ECONOMIC STATISTICS CENTRE OF EXCELLENCE

A COLLABORATION WITH Office for National Statistics

escoe.ac.uk @ESCoEorg

Towards Accountability in Machine Learning Applications: A System Testing Approach

Thies Lindenthal (University of Cambridge)

Starting at 11.30 AM

ESCoE ECONOMIC MEASUREMENT WEBINARS

WALKING DOWN A STREET...

So much to learn from looking at buildings, spaces, people...

RESEARCH AGENDA

Joint projects with Erik Johnson, Carolin Schmidt & Wayne Wan

1. Can we use street-level images to extract information about buildings?
2. What do the models we trained really "see"?
3. What is it that people pay attention to when looking at houses?

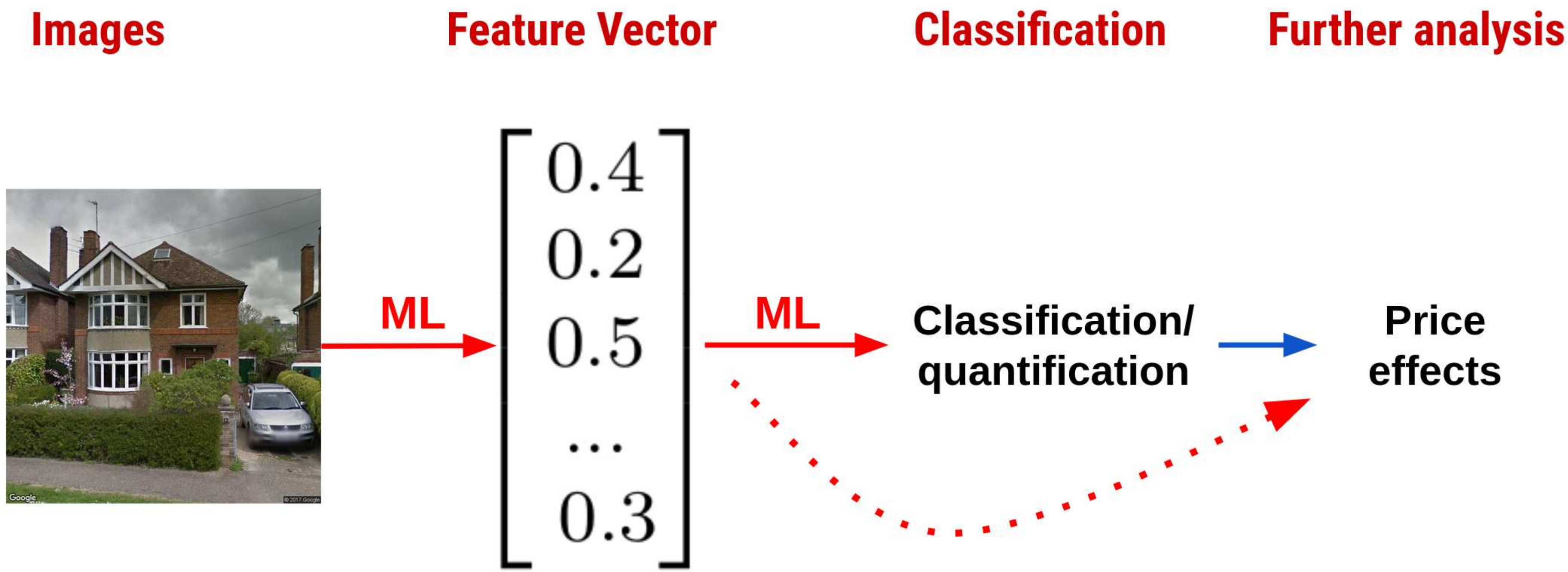

TEACH COMPUTERS TO “SEE”
Infer quality or style attributes from street-level images.

