How am I supposed to organize a protein database when I can't even organize my address book?

Jeremy Yang

UNM & IU

CINF Flash session - ACS National Meeting, March 25, 2012 – San Diego, CA

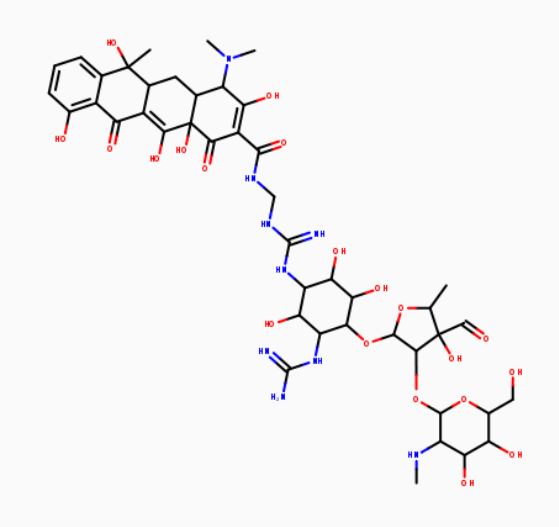
Alternate title (and take home message):

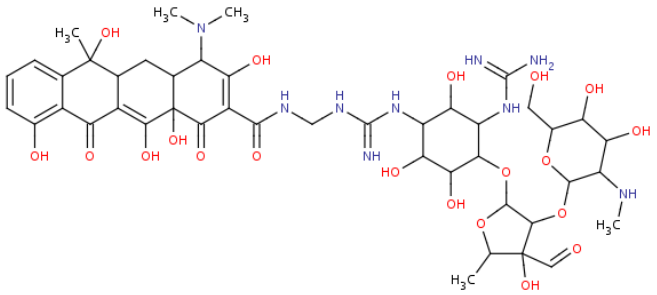
Cheminformatics is so great!

But is it too good to be (*transferably*) true?

How great is cheminformatics?

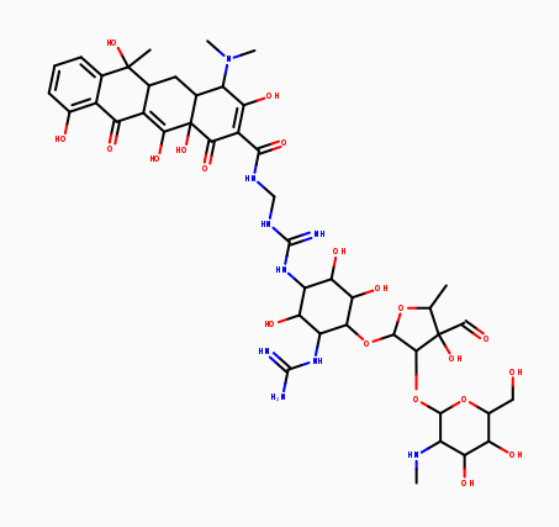
Example: Are these the same or different molecules?

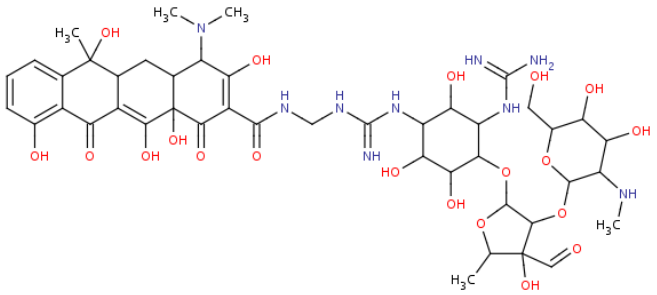




How great is cheminformatics?

Example: Are these the same or different molecules?



