From Purple Prose to Machine-Checkable Proofs:

Levels of rigor in family history tools

Dr. Luther A. Tychonievich, Ph.D.
Dept. Computer Science, University of Virginia
TSC Coordinator, Family History Information Standards Organisation

These slides: `http://www.cs.virginia.edu/luther/RT2015.pdf`
Me: `ltychonievich@fhiso.org` or `luther@virginia.edu`

Outline

- Two definitions of “proof”
- Can genealogy be rigorous?
- Can genealogy tools be rigorous?
- An upper bound on possible rigor
- Achieving rigor, and beyond

Two Definitions of Proof

- Persuasive writing intended to convince someone that an idea is beyond rational doubt
- A set of derivations that reduce a non-obvious claim to a set of other claims and derivation rules

Not Proof

- Purple prose is language that conceals a lack of substance behind a superfluity of description
- Structure can be that detail…
  - Try a genealogy search for Odin
- Potential for misuse ≠ bad

Outline

- Two definitions of “proof”
- Can genealogy be rigorous?
- Can genealogy tools be rigorous?
- An upper bound on possible rigor
- Achieving rigor, and beyond

Reasons not to trust genealogy

- Amateurs and fee-for-service
  - ...the two least-trusted sources
- Individuals don't yield to statistics
- Linked generations compound error
- Genes and probability offer little hope

Compound Error

- Suppose each parent-child link correct with 95% confidence
  - Great-grandparents = 14 links
  - $0.95^{14} < 49\%$ chance all correct
- Doubling per-link confidence
  - Doubles links we can trust
  - Gives **one** extra generation

Will genetics provide rigor?

- Makes statistical claims about overlap of ancestry of living people
- Illegitimacy and inbreeding
- Biological vs Social parent
- (see philogenetic tree research)

Can probability provide rigor?

- Probabilistic reasoning requires some subset of
  - Priors
  - Experiments or Ground truth
  - Well-defined populations
  - Independent distributions
- Not a silver bullet...

Outline

- Two definitions of “proof”
- Can genealogy be rigorous?
- **Can genealogy tools be rigorous?**
- An upper bound on possible rigor
- Achieving rigor, and beyond

Pop Quiz

- Which of the following is hardest?
  (1) Finding the right source
  (2) Citing the source right
  (3) Drawing the right conclusions
  (4) Communicating your reasoning
- Using tool _____, which is hardest?

Analogy: Proof-Carrying Code

- To prove code works is provably hard
  - ...or impossible (Rice's Theorem)
- To communicate a proof is easy
  - ...if you use the right tool

