Daring to Endeavor

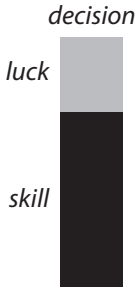
Part of an omni-channel strategy

Would you rather spend budget across all customers or more on high value customers?
Are you interested in high value customers or customers with the most value potential?
Do you think a single message will work on all value potential customers?
Are there value potential prospects you don't know about?
Are you a mortgage company, a technology company or a data company?
Is internal data enough to drive competitive advantage?

Part One: Data Acquisition

Data Engineering = Risk Management

70% of good judgement is skill, 30% is luck



if we have no luck and pure skill, we have a 7 in 10 chance

if we have 40 skill (out of 70), and pure luck (30%), we have a 7 in 10 chance

luck = risk!

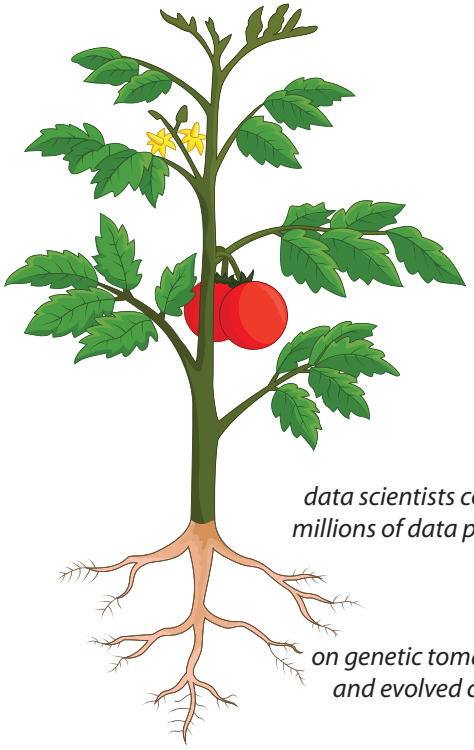
data engineering attempts to cut luck down by 25%+

it uses any numbers behind the skill judgement
and attempts to solve for the negative (anti-human)
Human's judgement is A, computer analyzes -A → *Risk Management*
This better informs the human's judgement → *what are the risks of making decision A?*

Human brains are very intuitive, but can't absorb and process millions of data points

Skill/Luck Example

The Flavr Savr Tomato



data scientists collect millions of data points

on genetic tomatoes and evolved crops

Problem: We need more tomatoes to feed our population
We need better tomatoes (ripen faster, stay red longer, insect-proof)

Currently: Natural selection (evolution over millions of years) and selective breeding

Proposed: DNA-splicing with a fish (genetically modified food) → *we did it for corn and soybeans why not tomatoes*

Skill: Did it work and are we producing more crops?
Did the new tomato kill anyone?

Luck: What's the risk of fast-tracking tomato evolution?

tastes great and no one died
put it in the win column!

not so fast
enter the data scientists
there is nothing data scientists won't analyze

Data engineering has no hypothesis, no proposed outcome
It follows the data and looks for anomalies

For example, the data may tell us that genetics are overtaking evolved
We grow less non-genetic tomatoes (why not, genetics are superior! ... ?)

from this to this → GMO

then disease strikes — *genetic tomatoes die*

→ *but one of these may have survived*

Only data engineering can model this potential outcome

Data Engineering Needs Unencumbered Access to Data

ListHub

5. Except as permitted by a separate agreement in place between the applicable listing content source and the applicable Industry Participant, *Network* participants shall not modify, redistribute, sell or create derivative works based on or containing the content received from the *Network*, provided however that *Network* participants may utilize such content in preparing averages, indexes, and trends.

REALTOR® MULTIPLE LISTING SERVICE MLS

2. Aggregate MLS feeds across the county – This route still requires that you aggregate hundreds of MLS feeds (there are roughly 900 MLS') to get comprehensive. Plus, you'll have to have a sponsoring real estate agent/broker in each market. Additionally, if you aggregate MLS feeds, you are bound by MLS rules that vary from MLS to MLS

