Day 5: Text as Data

Kenneth Benoit

CEU April 14-21 2011

April 20, 2011

Text as **data**

Inherit properties of statistics

Precise characterizations of uncertainty

Concern: Reliability

Concern: Validity

Past - Present - Future

Hand-coded content analysis

Sometimes computer-assisted

Extensively applied to manifestos

Past - Present - Future

"Text as data" scaling approaches

Classification approaches

and still: manifesto hand-coding

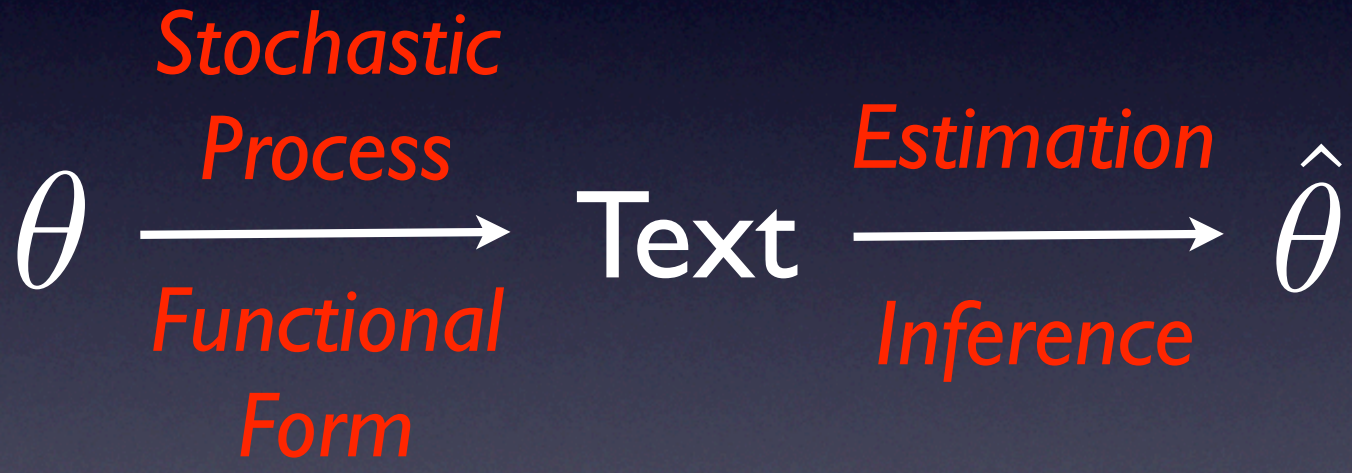
Past - Present - Future

Better uncertainty models

Advances in estimation methods

Advances in parametric scaling

"From Text to Policy Positions"



also: From Policy Positions to text

Problem 3: Now try this one!

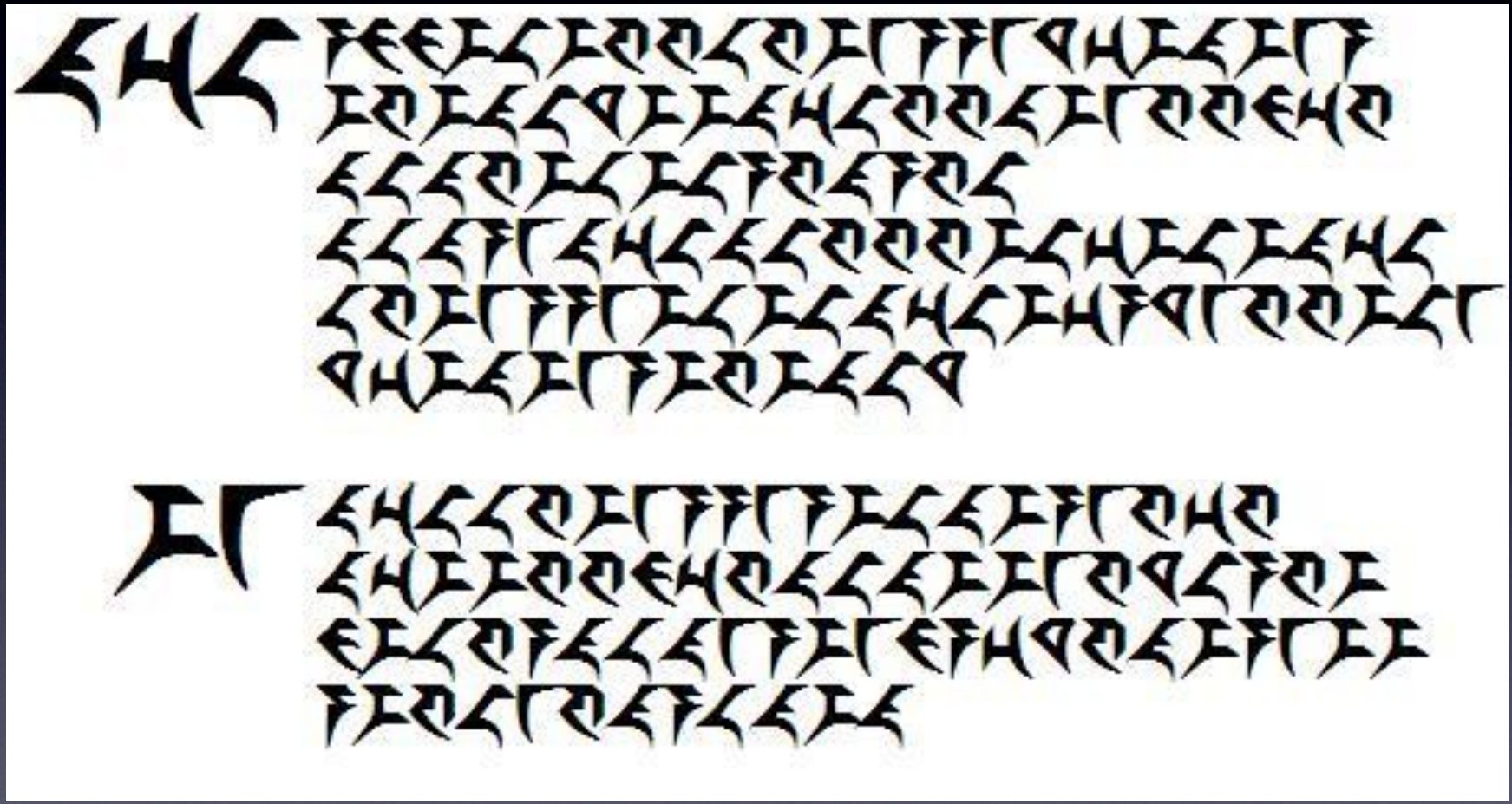
Kansainväliset uraaniyhtiöt ovat olleet kiinnostuneita Kainuussa sijaitsevista esiintymistä. Kainuun maakunta-kuntayhtymä on Perussuomalaisten valtuustoryhmän aloitteen pohjalta selvittänyt kainuulaisten suhtautumista mahdollisiin uraanikaivoksiin.

Sotkamossa sijaitsevan Talvivaaran kaivoksen sivutuotteena tulee myös uraania, joka aiotaan ottaa jätelietteestä talteen. Tässä uraanin talteenotossa syntyy niin paljon ydinvoimalaitosten polttoainetta, että se riittäisi noin 80 prosenttisesti Suomessa toimivien ydinvoimaloiden tarpeisiin.

Talvivaaran tapauksessa ei kaivoksen johdon mukaan ole kysymys varsinaisen uraanikaivoksen avaamisesta, vaan vain sivutuotteen talteenotosta. Valtioneuvosto tullee päättämään Talvivaara-asiasta uraanin osalta tämän vuoden aikana.

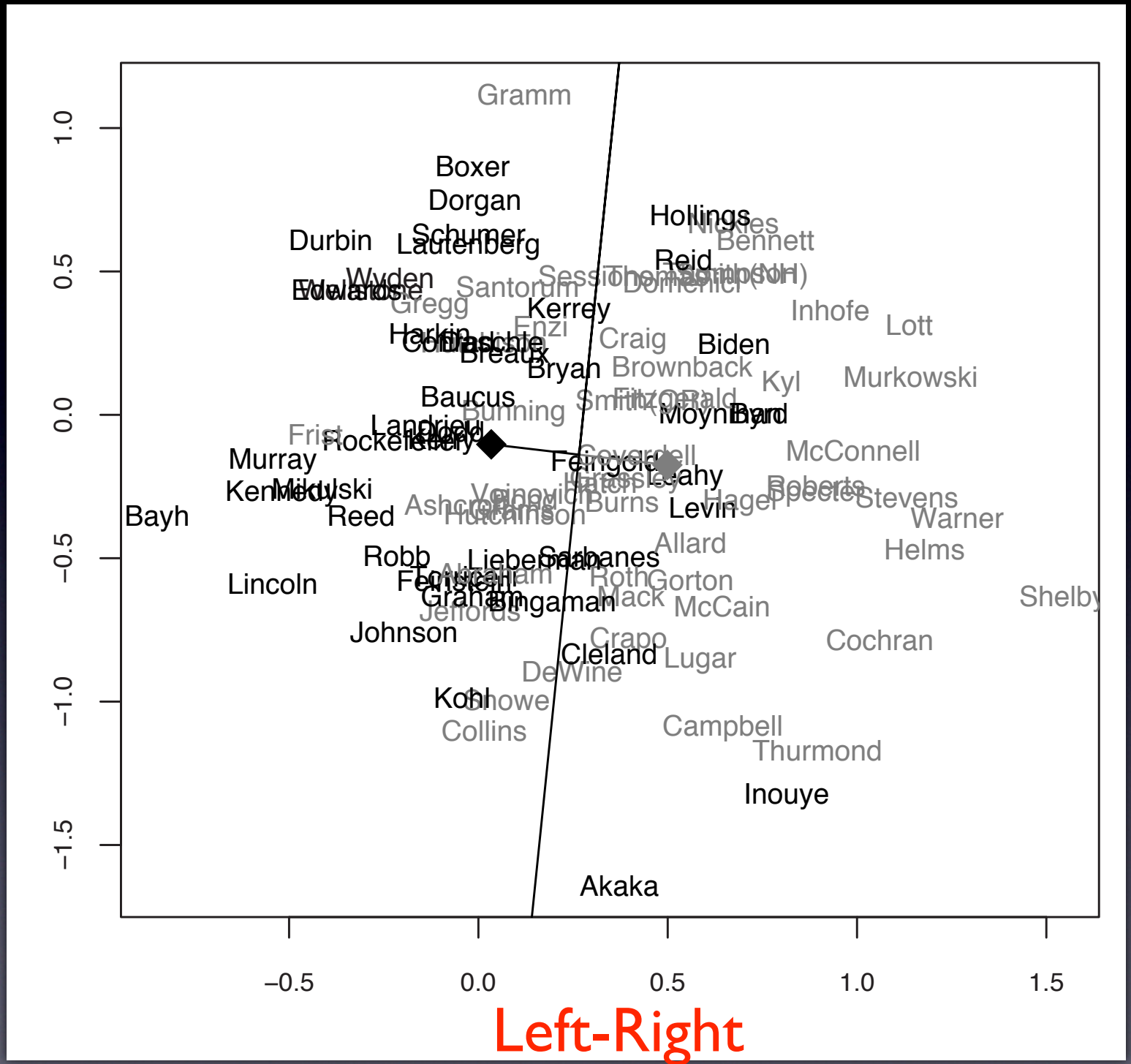
Perussuomalaiset

and this one???



