

Making Sense of Unstructured Data

Dan Roth

Department of Computer Science

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN



Information Access and Synthesis

- Most of the data today is unstructured
 - **books, newspaper articles, journal publications, reports, transcribed audio streams, images, and video streams.**
- **How to deal with the huge amount of unstructured data?**
- This is a collection of problems that everyone cares about:
 - Natural Scientists, social scientists, humanity scholars, engineers
 - Analysts of different sorts
 - All agencies
 - First responders
 - Business planning and intelligence
- **A Natural Language Perspective**

Information Access and Synthesis

In early November, American intelligence authorities say they learned from a communications intercept of Qaeda followers in Yemen that a man named “Umar Farouk” — the first two names of the jetliner suspect, Umar Farouk Abdulmutallab — had volunteered for a coming operation.

Different Names—same entity?

Different References—same entity?

Washington, Dec. 30 (ANI):

Agency (CIA) had received a lead on a person dubbed “The Nigerian” suspected for meeting “terrorist elements” in Yemen as early as August 2009, but it failed to establish a link when Umar Farouk Abdulmutallab’s father went to the U.S. Embassy in Nigeria in November to express concerns about his son’s ties with al Qaeda.

Information Access and Synthesis

The Christmas Day Nigerian bomber's failed attack on Northwest's Flight 253 makes us thankful that the plot was foiled by alert passengers, but shows that security in the U.S. is also a failure at critical times. Abdul Mudallad was known to be a threat, with possible al-Qaeda ties, for at least two years, was on a federal "watch list" but not on the "no-fly list." Would anything have made a difference? Mudallad boarded the flight in Amsterdam. Does it mean that....

In late December, more intercepts of Qaeda operatives focused their attacks in the region, mentioned the date of Dec. 25, and suggested that they were "looking for ways to get somebody out" or "for ways to move people to the West," one senior administration official said.

Almost—Natural Language Understanding (in context)

Information Access and Synthesis

But on Christmas Day, the final draft of the memorandum was still sitting in the computer of a junior C.I.A. analyst, waiting until a photo of the young Nigerian was located. Unbeknownst to the analyst, officials said, Mr. Abdulmutallab's photo had already been delivered to other counterterror

Can we integrate knowledge from TEXT with knowledge from IMAGES?

The American intelligence network was clearly listening in Yemen and sharing that

“The program not only can't connect the dots, it can't find the dots,” Representative Brad Miller, Democrat of North Carolina and chairman of a House panel that oversees the program, said at the time.

Information Access and Synthesis

- Most of the data today is unstructured
 - **books, newspaper articles, journal publications, reports, images, and audio and video streams.**
- How to deal with the huge amount of unstructured data **as if** it was organized in a database with a known schema.
 - **how to locate, organize, access and analyze unstructured data.**

Research Program:

- develop the theories, algorithms, and tools for analysts to
 - **Intelligently access a variety of data formats and models**
 - **integrate them with existing resources**
 - **transform raw **data** into useful and understandable **information.****

A view on Open Information Extraction

"as is, with all defects" basis, without maintenance, debugging, support or improvement. Licensee assumes the entire risk as to the results and performance of the Software and/or associated materials. Licensee agrees that

■ Given:

- A long contract that you need to **ACCEPT**

■ Determine: D

- Does it satisfy the 3 conditions that you really care about?

4.

5.

6.

7. This Agreement shall be construed and interpreted in accordance with the laws of the State of Illinois, U.S.A..

8. This Agreement shall be subject to all United States Government laws and regulations now and hereafter applicable to the subject matter of this Agreement, including specifically the Export Law provisions of the Departments of Commerce and State. Licensee will not export or re-export the Software without the appropriate United States or foreign government license.

By its registration below, Licensee confirms that it understands the terms and conditions of this Agreement, and agrees to be bound by them. This Agreement shall become effective as of the date of execution by Licensee.

Registration information: (We will not disclose any of this information. It is for internal use only.)

Name:

Email Address:

Organization:

Accept

Clear

ACCEPT?

Large Scale Understanding: Massive & Deep

Determine if Jim Carpenter works for the government



topic—intelligence



Jim Carpenter works for the U.S. Government.

The American government employed Jim Carpenter.

Jim Carpenter was fired by the US Government.

Jim Carpenter worked in a number of important positions. As a press liaison for the IRS, he made contacts in the white house.

Russian interior minister Yevgeny Topolov met yesterday with his US counterpart, Jim Carpenter.

Former US today...

Standard techniques cannot deal with the variability of expressing meaning nor with the ambiguity of interpretation

Needs:

- ❑ **Relations** and **Semantic Classes**, NOT keywords
- ❑ **Exhaustive** recovery of information needs
- ❑ Abstract over **variability** in natural language
- ❑ **Integrate** over large collections of text and DBs
- ❑ **Track** entities, events, etc.

