2019 Bradley Lecture

after dinner

Variable importance in statistics and real life

Art B. Owen

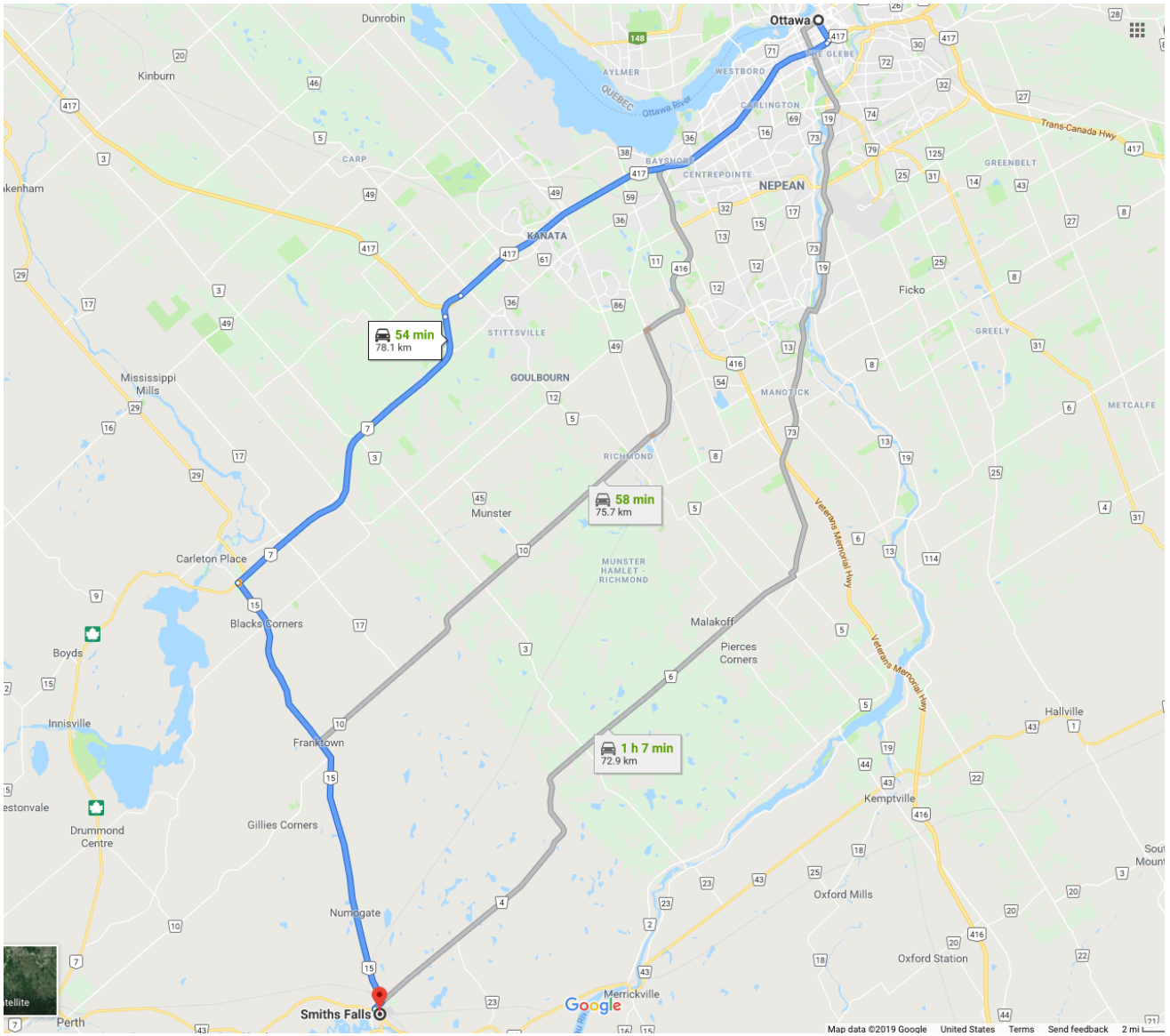
Stanford University

Context: Abhyuday Mandal invited me to the University of Georgia to give the 2019 Bradley lecture. I gave a technical lecture about some recent ideas in importance sampling. They also asked for a non-technical 'after dinner' talk. For that I decided to talk about 'variable importance' fleshing out some ideas I'd been kicking around since a meeting Banff International Research Station in September 2008.

Variable importance comes up a lot in quasi-Monte Carlo, global sensitivity analysis, statistics and machine learning. Since then I've added economics to the list, via Shapley value.

Ralph Bradley was born in Smiths Falls. That's not far from Ottawa where I grew up.

Ralph Bradley and I



Variable importance

- What is an important variable?
- Can we define it?
- Can we measure it?

Comes up in: statistics, economics, machine learning, engineering · · ·

Real world examples

- which lifestyle factors influence health?
- what affects loan repayment, ad clicking, appearing in court?
- what drives fuel efficiency of an airplane?
- or future arctic climate?

Warning: some toy examples and nerdy comments ahead.

Also some serious comments and hints of math.

Areas where it comes up

- predictive models
- experimental design
- algorithmic fairness
- uncertainty quantification
- causal inference

Technical terms

See if you spot some of these:

- p-value, correlation
- derivative, total variation
- Shapley value, Sobol' indices
- superset importance
- Breiman's importance measures
- interaction
- added variable plot

or fit them in.

References

If this becomes an article there will be many more references.

Notably

Wei, Lu & Song (2015)

"Variable importance analysis: a comprehensive review"

has 197 references focussed on "Uncertainty Quantification"

And Iooss & Lemaitre (2015)

Also

Causal inference literature

Algorithmic fairness literature

What is important?

A is important if changing A changes B

write $A \to B$

Assuming

that B is important.

But

why is B important?

Because

$$B \to C$$

Also

$$C \to D \to E \to F \to \cdots \to Z$$

Then what?

Maybe

$$Z \to A$$

That doesn't seem like enough.

Just assume that something is already important.

OR at least, important to you.

Upshot

Importance is transferred, not created.