Problem (last): Interpret these text scaling results!

Work-Horse/ Show-Horse



PRESENT

Scaling models:

Wordscores; "Wordfish"; others

Classification methods

Naive Bayes (e.g. McIntosh et al)

Readme (Hopkins and King)

Dynamic Topic Models (e.g. Quinn et al.)

Support Vector Machines (Hillard et al, Yu et al.)

Expressed Agenda Model (Grimmer 2010)

Laver and Garry (2000)

- Represents policy hierarchically, 300 categories:
  - Economy
  - Political system
  - Social system
  - External relations
  - General
- Each has an anti, neutral, positive
- Used a dictionary to classify each 10-word sequence
- Position defined as P = (R-L)/(R+L)
- Never used in a subsequent article, but much-cited

Wordscores in a nutshell

- *Wordscores* is a statistical method for extracting policy positions from political texts, implemented by computer
- Michael Laver, Kenneth Benoit, and John Garry. "Extracting Policy Positions From Political Texts Using Words as Data", *APSR* 2003
- Enables extraction of policy positions from texts without having to ascribe "meaning" to texts, or to read them, or even to be able to read them (works in English, German, French, and Italian so far)
- Because it is based on the statistics of relative word frequencies, Wordscores can generate estimates of uncertainty, something no existing methods of textual content analysis offer

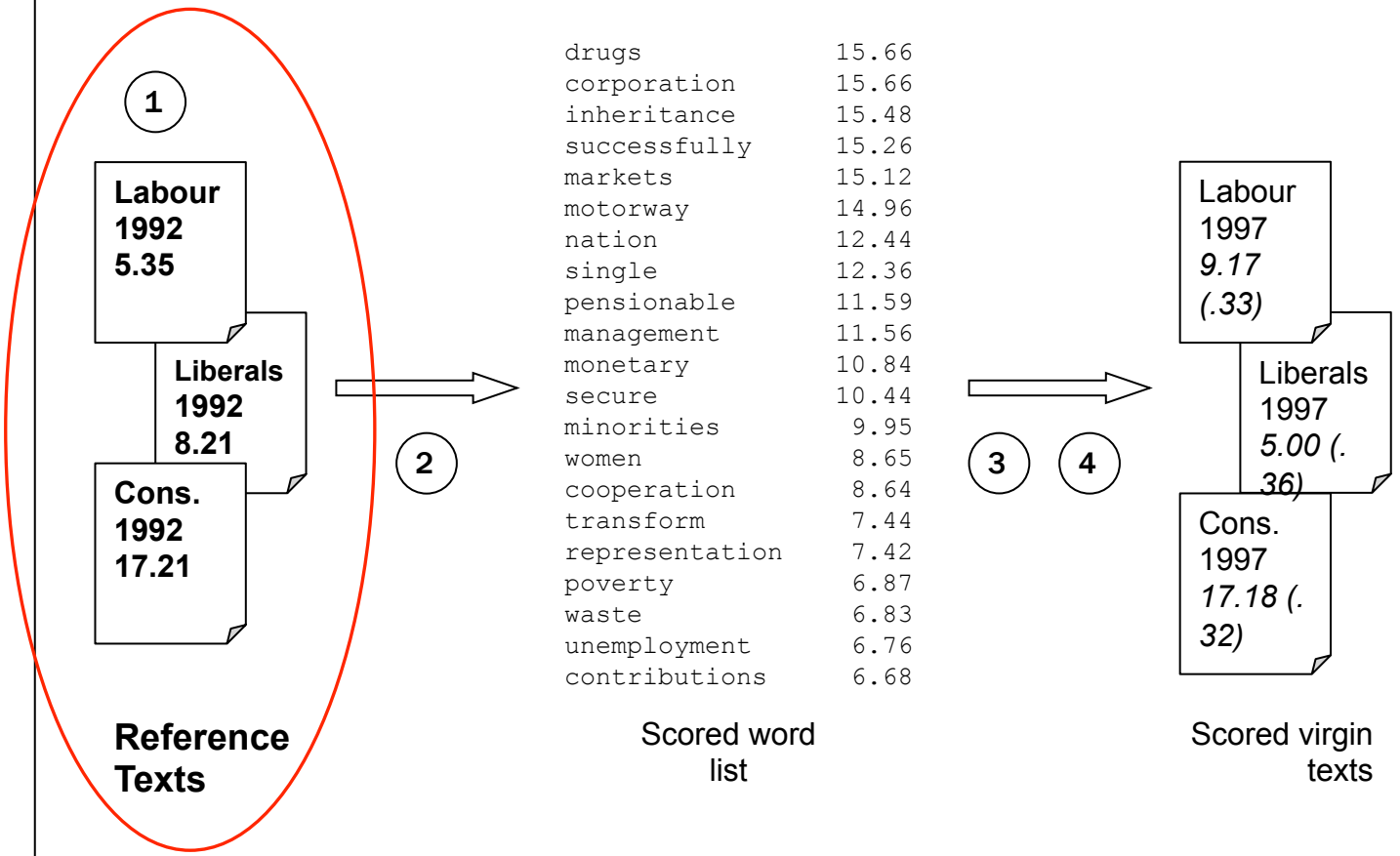
WORDSCORES conceptually

- "Reference" texts: texts about which we know something (a scalar dimensional score)
- "Virgin" texts: texts about which we know nothing (but whose dimensional score we'd like to know)
- Basic procedure:

1. Analyze reference texts to obtain word scores
2. Use word scores to score virgin texts

The Wordscore Procedure

(Using the UK 1997-2001 Example)

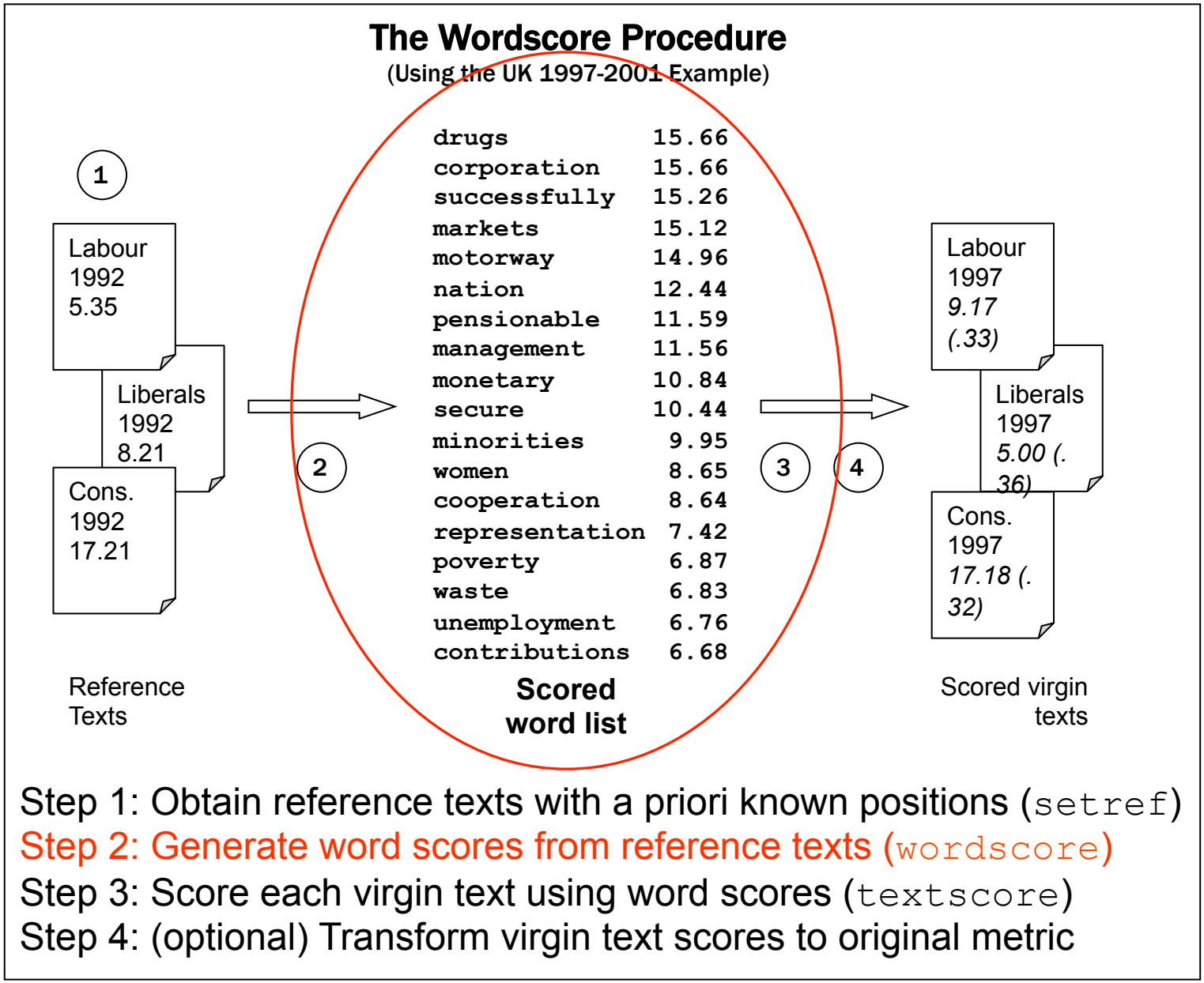


Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric



Step 1: Obtain reference texts with a priori known positions (`setref`)