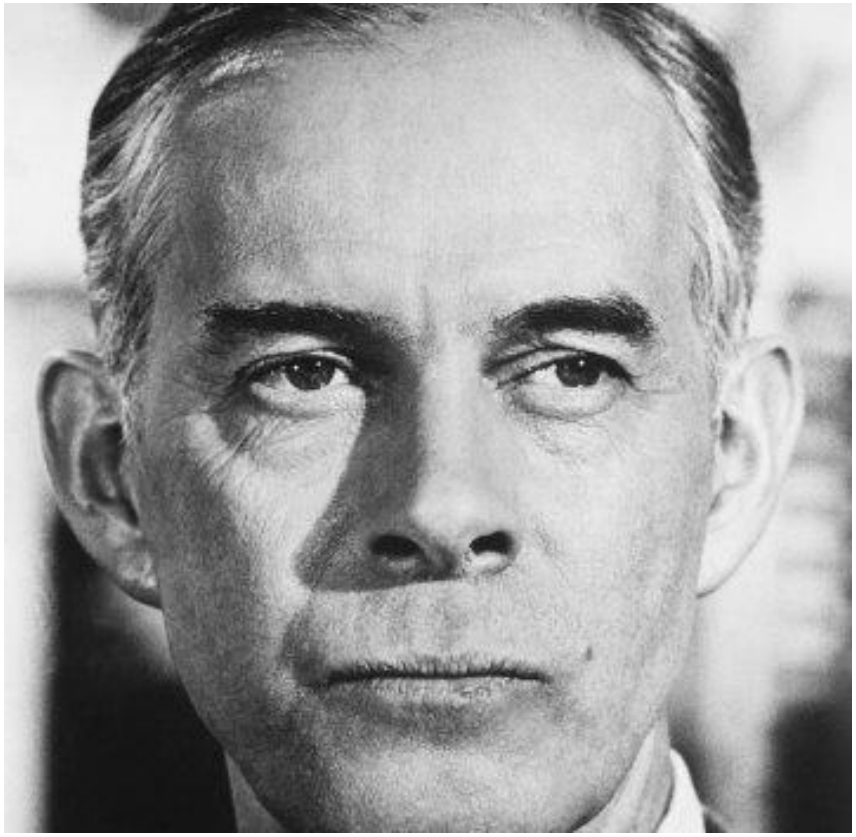


Answer: Same, that's easy, just use canonical graph algorithm via canonical SMILES:

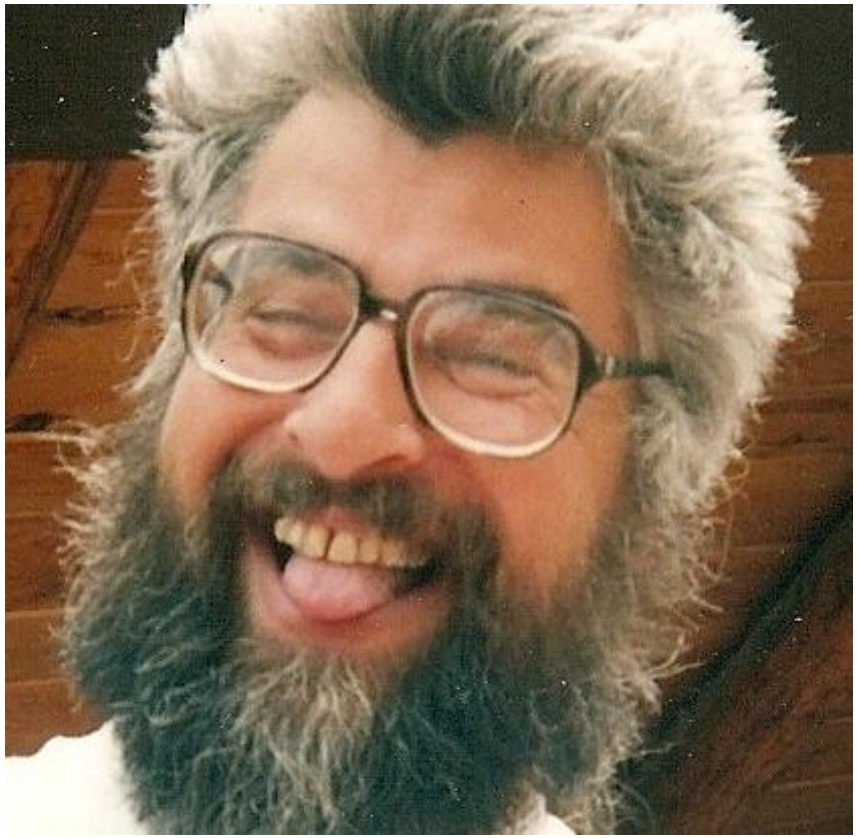
CNC1C(O)C(O)C(CO)OC1OC2C(OC(C)C2(O)C=O)OC7C(O)C(O)C(NC(=N)NCNC(=O)C4=C(O)C(C3CC6C(=C(O)C3(O)C4=O)C(=O)c5c(O)cccc5C6(C)O)N(C)C)C(O)C7NC(N)=N

