Computer Vision for Homes and Mortgages

Taka Tanaka
Managing Director and Head of Data

My own career path

- Undergrad (2003)
- Master's (2004)
- Jobs in education, finance (2004-2005)
- PhD (2006-2011)
- Postdoc in Germany (2011-2014)
- Research Faculty in US (2014-2016)

- Data science bootcamp (2017)
- Data Scientist => DS Manager at a consulting company (2017-2019)
- DS Team Head at WeightWatchers (2019-2020)
- Head of DS at Radish, a fiction app startup (2020-2021)
- Head of Data at Roc360, real estate fintech (2021-present)

Business Context

- **Roc360 is a financial institution specializing in residential real estate investments.**
- We fund business loans for investors—
  - "fix-and-flip" investors who purchase distressed properties, renovate them, and resell them;
  - buyers who purchase properties as rental investments.
- We source properties and leads for investors.
- We provide title and property insurance for properties.

Data Team

- 8 full-time data scientists, 5 part-time consultants.

- Work includes:
  - **Business Intelligence**—dashboards, reports on risks and trends
  - **Modeling**—e.g. cash flow forecasting, borrower segmentation
  - **Infrastructure**—e.g. automated lead-gen pipeline (search public record, run contact append services, update database)

Data Team


conf
Noah Krever
Viacheslav Sergeev
Farbod Jahandar
Leighann Irving
Dominic Sicilian
Bolu Olayemi
jeremy thomas
You

We use over a dozen data sources, including…

- **Property Listings**
  - Records of property listings, un-listings, and updates by agents
    - Price changes, rent, closing price
    - Property description text
    - Property images
- **Public record of home purchases**
  - County-level registered information of home sales
    - Transaction type, buyer, seller
    - Property type, beds/baths, build year…
    - Scans of documents

Challenges and Opportunities: data quality

Data comes from **many sources** (individual counties, agents, brokers);
is inherently **messy**
(often originating in manual and analog processes, e.g. mortgage documents);
not standardized…

This means that extracting value out of real estate data is high-lift.
Not many companies have fully mature data products.

Challenges and Opportunities: unstructured data

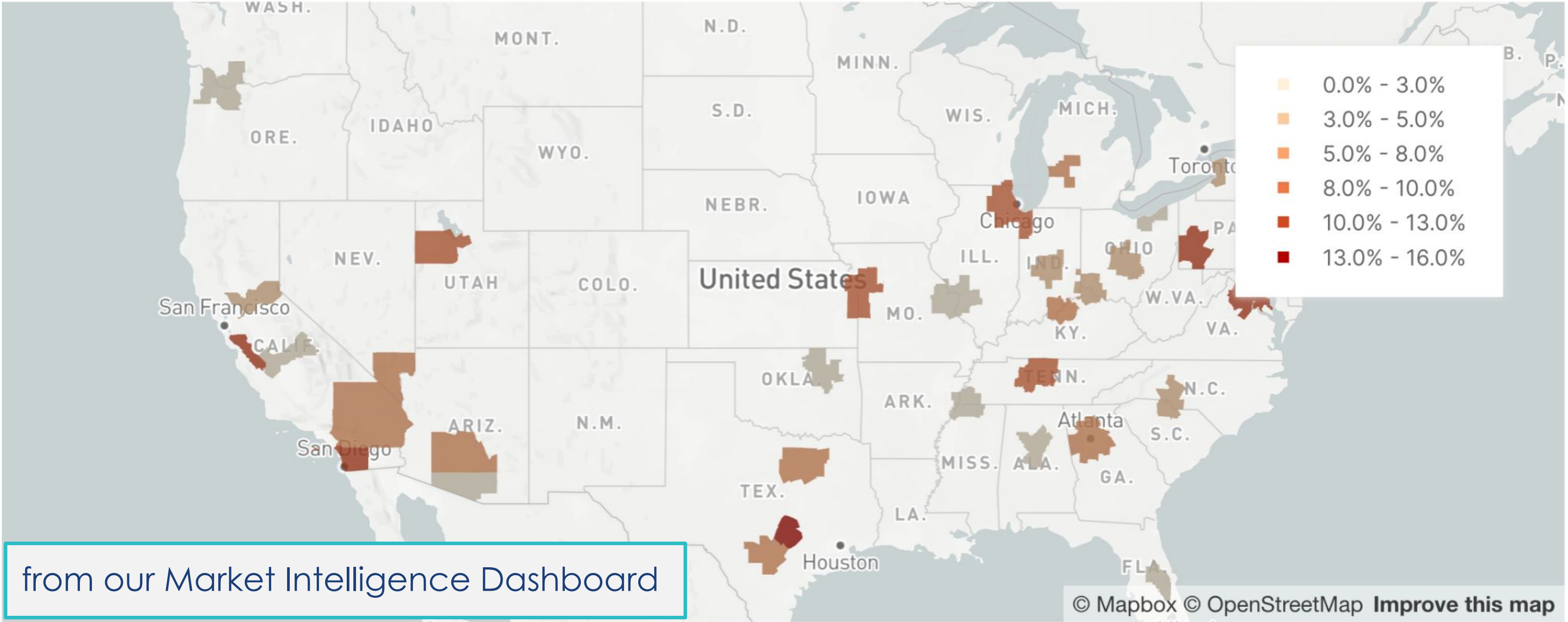
There is tons of value in **unstructured and highly idiosyncratic data**—e.g. scans of mortgage documents and records; property pictures, including of "distressed" properties.