VINTAGE

“Machine Learning, Architectural Styles and Property Values”, with Erik Johnson

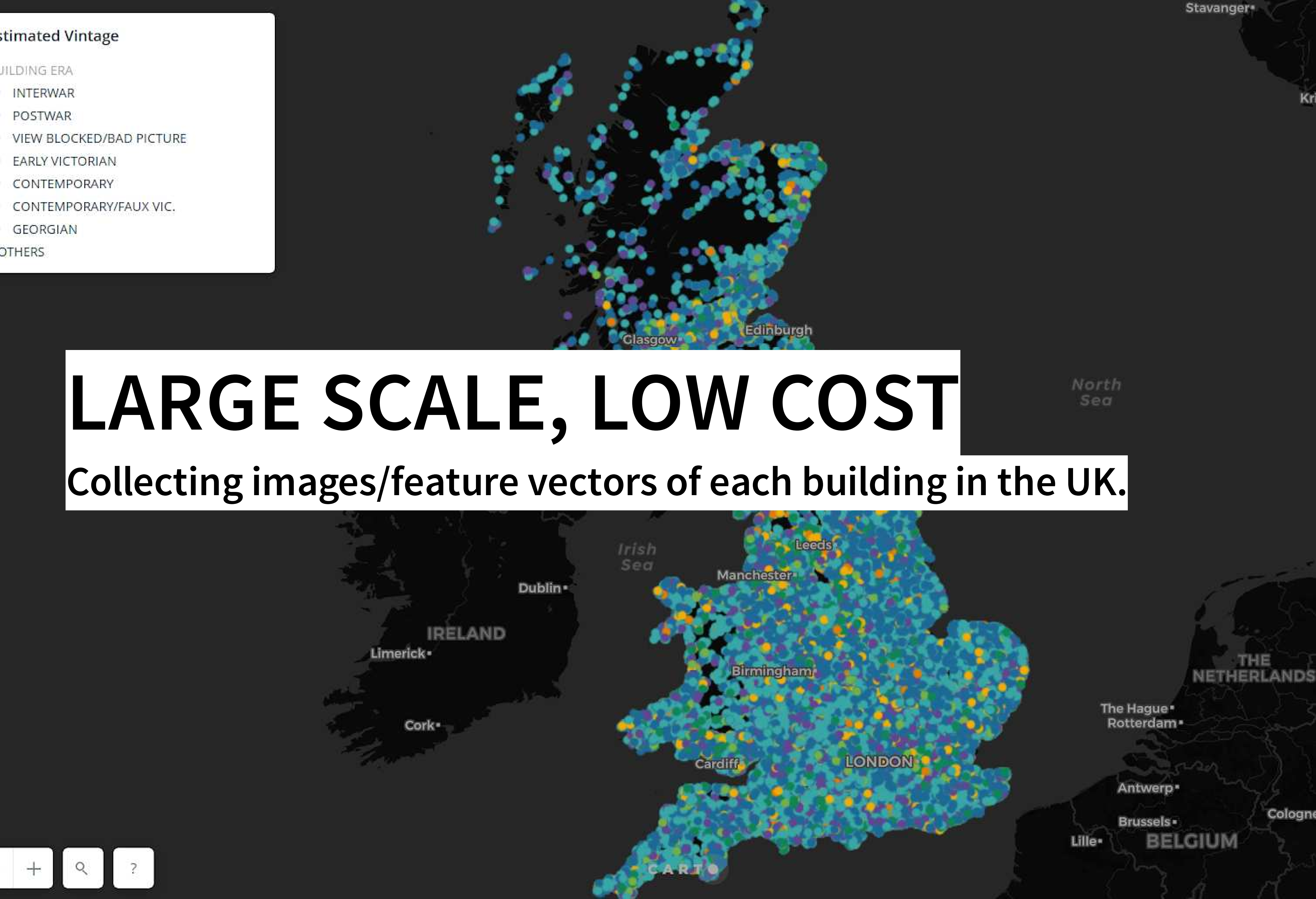
OUR TOOLBOX

Computer vision + ML classification + trad. econometrics



LARGE SCALE, LOW COST

Collecting images/feature vectors of each building in the UK.



KEY FINDING (besides proof of concept)

- No price premium for NEW buildings in traditional styles detectable.


ELECTRIC CAR!


SOLAR PANELS!


LOFT CONVERSION!

