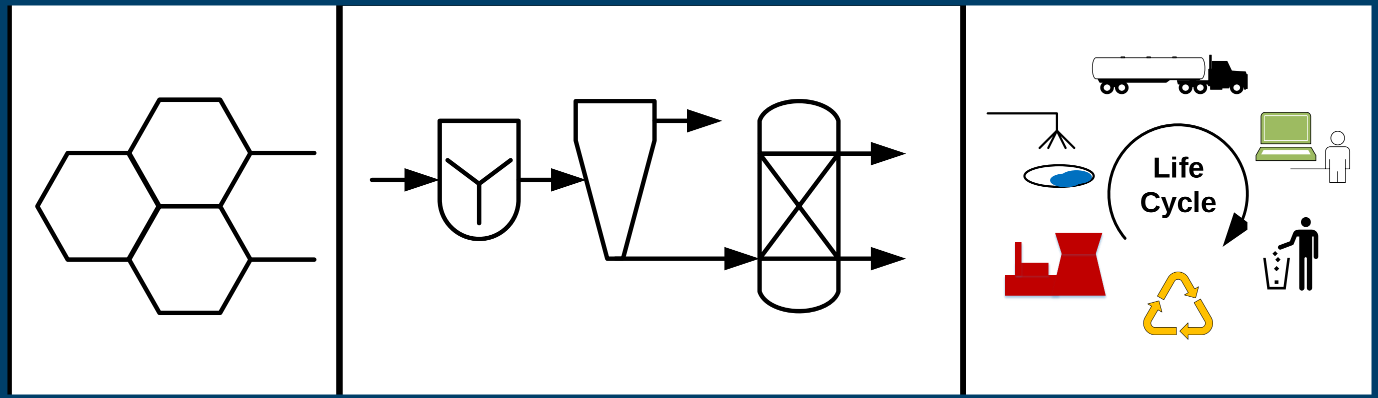
Applying Machine Learning to Estimate Releases from New Uses of Existing Chemicals

Raymond L. Smith, Jose D. Hernandez-Betancur, David E. Meyer, Gerardo J. Ruiz-Mercado, William M. Barrett, Michael A. Gonzalez, John P. Abraham



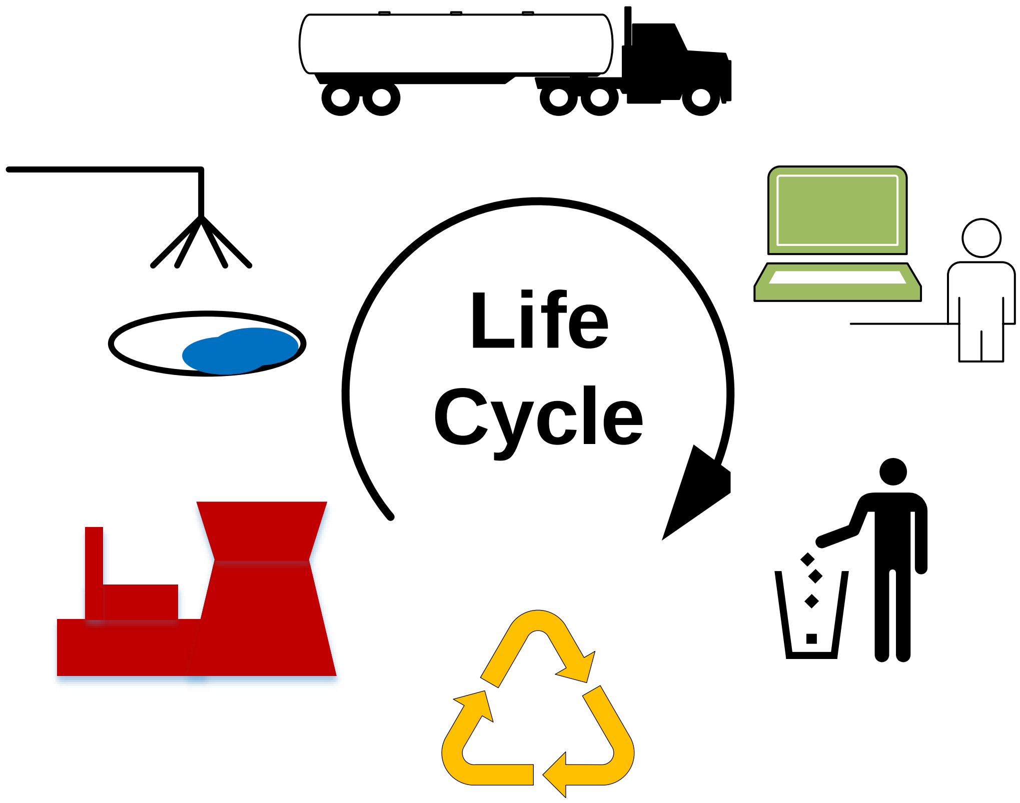
Office of Research and Development
Center for Environmental Solutions and Emergency Response
Chemical Safety for Sustainability

AIChE Annual Meeting
November 11, 2019

Disclaimer

The views expressed in this presentation are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

