Overcoming p-hacking and confirmation bias

Tony Cox

What can science do for us?

- Use data to test/challenge/replace assumptions and preconceptions, correct mistakes and…
- Discover how reality works
  - Reveal unexpected findings
  - Explain, attribute given assumptions
  - Predict
  - Discovery is not necessarily done best via hypothesis testing
- Discover and validate causal laws that enable trustworthy predictions, plausible explanations, effects estimates given assumptions

How can science harm us?

- Guaranteed false positives
  - P-hacking, modeling assumptions
- False confidence and arrogance
  - "Merchants of certainty"
  - Overconfidence, confirmation bias
- Prematurely shut down discussions
- Scientism: Looking to science for answers to non-science questions

How to get from data to causal predictions… objectively?

- Causal prediction
  - Deterministic causal prediction: Doing X will make Y happen to people of type Z
  - Probabilistic causal prediction: Doing X will change conditional probability distribution of Y, given covariates Z
    - Goal: Manipulative causation (vs. associational, counterfactual, predictive, computational, etc.)
- Data: Observed (X, Y, Z) values
- Challenge: How will changing X change Y?

Informed decisions require causal predictions

- How would cutting exposure concentration C in half affect future response rate R?

Community	Concentration , C	Income, I	Response rate, R
A	4	100	8
B	8	60	16
C	12	20	24

Informed decisions require causal predictions

- How would cutting exposure concentration C in half affect future response rate R?
  - $10M reward if answer is “Cutting C reduces R”

Community	Concentration , C	Income, I	Response rate, R
A	4	100	8
B	8	60	16
C	12	20	24

Informed decisions require causal predictions

- How would cutting exposure concentration C in half affect future response rate R?

Community	Concentration , C	Income, I	Response rate, R
A	4	100	8
B	8	60	16
C	12	20	24

Model: $R = 2C$

If this is a valid structural equation, then $\Delta R = 2\Delta C$

The corresponding DAG is: $C \rightarrow R$

Model-dependent associations undermine causal predictions from data

- How would cutting exposure concentration C in half affect future response rate R?
  - No way to determine from historical data

Community	Concentration , C	Income, I	Mortality rate, R
A	4	100	8
B	8	60	16
C	12	20	24

Model 1: $R = 2C$, $(I = 140 - 10C)$, DAG: $I \leftarrow C \rightarrow R$, $I \rightarrow C \rightarrow R$

Model 2: $R = 35 - 0.5C - 0.25*I$, DAG: $C \rightarrow R \leftarrow I$

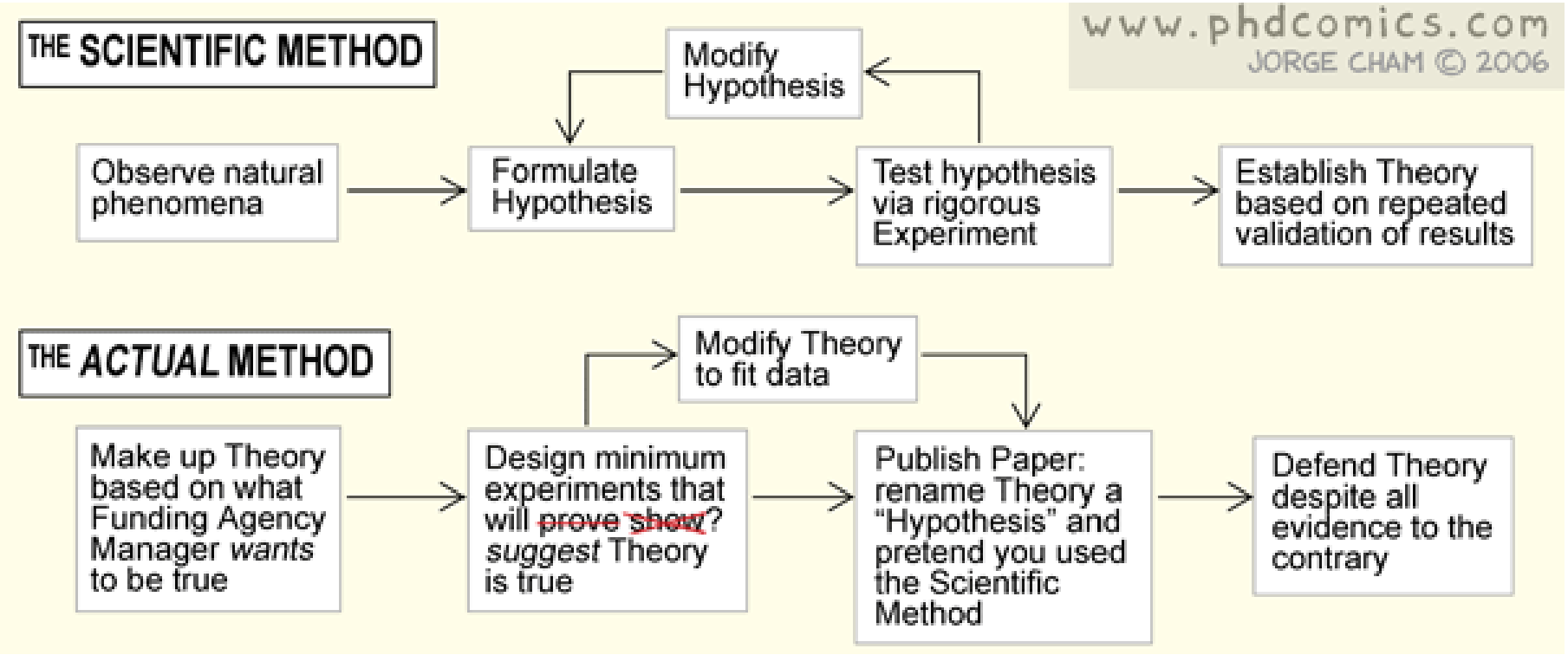
Model 3: $R = 28 - 0.2*I$, $(C = 14 - 0.1*I)$, DAG: $C \leftarrow I \rightarrow R$

So, decreasing C could decrease R, increase it, or leave it unchanged.

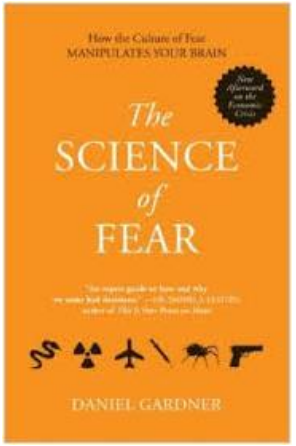
Implications

- Ambiguous associations obscure objective functions, make sound modeling and inference more difficult
  - Conclusions are not purely data-driven
    - *hypothesis* $\rightarrow$ *data* $\rightarrow$ *conclusion*
  - Instead, they conflate data and modeling assumptions
    - *hypothesis/model/assumptions* $\rightarrow$ *conclusions* $\leftarrow$ *data*
  - Undermines sound (objective, trustworthy, well-justified, independently repeatable, verifiable) inference
    - Undermined when conclusions rest on untested assumptions
  - Ambiguous associations are common in practice
- *Wanted:* A way to reach valid, robust (model-independent) conclusions *from data* that can be fully specified before seeing the data.
  - *Solution:* DAG discovery algorithms

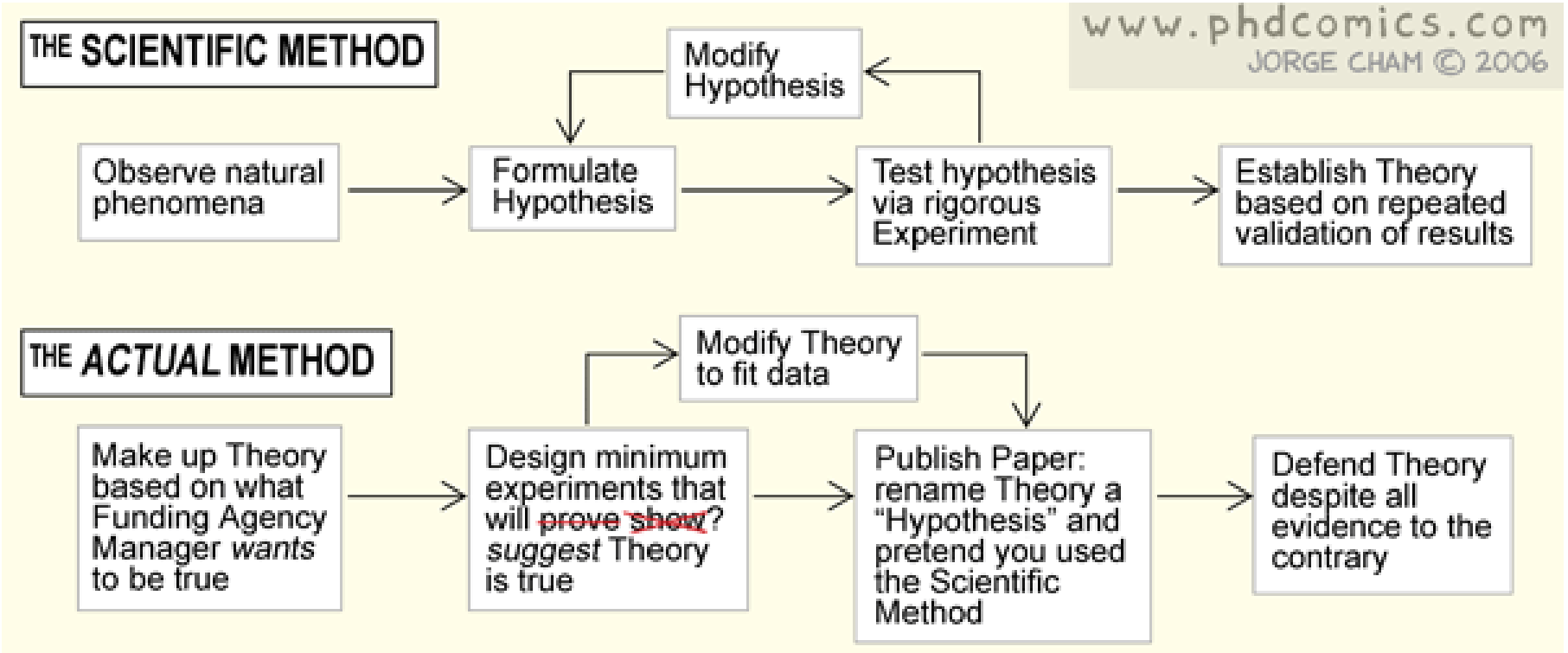
Scientific method: Theory vs. (bad) practice



www.cornwallalliance.org/wp-content/uploads/2014/12/Scientific-Method-vs-Non-Scientific-Method-Cartoon.png



Scientific method: Theory vs. (bad) practice



www.cornwallalliance.org/wp-content/uploads/2014/12/Scientific-Method-vs-Non-Scientific-Method-Cartoon.png

Statistical inference principles for causal discovery algorithms

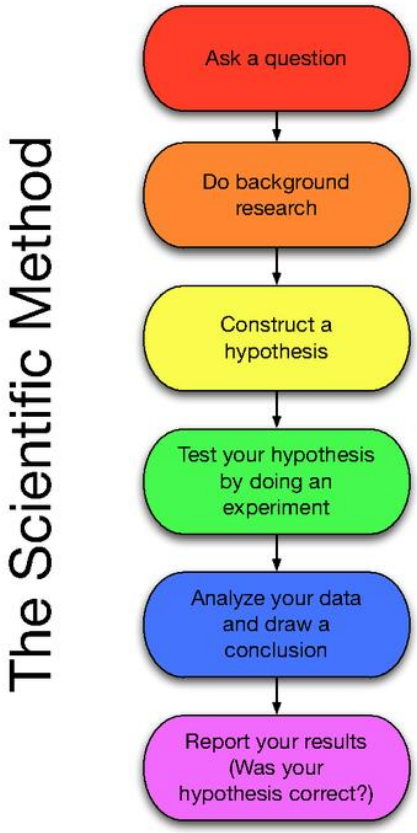
- Probabilistic
- Associational
- Attributive
- Counterfactual
- Structural
- Predictive
- Manipulative
- Mechanistic/explanatory

- Associational/attributive: Regression, RR
- Predictive
  - Conditional independence tests, $X \rightarrow Y \rightarrow Z$
  - Granger tests, transfer entropy
- Manipulative
  - Randomized control trial (RTC)
  - Generalization/transportability
- Mechanistic
  - Invariant laws (CPTs), well-behaved errors
  - Composition of effects

Principles for identifying causal DAGs from data are implemented in many R packages

- Conditional independence (constraint-based algorithms)
  - *dagitty*, *bnlearn* packages
- Likelihood principle (score-based algorithms)
  - Choose DAG model to maximize likelihood of data
  - Included among the algorithms in *bnlearn* package
- Composition principle: If $X \rightarrow Y \rightarrow Z$, then $dz/dx = (dz/dy)*(dy/dx)$
- Granger/transfer entropy principle: Predictively useful information flows from causes to their effects over time
  - Transfer entropy, Yin & Yao, 2016, www.nature.com/articles/srep29192
- Model error specification principle
  - effect = f(cause) + error
  - LiNGAM software, https://arxiv.org/ftp/arxiv/papers/1408/1408.2038.pdf
- Homogeneity/invariance principles for causal CPTs
  - Li et al., 2015, https://pdfs.semanticscholar.org/a051/9a2c6b85ca65d0df037142f550cf87d4e43f.pdf
  - Peters et al., 2015, *InvariantCausalPrediction* package http://stat.ethz.ch/~nicolai/invariant.pdf

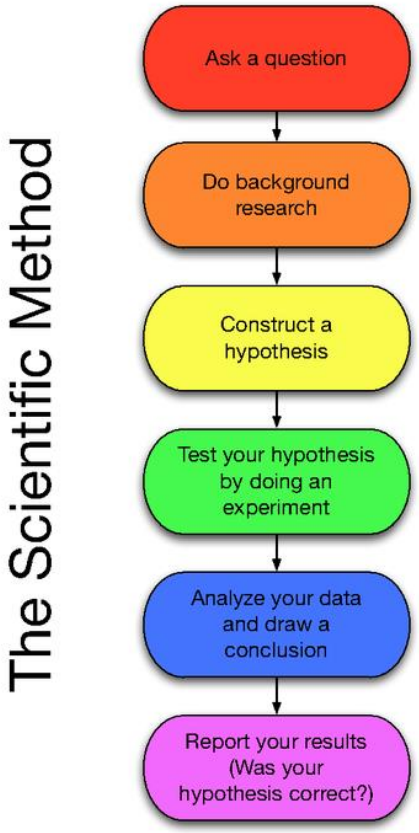
Example of Hypothetico-Deductive Scientific Method



- Observation: The sun rises and sets
- Question: Why?

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

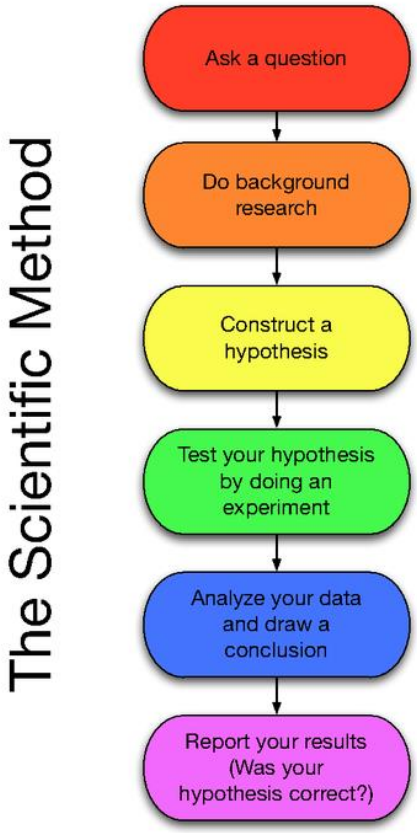
Example of Hypothetico-Deductive Scientific Method



- Observation: The sun rises and sets
- Question: Why?
- Alternative explanatory hypotheses:
  - Hypothesis A: Earth rotates (null hypoth)
  - Hypothesis B: Sun revolves about fixed earth (alternative hypothesis)

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

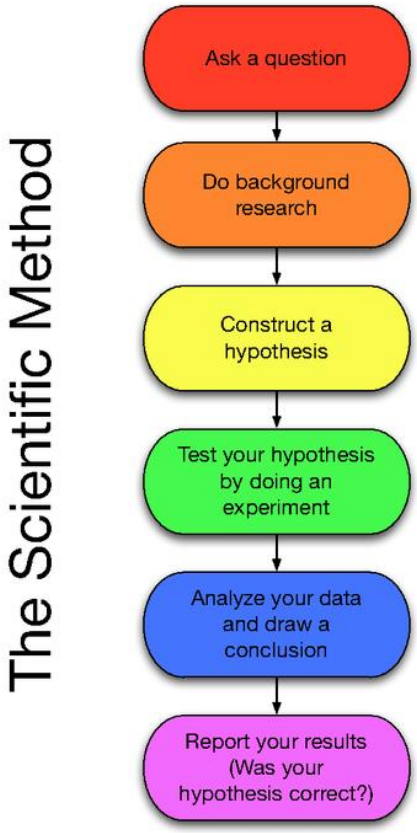
Example of Hypothetico-Deductive Scientific Method



- Observation: The sun rises and sets
- Question: Why?
- Alternative explanatory hypotheses:
  - Hypothesis A: Earth rotates (null hypoth)
  - Hypothesis B: Sun revolves about fixed earth (alternative hypothesis)
- Deduce testable implications
  - A implies strong wind, centrifugal force
  - B implies no strong wind, no centrifugal force

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

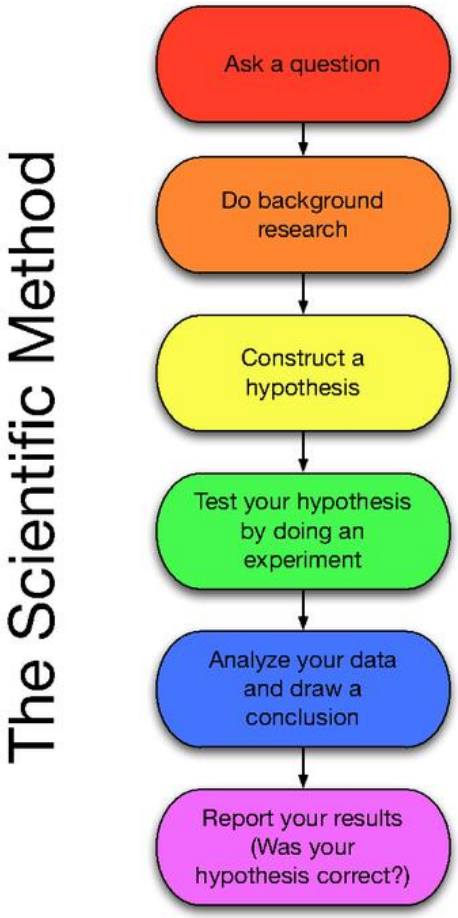
Example of Hypothetico-Deductive Scientific Method



- Alternative explanatory hypotheses:
  - A: Earth rotates (null hypoth)
  - B: Sun revolves about fixed earth
- Deduce testable implications
  - A implies strong wind, centrifugal force
  - B implies no strong wind, no centrifugal force
- Test hypotheses with data / observations
  - No strong wind, no centrifugal force observed
- Draw conclusion
  - Hypothesis B is consistent with reproducible observations
  - Hypothesis A is not

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

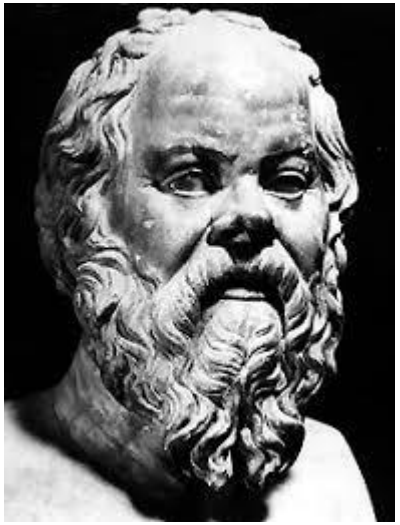
Hypothetico-deductive method makes *assumption-dependent* deductions



https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

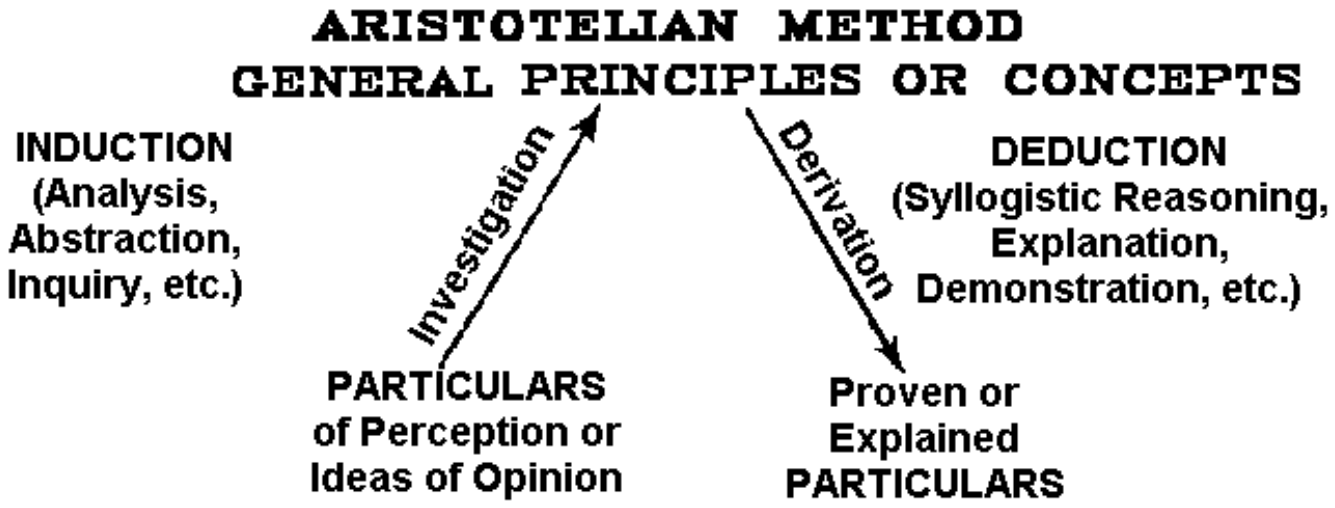
Scientific inference requires ability to reason soundly

- A: All men are mortal
- B: Socrates is mortal
- C: All men are Socrates



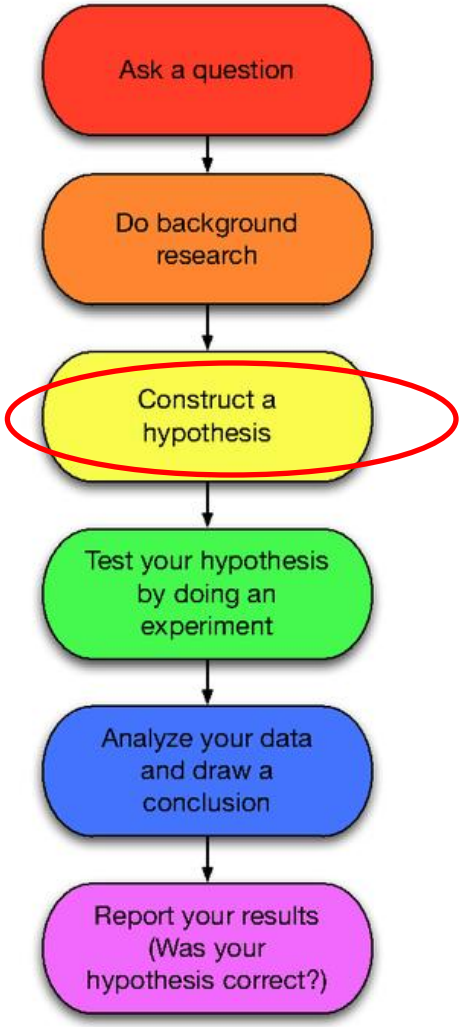
Scientific inference also requires ability to generalize… and no simple solution exists

- Correct *generalization* from specific data is a major challenge ("Problem of Induction")
  - Achilles heel of randomized control trials
  - Causal explanation is the cure, but not easy



http://stillunfold.com/science/top-10-contributions-of-aristotle-to-science

Let's try it!



