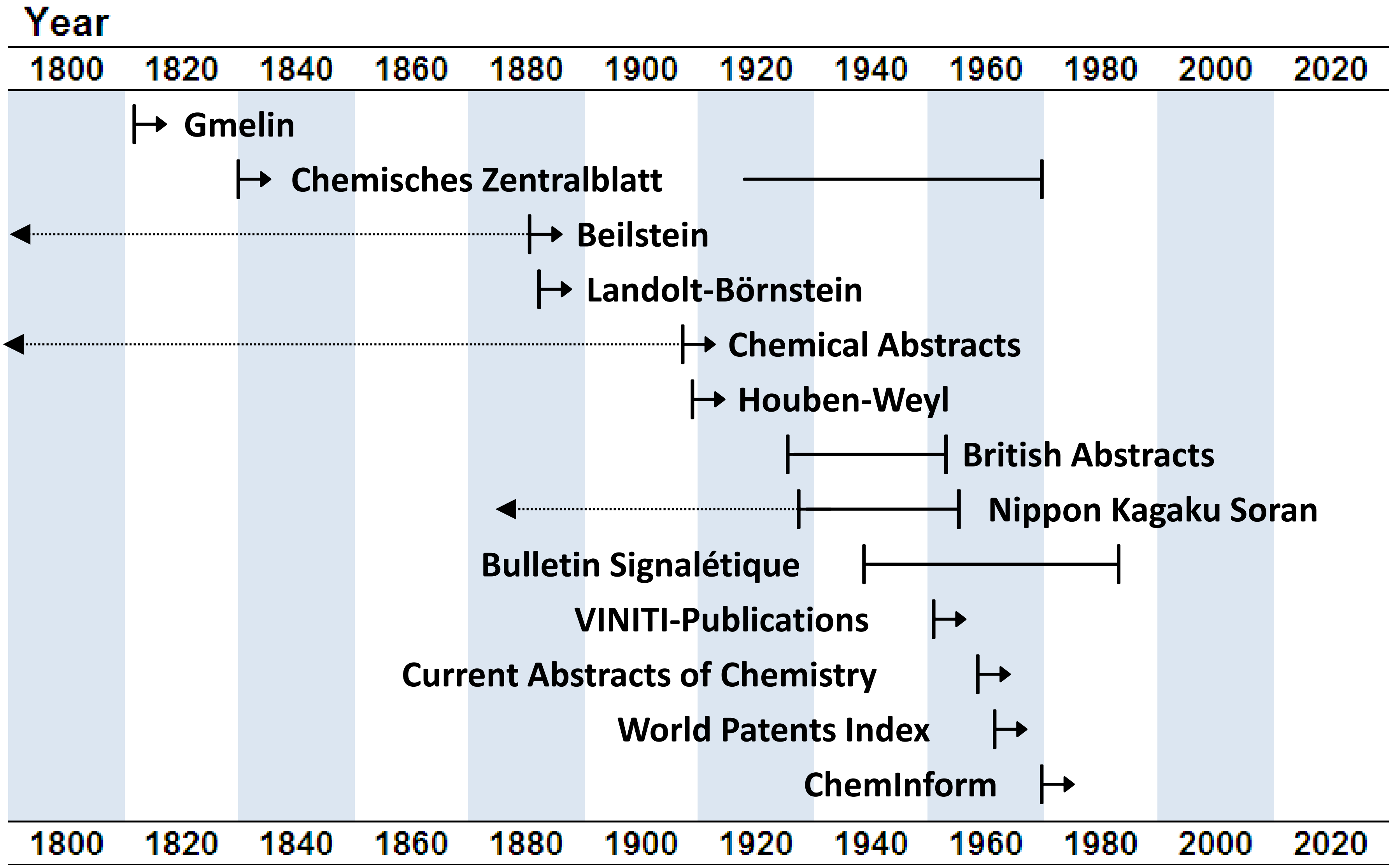
Molly Strausbaugh
Manager, CAS

# CAS substance matching technologies
# August 16, 2017

NIH meeting on IUPAC/InChI

"It's an inexact science."

