

Quantitative Research Methods IV - 17.806

Recitation, Week 3.

Topic: Unsupervised Learning II and Supervised Learning I.

Benjamín Muñoz

February 24, 2023

MIT

Table of contents

1. Supervised Learning

2. Problem Set Hints

1/ Supervised Learning

Supervised Learning

- Supervised ML: Predict the value of Y based on \mathbf{X} : $p(Y|\mathbf{X})$.

L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	g
dem	dlang	drel	durable	ef	ef2	ehet	elfo	elfo2	etdo4590	expgd	fedpol3	fuelexp	gdpgrowth	geo1	geo2	geo34	b
1	3	1	9	0.2549	0.064974	6	31	961	1	6.598187	0	8.721501	0.04505	0	0	0	0
1	3	1	10	0.2549	0.064974	6	31	961	1	6.303922	0	9.464923	0.044299	0	0	0	0
1	3	1	11	0.2549	0.064974	6	31	961	1	6.805417	0	10.48014	-0.04141	0	0	0	0
1	3	1	12	0.2549	0.064974	6	31	961	1	8.497777	0.13966	10.37251	0.041724	0	0	0	0
1	3	1	13	0.2549	0.064974	6	31	961	1	8.875194	0.13386	12.99692	0.067412	0	0	0	0
1	3	1	14	0.2549	0.064974	6	31	961	1	9.020614	0.132572	11.69704	0.025479	0	0	0	0
1	3	1	15	0.2549	0.064974	6	31	961	1	33.54913	0.132229	15.33493	-0.04604	0	0	0	0
0	23.1615	9.250124	16	0.440955	0.264102	38.93193	37.83801	2211.235	0.451153	33.58702	0.132183	15.31455	-0.01736	0.193761	0.097129	0.125969	(
1	6	6	20.70214	0.1338	0.017902	12	38.13281	2234.575	0.455181	33.70278	0.129773	15.65272	0.022349	0	1	0	0
1	6	6	0	0.1338	0.017902	12	38.20399	2243.33	0.453018	33.77172	0	15.76527	-0.42968	0	1	0	0
1	6	6	1	0.1338	0.017902	12	37.94931	2218.898	0.454227	33.74917	0	15.67645	-0.09924	0	1	0	0
1	6	6	2	0.1338	0.017902	12	37.95194	2220.186	0.453425	33.70349	0	15.58707	0.0495	0	1	0	0
1	6	6	3	0.1338	0.017902	12	37.85729	2211.733	0.453112	33.73692	0	15.50992	0.065588	0	1	0	0
0	6	6	0	0.1338	0.017902	12	37.94684	2219.891	0.451914	33.63824	0.132027	15.3354	0.054727	0	1	0	0
0	6	6	0	0.1338	0.017902	12	37.84507	2211.604	0.451948	33.5797	0.132613	15.31375	0.029936	0	1	0	0
0	6	6	1	0.1338	0.017902	12	37.81882	2208.832	0.452331	33.58608	0.132975	15.29686	0.070858	0	1	0	0
0	6	6	0	0.1338	0.017902	12	37.87026	2213.022	0.451946	33.61426	0.132769	15.28771	0.022549	0	1	0	0
0	23.15573	9.251361	1	0.441019	0.264049	38.94884	37.83476	2210.08	0.451206	33.59021	0.133037	15.31451	0.022583	0.194304	0.0974	0.126086	(
0	5	2	21.02088	0.148504	0.022053	12	37.89471	2211.006	0.45533	33.59899	0.132355	15.45635	-0.06047	0	0	0	0
1	5	2	0	0.148504	0.022053	12	37.88968	2211.608	0.454709	33.58667	1	15.35245	-0.04656	0	0	0	0
1	5	2	0	0.148504	0.022053	12	37.88434	2211.125	0.454627	33.58871	1	15.34972	0.010539	0	0	0	0
1	5	2	0	0.148504	0.022053	12	37.88583	2211.828	0.454255	33.57175	1	15.36617	0.045481	0	0	0	0
1	5	2	0	0.148504	0.022053	12	37.88069	2210.899	0.455037	33.56881	1	15.35526	0.038644	0	0	0	0
1	5	2	0	0.148504	0.022053	12	37.85563	2208.557	0.455008	33.58314	1	15.35867	0.048982	0	0	0	0
1	5	2	50	0.148504	0.022053	12	37.8391	2207.537	0.454194	33.60403	1	15.36008	-0.00988	0	0	0	0

Can you explain the following concepts in this example?

