Questions of accuracy and fairness in radicalisation research

what should we do about terrorism online?

'*Data Science in Action*' Lecture

Dr. Margeret Hall

26 October 2023

Problems of a barrier-free internet

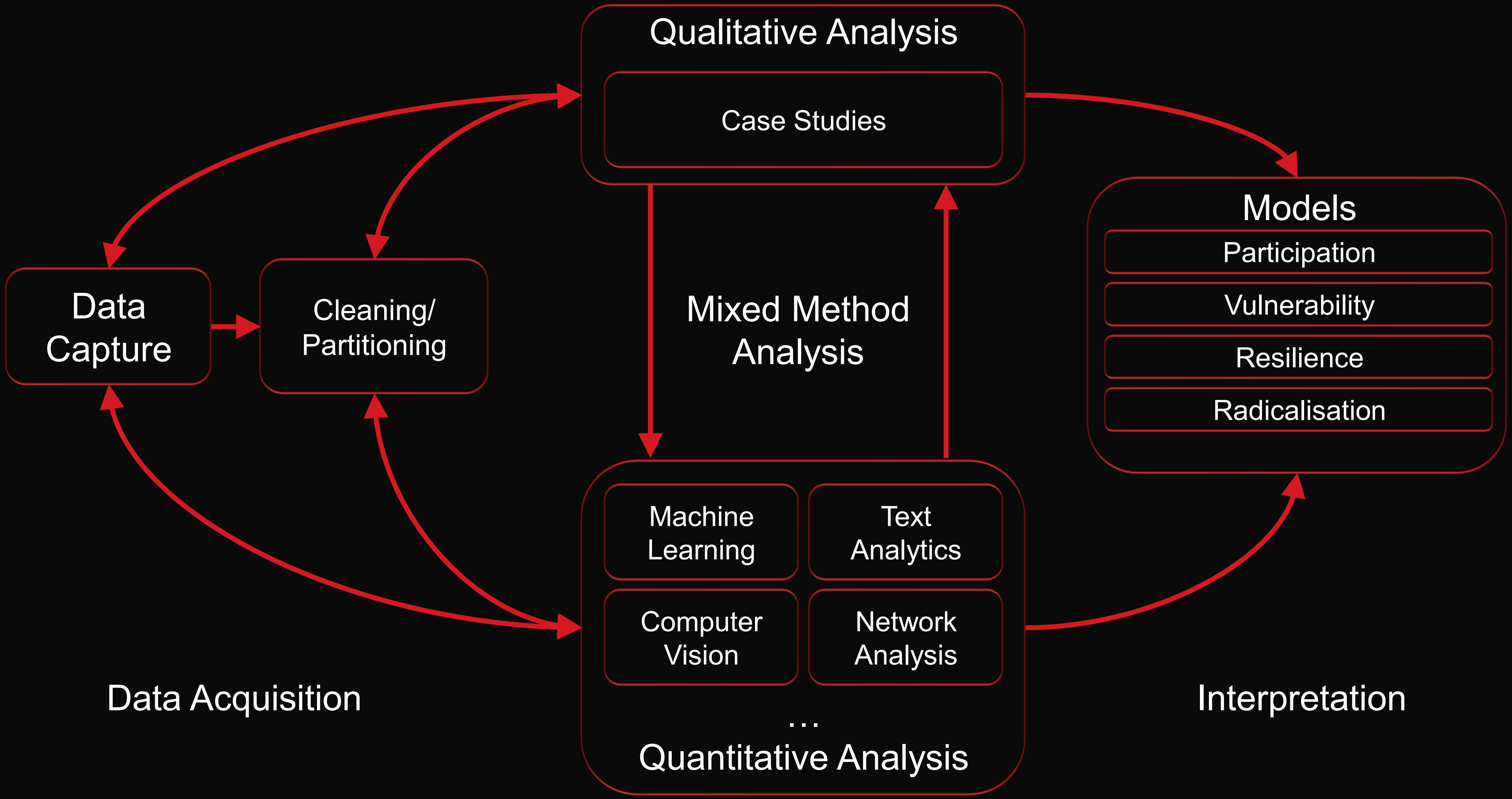
→ The internet has reduced barriers to communication for everyone.

→ Deviant groups are empowered to advertise and recruit online.

→ Black-box algorithmic detection regimes can inadvertently spread bias against similar but unrelated peoples.

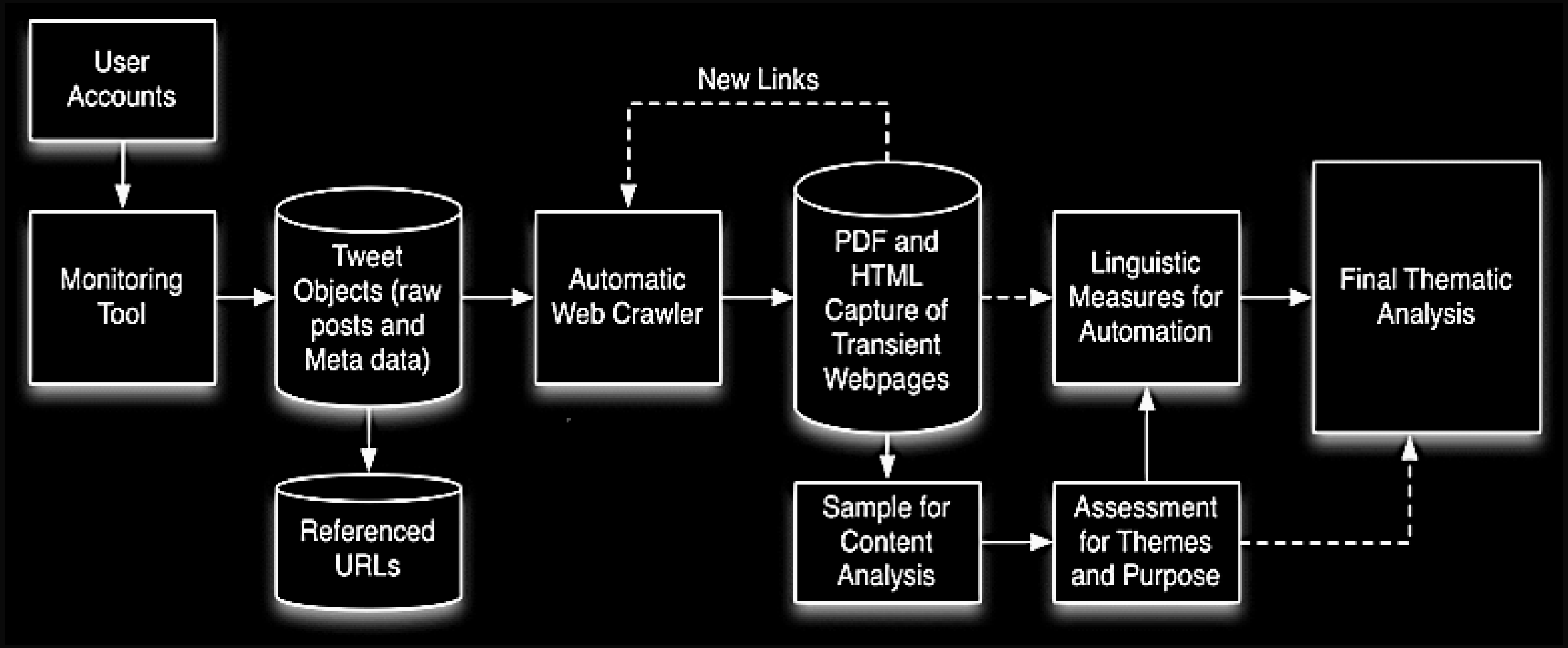
→ Unbiased, reliably coded data is particularly important to prevent biases in automated decisions, e.g., image classification

