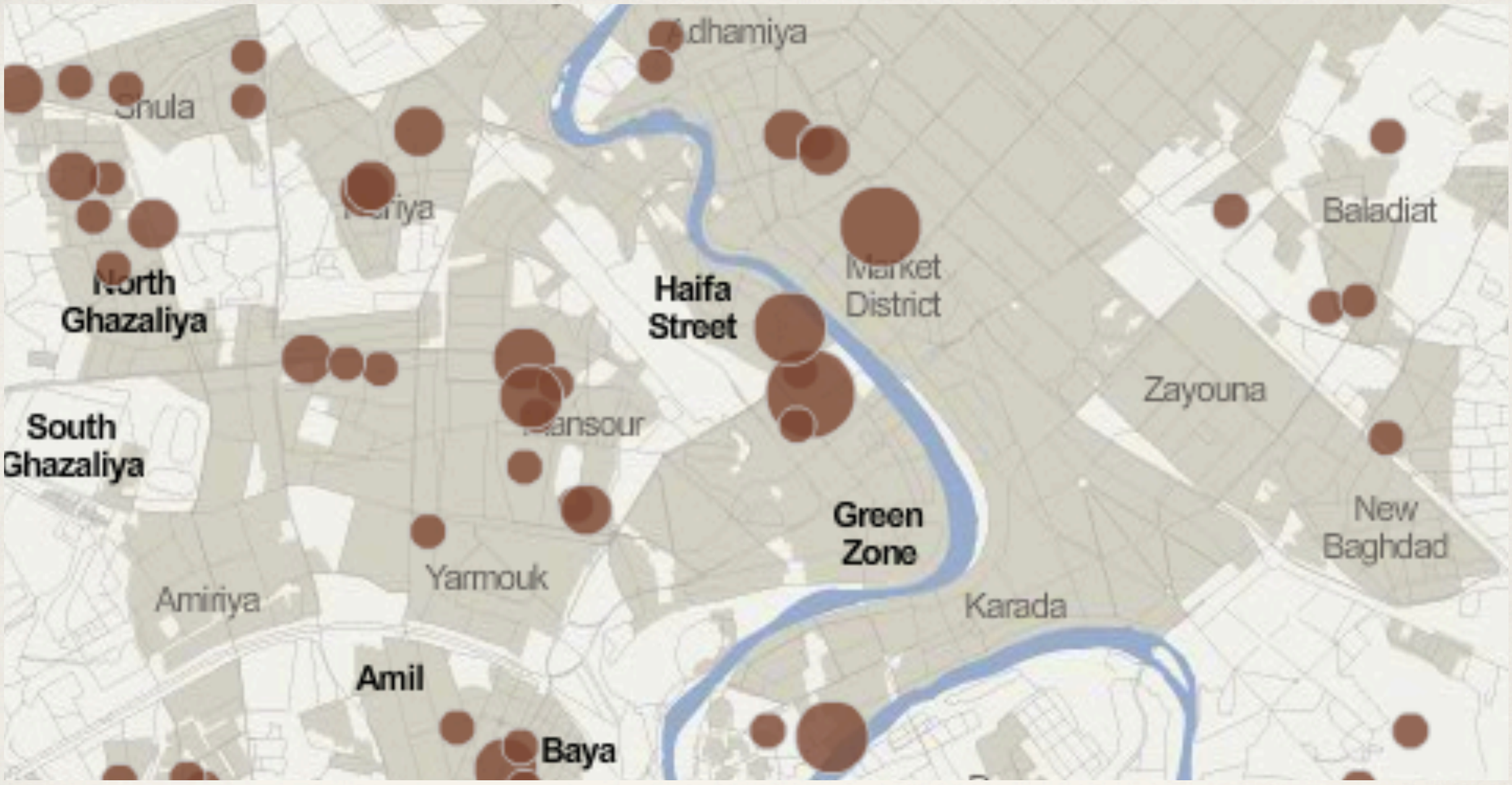


Reporting Wikileaks

Some Thoughts on Data Journalism

Jacob Harris // jharris@nytimes.com // @harrisj // nimblecode.com

Wednesday, March 9, 2011

Hello, my name is Jacob Harris and I work in the Interactive Newsroom Technologies group at the New York Times.

Most of my work is journalism-related, although I usually am a tool-maker rather than a reporter. For instance, I've worked on news apps for things like the elections and the olympics.

More recently, I was involved with the NYT coverage of the Wikileaks war diaries. I built some internal tools for our reporters to analyze the data.

I also got to do some reporting on my own. Which is what I'm going to be talking about today.

"If Afghanistan is a war of small cuts, Iraq was a gash. In the war's bloodiest months, according to the archive's reports, more than **3,000** Iraqi civilians were dying, more than 10 times the current civilian casualty rate in Afghanistan, a country with a larger population."

Sabrina Tavernise, *Mix of Trust and Despair Helped Turn Tide in Iraq*

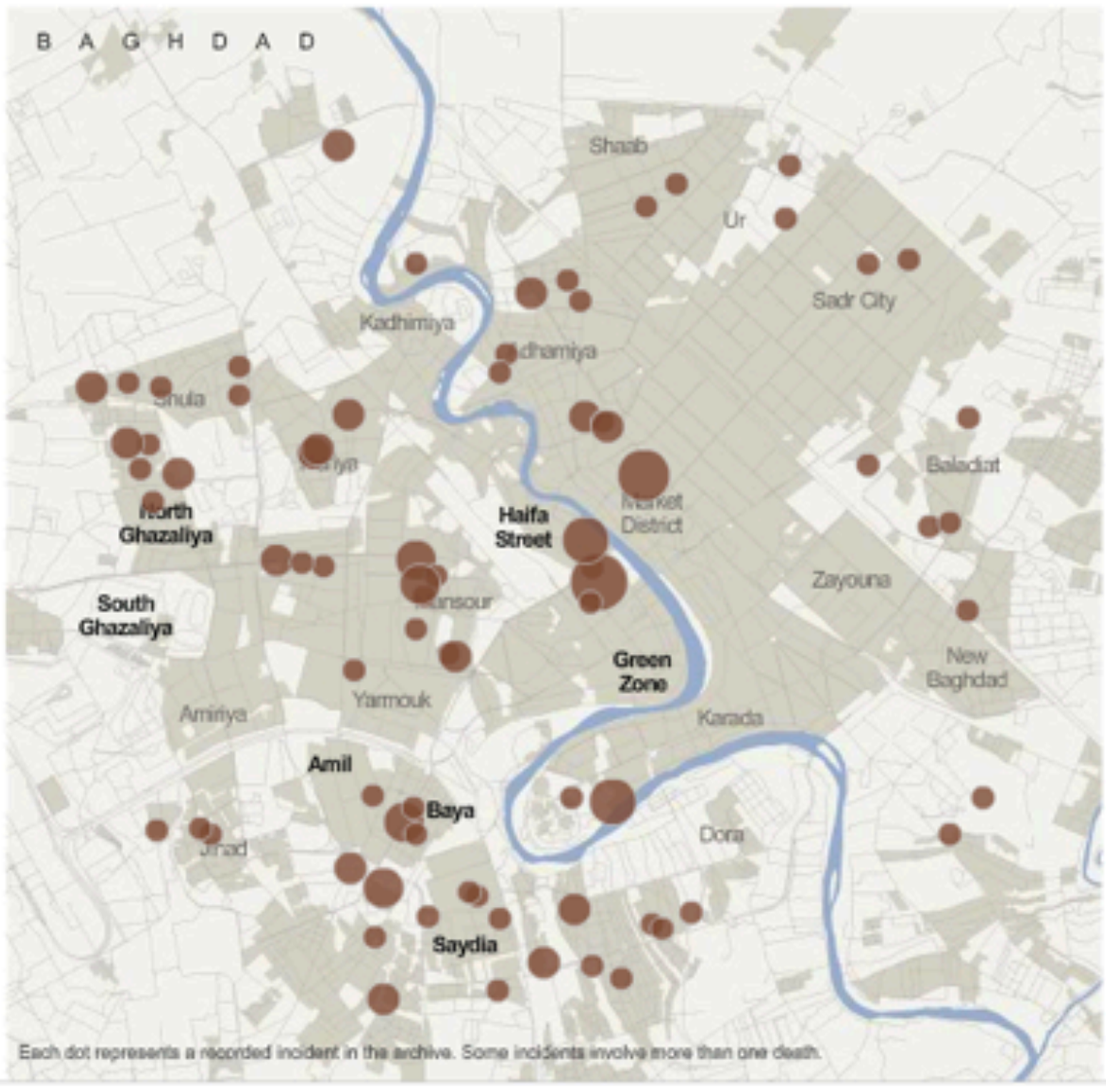
Wednesday, March 9, 2011

One of the hopes of the Wikileaks data is that it promised more detail on events that the Times reporters had witnessed firsthand. One of these stories was the surge of violence as ethnic cleansing swept across Baghdad.

Here is an excerpt from the recent Times story on the violence. The basic facts were the same as in the first reports 5 years ago, but we were able to better see the magnitude of the killings.

