Big Data

What it is, what it means, and where do we go from here..

Dr. Michael Stachiw

January 10, 2015

Big Data

- **Big data** is an all-encompassing term for any collection of **data** sets so large and complex that it becomes difficult to process them using traditional **data** processing applications *

- * http://en.wikipedia.org/wiki/Big_data

Big Data

- Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages. What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain.*

- Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "*

- *http://en.wikipedia.org/wiki/Big_data

Big Data

- **Examples of Big Data**
  - Crop Yield (GPS driven) data
  - Soil properties (GPS driven) data
  - Marketing/Sales information
    - Frequent buyer programs
  - Real time performance data. For example
    - Aircraft
    - Race cars
    - NASA rockets

Big Data

- **Examples of Big Data (cont.)**
  - Financial (stocks/bonds) trading information
  - Web real-time marketing (Amazon recommendation engine for example).
  - Live Purchase/selection suggestions: Netflix's recommendation engine.

Big Data

- **How does the data get so big?**
  - Tools now available to collect data in real time
  - More quantifiable measurements available
  - More opportunities to collect data
  - Just the fact that there are more people on the planet, interconnected, & global economy create more data opportunities

Big Data

- **Why is big data difficult to work with?**
  - Most tradition tools analyze data in a sequential/linear fashion. For example most statistical “tests” require multiple passes at the data.. Which if is 10x millions of records, can take hours on a pc to process.
  - New analysis paradigms often require “branching” type operations across multiple tables & sources/vendors of data.

Big Data

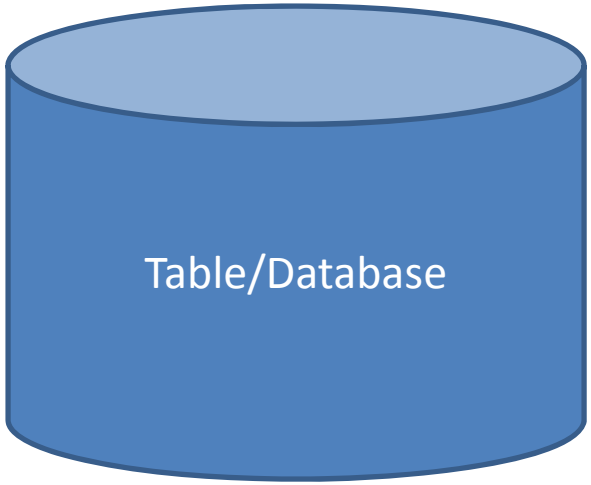
- **Issues other than size to consider:**
  - Data Structure
  - Privacy
  - Intellectual Property
  - Who owns what

Big Data

- Data Structure
  - The amount of data structure required is proportional to the size of the data.
  - More structure can reduce data size, but can increase processing requirements
  - Structure can add value/information to data

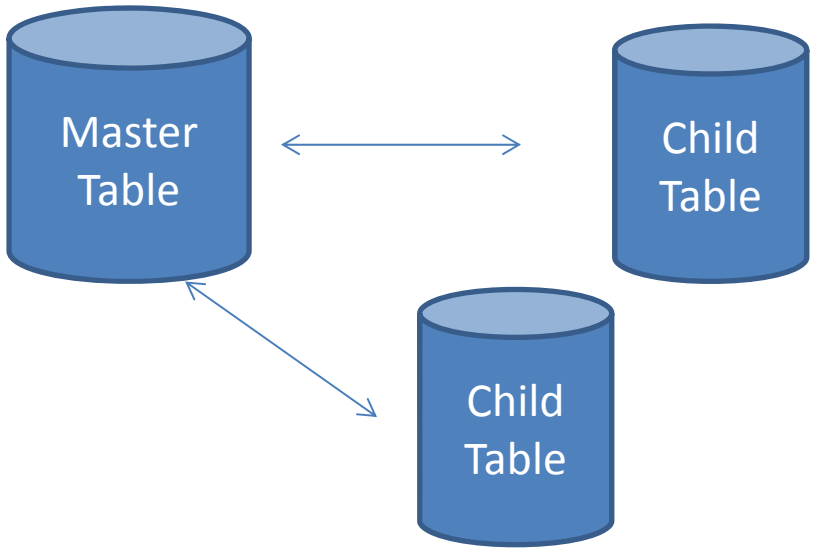
Big Data

- Structure.
  - Multiple-Tables vs one big table
- Unstructured data layout:



