Institute of Policy Studies

ENGAGING MINDS, EXCHANGING IDEAS

Talk by
Dr Marko Skoric
on
"Public Opinion Research in the Age of Social Media"

Friday, 10 April 2015

IPS Meeting Room

Lee Kuan Yew School of Public Policy

National University of Singapore

Public Opinion Research in the Age of Social Media

Marko M. Skoric
Department of Media and Communication
City University of Hong Kong
Hong Kong S.A.R., China

Institute of Policy Studies, Singapore, April 10th, 2015

Agenda for Today's Talk

- Evolution in social scientific method
- Nature of survey vs. social media data
- Do social media-based predictions of public opinion and political behavior work?
  - Representativeness vs. node importance (popularity)
  - Which methods work best? Why?
- Future research directions

Evolution in Social Sciences: From Scarcity to Abundance

- (R)evolution in social sciences: from data scarcity to data abundance (Lazer et al., 2009)
  - Human interactions increasingly happening in a technology-mediated context
    - Digitally recorded by default
  - Dramatic increase in the capacity to store, process and analyze data
- But, is it the same type of data?
  - Traditionally, social scientists “created” their data
  - Computational social scientists use “found” data

Surveys vs. Social Media Analytics

* Survey as a <u>structured</u>, <u>systematic</u> method of collecting fairly large number of <u>solicited</u> responses (“created data”)
  * Survey data criticized for being based on “self-reports”
    * Socially undesirable opinions and behaviors may be underreported
  * (Preferably) utilizing <u>probability</u> sampling
    * Ability to generalize based on sound statistical/mathematical principles
* Social media data as <u>unsolicited</u>, <u>unstructured</u> streams of social conversation in different modalities (“found data”)
  * Word-of-mouth (WoM) messages, diaries, ethnography
  * Non-probability data corpuses

Social Media Analytics

* Low(er) cost
* (Near) real-time analysis
* Greater variety of topics and contexts
* Unobtrusive measures
* Continuous, longitudinal, panel type data
* Cross-national/comparative data
* Captures <u>the structure</u>, not just the content
  * Behavioral logs
  * Social conversations (text, audio, & video)
  * Social networks and relationships
* Universe/corpus vs. sample

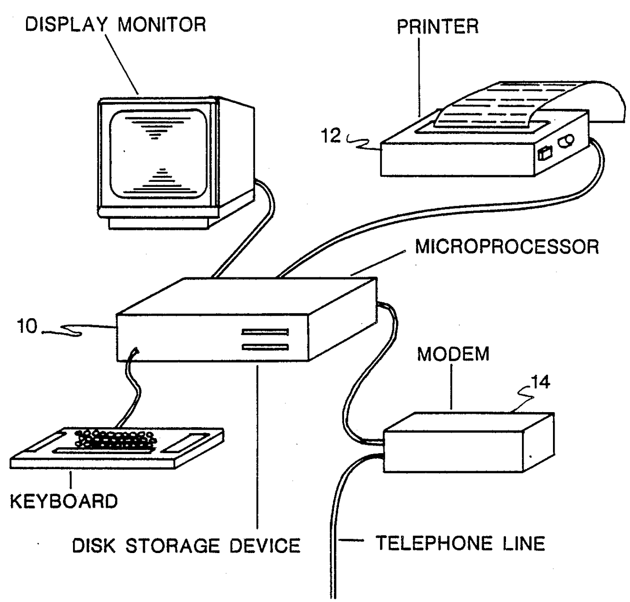
Public Opinion Research and Information Technology

* Almost 80 years of scientific public opinion research
* Co-evolution of technology and survey research
  * Random digital dialing (RDD), computer-assisted telephone interviewing (CATI)
* Still, two basic principles have not changed (until recently, at least)
  1. Probability-based sampling
     *But are all opinions worth the same?
  2. Structured, solicited nature of survey data
* Alternative approaches included content analysis

Fan, D. P. (1990). Information processing expert system for text analysis and predicting public opinion based on information available to the public. U.S. Patent 4,930,077.