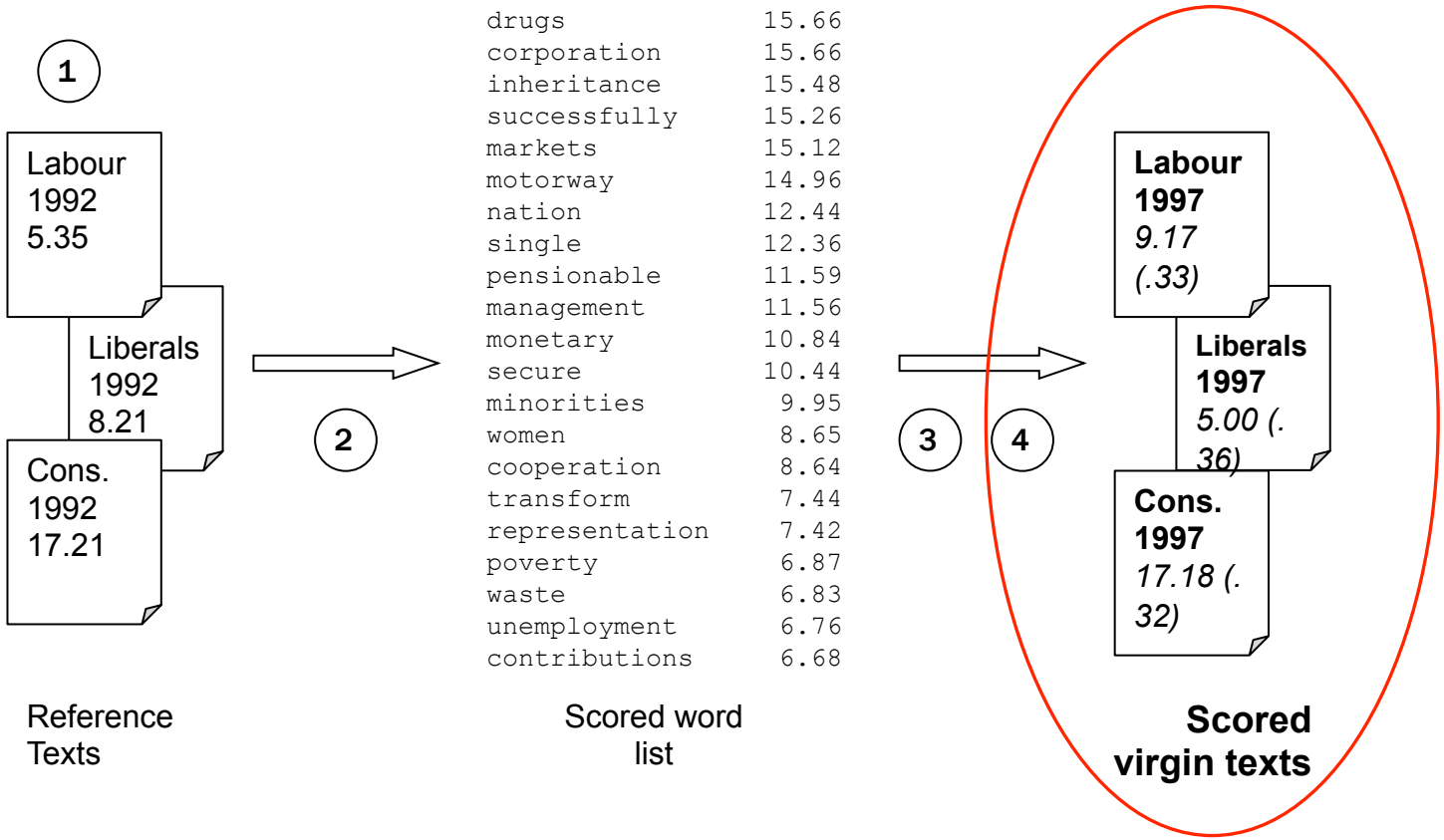
Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

The Wordscore Procedure

(Using the UK 1997-2001 Example)



Step 1: Obtain reference texts with a priori known positions (`setref`)

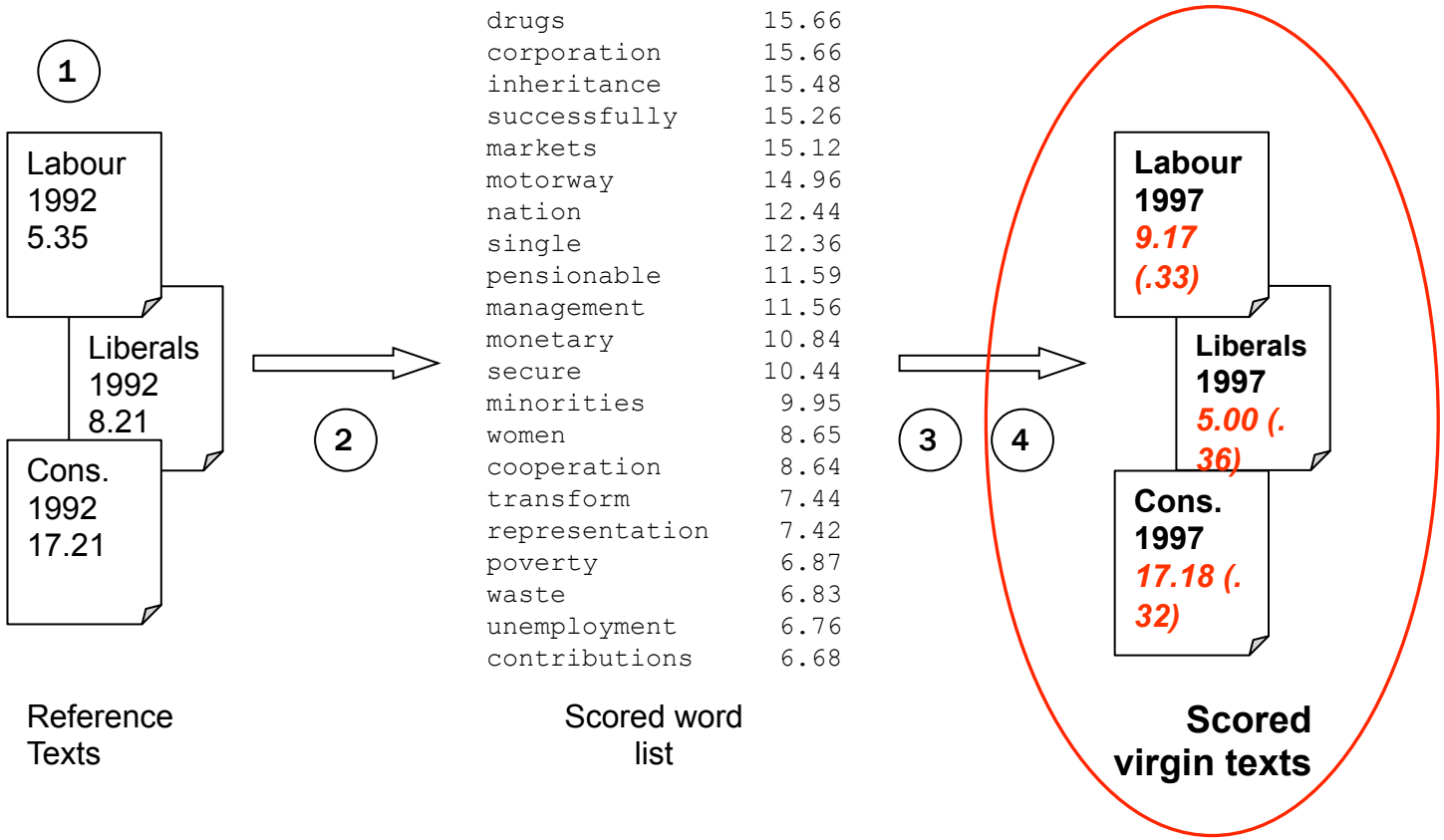
Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

The Wordscore Procedure

(Using the UK 1997-2001 Example)



Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

WORDSCORES mathematically

- Start with *R* reference texts and *V* virgin texts with with *W* words in common

  (using `wordcount` to generate a matrix of words and their relative frequencies in all reference texts)

- A_{rd} is assumed position of reference text *r* on policy dimension *d*

- F_{wr} is relative frequency of word *w* in text *r*

WORDSCORES mathematically

- Compute P_{wr} for each **reference text**: the probability that we are reading reference text *r* given that we are reading word *w*

$$P_{wr} = \frac{F_{wr}}{\sum_{r} F_{wr}}$$

- Example:

  Two reference texts, A and B. The word "choice" is used 10 times per 1,000 words in Text A and 30 times per 1,000 words in Text B.

  If we know only that we are reading the word "choice" in one of the two reference texts, then probability is 0.25 that we are reading Text A, and 0.75 that we are reading Text B.

WORDSCORES mathematically

- Compute S_{wd} for each **word**: the score of each word *w* on dimension *d*

$$S_{wd} = \Sigma_r (P_{wr} \cdot A_{rd})$$

- Example continued:

  We know (from independent sources) that Reference Text A has a position of –1.0 on dimension *d*, and Reference Text B has a position of +1.0.

  The score of the word "choice" is then:

  0.25 (– 1.0) + 0.75 (1.0) = – 0.25 + 0.75 = + 0.5

WORDSCORES mathematically

- Compute S_{vd} for each **virgin text**: the score of each virgin text *v* on dimension *d*

$$S_{vd} = \sum_{w} \left(F_{wv} \cdot S_{wd} \right)$$

- This score is the **mean** dimension score of all of the scored words that a virgin text contains, weighted by the frequency of the scored words
- *Uncertainty*: A weighted **variance** V_{vd} can also be computed for each virgin text, representing the uncertainty of the estimate S_{vd}. Because every words adds information to S_{vd}, more words reduce our uncertainty about S_{vd}. Also, the more consensus among the virgin words around S_{vd}, the more certain we are about S_{vd} .
- *Rescaling:* S_{vd} can be rescaled as S^*_{vd} for interpretation on the original metric of the reference text scores.