Big Data

- Typical structured data layout:



- **Structuring of Data**
- **Pro's:**
  - Less space (typically) required
  - Adds processing opportunities
- **Con's:**
  - More time (typically) required to produce reports (multiple searches across multiple tables)
  - More maintenance issues

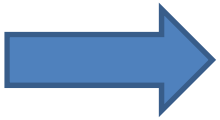
Big Data

Example of Big Data, collected by multiple sources

- **Two main Databases:**
  - Marketing database – 1,000,000 + records. 4.5Gb in size
  - Email Campaigns – 7,000,000 + records. 3.5Gb in size
  - Both databases are highly structured

Big Data

Snippet of tables list in main marketing database, contains in total 56 tables.



```
23:   horseMkt
24:   HumanMkt
25:   LlamaMkt
26:   MarketSegment
27:   NameAddress
28:   OutdoorWildlifeMkt
29:   PIN
30:   PlanningMkt
31:   PMINutritionMkt
32:   PoultryEzines
33:   PoultryMkt
34:   Prospector
35:   PureBredCattleMkt
36:   PureBredDairyMkt
37:   PurinaDays
38:   RabbitMkt
39:   RatiteMkt
40:   SheepMkt
```

Big Data

Fields within name/address table

No:	Field Name	Type	Width
1:	ID	Character	20
2:	COMPANY_NAme	Character	30
3:	PREFIX	Character	5
4:	FIRST_NAME	Character	20
5:	MIDDLE_INItial	Character	1
6:	LAST_NAME	Character	20
7:	ADDRESS1	Character	40
8:	ADDRESS2	Character	40
9:	ADDRESS3	Character	30
10:	CITY	Character	20
11:	STATE	Character	2
12:	ZIPCODE	Character	6
13:	ZIP_4	Character	4
14:	COUNTRY	Character	20
15:	PHONE	Character	20
16:	E_MAIL	Character	60
17:	OK_TO_EMAIL	Character	1
18:	OK_to_Contact	Character	1
19:	SOURCE	Character	60
20:	UPDATED	Date	8
21:	External_Id	Character	15
22:	Suffix	Character	15
23:	Referral_Code	Character	20
24:	Where_Hear	Character	30
25:	Friend_Count	Int32	4
26:	ReferredBy	Character	30
27:	Dealer_Id	Character	20
28:	Dealer_Name	Character	25
29:	Dealer_City	Character	20
30:	Dealer_State	Character	2
31:	BirthDate	Date	8
32:	MobilePhone	Character	20
33:	Self_Title	Character	20
34:	FIPS	Character	10
35:	Latitude	Double	8
36:	Longitude	Double	8

Big Data

Partial listing of fields in horse marketing table

12:	Id	Character	20
13:	Board_horses	Character	1
14:	Own_property	Character	1
15:	Num_Horses	Double	8
16:	Num_Older	Double	8
17:	Owned_Horses_1yr	Character	1
18:	Owned_Horses_1to6yr	Character	1
19:	Owned_Horses_6to10yr	Character	1
20:	Owned_Horses_11yr	Character	1
21:	Feed_All_PMI	Character	1
22:	Feed_Some_PMI	Character	1
23:	Feed_No_PMI	Character	1
24:	Supplement_No	Character	1
25:	Supplement_Herbal	Character	1
26:	Supplement_Joint	Character	1
27:	Supplement_Vit	Character	1
28:	Supplement_Hi_Fat	Character	1
29:	Deworm_0	Character	1
30:	Deworm_1	Character	1
31:	Deworm_3	Character	1
32:	Deworm_Daily	Character	1
33:	Vet_Sees_Horse_0yr	Character	1
34:	Vet_Sees_Horse_1yr	Character	1