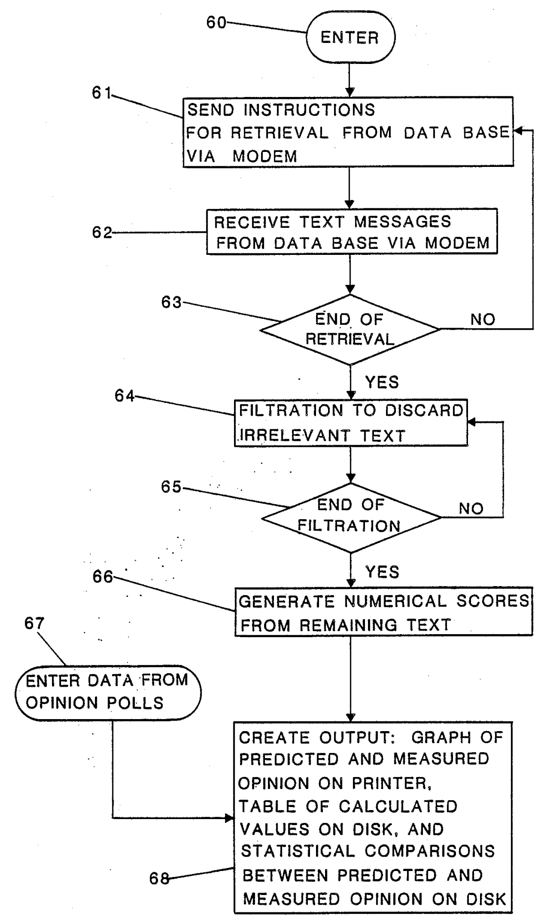
U.S. Patent May 29, 1990 Sheet 1 of 3 4,930,077

FIGURE 1



U.S. Patent May 29, 1990 Sheet 3 of 3 4,930,077

FIGURE 3



Surveys vs. Social Media Data

	Survey data	Social media data
Data description	Content (attention, affect, value, attitude)	Content (attention, affect, value, attitude), structure/network of communicators
Data collection	Structured data (reactive expressions solicited by researchers)	Organic data (unsolicited, unstructured, non-reactive expression: monologue or dialogue)
Social structure	Unknown	Generally known
Size of data	Limited size of data (thousands) that can be handled by traditional data processing applications	Large (millions to billions) that requires greater computational capacity
Statistic assumption	Probability sampling	Non-probability sampling or census

Types and Characteristics of Social Media Data

Category	Examples	Data characteristics	
		Content	Structure
Social network sites/microblogs	Twitter, Facebook, Sina Weibo	Attention, affect, value, attitude	Connections between people
Online forums	Usenet, Tinaya, Uwants	Attention, affect, value, attitude	Pathways of online discussion
Video/picture sharing websites	YouTube, Flicker, Panoramio, Instagram	Attention, implicit affect, attitude	Connections between people or content
Blogs	LiveJournal, BlogPulse, Sina Blogs	Attention, affect, value, attitude	Connections between people or content
Online encyclopedias	Wikipedia, Debatepedia	Attention(topics), opinion	Connections between content (hyperlink)

Challenges of Social Media Analytics

- Data collection
  - Know-how
  - Open vs. closed/limited/changing API
  - Scale
- Sampling and representativeness
  - Social media users are different from average citizens
  - Self-selection issues
  - Spam and astroturfing
- Analysis
  - Techniques and approaches (problems with black-boxing)
  - Statistical assumptions are often violated; which ones?
  - Scalability/computational issues
- Lack of good theories – data-driven research is dominant

Challenges of Social Media Analytics- continued

* Is “bigger data” better data?
  * Self-explanatory?
  * Methodologically sound?
    * Are messages on social media platforms genuine and authentic?
      * Or curated, managed, edited?
      * Long-tail of participation and content creation
        * 80/20 or 90/9/1 rule
    * Is Twitter data collected via APIs representative? If so, of what?
      * “Firehose”, “gardenhose” & “spritzer” types of access
      * API characteristics may shift over time (without warning)

Challenges of Social Media Analytics- continued

* (Really) big data is mostly proprietary or owned by governments (& national security agencies)
  * Lack of proper data-sharing norms, protocols and procedures
  * Only big players have the privilege of full access
  * Difficulty in replicating findings
* Supply of social media data varies highly across different societies and contexts
  * Dependent on level of technological/infrastructural development
    * "Big data" divide?