Application 1:
UK and Irish election manifestos

Party	*Liberal Democrats*	*Labour*	*Conservative*	*Mean Absolute Difference*
1992 reference texts				
A priori positions	**8.21**	**5.35**	**17.21**	
S.E (n = 34)	*0.425*	*0.377*	*0.396*	
Estimates				
1997 transformed virgin text scores	**5.00**	**9.17**	**17.18**	
S.E.	0.363	0.351	0.325	
1997 expert survey	**5.77**	**10.30**	**15.05**	
S.E (*n* =117)	0.234	0.229	0.227	
1997 standardized comparison scores				
Word scores	-0.88	-0.21	1.09	**0.13**
Expert survey	-0.99	-0.02	1.01	**--**
Hand coded content analysis	-0.83	-0.28	1.11	**0.17**
Dictionary based computer coding	-1.08	0.18	0.90	**0.13**

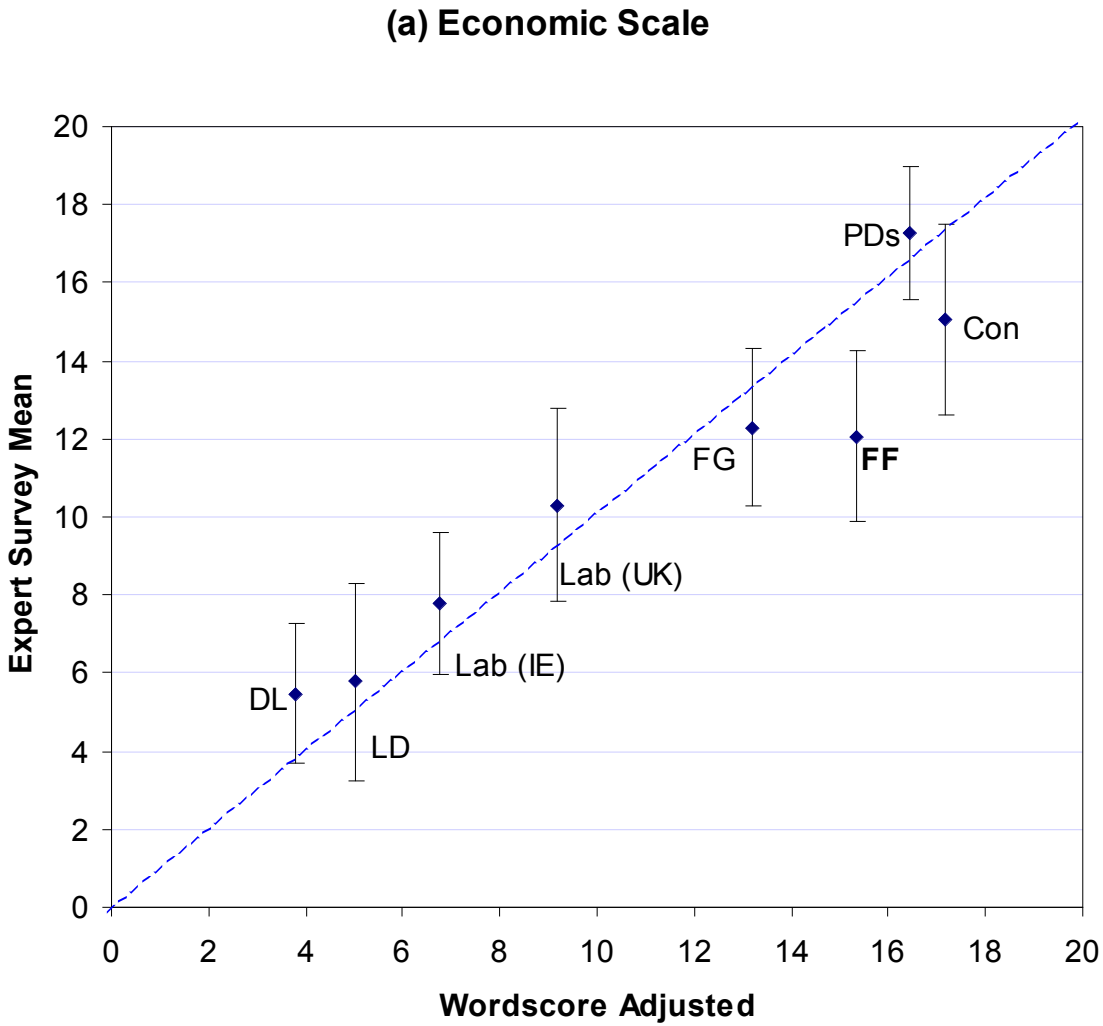
Application 1:
UK and Irish election manifestos

Party	*Liberal Democrats*	*Labour*	*Conservative*
Raw data			
1992 reference texts			
Length in words	17,077	11,208	28,391
No. of unique words	2,911	2,292	3,786
1997 virgin texts			
Raw mean word scores (*Svd*)	10.2181	10.3954	10.7361
S.E.	0.015	0.015	0.014
Length in words	13,709	17,237	20,442
Unique words scored	1,915	2,211	2,279
% words scored	94.9	96.2	95.5
Unique unscorable words	423	697	714
Mean frequency of unscorable words	1.23	1.26	1.29

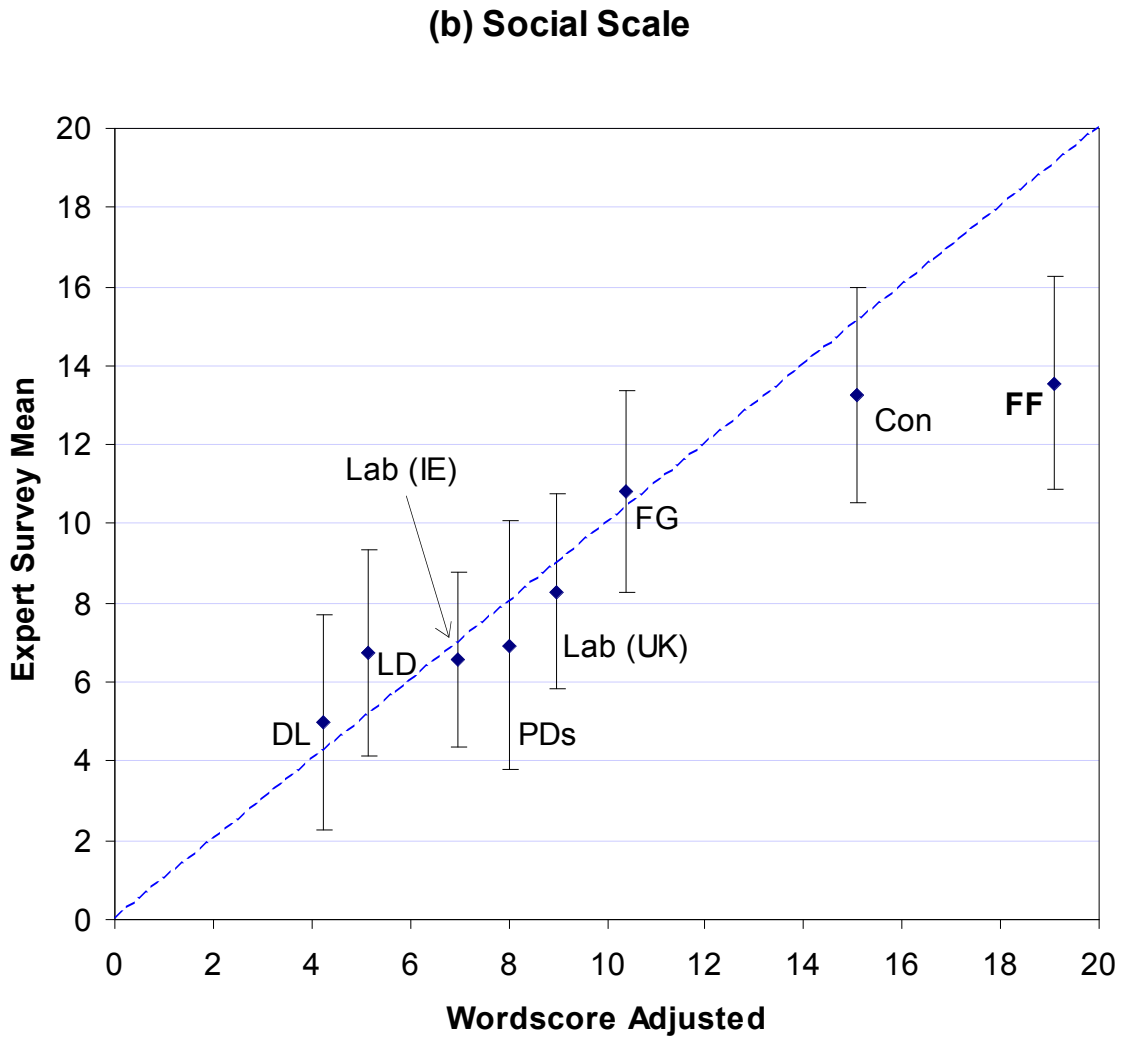
Application 2:
German election manifestos

Party	*1990 PDS*	*1994 PDS*	*1994 Green*	*1994 SDP*	*1994 CDU*	*1994 FDP*
Economic Policy Dimension						
1990 reference texts						
Economic Policy						
A priori positions (1991 expert	--	--	**5.21**	**6.53**	**13.53**	**15.68**
S.E. (*n* =19)	--	--	0.652	0.436	0.544	0.613
1994 transformed economic policy virgin text scores	**4.19**	**3.98**	**7.47**	**10.70**	**13.67**	**17.15**
S.E.	0.436	0.511	0.259	0.365	0.391	0.22
Social Policy Dimension						
A priori positions (1991 expert	--	--	**2.90**	**6.68**	**14.42**	**6.84**
S.E. (*n* =19)	--	--	0.908	0.856	0.537	0.603
1994 transformed social policy virgin text scores		**1.93**	**4.09**	**11.07**	**13.65**	**8.12**
S.E.	0.306	0.421	0.221	0.325	0.368	0.182

Applications 1 and 2: Manifesto Summary



Applications 1 and 2: Manifesto Summary



Application 3: Irish Daíl speeches

- Daíl Confidence debate 1991
- Haughey speech (6,771 words) given reference score of +1.0
- Opposition leader speeches coded -1.0: Bruton (FG, 4,375 words) and de Rossa (DL, 6,226 words)
- Words scored from these 3 speeches used to score virgin texts from 62 different speakers, mean speech length 2,368 words

Group	*N*	Raw Mean	Raw SD	Standard-ized Mean	Standard-ized SD
FF Ministers	12	-0.2571	0.0383	1.15	0.66
PD Minister	1	-0.2947	--	0.50	--
FF	10	-0.2999	0.0721	0.41	1.24
Independent	1	-0.3360	--	-0.21	--
Greens	1	-0.3488	--	-0.43	--
WP	2	-0.3501	0.0423	-0.46	0.73
FG	21	-0.3580	0.0306	-0.59	0.53
Labour	7	-0.3599	0.0220	-0.62	0.38

Application 4: Scoring legislative speeches in no-confidence debate (Irish legislature)

Group	*N*	Median Total Words	Median Unique Words	Raw Mean	Raw SD	Standard-ized Mean	Standard-ized SD
Reference Texts							
FF Prime Minister Haughey	1	6,711	1,617	1.0000	--	--	--
FG Opposition Leader Bruton	1	4,375	1,181	-1.0000	--	--	--
DL Leader de Rossa	1	6,226	1,536	-1.0000	--	--	--
Virgin Texts							
FF Ministers	12	3,851	727	-0.2571	0.0383	1.15	0.66
PD Minister	1	2,818	593	-0.2947	--	0.50	--
FF	10	1,553	397	-0.2999	0.0721	0.41	1.24
Independent	1	3,314	582	-0.3360	--	-0.21	--
Greens	1	1,445	415	-0.3488	--	-0.43	--
WP	2	2,001	455	-0.3501	0.0423	-0.46	0.73
FG	21	1,611	394	-0.3580	0.0306	-0.59	0.53
Labour	7	2,224	475	-0.3599	0.0220	-0.62	0.38