Life Cycle of a Chemical

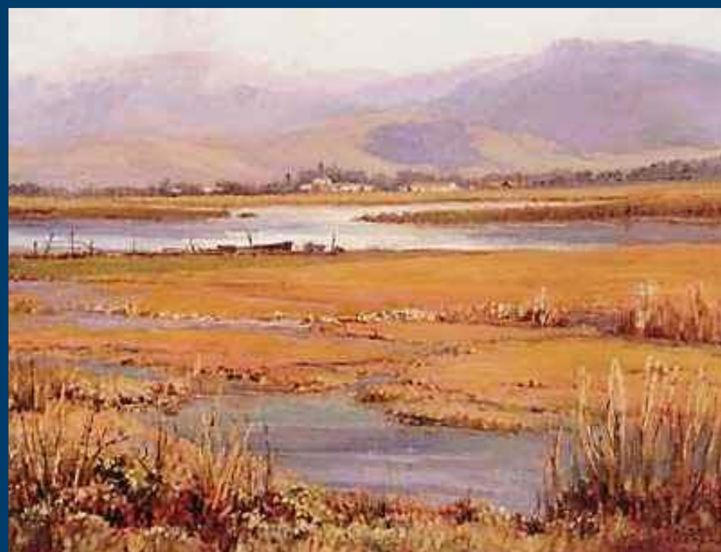


Motivation: Toxicity

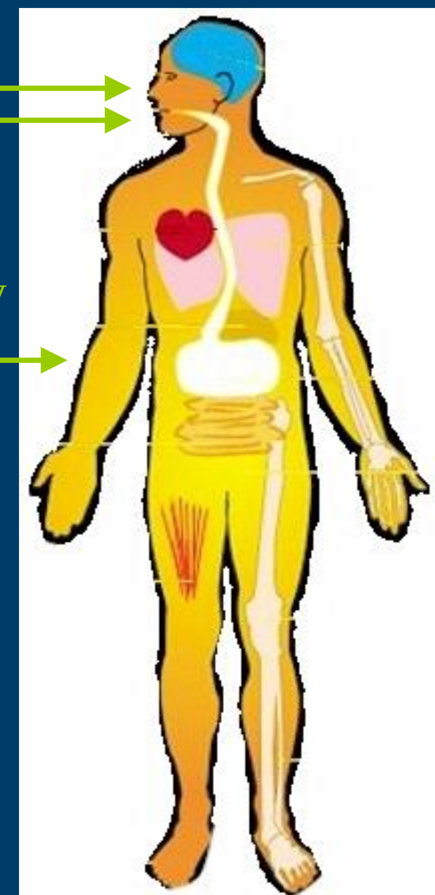
Air Emissions

Liquid Discharges

Solid Waste



Amount
Duration
Frequency



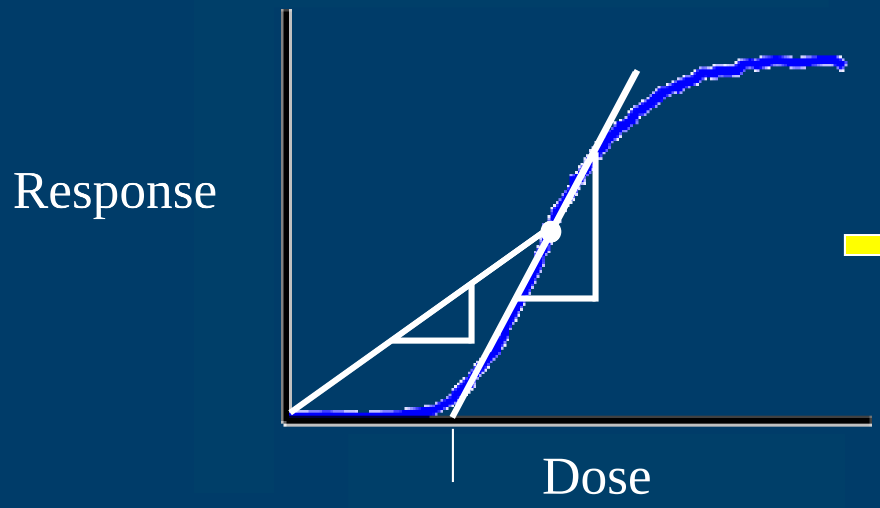
Known chemicals: ~100 million
TSCA inventory: ~85,000
Active chemicals (as of 6/19): 32,000+

