Challenges and Opportunities in eDiscovery and Information Governance:
Not Everything is Big Data

Jeremy Pickens

opentext™ | Legal Tech

Overview

1. Legal Information Retrieval (ESI)

    1. eDisclosure / eDiscovery

    2. Trial Preparation

    3. Governance

2. Principles of Legal Information Retrieval (ESI)

    1. Collection Completeness

    2. The Importance of Process

    3. Encouraging Curiosity

    4. Proprioception

    5. Trajectory

eDisclosure (eDiscovery): The process of identifying, preserving, collecting, processing, searching, reviewing, and producing Electronically Stored Information that may be relevant to a civil, criminal, or regulatory matter.

–The Grossman-Cormack Glossary of Technology Assisted Review, 2013

Electronically Stored Information (ESI)

Engineering designs (CAD/CAM drawings)
PDF / MS Word / Google docs
OneNote / Evernote records
Slack and Teams chats
Excel / Spreadsheets
Social media posts
Database dumps
System logs
Images
Email

Etc.

..in multiple languages!

Electronically Stored Information (ESI)

Typical Collection Size
50 thousand - 10 million documents

Find (almost) all relevant documents

Find the documents that make your case

**Find problematic documents
before they become problematic**

Find (almost) all relevant documents

EXAMPLES

- Large database company suing a large search engine over alleged nature of coffee-like programming language API usage

- Car manufacturer allegedly cheating on emissions tests

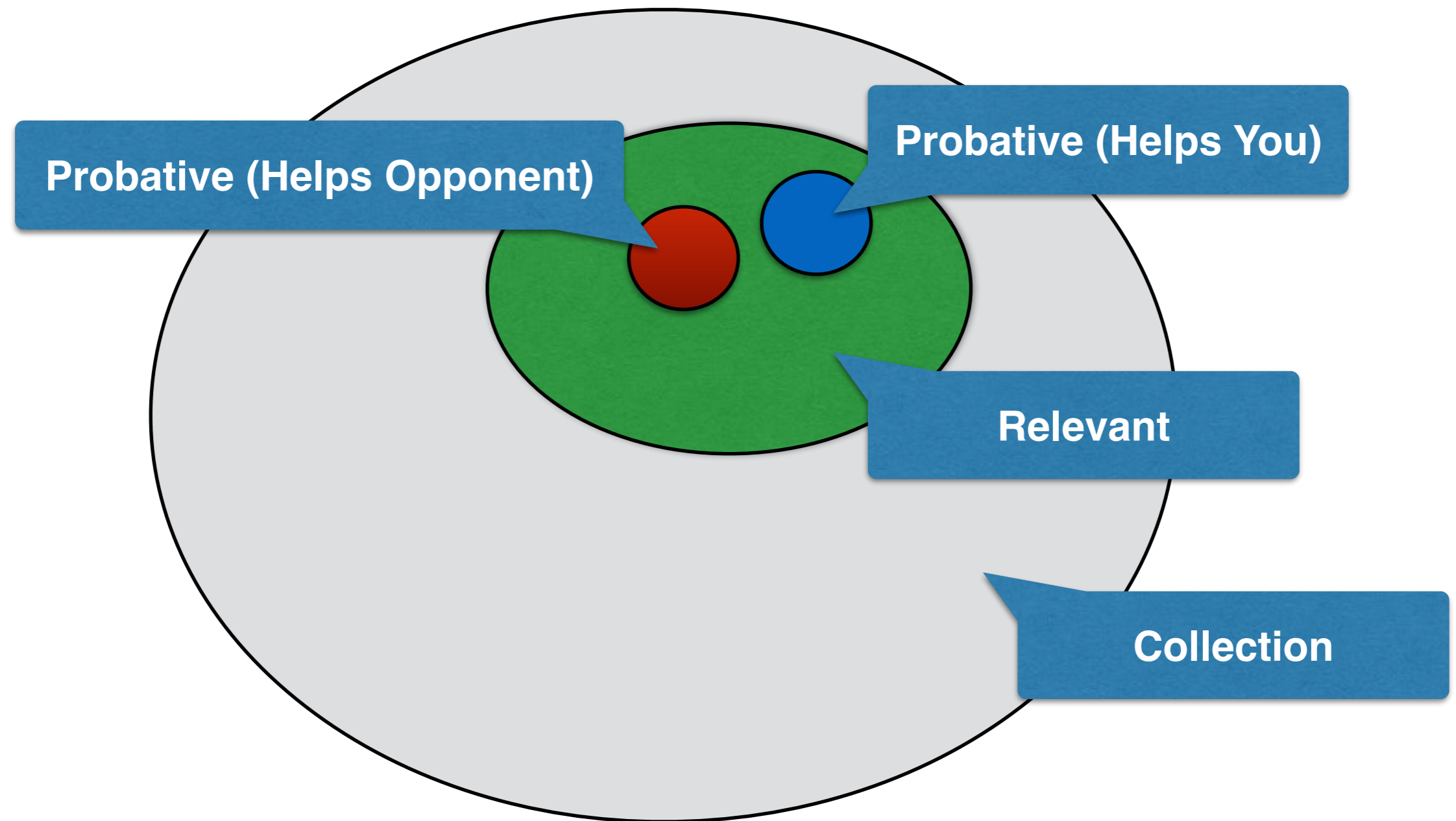- Milošević alleged war crimes (International Court of Justice)

EDISCLOSURE / EDISCOVERY

- ≥80% recall is required

- Every case (information need) is a cold start

- Machine learning models are ephemeral: The more a model is used, the less useful it becomes

Find the documents that make your case

PROBATIVE DOCUMENTS

Set #1 (helps you)		
Email from Frank saying: "I broke into the house and stole the fancy television"		

Set #1 (helps you)	Set #2 (helps you)	
Email from Frank saying: "I broke into the house and stole the fancy television"	Victim posted picture of his fancy television online. Frank is in that forum.	
	Frank has a sales receipt for lock-picking tools	
	Frank's cell phone pinged towers in burglary neighborhood	
	Sale of a fancy television on eBay by Frank	

Set #1 (helps you)	Set #2 (helps you)	Set #3 (helps opponent)
Email from Frank saying: "I broke into the house and stole the fancy television"	Victim posted picture of his fancy television online. Frank is in that forum.	Police report showing that Frank's house was broken into recently.
	Frank has a sales receipt for lock-picking tools	Receipt for local college course on physical security
	Frank's cell phone pinged towers in burglary neighborhood	Receipt for apartment rental deposit in neighborhood
	Sale of a fancy television on eBay by Frank	Real estate listing of Frank's old vs new apartment size

**Find problematic documents
before they become problematic**

REGULATORY COMPLIANCE

- General Data Protection Regulation (GDPR)
- UK Bribery Act
- Sexual Harassment
- Environmental Protection
- Insider Trading

EXAMPLE #1

Senior Yen Trader: the whole HF world will be kissing you instead of calling me if libor move lower

Yen Trader 1: ok, i will move the curve down 1bp maybe more if I can

Senior Yen Trader: maybe after tomorrow fixing hehehe

Yen Trader 1: fine will go with same as yesterday then

Senior Yen Trader: cool

Yen Trader 1: maybe a touch higher tomorrow

EXAMPLE #2

Cuomo Aide Who Called Bribes 'Ziti' Sentenced to Six Years in Prison

By Adam K. Raymond

Percoco was like a "brother" to Governor Cuomo. Photo: Louis Lanzano/Bloomberg via Getty Images

EXAMPLE #3

Traders at a company developed a sophisticated golf-based language for communicating insider information to trigger trades. E.g.:

- Birdie on hole 2
- Par on hole 5
- Hit a sand trap on hole 7

Investigator initially ignored these documents. Later noticed that the location on the documents was Boston, Massachusetts.

EXAMPLE #3

Traders at a company developed a sophisticated golf-based language for communicating insider information to trigger trades. E.g.:

- Birdie on hole 2
- Par on hole 5
- Hit a sand trap on hole 7

Investigator initially ignored these documents. Later noticed that the location on the documents was Boston, Massachusetts.

The date on the documents was February.

