STAT 8810 Lecture 1

Introduction

Introduction

- Syllabus
- Overview of topics
- Assignments, Midterms and Project
- Teams for each Assignment and for the Project
- Questions?

Modern Science, Statistics and Computing

Statistical learning tools are tools used for understanding data, or understanding patterns hidden within data.

Modern Science, Statistics and Computing

Statistical learning tools are tools used for understanding data, or understanding patterns hidden within data.

1. But they are *empirical* models. What does this mean?

Modern Science, Statistics and Computing

Statistical learning tools are tools used for understanding data, or understanding patterns hidden within data.

1. But they are *empirical* models. What does this mean?
   eg: They might usefully describe relationships in-sample.
   They sometimes usefully describe relationships out-of-sample.
   They don't "prove" anything. They don't confirm (or refute) theoretical models.

Modern Science, Statistics and Computing

Statistical learning tools are tools used for understanding data, or understanding patterns hidden within data.

1. But they are *empirical* models. What does this mean?
   eg: They might usefully describe relationships in-sample.
   They sometimes usefully describe relationships out-of-sample.
   They don't "prove" anything. They don't confirm (or refute) theoretical models.
2. At the same time we sometimes have theoretical models.
   What might these be?

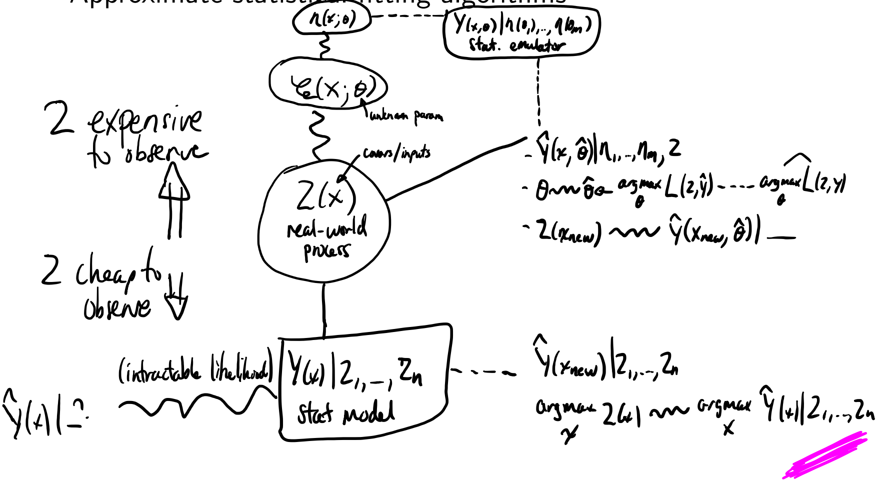
Modern Science, Statistics and Computing

Statistical learning tools are tools used for understanding data, or understanding patterns hidden within data.

1. But they are *empirical* models. What does this mean?
   eg: They might usefully describe relationships in-sample.
   They sometimes usefully describe relationships out-of-sample.
   They don't "prove" anything. They don't confirm (or refute) theoretical models.
2. At the same time we sometimes have theoretical models.
   What might these be?
   eg: Physics: $F = ma$ - a deterministic law;
   Differential Equations: $\frac{\partial \phi(\mathbf{s},t)}{\partial t} = D\nabla^2\phi(\mathbf{s},t)$ (heat, or diffusion equation) - can only be approximated numerically;
   Agent-based simulation models: stochastic laws.

Uncertainty Quantification in a Picture

- Statistical models of large, high-dimensional data
- Statistical emulation of simulators of complex processes
- Approximate statistical fitting algorithms



Modern Science, Statistics and Computing

- Computing becoming an ever-more key issue in our ability to analyze and interpret data. Big contrast from the historical statistical paradigm.
- Requires combining the diverse skillsets currently spread across multiple, previously independent, communities.
- We will be intested in learning a divserse set of statistical learning tools applicable to a diverse set of data, and in understanding to some degree the statistical properties of these “learning algorithms,” or statistical models.

From Physical Experiments...

- Traditional physical experiments involve developing an **empirical** model of a response conditional on covariate information and under an assumed error model
  - Exploratory data analysis $\rightarrow$ statistical empirical model $\rightarrow$ designed experiment $\rightarrow$ hypothesis testing $\rightarrow$ new knowledge $\rightarrow$ new experiment $\rightarrow$ ...
  - Classical examples include ANOVA, linear regression "$Y = X\beta$", ...

From Physical Experiments...

- Traditional physical experiments involve developing an **empirical** model of a response conditional on covariate information and under an assumed error model
  - Exploratory data analysis $\rightarrow$ statistical empirical model $\rightarrow$ designed experiment $\rightarrow$ hypothesis testing $\rightarrow$ new knowledge $\rightarrow$ new experiment $\rightarrow$ ...
  - Classical examples include ANOVA, linear regression "$Y = X\beta$", ...
- Model building: how can we usefully represent the observed process?

From Physical Experiments...

- Traditional physical experiments involve developing an **empirical** model of a response conditional on covariate information and under an assumed error model
  - Exploratory data analysis $\rightarrow$ statistical empirical model $\rightarrow$ designed experiment $\rightarrow$ hypothesis testing $\rightarrow$ new knowledge $\rightarrow$ new experiment $\rightarrow$ ...
  - Classical examples include ANOVA, linear regression "$Y = X\beta$", ...
- Model building: how can we usefully represent the observed process?
- Design: how can we efficiently estimate a model with limited data?

Physical Experiments

- Retrospective, Observational and Designed Experiments

Physical Experiments

- Retrospective, Observational and Designed Experiments
- Designed experiments started with Agricultural Field experiments, later moved to Industrial experiments - e.g. the quality revolution of the 70's, then into Clinical Trials.

