Artificial Intelligence (AI) & PDF Document Processing

Henry Sal
President Computing System Innovations

PDF DAY
WASHINGTON
DC

Todays Agenda

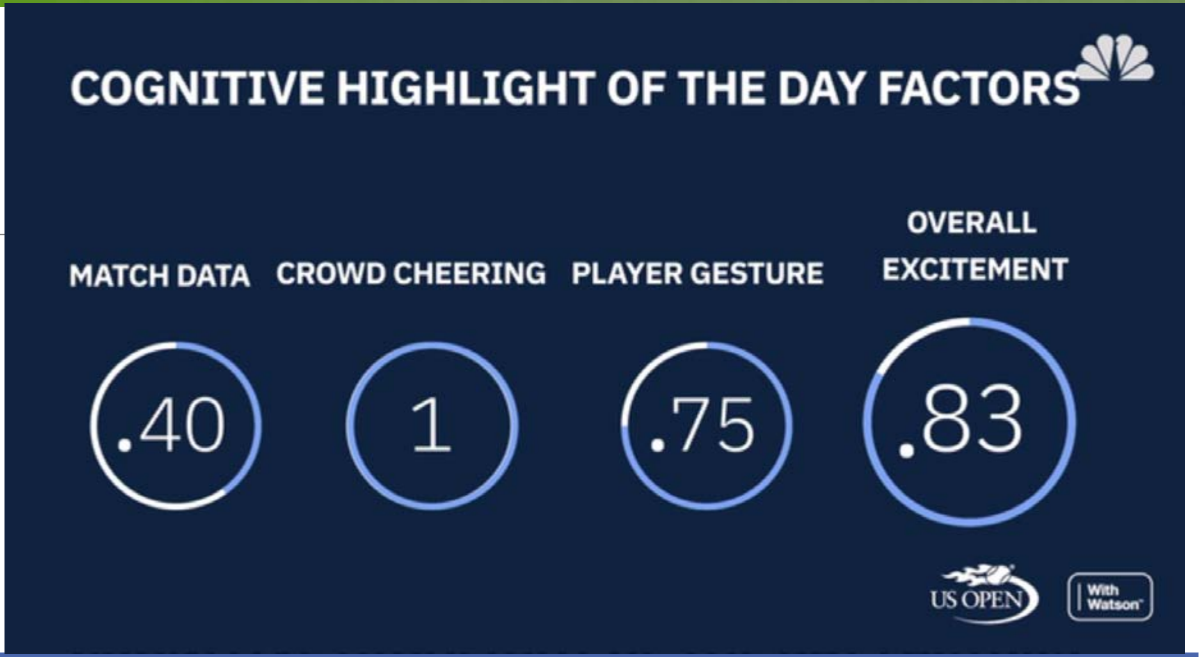
- What is artificial intelligence (AI)?
- How does machine learning work?
- What is good machine learning?
- Practical AI PDF applications
- Accuracies & Exceptions

Hobie Cat

Cognitive

Watson serves up Cognitive Highlights at the US Open

IBM's Watson - an AI platform, is doing a special job at this year's U.S. Open.



COGNITIVE HIGHLIGHT OF THE DAY FACTORS

MATCH DATA .40

CROWD CHEERING 1

PLAYER GESTURE .75

OVERALL EXCITEMENT .83

NO LINES. NO CHECKOUT
(NO. SERIOUSLY.)
JUST WALK OUT SHOPPING
amazon go

Types of AI

- **General AI** – simulated human intelligence, fabulous machines that have all our senses and more. Deep learning, AlphaGo, Watson, etc…

- **Narrow AI** – Technologies that are able to perform specific tasks as well as, or better than, we humans can
  - Electronic stock trading
  - Facial recognition
  - Self-driving vehicles
  - Unstructured document analysis

- **Machine learning** – state-of-the art approach to narrow AI. Given example data, machines teach themselves and then make predictions within specific domains.

  - **Supervised learning** – provide initial examples of data that machine algorithms train on and learn how to perform that task.

  - **Online learning** – automatic improvement of learned knowledge via real world usage.