Toxicity

Releases

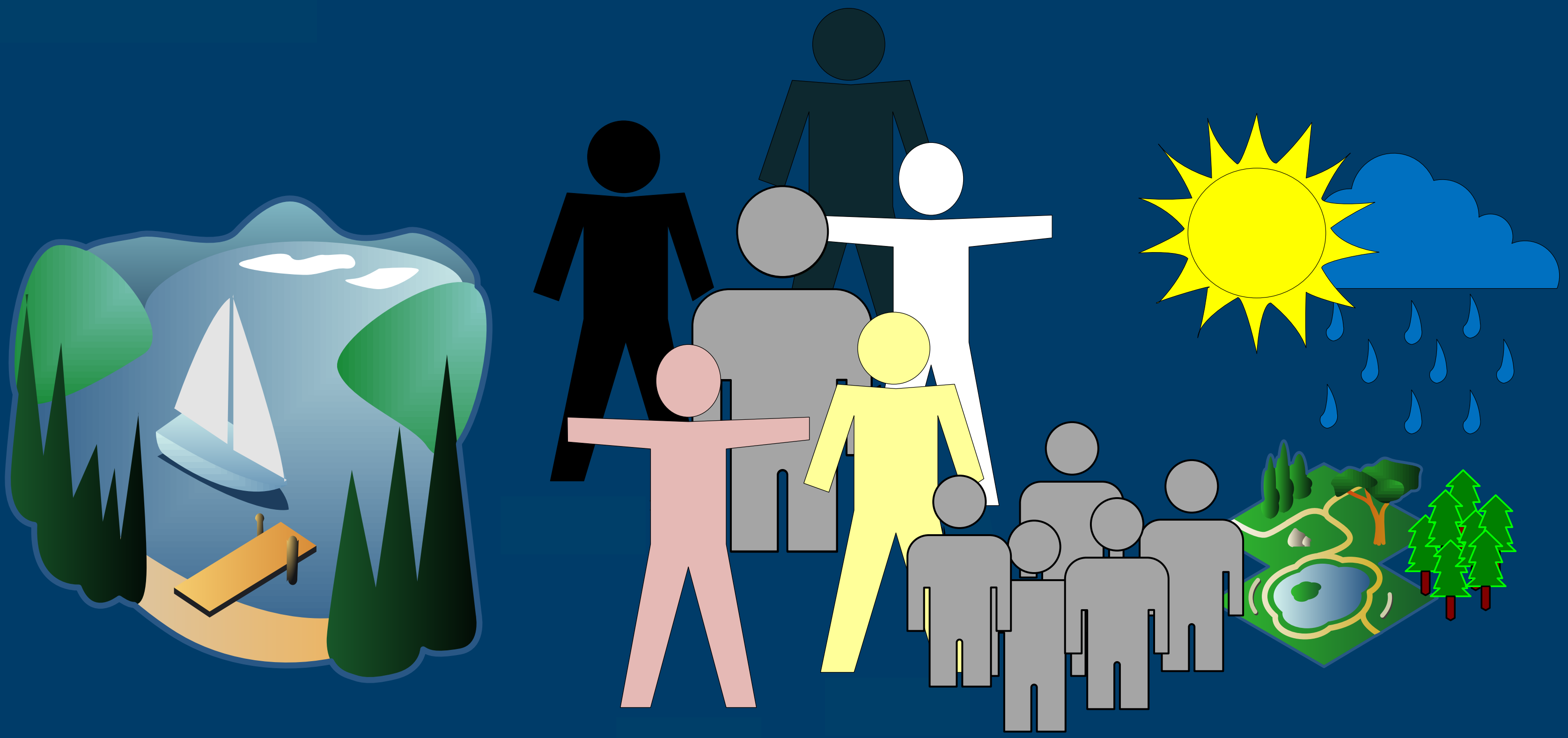
↳ Exposure Dose

↳ Internal (Organ/Tissue) Dose

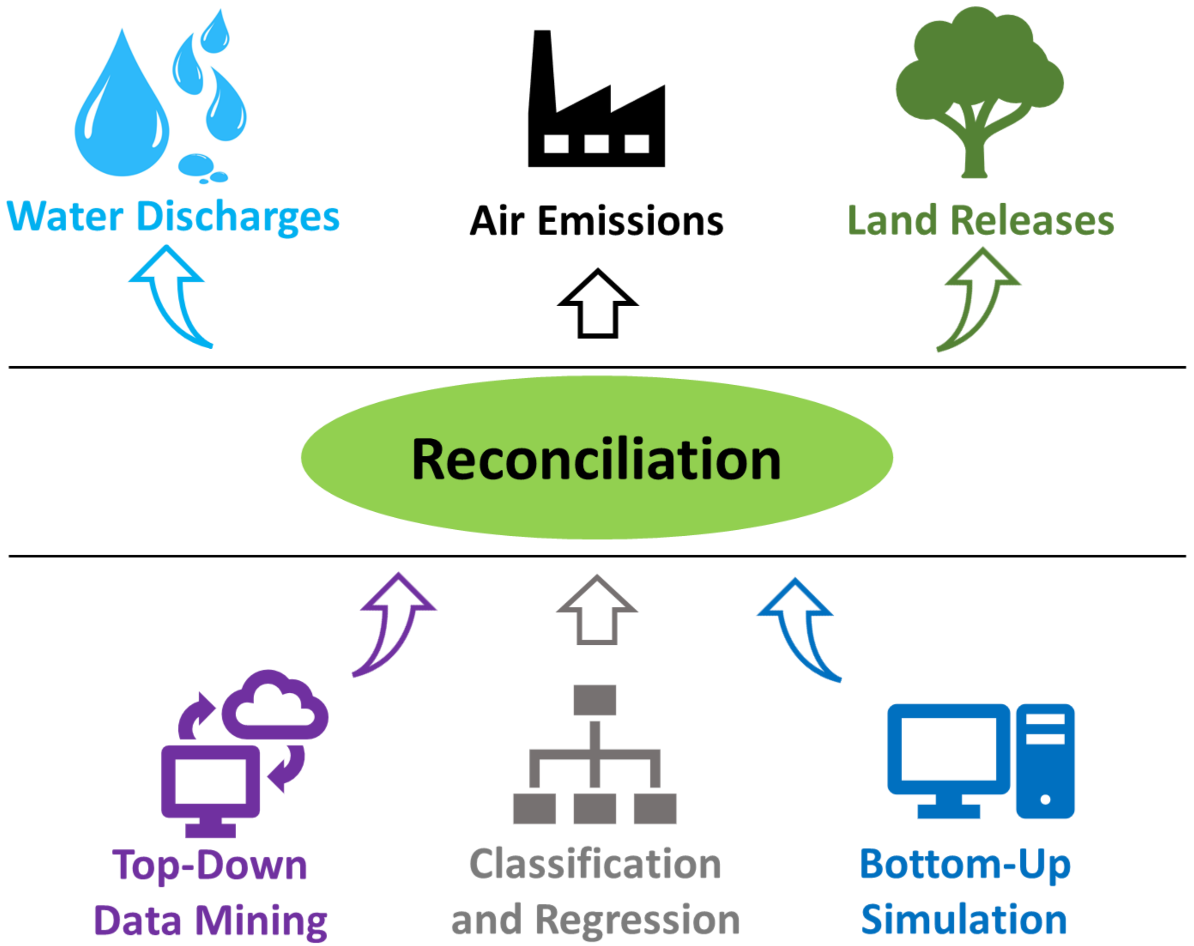


Adaptation,
Impaired Structure and/or Function,
Morbidity,
Mortality

Toxic Adverse Effects



Estimating Chemical Releases

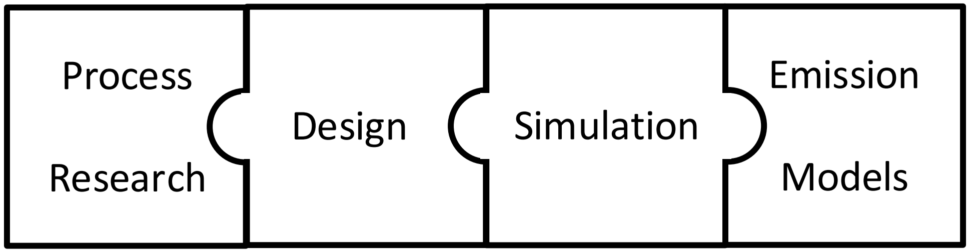


Rapid Estimation of Manufacturing Emissions

1. Existing inventory databases

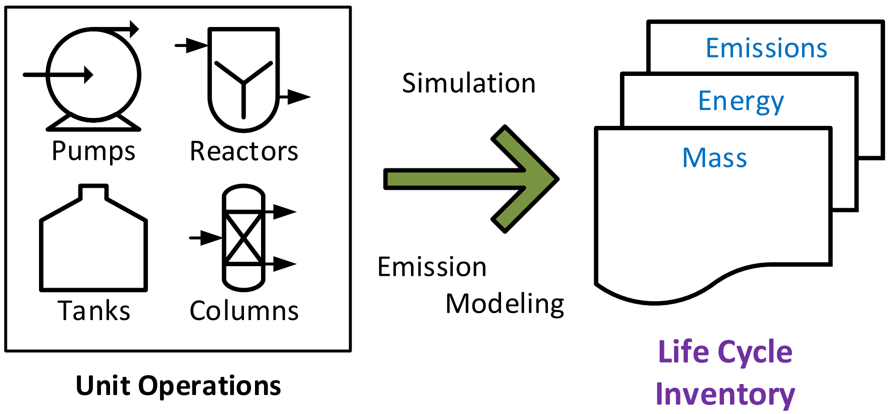
2. Top-down inventory data mining

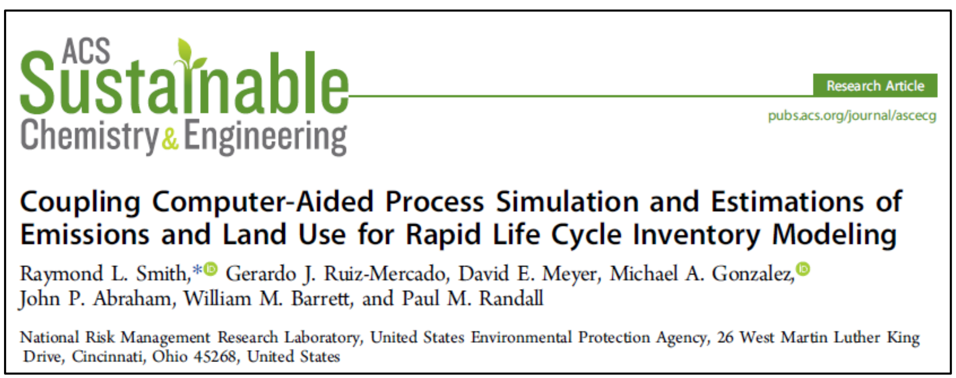
3. Bottom-up inventory development



Bottom-Up Simulation

Advantages: potential for improved Life Cycle Inventory; process specific; inputs naturally in results; storage, vent, and fugitive emissions included




ACS Sustainable Chemistry & Engineering

Research Article

pubs.acs.org/journal/ascecg

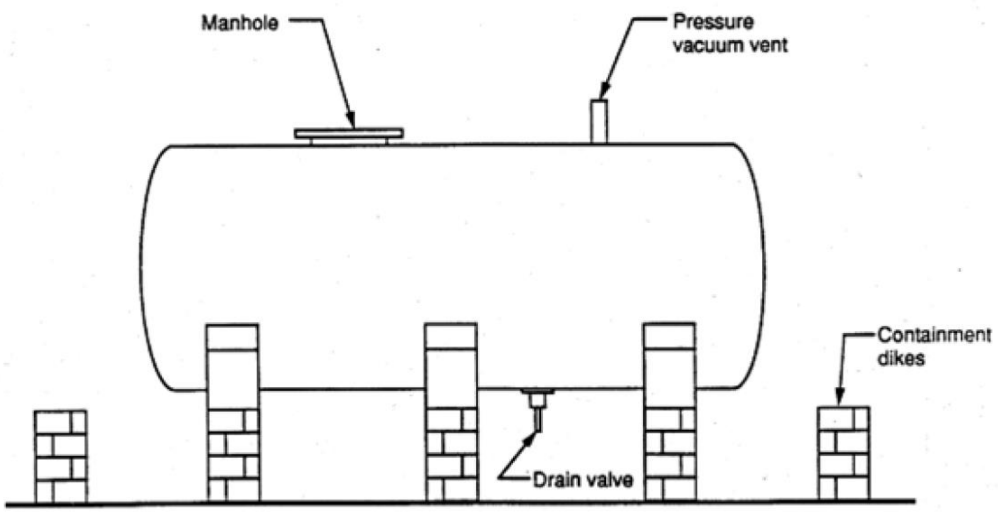
