Counting Votes and the Attempt to Replicate Human Interpretation

LABELNOISE'2017
December 1, 2017

Professor Daniel Lopresti
Computer Science and Engineering
Lehigh University
Bethlehem, PA, USA

Motivation

Attempt to interest you in an noisy label application area:

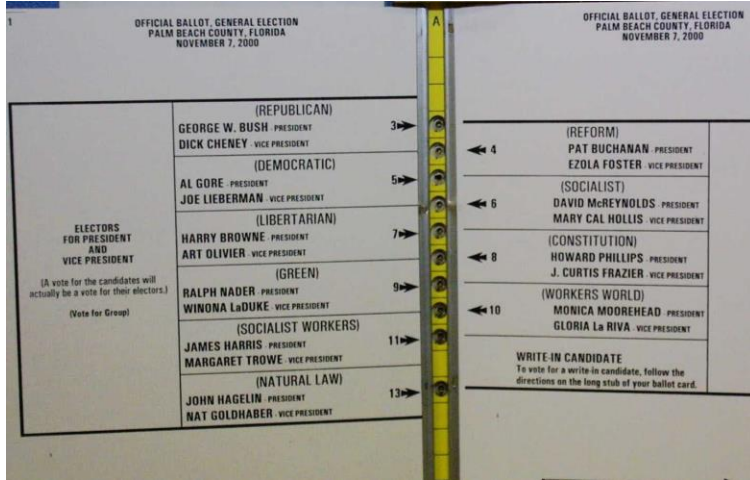- Counting votes – replicating human interpretation

Also, if there's time:

- Turing Test-inspired view of performance evaluation

Key features:

- Vitally important application, drawn from real life
- Valuable lessons to learn that can be applied elsewhere
- Wonderful opportunity to apply our talents
- Work still needed to frame problem, outline next steps
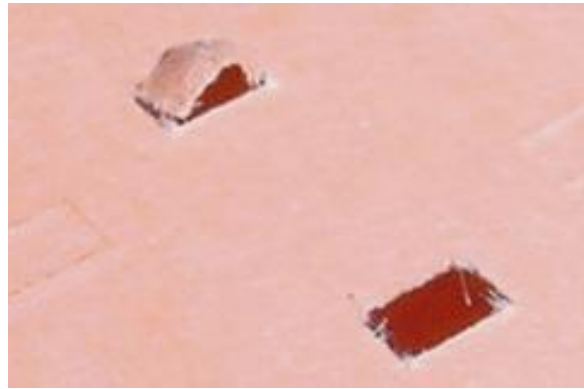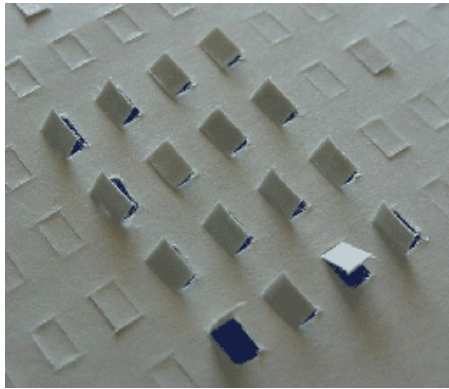
How did US get where we are today?

OFFICIAL BALLOT, GENERAL ELECTION
PALM BEACH COUNTY, FLORIDA
NOVEMBER 7, 2000

The infamous butterfly ballot from the 2000 Presidential election.

Bush v. Gore.

ELECTION 2000 EXTRA
WITH THE LATEST RESULTS
The Palm Beach Po
WEDNESDAY, NOVEMBER 8, 2000   FINAL EDITION

Florida's votes close enough for recount

IT'S NOT OVER

Hanging Chads and Voter Intent

Votomatic technology used in Florida was prone to paper jams. This led to hanging and dimpled chads, making it hard to determine voter intent.

http://www.cs.uiowa.edu/~jones/cards/chad.html
http://www.pushback.com/justice/votefraud/DimpledChadPictures.html

Next Big Step … Backward

Voting in the News: Take 3

The Voting Technology We Really Need? Paper

Software-independent backup systems are more important than ever.

LAWRENCE NORDEN | MAY 10, 2017 | TECHNOLOGY

f Share   **y** Tweet   ...   TEXT SIZE − +

In January, America's main intelligence agencies issued a report concluding that Russia interfered in the 2016 election, using a combination of cyber-intrusion, espionage, and propaganda. In addition to the details provided in this account, media outlets have since reported that several election databases were hacked before and after the election. While the Department of Homeland Security found no evidence any of these efforts manipulated vote tallies, the assaults have left many Americans asking: Just how safe are voting machines from cyberattack?

The answer is not reassuring.

Russian hacking fuels return to paper ballots

By Jenni Bergal   Oct 03, 2017

This article originally appeared in Stateline, an initiative of the Pew Charitable Trusts.

After the "hanging chad" fiasco during the 2000 presidential recount, many states and counties switched to electronic-only voting machines to modernize their systems.

Now, amid security concerns over Russian hackers targeting state voting systems in last year's election, there's a renewed focus on shifting to paper ballots.

In Virginia, election officials decided last month to stop using paperless touch-screen machines, in an effort to safeguard against unauthorized access to the equipment and improve the security of the state's voting system.

In Georgia, which uses electronic voting machines with no paper record, legislators are discussing getting rid of their aging equipment and using

MORE INFO

Virginia considers decertifying touchscreen voting machines
The state could accelerate the move to systems that provide a paper trail for election audits.
Read more.

https://www.theatlantic.com/technology/archive/2017/05/the-voting-technology-we-really-need-paper/524820/
https://gcn.com/articles/2017/10/03/return-paper-ballots.aspx

Voting in the News: Take 3

Virginia Politics

Paper ballots make a comeback in Virginia this fall

In face of hacks and attempted hacks, paper ballots make a comeback in Virginia this fall. (Paul J. Richards/Agence France-Presse Via Getty Images)

By Patricia Sullivan October 7

The Computer Scientist Who Prefers Paper

Barbara Simons believes there is only one safe voting technology.

JILL LEOVY | DECEMBER 2017 ISSUE | TECHNOLOGY

https://www.washingtonpost.com/local/virginia-politics/paper-ballots-make-a-comeback-in-virginia-this-fall/2017/10/07/
https://www.theatlantic.com/magazine/archive/2017/12/guardian-of-the-vote/544155/

A Simple Yet Vexing Case Study: Counting Votes Recorded on Paper

Topic of current interest where the legal need to respect voter intent transforms a seemingly trivial pattern recognition problem into much more complex task.

Counting Votes Not So Easy

INSTRUCTIONS TO VOTERS

To vote, completely fill in the oval(s) next to your choice(s) like this: ⬤

Is this a legal vote?

- Courts would probably say so ...
- ... but op-scan readers might not count it.

> *Increasing demands that machine's interpretation match a human's.*

Research Questions

Issues that arise from using paper ballots in elections:

- Accurate interpretation of marginal markings.
- Human cost, error rate, and bias in performing manual recounts.
- Failure modes in ballot imaging (e.g., paper jams).
- Systematic errors due to ballot layout (one candidate may be disadvantaged over another based on physical location on page).

Also keep in mind:

- U.S. elections can be complex (10's to 100's of choices).
- Impact of "voter error" (e.g., improper markings, erasures).
- Potential for traditional ballot-box stuffing.
- Computer hackers attempting to manipulate the vote.

Why isn't this a solved problem?

Students have been taking standardized tests using op-scan answer sheets for decades ...
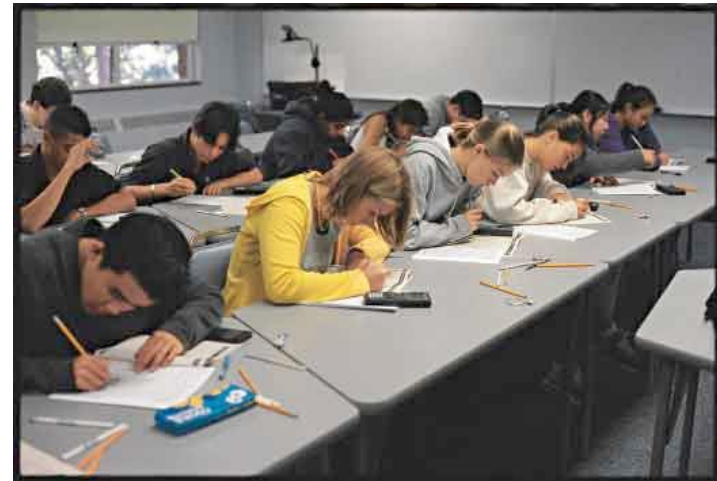
- While accuracy rates are very high, problems do occur.
- Compared to voters, students are a much more homogeneous (and well-educated) population.
- Standardized testing is NOT anonymous. Students can (and do) complain when they receive a lower score than they expect.

Connection to Forms Processing

Similarities to forms processing, but also some key differences:

- Much broader range of users (education level, literacy, etc.) than for traditional forms applications.
- Ballots must preserve a voter's anonymity.
- Demand to count votes and report results quickly.
- Elections are held infrequently, so voting equipment sits unused for long periods in storage.
- Poll workers often lack technical expertise.
- Maintaining chain-of-custody is a critical security requirement.
- No *financial* interest in making sure votes are counted accurately, but there is tremendous *public* interest.