Example: Scoring legislative speeches in no-confidence debate (Irish legislature)

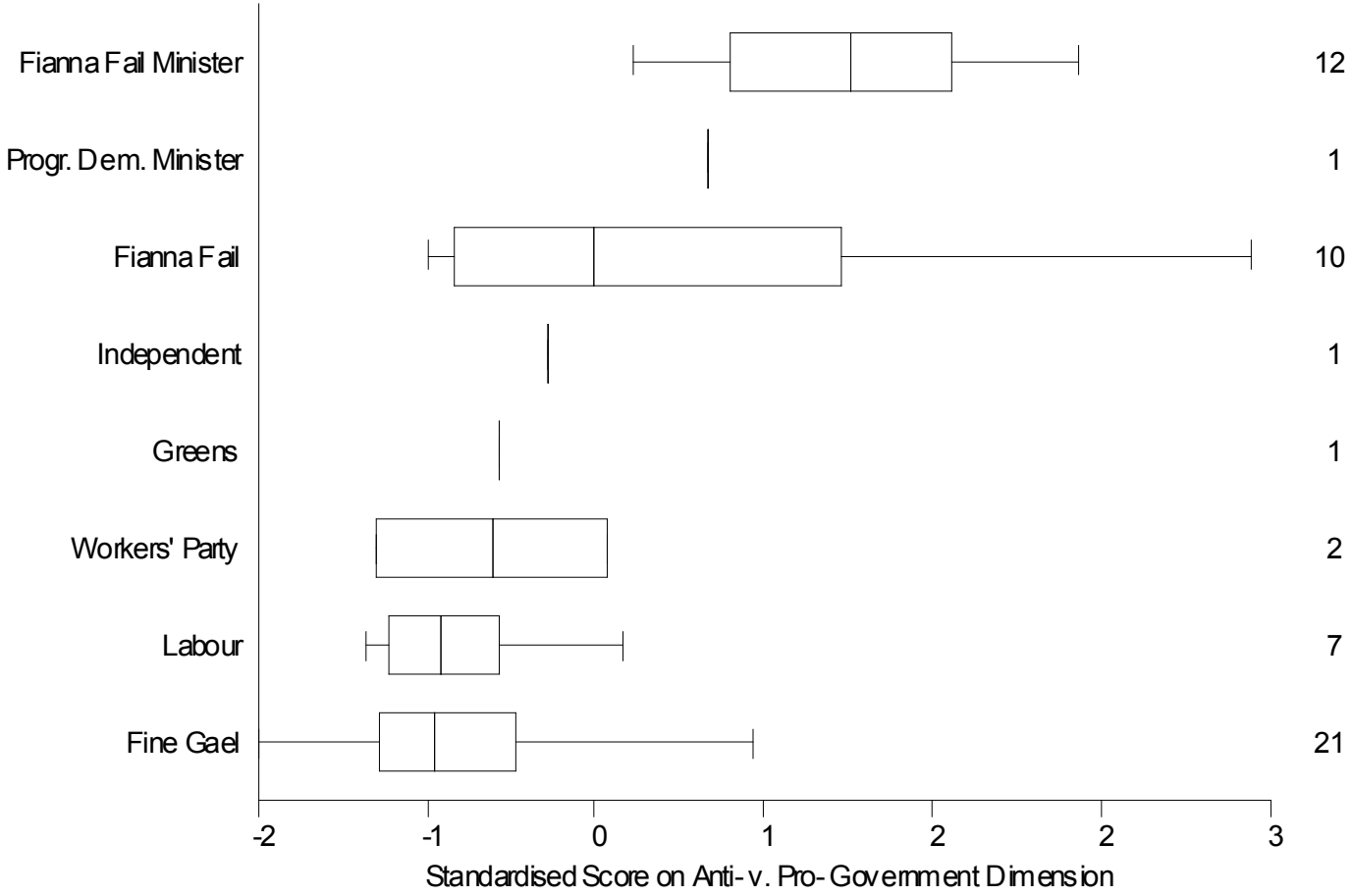
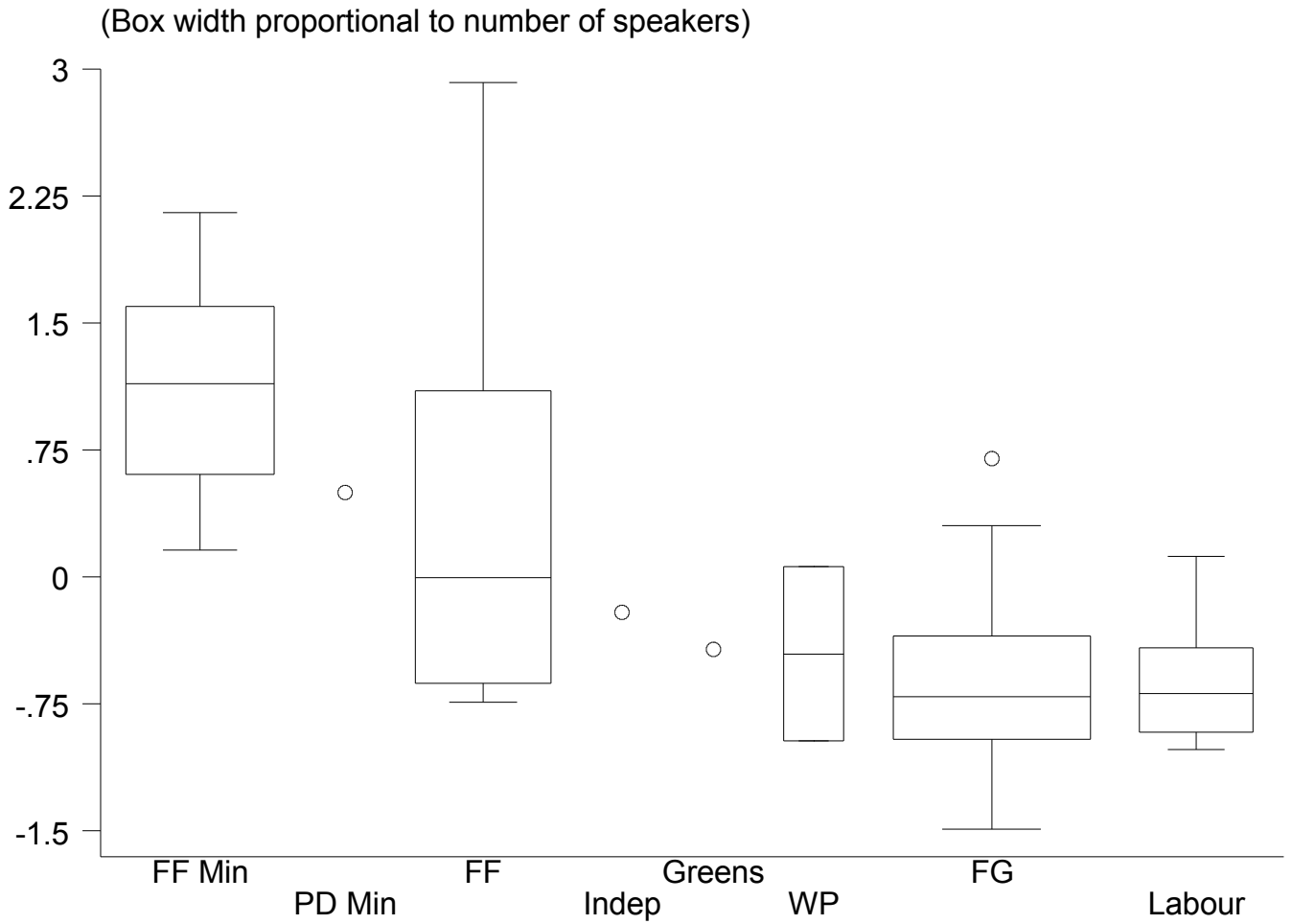


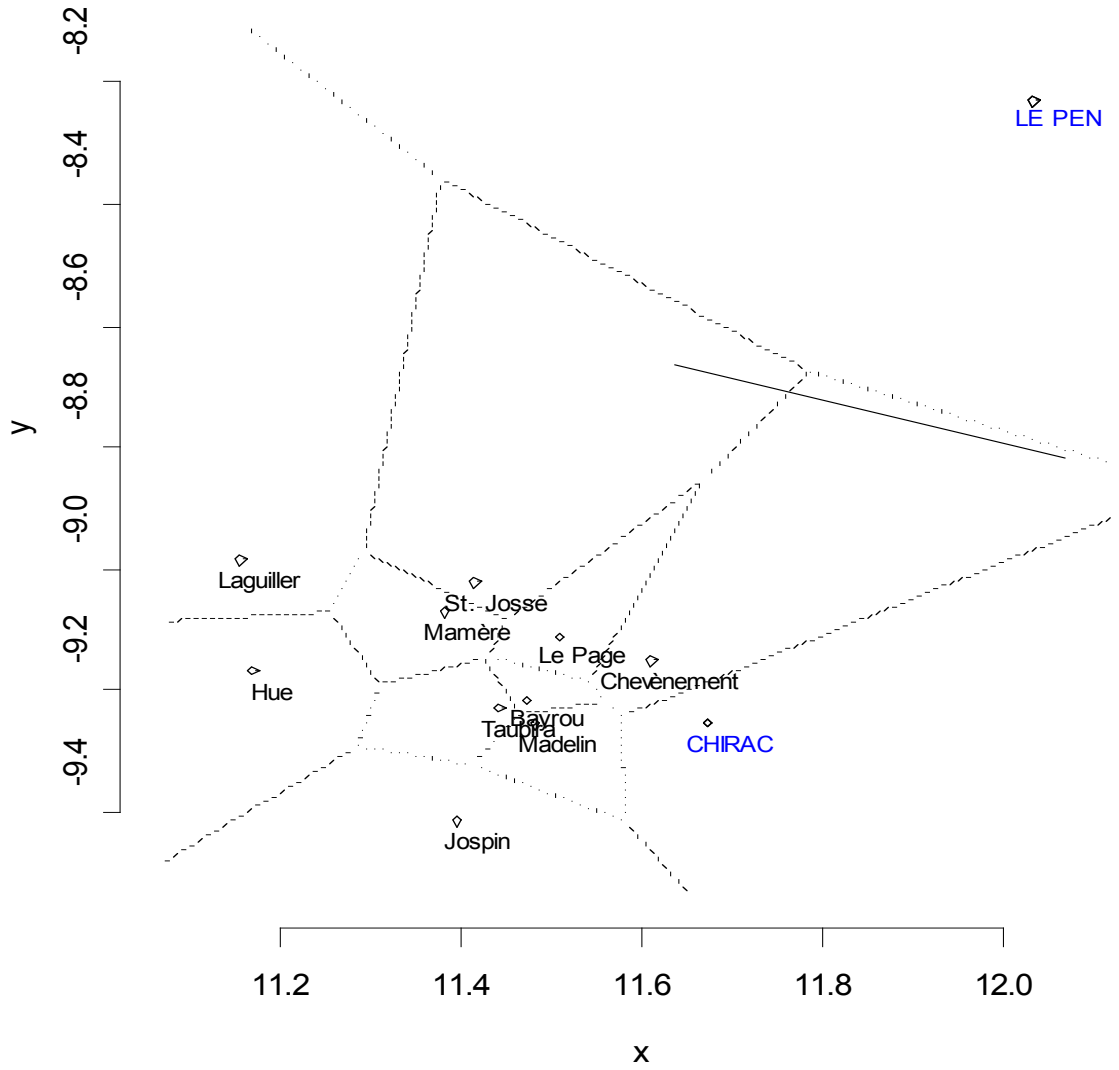
Figure 3. Box plot of standardised scores of speakers in 1991 confidence debate on "pro- versus anti- government" dimension, by category of legislator. Figures on the right indicate the number of legislators in each category.