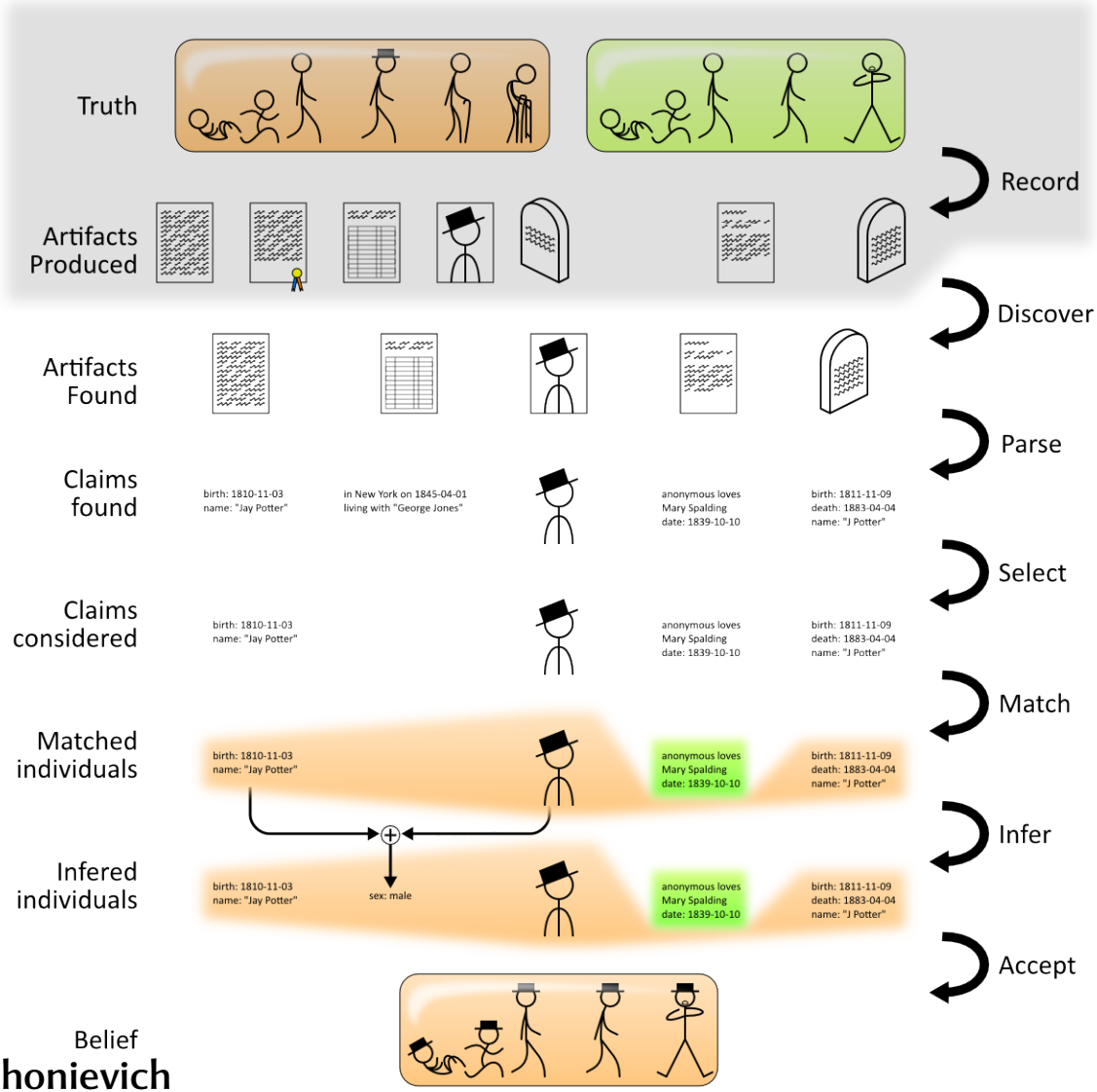
- Can we make the right tool for family history?

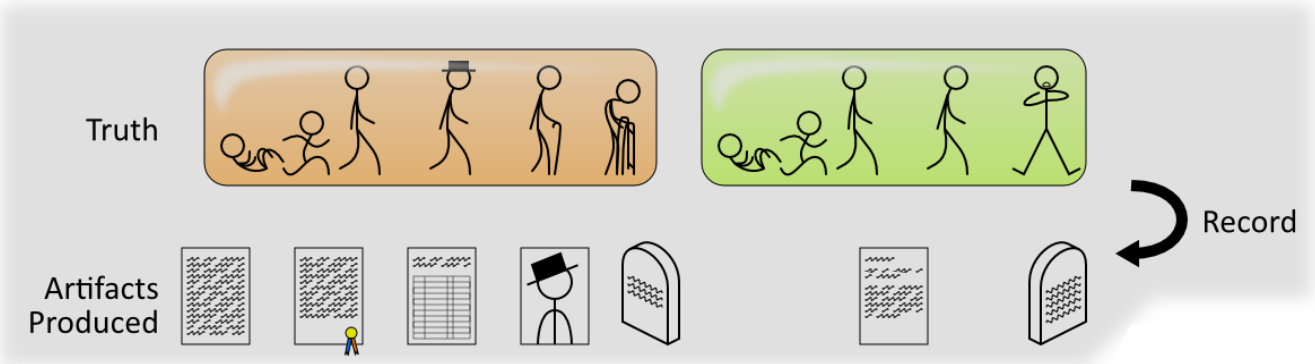
Outline

- Two definitions of “proof”
- Can genealogy be rigorous?
- Can genealogy tools be rigorous?
- **An upper bound on possible rigor**
- Achieving rigor, and beyond