# Historical Chemical Information – Secondary Literature



Gmelin

Chemisches Zentralblatt

Beilstein

Landolt-Börnstein

Chemical Abstracts

Houben-Weyl

British Abstracts

Nippon Kagaku Soran

Bulletin Signalétique

VINITI-Publications

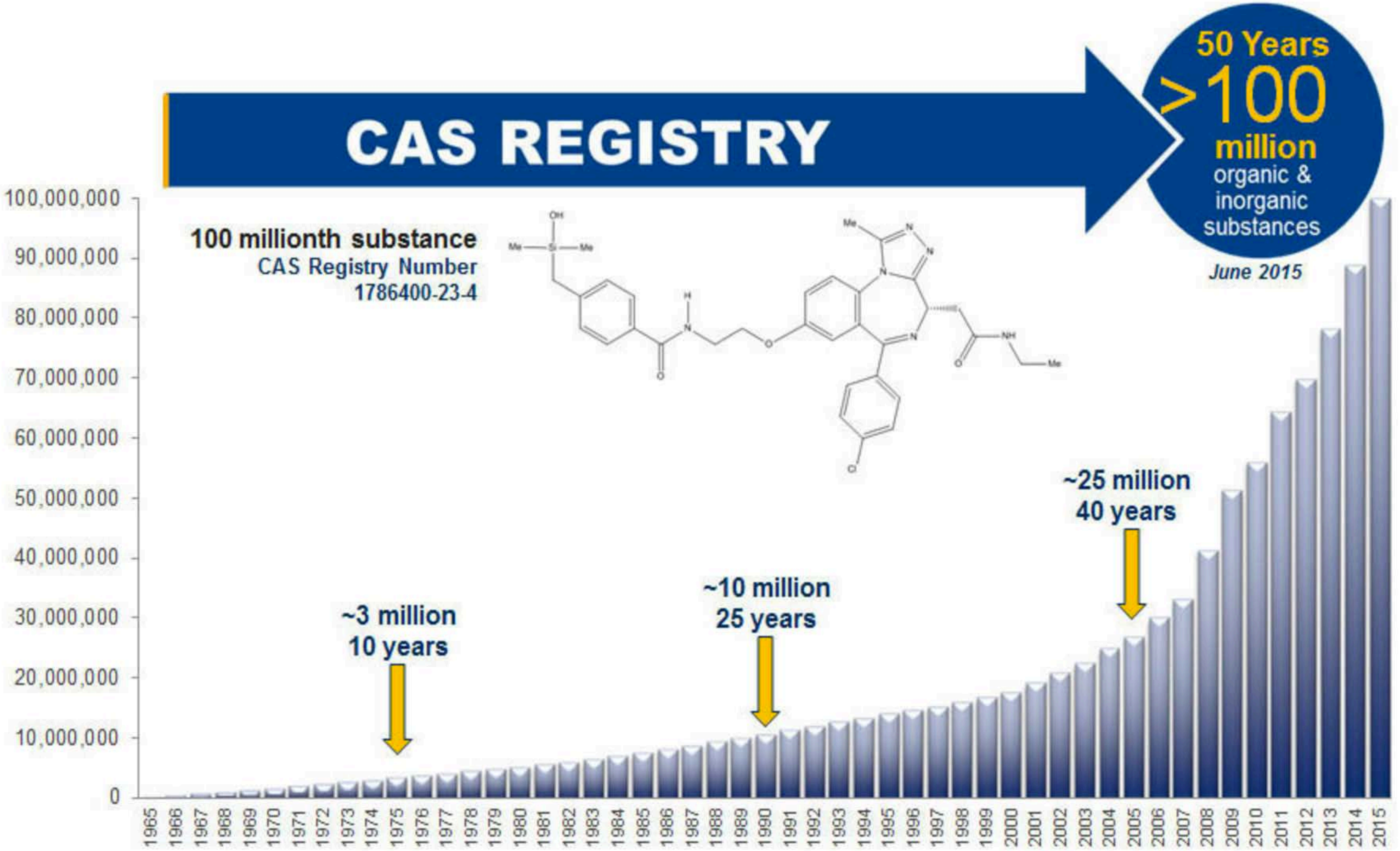
Current Abstracts of Chemistry

World Patents Index

ChemInform

*Data from Schulz, H. *From CA to CAS Online: Databases in Chemistry*, $2^{nd}$ ed.; Springer-Verlag: Berlin, 1994.

