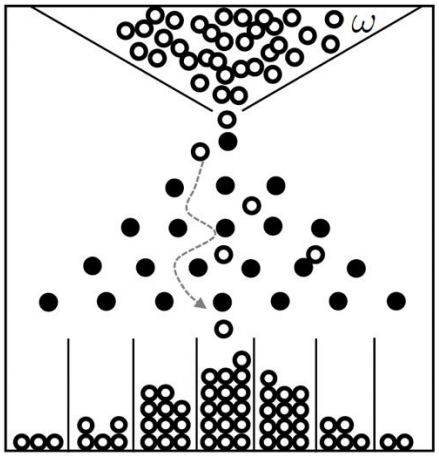
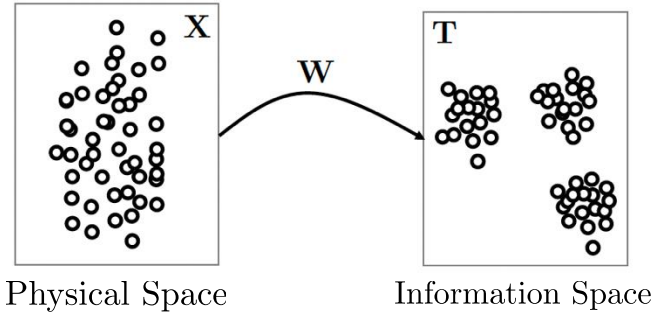


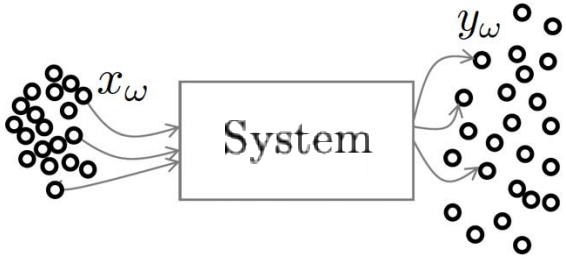
Experiences in Teaching Statistics & Data Science to ChemE Students at UW-Madison

Victor M. Zavala

Department of Chemical & Biological Engineering
University of Wisconsin-Madison

Mathematics and Computer Science Division
Argonne National Laboratory





Statistics for ChemEs: Why?

Statistics for ChemEs: Why?

Alumni surveys at UW-Madison indicated that **training in statistics** and related topics (e.g., data science) is **insufficient**.

ChemE undergrads used to take a **statistics for engineers course (Stat 324).** This is taught by the stats department.

ChemE undergrads often report that they **do not see connections** between Stat 324 and ChemE curriculum.

Stat 324 – Introductory Applied Statistics for Engineers

Contents

Descriptive statistics.

Probability, distributions, random variables.

Hypothesis tests, confidence intervals.

