Georgia Institute of Technology

ISyE INDUSTRIAL & SYSTEMS ENGINEERING

DOE in the High-Tech Age

C. F. Jeff Wu

School of Industrial and Systems Engineering,
Georgia Institute of Technology

- DOE (design of experiments): from past to future
- Two illustrative examples in the high-tech age
  - computational mechanical material design, data center thermal distribution (FEA runs, statistics-based meta modeling)
  - robust synthesis of nanostructures (modeling, robust process conditions)
- Drastic change in the mode of future design research

A brief history of DOE:

Scientific motivation and intellectual product

- Agricultural research (Fisher, Yates, etc.):
  randomization, blocking, combinatorial (block/ square) designs, ANOVA.
- Process modeling and optimization (Box):
  regression designs (e.g., central composite designs, optimal designs),response surface methodology.
- Manufacturing and quality improvement (Taguchi, etc.):
  robust parameter design, signal-to-noise ratio, response and performance measure modeling.

What's next? The high tech revolution

- Availability of massive data (thanks to advances in computing and hardware): no design issue, bad news for design research 
- Physical experiments replaced by computer experiments (savings in cost and time, more feasible): a definite opportunity. Are we going to miss it?
- Other opportunities abound (nanotechnology, molecular medicine, biotech devices, alternative fuel): unknown territory, tremendous promises.

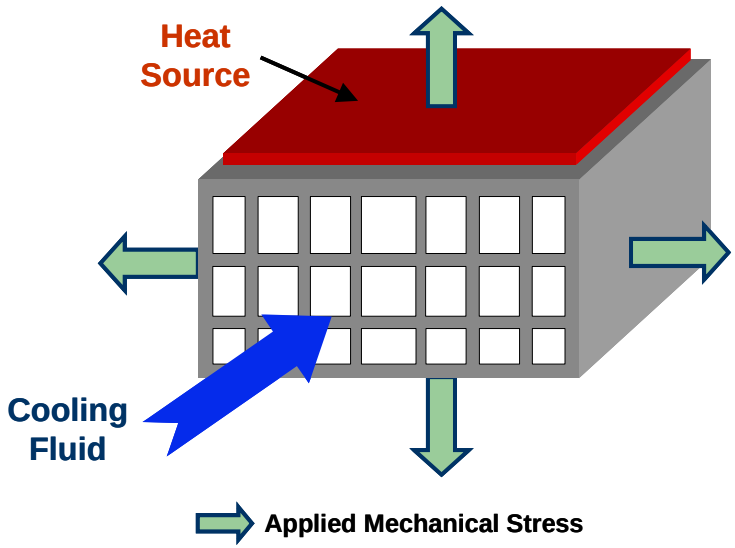
Two illustrative examples

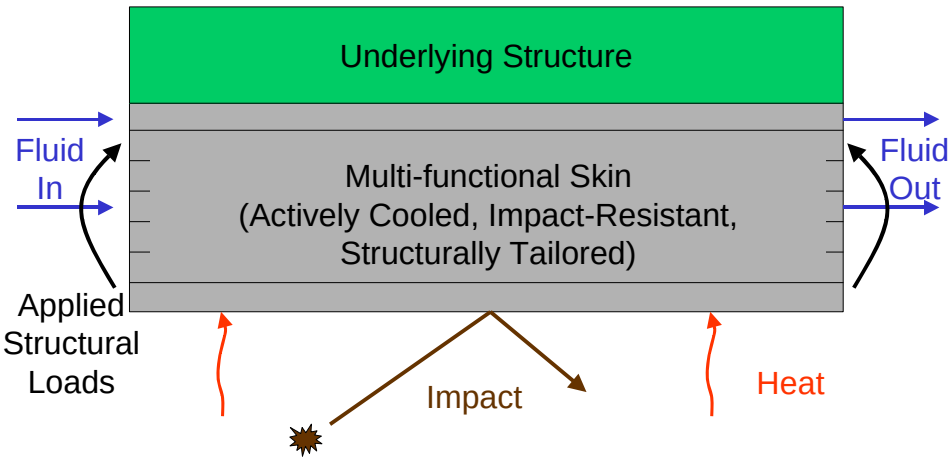
- Integration of low-accuracy and high-accuracy experiments: Bayesian hierarchical gaussian process modeling, computational techniques, application to linear cellular material design.
  (Joint work with ME colleagues at GT.)
- Robust synthesis of nanowires, nanosaws, nanobelts: a multinomial GLM, a new fitting algorithm, robust process conditions insensitive to inner noise.
  (Joint work with material science/engr. colleagues at GT.)

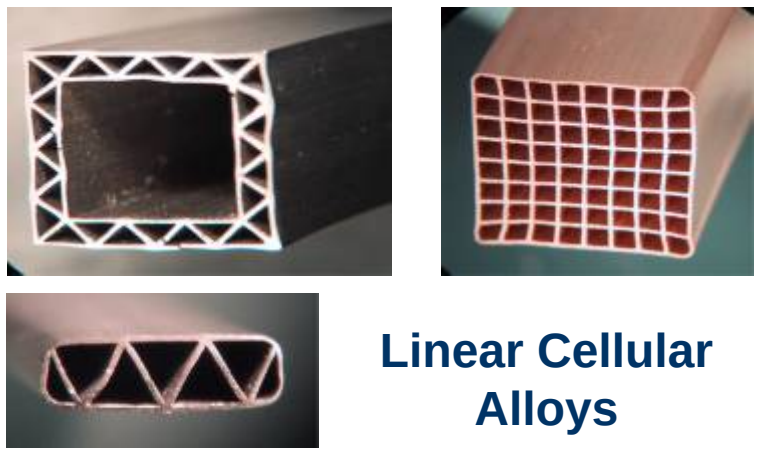
High-accuracy and Low-accuracy Experiments

- **Paradigm shift**: single experiment → multiple experiments with different levels of accuracy.
- A generic pair: **high-accuracy experiment** (HE) and **low-accuracy experiment** (LE).
- HE is more **accurate** but more **expensive** than LE.
- Examples:
  - physical experiment vs. computer simulation
  - detailed computer simulation vs. approximate computer simulation
- **Modeling**:
  How to model and analyze data from HE and LE?
  Experimental design:
  How to plan HE and LE?

Example 1: Designing Cellular Heat Exchangers for an Electronic Cooling Application







Linear Cellular Alloys

Important Factors:

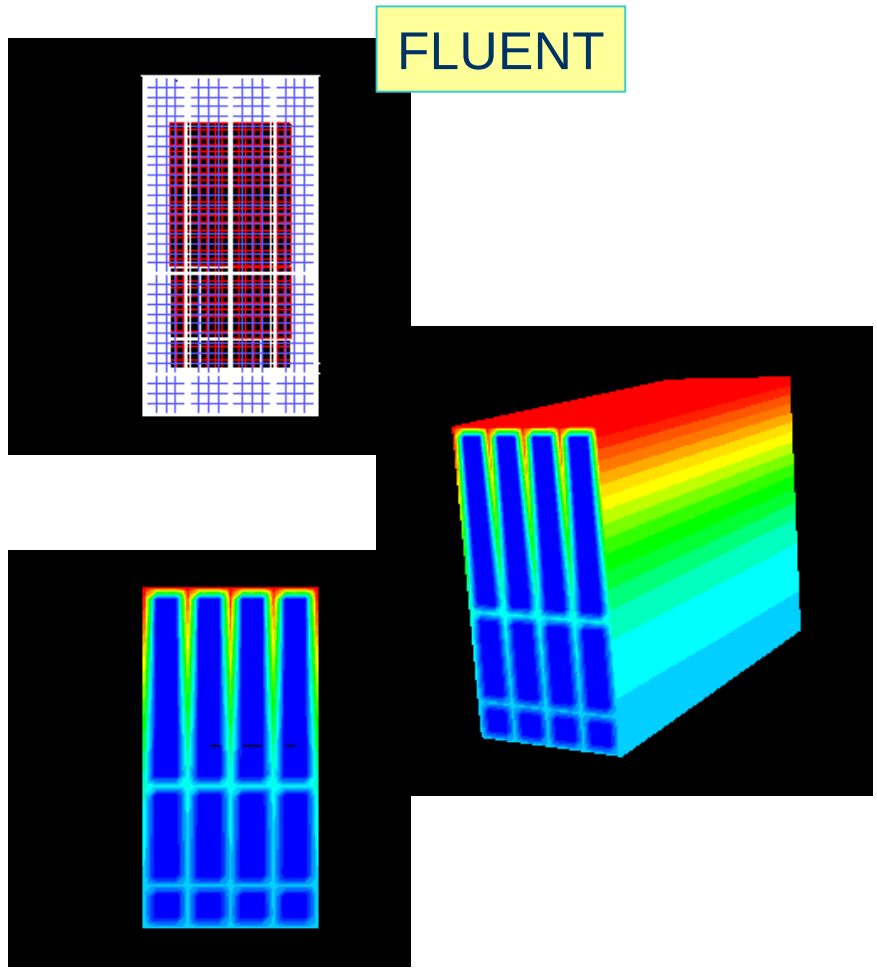
- Flow-rate of Air
- Inlet Temp of Air
- Conductivity of Solid
- Temp of Upper Wall

Response: Total Heat Transfer Rate from Solid to Air

Courtesy of Systems Realization Lab at Georgia Tech

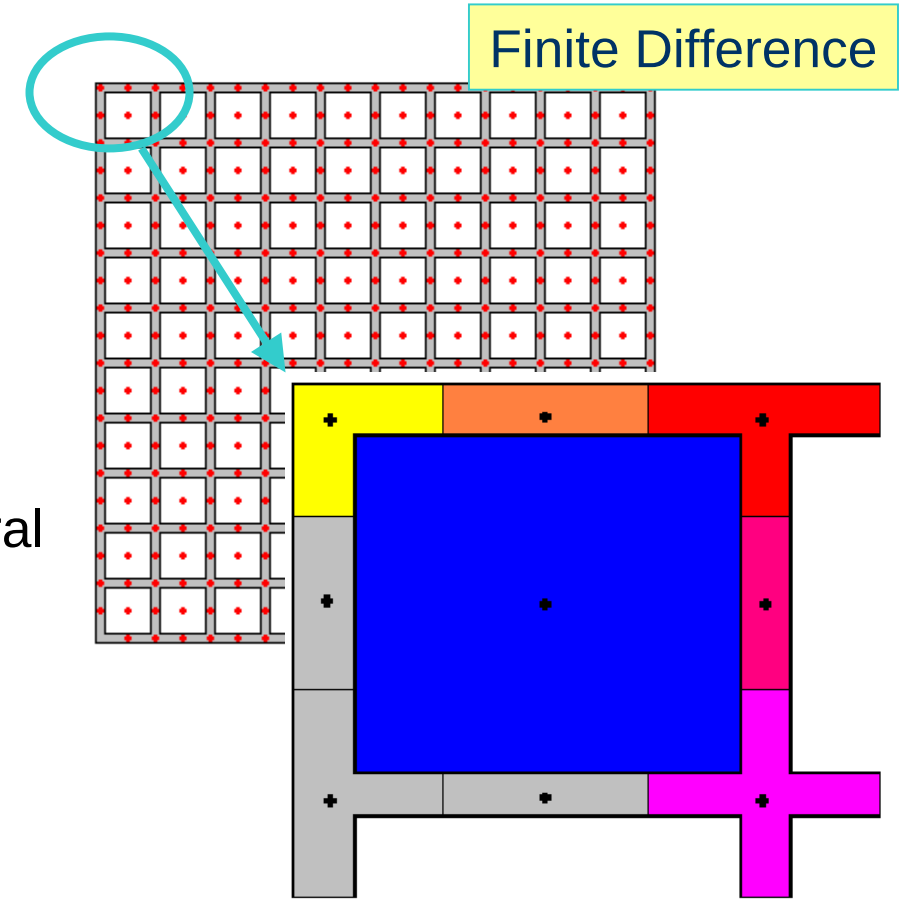
Heat Transfer Analysis
HE: Detailed Computer Simulation – Finite Element Analysis (FEA) Method

- Using the computational fluid dynamic solver FLUENT
- Problem domain is divided into thousands or millions of elements.
- Each run requires hours to days to complete.