ACCOUNTABILITY GAP

Joint work with Wayne Wan

- If all you need is predictive power then go ahead, increase training data, tweak models…
- Can we interpret the predicted values? Communicate what is causing an outcome?
  - Automatic valuations / property taxes
  - Causal inference? At least a little bit?
- How to ensure we are not breaking any laws (and not unethical in the first place)?
  - Discrimination based on protected characteristics is plain and simple illegal
  - Mortgage applications, tenant screening, valuations

AMAZON FAILED (I)

If even a tech giant struggles, caution might be warranted...

https://www.bbc.co.uk/news/technology-45809919

BBC | Sign in | Home | News | Sport | Weather | iPlayer | Sounds | Cl

NEWS

Home | Brexit | Coronavirus | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Technology

Amazon scrapped 'sexist AI' tool

10 October 2018

AMAZON FAILED (II)

Racist face recognition systems...

https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28

ACLU

About Issues Our work News Take action Shop **Donate**

Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots

By Jacob Snow, Technology & Civil Liberties Attorney, ACLU of Northern California
JULY 26, 2018 | 8:00 AM

TAGS: Face Recognition Technology, Surveillance Technologies, Privacy & Technology

Amazon's face surveillance technology is the target of growing opposition nationwide, and today, there are 28 more causes for concern. In a test the ACLU recently conducted of the facial recognition tool, called "Rekognition," the software incorrectly matched 28 members of Congress, identifying them as other people who have been arrested for a crime.

The members of Congress who were falsely matched with the mugshot

MICROSOFT FAILED

Racist chatbot...

https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

MICROSOFT \ WEB \ TL;DR

68

Twitter taught Microsoft's AI chatbot to be a racist a****e in less than a day

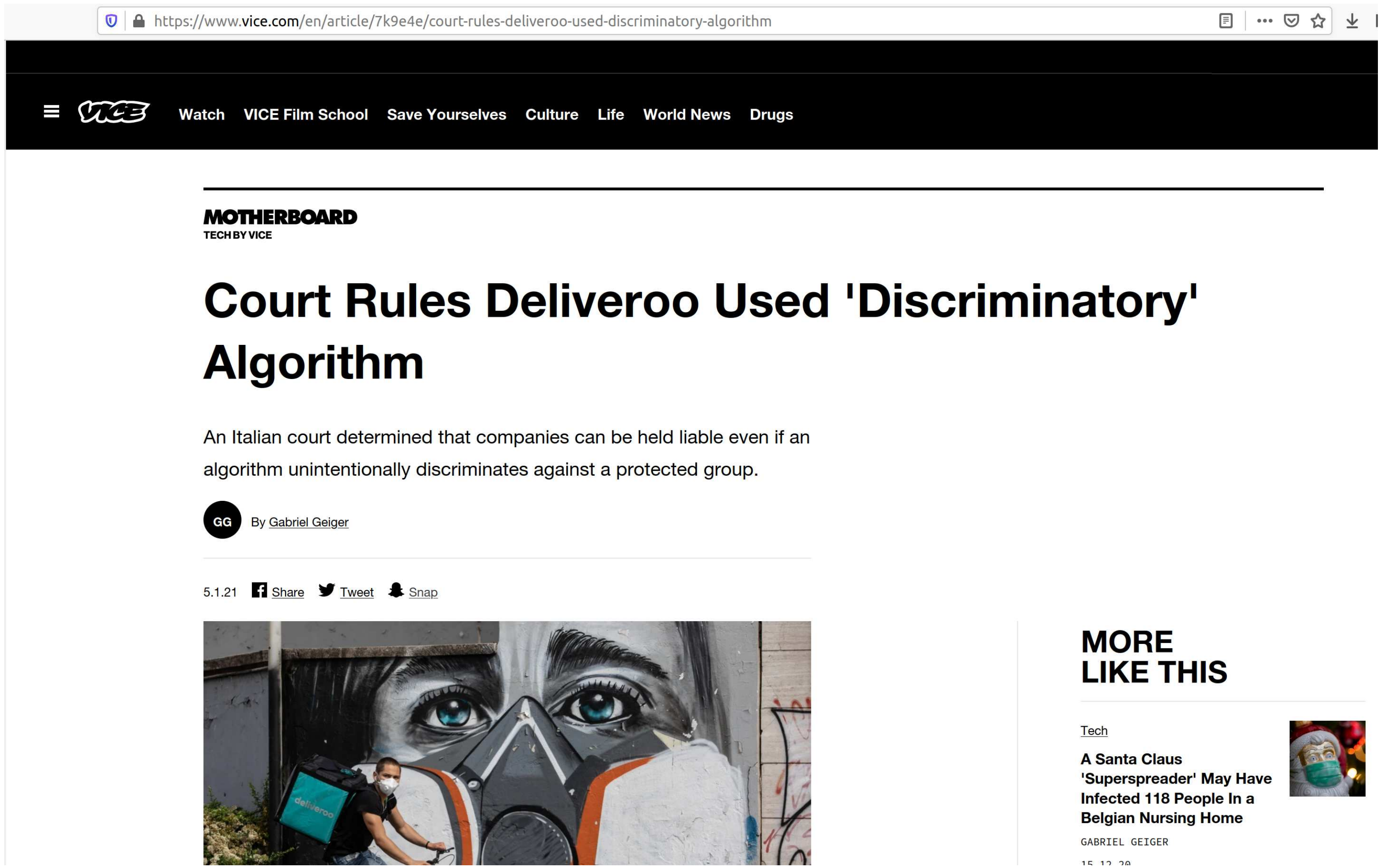
By James Vincent | Mar 24, 2016, 6:43am EDT