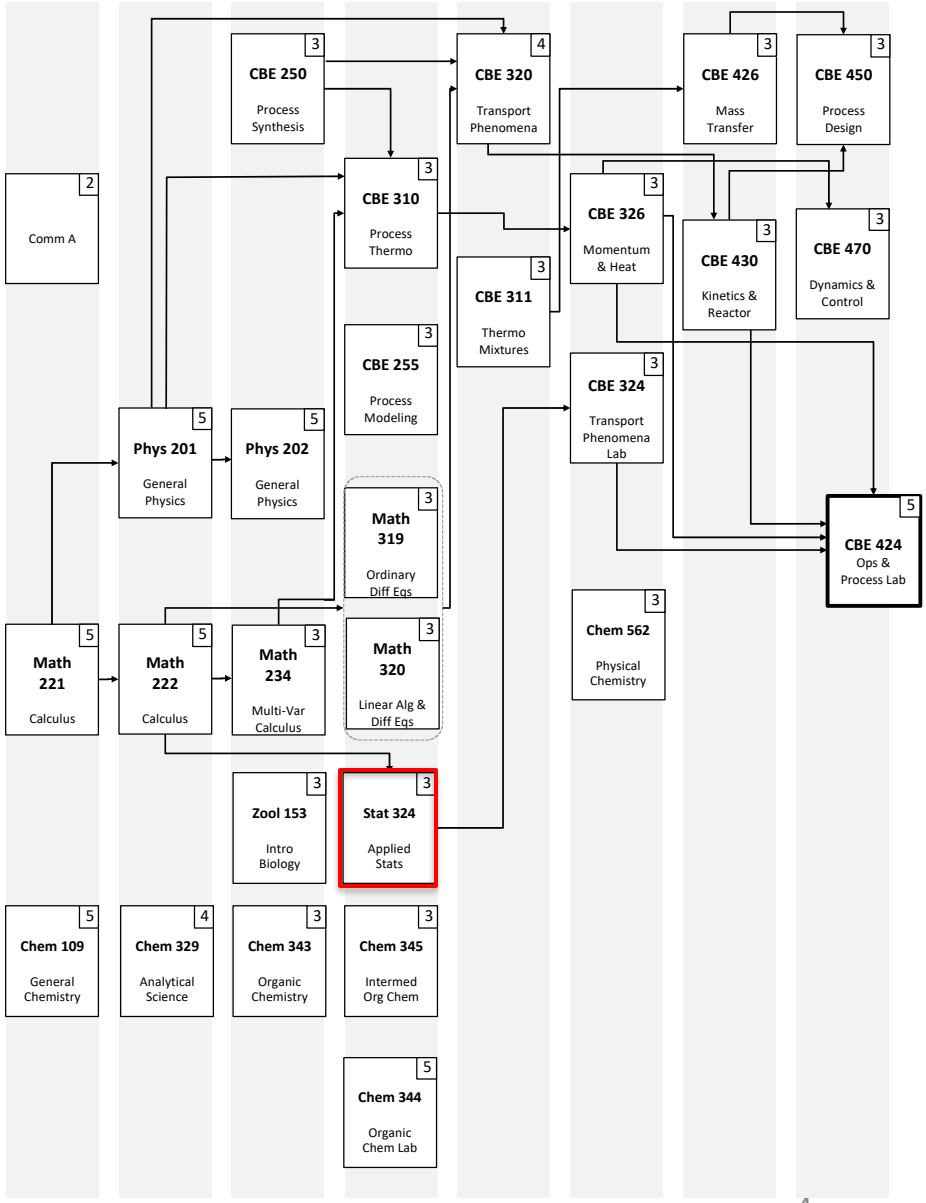
Linear regression, model checking, inference.

Analysis of variance, experimental design.

Observations (Personal):

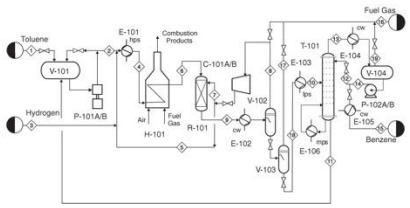
Concepts are essential but **too basic** and not particularly useful **without proper context**.

This type of course is **common** in ChemE programs around the world.



Statistics for ChemEs: How?

Stats for ChemEs: How?



ChemEs care about **physics & phenomena** (e.g., thermo, transport, kinetics).

ChemEs care about **decisions** (e.g., design, control, economics, sustainability, risk analysis)

ChemEs care about **systems** (e.g., how are things connected?)

ChemEs care about **data science & machine learning** (e.g., neural nets, computer vision)

CBE 562 – Stats for ChemEs

Module I: Intro (From Data to Models to Decisions)

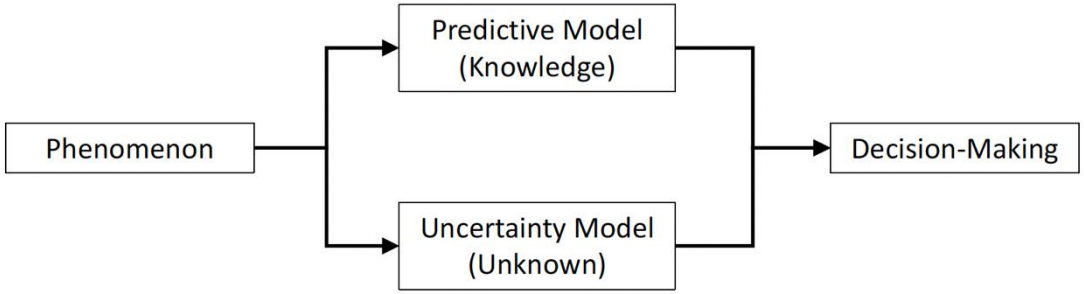
Module II: Random Phenomena

Module III: Estimation

Module IV: Statistical Learning

Module V: Decisions Under Uncertainty

CBE 562 – Stats for ChemEs



Module I: Intro (From Data to Models to Decisions)