Physical Experiments

- Retrospective, Observational and Designed Experiments
- Designed experiments started with Agricultural Field experiments, later moved to Industrial experiments - e.g. the quality revolution of the 70's, then into Clinical Trials.
- Designed experiments account for uncontrollable sources of variation in three ways: replication, randomization and blocking.
    - Replication: prevent measurement error from hiding treatment differences
    - Randomization: prevent unknown nuisance variables from systematically affecting the response in a way that confounds the true relationship between the response and a treatment
    - Blocking: Account for known nuisance variables by creating homogenous groups of experimental units

Limitations of Physical Experiments

- The real world is much more complex than μ_{ij} or μ_{ijk} or μ_{ijkl} or... even $\mu = X\beta$

Limitations of Physical Experiments

- The real world is much more complex than μ_{ij} or μ_{ijk} or μ_{ijkl} or... even $\mu = X\beta$
- Extrapolation is a problem - once we move away from having the support of the data, the predictive ability of statistical models is generally poor to none in all but the simplest of cases

Limitations of Physical Experiments

- The real world is much more complex than μ_{ij} or μ_{ijk} or μ_{ijkl} or... even $\mu = X\beta$
- Extrapolation is a problem - once we move away from having the support of the data, the predictive ability of statistical models is generally poor to none in all but the simplest of cases
- The number of inputs may be too large to efficiently explore

Limitations of Physical Experiments

- The real world is much more complex than μ_{ij} or μ_{ijk} or μ_{ijkl} or... even $\mu = X\beta$
- Extrapolation is a problem - once we move away from having the support of the data, the predictive ability of statistical models is generally poor to none in all but the simplest of cases
- The number of inputs may be too large to efficiently explore
- It is not always ethical to perform a designed experiment - e.g. the risk of cancer from exposure to radiation

Limitations of Physical Experiments

- The real world is much more complex than μ_{ij} or μ_{ijk} or μ_{ijkl} or... even $\mu = X\beta$
- Extrapolation is a problem - once we move away from having the support of the data, the predictive ability of statistical models is generally poor to none in all but the simplest of cases
- The number of inputs may be too large to efficiently explore
- It is not always ethical to perform a designed experiment - e.g. the risk of cancer from exposure to radiation
- It is not always cost-effective to perform a designed experiment - e.g. how a car's design or materials affect its crashworthiness

Limitations of Physical Experiments

- The real world is much more complex than μ_{ij} or μ_{ijk} or μ_{ijkl} or... even $\mu = X\beta$
- Extrapolation is a problem - once we move away from having the support of the data, the predictive ability of statistical models is generally poor to none in all but the simplest of cases
- The number of inputs may be too large to efficiently explore
- It is not always ethical to perform a designed experiment - e.g. the risk of cancer from exposure to radiation
- It is not always cost-effective to perform a designed experiment - e.g. how a car's design or materials affect its crashworthiness
- It is not always possible to perform a designed experiment - e.g. how much will sea-level rise if human CO_2 emissions increase by a prescribed amount over the next 50 years

From Physical Experiments.... to Computer Experiments

- To meet the challenges posed by these new types of experiments, the idea of a Computer Experiment (CE) was recently introduced
  - In a CE, we use a simulation model for the mean behavior of a process given a set of inputs
  - The simulator is **deterministic**; for any input x, the output is **always** $y(x)$.
  - As such, the notions of replication, randomization and blocking do not apply

From Physical Experiments.... to Computer Experiments

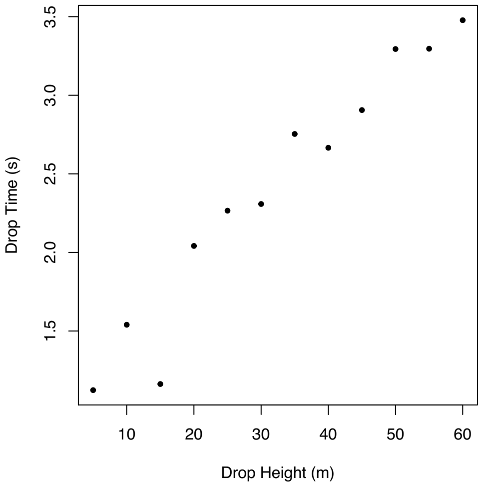
- To meet the challenges posed by these new types of experiments, the idea of a Computer Experiment (CE) was recently introduced
    - In a CE, we use a simulation model for the mean behavior of a process given a set of inputs
    - The simulator is **deterministic**; for any input x, the output is **always** $y(x)$.
    - As such, the notions of replication, randomization and blocking do not apply
- Yet the goals are often the same as Physical Experiments
    - develop a predictive model
    - optimize the response over the inputs
    - screen the inputs that that are most active
    - extract contours of the response surface
    - etc...

From Physical Experiments.... to Computer Experiments

- To meet the challenges posed by these new types of experiments, the idea of a Computer Experiment (CE) was recently introduced
    - In a CE, we use a simulation model for the mean behavior of a process given a set of inputs
    - The simulator is **deterministic**; for any input x, the output is **always** $y(x)$.
    - As such, the notions of replication, randomization and blocking do not apply
- Yet the goals are often the same as Physical Experiments
    - develop a predictive model
    - optimize the response over the inputs
    - screen the inputs that that are most active
    - extract contours of the response surface
    - etc...
- Challenges include computational expense of computer models, large number of inputs in computer models, combining computer models with physical observations, ...

Motivation: Of Bowling Balls and Basket Balls

- A grad student climbs many many stairs to carry a bowling ball to some height h. The bowling ball is dropped and we time how long it takes for it to hit the ground, T_f, in seconds



Motivation: Of Bowling Balls and Basket Balls

- Want a model of drop time so we can predict T_f for other drop heights h

Motivation: Of Bowling Balls and Basket Balls

- Want a model of drop time so we can predict T_f for other drop heights h
- A linear regression seems reasonable as a first try...

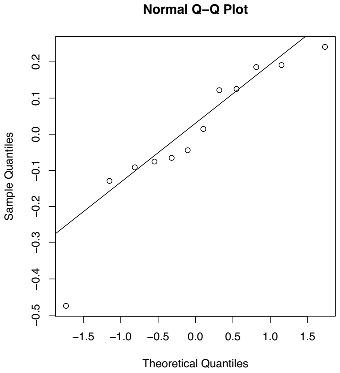
$$T_f = \beta_0 + \beta_1 h + \epsilon$$

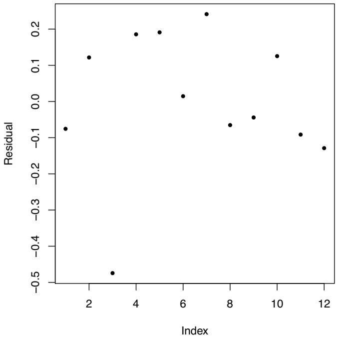
$$\epsilon \sim N(0, \sigma^2)$$

Motivation: Of Bowling Balls and Basket Balls

```
 > summary(fit1)
Call:  lm(formula = Tfobs   h0obs)
Residuals:  Min 1Q Median 3Q Max -0.47434 -0.07954
-0.01480 0.14041 0.24139
Coefficients:  Estimate Std.  Error t value Pr(>|t|)
(Intercept) 0.980347 0.126417 7.755 1.54e-05 ***
h0obs 0.043770 0.003435 12.741 1.66e-07 *** ---
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error:  0.2054 on 10 degrees of
freedom Multiple R-squared:  0.942, Adjusted
R-squared:  0.9362 F-statistic:  162.3 on 1 and 10
DF, p-value:  1.659e-07
```