Application 3:
Irish Daíl speeches



Applications 4

French presidential & party addresses 2002



Applications 5
Italian ministers / junior ministers

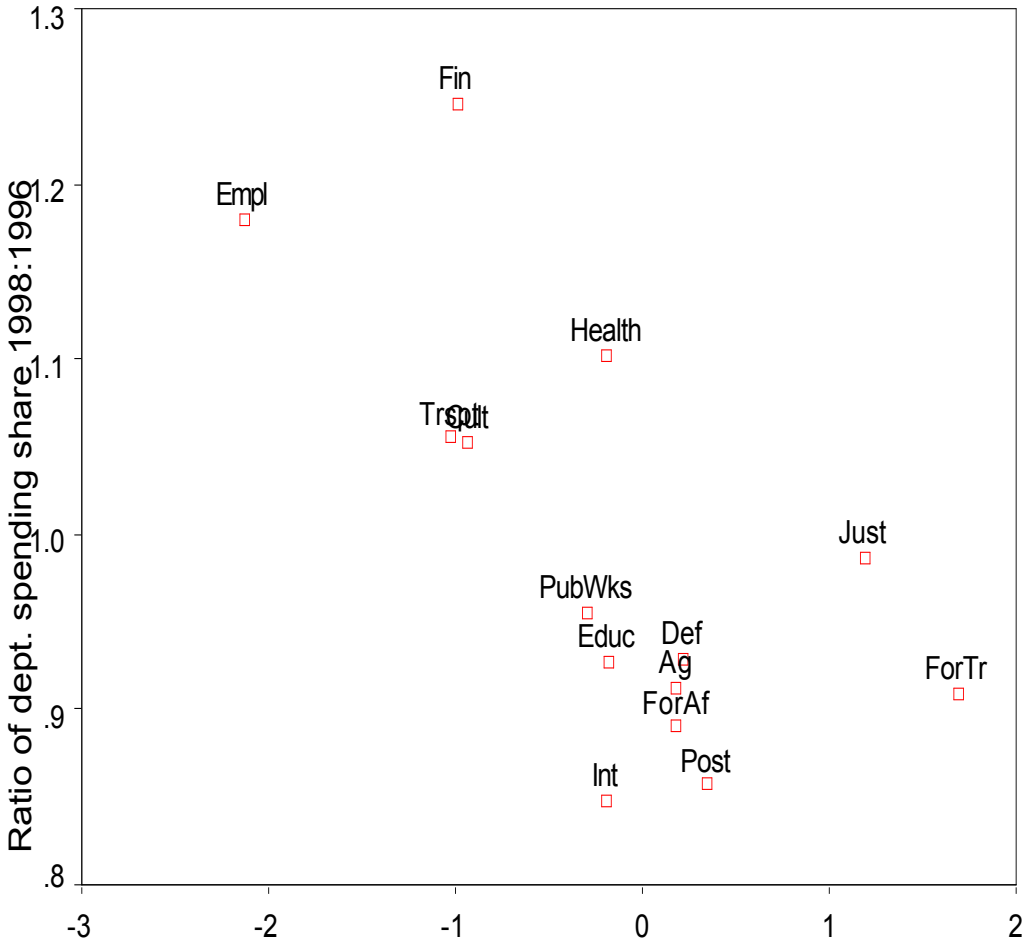
- Score Italian ministers' / junior ministers' legislative speeches over a year
  - Using party leaders' investiture speeches as reference texts
  - And expert surveys to estimate the positions of these
  - Can estimate relative positions of all individual members of the government.
  - To test hypotheses about junior ministers, and whether portfolio allocation makes a difference.

Applications 5

Italian ministers / junior ministers

	Minister to right of coalition	Minister to left of coalition
Junior minister to right of minister	**Agriculture** **Public administration**	**Employment** **Transport** **Finance** **Culture** **Public works** **Health** **Interior**
Junior minister to left of minister	**Defence** **Treasury** **Justice** **Foreign policy** **Posts** **Foreign trade**	**Education**

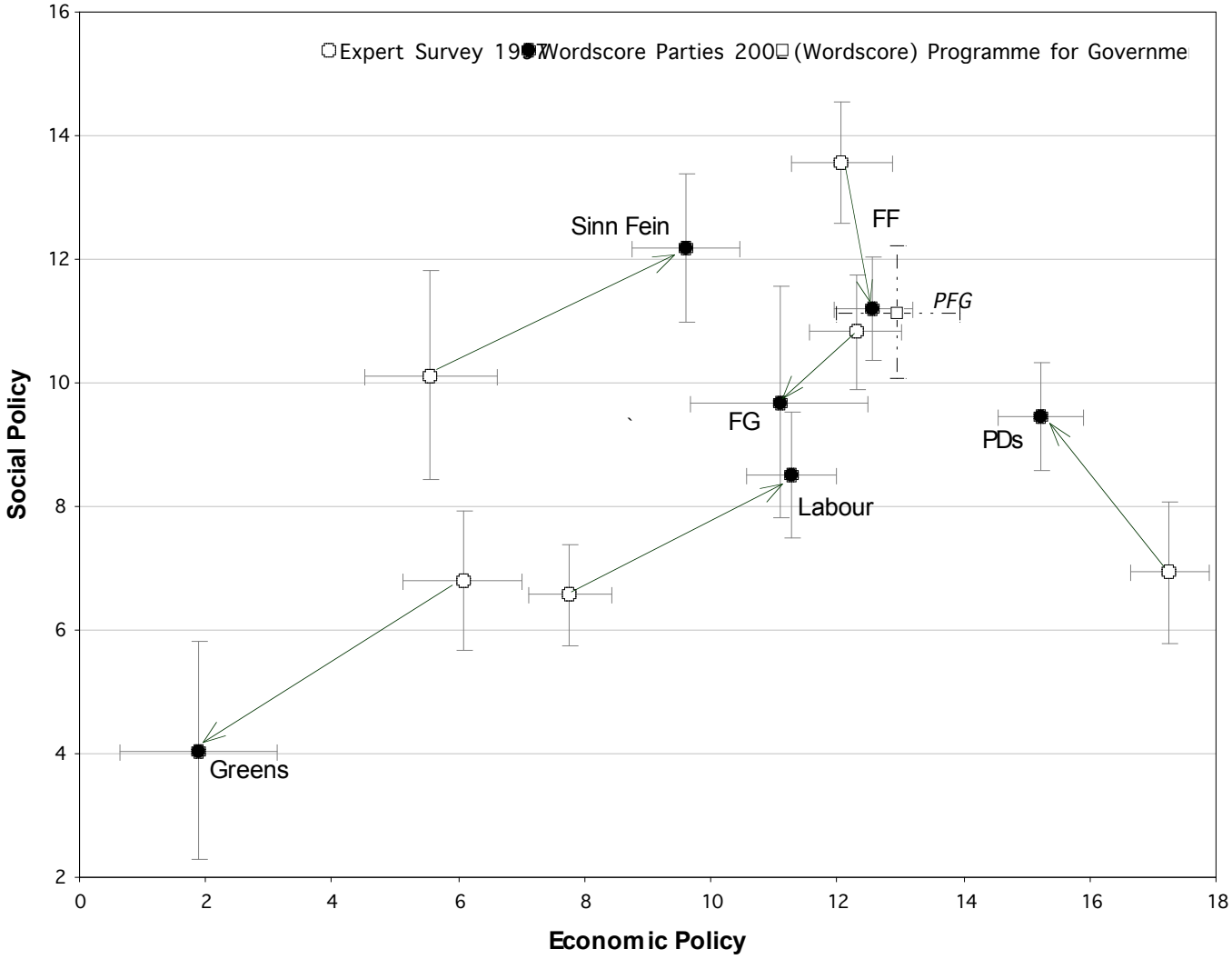
Applications 5
Italian ministers / junior ministers



