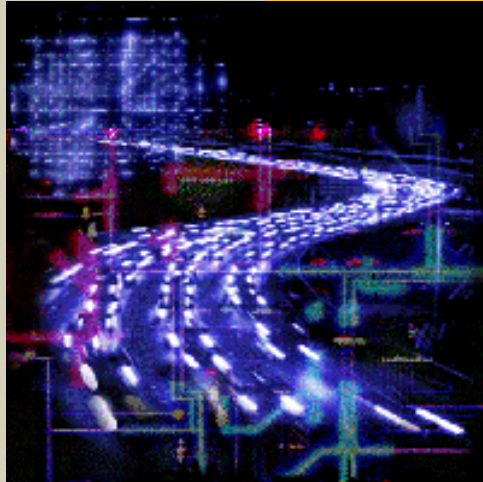# Mining *chemical* & *biomedical* information from the drug literature…

finding the *right stuff*!!!

Debra L. Banville

AstraZeneca Pharmaceuticals

Wilmington, Delaware, USA
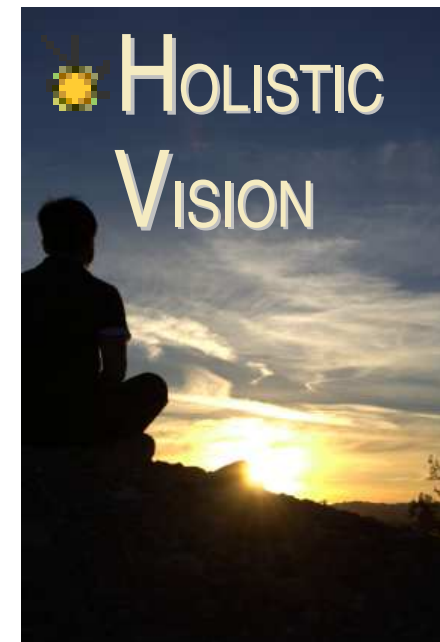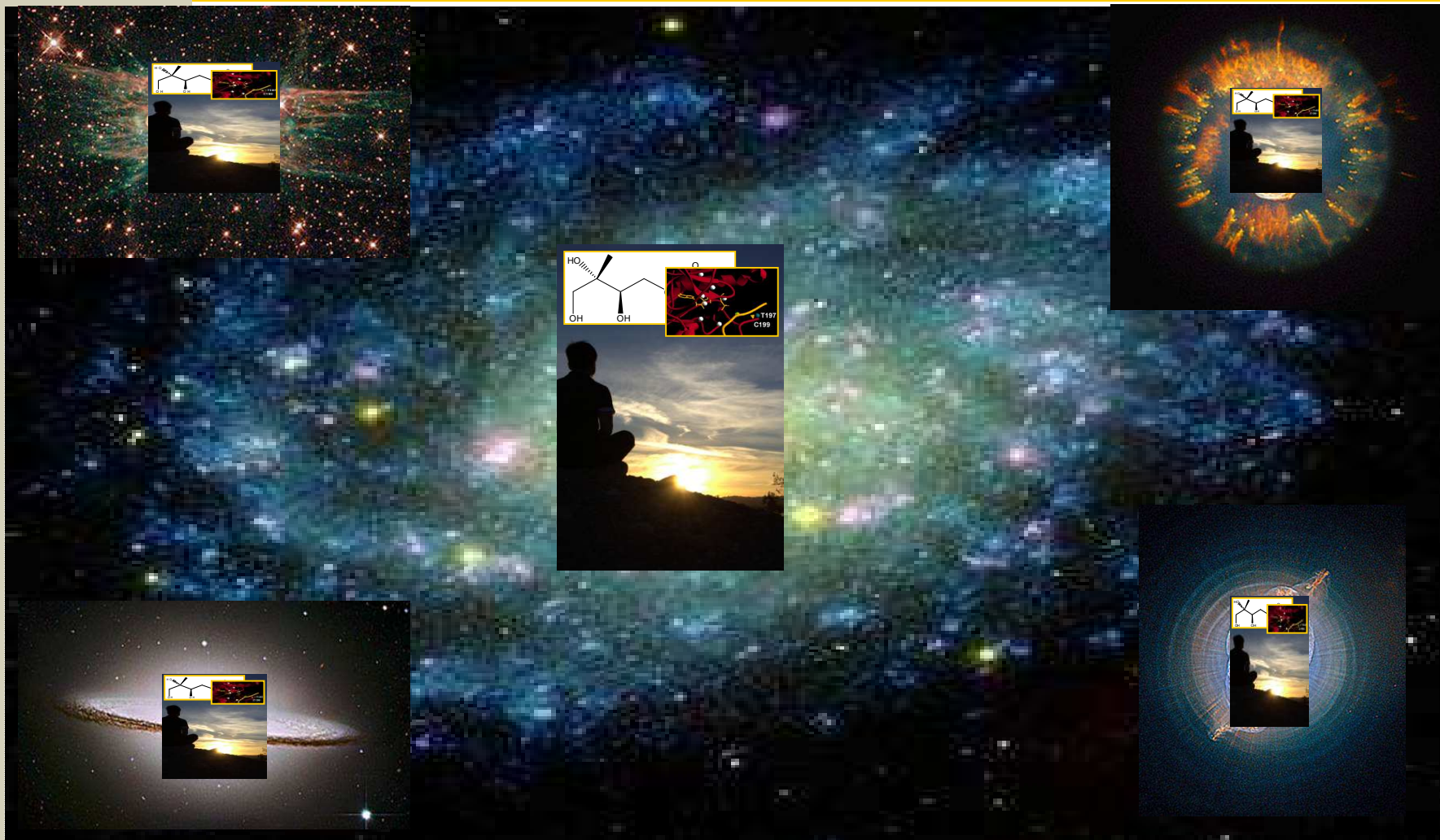
It's all about…