In statistical models

We use these three notions of importance:

1) Changing A changes Z **causally** (in real life)

   (this is the one we want)

2) Changing A changes **predicted** Z

   (all we changed was our own minds)

3) Modeling without A changes how well we can predict Z.

   E.g., R^2

Of course they are different.

Notion 2 $\not\Rightarrow$ 1

Correlation is not causation

Meehl 1960s: everything is correlated with everything in real data.

Notion 3 $\not\Rightarrow$ 1

My model does not use age, so how could it do age discrimination?

It does notice that Jones has been using Cobol for 30 years.

Importance

Suppose that

$$(A, B, C, D, E) \to Z$$

then we want to compare their importance levels.

To start

First we have to quantify importance of one thing.

For $A \to Z$, ***how much*** is A changing Z?

Then

Where are B, C, D, E when A changes?

Toy model

Health = 200 + 10 * Heredity - Age + 10 * Exercise - 30 * Smoking + 2 * Kale

$\cdots$ + 0.001 * Hot-sauce + Random

Hot sauce doesn't look that important.

But take it away and they stop eating the kale.

The formula misses a causal connection

If we don't measure Age

maybe it is effectively predicted by the $\cdots$ variables

and so we can predict Health almost as well as before.

Usual examples

replace numbers by β_j and variable names by x_j.

Basic example

$$Y = A + B$$

Is A or B more important? They look the same.

By the way

$$0 \leqslant A \leqslant 1 \qquad \text{and} \qquad 0 \leqslant B \leqslant 1000$$

Upshot

Not just the formula but the variable range matters.

Also the distribution in that range.

Next example

$$Y = A \times B$$

A is more important because when it is 0 you get nothing.

Then B must also be more important!

Eg.: people may claim soft skills are most important because

they amplify / multiply hard skills.

With ranges

$$0 \leqslant A \leqslant 1 \qquad \text{and} \qquad 100 \leqslant B \leqslant 200$$

Now A really does look more important.

B has no chance to zero it out.

If A and B are connected somehow, this can change.

Interaction

Sick	Immunized	Not immunized
Exposed to measles	2%	40%
Not exposed	0%	0%

(Hypothetical numbers.)

Either the row variable or the column variable could be more important.

- Most people are exposed (top row) $\Longrightarrow$ measles is about immunization
- Most people are not immunized (second column) $\Longrightarrow$ measles is about exposure

Current outbreaks

Largely about whether the people around you are immunized (herd immunity).

Defining importance

For some function f

$$y = f(a, b, \ldots, x)$$

Importance of x on y

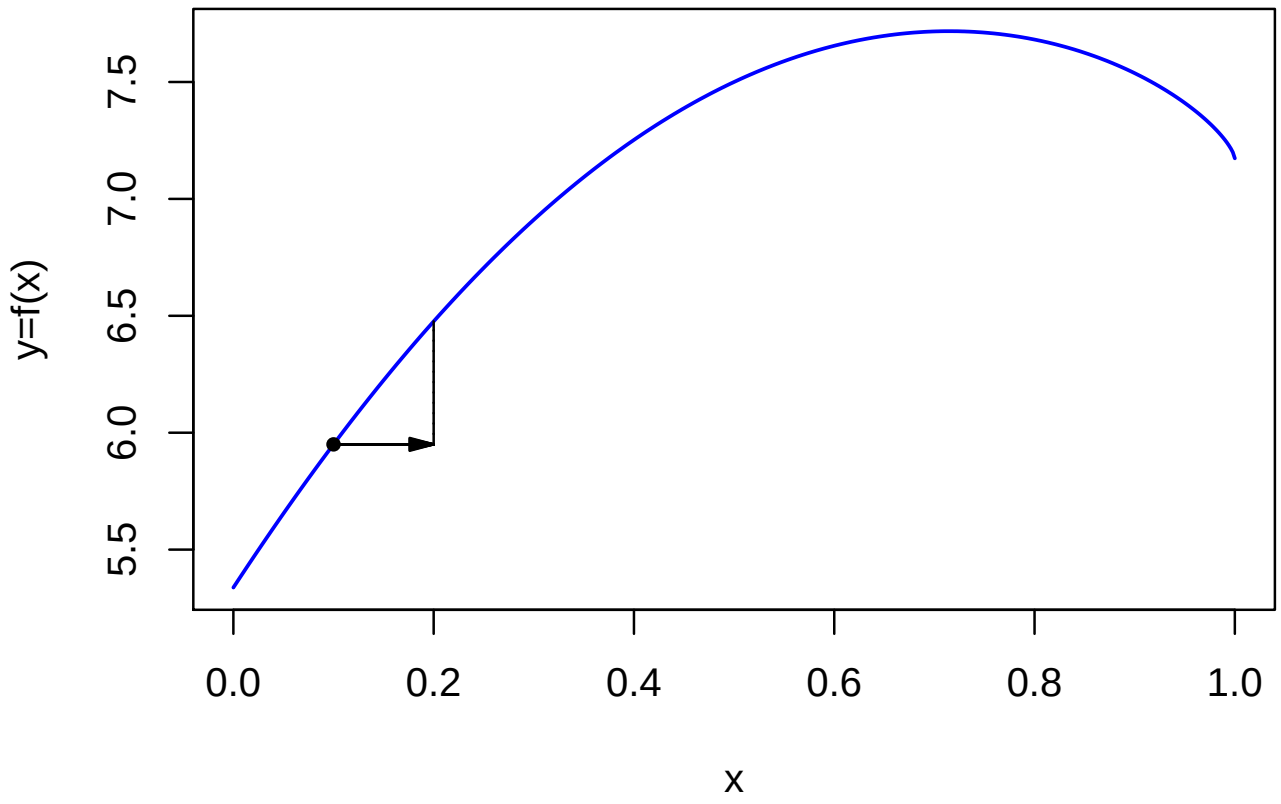
We change x and watch y respond.

1) Which value(s) of x do we ***start*** with?
2) What value(s) do we change it ***to***?
3) How to we ***score*** the change(s) in y?
4) Where are all the ***other*** variables while this is going on?