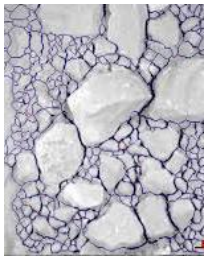
The Known (Predictable) vs. The Unknown (Not Predictable)

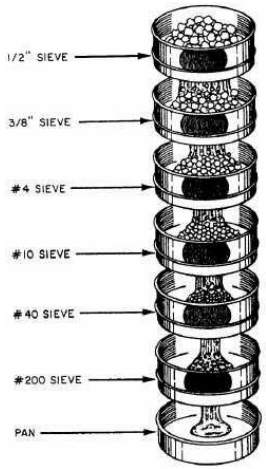
Obtaining Models from Data

Quantifying Uncertainty

Decisions under Uncertainty

CBE 562 – Stats for ChemEs





Module II: Random Phenomena

Univariate Random Variables

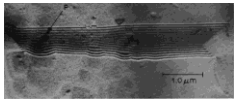
- Poisson
- Gaussian
- LogNormal
- Weibull,…

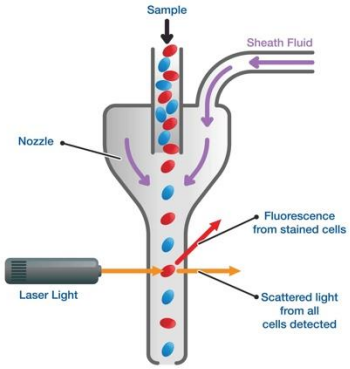
Multivariate Random Variables

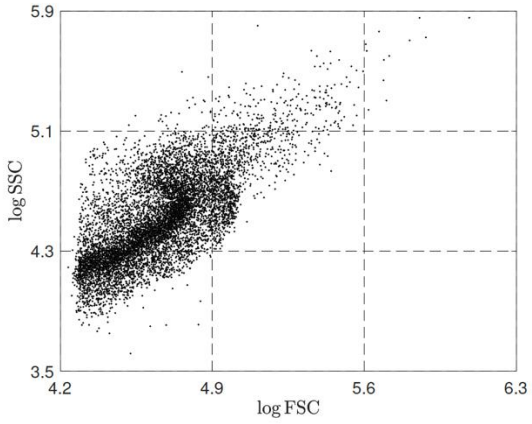
- Joint/Marginal/Conditional
- Correlation
- Events

Connections to Phenomena

- Particle Size Distributions
- Failure Times
- Random Walks
- Diffusion
- Residence Time







CBE 562 – Stats for ChemEs

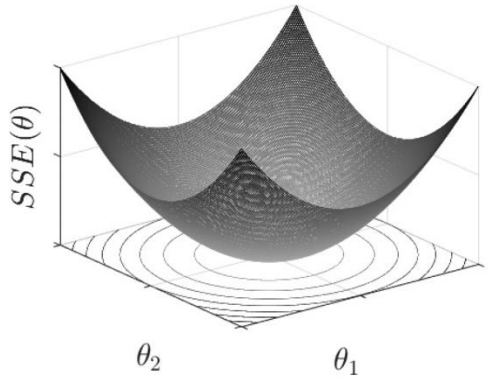
Module III: Estimation

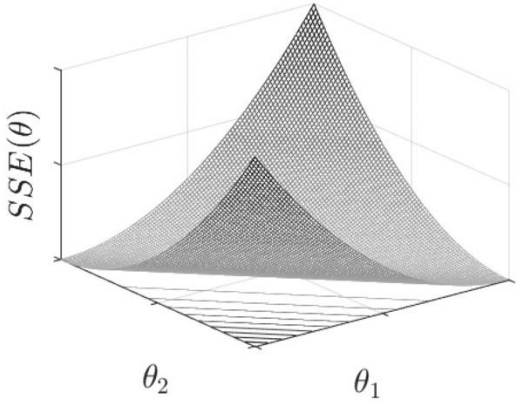
Estimation for Linear Models
- Max Likelihood
- Information Quantification
- Uncertainty Quantification

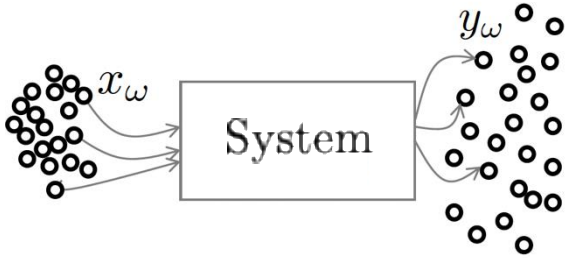
Estimation of Nonlinear Models
- Numerical Optimization
- Uniqueness
- Regularization