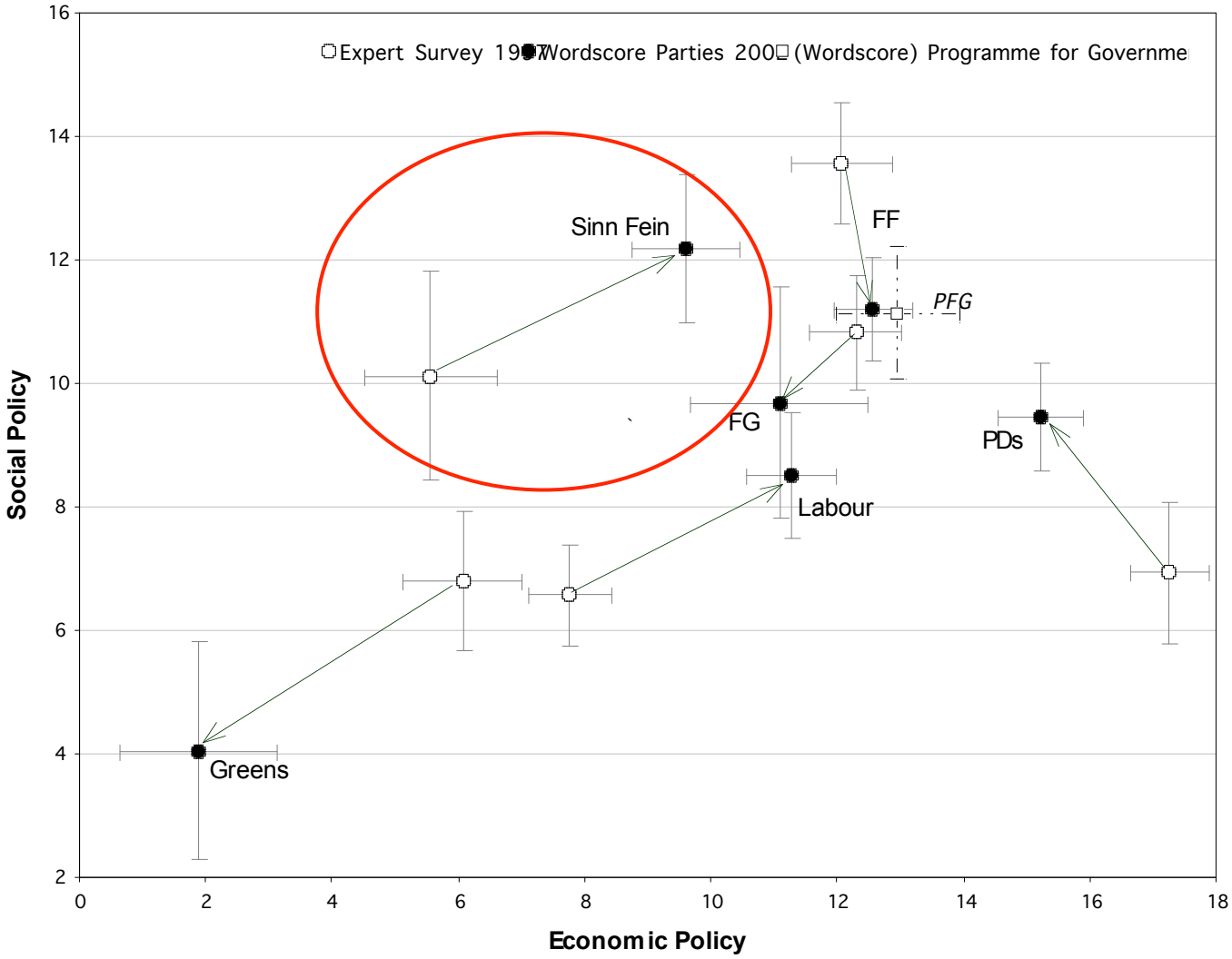
Application 6:
Irish 2002 Manifestos

- Reference scores are 1997 expert surveys
- During the 2002 election campaign, I downloaded each manifesto the day it became available, converted it to text, and scored it on 4 dimensions, getting immediate results
- Once the FF-PD coalition had issued its Programme for Government, I scored that too

Application 6: Irish 2002 Manifestos



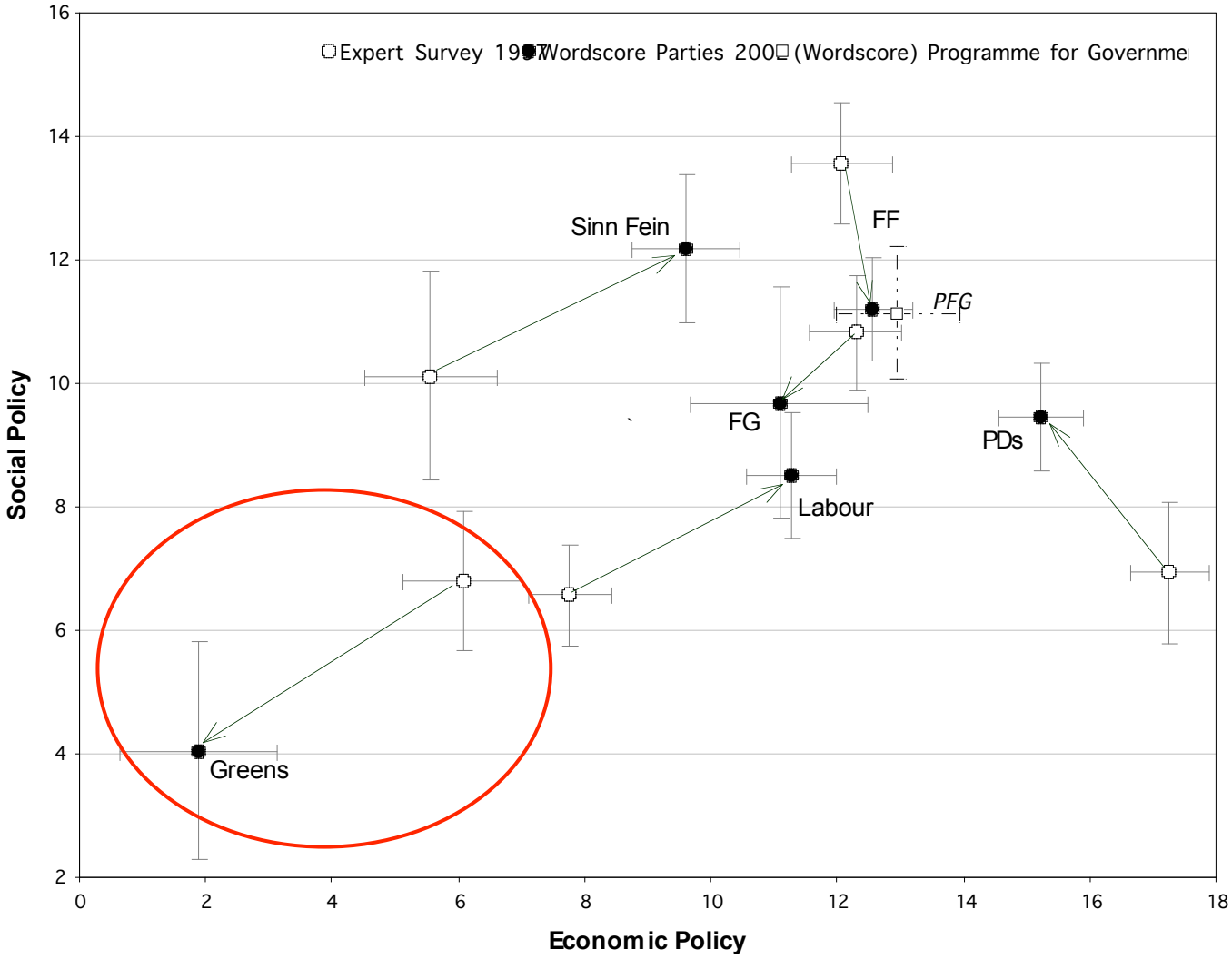
Application 4: Irish 2002 Manifestos



Economic and Social Results

- **SF to center**
- Greens to Left
- Labour and FG move together towards center
- PDS to center
- PFG between FF and PDs

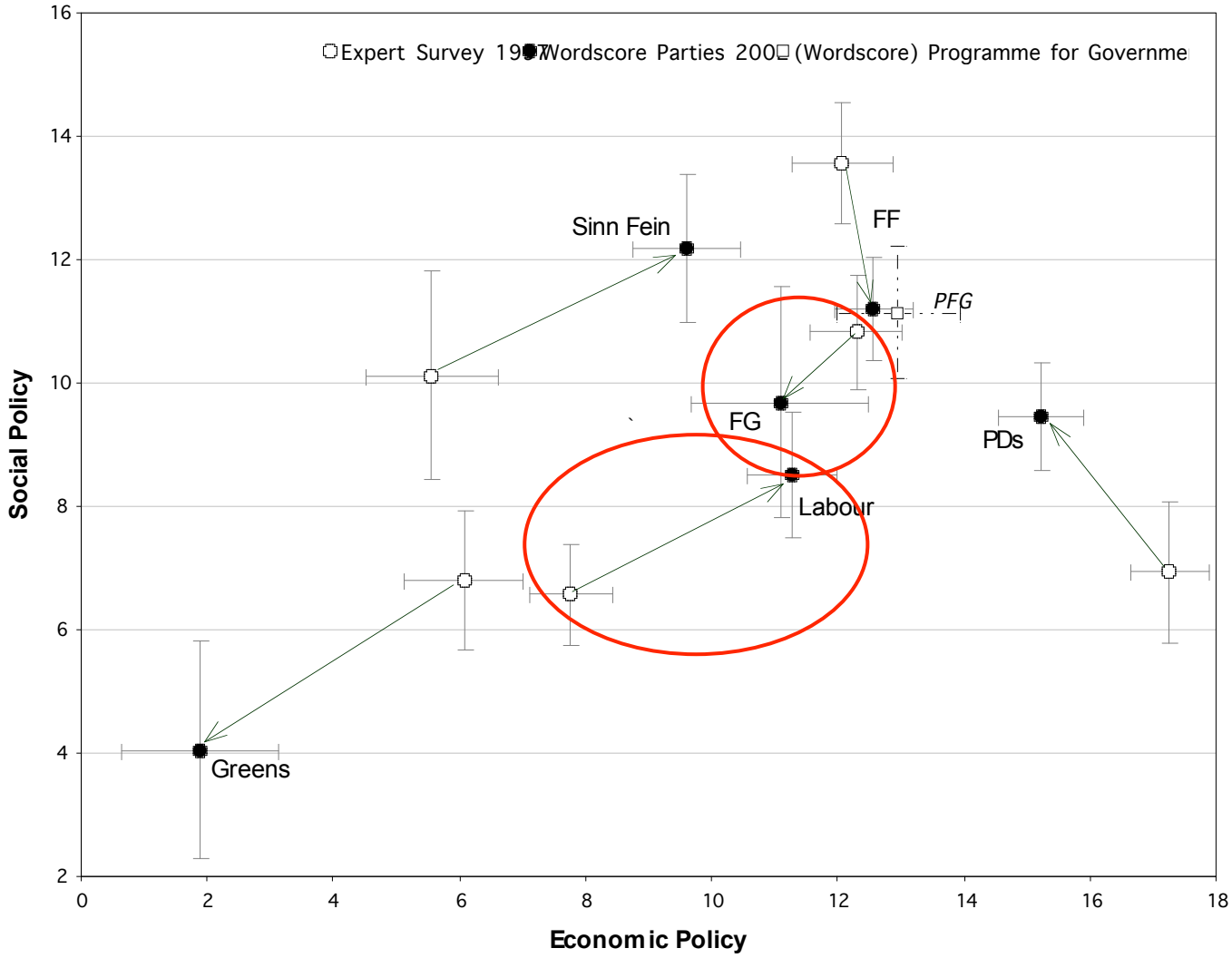
Application 4: Irish 2002 Manifestos



Economic and Social Results

- SF to center
- **Greens to Left**
- Labour and FG move together towards center
- PDS to center
- PFG between FF and PDs