Predicting Elections and Public Opinion Using Social Media Data: A Meta-Analysis

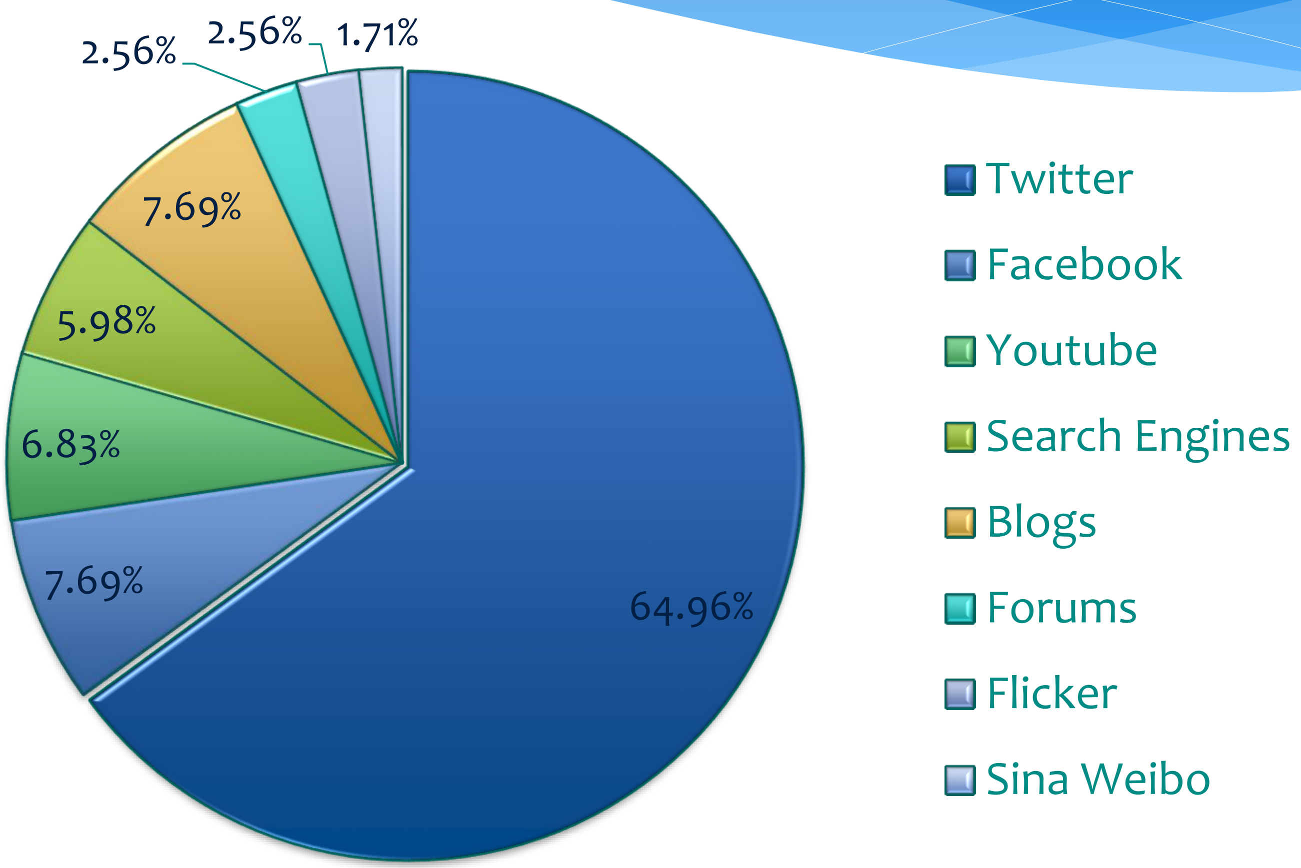
(Skoric et al., 2015)

Public Opinion, Political Behavior & Social Media: The Literature Search

44 studies identified in SCI, IEEE, ACM, AAAI & ComAbstracts.