(TETRACYCLINOMETHYLSTREPTOMYCIN)

Thanks to…

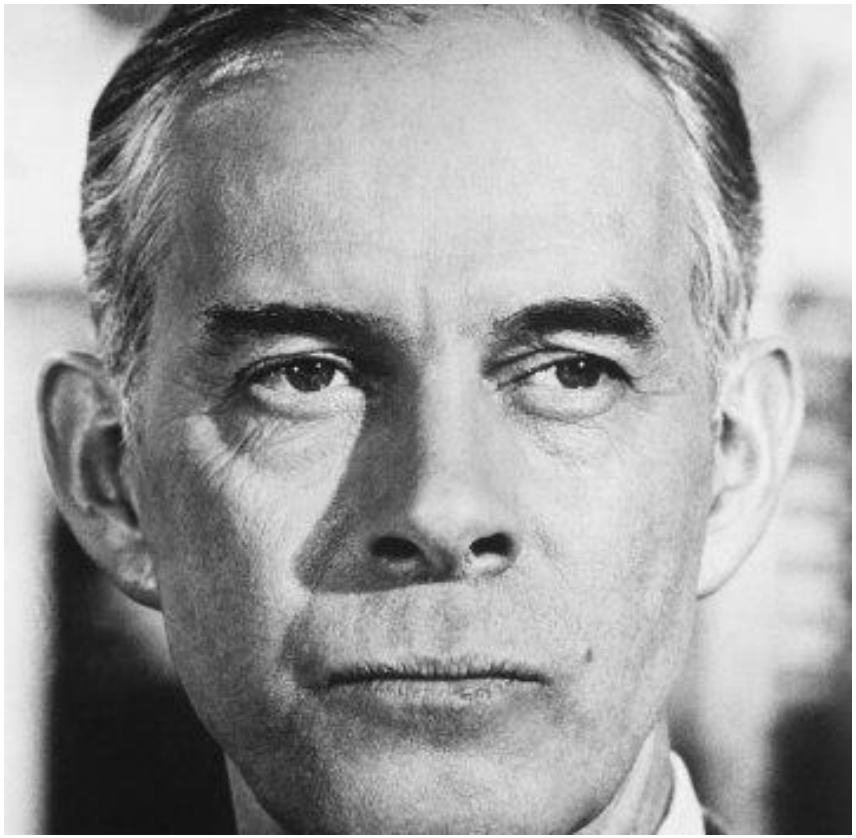


?



?

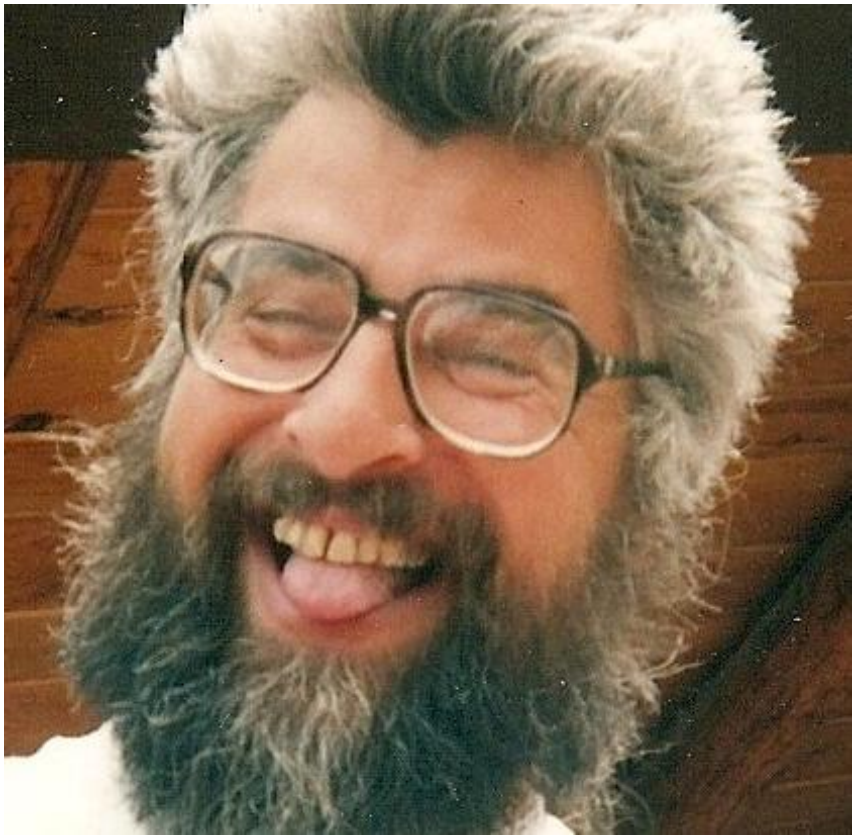
Thanks to…



Harry Morgan

Actor, “MASH”

(Hmmm…?)