Hashemi & Hall, 2018; Hall, Logan, Ligon, & Derrick, 2019; Hashemi & Hall, 2019; Hall & Haas, 2021 (*Best Paper Award*)



Müller et al, 2016; Hall et al, 2018

Harvesting Cross-media Data



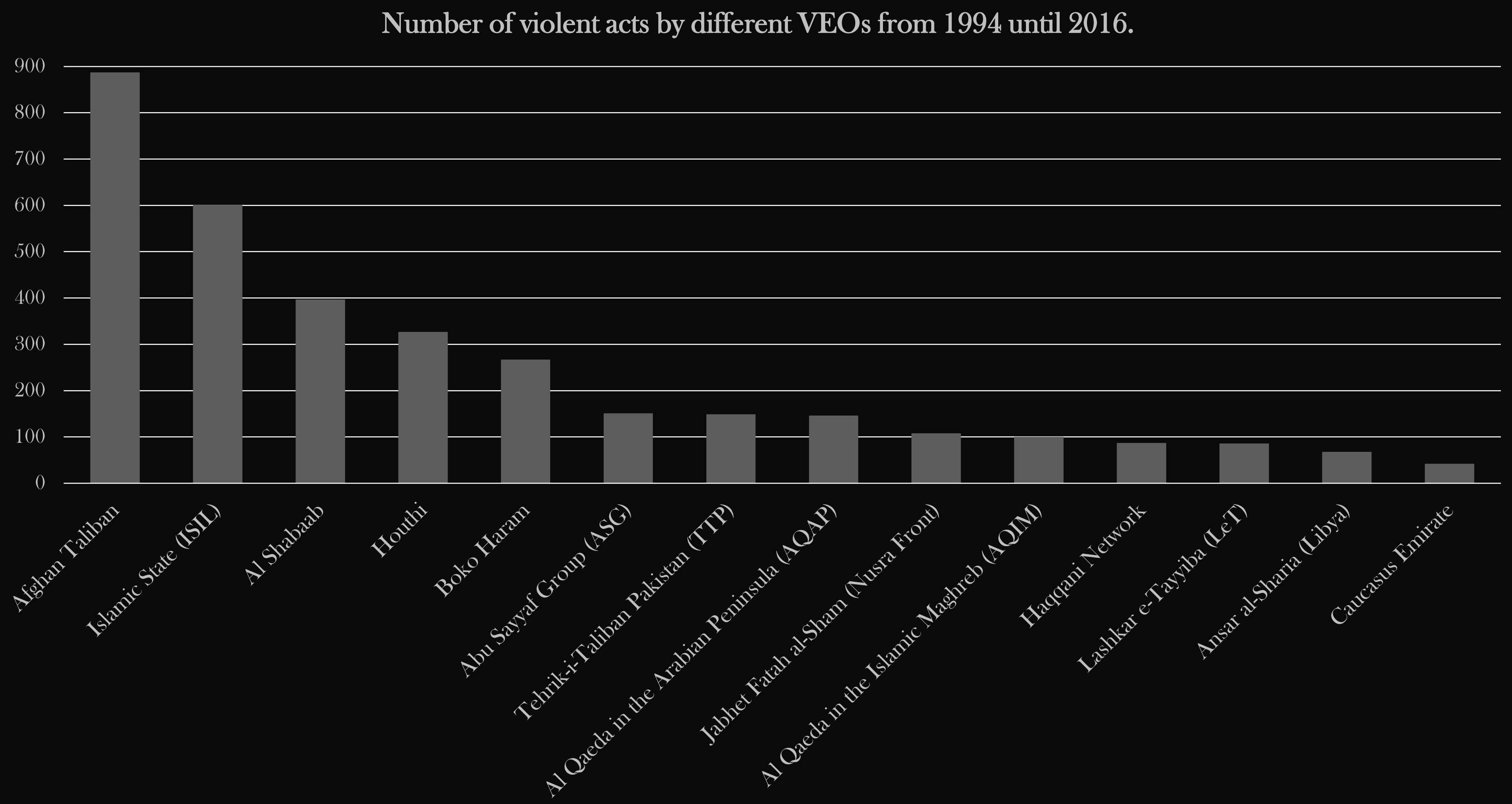
In 2018: Tweets – 8,181,302

Web pages – 48,297

Images - ~1.5 mil

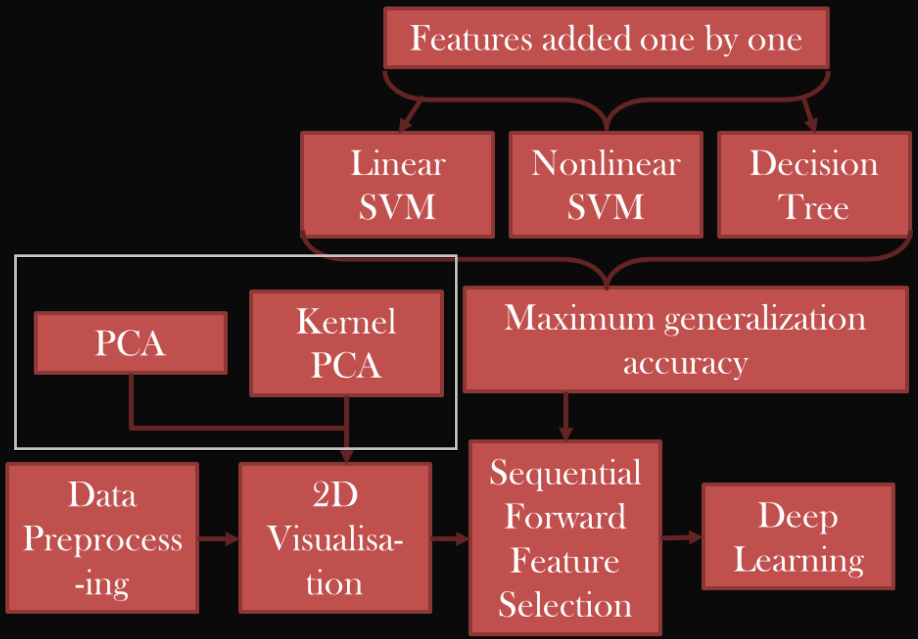
Hall, Logan, Ligon, & Derrick, 2019

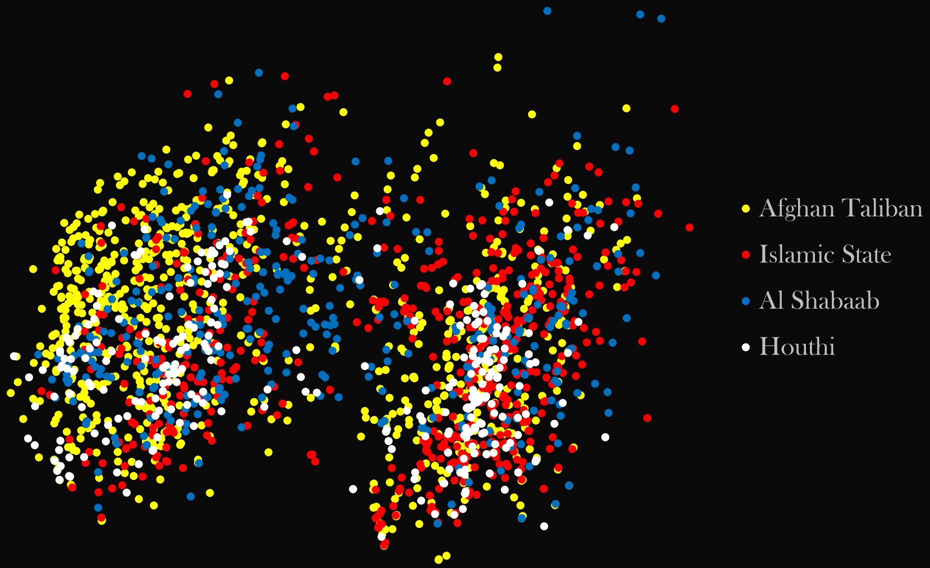
What should we do about terrorism online?

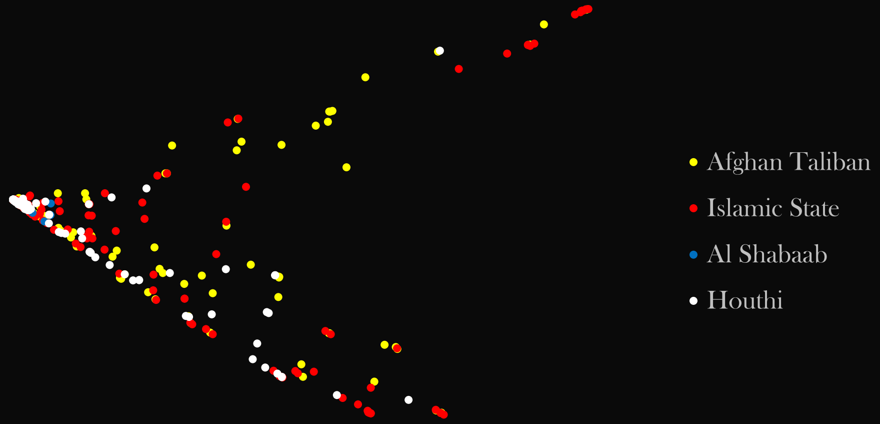


Hashemi & Hall, 2018

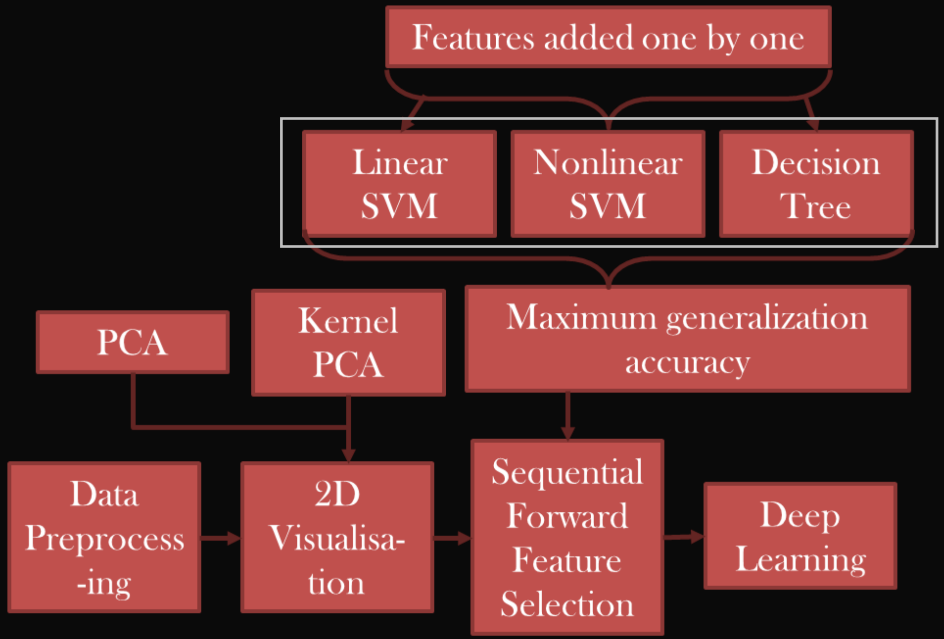
