

Zillow Group: Transforming Real Estate through Big Data and Data Science

Jasjeet Thind
Sr. Director, Data Science & Engineering

Agenda

- Zillow Group
- Data Lake
- Tech Platform
- Use Cases
- Deep Dive: Predicting Homeowners for 100M+ homes
- Deep Dive: Personalization
- Q&A

Zillow Group

Build the world's largest, most trusted and vibrant home-related marketplace.



Data Lake



Free Download

<http://www.zillow.com/data>

Zillow Home Value Index (ZHVI)

Zillow Rent Index (ZRI)

Home Related Metrics

Tech Platform



Zestimate



905 Hargis Way,
Santa Clara, CA 95054

4 beds • 3.5 baths • 2,083 sqft

Edit

♥ OFF MARKET

Zestimate®:

\$1,582,165

[Price this home](#)

Rent Zestimate®: \$4,700/mo

Edit home facts for a more accurate Zestimate.

Zestimate

Pricing Tool

Turbo Zest

ZHVI / ZRI

Rent Zest

Market Report

Similar Sales

Similar Homes for Sale

Zestimate Forecast

Rent Zest Forecast

Hot Homes / Listing

Best Time to List

ZHVI / ZRI Forecast

Personalization & Search

Seattle WA Real Estate 224 homes for sale.

[Homes for You](#) [Newest](#) [Cheapest](#) [More](#)



Search Relevance

Personalized
Homes for Sale

Personalized
Search

Personalized
Collections

Recommended
Favorites

Recommended
Homes on HDP /
People also viewed

Related Searches

People Recommendations



Home Owner Predictions

Pre-Seller Predictions


Pre-Investor Predictions

Home Buyer Predictions

Marketing

Premier Agent - Zillow & Trulia over a year ago

Create a plan for success with Zillow's 16-page business plan.



The graphic features a dark blue background. On the left, there are four white house icons of increasing size. A green line graph with an upward-pointing arrow starts from the first house and trends upwards across the others. Below the houses, a green button contains the text 'Download FREE Template'. Underneath the button, the text 'Agent Success Plan' is displayed. The Zillow logo is in the bottom right corner of the graphic.

Free Business Plan Template
Download and zoom into 2015 with a polished business plan.

WWW.ZILLOWPREMIERAGENT.COM

[Download](#)

1.9K 64 368

Social Media
Marketing

Search Engine
Optimization (SEO)

Search Engine
Marketing (SEM)

Automated
Audience Segments

Automated
Creatives

Automated Bidding

Mortgages

A blue banner for Zillow Mortgages. It features the Zillow logo on the left and the text 'Zillow Mortgages' in white. Below this, a line of text says 'We have hundreds of participating lenders with reviews from real customers. Quickly find one near Seattle, WA.' At the bottom, there is a white input field containing the ZIP code '98109' and a green 'Get Started' button.

Zillow Mortgages

We have hundreds of participating lenders with reviews from real customers. Quickly find one near Seattle, WA.

98109

Get Started

Use the ZIP code where you're home shopping.

Lender Recommendations

Borrower Risk (probability
borrower can pay mortgage)

Computer Vision



Listing Images

Listing Videos

Digs

Content Recommendations

What if the Housing Bubble Never Happened?

[Home](#) / [Real Estate and Rental Trends](#)

By [Jamie Anderson](#) on 8/16/2016


- The median U.S. home remains 4.9 percent less valuable than it was at its peak
- In Denver, home values are 45 percent above their April 2007 values, while in Las Vegas, they are 15 percent below peak.
- If home values had grown steadily at historic rates, rather than boom and bust, the median U.S. home would be worth 26 percent more today than it actually is.

Content Newsfeed

Digs Similar Photos



Similar Content

Data Quality

**tmcintyre689**
0 contributions

I have been trying to sell my home for approx. 2 1/2 months, its a very nice house,

my address is 115 califorina st. vallejo, I have seen many houses sell since I started this process, and it has not sold, I am with my second realtor now, there is a new post office almost done one block from my house, and a proposed new charter school across from the new post office, grades K thru 8, would appreciate any help, that you, Terry McIntyre,

20 hours ago · US  0 

[Email](#) [Share on Facebook](#) [Tweet](#) [Get email alerts](#) [RSS](#) [Post a Reply](#)

Answers (4)

SORT BY: Newest

**Christopher Phillips, "Christopher Phillips9"** Agent

427 contributions  [Write a review](#)

OCR for Public Record

UGC Fraud Detection

Listing Quality Analytics

Public Record Quality
Analytics

Clickstream Quality Analytics

USE CASE #1 - WHO OWNS THAT HOME?