EXAMPLE #4

Baseball	March to October
American Football	September to January
Basketball	October to April

Overview

1. Legal Information Retrieval (ESI)

    1. eDiscovery

    2. Trial Preparation

    3. Governance

2. Principles of Legal Information Retrieval (ESI)
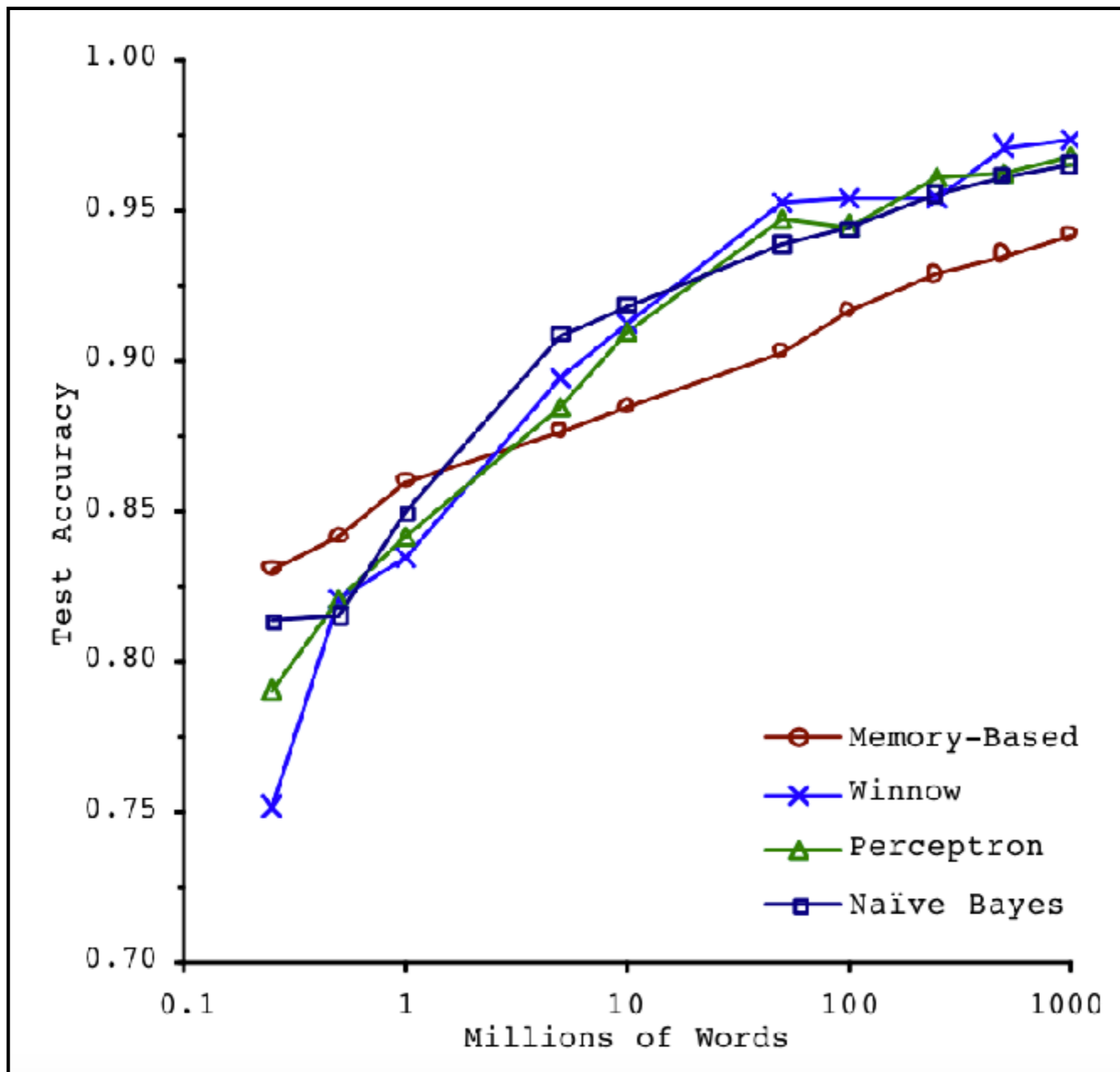
    1. Collection Completeness

    2. The Importance of Process
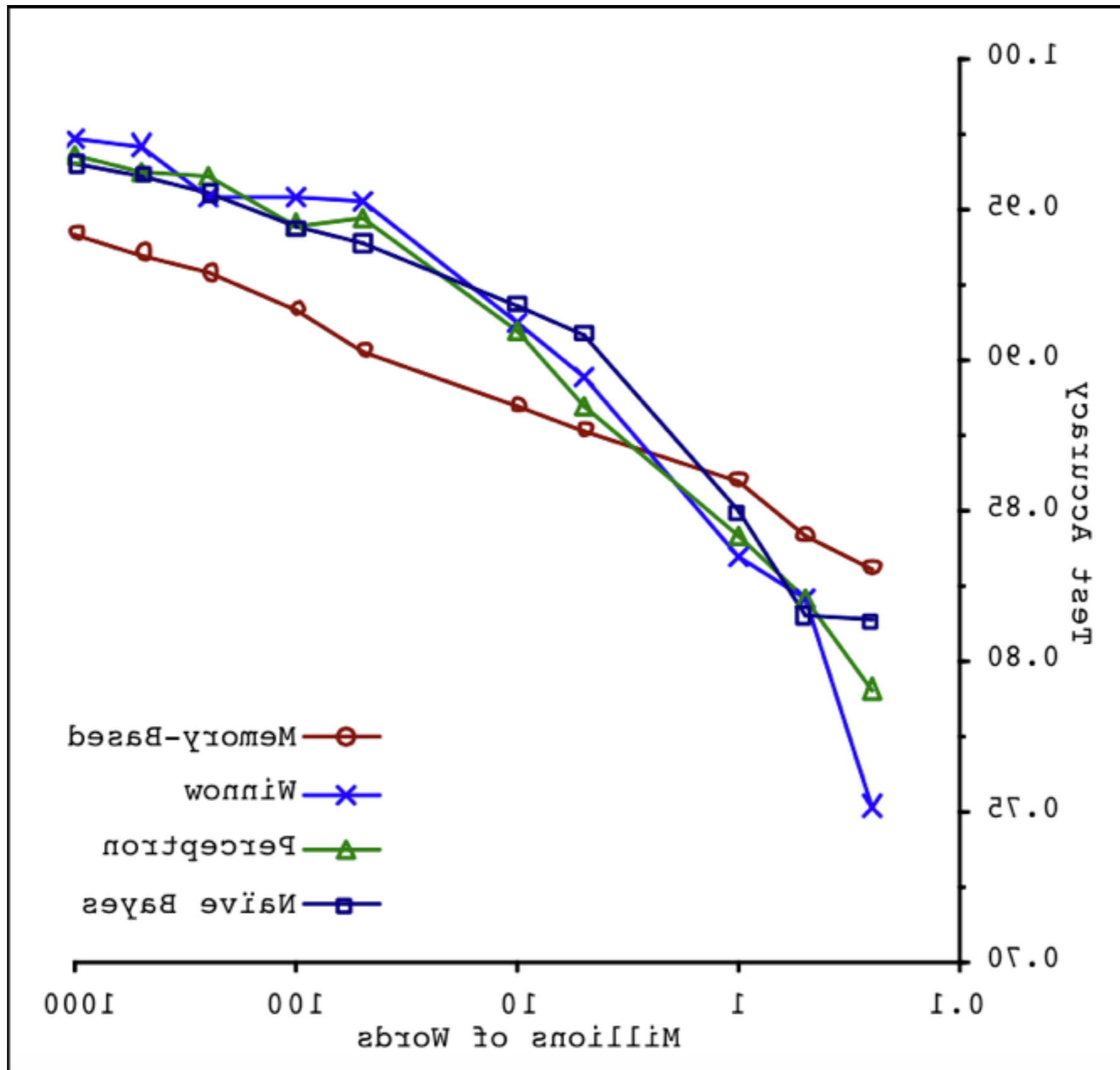
    3. Encouraging Curiosity

    4. Proprioception

    5. Trajectory

No simple answers.
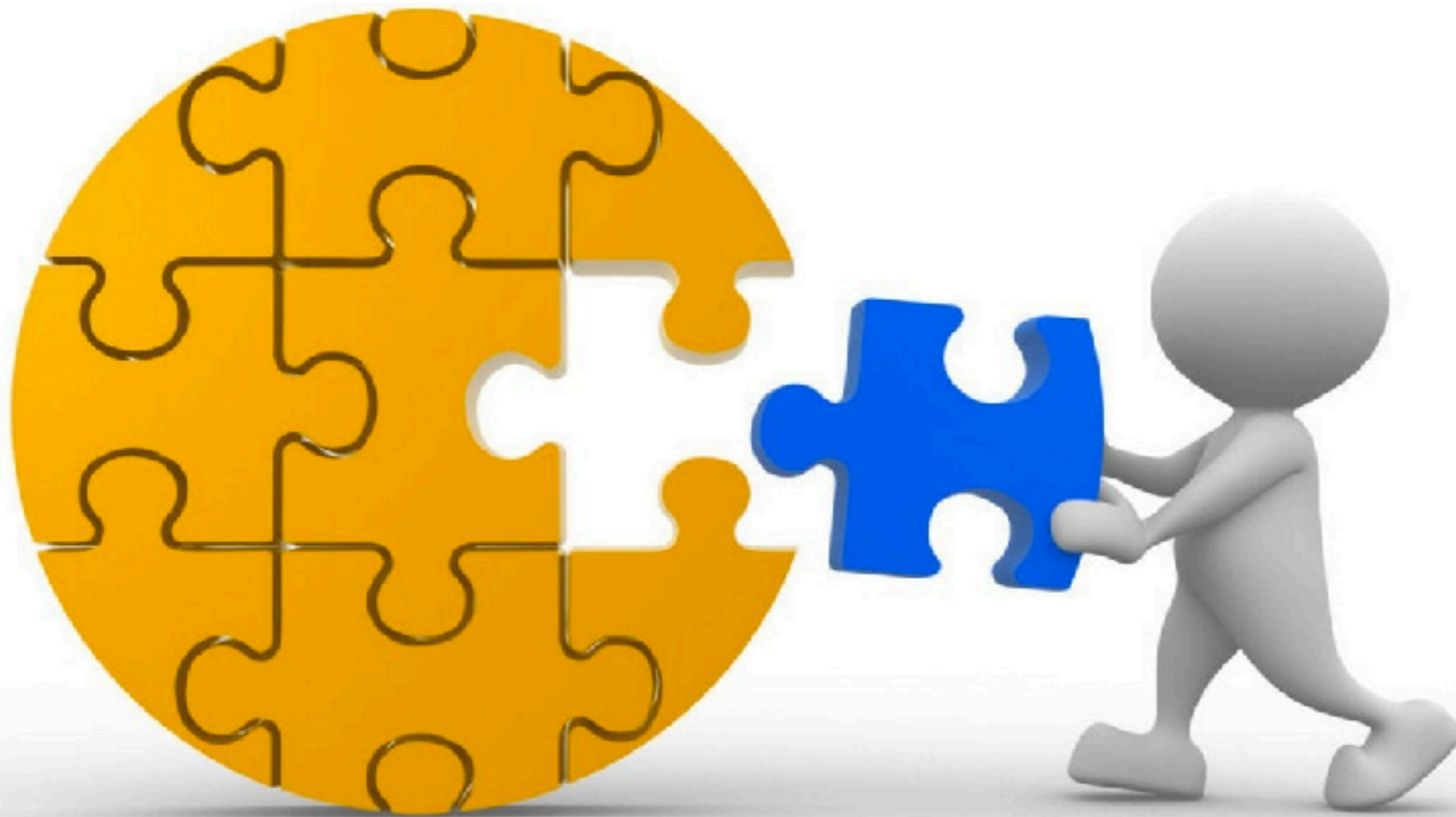Every task is its own journey.