- Examine the data on next slide
- Formulate a hypothesis
- Goal is to predict the value of the outcome for the next case (case 7)

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

Hypothesize a model and use it to predict Outcome for case 7

- Predictors are attributes, e.g., Predictor 1 = sex (0 = M, 1 = F); Predictor 2 = Age (0 = < 65, 1 = ≥ 65); Predictor 3 = income, etc.
- Outcome = Ever diagnosed with heart disease (0 = No, 1 = Yes)

Case	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Outcome
1	1	1	1	1	1
2	0	0	0	0	0
3	0	1	1	0	1
4	1	1	0	0	0
5	0	0	0	0	0
6	1	0	1	1	1
7	1	1	0	1	?

Different models make different predictions for case 7

- Model 1: Outcome = Predictor 3
- Model 2: Outcome = majority(Predictors 2-4)
- Model 3: Outcome = max(Predictors 3-4)

Case	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Outcome
1	1	1	1	1	1
2	0	0	0	0	0
3	0	1	1	0	1
4	1	1	0	0	0
5	0	0	0	0	0
6	1	0	1	1	1
7	1	1	0	1	?

Lesson 1: Best hypothesis is often under-determined by data

- Multiple models explain past data equally well
- But they make very different predictions
- No unique hypothesis is warranted by the data

Case	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Outcome
1	1	1	1	1	1
2	0	0	0	0	0
3	0	1	1	0	1
4	1	1	0	0	0
5	0	0	0	0	0
6	1	0	1	1	1
7	1	1	0	1	?

Lesson 2: Data may support contradictory hypotheses

- Multiple models explain past data equally well: P3 vs. max(P3, P4)
- But they make very different predictions
- No unique hypothesis is warranted by the data

Case	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Outcome
1	1	1	1	1	1
2	0	0	0	0	0
3	0	1	1	0	1
4	1	1	0	0	0
5	0	0	0	0	0
6	1	0	1	1	1
7	1	1	0	1	?

Model-dependence makes associations unreliable guides to causality

- Association often depends on choice of model
- Example: Observationally equivalent models with opposite associations between X and Y
  - Model 1: $Y = 50 + X$ (positive association)
  - Model 2: $Y = 150 – X – Z$, where $Z = 100 – 2X$
  - Choosing what to include on right-hand side of regression model changes size and direction of association between X and Y
  - In practice, model-dependent associations make published inferences about air pollution health effects unreliable in many cases (Dominici et al., 2014)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4206184/

Model dependence and omitted information

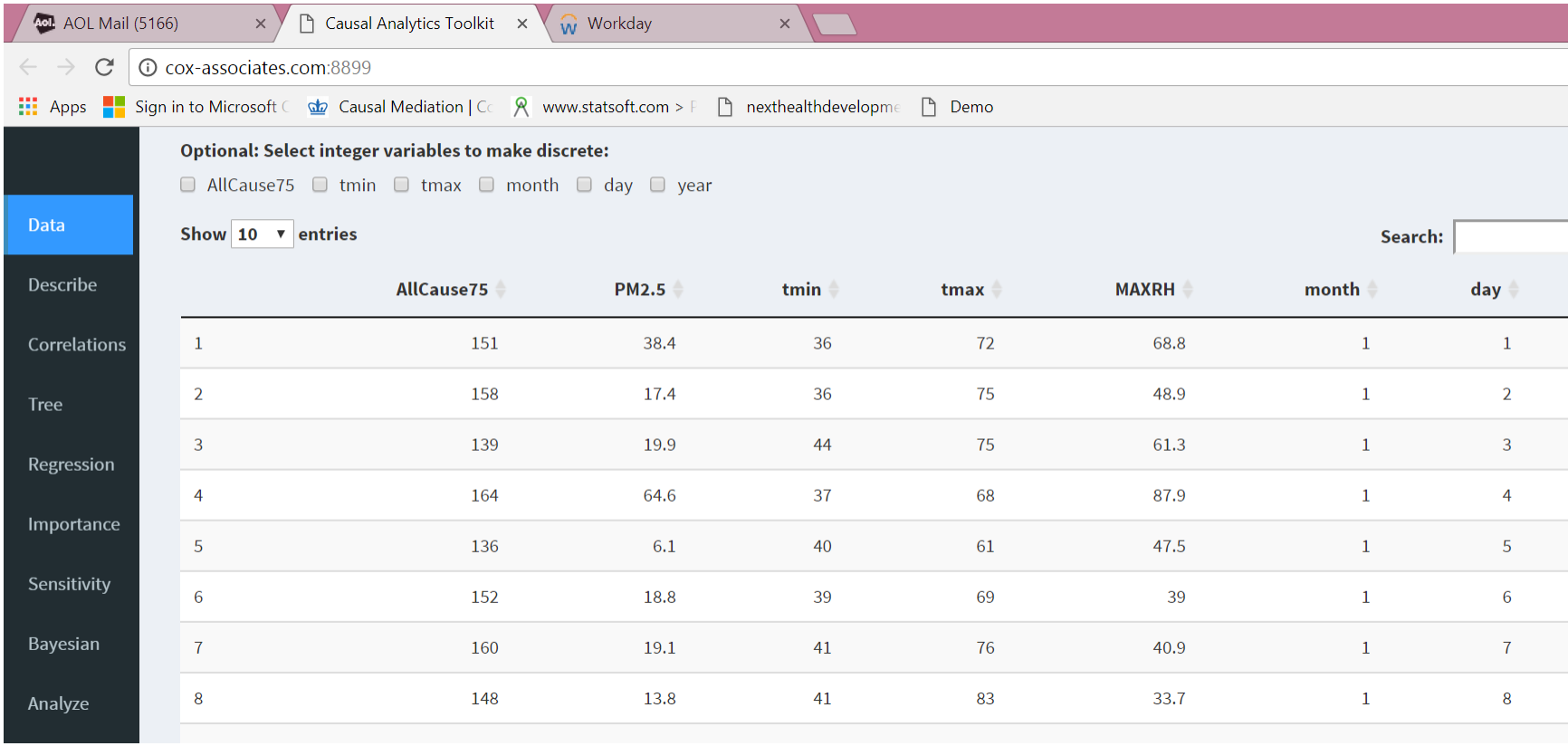
- Model-based measures of association (e.g., regression coefficients, odds ratios in logistic regression models) depend on modeling choices and assumptions
  - What functional form to assume (parametric)
  - Which variables to include on right side
- Changes in modeling choices can change directions, sizes, and statistical significance of associations
- Associational models usually leave out information on changes needed to study causality
  - Focus on association between historical levels of variables
  - This does not tell how future changes in one would change the other(s)
- Automated analysis can help

How to avoid p-hacking and model-dependent conclusions

- Automated (but appropriate) analyses
- Non-parametric methods
- Model ensembles

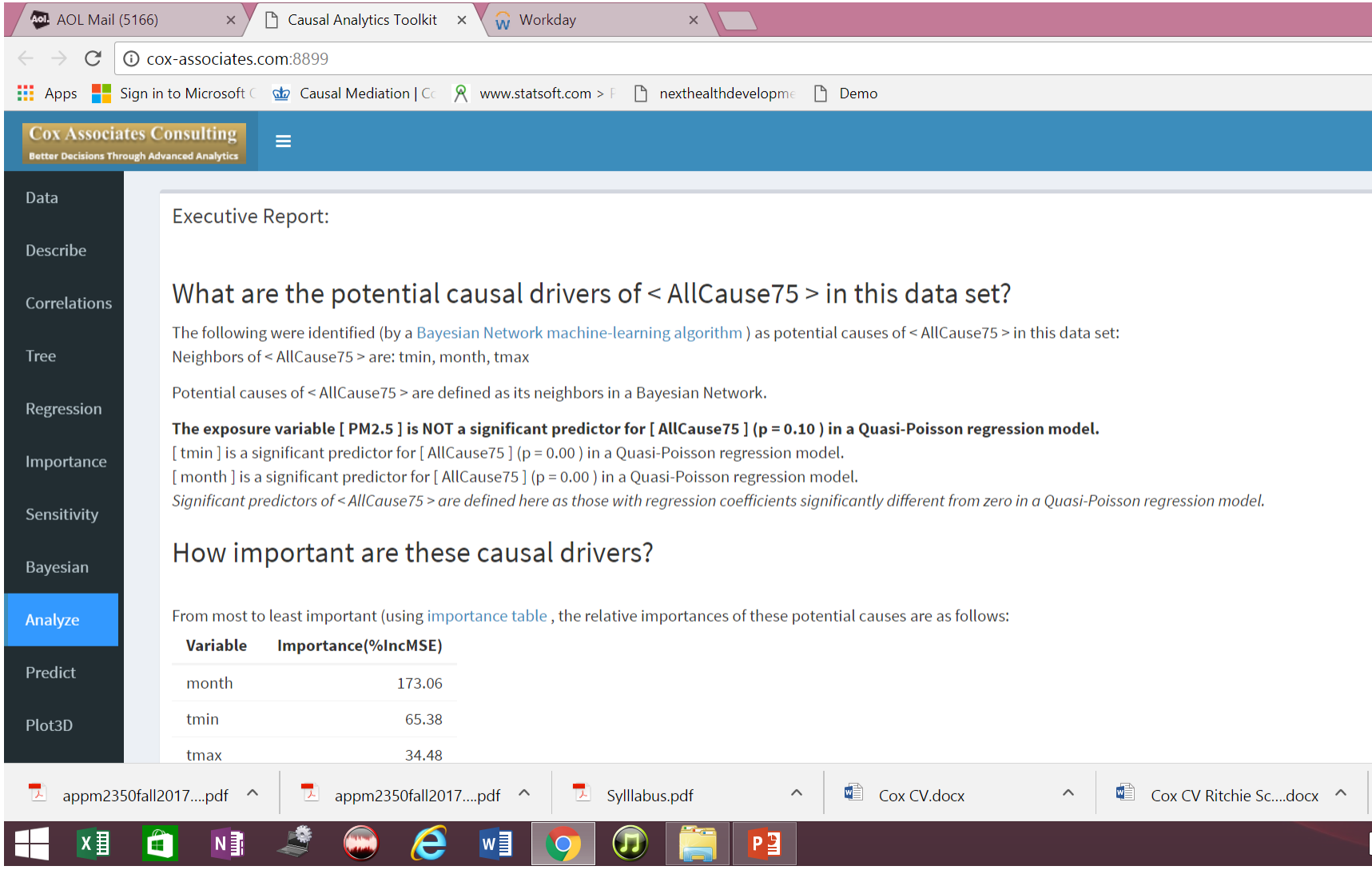
- Automated non-parametric model ensembles!
  - RandomForest
  - Causal DAGs

Automating analysis is now practical

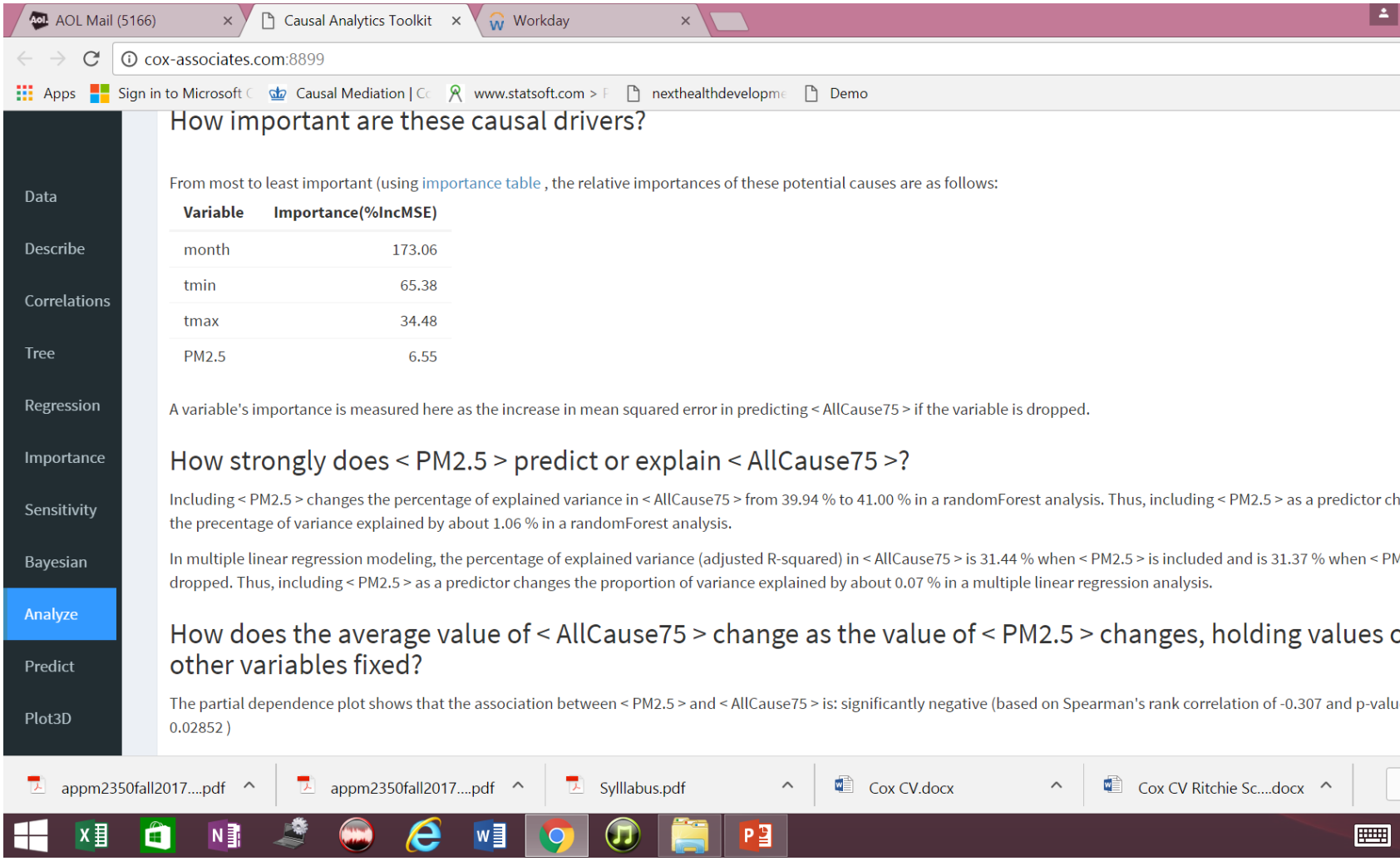


	AllCause75	PM2.5	tmin	tmax	MAXRH	month	day
1	151	38.4	36	72	68.8	1	1
2	158	17.4	36	75	48.9	1	2
3	139	19.9	44	75	61.3	1	3
4	164	64.6	37	68	87.9	1	4
5	136	6.1	40	61	47.5	1	5
6	152	18.8	39	69	39	1	6
7	160	19.1	41	76	40.9	1	7
8	148	13.8	41	83	33.7	1	8

Automating analysis is now practical



Automating analysis is now practical



Causal Analytics Toolkit

cox-associates.com:8899

Apps | Sign in to Microsoft | Causal Mediation | Cc | www.statsoft.com > F | nexthealthdevelopme | Demo

- Data
- Describe
- Correlations
- Tree
- Regression
- Importance
- Sensitivity
- Bayesian
- Analyze
- Predict
- Plot3D
- Granger

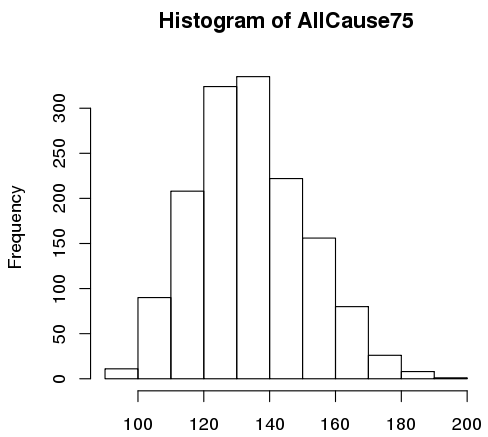
Summary statistics

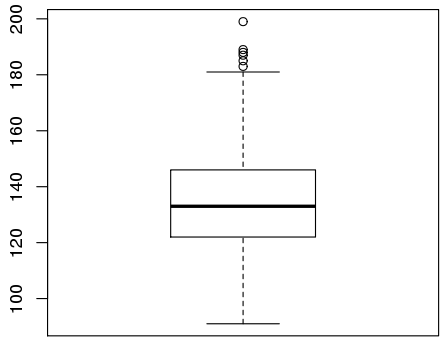
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
91.0	122.0	133.0	134.7	146.0	199.0

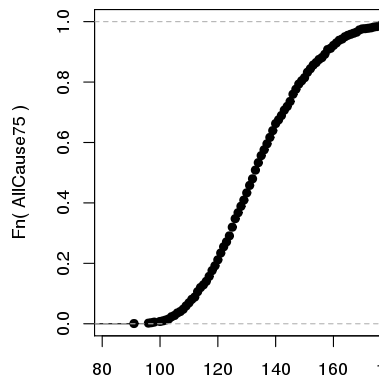
Descriptives, high and low data values

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1461	0	92	1	134.7	19.34	109	113	122	133	146	158	164

Histogram | Boxplot | ECDF



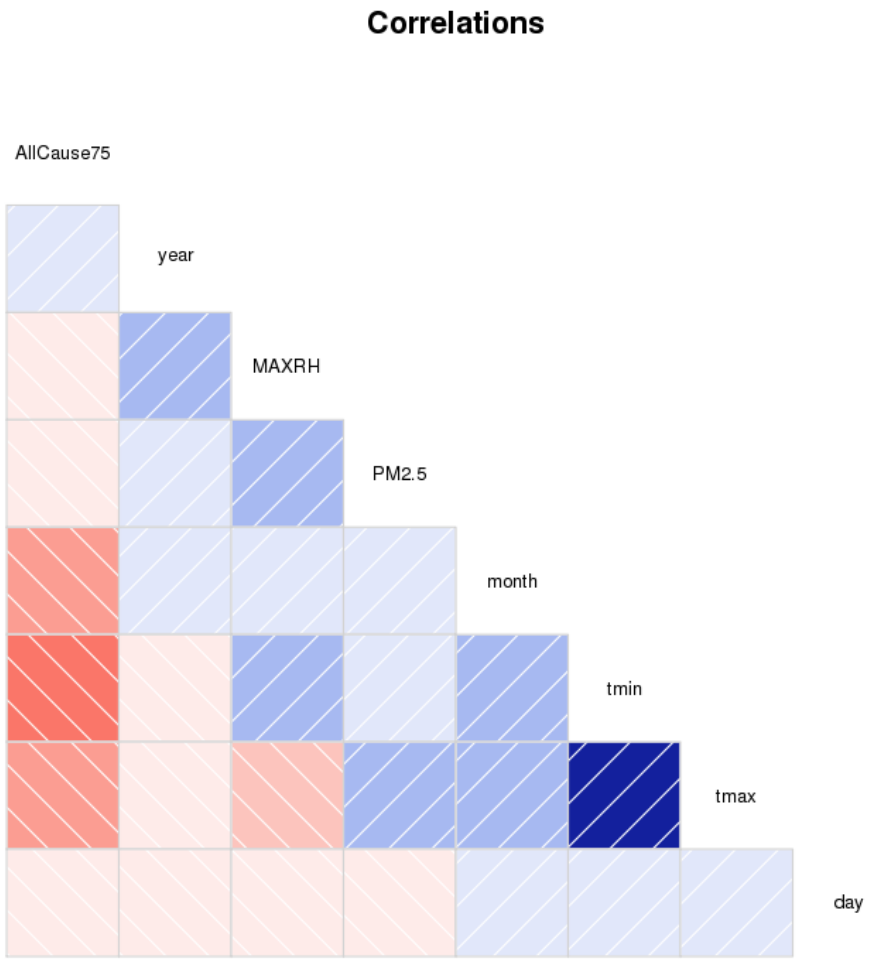




Cox Associates Consulting
Better Decisions Through Advanced Analytics

- Data
- Describe
- Correlations
- Tree
- Regression
- Importance
- Sensitivity
- Bayesian
- Analyze
- Predict
- Plot3D
- Granger

Corrgram display. Red = negative correlation,Blue = positive correlation.