Lots of choices

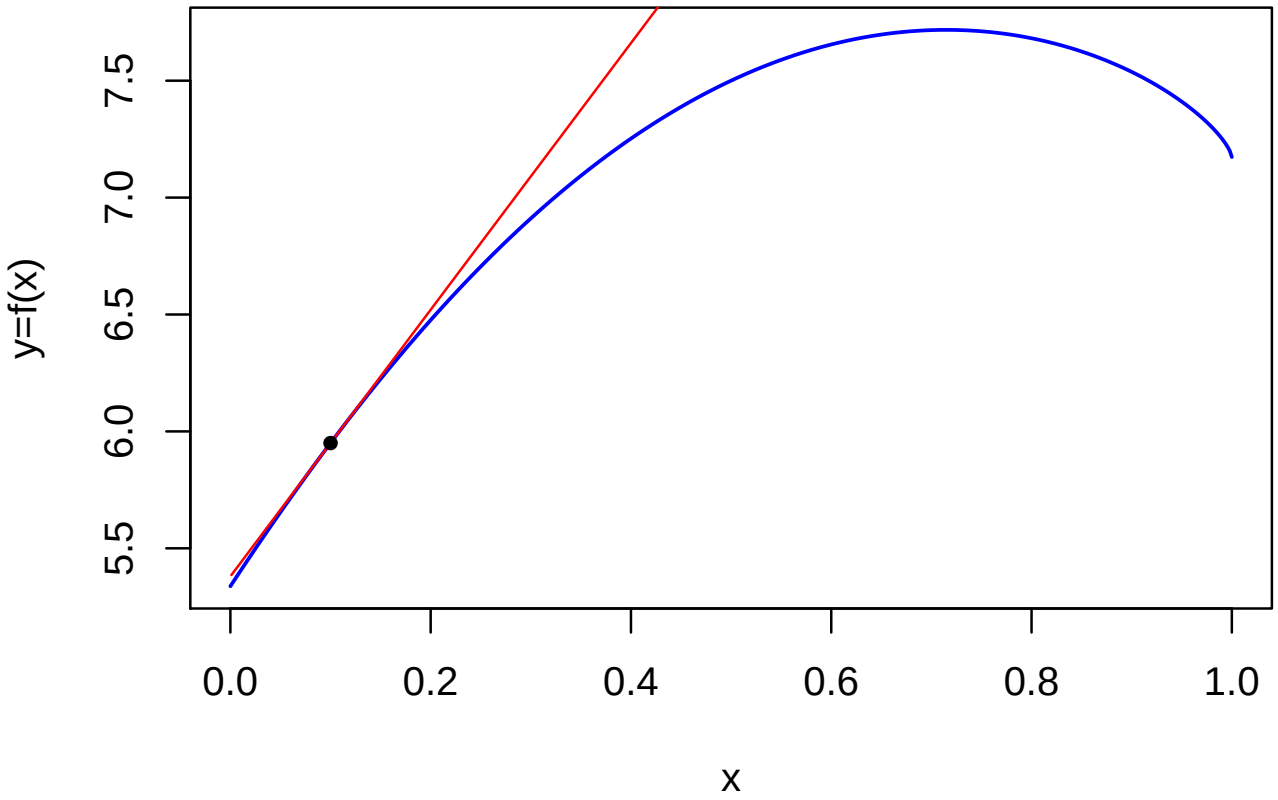
Let's not enumerate them all.

One start point & one end point



Get the vertical distance.

Tiny changes

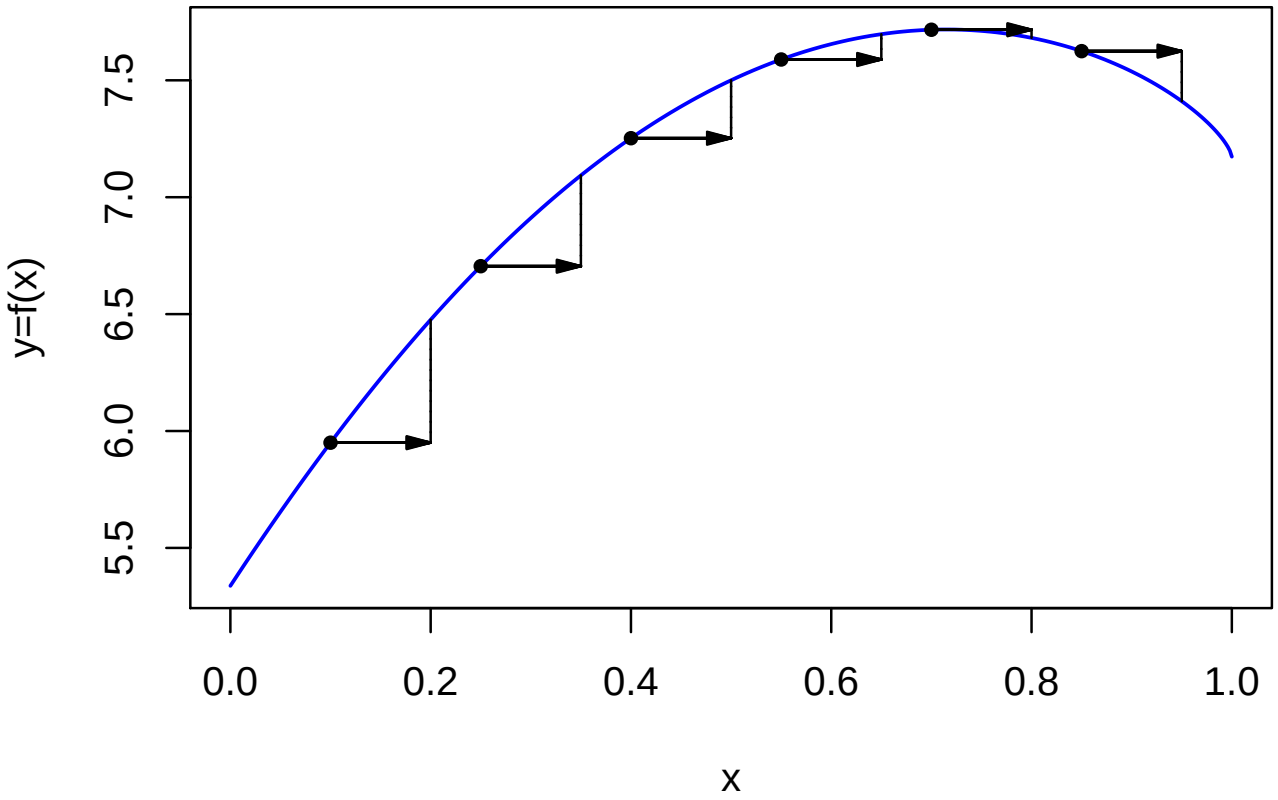


Get a slope.