Scaling to Very Very Large Corpora for Natural Language Disambiguation.
Michele Banko and Eric Brill, MSR, 2001

Scaling to Very Very Large Corpora for Natural Language Disambiguation.
Michele Banko and Eric Brill, MSR, 2001
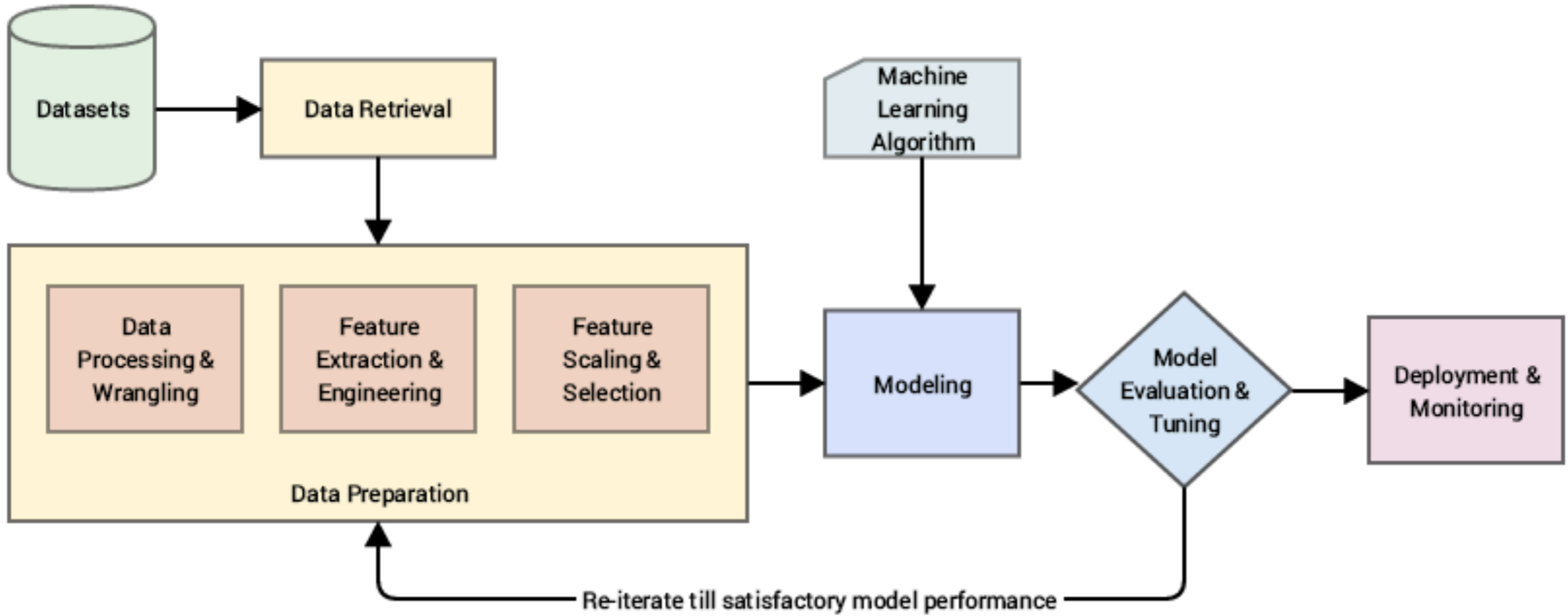
1. Collection Completeness

Gap Analysis

Missing emails in threads
Missing documents
Missing custodians
No "sandbox" allowed