A Deadly Day In Baghdad

Violence peaked in December 2006, just two months before American troops arrived as part of what was later called "the surge." At right are the details of one of the city's deadliest days, Dec. 20. There were 114 separate episodes of violence that day, resulting in the deaths of about 160 Iraqi citizens and police officers.



Each dot represents a recorded incident in the archive. Some incidents involve more than one death.

Haifa Street
The day's deadliest incidents came in the religiously mixed area near Haifa Street. A total of 11 people were found dead in the area, their identities unknown.

Green Zone
This heavily fortified area, controlled by American troops, has been the seat of Iraq's government and the country's diplomatic center since 2003. There have been sporadic rocket attacks there, but compared with its surrounding neighborhoods it has been relatively safe.

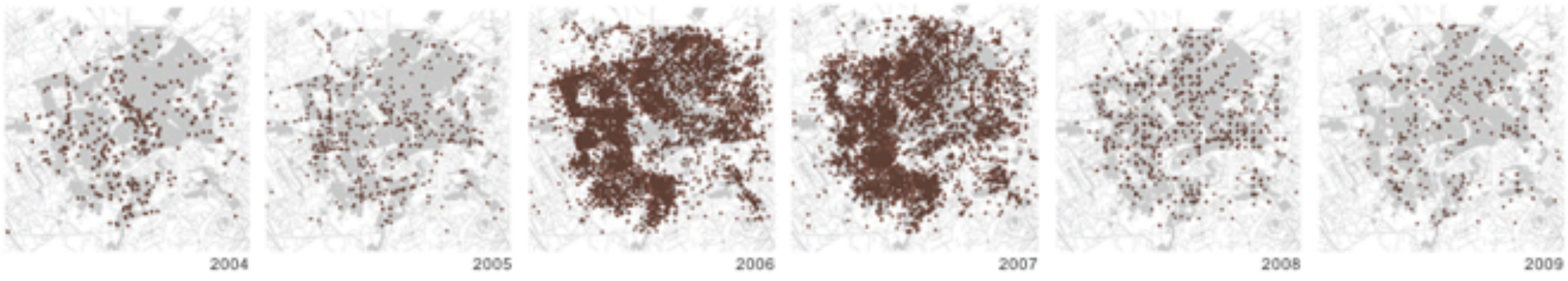
Saydia, Baya and Amil
These formerly mixed neighborhoods, now dominated by the Shiite Mahdi Army, were places of intense violence both that day and throughout the war. A total of 19 people were killed there in 13 separate incidents.

Ghazaliya
This Sunni enclave in western Baghdad was one of the first to experience systematic cleansing of Shiites. It remained violent through the worst periods of the war.

Extreme Sectarian Violence in Baghdad

Secret military field reports reveal more than 32,000 deaths in the Baghdad area between 2004 and 2009. Many of the deaths were Iraqi civilians, although more than a third of the bodies were unidentified. Violence exploded in 2006 and early 2007, with more than 2,700 deaths in Baghdad in December 2006, the most of any month.

Locations of fatalities in Baghdad, 2004-9



Wednesday, March 9, 2011

Here is a graphic we built to accompany this story. At the top, we look at a single day of violence in Baghdad; at the bottom, we present the overall trend. The Graphics department of the Times does many stellar graphics like these every day, but I'm excessively proud of this one. Because this one is mine. Sabrina fact checked it. Kevin Quealy made it beautiful. But I wrangled the data.

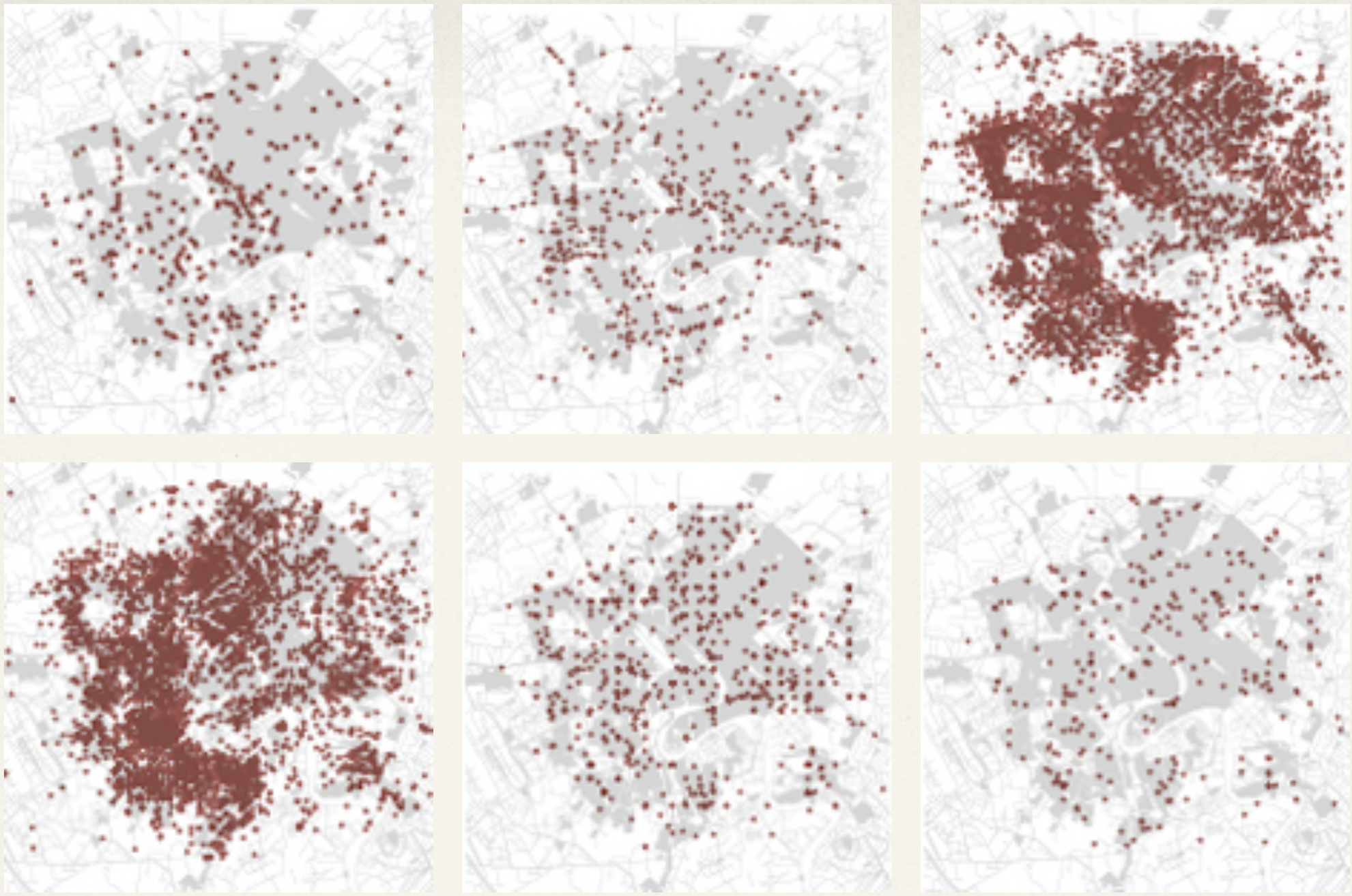
Of course, there is more we could've done with the graphic. For instance, I would've liked to let people drill down into a day-by-day view, but the sensitive circumstances of the data didn't allow it. Often we have to operate within external constraints.

Wednesday, March 9, 2011

Here is something we didn't do: a word cloud. Every time I see a word cloud, I die a little inside. Because it violates every single rule I have in my mind about what makes for good data journalism.

I didn't create this word cloud as a hypothetical example; I actually found it out there as part of someone's reporting. I hope you don't mind me using it as an example of WHAT not to do with data.

So, what are these rules of data journalism I'm talking about? I don't have a definitive list, but I'm going to talk about three that popped into my head when thinking about this talk. I'm sure there are more rules than 3. I'd like to hear your suggestions.



1. Find A Narrative

If you aren't going to look for a story in the data, don't call yourself a journalist.