Causal Analytics Toolkit

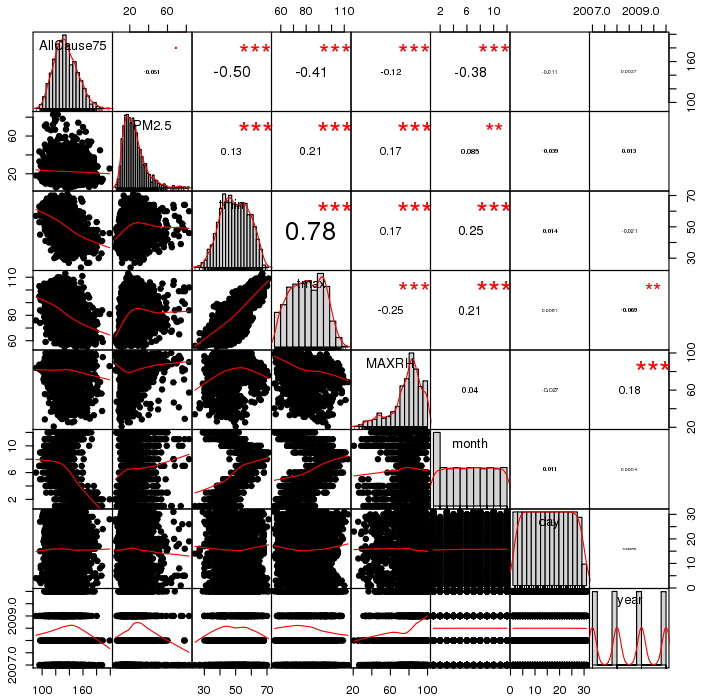
cox-associates.com:8899

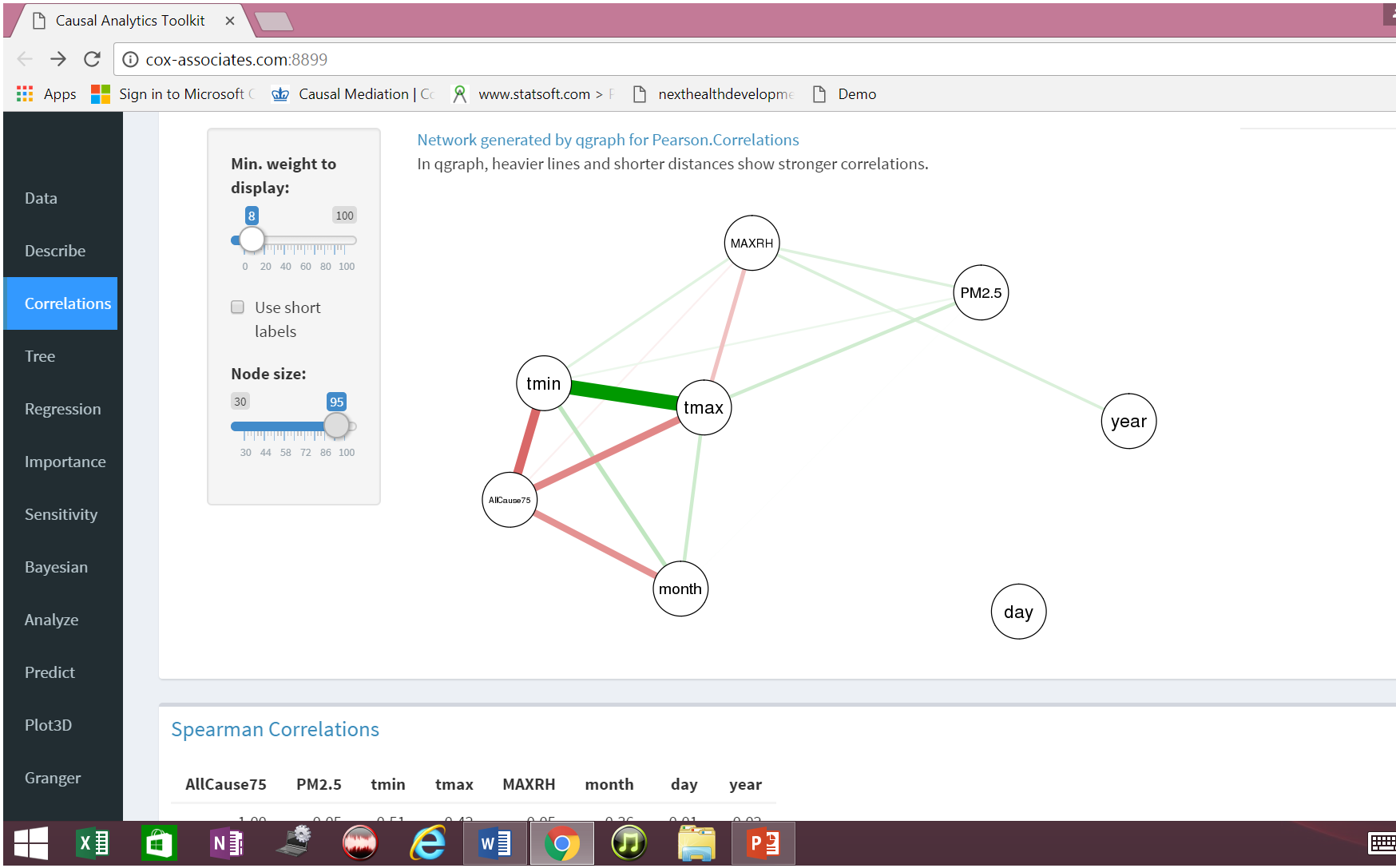
Apps | Sign in to Microsoft | Causal Mediation | Cc | www.statsoft.com > F | nexthealthdevelopme | Demo

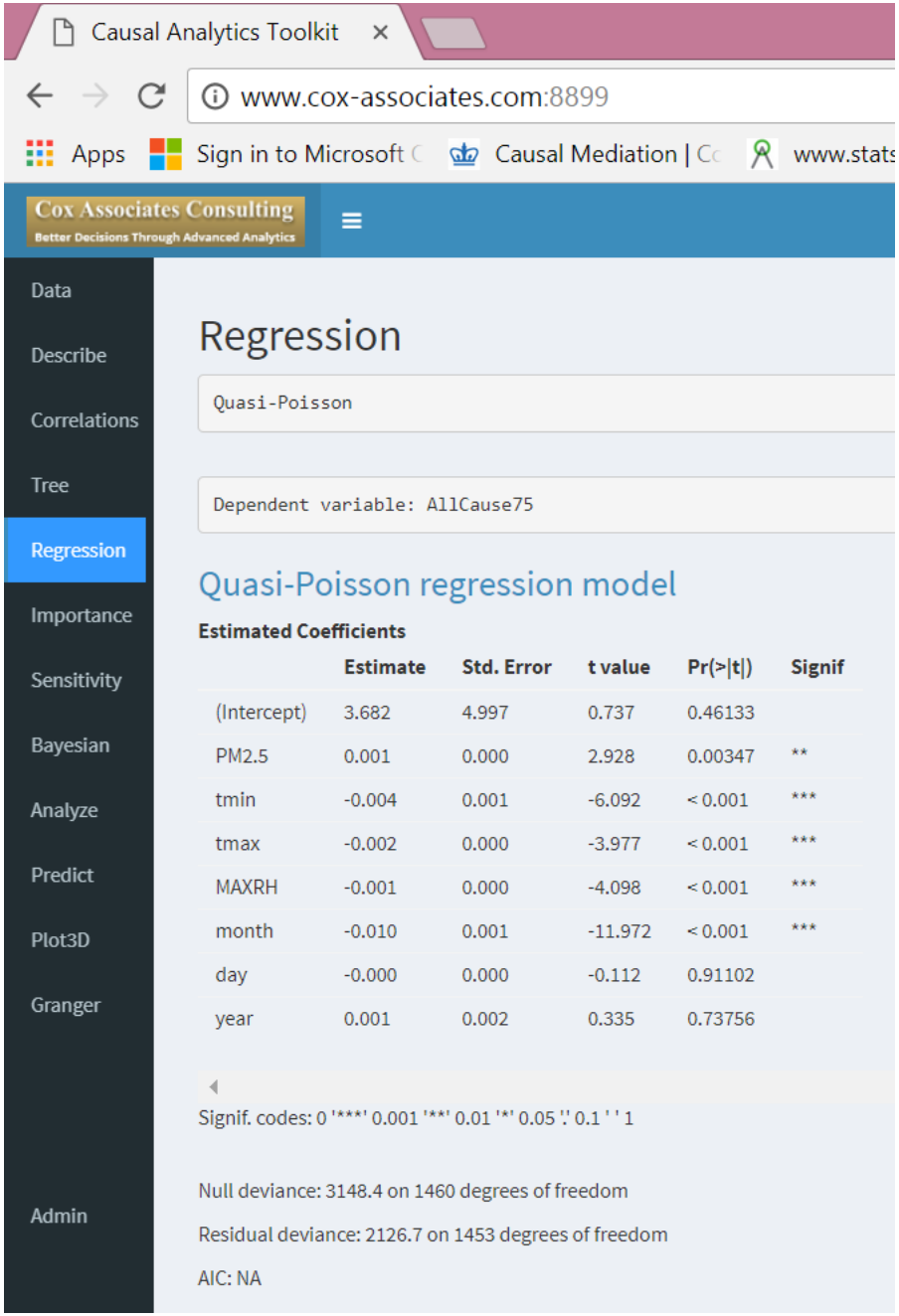
- Data
- Describe
- Correlations
- Tree
- Regression
- Importance
- Sensitivity
- Bayesian
- Analyze
- Predict
- Plot3D
- Granger

Scatterplot matrix

Scatterplot matrix, with histograms, kernel density overlays, absolute correlations, and significance asterisks (0.05, 0.01, 0.001)




Causal Analytics Toolkit
cox-associates.com:8899
Apps
Sign in to Microsoft
Causal Mediation | Cc
www.statsoft.com > F
nexthealthdevelopme
Demo
Data
Describe
Correlations
Tree
Regression
Importance
Sensitivity
Bayesian
Analyze
Predict
Plot3D
Granger
Min. weight to display:
8
100
0 20 40 60 80 100
Use short labels
Node size:
30
95
30 44 58 72 86 100
Network generated by qgraph for Pearson.Correlations
In qgraph, heavier lines and shorter distances show stronger correlations.
MAXRH
PM2.5
tmin
tmax
year
AllCause75
month
day
Spearman Correlations
AllCause75
PM2.5
tmin
tmax
MAXRH
month
day
year


Causal Analytics Toolkit
www.cox-associates.com:8899
Apps
Sign in to Microsoft
Causal Mediation | Cc
www.stats
Cox Associates Consulting
Better Decisions Through Advanced Analytics
Data
Describe
Correlations
Tree
Regression
Importance
Sensitivity
Bayesian
Analyze
Predict
Plot3D
Granger
Admin
Regression
Quasi-Poisson
Dependent variable: AllCause75
Quasi-Poisson regression model
Estimated Coefficients
Estimate
Std. Error
t value
Pr(>|t|)
Signif
(Intercept) 3.682 4.997 0.737 0.46133
PM2.5 0.001 0.000 2.928 0.00347 **
tmin -0.004 0.001 -6.092 < 0.001 ***
tmax -0.002 0.000 -3.977 < 0.001 ***
MAXRH -0.001 0.000 -4.098 < 0.001 ***
month -0.010 0.001 -11.972 < 0.001 ***
day -0.000 0.000 -0.112 0.91102
year 0.001 0.002 0.335 0.73756
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null deviance: 3148.4 on 1460 degrees of freedom
Residual deviance: 2126.7 on 1453 degrees of freedom
AIC: NA

Causal Analytics Toolkit

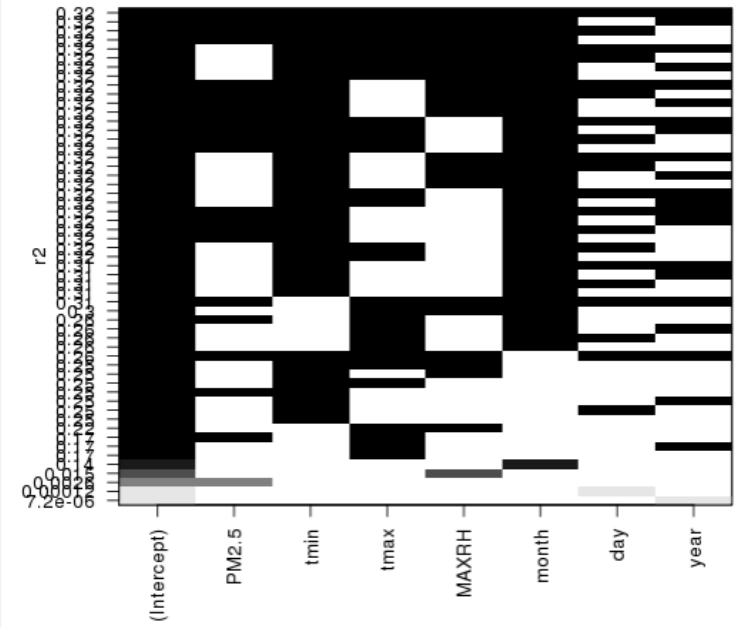
www.cox-associates.com:8899

Apps | Sign in to Microsoft | Causal Mediation | Cc | www.statsoft.com >

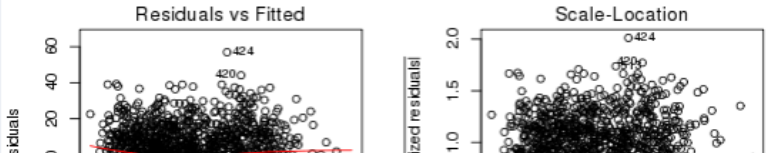
- Data
- Describe
- Correlations
- Tree
- Regression
- Importance
- Sensitivity
- Bayesian
- Analyze
- Predict
- Plot3D
- Granger
- Admin

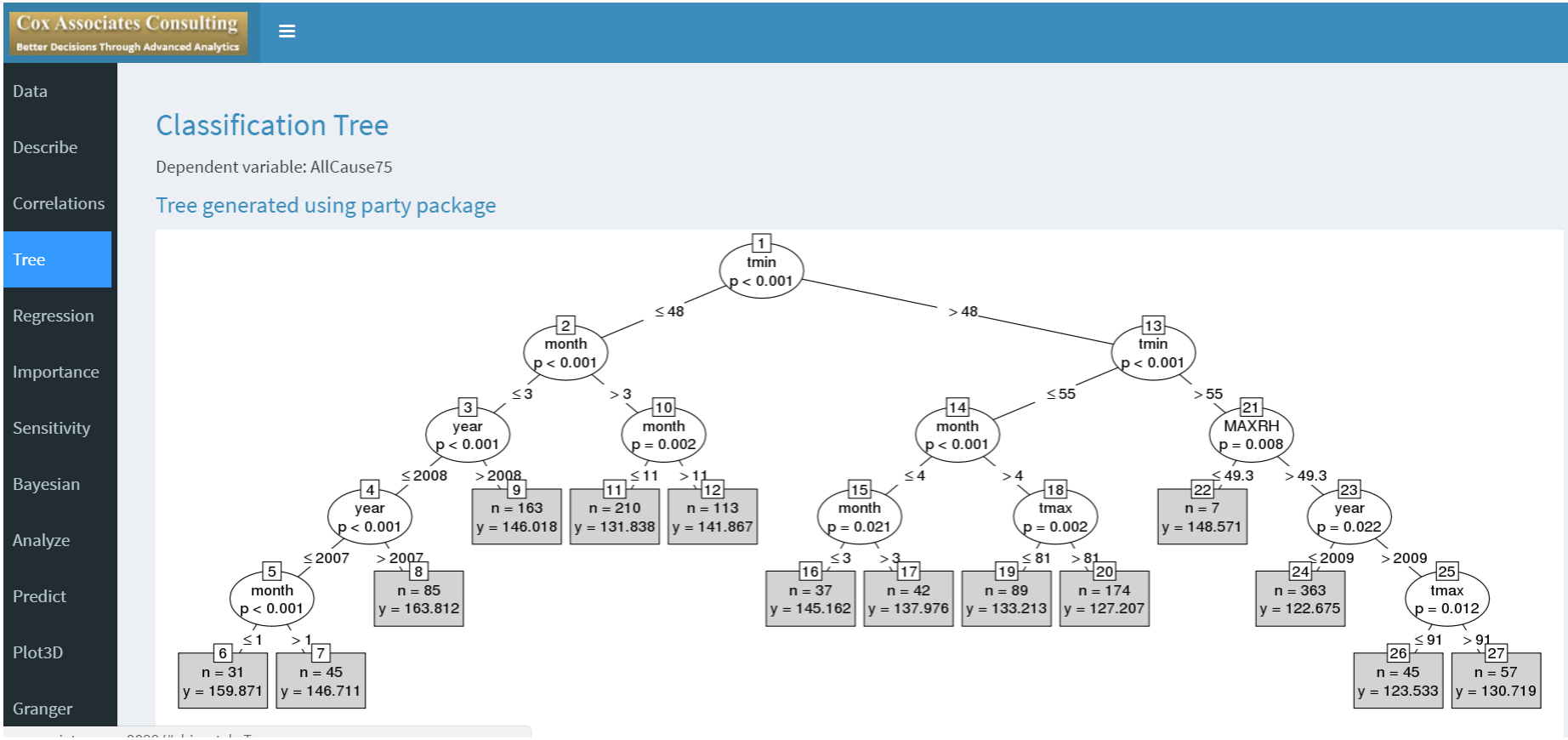
Variables used in Models

Presence of each variable in models of different sizes, and model r2, using all subsets regress



Regression diagnostic plots





Non-parametric methods

Causal Analytics Toolkit

cox-associates.com:8899

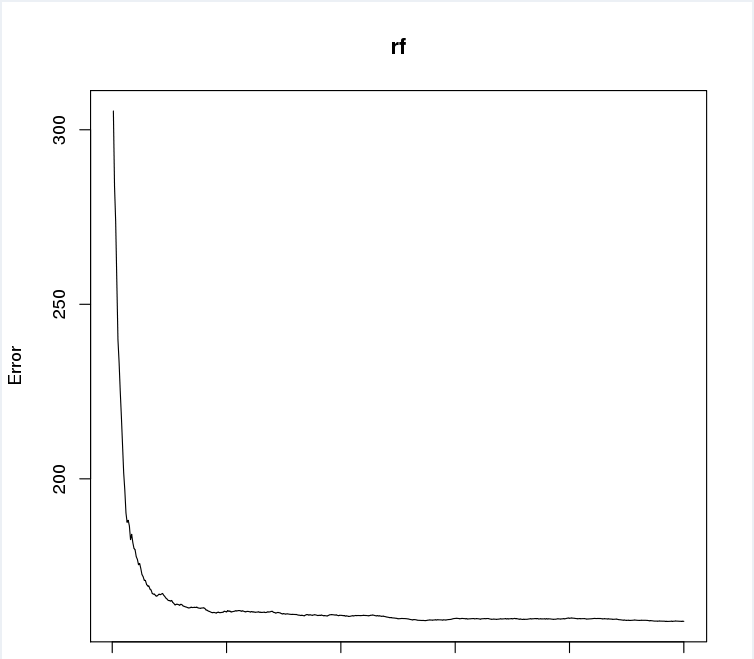
Apps | Sign in to Microsoft | Causal Mediation | | www.statsoft.com > | nexthealthdevelopme | Demo

- Data
- Describe
- Correlations
- Tree
- Regression
- Importance
- Sensitivity
- Bayesian
- Analyze
- Predict
- Plot3D
- Granger

randomForest summary

```
          Mean of squared residuals: 159.1731
                    % Var explained: 45.61
```

randomForest accuracy vs. number of trees



Model ensemble methods

Causal Analytics Toolkit

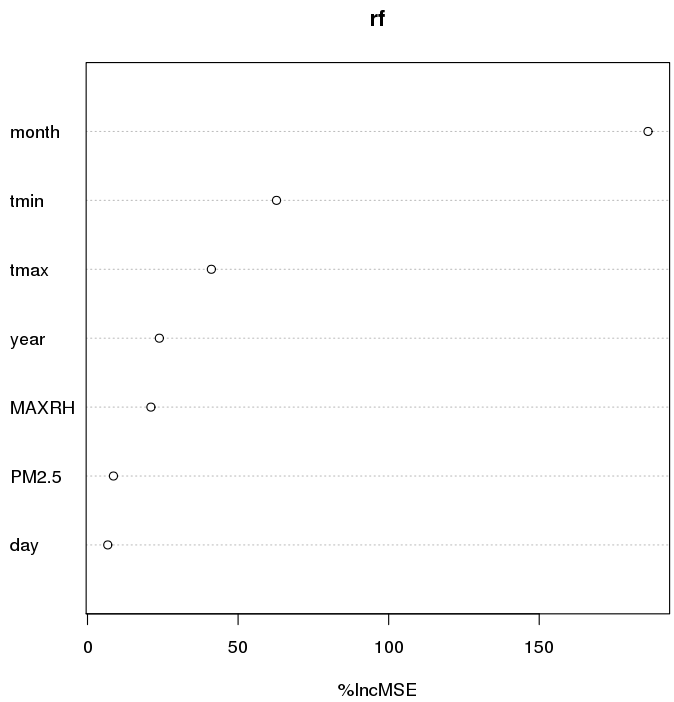
cox-associates.com:8899

Apps | Sign in to Microsoft | Causal Mediation | Cc | www.statsoft.com > F | nexthealthdevelopme | Demo