Finding Text

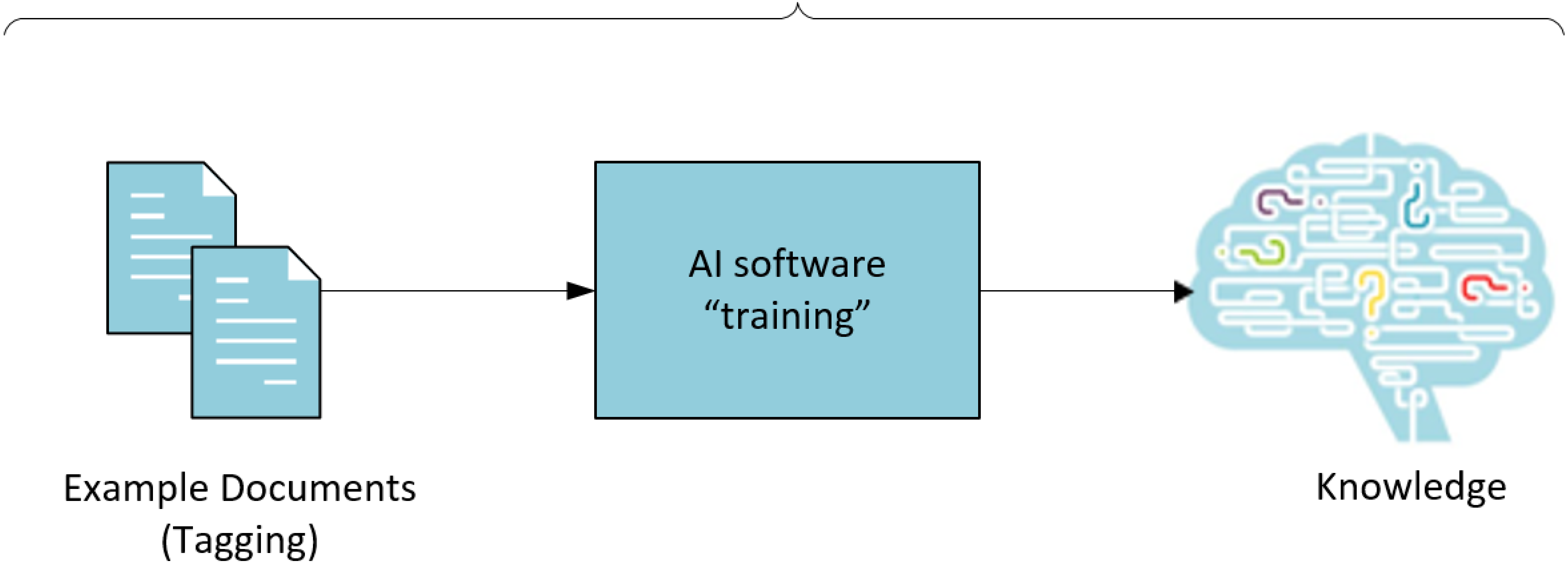
The Old Way

Regular Expressions Handprint SSN

```
([C][A]|[F][L]|[V][A]|<K california 90>|<K florida 95>|<K virginia 90>)[\ ]?[0-9]{3}[\-\ ][0-9]{2}[=\-][0-9]{4}|

([#:\.\ ]([0-9]{3}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{4})[,;\.\ ])|
(^([0-9]{3}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{4})[,;\.\ ])|
([#:\.\ ]([0-9]{3}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{4})^)|
(^([0-9]{3}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{4})^)|
([#:\.\ ]([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9]{4})[,;\.\ ])|
(^([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9]{4})[,;\.\ ])|
([#:\.\ ]([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9]{4})^)|
(^([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[\ ]?[=\*\-][\ ]?[0-9]{4})^)|
([#:\.\ ]([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9]{4})[,;\.\ ])|
(^([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9]{4})[,;\.\ ])|
([#:\.\ ]([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9]{4})^)|
(^([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\ ][0-9]{4})^)|
([#:\.\ ]([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9]{4})[,;\.\ ])|
(^([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9]{4})[,;\.\ ])|
([#:\.\ ]([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9]{4})^)|
(^([0-9][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9ABbCDGHIiJKkLOQRSUZXx&]{2}[=\-\.][0-9]{4})^)|
([Xx]{2,3}[=\-\.][Xx]{2}[=\-\.][0-9]{4})|
([#:\.\ ]([0-9]{3}[=\-\ ][0-9]{2}[=\-\ ][KkXx*]{4})[,;\.\ ])|
(^([0-9]{3}[=\-\ ][0-9]{2}[=\-\ ][KkXx*]{4})[,;\.\ ])|
([#:\.\ ]([0-9]{3}[=\-\ ][0-9]{2}[=\-\ ][KkXx*]{4})^)|
(^([0-9]{3}[=\-\ ][0-9]{2}[=\-\ ][KkXx*]{4})^)|
<K claim 95>[\ ]<K number 95>[:]?[\ ]{1,2}([0-9]{3}[\-]?[0-9]{2}[\-][0-9]{4}|[0-9]{3}[\-][0-9]{2}[\-]?[0-9]{4})
```