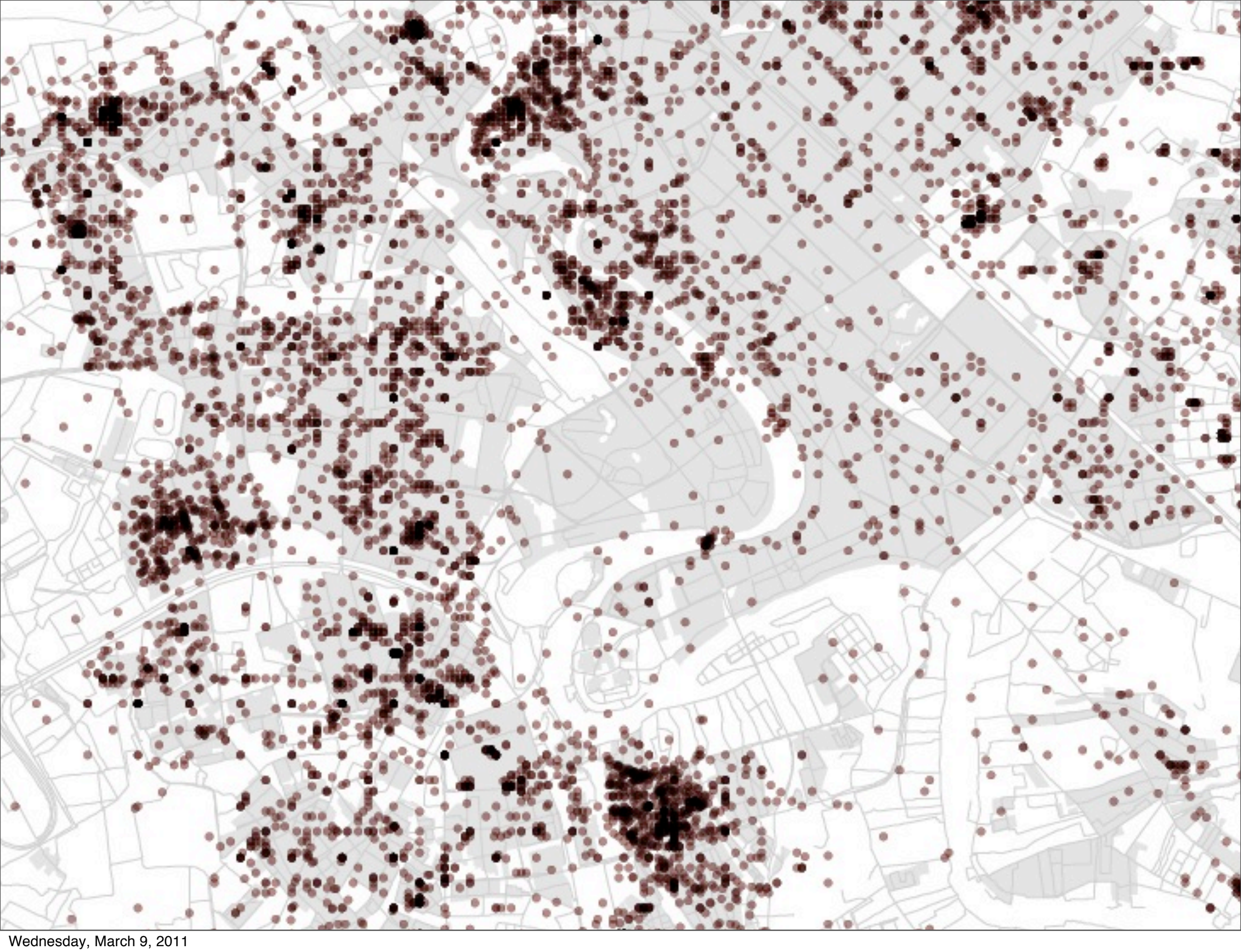
Wednesday, March 9, 2011

First rule. Find a narrative in the data. It doesn't have to be **THE** narrative (rich data sets have many narratives), but you need at least **A** narrative. Narratives are how we make sense of the data. We investigate by exploring questions inherent to then narrative. In science, this is called a hypothesis.

Narrative is also how we invite the reader into our work. We tell a story with the data. Otherwise, you're just posting a random spreadsheet or a meaningless PDF. And nobody cares.

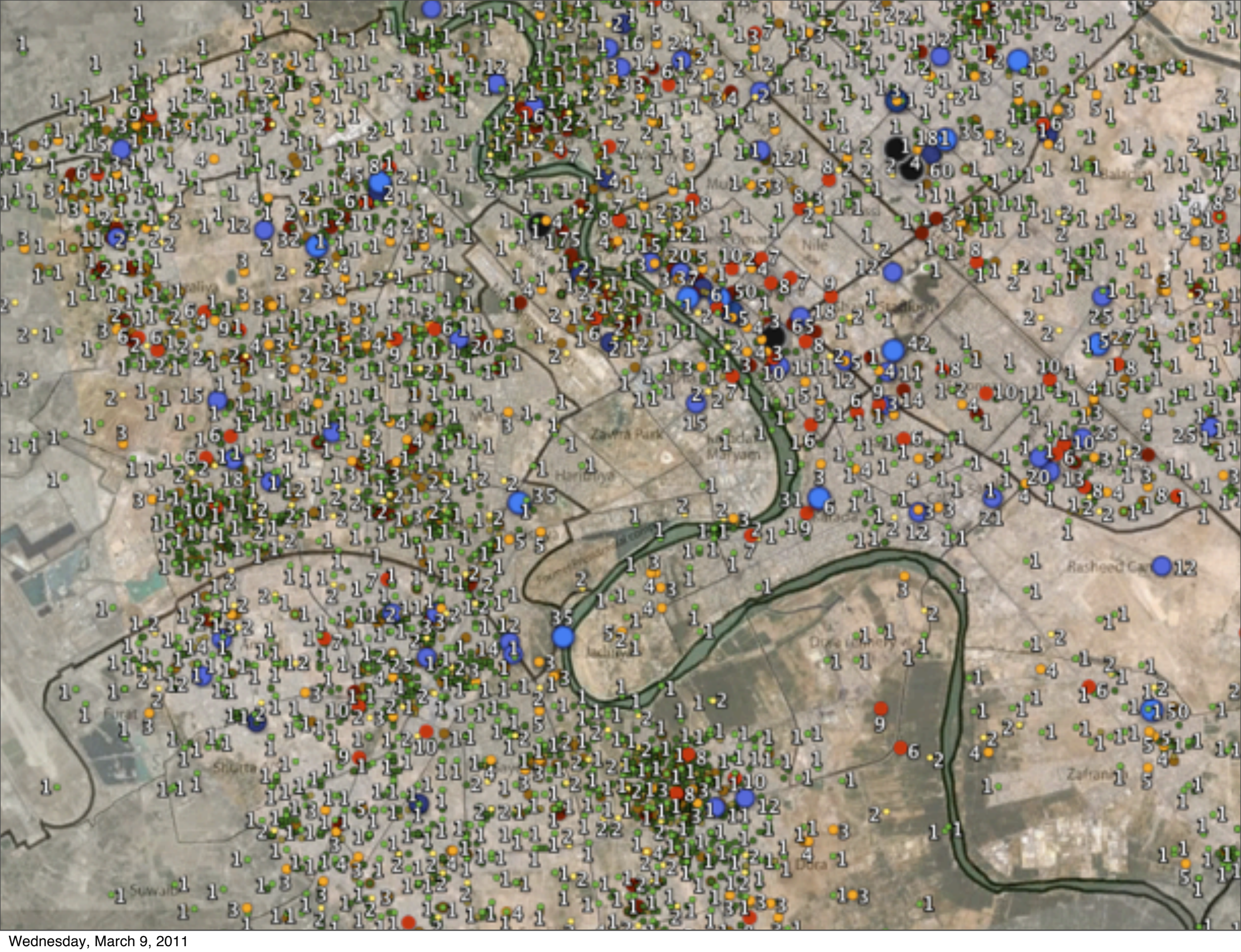
Word clouds don't have narrative. They just are.

In our graphic we decided to illustrate the narrative of violence at two timespans: the yearly picture and the carnage of a single day.



Wednesday, March 9, 2011

Here is an enlarged view of the incidents in 2006 from the graphic. Since it is meant to display overall trends in Baghdad, extra details like neighborhoods are left out. But you can see areas where some violence was more pronounced.



Wednesday, March 9, 2011

And this is the rough cut in Google Earth I created to pitch the graphic's narrative. It lacks any polish, but it does the job for me. The colors/and size correspond to the number of civilians killed in a single incident. As in the previous map, the violence is extremely widespread, but you can also see concentrations in certain religiously mixed neighborhoods: Dora in the south, Ghazaliya and Hurriya in the west. Adhamiya and Sheik Maruuf on the river.

Can you also guess where the Green Zone is?

```
Title: (CRIMINAL EVENT) MURDER RPT BY AL KHADRA
       POLICE IVO BAGHDAD (ZONE 37)
       (ROUTE UNKNOWN): 1 CIV KIA
Allied KIA: 0
Allied WIA: 0
Civilian KIA: 1
Civilian WIA: 0
Enemy KIA: 0
Enemy WIA: 0
Location: 44.32324982, 33.3067627
Type: Criminal Event
Category: Murder
SIGACT: (null)
```

Wednesday, March 9, 2011

Every incident in the war logs came with associated metadata. This was both illuminating and maddening, as every field was rife with errors. For instance, I had high hopes for the Type and Category fields, but soon discovered they were entered as freeform text and relying on them would entail far more false negatives than I could bear.