HOME Buyers MARKETING²

THE NATION'S LARGEST ONLINE REAL ESTATE COMPANY

6. Confidential Information. You will have a right to access and use HBM II Confidential Information when you are granted access to this Website. HBM II Confidential Information will mean the Confidential Information owned by or licensed to HBM II or its Affiliated Companies. You will only have the right to use the HBM II Confidential Information as provided for in this Agreement or as approved by HBM II in writing and you agree not to provide any person with your password or access to this Website or use any HBM II Confidential Information for any other purpose. You agree that you will not disclose, copy, reproduce, upload, decompile, transfer, modify, use, publish, display, reverse engineer, recreate, duplicate, resell, distribute, convert or exploit any HBM II Confidential Information except as expressly authorized in this Agreement or as approved in writing by HBM II. HBM II Confidential Information will mean and include (a) this password-protected Website and all web pages and all other password-protected websites and web pages designed, created and developed by HBM II or its Affiliated Companies, including all passwords, content, text, information, images, links, color schemes, graphics, images, illustrations, layout, look and feel, audio and video clips, metrics, methodology, programs, and functionality; (b) business information, business development concepts and systems, business practices, "know how," techniques, devices, formulas, procedures, methods of operation and processes of HBM II and its Affiliated Companies; (c) all marketing programs and methods of HBM II and its Affiliated Companies; (d) the manner, protocol, and/or method developed or utilized by HBM II and its Affiliated Companies to market home mortgage lenders and loan officers to Home Buyers and strategic business partners; (e) proprietary information developed, created or owned by HBM II and its Affiliated Companies; (f) trade secrets; (g) proprietary databases and related data, computer processes, computer systems, computer software programs, graphical user interfaces, methods of operation, processes, procedures, techniques and know-how embodied in or generated by the software, and all source or object codes for all computer software programs invented, developed, created, licensed and/or owned by HBM II and its Affiliated Companies (excluding commercially available off-the-shelf third-party software programs); (h) all Copyrighted Works that have not been publicly disclosed by HBM II and its Affiliated

All data vendors have usage restrictions...

Zillow®

- You may present the Zillow Data only on a transactional basis. You will not permit your users to access the Zillow Data in bulk.
- You may not retain any copies of the Zillow Data. Your license to Zillow Data is limited to making direct server calls to the Zillow API for the Zillow Data and distributing the Zillow Data to your end user(s) on your Site(s), immediately upon receipt by your servers.
- You may not present the Zillow Data (or permit the Zillow Data to be presented) so that it appears to be available from a third party website.

CoreLogic®

Confidential information that is legally required to be furnished; and (c) the receiving party shall use reasonable efforts to seek from the party to which the information must be disclosed confidential treatment of the disclosed Confidential Information. Notwithstanding that portions of the Data may be derived in whole or in part from publicly available sources, the Data and any of CoreLogic's databases used in deriving the Data are proprietary, copyrighted and trade secrets of CoreLogic and, for the avoidance of doubt, the restrictions on use and disclosure of the Data are not subject to the exceptions contained in this Section 5.3.

Data engineering can:
predict who is likely to refinance
predict which verbal complaints will become CFPBs
predict loans likely to become non-performing
predict customers likely to sell home soon
assess how customer experience solutions are performing
analyze chat logs and match up products with needs
offer analysis that further compels customer to refinance
... the list is endless...

... and it's getting worse

Realtors have finally started to push back against the questionable practices of the real estate listing aggregators (or "syndicators"), Zillow, Trulia, Realtor.com and others. For years, these deep-pocketed online media corporations have been growing exponentially at the expense of the professional real estate community. The aggregators take the intellectual property of individual Realtors, mix it with other content, re-brand the property as their own, and then sell related advertising rights back to the Realtors who they took the property from in the first place.

#1 challenge for data engineering? Math? Computing power? No. Data access.

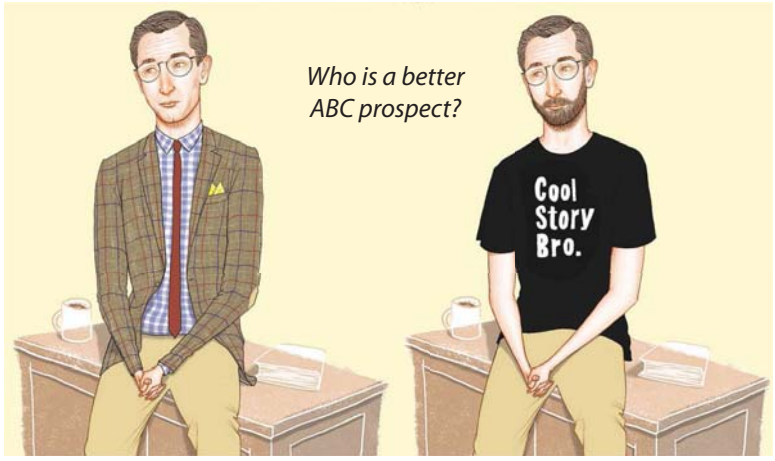
Zillow/HBM/MLS/etc are great if you are buying or selling a home
But not for a $1.5 billion mortgage powerhouse