INFORMATION **O**VERLOAD

HURDLES

PARADIGM **S**HIFT

SHIFT HAPPENS

HOLISTIC **V**ISION
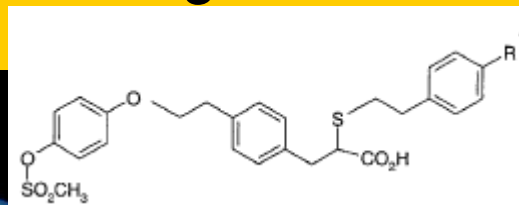
N_{EEDS:} User driven technologies

NEEDS: Current research paradigm

Retrieving

Finding

The present invention relates to certain novel 3-phenyl-2-arylalkylthiopropionic acid derivatives, to processes for preparing such compounds, to their the utility in treating clinical conditions including lipid disorders (dyslipidemias) whether or not associated with

NEWS AND VIEWS

When the *Clock* stops ticking, metabolic syndrome explodes

Nature Medicine **11**, 822 - 824 (2005)
doi:10.1038/nm0805-822

Bart Staels

Circadian control dep
and feeding. Recent s
resembling the metab
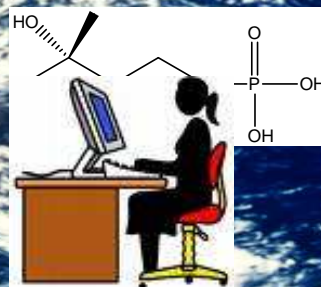
**TZDs and diabetes: testing the waters**

ARTICLE

Clay F S

Nature Medicine **11**, 191 - 198 (2005)
Published online: 30 January 2005; | doi:10.1038/nm1185

IKK-β links inflammation to obesity-induced insulin resistance

Melek C Arkan[1], Andrea L Hevener[2], Florian R Gret
Zhi-Wei Li[1, 3], Jeffrey M Long[4], Anthony Wynshaw
Poli[5], Jerrold Olefsky[2] & Michael Karin[1]
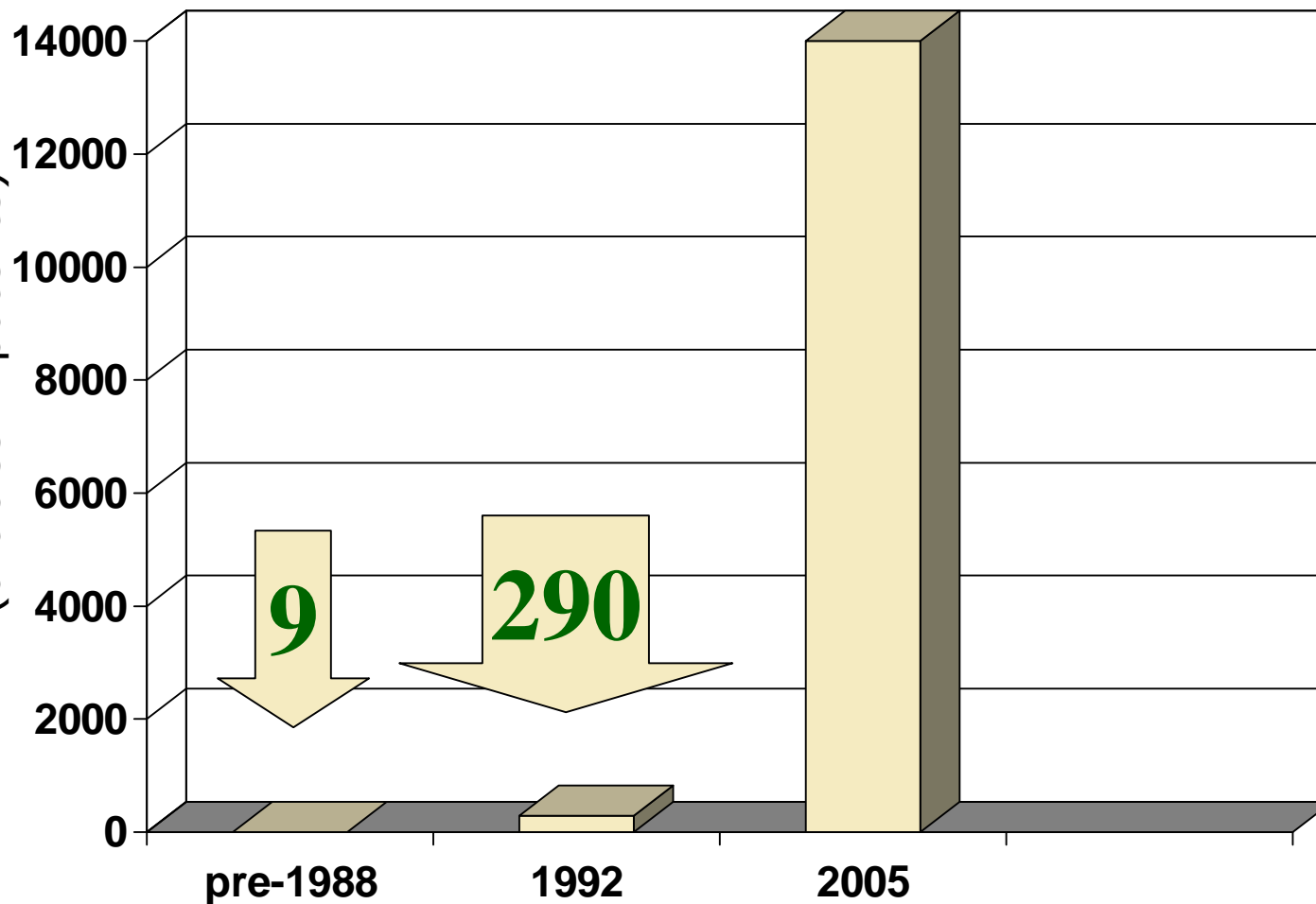
Deeper Analysis
Building Context
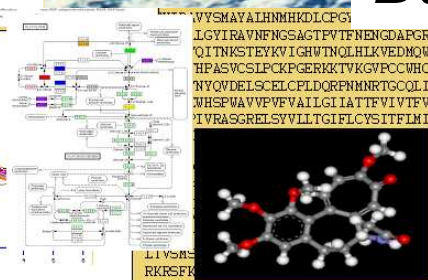
HURDLES:

Too much information

>14,000

Too many documents

Total # GPCR publications to date (articles + patents)

	pre-1988	1992	2005
	9	290	

HURDLES: Too much information, *too many interfaces….*

Multiple interfaces

Limited access

Diverse Output

Books **Conferences** **Dissertations**
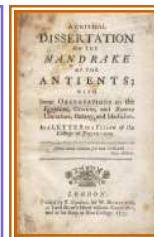
Patents documents

Journal Articles

News PR

Audios

Web Pages
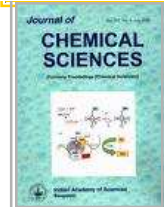
Videos

Wikis

Blogs

Aggregators e.g. SciFinder & OVID 🚫

PubChem

Patent Authorities e.g. USPTO, ESPACE etc…

Conference Web Sites 🚫

Publisher Sites 🚫

Biomed Central Chemistry Central

Etc…

Internet Search e.g. Google & OA etc…

PubChem

DrugBank

ChemSpider

Etc…

Open Access Databases

Note:

Provides Citations
Limited Full Text
Full Text

HURDLES: Inaccessible information

Some are images only!

Textual and/or images mixed together

Diverse document formats

Copyright / Licensing restrictions

HURDLES: Inaccessible information

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

☐ Patent/Publication Number: WO2006004532A1
☐ Publication Date: Jan, 12 2006

[54] NOVEL HYDANTOIN DERIVATIVES FOR THE TREATMENT OF OBSTRUCTIVE AIRWAY DISEASES

☐ Inventor(s):
GABOS, Balint; , AstraZeneca R & D Lund, S 221 87 Lund, SESE
RIPA, Lena; , AstraZeneca R & D Lund, S 221 87 Lund, SESE
STENVALL, Kristina; , AstraZeneca R & D Lund, S 221 87 Lund, SESE

☐ Assignee/Applicant:
ASTRAZENECA AB; ,
GABOS BALINT; ,
RIPA LENA; ,
STENVALL KRISTINA; ,

[30] Priority:
SE Jul, 5 2004 SE20041762A
SE Jul, 4 2005 WO2005SE1092A

[21] Application Number: WO2005SE1092A

[22] Application Date: Jul, 4 2005

[51] Int. Cl.^{8}: A61K0031506 A61K0031506 A61P001100 A61P001100 C07D040114 C07D040114

[52] ECLA: C07D040114+239B+233+211

[56] References Cited:

In accordance with the present invention, there

DINTSNH

(1) //\\Nit o wherein Is R represents C1 to 2 aLk
and R represents C1 to 3 alkyl; to and pharmace

The compounds of formula (I) may exist in ena
of the invention.

OCR conversion, removes formatting...

and generates *errors*!!

• Many patent are not text not images!!

HURDLES: Inaccessible information

Chemical nomenclature is not easily standardized
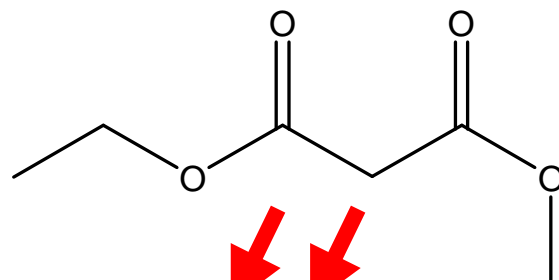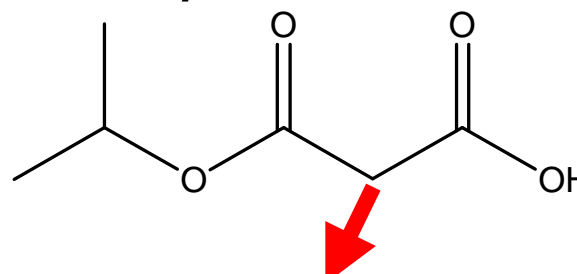
Systematic

Acetylsalicylic acid
salicylic acid, acetate
2-acetyloxybenzoate
2-carboxyl phenylacetate

Abbreviations

ASA

Common/generic

aspirin

Chemical Structures

How many ways can you say "aspirin"?

AC5230

Company Codes

Generic & fragmented

salicylates

Bayer

Trade names

Compound 10...

CAS RN 50-78-2

Anaphors

Index & Ref.