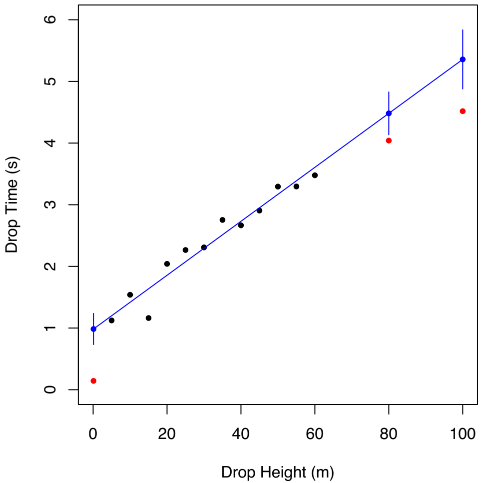
Motivation: Of Bowling Balls and Basket Balls





Motivation: Of Bowling Balls and Basket Balls

- What about prediction? Say we're interested in h=0.1, 80 and 100 meters.



Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals not scattered about zero

Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals not scattered about zero
- residuals still show a trend

Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals not scattered about zero
- residuals still show a trend
- qqplot suggests an outlier

Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals not scattered about zero
- residuals still show a trend
- qqplot suggests an outlier
- dropping from a height $h = 0$ should have a drop time of 0

Motivation: Of Bowling Balls and Basket Balls

Criticisms:

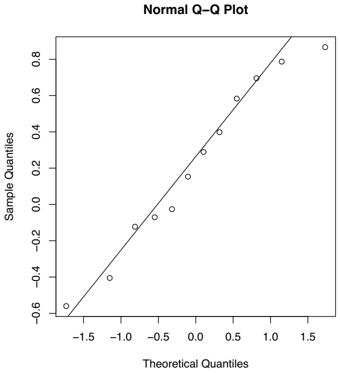
- residuals not scattered about zero
- residuals still show a trend
- qqplot suggests an outlier
- dropping from a height $h = 0$ should have a drop time of 0
- Try:

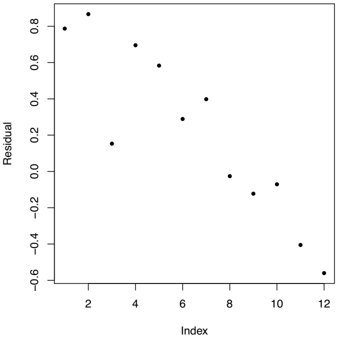
$$T_f = \beta_1 h + \epsilon$$

Motivation: Of Bowling Balls and Basket Balls

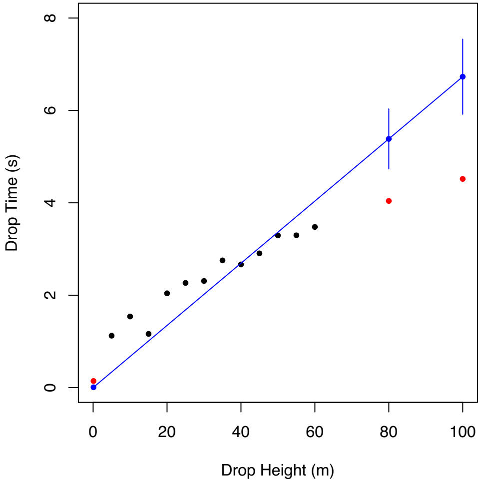
```
 > summary(fit2)
Call:  lm(formula = Tfobs   h0obs - 1)
Residuals:  Min 1Q Median 3Q Max -0.56028 -0.08371
0.22111 0.61128 0.86678
Coefficients:  Estimate Std.  Error t value Pr(>|t|)
h0obs 0.067299 0.004069 16.54 4.06e-09 *** ---
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error:  0.5187 on 11 degrees of
freedom Multiple R-squared:  0.9613, Adjusted
R-squared:  0.9578 F-statistic:  273.6 on 1 and 11
DF, p-value:  4.056e-09
```

Motivation: Of Bowling Balls and Basket Balls





Motivation: Of Bowling Balls and Basket Balls



Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals still show a trend

Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals still show a trend
- qqplot suggests lack of fit in right tail

Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals still show a trend
- qqplot suggests lack of fit in right tail
- dropping from a larger height $h > 60$ suggests increasingly worse fits

Motivation: Of Bowling Balls and Basket Balls

Criticisms:

- residuals still show a trend
- qqplot suggests lack of fit in right tail
- dropping from a larger height $h > 60$ suggests increasingly worse fits
- Try calling a physics dude?

Motivation: Of Bowling Balls and Basket Balls

Motivation: Of Bowling Balls and Basket Balls

- Physics dude says:

$$\frac{d^2 h}{dt^2} = g$$

$$\implies h'(t) = \frac{1}{2} g t + C$$

and

$$h(t) = \frac{1}{2} g t^2 + C t + D$$

Motivation: Of Bowling Balls and Basket Balls

- Physics dude says:

$$\frac{d^2 h}{dt^2} = g$$

$$\Longrightarrow \; h'(t) = \frac{1}{2} g t + C$$

  and

$$h(t) = \frac{1}{2} g t^2 + C t + D$$

- Assuming initial conditions (i.c.) $h(0) = h_0$ and $h'(0) = 0$ gives $C = 0$ and $D = h_0$ The physics-based model is therefore

$$h(t) = \frac{1}{2} g t^2 + h_0$$

Motivation: Of Bowling Balls and Basket Balls

- Physics dude says:

$$\frac{d^2h}{dt^2} = g$$

$$\implies h'(t) = \frac{1}{2}gt + C$$

and

$$h(t) = \frac{1}{2}gt^2 + Ct + D$$

- Assuming initial conditions (i.c.) $h(0) = h_0$ and $h'(0) = 0$ gives $C = 0$ and $D = h_0$ The physics-based model is therefore

$$h(t) = \frac{1}{2}gt^2 + h_0$$

- At $t = T_f$ we know $h(T_f) = 0 = \frac{1}{2}gT_f^2 + h_0$ which gives us

$$Tf = +\sqrt{\frac{-2h_0}{g}} = \sqrt{\frac{-2}{g}}\sqrt{h_0}$$

Motivation: Of Bowling Balls and Basket Balls

- Physics-based model:

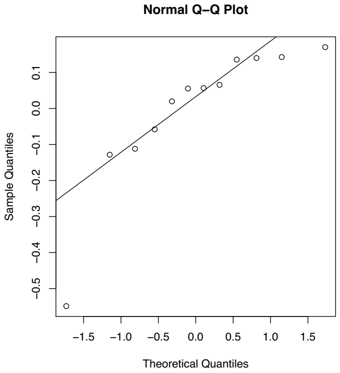
$$T_f = \beta\sqrt{h} + \epsilon$$

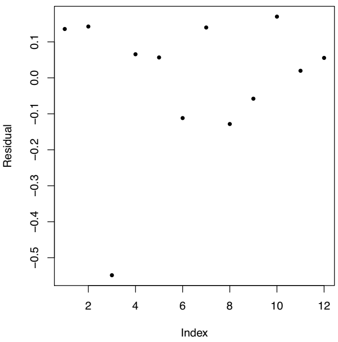
where β is related to gravity by $g = -2/\beta^2$

Motivation: Of Bowling Balls and Basket Balls

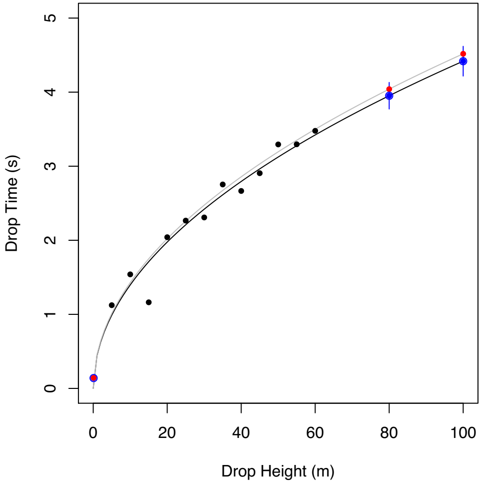
```
 > summary(fit3)
Call:  lm(formula = Tfobs   sqrt(h0obs) - 1)
Residuals:  Min 1Q Median 3Q Max -0.54856 -0.07141
0.05604 0.13679 0.17015
Coefficients:  Estimate Std.  Error t value Pr(>|t|)
sqrt(h0obs) 0.44181 0.01003 44.05 1.01e-13 *** ---
Signif.  codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error:  0.1981 on 11 degrees of
freedom Multiple R-squared:  0.9944, Adjusted
R-squared:  0.9939 F-statistic:  1941 on 1 and 11 DF,
p-value:  1.007e-13
```

Motivation: Of Bowling Balls and Basket Balls





Motivation: Of Bowling Balls and Basket Balls



Motivation: Of Bowling Balls and Basket Balls

- $\beta = 0.44181 + / - 0.01003$ gives $g \in (-9.79268, -10.72766)$

Motivation: Of Bowling Balls and Basket Balls

- $\beta = 0.44181 + / - 0.01003$ gives $g \in (-9.79268, -10.72766)$
- The outlier?
  "D. dropped the ball on his foot, then picked it up and threw it over the platform in anger" i.e. $h'(0) \neq 0$

Motivation: Of Bowling Balls and Basket Balls

- $\beta = 0.44181 + / - 0.01003$ gives $g \in (-9.79268, -10.72766)$
- The outlier?
  "D. dropped the ball on his foot, then picked it up and threw it over the platform in anger" i.e. $h'(0) \neq 0$
- Physics-based model has fewer deficiencies than linear regression models investigated, and likely has more reliable extrapolation ability (it would appear so at least - depends if the physics are right!)

Motivation: Ice Sheet Example

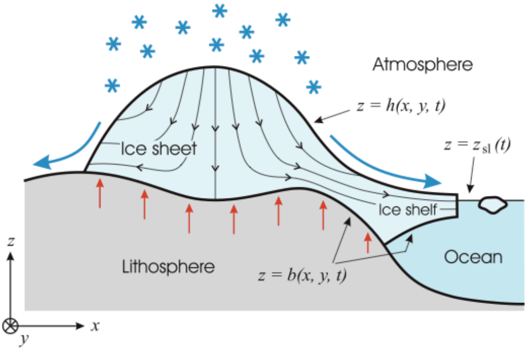
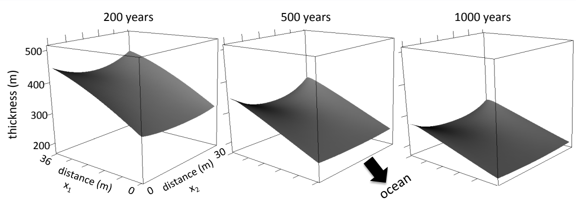
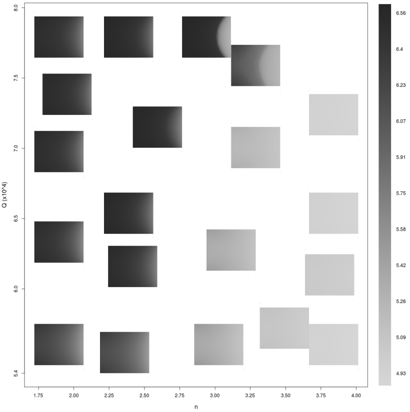


Figure 3.5: Ice-sheet geometry (with attached ice shelf) and Cartesian coordinate system. x and y span the horizontal plane, z is positive upward. $z = h(x, y, t)$ denotes the free surface, $z = b(x, y, t)$ the ice base and $z = z_{\rm sl}(t)$ the mean sea level. Interactions with the atmosphere, the lithosphere and the ocean are indicated. Vertical exaggeration factor $\sim 200 \ldots 500$.

R. Greve, *Dynamics of Ice Sheets and Glaicers*, Course Notes (2004).

Motivation: Ice Sheet Example





Motivation: Ice Sheet Example

- Ice sheet thickness h(x,y,t) evolves over a 2-D spatial domain (x,y) and over time t. Typically (x,y) are indexed at a fixed grid resolution, while snapshots in time can be output at user-specified intervals.

Motivation: Ice Sheet Example

- Ice sheet thickness h(x,y,t) evolves over a 2-D spatial domain (x,y) and over time t. Typically (x,y) are indexed at a fixed grid resolution, while snapshots in time can be output at user-specified intervals.
- Additional parameters include boundary conditions (basal condition b(x,y,t) and ice-ocean interface z(t)) and scalar simulator parameters Q and η. Q is related to thermal conductivity of the ice and η is a parameter involved in ice-sheet stress.