Bayesian Estimation
- Prior Knowledge
- Posterior Distribution

Monte Carlo Methods
- Law of Large Numbers
- Convergence

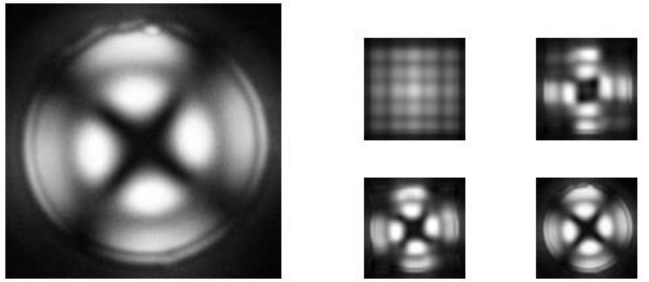






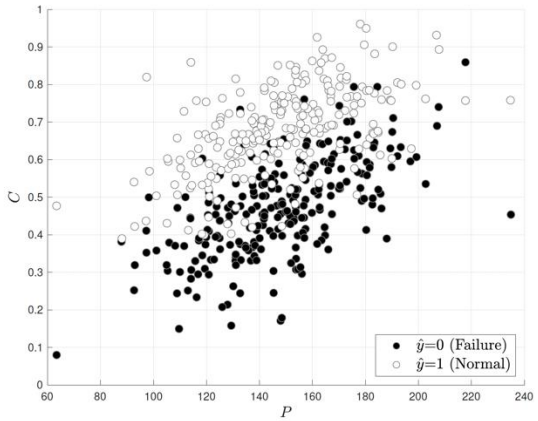
CBE 562 – Statistics for ChemEs

Module IV: Statistical Learning



Dimensionality Reduction
- Principal Component Analysis
- SVD

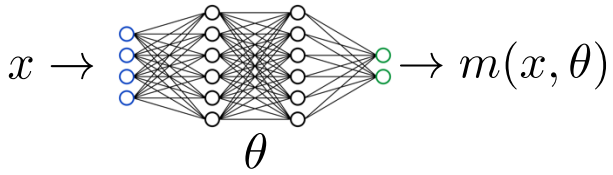
Data Processing
- Fourier, Convolution
- Computer Vision



Kernel Models
- Basis Functions
- Gaussian Processes

Experimental Design
- Bayesian Optimization
- Exploration vs. Exploitation

Machine Learning:
- Classification
- Neural Networks
- Combining Physics & ML



CBE 562 – Statistics for ChemEs

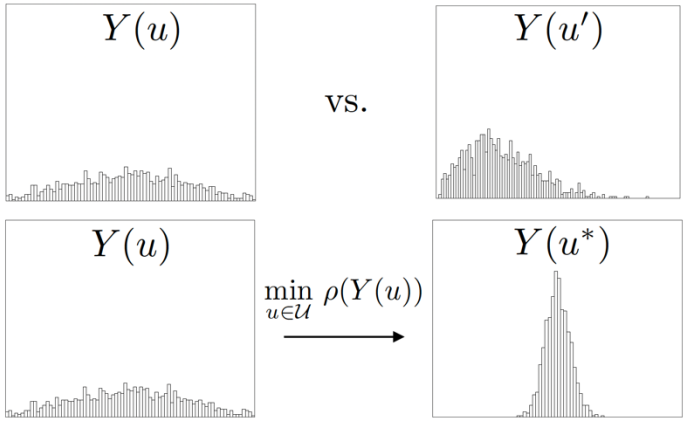
Module V: Decisions Under Uncertainty

Attitudes Towards Risk
- Risk Neutral
- Risk Taker
- Risk Averse

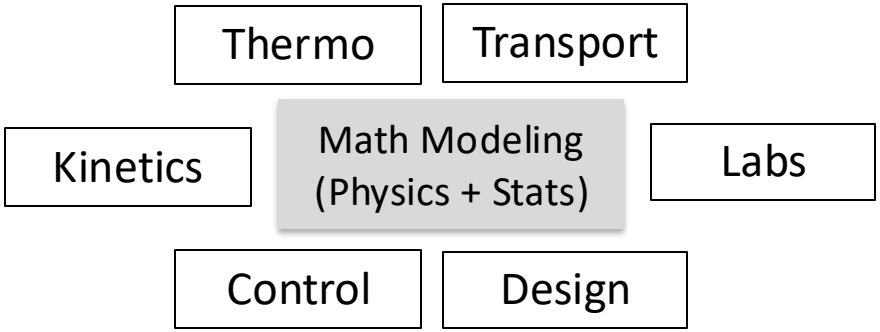
Modeling Risk
- Probability of Failure
- Conditional Value-at-Risk