- Classification/regression, dimensionality.
- OLS, overfitting, cross-validation, regularization, Ridge/LASSO.

2/ Problem Set Hints

Problem Set Reminders

- Collaborators: You must note all of your collaborators. If you work alone, please indicate that as well.
- Annotating code: Please annotate your code thoroughly so that other people including me can easily follow each step. This is an important practice especially when you do collaborative research.
- Embedding code: If you use Rmarkdown or Rsweave, please make sure that lines do not get cut off. You can avoid this by tidy=TRUE option in code chunks.

Pset 1: Webscraping and regular expression

- Problem 1 and 2: You will use regular expression and scraping tools to download, structure, and name files.
- Procedure:
 1. Analyze the html by extracting tags and attributes that contain data
 2. Download data and use regex to name each downloaded data.
- Some tips
 - Vectorize (`stringr` functions accept a vector as input).
 - Read html file carefully.
 - Start with one string and repeat (try & error).
 - Use `Sys.sleep` to randomize your wait time.

Data Collection: Chicago

```
4:1 (Top Level) ▾
Console Background Jobs ▾
R 4.2.1 · ~/Documents/PSet_N1/
> head(df_url_1519)
# A tibble: 6 × 2
  date      url
  <date>    <chr>
1 2015-12-15 https://chicago.lististar.com/MeetingDetail.aspx?ID=449760&GUID=55EAFEE8-98D1-4B35-A130-F30D98362ADS&Options=info1&Search=
2 2015-12-14 https://chicago.lististar.com/MeetingDetail.aspx?ID=448641&GUID=4C91FCF5-0D66-4AF5-B2AC-6D7FE699A05&Options=info1&Search=
3 2015-12-09 https://chicago.lististar.com/MeetingDetail.aspx?ID=445253&GUID=DA07B999-1303-4B0F-9FD0-FDC65AA04009&Options=info1&Search=
4 2015-12-09 https://chicago.lististar.com/MeetingDetail.aspx?ID=448681&GUID=2F6FF8B5-B8C0-425C-A5A0-AF712084F3FC&Options=info1&Search=
5 2015-12-08 https://chicago.lististar.com/MeetingDetail.aspx?ID=445254&GUID=591024C3-407E-401C-B7B7-F31D952C28B0&Options=info1&Search=
6 2015-12-08 https://chicago.lististar.com/MeetingDetail.aspx?ID=445802&GUID=F977B654-1646-4345-B391-7BCBB531DA1C&Options=info1&Search=
>
```

- Explore different URLs available in the dataset:
 - Link First row.
 - Link Third row.
- HTML tags and CSS elements:
 1. td.
 2. nth-child.
- Use **Selector Gadget** and/or Inspect.
- Handy functions in the rvest and xml2 packages:
 1. `read_html:: read HTML file.`
 2. `html_element: find HTML element using CSS selectors or XPath expressions.`
 3. `html_attr: gets a single attribute.`
 4. `html_text2: get element text.`

chicago.legistar.com/MeetingDetail.aspx?ID=449760&GUID=55EAFFEE8-98D1-4B35-A130-F30D98362AD5&Options=Info&Search=

RPub - Estimatio... Chapter 9 Two Le... Meetings Álgebra Lineal - M... POISGA_Quant_II... Diagnostics for fix... Chapter 17 Advan... Survey_Experime... SYN Personalizar y controlar Google Chrome

Legislative Information Center | Chicago City Clerk

Council Home Legislation Meetings Legislative Bodies Council Members

[Share](#) [RSS](#) [Alerts](#)