"My house is for sale"
Too late for approaching customer on refinancing

"My house is for sale"
Probably too late for recapturing that customer

"I just made an offer on a house"
Probably too late to originate a loan

If MLS+ knows it, the opportunity for ABC is in rear view mirror
ABC needs 10 times the (Zillow) data to predict the need before it goes public



Data engineering has the best shot at answering that question

Unencumbered Local Government Data

Property Tax Data

previous ownership
owner name
owner name
owner address
legal description
property name
property address
state class code
land use code
landArea
total living area
value date
land market value
land appraised value
improvement market value
improvement appraised value
total market value
total appraised value
year built
remodeled
type
quality
square feet
bedrooms
baths
neighborhood
neighborhood group
market area desc
map facet
key map
(and many others)

find large investors renting homes

data scientists need "boundaries" for analysis

what is the performance of this data versus data in the immediate area?

Permit Data

address
permit status
permit type
proj description
proj status
applicant name
contractor name
valuation amount
(and 20+ other fields)

new home construction
remodelling
electrical
mechanical
plumbing
etc...

Ownership Data

previous owners
sell date

Shapefile Data

latitude
longitude
address

GIS Examples:
boundaries
parks
schools
libraries
rail
crime
traffic
roadways
census data
fire
floodways
incomes

data scientists never say no to more data

correlations can be anywhere

etc...

marital status
birth and deaths
etc...

whatever the government tracks

For Sale Data

property url
address
city
state
zip
latitude
longitude
images url
photo count
beds
baths
days for sale
square feet
lot size
year built
price
price sqft

only basic data needed rest and much more from local government can get off internet link back to source(s)

Test Case One: Harris County (Houston, Texas)

Parsed government data

can communicate property tax record directly to customer with link
many customers/prospects are unaware of this site and will be impressed

Tax Year: 2015

HARRIS COUNTY APPRAISAL DISTRICT
REAL PROPERTY ACCOUNT INFORMATION
0382420000001

Print E-mail

File A Protest | Similar Owner Name | Nearby Addresses | Same Street Name | Related Map 5356A

Ownership History

Owner and Property Information

Owner Name & Mailing Address:	**DIMEO ARMANDO** **1105 CENTER ST** **DEER PARK TX 77536-3211**
Legal Description:	**TR 1 BLK 7** **CHERRYHURST & WINDSOR T/H U/R** **CHERRYHURST**
Property Address:	**1630 CHERRYHURST ST** **HOUSTON TX 77006**

State Class Code	Land Use Code
A1 -- Real, Residential, Single-Family	1001 -- Residential Improved

Land Area	Total Living Area	Neighborhood	Neighborhood Group	Market Area	Map Facet	Key Map®
1,100 SF	2,477 SF	8319	1629	163 -- 1F Montrose, Fourth Ward Areas	5356A	492V

Valuations

Value as of January 1, 2014			Value as of January 1, 2015		
	Market	Appraised		Market	Appraised
Land	74,250		Land	96,525	
Improvement	326,314		Improvement	369,877	
Total	400,564	400,564	Total	466,402	466,402

5-Year Value History

Land

Market Value Land

Line	Description	Site Code	Unit Type	Units	Size Factor	Site Factor	Appr O/R Factor	Appr O/R Reason	Total Adj	Unit Price	Adj Unit Price	Value
1	1001 -- Res Improved Table Value	SF1	SF	1,100	1.35	1.00	1.00	--	1.35	65.00	87.75	96,525.00

Building

Building	Year Built	Remodeled	Type	Style	Quality	Impr Sq Ft	Building Details
1	1973	2003	Residential Single Family	101 -- Residential 1 Family	Good	2,477 *	Displayed

Building Details (1)

Building Data	
Element	Details
Cost and Design	Extensive
Cond / Desir / Util	Very Good
Foundation Type	Slab
Grade Adjustment	B+
Heating / AC	Central Heat/AC
Physical Condition	Good
Exterior Wall	Brick / Veneer
Element	Units
Room: Total	7
Room: Rec	2
Room: Full Bath	2
Room: Bedroom	2
Fireplace: Masonry Firebrick	1