Counting Votes Not So Easy

Real ballot from an election in California:
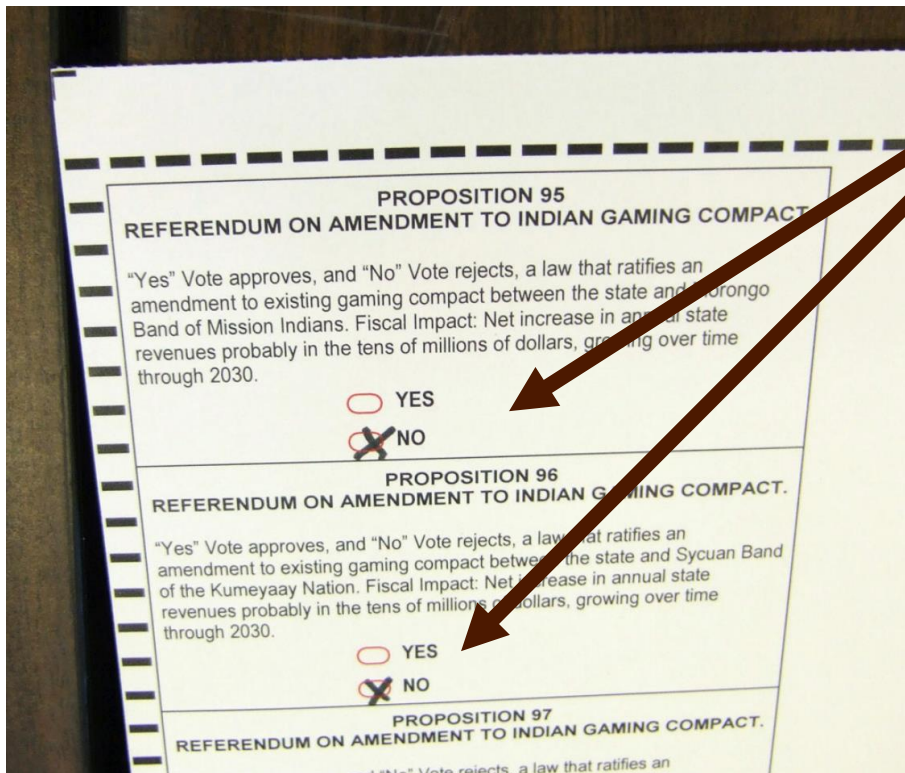
One of these votes was counted correctly by the op-scan equipment, the other was not.

Note: this does <u>not</u> mean voting on paper ballots is bad, just (1) manual audits should be mandatory, and (2) more research is needed.

"Improving California's 1% Manual Tally Procedure," Joseph Lorenzo Hall, UC Berkeley School of Information, EVT Workshop 2008.

Whole-Ballot Recognition

Can we capture voter intent via style-based techniques?

Style-Based Mark Recognition

Traditional Forms Processing

Style-Base Ballot Mark Recognition

Can the system interpret the voter's intent? (If a human judge would interpret a marking as an intended vote, then the voting machine should do the same.)

Can fail to record some votes simply because they do not satisfy an arbitrary criterion (e.g., a fixed threshold on the number of black pixels).

Assume a voter is self-consistent when marking his/her ballot.

Create a style-based classifier from a set of style-specialized classifiers to improve recognition accuracy.

Limiting

Promising

"Style-Based Ballot Mark Recognition," P. Xiu, D. Lopresti, H. Baird, G. Nagy, and E. Barney Smith, *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, July 2009, Barcelona, Spain, pp. 216-220.
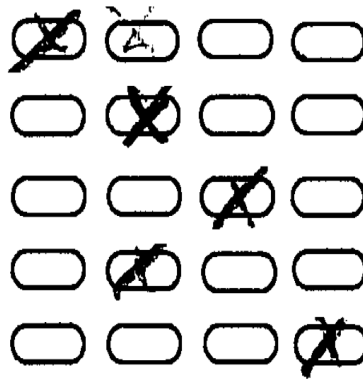
Challenging Cases
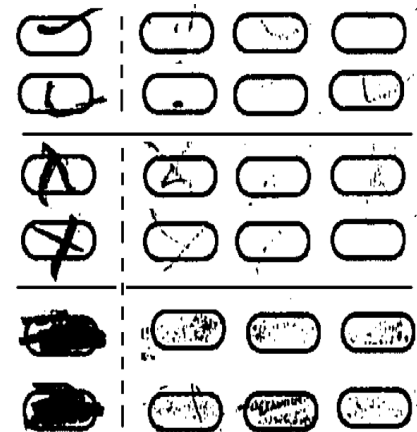
A Style-Consistent Field

Human Interpretation

VNNN,NVNN,NNVN,NVNN,NNNV

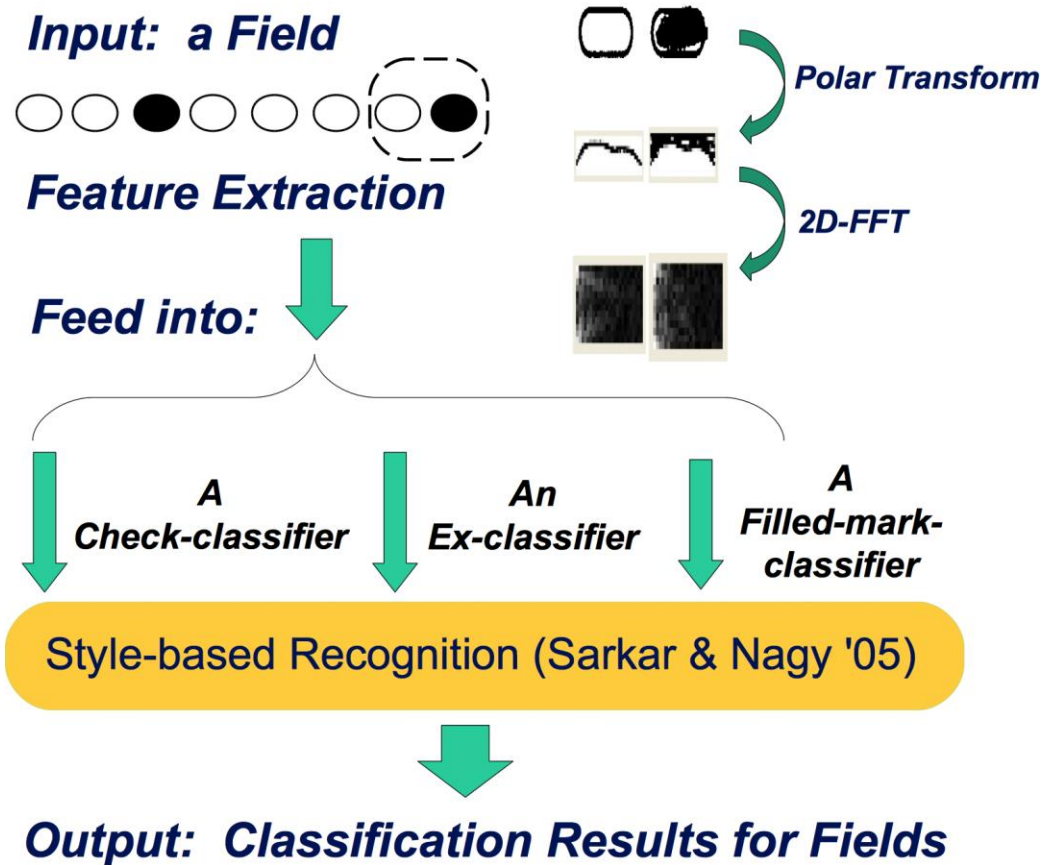**(V for Vote, N for Non-vote)**

Variations in Marking Style

**Check, ex, and filled marks (left)
vs. noisy non-votes (right)**

"Style-Based Ballot Mark Recognition," P. Xiu, D. Lopresti, H. Baird, G. Nagy, and E. Barney Smith, *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, July 2009, Barcelona, Spain, pp. 216-220.

System Design

Input: a Field

Feature Extraction

Feed into:

Polar Transform

2D-FFT

A Check-classifier   An Ex-classifier   A Filled-mark-classifier

Style-based Recognition (Sarkar & Nagy '05)

Output: Classification Results for Fields

"Style-Based Ballot Mark Recognition," P. Xiu, D. Lopresti, H. Baird, G. Nagy, and E. Barney Smith, *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, July 2009, Barcelona, Spain, pp. 216-220.

Style-Based Performance

Table 3. Target-level error rates (top) and field-level error rates (bottom).

Sample Set	Classifier					
	Check	Ex	Filled	Blend	Separate	Style-based
Check	2.36%	7.46%	25.00%	1.97%	4.35%	2.78%
Ex	0.40%	0.34%	16.16%	0.40%	0.40%	0.35%
Filled	2.75%	2.38%	1.10%	2.75%	2.50%	1.09%
Average	1.84%	3.39%	14.09%	1.70%	2.42%	1.41%

Sample Set	Classifier					
	Check	Ex	Filled	Blend	Separate	Style-based
Check	38.30%	83.25%	100.00%	33.43%	61.08%	42.85%
Ex	7.77%	6.70%	99.30%	7.77%	7.77%	6.75%
Filled	53.18%	46.07%	20.75%	53.18%	48.55%	20.63%
Average	33.08%	45.34%	73.35%	31.46%	39.13%	23.41%

"Style-Based Ballot Mark Recognition," P. Xiu, D. Lopresti, H. Baird, G. Nagy, and E. Barney Smith, *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, July 2009, Barcelona, Spain, pp. 216-220.

A Bit of Good Luck

But what we'd like to have is ballots from a real election. Even better, the ballots would be from an important election where the voter markings present serious pattern recognition challenges.

Extremely close U.S. Senate race in State of Minnesota: six days after election, unofficial results showed Republican Norm Coleman leading Democratic challenger Al Franken by 206 votes out of nearly 3 million cast, a difference of less than 0.01%.

"Document Analysis Issues in Reading Optical Scan Ballots," D. Lopresti, G. Nagy, and E. Barney Smith, *Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems*, June 2010, Boston, MA, pp. 105-112.

A Bit of Good Luck

- Minnesota uses op-scan ballots.

- Closeness of election triggers a manual recount.

- Both sides are allowed to challenge validity of "questionable" ballots.

- Openness laws make challenged ballots a matter of public record.

- Ballot images made available on MN public radio website.

- PDF files contain 300 dpi TIF images!

http://minnesota.publicradio.org/features/2008/11/19_challenged_ballots/

Minnesota Statutes

Remember that the guiding principle is *voter intent*. Here are a few key points to keep in mind when interpreting ballot markings:

- "A ballot shall not be rejected for a technical error that does not make it impossible to determine the voter's intent."

- "If a mark (X) is made out of its proper place, but so near a name or space as to indicate clearly the voter's intent, the vote shall be counted."

- "Misspelling or abbreviations of the names of write-in candidates shall be disregarded if the individual for whom the vote was intended can be clearly ascertained from the ballot."

https://www.revisor.mn.gov/statutes/?id=204C.22

Minnesota Statutes

... and ...

