# Extracting Knowledge from Reaction Databases: Developments from InfoChem

**CICAG meeting 3rd July 2013**

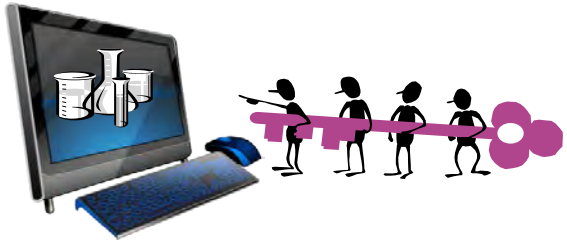
**Stephanie North, Allyl Consulting Ltd, representing InfoChem in the UK**

H. Kraut, H. Matuszczyk, H. Saller, J. Eiblmaier, P. Loew
InfoChem GmbH, Landsberger Strasse 408, Munich, 81241, Germany

# Outline

**Setting the scene: millions of reactions**

- Rapid growth of published reaction databases e.g. SPRESI, 4.2 M reactions
- Electronic Laboratory Notebooks now provide corporate reaction databases
- Reaction knowledge a key driver for chemists' innovation and decision making processes - but how best to unlock and exploit resource?
- InfoChem's CLASSIFY technology provides the solution



**Today's presentation**

- Introduce InfoChem
- Describe the reaction classification concept
- Show how CLASSIFY works and analyse the results
- Benefits of extending CLASSIFY to corporate reaction databases

# InfoChem at a Glance

## Company

- specializes in chemoinformatics
- founded in 1989
- based in Munich, Germany
- majority owned by Springer Science + Business Media
- subsidiary InfoChimia since 1990


## People

- 24 full time employees (Munich office)
- 2 representatives in UK
- 3 representatives in Sweden
- 60 freelance abstractors (residing offshore)

## Business Areas

- Software products
- Projects
- Text/Data mining
- Database building

## Revenues

Approx. 3.3 million Euros (2012):

- Software products (30%)
- Database building (35%)
- Projects (12%)
- Text/Data mining (23%)

# Business Areas

## Software

- IC*FSE*, IC*CARTRIDGE*, IC*CHEMDESK*
- IC*MAP*, CLASSIFY, IC*NameRXN*
- IC*SYNTH*, IC*EDIT*, IC*TOOLS*

## Services

- Project development
- Consulting
- Database building

InfoChem

## Content

- SPRESI[web], SPRESI*mobile*
- Chemisches Zentralblatt Structural Database

## Text/Data Mining

- Chemical entity recognition
- Name to structure conversion
- Image to structure conversion

# Fundamentals of reaction Classification Concept

- **What does it do?** Organises and analyses large reaction databases according to the type or ‘class’ of transformation

- **How does it work?** Groups similar transformation types together sharing same reaction centres and same environment