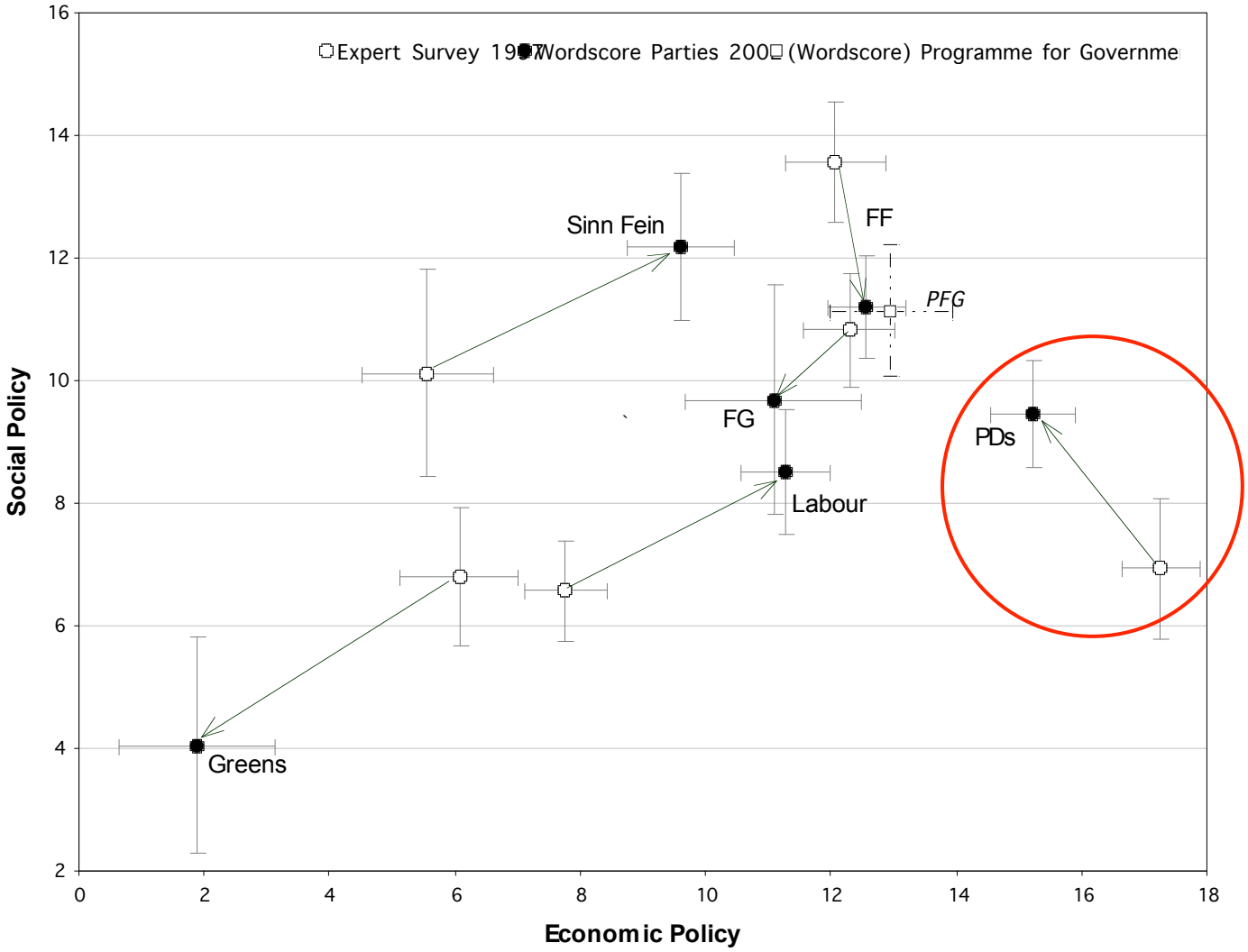
Application 4: Irish 2002 Manifestos



Economic and Social Results

- SF to center
- Greens to Left
- **Labour and FG move together towards center**
- PDS to center
- PFG between FF and PDs

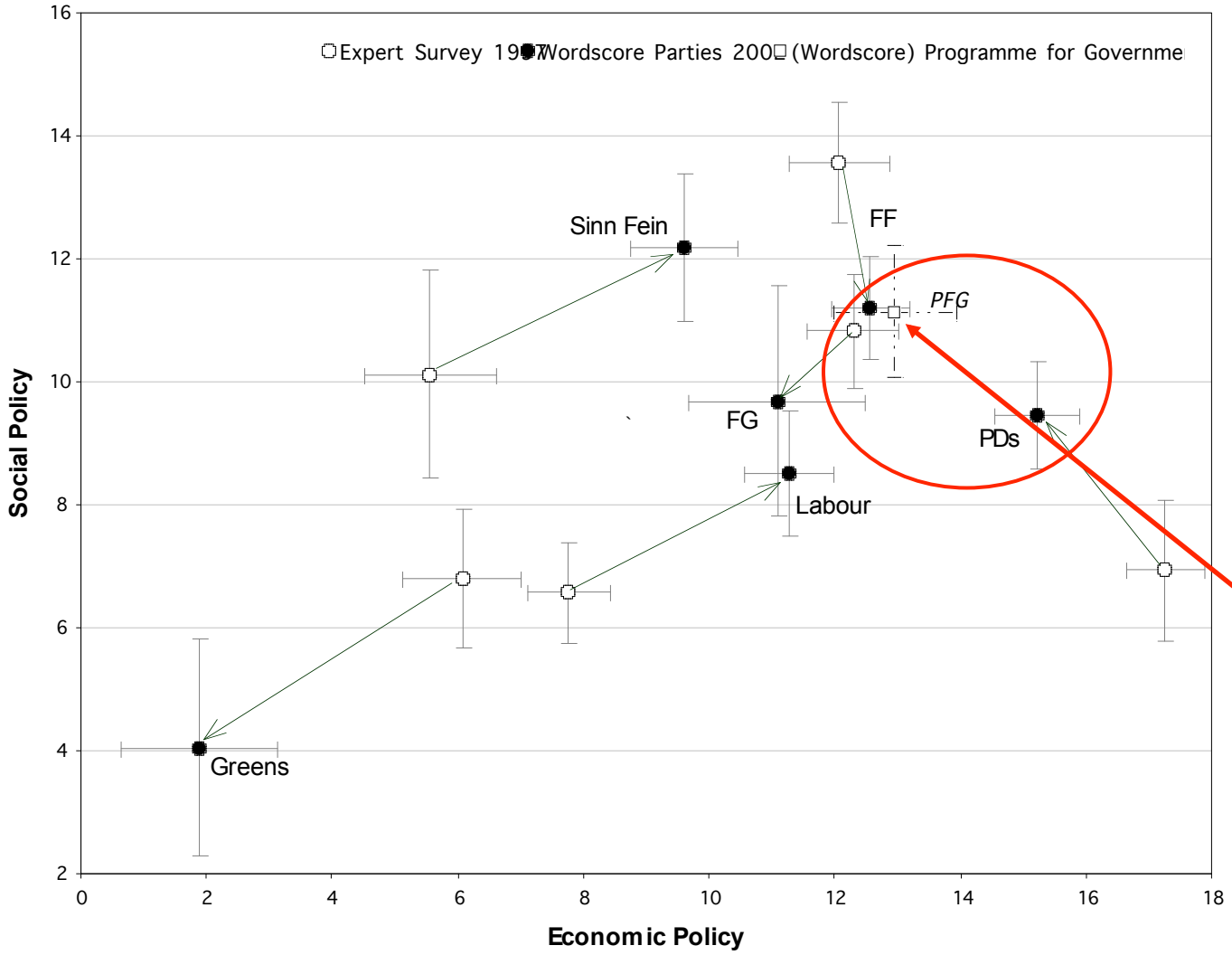
Application 4: Irish 2002 Manifestos



Economic and Social Results

- SF to center
- Greens to Left
- Labour and FG move together towards center
- **PDS to center**
- PFG between FF and PDs

Application 4: Irish 2002 Manifestos



Economic and Social Results

- SF to center
- Greens to Left
- Labour and FG move together towards center
- PDS to center
- **PFG between FF and PDs**

Meta-issues of Wordscores Approach

- Fully automated technique with minimal human intervention or judgment calls - only with regard to reference text selection
- Treats words as data; does not interpret them for meaning!!
  - Thus works in any language
  - And can generate statistical estimates of error
- Estimates unknown positions on *a priori* scales -- hence no inductive scaling with *a posteriori* interpretation of unknown policy space

Meta-issues of Wordscores Approach

- Does not burn information trying to find out what the key policy dimensions are
  - These are assumed in advance
  - An important aspect of research design
- Need valid and reliable estimates of – or confident assumptions about – positions of reference texts on dimension D
- May need transformation of raw text scores

Wordscores software

- `wordscores.pkg` for Stata, available from http://www.wordscores.com
- `Wordfreqj` command calls a Java runtime program to generate word frequency matrix
  - Can perform lemmatization
  - Can bootstrap sentences from texts
  - Extremely fast
- Very easy to install and use

Wordscores example

```
. clear
. net install http://www.politics.tcd.ie/wordscores/wordscores
. use http://www.tcd.ie/Political_Science/wordscores/files/APSR_uk9297
. setref lab92 5.35 ld92 8.21 con92 17.21  /* set ref. scores/texts */
. describetext *92 *97                     /* descr. stats on texts */
. wordscore economy                        /* score words for econ  */
. textscore economy lab97 ld97 con97       /* score virgin texts    */
```

Guidelines for use

- Texts need to contain information representing a clearly dimensional position
  - Dimension must be known *a priori*. Sources might include:
    - Survey scores or manifesto scores
    - Arbitrarily defined scales (e.g. -1.0 and 1.0 - more below)
  - Extreme texts on this dimension must be known - for reference anchors
  - Excludes things like scoring undergraduate essays, for instance, to validate grades!

Guidelines for use 2

- Need clearly defined reference values
  - Must be “known”
  - For any two texts, all scores are linear rescales - so might as well use [-1,1]
- Need reference texts that are carefully chosen
  - Same lexical universe as virgin texts
  - Should contain lots of words
  - Should be as discriminating as possible

Guidelines for use 3

- Guidelines for interpreting of virgin texts
  - May use LBG transformed scores if comparing to external sources of validation (e.g. expert surveys) - but not invariant to virgin text selection
  - With only two reference texts, may use Martin-Vanberg transformed scores - but not invariant to reference text word overlap
  - Raw scores are always invariant to virgin text selection and yield perfectly valid results for comparison
  - Key is to remember that virgin texts are *relative* scores
  - And ALWAYS consider confidence intervals