Coupling Computer-Aided Process Simulation and Estimations of Emissions and Land Use for Rapid Life Cycle Inventory Modeling

Raymond L. Smith,* Gerardo J. Ruiz-Mercado, David E. Meyer, Michael A. Gonzalez, John P. Abraham, William M. Barrett, and Paul M. Randall

National Risk Management Research Laboratory, United States Environmental Protection Agency, 26 West Martin Luther King Drive, Cincinnati, Ohio 45268, United States

Challenges: knowledge of engineering design; need for chemical synthesis details; uncontrolled emissions

Bottom-Up Simulation



Working Losses

$$L_W = \frac{\dot{V}}{22.4}\left(\frac{273.15}{T}\right)\left(\frac{P_i^{sat}}{760}\right)(MW)K_N K_P$$

Breathing Losses

$$L_B = 16.3 V_V \left(\frac{273.15}{T}\right)\left(\frac{P_i^{sat}}{760}\right)(MW)\left(\frac{T_R}{T}\right)$$

Process Vents

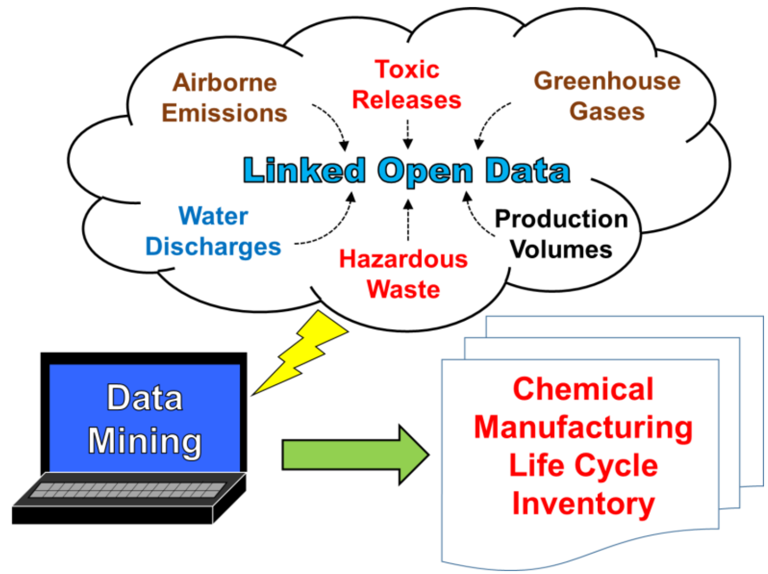
$$S_i = \frac{P_i^b}{x_i \gamma_i P_i^{sat}} = \frac{k_i A}{k_i A + F}$$