But, every incident had a location and timestamp associated with it, as well as summary totals for the wounded and the dead. I could do something with this.

```
<Placemark>
  <TimeStamp><when>2006-12-20</when></TimeStamp>
  <description>
    <![CDATA[<p>(CRIMINAL EVENT) MURDER RPT BY AL
KHADRA POLICE IVO BAGHDAD (ZONE 37) (ROUTE UNKNOWN):
1 CIV KIA</p>]]>
  </description>
  <name>1</name>
  <styleUrl>#green-icon</styleUrl>
  <Point>
    <coordinates>44.32324982,33.3067627</coordinates>
  </Point>
</Placemark>
```

Wednesday, March 9, 2011

Here is an incident transformed into a Google Earth Placemark. KML is a pretty neat thing from Google. It allows you to specify out points on a map with an XML file. I don't know a lot about maps, I haven't learned GIS, but I am pretty good at hacking scripts to dump text and this was a godsend to me.

```ruby
BOUNDS = {
  :north => 33.433590,
  :south => 33.192527,
  :east => 44.543875,
  :west => 44.203514
}

class Incident < ActiveRecord::Base
end

File.open("/tmp/kia.xml", "w") do |file|
  (2004..2009).each do |year|
    file.write("<Folder>\n<name>#{year}</name>\n<open>0</open>\n")

    (1..12).each do |month|
      file.write("  <Folder>\n  <name>#{month}</name>\n  <open>0</open>\n")
      start_date = Date.parse("#{month}/1/#{year}")
      end_date = start_date.end_of_month

      (start_date..end_date).each do |date|
        incidents = Incident.civilian_kia.all(:conditions => ['DATE(date_occurred) = :date AND lat >= :south AND lat <= :north
AND lng >= :west AND lng <= :east', {:date => date}.merge(BOUNDS)])
        if incidents.any?
          file.write("    <Folder>\n    <name>#{date.strftime('%m/%d/%Y')}</name>\n    <open>0</open>\n")

          incidents.each do |incident|
            next if incident.title.blank?
              file.write("      <Placemark><TimeStamp><when>#{incident.date_occurred.strftime("%Y-%m-%d")}</when></
TimeStamp><description><![CDATA[<p>#{incident.title.gsub('&', '&amp;')}</p>]]></description><name>#{incident.civilian_kia}</
name><styleUrl>#{icon_style(incident.civilian_kia)}</styleUrl><Point><coordinates>#{incident.lng},#{incident.lat}</
coordinates></Point></Placemark>\n")
          end
          file.write("    </Folder>\n")
        end
      end
      file.write("  </Folder>\n")
    end
    file.write("</Folder>")
  end
end
```

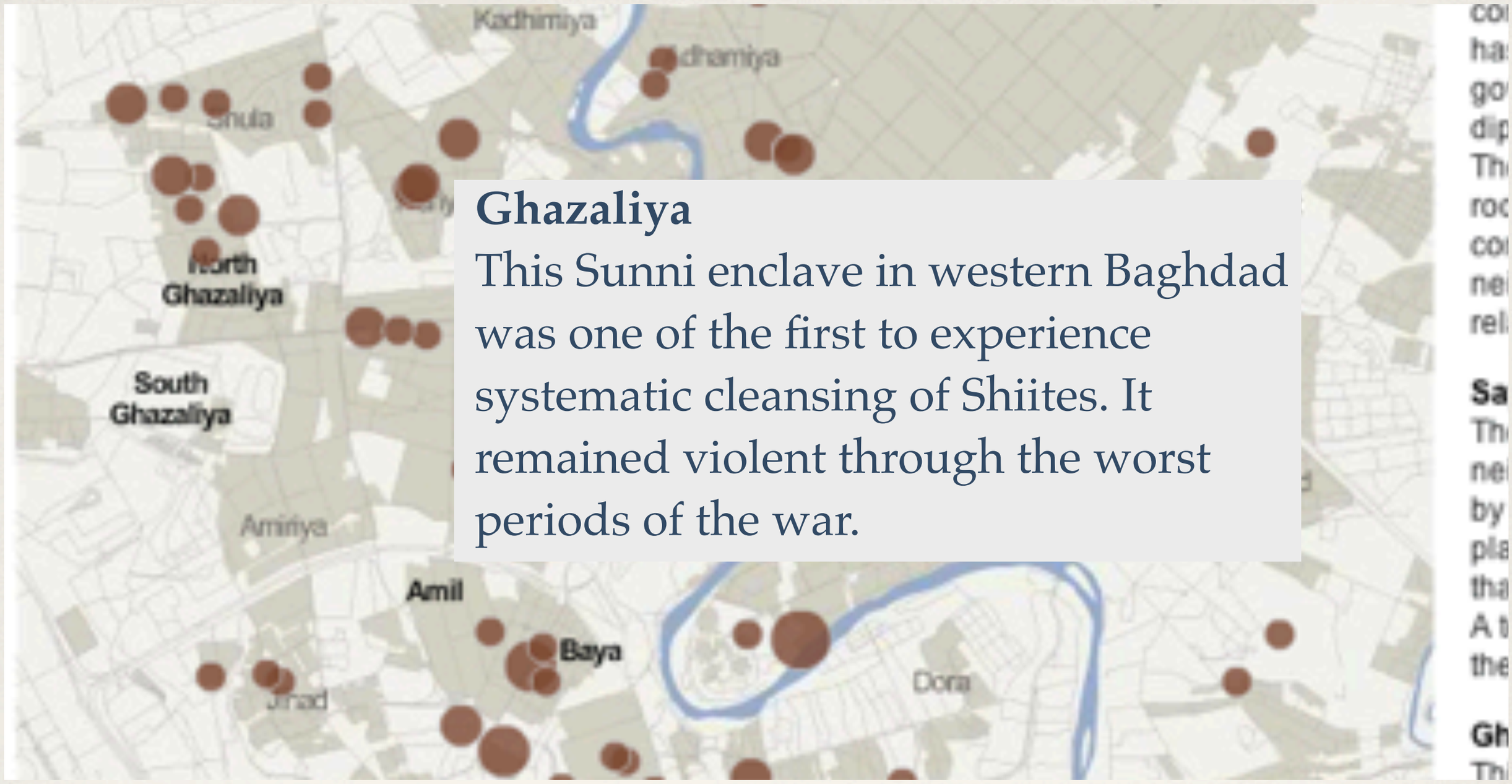
Wednesday, March 9, 2011

Because I feel like a fraud if I don't show you code, here is the ruby script I used to generate the KML file. Nothing particularly advanced here.

Wednesday, March 9, 2011

Putting it together, this was the final result. One thing not obvious in this screengrab was Google Earth provides you with a little date slider if you associate a time with your points. This allowed me to create a little animation that showed the violence move across Baghdad from the early days.

Still, something nagged at me. How did I know I hadn't made a fundamental mistake somewhere and was plotting things in the wrong location? How do I know this data matched reality at all? Paranoia is the best friend of the data journalist. More on that in a moment



2. Give Context

Empower readers to understand the data. Let them explore if possible.

Rule 2. Give Context

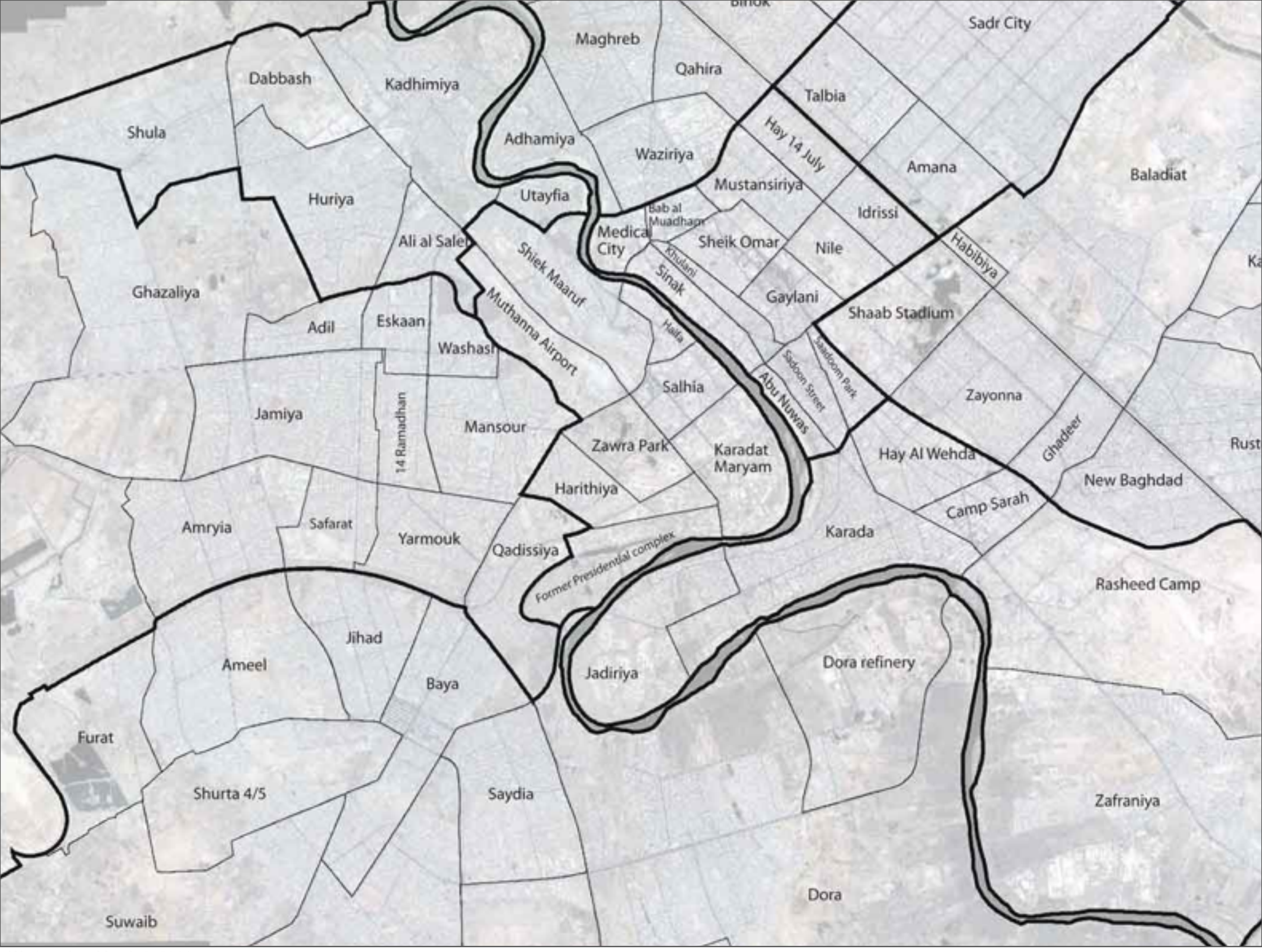
There are sadly a lot of graphics that are essentially infovizporn: you know, those abstract constructions with lots of edges and nodes and trippy colors and weird shapes.

