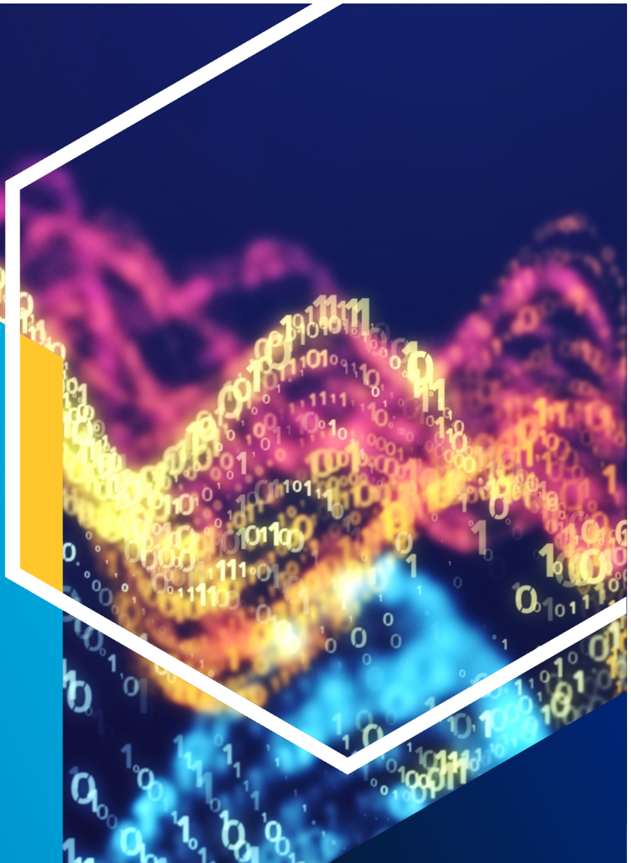
# THE IMPORTANCE OF DATA QUALITY IN SCIENTIFIC AI



Jacob Al-Saleem, Ph.D.

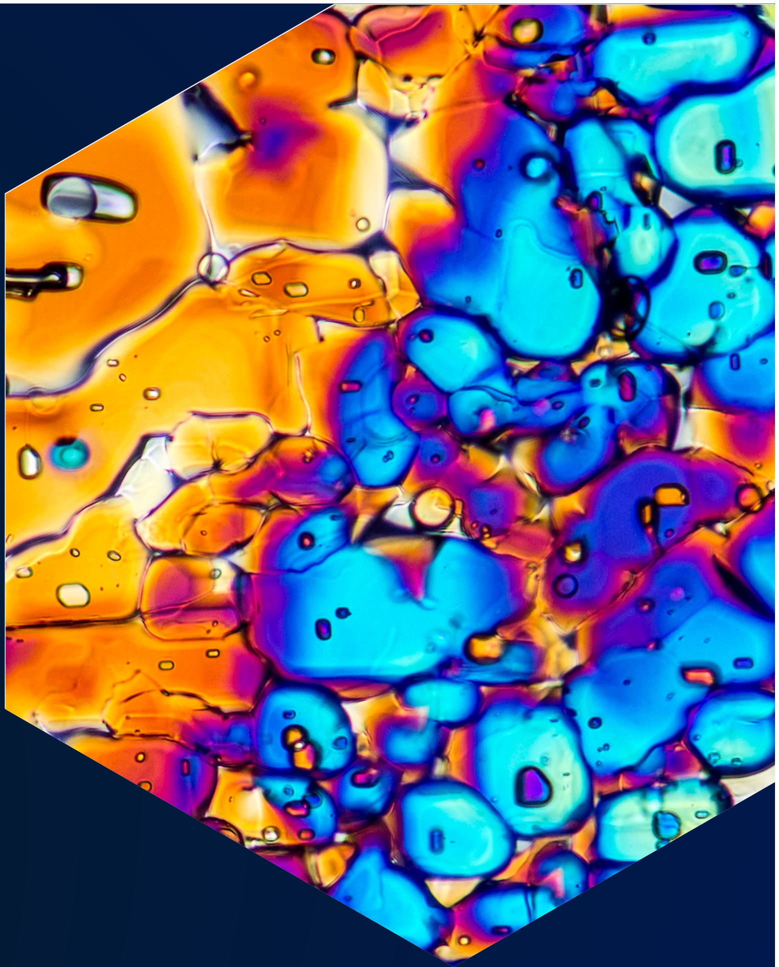
CAS
A division of the
American Chemical Society

# About CAS

CAS connects the world's scientific knowledge to accelerate breakthroughs that improve lives.