Supervised Machine Learning

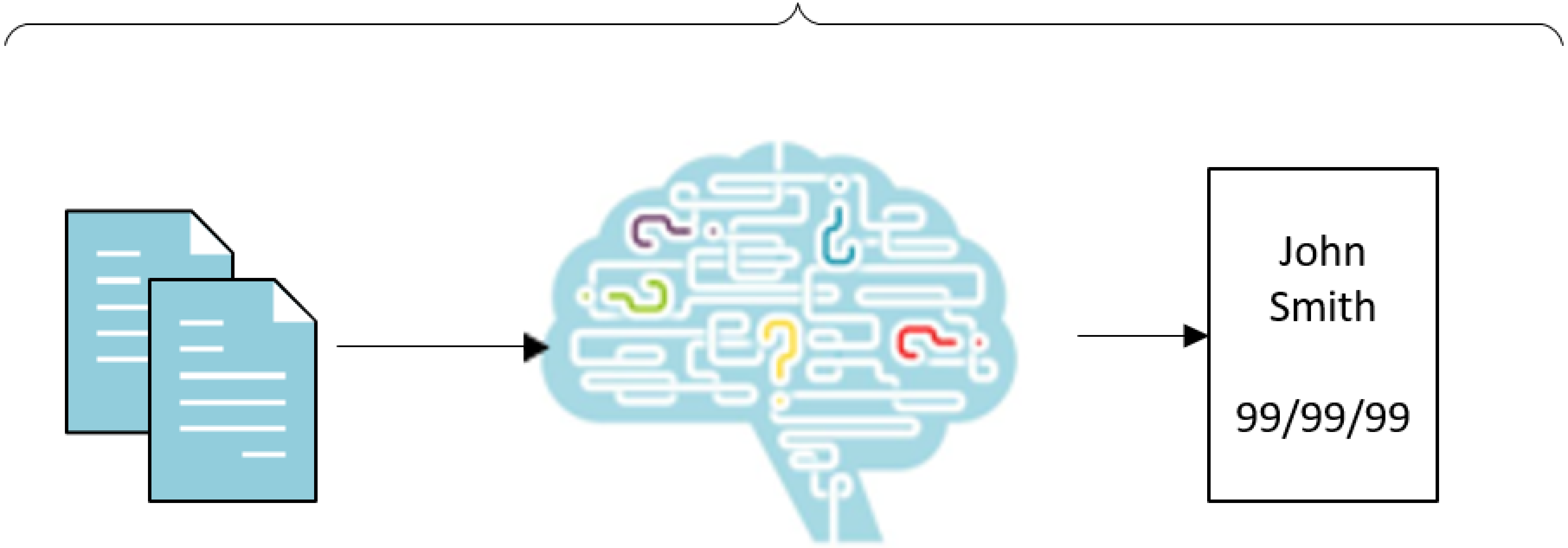


Example Documents (Tagging)

AI software "training"