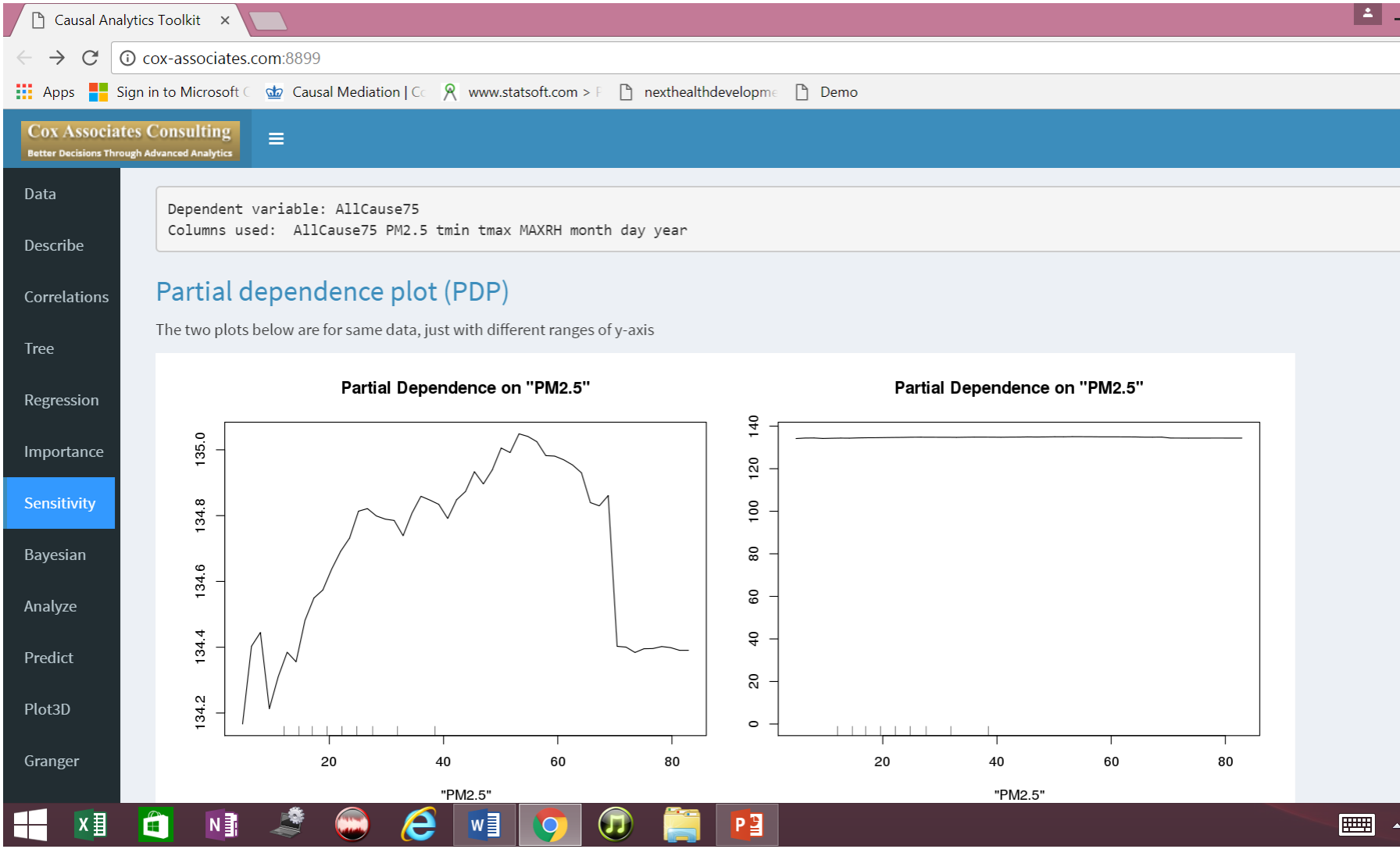
- Data
- Describe
- Correlations
- Tree
- Regression
- Importance
- Sensitivity
- Bayesian
- Analyze
- Predict
- Plot3D
- Granger

A variable's importance is measured here as the increase in mean squared error in predicting < AllCause75 > if the variable is dropped.

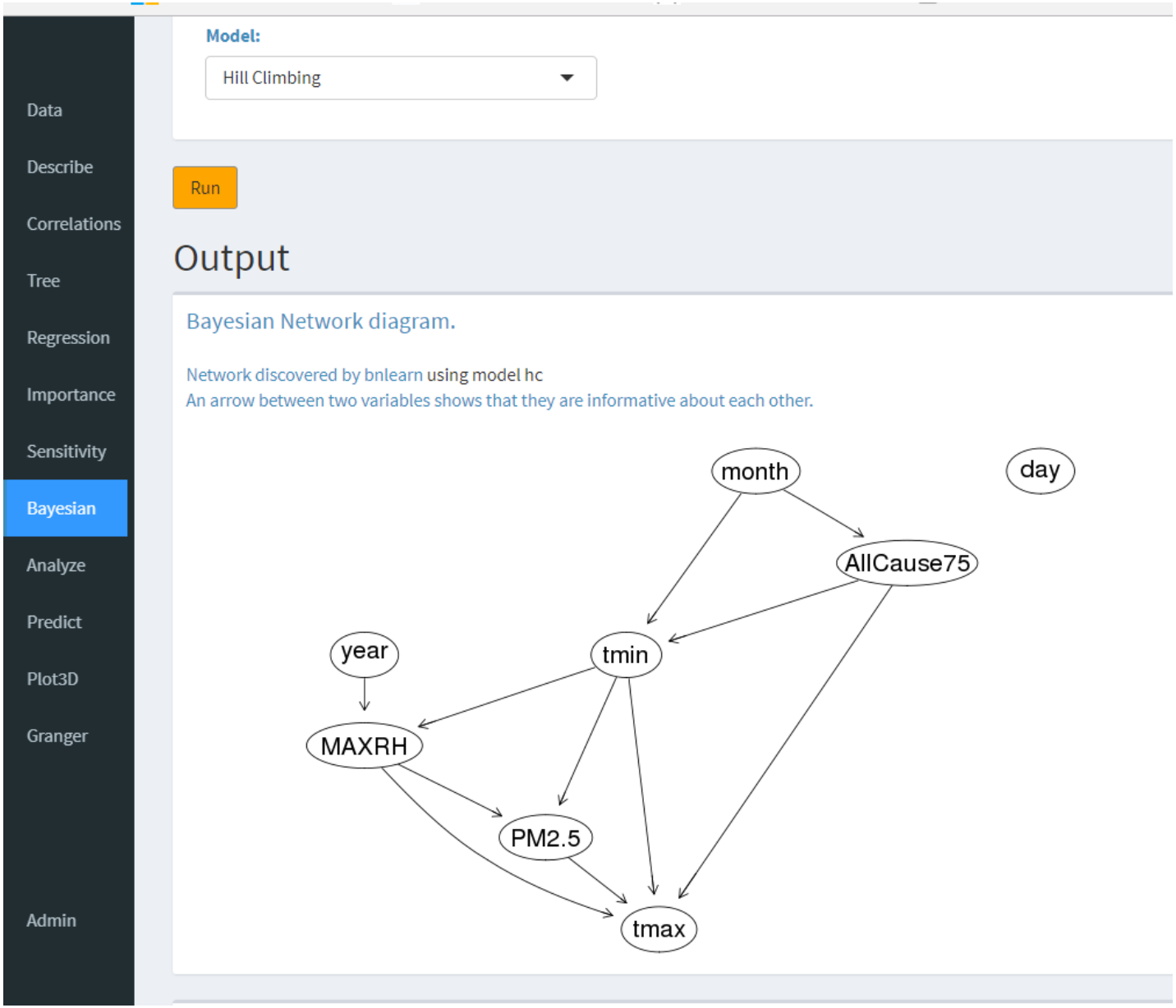
Variable Importance Plots



cox-associates.com:8899/#shiny-tab-Importance



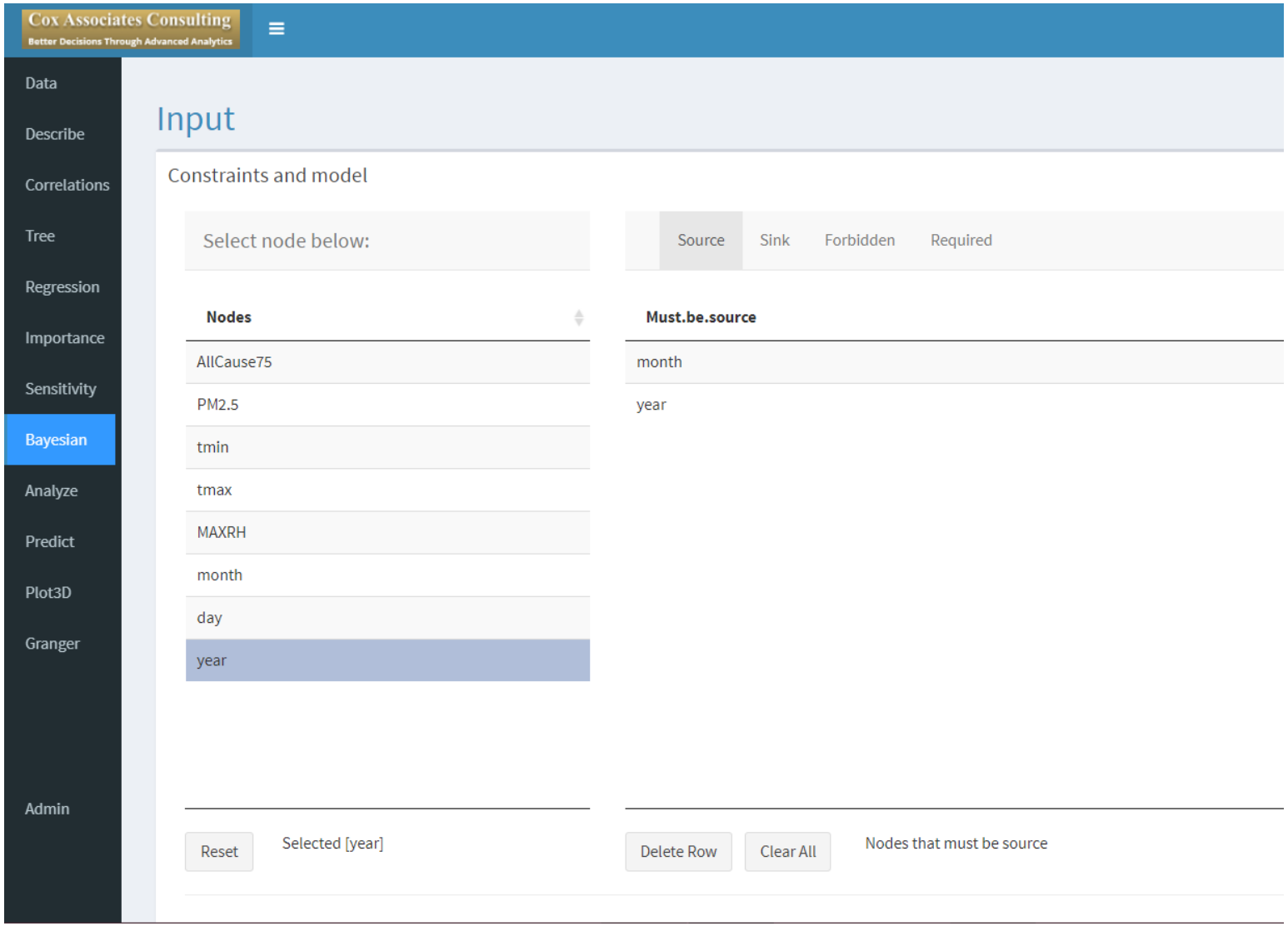
Automated non-parametric model ensembles

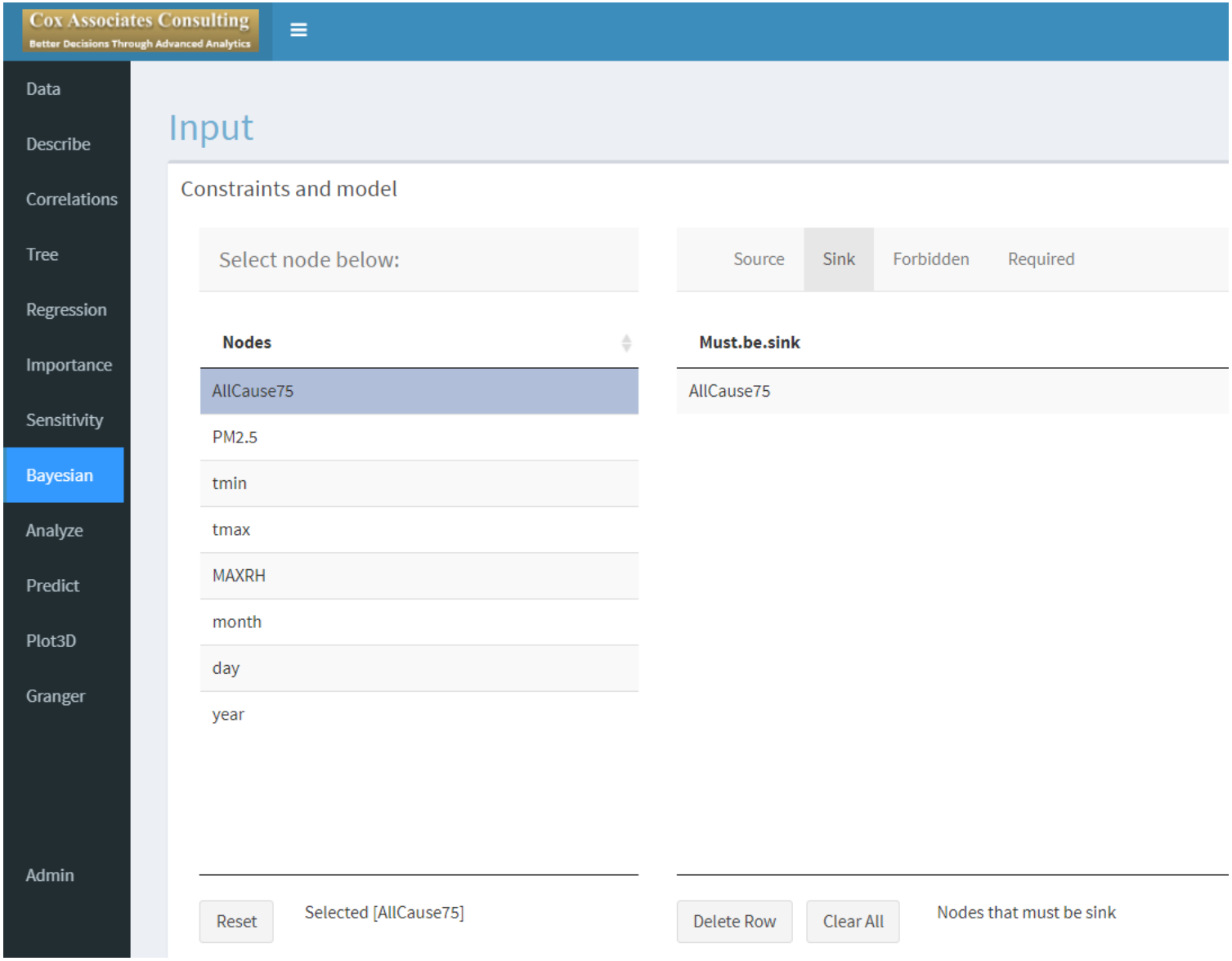


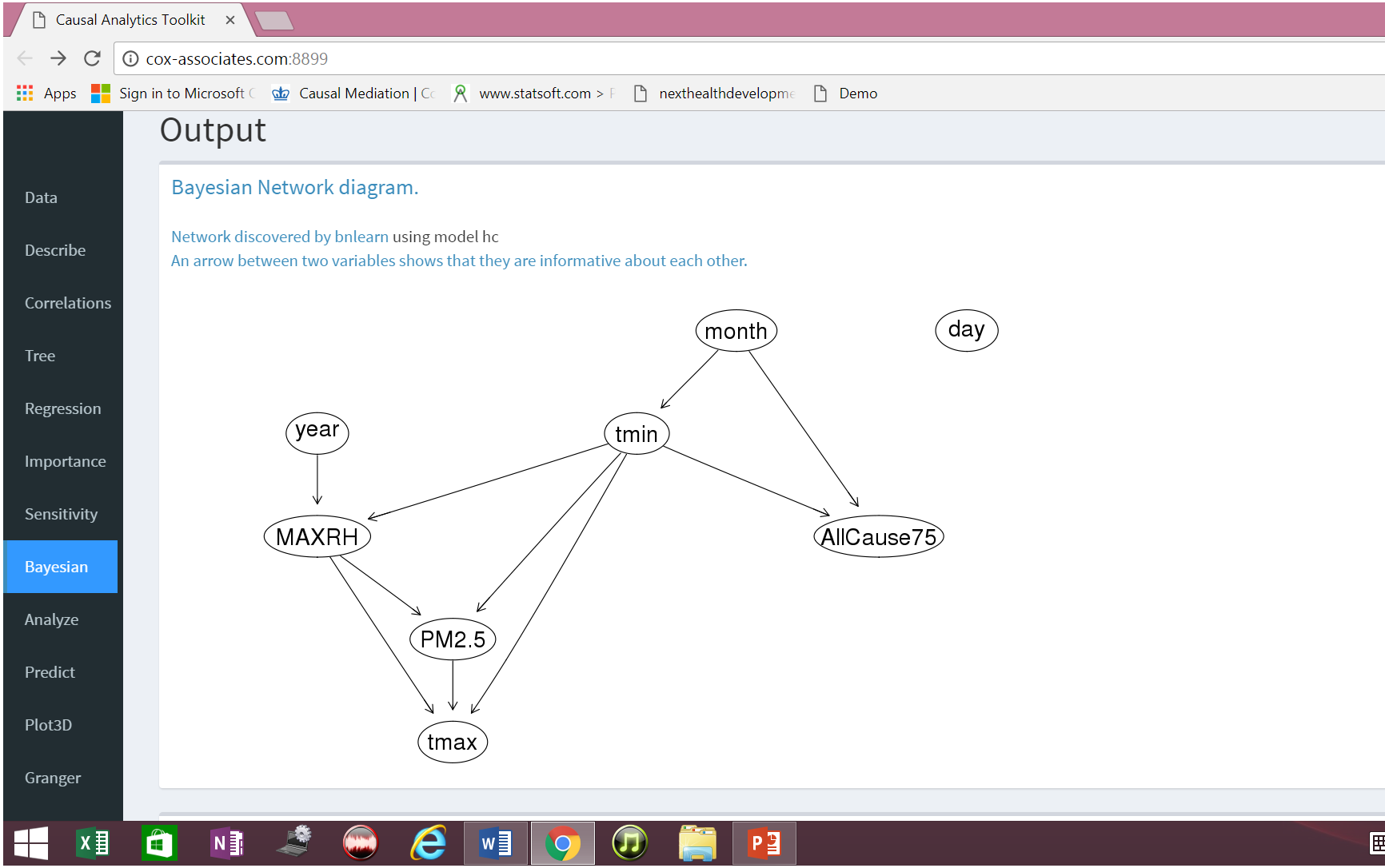
Automated non-parametric model (BN)

Knowledge-based constraints

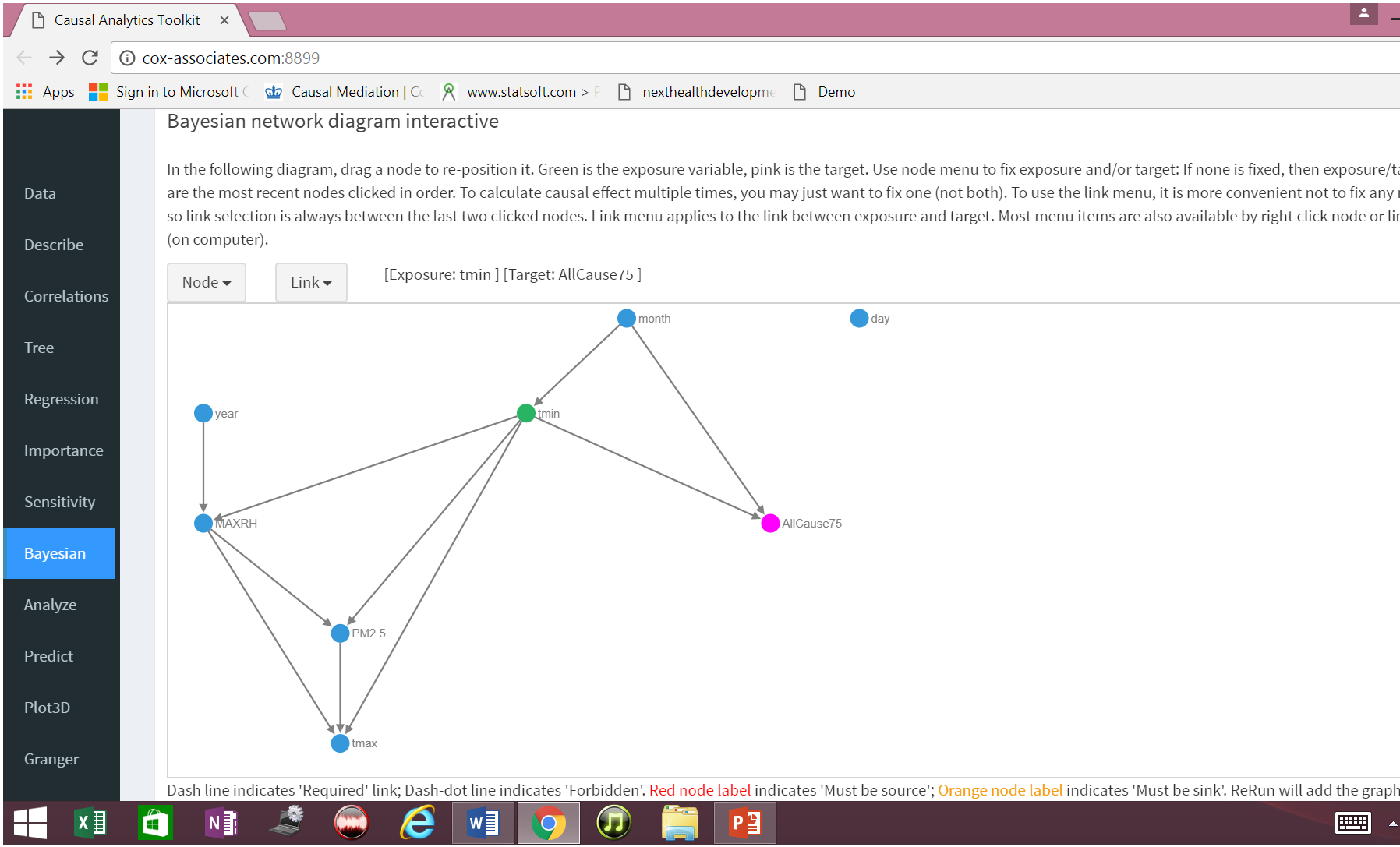
Potential p-hacking point, but controllable