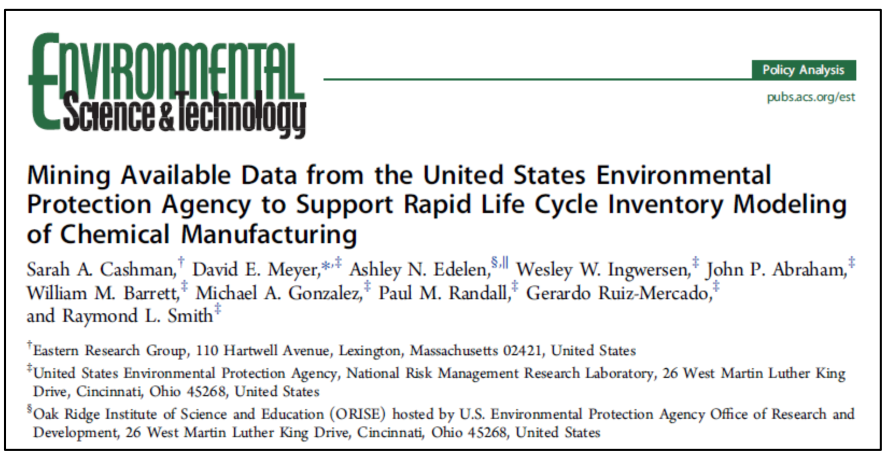
$$E_i = \frac{F x_i \gamma_i P_i^{sat}}{RT} S_i (MW_i)$$

Equipment Type	**Service**	**Emission Factor (kg/h/source)**
Pumps	Light liquid	0.0199
	Heavy liquid	0.00862
Compressors	Gas	0.228
Valves	Gas	0.00597
	Light liquid	0.00403
	Heavy liquid	0.00023
Connectors (e.g., flanges)	All	0.00183
Open-ended lines	All	0.0017
Sampling connections	All	0.0150
Pressure relief valves	Gas	0.104

Top-Down Data Mining

Advantages: primary data reported by industry and States; detailed release profiles; automation capabilities (linked open data)




Environmental Science & Technology

Policy Analysis

pubs.acs.org/est

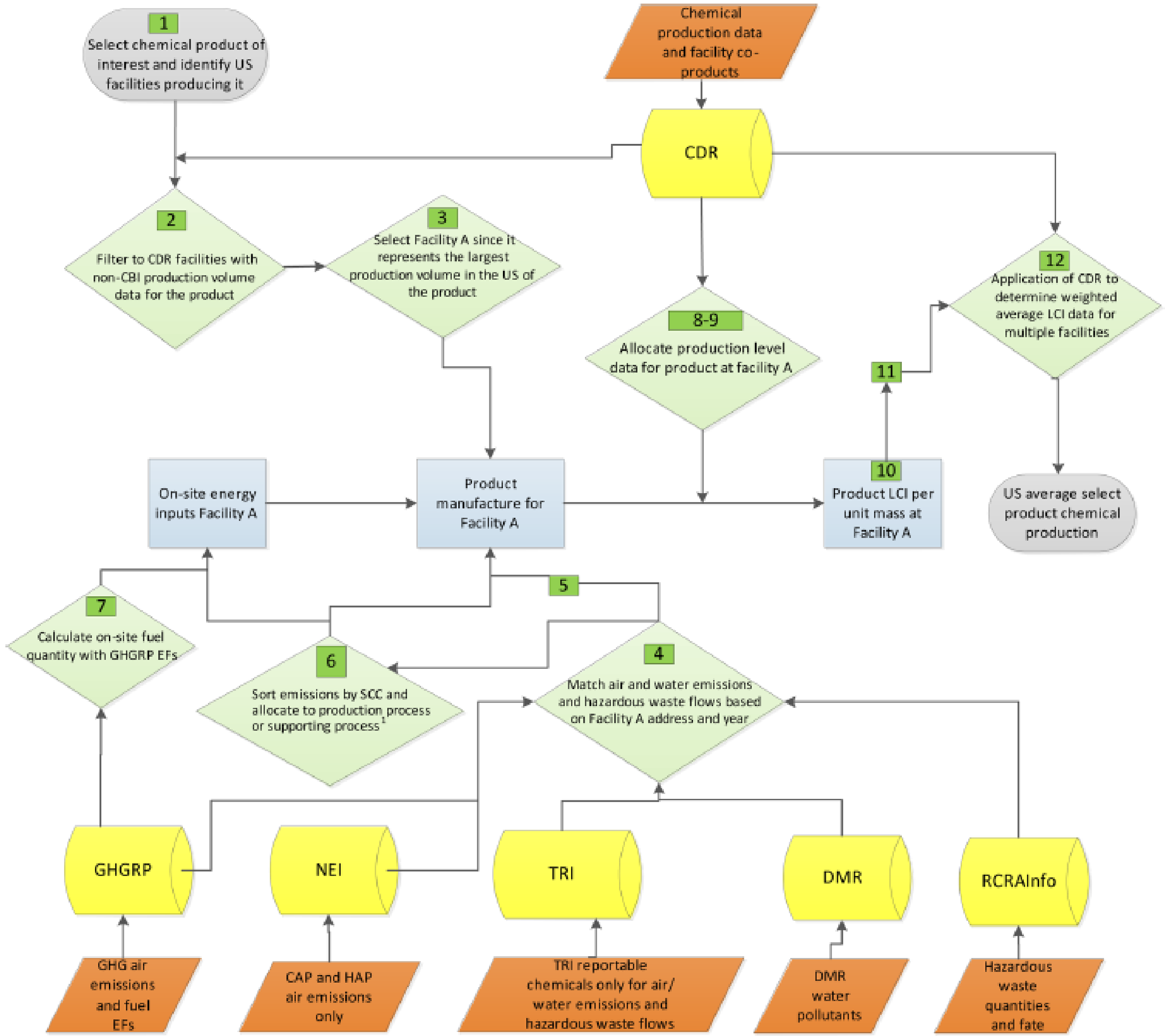
Mining Available Data from the United States Environmental Protection Agency to Support Rapid Life Cycle Inventory Modeling of Chemical Manufacturing