Known as “sensitivity analysis”

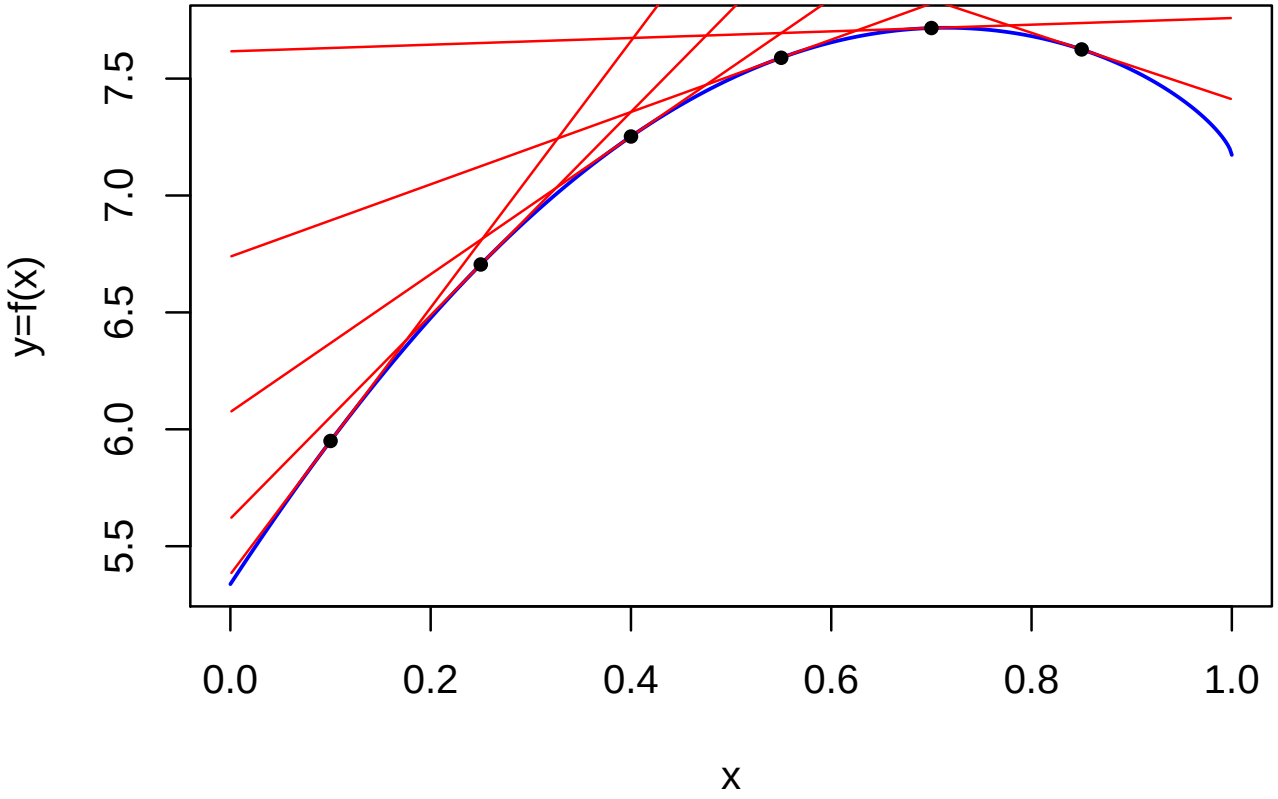
Describes small changes from nominal.

Lots of changes



We could average them.