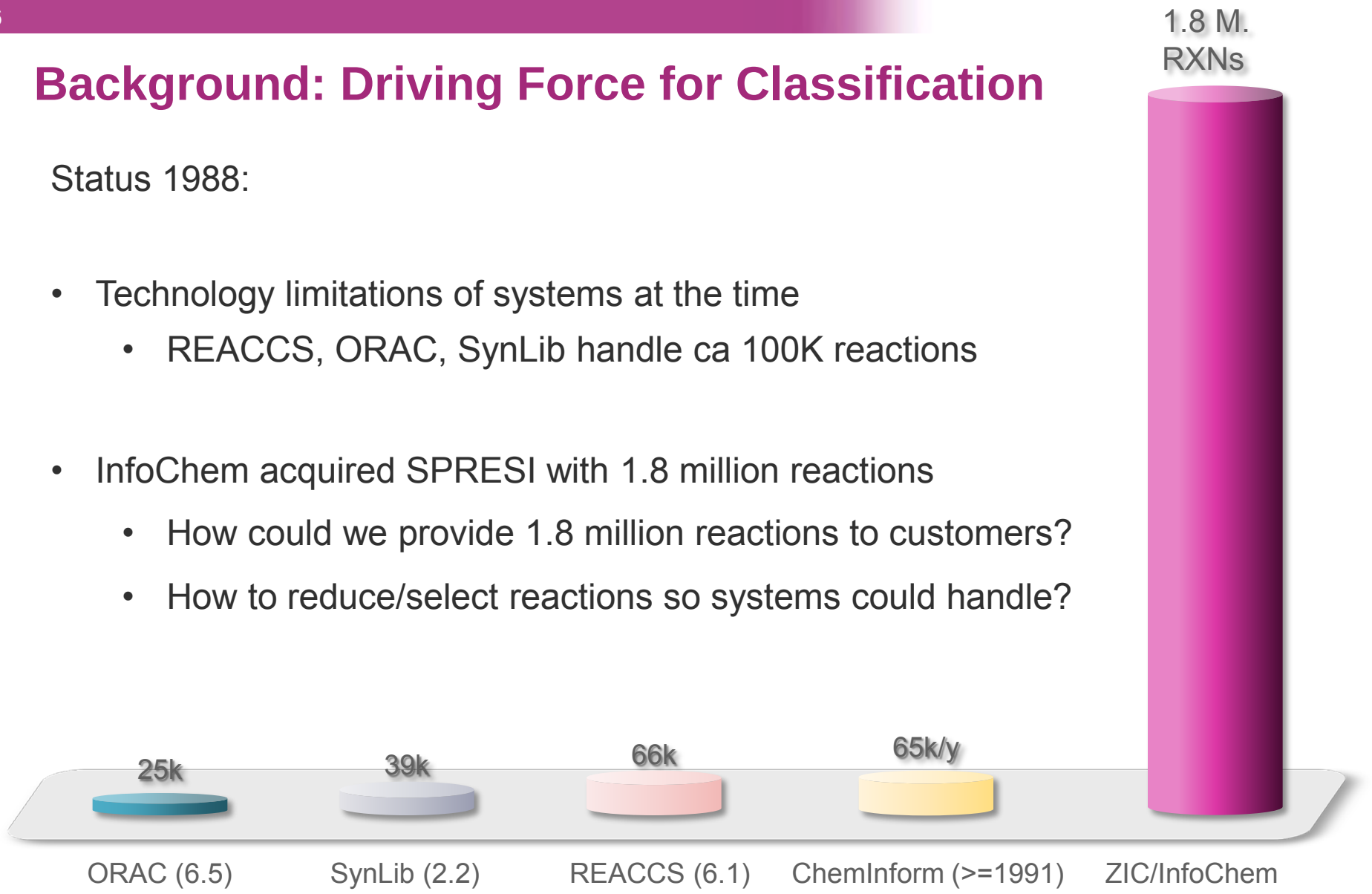
- **What are the benefits?** Gain knowledge about the type, quality and diversity of reactions to aid inventive process and decision making

# Background: Driving Force for Classification

Status 1988:

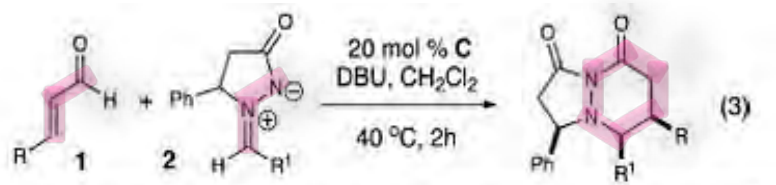
- Technology limitations of systems at the time
  - REACCS, ORAC, SynLib handle ca 100K reactions

- InfoChem acquired SPRESI with 1.8 million reactions
  - How could we provide 1.8 million reactions to customers?
  - How to reduce/select reactions so systems could handle?



# Concept of Reaction Classification

- Develop a chemically intelligent way to choose an example reaction from a given class of reactions
- Group similar reactions e.g. like chemistry is published