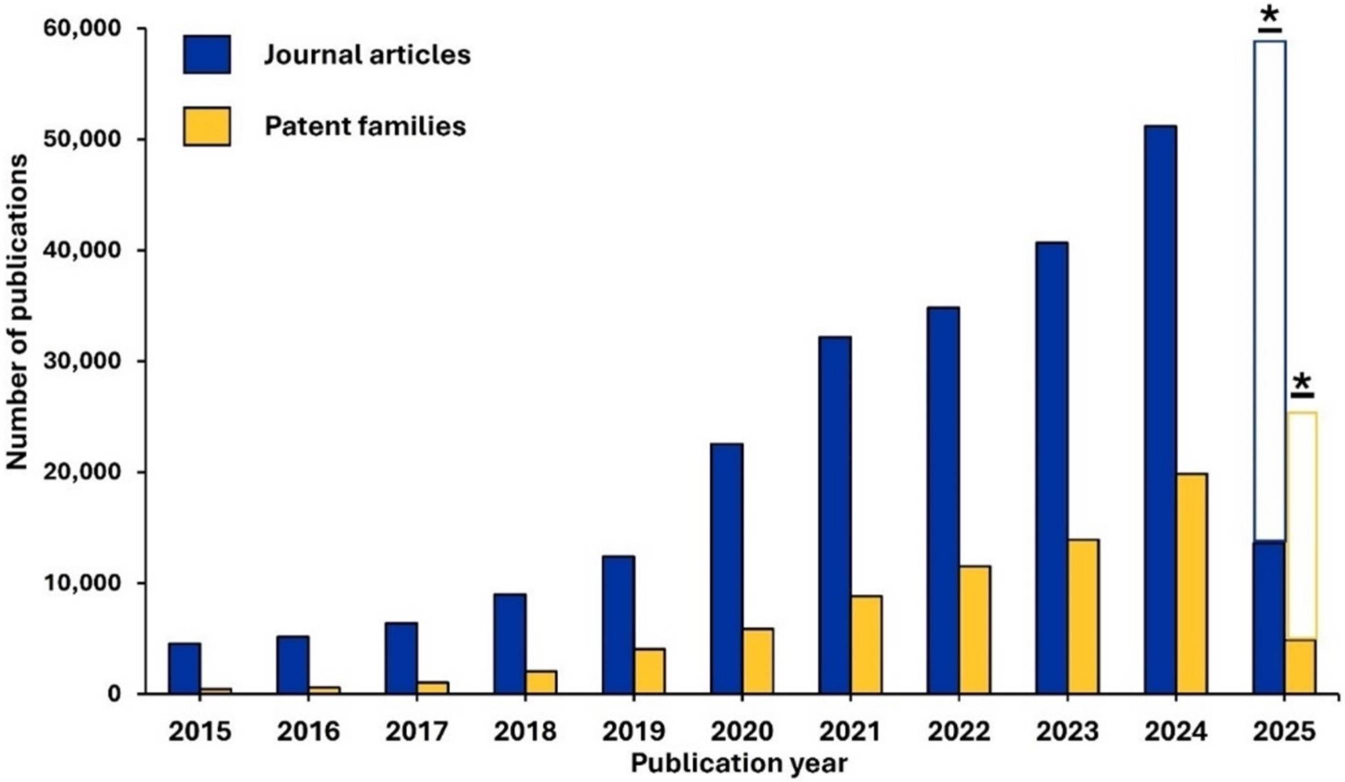
As a specialist in scientific knowledge management, we empower global innovators across industries with the essential data, solutions, services, and expertise needed to navigate today's complex digital information landscape and make more confident decisions.

CAS is a division of the American Chemical Society.