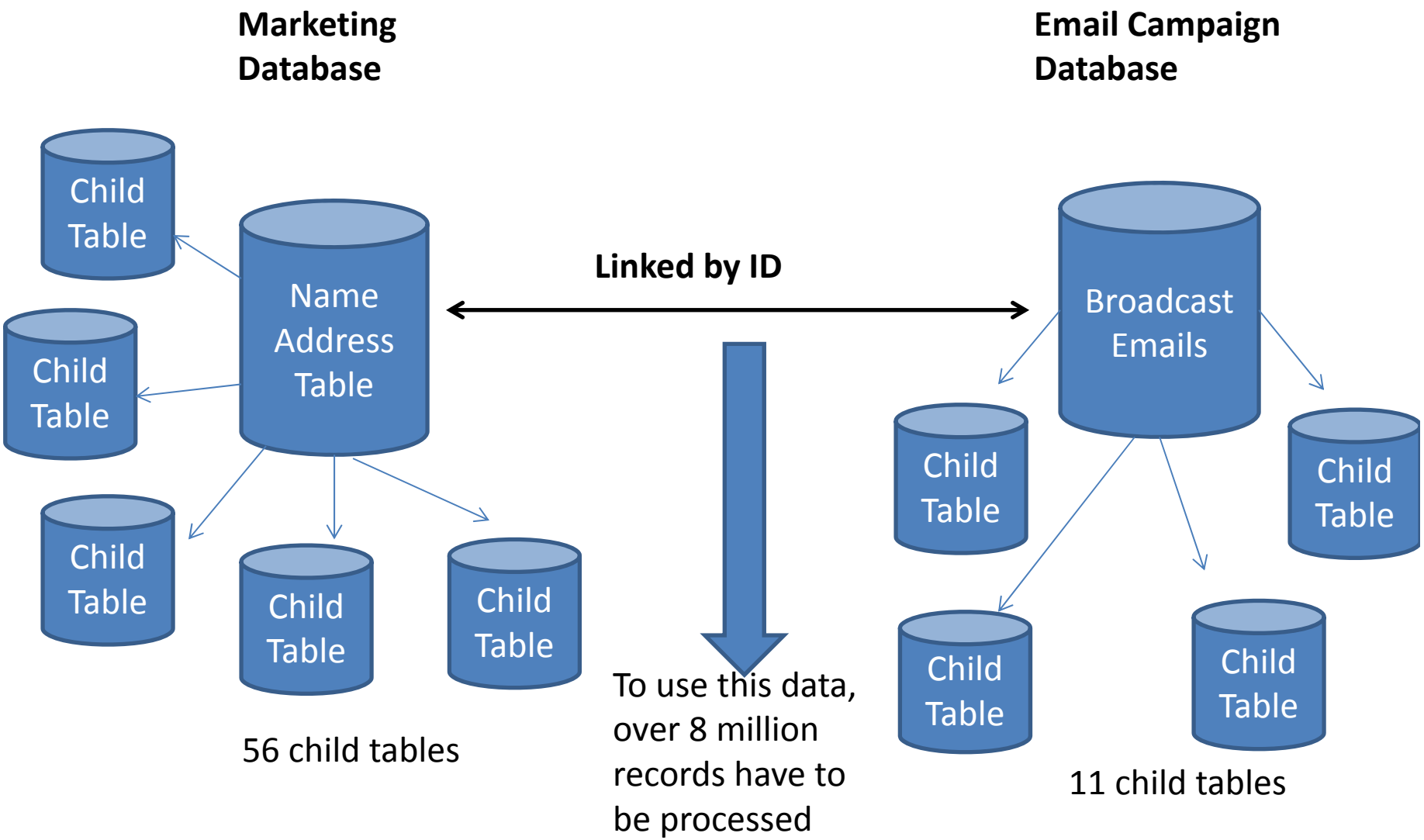
Big Data

List of tables in email marketing campaigns database

```
The file contains a total of 11 Tables:

1:      BouncedEMails
2:      broadcast
3:      CampaignEfforts
4:      Campaigns
5:      Campaign_Effort_Offers
6:      Clicks
7:      Coupon_CLicks
8:      Coupon_Data
9:      Coupon_Links
10:     Links
11:     OpenedEMails
```

Big Data


Marketing Database
Email Campaign Database
Child Table
Child Table
Child Table
Child Table
Child Table
Name Address Table
Linked by ID
Broadcast Emails
Child Table
Child Table
Child Table
Child Table
56 child tables
To use this data, over 8 million records have to be processed
11 child tables