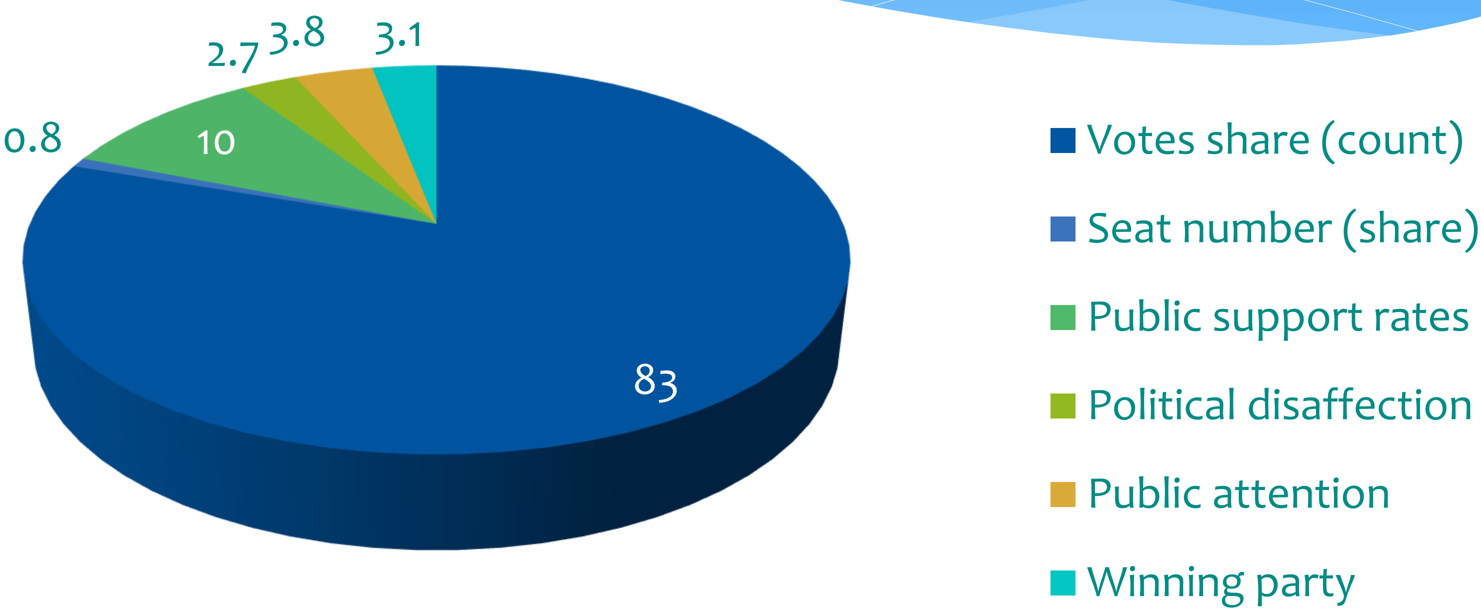
- Polls/surveys & offline behavior
  - **Opinion (7)**
    - Voting intentions
    - Political evaluations
  - **Behavior (37)**
    - Electoral outcomes
    - Protests/terror attacks

Data Sources



What Do These Studies Predict?

(percentage in estimates)



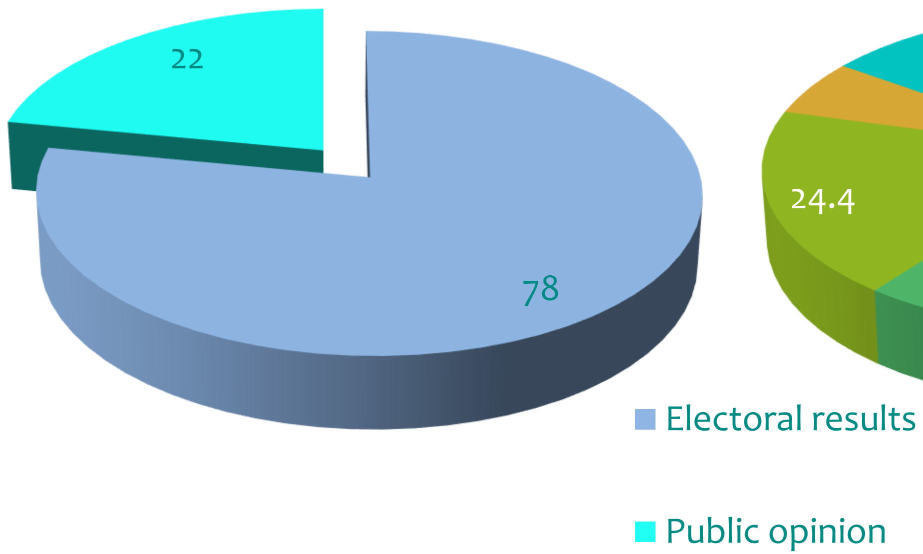
Within the 259 estimates, 215 predict the vote share(counts), 26 predict public support rates, 8 predict the winning party, 7 predict political disaffection, 2 predict seat number(share), 1 predict public attention.