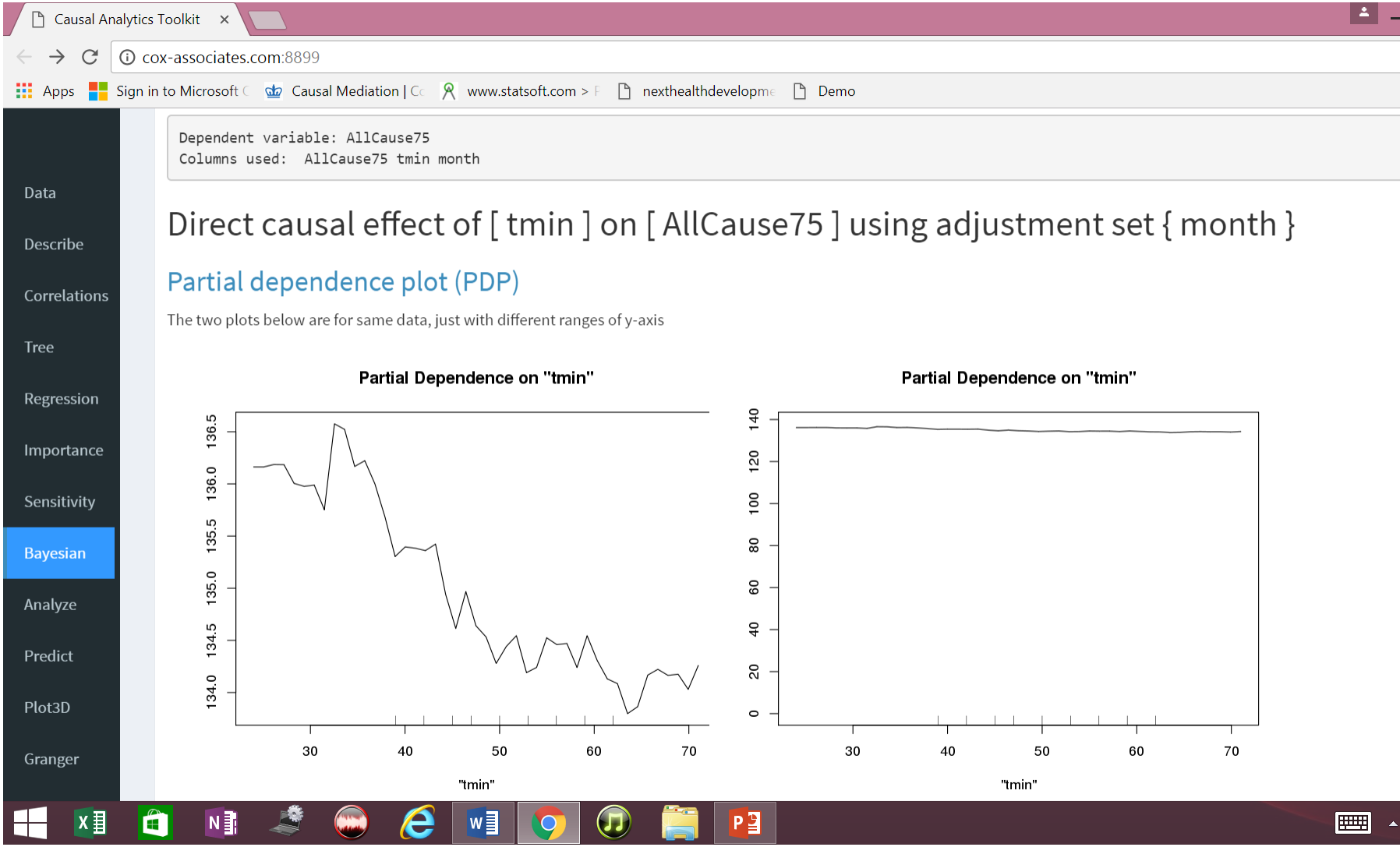



Cox Associates Consulting
Better Decisions Through Advanced Analytics
Data
Describe
Correlations
Tree
Regression
Importance
Sensitivity
Bayesian
Analyze
Predict
Plot3D
Granger
Admin
Input
Constraints and model
Select node below:
Source
Sink
Forbidden
Required
Nodes
AllCause75
PM2.5
tmin
tmax
MAXRH
month
day
year
Must.be.sink
AllCause75
Reset
Selected [AllCause75]
Delete Row
Clear All
Nodes that must be sink

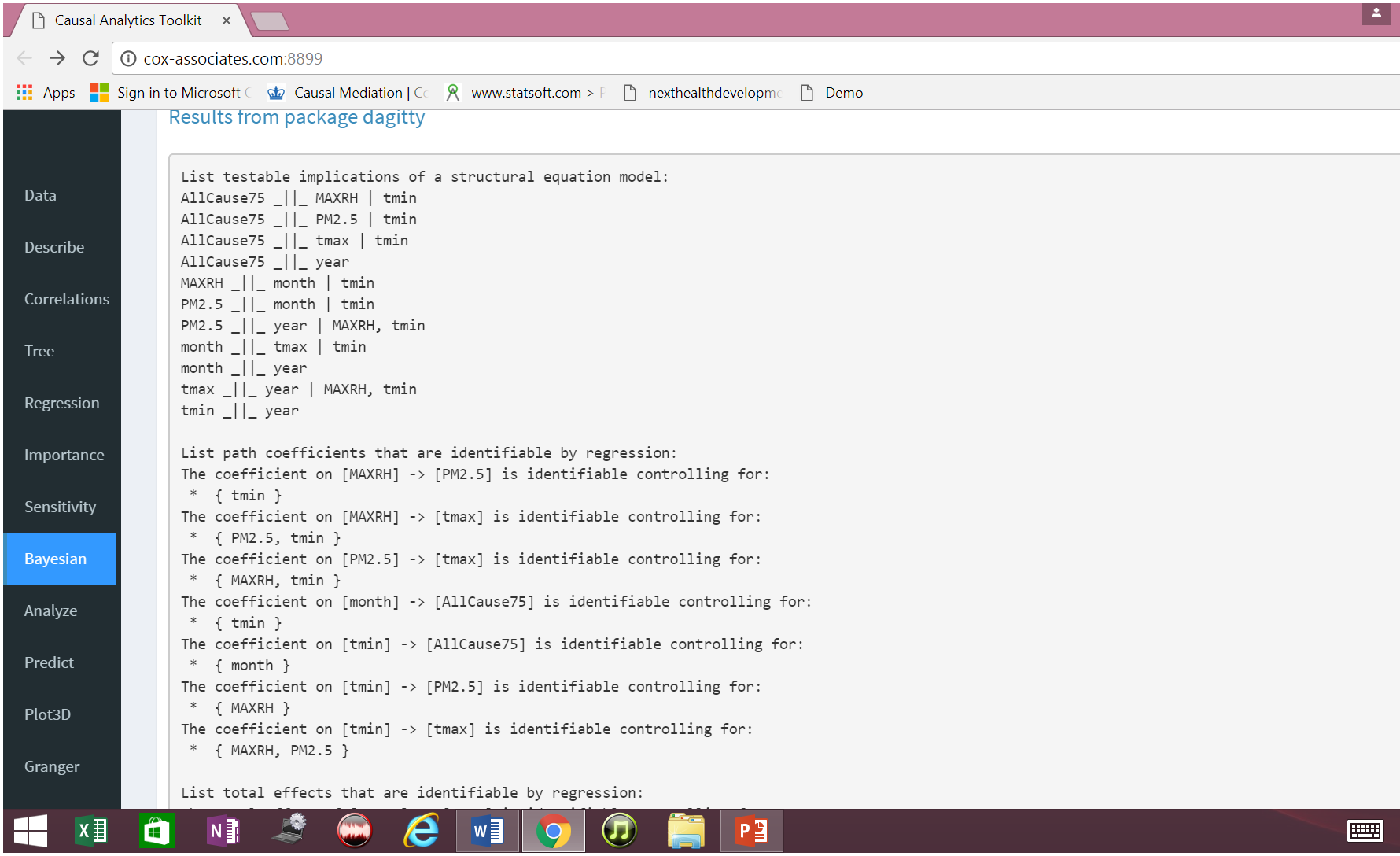


Constrained automated non-parametric causal model


Causal Analytics Toolkit
cox-associates.com:8899
Apps
Sign in to Microsoft
Causal Mediation | Cc
www.statsoft.com > F
nexthealthdevelopme
Demo
Bayesian network diagram interactive
In the following diagram, drag a node to re-position it. Green is the exposure variable, pink is the target. Use node menu to fix exposure and/or target: If none is fixed, then exposure/ta
are the most recent nodes clicked in order. To calculate causal effect multiple times, you may just want to fix one (not both). To use the link menu, it is more convenient not to fix any
so link selection is always between the last two clicked nodes. Link menu applies to the link between exposure and target. Most menu items are also available by right click node or lir
(on computer).
Node
Link
[Exposure: tmin] [Target: AllCause75]
Data
Describe
Correlations
Tree
Regression
Importance
Sensitivity
Bayesian
Analyze
Predict
Plot3D
Granger
month
day
year
tmin
MAXRH
AllCause75
PM2.5
tmax
Dash line indicates 'Required' link; Dash-dot line indicates 'Forbidden'. Red node label indicates 'Must be source'; Orange node label indicates 'Must be sink'. ReRun will add the graph



Constrained automated non-parametric causal model ensemble

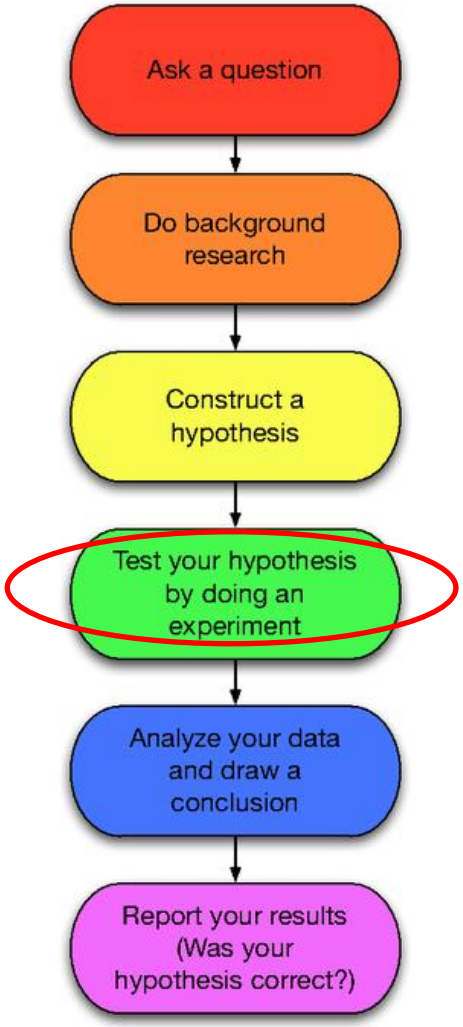

Causal Analytics Toolkit
cox-associates.com:8899
Apps
Sign in to Microsoft
Causal Mediation | C
www.statsoft.com > F
nexthealthdevelopme
Demo
Data
Describe
Correlations
Tree
Regression
Importance
Sensitivity
Bayesian
Analyze
Predict
Plot3D
Granger
Results from package dagitty
List testable implications of a structural equation model:
AllCause75 _||_ MAXRH | tmin
AllCause75 _||_ PM2.5 | tmin
AllCause75 _||_ tmax | tmin
AllCause75 _||_ year
MAXRH _||_ month | tmin
PM2.5 _||_ month | tmin
PM2.5 _||_ year | MAXRH, tmin
month _||_ tmax | tmin
month _||_ year
tmax _||_ year | MAXRH, tmin
tmin _||_ year
List path coefficients that are identifiable by regression:
The coefficient on [MAXRH] -> [PM2.5] is identifiable controlling for:
* { tmin }
The coefficient on [MAXRH] -> [tmax] is identifiable controlling for:
* { PM2.5, tmin }
The coefficient on [PM2.5] -> [tmax] is identifiable controlling for:
* { MAXRH, tmin }
The coefficient on [month] -> [AllCause75] is identifiable controlling for:
* { tmin }
The coefficient on [tmin] -> [AllCause75] is identifiable controlling for:
* { month }
The coefficient on [tmin] -> [PM2.5] is identifiable controlling for:
* { MAXRH }
The coefficient on [tmin] -> [tmax] is identifiable controlling for:
* { MAXRH, PM2.5 }
List total effects that are identifiable by regression:

Summary: Practical to avoid p-hacking and model-dependent conclusions via…

- Automated (but appropriate/intelligent)
- Non-parametric
- Causal model-constrained
- Ensembles

- Automated non-parametric causal model ensembles for causal DAG discovery!
  - Enabled by existing R packages: randomForest, bnlearn, dagitty, CompareCausalNeworks

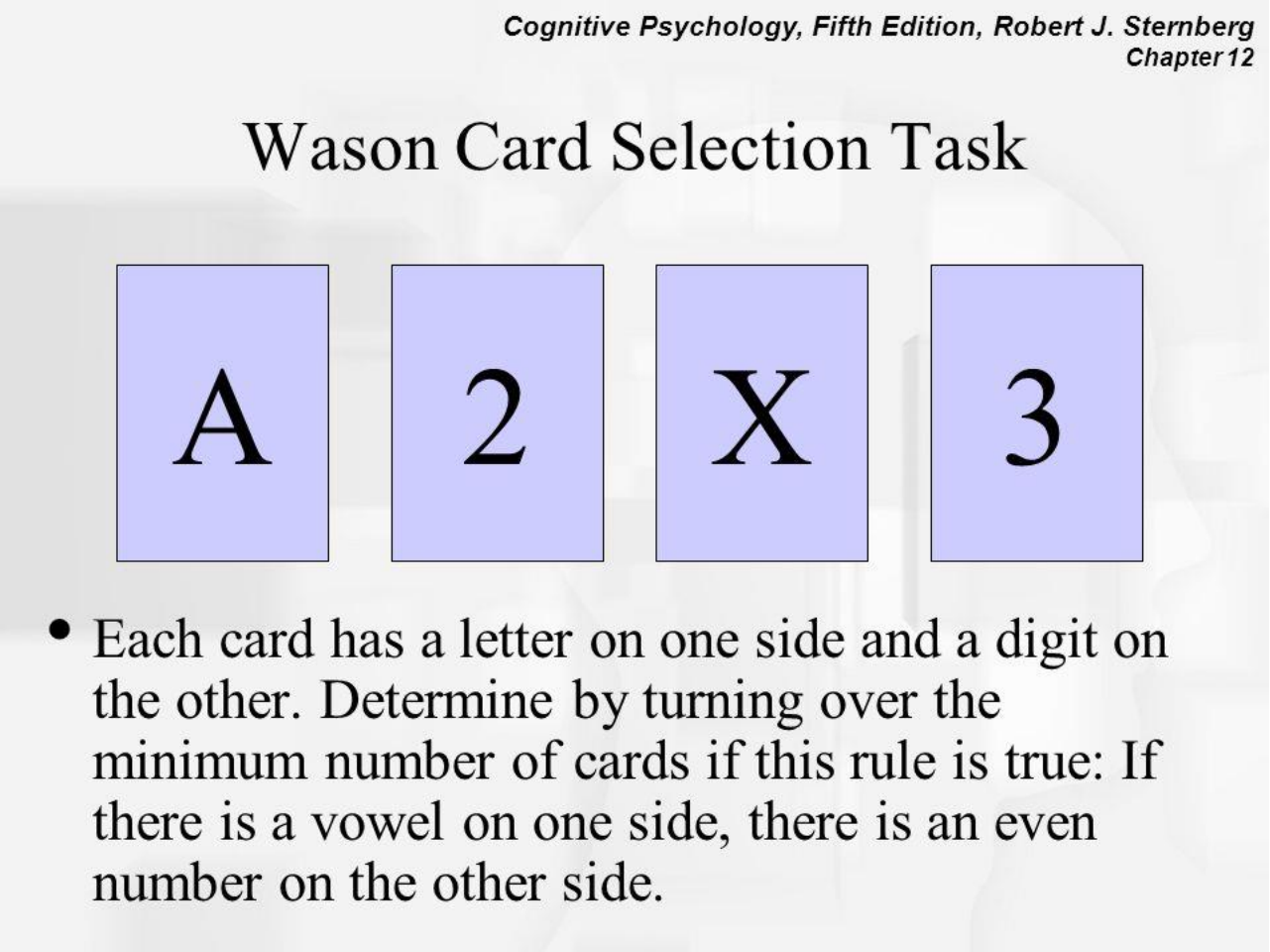
Designing an experiment



- Key question: What information do we need to test a hypothesis?

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

Example: Designing data collection to test a hypothesis

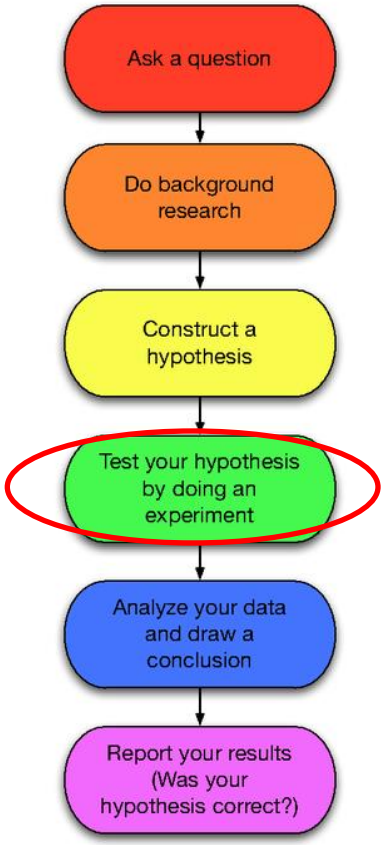


Vowels:
A, E, I, O, U

Even numbers:
0, 2, 4, 6, 8

http://images.slideplayer.com/16/5138662/slides/slide_43.jpg

Variation: Testing a more concrete hypothesis



There are four cards lying on a table. Each has a capital letter on one side and a single digit number on the other side. The exposed sides are shown below:

A	D	3	7

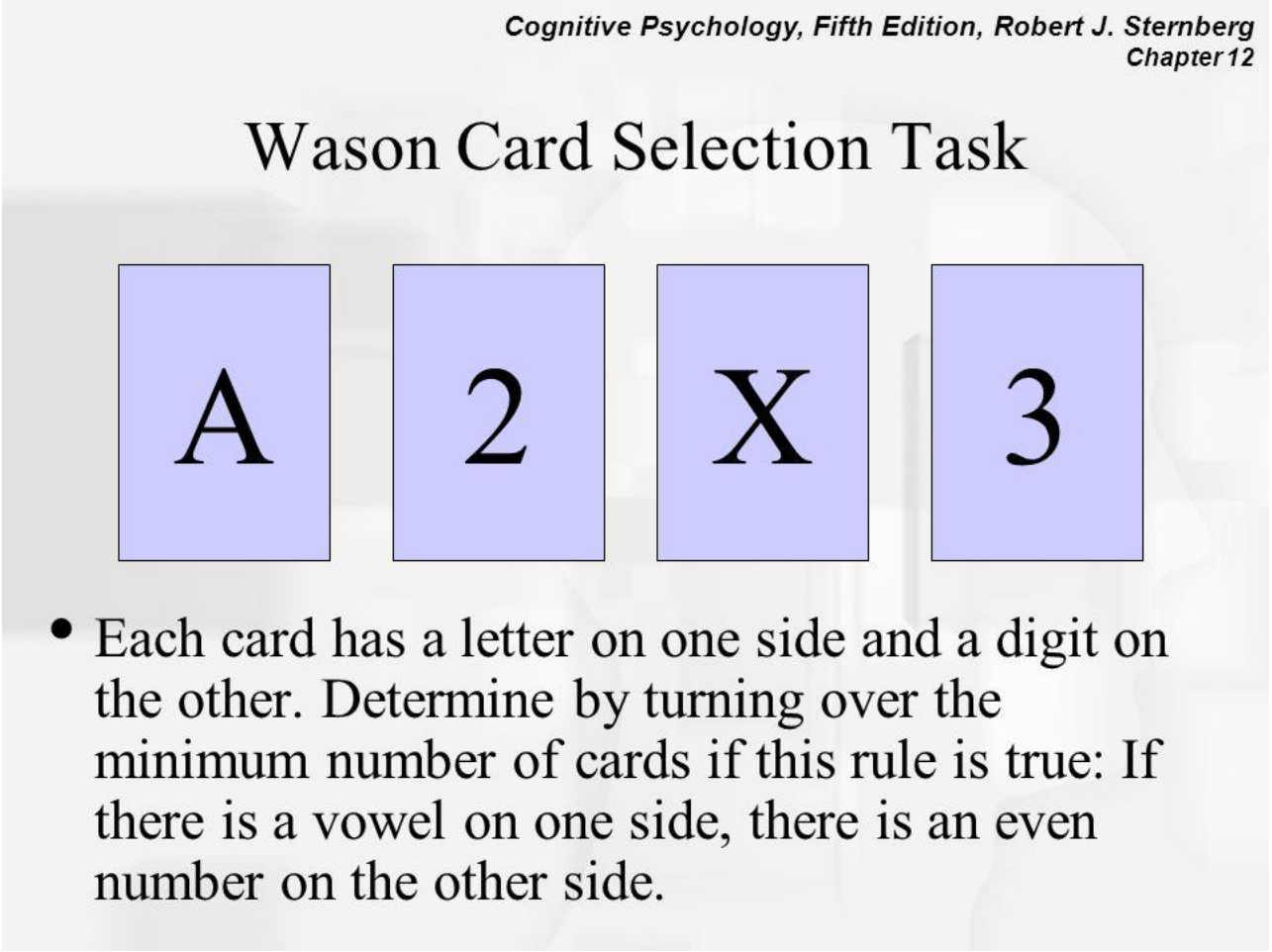
The rule shown below applies to these four cards and may be true or false:

> If there is an A on one side of the card, then there is a 3 on the other side of the card

Your task is to decide those cards, and only those cards, that need to be turned over in order to discover whether the rule is true or false.

www.researchgate.net/publication/258179748_Dual-Process_Theories_of_Higher_Cognition

Example: Designing data collection to test a hypothesis



Answer: A and 3. (Either one could *disconfirm* the hypothesis)

http://images.slideplayer.com/16/5138662/slides/slide_43.jpg

Variation: Testing a more concrete hypothesis

There are four cards lying on a table. Each has a capital letter on one side and a single digit number on the other side. The exposed sides are shown below:

The rule shown below applies to these four cards and may be true or false:

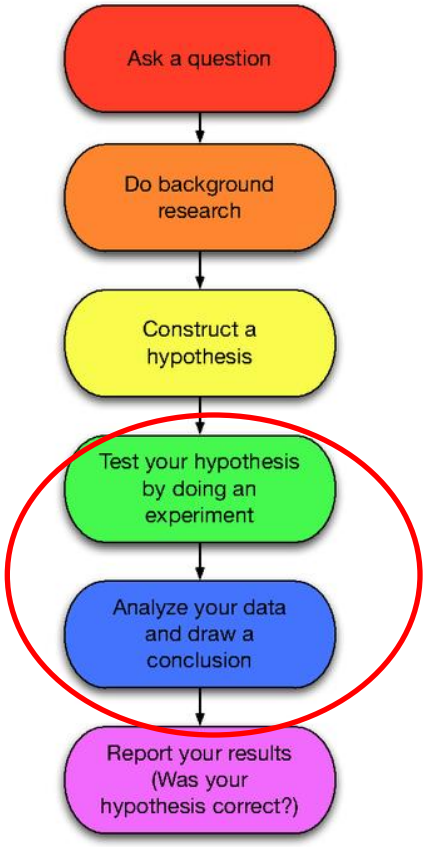
If there is an A on one side of the card, then there is a 3 on the other side of the card

Your task is to decide those cards, and only those cards, that need to be turned over in order to discover whether the rule is true or false.