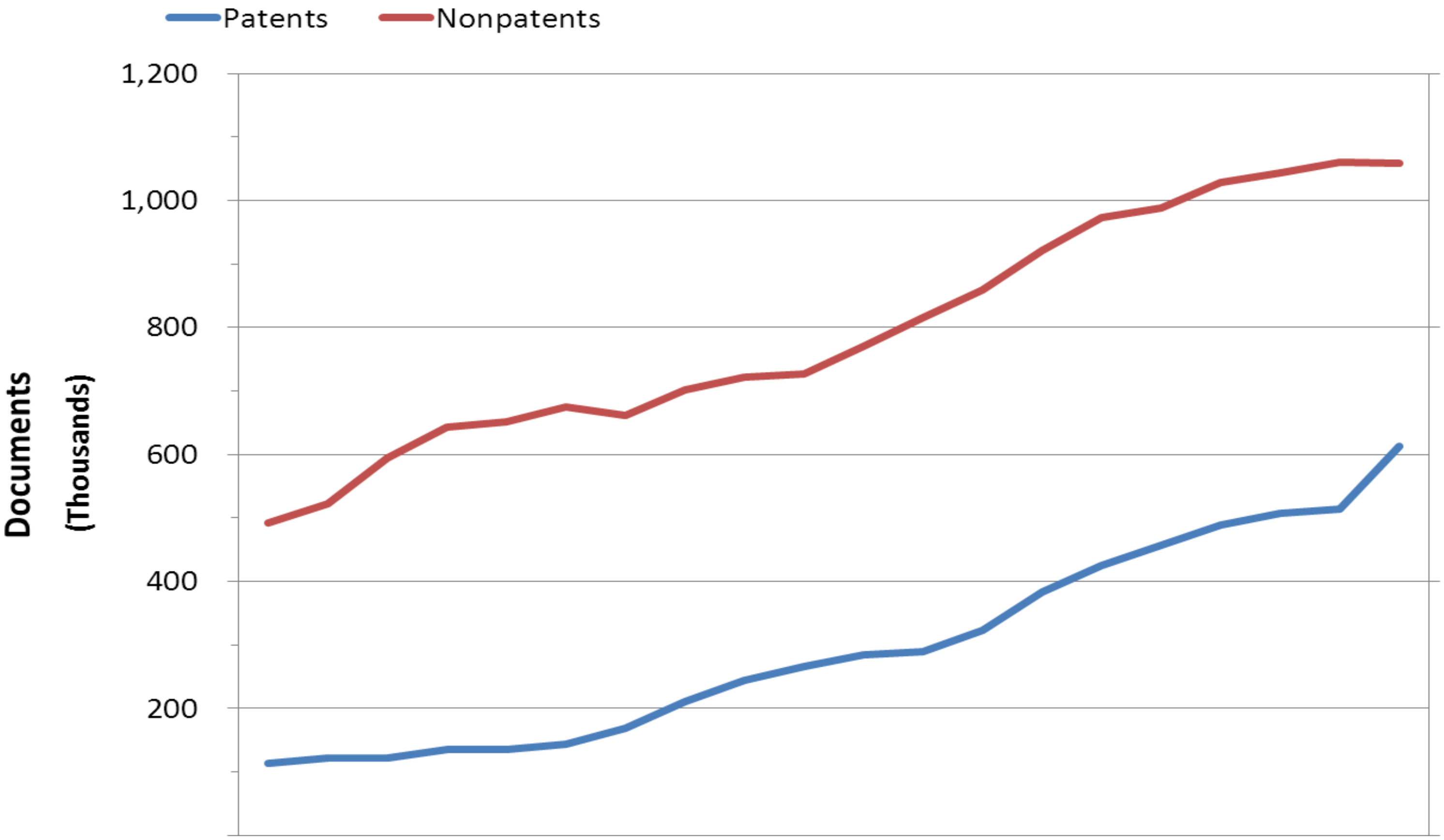
# CAS substance collection grew from ~20M to >130M in the last 20 years – document substance density is illustrative