- "If a voter uniformly uses a mark other than (X) which clearly indicates an intent to mark a name or to mark yes or no on a question, and the voter does not use (X) anywhere else on the ballot, a vote shall be counted for each candidate or response to a question marked.

- If a voter uses two or more distinct marks, such as (X) and some other mark, a vote shall be counted for each candidate or response to a question marked, unless the ballot is marked by distinguishing characteristics that make the entire ballot defective ..."

https://www.revisor.mn.gov/statutes/?id=204C.22

Minnesota Statutes

… and …

- "If the names of two candidates have been marked, and an attempt has been made to erase or obliterate one of the marks, a vote shall be counted for the remaining marked candidate."

- "A ballot shall not be rejected merely because it is slightly soiled or defaced."

- "If a ballot is marked by distinguishing characteristics in a manner making it evident that the voter intended to identify the ballot, the entire ballot is defective."

Goal here is to prevent coercion or vote selling.

https://www.revisor.mn.gov/statutes/?id=204C.22

Challenge: you be the judge

Who gets vote? Public opinion:

- Norm Coleman: 63% (7,626 votes)
- Al Franken: 4% (474 votes)
- Nobody: 33% (4,050 votes)

Challenge: you be the judge

Vote for Franken? Public opinion:
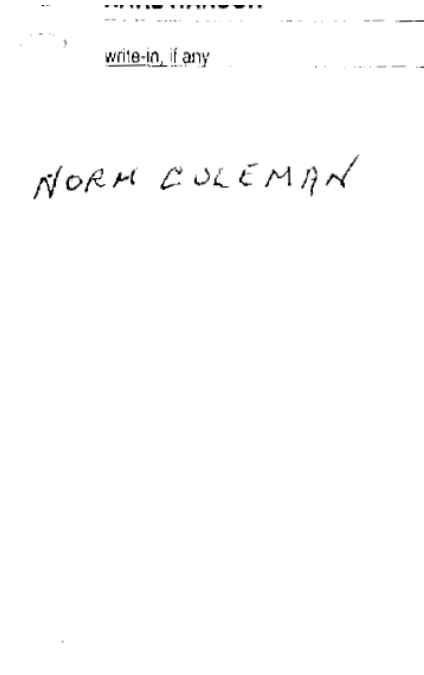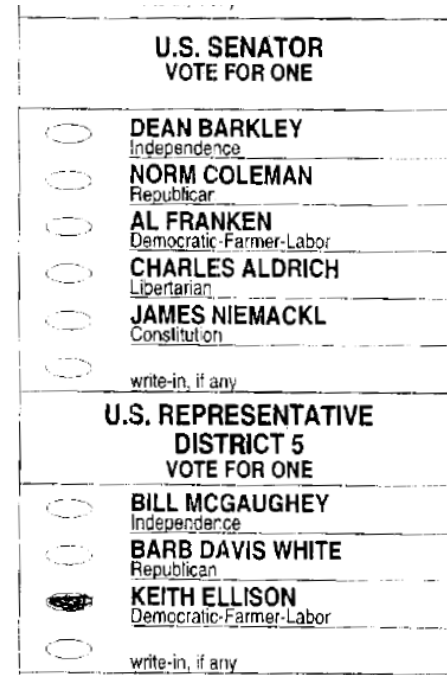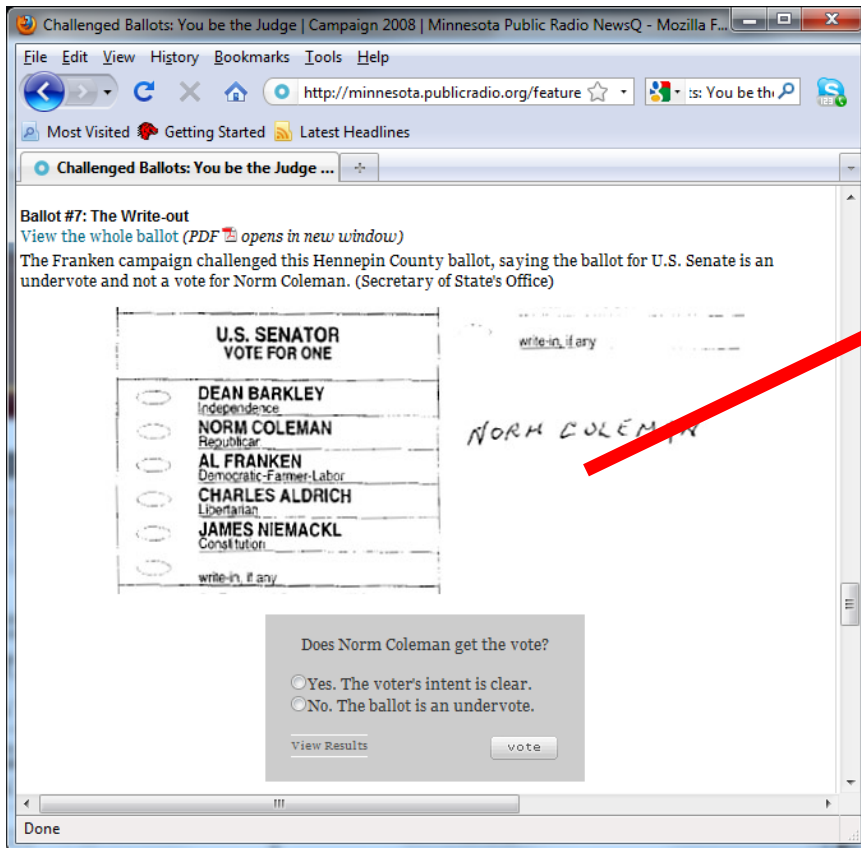- Yes: 92% (11,069 votes)
- No: 8% (1,012 votes)

Challenge: you be the judge

Vote for Franken? Public opinion:
- Yes: 96% (11,250 votes)
- No: 4% (452 votes)

Challenge: you be the judge

Vote for Coleman? Public opinion:

- Yes: 54% (6,080 votes)
- No: 46% (5,203 votes)

MN Challenged Ballot Collection

How the ballot collection was generated and harvested:

- Ballots photocopied and originals stored in a secure location.

- Copies scanned to PDF using auto-feeder flatbed scanner.

- Ballot was two-sided, with both sides scanned simultaneously.

- I wrote a simple web "crawler" that automatically downloaded all the files and extracted TIF images from PDF.

- A total of 6,737 ballots in the set.

- Examination of the TIF suggests that ballots were scanned at 300 dpi bitonal, and that lossy compression was never used.

- Hence, they form an ideal dataset for research purposes.