This Talk

- Describe a few key NLP and IE problems and sketch our solutions
- Can't be comprehensive nor deep
- Natural Language Processing is hard.
 - Not sufficient to do something; need to do it well, **use common benchmarks and quantify performance.**
- I will point to several state-of-the-art systems
 - all are publicly available for research purposes
 - Demo & Software page of **MIAS** & the **Cognitive Computation Group**
- **Know about it & Use it – many have.**
 - Recent customers: VACCINE

- **Technical Approach**
- Problems & Solutions
 - **Semantic Role Labeling**
 - NER
 - **Reference & Co-reference**
 - Textual Entailment
 - **Events**
 - Multilingual
 - **Trust**
 - Text & Images

Machine Learning + Inference based NLP

- **Modeling and learning algorithms** for different phenomena
 - Classification models
 - Structured models
 - Supervised as well as semi-supervised and unsupervised models
 - Study of adaptation of models to new domain
- **Inference** as a way to introduce domain & task specific constraints
 - Using constrained optimization

Extracting Relations via Semantic Analysis

Semantic Role Labeling Output

Input Text:

A car bomb that exploded outside the U.S. military base in Beniji killed 11 Iraqi citizens.

Result: Complete!

General Explanation of Argument Labels

A	bomb [A1]	killer [A0]
car		
bomb		
that	bomb (Reference) [R-A1]	
exploded	V: explode	
outside	location [AM-LOC]	
the		
U.S.		
military	temporal [AM-TMP]	
base		
in	location [AM-LOC]	
Beniji		
killed		V: kill
11		corpse [A1]
Iraqi		
citizens		

Screen shot from a CCG demo

<http://L2R.cs.uiuc.edu/~cogcomp>

– Semantic parsing reveals several **relations** in the sentence along with their **arguments**.

Top system in the CoNLL Shared Task Competition 2005

■ This level of analysis, however, cannot abstract over the inherent variability in expressing the relations. **Kill** and **Explode** can be expressed in many different ways.

Extended Semantic Role Labeling

Who did what to whom, why, where, when...

Jim Carpenter **works** for the U.S. Government.

Jim Carpenter, the Dept. of State spokesperson, **was seen** in Mexico City on Wednesday

Jim Carpenter **of** the US Department of State said today that

Russian interior minister Yevgeny Topolov **met** yesterday with *his US counterpart, Jim Carpenter*.

Jim Carpenter's employment with the American government ...

Standard Semantic Role Labeling is defined in terms of verbs only.

A	bomb [A1]	killer [A0]
car		
bomb		
that	bomb (Reference) [R-A1]	
exploded	V: explode	
outside	location [AM-LOC]	
the		
U.S.		
military base	temporal [AM-TMP]	
in	location [AM-LOC]	
Benji		
killed		V: kill
11		corpse [A1]
Iraqi citizens		

Prepositions, commas, nominalizations and possessives also express relations.

Task:

- ❑ **Generalize SRL** to support more linguistic phenomena
- ❑ **Joint inference** across these tasks
- ❑ Address variability of training data across domains: essential given no unified resources for all phenomena

Named Entity Recognition.

SOCGER - [PER BLINKER] BAN LIFTED .
[LOC LONDON] 1996-12-06 [MISC Dutch] forward
[PER Reggie Blinker] had his indefinite suspension
lifted by [ORG FIFA] on Friday and was set to make
[ORG Sheffield Wednesday] comeback against
[ORG Liverpool] on Saturday . [PER Blinker] missed
his club's last two games after [ORG FIFA] slapped a
worldwide ban on him for appearing to sign contracts for
both [ORG Wednesday] and [ORG Udinese] while he was
playing for [ORG Feyenoord].

■ Key contribution

- Best system available today.
[Ratinov & Roth, CoNLL-08]
- Uses Wikipedia to augment entities lists and statistics
 - Other resources possible
- Uses unlabeled text

■ Identify

- Organizations
- Locations
- People
- Quantities & Temporal (+reasoning)
- ...

■ Main algorithmic questions:

- Representing Entities
- Exploiting non-local information.
- Injecting world knowledge.

The Reference Problem



Kennedy

The same problem exists with other types of entities



[Li, Morie, Roth, NAACL'04, AACL'04, AI Magazine'05]

Document 1: *The Justice Department has officially ended its inquiry into the assassinations of **John F. Kennedy** and Martin Luther King Jr., finding "no persuasive evidence" to support conspiracy theories, according to department documents. The House Assassinations Committee concluded in 1978 that **Kennedy** was "probably" assassinated as the result of a conspiracy involving a second gunman, a finding that broke from the **Warren Commission**'s belief that Lee Harvey Oswald acted alone in **Dallas** on Nov. 22, 1963.*

Document 2: *In 1953, Massachusetts **Sen. John F. Kennedy** married Jacqueline Lee Bouvier in Newport, R.I. In 1960, Democratic presidential candidate **John F. Kennedy** confronted the issue of his Roman Catholic faith by telling a Protestant group in Houston, "I do not speak for my church on public matters, and the church does not speak for me."*

Document 3: ***David Kennedy** was born in Leicester, England in 1959. ... **Kennedy** co-edited *The New Poetry* (Bloodaxe Books 1993), and is the author of *New Relations: The Refashioning Of British Poetry 1980-1994* (Seren 1996).*

Wikification: Linking Entities to Wikipedia

- Context sensitive disambiguation of entities and concepts
- Mapping to an authority file (here: Wikipedia)
- Key question: what entities/concepts to link?