Building Areas	
Description	Area
BASE AREA PRI	895
OPEN MAS PORCH UPR	55
BASE AREA UPR	895
OPEN MAS PORCH UPR	45
MAS/BRK GARAGE LWR	475
BASE AREA UPR	191
BASE AREA LWR	496

detects home that are rentals
aggregated investor analysis

can limit to single family homes

can calculate price/square foot

shows remodel

Zillow-like information
In fact, Zillow probably
sources this same data

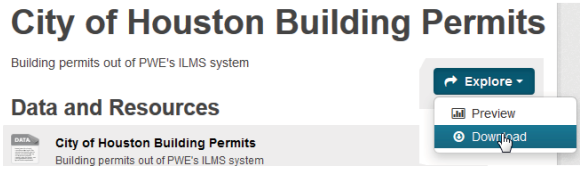
can track historical buy/sale

Ownership History: 0382420000001	
1630 CHERRYHURST ST HOUSTON TX 77006	
Owner	Effective Date
DIMEO ARMANDO	9/15/2015
MARKOFF ALAN S	8/17/2006
EASTHAM MATTHEW & VALERIE	6/14/2002
PRUDENTIAL RESIDENTIAL	12/15/2001
BULLARD CHRISTOPHER K &	8/27/1998
ASHCRAFT JOHNNY R	4/7/1997
MURRAY ROY III	1/2/1988

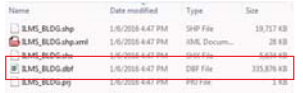
set boundaries for relative analysis

county appraised value is an excellent proxy for "automated value model" (which charges for this information)

many shapefiles are available for markets



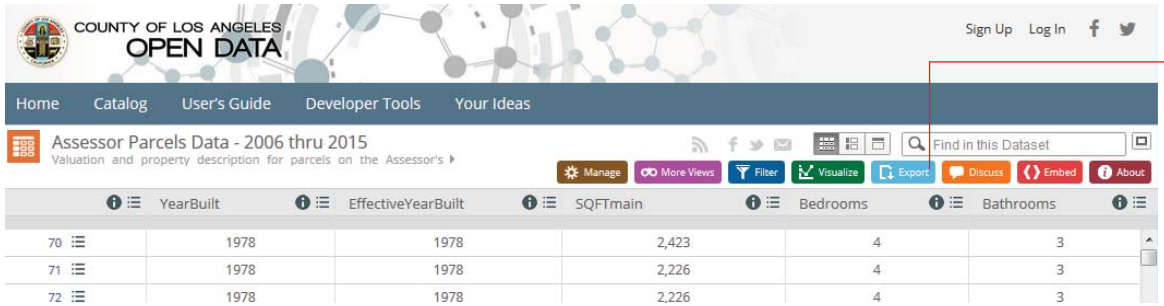
shapefiles have large database files inside + geocoordinates



dear customer
would you like to refinance? maybe take equity out?
did you know there is a lot of remodeling in your area?
did you know price/square foot going up 20%?
did you know houses are selling very quickly?
... now about that refinance question...

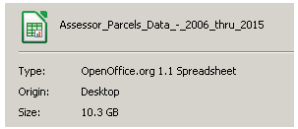
Test Case Two: Los Angeles County

Downloaded government data



While there are 23MM rows, this is for 10 years of data and includes some commercial & industrial permits
2,134,504 are residential permits, 1,884,961 of which are single family residences

Los Angeles County is a 10GB text file

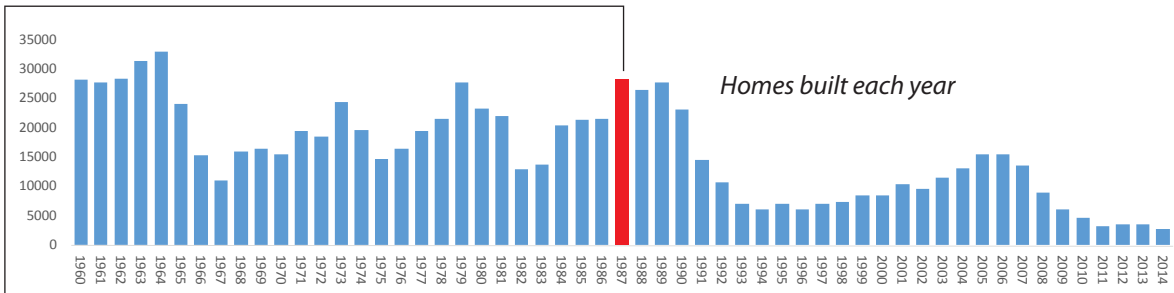


23,749,930 rows of raw data

51 data fields
x 23,749,930 rows
1,211,246,430 data values

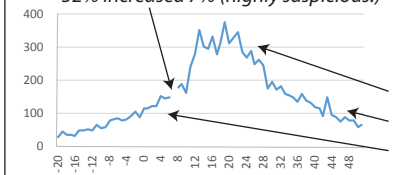
data scientist will use every it all
and this is only one county in U.S.