Do groups have a ‘tell’ in their destruction pattern?







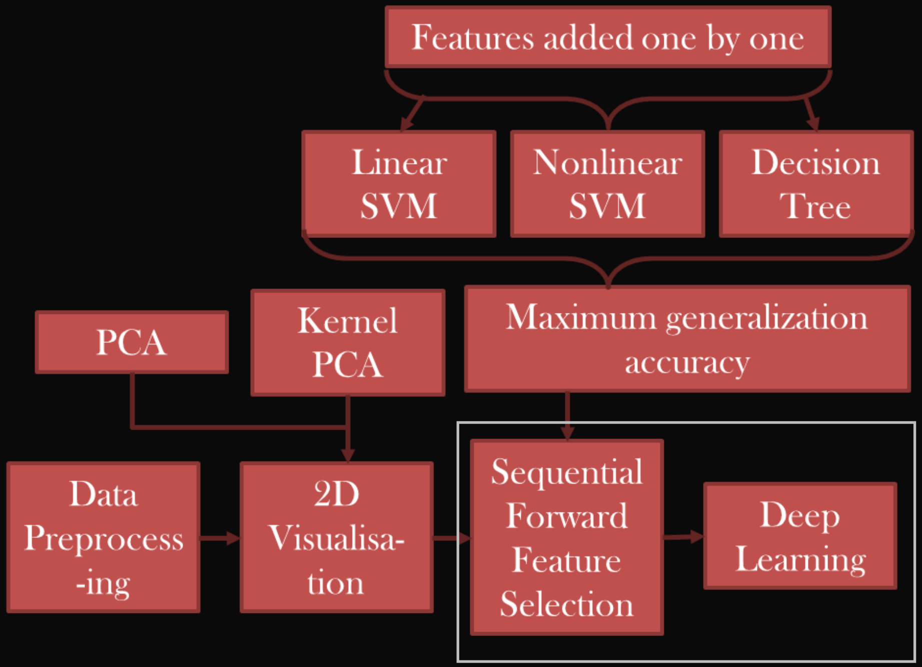
Hashemi & Hall, 2018



	Decision tree	Gaussian kernel SVM	Linear SVM
L	29.52%	12.16%	8.86%
L,N	31.45%	12.13%	5.74%
L,N,F	32.16 %	15.25%	9.66%
L,N,F,G	34.36 %	16.83%	11.77%
L,N,F,G,E	35.85%	16.94%	14.71%
L,N,F,G,E, M	36.06%	19.74%	15.76%
L,N,F,G,E, M,I	36.42%	20.72%	16.91%
All features	36.18%	24.76%	26.43%

Selected and unselected feature sets informs on aspects of attacks that are well-planned and the aspects that are more of random consequences.