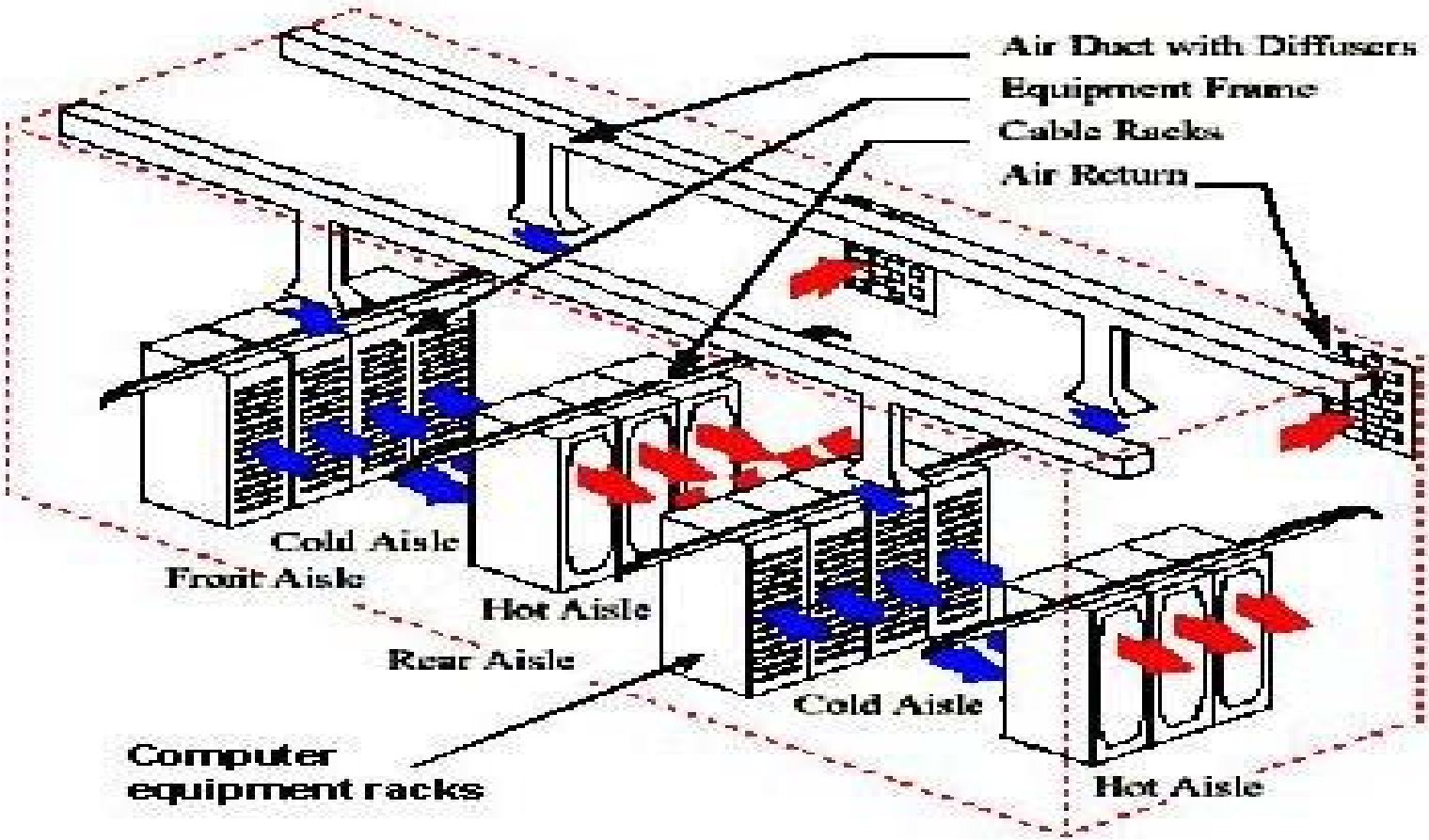


Heat Transfer Analysis
LE: Approximate Computer Simulation —Finite Difference Method

- The finite difference technique is a numerical technique for solving 2- or 3-D steady state heat transfer problems.
- Temperature distribution approximated via numerical solution of 3D heat transfer equations using forward or central difference methods.
- Each run takes minutes to complete.
- Less accurate than FEA.



Example 2: Modeling Thermal Distribution of a Data Center



Courtesy of IBM T. J. Waston Research Center

HE and LE for Modeling Data Center Thermal Distribution

HE: Physical experiment

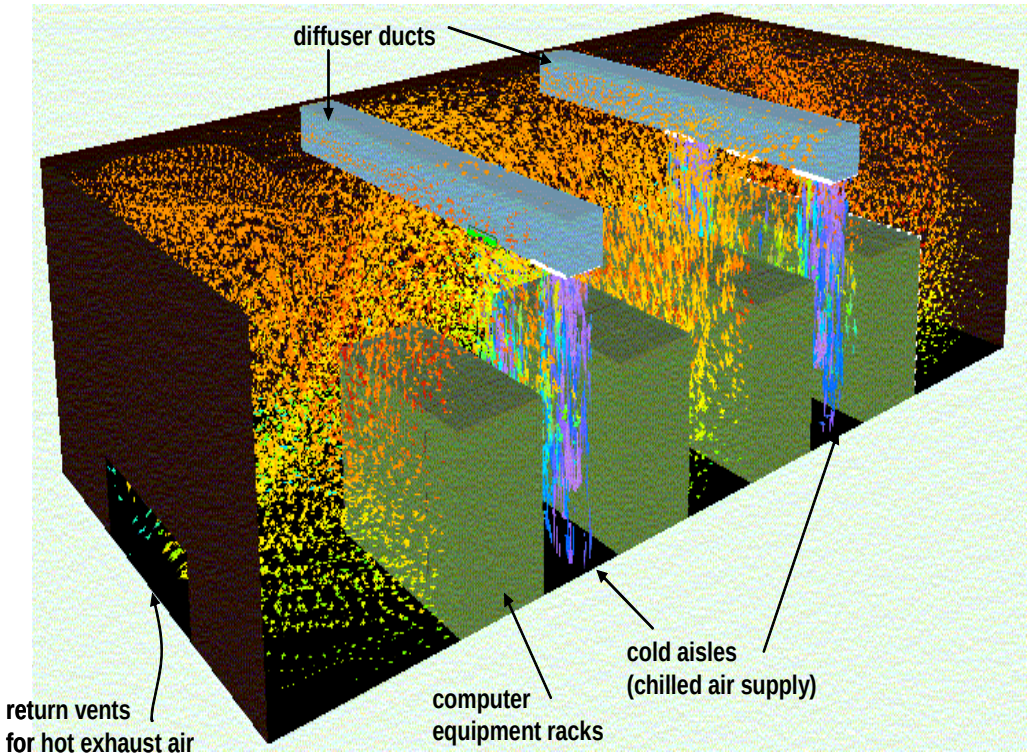
• Building a large data center costs a few million dollars and takes several months to complete.

LE: Computer simulation based on computational fluid dynamics, Flotherm

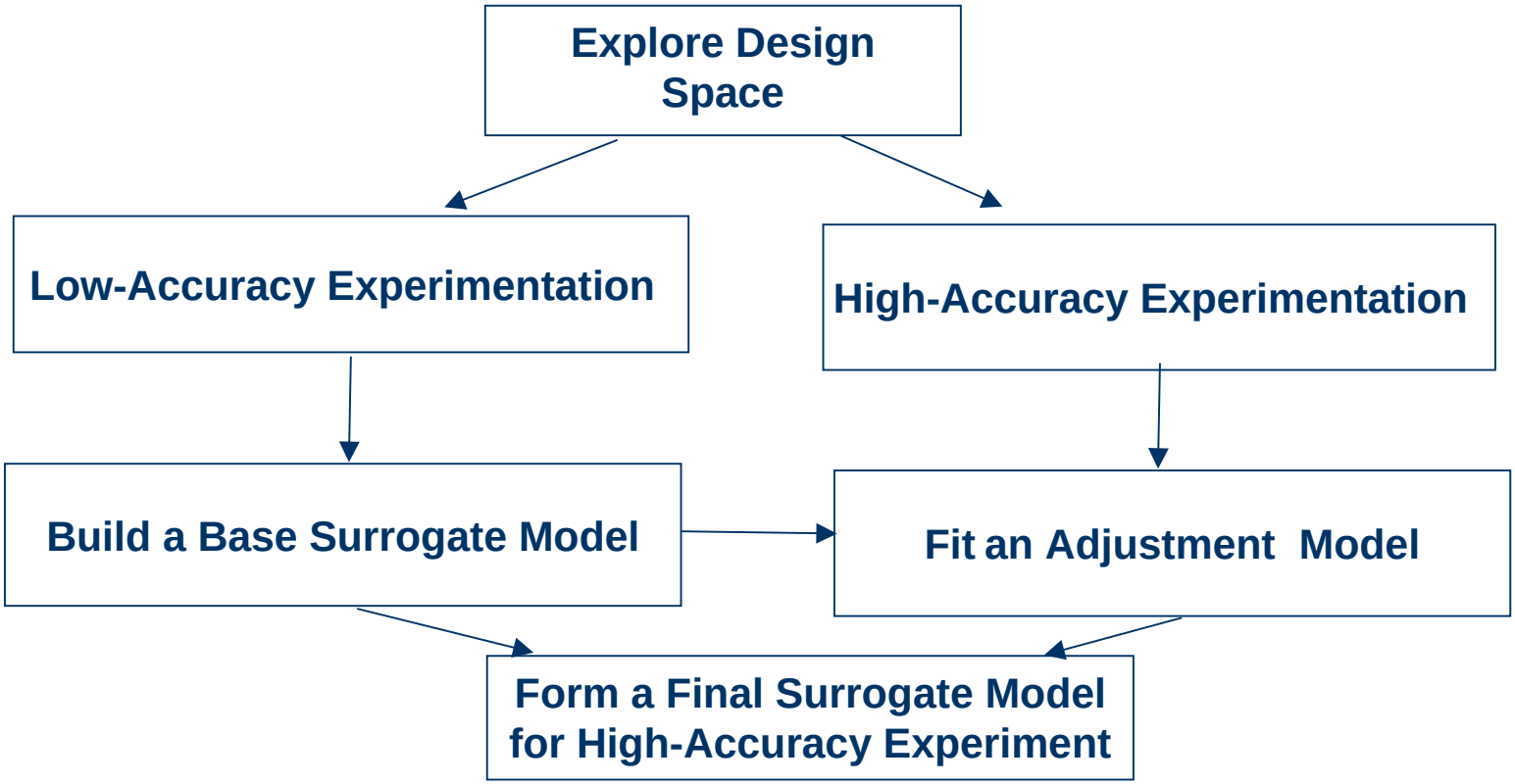
• Price for Flotherm: less than $2000.