- *ain*
- *assessorid*
- *bathrooms*
- *bedrooms*
- *latitude*
- *longitude*
- *city*
- *cluster*
- *effective year built*
- *fixture exemption*
- *fixture value*
- *general use type*
- *homeowners exemption*
- *house fraction*
- *house no*
- *imp base year*
- *improvement value*
- *is taxable parcel*
- *land base year*
- *land value*
- *net taxable value*
- *parcel boundary description*
- *personal property exemption*
- *personal property value*
- *property location*
- *property type*
- *property use code*
- *real estate exemption*
- *roll year*
- *specific use detail*
- *specific use type*
- *sqft main*
- *streetname*
- *tax rate area*
- *total exemption*
- *total land imp value*
- *total value*
- *year built*
- *zip code*
- *(etc)*



Homes built each year

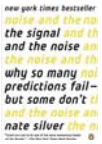
52% increased 7% (highly suspicious!) ← *Almost all of these were right at 6.89%* ← *Subject for another deck* ← *But data scientist would have a field day*



Home appreciation from 2010 to 2014 for homes built in 1987 (664 were outliers)

Is this data telling us something actionable?
Go after sweetspot?
Go after high appreciation homes?
Go after those on losing end?
And do you monitor negative appreciation?
at risk mortgages?!

Signal to noise problem



They write books about this stuff!
You don't know what's a signal until you study it
Some are turning to machine learning for help

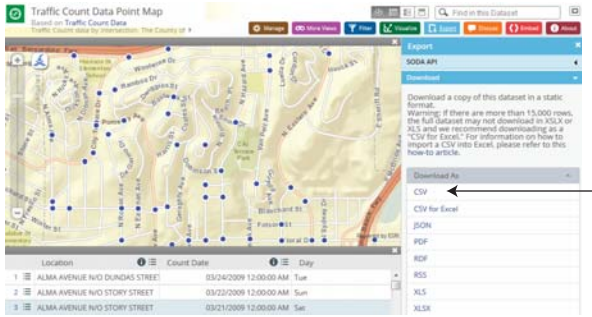
L.A. County Construction Permits

No one is harnessing this data, 23 downloads only

Why this data?
a) data scientists want all data (insatiable)
b) this will show new home construction
c) ABC pitch refinancing (take money out)

Bottom line: We don't know value until we try

Traffic data? Crime data? No idea what's useful...



But even maps can be exported as raw data

Homes for Sale in United States

1,960,310 from largest 750 counties in U.S. ← *counties by population size*

461,706 had MLS numbers, but no reported address ← *guess you have to contact realtor?*

411,217 were listings for land or other non-homes

1,087,387 were for dwellings with 1-10 bedrooms

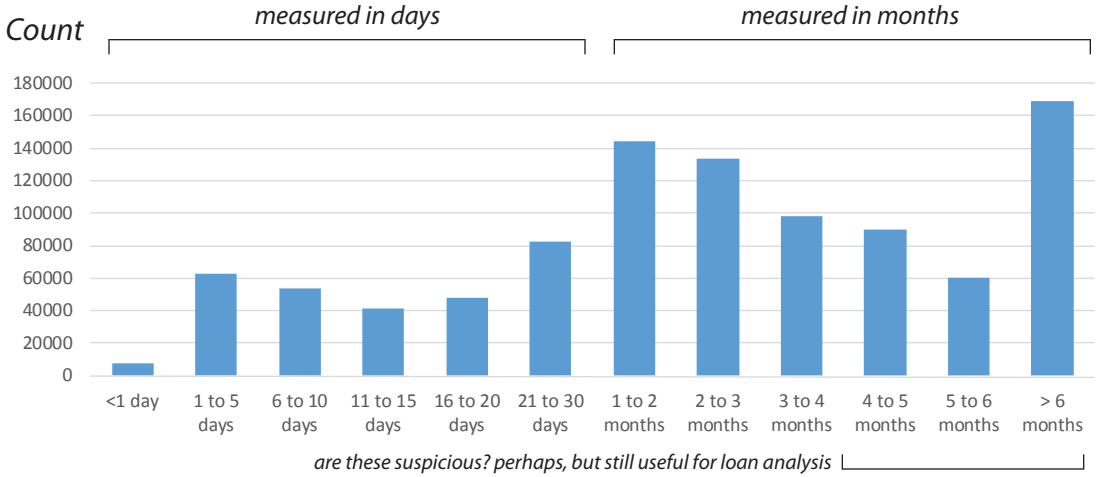
only basic information (search results)