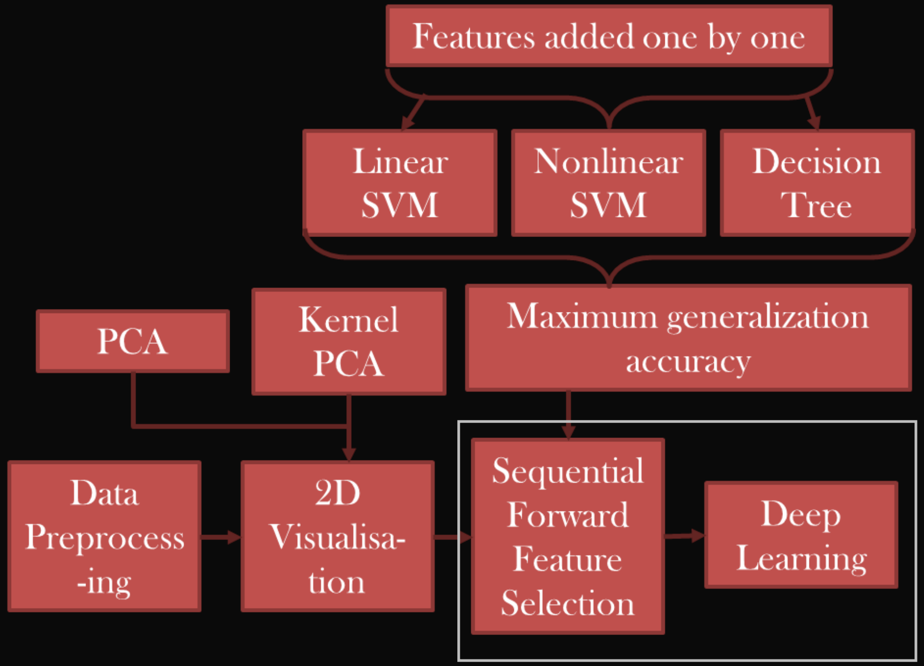
Hashemi & Hall, 2018



Algorithm 1. Sequential forward feature selection
l: number of features
for j = 1, …, l
for i = 1, …, l-j
temporarily include the i-th non-selected feature in the feature set
c_i = max{Accuracy of decision tree, Accuracy of nonlinear SVM, Accuracy of SVM}
$c^j = \max_i(c_i)$
if $c^j \leq c^{j-1}$
return the feature set
else
permanently include the i-th non-selected feature in the feature set

Feature selection techniques support more robust analyses

Hashemi & Hall, 2018



Number of hidden layers	Number of nodes in each hidden layer	Average test error	Std 10x CV
1	25	61.56%	2.05
1	50	61.53%	2.52
2	25	61.50%	2.43
2	50	61.80%	1.64
3	25	62.52%	1.58
3	50	62.39%	2.34

Multilayer perceptron accuracy based select features of the violent act reached 39%; 43% for all features; 42% with PCA generated-features

Hashemi & Hall, 2018

What should we do about terrorism online?

Removal options are still manually intensive.

- Freedom of speech
- Whack a mole problem
- Big data problem
- Burn out

How reliable are benchmark solutions of human coders taken from radical content on the open web?

Hall, Logan, Ligon, & Derrick, 2019; Hall & Haas 2021; Hall & Haas 2022

The Approach(1)

- Four-person team in jury format
- 4 main categories
  - 4 categories were added in initial QA round
  - 1 category change in final QA round

The Approach(2)

		Second Round Classification								
		Training Materials	*Hard	Non-ISIS	*None	O.C.	*Soft	*S.P.	Useless	
First Round Classification	Flags of ISIS	0	0	0	36	0	0	0	0	36
	*Hard Propaganda	0	35511	268	723	13	96	40	0	36651
	Non-ISIS Groups	0	17	1291	31	0	4	6	2	1351
	*None	452	520	586	18522	946	242	102	76	21446
	Official Communications	0	0	1	23	4411	0	0	0	4435
	*Soft Propaganda	0	123	739	4392	1	4163	40	0	9458
	*Symbolic Propaganda	0	193	10	134	4	249	933	0	1523
	Useless	0	6	4	18	0	6	1	20307	20342
Total		452	36370	2899	23879	5375	4760	1122	20385	95242

Images


تفقد خلايا بيوت النحل بين الحين والآخر


الاشتباك مع عناصر الجيش النصيري بالقرب من مطار الخير العسكري بالأسلحة الخفيفة
عاجل
٢٥ جمادى الآخرة ١٤٣٧
الخير


جانب من تدريبات كتيبة القناصين


دوّار مدخل البلدة


قصف صحوات الردة لعوام المسلمين في مدينة الحجر الأسود وشارع العروبة في جنوب دمشق
عاجل
١٧ ذو القعدة ١٤٣٦
دمشق