Knowledge

AI Data Extraction

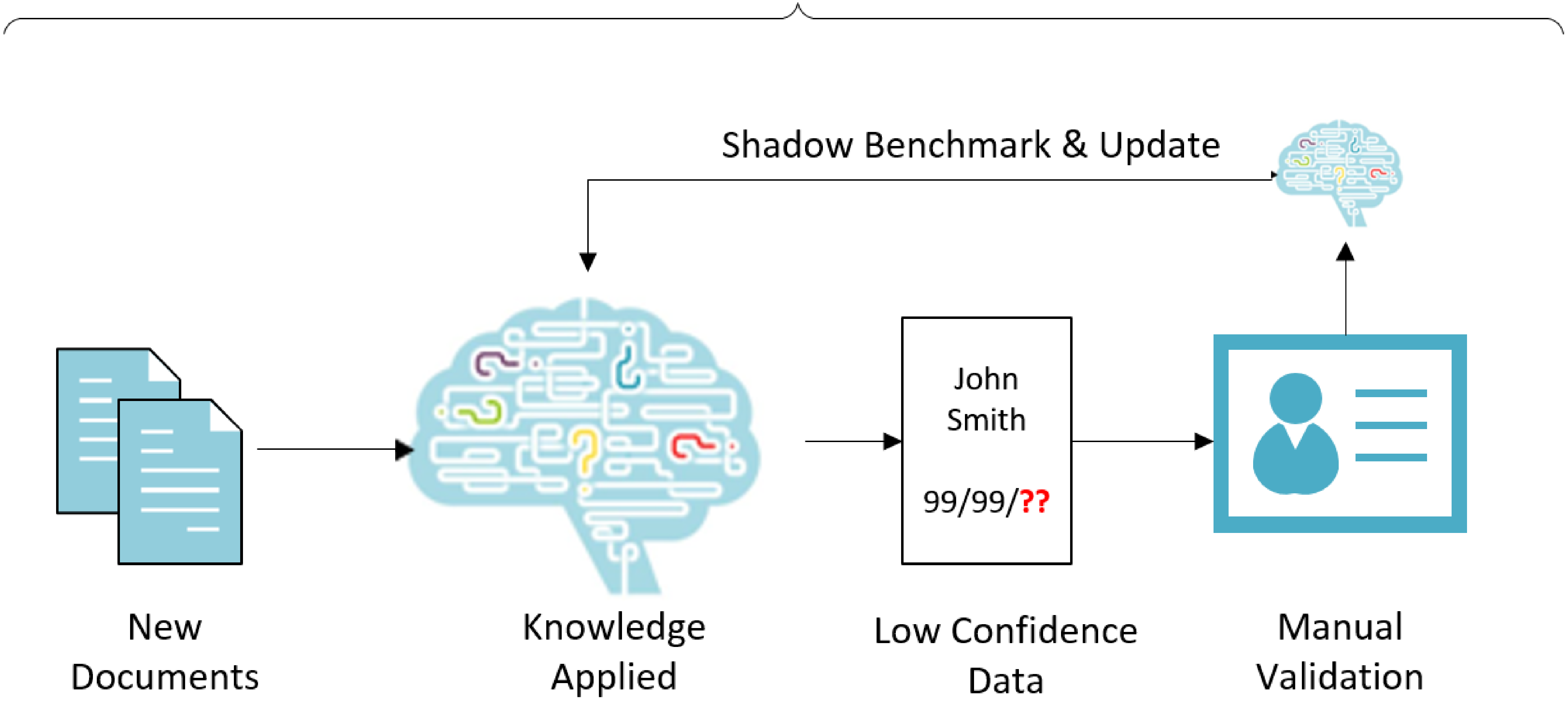


New Documents

Knowledge Applied

Data Extracted

Online Learning



Shadow Benchmark & Update

New Documents

Knowledge Applied

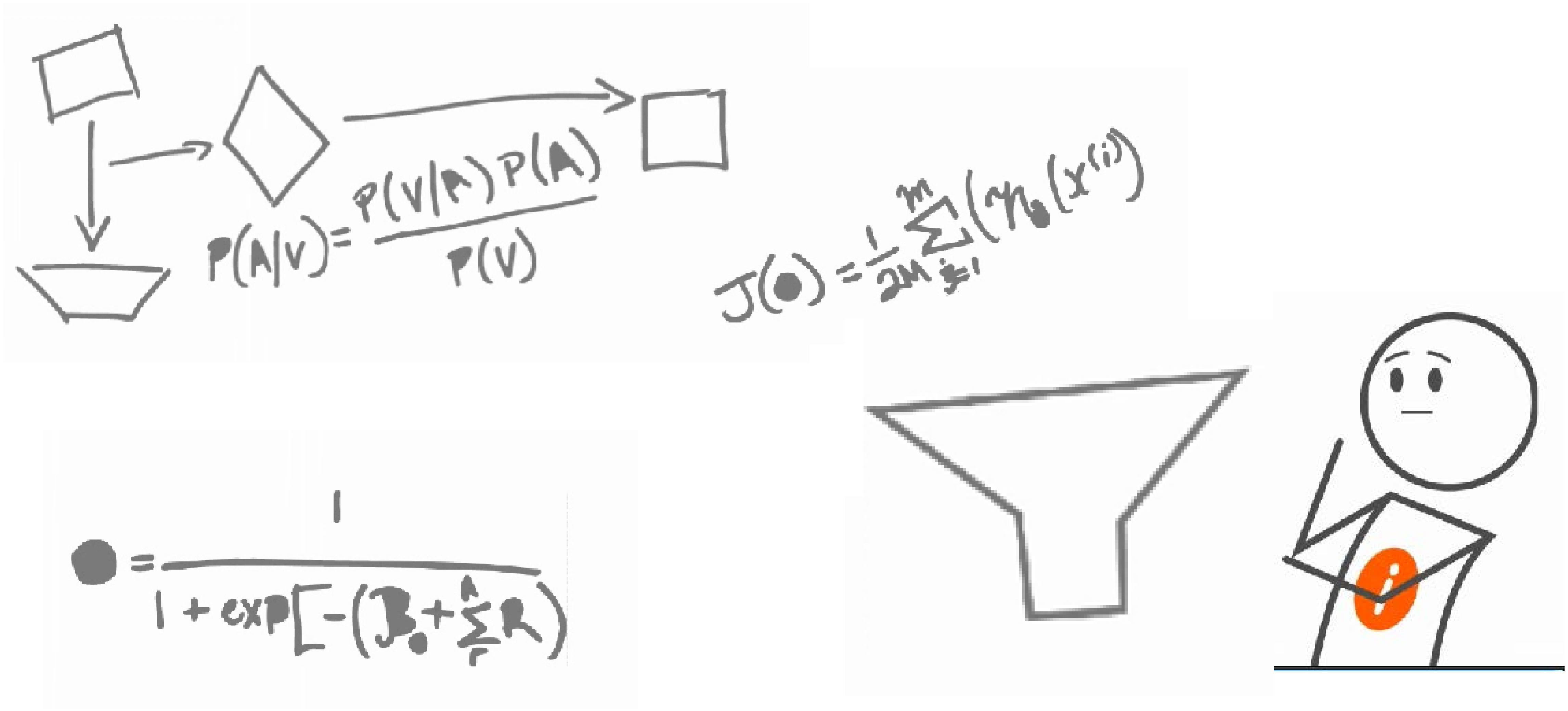
John Smith

99/99/??