A need and an opportunity to innovate with custom computer vision models, e.g. for parsing scanned documents and classifying images.

First, let me share some of our “standard” data capabilities…

Tracking market trends

% Decrease In Monthly Median Sale Price For SF Fair Market sale from Apr 2022 to Oct 2022

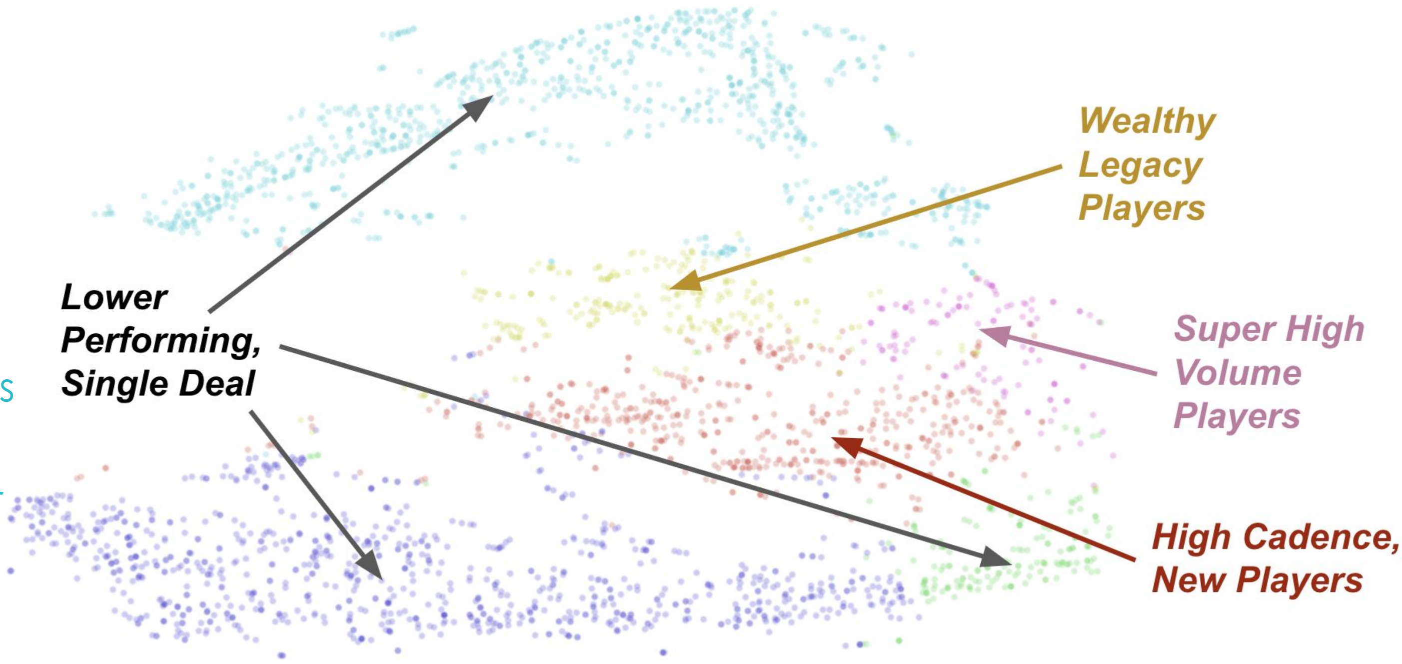


from our Market Intelligence Dashboard

Segmenting fix-and-flip investors

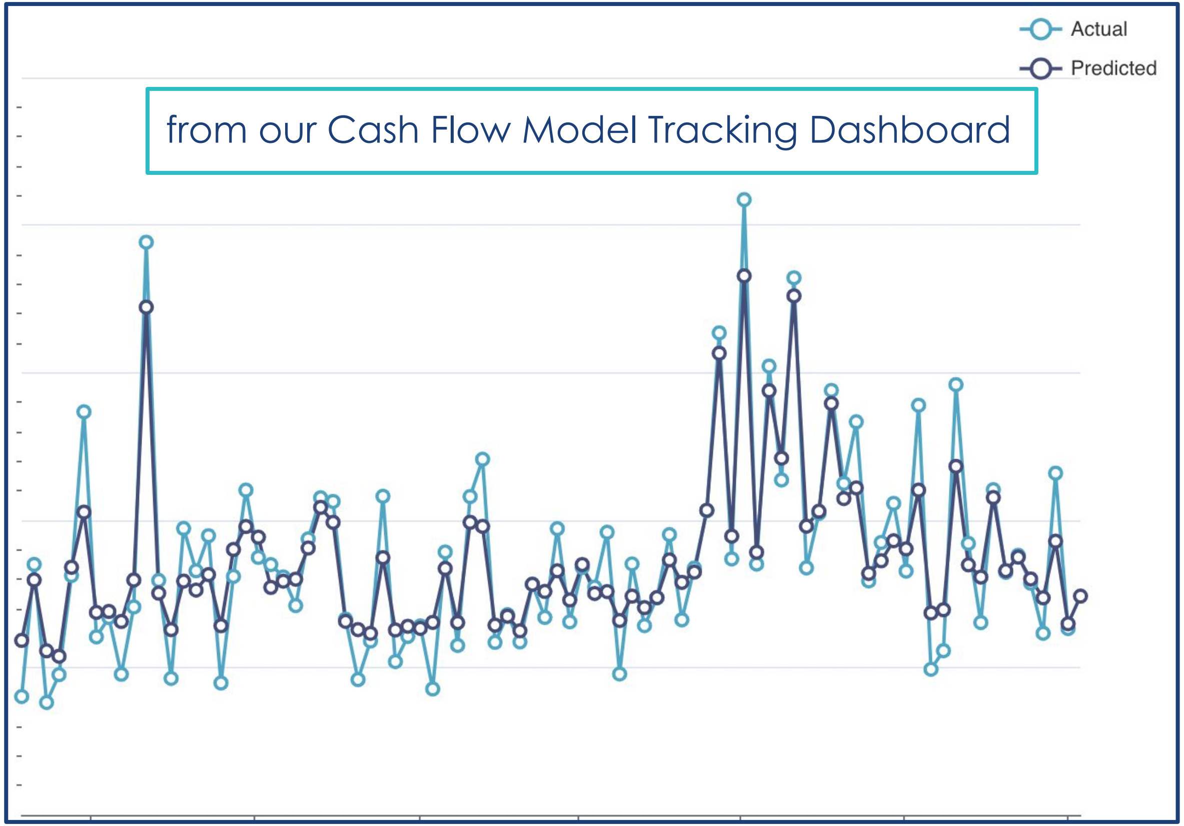
- K-Means clustering for segmentation of our borrower population.
- Feature space consists of many fix-and-flip related business variables.
- Use centroid locations and inter-cluster relationships to better understand our borrowers, and search for new ones.

T-SNE Embeddings of Roc Borrowers, Colored by Cluster



Cash flow prediction model

- Predict how much cash we need on-hand to fund loans in the upcoming weeks.
- Automated schedule:
  - Make predictions.
  - Record on cloud database.
  - Track performance against actual values in stakeholder-facing dashboard.



Now, for some really cool stuff…
(I think)

Documents!