Sarah A. Cashman,[†] David E. Meyer,[*,‡] Ashley N. Edelen,[§,∥] Wesley W. Ingwersen,[‡] John P. Abraham,[‡] William M. Barrett,[‡] Michael A. Gonzalez,[‡] Paul M. Randall,[‡] Gerardo Ruiz-Mercado,[‡] and Raymond L. Smith[‡]

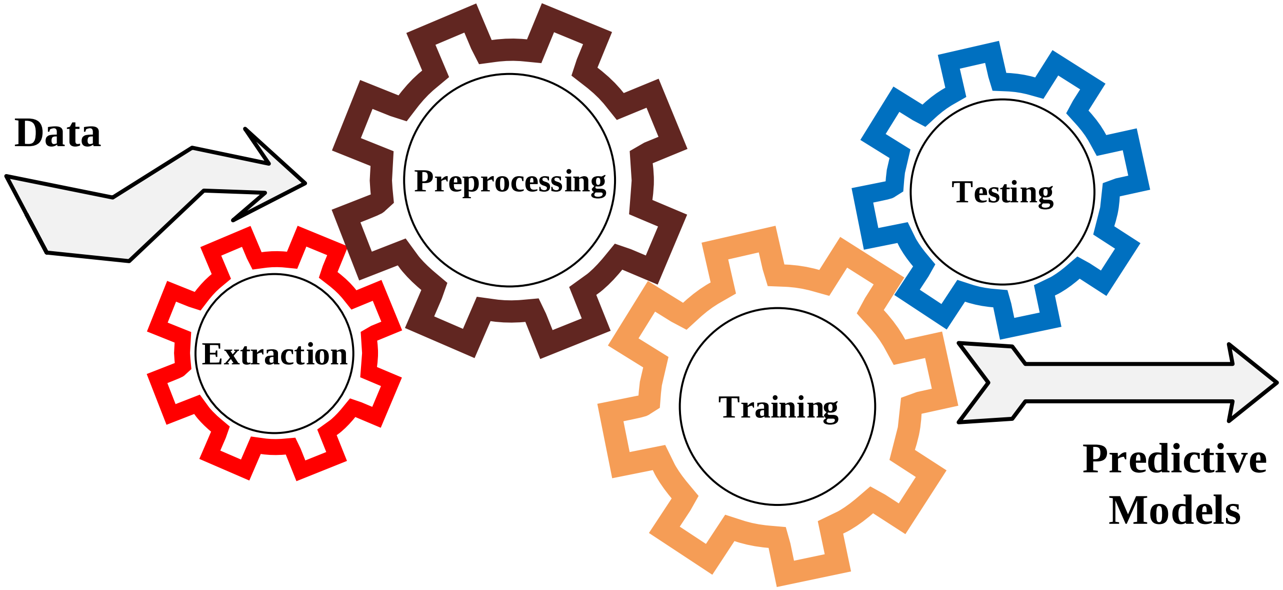
[†]Eastern Research Group, 110 Hartwell Avenue, Lexington, Massachusetts 02421, United States
[‡]United States Environmental Protection Agency, National Risk Management Research Laboratory, 26 West Martin Luther King Drive, Cincinnati, Ohio 45268, United States
[§]Oak Ridge Institute of Science and Education (ORISE) hosted by U.S. Environmental Protection Agency Office of Research and Development, 26 West Martin Luther King Drive, Cincinnati, Ohio 45268, United States

Challenges: multi-chemical facility-level allocation; input data gaps; currently limited to TSCA Chemical Data Reporting chemicals

Top-Down Data Mining



Machine Learning



Data Mining Literature

- Search Terms
- Searches
- Compilation of Search Results
- Applying Filters
- Scraping Desired Data

Acrylamide Case Study

Collection of Chemical Input Parameters

SMILES Structure: O=C(N)C=C

Molecular Weight	71.079 g/mol
XLogP3	-0.7
Hydrogen Bond Donor Count	1
Hydrogen Bond Acceptor Count	1
Rotatable Bond Count	1
Topological Polar Surface Area	43.1 A^2
Heavy Atom Count	5

19,048 rows of literature results about Acrylamide

Filtering Results

‘health’ and ‘exposure’

209 results

and ‘coffee’

18 results (**7 articles**)

‘release’ and ‘model’

64 results

and ‘exposure’

2 results

‘exposure’ and ‘cancer’ and ‘industr’

28 results

and

not ‘food’

3 results (**2 current articles**)

Tables as Results

United States Environmental Protection Agency

Table 1
Acrylamide levels (µg/kg) of foodstuffs monitored from 2007 to 2009 reported by EFSA. Adapted from (EFSA, 2011).

	2007				2008			
	N^b	Mean[a]	SD[c]	Max.	N^b	Mean[a]	SD[c]	Max.
Biscuits crackers	66	284	315	1526	131	204	178	1042
Biscuits infant	97	204	352	2300	88	110	147	1200
Biscuits not specified	291	303	433	4200	260	209	247	1940
Wafers	38	210	256	1378	48	252	416	2353
Bread crisp	153	228	328	2430	90	235	273	1538
Bread soft	123	70	116	910	191	49	56	528
Bread non specified	54	190	424	2565	17	23	19	86
Coffee instant	51	357	327	1047	58	502	285	1373
Coffee non specified	41	261	268	1158	10	241	215	720
Coffee roasted	151	253	203	958	253	208	182	1524
Gingerbread	357	425	494	3615	246	437	545	3307
Muesli and porridge	47	215	183	805	18	43	27	112
Other products not specified	378	271	355	2529	445	198	309	2592
Substitute coffee	59	800	1062	4700	73	1124	1138	7095
Breakfast cereals	132	152	184	1600	120	170	247	2072
Cereal-based baby food	92	69	72	353	96	45	81	660
Jarred baby food	87	44	35	162	128	35	39	297
Home cooked potato products deep fried	54	354	413	1661	39	228	253	1220
Home cooked potato products not specified	82	277	392	2175	100	192	402	3025
Home cooked potato products oven fried	8	385	342	941	94	235	268	1439
French fries	647	357	382	2668	521	280	279	2466
Potato crisps	273	565	259	4180	435	616	634	4382

[a] Values based on an upper bond scenario (values below LOD and values between LOD and LOQ were set to the LOD or the LOQ value, resp
[b] Number of individual samples analyzed for each food category.
[c] Standard deviation of the upper bond scenario. Standard deviation not available for the 2009 data.