Soccer -
 [LOC LON]
 [PER Regg]
 lifted by [OR]
 his [OR]
 [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord].



Regi Blinker		
Personal information		
Full name	Reginald Waldi Blinker	
Date of birth	4 June 1969	
Place of birth	Paramaribo, Suriname	
Height	1.73 m (5 ft 8 in)	
Playing position	Winger	
Club information		
Current club	Retired	
Senior career ¹		
Years	Club	App (Gls) ^a
1986–1996	Feyenoord	238 (45)
1988–1989	—Den Bosch (loan)	25 (6)
1996–1997	Sheffield Wednesday	42 (3)

Football career

Blinker began his career with Feyenoord Rotterdam in 1986. He stayed at *de Kuip* for 10 seasons, including one on loan at F. C. Den Bosch, and formed an efficient winger partnership with Gaston Taument (from 1991-95, the pair combined for 61 league goals).

Holland

From Wikipedia, the free encyclopedia

Holland is a name in common usage given to a region in the western part of the Netherlands. The name 'Holland' is also often informally used to refer to the whole of the country of the Netherlands. From the 10th century to the 16th century Holland proper was a unified political region, a county ruled by the Count of Holland. By the 17th century, Holland had risen to become a maritime and economic power, dominating the other provinces of the Dutch Republic. Today, the former County of Holland consists of the two Dutch provinces of North Holland and South Holland, which together include the Netherlands' three largest cities: country capital Amsterdam, seat of government The Hague, and Rotterdam, home of Europe's largest port.



Coordinates: 52.250°N 4.667°E



Udinese Calcio

From Wikipedia, the free encyclopedia

Udinese Calcio is an Italian football club based in Udine, Friuli-Venezia Giulia, and currently plays in the Serie A.

The traditional team home kit is black and white striped shirt, black shorts, and white socks. The club plays in the Stadio Friuli, which has a capacity of 41,652 (although it is currently limited to 30,900). It has a large number of fans in Friuli and surrounding areas, and it is sometimes seen as the best symbol of Friulian pride.



Udinese



Full name	Udinese Calcio SpA
Nickname(s)	<i>Bianconeri</i> ("White-blacks"), <i>Zebre</i> ("little zebras")
Founded	1896
Ground	Stadio Friuli, Udine, Italy (Capacity: 41,652 (current limit 30,900))

Current evaluation focuses on blogs data

Cross-document Co-reference: Dutch=Oranje?

Key Application: entity tracking

SOCCKER - [PER BLINKER] BAN LIFTED .

[LOC LONDON] 1996-12-06 [MISC Dutch] forward [PER Reggie Blinker] had his indefinite suspension lifted by [ORG FIFA] on Friday and was set to make his [ORG Sheffield Wednesday] comeback against [ORG Liverpool] on Saturday . [PER Blinker] missed his club's last two games after [ORG FIFA] slapped a worldwide ban on him for appearing to sign contracts for both [ORG Wednesday] and [ORG Udinese] while he was playing for [ORG Feyenoord] .

Regi Blinker played three matches for Oranje. The left winger started his pro career at Feyenoord and played 400 official matches for Feyenoord, Celtic and Sparta. He retired from football in 2003. Where is he now?



WIKIPEDIA
The Free Encyclopedia

Coordinates: 52.230°N 4.667°E

Holland

From Wikipedia, the free encyclopedia

Holland is a name in common usage given to a region in the western part of the Netherlands. The name 'Holland' is also often informally used to refer to the whole of the country of the Netherlands. From the 10th century to the 16th century Holland proper was a unified political region, a county ruled by the Count of Holland. By the 17th century, Holland had risen to become a maritime and economic power, dominating the other provinces of the Dutch Republic. Today, the former County of Holland consists of the two Dutch provinces of North Holland and South Holland, which together include the Netherlands' three largest cities: country capital Amsterdam, seat of government The Hague, and Rotterdam, home of Europe's largest port.

Contents

- 1 Etymology
- 2 Usage
- 3 Geography
- 4 Language
- 5 History



Co-reference Resolution (Within Documents)

The Problem:

identify phrases that refer to entities and cluster them according to those entities.

An American official announced that American President Bill Clinton met his Russian counterpart, Vladimir Putin, today. The president said that Russia was a great country.