Methyl+Ethyl+Malonate
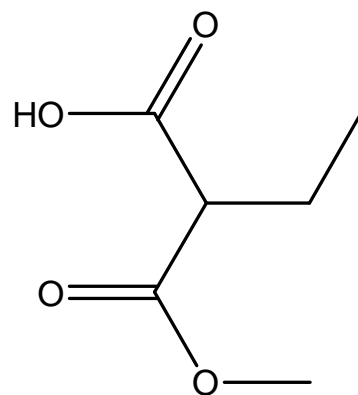
methyl ethyl malonate
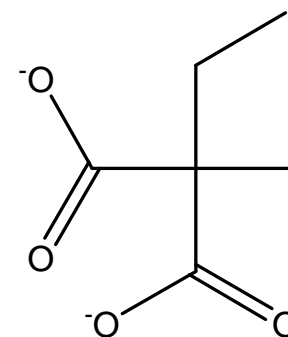
methylethyl malonate

methyl ethylmalonate

methylethylmalonate

Biomedical Ambiguity

Ambiguity Example
PCA has >60 expansions

```
"p-chloroamphetamine" "p-chloroaniline" "p-coumaric acid"
"p.rothrombin c.omplex a.ctivity" "para-chloramphetamine" "parietal
cell antibodies" "parietal cell autoantibodies" "paroxysmal cerebellar
ataxia" "passive cutaneous anaphylactic" "patient care appraisal"
"patient controlled analgesia" "patient controlled anesthesia" "pca"
"pentachloroanisole" "percent cortical area" "perchloracetic acid"
"perchloric acid" "percutaneous coronary angioplasty" "percutaneous
coronary atherectomy" "pericallosal artery" "peritoneal
carcinomatosis" "peritoneal carcinosis" "personal care attendant"
"phenazine-1-carboxylic acid" "phenylciclopentylacetic acid"
"phenylcyclohexylamine" "physical capacity assessment" "pig coronary
artery" "plate count agar" "pneumococcal capsular antigen" "pole
climbing avoidance" "polyclonal activator" "polyclonal antibody"
"polyclonal antisera" "polycyclic aromatic content" "porous coated
anatomic" "porous coated total hip arthroplasty" "porous-coated hip
arthroplasties" "porous-coated patellar component" "porta-caval
anastomosis" "portable clinical analyzer" "portacaval anastomosis"
"portacaval shunt" "post chigger attachment" "postconceptional age"
"posterior cerebral arteries" "posterior communicating artery"
"posterior cortical atrophy" "posterior crico-arytenoid" "potassium
channels activators" "presence of parietal cell" "primary cardiac
arrest" "primary congenital aphakia" "principal component analyses"
"procoagulant" "procoagulant activities" "procoagulant cellular
activity" "procoagulatory activity" "prostatic carcinoma" "protein c
activator" "prothrombin complex activity" "protocatechuate"
"protocatechuic acid" "protocaval anastomosis" "pulmonary corpora
amylacea" "pyroglutamic acid" "pyrrolidone carboxylic acid"
"pyrrolidone-2-carboxylic acid" "pyrrolidone-5-carboxylic acid"
"pyrroline-5-carboxylate"
```

Hᴜʀᴅʟᴇs:
Too much information

"There... How does that feel?"

- Limited access
- Multiple interfaces
- Diverse Output

Is it the right volume of information (coverage)?

Is it the right type of information (content)?

What are you doing with the information now (analysis for decision making)?

How do you manage this information over time (management)?

PARADIGM SHIFT *from*...

SHIFT HAPPENS

- **Hard copy => e-Journals/e-Books**

HighWire LIBRARY OF THE SCIENCES AND MEDICINE

- **Multiple search => Single search interfaces**

  – Federated Search & Mash-ups

- **Traditional publishing => e-Models**

  A leap forward into the semantic web gets award for RSC

  RSC | Advancing the Chemical Sciences

  14 September 2007

  A technological leap achieved by a team at the Royal Society of Chemistry has pushed forward the boundaries of information access, a feat which brings it a

- **Closed => Open systems**

  ChemistryCentral
  OA
  OD
  OS

- **Manual *only* => Automated curation**
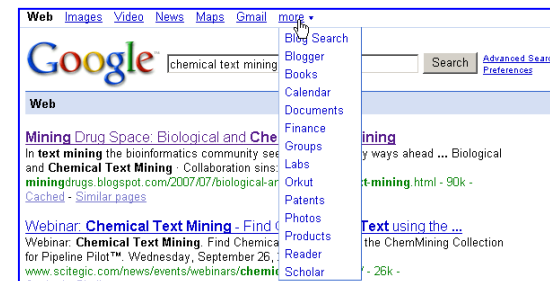
PARADIGM SHIFT:
Natural Language Processing (NLP) does not understand human speech…

☘ J. Weizenbaum (MIT) & *Eliza* in 1966

Eliza - computer program for…natural language communication between man and machine

Communications of the ACM 9 36-45 (1966)

Conversation between Researcher and his computer program:

Eliza: *How are you today?*

Res: *Not so good?*

Eliza: *Why do you say not so good ?*

Res : *I suffer from allergies*

Eliza: *That's interesting*

Etc…

Eliza: *you are being a bit negative*

Res : *No I'm not*

Eliza: *How long have you **been not***

•Ambiguities

PARADIGM SHIFT: NLP over 30 years later

Syntactic & semantic analysis is done by the researcher

Tokenization

Words & sentences

Nouns, Verbs etc...

Part of Speech tagging

Syntactic Analysis

The adverse events occurring in ≥ 5% of patients are shown in Table 4. The most commonly reported adverse event in both groups was somnolence, and this occurred in a higher proportion of patients with quetiapine than with risperidone (26.3% and 19.7%, respectively; p = .04). In both quetiapine and risperidone groups, the majority of

Is this a compound, a symptom/disease, a protein, etc...?

Identifying noun phrases

Name Entity Recognition

Noun Phrase Chunking

PARADIGM SHIFT: NLP today focused on NER

Royal Society of Chemistry's (RSC) *Project Prospect.*
Evidence of a shifting paradigm in the publishing world.
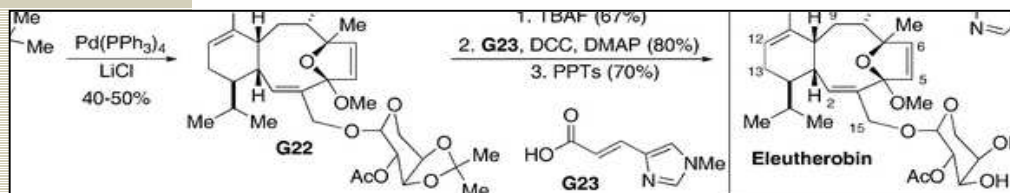
Scheme 19 Synthesis of eleutherobin.