Minnesota Ballot Front and Back

Sloppy-But-Valid Marks
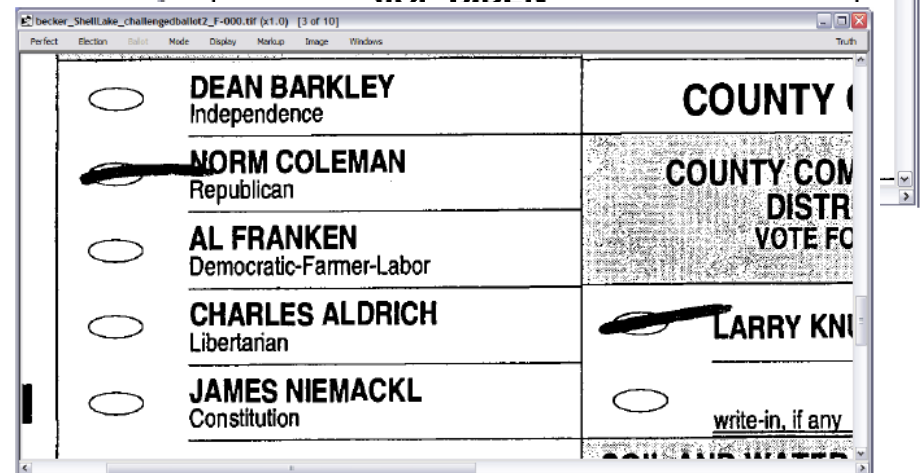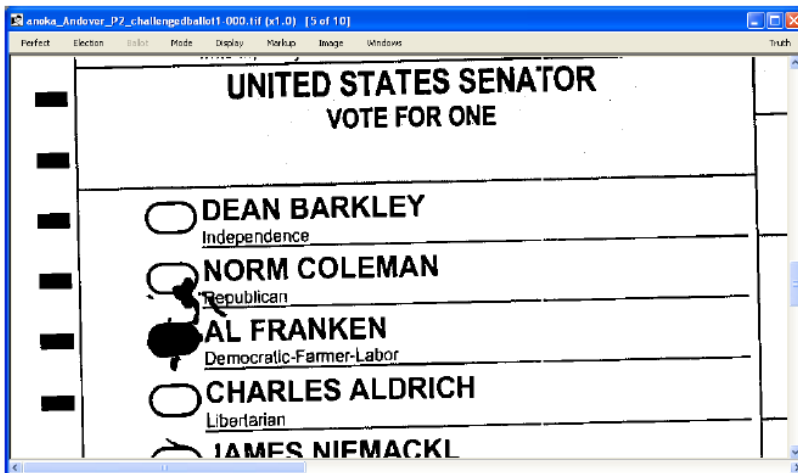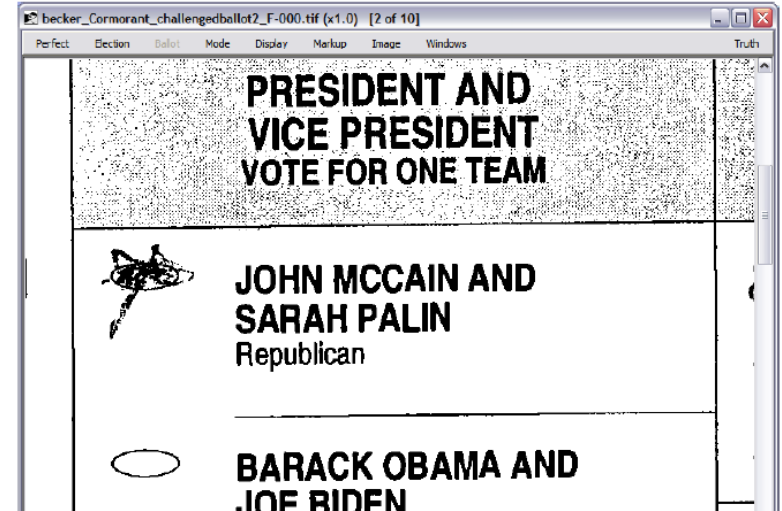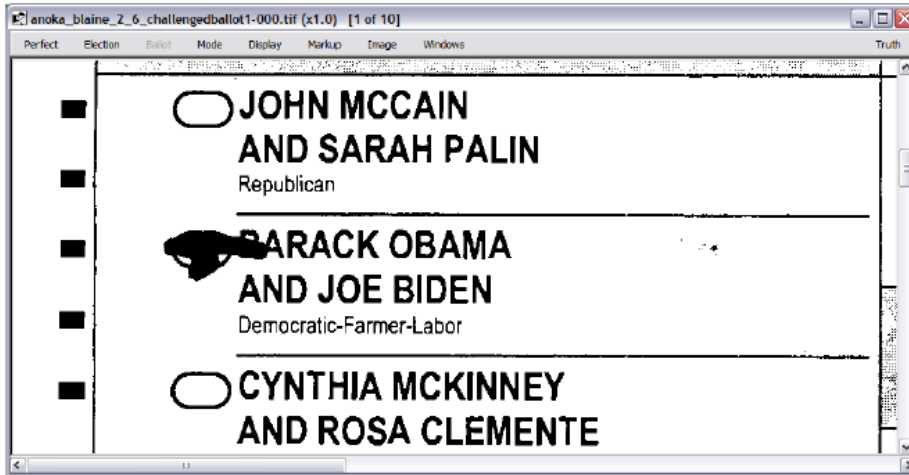
Non-Conforming Marking Styles
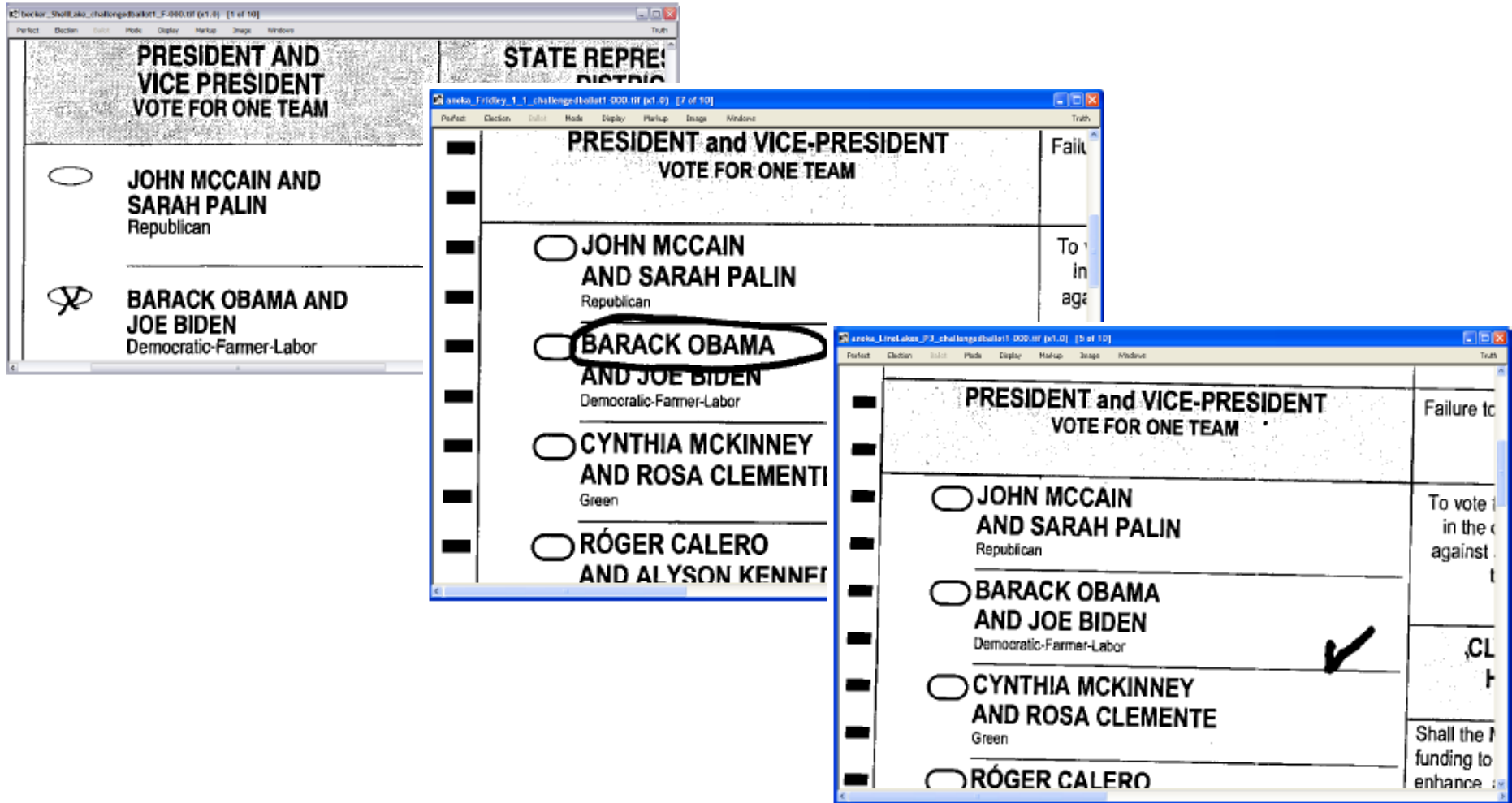
Attempts to Cancel a Vote