Challenge: Real estate is analog-first

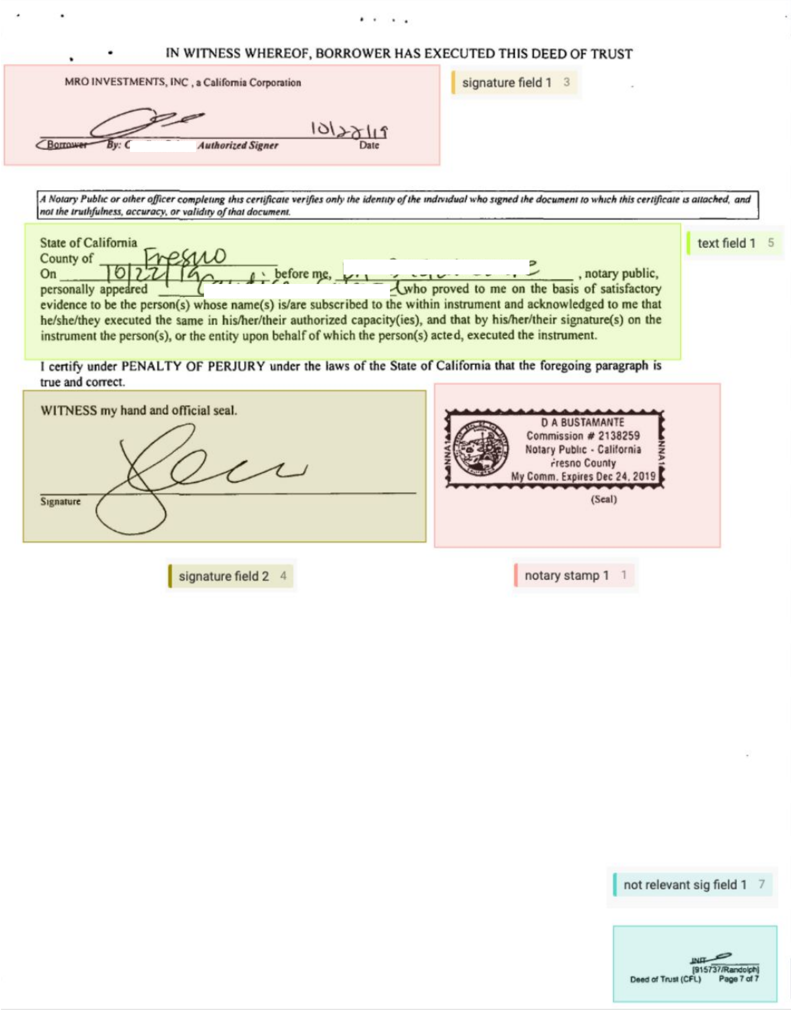
- Sources of truth, and critical pieces of information, are stored in manually produced, printed, and signed documents—e.g. mortgage documents, property appraisals.
- Mortgage documents are public record, maintained by each county in the United States.
- Having an up-to-date record of the persons and corporate entities who signed the documents is useful for understanding transaction trends and identifying the “players” in residential investment real estate.
  (There are ~100,000 individuals who actively flip properties.

Parsing documents at scale

- We leverage cloud compute resources (GCS) and directed acyclic graphs (Prefect) to define and schedule batch jobs which run at scale.
- Automatically scan the public record in cloud data warehouse for transactions of interest.
- Fetch documents for all relevant transactions asynchronously from data vendors (reading/writing to our cloud cache in the process).

Parsing documents at scale: OCR & token extraction

- Optical Character Recognition (OCR) is very resource intensive—seconds per page.
  Documents are 30-40 pages; we need to streamline.
  - Trained an object detection model (YOLOv5) to determine relevant portions of the document.
  - Per non relevant page: 10x speed up in inference time on CPU, 200x on GPU.
- After YOLO, we run OCR (tesseract/google vision) on relevant crops only, and write results to our data lake.
- Run fined tune Hugging Face models (BERT) to extract relevant information from the documents.
- Feed extracted names through downstream pipelines.



Parsing documents at scale
Named Entity Recognition

SUBSCRIBED AND SWORN TO BEFORE ME on the ae Ee wort plumber

by Taka Tanaka on behalf of Roc Capital, LLC, known or proved to me according to law
to be the person whose name is subscribed to the foregoing instrument, and acknowledged to me
that he/she/they
voluntarily executed the same for the purposes of consideration therein expressed, and in the
capacity stated.
ub day of odors . 20 2a).

Given under my hand and seal this

tN ies ig

borrower (1)

Taka Tanaka	0.999

entity (1)