- **Name Entity Recognition (NER)**
- **Open Source & Access**

- **Text Mining/NLP/NER** (SciBorg, OSCAR3)
- **Ontologies** (GO, SO, OBO ChEBI, Gold Book)
- **Chemical Structure information** (InChl, SMILES, CML)
  - Unilever Centre at U of Cambridge
  - Eur Bioinf Institute

PARADIGM SHIFT: NLP today focused on NER

Examples of licensed chemical NER capabilities…

- Accelrys (tool)

- Fraunhofer Institute/ChemInfo (tool)

- IBM (tool & patent database)

- ReelTwo/SureChem (patent database)

- TEMIS (tool)

PARADIGM SHIFT: NER requires conversion tools

✷ Name to Structure

– Licensed: ACD/Labs, CambridgeSoft, Lexichem, OpenEye

Example 22 Methyl 4-({[trans-4-(2,4-dihydroxyphenyl)cyclohexyl]amino}sulfonyl)benzoate

From Preparation 11 (150 mg, 344 .mu.mol) and methyl 4-chlorosulfonylbenzoate (97 mg, 414 .mu.mol) following method E to give the title compound (110 mg, 78%) as a white solid. gm (CD.sub.3 OD) 1.20-1.41 (m, 4H), 1.70-1.81 (m, 4H), 2.61-2.70 (m, 1H), 3.04-3.18 (m, 1H), 3.92 (s, 3H), 6.15-6.22 (m, 2H), 6.69 (d, 1H), 7.98 (d, 2H), 8.16 (d, 2H); m/z (ES.sup.+) 406 [M+H].sup.+.
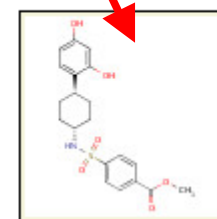
✷ Image to Structure

– Open Source: OSRA (NCI Frederick)

– Licensed: Keymodule/SimBioSys, Fraunhofer Inst.
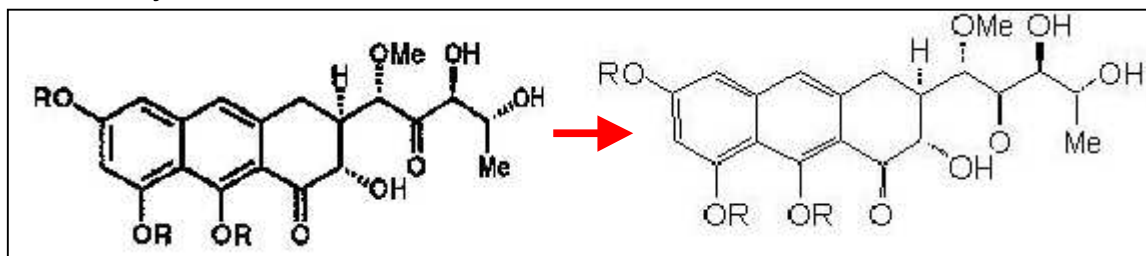
PARADIGM SHIFT:

Where *Semantic Web* & *Chemistry* meet— electronic way to capture structural information

Example 22 **Methyl 4-({[trans-4-(2,4-dihydroxyphenyl)cyclohexyl]amino}sulfonyl)benzoate**

From Preparation 11 (150 mg, 344 .mu.mol) and **methyl 4-chlorosulfonylbenzoate** (97 mg, 414 .mu.mol) following method E to give the title compound (110 mg, 78%) as a white solid. gm (CD.sub.3 OD) 1.20-1.41 (m, 4H), 1.70-1.81 (m, 4H), 2.61-2.70 (m, 1H), 3.04-3.18 (m, 1H), 3.92 (s, 3H), 6.15-6.22 (m, 2H), 6.69 (d, 1H), 7.98 (d, 2H), 8.16 (d, 2H); m/z (ES.sup.+) 406 [M+H].sup.+.
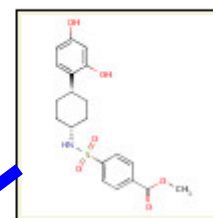
- **InChI &/or SMILES for chemical metadata inserted into a XML/CML formatted document:**

- InChI=1/C20H23NO6S/c1-27-20(24)14-4-9-17(10-5-14)28(25,26)21-15-6-2-13(3-7-15)18-11-8-16(22)12-19(18)23/h4-5,8-13,15,21-23H,2-3,6-7H2,1H3/t13-,15-
- SMILES: COC(=O)c(cc2)ccc2S(=O)(=O)NC3CCC(CC3)c1ccc(O)cc1O

