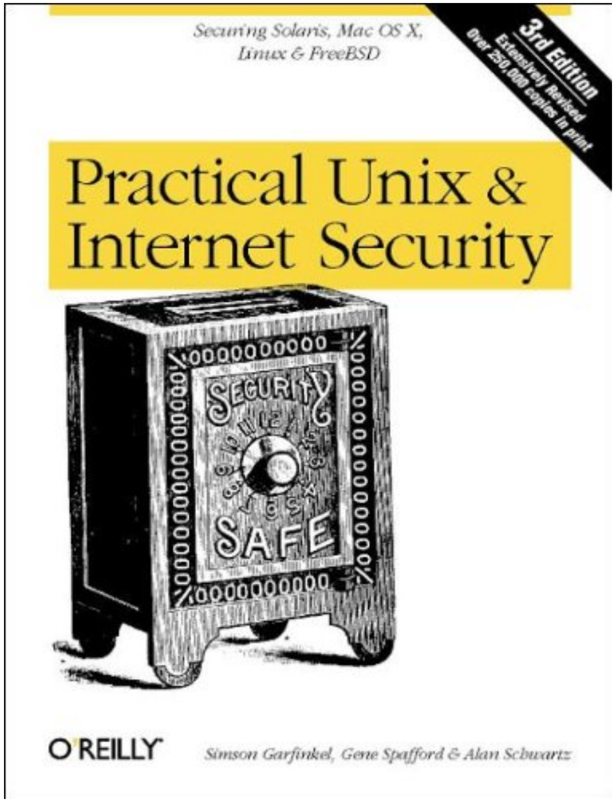
Tutorial: Text Analytics for Security

William Enck
North Carolina State University
http://www.enck.org
enck@cs.ncsu.edu

Tao Xie
University of Illinois at Urbana-Champaign
http://web.engr.illinois.edu/~taoxie/
taoxie@illinois.edu

What is Computer Security?

"A computer is secure if you can depend on it and its software to behave as you expect."



User Expectations

- User expectations are a form of *context*.
- Other forms of context for security decisions
  - Temporal context (e.g., time of day)
  - Environmental context (e.g., location)
  - Execution context
    - OS level (e.g., UID, arguments)
    - Program analysis level (e.g., control flow, data flow)

Defining User Expectations

- User expectations are difficult to formally (and even informally) define.
  - Based on an individual's perception the results from past experiences and education
  - ... so, we can't be perfect

- Starting place: look at the *user interface*

Why Text Analytics?

- User interface consists of graphics and *text*
  - End users: includes finding, installing, and running the software (e.g., first run vs. subsequent)
  - Developers: includes API documentation, comments in code, and requirements documents

- Goal: *process natural language textual sources to aid security decisions*

Outline

- Introduction
- Background on text analytics
- Case Study 1: App Markets
- Case Study 2: ACP Rules
- Wrap-up

Challenges in Analyzing NL Data

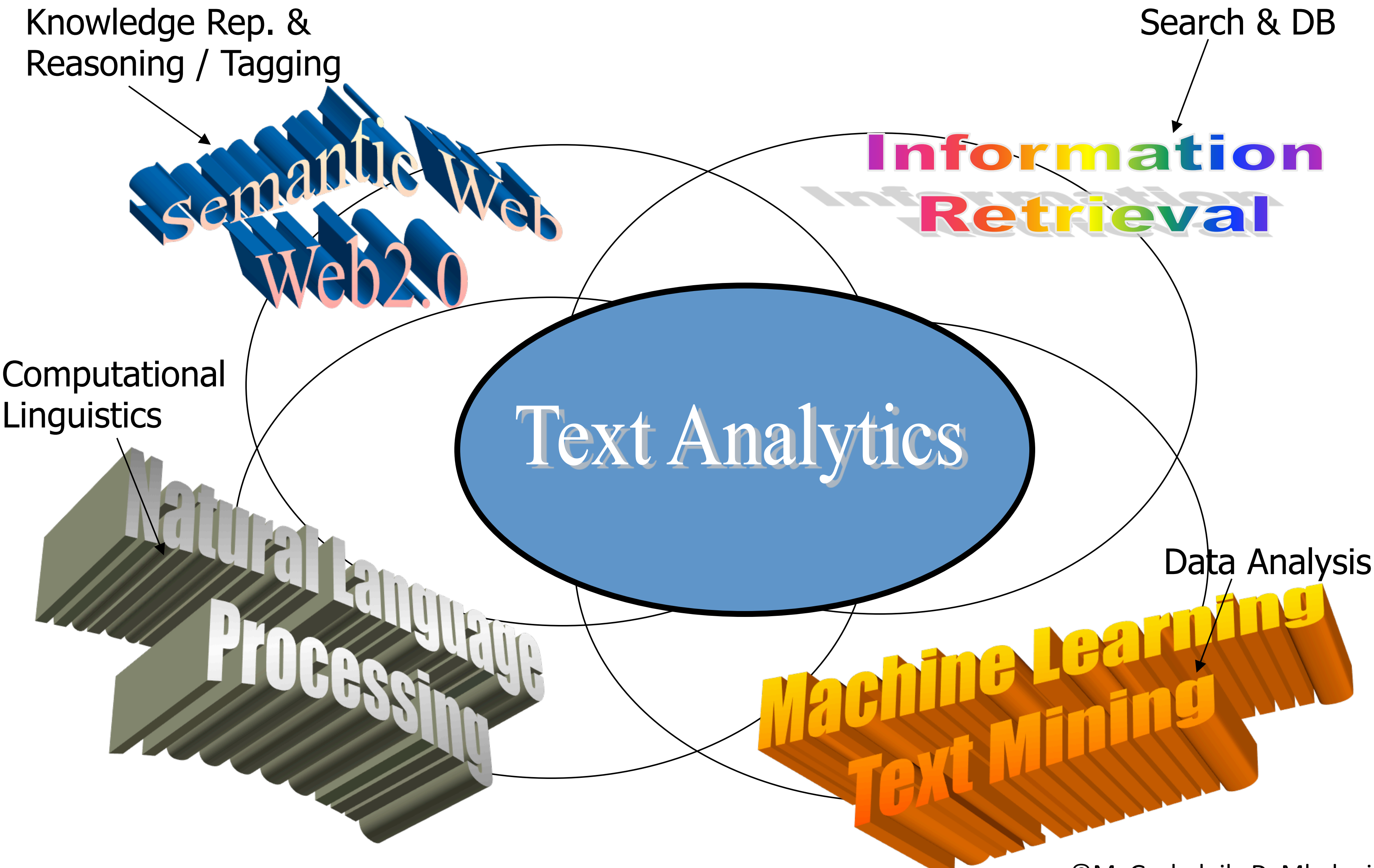
- Unstructured
  - Hard to parse, sometimes wrong grammar
- Ambiguous: often has no defined or precise semantics (as opposed to source code)
  - Hard to understand
- Many ways to represent similar concepts
  - Hard to extract information from

```
/* We need to acquire the write IRQ lock before calling ep_unlink(). */
/* Lock must be acquired on entry to this function. */
/* Caller must hold instance lock! */
```

Why Analyzing NL Data is Easy(?)