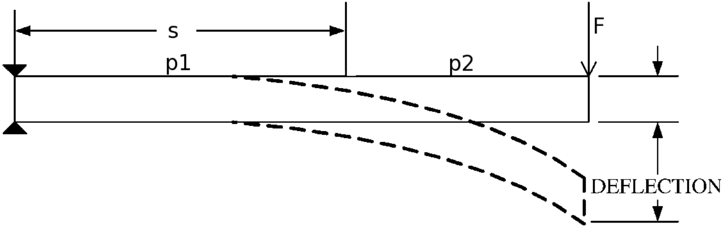
Motivation: Ice Sheet Example

- Ice sheet thickness h(x,y,t) evolves over a 2-D spatial domain (x,y) and over time t. Typically (x,y) are indexed at a fixed grid resolution, while snapshots in time can be output at user-specified intervals.
- Additional parameters include boundary conditions (basal condition b(x,y,t) and ice-ocean interface z(t)) and scalar simulator parameters Q and η. Q is related to thermal conductivity of the ice and η is a parameter involved in ice-sheet stress.
- The simulator is very slow/computationally expensive.

Motivation: Ice Sheet Example

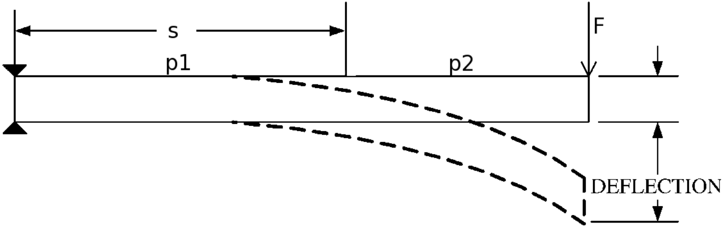
- Ice sheet thickness h(x,y,t) evolves over a 2-D spatial domain (x,y) and over time t. Typically (x,y) are indexed at a fixed grid resolution, while snapshots in time can be output at user-specified intervals.
- Additional parameters include boundary conditions (basal condition b(x,y,t) and ice-ocean interface z(t)) and scalar simulator parameters Q and η. Q is related to thermal conductivity of the ice and η is a parameter involved in ice-sheet stress.
- The simulator is very slow/computationally expensive.
- Scientific interest in $E(h(x, y, t_0 + 100))$ or $P\left(\int_{x,y} h(x, y, t_0 + 100) - h(x, y, t_0) > c\right)$

Motivation: Cantilever Beam Example



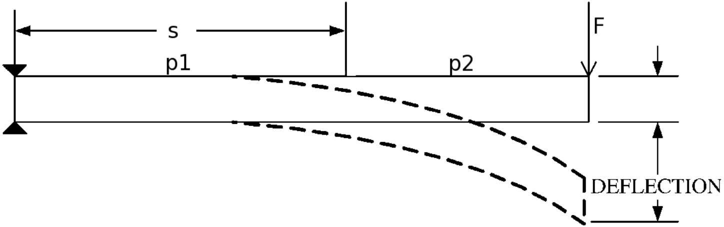
- Beam fixed at one end and free to deflect at other. Constructed from two materials with densities ρ_1, ρ_2 used in fraction s

Motivation: Cantilever Beam Example



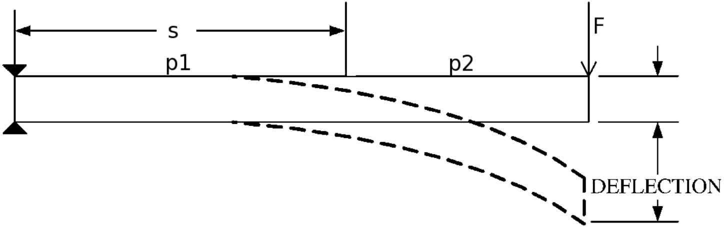
- Beam fixed at one end and free to deflect at other. Constructed from two materials with densities ρ_1, ρ_2 used in fraction s
- A force F is applied to the tip

Motivation: Cantilever Beam Example



- Beam fixed at one end and free to deflect at other. Constructed from two materials with densities ρ_1, ρ_2 used in fraction s
- A force F is applied to the tip
- Simulator calculates $Y(\rho_1, \rho_2, s, F)$, the strain energy of the beam under load using Finite Element Analysis (FEA)

Motivation: Cantilever Beam Example



- Beam fixed at one end and free to deflect at other. Constructed from two materials with densities ρ_1, ρ_2 used in fraction s
- A force F is applied to the tip
- Simulator calculates $Y(\rho_1, \rho_2, s, F)$, the strain energy of the beam under load using Finite Element Analysis (FEA)
- Treat F as a noise variable having normal density $g(\cdot)$ with known mean (10) and variance (2).

Motivation: Cantilever Beam Example

- Run the code at $n = 20$ values of the input variables ρ_1, ρ_2, s, f:

ρ_1	ρ_2	s	f	$Y(\rho_1, \rho_2, s, f)$
0.7	0.7	0.5	-13.90	350.95
0.1	0.1	0.7	-2.26	58.66
0.9	0.9	0.3	-7.24	450.12
0.3	0.5	0.1	-6.10	162.33
0.5	0.3	0.9	-4.91	160.33
0.7	0.9	0.1	-3.65	360.07
0.9	0.3	0.9	-10.55	180.26
0.5	0.5	0.3	-12.76	252.21
0.1	0.1	0.5	-9.45	201.35
0.3	0.7	0.7	-16.35	306.77
0.5	0.7	0.3	-15.09	283.09
0.9	0.5	0.1	-8.35	430.16
0.3	0.3	0.7	-0.62	150.02
0.1	0.1	0.5	-19.38	685.99
0.7	0.9	0.9	-17.74	441.55
0.1	0.7	0.1	-11.65	309.72
0.7	0.3	0.5	-16.35	251.32
0.9	0.1	0.7	-4.91	170.06
0.3	0.9	0.9	-8.35	424.35
0.5	0.5	0.3	-10.55	251.51
0.3	0.3	0.3	-19.38	173.56

Motivation: Cantilever Beam Example

- Interest in minimizing the expected strain energy,

$$\mu(\rho_1, \rho_2, s) = \int Y(\rho_1, \rho_2, s, f) g(f) df$$

over $0 \leq \rho_1, \rho_2, s \leq 1$.

Motivation: CO_2 Plume Example

- Stack emissions from the Four-Corners power plant in New Mexico are simulated in 2-D dependent on 2 emission rate parameters and 2 wind-forcing parameters.