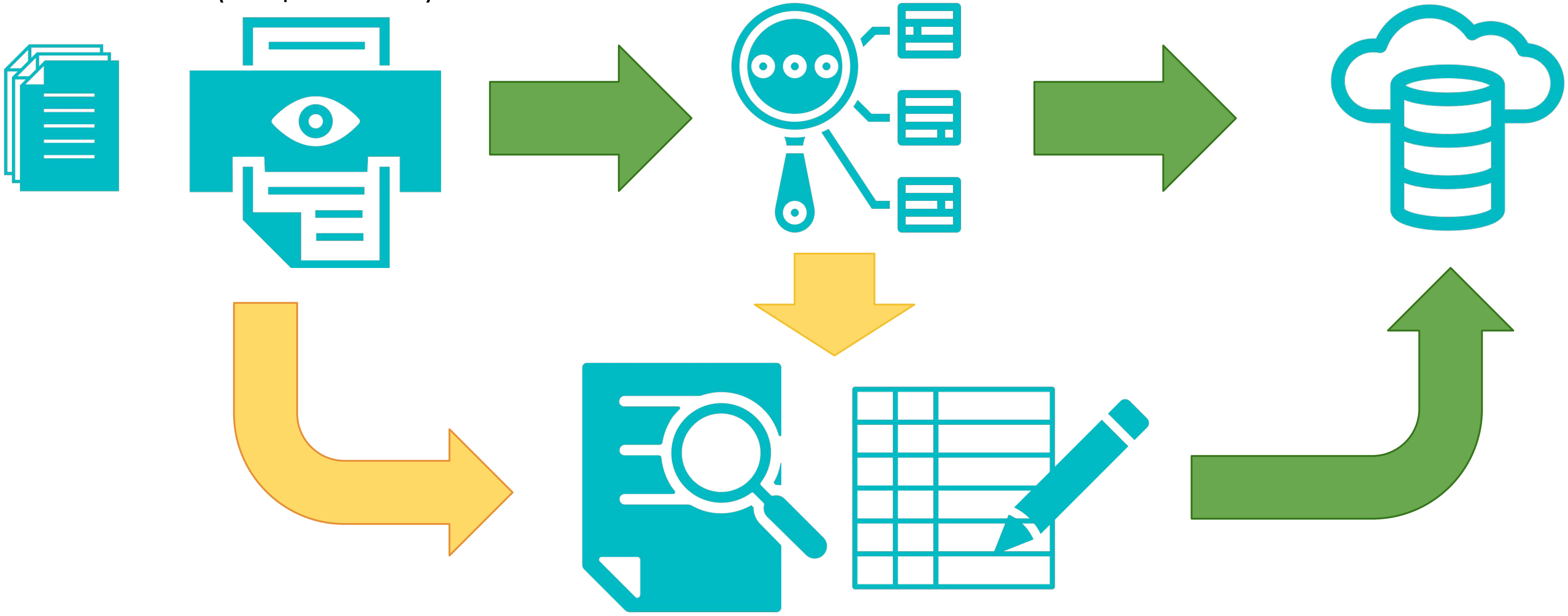
Roc Capital, LLC	0.999

We deploy this as an automated pipeline

Batch-process mortgage documents (crop + OCR)…

…then named entity recognition

Record into database



Failed scans and low-confidence scores are sent to human team for follow-up and data entry.

Another example: property appraisals

Using computer vision to scan tables

INPUT

FEATURE	SUBJECT	COMPARABLE SALE # 1		COMPARABLE SALE # 2		COMPARABLE SALE # 3	
Address	942 Gulf Shore Dr Carrabelle, FL 32322	942 Gulf Shore Dr Carrabelle, FL 32322		367 Gulf Shore Dr Carrabelle, FL 32322		922 Gulf Shore Dr Carrabelle, FL 32322	
Proximity to Subject		0.00 miles		2.75 miles NE		0.10 miles NE	
Sale Price	$		$ 340,000		$ 215,000		$ 265,000
Sale Price/Gross Liv. Area	$ 194.48 sq.ft.	$ 273.31 sq.ft.		$ 297.37 sq.ft.		$ 186.62 sq.ft.	
Data Source(s)		RAFGC Sold Listing #307238;DOM 0		RAFGC Sold Listing #312094;DOM 13		RAFGC Sold Listing #310908;DOM 4	
Verification Source(s)		ORB 1338 Page 546		ORB 1348 Page 248		ORB 1337 Page 2660	
VALUE ADJUSTMENTS	DESCRIPTION	DESCRIPTION	+(-) $ Adjustment	DESCRIPTION	+(-) $ Adjustment	DESCRIPTION	+(-) $ Adjustment
Sales or Financing Concessions		ArmLth Conv;0		ArmLth Cash;0		ArmLth Cash;0	
Date of Sale/Time		s06/22;Unk		s10/22;Unk		s05/22;Unk	
Location	N;Res;Dog Island/Gulf	N;Res;Dog Island/Gulf		N;Res;Dog Island/Gulf		N;Res;Dog Island/Gulf	
Leasehold/Fee Simple	Fee Simple	Fee Simple		Fee Simple		Fee Simple	