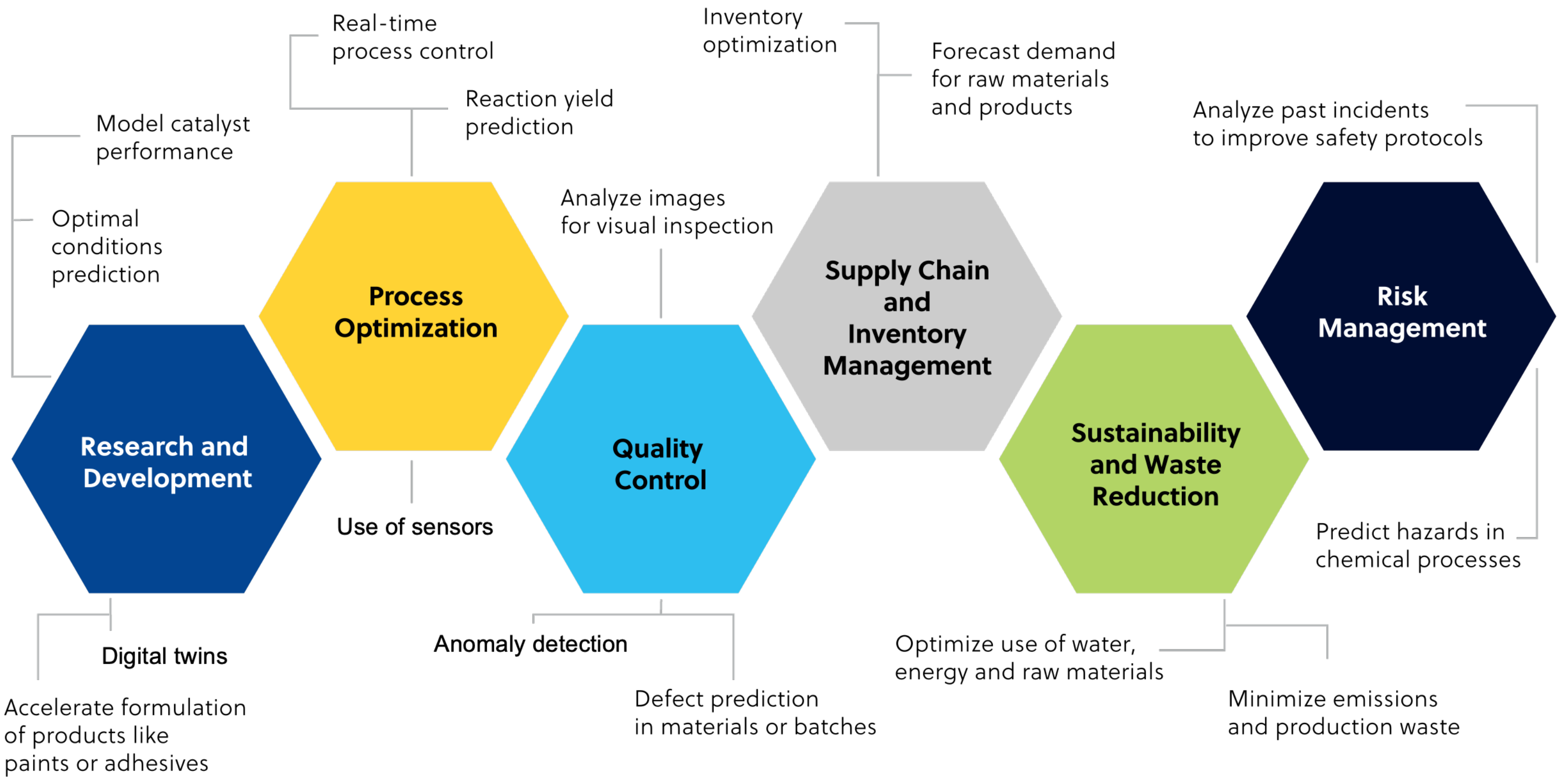


CAS
A division of the
American Chemical Society

# Publications featuring AI grows year over year



# Emerging applications touch every phase of innovation



Research and Development

Model catalyst performance

Optimal conditions prediction

Digital twins

Accelerate formulation of products like paints or adhesives

Process Optimization

Real-time process control

Reaction yield prediction

Use of sensors

Quality Control

Analyze images for visual inspection