Of course, we like to make our stuff look pretty, but it needs to be informative first.

We strive to involve the readers with our work. In the best cases like Toxic Waters or Election Results, we let the reader explore the data on their own. This was sadly not an option for the Wikileaks data.

But, in any event, it's **essential** to provide context. Every data set has its own nuances, its own jargon, its own quirks. It's important to give the reader context with every data piece. We want to bring them up to speed.

Here's an example of the sidebar text from the graphic, explaining why the violence was bad in Ghazaliya.



Wednesday, March 9, 2011

Sabrina had spent several brave years reporting from Baghdad, and she knew firsthand which neighborhoods had suffered the worst sectarian violence. Google Maps had a pretty atrocious knowledge of Iraq, so my first step was to find a map of the neighborhoods of Baghdad and overlay that on the map grid.

This gave me **context** to where the violence was occurring and also a way to **validate** that my geocoding was working and the data corresponded with what Sabrina saw. And we saw the violence from the data showing up in the same areas Sabrina had witnessed it several years earlier. Dora in the south. Ghazaliya and Hurriya in the west, Sheik Maaruf in the center.

Secret Dispatches From the War in Iraq

Below are a selection of the reports from a six-year archive of classified military documents to be published by WikiLeaks. These examples provide an unvarnished, ground-level picture of the war in Iraq. Some names and details have been redacted by The Times to conceal suspects' identities, or because they might put people in danger or reveal key tactical military capabilities. See below for more details on the redactions.

Key

— Redacted text

REL Roll over underlined military terms for definitions

▸ Iran

▾ Casualties

7/16/07
Helicopter Attack

2/22/07
Attempted Surrender

7/22/05
Checkpoint Shootings

2/20/06
Interpreter Killed

▸ Prisoner Abuse

▸ Kidnapped Hikers

▸ Contractors

▸ Country in Chaos

▸ Kurds

BLUE ON WHITE BY 1ST RECON S OF NASSER WA AL SALEM: 1 *CIV* KILLED, 0 *CF* INJ/DAMA

MAM
Military-Aged Male

AT 200100C FEB C[illegible]IPER TEAM WHILE CONDUCTING CLANDESTINE SNIP[illegible] HAJI RD IN THE ZAIDON ENGAGED (1) *MAM* WITH (4) 5.56M[illegible] ROUNDS *IVO* (38S MB 09971 79804) 4KM S OF NASSER WA AL SALEM. THE *MAM* WAS *PID* W/ AK-47 CREEPING UP BEHIND THEIR SNIPER POSITION AND WAS SHOT IN THE CHEST W/ (2) 5.56MM ROUNDS AT 15M. *QRF* WAS LAUNCHED TO EXTRACT THE SNIPER TEAM. THE *MAM* WAS SEARCHED BY TEAM AND RECOVERED (1) AK-47, (2) MAGAZINES OF 7.62MM, DOUBLE TAPED, (1) LARGE KNIFE, (1) ID CARD WITH "——— ———" WRITTEN ON CARD. *MAM* WAS ALSO NOTED TO BE WEARING A TRACKSUIT AND SEVERAL WARMING LAYERS TO INCLUDE (2) PAIRS OF SOCKS. THE BODY WAS LEFT BEHIND AT (38S MB 09971 79804) UPON EXTRACT OF THE SST. PIONEER OBSERVING ON SITE W/ *NSTR*.

UPDATE: UPON FURTHER INVESTIGATION THE *KIA* TURNED OUT TO BE THE PLATOON'S INTERPRETER THAT WAS SEPARATED FROM UNIT. THE BODY WAS RECOVERED AND IS CURRENTLY LOCATED AT FALLUJAH SURGICAL. THIS ACTION IS NOW CONSIDERED A BLUE ON GREEN. IT RESULTED IN (1) *IZ* *KIA* (IRAQI INTERPRETER EMPLOYED BY TITAN.

SUMMARY If the war was dangerous for Americans, it was far worse for the Iraqis who worked for them. One Iraqi interpreter was killed by an American sniper from his own unit, who mistook him for a militant when the Iraqi became separated from his platoon.

About the Redactions The types of information that have been removed from the documents include: Names of public figures (generals,

Wednesday, March 9, 2011

Here's another example of the War Logs where you can see several of the ways we provide context to the reader:

* The intro at the top for the whole list of documents.
* The summary on the right (and the navigational bar on the left)
* The pop-up inline Jargon translator we call the Jargonator

I wanted to call it JSON and the Jargonauts, but I was overruled.

Toxic Waters *A series about the worsening pollution in American waters and regulators' response*

SHARE | FEEDBACK | E-M

Examine whether contaminants in your water supply met two standards: the legal limits established by the Safe Drinking Water Act, and the typically stricter health guidelines. The data was collected by an advocacy organization, the Environmental Working Group, who shared it with The Times.

Browse All States | Understanding the Data | Related Story | Join the Discussion | All Stories in the Series: Toxic Waters

New York City-Catskill/Delaware System

View State...

New York (c), New York. Serves 6,552,718 people.

CHART KEY: 1 or more tests taken in the month | 1 or more positive detections | 1 or more tests above health limit | 1 or more tests above legal li

1 contaminant below legal limits, but above health guidelines

In some states a small percentage of tests were performed before water was treated, and some contaminants were subsequently removed or diluted. As a result, some repor levels of contamination may be higher than were present at the tap.

Contaminant	Average result	Maximum result	Health limit	NUMBER OF TESTS: Legal limit	Total #	Positive result	Above health	Above legal	MONTHLY TESTING HISTORY (2004–2009)	E.P. regula
Trichloroacetic acid	24.26 ppb	56.80	20	60	24	24	20	0		Yes

11 contaminants found within health guidelines and legal limits

Contaminant	Average result	Maximum result	Health limit	NUMBER OF TESTS: Legal limit	Total #	Positive result	Above health	Above legal	MONTHLY TESTING HISTORY (2004–2009)	E.P. regula
Bromodichloromethane	3.69 ppb	6.90	100	80	25	25	0	0		Yes
Chloroform	33.82 ppb	59.30	70	80	25	25	0	0		Yes
Copper	4.64 ppb	20	1300	1300	14	14	0	0		Yes
Dibromoacetic acid	0.00 ppb	0.37	-	60	24	1	0	0		Yes
Dibromochloromethane	0.01 ppb	0.80	60	80	25	3	0	0		Yes
Dichloroacetic acid	17.88 ppb	32.80	70	60	24	24	0	0		Yes
Lead (total)	0.01 ppb	2	0.20	15	15	2	0	0		Yes
Monobromoacetic acid	0.00 ppb	0.65	-	60	24	1	0	0		Yes

Wednesday, March 9, 2011

And breaking out of Wikileaks, here is how we give context for a different complex data set like Toxic Waters. This is showing testing for various chemicals in NYC's drinking water. There is a lot of complex data here, it's easy to overwhelm the reader. Context is essential

Show results for: All Districts

GAIN WIN LEAD

Democrats expected to win easily

District	Dem.	Rep.	% Rpt.
Ala. 7	72%	28%	100%
Ariz. 4	67%	28%	100%
Calif. 1	63%	31%	100%
Calif. 5	72%	25%	100%
Calif. 6	66%	30%	100%
Calif. 7	68%	32%	100%
Calif. 8	80%	15%	100%
Calif. 9	84%	11%	100%
Calif. 10	59%	38%	100%
Calif. 12	76%	22%	100%
Calif. 13	72%	28%	100%
Calif. 14	69%	28%	100%
Calif. 15	68%	32%	100%
Calif. 16	68%	24%	100%
Calif. 17	67%	26%	100%
Calif. 23	58%	38%	100%
Calif. 27	65%	35%	100%
Calif. 28	70%	22%	100%
Calif. 29	65%	32%	100%
Calif. 30	65%	32%	100%
Calif. 31	84%	16%	100%
Calif. 32	71%	29%	100%
Calif. 33	86%	14%	100%
Calif. 34	77%	23%	100%
Calif. 35	79%	21%	100%
Calif. 36	60%	35%	100%