905 Hargis Way,
Santa Clara, CA 95054

\$1,582,131 | 4 Beds | 3.5 Baths | 2,083 Sq Ft

● OFF MARKET


EDIT FACTS


LIST HOME


MORE

Your Zestimate

[Add owner estimate](#)

Zestimate 

\$1,582,131

+\$6,563 Last 30 days

\$1.50M | \$1.66M

Zestimate range


Rent Zestimate 

\$4,700/mo

-\$48 Last 30 days

\$4.0K | \$5.3K

Zestimate range


Zestimate forecast 

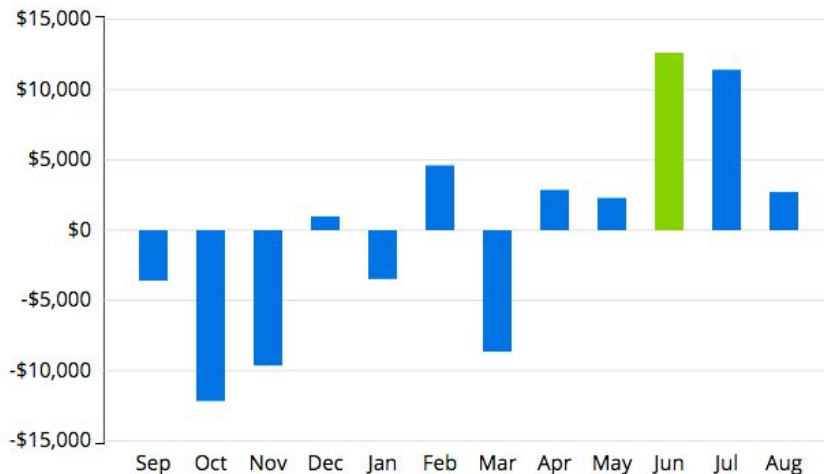
\$1,606,496

 +1.5%
One year

[See Zestimate chart](#)

BEST TIME TO LIST

June has historically been the best month to list a home in your area. We estimate that listing your home in June would increase your sale price by \$16,216 over listing in September. 



WHO OWNS THAT HOME? - MODELS

Ensemble of Multiple Models: Random Forest, Linear Regression

- Property Centric Model
- User-Centric Model

WHO OWNS THAT HOME? - FEATURES

Property Centric Model - 45 features, past 90 days of clickstream events

- # times user triggered edit home fact event
- # of times user arrived at property via sell tab on search from homepage
- # of times user triggered post as for-sale-by-owner event
- # of times user searched for single address for target home
- # of times user expanded price history
- average time on home detail page

Label - Explicit home claims

WHO OWNS THAT HOME? - SYSTEM ENGINEERING

Apache Spark - PySpark / Spark SQL for ETL and Feature Generation

Streaming / Real-Time Data Ingestion

MongoDB / SQS - Low latency serving API

CI / CD - Unit Tests, Regression Tests, Automated Deployment

WHO OWNS THAT HOME? - NEXT STEPS

More signals

New Models

Real-Time Scoring API - predict homeowner in-session

USE CASE #2: PERSONALIZED HOMES



● FOR SALE

\$745,000

3 bd, 4.0 ba, 2,290 sqft

3011 Humes Pl W, Seattle, WA

Reduced from: **\$775,000**

Open: Sat. 1-4pm



● FOR SALE

\$879,000

3 bd, 3.0 ba, 1,720 sqft

215 W Lee St, Seattle, WA

Open: Sat. 1-4pm

Seattle WA Real Estate

148 homes for sale.