ACS
Chemistry for Life®

CAS®
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

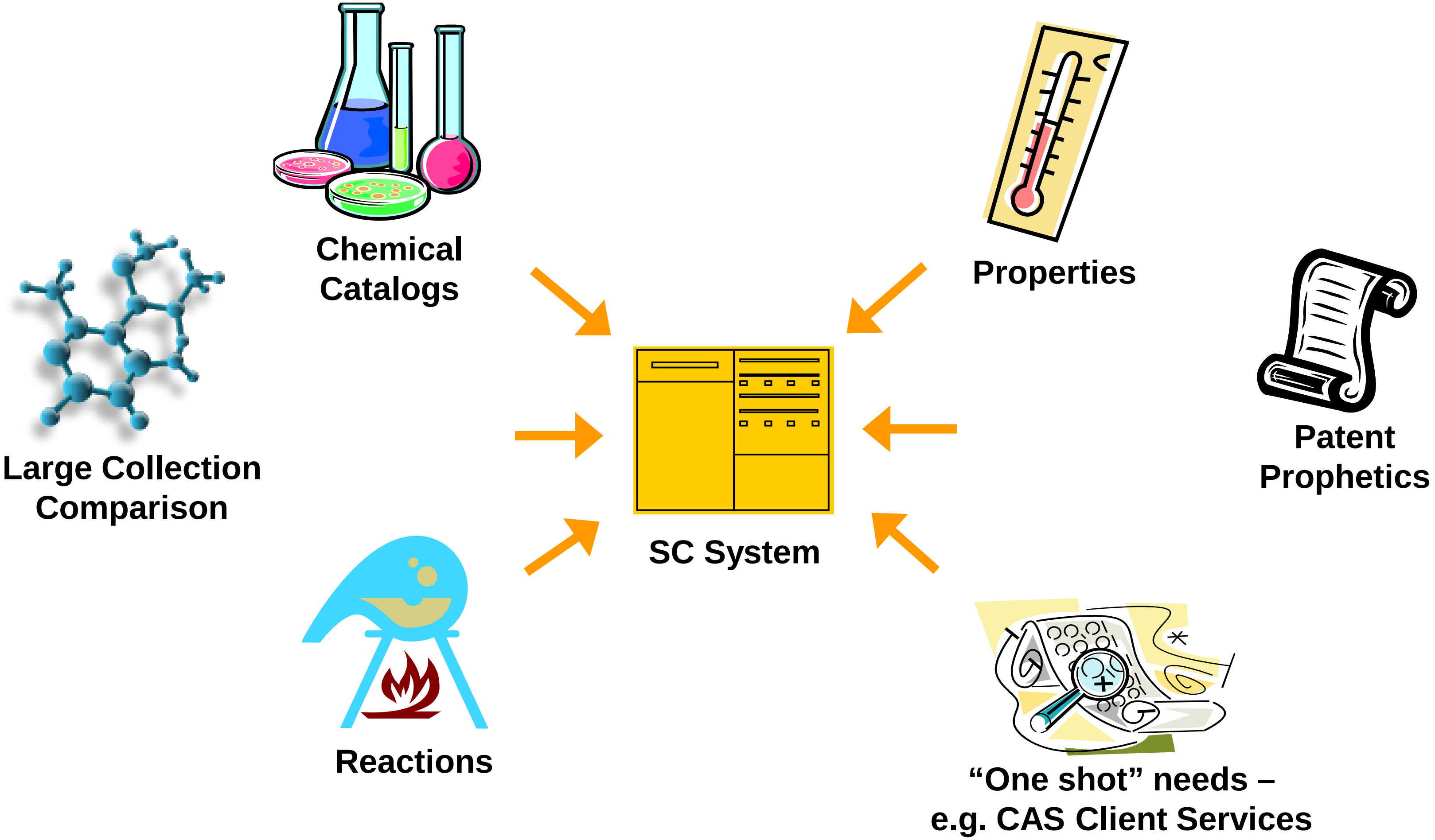
# CAS substance collection grew from ~20M to >130M in the last 20 years – document substance density is illustrative



ACS Chemistry for Life®

CAS® A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

# Shared Substance Collection Systems



Chemical Catalogs

Properties

Large Collection Comparison

Patent Prophetics

SC System

Reactions

“One shot” needs – e.g. CAS Client Services

ACS Chemistry for Life®

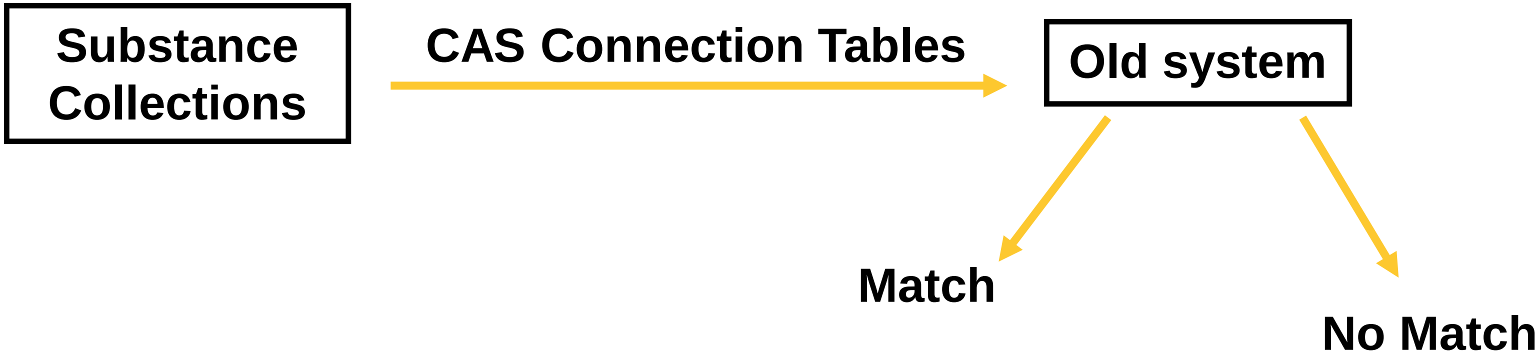
CAS® A DIVISION OF THE AMERICAN CHEMICAL SOCIETY

# A project to create a new chemistry identification technology for CAS

## Start with some history….

1. CAS perfected the connection table as a unique molecular representation for CAS Registry beginning in the 1960s.
2. CAS Online structure search of Registry in the early 1980s demonstrated the insight and time savings of searching connection tables.
3. Other molecular representations developed in the 1980s, 1990s, and 2000s and were used to build some private or public substance collections.
4. The Substance Collection project created, beginning in 2006, CAS technology that enables other molecular representations to be compared to CAS’s substance connection tables…like a translation.