Democrats expected to win narrowly

District	Dem.	Rep.	% Rpt.
Ark. 4	58%	40%	100%
Calif. 18	58%	42%	100%
Calif. 20	52%	48%	100%
Calif. 47	53%	39%	100%
Colo. 7	53%	42%	100%
Conn. 4	53%	47%	100%
Conn. 5	54%	46%	100%
Del. 1	57%	41%	100%
Ga. 12	57%	43%	100%
Iowa 1	50%	48%	100%
Iowa 2	51%	46%	100%
Iowa 3	51%	47%	100%
Ill. 8	48%	48%	100%
Ill. 10	49%	51%	100%
Ky. 3	55%	44%	100%
La. 2	65%	33%	100%
Mass. 4	54%	43%	100%
Me. 1	57%	43%	100%
Me. 2	55%	45%	100%
Mich. 9	50%	47%	100%
Mich. 15	57%	40%	100%
Minn. 1	49%	44%	100%
Mo. 4	45%	50%	100%
N.C. 2	49%	49%	100%
N.C. 7	54%	46%	100%
N.C. 11	54%	46%	100%

Tossup seats

District	Dem.	Rep.	% Rpt.
Ala. 2	49%	51%	100%
Ariz. 5	43%	52%	100%
Ariz. 7	50%	44%	100%
Ariz. 8	49%	47%	100%
Calif. 11	48%	47%	100%
Colo. 3	46%	50%	100%
Fla. 22	46%	54%	100%
Fla. 25	43%	52%	100%
Ga. 2	51%	49%	100%
Ga. 8	47%	53%	100%
Hawaii 1	53%	47%	100%
Idaho 1	41%	51%	100%
Ill. 14	45%	51%	100%
Ill. 17	43%	53%	100%
Ind. 2	48%	47%	100%
Ind. 9	42%	52%	100%
Ky. 6	50%	50%	100%
Mass. 10	47%	42%	100%
Mich. 7	45%	50%	100%
Miss. 4	47%	52%	100%
N.C. 8	53%	44%	100%
N.D. 1	45%	55%	100%
N.H. 2	47%	48%	100%
N.J. 3	47%	50%	100%
Nev. 3	47%	48%	100%
N.Y. 19	47%	53%	100%

Republicans expected to win narrowly

District	Dem.	Rep.	% Rpt.
Ark. 1	44%	52%	100%
Ariz. 1	44%	50%	100%
Ariz. 3	41%	52%	100%
Calif. 3	43%	50%	100%
Colo. 4	41%	52%	100%
Fla. 2	42%	54%	100%
Fla. 8	38%	56%	100%
Fla. 24	40%	60%	100%
Ill. 11	43%	57%	100%
Md. 1	42%	54%	100%
Mich. 1	41%	52%	100%
Minn. 6	40%	53%	100%
Miss. 1	41%	55%	100%
Neb. 2	39%	61%	100%
N.H. 1	42%	54%	100%
N.M. 2	45%	55%	100%
Ohio 1	46%	51%	100%
Ohio 15	41%	54%	100%
Pa. 3	44%	56%	100%
Pa. 6	43%	57%	100%
Pa. 7	44%	55%	100%
Pa. 11	45%	55%	100%
Pa. 15	39%	54%	100%
Tex. 17	37%	62%	100%
Va. 2	42%	53%	100%
Va. 5	47%	51%	100%

Republicans expected to win easily

District	Dem.	Rep.	% Rpt.
Alaska 1	31%	69%	100%
Ala. 1		83%	100%
Ala. 3	41%	59%	100%
Ala. 4		*Unc.*	
Ala. 5	42%	58%	100%
Ala. 6		*Unc.*	
Ark. 2	38%	58%	100%
Ark. 3	28%	72%	100%
Ariz. 2	31%	65%	100%
Ariz. 6	29%	66%	100%
Calif. 2	43%	57%	100%
Calif. 4	31%	61%	100%
Calif. 19	35%	65%	100%
Calif. 21		*Unc.*	
Calif. 22		*Unc.*	
Calif. 24	40%	60%	100%
Calif. 25	38%	62%	100%
Calif. 26	37%	54%	100%
Calif. 40	33%	67%	100%
Calif. 41	37%	63%	100%
Calif. 42	32%	62%	100%
Calif. 44	44%	56%	100%
Calif. 45	42%	51%	100%
Calif. 46	38%	62%	100%
Calif. 48	36%	60%	100%
Calif. 49	31%	63%	100%

Wednesday, March 9, 2011

Here is an example from the recent elections, showing how context can reveal a narrative.

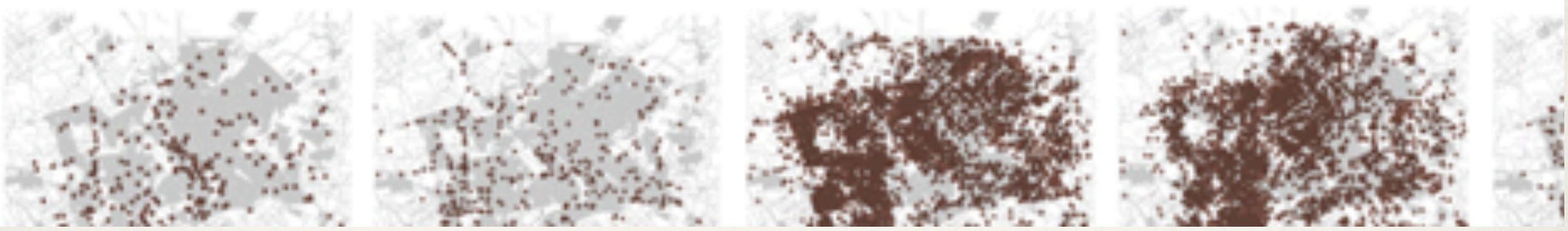
Normally, it's good to anchor context around familiar tropes like a map or a timeline. And we had an election map ourselves.

But on the Big Board, we ordered races by their expected outcome and you can just see how the Republicans took back the House by reading down the center and right columns. Those dark red boxes with black stripes are a Republican gain of a formerly Democratic seat (yes, more context)

Extreme Sectarian Violence in Baghdad

Secret military field reports reveal more than 32,000 deaths in the Baghdad area between 2004 and 2009. M
although more than a third of the bodies were unidentified. Violence exploded in 2006 and early 2007, with r
December 2006, the most of any month.

Locations of fatalities in Baghdad, 2004-9



3. Work The Data

Understand its strength. Don't sugarcoat its flaws.

Wednesday, March 9, 2011

Rule 3. Work The Data

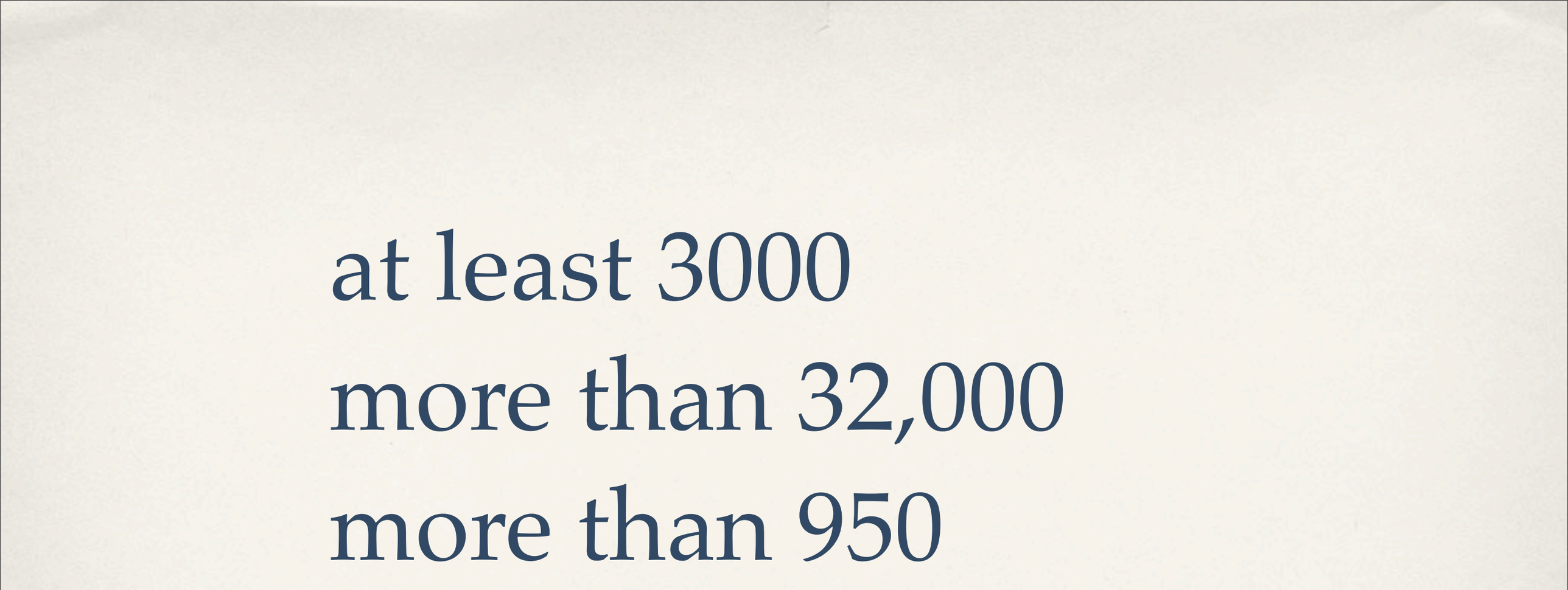
The final rule I'm going to talk about is you need to really grok the data before you report it. This makes it hard if you are thinking about deadlines, and I worry about the intersection of data journalism and breaking news, but data is special.

