



Opportunities and Challenges in an Era of Big Data

Ginger Zhe Jin
University of Maryland & NBER



http://www.

Data and model structure are substitutes.

--- Thomas Hubbard, 1996

Structure is an inferior substitute to data,
though sometimes a necessary and informative
substitute.

--- Thomas Hubbard, 2012

Where to find data?



- Giant internet companies
 - Google, eBay, Amazon, Yahoo, Microsoft
- Price comparison websites
 - Priceline, Orbitz, Expedia
- User-generated reviews
 - Yelp, Tripadvisor
- Companies that collect and sell data from the Internet
 - Keywordspy, Terapeak
- Open Government
 - FAA: air traffic data, airline on-time ratings
 - FCC: spectrum auctions, internet access
 - FTC: Do not call list
 - EPA: pollution data
 - Dept of Education: NCES
 - CDC: National Center for Health Statistics
 - CMS: Medicare health utilization data
 - AHRQ: Hcup data
 - FDA: orange book, Adverse Event Reporting System

Our roles in data



- As a consumer of data
- More active roles:
 - Design data structure and interface (Ghose et al. 2012)
 - Make data available to others (Allcott and Rogers 2012)
 - Translate data into strategy and policy
 - Study transparency itself (Harvard's Transparency Policy Project)
- Private sector is meeting the challenge
 - Proprietary data
 - more on a single firm than on the whole picture of an industry/market
- Public sector is another story
 - Potential open to the public via FOIA
 - On many firms/consumers/regulators
 - Generating, organizing and disseminating data is not the main business of governments

An illustrative example



On-going project:

Digital Disclosure of a Nationally Standardized Database of Restaurant Food Safety Inspections

PI: Ben Bederson, Ginger Jin, and Phillip Leslie

Funded by: the Alfred P. Sloan Foundation

Challenge #1

http://www.



- Data Quality
 - Sample coverage
 - Variable Definition
 - Consistency
 - Identity tracking
 - Interpretation

Example of New York City

<http://www.nyc.gov/health>



- Count violations
- 0-13 points = A
- 14-27 points = B
- 28 or more points = C
- A restaurants inspected every 12 months
- B restaurants inspected every 6 months
- C restaurants inspected every 4 months



Response from Industry



- “Restaurateurs Beg City to Drop Report Card Plan” (NY Times Feb. 5, 2010)
- Peter Hansen at Benchmark Restaurants
 - “It could be a cracked toilet cover; it could be a gap around a pipe, but what your customer is thinking is: old tuna.” (NY Times Mar. 6, 2012)
- Elizabeth Meltz at Mario Batali’s Italian restaurants
 - some city health inspectors seemed inconsistent in their standards, sometimes appeared unfamiliar with complex dishes (NY Times Mar. 6, 2012)
- Per Se
 - consistently rated among the city’s best restaurants, had avoided a “B” grade with a telephone call to a city official, in which the restaurant successfully argued that the inspection report had contained errors. (New York Post Mar. 3 2012)

Response from consumers and local government



- Bloomberg: denounced critics as “people that complain because they don’t want to keep their restaurants clean.” (NY Times Mar. 6, 2012)
- NY local health department: “Restaurant Grading in New York City at 18 Months”
 - 81% report seeing the letter grades
 - 88% consider grades in dining decisions
 - 70% concerned about getting sick from eating out
 - Seeing A makes 76% more confident about the safety of eating at a restaurant
 - 21% believe NYC inspected four or more times per year
 - 88% approve frequent inspection of poorly performing restaurants
 - % of no trained food protection supervisor drops from 13% to 7%
 - Presence of mice drops from 32% to 22%
 - Inadequate hand washing facilities drop from 11% to 5%
 - Reported cases of Salmonella infections drop 14%

Response from researchers



- Daniel Ho “Improve Restaurant Report Cards” (NY Times Opinion Mar. 7, 2012)

“Along with researchers at New York University, Stanford and Yale law schools, I have studied data from more than 500,000 inspections of more than 100,000 restaurants from the last few years in nine jurisdictions. ... Our research examined the process for tallying violations and the consistency of inspections across repeat, unannounced visits to the same restaurant. In a critical dimension, New York performed the worst of the nine.”

Some Questions



- Should data be disclosed to the public even if they are imperfect?
 - No, because they could be misleading
 - Yes, the best way to improve data quality is transparency
- Users may have different and even conflicting demand for data quality
 - How to balance between users?
 - Target which users?