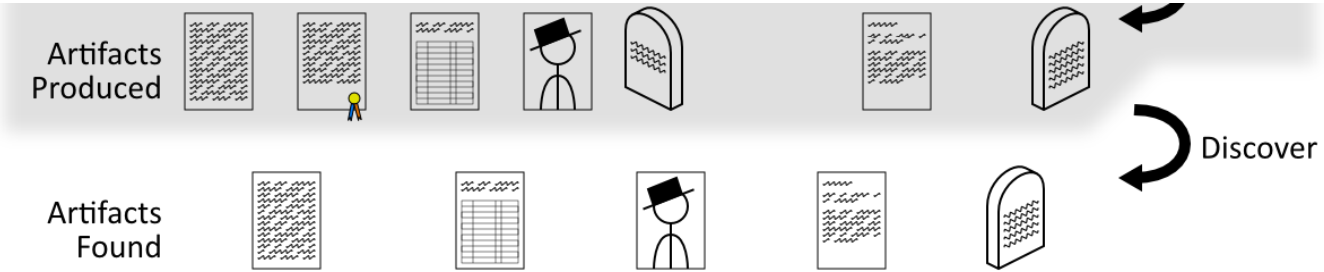
On the next slide we'll see...

- A recorder's perspective of the logical flow of genealogy
- **Not** a suggested research work-flow

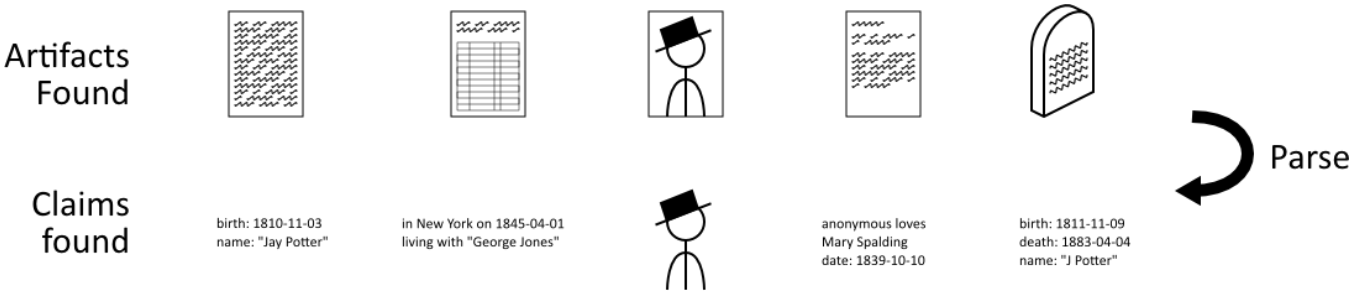

Truth
Artifacts Produced
Record
Artifacts Found
Discover
Claims found
birth: 1810-11-03
name: "Jay Potter"
in New York on 1845-04-01
living with "George Jones"
anonymous loves
Mary Spalding
date: 1839-10-10
birth: 1811-11-09
death: 1883-04-04
name: "J Potter"
Parse
Claims considered
Select
Matched individuals
Match
Infered individuals
sex: male
Infer
Belief
Accept



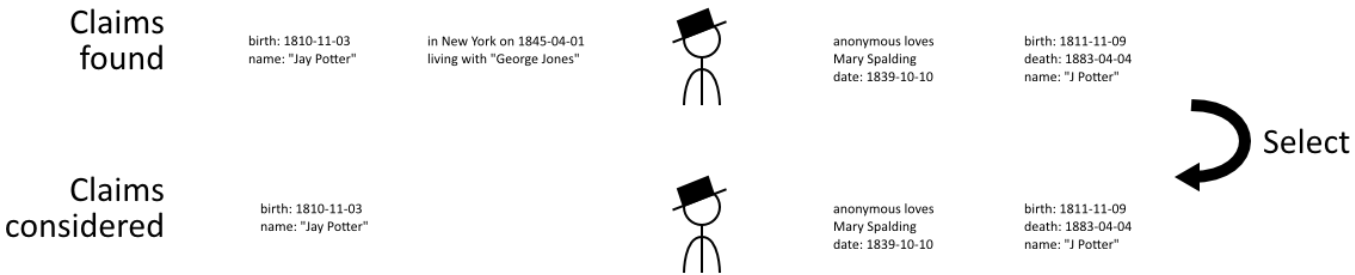
- History happened, creating artifacts (documents, monuments, memories…)
- **Not** part of research; should **not** appear in tools