Via The Guardian | Source TayandYou (Twitter)

SHARE

DELIVEROO FAILED

Companies are liable for biased ML systems.



https://www.vice.com/en/article/7k9e4e/court-rules-deliveroo-used-discriminatory-algorithm

Watch VICE Film School Save Yourselves Culture Life World News Drugs

MOTHERBOARD
TECH BY VICE

Court Rules Deliveroo Used 'Discriminatory' Algorithm

An Italian court determined that companies can be held liable even if an algorithm unintentionally discriminates against a protected group.

GG By Gabriel Geiger

5.1.21 Share Tweet Snap

MORE LIKE THIS

Tech

A Santa Claus 'Superspreader' May Have Infected 118 People In a Belgian Nursing Home

GABRIEL GEIGER

PREDICTIVE POWER NOT GOOD ENOUGH!

Transparency and due diligence needed

- *System testing* is a concept from software engineering:
  - While developing a system, engineers define tests that check whether the outcome remains in pre-defined range.
- Training an ML system is software development.
  - Follow best practices, define and implement system tests.
  - Such tests should be independent of the training process.
  - Tests are customised to task at hand. We give two examples.

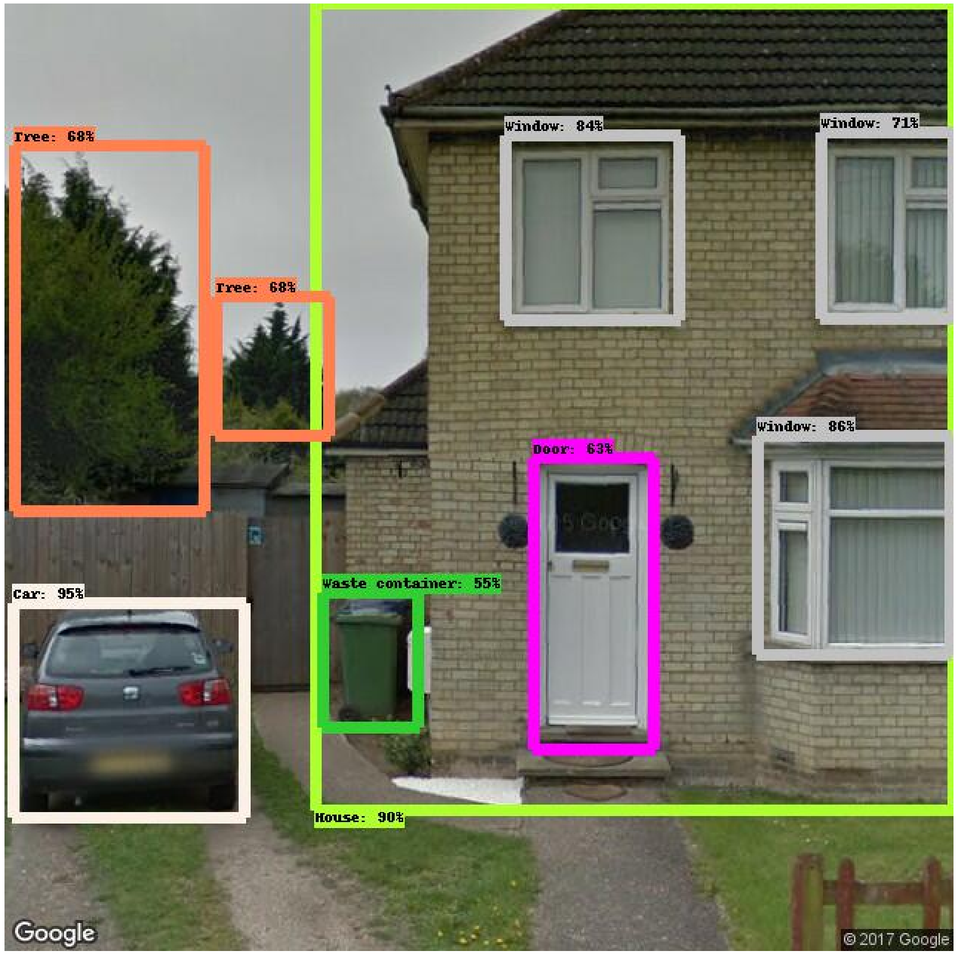
VINTAGE CLASSIFICATION TEST

Which aspects lead to a classification?

- Architects told us: Focus on windows, doors, rooflines, ratios, brickwork
- Test definition: "Good" model should emphasise informative aspects and ignore background, trees, cars, people.
- Test should be implementable on large sample – fully automated.

RELEVANT OBJECTS

Off-the-shelf object detection reveals areas that *should* matter...



RELEVANT PIXELS

LIME algorithm detects areas that matter for classification.

- Local interpretable model-agnostic explanation algorithm (LIME) identifies "super pixels" (Ribeiro et al., 2016).

Interwar: 0.4119

RELEVANT PIXELS

Super pixels depend on classification

- For competing (incorrect) classifications, different sets of super pixels are detected.

Postwar: 0.2112

Edwardian: 0.0666

RIGHT EMPHASIS?

Kind of: focus on doors, windows – but also cars.