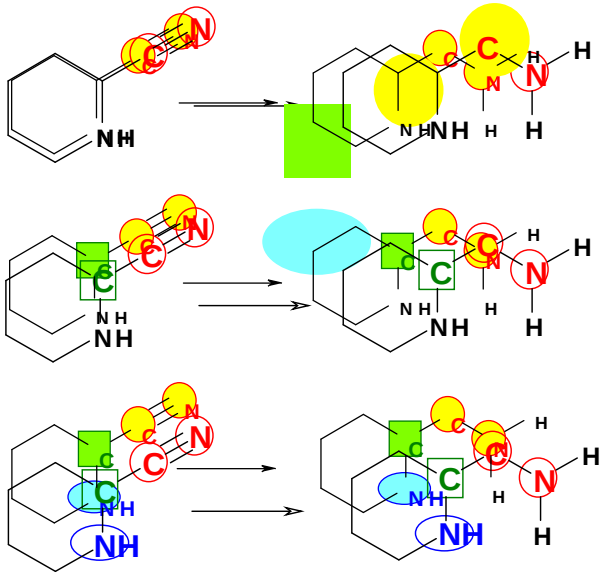



| entry | R | R¹ | product | yield (%)[b] | dr[c] |
|---|---|---|---|---|---|
| 1 | Ph (1a) | Ph (2a) | 4 | 79 | >20:1 |
| 2 | 4-OMePh (1b) | Ph (2a) | 5 | 76 | >20:1 |
| 3 | 3-OMePh (1c) | Ph (2a) | 6 | 79 | >20:1 |
| 4 | 2-OMePh (1d) | Ph (2a) | 7 | 94 | >20:1 |
| 5 | 2-naphthyl (1e) | Ph (2a) | 8 | 77 | >20:1 |
| 6 | $CH_3CH_2CH_3$ (1f) | Ph (2a) | 9 | 67 | >20:1 |
| 7 | $HC{=}CHCH_3$ (1g) | Ph (2a) | 10 | 51 | >20:1 |

- Select one example transformation to represent the group
- The result is a smaller selective database consisting of one example for each transformation type

# CLASSIFY Classification algorithm: how does it work?

- For each reaction, determine and map reaction centers
- Calculate atom hash codes for the reaction centre
- Atoms included depend on extent of reaction centre's chemical environment considered, resulting in different sized search hit lists (clusters):



**0-Sphere (BROAD)**

**Reaction centers only**

**1-Sphere (MEDI**

**Reaction centers plus alpha atoms, excluding hydrogens**

**2-Sphere (NARROW)**

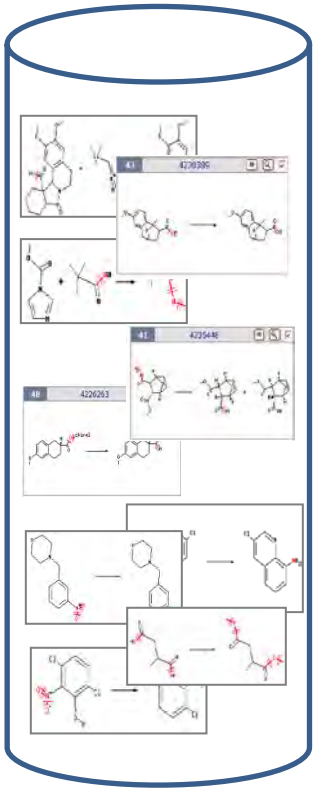
**Reaction centers plus beta atoms, excluding hydrogens and consecutive $sp^3$-atoms**

Large – sized clusters

Small-sized clusters

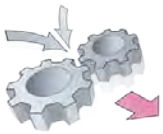
- The sum of the hash codes from these calculations provides the unique InfoChem Reaction Classification Code. Three codes are assigned to each reaction

# Classification results



Large database
e.g. SPRESI

*CLASSIFY*



Each reaction processed

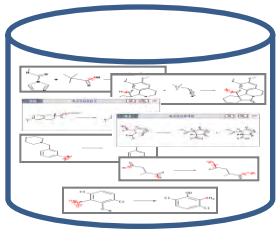
ClassCode determined

Cluster according to ClassCodes

| Codes | # RXNs |
|---|---|
| 3257410 82969498 | 13,000 |
| 2945604 35478524 | 12,000 |
| 2969666 48127528 | 11,000 |
| 4257410 82969497 | 10,000 |
| 5945604 35478523 | 9,500 |
| 6969666 38127527 | 9,000 |
| 3257410 72969496 | 8,500 |
| 5945604 25478522 | 8,000 |
| 4969666 38127526 | 7,000 |
| 2257410 72969495 | 6,000 |
| 4945604 25478521 | 5,000 |
| 3969666 38127524 | 4,000 |
| 3257410 72969494 | 3,000 |
| 2945604 25478522 | 2,000 |
| 2969666 38127523 | 1,000 |
| 4257410 72969495 | 500 |
| 5348521 83070506 | 90 |
| 7560743 05292708 | 25 |
| 8671845 16303819 | 1 |
| ..... | ..... |

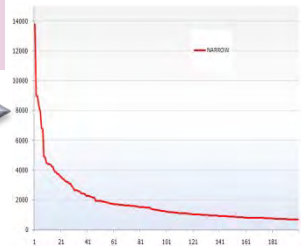
Groups of reaction types according to ClassCodes