- Redundant data
- Easy to get “good” results for simple tasks
  - Simple algorithms without much tuning effort
- Evolution/version history readily available
- Many techniques to borrow from text analytics: NLP, Machine Learning (ML), Information Retrieval (IR), etc.

Text Analytics



Why Analyzing NL Data is Hard(?)

- Domain specific words/phrases, and meanings
  - “**Call** a function” vs. call a friend
  - “Computer **memory**” vs. human memory
  - “This method also returns **false** if path is **null**”
- Poor quality of text
  - Inconsistent
  - grammar mistakes
    - “**true** if path is an absolute path; otherwise false” for the File class in .NET framework
  - Incomplete information

Some Major NLP/Text Analytics Tools

IBM jStart

BIGSHEETS

sas THE POWER TO KNOW.

Text Miner

SPSS AN IBM® COMPANY

Text Analytics for Surveys

Stanford University Natural Language Processing NLP

Stanford Parser

http://nlp.stanford.edu/software/lex-parser.shtml

UIMA Unstructured Information Management Architecture An Apache Project

http://uima.apache.org/

http://nlp.stanford.edu/links/statnlp.html
http://www.kdnuggets.com/software/text.html

Dimensions in Text Analytics

- Three major dimensions of text analytics:
  - Representations
    - …from words to partial/full parsing
  - Techniques
    - …from manual work to learning
  - Tasks
    - …from search, over (un-)supervised learning, summarization, …

Major Text Representations

- Words (stop words, stemming)
- Part-of-speech tags

- Chunk parsing (chunking)
- Semantic role labeling
- Vector space model

Words' Properties

- Relations among word surface forms and their senses:
  - **Homonymy**: same form, but different meaning (e.g. bank: river bank, financial institution)
  - **Polysemy**: same form, related meaning (e.g. bank: blood bank, financial institution)
  - **Synonymy**: different form, same meaning (e.g. singer, vocalist)
  - **Hyponymy**: one word denotes a subclass of an another (e.g. breakfast, meal)
- General thesaurus: WordNet, existing in many other languages (e.g. EuroWordNet)
  - http://wordnet.princeton.edu/
  - http://www.illc.uva.nl/EuroWordNet/

Stop Words

- Stop words are words that from non-linguistic view do not carry information
  - …they have mainly functional role
  - …usually we remove them to help mining techniques to perform better

- Stop words are language dependent – examples:
  - **English**: A, ABOUT, ABOVE, ACROSS, AFTER, AGAIN, AGAINST, ALL, ALMOST, ALONE, ALONG, ALREADY, ...

Stemming

- Different forms of the same word are usually problematic for text analysis, because they have different spelling and similar meaning (e.g. learns, learned, learning,…)
- Stemming is a process of transforming a word into its stem (normalized form)
  - …stemming provides an inexpensive mechanism to merge

Stemming cont.

- For English is mostly used Porter stemmer at http://www.tartarus.org/~martin/PorterStemmer/
- Example cascade rules used in English Porter stemmer
  - ATIONAL -> ATE relational -> relate
  - TIONAL -> TION conditional -> condition
  - ENCI -> ENCE valenci -> valence
  - ANCI -> ANCE hesitanci -> hesitance
  - IZER -> IZE digitizer -> digitize
  - ABLI -> ABLE conformabli -> conformable
  - ALLI -> AL radicalli -> radical
  - ENTLI -> ENT differentli -> different
  - ELI -> E vileli -> vile
  - OUSLI -> OUS analogousli -> analogous

Part-of-Speech Tags

- Part-of-speech tags specify word types enabling to differentiate words functions
  - For text analysis, part-of-speech tag is used mainly for “information extraction” where we are interested in e.g., named entities (“noun phrases”)
  - Another possible use is reduction of the vocabulary (features)
    - …it is known that nouns carry most of the information in text documents
- Part-of-Speech taggers are usually learned on manually tagged data

Part-of-Speech Table

part of speech	function or "job"	example words	example sentences
Verb	action or state	(to) be, have, do, like, work, sing, can, must	EnglishClub.com **is** a web site. I **like** EnglishClub.com.
Noun	thing or person	pen, dog, work, music, town, London, teacher, John	This is my **dog**. He lives in my **house**. We live in **London**.
Adjective	describes a noun	a/an, the, 69, some, good, big, red, well, interesting	My dog is **big**. I like **big** dogs.
Adverb	describes a verb, adjective or adverb	quickly, silently, well, badly, very, really	My dog eats **quickly**. When he is **very** hungry, he eats **really** quickly.
Pronoun	replaces a noun	I, you, he, she, some	Tara is Indian. **She** is beautiful.
Preposition	links a noun to another word	to, at, after, on, but	We went **to** school **on** Monday.
Conjunction	joins clauses or sentences or words	and, but, when	I like dogs **and** I like cats. I like cats **and** dogs. I like dogs **but** I don't like cats.
Interjection	short exclamation, sometimes inserted into a sentence	oh!, ouch!, hi!, well	**Ouch**! That hurts! **Hi**! How are you? **Well**, I don't know.

http://www.englishclub.com/grammar/parts-of-speech_1.htm http://www.clips.ua.ac.be/pages/mbsp-tags

Part-of-Speech Examples

verb
Stop!

noun	verb
John	works.

noun	verb	verb
John	is	working.

pronoun	verb	noun
She	loves	animals.

noun	verb	adjective	noun
Animals	like	kind	people.

noun	verb	noun	adverb
Tara	speaks	English	well.

noun	verb	adjective	noun
Tara	speaks	good	English.

pronoun	verb	preposition	adjective	noun	adverb
She	ran	to	the	station	quickly.

pron.	verb	adj.	noun	conjunction	pron.	verb	pron.
She	likes	big	snakes	but	I	hate	them.