Answer: A and 7. (Either one could *disconfirm* the hypothesis)

www.researchgate.net/publication/258179748_Dual-Process_Theories_of_Higher_Cognition

Testing a hypothesis of discrimination

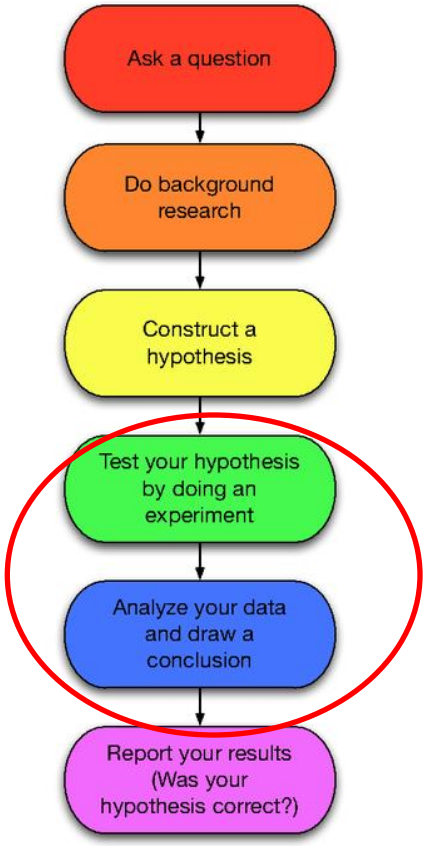


- You are hearing a case on discrimination in admissions at a state university
- The data before you are as follows
  - Assume men and women are identically qualified

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

Testing a hypothesis of discrimination

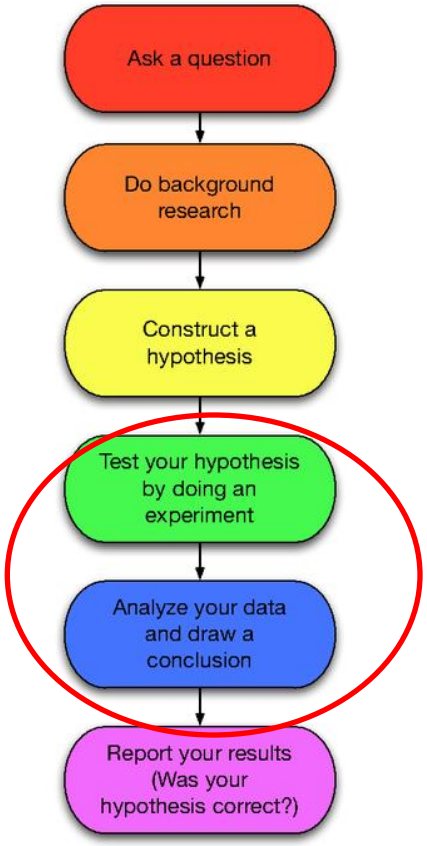


- Do these data allow a test of the null hypothesis of no discrimination?
- If a statistician finds this discrepancy not likely to be due to chance, can we conclude that discrimination is likely?

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

Testing a hypothesis of discrimination

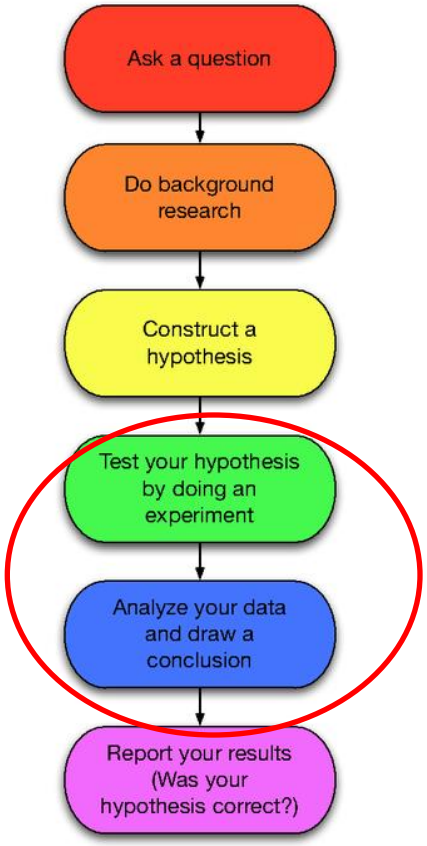


- What would be the probable effect on admission rates of instructing all departments to change women's admission rate to equal that of (equally qualified) men?

	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

Testing a hypothesis of discrimination

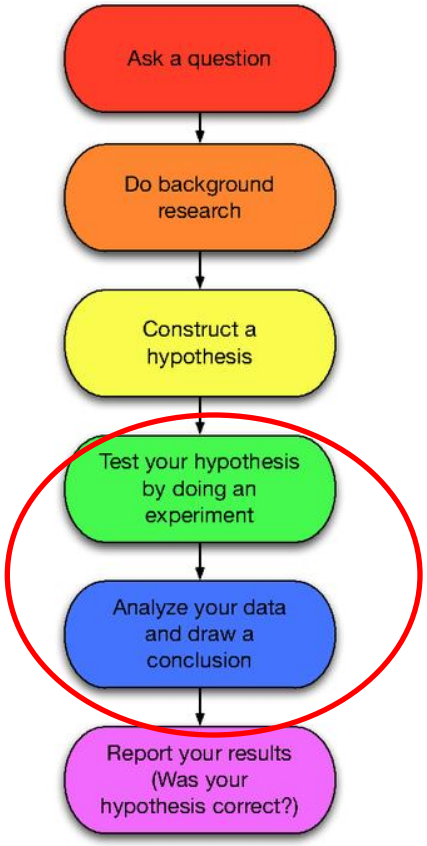


- *Answer:* No conclusions can be drawn from these data. Women may have higher acceptance rates than men in every department, yet apply to departments with lower admission rates.

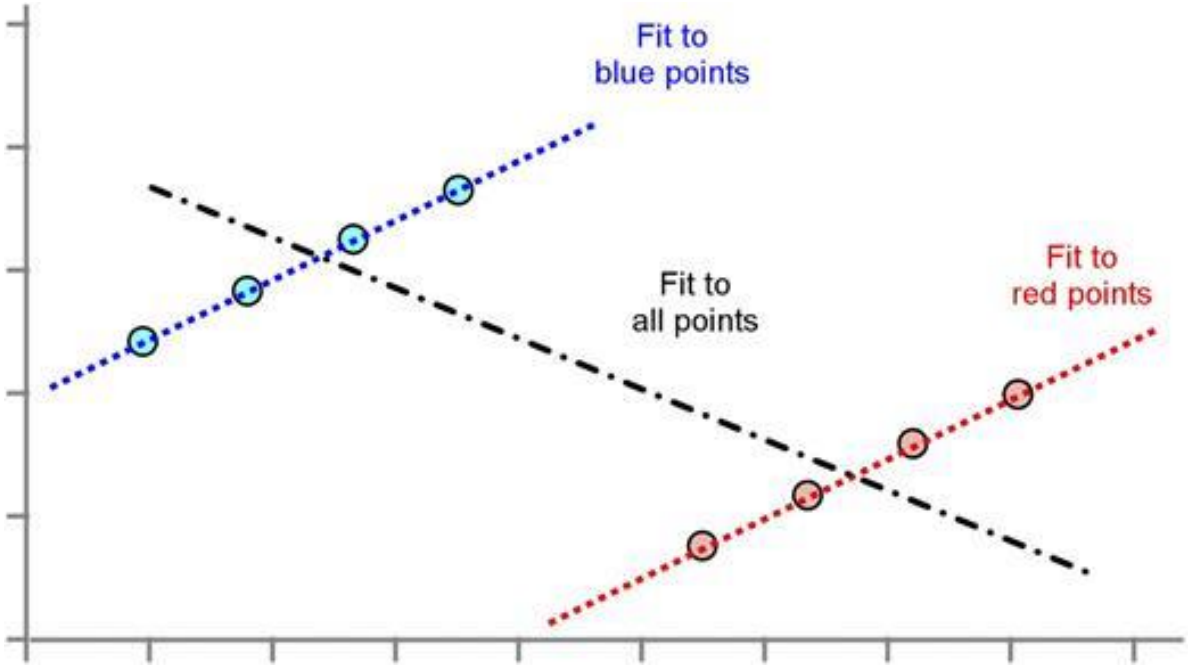
	Applicants	Admitted
Men	8442	**44%**
Women	4321	35%

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

Simpson's Paradox

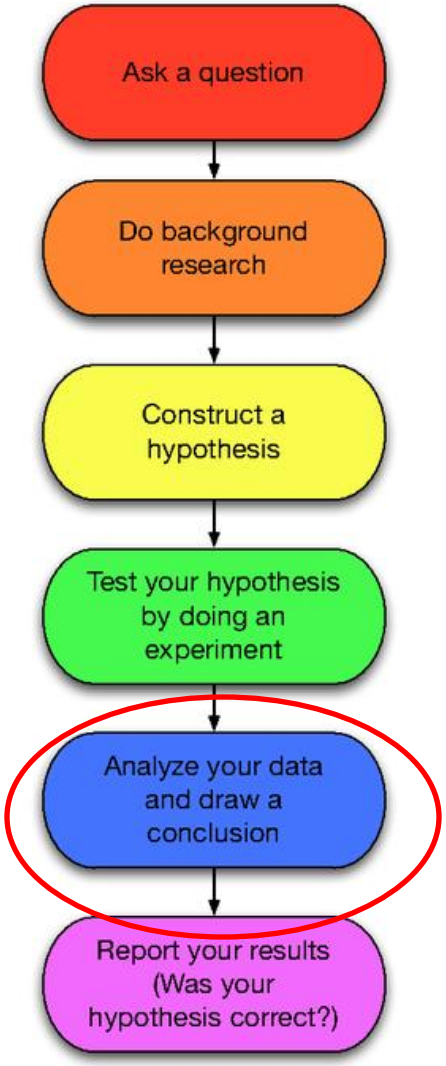


- Direction of association depends on how we aggregate the data



https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

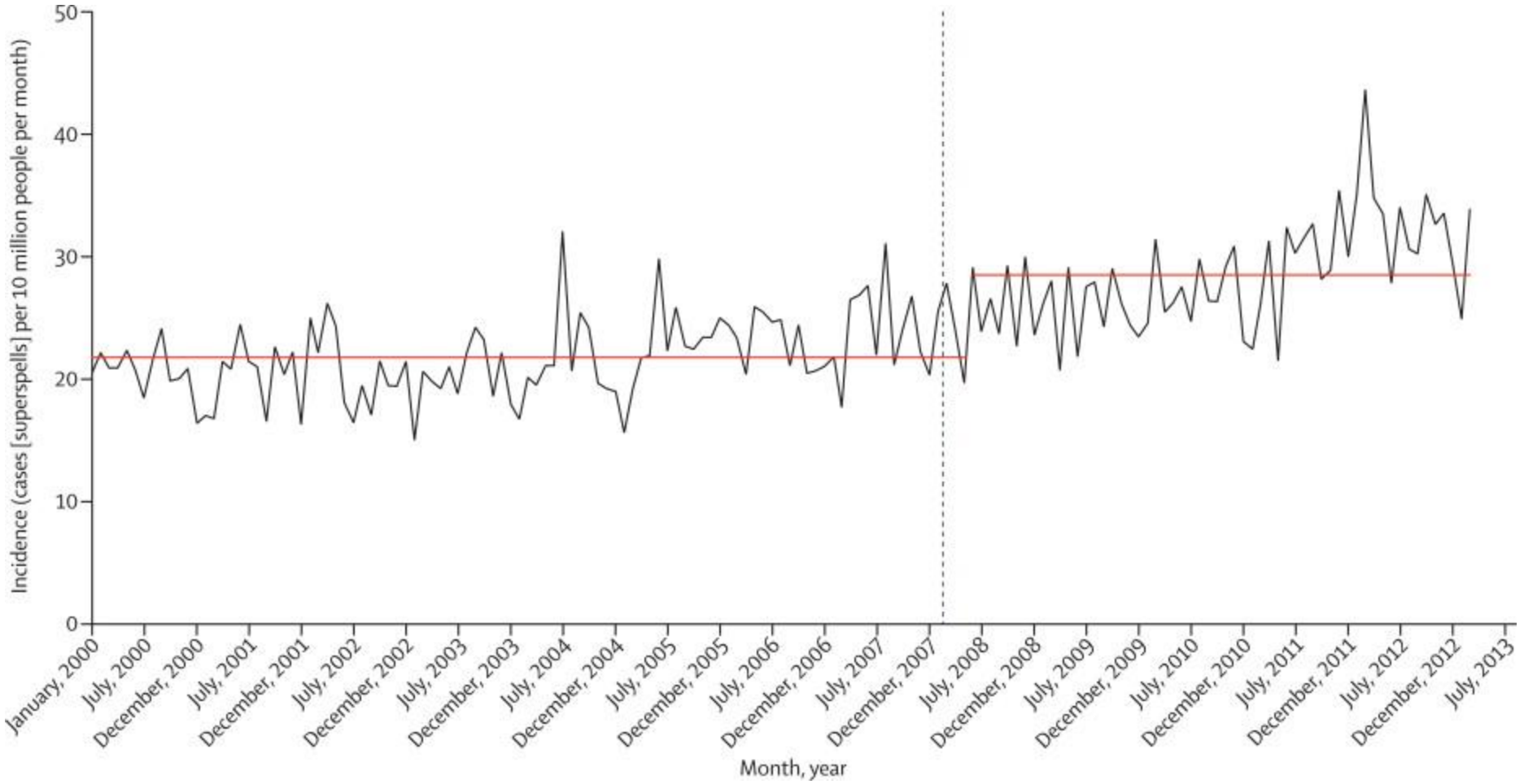
Analyzing data to draw a conclusion



- Suppose we collect data for several years before and after an intervention to examine how much difference it makes in outcomes
- Quasi-experimental design: Use control groups to confirm no change without intervention

https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

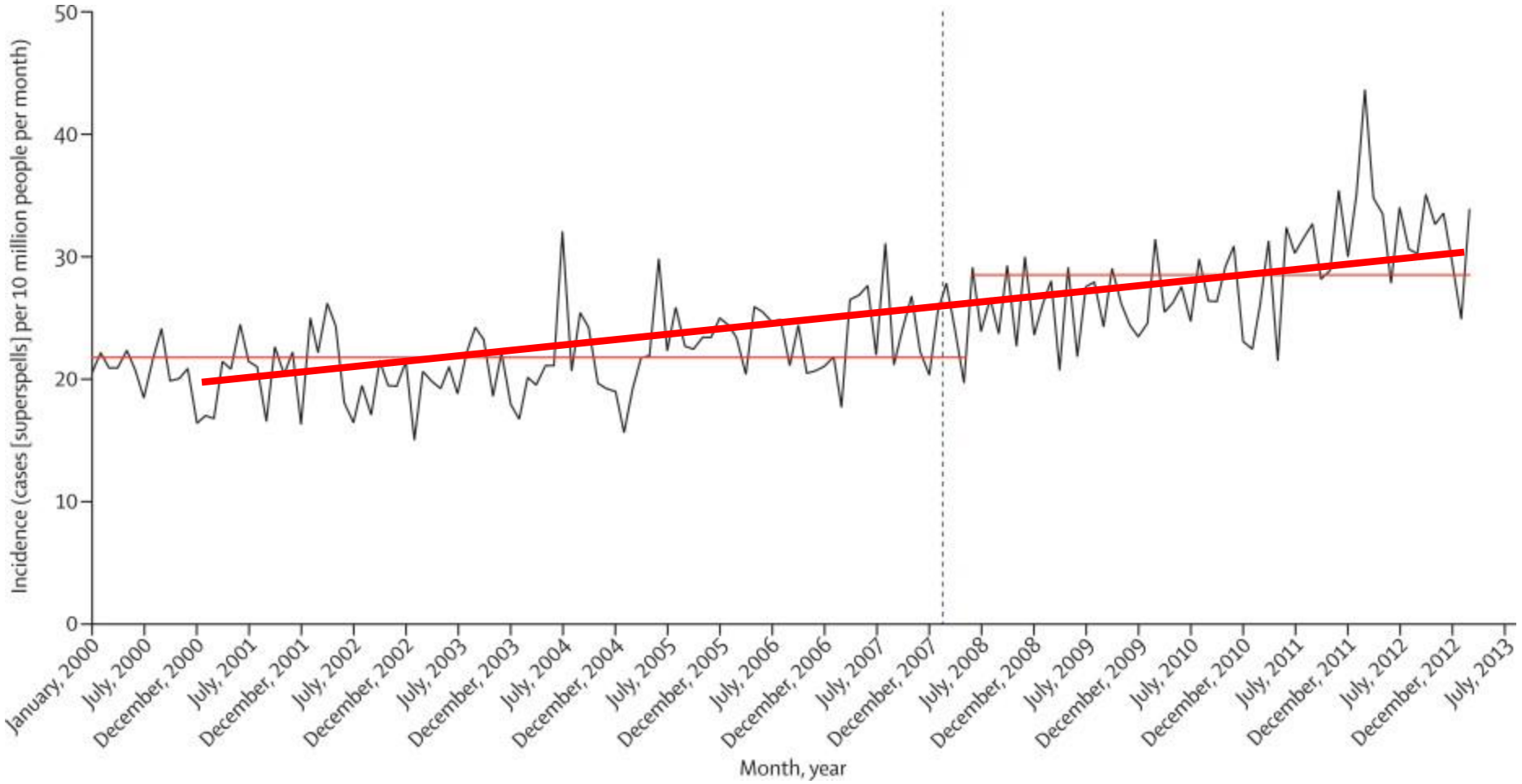
How did U.K. National Institute for Health and Clinical Excellence (NICE) recommendation of complete cessation of antibiotic prophylaxis for prevention of infective endocarditis in March, 2008 affect incidence of infective endocarditis?



www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)62007-9/fulltext?rss=yes

Different models yield different conclusions.
So, *how to deal with model uncertainty*?