Motivation: CO_2 Plume Example

- Stack emissions from the Four-Corners power plant in New Mexico are simulated in 2-D dependent on 2 emission rate parameters and 2 wind-forcing parameters.
- The simulation model, known as HIGRAD, models the hydrodynamic evolution of the emitted plume over time via the Navier-Stokes equations.

Motivation: CO_2 Plume Example

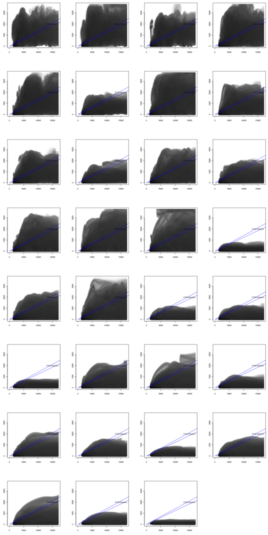
- Stack emissions from the Four-Corners power plant in New Mexico are simulated in 2-D dependent on 2 emission rate parameters and 2 wind-forcing parameters.
- The simulation model, known as HIGRAD, models the hydrodynamic evolution of the emitted plume over time via the Navier-Stokes equations.
- The simulator outputs two responses, one a Eulerian representation of the physical process and one a Lagrangian representation of the physical process.

Motivation: CO_2 Plume Example

- Stack emissions from the Four-Corners power plant in New Mexico are simulated in 2-D dependent on 2 emission rate parameters and 2 wind-forcing parameters.
- The simulation model, known as HIGRAD, models the hydrodynamic evolution of the emitted plume over time via the Navier-Stokes equations.
- The simulator outputs two responses, one a Eulerian representation of the physical process and one a Lagrangian representation of the physical process.
- Goal is to estimate the unknown parameters by solving an inverse problem with limited runs of the simulator e.g. $E[\theta|data]$

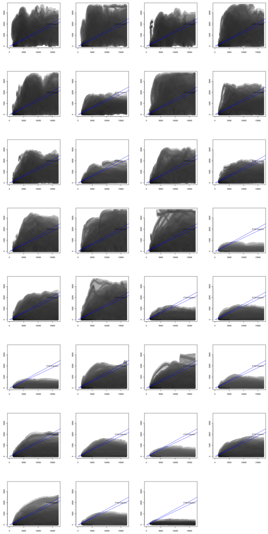
Motivation: CO_2 Plume Example

(Eulerian)



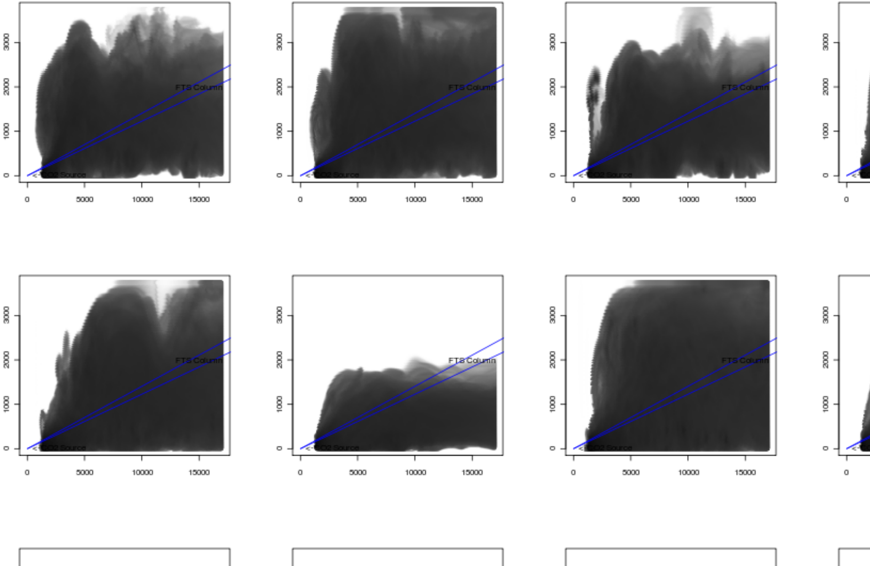
Motivation: CO_2 Plume Example

(Lagrangian)



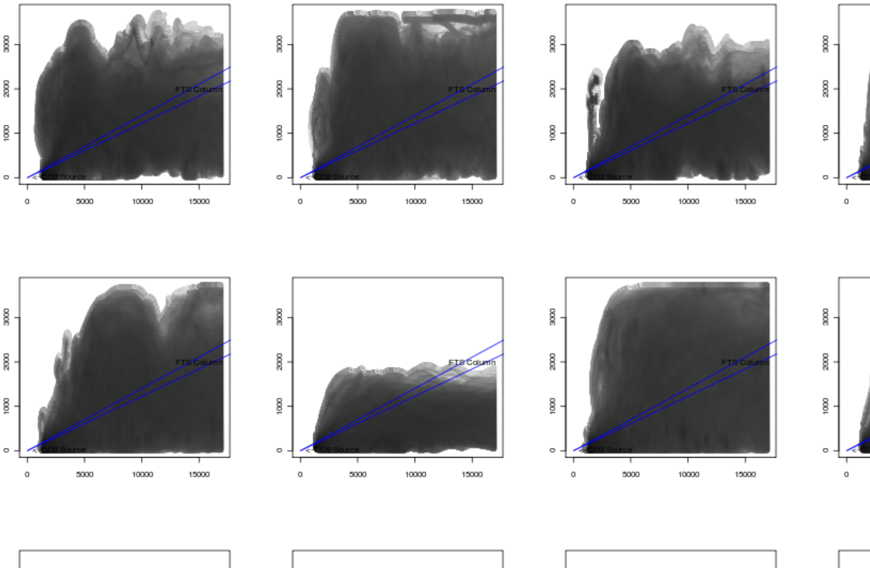
Motivation: CO_2 Plume Example

(Eulerian)



Motivation: CO_2 Plume Example

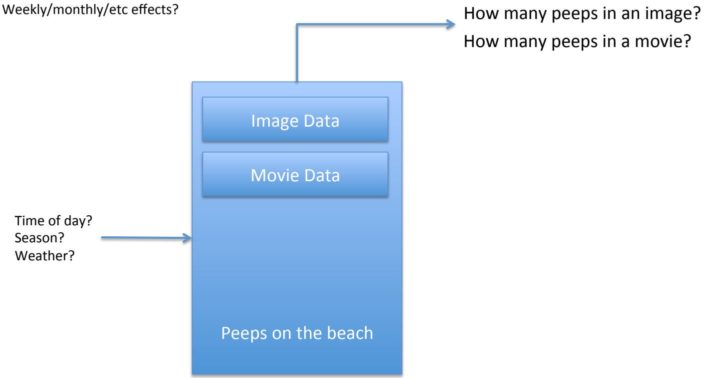
(Lagrangian)



Motivation: Beach Counts

Beachgoers Counting Problem

How many visitors on a beach?
Spatial density of visitors?
Temporal density of visitors?
Weekly/monthly/etc effects?