Comparing Random Variables
- Stochastic Dominance
- Wasserstein Distance

Stochastic Optimization
- Controlling Risk
- Recourse & Feedback



Teaching Philosophy



This is a **modeling (statistical)** course that complements **modeling (physics)** courses

This **is NOT** a data science & ML course

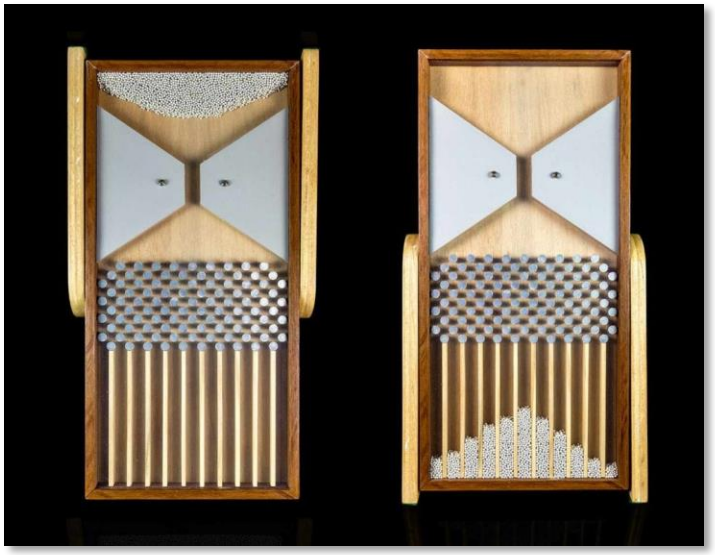
- This is about **foundations** (broadly applicable) → ML is **important** application

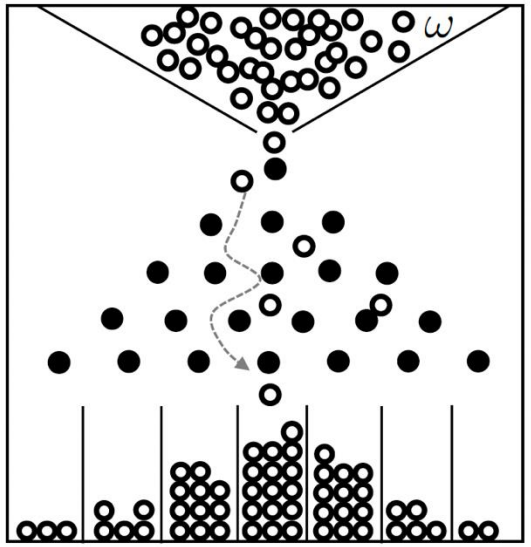
Data provides great venue to emphasize relevance of math:

- Optimization, Eigenvalues, Fourier…
- e.g., Why should I care about **2nd derivatives**? Because this is a **measure of information**.
- e.g., Why should I care about the **rank of a matrix**? Because this indicates **data redundancy**.