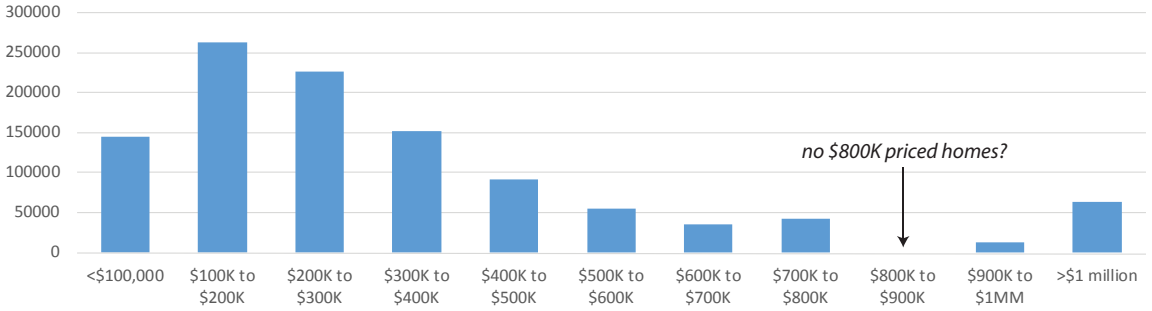
address
beds
city
dayslisted
full_baths
half_baths
images_url
latitude
listingid
longitude
price
pricesqft_reported
property_url
sqft
state
substr
yearbuilt
zip

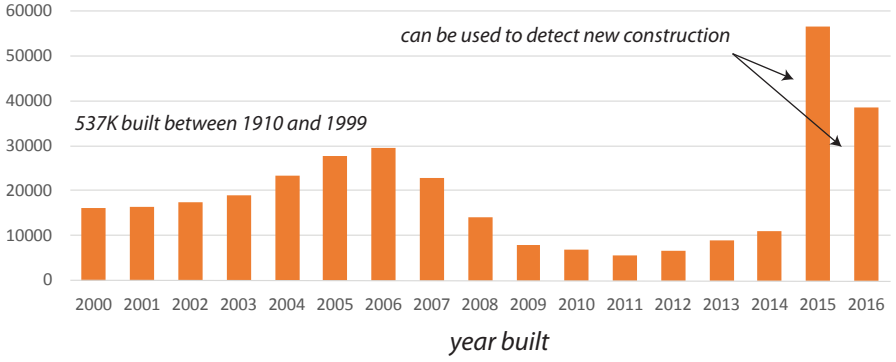
Convert to geohash

once connected to loan data more detailed information can be parsed

suspect nightly overwrite of mls data

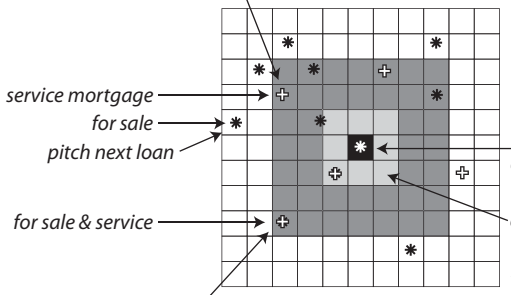






if there is increased home sales activity in vicinity of loan, pitch refinance

Convert lat/long to geohash



if for sale and serviced, pitch recapture
if on market for 6 months, pitch refinance

can also increase hash size

Length	Width		Height
1	≤ 5,000km	×	5,000km
2	≤ 1,250km	×	625km
3	≤ 156km	×	156km
4	≤ 39.1km	×	19.5km
5	≤ 4.89km	×	4.89km
6	≤ 1.22km	×	0.61km
7	≤ 153m	×	153m
8	≤ 38.2m	×	19.1m
9	≤ 4.77m	×	4.77m

dr4zy9wrk length = 9

Bedrooms	Count	Percent
1	31,683	2.91%
2	175,811	16.17%
3	448,759	41.27%
4	309,261	28.44%
5	92,353	8.49%
6	20,811	1.91%
7	4,698	0.43%
8	2,505	0.23%
9	1,030	0.09%
10	476	0.04%
	1,087,387	100.00%

some do not report baths

Full Baths	Count	Percent
1	213,775	19.84%
2	558,936	51.86%
3	213,682	19.83%
4	65,934	6.12%
5	19,020	1.76%
6	6,380	0.59%
	1,077,727	100.00%