Big Data

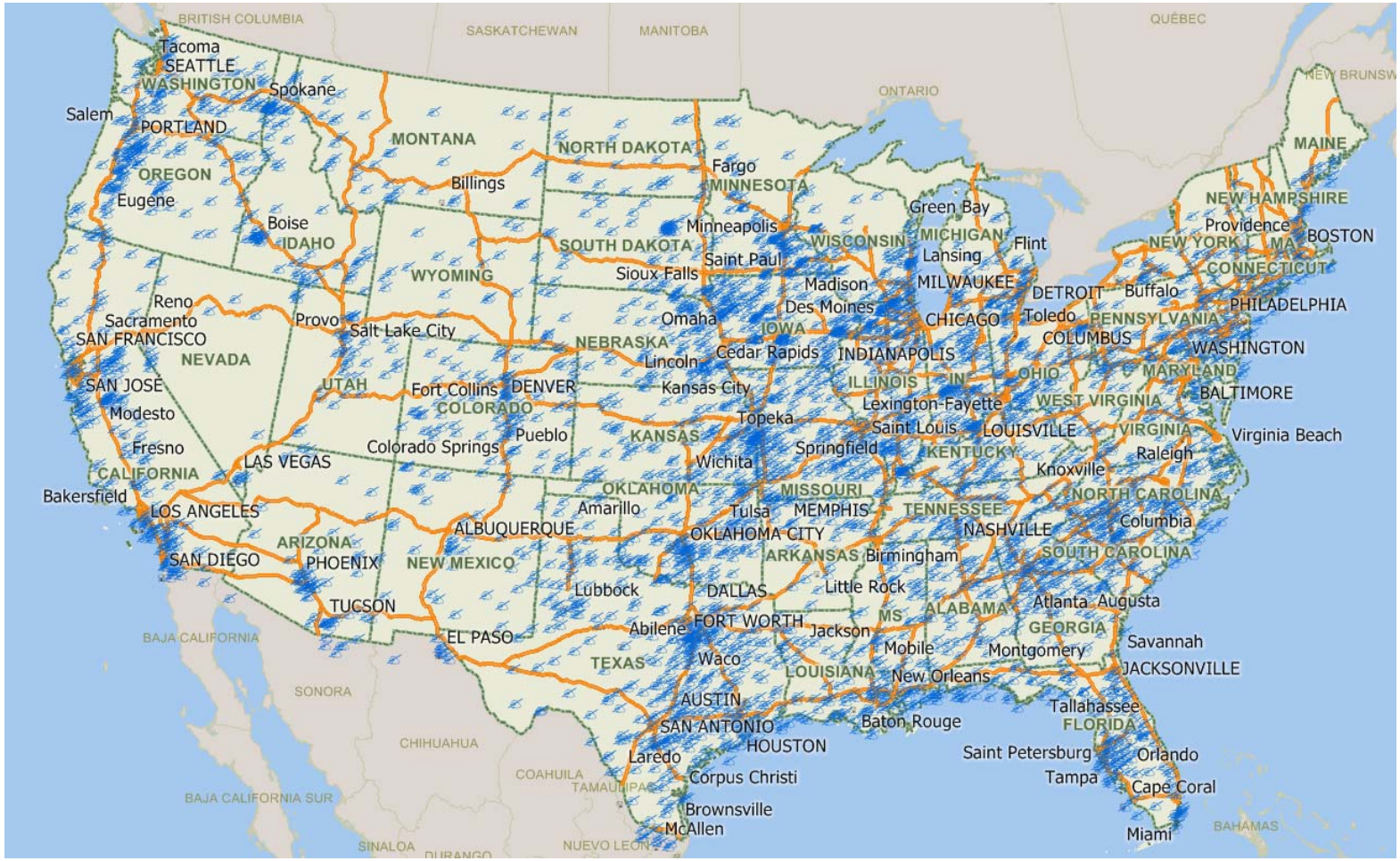
- Our Customer (both corporate & independent dealers) have us: Using Geospatial & prospect/customer demographic restrictions
  - Send out over 100 email blasts/month with over 2 million emails in those blasts
  - Prepare 50-100 mailing label sets/month
  - Prepare 30-40 maps/reports each month to support marketing & sales campaigns

Big Data

- **Multiple Issues exemplified by this example**
  - Who owns the data
  - Stepping on neighbors toes
  - Processing requirements

Big Data

Map of independent dealers who assist in collecting and using the databases



Big Data

- **US is Approximately**
  - 2,600 miles wide
  - 1,500 miles tall
  - Representing 3,900,000 sq miles
  - Approximately 5,000 independent dealers associated with my customer
  - That means on average, each dealer covers 780 sq miles