Statistics for ChemEs: Examples

Random Variable (Uncertainty) Models

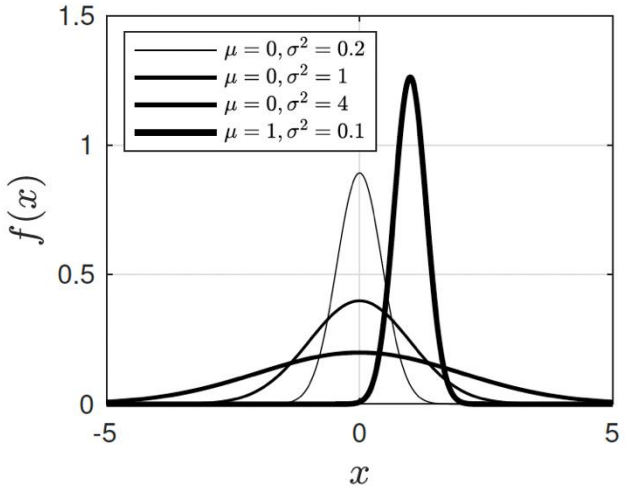




$$\frac{\partial f(x,t)}{\partial t} = D\frac{\partial^2 f(x,t)}{\partial x^2}$$

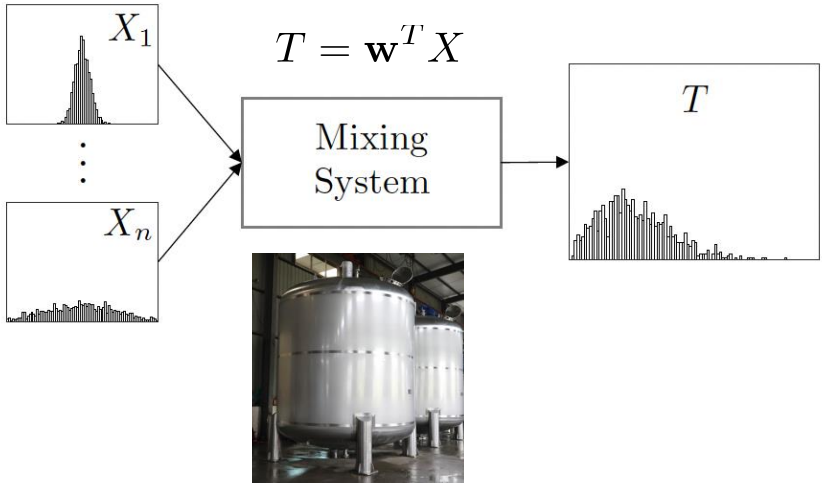
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$\sigma^2 = 2Dt$$

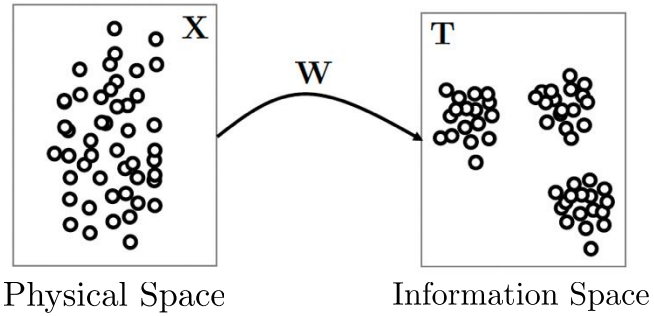


Key: Nature Exhibits Random Behavior (e.g., Random Walks, Collisions, Failure)

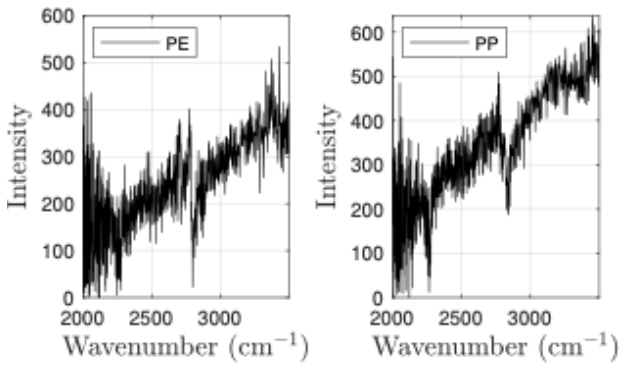
Principal Component Analysis (PCA)

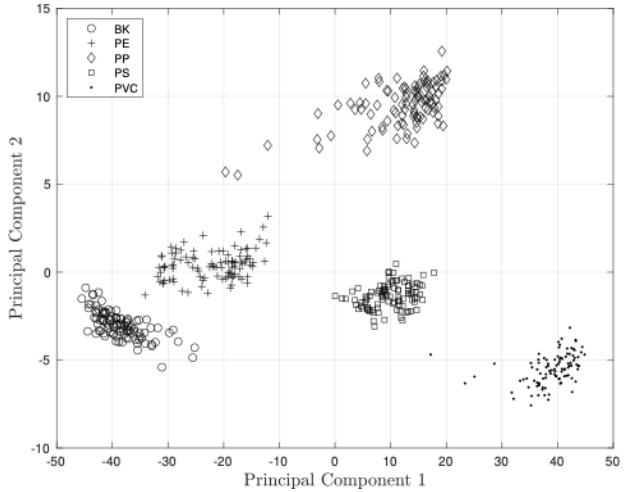


$$\mathbf{T} = \mathbf{XW}$$



$$\mathbf{X}^T\mathbf{X} = \lambda_1\mathbf{w}_1\mathbf{w}_1^T + \cdots + \lambda_n\mathbf{w}_n\mathbf{w}_n^T.$$