Select one sample for each ClassCode



Smaller representative database e.g. ChemReact

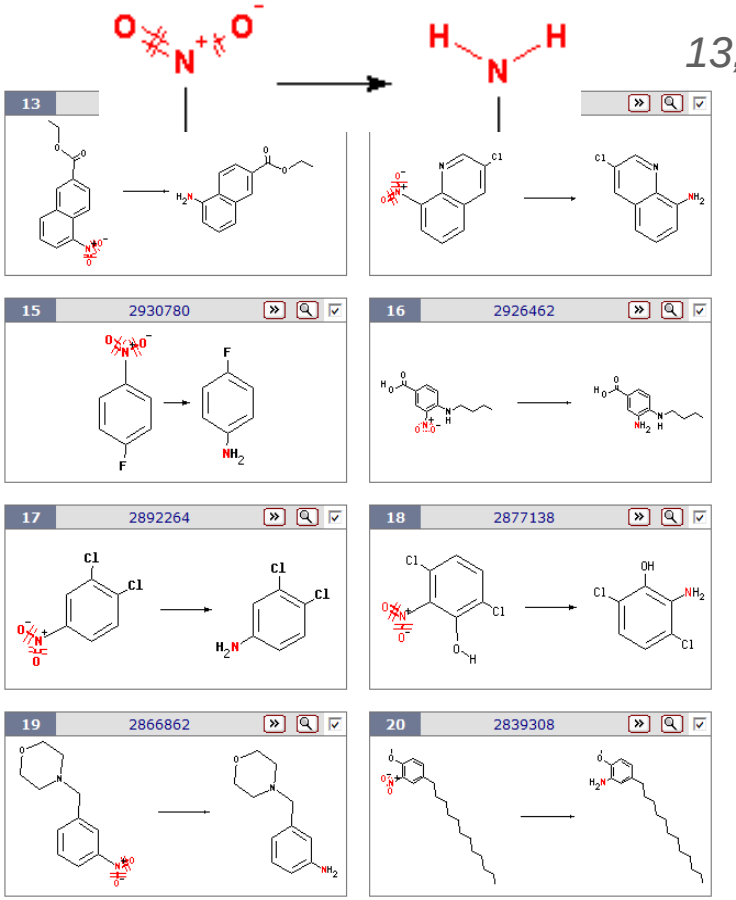
Further analysis of clusters



Extract knowledge from classification

# ClassCodes: representative database creation

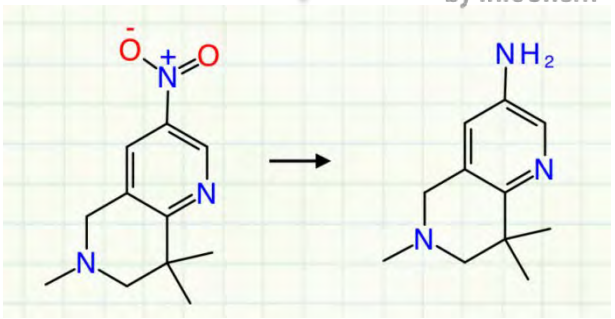
- Apply NARROW Classification to SPRESI
- Most frequent ClassCode reduction of a nitro group:




*13,794 examples*

ONE representative sample reaction per ClassCode from SPRESI used to create ChemReact database.
Selection criteria: yield, journal, year of publication etc