Anomaly detection

Defect prediction in materials or batches

Supply Chain and Inventory Management

Inventory optimization

Forecast demand for raw materials and products

Sustainability and Waste Reduction

Optimize use of water, energy and raw materials

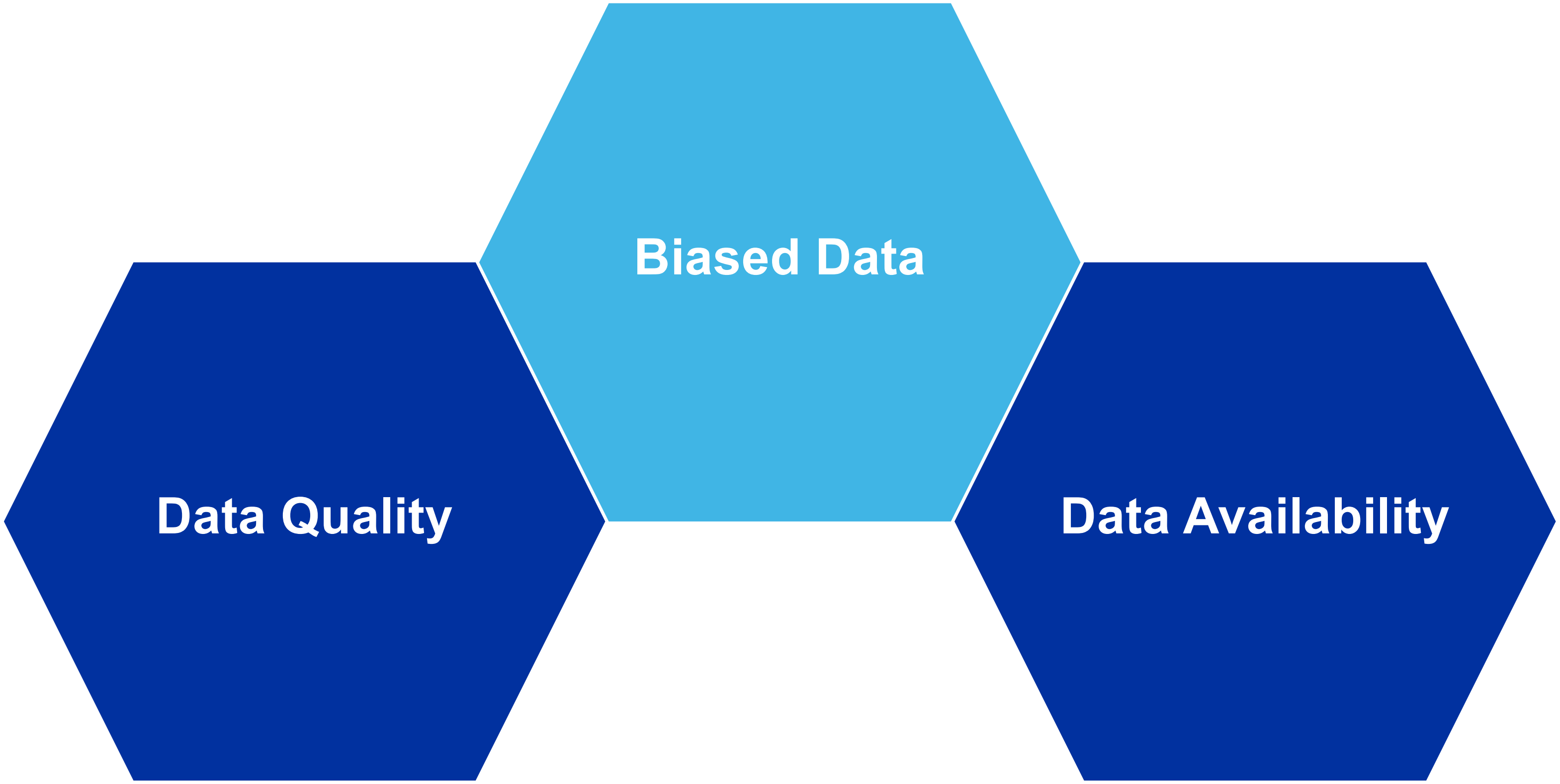
Minimize emissions and production waste

Risk Management

Analyze past incidents to improve safety protocols

Predict hazards in chemical processes

# What challenges exist in scientific AI?



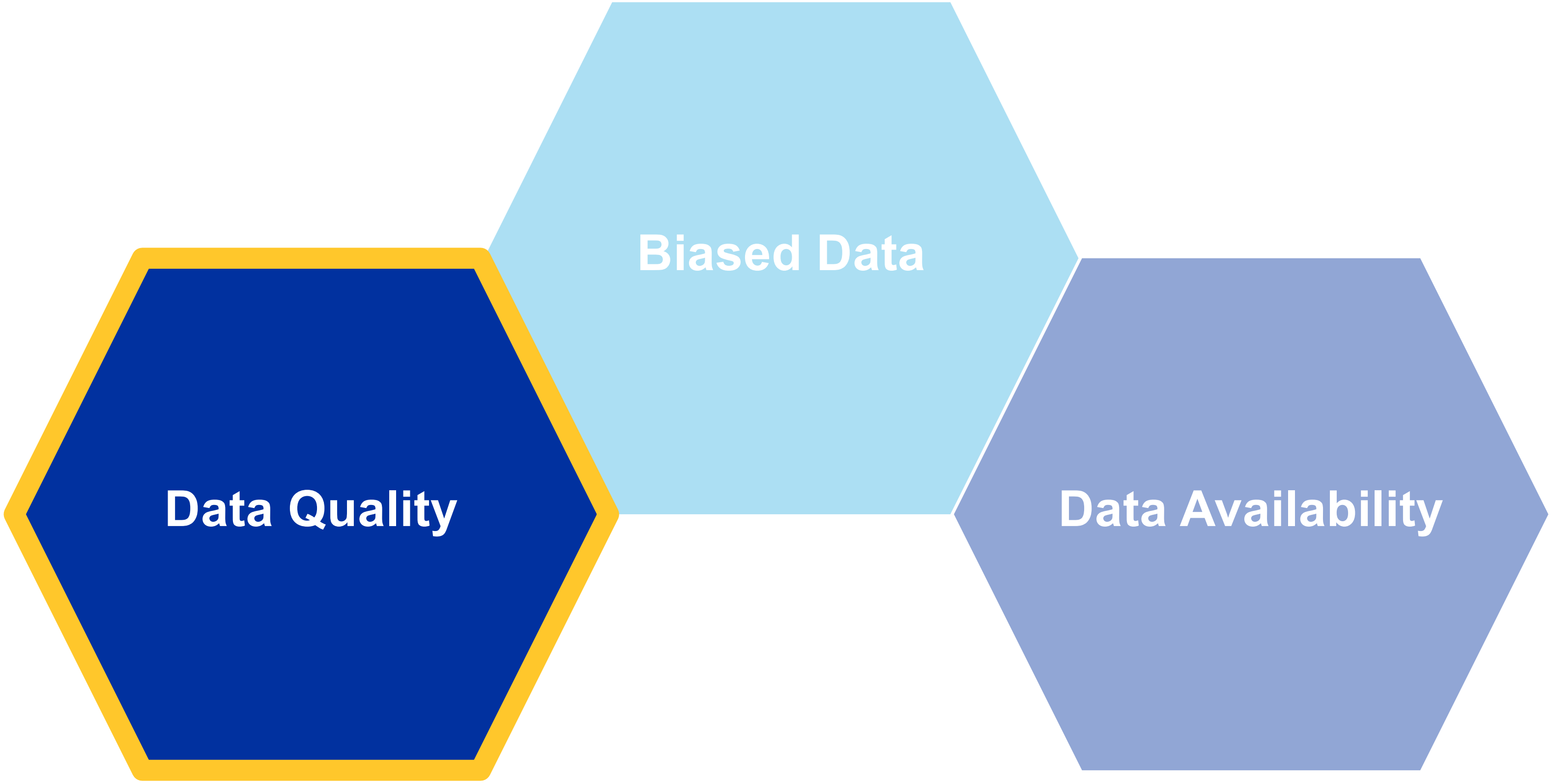