[Homes for You](#)

[Newest](#)

[Cheapest](#)

[More](#)



● HOUSE FOR SALE
\$1,599,988 4 bds • 4 ba • 4,869 sqft
41 days on Zillow • 362 Ward St, Seattle, WA



● HOUSE FOR SALE
\$1,335,000 3 bds • 3 ba • 3,050 sqft
• \$64,000 (Jul 9) • 1719 Bigelow Ave N, Seattle, WA



● TOWNHOUSE FOR SALE
\$599,000 2 bds • 2 ba • 1,200 sqft
Open: Sun. 1-4pm • 1910 11th Ave W # B, Seattle, WA



● HOUSE FOR SALE
\$2,350,000 2 bds • 3 ba • 3,140 sqft
Open: Sun. 2-4pm • 128 Highland Dr, Seattle, WA



● CONDO FOR SALE
\$339,000 1 bd • 1 ba • 585 sqft
2 days on Zillow • 275 W Roy St APT 109, Seattle, WA
Savvy Lane Inc



● APARTMENT FOR SALE
\$989,000 4 bds • 2 ba • 2,180 sqft
Open: Sat. 1-4pm • 1906 3rd Ave W, Seattle, WA

PERSONALIZED HOMES - MODELS

Collaborative Filtering Model

Content-Based Filtering Model

- User Profile Model
- Content Model
- Ranking Model

Combination Model

PERSONALIZED HOMES - COLLABORATIVE FILTERING (1 of 3)

Predict personalized homes for sale per user using behavior of similar users

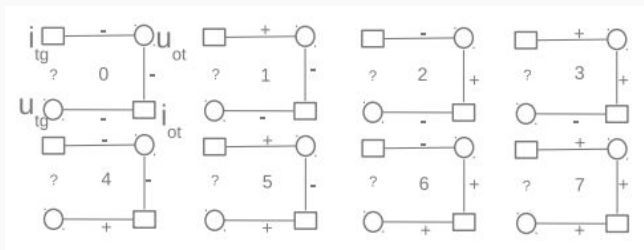
Like vs. Dislike

- Like = user actively engaged with the property
- Dislike = user viewed a property but engagement signals are weak

PERSONALIZED HOMES - COLLABORATIVE FILTERING (2 of 3)

Feature Generation - Wedge Counting Algo

- For properties (i_{ot}) intersecting target user (u_{tg}) and other user (u_{ot}), label 3 edges
- For all u_{tg} and i_{tg} pairs, populate counts for 8 like/dislike configurations (features)



PERSONALIZED HOMES - COLLABORATIVE FILTERING (3 of 3)

Regression Model - ranked list of properties per user

- Train - 8 features as variables, label if if target user likes / dislikes a property (i_{tg})
- Score - all u_{tg} / i_{tg} pairs.
- Filter out popular users and items.

PERSONALIZED HOMES - CONTENT-BASED FILTERING (1 of 3)

User Profile Model - generate set of features with scores per user

- Key features: zipcode, type, bedrooms, bathrooms, square feet, lot size, built year, list price
- Convert numerical values into categorical variables
- Regression model that creates user vector from raw signals

PERSONALIZED HOMES - CONTENT-BASED FILTERING (2 of 3)

Content Model - generate features with scores per property

- Same feature set as user profile model

PERSONALIZED HOMES - CONTENT-BASED FILTERING (3 of 3)

Ranking Model - match homes for sale with user profiles

- element-wise product of user vector and content matrix
- freshness / age decay
 - vector of time on market decay
 - vector of last updated decay

PERSONALIZED HOMES - COMBINATION MODEL

Overcome the problem of

- cold start (collaborative filtering)
- sparsity (content-based filtering)

Combination Model

- Weighted: scores of collaborative filtering results and content-based filtering results are combined numerically

PERSONALIZED HOMES - SYSTEM ENGINEERING

Apache Spark - PySpark, Spark SQL - Model training / scoring with scikit-learn

Streaming Data Ingestion - Clickstream, Public Record, MLS, UGC

API - Python, Flask, Redis

CI / CD - Unit Tests and Regression Tests, Automated Deployment

THANK YOU!



<http://www.zillow.com/data-science/>

We are hiring!

- Big Data Engineer
- Software Dev Engineer, ML
- Data Scientists
- Director, ML Apps