# Prior to 2006, large datasets could not be processed quickly or efficiently



Substance Collections

CAS Connection Tables

Old system

Match

No Match

*Significant  manual effort was needed*

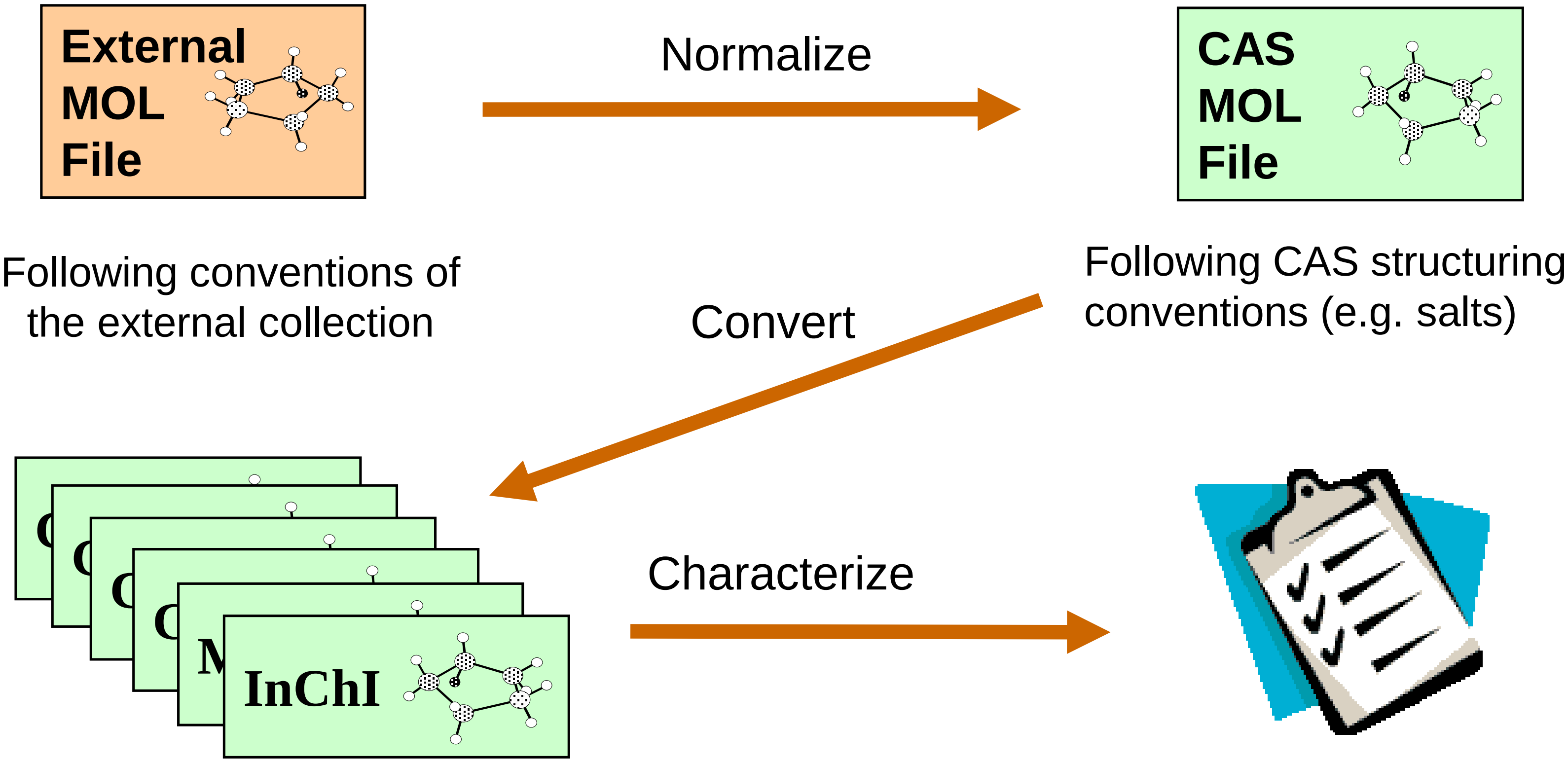
*Computer matching and registration required CAS connection tables*

*System limits restricted the size of datasets that could be processed*

# Substance Collection Project: Goals and Challenges

- Goal: Create technology to compare CAS' small molecule collection to large substance sets and to validate reputable/verifiable substances not in the CAS database
- Challenges:
  - **Compare CAS Substances** to **Large** Substance Sets (**millions** of substances)
  - **Automate** as much as possible
  - **Matching** different substance representations ("translation" required)
  - **Register reputable/verifiable substances** – establish rules
  - **Design and build software that can meet future needs** – shared substance services