- Score of 1 represents a proportional representation. Low score for trees is good!

	Architect's Classification
	(1) All
House	1.103
Window	1.336
Door	1.399
Tree	0.787
Car	1.167

Notes: In Panel A, the *verification test score* presents the proportion of the interpretable area (super-pixels) that overlap with objects detected in the image (e.g., house or window), by vintage. Standard errors are reported in parentheses. In Panel B, the *verification test ratio* normalizes the verification test scores by dividing by the share each object takes up of the entire image. A ratio larger than 1 means that the ML model uses relatively much information from the object type to classify building styles, a score below 1 indicates a lack of emphasis. The ratios for the facade, windows, and doors are larger than 1 overall, lower than 1 for trees, and mixed for cars.

RIGHT EMPHASIS?

Kind of: focus on doors, windows – but also cars.

- Overall, a consistent pattern across styles. But there is a strange emphasis on cars for Georgian homes.

	Architect's Classification							
	(1) All	(2) Georgian	(3) Victorian	(4) Edwardian	(5) Interwar	(6) Postwar	(7) Contemp.	(8) Revival
House	1.103	1.114	1.068	1.127	1.140	1.080	1.061	1.176
Window	1.336	1.493	1.024	1.581	1.524	1.414	1.081	1.441
Door	1.399	1.765	1.515	1.406	1.062	1.579	1.211	1.304
Tree	0.787	0.774	0.769	0.701	0.782	0.764	0.985	0.760
Car	1.167	2.065	1.359	1.181	1.126	1.248	0.968	0.992

Notes: In Panel A, the *verification test score* presents the proportion of the interpretable area (super-pixels) that overlap with objects detected in the image (e.g., house or window), by vintage. Standard errors are reported in parentheses. In Panel B, the *verification test ratio* normalizes the verification test scores by dividing by the share each object takes up of the entire image. A ratio larger than 1 means that the ML model uses relatively much information from the object type to classify building styles, a score below 1 indicates a lack of emphasis. The ratios for the facade, windows, and doors are larger than 1 overall, lower than 1 for trees, and mixed for cars.

