262	0	0	Judy C.	Lewent	I	0.0
DELL	2001	38942	0	0	Klaus S.	Luft	I	0.0
DELL	2001	38943	1	1	Donald J.	Carty	I	0.0
DELL	2001	40255	0	0	Sam	Nunn	I	0.0
DELL	2001	38929	0	0	Morton L.	Topfer	E	0.0
DELL	2002	31702	0	0	Michael H.	Jordan	I	0.0
DELL	2002	32244	0	0	William H.	Gray III	I	0.0
DELL	2002	33232	1	1	Michael S.	Dell	E	11.8
DELL	2002	33233	0	0	Thomas W.	Luce III	L	0.0
DELL	2002	34718	0	0	Alex J.	Mandl	I	0.0

Datastream/Worldscope

- Datastream is a comprehensive international database:
 - stock returns (>50,000 stocks in >60 countries)
 - company financial data (provided within Datastream by Worldscope)
 - bond returns
 - stock and bond indices
 - foreign exchange, commodity prices, and economic data.
- Indispensable for studies needing international data.
- However, the data interface is not user-friendly and there are many errors in the dataset. See for eg., Ince and Porter (2004). <http://papers.ssrn.com/id=486523> or appendix of Griffin, Nadari, and Kelly (2006).

SDC

- Thomson's Securities Data Corporation.
- Corporate issuance database
 - New Issues (US and International)
 - IPOs and secondary issues
 - Bond offerings
 - Rights offerings
 - Mergers and Acquisitions
 - Hostile/friendly takeovers
 - Successful and unsuccessful deals
 - US and non-US targets
- SDC also contains some errors. See:
 - <http://pages.stern.nyu.edu/~aljungqv/research.htm> (Alexander Ljungqvist's website)
 - <http://bear.cba.ufl.edu/ritter/ipodata.htm> (Jay Ritter's website)

Overview

Importance of Data in Financial Research

Describe each Database

How to Access the Data?

Practice sessions in Fisher 606

Accessing WRDS

- Wharton Research Database Services (WRDS) is the interface to access most of the databases.
- Only Datastream and SDC cannot be accessed through WRDS.
- 3 ways to access WRDS
 - Web queries (simplest way, but the most limited)
 - Unix SAS (through SSH Secure Shell, download from <http://osusls.osu.edu/upgrades/stg2wnx.html>)
 - SAS PC Connect (need SAS installation on your PC)

25

Why use SAS PC Connect?

- For research intensive tasks, Unix and PC connect are superior to the web interface.
- SAS PC connect has the following advantages:
 - No need to learn unix code. More user-friendly but just as powerful as the unix interface.
 - Easier to edit programs and debug.
 - Can move across windows to view program, output and log file.
- One issue with PC Connect: jobs that take more than 1 hour to run require a SSH tunnel connection to WRDS.
- You may like to use Unix for very large jobs (e.g. those that take more than 1 day to run). Although doing everything on one platform is usually preferable.

26

Steps to Use SAS PC Connect

1. From <http://wrds.wharton.upenn.edu/support/dslist/dslist.shtml> Determine wrds library name for dataset. Eg. CRSP monthly stock returns library is "crsp.msf".
2. Determine variable names needed. E.g., permno, date, ret, prc, vol from crsp.msf dataset.
3. Write your access program and sandwich it with the PC connect commands.
4. Run the program. SAS will connect to WRDS and run the code on the WRDS Unix server. When the program completes, the resulting output will be downloaded to your PC if you have a proc download statement.

27

SAS PC Connect Commands

Sandwich your SAS program between the remote submission commands

```
/*Remote submission to wrds*/
%let wrds = wrds.wharton.upenn.edu 4016;
options comamid=TCP remote=WRDS;
signon username=_prompt_;
rsubmit;

/*Get data from CRSP monthly*/
Data crsp1;
  set crsp.msf;
  keep permno date year ret prc hsiccd;
run;
Proc download data=crsp1 out=crsp1; run;

/*End wrds submission*/
endrsubmit;
```

28

Datastream/Worldscope

- Requires a software installation on your desktop. Important to log out after use because only one person can be logged in.
- 2 ways of access:
 - Datastream Advance Excel add-in (simple but limited usefulness)
 - DS Windows macros (useful website, <http://www.princeton.edu/~econlib/ds/samplemac.htm>)
 - Cross-sectional data (use 900A macro)
 - Time-series data (use 900B macro)
- If you have many macros to run, DS Agenda can help you to run them sequentially.