# Shared Substance Services provide:
# Substance Normalizing, Converting & Characterizing



External MOL File

Normalize

CAS MOL File

Following conventions of the external collection

Following CAS structuring conventions (e.g. salts)

Convert

InChI

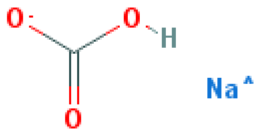
Characterize

Six different structure formats used at CAS

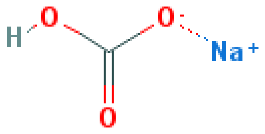
Flags indicating unusual structure types that may need special Chemist-Assisted processing (e.g. charges)

# Automated InChI matching reduces need for manual matching

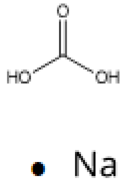
- The IUPAC International Chemical Identifier, InChI, is a protocol for converting a chemical structure (such as a MOL file) to a unique, unambiguous text string.
- CAS enhances the input MOL file so that the resulting InChI follows CAS chemical conventions (such as stereo, charges, etc.) and thus improves hit rates



InChI=1S/CH2O3.Na/c2-1(3)4;/h(H2,2,3,4);/p-1

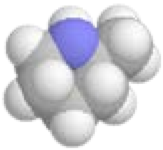


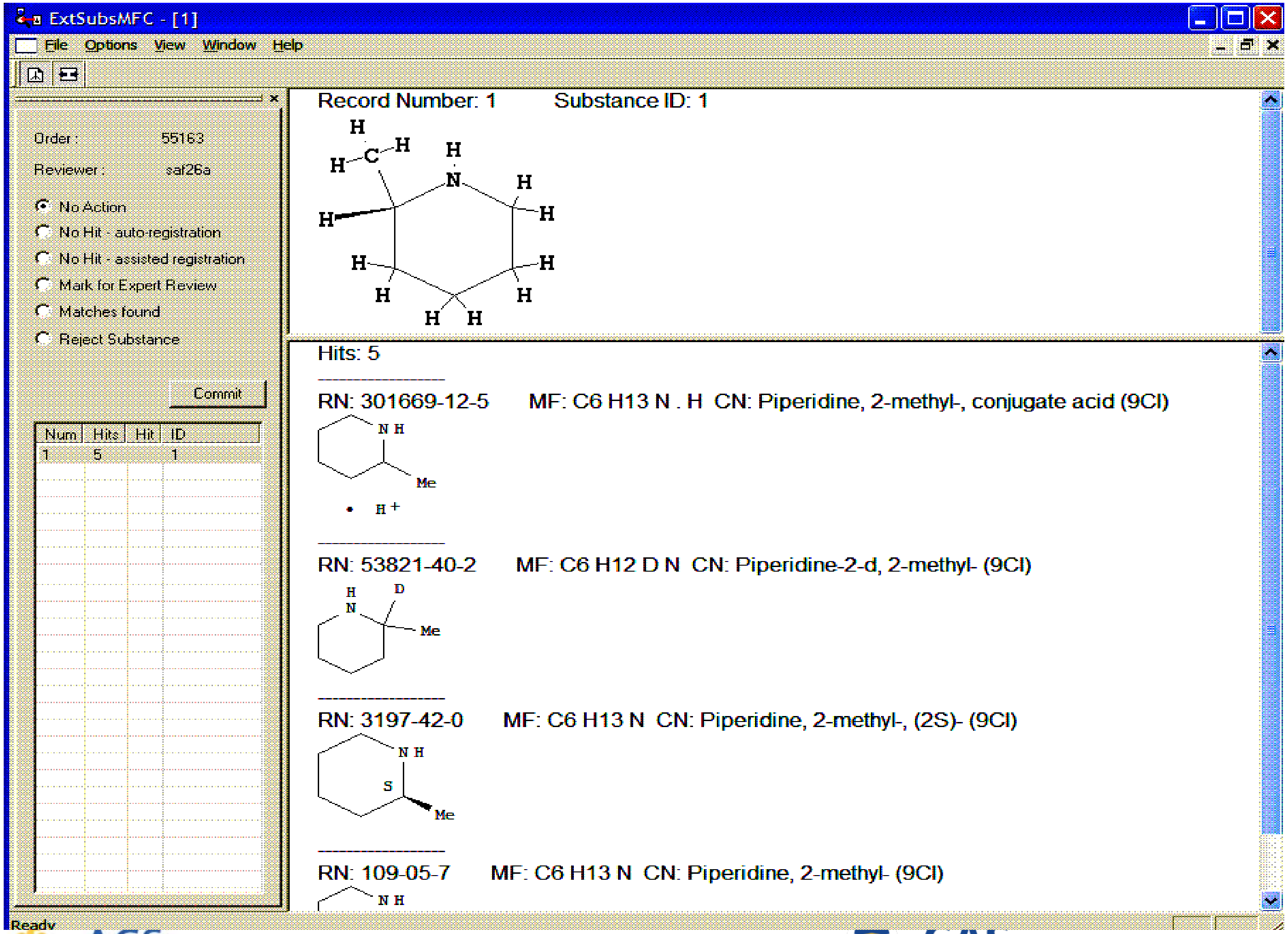
InChI=1S/CH2O3.Na/c2-1(3)4;/h(H2,2,3,4);/q;+1/p-1)