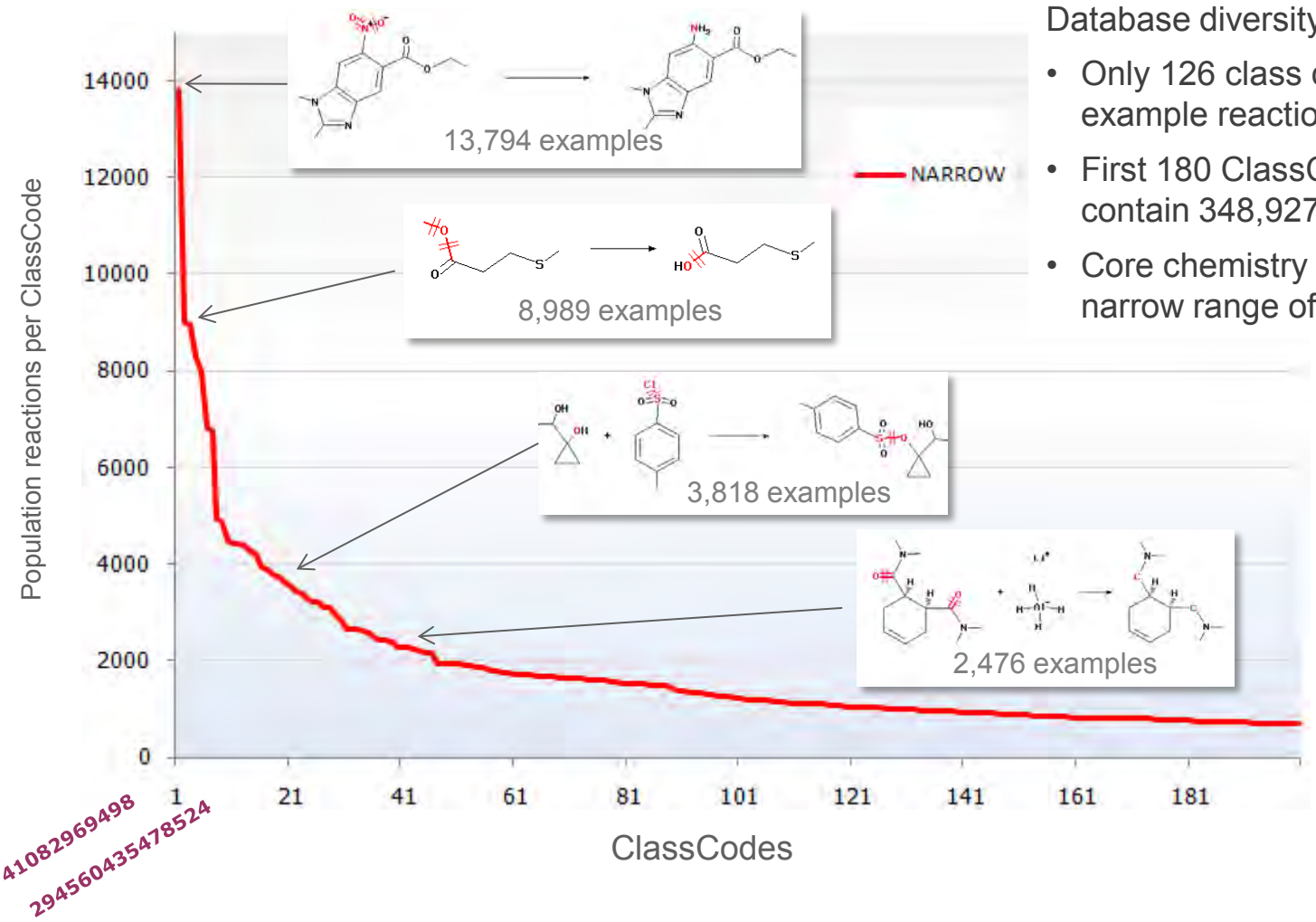
SPRESI*mobile* by InfoChem



# Sample reactions from Classcodes with high popluations

| ClassCode | Reaction examples in SPRESI (Narrow) | ChemReact sample reaction |
|---|---|---|
| 3257410<br>82969498 | 13,794 … | |
| 2945604<br>35478524 | 8,989 … | |
| 2969666<br>48127528 | 8,947 … | |

# Analysis: Frequency distribution of SPRESI ClassCodes (narrow)

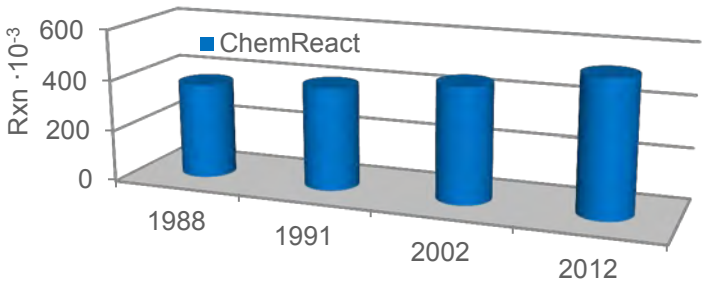



Database diversity

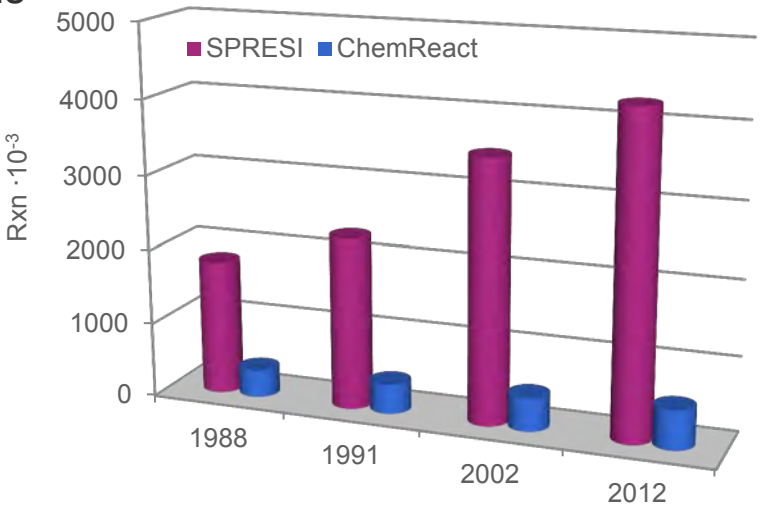
- Only 126 class codes > 1,000 example reactions
- First 180 ClassCodes combined contain 348,927 reactions
- Core chemistry within fairly narrow range of reaction types