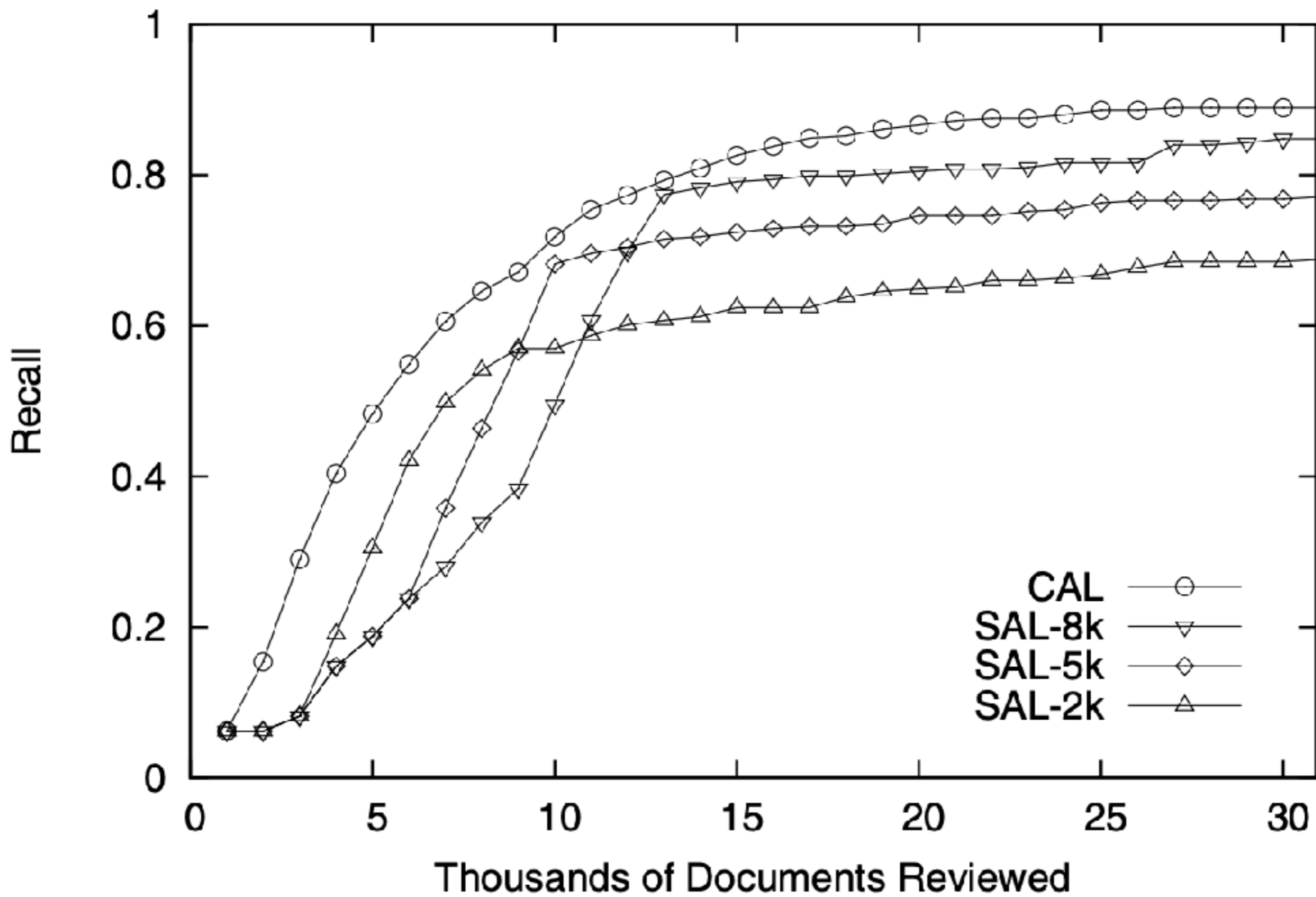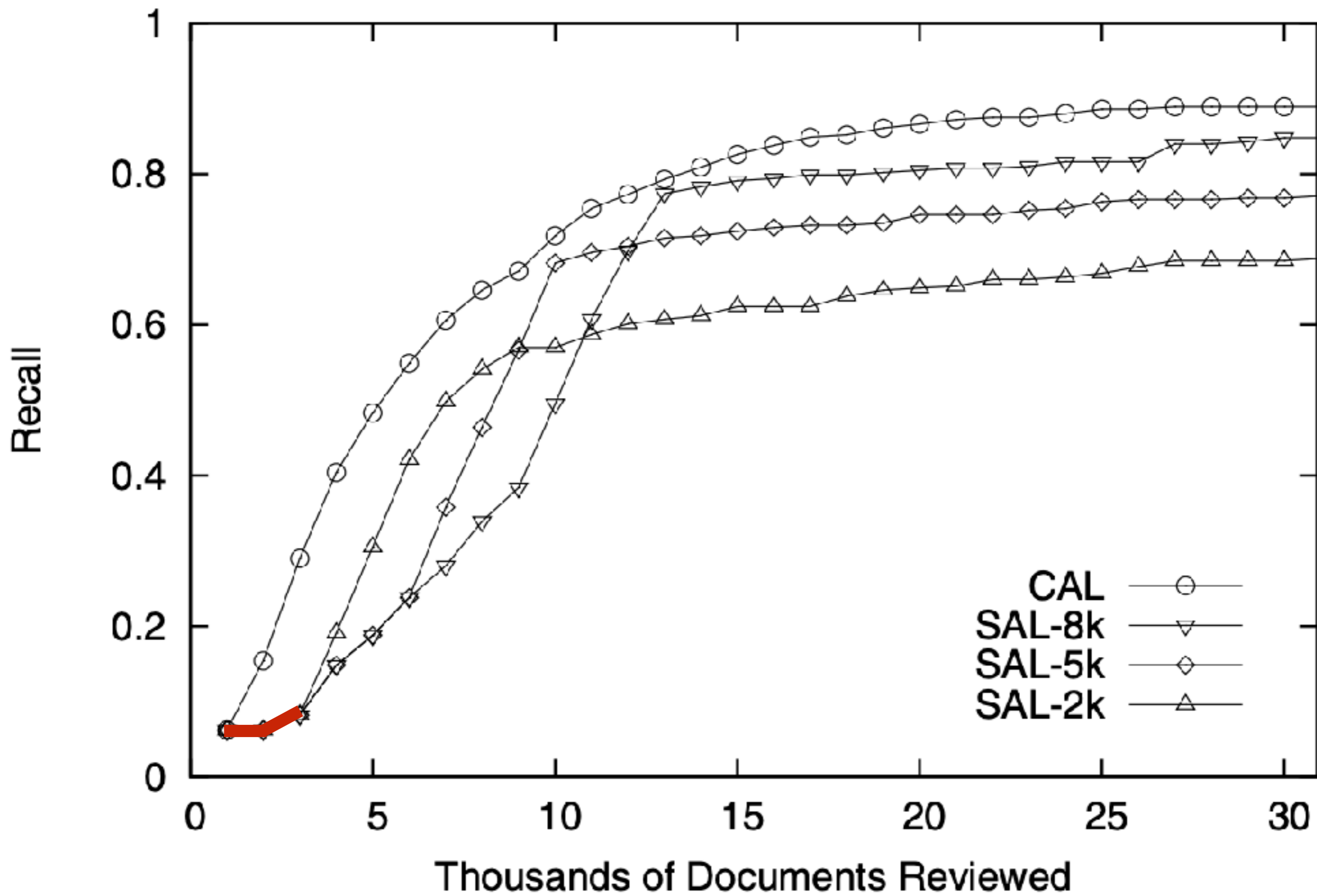| From	Subject	CUST1	CUST2	CUST3	CUST4	CUST5	CUST6	CUST7	CUST8	CUST9	CUST10	CUST11	CUST12	CUST13	CUST14	CUST15	CUST16	CUST17	CUST18	CUST19	CUST20
CUST16	CUST21 - $3B TEST Pas...	●		●	●					●	●						●		●		
MISSING																					
MISSING																					
MISSING																					
CUST2	RE: CUST21 - $3B TEST...	●	●	●	●					●			●	●			●		●		●
CUST16	RE: CUST21 - $3B TEST...	●	●	●	●					●			●	●			●		●		●
CUST2	RE: CUST21 - $3B TEST...	●	●	●	●					●			●	●			●		●		●
CUST16	RE: CUST21 - $3B TEST...	●	●	●	●					●			●	●			●		●		●
CUST1	FW: CUST21 - $3B TEST...	●	●	●	●								●	●			●		●		
CUST4	RE: CUST21 - $3B TEST...	●			●																
CUST9	Automatic Reply: CUST21...	●			●					●			●	●							
CUST4	Automatic Reply: CUST21...	●			●					●									●		●
CUST9	Automatic Reply: CUST21...	●	●		●					●			●	●			●		●		●
CUST1	FW: CUST21 - $3B TEST...	●	●							●			●	●			●		●		
CUST9	RE: CUST21 - $3B TEST...	●								●			●	●			●				●
CUST1	FW: CUST21 - $3B TEST...	●	●	●						●			●	●			●		●		
CUST13	RE: CUST21 - $3B TEST...	●			●					●	●		●	●		●	●	●	●		
CUST16	RE: CUST21 - $3B TEST...	●	●	●	●					●	●				●	●	●	●	●		●
CUST4	RE: CUST21 - $3B TEST...	●	●	●	●					●	●		●	●	●	●	●	●	●	●	●
CUST16	RE: CUST21 - $3B TEST...	●	●	●	●					●	●		●	●	●	●	●		●	●	●
MISSING																					
MISSING	RE: CUST21 - $3B TEST...																				
CUST4	RE: CUST21 - $3B TEST...	●	●	●	●					●	●		●				●		●		●
CUST18	RE: CUST21 - $3B TEST...	●		●						●	●						●		●		
MISSING																					
MISSING																					
CUST9	FW: CUST21 - $3B TEST...	●	●							●		●									
MISSING																					

2. The Importance of Process

"Statistics can never be dialectics."

–Jean-Paul Sartre

```
Datasets → Data Retrieval

Machine Learning Algorithm

Data Preparation
    Data Processing & Wrangling
    Feature Extraction & Engineering
    Feature Scaling & Selection

Modeling → Model Evaluation & Tuning → Deployment & Monitoring

Re-iterate till satisfactory model performance
```
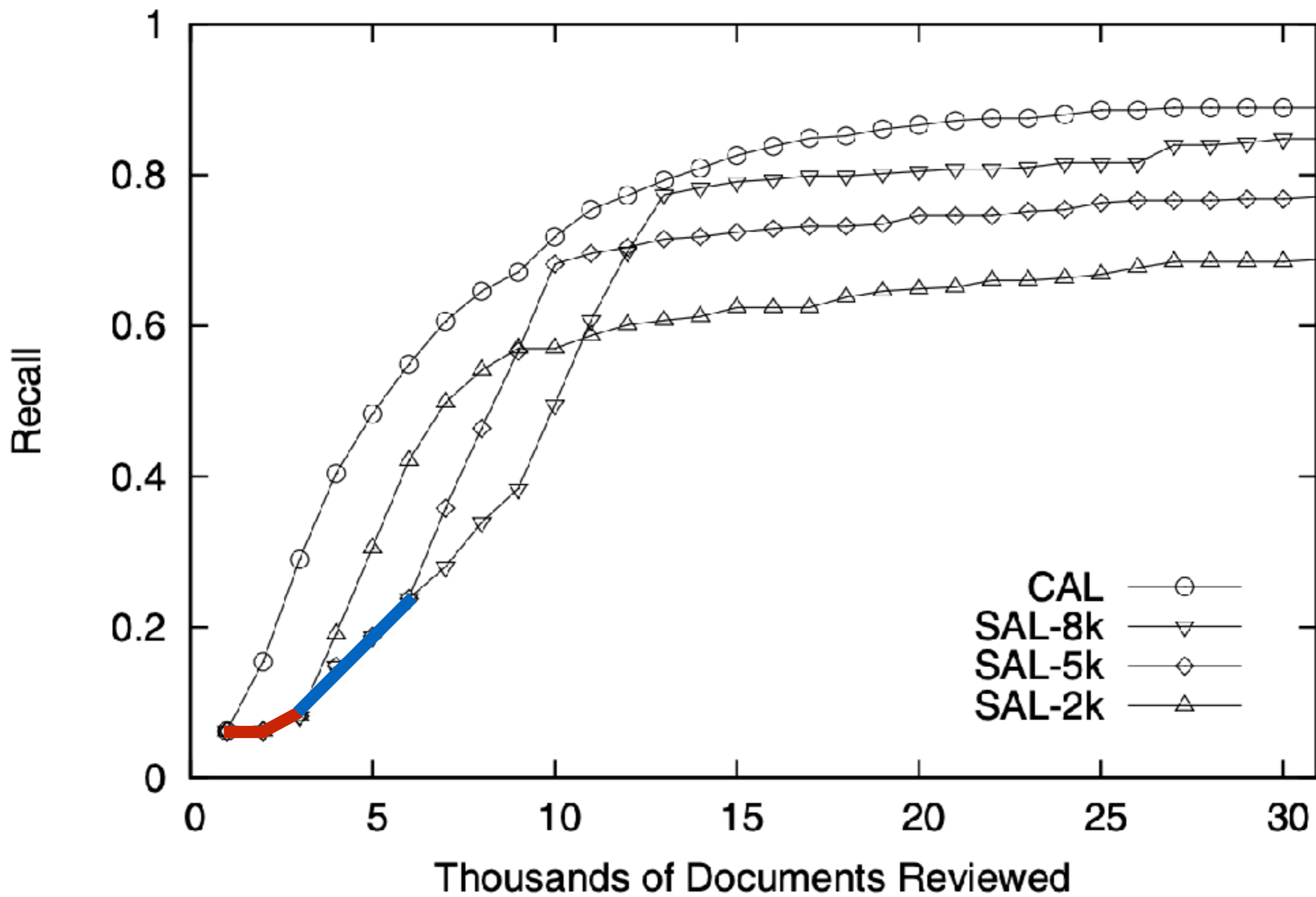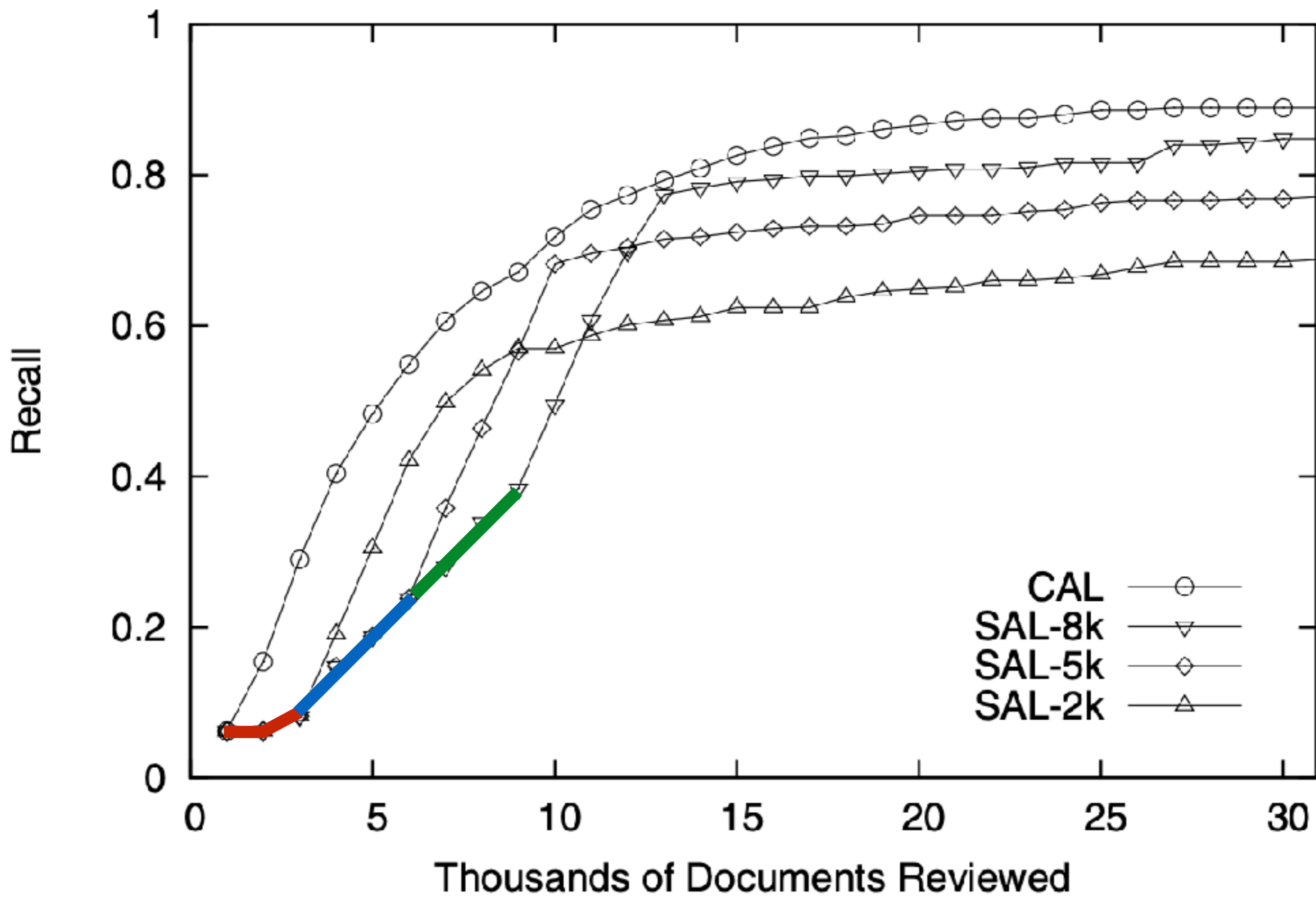
Optimize the Model