Details

Meeting Name: Joint Committee: Human Relations; Public Safety (inactive) Agenda status: Final

Meeting date/time: 12/15/2015 10:00 AM Minutes status: Final

Location: Council Chambers -- City Hall -- Chicago, Illinois

Published agenda: [Amended Agenda](#) Published summary: [Summary](#) Meeting Extra: [Notice](#)

Meeting video: Not available/Not available

Attachments: [Original Agenda](#)

Meeting Items (5)

Record #	Ver.	Agenda #	Type	Title	Action	Result	Action Details	Captions
R2015-978	1	1	Resolution	Reassertion of need for examination of Department of Police practices and procedures	Recommended to Pass	Pass	Action details	Not available
R2015-974	1	2	Resolution	Call for public hearing(s) concerning Laquan McDonald case	Recommended to Pass	Pass	Action details	Not available
R2015-975	1	3	Resolution	Call for appointment of special prosecutor to represent Cook County in case against Jason Van Dyke on murder of Laquan McDonald	Recommended to Pass	Pass	Action details	Not available
R2015-976	1	4	Resolution	Call to hold hearing(s) and engage communities and leaders on police task force accountability	Recommended to Pass	Pass	Action details	Not available
O2015-8884	1	5	Ordinance	Amendment of Municipal Code Chapter 2-84-050 to establish and monitor police training programs	Recommended to Pass	Pass	Action details	Not available

← → ⌂ chicago.legistar.com/MeetingDetail.aspx?ID=449760&GUID=55EAFEE8-98D1-4B35-A130-F30D98362AD5&Options=Info&Search=

RPub - Estimatio... Chapter 9 Two Le... Meetings Álgebra Lineal - M... POLSGA_Quant_II... Diagnostics for fix... Chapter 17 Advan... Survey_Experime... SYNTH_R

Legislative Information Center

Chicago City Clerk - Office of the City Clerk - Action Details

Details

Record #: R2015978 Version: 1

Type: Resolution

Title: Reassertion of need for examination of Department of Police practices and procedures

Mover:

Result: Pass

Seconder:

Agenda note:

Action: Recommended to Pass

Action text: Recommended to Pass BY VOICE VOTE

Consent Votes (0:0)

0 records

Person Name	Vote
No records to display.	

Result	Action Details	Captions
Pass	Action details	Not available
Pass	Action details	Not available
Pass	Action details	Not available
Pass	Action details	Not available
Pass	Action details	Not available

<https://chicago.legistar.com/MeetingDetail.aspx?ID=449760&GUID=55EAFEE8-98D1-4B35-A130-F30D98362AD5&Options=Info&Search=#>

← → C chicago.legistar.com/MeetingDetail.aspx?ID=449760&GUID=55EAFFEE8-98D1-4B35-A130-F30D98362AD5&Options=info&Search=

RPPubs - Estimatio... Chapter 9 Two Le... Meetings Álgebra Lineal - M... POLSGA_Quant_I... Diagnostics for fix... Chapter 17 Advan... Survey_Experime... SYNTH_R

Legislative Information Center Chicago City Clerk

Council Home Legislation Meetings Legislative Bodies Council Members

Details

Meeting Name: Joint Committee: Human Relations; Public Safety. (inactive) **Agenda status:** Final

Meeting date/time: 12/15/2015 10:00 AM **Minutes status:** Final

Location: Council Chambers -- City Hall -- Chicago, Illinois

Published agenda: [Amended Agenda](#) **Published summary:** [Summary](#) **Meeting Extra1:** [Notice](#)

Meeting video: Not available/Not available

Attachments: [Original Agenda](#)

Meeting Items (5)