Lots of tiny changes



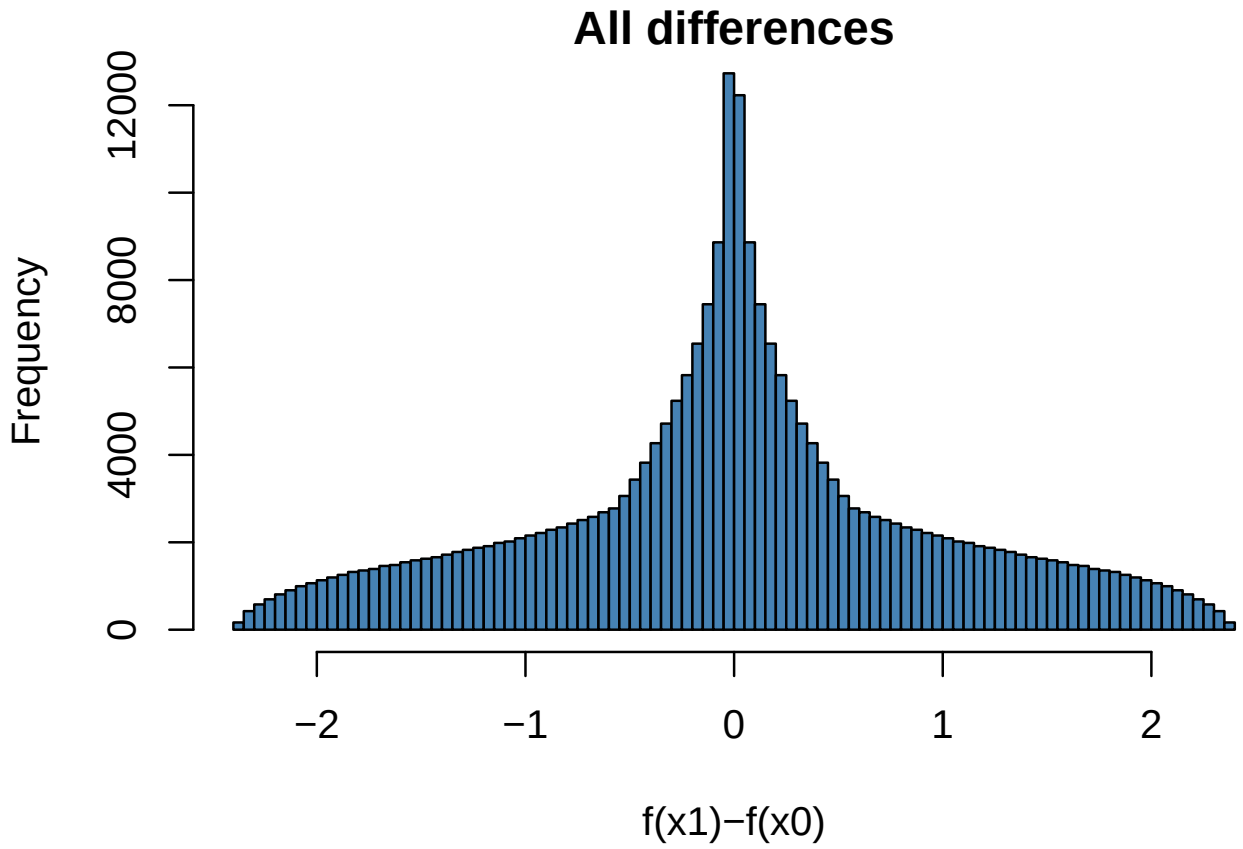
Averaging them gets total variation.