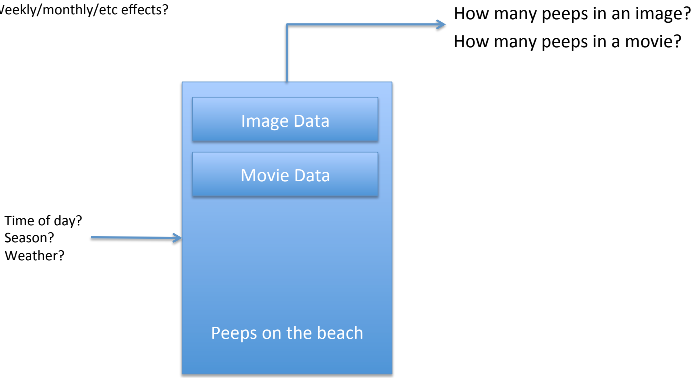
Motivation: Beach Counts

Aerial Movie SanDiego Beach

Motivation: Beach Counts

Beachgoers Counting Problem

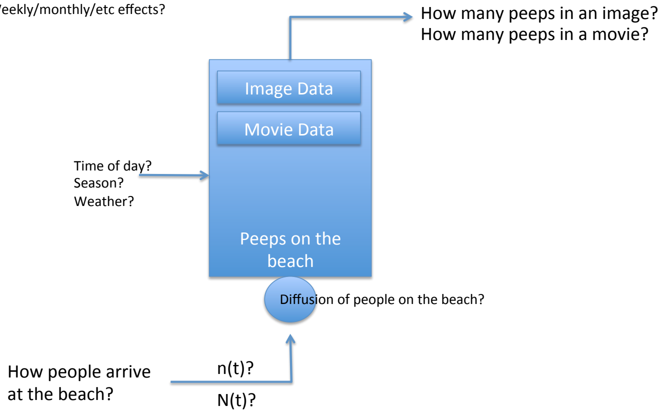
How many visitors on a beach?
Spatial density of visitors?
Temporal density of visitors?
Weekly/monthly/etc effects?

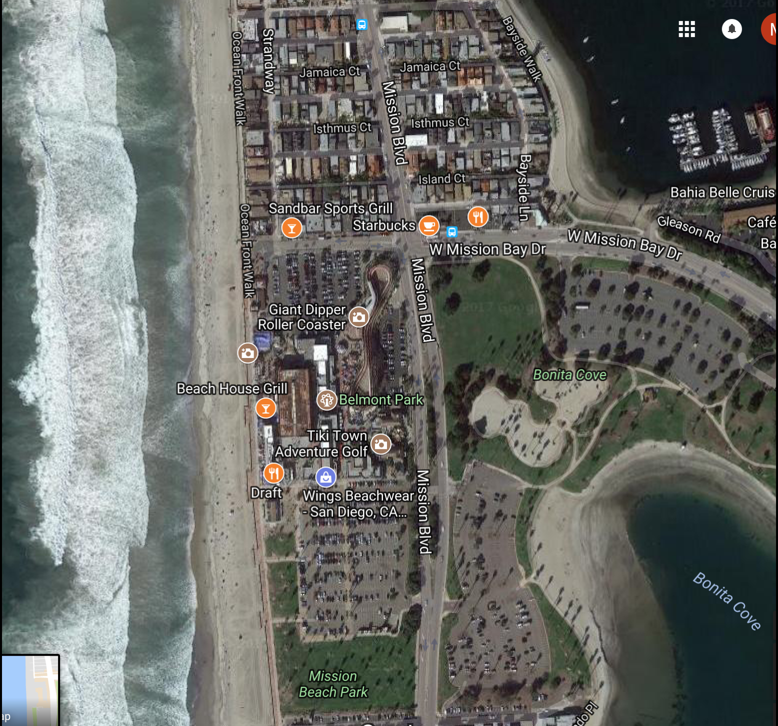


Motivation: Beach Counts

Beachgoers Counting Problem

How many visitors on a beach?
Spatial density of visitors?
Temporal density of visitors?
Weekly/monthly/etc effects?

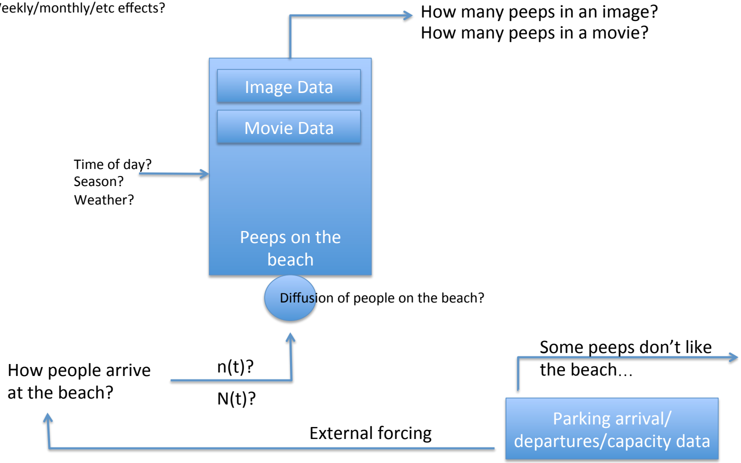



Ocean Front Walk
Strandway
Jamaica Ct
Jamaica Ct
Mission Blvd
Bayside Walk
Isthmus Ct
Isthmus Ct
Island Ct
Bayside Ln
Bahia Belle Cruis
Sandbar Sports Grill
Starbucks
Gleason Rd
Café
W Mission Bay Dr
W Mission Bay Dr
Mission Blvd
Giant Dipper
Roller Coaster
Bonita Cove
Beach House Grill
Belmont Park
Tiki Town
Adventure Golf
Draft
Wings Beachwear
- San Diego, CA...
Mission Blvd
Bonita Cove
Mission
Beach Park