Biased Data

Data Quality

Data Availability

CAS
A division of the
American Chemical Society

# What Challenges exist in scientific AI?



Biased Data

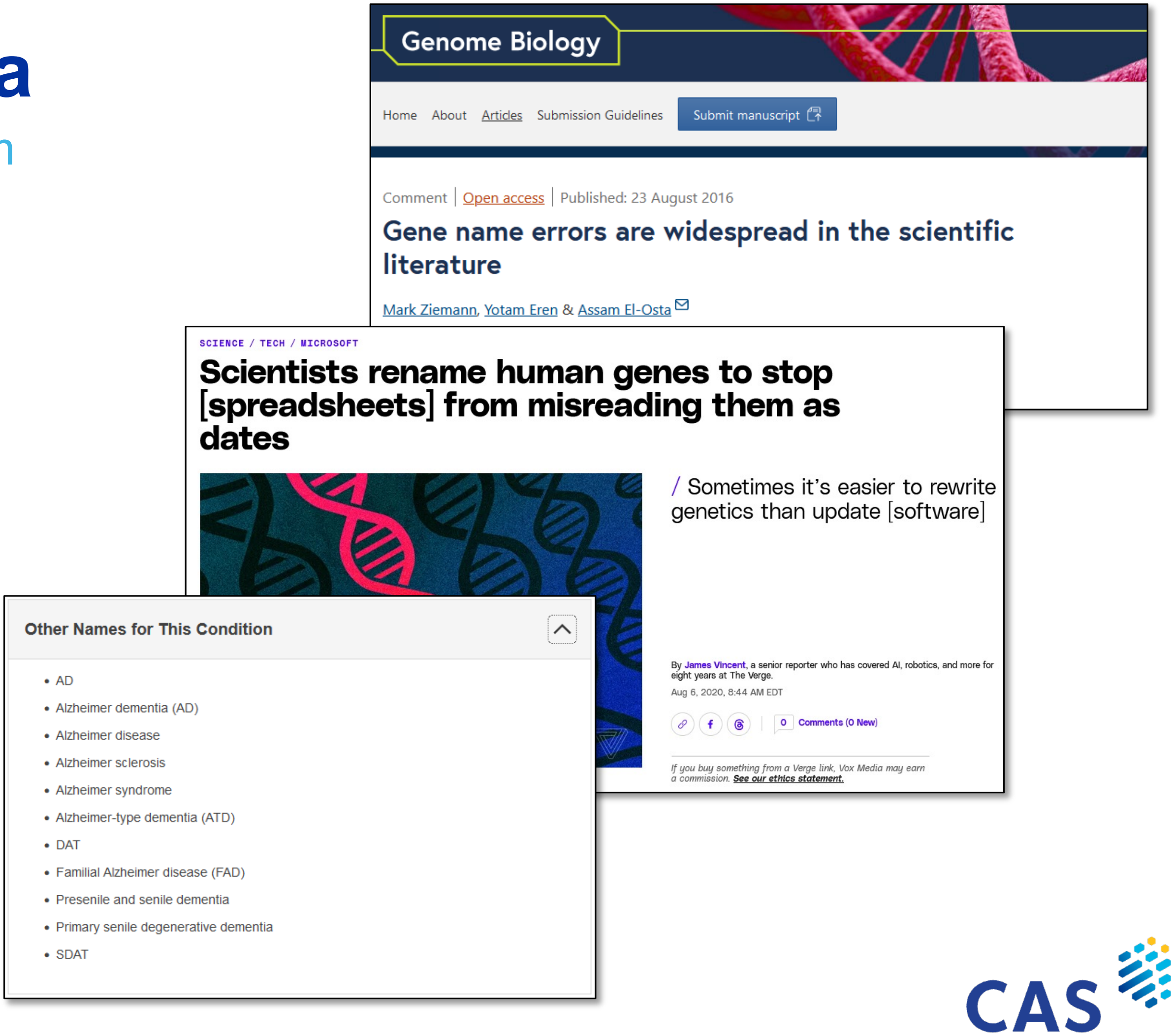
Data Quality

Data Availability

# Inconsistency in data

## Leading to problems downstream

- Many names for the same entity
- Different labs prefer different names
- Author errors
- Technology errors
- Different databases and IDs



Comment | Open access | Published: 23 August 2016