The screenshot shows a web-based interface for Co-reference Resolution. At the top, there is a text input field containing the sentence: "An American official announced that American President Bill Clinton met his Russian counterpart, Vladimir Putin, today. The president said that Russia was a great country." Below the input field is a "Go" button. To the right of the input field is a list of "Entities Detected" with their counts: Russia(7), a great |country|(12), his Russian(0), Vladimir Putin(10), An American |official|(4), that(8), American President |Bill Clinton|(2), The |president|(1), American |President|(5), that(3), his(9), American(6), and American(11). Below the input field is a graph showing the relationships between mentions. The graph nodes are: "An American |official|", "that", "American", "American |President|", "American President |Bill Clinton|", "met", "his", "his Russian", "counterpart, Vladimir Putin, today.", "The |president|", "said", "that", "Russia", "was", and "a great |country|". Edges connect these nodes to show co-referent clusters.

The Solution:

- ❑ A Classification based pairwise model
- ❑ Followed by a decoding (inference) step with a single constraint,
- ❑ A thorough feature engineering

Details:

Coref classifier $C : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$ takes two mentions m_1 and m_2 as input. We allow each mention to be linked to at most one previous mention while maximizing the sum of linked scores.

Best system available today.
(Bengston & Roth, EMNLP'08)

Current work:

- ❑ Integrating within and across document co-ref
- ❑ Incorporating background knowledge into co-ref

Recognizing Textual Entailment

- Alternate formulation of Natural Language Understanding
- Instead of mapping text to a canonical (FOL) representation, answer a different question:
 - Given two text fragments, does the meaning of one fragment follow from the meaning of the other? E.g.

Is it true that...?

Eyeing the huge market potential, currently led by Google, Yahoo took over search company ~~Overture Services Inc.~~ last year

Yahoo acquired Overture
Overture is a search company
Google is a search company
Google owns Overture
.....

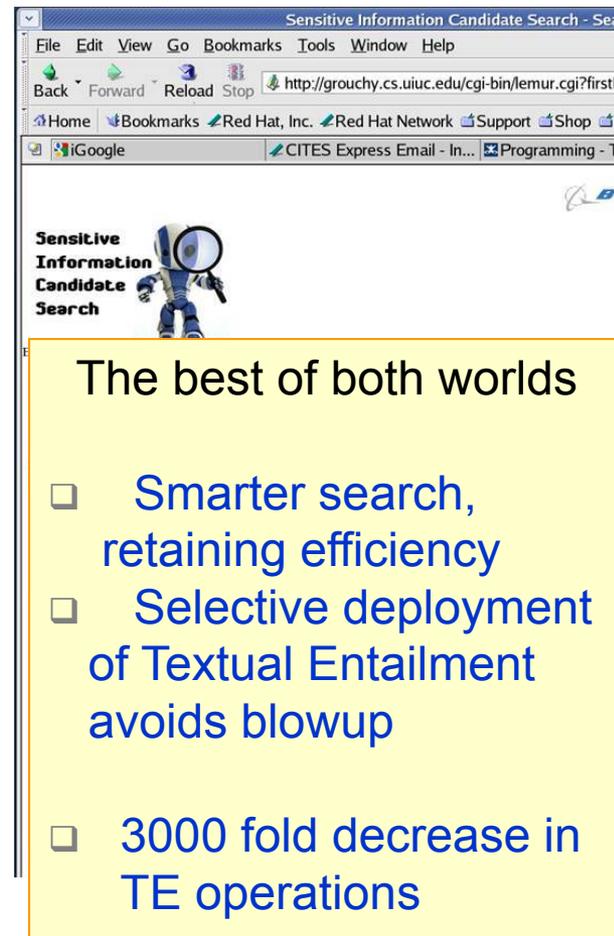
Application: Discovering Sensitive Information

Recent work: **Integrating Entailment and Search**

Need: Exhaustive extraction of information need

Open-ended information need

- **First Stage:** use **enhanced** Information Retrieval techniques to return K most relevant text spans – **Semantic Search (SR)**
- Multiple NLP technologies support richer indexing
- Index a range of **semantic units** (Named Entities, Phrasal Verbs, Multi-word expressions...)
- Use **specialized (Named Entity, phrasal verb, distributional) similarity measures** to build multiple IR indexes, and expand query
- **Second Stage:** use Textual Entailment (TE) to **filter** results



Sensitive Information Candidate Search

The best of both worlds

- Smarter search, retaining efficiency
- Selective deployment of Textual Entailment avoids blowup
- 3000 fold decrease in TE operations

Event Recognition

Tasks:

- Recognizing and classifying events in texts.
- De-duplicating recognized events.
- Tracking events.

BBC, Nov. 6th, 2009: Thirteen people have been killed and 30 injured in a shooting at a military base in Texas. News of the shooting has dominated the US media.