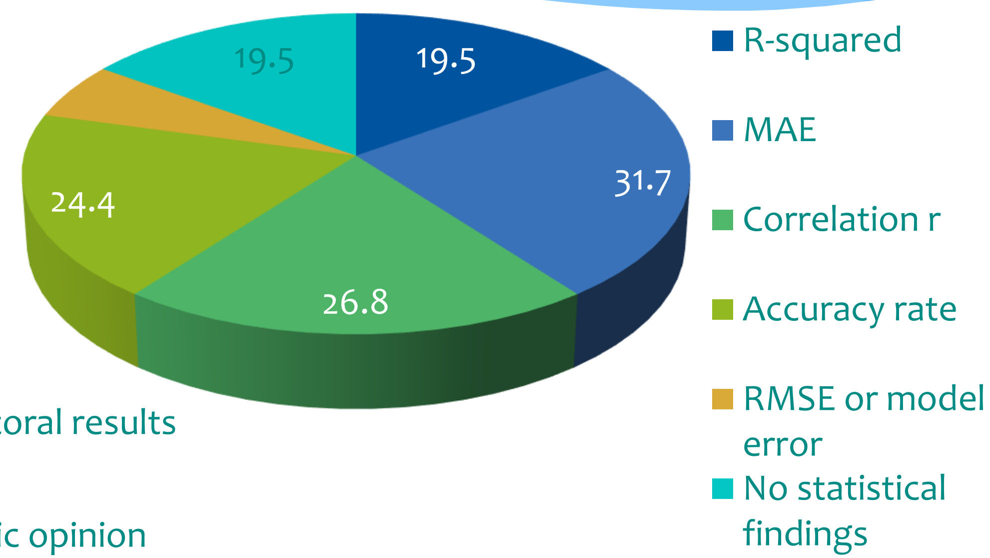
How Do They Make Predictions?

* Volume-based analysis (frequency counts)
  * Number of tweets, retweets, mentions, likes, dislikes, frequency/growth rate of posts, etc.
* Sentiment analysis
  * Extracting sentiment, emotions, affect via the use of lexicons or machine learning
* Network analysis
  * Opinion leadership: network of mentions, follower/followee relationships, centrality of nodes

Measurement of Predictive Power

(at a study level, in %)





41 studies: 32 predict election results, while 9 predict public opinion.
41 studies: 8 studies report (adjusted) R^2; 13 studies report MAE; 11 report a correlation coefficient; 10 report accuracy rate; 3 report other measures (RMSE, model error, etc.); and 8 report no statistical findings.

Predictions, Predictive Power and Predictive Models

Predictors	Predictions	Predictive power	Predictive model
Volume-based analysis	Wining party or candidate	Mean Absolute Error (MAE)	Correlation
Sentiment analysis	Candidates' vote share, Parties' number of seats	R-squared	OLS regression
Network analysis	Presidential approval rates Candidates' popularity Political disaffection	Correlation coefficient Race-based accuracy rate classification accuracy RMSE, model error, etc.	ANOVA