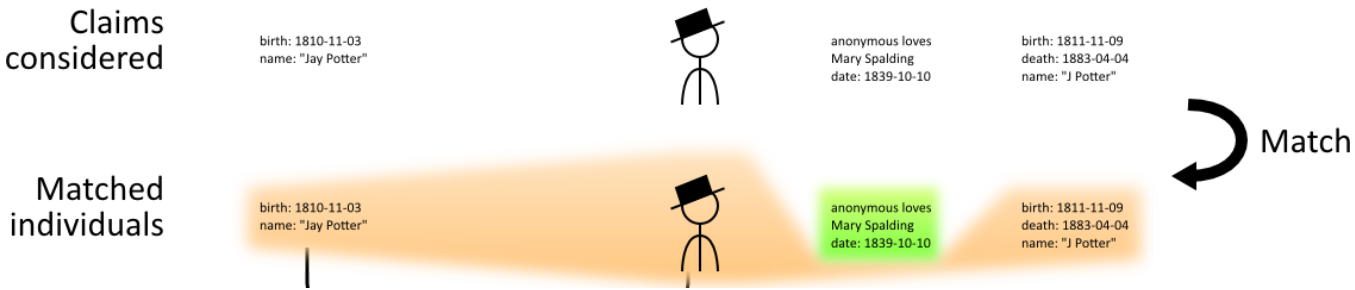
- Step 1: discover artifacts
- Digitization helps us all
- Citations are getting better…
- Logs (negative evidence) need work



- Many tiers: shapes → glyphs → words → meaning → **claims**
- Single version = data pollution
- Transcription community much better at this than us



- We don't consciously consider every claim we see
- How do you detect, let alone store, subconscious selection?
- Computer could notice for us...



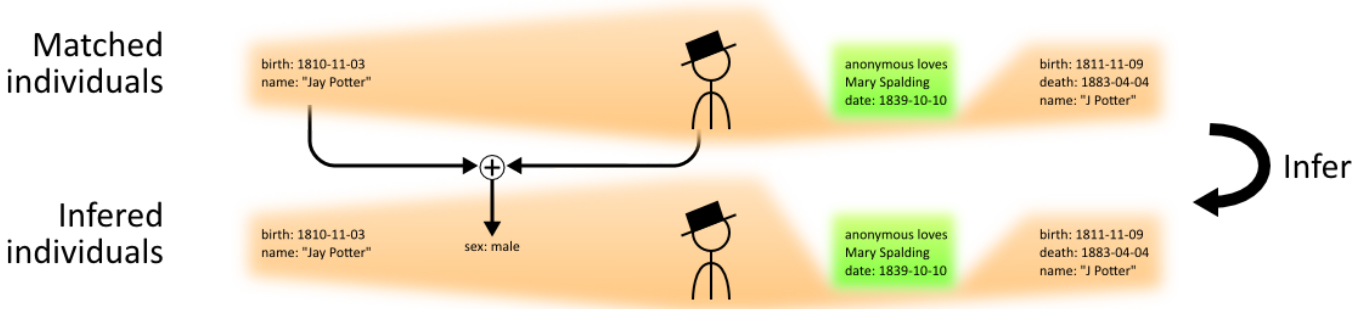
- "Is that *my* John?"
- Turns claims into people
- The crux of genealogy; most errors I've seen happen here

- Deserves a second slide...