Low Confidence Data

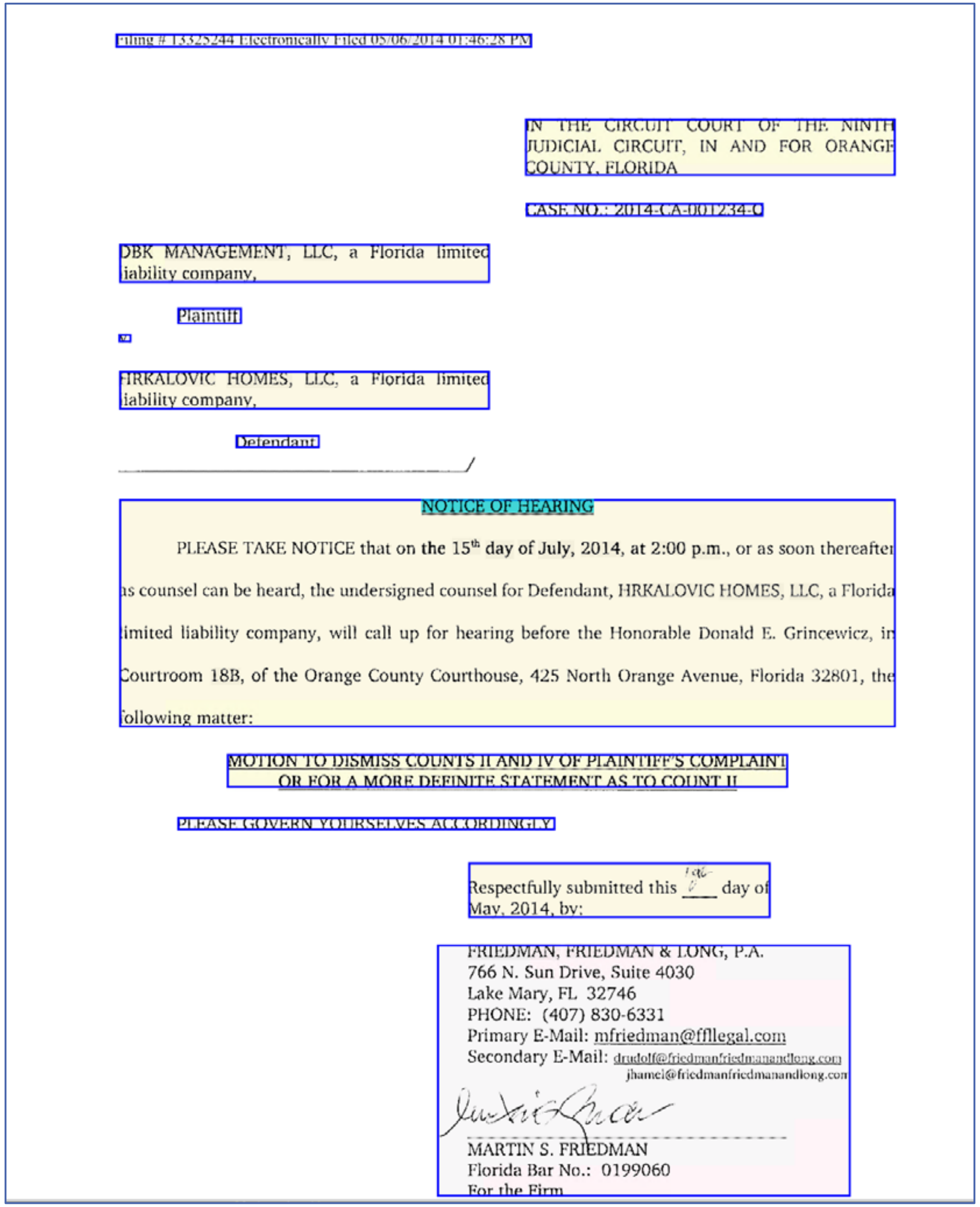
Manual Validation

Content Locators, Role Determination, Conditional Random Fields (CRFs), Natural Language Processing (NLP), Parts of Speech Tagging, etc…


$P(A|V) = \frac{P(V|A)P(A)}{P(V)}$
$J(●) = \frac{1}{2m}\sum_{i=1}^{m}(h_●(x^{(i)})$
$● = \frac{1}{1+\exp[-(B_0+\sum_{r}^{A}R)}$
i

Visible features

- Format
- Relative location
- What's around it (that matters)



Filing # 13325244 Electronically Filed 05/06/2014 01:46:28 PM

IN THE CIRCUIT COURT OF THE NINTH JUDICIAL CIRCUIT, IN AND FOR ORANGE COUNTY, FLORIDA

CASE NO.: 2014-CA-001234-O

DBK MANAGEMENT, LLC, a Florida limited liability company,

Plaintiff

v.