$$\int |f'(x)| \, \mathrm{d}x$$

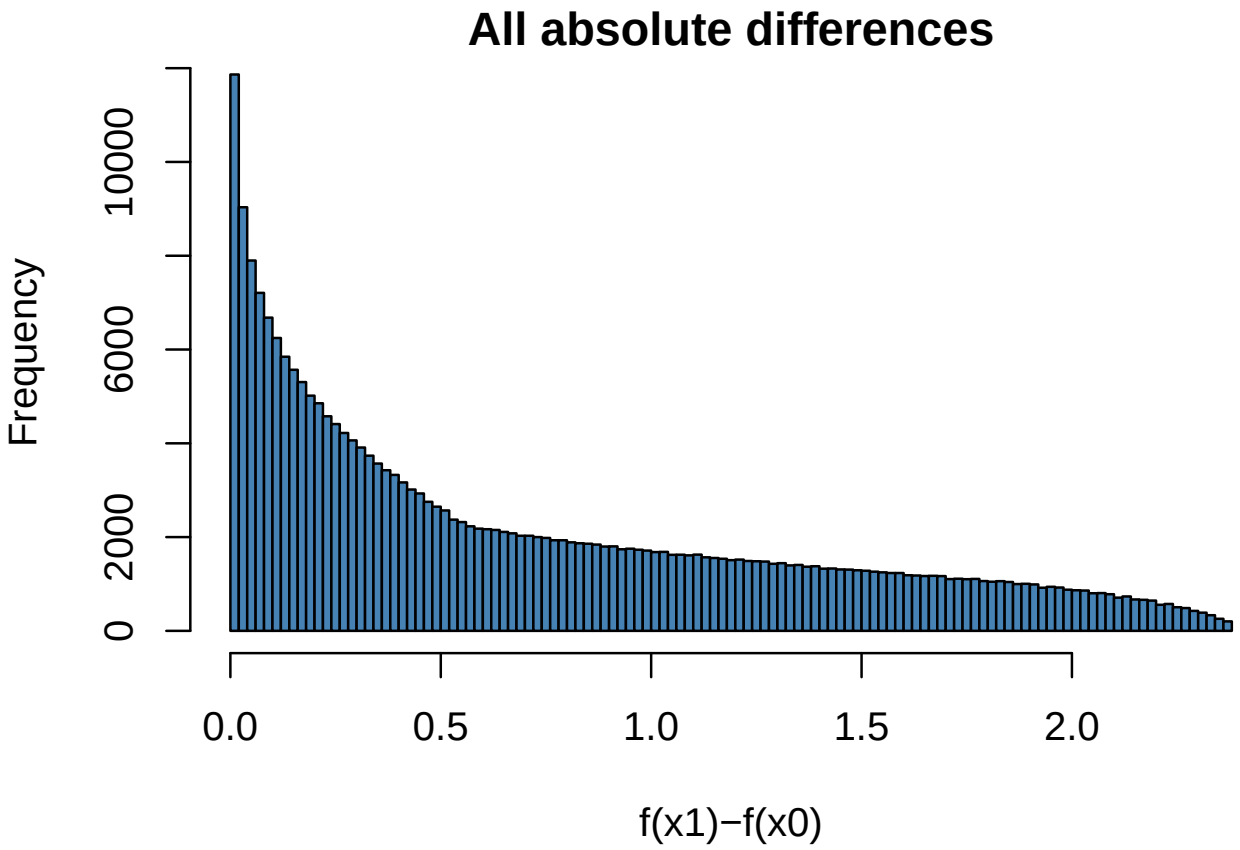
All changes

From everywhere **to** everywhere

500 points in $0 \leqslant x \leqslant 1$



Global sensitivity



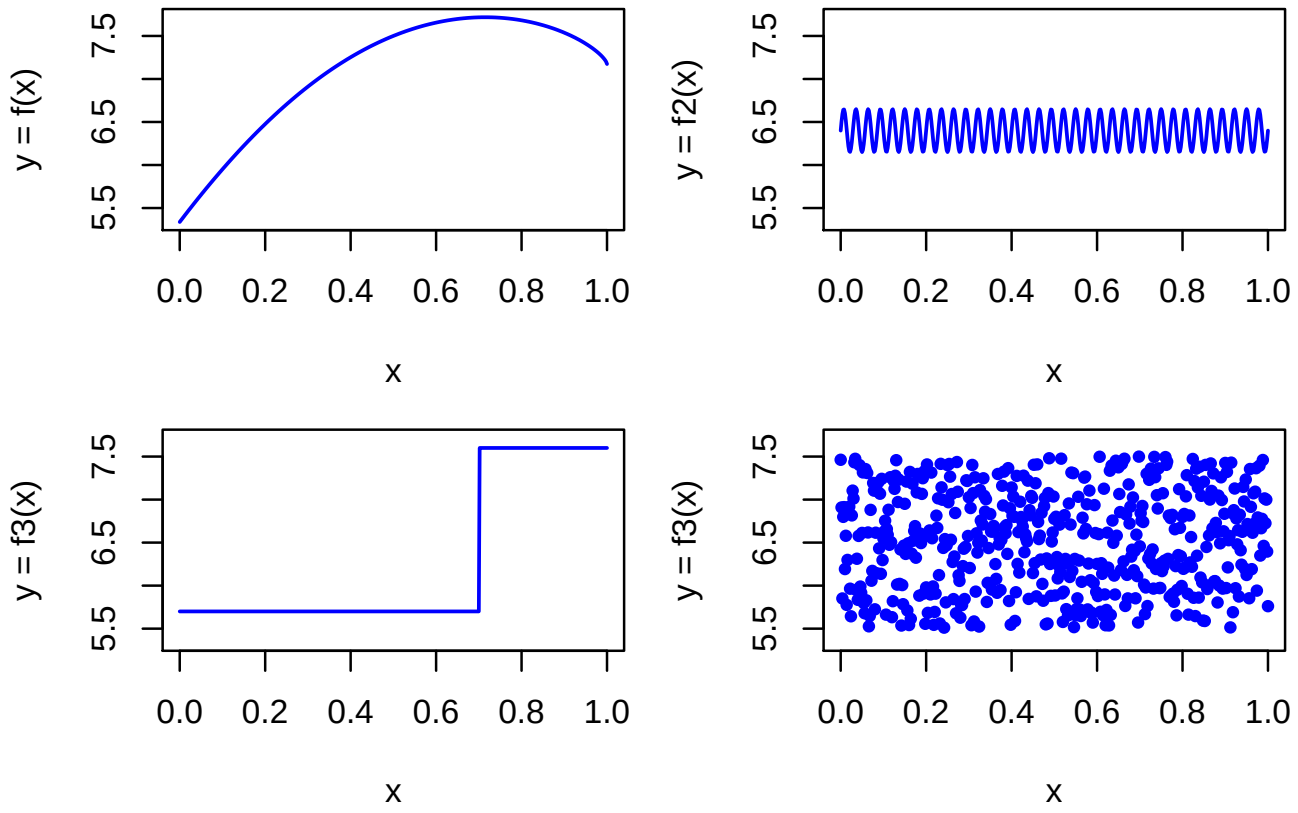
We could take the mean.

Or mean square

From variation to variance

Other choices.

When is x is most influential?



Depends on how you want to keep score,

· · · which depends on your goals.

There's not just one x

For $(a, b, c, \ldots, w, x) \rightarrow y$ where are a, b, $\ldots$, w while we change x?

They could be local or ***global***.

Define:

$$\text{not-x} = (a, b, c, \ldots, w)$$

$$y = f(x, \text{not-x})$$

x could be very important at some values of not-x and less important at others

Global sensitivity analysis books

Fang, Li & Sudijanto (2010), Saltelli, Chan & Scott (2009), Saltelli, Ratto & Andres (2008), Cacuci, Ionescu-Bujor & Navon (2005), Saltelli, Tarantola & Campolongo (2004), Santner, Williams & Notz (2003)

and there are many more articles.

Saint-Venant flood model

Lamboni, Iooss, Popelin, Gamboa, (2012)

Overflow in meters at a dyke

$$S = Z_v + H - H_d - C_b, \quad \text{where}$$

$$H = \left(\frac{Q}{BK_s\sqrt{(Z_m - Z_v)/L}} \right)^{3/5} \qquad \text{(max annual river height)}$$

Q	Maximal annual flow	m^3/s	Gumbel$(1013, 558) \cap [500, 3000]$
K_s	Strickler coefficient	$m^{1/3}/s$	$\mathcal{N}(30, 8) \cap [15, \infty)$
Z_v	River downstream level	m	Triangle$(49, 50, 51)$
Z_m	River upstream level	m	Triangle$(54, 55, 56)$
H_d	Dyke height	m	$\mathbf{U}[7, 9]$
C_b	Bank level	m	Triangle$(55, 55.5, 56)$
L	Length of river stretch	m	Triangle$(4990, 5000, 5010)$
B	River width	m	Triangle$(295, 300, 305)$

Reduced from a Navier-Stokes model; Usually we don't see a formula.

One at a time

Morris (1991)

Tiny increases: x to $x + \epsilon$

Average

$$f(x, \text{not-x}) - f(x + \epsilon, \text{not-x})$$

over all x and all "not-x"

Compolongo, Cariboni, Saltelli (2007)

Average $\left|f(x, \text{not-x}) - f(x + \epsilon, \text{not-x})\right|$ over x and not-x

Sobol' indices

Sobol' (1993) changes x at random. Let x_{old} and x_{new} be two values.

Closed index

$$\text{Average} \quad \frac{1}{2}\Big(f(x_{\text{old}}, \text{not-x}) - f(x_{\text{new}}, \text{not-x})\Big)^2$$

over all x_{old} all x_{new} and all not-x

Pick-freeze

Pick two values for x: $x_{\text{old}}, x_{\text{new}}$

Freeze "not-x" $= (a, b, c, \ldots, w)$

Technicalities

The inputs are assumed independent

People usually normalize by the variance

It estimates an ANOVA main effect of x.