InChI=1/CH2O3.Na/c2-1(3)4;/h(H2,2,3,4)

# For the substances that don’t auto match, closest possible matches are grouped for review





Order : 55163

Reviewer : saf26a

- ● No Action
- ○ No Hit - auto-registration
- ○ No Hit - assisted registration
- ○ Mark for Expert Review
- ○ Matches found
- ○ Reject Substance

Commit

| Num | Hits | Hit | ID |
|---|---|---|---|
| 1 | 5 |  | 1 |

Record Number: 1 Substance ID: 1

Hits: 5

RN: 301669-12-5 MF: C6 H13 N . H CN: Piperidine, 2-methyl-, conjugate acid (9CI)

RN: 53821-40-2 MF: C6 H12 D N CN: Piperidine-2-d, 2-methyl- (9CI)

RN: 3197-42-0 MF: C6 H13 N CN: Piperidine, 2-methyl-, (2S)- (9CI)

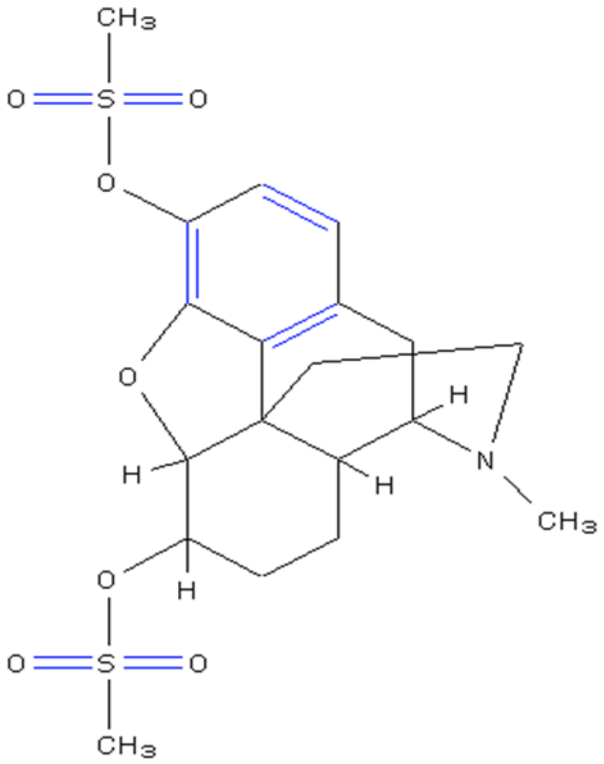
RN: 109-05-7 MF: C6 H13 N CN: Piperidine, 2-methyl- (9CI)