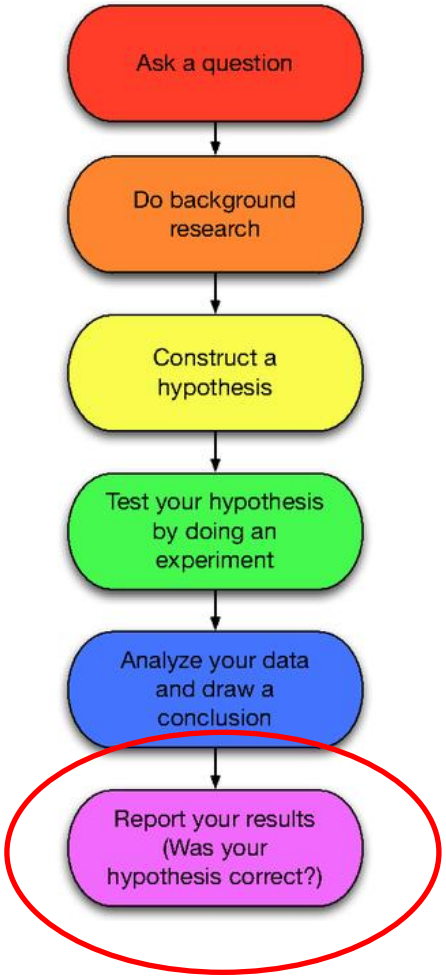
Technical solution: Model ensembles

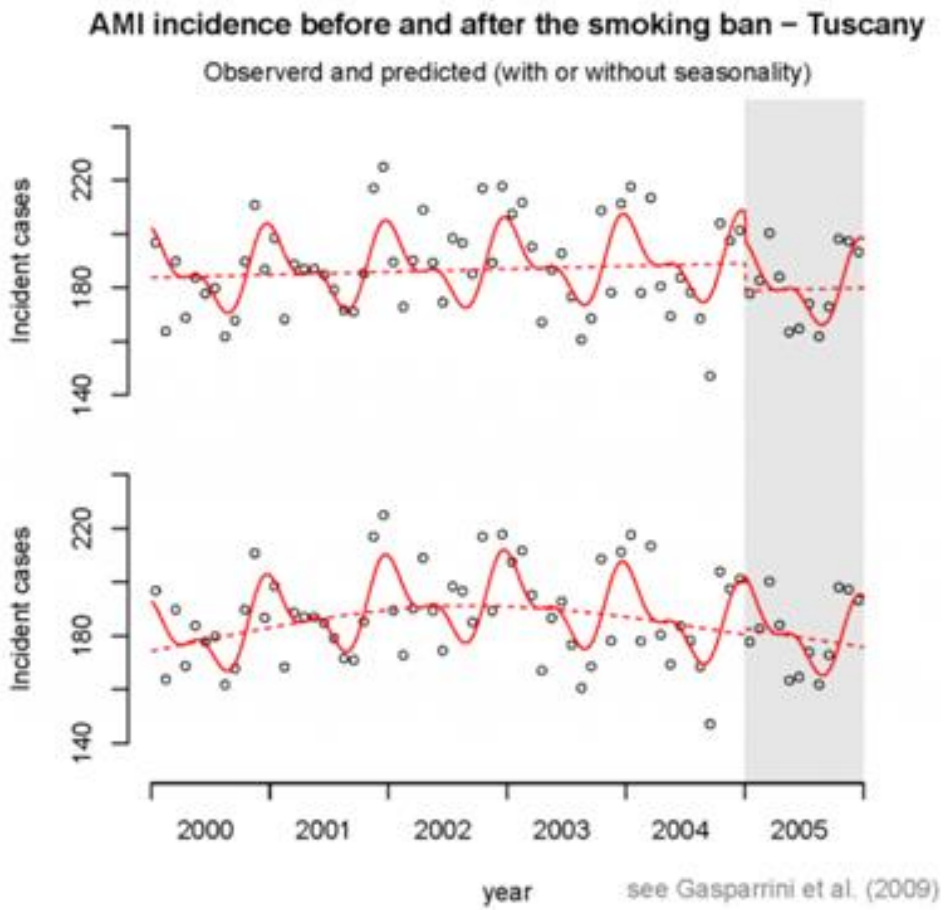


www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)62007-9/fulltext?rss=yes

Nonlinear models complicate inference of intervention effects

Technical solution: Non-parametric model ensembles



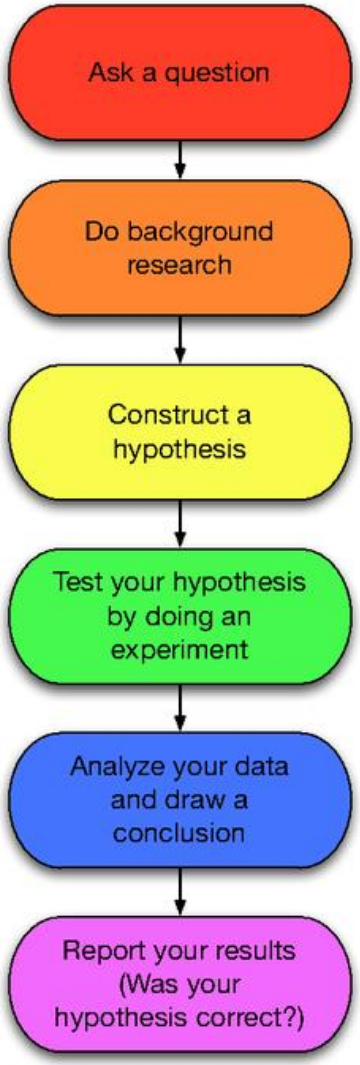


Results depend on modeling choices

Lessons

- The “statistically significant” results that are reported may depend on modeling choices
- Different modeling choices often give different (and even opposite) results
- Usually, only one set of modeling choices and results is reported
- Technically, this is no longer necessary.
  - Ask about results from non-parametric model ensembles. (Do not settle for sensitivity analyses or best-fitting models.)

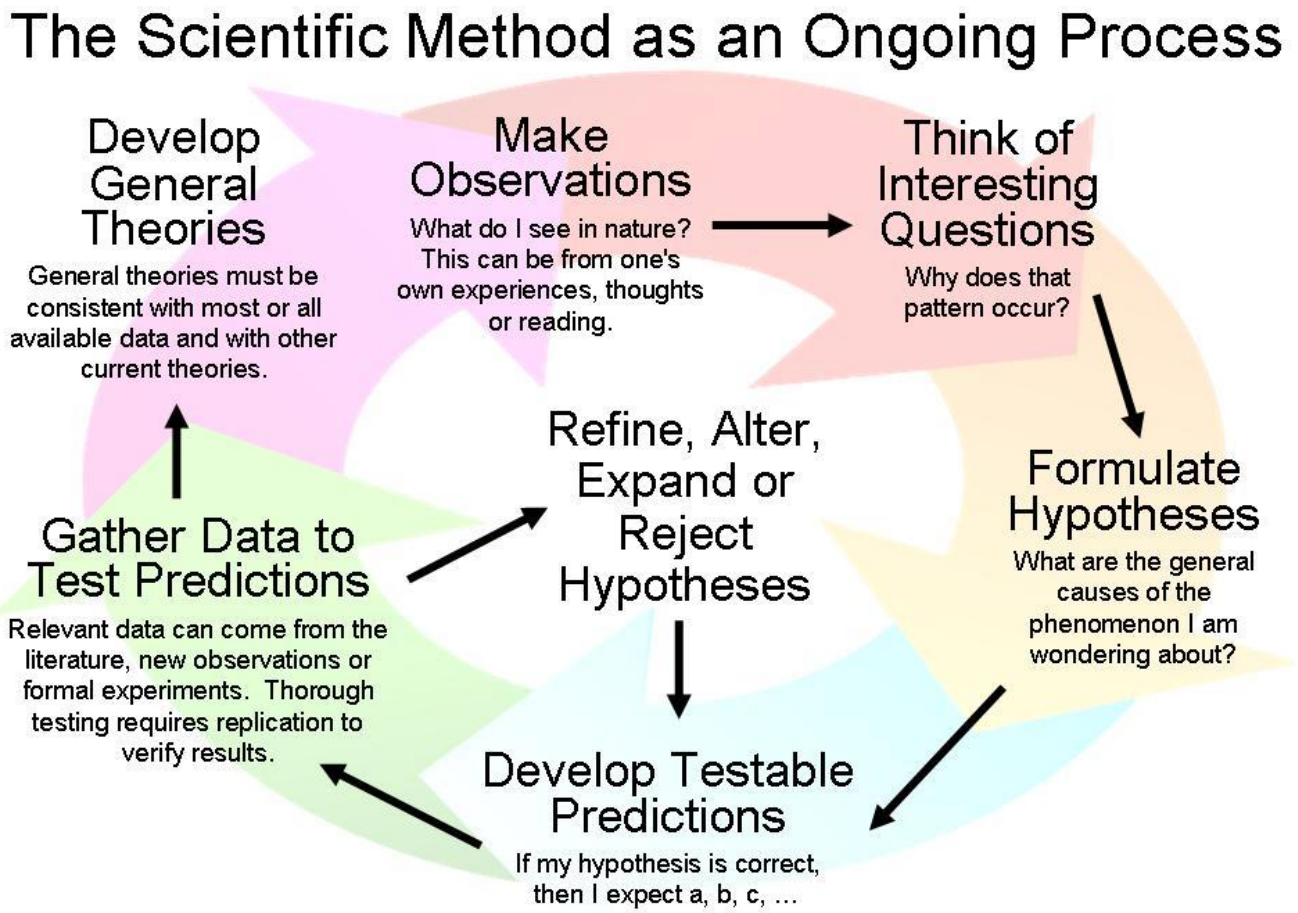
Congratulations on making it through one iteration!



- A true scientist's lot is not a happy one
- Much work, ambiguous data, usually weak and ambiguous conclusions
- Occasional surprises, breakthroughs, and definitive answers are rare and wonderful
- Pseudo-science is much easier, more common, and more gratifying to those impatient for sensational results
- Statistics can importantly help (or harm) at every step

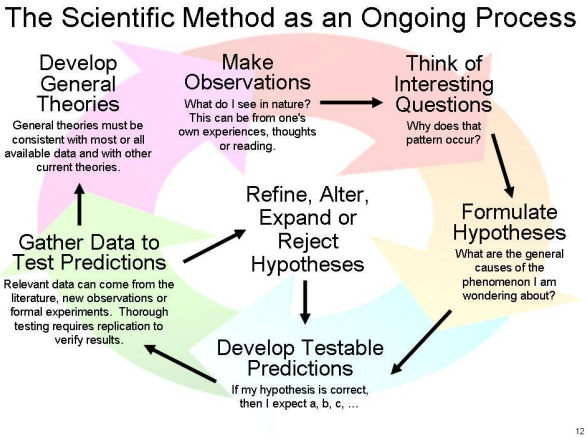
https://sites.google.com/a/wcpss.net/zelenakas-shs/earth-and-environmental-science/unit-1-introduction-scientific-method

In reality, iterations would now begin



http://healthjournalism.org/blog/tag/scientific-method/

We might be doing it all wrong



- Formulating and testing hypotheses is not necessarily the best (quickest, most objective, least error-prone) way to discover what is true
- Modern causal discovery algorithms (machine-learning, AI) typically make little use of hypothesis testing and much of invariance, constraints, and conflict resolution

http://healthjournalism.org/blog/tag/scientific-method/

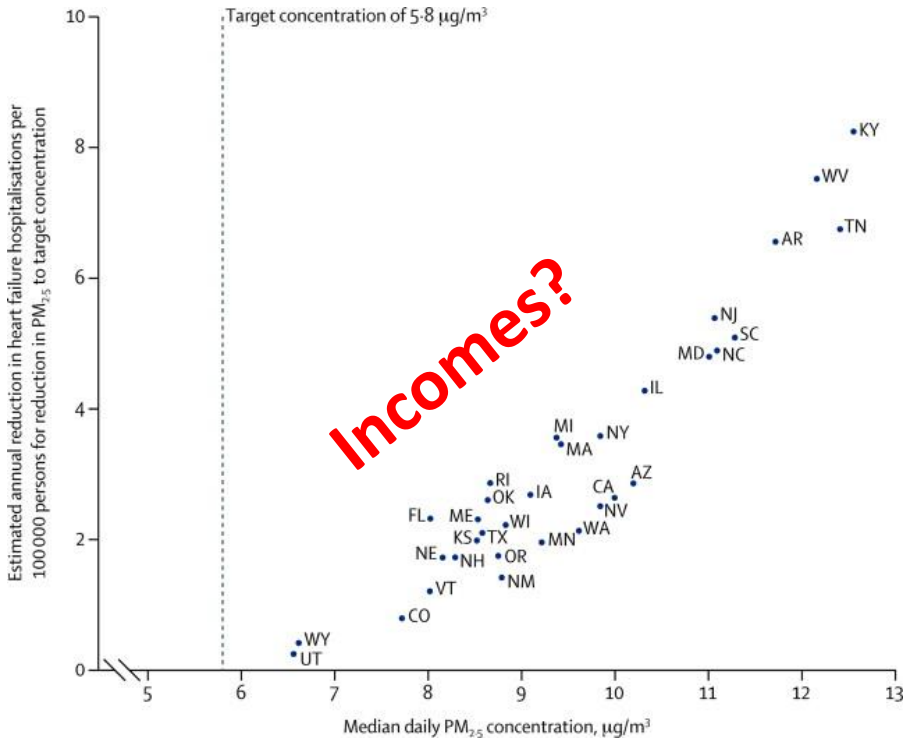
How can we harm science?

- Fund (only) the groups that most consistently and productively generate politically desired answers
  - Use p-hacking to guarantee desire results

- Publish assumptions, advocacy, ideology, and unjustified conclusions masquerading as science

Be skeptical of over-interpreted findings

Estimated benefit (decreased heart failure hospitalizations) from tighter PM2.5 regulation)



Example: Shah et al., 2013, *Lancet* meta-analysis

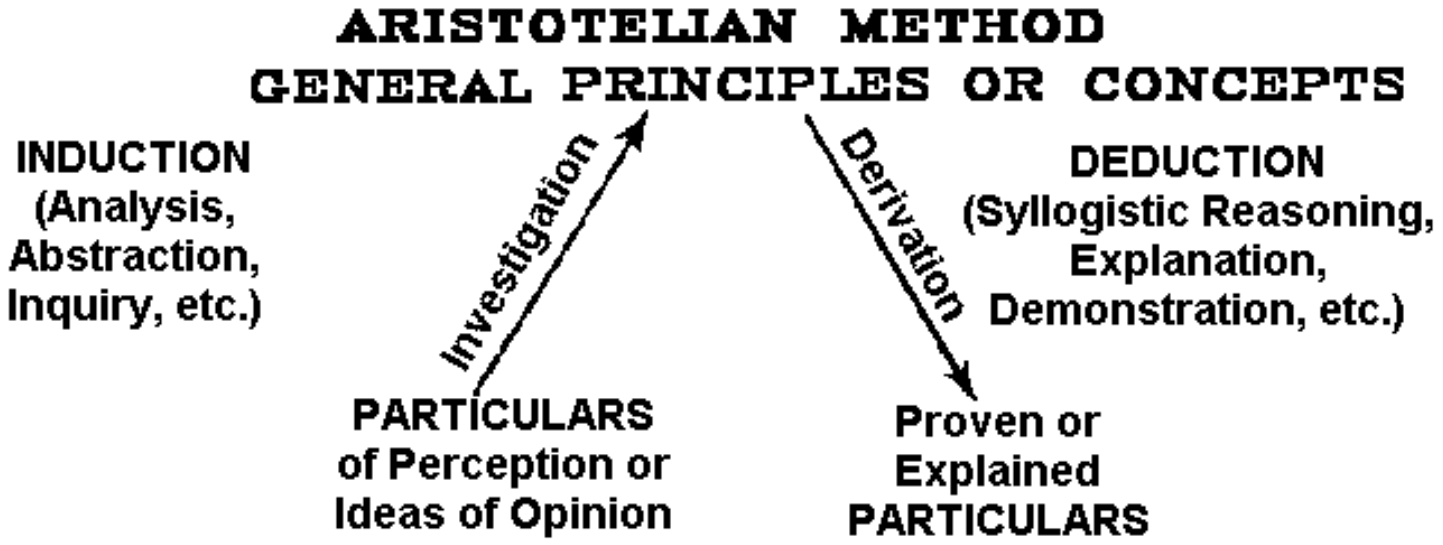
"Findings: Increases in particulate matter concentration were associated with heart failure hospitalisation or death (PM2·5 2·12% per 10 µg/m3, 95% CI 1·42–2·82… In the USA, we estimate that a mean reduction in PM2·5 of 3·9 µg/m3 would prevent 7978 heart failure hospitalisations and save a third of a billion US dollars a year."

Scientific method: Vision

- A reliable process for discovering objective scientific truth from independently reproducible data and experiments
- Application: Predict consequences of actions

Scientific method: Vision

- A reliable process for discovering objective scientific truth from data and reproducible experiments
- Application: Predict consequences of actions

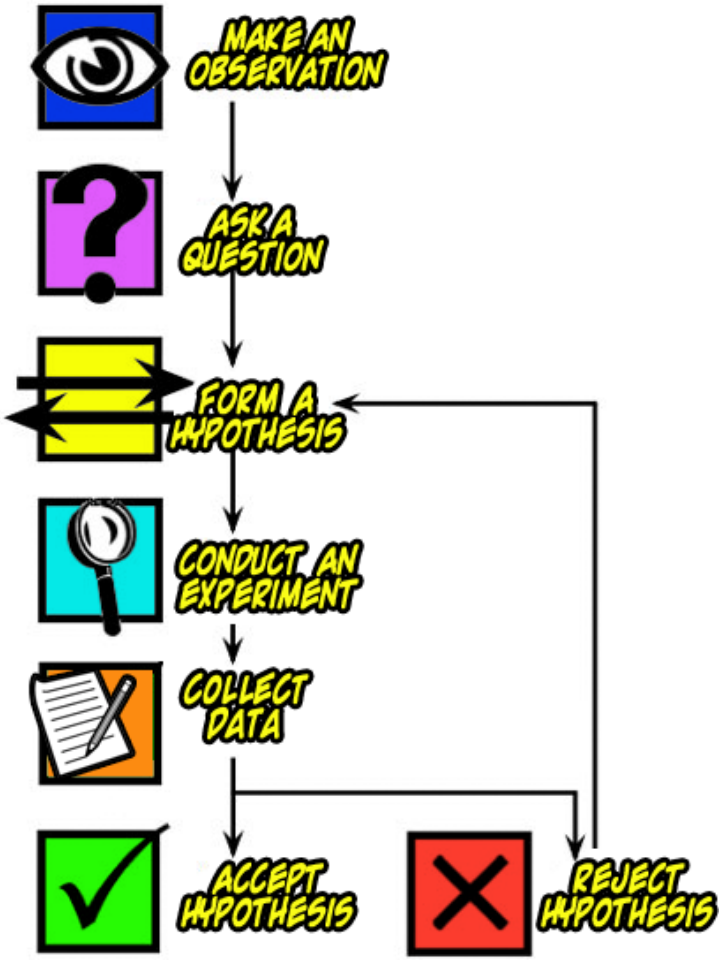


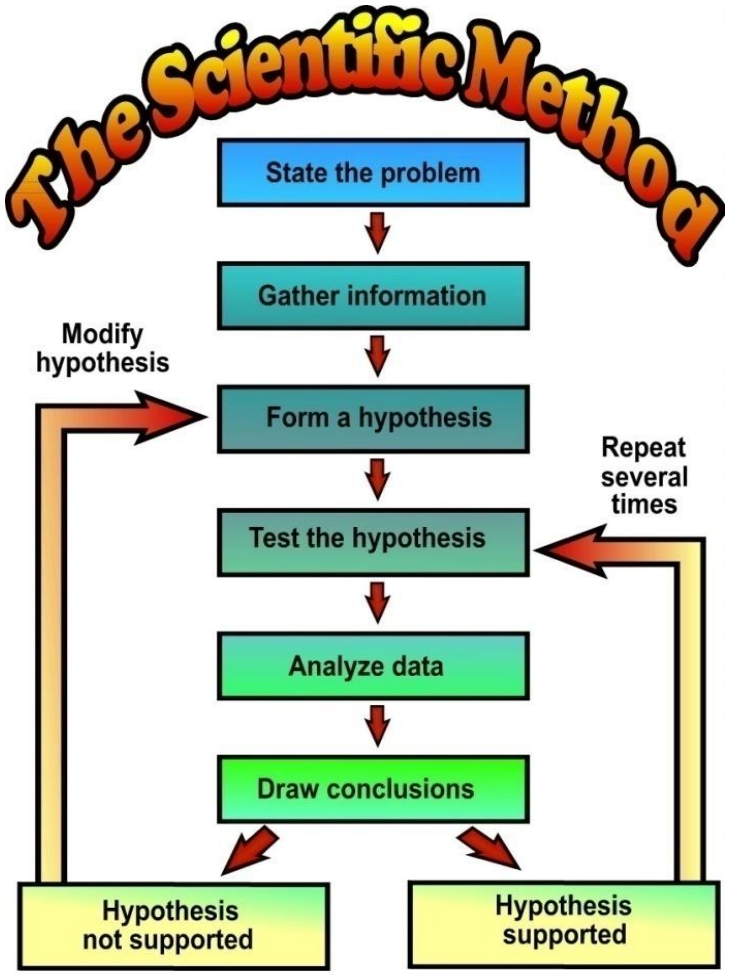
http://stillunfold.com/science/top-10-contributions-of-aristotle-to-science

Scientific method: Vision

- A reliable process for discovering objective scientific truth from data and reproducible experiments
  - Causal realism: The truth is out there!
  - Causal laws, mechanisms, networks
- Discover what causes what, and how
  - Events, conditions, probabilities
  - Causal explanations, predictions, effects
- Objective, self-correcting: Learn from reality
  - Falsifiable theories, testable predictions

The scientific method (A and B)

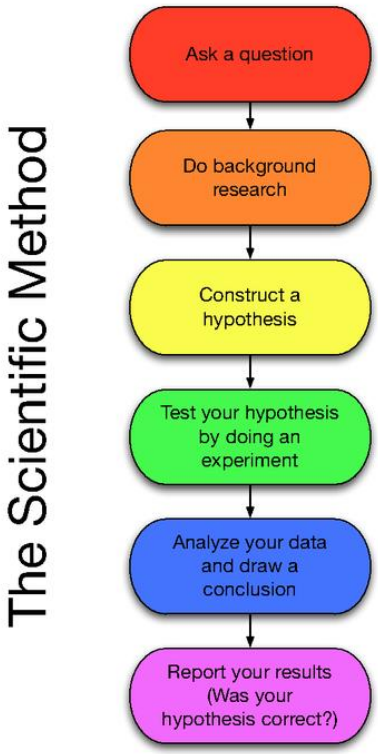


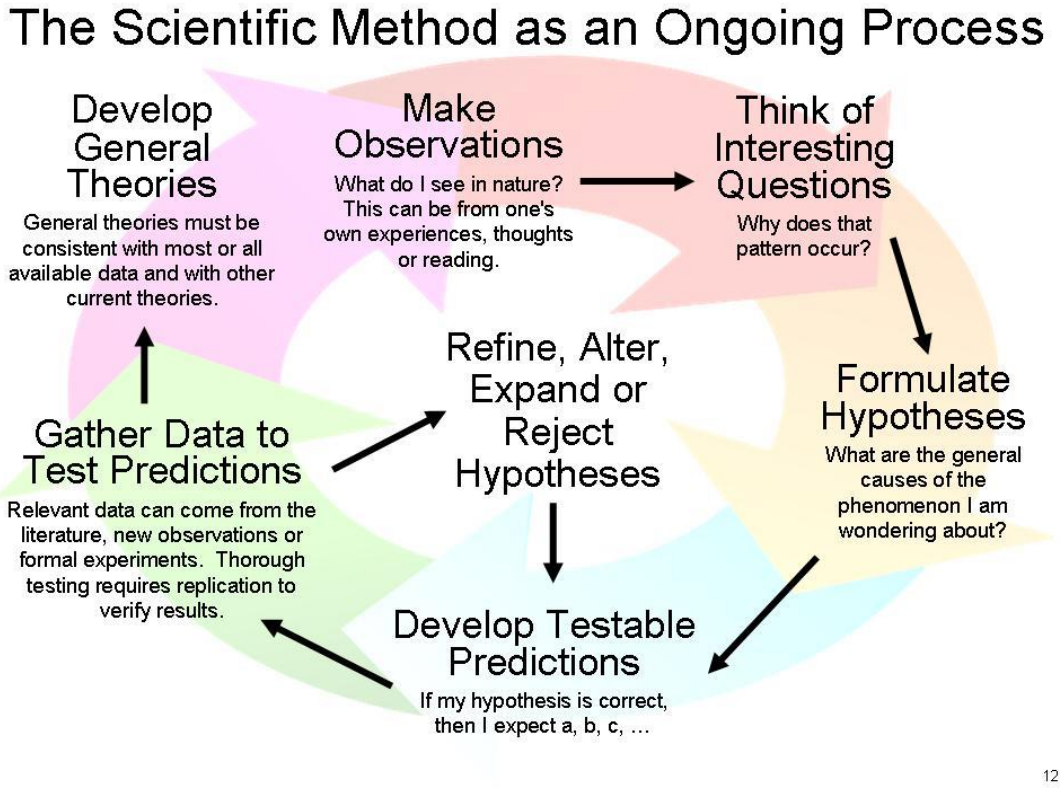


www.irc.vbschools.com/ForTheWeb/Science/pages/ScientificMethod_jpg.htm

https://students.ga.desire2learn.com/d2l/lor/viewer/viewFile.d2lfile/1798/12842/scientific-methods_print.html