• Each run takes hours to complete.

• Result is not as accurate as the physical experiment.



Schematic of Integrated Analysis of HE and LE Data



Gaussian Process Modeling

- Popular tool in computer experiment, global optimization and machine learning. Suitable for modeling **non-linear** phenomena.
- Data: k: the number of variables, n: the number of points, $\mathbf{x}_i = (x_{i1}, \cdots, x_{ik})$: sampled point i, $y_i = y(\mathbf{x}_i)$: response value.
- Model:

$$y(\mathbf{x}_i) = \sum_m \beta_m f_m(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i),\ i = 1, \ldots, n,$$

  - $f_m(\mathbf{x})$'s: functions of $\mathbf{x}$, β_m's: unknown coefficients,
  - $\varepsilon \sim GP(0, \sigma^2, \phi)$: Gaussian process with mean zero, variance σ^2 and correlation parameters ϕ.
- Gaussian correlation function:

$$\mathrm{corr}[\varepsilon(\mathbf{x}_i), \varepsilon(\mathbf{x}_j)] = \exp[-\sum_{h=1}^{k} \phi_h |\mathbf{x}_{ih} - \mathbf{x}_{jh}|^2].$$

- Its predictor can **interpolate** observed data points $y_i, 1 = 1, \ldots, n$.

Method 1: Frequentist Approach in Qian, Seepersad, Joseph, Allen and Wu (2006)

- $\mathbf{x} = (x_1, \ldots, x_k)$: design variables in $[0,1]^k$.
  D_l and D_h: sets of design points ($\mathbf{x}_i$) for LE and HE.
  $D_h \subset D_l$. y_h and y_l: outputs from HE and LE.
- **Base Surrogate Model**: $y_l = \beta_{l0} + \sum_j \beta_{lj}\mathbf{x}_i + \varepsilon_l(\mathbf{x}_i), \ \ \mathbf{x}_i \in D_l,$
  $\varepsilon_l \sim GP(0, \sigma_l^2, \phi_l)$.
- **Adjustment Model**: $y_h(\mathbf{x}_i) = \rho(\mathbf{x}_i) y_l(\mathbf{x}_i) + \delta(\mathbf{x}_i), \ \ \mathbf{x}_i \in D_h,$
  - **scale adjustment**: $\rho(\mathbf{x}_i) = \rho_0 + \sum_j \rho_j x_{ij}$;
  - **location adjustment**: $\delta \sim GP(\delta_0, \sigma_\delta^2, \phi_\delta)$.
- Fitting:

$$\left.\begin{array}{c}\text{Base surrogate model: } \widehat{y}_l \\ \text{Adjustment model: } \widehat{\rho} \text{ and } \widehat{\delta}\end{array}\right\} \Longrightarrow \textbf{Final surrogate model}: \widehat{y}_h = \widehat{\rho}\widehat{y}_l + \widehat{\delta}.$$

- Desirable property: $\widehat{y}_h$ **interpolates** $y_h(\mathbf{x}_i), \mathbf{x}_i \in D_h$ if HE is deterministic.

Method 2: Bayesian Approach in Qian and Wu (2006)

- Use flexible **Bayesian hierarchical Gaussian process model (BHGP)**.
- Provide more **flexible adjustment**. Can accommodate **parameter uncertainty** and measurement error of HE.
- BHGP:
  - **Base Surrogate Model**: $y_l = \beta_{l0} + \sum_j \beta_{lj} \mathbf{x}_i + \varepsilon_l(\mathbf{x}_i), \ \ \mathbf{x}_i \in D_l$, $\varepsilon_l \sim GP(0, \sigma_l^2, \phi_l)$.
  - **Flexible Adjustment Model**:

$$
\begin{aligned}
y_h(\mathbf{x}_i) &= \rho(\mathbf{x}_i) y_l(\mathbf{x}_i) + \delta(\mathbf{x}_i) + \varepsilon(\mathbf{x}_i), \ \ \mathbf{x}_i \in D_h, \\
\text{scale adjustment } \rho &\sim GP(\rho_0, \sigma_\rho^2, \phi_\rho), \\
\text{location adjustment } \delta &\sim GP(\delta_0, \sigma_\delta^2, \phi_\delta), \\
\varepsilon(\mathbf{x}) &\sim N(0, \sigma_\varepsilon^2). \ \varepsilon = 0 \Leftrightarrow \text{No measurement error.}
\end{aligned}
$$

Model Parameters and Priors for the BHGP Model

- Model parameters
  - Mean parameters $\theta_1 = (\beta_l, \rho_0, \delta_0)$.
  - Variance parameters $\theta_2 = (\sigma_l^2, \sigma_\rho^2, \sigma_\delta^2, \sigma_\varepsilon^2)$.
  - Correlation parameters $\theta_3 = (\phi_l, \phi_\rho, \phi_\delta)$.
- Priors
  - Normal for θ_1.
  - Inverse-gamma for θ_2.
  - Gamma for θ_3.

Step 1: Fitting Correlation Parameters θ_3

- Obtain posterior mode $\widehat{\theta}_3 = (\widehat{\phi}_l, \widehat{\phi}_\rho, \widehat{\phi}_\delta)$ by solving

$$(P): \max_{\phi_l, \phi_\rho, \phi_\delta} p(\theta_3 | \mathbf{y}_h, \mathbf{y}_l).$$

- (P) is equivalent to two **separable** problems:
  - (P_1) : a non-linear program for ϕ_l,
  - (P_2) : a problem for ϕ_ρ and ϕ_δ. Its objective function involves an integral.
- $$(P_2) \Leftrightarrow \text{a } \textbf{stochastic program } (P_2'): \max_{\phi_\rho, \phi_\delta} L_2 = E_{p(\tau_1, \tau_2)} f(\tau_1, \tau_2).$$
- Solve P_2' by the **Sample Average Approximation (SAA) algorithm**:

  1. Generate random samples (τ_1^s, τ_2^s) from $p(\tau_1, \tau_2), s = 1, \cdots, S$.

  2. Solve the approximate problem:

$$(\widetilde{\phi}_\rho, \widetilde{\phi}_\delta) = \arg\max_{\phi_\rho, \phi_\delta} [\widehat{L}_2 = \frac{1}{S} \sum_{s=1}^{S} f(\tau_1^s, \tau_2^s)].$$

Step 2: Markov Chain Monte Carlo (MCMC) Sampling from $p(\theta_1, \theta_2 | \mathbf{y}_l, \mathbf{y}_h, \theta_3)$