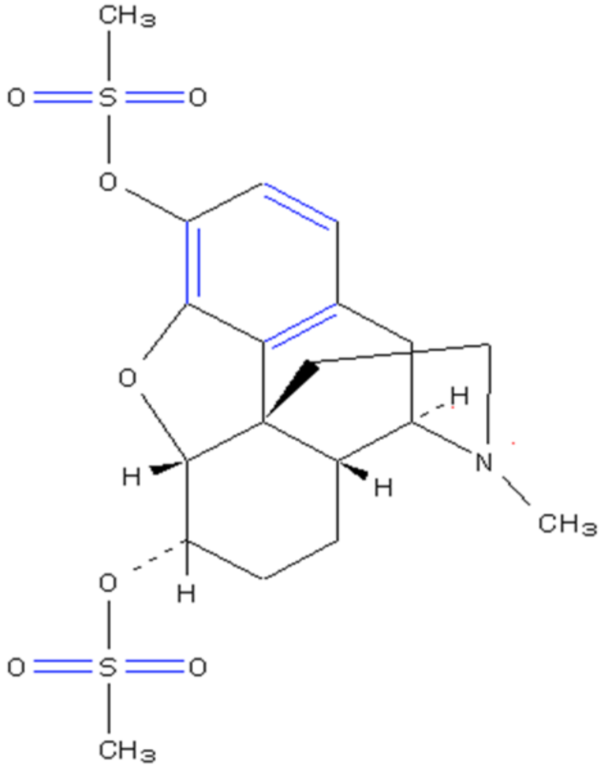
Ready

# Assisted Chemist Review - Illustration

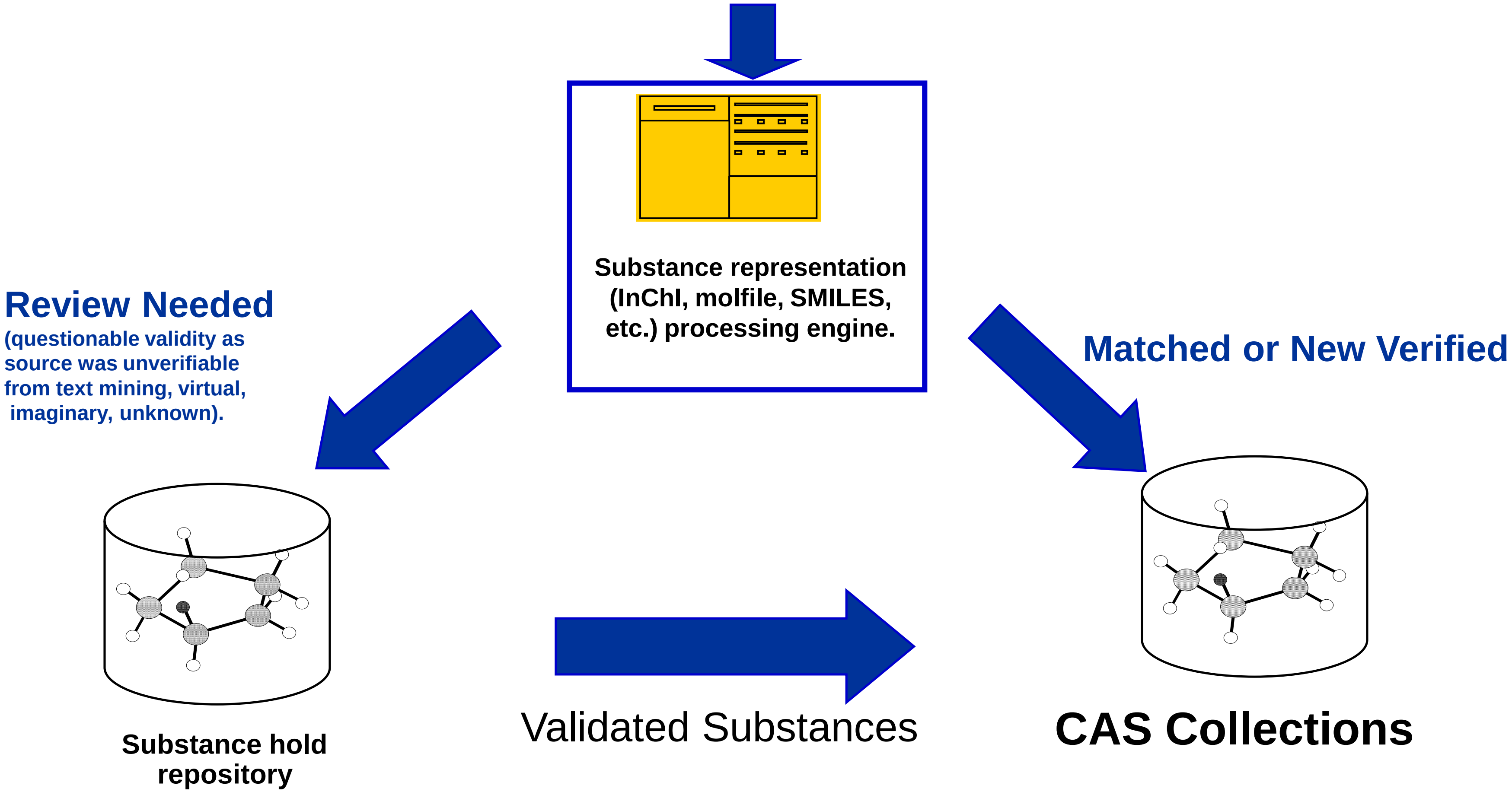
- Molecular representation of a morphine derivative is incomplete: Stereochemical information is missing on 5 atoms

- Chemist-Assisted Process adds the missing stereochemical information





# Outside collections are thoroughly reviewed and eligible substances captured



Substance representation (InChI, molfile, SMILES, etc.) processing engine.

**Review Needed**

(questionable validity as source was unverifiable from text mining, virtual, imaginary, unknown).

**Matched or New Verified**

Substance hold repository

Validated Substances

**CAS Collections**

# CAS substance collection technology provides…

- Ability to compare and analyze substance databases vs. CAS's collections
- Automated ability to handle and qualify new sources of substance information from collections
- Faster and more complete matching via
  - InChI**,** SMILES, molfile
  - Automation of many formerly manual steps
- Foundation software for other substance efforts
  - **e.g.** Matching substances in purchased chemical properties databases

# Thank you

# Questions?