29

Datastream/Worldscope

- Categories of Data that will be useful
 - Equities
 - Equity Lists
 - Equity & Misc Indices
 - Exchange Rates
 - Economic Series
 - Interest Rates
- Variables available
 - Datatypes
 - Worldscope Data Items

Base-Date	Mnemonic	Code	DS	ISS	MS	Mkt
09-11-03		275640	N			IR
06-25-03		272110	N			SD
12-22-05		325940	N			UK
06-02-05		311622	N			ME
07-17-04	D:EN2X	296896	N			BD
07-16-04	D:PLXK	292451	Y			
07-16-04	D:PRGX	29245X	Y			
05-03-05		309258	N			IR
03-23-05		30727F	N			UK
08-09-04		293263	N			IR
02-22-05	800A	30379R	N			US
04-13-06	D:BSW	35703C	Y			BD
02-07-06	D:SGP	32794U	Y			BD
02-08-06	D:EXQ	32801N	Y			BD
02-09-06	D:FQJ	32803B	Y			BD

30

Eg. Get prices for all 30 Dow Jones Index stocks

- Open DS Windows
- Get dscodes (identifier codes) for all 30 firms. Under "equity lists" category, search for Dow Jones Industrial and we find the code LDJINDUS.
- Enter LDJINDUS into the Codes: section of the 900A macro (cross-section macro).
- Run 900A macro to get dscodes.
- Collect the output dscodes and paste them onto the 900B macro.
- Run 900B macro to get the monthly price series for each of the 30 firms.

31

900B – Time-series macro

```

STARTDC(CSVFILE, "C:\dataosu\900bdata.csv")
OpenData Codes
Loop:
If &endOfData = FALSE Then
Input code
Send( "900B " + code + "(P, 1-1-2002, 12-31-2003, D//C" )
Send("[CLEAR]*")
Goto Loop
EndIf
ENDDC
End

Codes:
DATA
"902172"
"945388"
"905113"
"904853"
"906156"
"916305"
"904818"

ENDDATA
    
```

32

Making downloaded Datastream data usable

- Data extracted is in a panel format.
- But we need it in a stacked format for analysis. Need to write a program (e.g. in SAS) to transpose the data.

	A	B	C	D	E	F	G
1	Name	3M	AT&T	ALCOA	ALTRIA GI	AMERICA	AMERICA
2	Code	902172(P)	945388(P)	905113(P)	904853(P)	906156(P)	916305(P)
3	Currency	US	US	US	US	US	US
4	1/1/2002	59.105	39.17	35.55	45.85	31.2412	79.4
5	1/2/2002	58.57	39.9	35.68	46.64	31.4075	78.75
6	1/3/2002	58.375	40.17	36.14	46.59	31.9064	78.58
7	1/4/2002	58.55	39.99	37.3	46.09	33.0181	77.8
8	1/7/2002	57.85	39.91	38.16	46.58	32.9131	76.8

	A	B	C	D
1	Date	Name	dscode	Price
2	1/1/2002	3M	902172	59.105
3	1/2/2002	3M	902172	58.57
4	1/3/2002	3M	902172	58.375
5	1/4/2002	3M	902172	58.55
6	1/7/2002	3M	902172	57.85
7	1/8/2002	3M	902172	57.525
8	1/9/2002	3M	902172	57.325
9	1/10/2002	3M	902172	56.6
10	1/1/2002	AT&T	945388	39.17
11	1/2/2002	AT&T	945388	39.9
12	1/3/2002	AT&T	945388	40.17
13	1/4/2002	AT&T	945388	39.99
14	1/7/2002	AT&T	945388	39.91
15	1/8/2002	AT&T	945388	39.74
16	1/9/2002	AT&T	945388	38.17
17	1/10/2002	AT&T	945388	38.5

Overview

Importance of Data in Financial Research

Describe each Database

How to Access the Data?

Practice sessions in Fisher 606