# Classification of SPRESI to ChemReact since 1988

| Release | SPRESI | ChemReact |
|---|---|---|
| 1988 | 1,800 K | 370 K |
| 1991 | 2,300 K | 400 K |
| 2002 | 3,500 K | 450 K |
| 2012 | 4,260 K | 524 K |




- Steady increase in diversity of organic synthesis specific reactions
- Growth of ChemReact not proportional to SPRESI database (overall ca 1.4x vs. 2.4x)
- New reaction references in literature do not necessarily generate new ClassCodes (chemistry)
- Organic synthesis is described by a comparatively small amount of reaction types

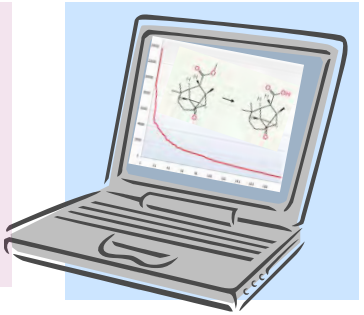


# Classfication for Corporate Reaction Databases

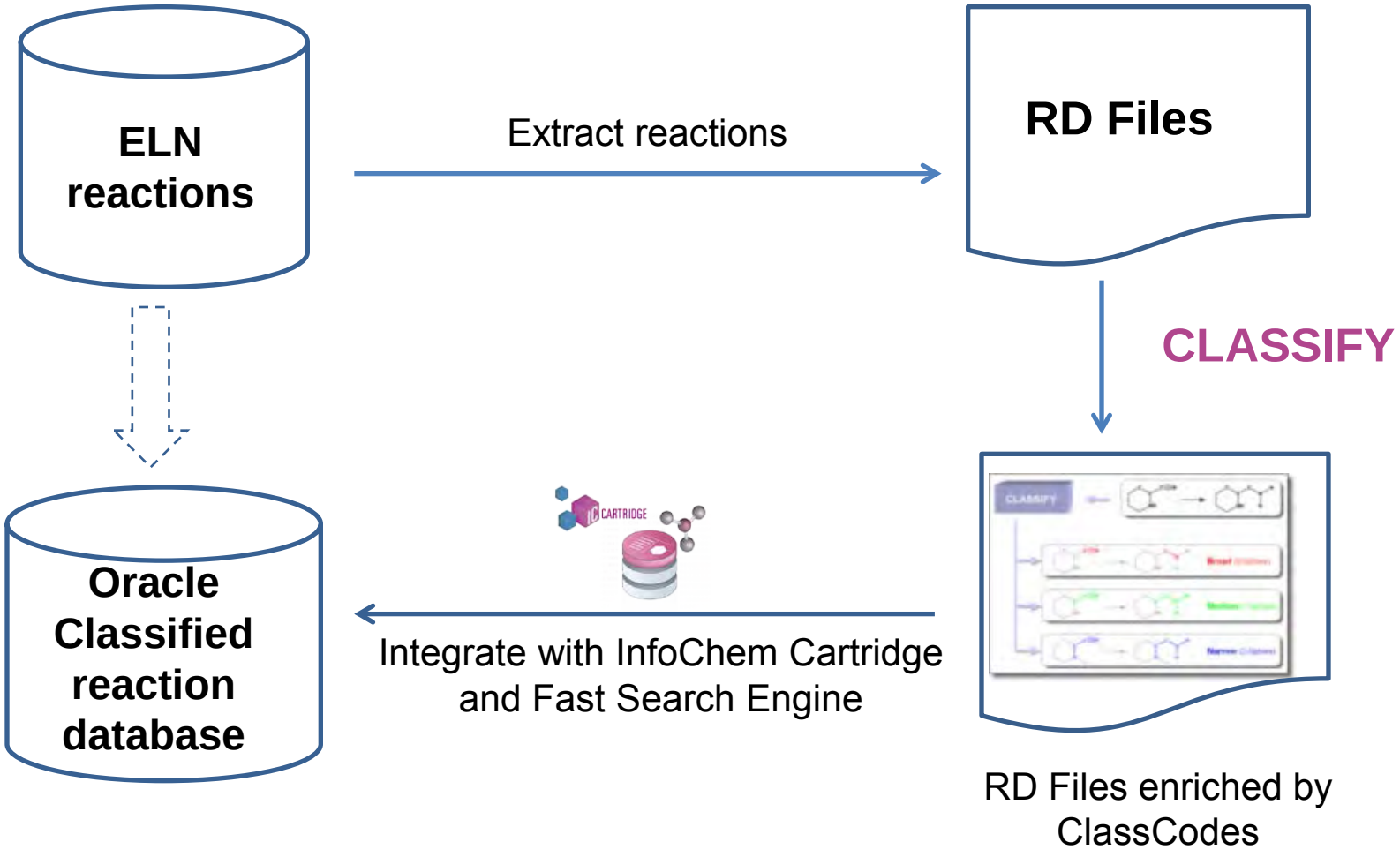
- Organisations are making significant investments in ELNs to capture intellectual property
- ELNs generate large repositories of corporate reactions
- Accessibility of reactions depends on type of ELN