Record #	Ver.	Agenda #	Type	Title	Action	Result	Action Details	Captions
R2015-978	1	1	Resolution	Reassertion of need for examination of Department of Police practices and procedures	Recommended to Pass	Pass	Action details	Not available
R2015-974	1	2	Resolution	Call for public hearing(s) concerning Laquan McDonald case	Recommended to Pass	Pass	Action details	Not available
R2015-975	1	3	Resolution	Call for appointment of special prosecutor to represent Cook County in case against Jason Van Dyke on murder of Laquan McDonald	Recommended to Pass	Pass	Action details	Not available
R2015-976	1	4	Resolution	Call to hold hearing(s) and engage communities and leaders on police task force accountability	Recommended to Pass	Pass	Action details	Not available
Q2015-8884	1	5	Ordinance	Amendment of Municipal Code Chapter 2-84-050 to establish and monitor police training programs	Recommended to Pass	Pass	Action details	Not available

#ctl00_ContentPlaceHolder1_gridMain_ctl04_hyp Clear (1) Toggle Position XPath ? X

[https://chicago.legistar.com/MeetingDetail.aspx?ID=449760&GUID=55EAFFEE8-98D1-4B35-A130-F30D98362AD5&Options=info&Search#](https://chicago.legistar.com/MeetingDetail.aspx?ID=449760&GUID=55EAFFEE8-98D1-4B35-A130-F30D98362AD5&Options=info&Search=)

RPubs - Estimatio... Chapter 9 Two Le... Meetings Algebra Lineal - M... POLSGA_Quant_II... Diagnostics for f... Chapter 17 Advan... Survey_Experime... SYNTH_R

Dimensions: Responsive 583 x 568 100% No throttling

DevTools is now available in Spanish Always match Chrome's language Switch DevTools to Spanish Don't show again

Elements Console Sources Network Performance Memory Application

Styles >

City of Chicago * Office of the City Clerk

Legislative Information Center Chicago City Clerk

Council Home Legislation Meetings Legislative Bodies Council Members

Social Share RSS Alerts

Meeting Name: Joint Committee Human Relations Public Agenda status: Final

Meeting date/time: 12/13/2013 10:00 AM Minutes status: Final

Location: Council Chambers - City Hall - Chicago, Illinois

Published agenda: [Approved Agenda](#) Published summary: [Summary](#)

Meeting Status: [Not Available](#)

Meeting video: Not available/not available

Attachments: [Official Agenda](#)

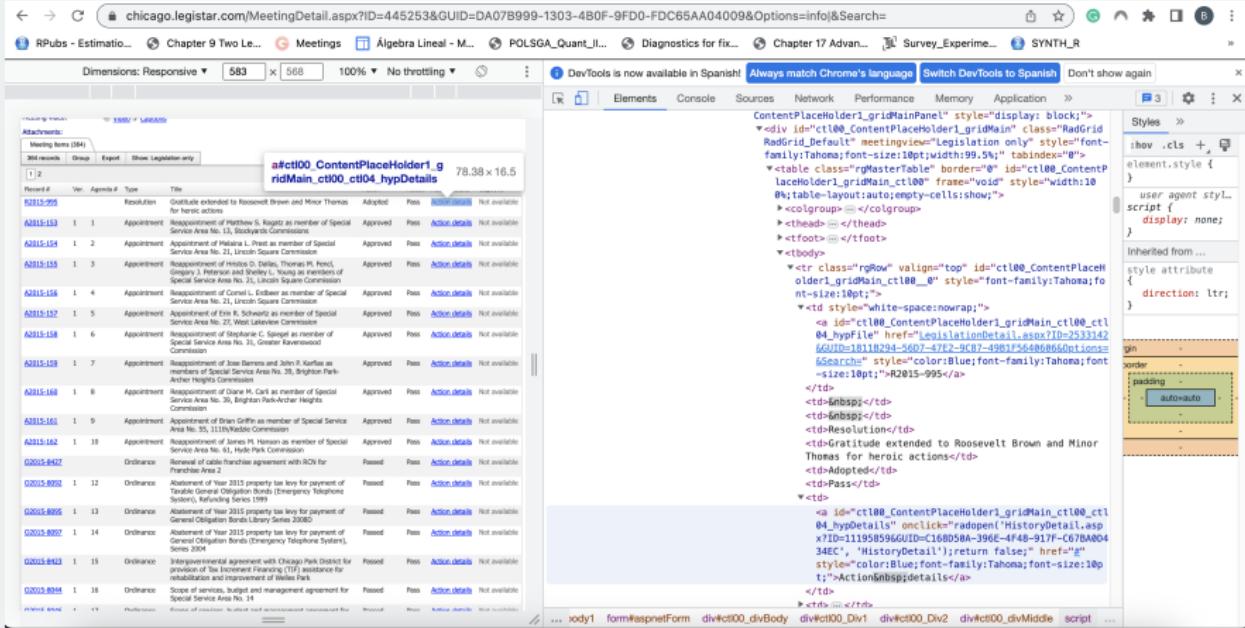
Meeting Items (5)