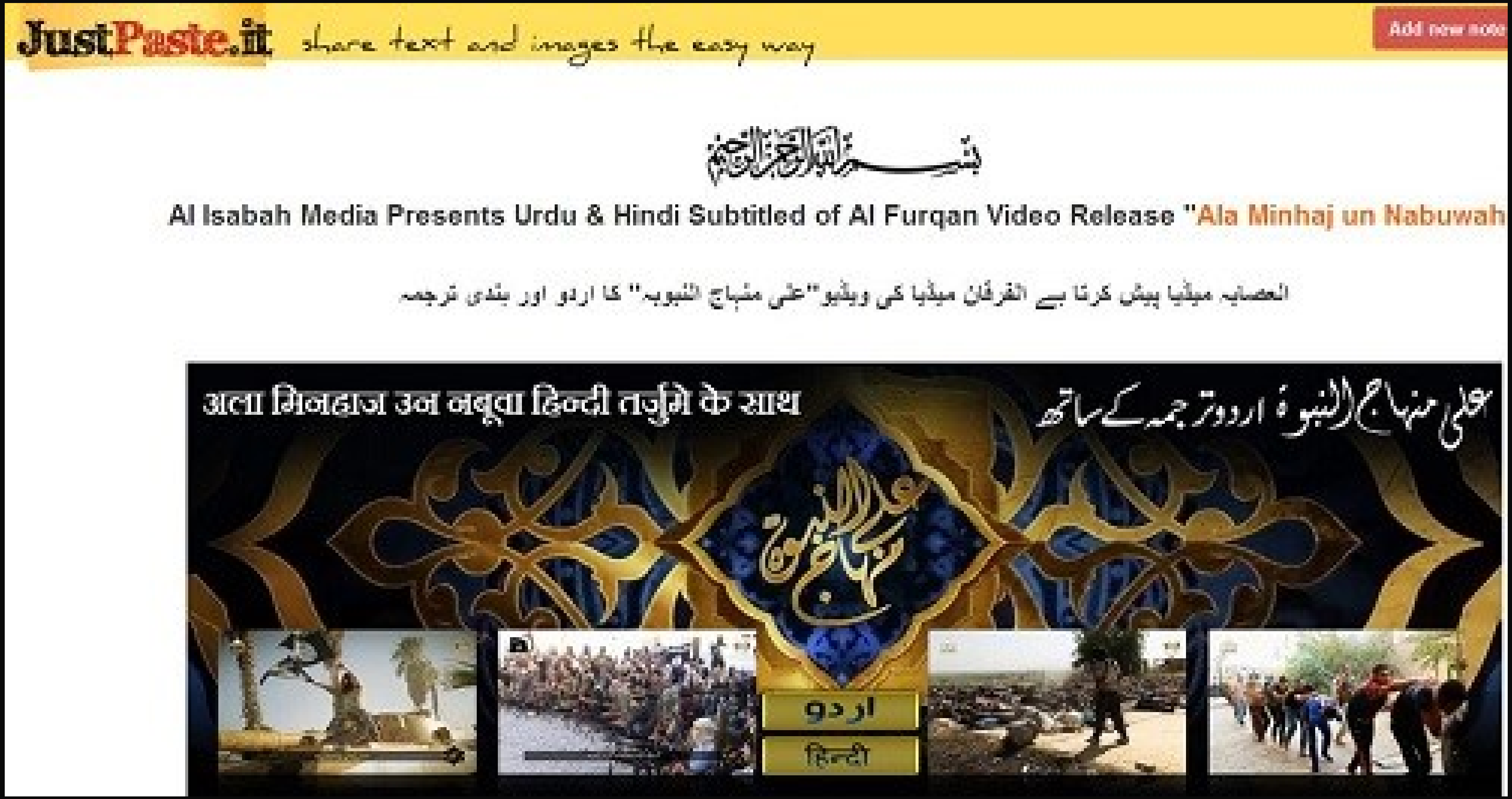

تلاوة ما تيسر من القرآن

Tweets

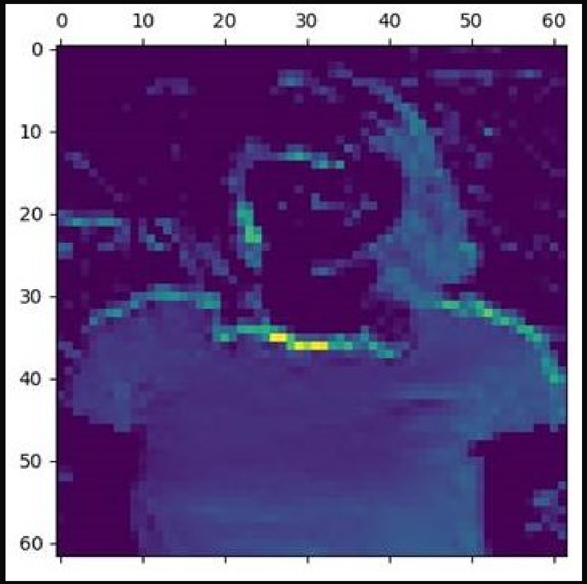
إعلامي مؤسسة الفرقان @alfurqan2013 · 1m

غطية مصورة لجانب من العمليات العسكرية في ناحية البغدادي غربي #ولاية_الأنبار
justpaste.it/albaghdadi1435
#الدولة_الإسلامية