Here is a sentence that contains every part of speech:

interjection	pron.	conj.	adj.	noun	verb	prep.	noun	adverb
Well,	she	and	young	John	walk	to	school	slowly.

http://www.englishclub.com/grammar/parts-of-speech_2.htm

Part of Speech Tags

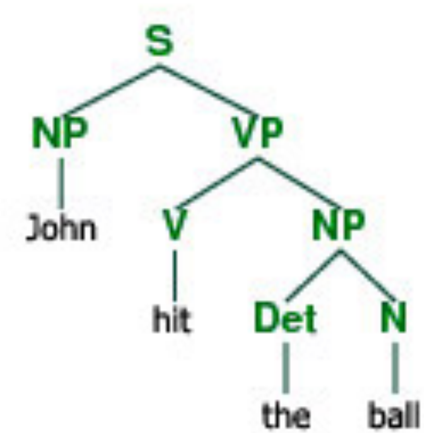
Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	$
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	“	Left quote	(‘ or “)
POS	Possessive ending	*'s*	”	Right quote	(’ or ”)
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	(],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

http://www2.sis.pitt.edu/~is2420/class-notes/2.pdf

Full Parsing

- Parsing provides maximum structural information per sentence
- Input: a sentence → output: a parse tree
- For most text analysis techniques, the information in parse trees is too complex

- Problems with full parsing:
  - Low accuracy
  - Slow
  - Domain Specific



Chunk Parsing

- Break text up into non-overlapping contiguous subsets of tokens.
  - aka. partial/shallow parsing, light parsing.
- What is it useful for?
  - Entity recognition
    - people, locations, organizations
  - Studying linguistic patterns
    - gave NP
    - gave up NP in NP
    - gave NP NP
    - gave NP to NP
  - Can ignore complex structure when not relevant

Chunk Parsing

Goal: divide a sentence into a sequence of chunks.

- Chunks are non-overlapping regions of a text

  [I] saw [a tall man] in [the park]

- Chunks are non-recursive
  - A chunk cannot contain other chunks
- Chunks are non-exhaustive
  - Not all words are included in the chunks

Chunk Parsing Techniques

- Chunk parsers usually ignore lexical content
- Only need to look at part-of-speech tags

- Techniques for implementing chunk parsing
  - E.g., Regular expression matching

Regular Expression Matching

- Define a regular expression that matches the sequences of tags in a chunk
  - A simple noun phrase chunk regrexp:

    <DT> ? <JJ> * <NN.?>
- Chunk all matching subsequences:

The /DT little /JJ cat /NN sat /VBD on /IN the /DT mat /NN

[The /DT little /JJ cat /NN] sat /VBD on /IN [the /DT mat /NN]

- If matching subsequences overlap, the first one gets priority

DT: Determinner   JJ: Adjective   NN: Noun, sing, or mass
VBD: Verb, past tense   IN: Prepostion/sub-conj Verb

Semantic Role Labeling

Giving Semantic Labels to Phrases

- [$_{\text{AGENT}}$ John] **broke** [$_{\text{THEME}}$ the window]
- [$_{\text{THEME}}$ The window] **broke**
- [$_{\text{AGENT}}$Sotheby's] .. **offered** [$_{\text{RECIPIENT}}$ the Dorrance heirs] [$_{\text{THEME}}$ a money-back guarantee]
- [$_{\text{AGENT}}$ Sotheby's] **offered** [$_{\text{THEME}}$ a money-back guarantee] to [$_{\text{RECIPIENT}}$ the Dorrance heirs]
- [$_{\text{THEME}}$ a money-back guarantee] **offered** by [$_{\text{AGENT}}$ Sotheby's]
- [$_{\text{RECIPIENT}}$ the Dorrance heirs] will [$_{\text{ARM-NEG}}$ not] be **offered** [$_{\text{THEME}}$ a money-back guarantee]

Semantic Role Labeling Good for
Question Answering

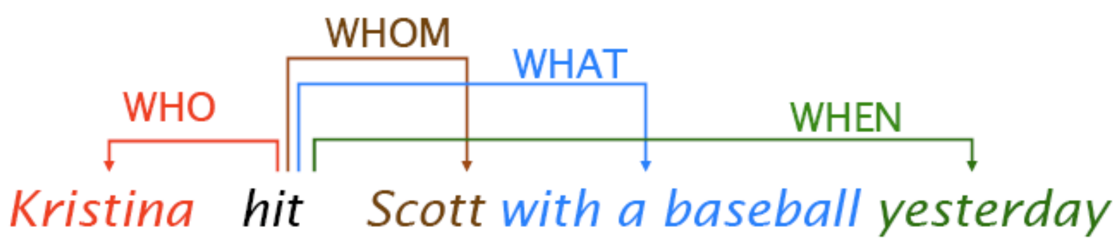
Q: What was the name of the first computer system that defeated Kasparov?

A: [PATIENT Kasparov] was defeated by [AGENT Deep Blue] [TIME in 1997]

Q: When was Napoleon defeated?

Look for: [PATIENT Napoleon] [PRED **defeat-synset**] [ARGM-TMP *ANS*]

More generally:



- Who hit Scott with a baseball?
- Whom did Kristina hit with a baseball?
- What did Kristina hit Scott with?
- When did Kristina hit Scott with a baseball?

Typical Semantic Roles

Thematic Role	Definition
AGENT	The volitional causer of an event
EXPERIENCER	The experiencer of an event
FORCE	The non-volitional causer of the event
THEME	The participant most directly affected by an event
RESULT	The end product of an event
CONTENT	The proposition or content of a propositional event
INSTRUMENT	An instrument used in an event
BENEFICIARY	The beneficiary of an event
SOURCE	The origin of the object of a transfer event
GOAL	The destination of an object of a transfer event

Example Semantic Roles

Thematic Role	Example
AGENT	*The waiter* spilled the soup.
EXPERIENCER	*John* has a headache.
FORCE	*The wind* blows debris from the mall into our yards.
THEME	Only after Benjamin Franklin broke *the ice*...
RESULT	The French government has built a *regulation-size baseball diamond*...
CONTENT	Mona asked *"You met Mary Ann at a supermarket?"*
INSTRUMENT	He turned to poaching catfish, stunning them *with a shocking device*...
BENEFICIARY	Whenever Ann Callahan makes hotel reservations *for her boss*...
SOURCE	I flew in *from Boston*.
GOAL	I drove *to Portland*.

Outline

- Introduction
- Background on text analytics
- Case Study 1: App Markets
- Case Study 2: ACP Rules
- Wrap-up

Case Study: App Markets

- App Markets have played an important role in the popularity of mobile devices
- Provide users with a textual description of each application's functionality

Apple App Store

Google Play

Microsoft Windows Phone

Current Practice

- Apple: market’s responsibility
  - Apple performs manual inspection
- Google: user’s responsibility
  - Users approve permissions for security/privacy
  - Bouncer (static/dynamic malware analysis)
- Windows Phone: hybrid
  - Permissions / manual inspection

Is Program Analysis Sufficient?