HRKALOVIC HOMES, LLC, a Florida limited liability company,

Defendant

_______________________________________/

NOTICE OF HEARING

PLEASE TAKE NOTICE that on the 15th day of July, 2014, at 2:00 p.m., or as soon thereafter as counsel can be heard, the undersigned counsel for Defendant, HRKALOVIC HOMES, LLC, a Florida limited liability company, will call up for hearing before the Honorable Donald E. Grincewicz, in Courtroom 18B, of the Orange County Courthouse, 425 North Orange Avenue, Florida 32801, the following matter:

MOTION TO DISMISS COUNTS II AND IV OF PLAINTIFF'S COMPLAINT
OR FOR A MORE DEFINITE STATEMENT AS TO COUNT II

PLEASE GOVERN YOURSELVES ACCORDINGLY.

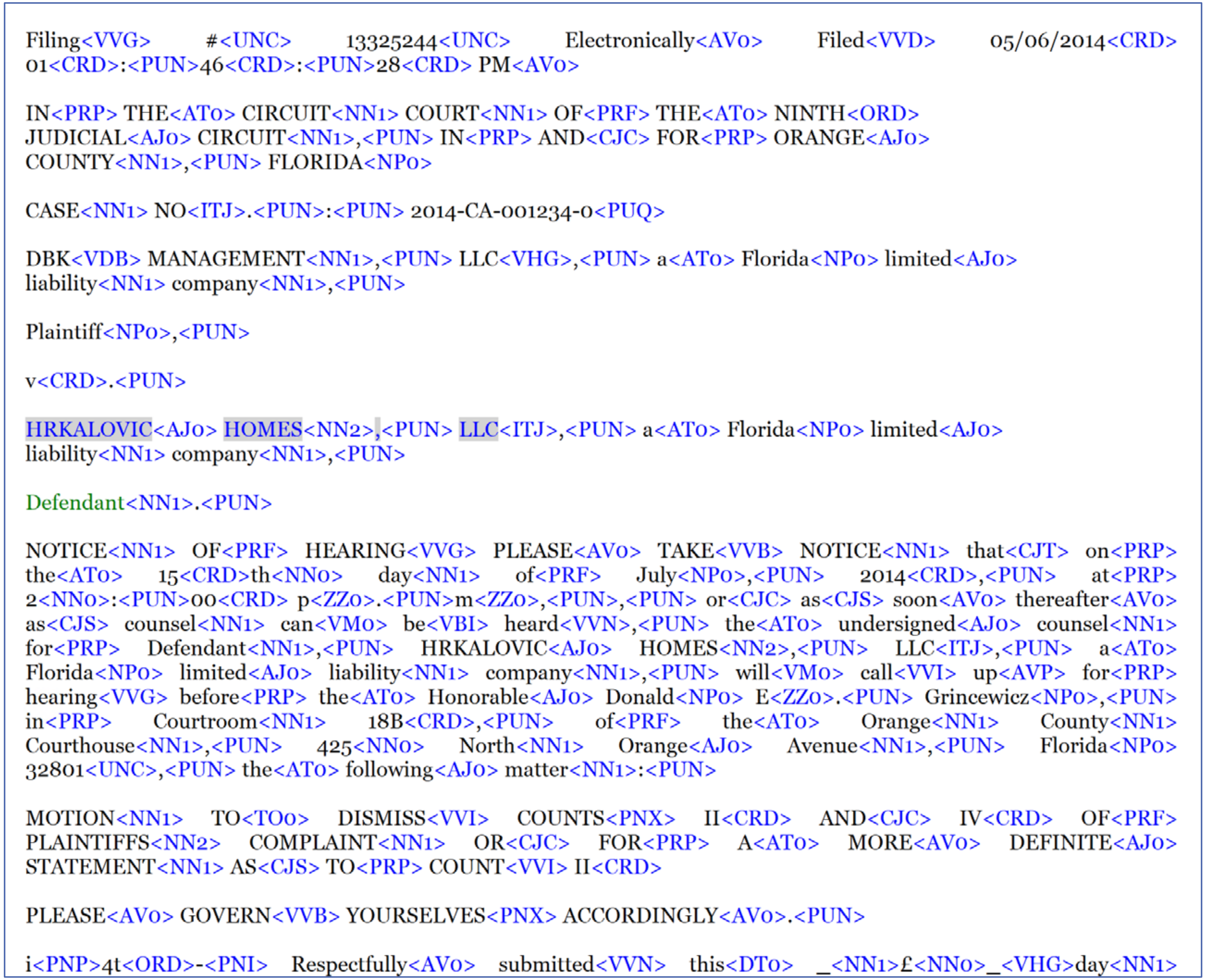
Respectfully submitted this 1st day of May, 2014, by:

FRIEDMAN, FRIEDMAN & LONG, P.A.
766 N. Sun Drive, Suite 4030
Lake Mary, FL 32746
PHONE: (407) 830-6331
Primary E-Mail: mfriedman@ffllegal.com
Secondary E-Mail: drudolf@friedmanfriedmanandlong.com
jhamel@friedmanfriedmanandlong.com