Record #	Group	Expert	Show Legislati... <th>only</th>	only
R2013-970	1	Agenda	Type	Title
R2013-970	1	Resolution	Resolution in support of need for reorganization of Department of Police practices and procedures.	
R2013-970	1	Resolution	Call for public hearing(s) concerning Laquan McDonald	
R2013-970	1	Resolution	Call for appointment of special prosecutor to re-examine Cook County in case against Jason Van Dyke in murder of Laquan McDonald	
R2013-970	1	Resolution	Resolution to engage communities and leaders on police task force accountability	
R2013-970	1	Ordinance	Amendment of Municipal Code Chapter 2-84-090 to establish and monitor police training programs	

aeft100_ContentPlaceHolder1_gridMain_ctl00_c104_hypDetails 78.38 x 16.5

... pageBottom.RadMultiPage.RadMultiPage_Default ...

Inherited from table



[Sign In](#)



City of Chicago ★ Office of the City Clerk

[Legislative Information Center](#)

[Chicago City Clerk](#)

[Council Home](#) [Legislation](#) [Meetings](#) [Legislative Bodies](#) [Council Members](#)

Details

Record #: [R2015995](#) Version:

Type: Resolution

Title: Gratitude extended to Roosevelt Brown and Minor Thomas for heroic actions

Mover: [Burke, Edward M.](#) Seconder: [Beale, Anthony](#)

Result: Pass

Agenda note:

Action: Adopted

Action text: On motion of Alderman Burke, and unsuccessful motion to reconsider the vote by Alderman Beale, the Resolution was Adopted by yeas and nays as follows:

[Votes \(50:0\)](#)

[50 records](#)

Person Name

Vote

[Moreno, Proco Joe](#)

Yea

[Hopkins, Brian](#)

Yea

[Dowell, Pat](#)

Yea

← → C chicago.legistar.com/HistoryDetail.aspx?ID=11195859&GUID=C168D50A-39E6-4F48-917F-C67BA0D434EC

RPub - Estimatio... Chapter 9 Two Le... Meetings Álgebra Lineal - M... PDLSCG_Quant_I... Diagnostics for... Chapter 17 Advan... Survey_Experime...

Dimensions: Responsive 423 x 481 100% ▾

DevTools is now available in Spanish Always match Chrome's language Switch DevTools to Spanish Don't show again

Elements Console Sources Network Performance Memory Application

Styles

:host .cls + ↴ element.style { }

user agent style... div { display: block; }

Inherited from ... style attribute { direction: ltr; }

grid - border - padding -

980x1732.070 -

City of Chicago • Office of the City Clerk

Lawsuit Details

Case Name: **Veron**

Version: Resolution

Title: Subtitle submitted to Roosevelt, Brooks and Mayor Rahm for heroic actions

Author: **Sonia_Eduard.S.** Seconded: **Beth_Arturo**

Agenda note:

Action: **Accepted**

Comments: On motion of Alderman Burke, and unsuccessful motion to reconsider the vote by Alderman Brooks, the Resolution was adopted by year and read as follows:

View History 30 months Filter

Author Status Date

Brooks, Brian Yes 7/26/2012

Givens, Bill Yes 7/26/2012

Harold, Leslie A. Yes 7/26/2012

Hartman, Leslie A. Yes 7/26/2012

Severin, Stephen T. Yes 7/26/2012

Shedd, Daniel L. Yes 7/26/2012

Watts, Michael A. Yes 7/26/2012

White, Anthony Yes 7/26/2012

Daley, Richard J. Sponsoring Member Since: 7/26/2012

Thomason, Patricia D. Yes 7/26/2012

Garcia, Dennis A. Yes 7/26/2012

Burke, Edward H. Yes 7/26/2012

Loew, Sherman A. Yes 7/26/2012

Ortiz, David J. Yes 7/26/2012

Rahm, David E. Yes 7/26/2012

Garrett, Michael J. Yes 7/26/2012

Cohen, Matthew J. Yes 7/26/2012

White, B. Debbie Yes 7/26/2012

Brickley, Michael Yes 7/26/2012

Flanagan, Michael J. Yes 7/26/2012

Jackson, Michael J. Yes 7/26/2012

Smith, Michael J. Yes 7/26/2012

html:chrome,t-chrome110 body#ctl00_body1 form#aspnetForm div#ctl00_divBody

```
</div>
<div id="ctl00_divLeft"></div>
<div id="ctl00_Div1">
<div id="ctl00_Div2">
<div id="ctl00_divMiddle">
<script type="text/JavaScript" src="GridFilter.js"></script>
<div id="ctl00_ContentPlaceHolder1_RadAjaxManager1SU" style="display: none;"></div>
</div>
<div id="ctl00_ContentPlaceHolder1_RadAjaxLoadingPanel1" style="display:none;height:75px;width:75px;"></div>
<div id="ctl00_ContentPlaceHolder1_tabTop" class="RadTabStrip RadTabStrip_Default R adTabStrip_DotNet RadTabStripTop" style="background-color:White;font-family:Tah om;font-size:10pt;width:100%;"></div>
<div id="ctl00_ContentPlaceHolder1_MultiPageTop" class="RadMultiPage RadMultiPage_D efault" style="width:99.8%;">
<div id="ctl00_ContentPlaceHolder1_pageTop1" class="rmpView" style="background-co lor:White;border-color:Silver;">
<table border="0" cellpadding="2" cellspacing="2" width="100%">
<tbody>
<tr>
<td valign="top" style="width: 150px"></td>
<td valign="top">
<a href="LegislationDetail.aspx?ID=11195859&GUID=C168D50A-39E6-4F48-917F-C67BA0D434EC" style="color:Blue;font-family:Tahoma;font-size:10pt;">R2015995</a>
<br/>
<span id="ctl00_ContentPlaceHolder1_lblVersion1" style="color:Navy;font-family:Tahoma;font-size:10pt;">-Version</span>
<span id="ctl00_ContentPlaceHolder1_lblVersion2" style="color:Black;font-family:Tahoma;font-size:10pt;"></span>

```

chicago.legistar.com/HistoryDetail.aspx?ID=11195859&GUID=C168D50A-396E-4F48-917F-C76BA0D434EC

RPubs - Estimatio... Chapter 9 Two Le... Meetings Algebra Lineal - M... POL5GA_Quant_I... Diagnostics for fix... Chapter 17 Advan... Survey_Experime...

Dimensions: Responsive ▾ 423 x 481 100% ▾

DevTools is now available in Spanish! Always match Chrome's language Switch DevTools to Spanish Don't show again

Elements Console Sources Network Performance Memory Application

City of Chicago Office of the City Clerk
Legislative Information Center Classes On Click
Committee Agendas Meeting Legislative Rules Council Members
Record ID: FASCE000000 Version: 1
Type: Resolution
Title: Continue extension to Roosevelt, Brown and River Thomas for heroic actions

div#ctl00_ContentPlaceHolder
1_MultiPageBottom.RadMultiP
age.RadMultiPage_Default 978.04 x 1254

