# The Human Element in Document Analysis Systems Daniel Lopresti <sup>[0000-0003-2129-4223]</sup> Lehigh University, Bethlehem PA 18015, USA DAS 2022 lopresti@cse.lehigh.edu



## Summary

• Humans are a vital component in document analysis systems. A more intentional inclusion of the human element would yield better outcomes for those impacted by what we build, and also lead to interesting research questions.

### Motivation

- The "S" in "DAS" refers to systems.
- This definition should include humans at all levels, including those who design, implement, and use document analysis systems.
- It should also include those who are indirectly impacted by the systems we build.
- To date, little attention has been paid to this, beyond attempts to facilitate ground-truthing and a few applications to help the disabled.
- Elsewhere, the broader field of AI is confronting serious issues of bias and fairness (e.g., facial recognition, large language models).
- Is the document analysis community immune to such concerns?

Ec]anoka_blaine_Z_6_challengedbalbt1-000.tif (x1.0) [1 of 10]           Perfect         Exction           Bection         Exitor		E becker_Cormorant_challengedballot2_F-000.tif (x1.0) [2 of Fefect Dection Data Note Display Markup Ineer	10] La Carlo
JOHN MCCAIN AND SARAH PALIN Republican		PRESIDE VICE PRI VOTE FOR	NT AND SIDENT
CARACK OBAMA     AND JOE BIDEN     Democratic-Farmer-Labor		JOHN MCC SARAH PA Republican	
CYNTHIA MCKINNEY AND ROSA CLEMENTE			BAMA AND
naka_Ande var_P?shellengedizelle11-000.htf (p1.0) [5 el 10]	E becker_ShellLake_chall Perfect Bectian Bala	illengedballut2_F-000.tlf (x1.0) [3 of 10] of Node Display Markup Image Vindaws	L C C C C C C C C C C C C C C C C C C C
UNITED STATES SENATOR	0	DEAN BARKLEY	COUNTY
		NORM COLEMAN Republican	COUNTY CON -
		AL FRANKEN Democratic-Farmer-Labor	UISTR Vote FC
		CHARLES ALDRICH Libertarian	
CHARLES ALDRICH			write-in, if any
			CONTRACTOR OF CAREFUL AND A CONTRACTOR

#### Sloppy-but-valid marks

## **Three Illustrative Cases**

Where might we start looking to see whether our systems and our policies negatively impact segments of the population?

#### **Reliably Reading Hand-Marked Paper Ballots**

- Optical Mark Reading has a long history and appears to be a solved problem.
- So simple it does not attract much attention.
- When examining a real population of voters, however, representing all demographics from across society, the problem becomes much more challenging.
- Voter intent is the overriding consideration.
- A system that fails to count certain voters' ballots accurately is a serious problem. A ballot-reading system that looks like it is doing a good job on average may still be unfairly disenfranchising certain groups of voters.
- Images below are from ballots challenged during the 2008 U.S. presidential election.
- we know whose votes are not being Do counted correctly? Should we worry?

PRESIDENT AND VICE PRESIDENT	anska, Fridey, 1. 1. chalengedholarti 500 tif (pl.4). [7 of 10]	<b>158</b>		
JOHN MCCAIN AND SARAH PALIN	Difference     Partice     Table     Table     Table     Table       DHN MCCAIN AND NRAH PALIN     PRESIDENT and VICE-PRESIDENT     Fail			
BARACK OBAMA AND JOE BIDEN	JOHN MCCAIN AND SARAH PALIN Republican	To in aga		
Democratic-Farmer-Labor BA Demo CY AN Green RÓ		Better Date PL challesgefielden 202 of [b1.0] [b of 10]     Better Date Place Date Marke Date Marke      PRESIDENT and VICE-PRESIDENT     VOTE FOR ONE TEAM	Failure tc	
		JOHN MCCAIN     AND SARAH PALIN     Republican     BARACK OBAMA     AND JOE BIDEN	To vote i in the c against t	
Non-con marking	-	CYNTHIA MCKINNEY AND ROSA CLEMENTE Green RÓGER CALERO	CL Shall the M funding to enhance	