CNN, Nov. 6th,

Collaborations:

- START – Terrorism Events
- Democracy Institute: Societal Stability & Climate Change

injured more than his fellow soldiers. Another mass shooting in the United States in Orlando, Florida. But the deadliest terror-event particular at the US military base is shocking for every one, belonging to any religious community.

CNN, Nov. 7th, 2009: It was Munley who arrived quickly Thursday at the scene of the worst massacre at an Army base in U.S. history, where 13 people were killed. She confronted the alleged gunman, Maj. Nidal Malik Hasan, and shot him four times. Munley was wounded in the exchange.

Event Recognition

Technical Challenges:

- Recognition of **Entities**
- **Co-reference** resolution
- **Semantics** of event triggers and contexts

BBC, Nov. 6th, 2009: **Thirteen** people have been killed and **30** injured in a shooting at a military base in **Texas**. News of the shooting has dominated the **US** media.

CNN, Nov. 6th, 2009: **Major Nidal Malik Hasan**, killed **13** soldiers and injured more than **two dozen** on **November 5** by opening fire on **his** fellow soldiers at **Fort Hood** in a **US** military base. After fourteen hours, **two** people were killed and **six** were hurt in another mass shooting in the **United States** in **Orlando, Florida**. But the deadliest terror-event particular at the **US** military base is shocking for every one, belonging to any religious community.

CNN, Nov. 7th, 2009: It was **Munley** who arrived quickly **Thursday** at the scene of the worst massacre at an **Army** base in **U.S.** history, where **13** people were killed. **She** confronted the alleged gunman, **Maj. Nidal Malik Hasan**, and shot **him** four times. **Munley** was wounded in the exchange.

Entities:

Number

Date

Person

Organization

Location

Co-reference:



Event Recognition

Technical Challenges:

- Recognition of **Entities**
- **Co-reference** resolution
- **Semantics** of event triggers and contexts

BBC, Nov. 6th, 2009: **Thirteen people have been killed and 30 injured in a shooting at a military base in Texas.** News of the shooting has dominated the US media.

massacre

CNN, Nov. 6th, 2009: **Major Nidal Malik Hasan, killed 13 soldiers and injured more than two dozen on November 5 by opening fire on his fellow soldiers at Fort Hood in a US military base.** **After fourteen hours, two people were killed and six were hurt in another mass shooting in the United States in Orlando, Florida.** But the deadliest terror-event particular at the US military base is shocking for every one, belonging to any religious community.

massacre

massacre

CNN, Nov. 7th, 2009: It was Munley who arrived quickly Thursday at the scene of the worst massacre at an Army base in U.S. history, where 13 people were killed. **She confronted the alleged gunman, Maj. Nidal Malik Hasan, and shot him four times.** **Munley was wounded in the exchange.**

shooting

wound

Event Recognition

Technical Challenges:

- Recognition of **Entities**
- **Co-reference** resolution
- **Semantics** of event triggers and contexts

BBC, Nov. 6th, 2009: **Thirteen people have been killed and 30 injured in a shooting at a military base in Texas.** News of the shooting has dominated the US media.

massacre

CNN, Nov. 6th, 2009: **Major Nidal Malik Hasan, killed 13 soldiers and injured more than two dozen on November 5 by opening fire on his fellow soldiers at Fort Hood in a US military base. After fourteen hours, two people were killed and six were hurt in another mass shooting in the United States in Orlando, Florida.** But the deadliest terror-event particular at the US military base is shocking for every one, belonging to any religious community.

massacre

massacre

CNN, Nov. 7th, 2009: It was Munley who arrived quickly Thursday at the scene of the worst massacre at an Army base in U.S. history, where 13 people were killed. **She confronted the alleged gunman, Maj. Nidal Malik Hasan, and shot him four times. Munley was wounded in the exchange.**

shooting

wound

Named Entity Transliteration

- The process of transcribing a NE to a different language
- Key step in
 - transferring resources across languages
 - supporting multilingual access to information



Find Obama in the Hebrew Wikipedia

ברק אובאמה (הופנה מהדף אובמה)
 ברק חוסיי אובאמה השני (באנגלית: Barack Hussein Obama II; מולד ב-4 באוגוסט 1961) הוא נשיא ארצות הברית המכחי והארבעים וארבעה במניין. שימש גם כסנאטור אמריקני מטעם מדינת אילינוי.
 אובמה הושבע לנשיאה ה-44 של ארצות הברית ב-20 בינואר 2009, ובכך הפך לנשיא האפרו-אמריקאי^[1] הראשון בהיסטוריה של ארצות הברית.



- This process should capture phonetic and orthographic differences across the two languages
- **Not easy.** Context sensitive; time sensitive

Best system available, both unsupervised and supervised English—[Russian; Hebrew; Chinese] (NAACL'09, CIKM'09)

Application: Trust in on-line information



WIKIPEDIA
The Free Encyclopedia

navigation

- Main page
- Contents
- Featured content
- Current events
- Random article

Your *continued donations* keep Wikipedia

[article](#) [discussion](#) [view source](#)

Cancer

From Wikipedia, the free encyclopedia

For other uses, see *Cancer* (disambiguation)



Cancer (medical term: *malignant neoplasm*) is a group of diseases characterized by abnormal cell growth with the potential to invade or destroy other normal tissues of the body. These three malignant behaviors are commonly known as *metastasis*. Most cancers form from



WebMD
Better information. Better health.