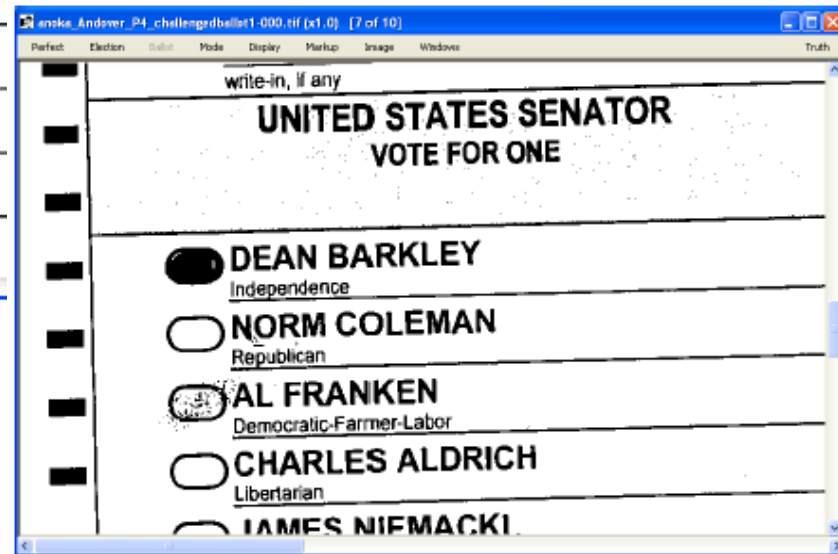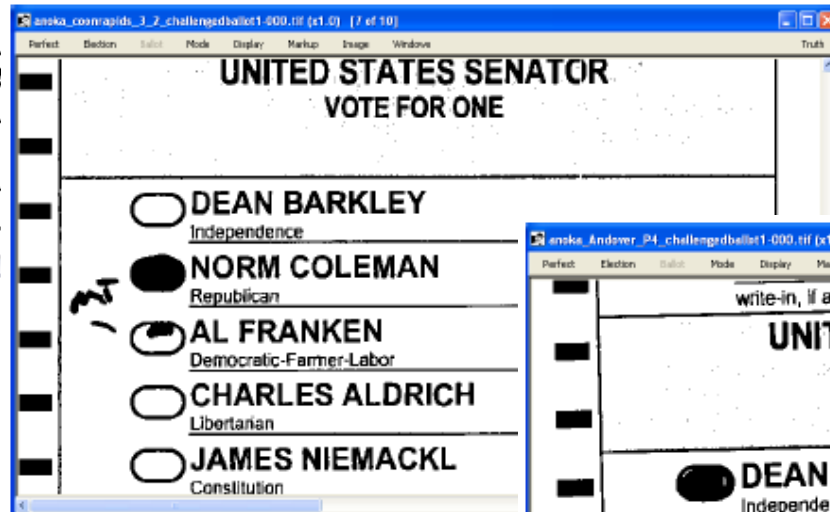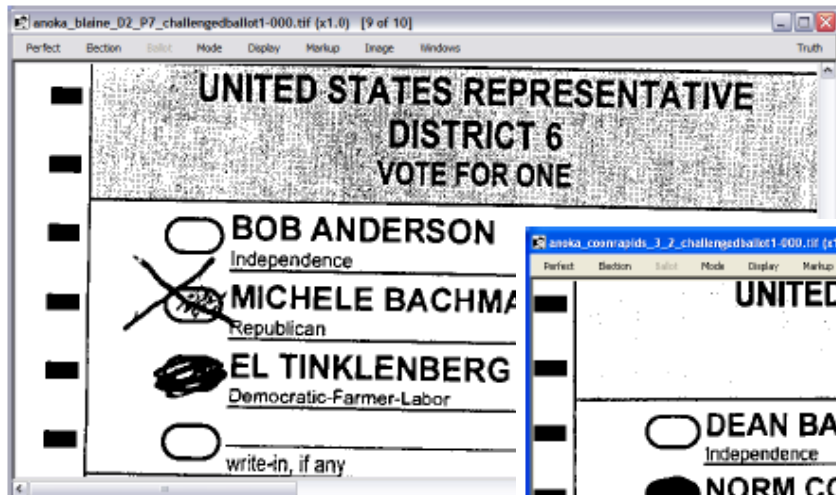
Votes that Look Cancelled
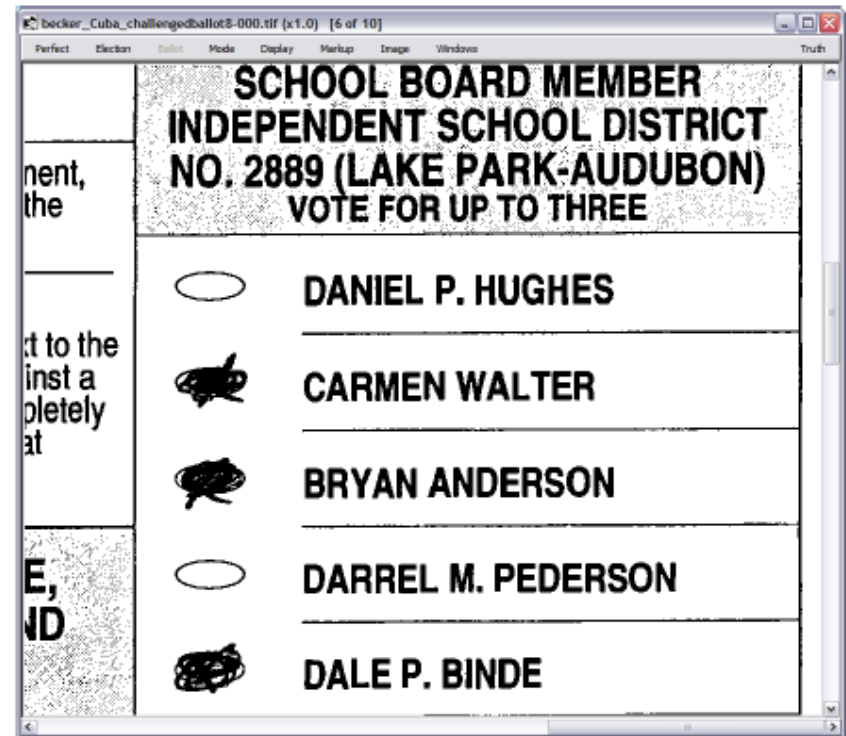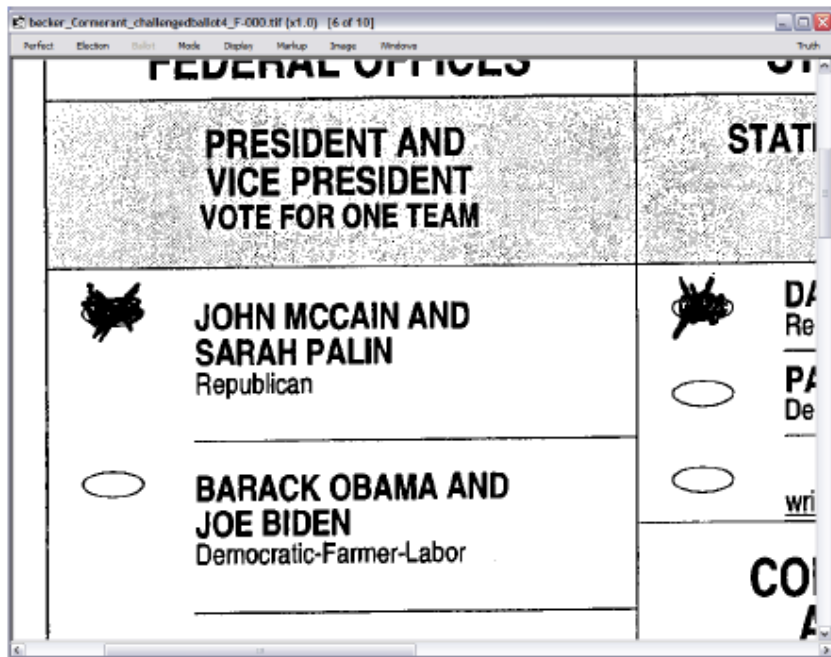
Stray Marks and Bleedthrough
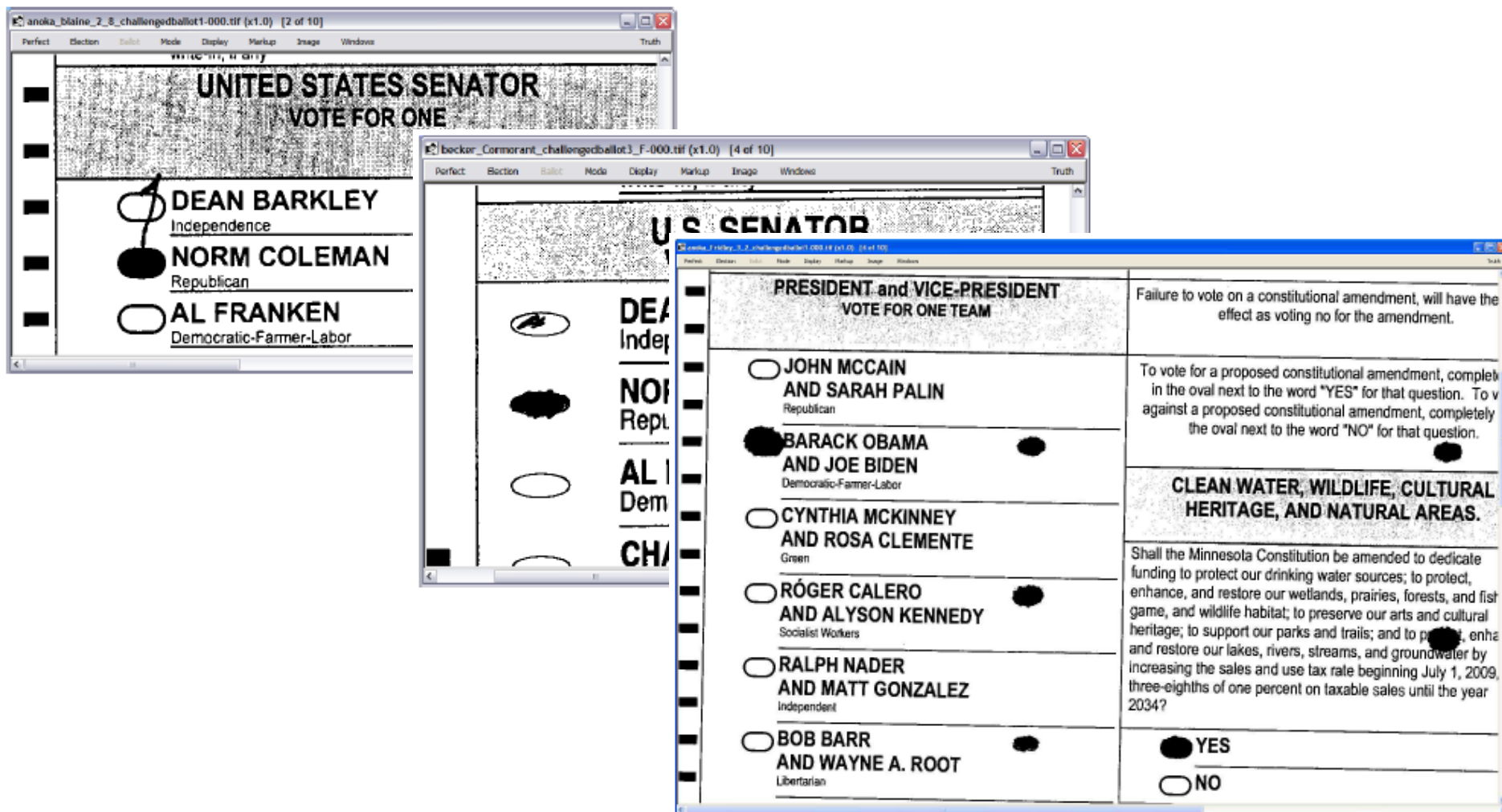
Invalidating Markings
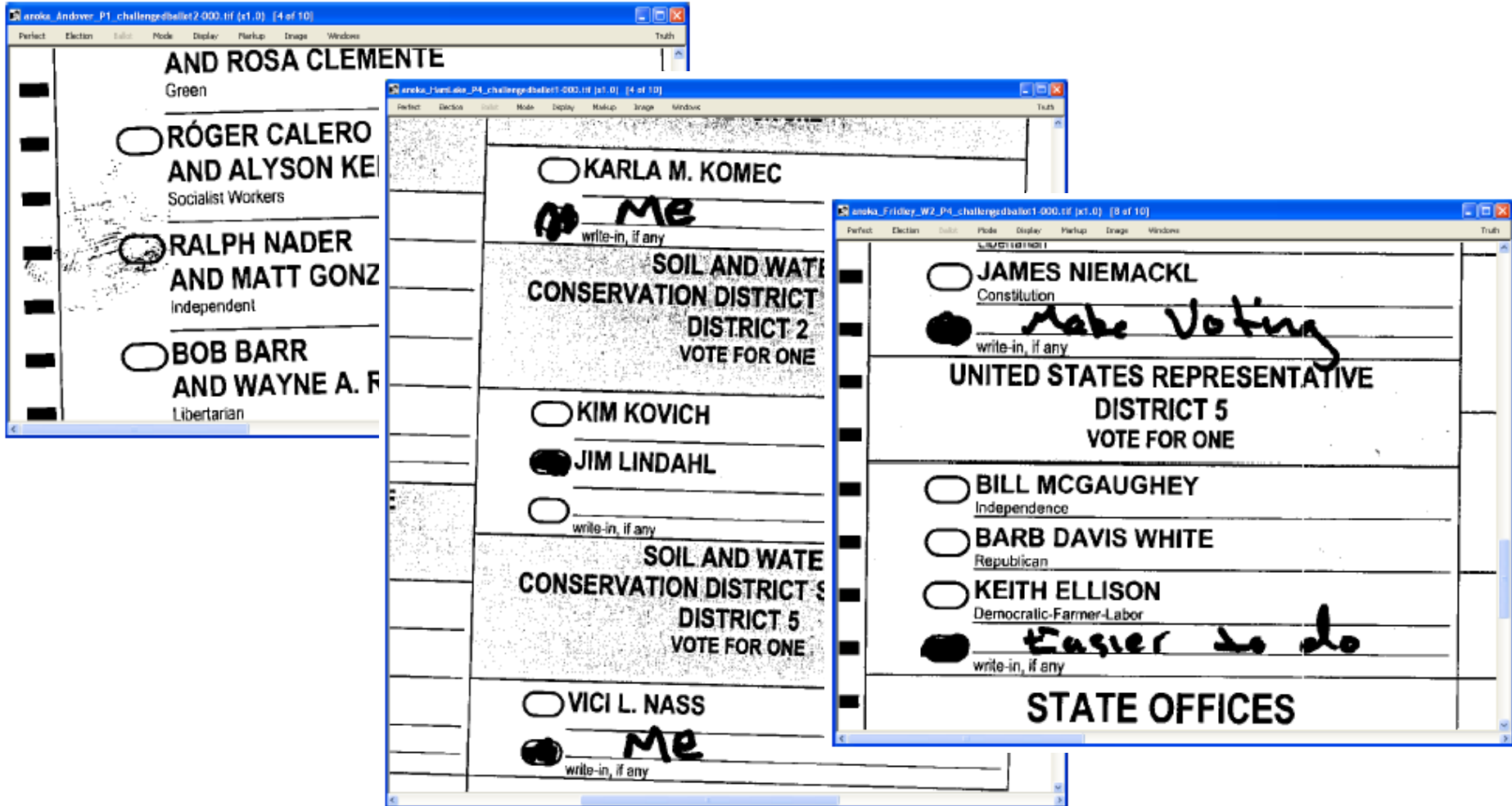
Another Example of Recent Interest