- Some **full conditional distributions** for θ_1, θ_2 are not regular:
  - $p(\beta_l | \mathbf{y}_l, \mathbf{y}_h, \theta_3, \overline{\beta_l}) \sim Normal,$
  - $p(\rho_0 | \mathbf{y}_l, \mathbf{y}_h, \theta_3, \overline{\rho_0}) \sim Normal,$
  - $p(\delta_0 | \mathbf{y}_l, \mathbf{y}_h, \theta_3, \overline{\delta_0}) \sim Normal,$
  - $p(\sigma_l^2 | \mathbf{y}_l, \mathbf{y}_h, \theta_3, \overline{\sigma_l^2}) \sim IG,$
  - $p(\sigma_\rho^2 | \mathbf{y}_l, \mathbf{y}_h, \theta_3, \overline{\sigma_\rho^2}) \sim IG,$
  - $p(\tau_1, \tau_2 | \mathbf{y}_l, \mathbf{y}_h, \overline{\tau_1, \tau_2}) \propto \frac{1}{\tau_1^{\alpha_\delta + \frac{3}{2}}} \frac{1}{\tau_2^{\alpha_\varepsilon + 1}} \exp\{-\frac{1}{\tau_1}(\frac{\gamma_\delta}{\sigma_\rho^2} + \frac{(\delta_0 - u_\delta)^2}{v_\delta \sigma_\rho^2}) - \frac{\gamma_\varepsilon}{\tau_2 \sigma_\rho^2} \frac{1}{|\mathbf{M}|^{\frac{1}{2}}}\}$
    $\cdot \exp\{-\frac{(\mathbf{y}_h - \rho_0 \mathbf{y}_{l_1} - \delta_0 \mathbf{1}_{n_1})^t \mathbf{M}^{-1} (\mathbf{y}_h - \rho_0 \mathbf{y}_{l_1} - \delta_0 \mathbf{1}_{n_1})}{2\sigma_\rho^2}\}$
- Use the **Metropolis-within-Gibbs algorithm.**

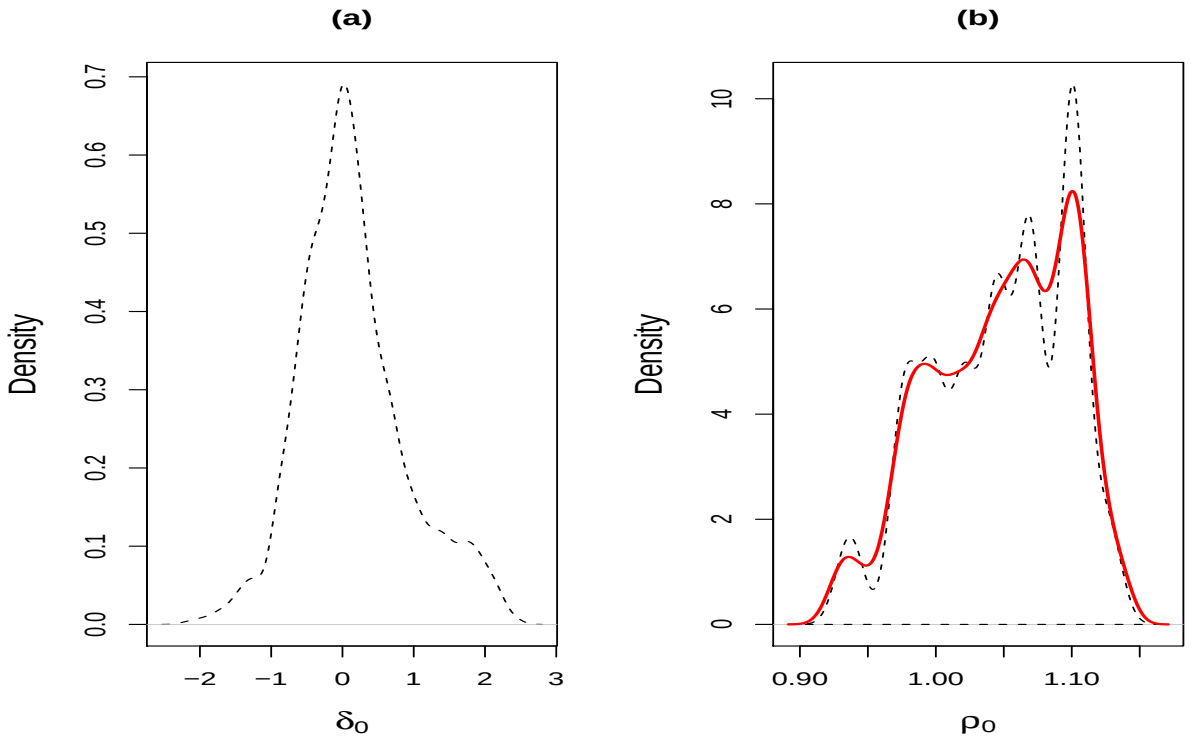
Example: Detailed and Approximate Computer Experiments for Cellular Heat Exchangers

- Response: Q = total heat transfer rate from solid to air.
- 4 input variables

$\dot{m}$ (flow-rate of air)	[0.00055, 0.001]
Tin (inlet temp of air)	[270.00, 303.15]
k (thermal conductivity of solid)	[202.4, 360.0]
T_{wall} (temp of upper wall)	[330, 400]

- 32 HE runs (using Finite Difference software).
- 32 LE runs (using FLUENT software).
- Training set: 24 randomly selected HE runs + 32 LE runs.
- Objective: Predict y_h at the remaining 8 runs.

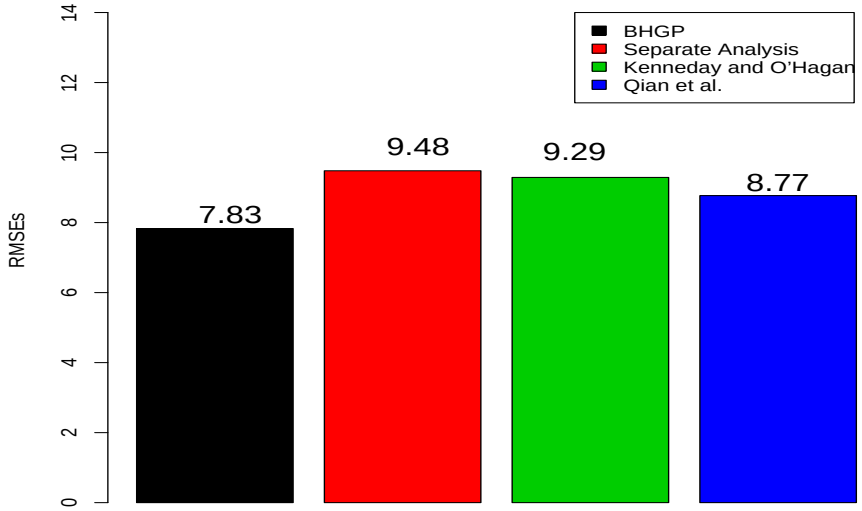
Intricate Location and Scale Change



- δ_0 is symmetric with center at -0.14.
- ρ_0 has multiple modes. ρ_0 may come from a **mixture model**.
- Advantage of the proposed Bayesian approach. This mixture model cannot be captured by using a frequentist approach.

Prediction Results

- Four methods for predicting y_h for the eight testing runs: 1. BHGP model, 2. Separate analysis, 3. Kennedy-O'Hagan (2000) and 4. Qian et al. (2005).
- Compute RMSE (root-mean-square-error)$= \sqrt{\sum_{j=1}^{8}(y_h(\mathbf{x}_j) - \widehat{y}_h(\mathbf{x}_j))^2/8}$.



- Three combined methods beat the separate analysis. BHGP outperforms the other two by 16% and 11% respectively.

Planning of HE and LE

- Key to efficiently allocate resources and acquire information from HE and LE.
- $\mathbf{x} = (x_1, \ldots, x_k)$: design variables in $[0,1]^k$.
  D_l: set of design points ($\mathbf{x}_i$) for LE. D_h: set of design points ($\mathbf{x}_i$) for HE.
- Three principles for constructing D_l and D_h:

  Principle of economy: The size of D_h is less than the size of D_l.