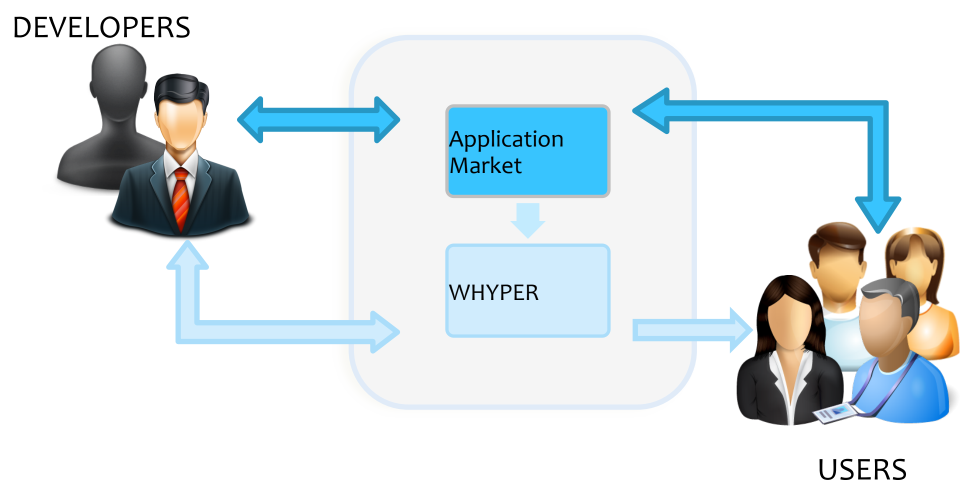
- Previous approaches look at permissions, code, and runtime behaviors
- Caveat: *what does the user expect?*
  - GPS Tracker: record and send location
  - Phone-call Recorder: record audio during call
  - One-Click Root: exploit vulnerability
  - Others are more subtle

Vision

- Goal: *bridge gap between user expectation and app behavior*
- WHYPER is a first step in this direction
- Focus on permission and app descriptions
  - Limited to permissions that protect "user understandable" resources

Use Cases

- Enhance user experience while installing apps
- Functionality disclosure to during application submission to market
- Complementing program analysis to ensure more appropriate justifications



Straw man: Keyword Search

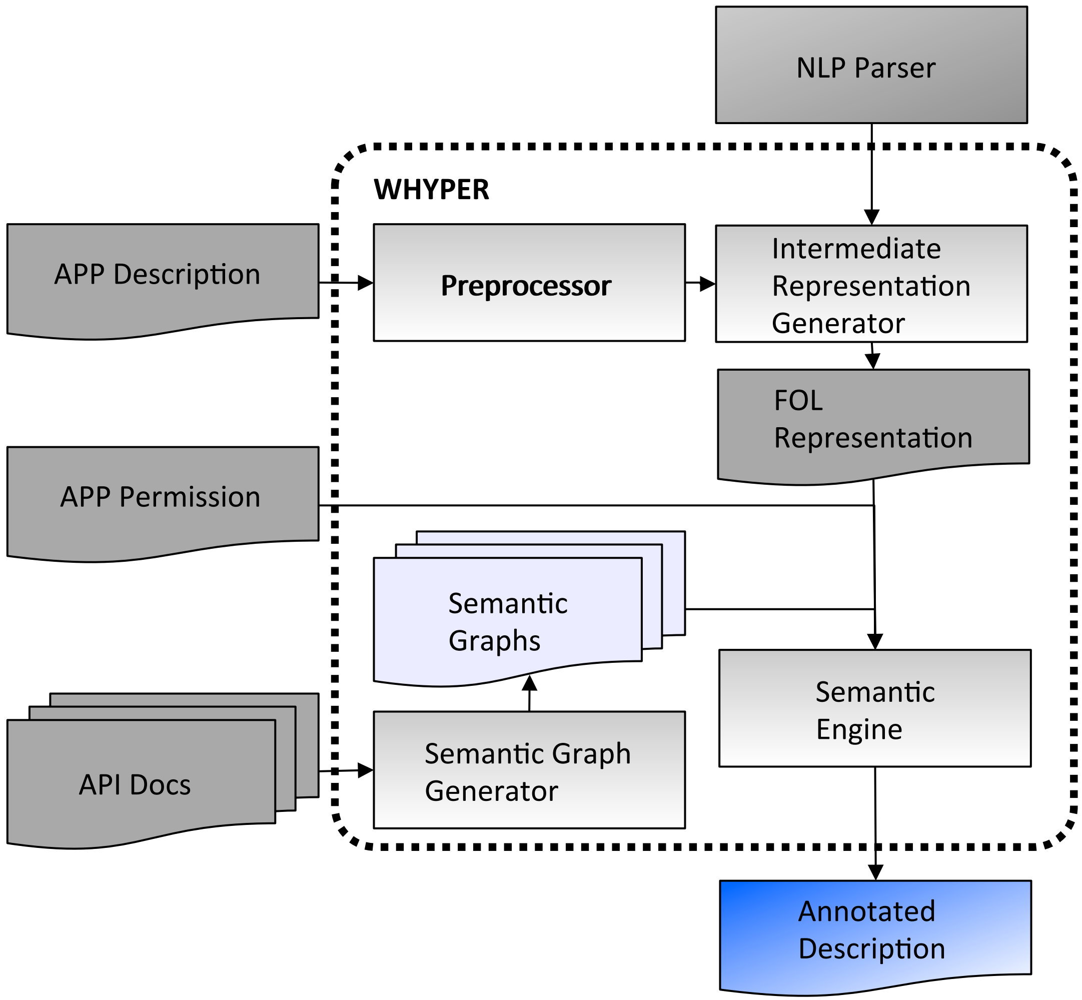
- Confounding effects:
  - Certain keywords such as “contact” have a confounding meaning, e.g.,

    “... displays user contacts, ...” vs “... contact me at abc@xyz.com”

- Semantic Interference:
  - Sentences often describe a sensitive operation such as reading contacts without actually referring to the keyword “contact”, e.g.,

    “share yoga exercises with your friends via email, sms”