Note that ballots were counted by hand in this case.

See *Dealing with doubtful paper ballots in GB* :

http://www.electoralcommission.org.uk/__data/assets/pdf_file/0012/87699/UKPE-doubtfuls-booklet.pdf

Why isn't this an easy problem?

After all, ballots are just a simple type of form. We must read votes correctly, but we aren't expected to recognize write-ins.

Can't we just push up reject rate until accuracy reaches 100%?

Remember, we can't change rules in ways that violate the law. VOTER INTENT is the definition we must always follow.

To do this right, we must be prepared to:

- Reject any ballot that may contain "identifying marks."
- Recognize intent when mark is atypical or far from target.
- Accurately identify when a vote has been cancelled.

Status

- Interpretations collected from 8 test subjects, 980 ballot sides.

- All 6,737 ballots now online on DAE server (see URL for more details on server and its capabilities: joint work with Bart).

- Approach is a bit traditional, so far …

http://dae.cse.lehigh.edu/DAE/

"An Open Architecture for End-to-End Document Analysis Benchmarking," B. Lamiroy and D. Lopresti, Proceedings of the Eleventh International Conference on Document Analysis and Recognition (ICDAR 2011), September 2011, Beijing, China, pp. 42-47.

Very Close Indeed

That's all well and good. But what <u>really</u> happened in Minnesota?

Date	Description	Votes for Coleman	Votes for Franken
11/18/08	Initial State Canvassing Board meeting.	1,211,590	1,211,375
12/5/08	After hand recount, not including challenged ballots.	1,209,240	1,209,228
12/20/08	After review of challenged ballots by State Canvassing Board.	1,211,901	1,211,950
1/5/09	After counting of improperly rejected absentee ballots by order of Minnesota Supreme Court. This total was certified by State Canvassing Board.	1,212,206	1,212,431
4/13/09	After counting of improperly rejected absentee ballots by order of three-judge panel during the election contest.	1,212,317	1,212,629

http://www.sos.state.mn.us/elections-voting/2008-general-election-results/2008-state-recounts/

FAQ for Official Recount

WHEN DID THE RECOUNT BEGIN?

WHO SITS ON THE STATE CANVASSING BOARD?

HOW WILL THE RECOUNT WORK?

Every single vote cast for the U.S. Senate candidates will be recounted by hand.

The official recount is being conducted in approximately 110 locations throughout the state, generally in every county courthouse and in the city halls of major cities. In some locations more than one recount "station" will be used depending on the size of the jurisdiction.

The people doing the recounting are county election officials and election judges. Teams of recounters will examine each ballot and record the vote.

As many as four, perhaps even more, observers have been present as each ballot is recounted -- the election judge doing the recounting, representatives from each candidate's campaign, and any other interested parties. The recounts and canvassing board meetings are all open to the public.

https://www.mprnews.org/story/2008/11/06/recount_faq

WHAT ARE THE RECOUNT OFFICIALS LOOKING FOR?

The recounters are trying to determine the intent of the voter when they encounter problem ballots.

Most voters fill in the circle next to the candidate they choose. But sometimes an individual will put a check mark or an X next to a name. Others will circle a name. Ballots marked in that way cannot be scanned by the voting machines, so they wouldn't have been counted the first time around.

If a voter's intentions aren't clear by looking at a ballot, or if there is any objection to the decision being made by the election official by either one or both of the candidates' representatives, the ballots in dispute become "challenged" ballots that will go to the State Canvassing Board for review.

HOW LONG WILL THIS PROCESS TAKE?

HOW MUCH WILL THIS COST?

COULD THIS END UP BEING TAKEN TO COURT?

HOW WOULD THE CASE PROCEED?

WHY WOULD A CANDIDATE CHOOSE THIS PROCESS INSTEAD OF GOING WITH THE SECRETARY OF STATE'S RECOUNT?

More Details on Official Recount

POLITICS & POLICY

Judges' 'three votes' give Al Franken convincing win in Senate recount trial

By Jay Weiner | 04/13/09   ✉ Email   f Share   🐦 Tweet   🖨 Print

When this 2008 U.S. Senate race is finally over, when all the appeals are exhausted, when its history is written – based on what we know today -- it will be said that Al Franken won the election and the seat of the junior senator from Minnesota by 312 votes.

But that will not be exactly correct.

After Monday's long-awaited final legal ruling (PDF), add three more votes to Franken's tally: those of Judges Elizabeth Hayden, Kurt Marben and Denise Reilly.

Technically, the three-judge panel that oversaw a seven-week-long trial that generated 19,181 pages of legal filings "voted" against Norm Coleman in their unanimous 56-page opinion, with another 12 pages of exhibits.

It was Coleman's case to prove, and now he'll get another chance when he appeals today's ruling to the Minnesota Supreme Court within 10 days.

https://www.minnpost.com/politics-policy/2009/04/judges%E2%80%99-%E2%80%98three-votes%E2%80%99-give-al-franken-convincing-win-senate-recount-trial

Deciphering Official Recount Results

Freely available as MS Excel file. But note ambiguity: work is needed to translate this into decisions on a ballot-by-ballot basis.

What Can Be Learned Here?

Counting ballots not just an <u>abstract</u> pattern recognition problem:

- A real task defined by pre-determined laws and processes.

- Important to society (not just labeling "cute cat" photos).

- Inherently political, but designed to be as fair as possible.

- Expressed in terms of <u>human</u> <u>interpretation</u>.

- Ambiguity is utterly inherent (real world is messy).

- "<u>Noisy</u> <u>labeling</u>" is utterly inherent.

- AI (pattern recognition) <u>can</u> <u>and</u> <u>must</u> do better.

In other words, this is a perfect problem to study for those of us who want our research to have an impact in the real world.

A Sad Epilogue ...

Adapting the Turing Test for Declaring a Problem Solved

An interesting thought experiment, given the demand for algorithms that can perform at human levels when users are free to act in ways that confound the system.

When is a Problem Solved?

The Turing Test:

> "A problem is solved if there is a method which has been widely publicized and documented and freely available to the community which generates output for a given input that a human judge cannot reliably distinguish from the output of a human expert."

Differs significantly from employing ground-truth provided by a human expert in advance.

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

The Imitation Game

VOL. LIX. No. 236.]   [October, 1950

MIND

A QUARTERLY REVIEW
OF
PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND
INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?'
This should begin with definitions of the meaning of the terms
'machine' and 'think'. The definitions might be framed so as to
reflect so far as possible the normal use of the words, but this
attitude is dangerous. If the meaning of the words 'machine'
and 'think' are to be found by examining how they are commonly
used it is difficult to escape the conclusion that the meaning
and the answer to the question, 'Can machines think?' is to be
sought in a statistical survey such as a Gallup poll. But this is
absurd. Instead of attempting such a definition I shall replace the
question by another, which is closely related to it and is expressed
in relatively unambiguous words.

The new form of the problem can be described in terms of
a game which we call the 'imitation game'. It is played with
three people, a man (A), a woman (B), and an interrogator (C) who
may be of either sex. The interrogator stays in a room apart
from the other two. The object of the game for the interrogator
is to determine which of the other two is the man and which is
the woman. He knows them by labels X and Y, and at the end
of the game he says either 'X is A and Y is B' or 'X is B and Y
is A'. The interrogator is allowed to put questions to A and B
thus:

C: Will X please tell me the length of his or her hair?
Now suppose X is actually A, then A must answer. It is A's

28   433

---

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?'
This should begin with definitions of the meaning of the terms
'machine' and 'think'. The definitions might be framed so as to

---

The new form of the problem can be described in terms of
a game which we call the 'imitation game'. It is played with
three people, a man (A), a woman (B), and an interrogator (C) who
may be of either sex. The interrogator stays in a room apart
from the other two. The object of the game for the interrogator
is to determine which of the other two is the man and which is
the woman. He knows them by labels X and Y, and at the end

We now ask the question, 'What will happen when a machine
takes the part of A in this game?' Will the interrogator decide
wrongly as often when the game is played like this as he does
when the game is played between a man and a woman? These
questions replace our original, 'Can machines think?'

A. M. Turing, "Computing Machinery and Intelligence,"
Mind, vol. 59, no. 236, October 1950, pp. 433-460.

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, *Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012)*, March 2012, Gold Coast, Australia, 5 pages.