Evolve the Process

- How documents are selected for review (greedy, uncertainty, diversity, manual (search), random, etc.)

- The rate at which retraining is done

- Whether "related" documents are reviewed together (e.g. email families, threads)

- Who does the review/training (senior vs junior attorneys)

- How, and how much, quality control is done (how to handle "wrong" judgments). And when should it be done?

- How to handle multilingual workflows

- How to handle multi-issue workflows

- How to handle dynamic document collection (gap filling while review is ongoing)

3. Encouraging Curiosity (Friction is Good)

PROBATIVE DOCUMENTS
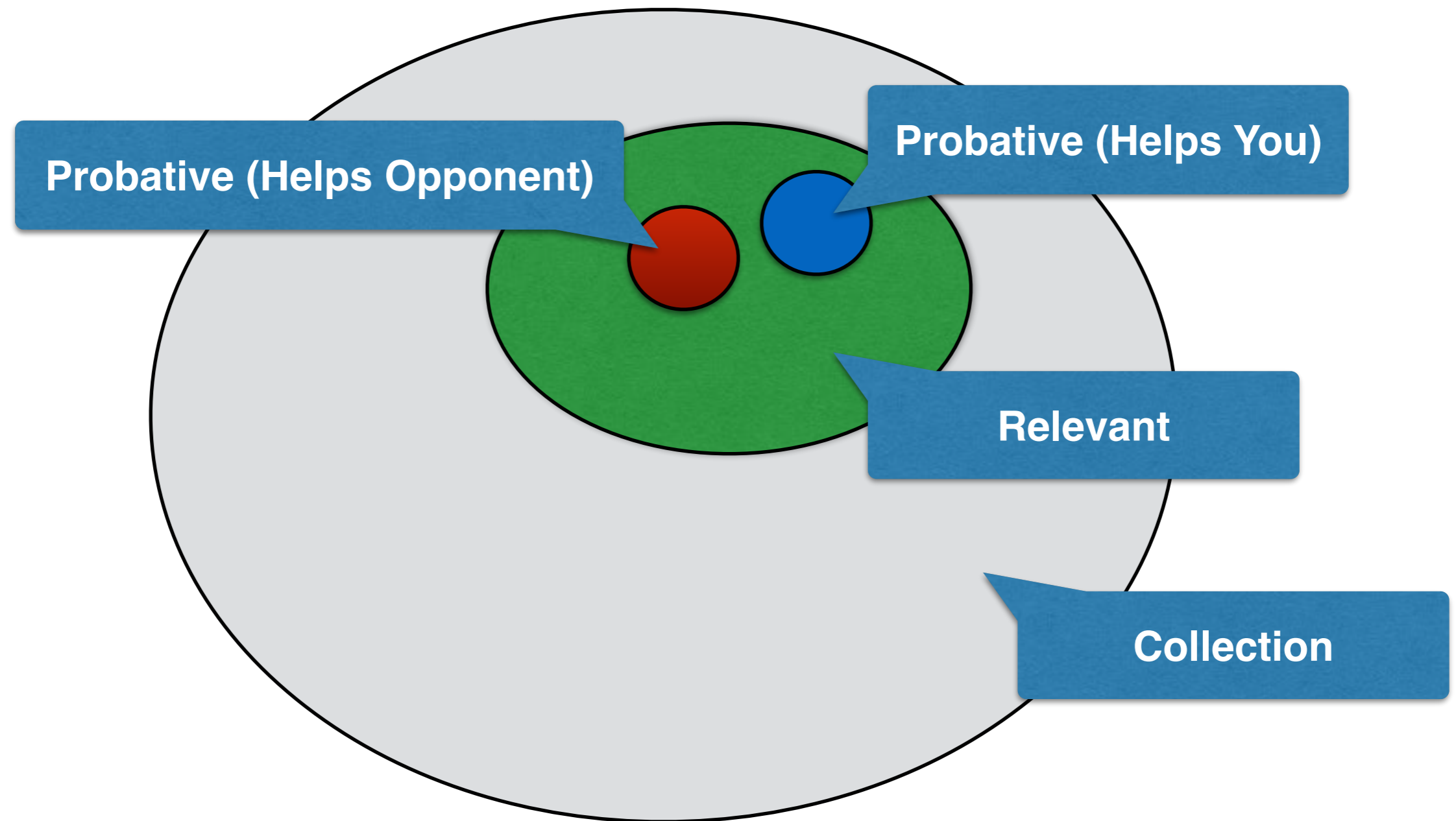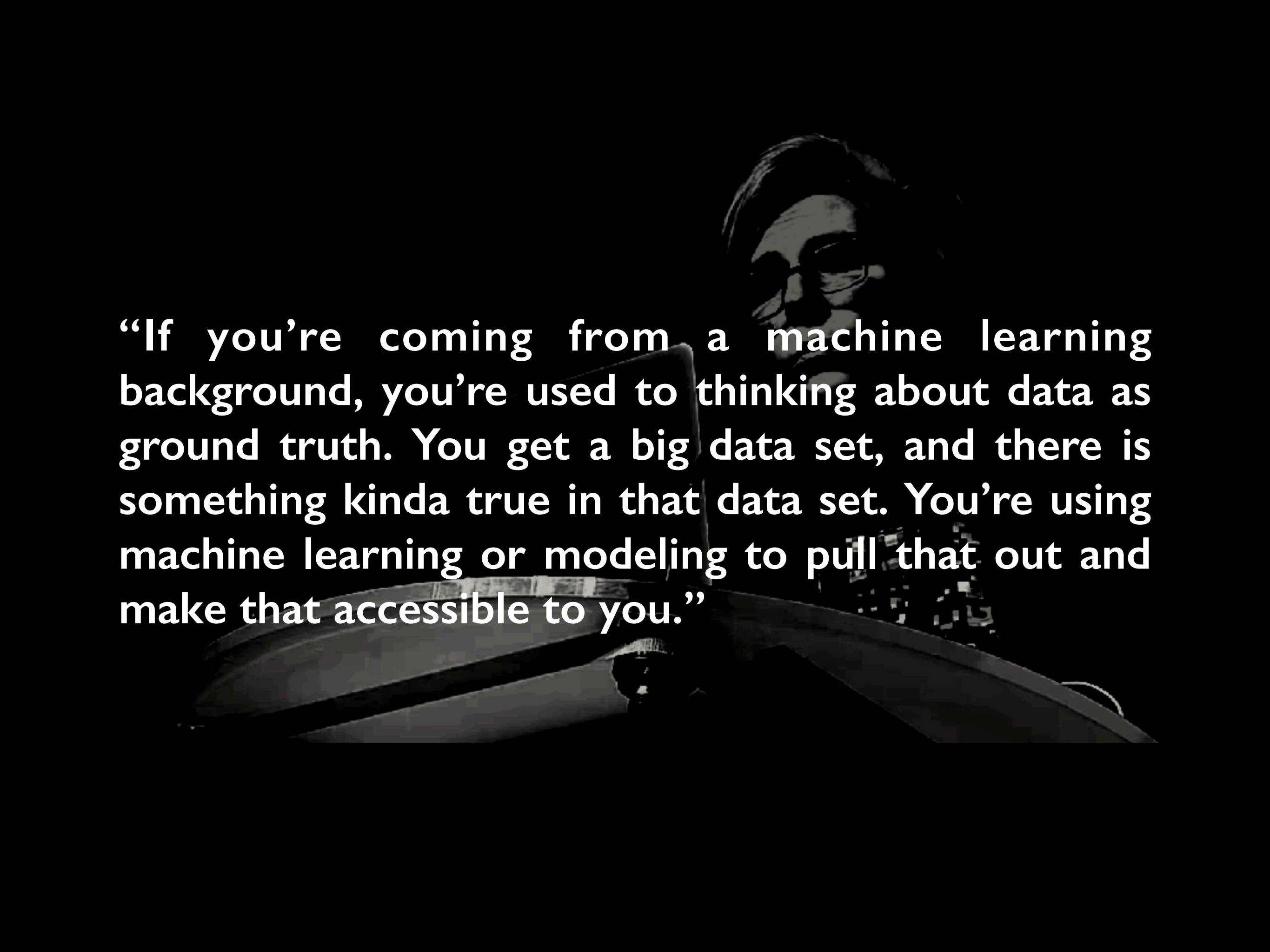
"I may not know, at the beginning, whether a seemingly straightforward line of inquiry will expand beyond recognition."