How to benefit and unlock this valuable resource for improved decision making and reaction planning?

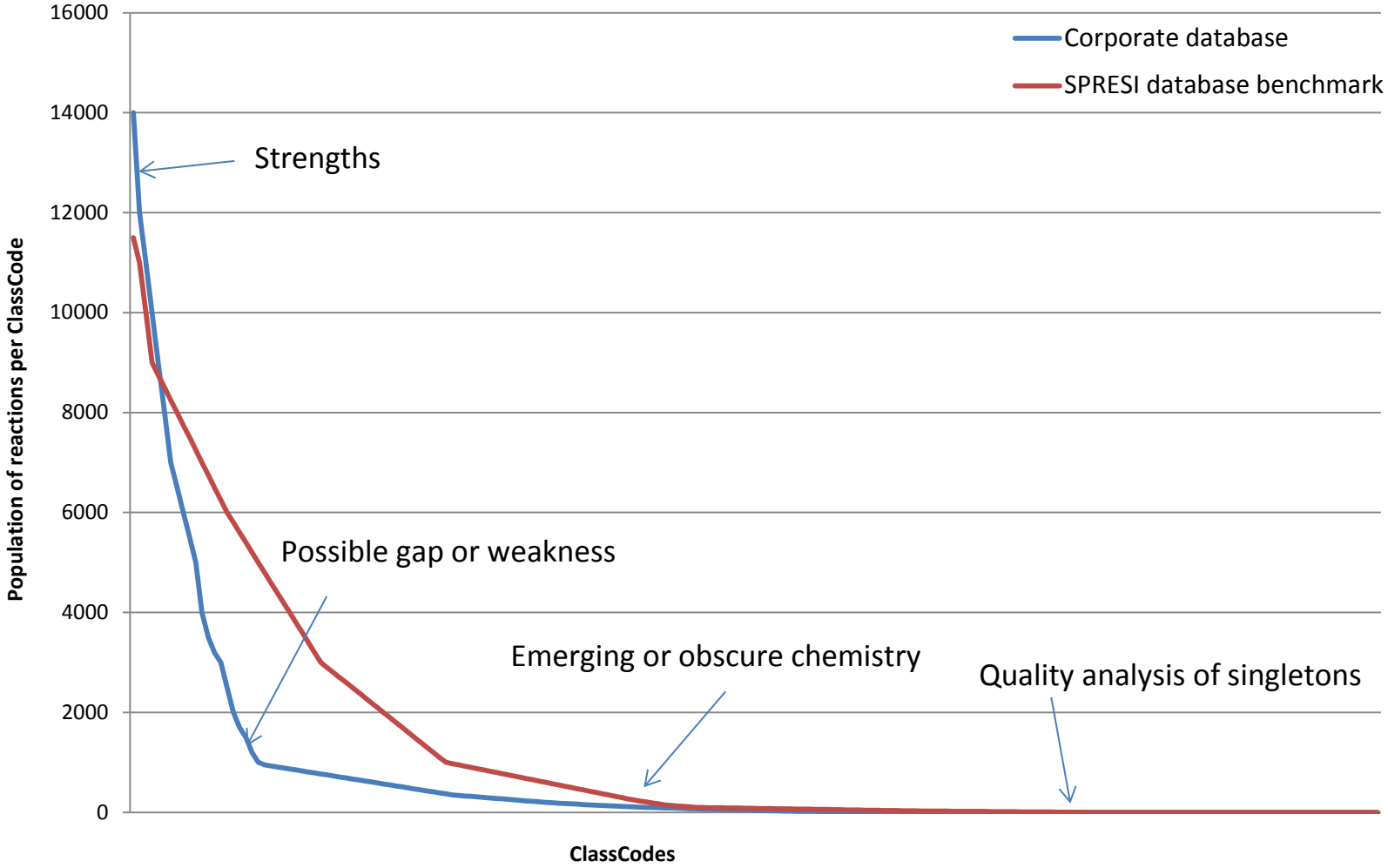

- CLASSIFY solution to automatically categorize and analyse corporate reactions
- ICCartridge and Fast Search Engine for storage and fast retrieval of corporate chemical reactions and data