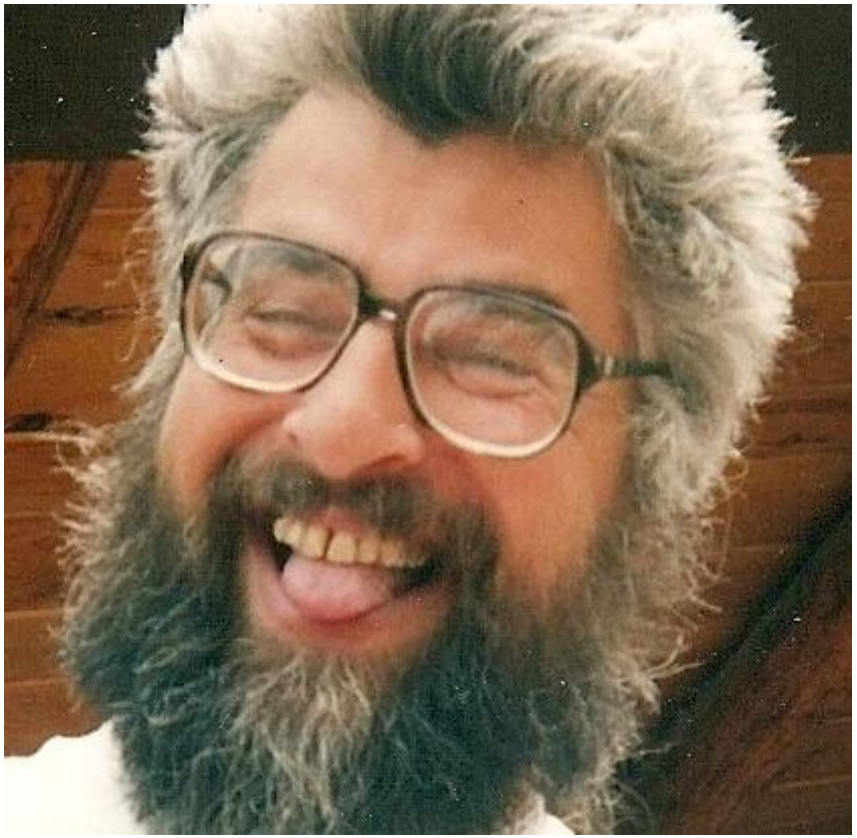
Dave Weininger

Daylight

(SMILES)

Thanks to…

Harry Morgan
ACS CAS
(Morgan Algorithm)



Dave Weininger
Daylight
(SMILES)

Et al., et al….

Now about those proteins…

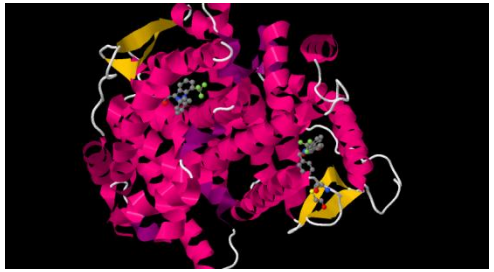
- Example: Are these the same or different proteins?

1YIN:

ALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKRVPGFVDLTLHDQVHLLECAWLEI
LMIGLVWRSMEHPGKLLFAPNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLNSGVYTFLSSTL
KSLEEKDHIHRVLDKITDTLIHLMAKAGLTLQQQHQRLAQLLLILSHIRHMSNKGMEHLYSMKCKNVVPLYDLLLEMLDA
HRLHAPTS

3OS8:

SNAKRSKKNSLALSLTADQMVSALLDAEPPILYSEYDPTRPFSEASMMGLLTNLADRELVHMINWAKRVPGFVDLTRHDQ
VHLLECAWLEILMIGLVWRSMEHPGKLLFAPNLLLDRNQGKCVEGMVEIFDMLLATSSRFRMMNLQGEEFVCLKSIILLN
SGVYTFLSSTLKSLEEKDHIHRVLDKITDTLIHLMAKAGLTLQQQHQRLAQLLLILSHIRHMSNKGMEHLYSMKCKNVVP
SYDLLLEMLDAHRLHAPT



PAM250 alignment score:
(gap: -3; extend: -10)

1156 (1156/1260 = 92%)

Ergo, um… Maybe.