...
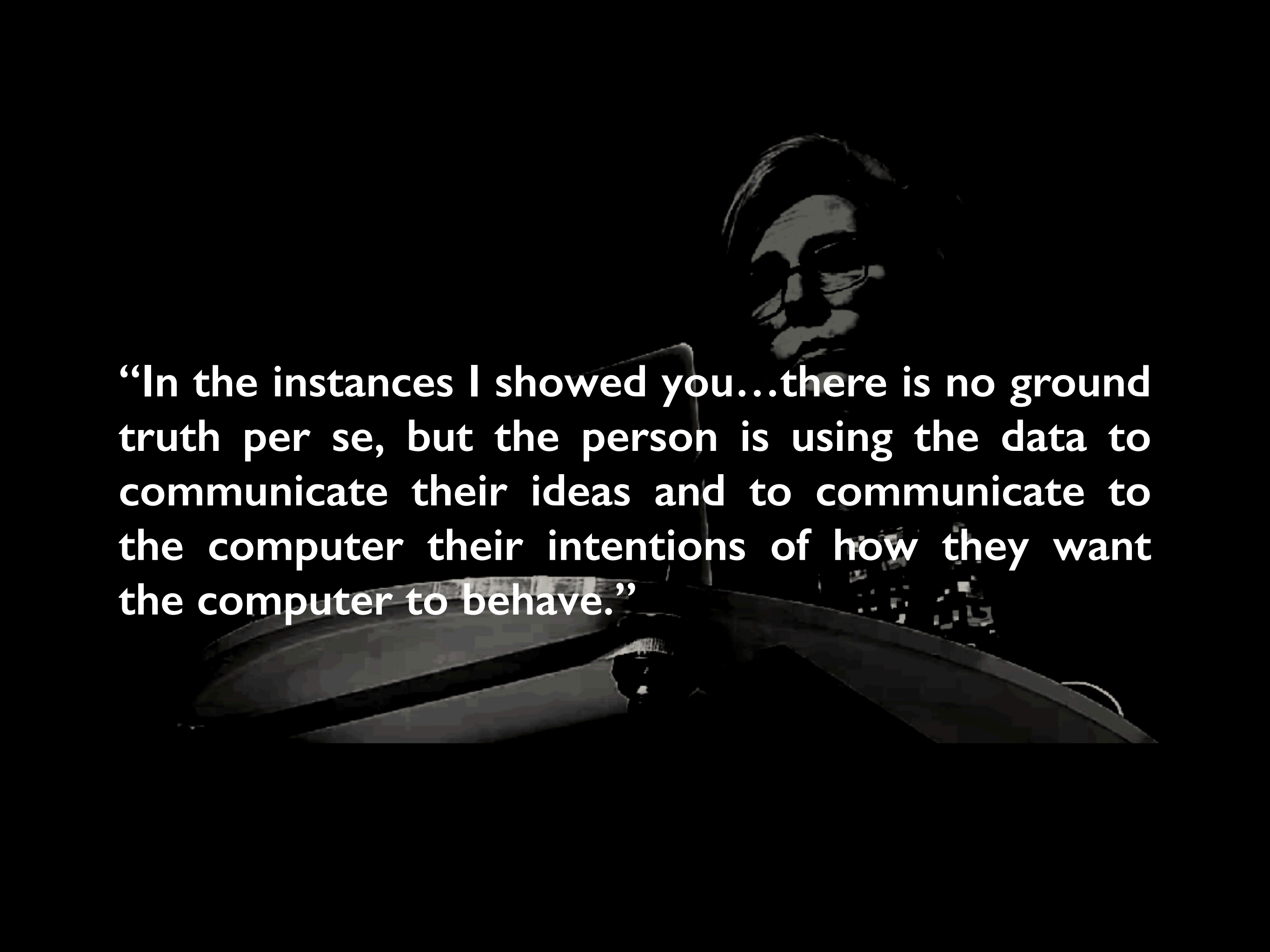
I can't determine, objectively, whether the exploration has been a success. There is usually no "right answer" against which I can measure my progress.

–Musings on Information Seeking Support Systems.
Michael Levi, 2008

"If you're coming from a machine learning background, you're used to thinking about data as ground truth. You get a big data set, and there is something kinda true in that data set. You're using machine learning or modeling to pull that out and make that accessible to you."

"In the instances I showed you…there is no ground truth per se, but the person is using the data to communicate their ideas and to communicate to the computer their intentions of how they want the computer to behave."

"Optimisation hinders evolution."

–Alan Perlis (as quoted by Gary Kasparov
in his book Deep Thinking)

4. Proprioception

"A sense or perception, usually at a subconscious level, of the movements and position of the body and especially its limbs, independent of vision."