IS FOCUS GOOD?

Yes, *correct* classifications emphasise doors, windows *more*...

- ... but not trees, cars.

	Y: Verification Test Score		
	Correct Classifications	Incorrect Classifications	Difference
	(1)	(2)	(3)
House	0.8186	0.7624	0.0562***
	(0.0052)	(0.0093)	(0.0107)
Window	0.1984	0.1640	0.0344***
	(0.0042)	(0.0064)	(0.0077)
Door	0.0392	0.0271	0.0121***
	(0.0024)	(0.0029)	(0.0038)
Tree	0.0853	0.1095	-0.0242***
	(0.0042)	(0.0073)	(0.0084)
Car	0.0485	0.0609	-0.0123**
	(0.0028)	(0.0050)	(0.0057)

Notes: Column (1) reports the model verification score for the correctly classified sampled. Column (2) reports the model verification score for the incorrectly classified sampled. Standard errors are reported in parentheses.
*** $p<0.01$, ** $p<0.05$, * $p<0.1$

NEW MODEL, NEW TEST

Automatic valuation: Shift in focus?

- Wealth confounding factor for car brands and home values? Externalities?

ML IN AVM

Automatic valuation, incorporating image data.

- $log(Price_{it}) = \alpha_0 + X_{it}^{'}\beta + \varphi_t + \omega_i + \epsilon_{it}$
- Classify residuals based on images (out of sample), then re-estimate:
- $log(Price_{it}) = \alpha_0 + X_{it}^{'}\beta + PredResidual_i\gamma + \varphi_t + \omega_i + \epsilon_{it}$

Models	(1) RMSE	(2) MAE
Model 1 (without Architectural Style)	0.2229	0.1613
Model 1 (without Architectural Style) + Predicted Residuals	0.2078	0.1492
Model 2 (with Architectural Style)	0.2169	0.1562
Model 2 (with Architectural Style) + Predicted Residuals	0.2018	0.1445

BLACK-BOX BEHAVIOUR

Automatic valuation: Shift in focus?

- More weight on cars! If vintage is explicitly controlled for (3) then loading on windows and doors decreases.