Challenge #2

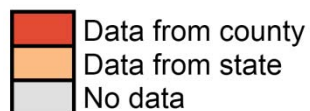
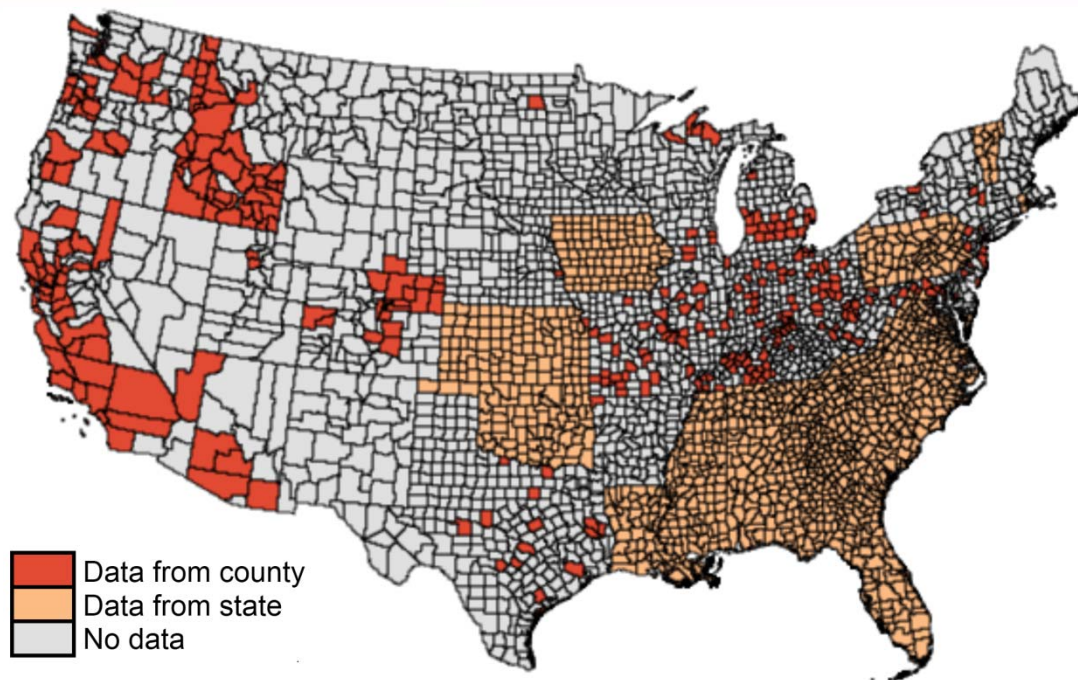
<http://www.>



Diversity and fragmentation
across local jurisdictions

Data available from local jurisdictions

<http://www.>



Available in:

16 states
1534 counties,

Cover 61.4%
population

Existing efforts in standardization



- FDA efforts
 - FDA food code since 1993 (as a guidance instead of requirement)
 - FDA sample inspection form
 - FDA training of inspectors
 - Periodical study nationwide
 - Food Safety and Modernization Act (January, 2011)
- Industry efforts
 - The Association of Food and Drug Officials has existed since 1896
 - The Conference for Food Protection
- Third party efforts
 - Findthedata.org, city-data.com, etc.

Our Approach



- Define eight crude categories
 - minimize controversy
- Let users do sophisticated and customized searches

Our sample so far

http://www.



Restaurant Food Safety: Stats - Mozilla Firefox

Restaurant Food Safety: Stats

hcl2.cs.umd.edu/aq/rfs/stats/

CONFIDENTIAL - INTERNAL USE ONLY

restaurantinspections [about](#)

Latest Stats

all time recent

	availability	days	establishments	inspections	violations per inspection
Anne Arundel	1/2012~1/2012	28	392	414	0.12
Chicago	1/2010~2/2012	778	16,446	38,314	3.13
District of Columbia	3/2009~3/2012	1,092	3,777	8,620	3.78
Florida	7/2011~9/2012	431	47,537	98,273	4.66
King	1/2006~2/2012	2,231	9,248	102,814	0.94
Los Angeles	2/2009~2/2012	1,087	16,823	48,744	1.37
New York	9/2007~1/2012	1,583	23,504	144,483	3.46
San Francisco	9/1998~3/2012	4,926	6,243	96,326	1.16
Virginia	7/2002~10/2012	3,720	23,078	144,581	3.87
TOTAL			147,048	682,569	2.85

Full sample comparison

<http://www.>



	# of Violations per inspection	Violation Type Distribution				
		Employee	Temperature	Contamination	Physical facilities	Other 4 Categories
Ann Arundel	0.12	2%	23%	21%	54%	0%
Chicago	3.13	3%	6%	32%	55%	3%
Florida	4.66	8%	11%	43%	25%	12%
King County	0.94	21%	26%	40%	8%	5%
Los Angeles	1.37	2%	7%	46%	42%	3%
New York City	3.44	6%	15%	43%	27%	6%

What about McDonalds only?