October 07, 2009

Search

Other search tools: Symptoms | Doctors

[WebMD Home](#) > [Cancer Health Center](#) > [Lung Cancer Health Center](#) > [Lung Cancer News](#)

[Lung Cancer Health Home](#)

[Lung Cancer News](#)

[Lung Cancer Videos](#)

[Talk With Others About](#)

Lung Cancer Health Center

Erbitux Helps Treat Advanced Lung Cancer

Study Shows Benefits for Patients With Non-Small-Cell Lung Cancer

by [Charlene Laino](#)
WebMD Health News

Reviewed by [Louise Chang, MD](#)

FONT SIZE
A A A

[Home](#) [Privacy Statement](#) [Terms of Use](#) [Site Map](#)

Search

Sept. 23, 2009 (Berlin) -- Adding the targeted drug [Erbitux](#) to standard chemotherapy drugs significantly cuts the risk of death for advanced non-small-cell

[Contact Us / Careers](#) | [Site Map](#) | [Sign up for Site Updates](#)

Search

[text zoom](#) | [email page](#) | [print page](#)



chemotherapy.com
easing the chemotherapy journey

About Cancer
· [Understanding Cancer](#)

About Chemotherapy

» [Treating Cancer with Chemotherapy](#)

- [How People Receive Chemotherapy](#)
- [Chemotherapy Cycles and Schedules](#)
- [Remission and Goals of Cancer Therapy](#)

- [Chemotherapy Side Effects](#)
- [Treating Cancer in Other Ways](#)
- [Tracking Your Test Results](#)

Tools and Resources

- [Understanding Insurance and Tax Issues: Insurance Tips](#)
- [Web Resources and](#)

Treating Cancer With Chemotherapy

Many people fear chemotherapy because they have heard that it has uncomfortable side effects. But side-effect management has improved over the last few decades. Today, many side effects once associated with chemotherapy can be prevented or controlled. With some types of chemotherapy, you may experience only minimal side effects. For many people, chemotherapy may be your best option for a successful outcome. To achieve a successful outcome by understanding how side effects affect your treatment. Learn how best to manage chemotherapy side effects.

Chemotherapy is the general term for any treatment involving the use of chemical agents to stop cancer cells from growing. Chemotherapy eliminates cancer cells at sites great distances from the origin of the tumor. As a result, chemotherapy is considered a *systemic* treatment.

More than half of all people diagnosed with cancer receive chemotherapy. For millions of people, chemotherapy helps treat their cancer effectively and allows them to enjoy full, productive lives.

A chemotherapy *regimen* (a treatment plan and schedule) uses a variety of drugs to fight cancer plus drugs to help support completion of treatment.²⁻⁸ To get the most from chemotherapy, it's important to follow the schedule of treatment. Find out more about [chemotherapy cycles and schedules](#).



Genentech
BIOONCOLOGY

[Home](#) | [Products](#)

[Home Page](#)

[Innovative Research Focus](#)

[Clinical Trials](#)

[Professional Resources](#)

[Products](#)

[Scientist Profiles](#)

[Patient Access](#)

[Contact Us / Careers](#)

[Pharmacists Center](#)

The Pharmacists Center features helpful resources, links, and tools supporting healthcare provider and patient education as well as useful information for practicing oncology pharmacists.

[Pharmacists Center](#)

Products

At Genentech BioOncology, we are dedicated to defining the molecular basis of cancer and investigating multiple approaches to treating the disease and improving patients' lives. Some groundbreaking developments of Genentech BioOncology include:

[Avastin website](#)

[Herceptin website](#)

[Rituxan website](#)

[Tarceva website](#)



AVASTIN[®]
bevacizumab

[Click here to view all Safety and Indication Information](#)



Herceptin[®]
trastuzumab

[Click here to view all Safety and Indication Information](#)



Rituxan[®]
Rituximab

[Click here to view all Safety and Indication Information](#)



Tarceva[®]
erlotinib
tablets

[Click here to view all Safety and Indication Information](#)

What to believe?

“A review article of the latest studies looking at red wine and cardiovascular health shows drinking two to three glasses of red wine daily is good for the heart.”

Excerpt from <http://www.sciencedaily.com/>

- Is this information trustable?
- **What do others think about this information?**
- Is this supported by “reliable” sources?