The Turing Test

(A) Man pretending to be woman

(B) Woman trying to help interrogator

(C) Interrogator trying to make right guess

$SuccessRate_1$

(A) Machine pretending to be woman

(B) Woman trying to help interrogator

(C) Interrogator trying to make right guess
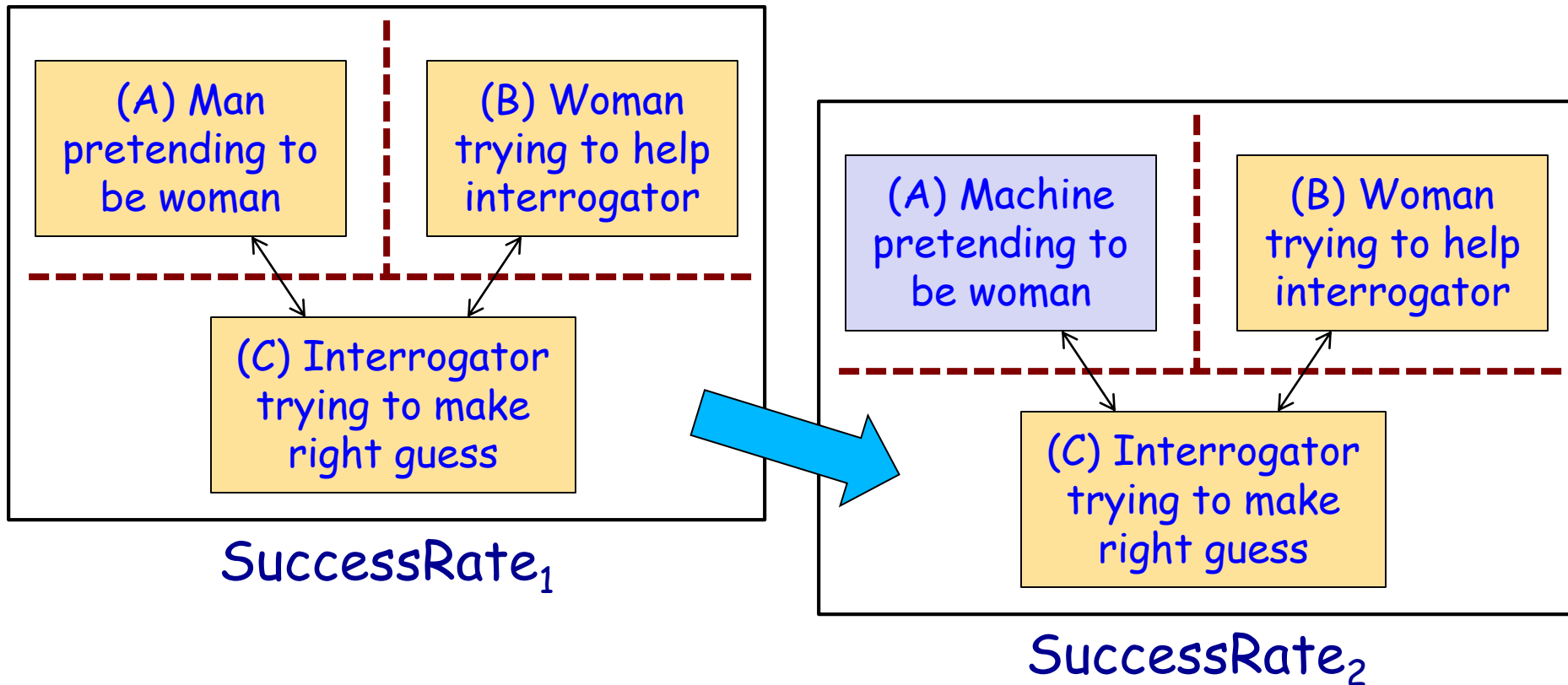
$SuccessRate_2$

Is $SuccessRate_2 \approx SuccessRate_1$?

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

The Turing Test

The Turing Test is an elegantly simple idea, so it should be simple to implement, right?

(A) Machine performing some task		(B) Human performing same task
	(C) Interrogator trying to make right guess	

Is SuccessRate no better than random chance ?

- Note this differs from Turing's original formulation.

- When considering a real implementation, other, more serious complications arise.

Long Bet*

"By 2029 no computer – or 'machine intelligence' – will have passed the Turing Test."

PREDICTOR:
Mitchell Kapor
CHALLENGER:
Ray Kurzweil
STAKES: $20,000

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

Long Bet Rules

Turing was nonspecific about how to administer his Test, but concreteness is needed when $20,000 is at stake.

- Each of three Turing Test judges is to conduct an online interview ("chat") with each of four human players as well as the machine for two hours.

- At the end of these interviews, the judges indicate whether or not each candidate is human and also rank them from "least human" to "most human."

- The machine is said to pass the Turing Test if it fools two or more judges and if its median rank is equal to or greater than at least two of the human players.

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.
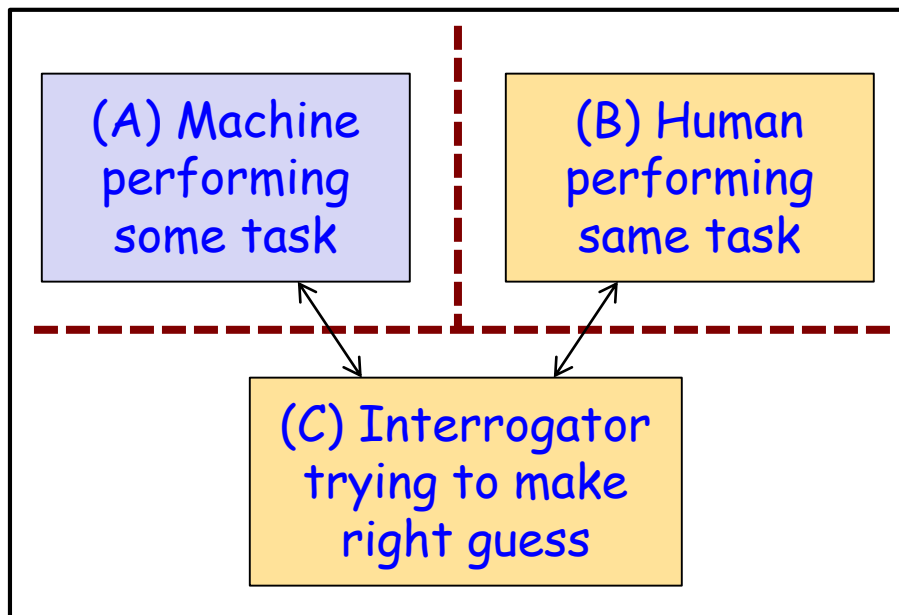
Adapting the Turing Test

The Long Bet is a one-time event with a significant amount of prize money involved. As a result, it makes sense to employ a heavy-weight protocol for the test.

How can the Turing Test be applied in document analysis?

- What are the essential qualities to preserve?

- What can be dispensed with, or at least simplified?

- When implemented, how would the test "look"?

- When might such a test be appropriate?

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

Properties to Preserve #1

Human judgment is applied to determine a simple machine/human distinction and nothing more complex than this. Automated evaluation (i.e., a computation to determine how "similar" a machine output is to some predefined human "ground truth") is ruled out.
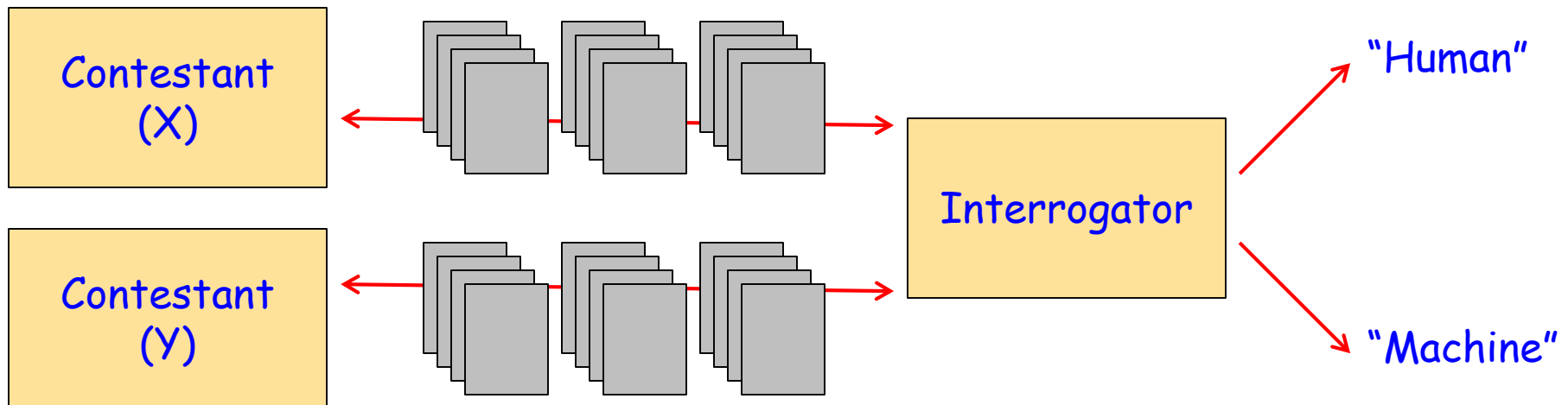
"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

Properties to Preserve #2

A judge may ask any number of questions before making a determination. A "question" here is a challenge that requires a response from the player. For document analysis applications, this will normally consist of a page image to be processed in some way.
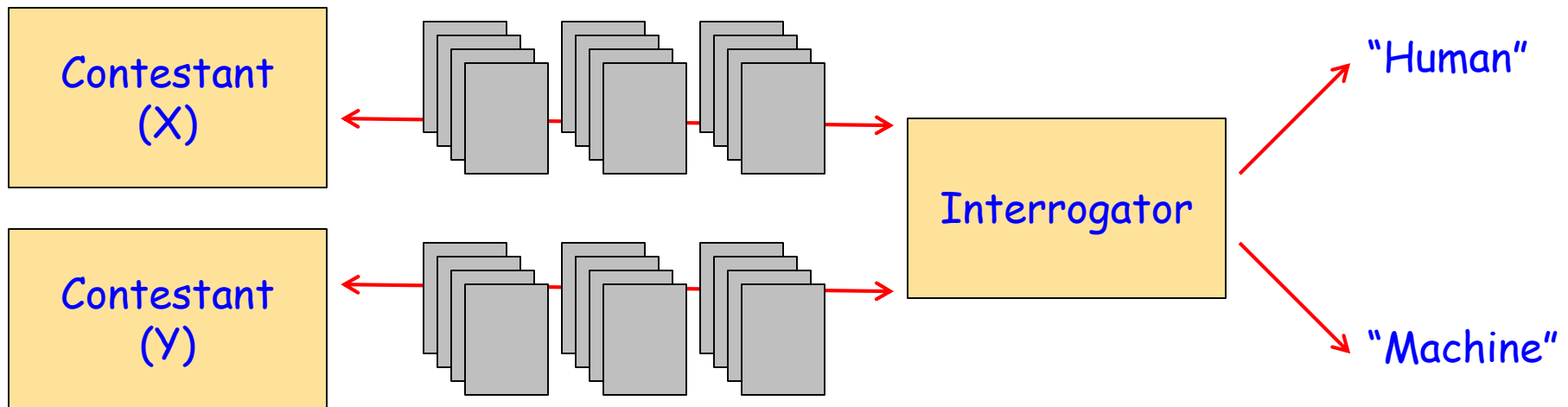
"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