CML Source

```
<?xml version="1.0"?>
<!DOCTYPE cml SYSTEM "http://www.xml-cml.org/dtd/cml1_0_1.dtd">

<cml title="ethanol" id="cml_ethanol_karne" xmlns=
"x-schema:cml_schema_ie_02.xml">
  <molecule title="ethanol" id="mol_ethanol_karne">
   <formula>C2 H6 O</formula>
```
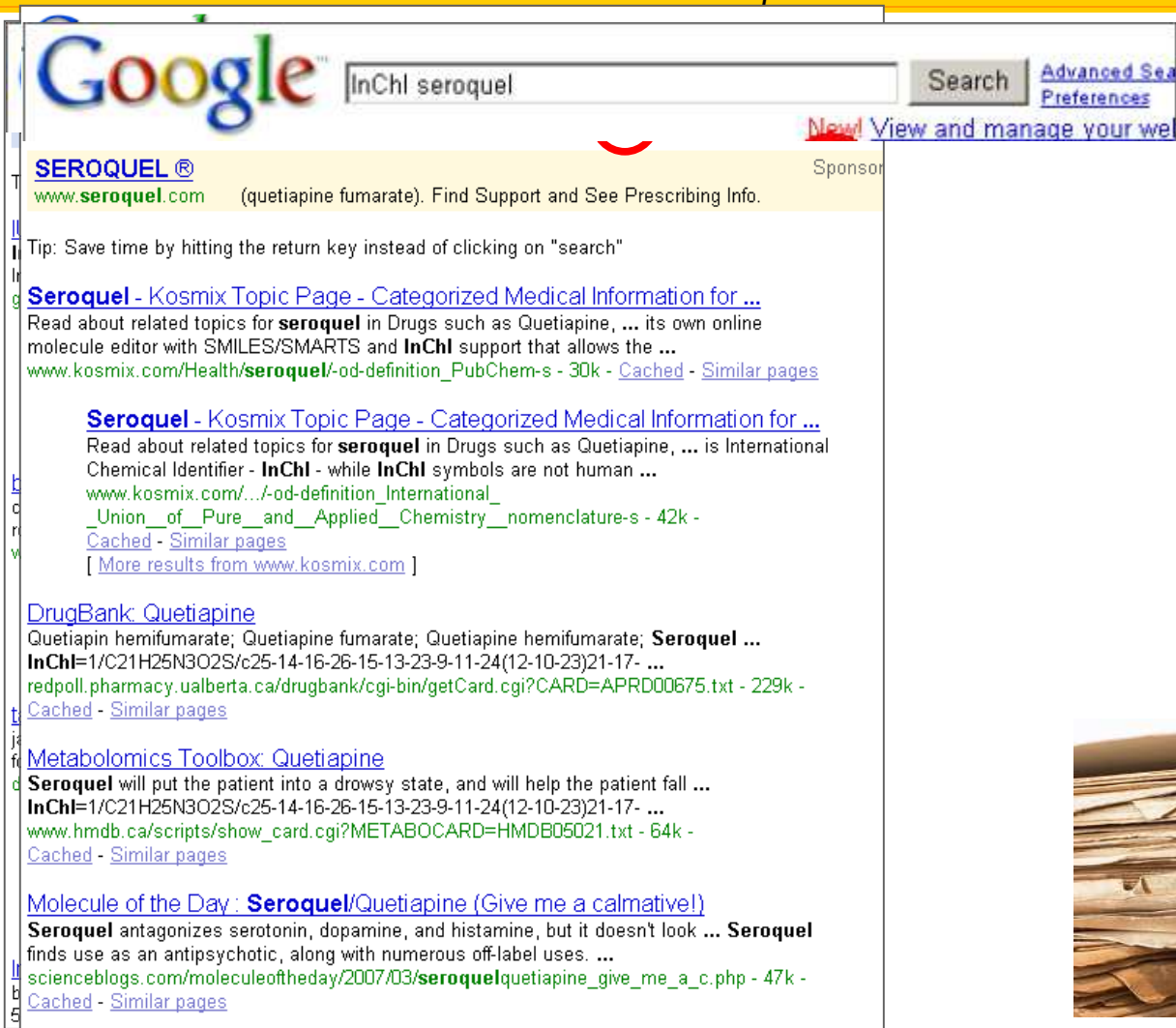
XML/CML formatted documents are great for information exchange!

PARADIGM SHIFT:

Using *Google* to search on chemical structures?

Benzene & *Seroquel*…

InChI internet searches just starting to be possible

Google

InChI seroquel — Search — Advanced Search / Preferences

New! View and manage your web

SEROQUEL ® Sponsor
www.**seroquel**.com (quetiapine fumarate). Find Support and See Prescribing Info.

Tip: Save time by hitting the return key instead of clicking on "search"

Seroquel - Kosmix Topic Page - Categorized Medical Information for ...
Read about related topics for **seroquel** in Drugs such as Quetiapine, **...** its own online
molecule editor with SMILES/SMARTS and **InChI** support that allows the **...**
www.kosmix.com/Health/**seroquel**/-od-definition_PubChem-s - 30k - Cached - Similar pages

Seroquel - Kosmix Topic Page - Categorized Medical Information for ...
Read about related topics for **seroquel** in Drugs such as Quetiapine, **...** is International
Chemical Identifier - **InChI** - while **InChI** symbols are not human **...**
www.kosmix.com/.../-od-definition_International_
_Union__of__Pure__and__Applied__Chemistry__nomenclature-s - 42k -
Cached - Similar pages
[More results from www.kosmix.com]

DrugBank: Quetiapine
Quetiapin hemifumarate; Quetiapine fumarate; Quetiapine hemifumarate; **Seroquel ...**
InChI=1/C21H25N3O2S/c25-14-16-26-15-13-23-9-11-24(12-10-23)21-17- **...**
redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/getCard.cgi?CARD=APRD00675.txt - 229k -
Cached - Similar pages

Metabolomics Toolbox: Quetiapine
Seroquel will put the patient into a drowsy state, and will help the patient fall **...**
InChI=1/C21H25N3O2S/c25-14-16-26-15-13-23-9-11-24(12-10-23)21-17- **...**
www.hmdb.ca/scripts/show_card.cgi?METABOCARD=HMDB05021.txt - 64k -
Cached - Similar pages

Molecule of the Day : **Seroquel**/Quetiapine (Give me a calmative!)
Seroquel antagonizes serotonin, dopamine, and histamine, but it doesn't look **... Seroquel**
finds use as an antipsychotic, along with numerous off-label uses. **...**
scienceblogs.com/moleculeoftheday/2007/03/**seroquel**quetiapine_give_me_a_c.php - 47k -
Cached - Similar pages