Many support groups and medical forums

HealthBoards
HEALTH MESSAGE BOARDS

HOME MESSAGE BOARDS HEALTH GUIDE JOIN FOR FREE

SEARCH Go

Register FAQ Doctor Delivery Today's Posts Advanced Search

Breast Cancer Mailing List Archives

YAHOO! HEALTH Groups [Sign In](#) [New User? Sign Up](#)

Health - Groups

378 messages
sort by: [a] [d] [r]
Nearby: [A] [L] [R]

 LIVING WITH lymphoma
INFORMATION, INSPIRATION AND DETERMINATION

Find Inspiration
Through customized emails

Join the program now 

Lung_Cancer_Online_Support · Lung Cancer Online Support Group

Search for other groups...

Search

- [regular reports](#)
- [BC screening](#)
 - [Re: E](#)
 - [Re: E](#)
- [DISH Jackie](#)
- [vitamin D](#)
 - [Re: v](#)
 - [Re: v](#)
 - [Re: v](#)
- [OT - Made](#)
 - [Re: C](#)
- [Birthday A](#)
- [OT Help re](#)
 - [Re: C](#)
 - [Re: C](#)

Home
Attachments

Members Only
Messages
Post
Files
Photos
Links
Database
Calendar
Promote
Groups Labs (Beta)

Info Settings

Group Information
Members: 367
Category: [Cancers](#)
Founded: Oct 25, 2004
Language: English

 Visit the [Groups blog](#) for the latest Yahoo! Groups information

Home

Join This Group

Activity within 7 days: **1** New Link - **82** New Messages - **1** New File - [New Questions](#)

Description

CANCER! You have lung cancer or a loved one was just told the news. Now what?

This is an online support group for lung cancer patients, their families and friends. It is a closed group. We are not medical experts and advocate following your doctor's advice and encourage people to get second opinions.

Members in this group can exchange information about clinical trials, diagnosis, treatments, concerns, and share their fears and hopes in a spam-free environment. When you are communicating please no personal attacks, name calling, and/or challenging the beliefs of others. Members need support, not harassment.

You will have access to features such as archived messages, databases, files, links, photos, and can post in celebration of or in memory of the battle against lung cancer. We encourage uploading of photos.

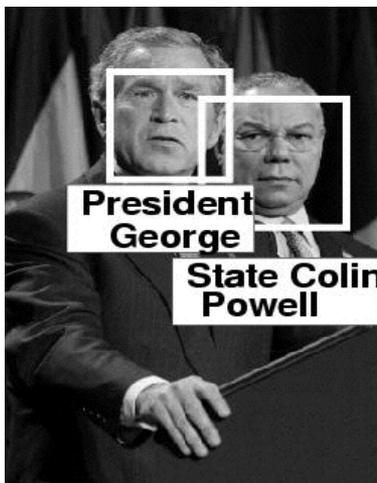
Please indicate "WHY" you want to join this group. As a spam guard all messages are moderated. The public cannot view your posts. This group is not open to researchers who want to survey members on what it is like to have lung cancer, or fundraisers who want to solicit donations. This is a lung cancer support group!



Trustworthiness

- Given:
 - Multiple content sources: websites, blogs, forums, mailing lists
 - Some target relations (“facts”)
 - E.g. [disease, treatments], [treatments, side-effects]
 - Prior beliefs and background knowledge
- We can:
 - **Score trustworthiness of relations** (“facts”) based on
 - support across multiple (trusted) sources
 - source characteristics:
 - reputation, interest-group (commercial / govt. backed / public interest), verifiability of information (cited info)
 - **Rate databases/sources** as more/less trustworthy based on the verifiability scores of the facts in database
 - **Track** how the trustworthiness of fact / database varies with time as the text corpus grows over time

Semantic Data Enrichment: Text and Images



US President George W. Bush (L) makes remarks while Secretary of State Colin Powell (R) signs the U.S. National HIV/AIDS, and STD Prevention Act of 2003 at the State in Washington, DC. The five-year plan is designed to help prevent and treat AIDS, especially in more than a dozen African and Caribbean nations (AFP/Luke Frazza)



German supermodel Claudia Schiffer gave birth to a baby boy by Caesarian section January 30, 2003, her

Schiffer is seen on the German television show 'Bet It...?!' ('Wetten Dass...?!') in Braunschweig, on January 26, 2002. (Alexandra Winkler/Reuters)



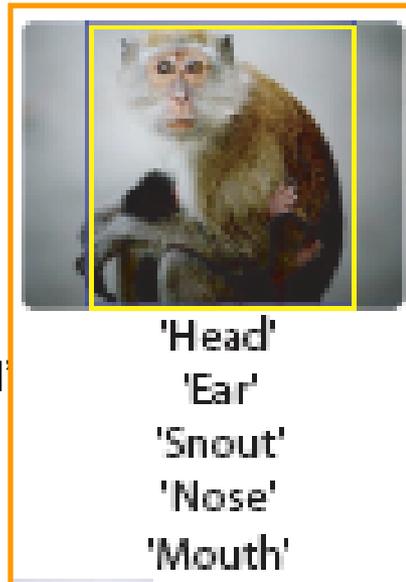
British director Sam Mendes and his partner actress Kate Winslet

man who has a separate family life and co-stars Paul Newman and Jude Law. REUTERS/Dan Chung

The kinds of information contained in pictures may be very difficult to access in other ways.