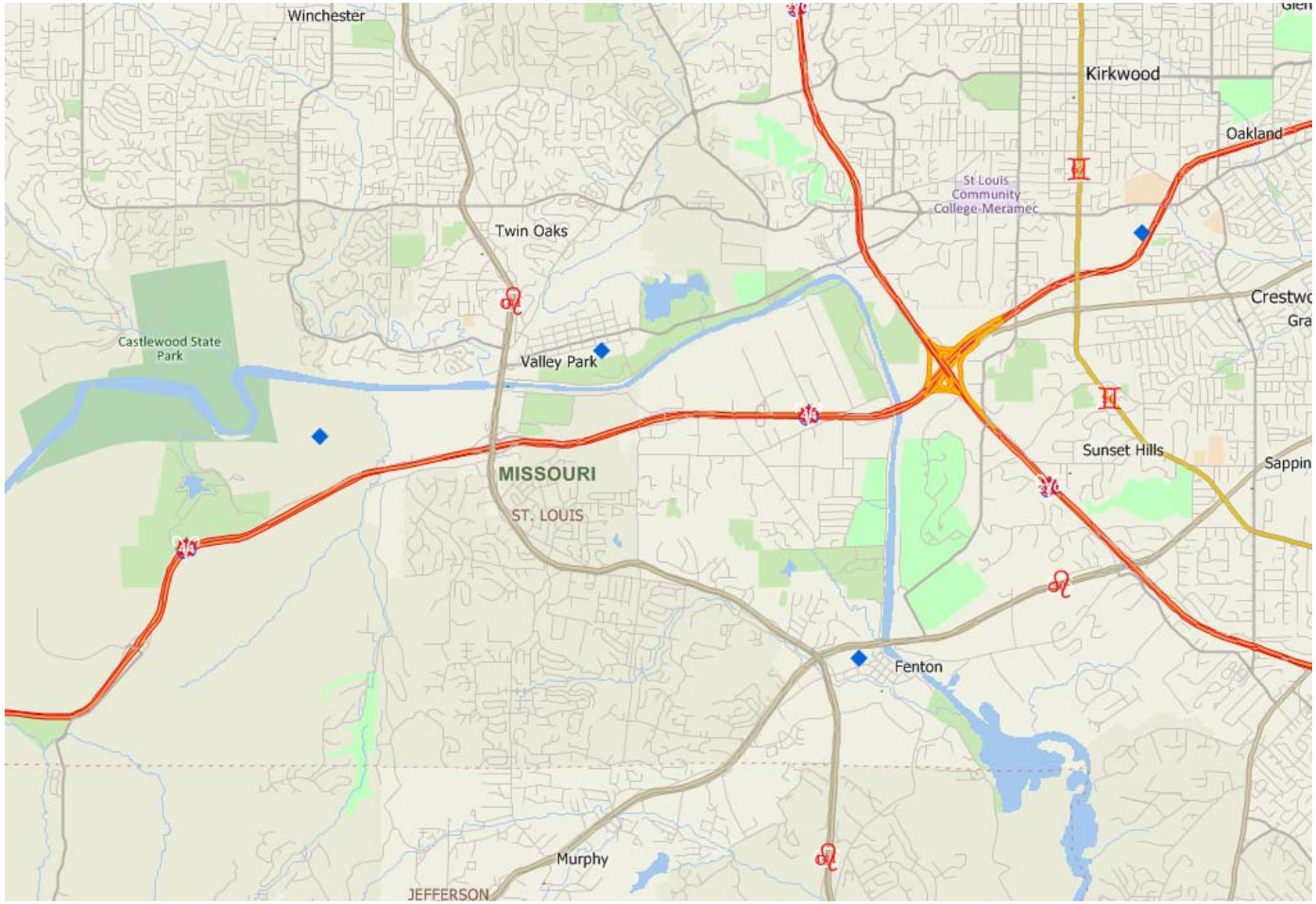
Big Data

- As big as that square mile coverage sounds, its only 27.9 x 27.9 miles in size.
- Very easy to have more than one dealer in the same space

- Issues:
  - Dealer vs Corporate data ownership
  - Dealer vs Dealer data usage
  - Vendor vs corporate intellectual property