WHYPER Framework

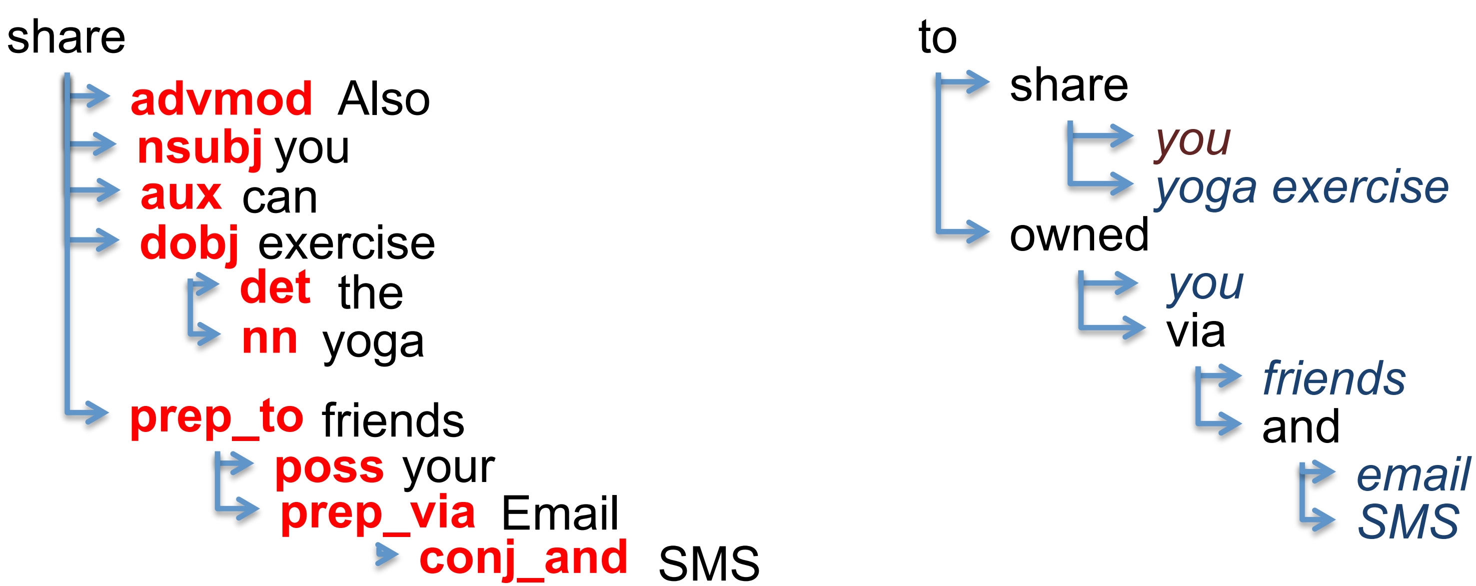


Preprocessor

- Period Handling
  - Decimals, ellipsis, shorthand notations (Mr., Dr.)
- Sentence Boundaries
  - Tabs, bullet points, delimiters (:)
  - Symbols (*,-) and enumeration sentence
- Named Entity Handling
  - E.g., “Pandora internet radio”
- Abbreviation Handling
  - E.g., “Instant Message (IM)”

Intermediate Representation Generator

Also	you	can	share	the	yoga	exercise	to	your	friends	via	Email	and	SMS
RB	**PRP**	**MD**	**VB**	**DT**	**NN**	**NN**		**PRP**	**NNS**		**NNP**		**NNP**



RB: adverb; PRP: pronoun; MD: verb, modal auxillary; VB: verb, base form; DT: determiner; NN: noun, singular or mass; NNS: noun, plural; NNP: noun, proper singular
http://www.clips.ua.ac.be/pages/mbsp-tags

Semantic-Graph Generator

public static class Developers

Summary: Nested Classes | Constants | Inherited Constants | Fields | Methods | Inherited Methods | [Expand All]

Added in API level 5

ContactsContract.Contacts

extends Object
implements BaseColumns ContactsContract.ContactNameColumns ContactsContract.ContactOptionsColumns ContactsContract.ContactStatusColumns ContactsContract.ContactsColumns

java.lang.Object
↳android.provider.ContactsContract.Contacts

Class Overview

Constants for the contacts table, which contains a record per aggregate of raw contacts representing the same person.

Operations

Insert

A Contact cannot be created explicitly. When a raw contact is inserted, the provider will first try to find a Contact representing the same person. If one is found, the raw contact's `CONTACT_ID` column gets the _ID of the aggregate Contact. If no match is found, the provider automatically inserts a new Contact and puts its _ID into the `CONTACT_ID` column of the newly inserted raw contact.

Update

Only certain columns of Contact are modifiable: `TIMES_CONTACTED`, `LAST_TIME_CONTACTED`, `STARRED`, `CUSTOM_RINGTONE`, `SEND_TO_VOICEMAIL`. Changing any of these columns on the Contact also changes them on all constituent raw contacts.

Delete

Be careful with deleting Contacts! Deleting an aggregate contact deletes all constituent raw contacts. The corresponding sync adapters will notice the deletions of their respective raw contacts and remove them from their back end storage.

Query

- If you need to read an individual contact, consider using `CONTENT_LOOKUP_URI` instead of `CONTENT_URI`.
- If you need to look up a contact by the phone number, use `PhoneLookup.CONTENT_FILTER_URI`, which is optimized for this purpose.
- If you need to look up a contact by partial name, e.g. to produce filter-as-you-type suggestions, use the `CONTENT_FILTER_URI` URI.

Semantic-Graph Generator

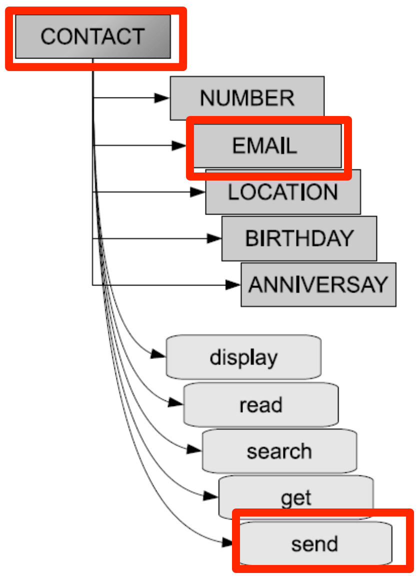
- Systematic approach to infer graphs
  - Find related API documents using Pscout [CCS'12]
  - Identify resource associated with permissions from the API class name
    - `ContactsContract.Contacts`
  - Inspect the member variables and member methods to identify actions and subordinate resources
    - `ContactsContract.CommonDataKinds.Email`

Semantic Engine

“Also you can share the yoga exercise to your friends via Email and SMS.”