**Gene name errors are widespread in the scientific literature**

Mark Ziemann, Yotam Eren & Assam El-Osta

**Scientists rename human genes to stop [spreadsheets] from misreading them as dates**

/ Sometimes it's easier to rewrite genetics than update [software]

By James Vincent, a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Aug 6, 2020, 8:44 AM EDT

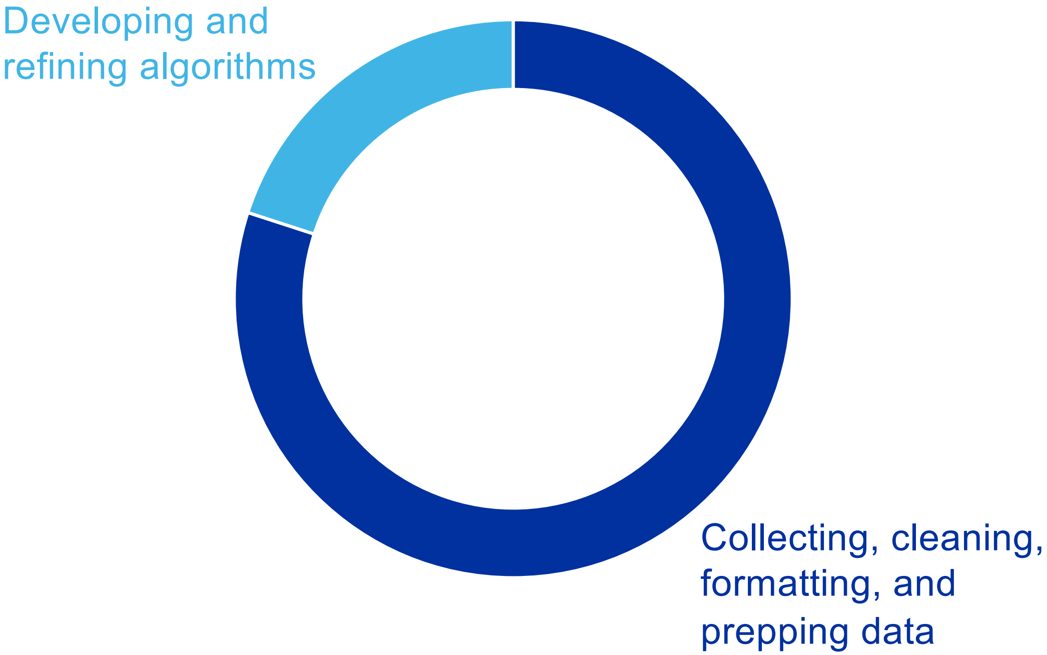
*If you buy something from a Verge link, Vox Media may earn a commission.* ***See our ethics statement.***