Site	24,000 sf	24,000 sf		10,759 sf	+10,639	1.15 ac	-20,966
View	N;Res;CtyStr/Gulf	N;Res;CtyStr/Gulf		N;Res;CtyStr/Gulf		N;Res;CtyStr/Gulf	
Design (Style)	DT1;Coastal/Vyl/Stn	DT1;Coastal/Vyl/Stn		DT1;Coastal/Vyl	+3,500	DT2;Coastal/Wd	-10,000
Quality of Construction	Q4	Q4		Q4		Q4	
Actual Age	26	26		50	+24,000	42	+16,000
Condition	C3	C3		C5	+40,000	C4	+20,000
Above Grade	Total / Bdrms. / Baths	Total / Bdrms. / Baths		Total / Bdrms. / Baths	+20,000	Total / Bdrms. / Baths	+20,000
Room Count	7 / 3 / 2.0	7 / 3 / 2.0		6 / 2 / 1.0	+10,000	6 / 2 / 2.0	0
Gross Living Area	1,244 sq.ft.	1,244 sq.ft.		723 sq.ft.	+52,100	1,420 sq.ft.	-17,600
Basement & Finished Rooms Below Grade	0sf	0sf		0sf		0sf	
Functional Utility	Average	Average		Average		Average	
Heating/Cooling	Central HVAC	Central HVAC		Central HVAC		Central HVAC	
Energy Efficient Items	Average Package	Average Package		Average Package		Average Package	
Garage/Carport	1cp4dw	1cp4dw		4dw	+5,000	4dw	+5,000
Porch/Patio/Deck	Cov Porch/Deck	Cov Porch/Deck		Cov Porch/Deck		Cov Porch/Deck	
Int. Amenity	1 - Fireplace	1 - Fireplace		None	+3,500	None	+3,500
Ext. Amenity	Storage	Storage		Fence	+3,000	None	+5,000
Ext. Amenity	No driveable access	No driveable access		None	-20,000	None	-20,000
Net Adjustment (Total)		☐ + ☐ -	$ 0	☒ + ☐ -	$ 151,739	☒ + ☐ -	$ 934
Adjusted Sale Price of Comparables		Net Adj. 0.0 % Gross Adj. 0.0 %	$ 340,000	Net Adj. 70.6 % Gross Adj. 89.2 %	$ 366,739	Net Adj. 0.4 % Gross Adj. 52.1 %	$ 265,934