Motivation: Beach Counts

Beachgoers Counting Problem

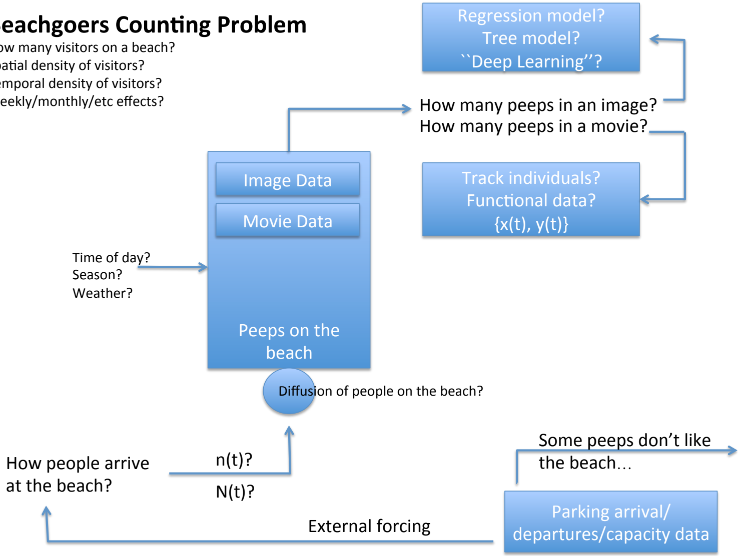
How many visitors on a beach?
Spatial density of visitors?
Temporal density of visitors?
Weekly/monthly/etc effects?



Motivation: Beach Counts

Beachgoers Counting Problem

How many visitors on a beach?
Spatial density of visitors?
Temporal density of visitors?
Weekly/monthly/etc effects?



Motivation: Beach Counts

- Arrival model: $N(t) | \text{Parking, Historical, Covariates}$

Motivation: Beach Counts

- Arrival model: $N(t)|\text{Parking, Historical, Covariates}$
- Beach model: $Y(s,t)|N(t), \eta(t,s), \texttt{Covariates}$ where $\eta(t,s)$ might be a diffusion (differential equation) model, or an agent-based (particle) model

Motivation: Beach Counts

- Arrival model: $N(t)|$Parking, Historical, Covariates
- Beach model: $Y(s,t)|N(t), \eta(t,s), \texttt{Covariates}$ where $\eta(t,s)$ might be a diffusion (differential equation) model, or an agent-based (particle) model
- Data: $Z(s,t)$ is our images or movies

Motivation: Beach Counts

- Arrival model: $N(t)|\text{Parking, Historical, Covariates}$
- Beach model: $Y(s,t)|N(t),\eta(t,s),\texttt{Covariates}$ where $\eta(t,s)$ might be a diffusion (differential equation) model, or an agent-based (particle) model
- Data: $Z(s,t)$ is our images or movies
- Likelihood: $Z(t,s)|Y(s,t),\texttt{covariates}$ (intractible)

Tools

- Many of our motivating problems are obviously incredibly difficult!

Tools

- Many of our motivating problems are obviously incredibly difficult!
- We will learn some tools which, hopefully, will help make at least sub-problems of these overall problems solveable (at least approximately).

Inference Tools

- Say we want to compute MLE's. Sometimes modern data is so huge (n large) that this is computationally difficult.
- Say we want to sample the posterior in a Bayesian setting. Sometimes modern data is so huge (n large) that this is computationally difficult.
- Say we want to sample the posterior in a Bayesian setting. Sometimes the likelihood is not available in closed form, but we can draw realizations from the model represented by the likelihood.

- Markov Chain Monte Carlo (MCMC)
- Gibbs Sampler
- Metropolis-Hastings (MH)
- Approximate Bayesian Computation (ABC)
- Stochastic Gradient Descent (SGD)

Emulation Tools

- Often we have some theoretical motivations about how a real-world process should behave under some, hopefully weak, assumptions.
- Such theoretical models may be implemented as a computer simulator, but it may be computationally expensive to run.
- Often we also have observational data of the process. How can we link these two sources of information together in a thoughtful and scientifically meaningful way?
- Think fancy non-linear regression.

- Gaussian Processes (GP's)
- GP Emulation
- Local approximate Gaussian Processes
- Space-filling designs
- Sensitivity Analysis
- Model Calibration

Statistical Modeling Tools

- Often we have some observational data of some process, but the data is complex, high-dimensional, huge (large *n*).
- Such data may not even fit on a single computer, it may only exist in decentralized databases spread over multiple computers.
- To capture the behavior in the data may require a very flexible model, and we don't know the form of this flexibility a-priori. $X\beta$? Unlikely.

- GP's + Dimension Reduction
- Local approximate Gaussian Processes
- Bayesian Additive Regression Trees (BART)
- BART Heteroscedasticity, Influence, Scalability
- Artificial Neural Networks ("Deep Learning")

Uncertainty Quantification (UQ)

- All the tools outlined exist for dealing with uncertainties.
- Uncertainties in arriving at the posterior distribution in a Bayesian analysis because it is not available in closed form, instead it is approximated by drawing samples.
- Uncertainties in centering our statistical model on a theoretically-motivated simulator because the simulator is expensive to evaluate, so must also be approximated.
- Uncertainties in the statistical model of our real-world observations because the data-generating mechanism is complex, high-dimensional and unknown.

Uncertainty Quantification (UQ)

Wikipedia: *Uncertainty quantification (UQ) is the science of quantitative characterization and reduction of uncertainties in both computational and real world applications. It tries to determine how likely certain outcomes are if some aspects of the system are not exactly known. An example would be to predict the acceleration of a human body in a head-on crash with another car: even if we exactly knew the speed, small differences in the manufacturing of individual cars, how tightly every bolt has been tightened, etc., will lead to different results that can only be predicted in a statistical sense.*

Uncertainty Quantification (UQ)

SIAM Journal on UQ: *The SIAM/ASA Journal on Uncertainty Quantification publishes research articles presenting significant mathematical, statistical, algorithmic, and application advances in uncertainty quantification, defined as the interface of complex modeling of processes and data, especially characterizations of the uncertainties inherent in the use of such models. The journal also focuses on related fields such as sensitivity analysis, model validation, model calibration, data assimilation, and code verification. The journal also solicits papers describing new ideas that could lead to significant progress in methodology for uncertainty quantification as well as review articles on particular aspects. The journal is dedicated to nurturing synergistic interactions between the mathematical, statistical, computational, and applications communities involved in uncertainty quantification and related areas.*

Reading

Sacks, Welch, Mitchell and Wynn: Design and Analysis of Computer Experiments, Statistical Science, Vol 4, pg 409–435 (1989).