This is not theoretical technology

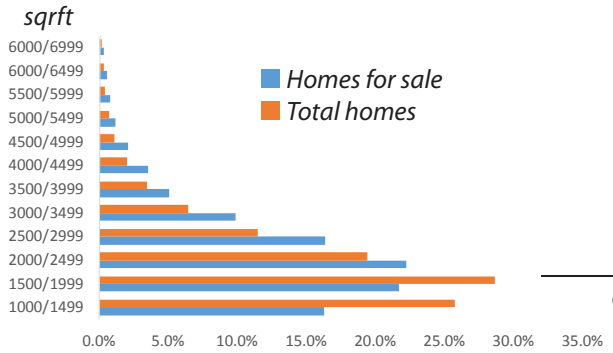
1,188,922 Harris County web pages have been parsed
902,605 are real, residential, single-family
704,968 single family homes are owner occupied
197,537 single family homes are owned by investors

44 different classifications
78K are vacant lots, 45K condos, 26K government buildings, 7K religious
52K are commercial, 27K more in vacant commercial

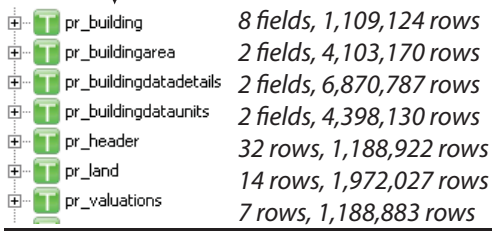
Top 10 single family home investors in Harris County

Tah Holding LP	617
Tarbert LLC	568
Sfr-Hou I LLC	491
American Residential Leasing C/O Altus Group	479
Sway 2014 1 Borrower LLC	476
Pfirman Richard L	295
Amh 2014 3 Borrower LLC	294
Arp 2014-1 Borrower LLC C/O Altus Group	284
Camillo Properties LTD	238
Red Door Housing LLC	215

13,125 houses currently for sale



Larger homes more likely to be for sale (refi targets?)



8 fields, 1,109,124 rows
2 fields, 4,103,170 rows
2 fields, 6,870,787 rows
2 fields, 4,398,130 rows
32 rows, 1,188,922 rows
14 rows, 1,972,027 rows
7 rows, 1,188,883 rows

113,393,229 pieces of Harris County data

different perimeters can be analyzed

Boundary data great for multi-level analysis

Analysis of 815 Rosine Street, Houston 77019

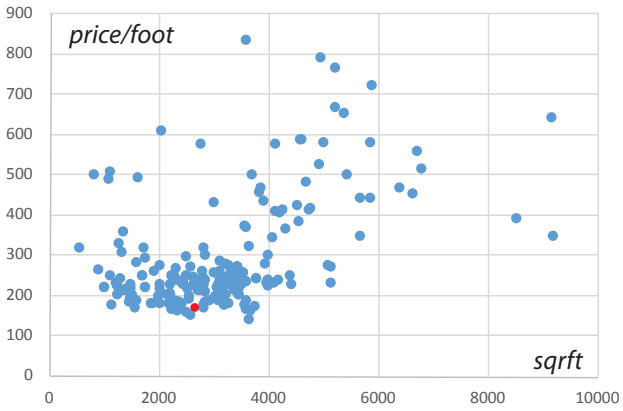
Map Facet = 5357A	*1,351 records*
Property Zip Code = 77019	*1,296 records*
Keymap = 492M	*772 records*
Market Area Code = 163	*142 records*
Neighborhood Group = 1629	*67 records*
Neighborhood = 8319	*25 records*

Radial one that you want to present to customer

Millennials love (and expect) visuals
If you don't reveal visuals (e.g. SmartWatch) to end users, would still use for math part

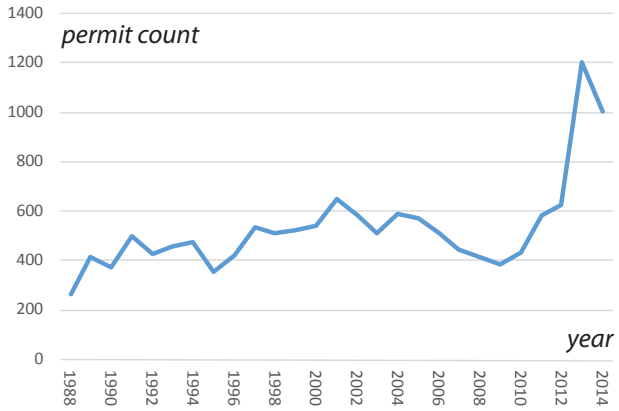
Data engineering helps sell refinancing...