- to
  - share
    - *you*
    - *yoga exercise*
  - owned
    - *you*
    - via
      - *friends*
      - and
        - *email*
        - *SMS*

WordNet Similarity



Evaluation

- Subjects
  - Permissions: READ_CONTACTS, READ_CALENDAR, RECORD_AUDIO
  - 581/600* application descriptions (English only)
  - 9,953 sentences
- Research Questions
  - RQ1: What are the precision, recall, and F-Score of WHYPER in identifying permission sentences?
  - RQ2: How effective is WHYPER in identifying permission sentences, compared to keyword-based searching

Subject Statistics

Permissions	#N	#S	S_p
READ_CONTACTS	190	3,379	235
READ_CALENDAR	191	2,752	283
RECORD_AUDIO	200	3,822	245
TOTAL	581	9,953	763

RQ1 Results: Effectiveness

Permission	S_I	TP	FP	FN	TN	Prec.	Recall	F-Score	Acc
READ_CONTACTS	204	186	18	49	2,930	91.2	79.2	84.8	97.9
READ_CALENDAR	288	241	47	42	2,422	83.7	85.2	84.5	96.8
RECORD_AUDIO	259	195	64	50	3,470	75.3	79.6	77.4	97.0
TOTAL	751	622	129	141	9,061	82.8	81.5	82.2	97.3

- Out of 9,061 sentences, only 129 flagged as FPs
- Among 581 apps, 109 apps (18.8%) contain at least one FP
- Among 581 apps, 86 apps (14.8%) contain at least one FN

R2 Results: Comparison to Keyword-based search

Permission	Keywords
READ_CONTACTS	contact, data, number, name, email
READ_CALENDAR	calendar, event, date, month, day, year
RECORD_AUDIO	record, audio, voice, capture, microphone

Permission	Delta Precision	Delta Recall	Delta F-score	Delta Accuracy
READ_CONTACTS	50.4	1.3	31.2	7.3
READ_CALENDAR	39.3	1.5	26.4	9.2
RECORD_AUDIO	36.9	-6.6	24.3	6.8
WHYPER Improvement	41.6	-1.2	27.2	7.7

Results Analysis: False Positives

- Incorrect Parsing
  - “MyLink Advanced provides full synchronization of all Microsoft Outlook emails (inbox, sent, outbox and drafts), contacts, calendar, tasks and notes with all Android phones via USB”
- Synonym Analysis
  - “You can now turn recordings into ringtones.”

Results Analysis: False Negatives

- Incorrect parsing
  - Incorrect identification of sentence boundaries and limitations of underlying NLP infrastructure
- Limitations of Semantic Graphs
  - Manual Augmentation
    - Microphone (*blow into*) and call (*record*)
    - Significant improvement of delta recalls: -6.6% to 0.6%
  - Future: automatic mining from user comments and forums

Broader Applicability

- Generalization to other permissions
  - User-understandable permissions: calls, SMS
  - Problem areas
    - Location and phone identifiers (widely abused)
    - Internet (nearly every app requires)

Dataset and Paper

- Our code and datasets are available at https://sites.google.com/site/whypermission/

- Rahul Pandita, Xusheng Xiao, Wei Yang, William Enck, and Tao Xie. **WHYPER: Towards Automating Risk Assessment of Mobile Applications.** *In Proc. 22nd USENIX Security Symposium (USENIX Security 2013)* *http://www.enck.org/pubs/pandita-sec13.pdf*

Outline

- Introduction
- Background on text analytics
- Case Study 1: App Markets
- Case Study 2: ACP Rules
- Wrap-up

Access Control Policies (ACP)

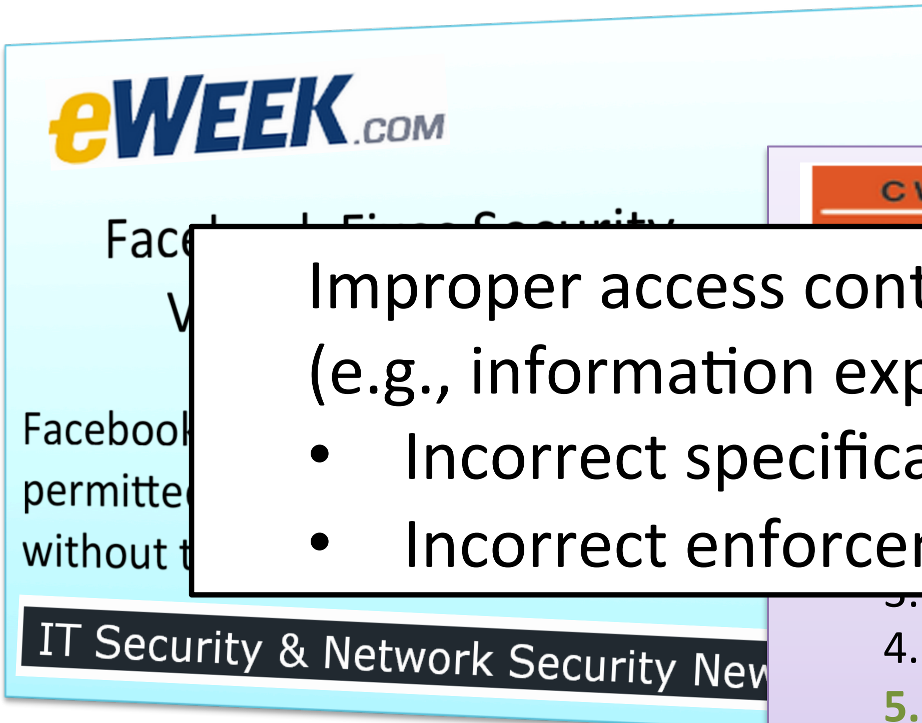
- Access control is often governed by security policies called Access Control Policies (ACP)
  - Includes rules to control which <u>principals</u> have <u>access</u> to which <u>resources</u>

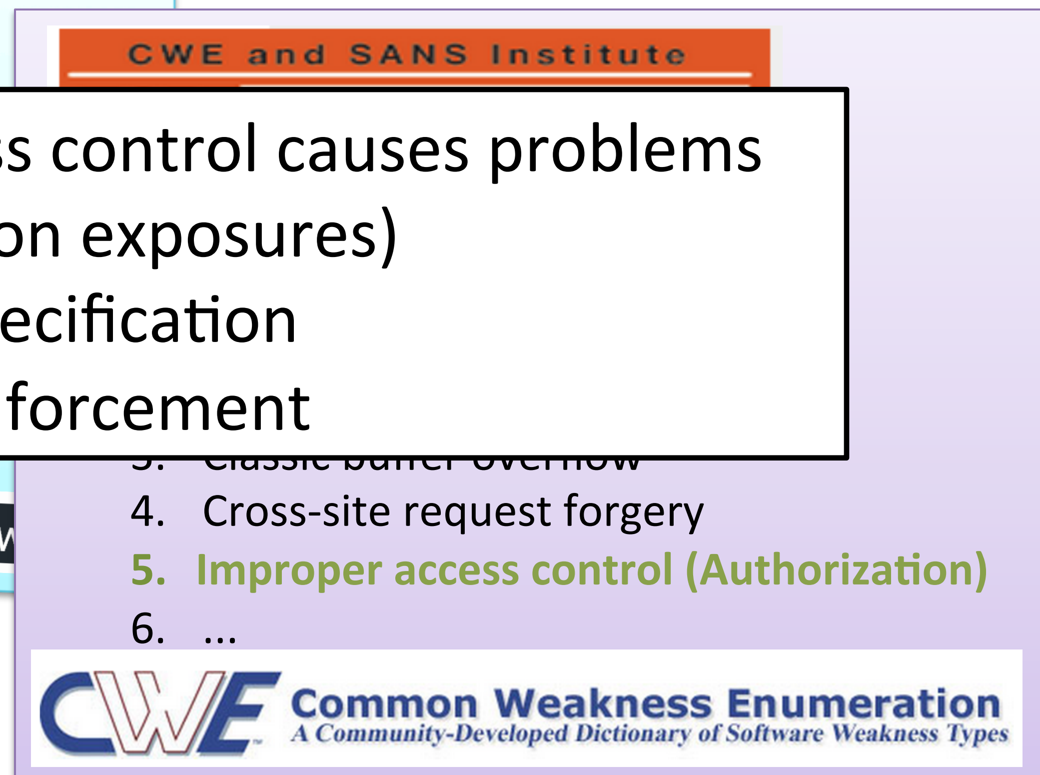
ex.