PARADIGM SHIFT:

Assessing current *state of the art* with NER
MDL Reading Machine model for NER is proving to be valid

IE requires manual analysis

3.2.1

2,2-Dimethyl-5-(4-pentenyl)-1,3-dioxane-4,6-dione (10)

Ethylene diammonium diacetate (EDDA) (248 mg, 1.4 mmol) was added to a solution of Meldrum's acid (1.188 g, 8.5 mmol) and 4-pentenal (460 mg, 5.5 mmol) in absolute EtOH (6.6 mL) at rt and the resulting solution was stirred for 1 h. $BH_3 \cdot NH(CH_3)_2$ (326 mg, 5.5 mmol) was added to the reaction, which was stirred for 17.5 h. Water (35 mL) and 5% aqueous HCl solution (3 mL) were added to the reaction, which was extracted with CH_2Cl_2 (3×20 mL). The combined organic layers were dried ($MgSO_4$) and concentrated to give 1.47 g of crude product. Flash chromatography (9:1 hexanes/EtOAc containing 0.5% HOAc) gave 805 mg (69%) of pure 10 as a white solid: mp 48-50°C; ^{1}H NMR 5.80 (ddt, 1, J=17.1, 10.4, 6.1 Hz), 5.04 (d, 1, J=17.1 Hz), 4.99 (d, 1, J=10.4 Hz), 3.52 (t, 1, J =4.9 Hz), 2.17-2.10 (m, 4), 1.79 (s, 3), 1.76 (s, 3), 1.61-1.53 (m, 2); ^{13}C NMR 165.5 (2C), 137.7, 115.2, 104.8, 46.1, 33.5, 28.4, 26.9, 26.1, 25.6.

Identify:

- **Formal structure**
- **Physical properties**
- **Name Candidates**
- **Labels, Abbreviations, Anapher**
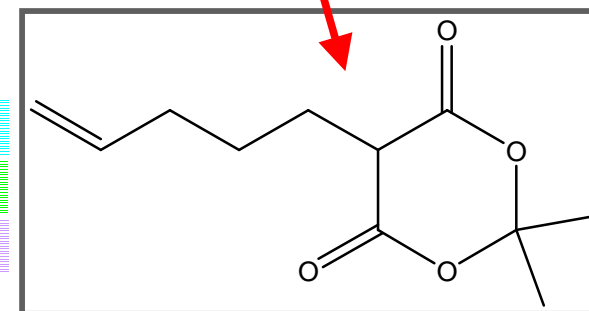
PARADIGM SHIFT: Assessing NLP conversion tools for information extraction...

☀ Ability to automatically identify a...
- ✓ *Chemical & Biological names* is fairly successful
- ✓ *Image are* more difficult requires manual intervention

☀ Ability to convert a chemical name/image to structure
- ✓ *Name conversion* fairly successful
- ✓ *Image conversion* more difficult

☀ Syntactic analysis of biological and chemical information manual inspection.

NER tools can improve information extraction immensely.

IN A Nutshell

HOLISTIC VISION:

Multiple Views of the same information allow a cross-over into key areas of exploration: Chemistry & Biology Space

Trends

❑Can I find trends in research or IP?

Drug DX

❑Can I find compounds linked to protein targets/diseases?

Drug Safety

❑Are these drugs expected to have toxicity/safety issues?

Chemistry Space

❑Can I find the compounds of interest?

Biology Space

T197
C199

❑Can I find protein targets/diseases of interest?

HOLISTIC VISION: A Chemistry query can branch into Biology
Space, to filter information…

Can I find key information on a compound of interest?

Search

Mol. Wt. 474,61

GVK ID: 4198254 | 3D   GVK ID

In a further embodiment, R may be 2-fluorophenyl, and the compound of formula I may be 4-[(3-aminophenyl)[4-[(2-fluoro]

further embodiment, R ma

aminophenyl)[4-[(3-fluoro]

further embodiment, R ma

aminophenyl)[4-[(4-fluoro]

Claims….

16. A compound selected from enantiomerically pure 4-[(S)-(3-aminophenyl) [4 (4-fluorobenzyl)piperazin-1-yl]methyl]-N,N-diethylbenzamide ; enantiomerically pure 4-[(R)-(3-aminophenyl)[4-(4-fluorobenzyl)piperazin-1-yl]methyl]-N,N diethylbenzamide ; enantiomerically pure 4-[(R)-(3-aminophenyl)[4-[(2-fluorophenyl)methyl]-1-piperazinyl] methyl]-N,N-diethylbenzamide ; enantiomerically pure 4-[(R)-(3-aminophenyl)[4-[(3-fluorophenyl)methyl]-1 piperazinyl] methyl]-N,N-diethyl-benzamide and pharmaceutically acceptable salts thereof.

Select Clusters of most interest and analyze them

Clustered Results
TiAbMh | TiAb | Mh | Au | Ad
Dp

▶ **GPCR biomarker** (47)
⊕▶ **Kinase, Signaling** (9)
⊕▶ **Cancer, Tumor** (8)
⊕▶ **CHO Cells, Cricetinae** (5)
⊕▶ **C5a, Compleme**
▶ **Lymphocyte Act Immunology** (4)
⊕▶ **HIV-1** (5)
▶ **Immunoprecipitates, Cell surface** (4)
▶ **Dipeptidyl, Strategy** (2)
▶ **Somatostatin, Brain** (2)
▶ **Molecular Evolution** (2)
▼ More

Review context w/in the literature…

Too many hits to analyze?
e.g. >14,000.
=>*View them in topical clusters/categories or based on toxicology data extracted etc…*

HOLISTIC VISION:

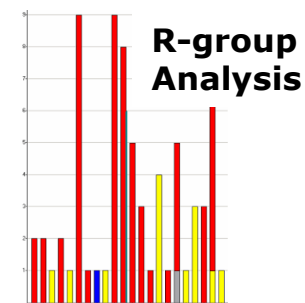A Biology query can branch out into Chemistry Space, to resolve toxicity question…

What compounds are known to cause *acute liver failure*?