244 homes are for sale in zip code 77019



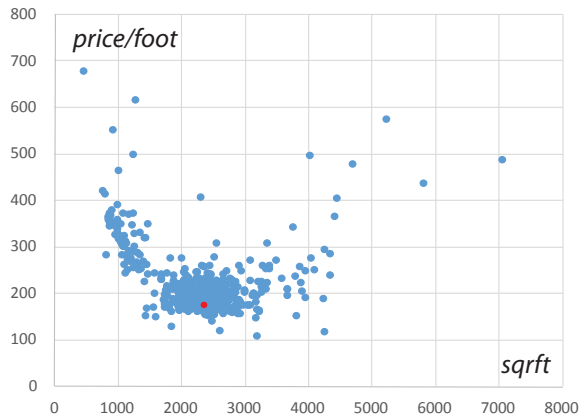
Data can be used for automatic valuation modelling

Permitting is up 100% for zip code 77019



Great time to do a remodel (using a refinanced loan)

Price/Square foot of 815 Rosine = $185, on the low end



May consider a refinance and/or home improvements

Getting the Data

Property tax records and construction permits are only the tip of the iceburg
429 different Los Angeles datasets, 241 datasets in Houston, 194,354 in data.gov

Property tax data
Distance to crime
School systems
Demographics
Gentrification
Home sales

can never have enough metadata

and you can't guess what will predict and what won't

Need millions of rows

Some sample datasets

Albany, NY	https://data.albanyny.gov/City-Finances/City-of-Albany-Property-Tax-Assessment/5592-nkj8
Atlanta , Ga.	http://share.myfultoncountyga.us:8080/geoportal/catalog/search/browse/browse.page
Austin, Tex.	https://tax-office.traviscountytx.gov/reports-data
Baltimore, Md.	https://data.baltimorecity.gov/Financial/Real-Property-Taxes/27w9-urtv
Boston, Mass.	https://data.cityofboston.gov/dataset/Property-Assessment-2014/qz7u-kb7x
Chicago, Ill.	https://catalog.data.gov/dataset/ssmma-property-tax-29d46
Des Moines, Iowa	https://data.iowa.gov/Communities-People/Assessed-Property-Values-By-Tax-District-and-Year/p8sk-8ig7
Everett, WA	http://snohomishcountywa.gov/372/GIS-FTP-Data
Houston, Tex.	http://hcad.org/records/real/default.asp
Los Angeles, Calif.	https://data.lacounty.gov/Parcel-/Assessor-Parcels-Data-2006-thru-2015/9trm-uz8i
Massachusetts	https://data.mass.gov/Permitting/Real-Property-Assessment-2015/ybiy-gfra
New York, N.Y.	http://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page
Philadelphia, Pa.	https://www.opendataphilly.org/dataset/opa-property-assessments
Phoenix, Ariz.	http://mcassessor.maricopa.gov/reports-data-sales/data-sales/data-policies-pricing/
Pittsburgh, Pa	https://data.wprdc.org/dataset/property-assessments
Providence, R.I.	https://data.providenceri.gov/Finance/2013-Residential-Assessment/apvu-8y72
San Jose, Calif.	https://www.sccassessor.org/index.php/online-services/property-search/real-property
Seattle , Wash.	http://info.kingcounty.gov/assessor/DataDownload/default.aspx
St. Louis, MO	http://data.stlouis-mo.gov/downloads.cfm
Washington, DC	https://catalog.data.gov/dataset/levy-year-2014-real-property-tax-rates-449c7

External data tagged with internal data

815 Rosine Street, Houston 77019
Refinance with ABC?
Late payment history?
Open emails at all?
Use call center?

Model will predict who will pay late
Model will predict who will refi
Model will predict who will call

Machine Learning (Primer)

I know how profitable each customer is to me
I have information (metadata) about each customer
I want to predict which metadata correlates to profitability
I then want to go after prospects most like my profitable customers

Use for targeted email campaigns
Use for targeted advertising
Refine targeting messages

Need more than ABC data alone
Need external metadata

Data Collection (this deck)
Acquire sufficient data
garbage in, garbage out

Prepare the data
Data transfer
Missing & outlier values
Normalize and band data

Develop & Train Model
Can use a cloud based solution - algorithms already baked in

Can outsource data classification to vendor
Pay as you go, but data in the cloud
Vendors have different pluses and minuses

IBM Watson Analytics
Launched: December 2014

TensorFlow
Launched: November 2015

amazon web services Machine Learning
Launched: April 2015

Azure Machine Learning
Launched: February 2015