> "The Health Care Personnel (HCP) does not have the ability to edit the patient's account."

- A policy rule includes four elements
  - Subject – HCP
  - Action – edit
  - Resource - patient's account
  - Effect - deny

Access Control Vulnerabilities





Improper access control causes problems (e.g., information exposures)

- Incorrect specification
- Incorrect enforcement

Problems of ACP Practice

- In practice, ACPs
  - Buried in requirement documents
  - Written in NL and not checkable
- NL documents could be large in size
  - Manual extraction is labor-intensive and tedious

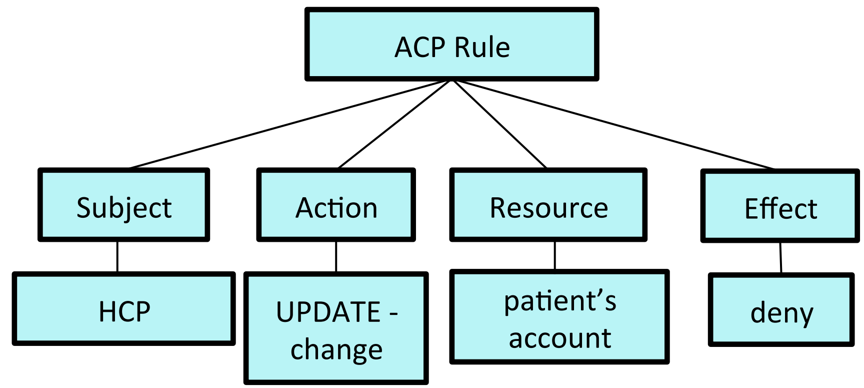
Overview of Text2Policy

Linguistic Analysis

A HCP should not change patient's account.

An [*subject*: HCP] should not [*action*: change] [*resource*: patient's account].

Model-Instance Construction




```xml
<Policy PolicyId="ACP2" RuleCombAlgId="deny-overrides">
  <Target/>
  <Rule Effect="Deny" RuleId="rule-1">
    <Target>
      <Subjects><Subject><SubjectMatch MatchId="string-equal">
            <AttrValue>HCP</AttrValue>
            <SubjectAttrDesignator.../></SubjectMatch></Subject>
      </Subjects>
      <Resources><Resource><ResourceMatch MatchId="string-equal">
            <AttrValue>patient.account</AttrValue>
            <ResourceAttrDesignator.../></ResourceMatch></Resource>
      </Resources>
      <Actions><Action><ActionMatch MatchId="string-equal">
            <AttrValue DataType="string">UPDATE</AttrValue>
            <ActionAttriDesignator.../></ActionMatch></Action>
      </Actions>
    </Target></Rule></Policy>
```

Transformation

Linguistic Analysis

- Incorporate syntactic and semantic analysis
  - **syntactic** structure -> noun group, verb group, etc.
  - **semantic** meaning -> subject, action, resource, negative meaning, etc.

- Provide New techniques for model extraction
  - Identify ACP sentences
  - Infer semantic meaning

Common Techniques

- Shallow parsing
- Domain dictionary
- Anaphora resolution

NP VG PNP

An HCP can view patient’s account.
He is disallowed to change the patient’s account.

HCP UPDATE

Subject Main Verb Group Object

http://www.clips.ua.ac.be/pages/mbsp-tags

NP: noun phrase
VG: verb chunk
PNP: prepositional noun phrase

Technical Challenges (TC) in ACP Extraction

ACP 1: An HCP cannot change patient's account.
ACP2: An HCP is disallowed to change patient's account.

- TC1: Semantic Structure Variance
  - different ways to specify the same rule
- TC2: Negative Meaning Implicitness
  - verb could have negative meaning

Semantic-Pattern Matching

- Address TC1 Semantic Structure Variance

- Compose pattern based on grammatical function

ex. An HCP is disallowed to change the patient's account.

passive voice *followed by* to-infinitive phrase

Negative-Expression Identification

- Address TC2 Negative Meaning Implicitness

- Negative expression
  - "not" in subject:

    ex. **No** HCP can edit patient's account.
  - "not" in verb group:

    ex. HCP can **not** edit patient's account.
    HCP can **never** edit patient's account.

- Negative meaning words in main verb group

ex. An HCP is **disallowed** to change the patient's account.

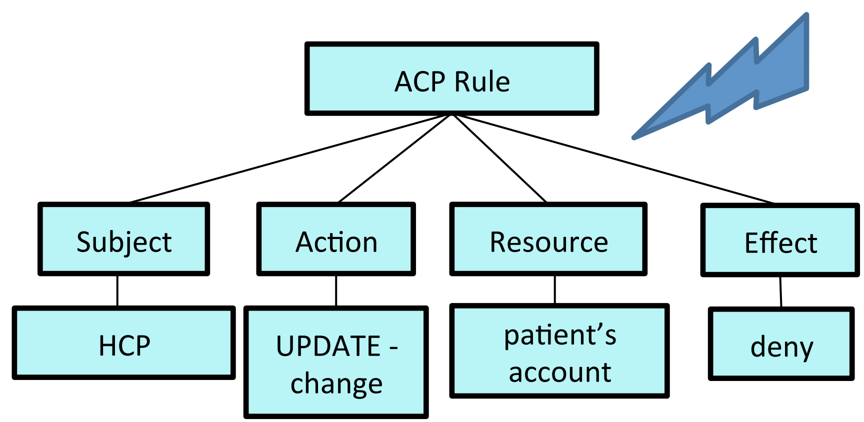
Overview of Text2Policy

Linguistic Analysis

A HCP should not change patient's account.

An [*subject*: HCP] should not [*action*: change] [*resource*: patient's account].