–Farlex Partner Medical Dictionary

(((master settlement agreement OR msa) AND NOT (medical savings account OR metropolitan standard area)) OR s. 1415 OR (ets AND NOT educational testing service) OR (liggett AND NOT sharon a. liggett) OR atco OR lorillard OR (pmi AND NOT presidential management intern) OR pm usa OR rjr OR (b&w AND NOT photo*) OR phillip morris OR batco OR ftc test method OR star scientific OR vector group OR joe camel OR (marlboro AND NOT upper marlboro)) AND NOT (tobacco* OR cigarette* OR smoking OR tar OR nicotine OR smokeless OR synar amendment OR philip morris OR r.j. reynolds OR ("brown and williamson") OR ("brown & williamson") OR bat industries OR liggett group)
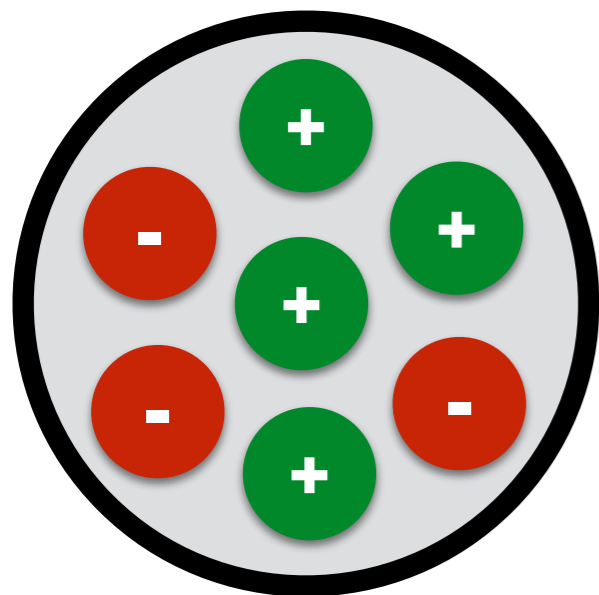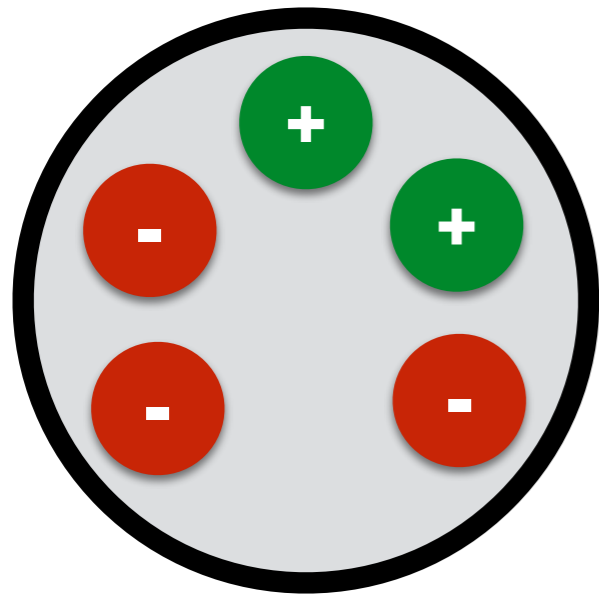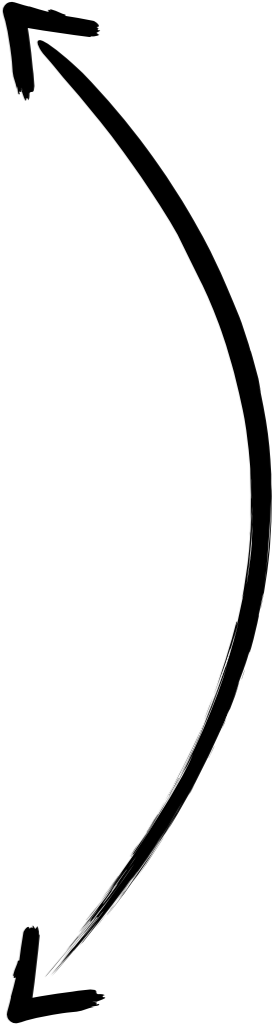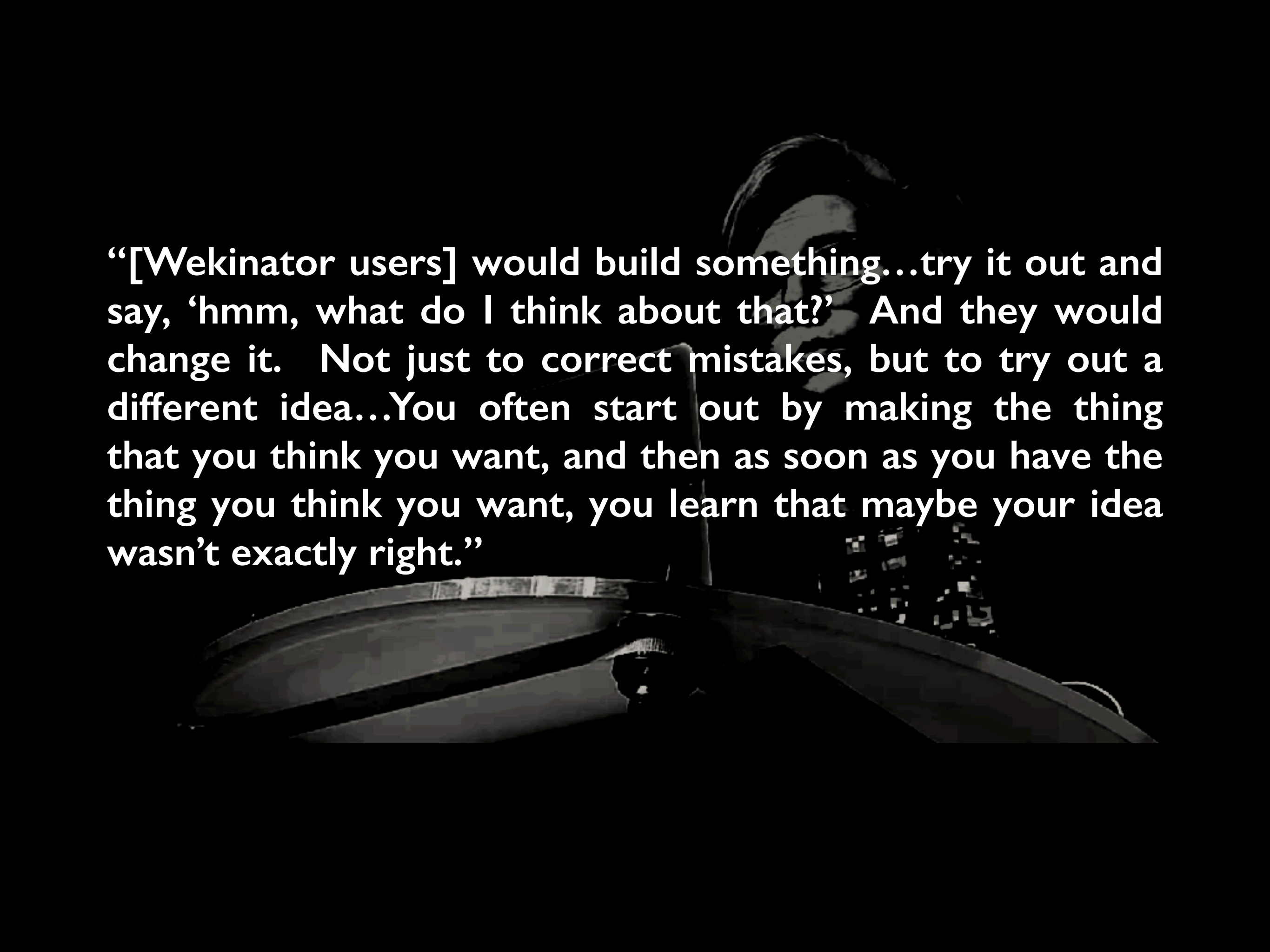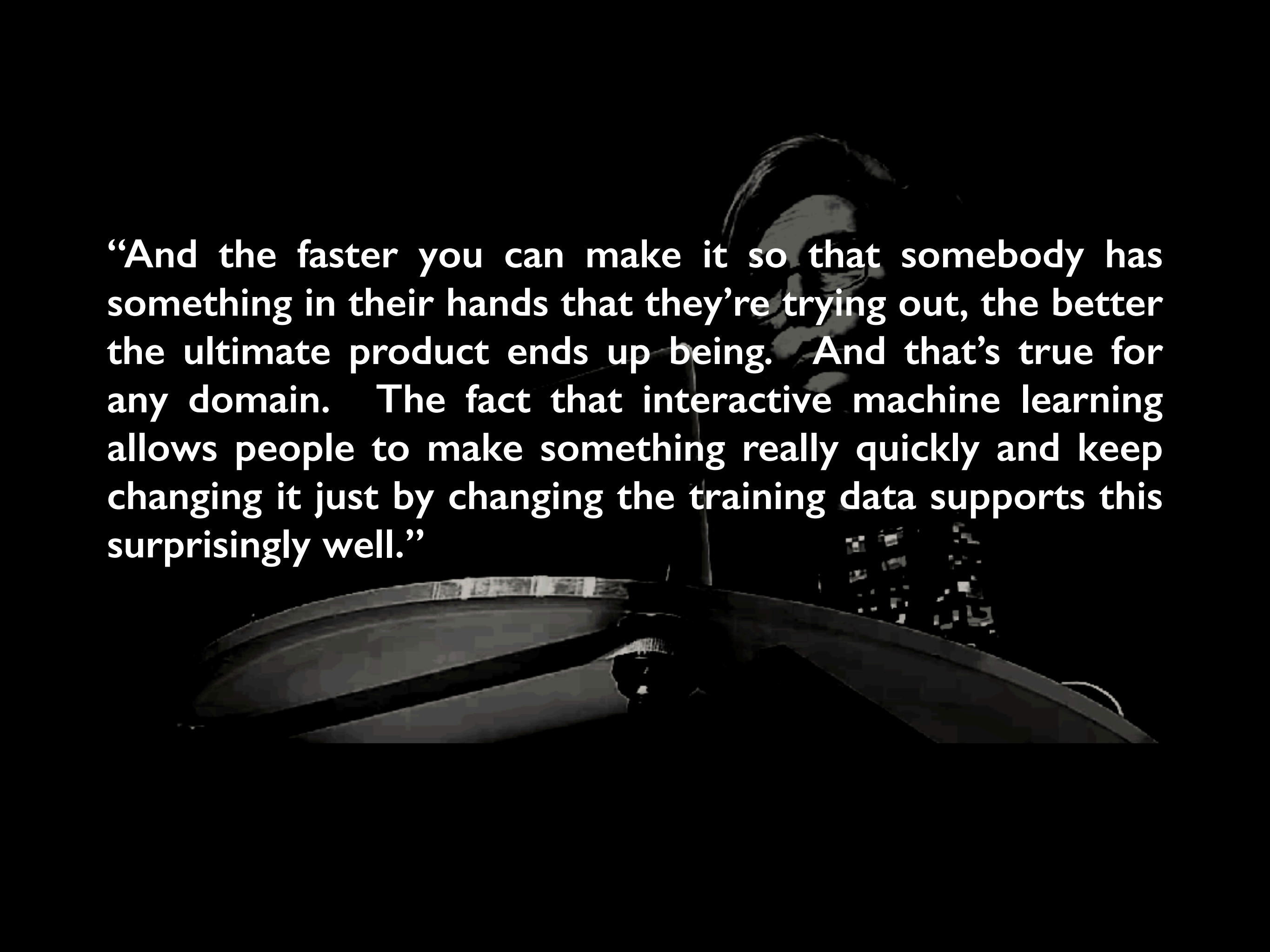
"[Wekinator users] would build something…try it out and say, 'hmm, what do I think about that?' And they would change it. Not just to correct mistakes, but to try out a different idea…You often start out by making the thing that you think you want, and then as soon as you have the thing you think you want, you learn that maybe your idea wasn't exactly right."