Adding Annotation [Forsyth, Hoiem et. al]

- Problem: How to tag pictures with text-like summaries of **what is important in the picture**
 - Method: Analyze relations between text and images in illustrated documents
- Key Challenge:
 - make sensible statements about **unknown objects**
- Strategy
 - **Represent objects as collections of attributes**
 - “Red”; “has a head”; “furry”; “has a wing” ; “made of metal”
 - **Build classifiers to detect attributes**
 - Right bag of attributes → object
 - **funny attributes? say so**
 - Otherwise - report attributes



- State-of-the-art recognition for **known objects**
- Accurate reports of **attributes of unknown objects** Recognition from text description



Bird
"Leaf"



Motorbike
"cloth"



DiningTable
"skin"



Sofa
"wheel"



Bike
"Horn"



Aeroplane
No "wing"



Car
No "window"

- Predict ways in which recognized object is **special**
- I.e. extra/missing attributes



Bicycle
No "wheel"



Sheep
No "wool"



Train
No "window"

Information Access & Synthesis [Processes & Tools]

- Focused data retrieval and integration,
 - Identify and collect relevant data from multiple sources
-  Semantic data enrichment,
 - Infer semantics from unstructured data and images;
 - Allow navigation and search across disparate data modalities;
-  Entity identification and relations discovery,
 - Identify real-world entities and relations among them
 - Relate them to existing institutional resources for information i
- Knowledge discovery and hypotheses generation and verification
 - Construct the rich semantic structure and hidden networks of e
- Foundations
 - Machine learning, database and data mining, natural language inference and optimization and computer vision techniques
 - Called for and driven by the aforementioned problems.

Tools

Text
Processing &
Analysis

Semantic
Analysis &
Information
Extraction

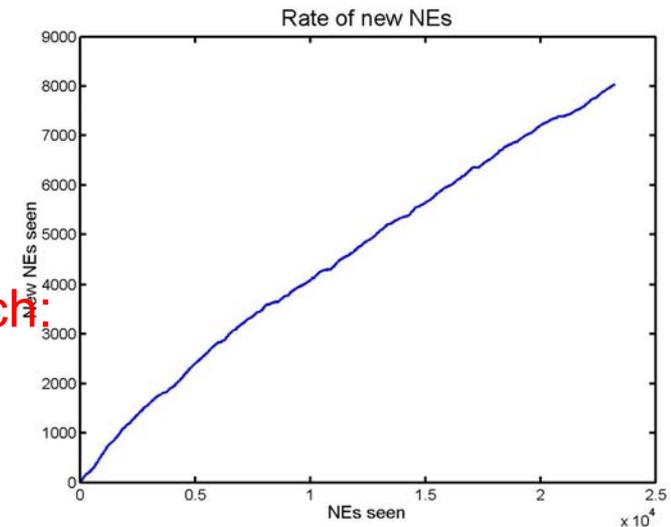
Information
Integration

Machine
Learning &
Data Mining

Integrating
Text &
Images

Example: Named Entity Recognition

- Entities are inherently ambiguous (e.g. **JFK** can be both location and a person depending on the context)
 - Can appear in various forms ; Can be nested.
 - Using lists is not sufficient
 - New entities are always being introduced
- **Necessary to use a Machine Learning approach:**
- Significant over fitting
- Adaptation to:
 - New domains/corpora
 - Slightly new definition of an entity
 - New languages; New types of entities
- Need to:
 - reduce the requirements on resources (training)
 - Incorporate knowledge



Constraints Conditional Models (CCMs)

aka Integer Linear Programming for NLP

- **Informally:** Global decisions with learned models, in the presence of constraints
- **Why Constraints?**
 - A effective way to inject expressive prior knowledge into models.

- **Issues to attend to:**
 - While we formulate the problem as an **ILP problem**, Inference can be done multiple ways
 - Search; sampling; dynamic programming; SAT; ILP
 - The focus is on **joint global inference**
 - **Learning** may or may not be joint.
 - Decomposing models is often beneficial

tutorial on my web page and ILPNLP workshop]

ee

Formal Model

$$\operatorname{argmax}_y \lambda \cdot F(x, y)$$

Weight Vector for
“local” models

A collection of Classifiers;
Log-linear models (HMM,
CRF) or a combination

Subject to constraints

Penalty for violating
the constraint

(soft) constraints
component

How far y is from
a “legal” assignment

How to solve?

This is an Integer Linear Program

Solving using ILP packages gives an
exact solution.

Search techniques are also possible

How to train?

How to decompose the global
objective function?

Should we incorporate constraints
in the learning process?