**Other Names for This Condition**

- AD
- Alzheimer dementia (AD)
- Alzheimer disease
- Alzheimer sclerosis
- Alzheimer syndrome
- Alzheimer-type dementia (ATD)
- DAT
- Familial Alzheimer disease (FAD)
- Presenile and senile dementia
- Primary senile degenerative dementia
- SDAT

# The 80:20 data science rule

## Messy data consumes time



**90%**

**Data generated by humans created in the last 2 years (est.)**

*Stobierski T. Harvard Business School Online. 2021*

# Harmonization creates consistency

From chaos to order

- Corrects inconsistencies and errors
- Combines different types of data
- Connects across data sources

# Good harmonization is human driven

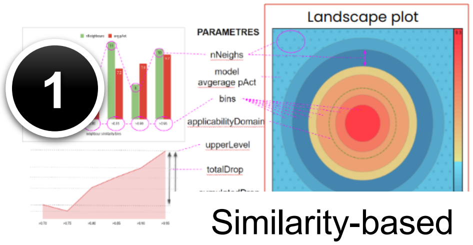
## Human effort prior to AI efficiency

- Human input ensures alignment to authority constructs
- Human assisted curation creates training data at the standards set by data scientists
- CAS employs a global team of hundreds of scientists and technologists to accomplish this

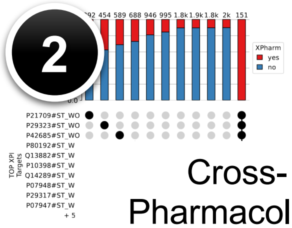
CAS
A division of the
American Chemical Society

# Drug-target activity prediction

## Estimating the pActivity of ligands towards protein targets



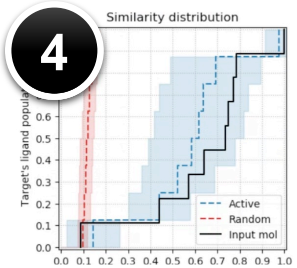
Similarity-based method (**SIM**)



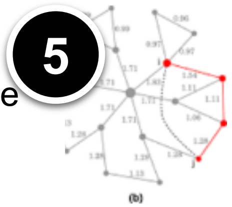
Cross-Pharmacology Index (**XPI**)

3
Training set
Features combinations

Machine Learning Method (**MLM**)

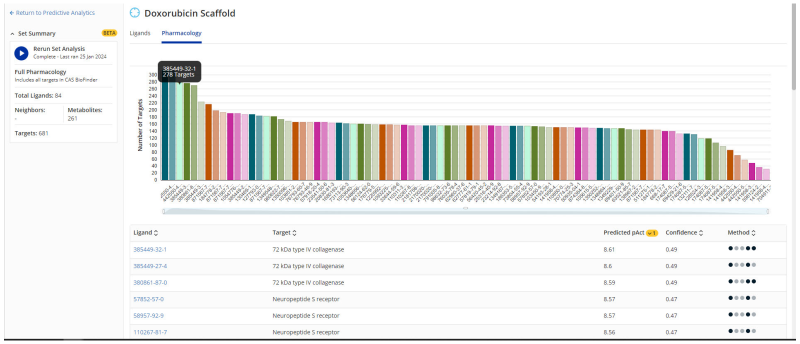


Similarity ensemble approach (**SEA**)



Simplest Active Substructure (**SAS**)

Estimated pActivity for a given ligand-target pair obtained by aggregating the outputs of an ensemble of individual models into a **single consensus value**