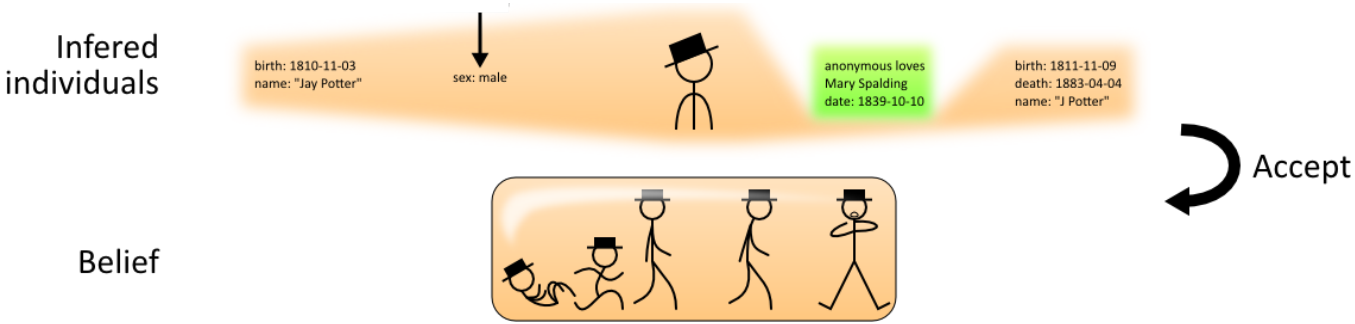
- We need to
  1. record matches (and non-matches)
  2. allow changes
  3. record reason for each match
  4. embrace ambiguity
- More tools gaining 1, a few have 2; 3 is unusual, 4 yet to appear

- Deserves a third slide...

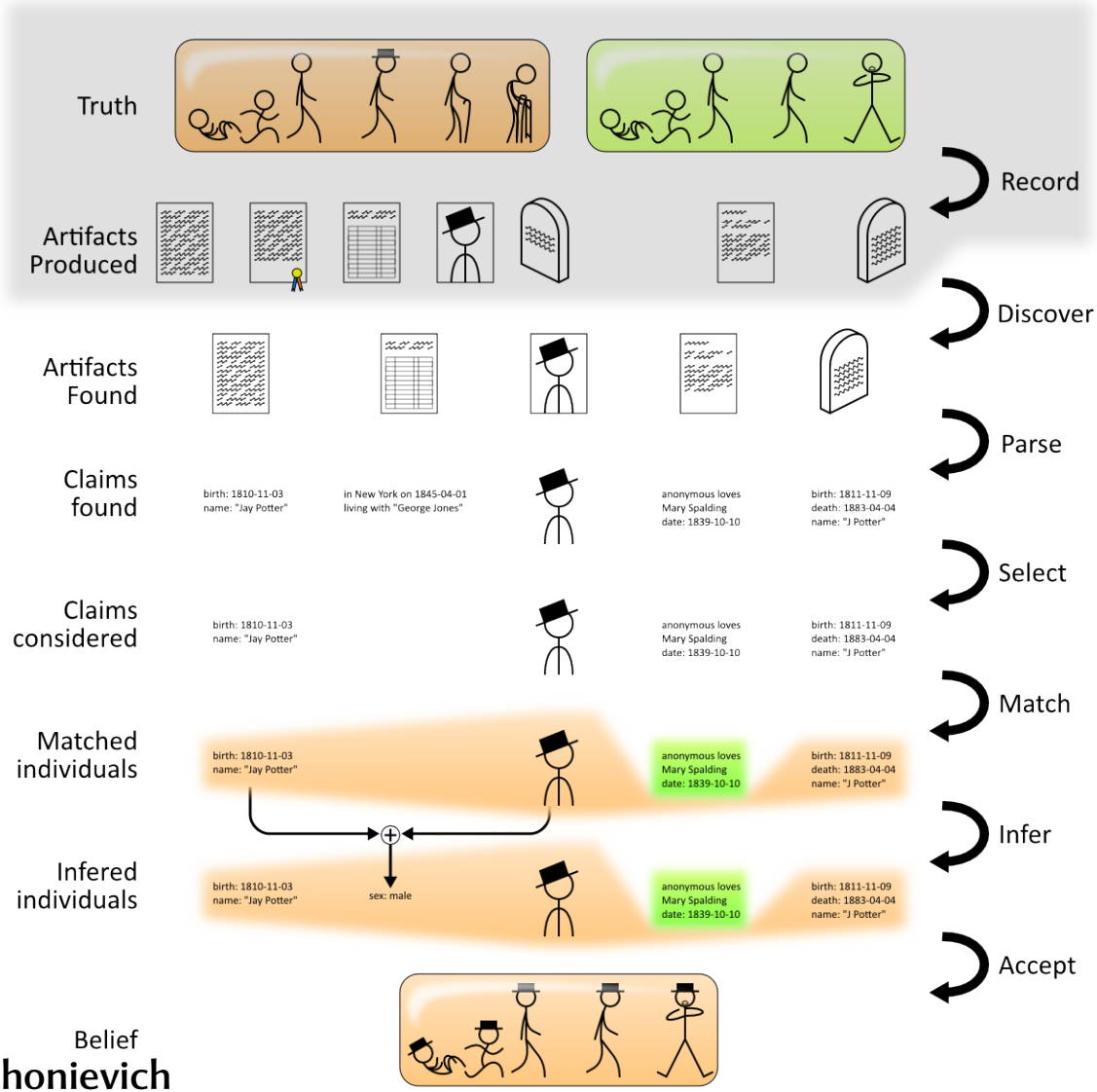
- Match algebra:
  - A = B
  - A ≠ B
  - {A, B, C, D, ...} ≥ 2 individuals
- Matches of any noun (person, place, event, etc.)
- Often evidence for A = B *and* A ≠ B