Using computer vision to scan tables

OUTPUT

	SUBJECT	COMP #1	COMP #2
Address 942 Gulf Shore Dr Carrabelle, FL 32322	942 Gulf Shore Dr Carra	942 Gulf Shore Dr Carrabelle, FL 32322	367 Gulf Shore Dr Carrabelle, FL 32322
Proximity to Subject	nan	0.00 miles	2.75 miles NE
Sale Price	$	$ 340,000	$ 215,000
Sale Price/Gross Liv. Area	$194.48 sq.ft.	$ 273.31sq.ft.	$ 297.37sq.ft.
Data Source(s)	nan	RAFGC Sold Listing #307238;DOM 0	RAFGC Sold Listing #312094;DOM 13
Verification Source(s)	nan	ORB 1338 Page 546	ORB 1348 Page 248
Date of Sale/Time	nan	{'DESCRIPTION': 's06/22;Unk', 'ADJUSTMENT': nan}	{'DESCRIPTION': 's10/22;Unk', 'ADJUSTMENT': nan}
Location	N;Res;Dog Island/Gulf	{'DESCRIPTION': 'N;Res;Dog Island/Gulf', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'N;Res;Dog Island/Gulf', 'ADJUSTMENT': nan}
Leasehold/Fee Simple	Fee Simple	{'DESCRIPTION': 'Fee Simple', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'Fee Simple', 'ADJUSTMENT': nan}
Site	24,000 sf	{'DESCRIPTION': '24,000 sf', 'ADJUSTMENT': nan}	{'DESCRIPTION': '10,759 sf', 'ADJUSTMENT': '+10,639'}
View	N;Res;CtyStr/Gulf	{'DESCRIPTION': 'N;Res;CtyStr/Gulf', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'N;Res;CtyStr/Gulf', 'ADJUSTMENT': nan}
Design (Style)	DT1;Coastal/Vyl/Stn	{'DESCRIPTION': 'DT1;Coastal/Vyl/Stn', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'DT1;Coastal/Vyl', 'ADJUSTMENT': '+3,500'}
Quality of Construction	Q4	{'DESCRIPTION': 'Q4', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'Q4', 'ADJUSTMENT': nan}
Actual Age	26	{'DESCRIPTION': '26', 'ADJUSTMENT': nan}	{'DESCRIPTION': '50', 'ADJUSTMENT': '+24,000'}
Condition	C3	{'DESCRIPTION': 'C3', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'C5', 'ADJUSTMENT': '+40,000'}
Functional Utility	Average	{'DESCRIPTION': 'Average', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'Average', 'ADJUSTMENT': nan}
Heating/Cooling	Central HVAC	{'DESCRIPTION': 'Central HVAC', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'Central HVAC', 'ADJUSTMENT': nan}
Energy Efficient Items	Average Package	{'DESCRIPTION': 'Average Package', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'Average Package', 'ADJUSTMENT': nan}
Garage/Carport	1cp4dw	{'DESCRIPTION': '1cp4dw', 'ADJUSTMENT': nan}	{'DESCRIPTION': '4dw', 'ADJUSTMENT': '+5,000'}
Porch/Patio/Deck	Cov Porch/Deck	{'DESCRIPTION': 'Cov Porch/Deck', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'Cov Porch/Deck', 'ADJUSTMENT': nan}
Int. Amenity	1 - Fireplace	{'DESCRIPTION': '1 - Fireplace', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'None', 'ADJUSTMENT': '+3,500'}
Ext. Amenity	Storage	{'DESCRIPTION': 'Storage', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'Fence', 'ADJUSTMENT': '+3,000'}
I	nan	{'DESCRIPTION': 'did', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'did not research the sale or transfer history of
subject property and a 36 month sales /transfer	nan	{'DESCRIPTION': nan, 'ADJUSTMENT': nan}	{'DESCRIPTION': nan, 'ADJUSTMENT': nan}
Sales or Financing	nan	{'DESCRIPTION': 'ArmLth', 'ADJUSTMENT': nan}	{'DESCRIPTION': 'ArmLth', 'ADJUSTMENT': nan}


```
"COMP #1" : {
  "Address 942 Gulf Shore Dr Carrabelle, FL 32322" :
  "942 Gulf Shore Dr Carrabelle, FL 32322"
  "Proximity to Subject" : "0.00 miles"
  "Sale Price" : "$ 340,000"
  "Sale Price/Gross Liv. Area" : "$ 273.31sq.ft."
  "Data Source(s)" : "RAFGC Sold Listing #307238;DOM 0"
  "Verification Source(s)" : "ORB 1338 Page 546"
  "Date of Sale/Time" : {
    "DESCRIPTION" : "s06/22;Unk"
    "ADJUSTMENT" : NULL
  }
  "Location" : {
    "DESCRIPTION" : "N;Res;Dog Island/Gulf"
    "ADJUSTMENT" : NULL
  }
  "Leasehold/Fee Simple" : {
    "DESCRIPTION" : "Fee Simple"
    "ADJUSTMENT" : NULL
  }
  "Site" : {
    "DESCRIPTION" : "24,000 sf"
    "ADJUSTMENT" : NULL
  }
  "View" : {
    "DESCRIPTION" : "N;Res;CtyStr/Gulf"
    "ADJUSTMENT" : NULL
  }
  "Design (Style)" : {
    "DESCRIPTION" : "DT1;Coastal/Vyl/Stn"
```