MARTIN S. FRIEDMAN
Florida Bar No.: 0199060
For the Firm

Invisible features

- Semantics (parts of speech)



Filing<VVG> #<UNC> 13325244<UNC> Electronically<AV0> Filed<VVD> 05/06/2014<CRD> 01<CRD>:<PUN>46<CRD>:<PUN>28<CRD> PM<AV0>

IN<PRP> THE<AT0> CIRCUIT<NN1> COURT<NN1> OF<PRF> THE<AT0> NINTH<ORD> JUDICIAL<AJ0> CIRCUIT<NN1>,<PUN> IN<PRP> AND<CJC> FOR<PRP> ORANGE<AJ0> COUNTY<NN1>,<PUN> FLORIDA<NP0>

CASE<NN1> NO<ITJ>.<PUN>:<PUN> 2014-CA-001234-0<PUQ>

DBK<VDB> MANAGEMENT<NN1>,<PUN> LLC<VHG>,<PUN> a<AT0> Florida<NP0> limited<AJ0> liability<NN1> company<NN1>,<PUN>

Plaintiff<NP0>,<PUN>

v<CRD>.<PUN>

HRKALOVIC<AJ0> HOMES<NN2>,<PUN> LLC<ITJ>,<PUN> a<AT0> Florida<NP0> limited<AJ0> liability<NN1> company<NN1>,<PUN>

Defendant<NN1>.<PUN>

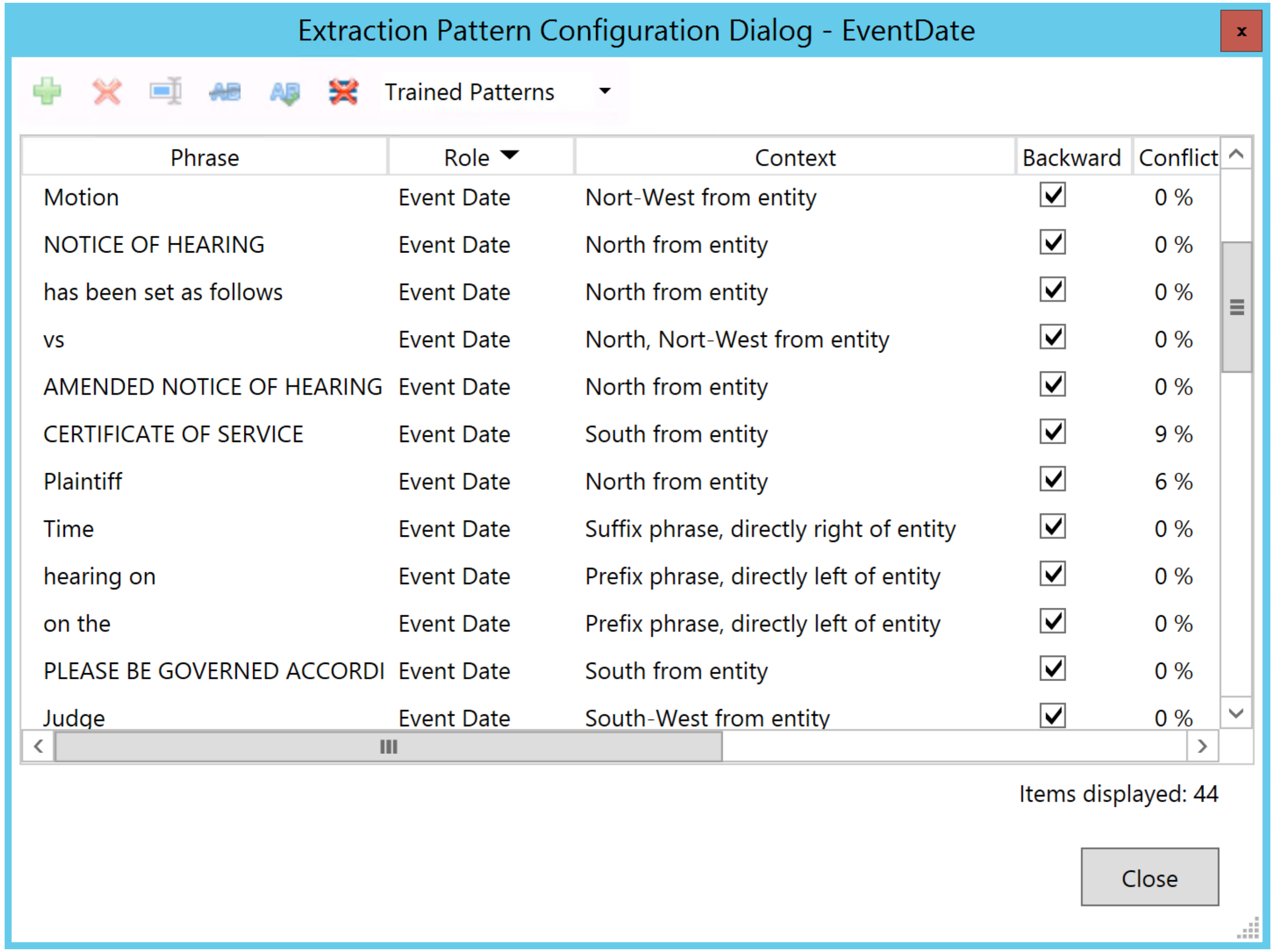
NOTICE<NN1> OF<PRF> HEARING<VVG> PLEASE<AV0> TAKE<VVB> NOTICE<NN1> that<CJT> on<PRP> the<AT0> 15<CRD>th<NN0> day<NN1> of<PRF> July<NP0>,<PUN> 2014<CRD>,<PUN> at<PRP> 2<NN0>:<PUN>00<CRD> p<ZZ0>.<PUN>m<ZZ0>,<PUN>,<PUN> or<CJC> as<CJS> soon<AV0> thereafter<AV0> as<CJS> counsel<NN1> can<VM0> be<VBI> heard<VVN>,<PUN> the<AT0> undersigned<AJ0> counsel<NN1> for<PRP> Defendant<NN1>,<PUN> HRKALOVIC<AJ0> HOMES<NN2>,<PUN> LLC<ITJ>,<PUN> a<AT0> Florida<NP0> limited<AJ0> liability<NN1> company<NN1>,<PUN> will<VM0> call<VVI> up<AVP> for<PRP> hearing<VVG> before<PRP> the<AT0> Honorable<AJ0> Donald<NP0> E<ZZ0>.<PUN> Grincewicz<NP0>,<PUN> in<PRP> Courtroom<NN1> 18B<CRD>,<PUN> of<PRF> the<AT0> Orange<NN1> County<NN1> Courthouse<NN1>,<PUN> 425<NN0> North<NN1> Orange<AJ0> Avenue<NN1>,<PUN> Florida<NP0> 32801<UNC>,<PUN> the<AT0> following<AJ0> matter<NN1>:<PUN>