Properties to Preserve #3

The judge decides which questions to use, and is free to conduct the questioning of the players without constraint on the choice, sequence, and number of questions.
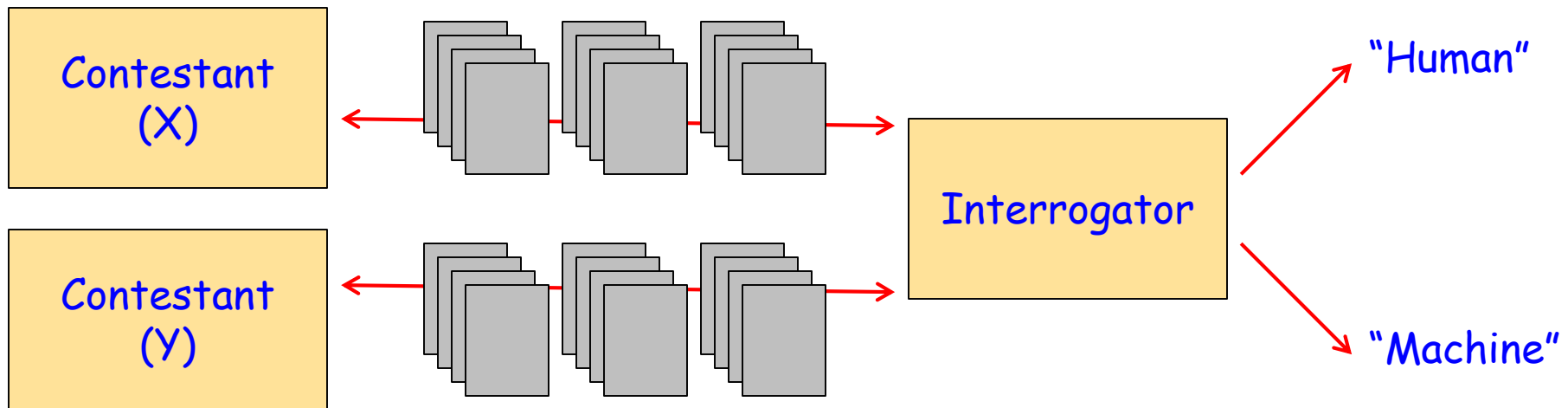
"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

Properties to Preserve #4

A series of such evaluations, with anyone being allowed to volunteer to serve as judge or as the human player, is conducted before declaring a problem "solved" (if/when the success rates of the best-performing judges are statistically no better than random).

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.
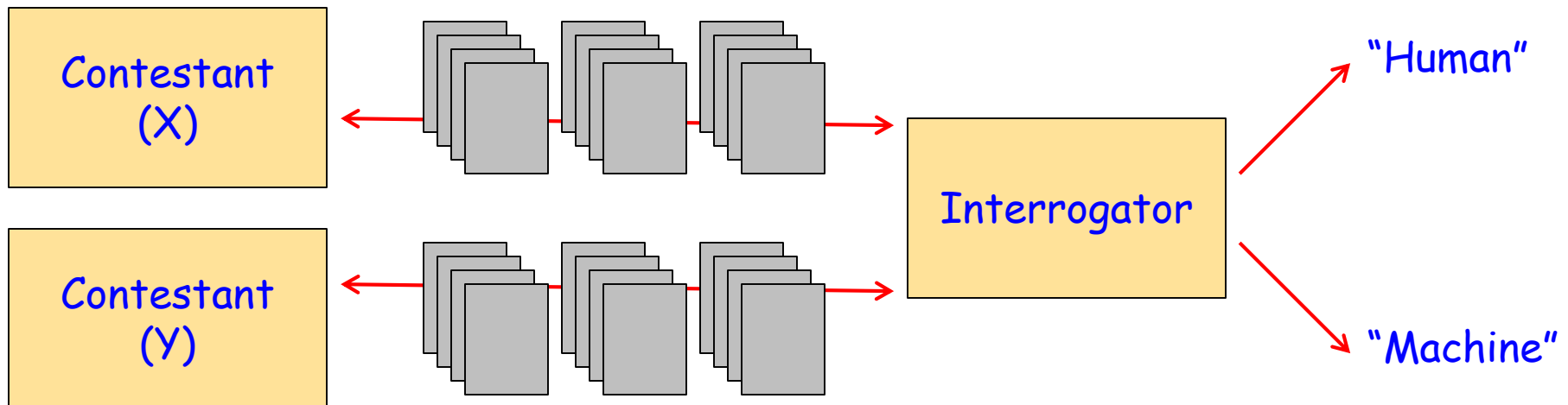
Properties to Adapt

Some aspects of Turing's original Test must be updated:

- The judge and players do not interact via a natural language question-and-answer process. Instead, they employ a graphical user interface which supports the upload of image files and visual inspection of results.

- The domain of discourse is no longer open-ended. Note that this replaces Turing's original question "Can machines think?" with our "Is this problem solved?"

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

GUI from Judge's Perspective

Task is: Logo Detection   **Current Challenge is #12**

Pre-defined Challenge Library

Create New Challenge

File name	Upload

Submit to Player A	Submit to Player B

Responses

Determination: | A human, B machine | | A machine, B human |

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.
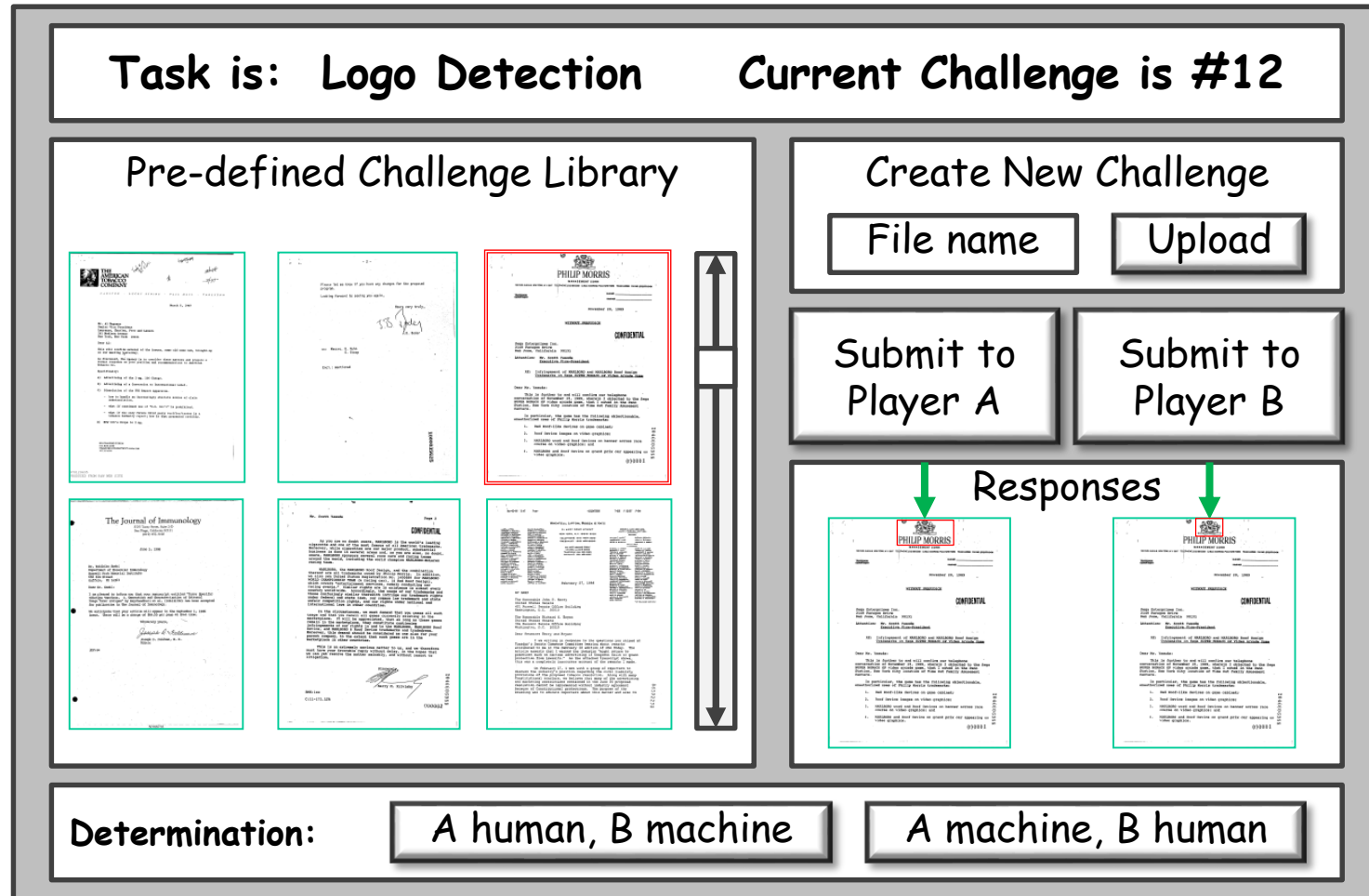
Other Considerations

Additional details to be addressed, some easy, some hard:

- Anyone should be permitted to volunteer at any point in time to serve as the judge or the human player.

- The need to pair a judge with a human player can be addressed through crowdsourcing (e.g., using micro-payments to recruit subjects like Mechanical Turk).

- How can we eliminate out-of-scope querying / collusion?

- Which problems are appropriate to test this way? (Avoid tedious tasks where machines are "too good.")

- How can learning (by human, by machine) be included?

"Adapting the Turing Test for Declaring Document Analysis Problems Solved," D. Lopresti and G. Nagy, Proceedings of the Tenth IAPR International Workshop on Document Analysis Systems (DAS 2012), March 2012, Gold Coast, Australia, 5 pages.

Hopefully this has given you some points to think about … if you're interested in collaborating to turn MN ballots into a community resource for exploring interesting and important "noisy labeling" problem, let me know!

Thank you! Merci!!