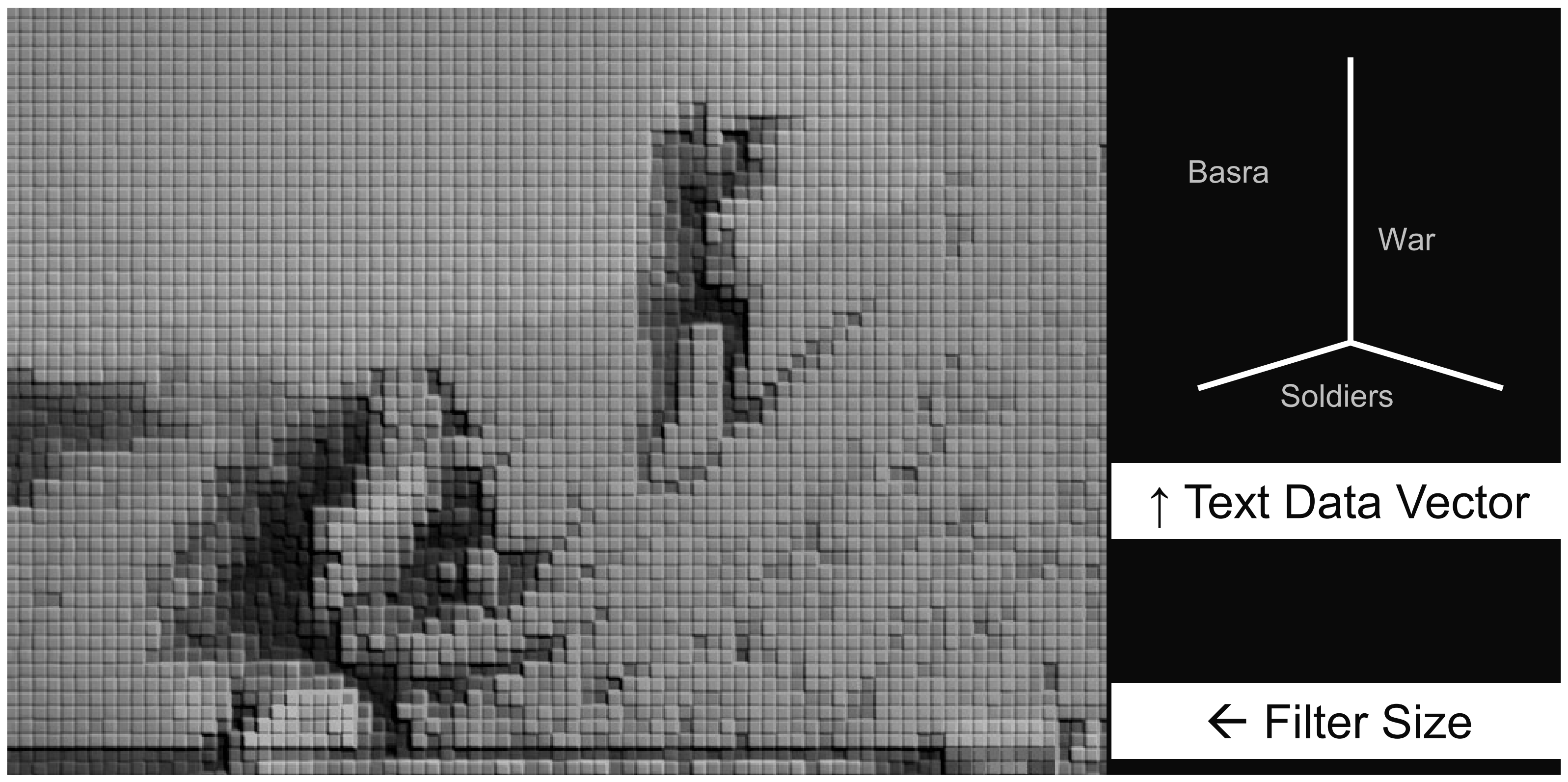
Pages


JustPaste.it
share text and images the easy way
Add new note
بسم الله الرحمن الرحيم
Al Isabah Media Presents Urdu & Hindi Subtitled of Al Furqan Video Release "Ala Minhaj un Nabuwah
العصابہ میڈیا پیش کرتا ہے الفرقان میڈیا کی ویڈیو"علی منہاج النبوہ" کا اردو اور ہندی ترجمہ
अला मिनहाज उन नबूवा हिन्दी तर्जुमे के साथ
علی منہاج النبوۃ اردو ترجمہ کے ساتھ
اردو
हिन्दी

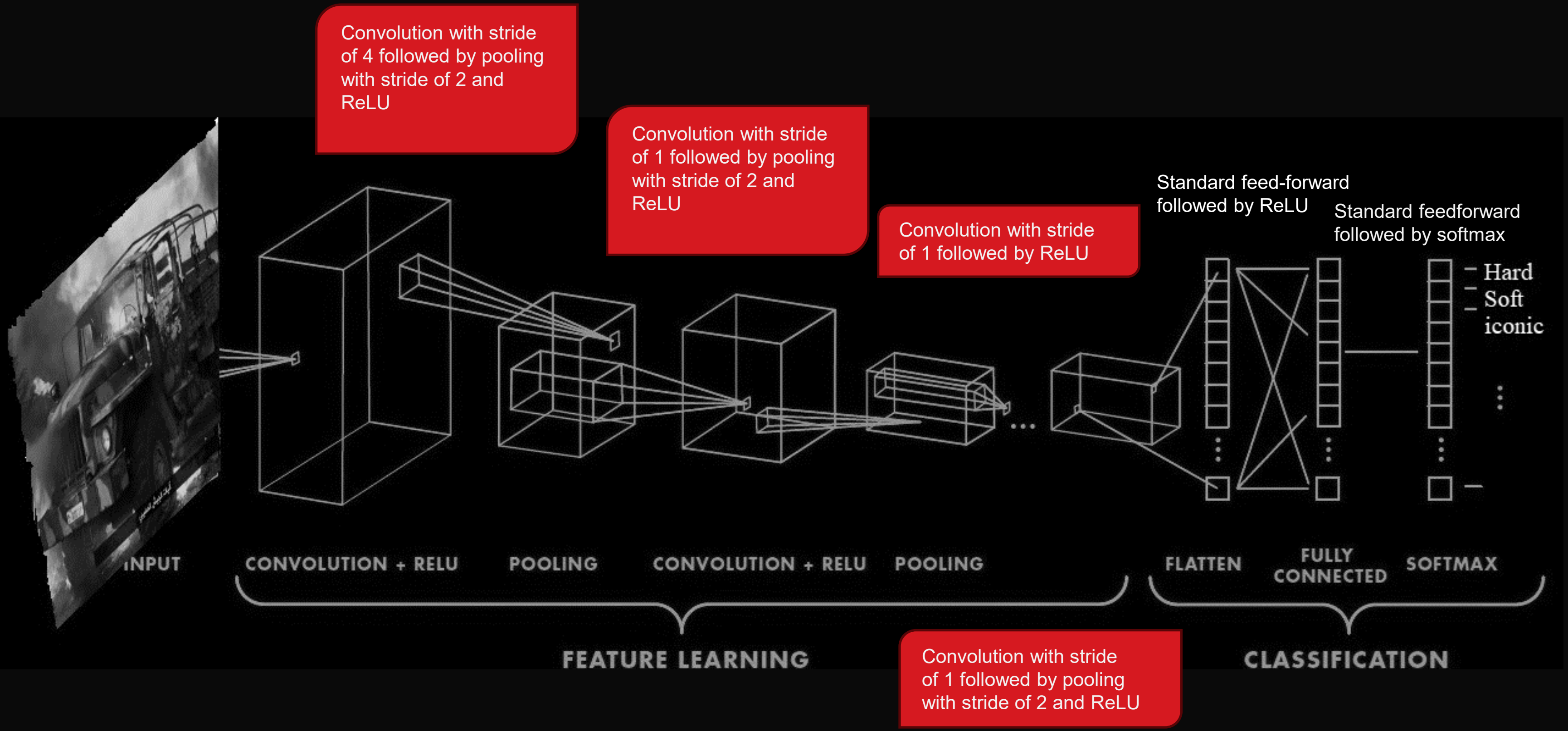
Typical Max Pooling Layer



Proposed Initial Layer of a Combined Model



Stylized CNN



Convolutional layers:

$f_1 = 42\times42\times2$, $s_1 = 3$, $n_1 = 74$

$f_2 = 15\times15\times5$, $s_2 = 3$, $n_2 = 26$

$f_3 = 13\times13\times24$, $s_3 = 2$, $n_3 = 46$

$f_4 = 12\times12\times84$, $s_4 = 2$, $n_4 = 89$

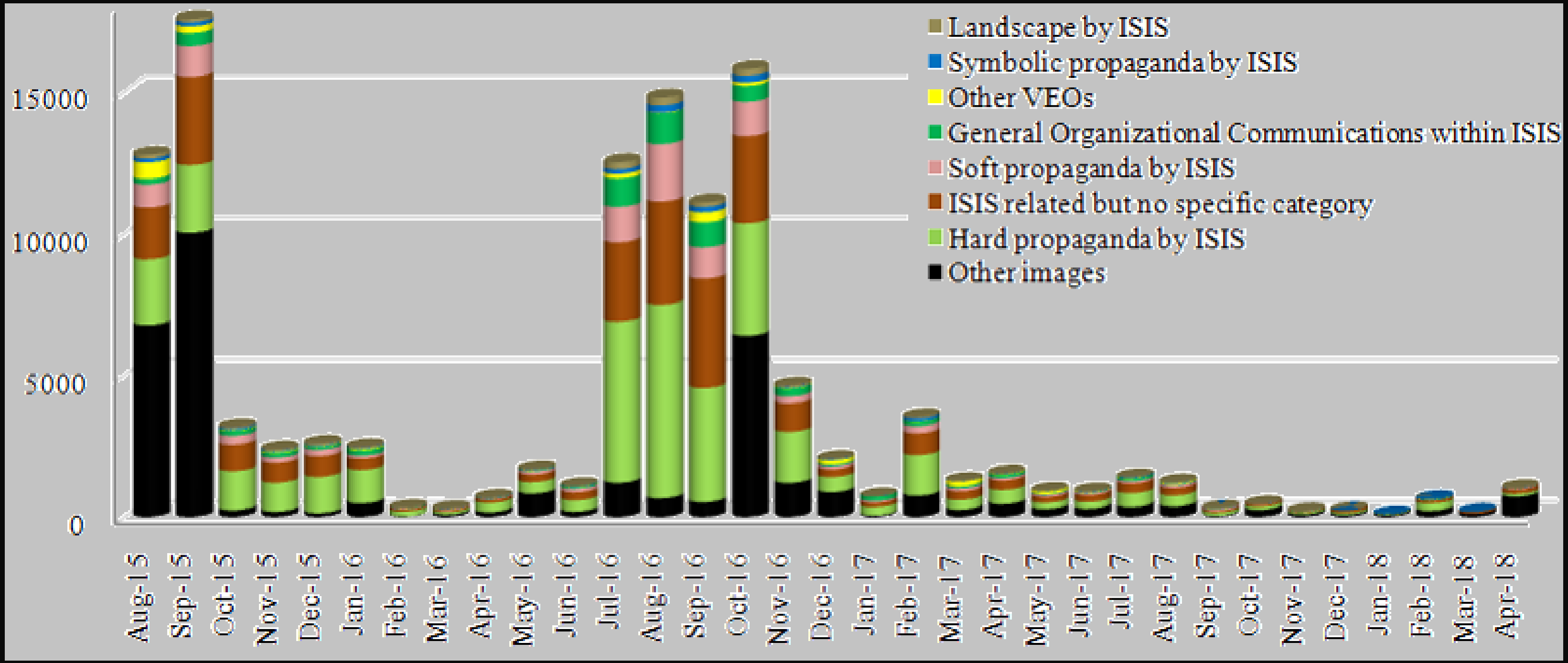
$f_5 = 5\times5\times15$, $s_5 = 2$, $n_5 = 62$

where f_m, s_m, and n_m denote the stride, size, and number of windows of the m-th layer, respectively.

The first two fully-connected layers have 4096 neurons each and the third fully-connected layer has as many neurons as the number of classes (8 here).

Hashemi & Hall, 2019

How well do machines identify extremist propaganda?



Class	Hard Propaganda	Soft Propaganda	Symbolic Propaganda	Organizational Communications	Landscapes	ISIS-other	Other Groups	None
Absolute size	38,460	9,933	1,605	5,960	1,569	26,360	2,151	34,597
Relative size	31.88%	8.23%	1.33%	4.94%	1.30%	21.85%	1.78%	28.68%
Above: Labelled Images; Below: All Images								
Absolute size	97964	25611	6886	16906	2532	91277	6330	941820
Relative size	8.24%	2.15%	0.58%	1.42%	0.21%	7.67%	0.53%	79.19%

Hashemi & Hall, 2019

Visual propaganda from OSM can be reliably detected

Binary Classifier

Overall generalization accuracy

97.02%

Overall generalization ***F1***

97.89%

8-way Classifier (intent-based)

Overall generalization accuracy

86.08%

Overall generalization ***F1***

85.76%

Class	Hard Propaganda	Soft Propaganda	Symbolic Propaganda	Organizational Communications	Landscapes	ISIS-other	Other Groups	None
Precision	84.47%	73.50%	66.80%	98.86%	75.00%	79.56%	94.43%	93.84%
Recall	92.78%	61.12%	57.79%	95.50%	61.56%	73.50%	77.40%	96.01%
F1	88.43%	66.74%	61.97%	97.67%	67.62%	76.41%	85.07%	94.91%

Hashemi & Hall, 2019

Evaluation of Approach

- Two κ measurements in use
  - Cohen's κ
  - Fleiss' κ

	Cohen's Kappa	Asymptotic			Asymptotic 95% Confidence Interval	
		Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	0.857	0.002	483.591	0.000	0.854	0.860

Fleiss' κ

Rating Category	Conditional Probability	Kappa	Asymptotic			Asymptotic 95% Confidence Interval	
			Standard Error	z	Sig.	Lower Bound	Upper Bound
Flags of ISIS	0.000	0.000	0.003	-0.058	0.953	-0.007	0.006
General Training Materials	0.000	-0.002	0.003	-0.734	0.463	-0.009	0.004
Hard Propaganda	0.973	0.956	0.003	294.913	0.000	0.949	0.962
Non-ISIS Groups	0.608	0.599	0.003	184.727	0.000	0.592	0.605
None	0.817	0.760	0.003	234.623	0.000	0.754	0.767
Official Communications	0.899	0.894	0.003	275.844	0.000	0.887	0.900
Soft Propaganda	0.586	0.552	0.003	170.407	0.000	0.546	0.559
Symbolic Propaganda	0.705	0.701	0.003	216.441	0.000	0.695	0.708
Useless	0.997	0.996	0.003	307.524	0.000	0.990	1.003

a. Sample data contains 95242 effective subjects and 2 raters.

Lessons Learned

Reliable and Explainable Detection

Machines can detect terror propaganda	It is possible to separate religious, cultural, and co-opted messaging

Generalizable Lessons and Future Research Directions

Machines can learn nuance	Low-bias management regimes are possible but un(der) developed

Hall & Haas, 2021; Hashemi & Hall, 2019; Caton, Hall, & Weinhardt, 2015

Discussion

How reliable are benchmark solutions of human coders taken from radical content on the open web?

→ Soft propaganda is hard to define.

→ Good performance on frequent samples; lesser performance on the margins.

→ Necessary tradeoff between accuracy and scale.

Next Steps

How reliable are benchmark solutions of human coders taken from radical content on the open web?