<http://www.>



	# of Violations per inspection	Violation Type Distribution				
		Employee	Temperature	Contamination	Physical facilities	Other 4 Categories
Chicago	3.19	4%	8%	32%	56%	1%
Florida	4.85	9%	7%	47%	28%	8%
King County	0.66	19%	22%	32%	21%	5%
Los Angeles	1.22	3%	6%	48%	42%	2%
New York City	2.81	4%	12%	41%	34%	4%

Anova Analysis over all chains (R2)



Controlled factors	# of violations	% of Violations Per Category			
		Employee	Temperature	Contamination	Facilities
Season	0.01	0.00	0.00	0.01	0.00
Chain	0.17	0.02	0.05	0.05	0.06
Jurisdiction	0.22	0.08	0.11	0.22	0.22
Jurisdiction, chain, season	0.28	0.09	0.13	0.23	0.24
City/county, chain, season	0.32	0.11	0.14	0.25	0.24
Restaurant, season	0.51	0.14	0.16	0.30	0.32

Some Questions



- Why are there so many variations across jurisdictions despite FDA-standardized guidance and training?
- Why are some jurisdictions by state, and some by county/city?
- A need for standardization?
 - No, local jurisdictions have better knowledge on local food safety problems
 - Yes, chains strive to standardize, and fragmentation will generate inefficient compliance and frustration among data users
- Who should standardize the data and how to?

Challenge #3: User-friendliness



- Open to whom?
- In what format?
- By what time frame?
- Examples:
 - [Macomb, MI](#)
 - [North Carolina](#)
 - [New York City, NY](#)

Macomb County, MI

http://www.

Macomb County Michigan : Health Department: food inspections - Mozilla Firefox

ode Anova - zouben@... Macomb County Michigan : ... New Tab

www.macombcountymi.gov/publichealth/eh/FoodInspections.htm

Google

Departments Business Community Elected Officials News

MACOMB COUNTY PUBLIC HEALTH

Food Service Inspections

The local Health Department is required to conduct routine, unannounced inspections of licensed food service facilities as follows:

- At least once every six (6) months
- At least once every twelve (12) months at facilities open nine months or less.

Follow-up inspections are done as needed to verify correction of any "critical violation" found during the routine inspection.

A "critical violation" is a condition that is more likely than other violations to cause food contamination, illness, or environmental hazards. This site allows you to view the "critical violations" marked on the most recent routine inspection.

To locate a restaurant, either select a zip code from the drop-down box below or view the [zip code map](#) and click on a zip code area.

Zip Code:

Questions?

If you have any questions, call us at 586-469-5236, Monday-Friday, 8:30 a.m. - 5:00 p.m. or [E-mail us now](#).

[Back To Top](#)

Google Translate

Quick Links

- [New: Health & Community Services](#)
- [Public Health Home](#)
- [About the Health Department](#)
- [Health Events](#)
- [Help Me Find It](#)
- [HIPAA](#)
- [Public Health Media](#)

Programs

- [Animal Shelter](#)
- [Emergency Preparedness](#)
- [Environmental Health](#)
- [Family Health Services](#)
- [Health Promotion / Disease Control](#)
- [Medical Examiner](#)

Resource

- [Contact Health Department](#)
- [Employment](#)
- [Fees](#)
- [Forms](#)
- [Health Links](#)

North Carolina

http://www.

Restaurant Ratings, Fridays @ 5:00 on WRAL-TV :: WRAL.com - Mozilla Firefox

New code Anova - zoub... Restaurant Ratings, Fri... ShowESTABLISHMENTT... DHSS - Division of Publi... New York City Departm... [conquer] 什么意思...

www.wral.com/Sonyourside/restaurants/page/1001540/

WRAL.com
RALEIGH • DURHAM • FAYETTEVILLE

SmartShopper
Save with Faye!
CLICK HERE
for recipes, coupons, sales and tips.