MOTION<NN1> TO<TO0> DISMISS<VVI> COUNTS<PNX> II<CRD> AND<CJC> IV<CRD> OF<PRF> PLAINTIFFS<NN2> COMPLAINT<NN1> OR<CJC> FOR<PRP> A<AT0> MORE<AV0> DEFINITE<AJ0> STATEMENT<NN1> AS<CJS> TO<PRP> COUNT<VVI> II<CRD>

PLEASE<AV0> GOVERN<VVB> YOURSELVES<PNX> ACCORDINGLY<AV0>.<PUN>

i<PNP>4t<ORD>-<PNI> Respectfully<AV0> submitted<VVN> this<DT0> _<NN1>£<NN0>_<VHG>day<NN1>

Example Knowledge



Extraction Pattern Configuration Dialog - EventDate

Phrase	Role	Context	Backward	Conflict
Motion	Event Date	Nort-West from entity	☑	0 %
NOTICE OF HEARING	Event Date	North from entity	☑	0 %
has been set as follows	Event Date	North from entity	☑	0 %
vs	Event Date	North, Nort-West from entity	☑	0 %
AMENDED NOTICE OF HEARING	Event Date	North from entity	☑	0 %
CERTIFICATE OF SERVICE	Event Date	South from entity	☑	9 %
Plaintiff	Event Date	North from entity	☑	6 %
Time	Event Date	Suffix phrase, directly right of entity	☑	0 %
hearing on	Event Date	Prefix phrase, directly left of entity	☑	0 %
on the	Event Date	Prefix phrase, directly left of entity	☑	0 %
PLEASE BE GOVERNED ACCORDI	Event Date	South from entity	☑	0 %
Judge	Event Date	South-West from entity	☑	0 %

Items displayed: 44

What is good machine learning?

- Highly accurate results
- Doesn’t require enormous volume of examples to train on
- Learns from its mistakes, not a static fragile system
- Flexible! Users can easily establish new data items without software development
- Not a black box. Analysts can review, debug, and refine knowledge
- Most steps performed by software (i.e. initial tagging, model refinement, online learning, etc…)

Practical Application to PDF Documents

- Automatically locate and extract (or redact) data
- Auto tag PDF documents for accessibility
- Auto separate and bookmark embedded PDFs
- Transform unstructured content to structured output(s)
- Eliminate human document review & data entry
- Create 24x7x365 “lights out” document workflows

Accuracies
&
Exceptions

- OCR engines are pattern recognition engines
  (Machine learning from the '70s)
- OCR full page reads 90%+
  (Image quality, skew, background, form dropouts, color, etc…)
- Field level accuracy 85% - 100%
- Transaction accuracy 85%+ — end user gold standard
- Digitally born PDFs accuracy game changer — no OCR needed

Questions?

Thank You!

hsal@csisoft.com