"And the faster you can make it so that somebody has something in their hands that they're trying out, the better the ultimate product ends up being. And that's true for any domain. The fact that interactive machine learning allows people to make something really quickly and keep changing it just by changing the training data supports this surprisingly well."

Transparency

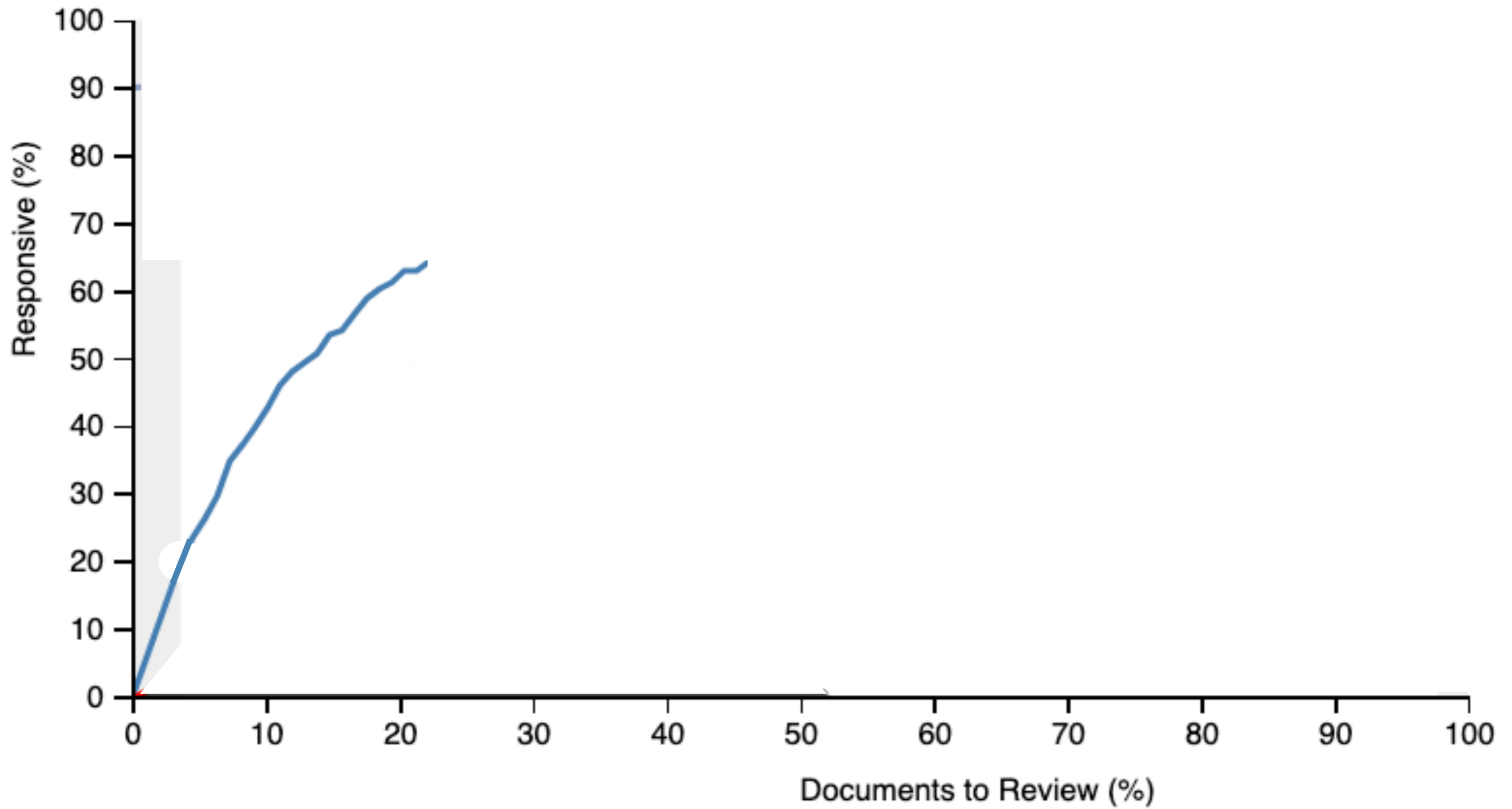Explainability

Proprioception

Transparency

<

Explainability

<

Proprioception

5. Trajectory

Visualization of the progress and/or change of state as a user *moves through* a collection
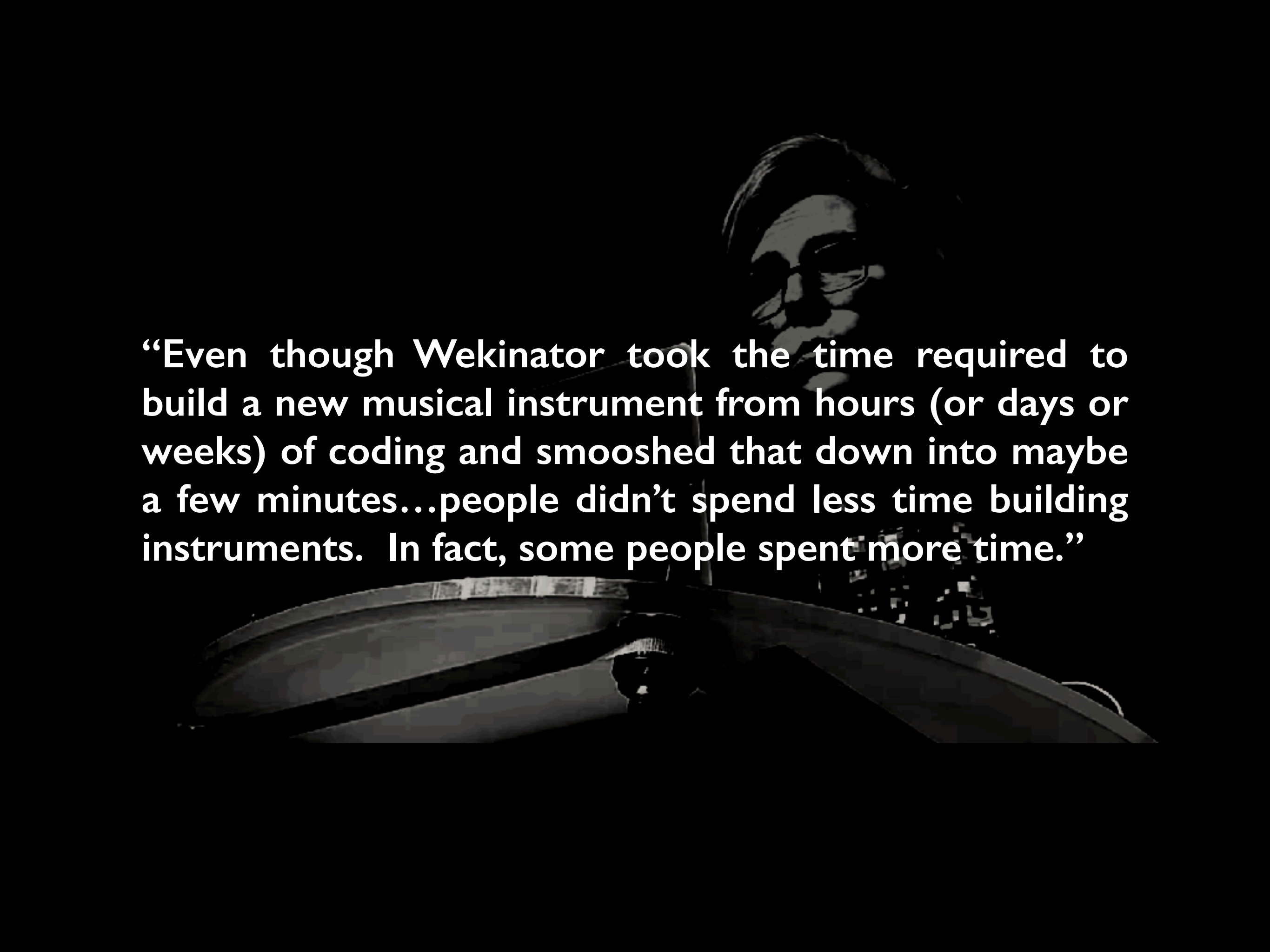
Gain Curve

Summary

Principles of Legal Information Retrieval (ESI)

    1. Collection Completeness

    2. The Importance of Process

    3. Encouraging Curiosity

    4. Proprioception

    5. Trajectory

Must you optimize for engagement?

"Even though Wekinator took the time required to build a new musical instrument from hours (or days or weeks) of coding and smooshed that down into maybe a few minutes…people didn't spend less time building instruments. In fact, some people spent more time."

Thank you!

jpickens@opentext.com