Importance of multiple variables

Freeze: $a, b, c, \cdots, u, v$

Pick: $(w_{\rm old}, x_{\rm old})$ and $(w_{\rm new}, x_{\rm new})$

Average $\displaystyle \frac{1}{2}\Big(f(a,\ldots,v,\boldsymbol{w_{\rm old}},\boldsymbol{x_{\rm old}}) - f(a,\ldots,v,\boldsymbol{w_{\rm new}},\boldsymbol{x_{\rm new}})\Big)^2$

Technicalities

This counts w and x and their interaction.

Total index

His total index includes that and everything that interacts with x.

Small total index $\Longrightarrow$ unimportant.

Sobol' total indices

Percent	Q	K_s	Z_v	Z_m	H_d	C_b	L	B
Height H	0.72	0.29	0.0078	0.0077	0	0	7.4×10^{-7}	0.00021
Overflow S	0.35	0.14	0.19	0.0038	0.28	0.036	3.6×10^{-7}	0.00010
Cost C_p	0.48	0.25	0.23	0.0077	0.17	0.039	6.8×10^{-7}	0.00019

Variables

Q	Maximal annual flow	m^3/s	Gumbel$(1013, 558) \cap [500, 3000]$
K_s	Strickler coefficient	$m^{1/3}/s$	$\mathcal{N}(30, 8) \cap [15, \infty)$
Z_v	River downstream level	m	Triangle$(49, 50, 51)$
Z_m	River upstream level	m	Triangle$(54, 55, 56)$
H_d	Dyke height	m	$\mathbf{U}[7, 9]$
C_b	Bank level	m	Triangle$(55, 55.5, 56)$
L	Length of river stretch	m	Triangle$(4990, 5000, 5010)$
B	River width	m	Triangle$(295, 300, 305)$

Breiman's permutation

Random forests: Breiman (2001)

f is a prediction fit to n data points

He replaces those n by randomly reordered values.

old w	old x	new w	new x
w_1	x_1	w_1	x_3
w_2	x_2	w_2	x_1
w_3	x_3	w_3	x_5
w_4	x_4	w_4	x_4
w_5	x_5	w_5	x_1

So it is very much like a Sobol' index.

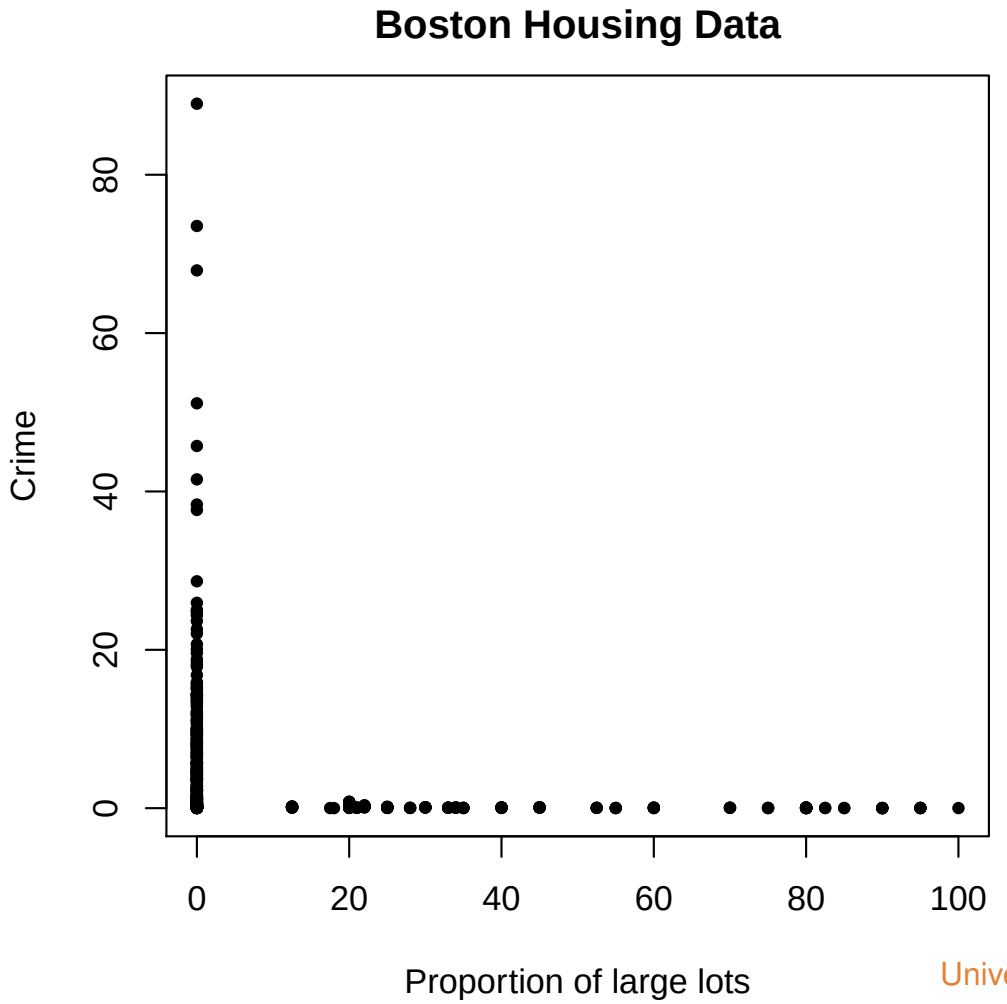
He keeps track of what changes when x gets shuffled:

old f vs new f

Boston housing data

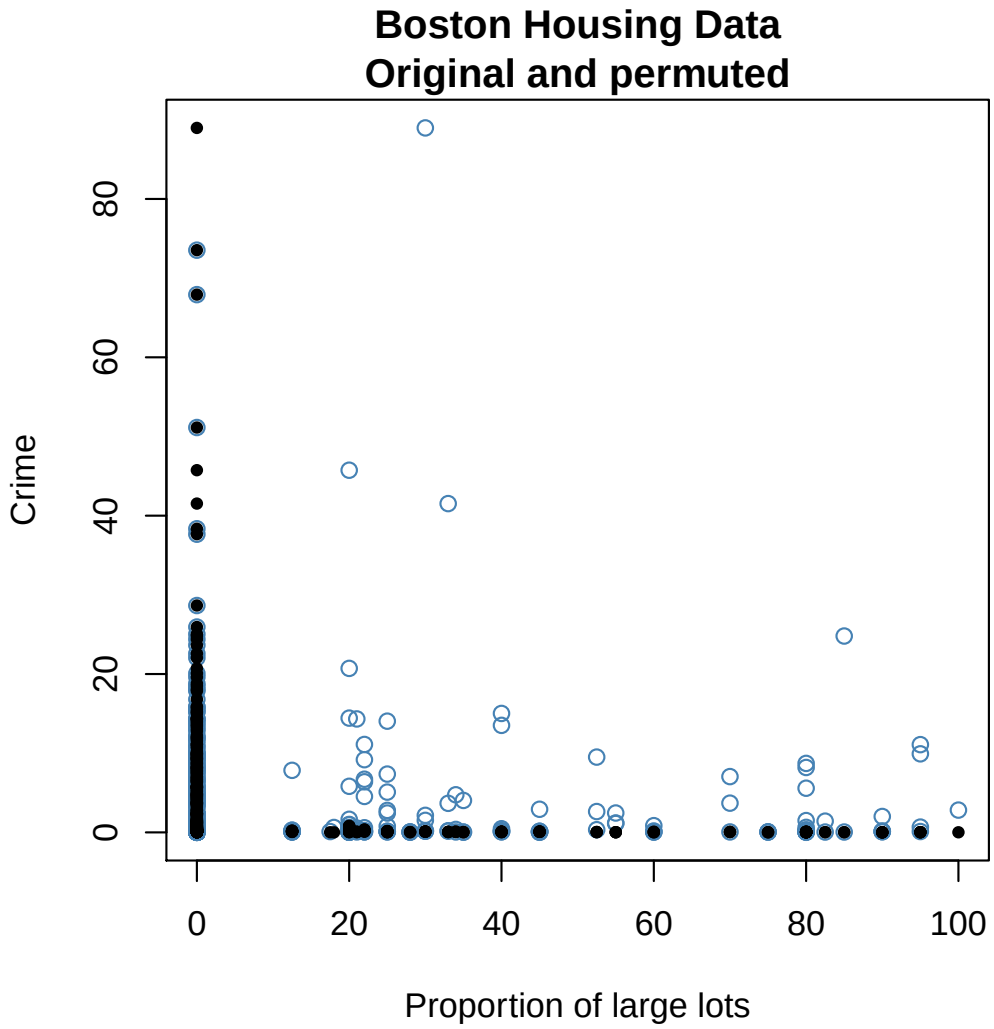
Predict median housing value: 506 regions and 13 predictors

Harrison & Rubinfeld (1978)



Boston housing data

With shuffled values. Too many improper combinations. Not truly Bostonian.



Hooker (2007) ameliorates using a dependent data ANOVA from Stone (1994).

From economics

How to attribute a reward among multiple causes or team members.

Solved by Shapley (1953)

$15 million

Shapley's (1953) value measures contributions of team members.

We need to know what each subset of the team would have accomplished.

Example from Bank of International Settlement

Team	Output value
$\varnothing$	0
A	4,000,000
B	4,000,000
C	4,000,000
A,B	9,000,000
A,C	10,000,000
B,C	11,000,000
A,B,C	15,000,000

Q: How should we split the $15,000,000 earned by A, B, C among them?

$15 million

Example from Bank of International Settlement

Team	Output value
$\varnothing$	0
A	4,000,000
B	4,000,000
C	4,000,000
A,B	9,000,000
A,C	10,000,000
B,C	11,000,000
A,B,C	15,000,000

Q: How should we split the $15,000,000 earned by A, B, C among them?

A: Shapley (1953) says: A gets $4,500,000, B gets $5,000,000, C gets $5,500,000

Shapley setup

Let team $u \subseteq \mathcal{D} \equiv \{1, 2, \ldots, d\}$ create value **val**(u).

Total value is **val**$(\mathcal{D})$.

We attribute ϕ_j of this to $j \in \mathcal{D}$.

Shapley axioms

Efficiency	$\sum_{j=1}^{d} \phi_j =$ **val**$(\mathcal{D})$
Dummy	If **val**$(u \cup \{i\}) =$ **val**(u), all u then $\phi_i = 0$
Symmetry	If **val**$(u \cup \{i\}) =$ **val**$(u \cup \{j\})$, all $u \cap \{i, j\} = \varnothing$ then $\phi_i = \phi_j$
Additivity	If games **val**, **val**$'$ have values ϕ, ϕ' then **val** + **val**$'$ has value $\phi_j + \phi'_j$

There is one and only one fair way to share it.

For variable importance

Let variables $x_1, x_2, \ldots, x_d$ be team members trying to explain f.

The value of any subset u is how much can be explained by $\boldsymbol{x}_u$.

This leads to a (non-causal) variable importance meaure.

For linear models

Lendeman, Merenda & Gold (1980) use it on R^2

For nonlinear models

See O (2014), Song, Nelson & Staum (2016), O & Prieur (2017)

Solves the conceptual problem about **Zone** vs **Crime** for Boston housing.

Hard computations.

More directions

Variable importance in regression:

Gromping (2007)

Using derivatives to bound Sobol' indices:

Sobol' & Kucherenko (2010)

among others

More about Breiman's measure:

Bouleisteix and co-authors.

What if we care mostly about extreme moves?

O, Dick, Chen (2014)

Changes in changes

Interactions, Vitali variation, mixed partial derivatives, differences in differences

Conclusions

1) Importance is transferred, not created
2) It is about changing inputs and outputs
3) Change **from** someplace(s) **to** someplace(s)
4) Track what the other inputs are doing
5) Score / combine the output changes

Thanks

- University of Georgia statistics
- Abhyuday Mandal
- Nikki Rowden
- NSF: DMS-1521145, DMS-1407397, DMS-0906056, DMS-0604939