Tweets and Votes in 2011 Singapore General Election

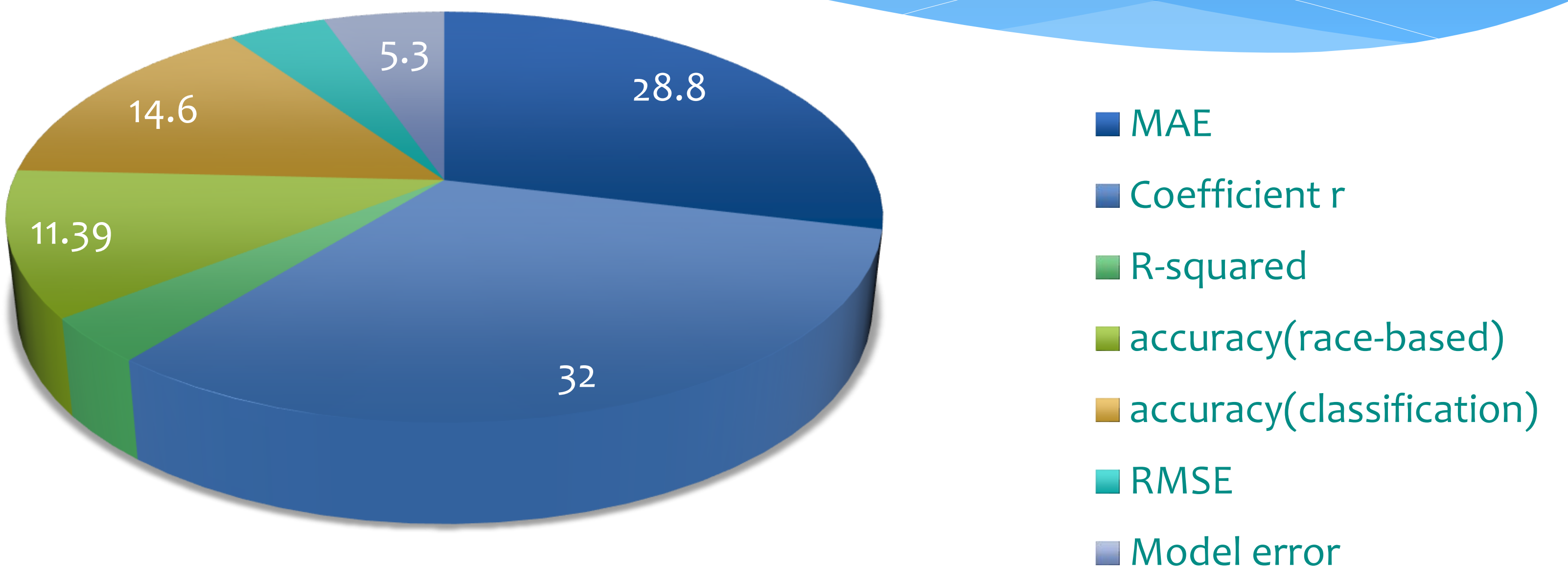
(Skoric et al., 2012)

Party	% Tweets	% Votes	Error
PAP	42.80 (1)	60.14 (1)	-17.34
WP	20.83 (2)	12.83 (2)	8.00
NSP	13.86 (3)	12.04 (3)	1.82
SDP	11.07 (4)	4.83 (4)	6.24
RP	5.22 (5)	4.28 (5)	0.94
SPP	4.41 (6)	3.11 (6)	1.30
SDA	1.81 (7)	2.78 (7)	-0.97
MAE			**5.23**

Numbers in parentheses indicate relative rank.
MAE = mean absolute error.

Measurement of Predictive Power

(estimates, in %)



259 estimates: 81 report MAE; 90 report coefficient; 10 report R^2; 32 report race-based accuracy; 41 report classification accuracy; while 12 report RMSE and 15 report model error (based which we computed 5 estimates reporting MAE).

Predictive Power of Social Media:
Elections and Public Opinion

Predictors	MAE (% error)			R^2			Coefficient r			Race-based accuracy (%)			Classification accuracy (%)		
	Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD
Volume-based analysis	**6.80**	38	6.18	.41	3	.25	.69	54	.33	63.79	10	11.07	66.90	24	11.06
Sentiment analysis	8.04	48	3.24	.72	5	.12	.44	24	.30	63.74	11	16.77	67.52	5	24.93
Network analysis										66.99	8	7.70	72.68	8	11.61
Volume-based + network analysis										**85.83**	3	2.02	**79.40**	4	12.47
Volume-based + sentiment analysis				**.89**	2	.00	.77	12	.21						

Statistical Comparison of the Predictive Power across Different Outcomes

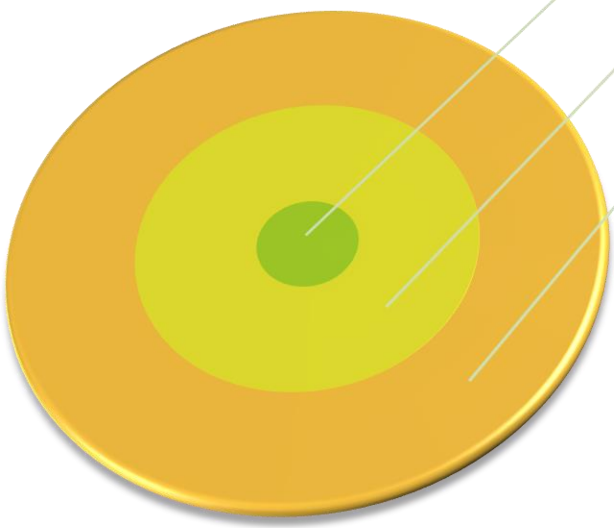
Predictions	MAE (% error)			R^2			Coefficient r			Race-based accuracy (%)			Classification accuracy (%)		
	Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD	Mean	N	SD
Votes share	7.54	77	4.89	.51	4	.28	.64	69	.32	65.11	27	13.70	68.24	38	12.82
Seat number	**1.67**	2	0.47												
Public support	8.59	7	2.86	**.76**	6	.14	.48	13	.36						
Disaffection							.80	7	.13						
Public attention							**.97**	1							
Winning party										**74.88**	5	9.15	**83.00**	3	18.33
Total	7.49	86	4.78	.66	10	.23	.63	90	.33	66.64	32	13.46	69.32	41	13.57