Most data journalism is not something you can just plug into some magical technology and go get a coffee while the computer figures the story out. It's an iterative process. It requires work, sometimes lots of it. Especially with something like the War Logs, where you get the sloppiness of hasty individual field reports combined with the explosive story of their release. We trod very carefully here.

```
select sum(civilian_kia)
from incidents
where lat >= 33.192527
and lat <= 33.433590
and lng >= 44.203514
and lng <= 44.543875
```


For instance, it is tempting to think you can just run a query like this and declare you now know **definitively** how many people died in Baghdad during the occupation. And I think some news outlets did. But there are big problems:

* We saw reports where the tally didn't match the numbers in the report text.
* The tally only includes what units saw at the time. If many people died later or were uncounted, we wouldn't see that.
* At the beginning of the war, many civilian deaths were counted as enemy combatants
* We don't know the methodology under which this data was collected.
* We weren't even sure if the leak was complete; the leaker could've forgotten to copy some files.
* This leak only includes things at SECRET level or below. Embarrassing events are often hidden behind higher security levels.



Wednesday, March 9, 2011

And so, we hedge. In the excerpt at the beginning we said, **at least 3000** died in the bloodiest months of the occupation. We didn't give an exact count like 3784, because that would've suggested a precision about the facts we didn't get from this data. And precision implies certainty. Sure, the data has an exact number, but always remember you're reporting on reality, not the data. The data is just a source.

Of course, we do have exact counts for the deaths on Dec. 20, 2006. This is because Sabrina painstakingly went through and vetted the data for the day, weeding about 15% of the records as duplicates and such.

(CRIMINAL EVENT) MURDER *RPT* BY NOT PROVIDED *IVO* BAGHDAD (ZONE 1) (ROUTE UNKNOWN): 1 *ISF KIA* 27 *CIV KIA*
28X CORPSES WERE FOUND THROUGHOUT BAGHDAD:

2X HANDCUFFED, BLINDFOLDED, AND SHOT IN THE HEAD IN AL JIHAD (MB393859, MAHALAH #887, 1136 HRS, HY ALAMIL PS)

2X SHOT IN THE HEAD IN AL HURRIYA (MB367918, 1340 HRS, AL HURIYA PS)

1X SHOT IN THE HEAD IN AL ALAMIL (MB374824, 1400 HRS, HY ALAMIL PS)

1X SHOT IN THE HEAD IN AL JIHAD (MB332816, MAHALAH #895, 1245 HRS, HY ALAMIL PS)

1X SHOT IN THE HEAD IN SADR CITY (MB502242, 1500 HRS, AL RAFIDIAN PS)

6X SHOT IN THE HEAD IN SHEIKH MAARUF (MB425880, MAHALAH #212, 1620 HRS, AL JAAIFER PS)

Wednesday, March 9, 2011

Here's a more complex example of the work that goes into data journalism. Looking at the map of the deaths, I noticed there would be a massive dump of bodies at a different single spot every day. Such gruesome things did indeed happen from time to time, but daily?

This was the cause. As the violence intensified, the military took to instead publishing rollup reports like this that summarized the day's crimes scenes. If you plotted this with the death toll and location of the report, it would show up as a single point with 28 corpses, but in actuality, the locations were scattered around Baghdad.

What to do?

(CRIMINAL EVENT) MURDER *RPT* BY NOT PROVIDED *IVO* BAGHDAD (ZONE 1) (ROUTE UNKNOWN): 1 *ISF KIA* 27 *CIV KIA*
28X CORPSES WERE FOUND THROUGHOUT BAGHDAD:

2X HANDCUFFED, BLINDFOLDED, AND SHOT IN THE HEAD IN AL JIHAD (**MB393859**, MAHALAH #887, 1136 HRS, HY ALAMIL PS)

2X SHOT IN THE HEAD IN AL HURRIYA (**MB367918**, 1340 HRS, AL HURIYA PS)

1X SHOT IN THE HEAD IN AL ALAMIL (**MB374824**, 1400 HRS, HY ALAMIL PS)

1X SHOT IN THE HEAD IN AL JIHAD (**MB332816**, MAHALAH #895, 1245 HRS, HY ALAMIL PS)

1X SHOT IN THE HEAD IN SADR CITY (**MB502242**, 1500 HRS, AL RAFIDIAN PS)

6X SHOT IN THE HEAD IN SHEIKH MAARUF (**MB425880**, MAHALAH #212, 1620 HRS, AL JAAIFER PS)

Wednesday, March 9, 2011

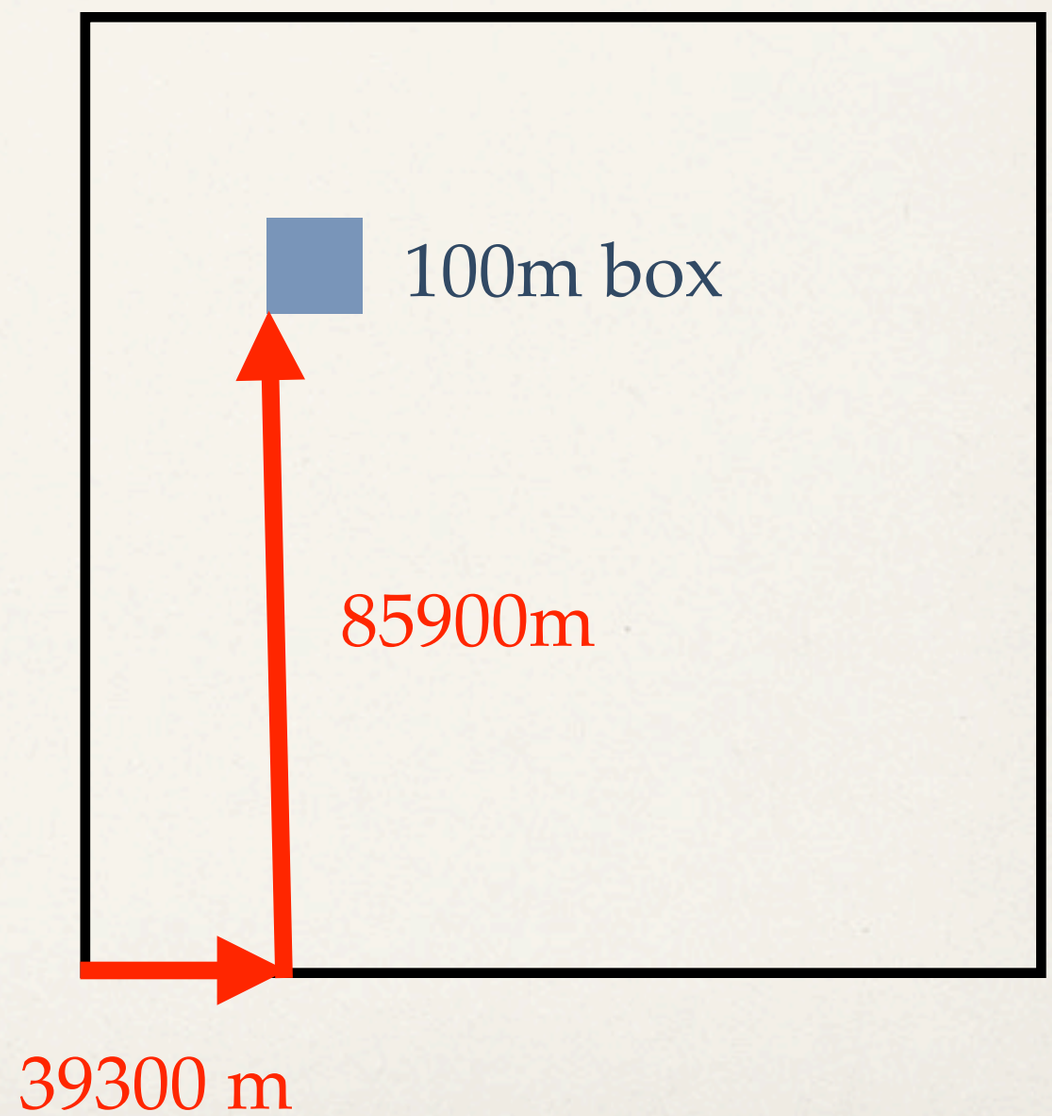
The text however contained location information in the form of **Military Grid Reference System** coordinates. I was able to pull these locations out with a bit more hacking. MGRS coordinates are highlighted in red here.