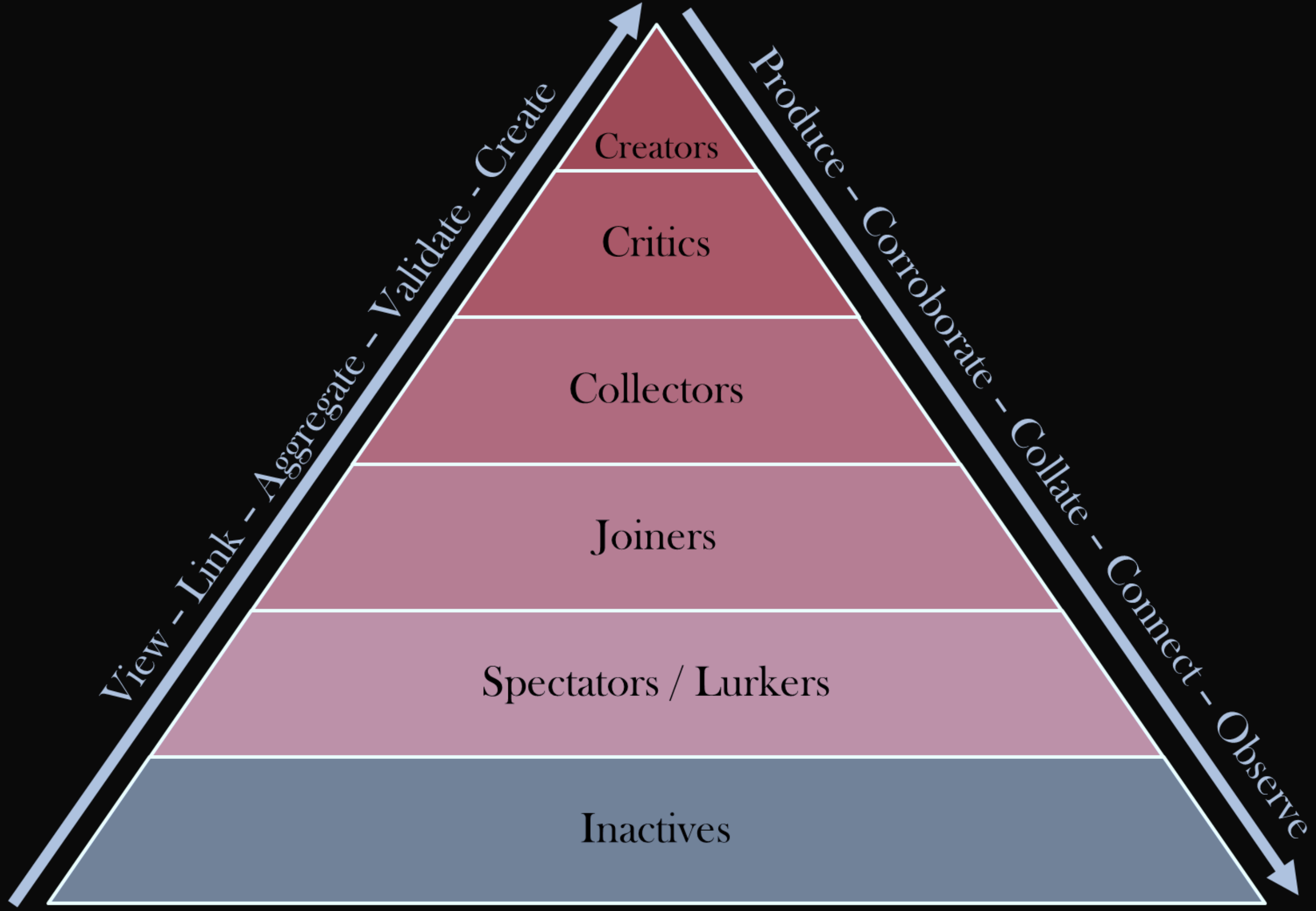
→ Transfer learning to pre-sort Hard Propaganda.

→ Separation/further refinement of Soft Propaganda.

→ Overarching need for method to combine text and images (Crowe, Ricks & Hall, 2024).

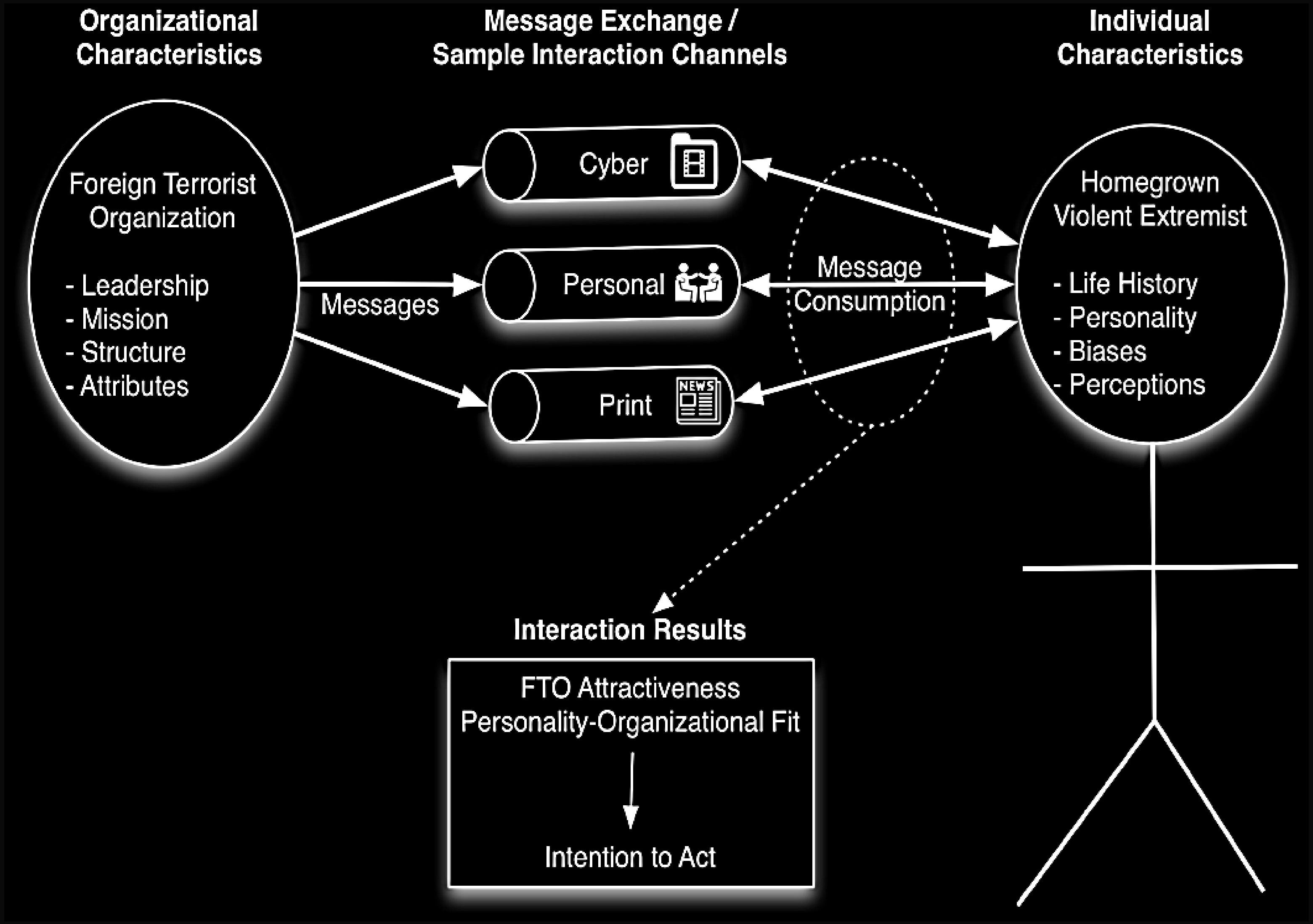
Questions?

(Digital) Participation theory can address individual online extremism



Li & Bernoff, 2011; Ligon, Hall, & Braun, 2018

(Digital) Participation theory can address individual online extremism



Li & Bernoff, 2011; Ligon, Hall, & Braun, 2018