Now about those proteins…

- Example: Are these the same or different proteins?

Human estrogen receptor alpha ligand-binding domain in complex with compound 3F

Estrogen Receptor

Answer: Same… but what does that even mean?

Why protein identification is hard

- Proteins are large, complex, dynamic
- PDB is database of crystallography experiments, not molecules
- Ligands, co-crystals, waters
- Protein crystallography & NMR is hard
- History, culture…

COX-2 ARF-1 ER-α

How about human identification?

(Should be easier, may shed light…)

Voting Rights Advocates Challenge Florida Registration Law in Federal Court

PRESS RELEASES

– 09/17/07

Security systems. Common database errors, however, make "matching" unreliable, jeopardizing the status of up to 30% of new voters. A 2006 study by the Brennan Center for

Florida and a handful of other states refuse to place eligible citizens on the rolls unless they clear a series of extra bureaucratic hurdles largely dependent on "matching" registration information on a new statewide voter list with information in the state motor vehicle or Social Security systems. Common database errors, however, make "matching" unreliable, jeopardizing the status of up to 30% of new voters. A 2006 study by the Brennan Center for Justice, one of the voting rights groups that brought today's suit, found that such a procedure misinterpreted the federal Help America Vote Act (HAVA), which told states to create the statewide lists.

Plaintiffs today argued that there are several ways the bureaucratic process, embodied in Florida's state law (Subsection 6 of Section 97.053), will disenfranchise tens of thousands of eligible voters in the 2008 election cycle, especially in trying to match registration forms with

Human identification hard too, apparently…

Homeland security

MISTAKEN IDENTITY: Garcia stuck on watch list

Schools chief raises red flag when he flies

By LISA KIM BACH
REVIEW-JOURNAL

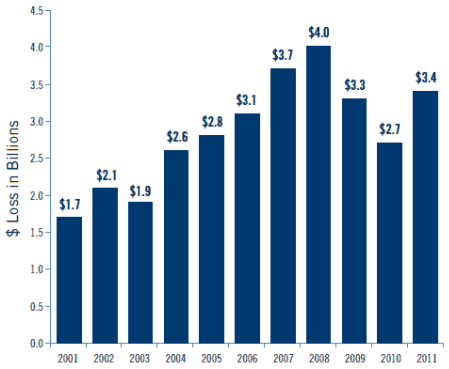
The superintendent of the Clark County School Dis

It's a point local public schools leader Carlos Garci
Security Administration and Southwest Airlines for

He's provided the government with notarized copie
school district identification in an attempt to show
international terrorist.

Credit card fraud

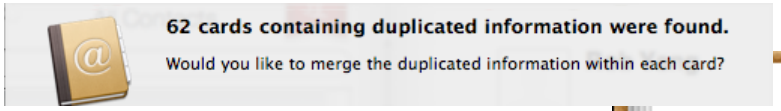
Online Revenue Loss Due to Fraud
Estimated $3.4B in 2011

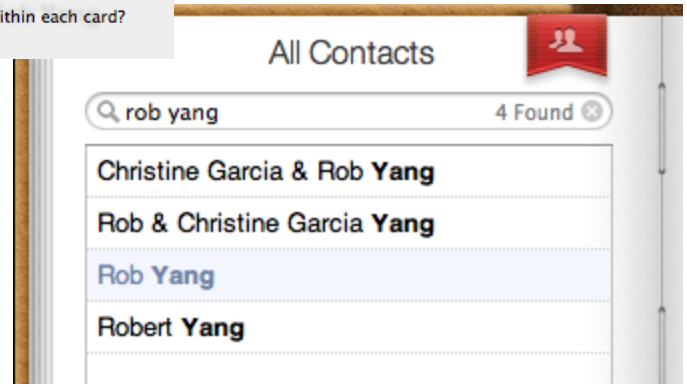


http://forms.cybersource.com/forms/NAFRDQ12012whitepaperFraudReport2012CYBSwww2012

(Which brings us to…)