Images!

Challenge: images are hard!

- May be biased by who takes the images, and in what context (e.g. listing vs. appraisal, luxury condo vs. fixer-upper)
  - Different lenses, lighting; physical staging; digital staging.

- The same room or amenity may have different connotations in different contexts and locations—e.g. in the Upper West Side of Manhattan vs. rural Minnesota.

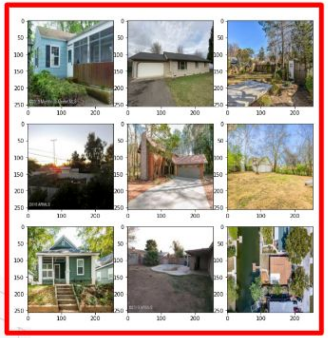
Auto-clustering room types

[inception_v3; ImageNet weights]

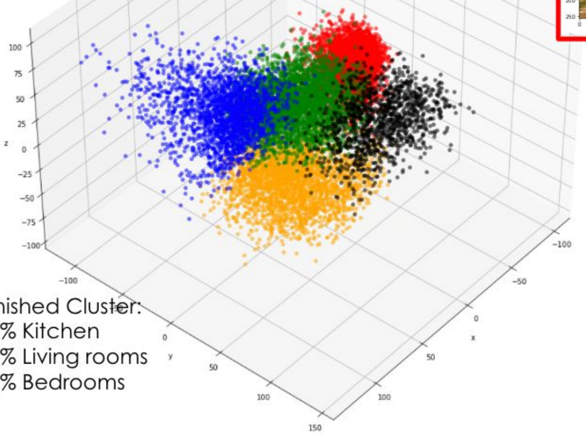
K-means clustering for MLS photos

Mixed cluster of basements + poor condition houses

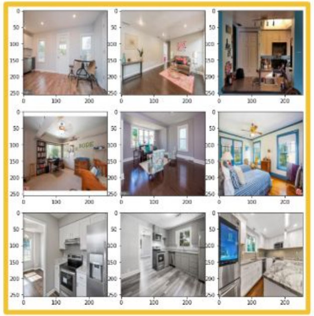
Empty Rooms cluster
Accuracy: 93%



Outdoor cluster
Accuracy: 99%



Bathroom cluster
Accuracy: 93%



Furnished Cluster:
~70% Kitchen
~20% Living rooms
~10% Bedrooms

Predicting property condition

Input image

Google Cloud Storage

Direct reading of encoded images

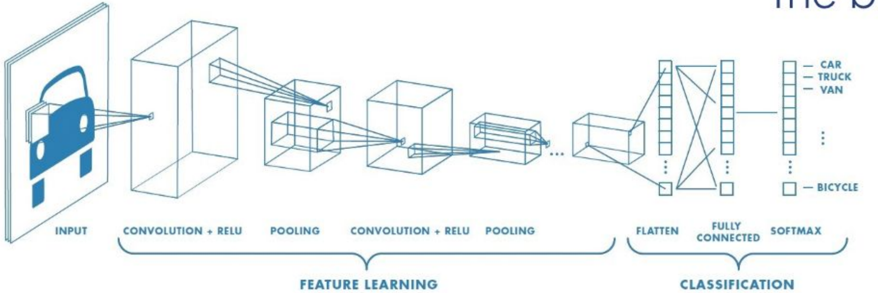
vertex.ai

Predicted Condition: Good

Probability Score: 0.99

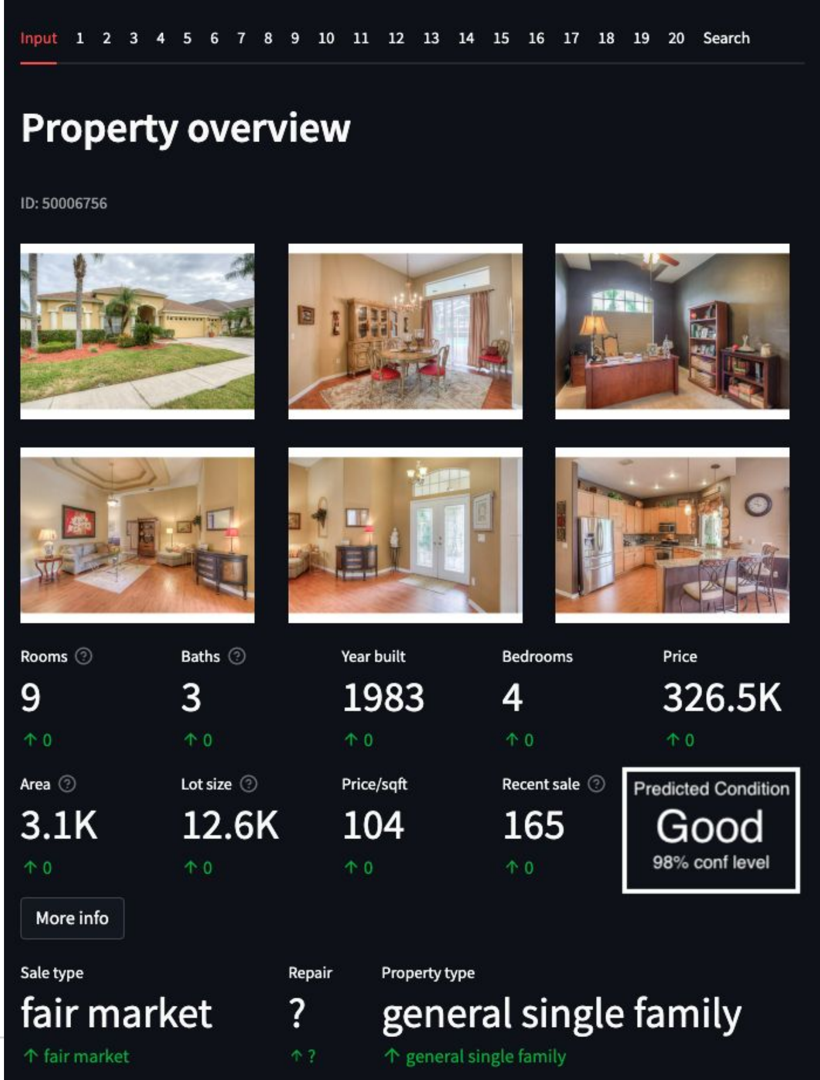
Image Name: filename.jpg

The best trained vision model



Deploy as an app

- Enter an input property.
- Predict comparable properties by weighing available features from available data.
- For the input and output properties, provide images and the algorithmically inferred conditions.



For all algorithms…

- Successful parsing = victory at scale…

- … but faults and mistakes can be disastrous.
  (Recall the classic example of "wolf or dog?" algorithm.)

- We meet often to discuss sub-segments of algorithmic tasks, and perform human sanity checks both within the team and with business users.

Summary

- Real estate is an analog and manual industry.
- There are many third-party sources of critical truth—getting full value out of the data is more difficult than in industries that are digital-native and/or driven by first-party data.
- We are building custom computer vision and NLP algorithms for parsing public-record documents (listings and county-level transaction data).
- We are also building image classification algorithms to predict property conditions.
- As with any business-centered science, it is important to deliver solutions in robust, automated, intuitive vehicles.

Thank you! Questions?

"Astronomy to Data Science" resource:

https://github.com/taka-tanaka/astronomy-to-data-science

linkedin.com/in/takatanaka

Twitter: @astrobassball

Mastodon: astrobassball@mastodon.social