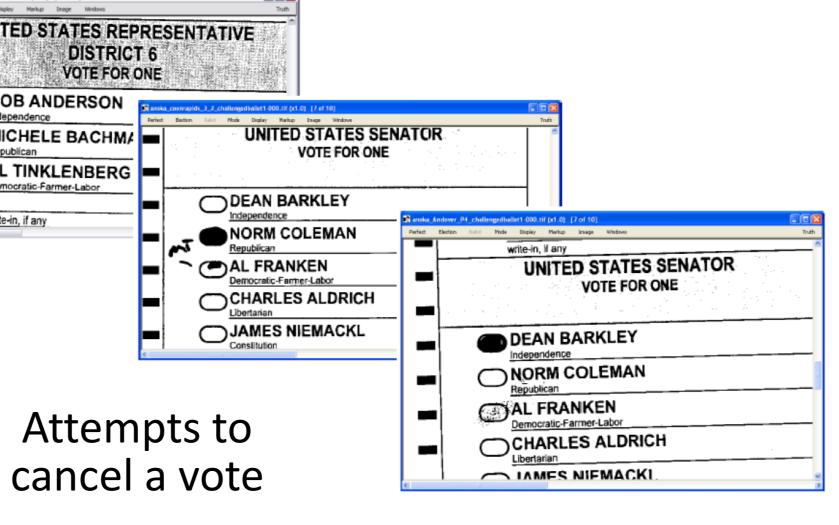
R anoka_	blaina (1)	97 cha	llenseth	allee1.
Perfect	Bection	Balot	Mode	Disple
			UN	11
=		$\mathbf{x}$		BO ndep MIC Repu
		4		EL
6	Ľ.,	_	<u> </u>	vrite-

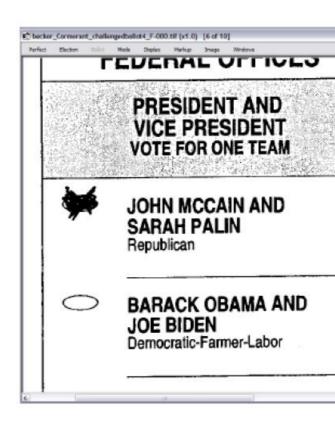
#### **Robust Signature Verification for Elections** • Voters are sometimes required to sign their name as proof of identity.

- decades earlier.

#### **Under-Resourced** Analysis Document for Languages

- contribution."







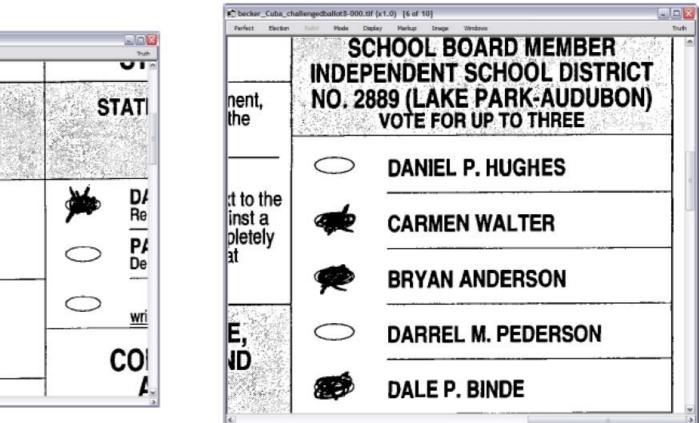
• This signature is compared to one collected when the voter first registered, which could be

• These comparisons are normally made by election officials, but if there is a push to automate this process, who will be hurt? • Voters with a name-change due to marriage, a hand injury, a stroke, or forgetfulness about how they signed so many years ago?

• Thousands of languages in the world today, but most research reported at events like DAS reflects only a small percentage of these.

• Often, submissions applying known techniques to a new language will be rejected for "lack of

Existing methods are often easily adapted, but why should first language "win the race"? What do we lose by erecting such barriers?



Valid votes that look cancelled