Model-Instance Construction



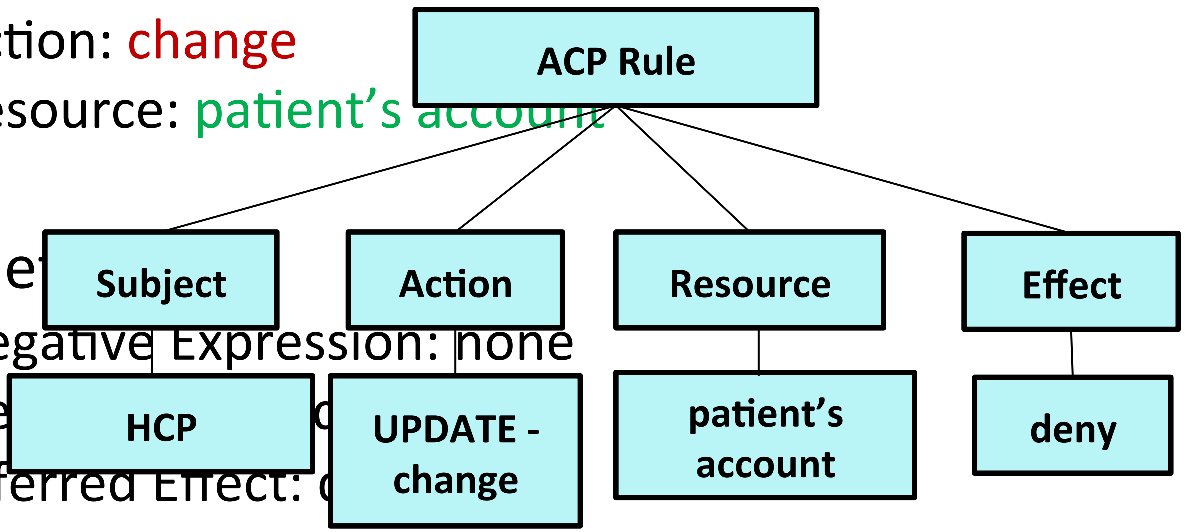

```xml
<Policy PolicyId="ACP2" RuleCombAlgId="deny-overrides">
  <Target/>
  <Rule Effect="Deny" RuleId="rule-1">
    <Target>
      <Subjects><Subject><SubjectMatch MatchId="string-equal">
            <AttrValue>HCP</AttrValue>
            <SubjectAttrDesignator.../></SubjectMatch></Subject>
      </Subjects>
      <Resources><Resource><ResourceMatch MatchId="string-equal">
            <AttrValue>patient.account</AttrValue>
            <ResourceAttrDesignator.../></ResourceMatch></Resource>
      </Resources>
      <Actions><Action><ActionMatch MatchId="string-equal">
            <AttrValue DataType="string">UPDATE</AttrValue>
            <ActionAttriDesignator.../></ActionMatch></Action>
      </Actions>
    </Target></Rule></Policy>
```

Transformation

ACP Model-Instance Construction

ex. An HCP is disallowed to change the patient's account.

- Identify subject, action, and resource:
  - Subject: HCP
  - Action: change
  - Resource: patient's account
- Infer e
  - Negative Expression: none
  - Ne
  - Inferred Effect: d



- Access Control Rule Extraction (ACRE) approach [ACSAC'14] discovers more patterns
  - Able to handle existing, unconstrained NL texts

Evaluation – RQs

- RQ1: How effectively does Text2Policy identify ACP sentences in NL documents?
- RQ2: How effectively does Text2Policy extract ACP rules from ACP sentences?

Evaluation – Subject

- iTrust open source project
  - http://agile.csc.ncsu.edu/iTrust/wiki/
  - 448 use-case sentences (37 use cases)
  - preprocessed use cases

- Collected ACP sentences
  - 100 ACP sentences
  - From 17 sources (published papers and websites)

- A module of an IBMApp (financial domain)
  - 25 use cases

RQ1 ACP Sentence Identification

- Apply Text2Policy to identify ACP sentences in iTrust use cases and IBMApp use cases

Subjects	# Sent.	# ACP Sent.	# Ident.	FP	FN	$Prec$	Rec	F_1
iTrust	448	117	119	16	14	86.6%	88.0%	87.3
IBMApp	479	24	23	0	1	100.0%	95.8%	97.9
Total	927	141	142	16	15	88.7%	89.4%	89.1

- Text2Policy effectively identifies ACP sentences with precision and recall more than 88%

- Precision on IBMApp use cases is better
  - proprietary use cases are often of higher quality compared to open-source use cases

Evaluation – RQ2 Accuracy of Policy Extraction

- Apply Text2Policy to extract ACP rules from ACP sentences

Subjects	# ACP Sent.	# Extracted	Accu.
iTrust	217	187	86.2%
IBMApp	24	21	87.5%
Total	241	208	86.3%

- Text2Policy effectively extracts ACP model instances with accuracy above 86%

Dataset and Paper

- Our datasets are available at https://sites.google.com/site/asergrp/projects/text2policy

- Xusheng Xiao, Amit Paradkar, Suresh Thummalapenta, and Tao Xie. **Automated Extraction of Security Policies from Natural-Language Software Documents**. In *Proc. 20th ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE 2012)* *http://web.engr.illinois.edu/~taoxie/publications/fse12-nlp.pdf*

- John Slankas, Xusheng Xiao, Laurie Williams, and Tao Xie. **Relation Extraction for Inferring Access Control Rules from Natural Language Artifacts**. In *Proc. 30th Annual Computer Security Applications Conference (ACSAC 2014)* *http://web.engr.illinois.edu/~taoxie/publications/acsac14-nlp.pdf*

Outline

- Introduction
- Background on text analytics
- Case Study 1: App Markets
- Case Study 2: ACP rules
- Wrap-up

Take-away

- Computing systems contain textual data that partially represents expectation context.
- Text analytics and natural language processing offers an opportunity to automatically extract that semantic context
  - Need to be careful in the security domain (e.g., social engineering)
  - But potential for improved security decisions

Future Directions

- Only beginning to study text analytics for security
  - Many sources of natural language text
  - Many unexplored domains
  - Use text analytics in software engineering as inspiration
    - https://sites.google.com/site/text4se/
- Hard problem: to what extent can we formalize "expectation context"?
- Creation of open datasets (annotation is time intensive)
- Apply to real-world problems

Thank you!

Questions?

William Enck
North Carolina State University
http://www.enck.org
enck@cs.ncsu.edu

Tao Xie
University of Illinois at Urbana-Champaign
http://web.engr.illinois.edu/~taoxie/
taoxie@illinois.edu

Acknowledgment: We thank authors of the original slides that some slides from this tutorial were adapted from. The work is supported in part by a Google Research Faculty Award, NSA Science of Security Lablet grants, NSF grants CCF-1349666, CCF-1409423, CNS-1434582, CCF-1434596, CCF- 1434590, CNS-1439481, CNS-1253346, CNS-1222680