Table 1
Experimental limits for AA toxicity related to human exposure

Effect	Experimental limit (µg/kg bw/day)	Margin of exposure		RfD or TDI (µg/kg bw/day)	Uncertainty factor
		Average intake[a]	High intake[b]		
Neurotoxicity	200 (or 500) (NOAEL)	200	50	0.67	300
Toxicity to reproduction and development	2000 (NOAEL)	2000	500	20	100
Carcinogenesis	300 (BMDL)	300	75	1	300
Carcinogenesis	440 (POD as LED_{10})	440	110	1.4	300

[a] Average intake: 1 µg/kg/day.
[b] High intake: 4 µg/kg/day.

Table 3
Human exposure to acrylamide from caffeinated beverages. Dietary intake as well as the Margins of Exposure for neurotoxic risk assessment (MOE_N) and carcinogenic risk assessment (MOE_C) are reported. Data are expressed as µg/kg-bw/day for dietary intake. MOE_N values are reported for $BMDL_{10}$ (0.2 mg/kg-bw/day). MOE_C values are reported for $BMDL_{10}$ (0.31 and 0.18 mg/kg-bw/day).

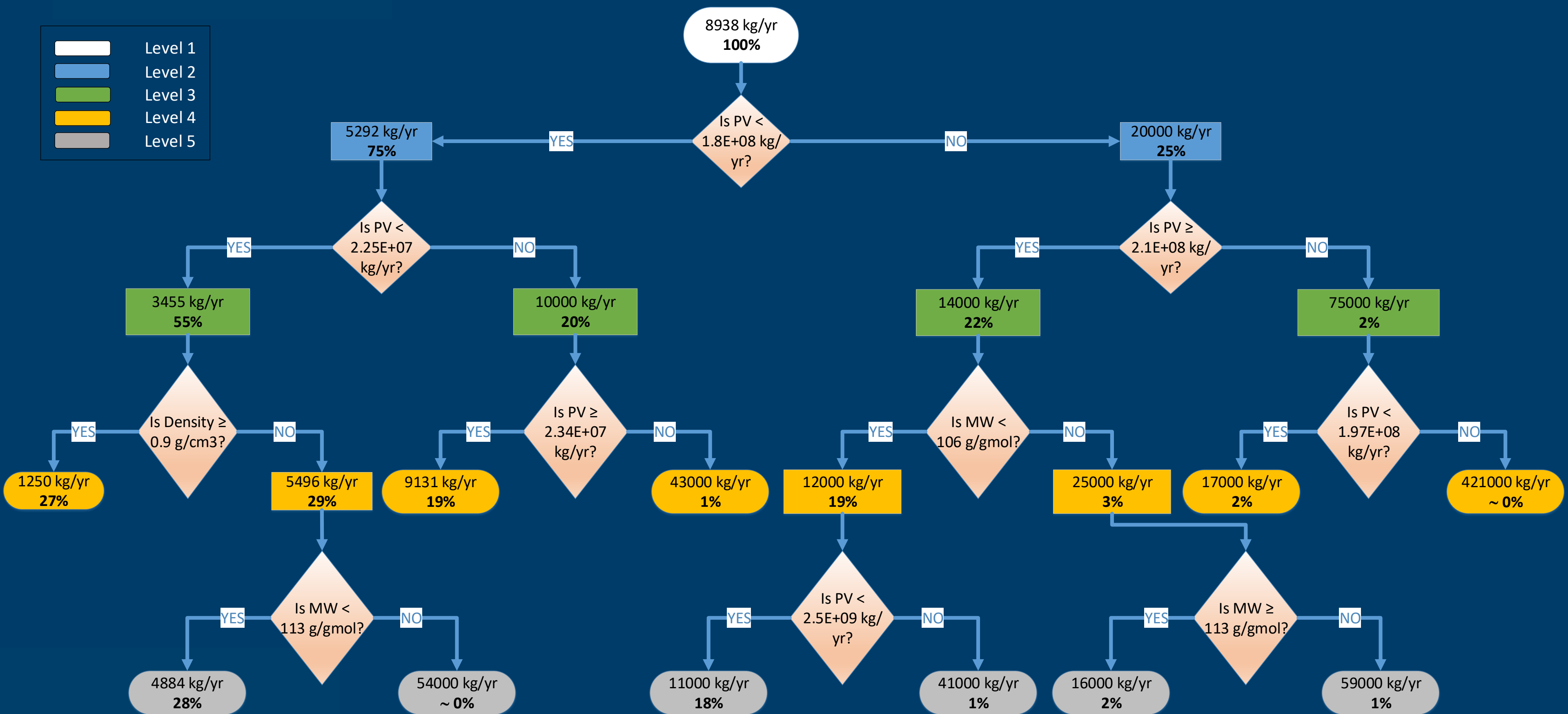
Mean Acrylamide Intake; µg/kg-bw/day				
Age (years)	Lebanese	American	Chocolate	Espresso
Population (3–75)				
Dietary Intake	10.9 ± 6.5	0.37 ± 0.24	1.2 ± 0.8	7.4 ± 4.4
MOE_N	18	535	163	27
[a]MOE_C	28(17)	829(481)	252(139)	42(24)
Children/Teens (3–18)				
Dietary Intake	8.5 ± 5.5	0.26 ± 0.15	1.3 ± 1.0	6.1 ± 3.7
MOE_N	24	775	148	33
[a]MOE_C	37(21)	1202(698)	230(134)	51(30)
Young Adults (18–30)				
Dietary Intake	9.6 ± 6.9	0.3 ± 0.2	1.2 ± 0.7	6.3 ± 4.1
MOE_N	21	633	172	32
[a]MOE_C	32(19)	981(570)	266(155)	49(29)
Adults (31–40)				
Dietary Intake	11.1 ± 5.9	0.37 ± 0.23	1.3 ± 1.0	9.5 ± 4.6
		548	151	21
	16)	849(493)	234(136)	33(19)
	5 ± 7.2	0.40 ± 0.25	1.1 ± 0.6	7.5 ± 3.8
		474	182	27
	14)	735(427)	282(164)	42(24)
	9 ± 5.9	0.45 ± 0.35	1.3 ± 1.1	7.7 ± 4.7
		441	144	25
	17)	683(396)	223(129)	39(22)

spond to $BMDL_{10}$ 0.31 (0.18 mg/kg-bw/day).