# Generation of ELN Classified Reaction database



# Possible Frequency Distribution of a Corporate database



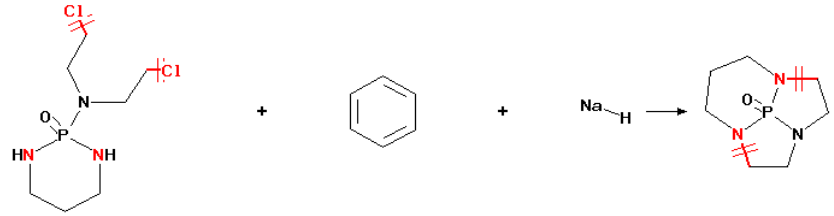
# CLASSIFY unlocks corporate reaction knowledge

Analysis of ClassCodes and frequency distribution provides knowledge about:

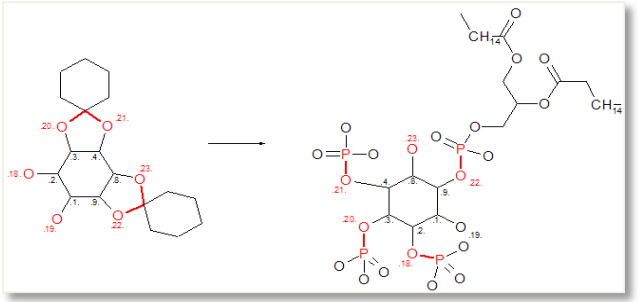
- Diversity, strengths and gaps in corporate reactions
- Optimum yields, reagents and conditions for a reaction type
- Low frequency reactions may indicate
  - Emerging valuable new reactions
  - Obscure but useful chemistry not often repeated
- Database Quality
  - Few Singletons (one reaction in a ClassCode) indicate database has a certain quality standard
  - Many singletons may indicate data quality problems

# Singleton reactions: deeper look

- New/unique/weird but valid chemistry



- Erroneous or inconsistent reaction or data entry
  - roles, missing reactants, different solvent/catalyst/reagent/reactant definitions
- Inconsistent normalization rules
- Inadequate or incorrect mapping
- Multistep reactions as overall reactions e.g. overall step in 13 step reaction:



Da-Ming Gou and Ching-Shih Chen: Synthesis of L-α phosphatidyl-D-myo-inositol 3,4,5-trisphosphate, an important intracellular signalling molecule, *J. Chem. Soc., Chem.Commun.*, 1994, 2125

**Spresi Doc-ID 3396558**

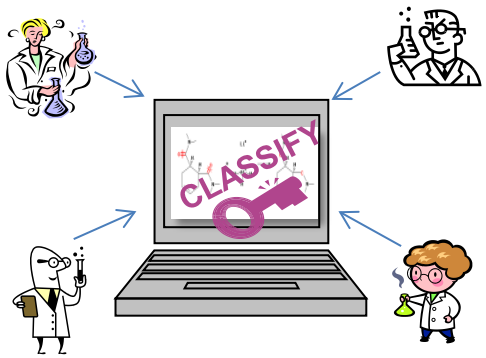
# Summary of CLASSIFY Benefits

**Diversity** of reaction types identifies strengths and gaps in corporate knowledge
- Does the business need to focus on alternative reaction classes?
- Location of failed internal reactions (how/why) to drive reaction improvement

**Quality**: singleton reaction types may indicate erroneous data
- Drive data quality improvements and business rule compliance

**Improved route design and innovation**
- Maximum value obtained from corporate reaction expertise
- Analysis of reactants and conditions within a reaction class to determine best practice leading to simplification and economy of processes
- Chemists alerted to emerging or novel chemistry, thus speeding discovery
- Enables selection of viable reactions and efficient route maps

**Knowledge sharing**
- Only way to link and compare reaction databases
- Reaction knowledge and expertise shared across organisation



## Acknowledgments: InfoChem Team


- Hans Kraut
- Henry Matuszczyk
- Heinz Saller
- Josef Eiblmaier
- Valentina Eigner-Pitto
- Ulf Frieske
- Peter Loew


Look out for publication underway

# Thank you