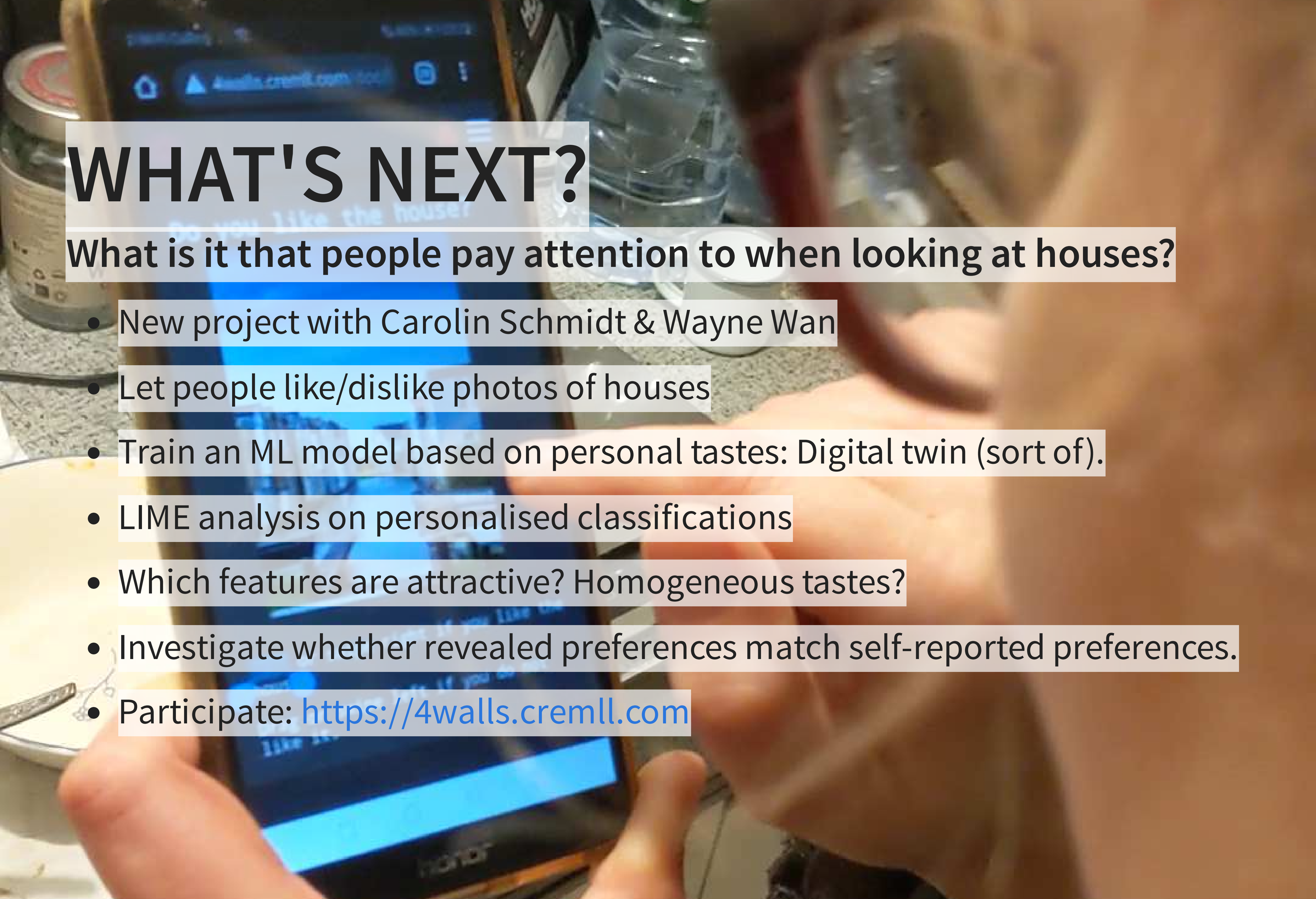
	(1)	(2)	(3)	(4)	(5)	(6)
	Y: Verification Test Ratio					
	Model 1: Without Style		Model 2: With Style		t-test	
	Mean	Std. Dev.	Mean	Std. Dev.	Diff (1)-(3)	Std. Err.
House	1.0664	0.2857	1.0746	0.2784	-0.0082	0.0103
Window	1.3615	1.3884	1.0153	1.1808	0.3465***	0.0502
Door	1.4063	2.0807	1.1242	1.9057	0.2821**	0.1168
Tree	0.8057	1.0469	0.8066	1.5200	-0.0009	0.0622
Car	1.6267	1.8177	1.6325	1.6652	-0.0058	0.1083

Notes: Model 1 refers to the hedonic price model without controls for architectural styles. Model 2 refers to the hedonic price model with controls for architectural styles. In Columns (1) and (3), the *verification test ratio* equals the verification test score over the benchmark score, and the benchmark score is the ratio of the object size to the image size. If the verification test ratio is larger than one, it means that the ML model intentionally uses information from the object (e.g., window or door) to classify price residuals, and vice versa. A positive difference in Column (5) means that Model 1 uses more information of the object to predict the price residuals than Model 2 does, and vice versa. *** p<0.01, ** p<0.05, * p<0.1

WHAT DID WE LEARN?

Better understanding of inner workings of ML black boxes.

- Far from perfect, still, but models behave similarly to human experts.
- Findings are helpful for improving the classifications.
  - Windows and doors should be visibile on images. No trees!
  - Some segments appear to be problematic: Georgian/cars.
  - We would not have known that without testing. Now we can re-train models – or be aware of limitations.



WHAT'S NEXT?

What is it that people pay attention to when looking at houses?

- New project with Carolin Schmidt & Wayne Wan
- Let people like/dislike photos of houses
- Train an ML model based on personal tastes: Digital twin (sort of).
- LIME analysis on personalised classifications
- Which features are attractive? Homogeneous tastes?
- Investigate whether revealed preferences match self-reported preferences.
- Participate: https://4walls.cremll.com

LET'S TALK?

I WOULD LOVE TO HEAR FROM YOU!

E-mail: htl24@cam.ac.uk
Twitter: @ThiesLindenthal