  Principle of nested relationship: D_h is a subset of D_l.

  Principle of uniformity: Points in D_h and D_l are uniformly distributed.
- How to construct **multiple experiments** with respect to **multiple requirements**?

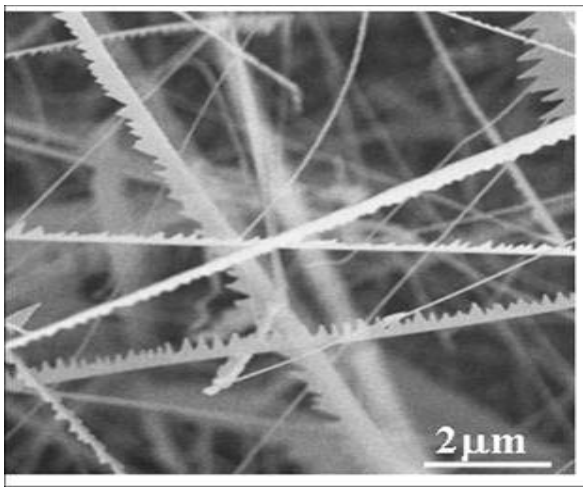
Role of statistics in nanotechnology research

- Nanotechnology research
  - Shift from laboratory-level experimentation to controlled and large scale synthesis.
  - High yield and reproducibility.

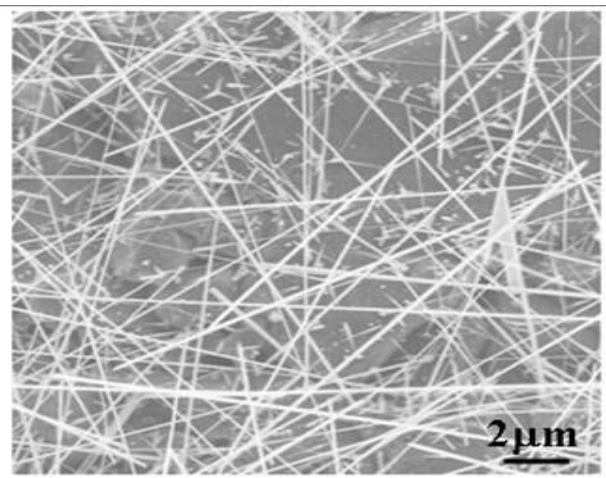
- Role of statistical methodology
  - Systematically investigating the experimental conditions for achieving the desired nanostructures.
  - Building empirical models to express yields and properties of various types of nanostructures as functions of process variables.
  - Developing robust synthesis processes for producing nanostructures with high yield and minimal variation.

Dasgupta, Ma, Joseph, Wang and Wu (2006)