My address book problems





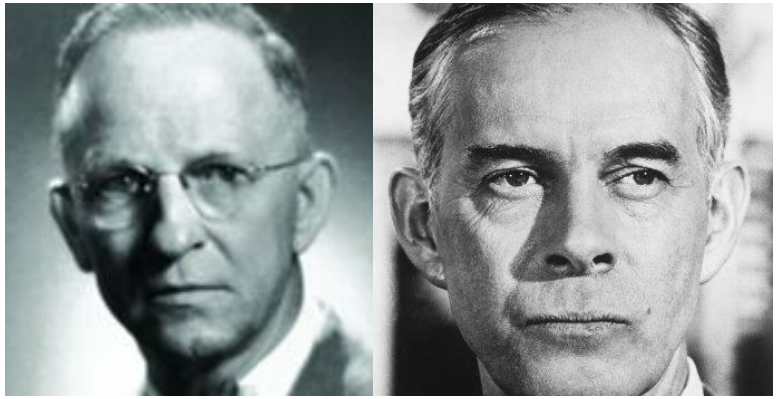
Merge duplicate contacts

We have found 640 contacts with duplicate data.

How many Rob Yangs?

(Philosophical tangent:)

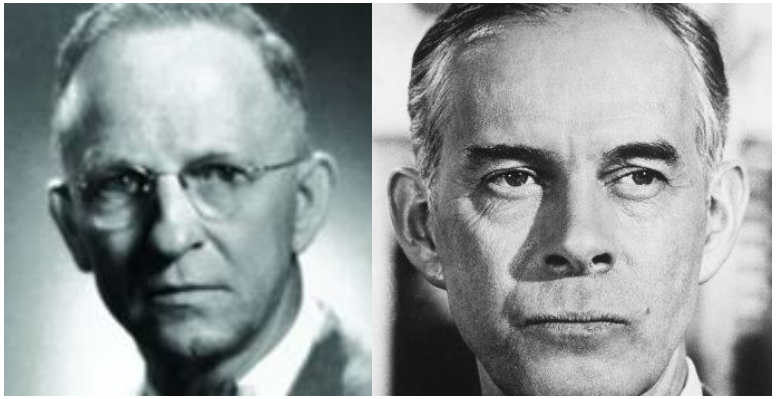
Are human entities actually identifiable?



One Harry Morgan or two?
How can we know?

(Philosophical tangent:)

Are human entities actually identifiable?

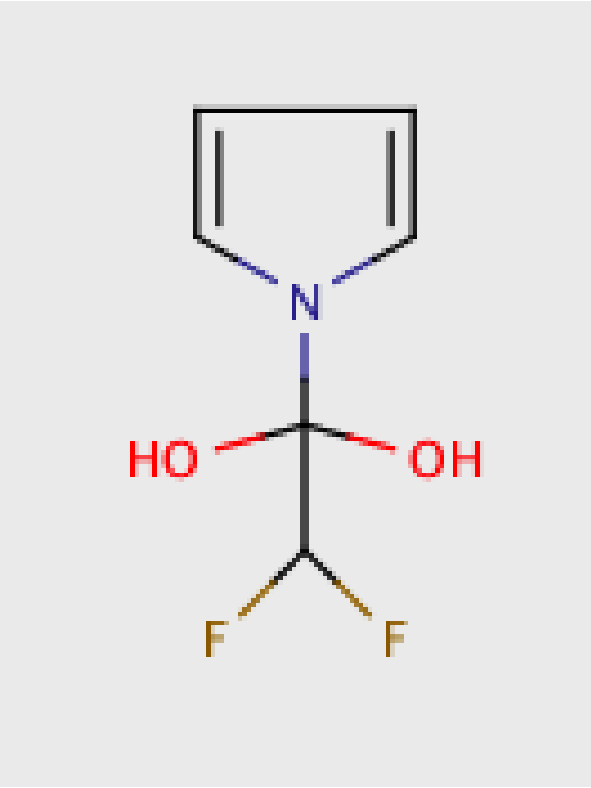


One Harry Morgan or two?
How can we know?

Individuality may be contextual.

Could I organize my address book using cheminformatics?

What would the algorithm look like?



Conclusions

- “CINF” (cheminformatics) is awesome.
- But some CINF-awesomeness is not readily transferable to other domains.
- Cannot automate logic if not logical (How many Harry Morgans?).
- Perhaps CINF-awesomeness can be used as an indexing approach for chem-related domains.

“Chester”