← Return to Predictive Analytics

Doxorubicin Scaffold

Set Summary BETA

Rerun Set Analysis
Complete - Last ran 25 Jan 2024

Full Pharmacology
Includes all targets in CAS BioFinder

Total Ligands: 84

Neighbors: -

Metabolites: 261

Targets: 681

| Ligand | Target | Predicted pAct | Confidence | Method |
|---|---|---|---|---|
| 385449-32-1 | 72 kDa type IV collagenase | 8.61 | 0.49 | ●●●●● |
| 385449-27-4 | 72 kDa type IV collagenase | 8.6 | 0.49 | ●●●●● |
| 380861-87-0 | 72 kDa type IV collagenase | 8.59 | 0.49 | ●●●●● |
| 57852-57-0 | Neuropeptide S receptor | 8.57 | 0.47 | ●●●●● |
| 58957-92-9 | Neuropeptide S receptor | 8.57 | 0.47 | ●●●●● |
| 110267-81-7 | Neuropeptide S receptor | 8.56 | 0.47 | ●●●●● |

CAS
A division of the American Chemical Society

# The difference harmonized data makes

## Higher quality data lead to higher quality results

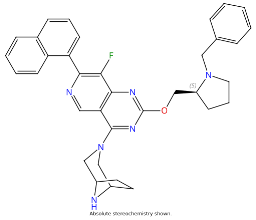
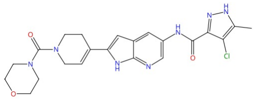
- CAS has an extensive set of harmonized bioactivity data
- This harmonized data was used to retrain existing models

**56%**

Reduction of difference in Predicted vs. Experimental

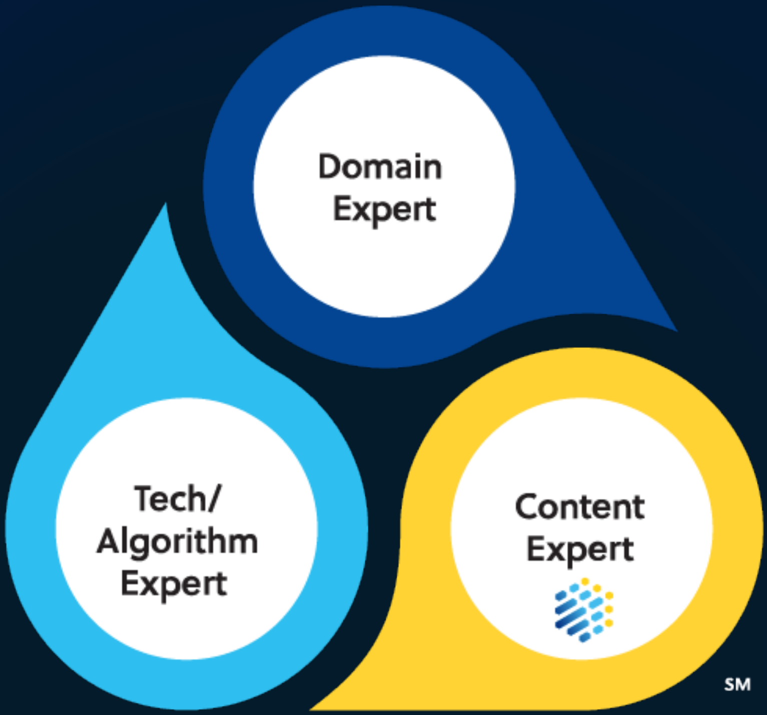
**23%**

Reduction of Std. Dev. of Predicted vs. Experimental

| Ligand | Target | Experimental pActivity | Original Model | Retrained Model |
|---|---|---|---|---|
|  | KRas | 4.2 | 11.1 | 4.8 |
|  | KIT | 10.0 | 4.4 | 9.9 |

# Effective partnering for scientific AI solutions

## Requires the “Triangle for Success”



**Domain Expert**
Deep understanding of the science and workflows
– Brings expertise as well as real-world data

**Tech/Algorithm Expert**
Mathematical and computational capabilities
– Brings expertise as well as unique technologies

**Content Expert**
Expertise in data modeling, curation, harmonization
– Brings expertise, data, and information management tech

CAS
A division of the
American Chemical Society

# Thank you

Connect with us at cas.org

in linkedin.com/company/cas

𝕏 @CASchemistry

**Jacob Al-Saleem PhD**
JAl-Saleem@cas.org

CAS
A division of the
American Chemical Society