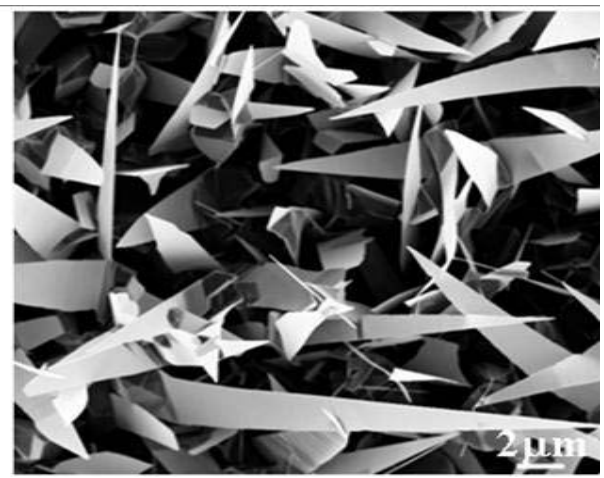
Different CdSe nanostructures



NANOSAWS



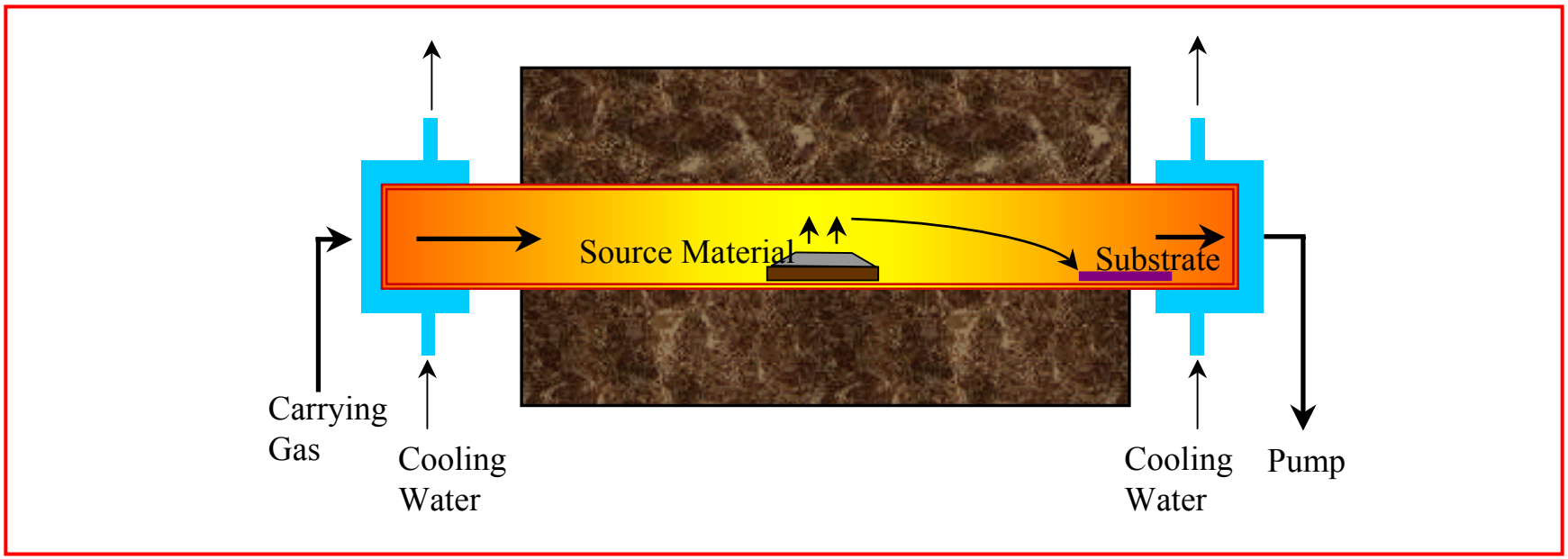
NANOWIRES



NANOBELTS

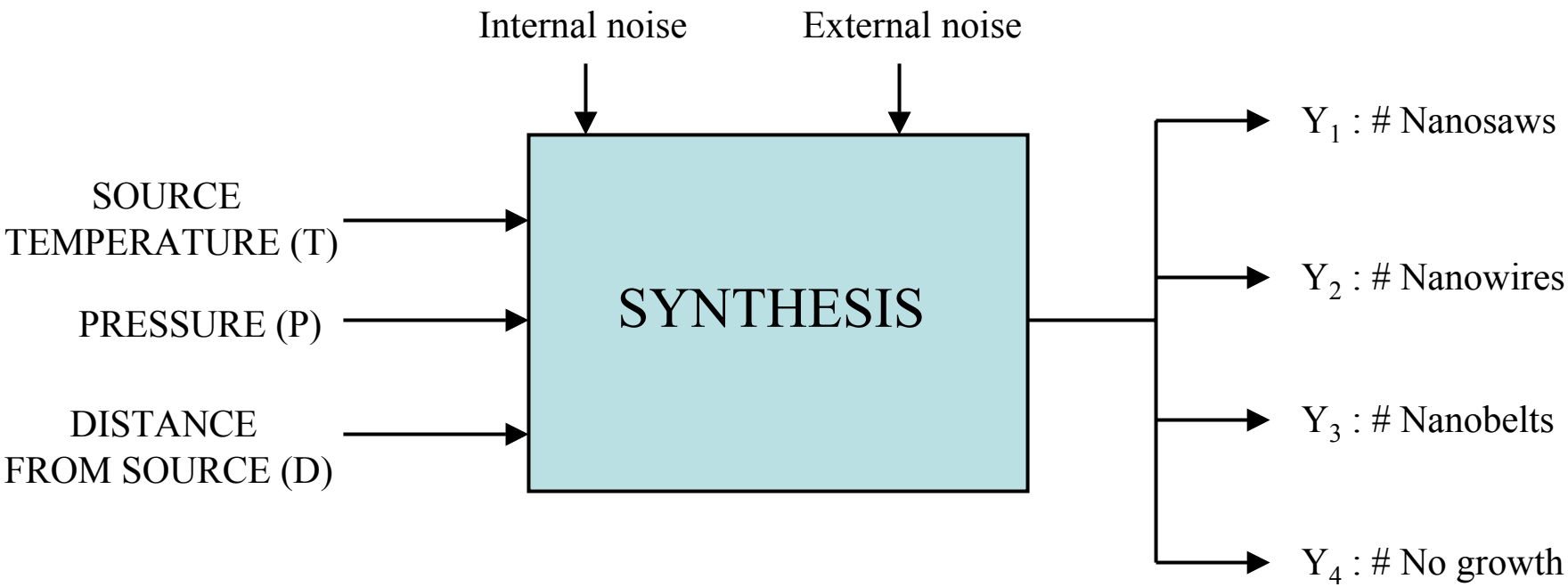
•Synthesized through a thermal evaporation process in a single-zone horizontal tube furnace.

Synthesis process



- Two main control variables
  - Source temperature (T)
  - Pressure (P)
- Distance (D) of the substrate from the source is a covariate.
- On each substrate
  - A deposition is obtained.
  - 180 individual nanostructures counted using Scanning Electron Microscopy (SEM).

A schematic description



- $Y_1 + Y_2 + Y_3 + Y_4 = 180$.
- (Y_1, Y_2, Y_3, Y_4) is multinomial $(180, p_1, p_2, p_3, p_4)$.

Modeling strategy : multinomial GLM

- Multinomial logits : $\eta_j = \log(\frac{p_j}{p_4}), j = 1,2,3.$
- Interpretation of η_j : The logg-odds ratio of obtaining the j^{th} morphology as compared to no nanostructure, with $\eta_4 = 0$.
- Model η_j as functions of T, P, D (each process variable scaled to [-1,1]).

Existing methods

- Use a Poisson surrogate model
  - Create a pseudo factor with a level for each data point.
  - Cumbersome for large datasets (Faraway 2006).

- Direct maximization of multinomial likelihood using neural network (Venebles and Ripley, 2002)
  - S-PLUS and R modules available.
  - Separate inference for sub-models is not possible.

New iterative method

- $\eta_{ij} = \mathbf{x}_i' \beta_j, \quad j = 1, 2, 3.$
- Define $\exp(\gamma_{i23}) = \frac{1}{1+\exp(\eta_{i2})+\exp(\eta_{i3})}$.
- From the ML equations of the multinomial GLM,

$$\sum_{i=1}^{N} \mathbf{x}_i \left(y_{i1} - n_i \frac{\exp(\eta_{i1} + \gamma_{i23})}{1 + \exp(\eta_{i1} + \gamma_{i23})} \right) = \mathbf{0}$$

- The above is the ML equation of a binomial GLM with logit link.
- The following iterative procedure is thus used
  - Obtain initial estimates of β_2 and β_3 (use binomial GLM).
  - Compute η_2, η_3 and consequently γ_{23}.
  - Estimate β_1 by fitting a binomial GLM of Y_1 on $\mathbf{X}$.
  - Repeat the process with γ_{13} and γ_{12} defined similarly.

Fitted models

$$\widehat{\eta_1} = 0.42 - 0.12\ T - 3.08\ P - 3.68\ D - 1.84\ T^2 - 1.52\ P^2 - 9.09\ D^2 + 0.60\ TP - 2.31\ PD + 5.75\ TD \quad (\textit{Generalized}\ R^2 = 61\%), \qquad (1)$$

$$\widehat{\eta_2} = 0.54 + 0.88\ T - 3.85\ P - 3.13\ D - 1.21\ T^2 - 2.28\ P^2 - 5.26\ D^2 + 1.83\ TP - 2.62\ PD + 2.07\ TD \quad (\textit{Generalized}\ R^2 = 50\%), \qquad (2)$$

$$\widehat{\eta_3} = -\ 0.10 + 0.39\ T - 3.67\ P - 2.51\ D - 2.51\ T^2 - 1.12\ P^2 - 7.07\ D^2 + 1.72\ TP - 2.38\ PD + 4.47\ TD \quad (\textit{Generalized}\ R^2 = 76\%). \qquad (3)$$

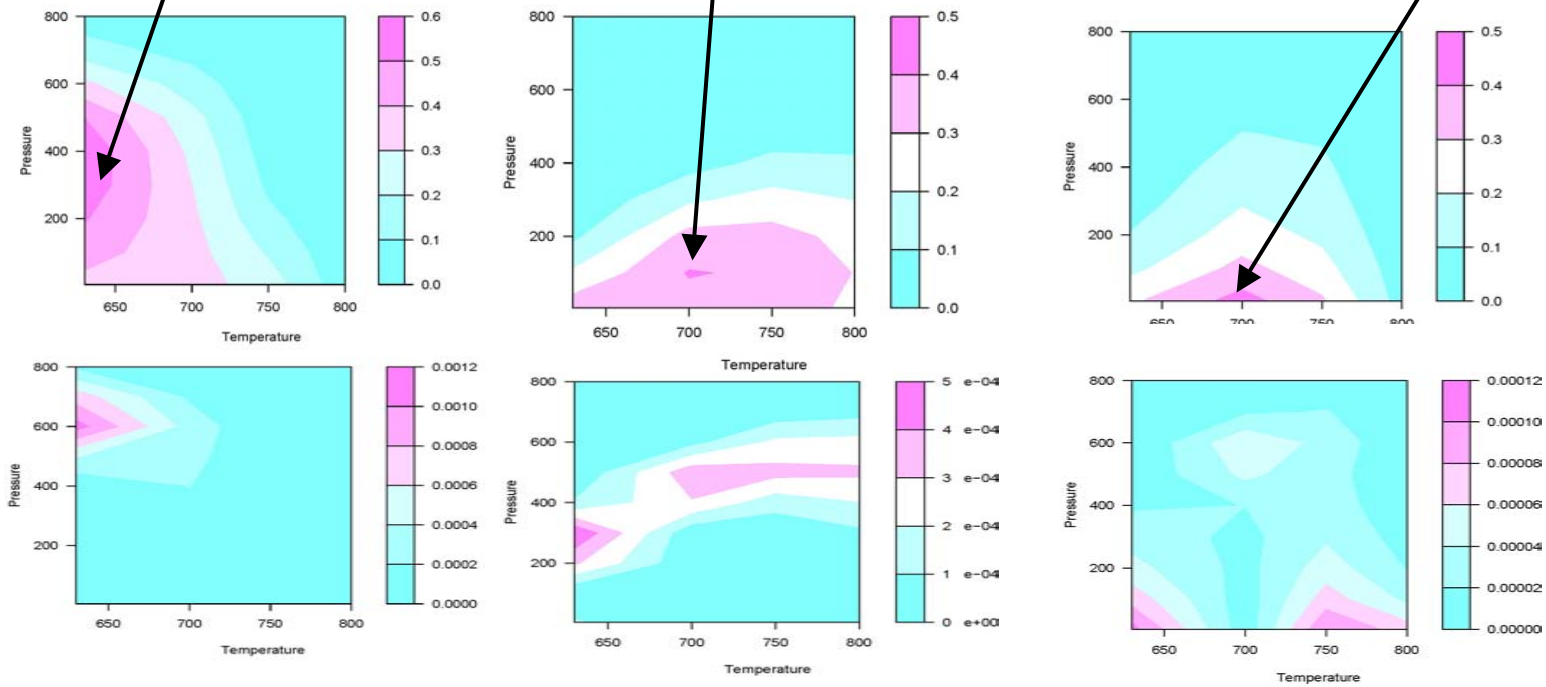
- Convergence of algorithm established.
- The algorithm enables binomial GLM diagnostic tools to be used for inference.
- Separate evaluation of the three sub-models is possible.