Tentative Process to Obtain Results

- Developing input data is time intensive
- While efforts continue, use input parameters from EPA's ChemSTEER program:
    - Density
    - Molecular Weight
    - Production Volume
    - Solubility
    - Vapor Pressure

Regression Tree Model

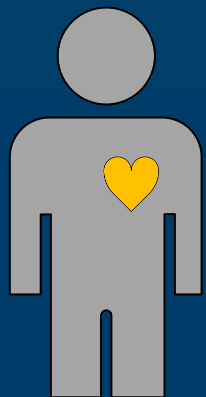


Meyer, D.E. et al. "Purpose-Driven Reconciliation of Approaches to Estimate Chemical Releases," *ACS Sustainable Chemistry & Engineering*, 7, 1260-1270 (2019).

Case Study: Cumene Emissions

Approach	Emission Factor (kg/kg)
Top-Down Data Mining	$2.0x10^{-5}$
Bottom-Up Simulation	$1.3x10^{-4}$
Regression Tree	$9.3x10^{-5}$
Random Forest	$2.0x10^{-4}$

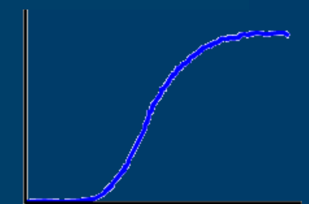
Summary

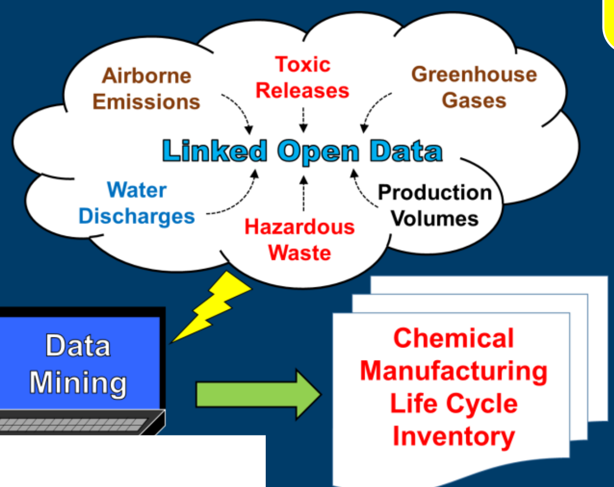


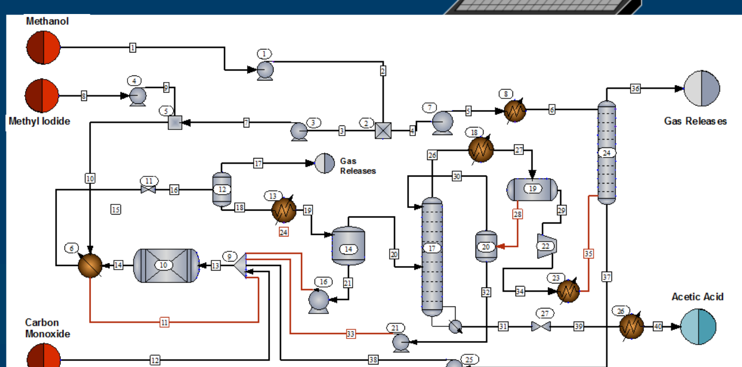
Releases

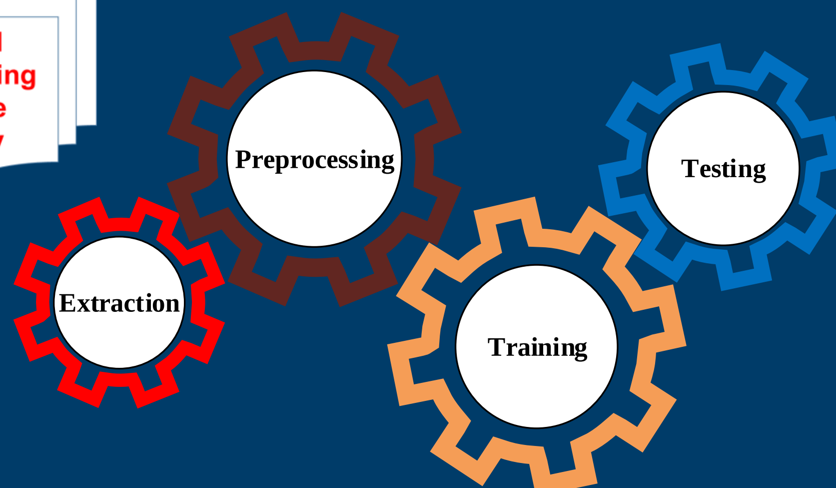
Exposure Dose

Internal (Organ/Tissue) Dose









Contact Info

smith.raymond@epa.gov

References

S.A. Cashman et al. (2016). "Mining Available Data from the United States Environmental Protection Agency to Support Rapid Life Cycle Inventory Modeling of Chemical Manufacturing," *Environmental Science & Technology*, 50(17), 9013-9025.

R.L. Smith et al. (2017). "Coupling Computer-Aided Process Simulation and Estimations of Emissions and Land Use for Rapid Life Cycle Inventory Modeling," *ACS Sustainable Chem. Eng.* 5, 3786-3794.

D.E. Meyer et al. (2019). "Purpose-Driven Reconciliation of Approaches to Estimate Chemical Releases," *ACS Sustainable Chem. Eng.*, 7, 1260-1270.