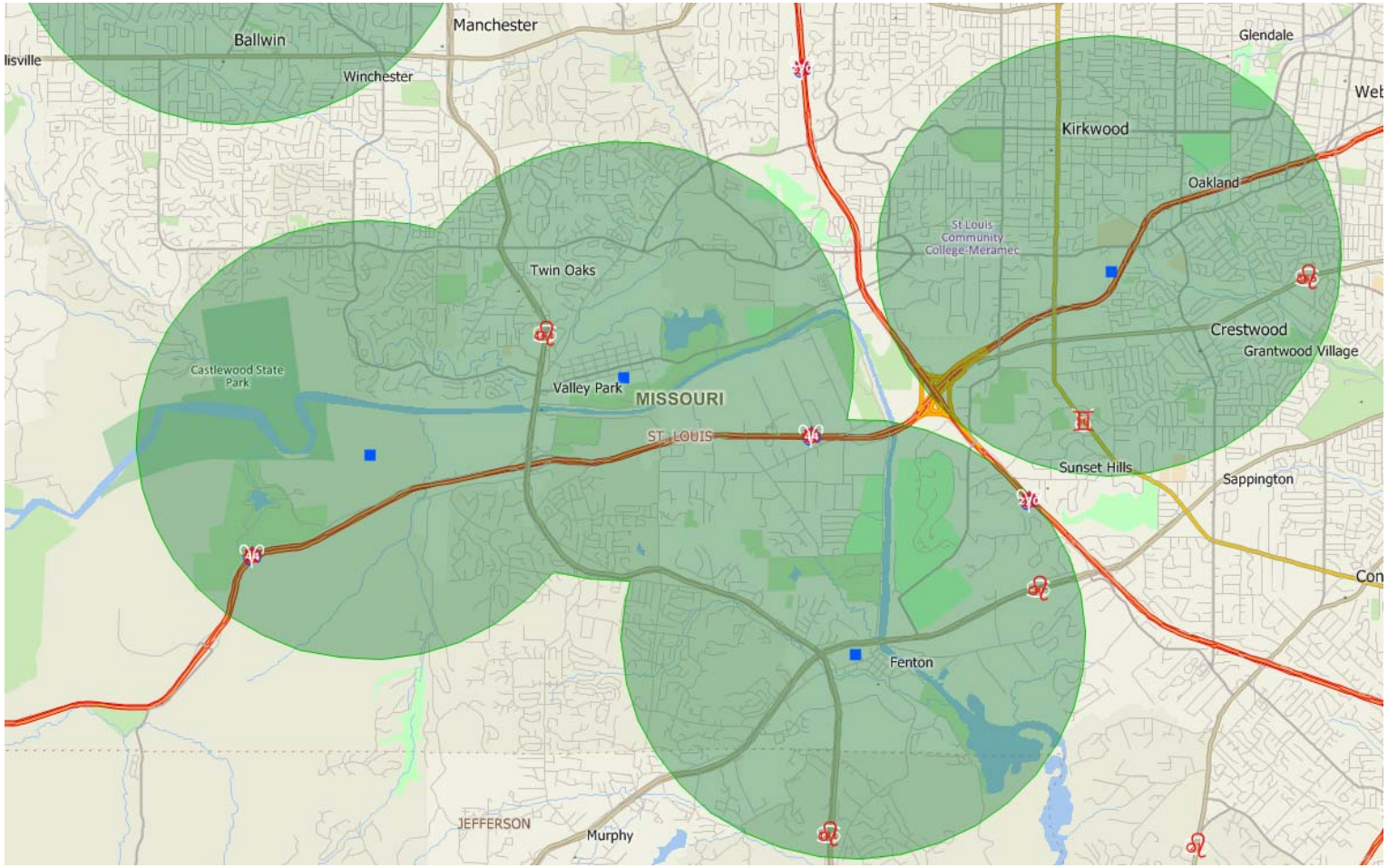
Big Data

Example of proximity of independent dealers to each other



Big Data

Buffers (1 mile) around each independent dealer



Big Data

- **To contact me:**

Dr. Michael Stachiw
Strategic Mapping & Data Services LLC
10715 Kahlmeyer Dr.
St Louis, MO 63132
314-428-3156

Dr.Mike@FeedDealer.com