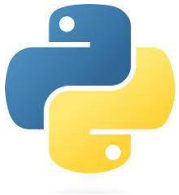


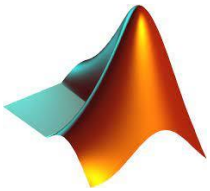
Key: - PCs are mixtures of random variables (find "best" mixtures).

- PCA projects data from physical to information space.

Statistics for ChemEs: Final Thoughts

What About Software?



So far, I have used **Matlab & Python** (scripts/notebooks used as templates)

Avoid using **built-in functions** of packages (avoid black-box thinking).

Not a fan of **domain-specific** software (Minitab, R)

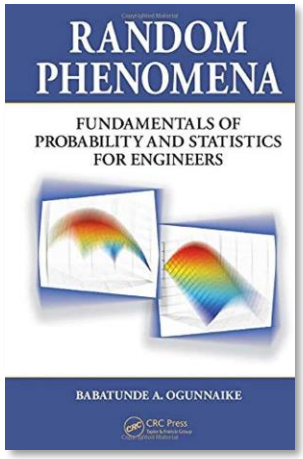
Can be easily learned if you know the fundamentals

What About Textbook?

As usual, **no book satisfies the instructor**

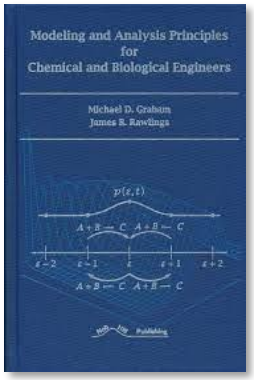
I find "**Stats for Engineers**" books to be too "applied":

- No context of physics



I really like "**Random Phenomena**" by Ogunnaike

- First-principles derivations of models

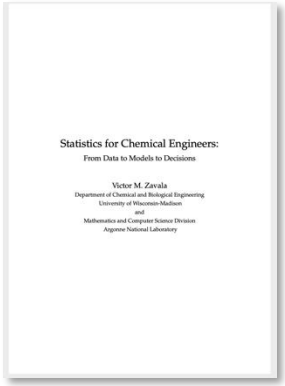


I really like "**Modeling and Analysis Principles for Chemical and Biological Engineers**" by Graham & Rawlings

- Focuses on math foundations

I decided to write my **own textbook**:

- "Statistics for Chemical Engineers: From Data to Models to Decisions"
- 500 +/- 2 pages of statistical fun



Acknowledgements

Students:

- Alex Smith
- Bruce Jiang
- Amy Qin
- Sungho Shin
- CBE 562 Students





Colleagues & Collaborators:

- Thatcher Root (UW-Madison)
- Reid Van Lehn (UW-Madison)
- Mike Graham (UW-Madison)
- Nick Abbott (Cornell)
- Jim Rawlings (UCSB)
- Joe Qin (Lingnan)
- Larry Biegler (CMU)
- Richard Braatz (MIT)
- Sal Garcia (Lilly)
- Swee-Teng Chin (Dow)



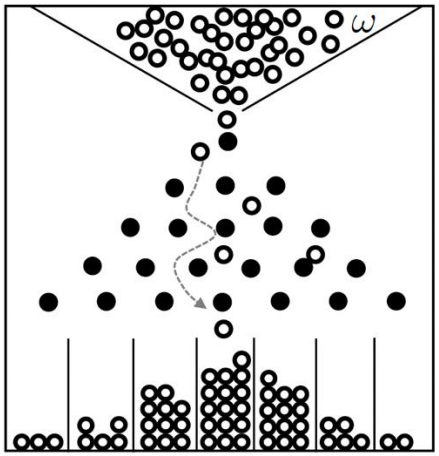






Funding:

- NSF CAREER (CBET-1748516)
- NSF BIGDATA (IIS- 1837812)
- TWCCC Consortium



Experiences in Teaching Statistics & Data Science to ChemE Students at UW-Madison

Victor M. Zavala

Department of Chemical & Biological Engineering
University of Wisconsin-Madison

Mathematics and Computer Science Division
Argonne National Laboratory