Log In
Register
Subscribe

7:22 p.m. • 11-14-12
45°
Currently
Next 3 Days
7 Day Forecast

Triangle 411 Autos Classifieds Service Alley Jobs Real Estate Shop WRAL Coupons

Home News Traffic Weather Sports Business 5 On Your Side Health & Life Entertainment WRAL-TV Out & About Share

Restaurant Ratings Consumer Complaint Form Here to Help Grocery Cart Tracker SmartShopper Blog

WRAL.com's restaurant sanitation information comes from the N.C. Division of Environmental Health. Wake, Mecklenburg, Alleghany, Ashe, and Watauga counties post scores on their own web sites. Links are provided to those sites. Some counties do not provide specific information about the violations on their sites. That information is public record. To see a sanitation report for a particular restaurant, contact the county Department of Environmental Health directly. Tune in to WRAL News at 5 every Friday for the latest Restaurant Ratings.

What is your favorite restaurant? Would you like to know how clean it is? Submit your request to Five On Your Side to see how it scores with county health inspectors.

Restaurant Ratings

Alamance	Chowan	Guilford	Mitchell	Rutherford
Alexander	Clay	Halifax	Montgomery	Sampson
Alleghany	Cleveland	Harnett	Moore	Scotland
Anson	Columbus	Haywood	Nash	Stanly
Ashe	Craven	Henderson	New Hanover	Stokes
Avery	Cumberland	Hertford	Northampton	Surry
Beaufort	Currituck	Hoke	Onslow	Swain
Bertie	Dare	Hyde	Orange	Transylvania
Bladen	Davidson	Iredell	Pamlico	Tyrrell
Brunswick	Davie	Jackson	Pasquotank	Union
Buncombe	Duplin	Johnston	Pender	Vance
Burke	Durham	Jones	Perquimans	Wake
Cabarrus	Edgecombe	Lee	Person	Warren
Caldwell	Forsyth	Lenoir	Pitt	Washington
Camden	Franklin	Lincoln	Polk	Watauga
Carteret	Gaston	Macon	Randolph	Wayne
Caswell	Gates	Madison	Robeson	Wilkes
Catawba	Graham	Martin	Rockingham	Wilson
Chatham	Granville	McDowell	Rowan	Yadkin
Cherokee	Greene	Mecklenburg		Yancey

Public Health Inspector
Pursue Your Passion as a Public Health Inspector. Earn Your Degree.
www.APUS.edu/Public-Health

See Nursing Home Pricing
Get Online info, prices, & Options View Pictures and Floor Plans Today
www.TopNursingHomes.com

Long Term Care Insurance
Maine residents can compare & save with quotes from the top providers
www.completelongtermcare.com

AdChoices

RESTAURANT RATINGS

RESTAURANT RATINGS by Request Restaurant ratings by request (Nov.)	RESTAURANT RATINGS Restaurant ratings (Oct. 26, 2012)	RESTAURANT RATINGS Restaurant ratings (Oct. 19, 2012)
--	--	--

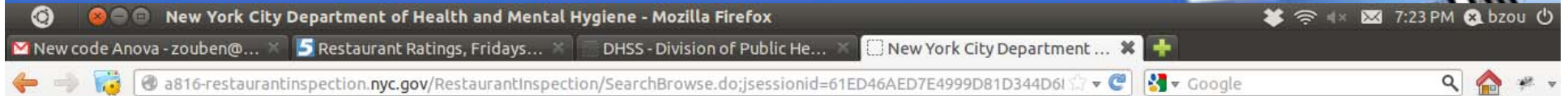
TRIANGLE AREA SPECIAL OFFERS

AlarmForce VideoRelay Security as low as \$10/mo
50 Vehicles for Under \$10,000!

www.wral.com/Sonyourside/smartshopper/?sc_int_cid=2012_Save_with_Faye_226

New York City

http://www.



Restaurant Inspection Information

Home

Search for Inspection Results

Restaurant Name (Type the beginning, full or portion of name)

- or -

Alphabetical List(s):

#	A	B	C	D	E	F	G	H	I	J	K	L	M
N	O	P	Q	R	S	T	U	V	W	X	Y	Z	

You may also browse or narrow your search by choosing any of the following:

Zip Code:

☐ Include Surrounding Zip Codes

- or -

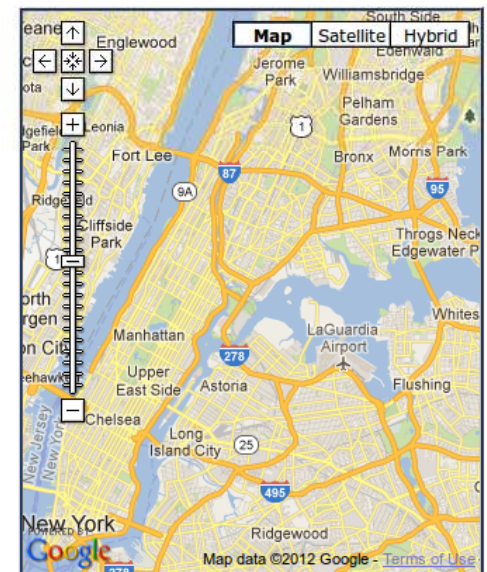
Borough(s):

The Bronx	Brooklyn	Manhattan	Queens	Staten Island
-----------	----------	-----------	--------	---------------

Cuisine:

Grade:

A	B	C	GRADE PENDING	CLOSED
---	---	---	---------------	--------



[Feedback](#) | [Search Tips](#) | [Learn about how we grade](#) | [Help](#)

Copyright 2010 The City of New York

[Contact Us](#) | [FAQs](#) | [Privacy Statement](#) | [Site Map](#)

Our preliminary attempt

http://www.



restaurantinspections

[about](#)

Establishment name:

City:

State:

Zip code:

Jurisdiction:

Violations:

holding temperature

Submit

[Latest Stats](#)

[Jurisdiction Links](#)

Showing results 1-100 of 9890 | [Next](#)

<i>Jurisdiction</i>	<i># of restaurants</i>	<i># of violations</i>
Anne Arundel County	22	26
Chicago City	402	490
District of Columbia City	565	675
San Francisco County	4	5
Virginia State	8,897	30,922

1. [!Aqui Las Tortas!](#)

4828 E Whittier Blvd, Los Angeles, CA 90022

Rule: **holding** of phf - multiple servings (major)

Rule: improper concentration / **temperature** / missing supplies

2. [*Country Inn and Suites Continental Breakfast!](#)

7135 Pocahontas Trail, Williamsburg, VA

Rule: provide a **temperature** measuring device that is easily readable and accurate to +/-3 degrees fahrenheit in the intended range of use.

Note: "the ambient air **temperature** measuring device (degrees f) located in the reach in cooler is broken."

Rule: provide a **temperature** measuring device in all hot or cold **holding** equipment used to store potentially hazardous foods so that employees can routinely monitor the ambient air **temperature**.

Note: "the **temperature** measuring device located egg ri cooler is not easily readable."

Rule: discard food that is not cooked and served or served if ready-to-eat within 4 hours from the point in time when the food is removed from **temperature** control. clearly label or otherwise mark the food with a 4-hour cook-serve or serve by time.

Note: "no time control procedure observed posted for any of the potentially hazardous food(s) that are being held using time control instead of **temperature** control (ex. oatmeal, cut melons, cheese, chopped ham, eggs....)"

3. [-Kumo Sushi](#)

2338 Woodland Crossing, Herndon, VA

Rule: (placed on proper shelves) food shall be protected from cross contamination by separating raw animal foods like poultry, beef, pork, seafood, and eggs during **holding** from all other foods such as cooked foods, ready-to-eat foods, washed produce, etc.

Rule: if raw or partially cooked fish are served or sold in ready-to-eat form, the person in charge shall record the freezing **temperature** and time to which the fish are subjected and shall retain the records for the frozen fish at the food establishment for 90 calendar days beyond the time of service or sale to a consumer. if the fish are frozen by a supplier, a letter the fish

Challenge #4: Long run sustainability



- Need on-going efforts
 - Maintain data structure
 - Ensure data quality
 - Technical support for online, real-time access
 - Support for end users
- Who pays for this?

Data issues are more important than just being an inconvenience to researchers



- Who should do what in this push for more and better data?
 - Government (openspending.gov)
 - Non-profits (Sunlight Foundation)
 - Industry (Google, Yelp,....)
 - Academics (**what role should economists play?**)
- Research on data issues
 - Who is using the data, when and for what?
 - How do data format, content and interface affect data usage?
 - Does transparency improve data quality?
 - How does data quality change when we engage end users?
 - What is the effect of standardization on data quality?
 - What danger does data closure impose on privacy protection (and national security)?
 - What can we learn from UK, Brazil and other countries?

Last Words



Can more and better data complement model structure?

- Help justify model assumptions
- Test model fit and out-of-sample predictions
- Facilitate model choice
- With the aid of data, model structure can better decipher mechanisms and provide more precise counterfactuals