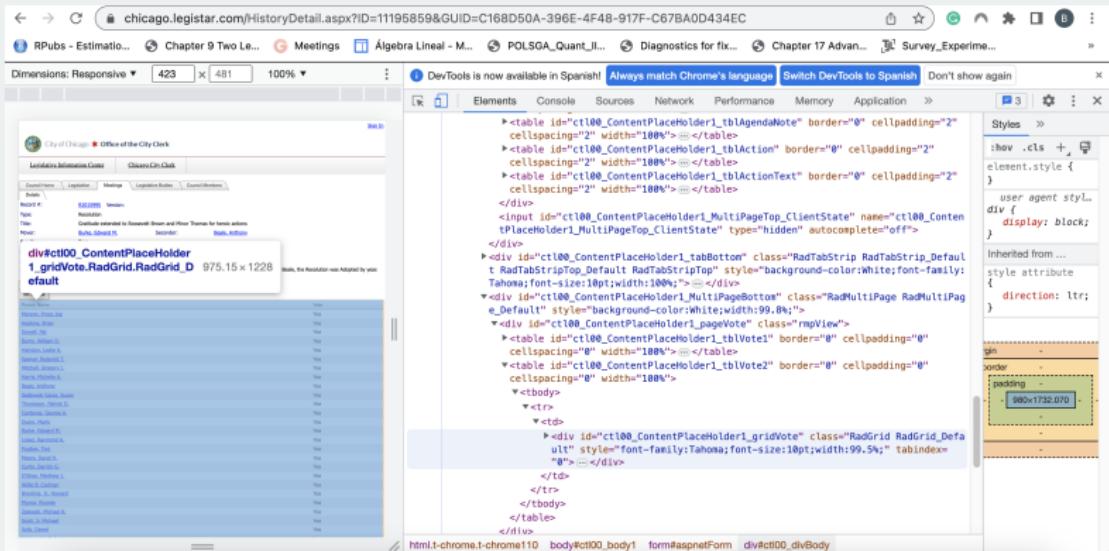
Make the resolution was adopted by your

Person Name Title
Hanson, Peter J.D. Hon.
Hanson, Peter Hon.
Gerrard, Mr. Hon.
Bartek, Michael Hon.
Lindquist, John G. Hon.
Sorenson, Andrew J. Hon.
Mehlman, Daniel J. Hon.
Hicks, Henry A. Hon.
Rosen, Andrew Hon.
Kaufman, David S. Hon.
Thompson, Steve D. Hon.
Gordon, Jason A. Hon.
Jones, Brian Hon.
Bolton, Jennifer E. Hon.
Lyon, Kenneth R. Hon.
Lindquist, John G. Hon.
Hanson, Peter A. Hon.
Gillis, Dennis L. Hon.
Cohen, Michael A. Hon.
Wilkis, Justine Hon.
Bartek, Michael A. Hon.
Jones, Brian Hon.
Gordon, Michael A. Hon.
Bolton, Jennifer E. Hon.

html:1 chrome:1 chrome:110 body:ctl00_body1 form#aspnetForm div#ctl00_divBody

Styles

```
:host .cls + _ element.style { }  
user agent styl...  
div { display: block; }  
Inherited from ...  
style attribute { direction: ltr; }  
span { border: 1px solid black; padding: 2px; width: 980px; height: 1732.07px; }  
border - padding - width - height -
```



Data Collection: New York

- Steps to collect data with API:
 1. Sign up for an app token/key.
 2. Identify the API endpoint.
 3. Filter the dataset (select columns and rows).
 4. Download selected data.
- Client: nyc.
- Explore Legistar Help.
- Example:
 - "<https://webapi.legistar.com/v1/nyc/events?token=XXXX>"

Principal Component Analysis

- Datasets have different types of information (and columns have different names).
- The New York dataset contains rows with different EventItems (not only votes). It is preferable to filter them.
- Original data: long format. PCA: wide format.
- Many votes are unanimous (or the minority is less than 10Methodological decisions:
 - Number of authorities.
 - Number of votes (the important thing is plenary, over the committees).
 - Management of missing values.
 - Substantive interpretation.