Achieving robustness : optimization of process parameters

- T, P, D treated as random variables due to inner noise.
- $T \sim N(\mu_T, \sigma_T^2)$, $P \sim N(\mu_P, \sigma_P^2)$, $D \sim N(\mu_D, \sigma_D^2)$.
- σ_T^2, σ_P^2, σ_D^2 estimated from data.
- Expectation and variance of binomial log-odds ratio (ζ_j) for each nanostructure expressed as functions of μ_T, μ_P, μ_D using Monte-Carlo simulations.
- Maximization of $E(\zeta_j)$ with respect to μ_T, μ_P, μ_D subject to constraints over $Var(\zeta_j)$.

Optimal conditions

Nanostructure	Temperature	Pressure	Distance
Nanosaws	630	307	15.1
Nanowires	695	113	19.0
Nanobelts	683	4	17.0



Impact of the study

- First instance of application of statistical techniques in nanotechnology research.
- Significant advancement over the rudimentary data analysis methods that have been reported in nanotechnology.
  - Slight changes in the growth can be overlooked in the current methodology of nanomaterial characterization, possibly leading to inaccurate conclusions regarding the control of growth mechanism.
  - Offers the advantage of observing and quantifying subtle changes in the growth of a particular nanostructure as a function of the processing variables.
- An important step towards large-scale controlled synthesis of CdSe nanostructures.

Further Challenges

- The profile of some of the experimental factors (e.g., temperature) change over time and this plays a crucial role in synthesis of nanostructures.
  - From functional response to functional factors ? Challenges in both design and analysis.
- Complete disapperance of morphology in some experimental regions makes exploration of optimal extremely difficult.
  - This would require a new design strategy. A combination of sequential and space-filling designs?

Future of DOE research

- Less emphasis on carefully controlled collection of data:
  Combinational designs less popular?
  Optimal designs and algorithms more popular for problem-specific applications.
- Role of DOE research in computer experiments:
  design issues need to be addressed; bigger impact in modeling (meta-modeling above physical or FEA runs).
- Any design research in massive data?
  - Use of sequential design in data query.
  - Active learning in machine learning (sequentially generating data for learning): ongoing work with Bank of America on fraud detection using adaptive sequential design (better than existing active learning method).

Drastic change in thinking and practice

- Diminishing return from work on theoretical design with no clear motivation.
- Rigidity of combinatorial designs makes them less applicable in novel situations; little interface with rest of statistics
  ($\Rightarrow$ Less visibility, student placement, funding).
- More work on computational algorithms.
- **Move from DOE to DAE** (Design and Analysis of Experiments):
  gradually put more weights on *modeling* and *analysis*. Connected better with other statistical disciplines.

Bayesian Prediction

- Main goal: make prediction of y_h at an untried point $\mathbf{x}_0$.
- WLOG, assume $\mathbf{x}_0 \in D_l/D_h$. Observe $y_l(\mathbf{x}_0)$ but need to predict $y_h(\mathbf{x}_0)$.
- **Important observation:** Given correlation parameters θ_3, mean parameters θ_1 and variance parameters θ_2 can be viewed as coming from **a general linear model**.
- Two-step procedure:
  - 1. Fit θ_3 at their posterior modes. **Use Stochastic programming methods**.
  - 2. Make prediction conditionally on θ_3 based on Bayesian predictive density function:

    $$p(y_h(\mathbf{x}_0)|\mathbf{y}_h,\mathbf{y}_l) = \int_{\theta_1,\theta_2} p(y_h(\mathbf{x}_0)|\mathbf{y}_l,\mathbf{y}_h,\theta_1,\theta_2,\theta_3)p(\theta_1,\theta_2|\mathbf{y}_l,\mathbf{y}_h,\theta_3)d\theta_1 d\theta_2.$$

    Use Markov chain Monte Carlo methods.
- Desirable property: $\widehat{y_h}$ **interpolates** $y_h(\mathbf{x}_i), \mathbf{x}_i \in D_h$ if HE is deterministic.

Data of Cellular Heat Exchangers Example

Run #	$\dot{m}(kg/s)$	$T_{in}(K)$	$k(W/mk)$	$T_{wall}(K)$	y_l	y_h	Status
1	0.0005	293.15	362.73	393.15	25.601	23.54	Test
2	0.00055	315	310	365	21.23	20.15	Train
3	0.000552	293.53	318.63	388.29	11.44	10.17	Train
4	0.000557	290.18	298.27	377.49	15.03	15.29	Test
5	0.00056	277.01	354.98	374	18.55	18.39	Train
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
31	0.000751	287.99	326.02	354.08	36.56	47.05	Train
32	0.000757	300.64	235.03	391.68	27.24	25.82	Train

- Input values are standardized.

Hyperparameters of Priors

Par	Value	Par	Value
α_l	2	u_δ	0
γ_l	1	v_δ	1
α_ρ	2	a_l	2
γ_ρ	1	b_l	0.1
α_δ	2	a_ρ	2
γ_δ	1	b_ρ	0.1
$\mathbf{u}_l$	$(0,0,0,0,0)^t$	a_δ	2
v_l	1	b_δ	0.1
u_ρ	1		
v_ρ	1		

Posterior Modes of θ_3

- Results:

Parameter	Posterior Mode
ϕ_l	(2.83, 2.13,22.65,12.87)
ϕ_ρ	(3.22,7.23,1.26,1.38)
ϕ_δ	(2.26,0.74,6.92,7.24)

MCMC Samples of θ_1 and θ_2

- Implemented in WinBugs.
- 5000 runs as burn-in.
- Another 5000 runs for calculation.
- Posterior means and 95% confidence intervals:

	Posterior mean	Lower Bound	Upper Bound
β_{l1}	-5.334	-7.141	-3.504
β_{l2}	2.379	-1.171	5.78
β_{l3}	0.09424	-1.128	1.255
β_{l4}	5.896	3.639	8.373
ρ_0	1.05	0.94	1.13
σ^2_ρ	0.29	0.16	0.49
δ_0	0.14	-1.24	1.93
σ^2_δ	0.78	0.18	2.70