MGRS: MB393859

Normalized: 38S MB 393 859

Position: 33.31096 lat,
44.34846 lng

Accuracy: 100m



Wednesday, March 9, 2011

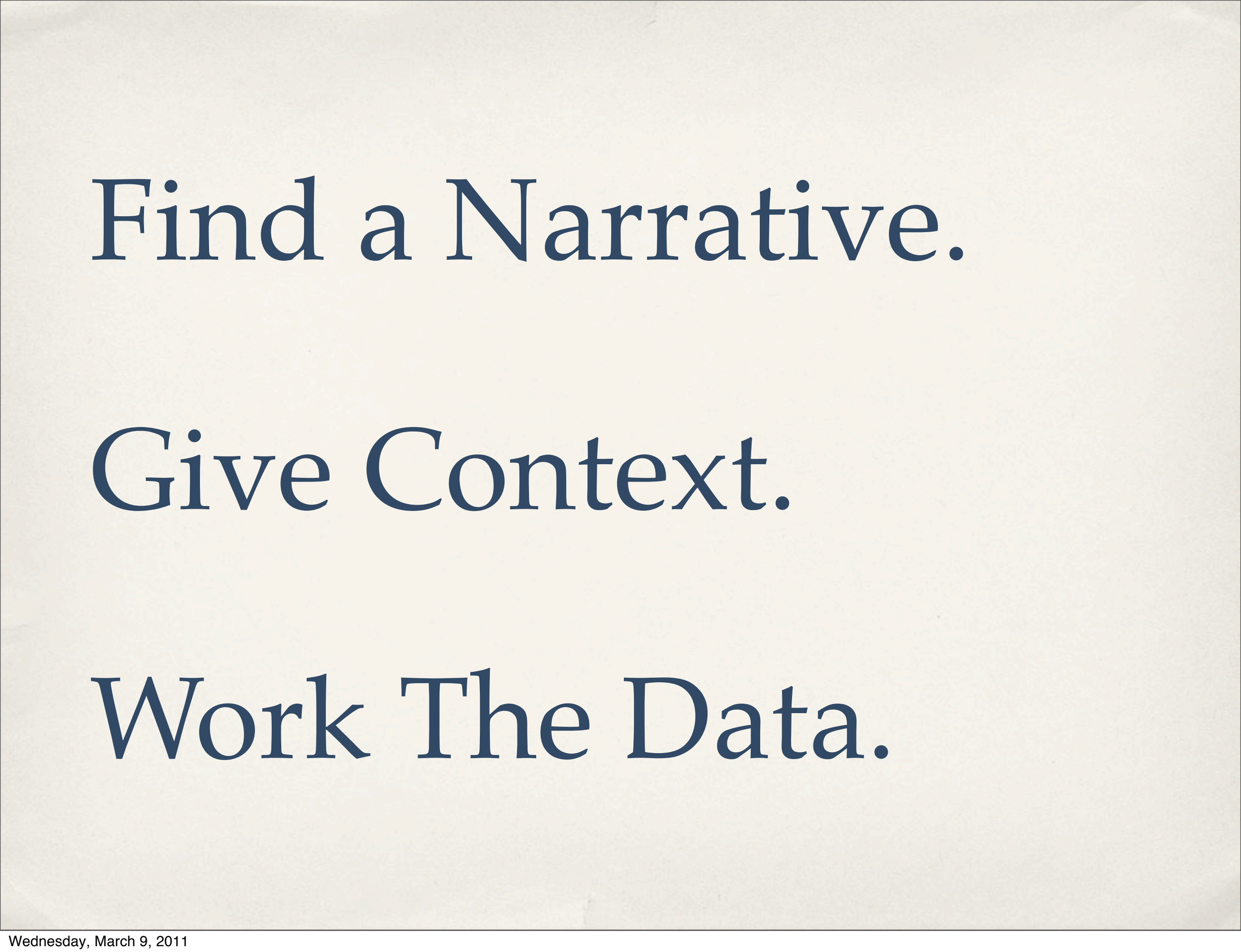
And here is that coordinate geocoded to a lat/lng.

MGRS is an interesting system that specifies locations as offsets within 100,000m squares on the earth. The cool bit is that it can handle variable accuracies of 1m all the way up to 100 kilometers just by changing the string representation.

Look it up on Wikipedia

Wednesday, March 9, 2011

Following up, here are the coordinates associated with that report. The red dot in the center is the location for the report. All the other dots are locations of individual crime scenes around Baghdad extracted from the report.



Wednesday, March 9, 2011

Wrapping up, here are the those three rules again.

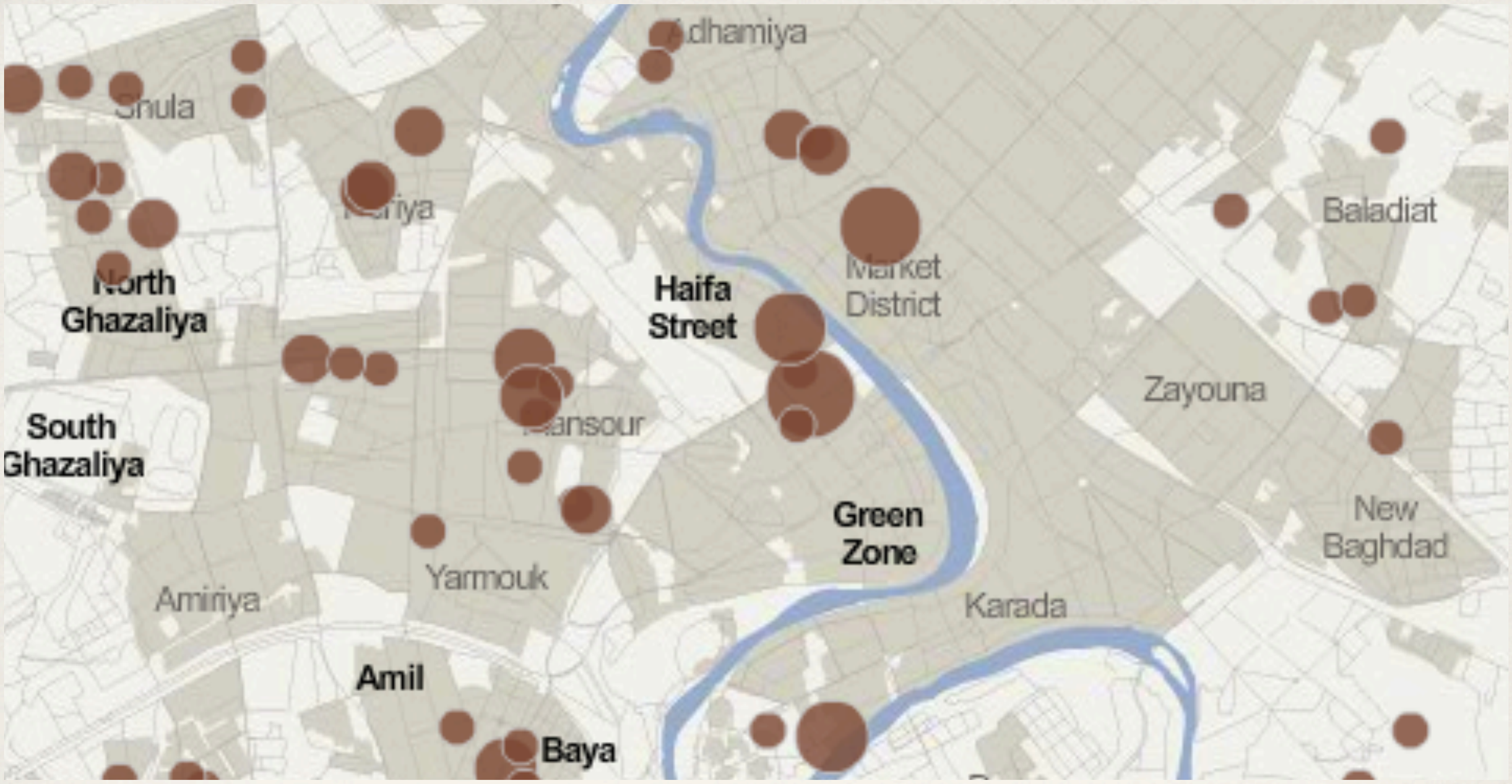
I'm sure there are more rules, and there are more artful ways to express these rules, but these three work pretty well for me.

Wednesday, March 9, 2011

Let me bring back my friend the Word Cloud. And now I hope you can see why I hate them. There is no narrative here, just a jumble of words. There is no context here. How are you supposed to know that COP is short for Combat Outpost or LN means Local National? And there is only a superficial analysis of a promising and problematic data set. The word cloud pretends that counting word frequencies is somehow a meaningful analysis of the Iraq War.

If there are two things I want you to take from this talk, they are this: that data contains narratives that are the basis of journalism, but it is work to report them.

And two, apart from the one case where they are actually appropriate, word clouds are an abomination that reveal nothing more than that you have given up on yourself and your readers.



Thank You

Jacob Harris // jharris@nytimes.com // @harrisj // nimblecode.com

Thank you. Questions?