- The “hidden” step today: from “name: Jane” we infer “gender: female” but don’t record that we made an inference.
- Data simple, missing in our tools



- The final belief is logically uninteresting, but socially vital
- Just one is a problem
- Many tools store **only** beliefs and artefacts found


Truth
Artifacts Produced
Record
Artifacts Found
Discover
Claims found
birth: 1810-11-03
name: "Jay Potter"
in New York on 1845-04-01
living with "George Jones"
anonymous loves
Mary Spalding
date: 1839-10-10
birth: 1811-11-09
death: 1883-04-04
name: "J Potter"
Parse
Claims considered
Select
Matched individuals
Match
Infered individuals
sex: male
Infer
Belief
Accept

Outline

- Two definitions of “proof”
- Can genealogy be rigorous?
- Can genealogy tools be rigorous?
- An upper bound on possible rigor
- **Achieving rigor, and beyond**

Do we want rigorous tools?

- Depends on the audience
- Option of rigor seems good
- Can we make a “proof-carrying GEDCOM”?

To achieve rigor:

- Embrace ambiguity
  - 2+ conflicting versions of each step
- Adopt best-practice transcription
- Match algebra
- Support reasons for matches
- Add inferences explicitly

...and beyond

- Track the "select" step
- Inference rules could allow
  - translation-free rationales
  - statistically-sound probability
- Mid-research belief: $A = B = C \neq A$
- Time-varying rigor could be revealing data

Outline

- Two definitions of “proof”
- Can genealogy be rigorous?
- Can genealogy tools be rigorous?
- An upper bound on possible rigor
- Achieving rigor, and beyond
- Extra Slides

More on Inferences

- Inference step:
  rule + antecedents = consequent
- Antecedents, consequent: claims
- **source** of consequent: the inference
- **support** of conseq.: the antecedents
- **rationale** of consequent: the rule
- Support may be enough…

Inference “rule”s

- A predicate over antecedents
- A claim built from antecedents
- Could be probabilistic (“trend” not “rule”)
- Consequent could be claim, match, non-match, etc.
- Encodes “why” as pure data

More on Probability

- Probability a claim is true: 0 or 1
- Probability a rule's consequent is true: $\frac{\text{validated}}{\text{validated} + \text{invalidated}}$ (approximate)
- Is "83% chance Jane is John's mother" useful?
- What about "99.7% chance either Jane, Sue, or Anna is John's mother"?

Outline

- Two definitions of “proof”
- Can genealogy be rigorous?
- Can genealogy tools be rigorous?
- An upper bound on possible rigor
- Achieving rigor, and beyond
- Extra Slides

Questions?

These slides: `http://www.cs.virginia.edu/luther/RT2015.pdf`

Me: `ltychonievich@fhiso.org` **or** `luther@virginia.edu`