Findings and Discussion

* Multiple approaches outperform single approach in predicting real-life public opinion and electoral results
* Network analysis outperforms both volume-based and sentiment analysis
  * Social structure tells more than the content?
    * More stable, less error?
* Why are volume-based analysis and sentiment analysis less accurate?
  * Problems in the measurement or conceptualization?

Why Could Structure Tell Us More Than the Content?

Getting data from participants



Active users

Inactive but connected users

Inactive and unconnected users

Analysis of content

Analysis of structure

Solicited data (surveys)

For example, the number of registered users of Sina Weibo exceeded 500 million , while the monthly active users is 156.5 million as of June, 2014.

Future Opportunities

* Triangulation of traditional and social media mining methods
  * Establishing validity of measures
    * Surveys combined with social media analytics
  * Combining human coders and machine-learning algorithms (crowdsourcing, e.g. Amazon's Mechanical Turk)
* Developing standardized sets of methods and procedures for data collection, processing and analysis
  * Preserving comparability and allowing for replicability
    * Data sharing
* Doing RQ or theory-driven research; opinion dynamics

Future Challenges

* From "public-by-default" to "private-by-default"
  * SNS (Facebook, Twitter) → IM (WhatsApp, KakaoTalk)
    * Loss of open APIs
* Will we be able to find "found" data in the future?
* Should access to social media data be legally mandated?
  * Who should mandated it?
    * National or global level?
  * For research purposes only? Aggregated and anonymized?

Conclusion

* The scientific puzzle and the methodology designed to solve it are intimately linked (Kuhn, 1962)
* Social media-based predictions should not substitute, but complement existing methods
* Social media analytics may be better suited for understanding the dynamics of opinion change and for identifying opinion leaders
  * Predicting the future, rather than assessing the present

Political expression, exposure to disagreement and opinion shielding on social media:

Survey evidence from Singapore and Hong Kong

Exposure to Political Disagreement and Citizen Participation

* Homogeneous environments are ideal for encouraging political participation (i.e. voting), by reinforcing opinions and promoting recognition of common problems
  * Exposure to political difference depresses voting because of increased social costs and political ambivalence (Mutz, 2002)
* Exposure to countervailing views has a negative impact on the likelihood of voting but encourages other forms of participation
  * e.g., voluntary activities and future involvement in either political activism or party politics (Pattie and Johnston, 2009)
* Exposure to a cross-cutting online network may yield different impact, depending on the forms of participation
  * Partisan-related activities vs. community-related engagement

Political Expression, Exposure to Disagreement and Opinion Shielding in Singapore

(Skoric et al, 2014)

* Political (partisan) participation is linked with both political expression and shielding oneself from disagreement on Facebook
* Exposure to political disagreement on Facebook is positively associated with certain forms of civic engagement
  * Donating money and doing voluntary work
    * Marginally associated with boycotting and signing a petition
  * Shielding oneself negatively associated with donating money to civic groups

Political Expression, Polarization and Opinion Shielding during 2014 Hong Kong Protests

(Skoric et al., 2015)

- Students who spent more days and nights at the protests and engaged in more protest activities used Facebook and online forums more extensively to mobilize, express their opinions and discuss protest-related issues
  - Frequency of hiding posts and comments with dissenting views, deleting Facebook friends, and using uncivil language was generally low, but was still significantly higher among the more engaged group of student protesters
    - High levels of offline participation during the protests was related to intensified shielding from dissenting or critical views online

Thank you!

mskoric@cityu.edu.hk

Questions? Comments?