Search on terms related to acute liver failure.

2516 citations

R-group Analysis

Extract my compounds and merge for analysis

Which articles link drugs to *acute liver failure*?

Select citations with drug references.

75

Review & extract compounds from 75 full text articles

How do these compounds compare structurally with compounds from my database?
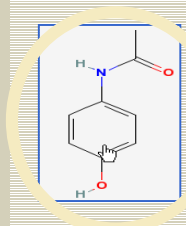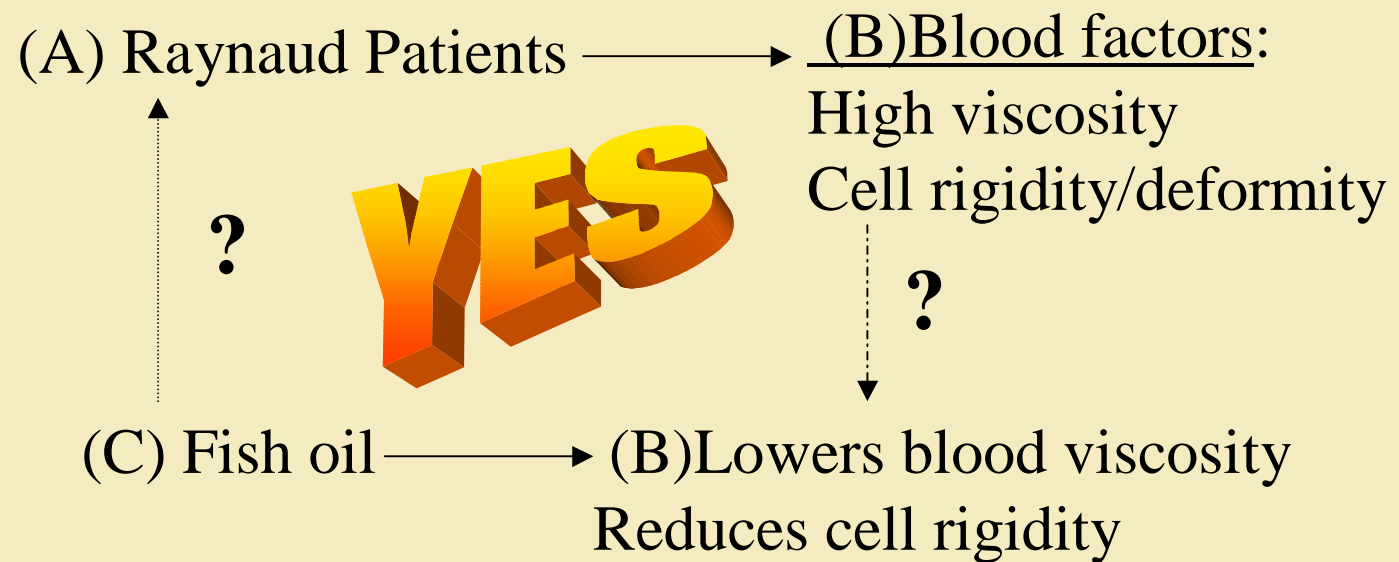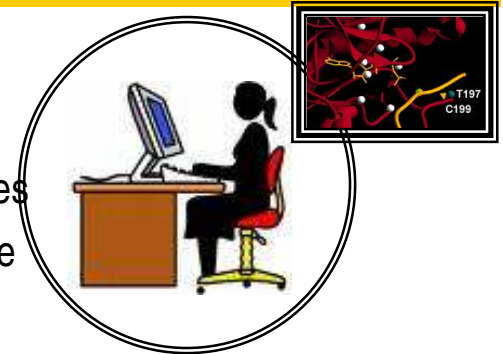
Acetaminophen, highly active antiretroviral therapy and **drugs** for tuberculosis retained their preeminent position as the most commonly reported agents causing **drug-induced liver disease**, with **acetaminophen** continuing to be the leading cause of **acute liver failure** in the USA.

large case series of acute **drug-induced liver disease** from Spain and Sweden supported the observation that **acute hepatocellular jaundice** from a **drug** is associated with **death** or the need for transplant in at least 10% (known as Hy's Law). With respect to using potentially **hepatotoxic** medications in patients with underlying liver disease, **statins** and second-generation **thiazolidinediones** were shown to be safe when used in patients with elevated baseline **alanine aminotransferase** or **aspartate aminotransferase** levels.

HOLISTIC VISION: First example of biomedical text mining from the literature (1986)

- David R. Swanson
  - *Raynaud's Syndrome* Episodic shutting off of blood to fingers or toes
  - *Syllogism* -- _Unknown_ connections bet/ symptoms/causes & possible cures

(A) Raynaud Patients ⟶ (B)Blood factors:
High viscosity
Cell rigidity/deformity

YES

?

?

(C) Fish oil ⟶ (B)Lowers blood viscosity
Reduces cell rigidity

"Take Home" Messages…

🌼 Remember researcher…

🌼 Paradigm Shift…
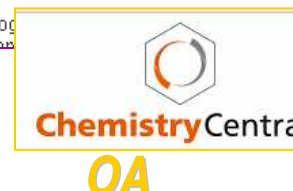
Prediction

✺ Federated Search & Mash-ups

−   Researchers *beware*…

+   Researchers can view diverse information sets with…

…new eyes

The End

Great!

"There... How does that feel?"

Acknowledgments:
James Rosamond
James Damewood
Bob Stumpo
And many other colleagues...

many thanks to you for your kind attention!!