The scientific method (C and D)





https://www.thinglink.com/scene/427479475416989696

http://healthjournalism.org/blog/tag/scientific-method/

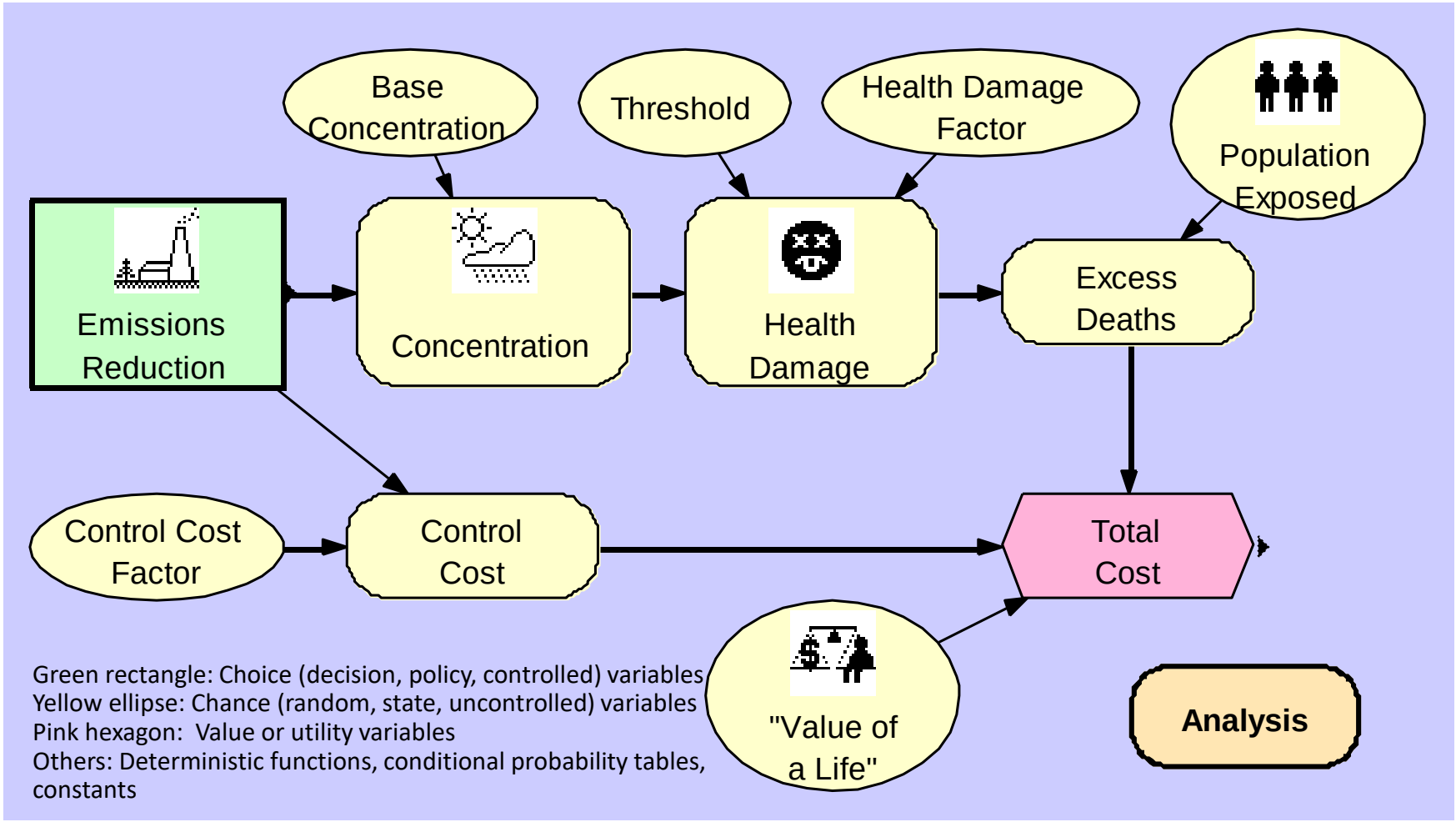
Causal analytics informs the rest of the policy analytics cycle

Analytics Goal: Discover how to act more effectively

1. **Descriptive analytics:** What's happening? What's new? How have causes or effects changed? What to worry about?
2. **Predictive analytics:** What will (probably) happen next if we don't change what we're doing?
3. **Causal analytics:** What can we do about it? What will (probably) happen next if we do things differently?
4. **Prescriptive analytics:** What should we do?
5. **Evaluation analytics:** How well is it working?
6. **Learning analytics:** How might we do better?
7. **Collaborative analytics:** How might we do better together?

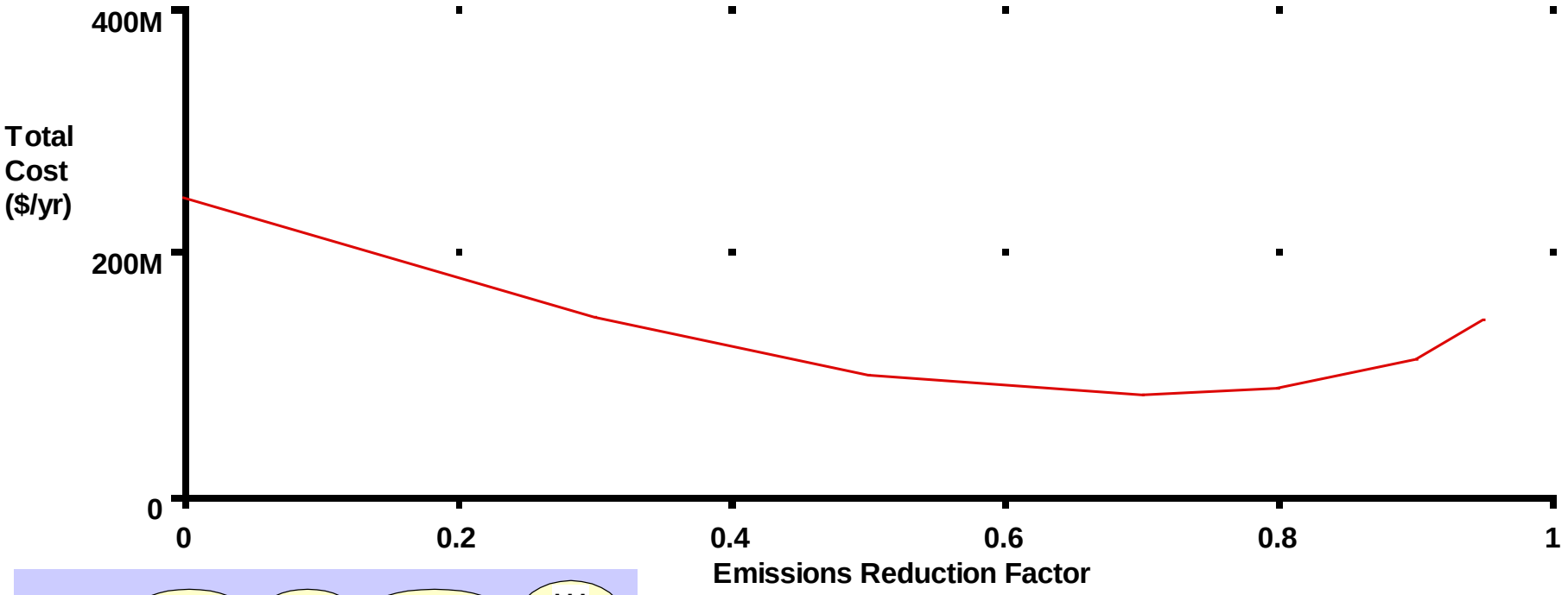
Example: *Analytica* Influence Diagram (ID) causal model

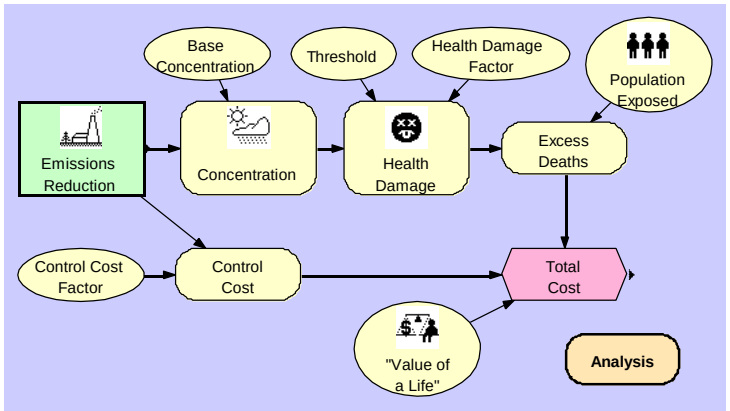
Total Cost to society = *Control Cost* of *Emissions Reduction* + *"Value of a Life"*Excess Deaths*



http://www.lumina.com/products/free101

Analytica output: Clicking on "Total Cost" value node and selecting mean value yields:





Simulation-based partial dependence plot: Decision variable is varied over a range of alternative (counterfactual) values. Other variables are drawn from distributions, for each value of the decision variable.

Causal analytics

1. What works? (policies, interventions, acts)
   - How well? (effect size estimation)
   - For whom, under what conditions? (effects of covariates)
2. How do changes in inputs affect outputs?
   - Causal explanation, mediation analysis, path analysis
3. What might work better? (trials, learning)
   - What is the best achievable result? (optimization)
   - What will happen if we make changes? (causal prediction)
   - How sure can we be? (uncertainty analysis)
4. How to cause desired changes? (decision analysis)
5. What information would improve answers? (value of information (VOI) analysis)

Causal analytics informs the rest of the policy analytics cycle

Analytics Goal: Discover how to act more effectively

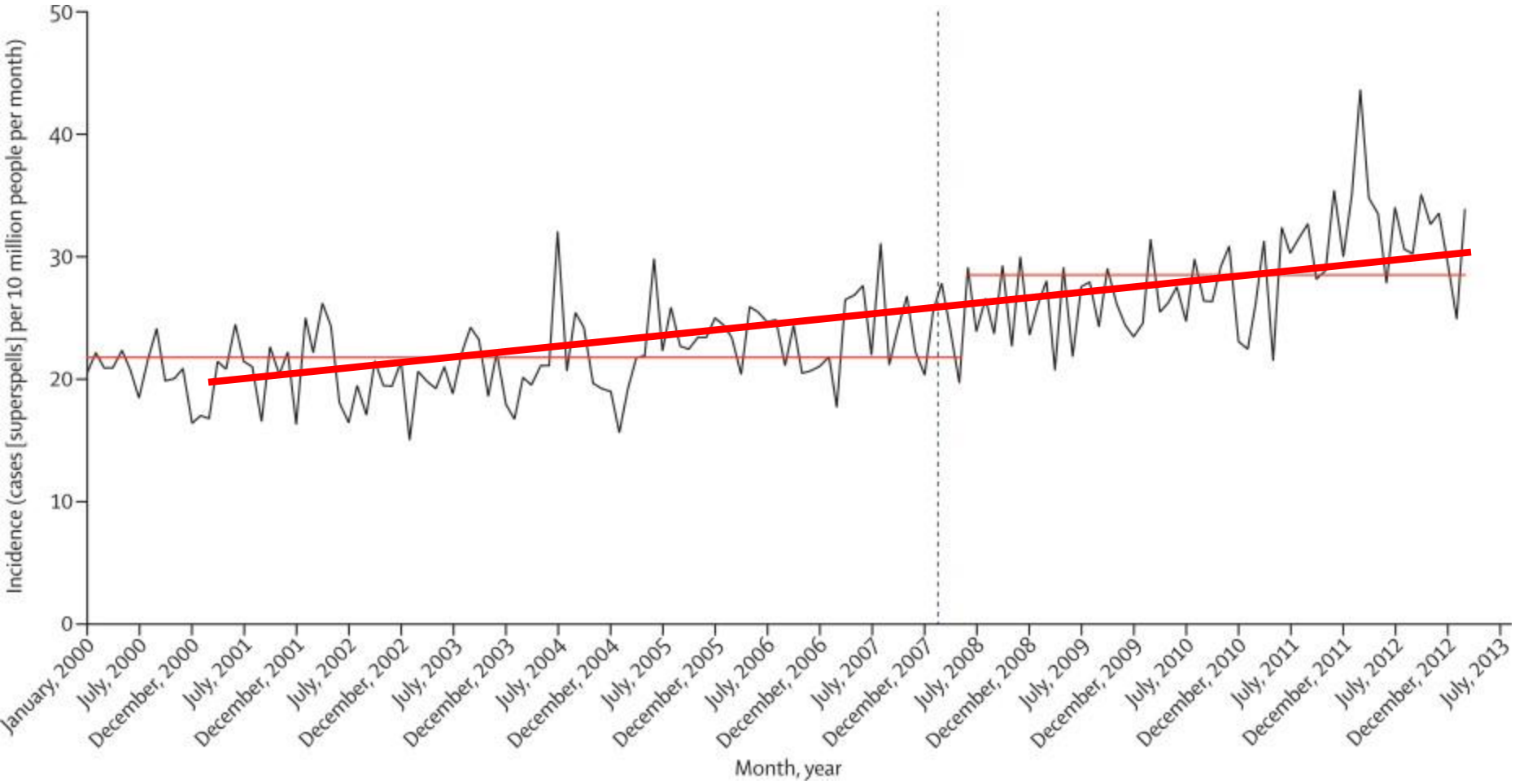
1. **Descriptive analytics:** What's happening? What's new? How have causes or effects changed? What to worry about?
2. **Predictive analytics:** What will (probably) happen next if we don't change what we're doing?
3. **Causal analytics:** What can we do about it? What will (probably) happen next if we do things differently?
4. **Prescriptive analytics:** What should we do?
5. **Evaluation analytics:** How well is it working?
6. **Learning analytics:** How might we do better?
7. **Collaborative analytics:** How might we do better together?

Evaluation analytics: How well are policies working?

- Algorithms for evaluating effects of actions, events, conditions
  - Intervention analysis/interrupted time series
    - Key idea: Compare predicted outcomes with no action to observed outcomes with it
      - Counterfactual causal analysis
      - Google's new CausalImpact algorithm
- Quasi-experimental designs and analysis
  - Refute non-causal explanations for data
  - Compare to control groups to estimate effects

Different models yield different conclusions.
So, *how to deal with model uncertainty*?

Solution: Model ensembles, Bayesian Model Averaging (BMA)



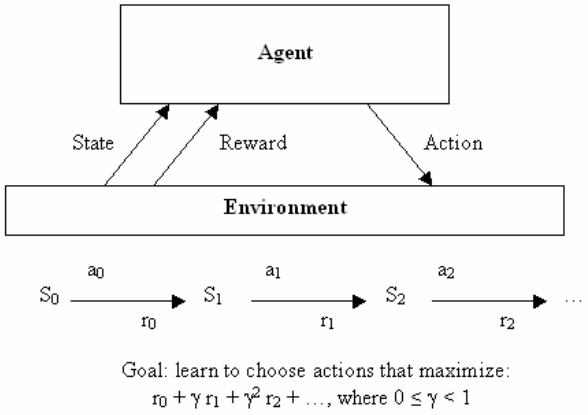
www.thelancet.com/journals/lancet/article/PIIS0140-6736(14)62007-9/fulltext?rss=yes

Algorithms for evaluating effects of combinations of factors

- Classification trees
  - Boosted trees, Random Forest, MARS
- Bayesian Network algorithms
  - Discovery
    - Conditional independence tests
  - Validation (e.g., train-test, cross-validation)
  - Inference and explanation
- Response surface algorithms
  - Adaptive learning, design of experiments

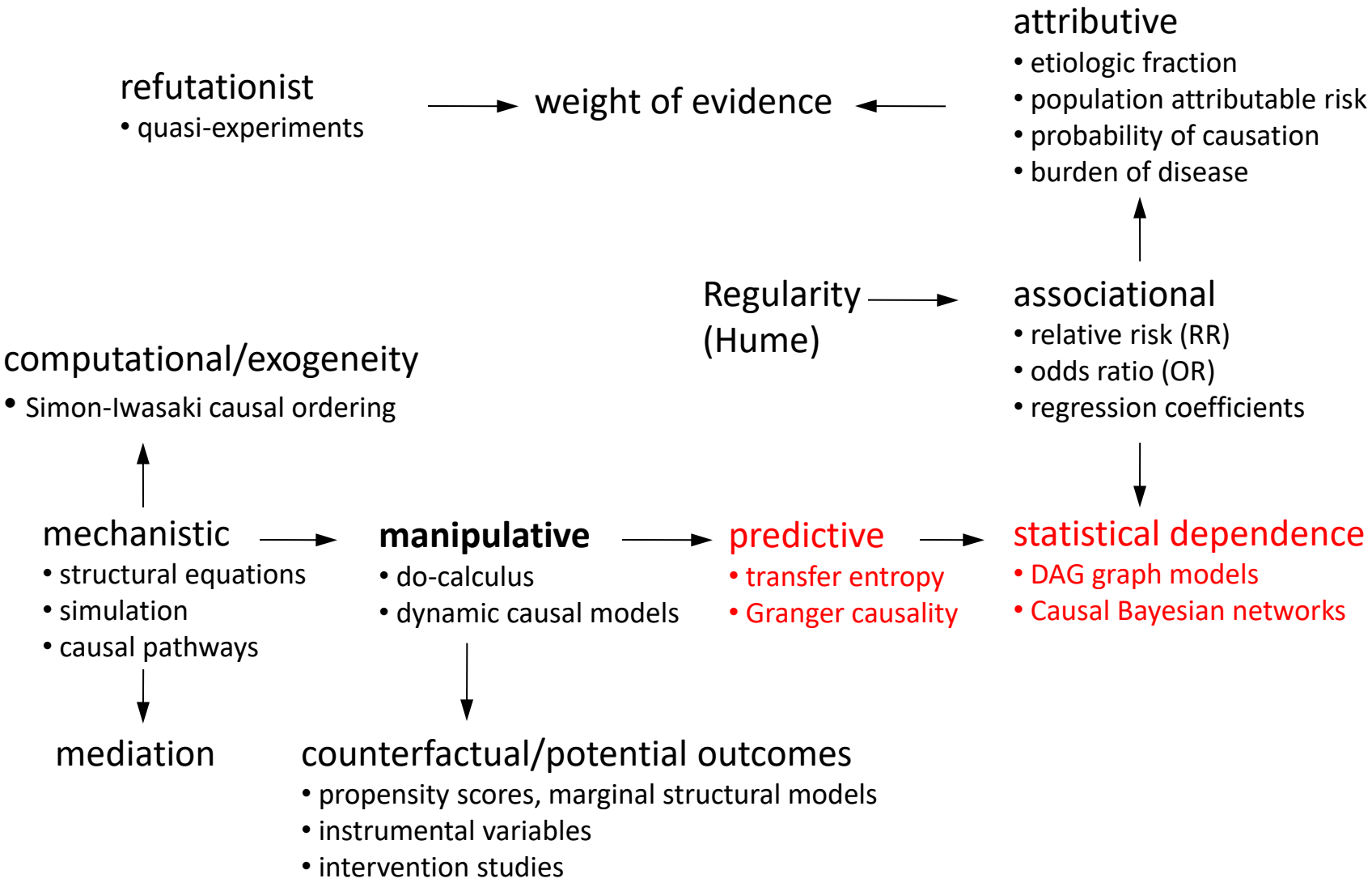
Learning analytics

- Learn to predict better
  - Create ensemble of models, algorithms
    - Use multiple machine learning algorithms
      - Logistic regression, Random Forest, SVM, ANN, deep learning, gradient boosting, KNN, lasso, etc.
  - "Stack" models (hybridize multiple predictions)
    - Cross-validation assesses model performance
      - Meta-learner combines performance-weighted predictors to produce an improved predictor
    - Theoretical guarantees, practical successes (Kaggle competitions)
- Learn to decide better
  - Low-regret learning of decision rules
    - Theoretical guarantees (MDPs)
    - Practical performance



http://www2.hawaii.edu/~chenx/ics699rl/grid/rl.html

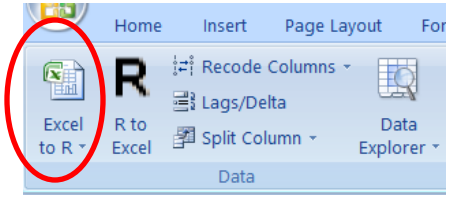
Implications among types of causation: Manipulative implies predictive but not associational



Key methods for causal inference from time series and longitudinal data

- Quasi-experiments
- Panel data analysis
- Granger causality testing and generalizations
- Intervention analysis
- Change point analysis
- Counterfactual/potential outcomes methods
- Causal network methods
- Negative controls

Air pollution example in CAT*



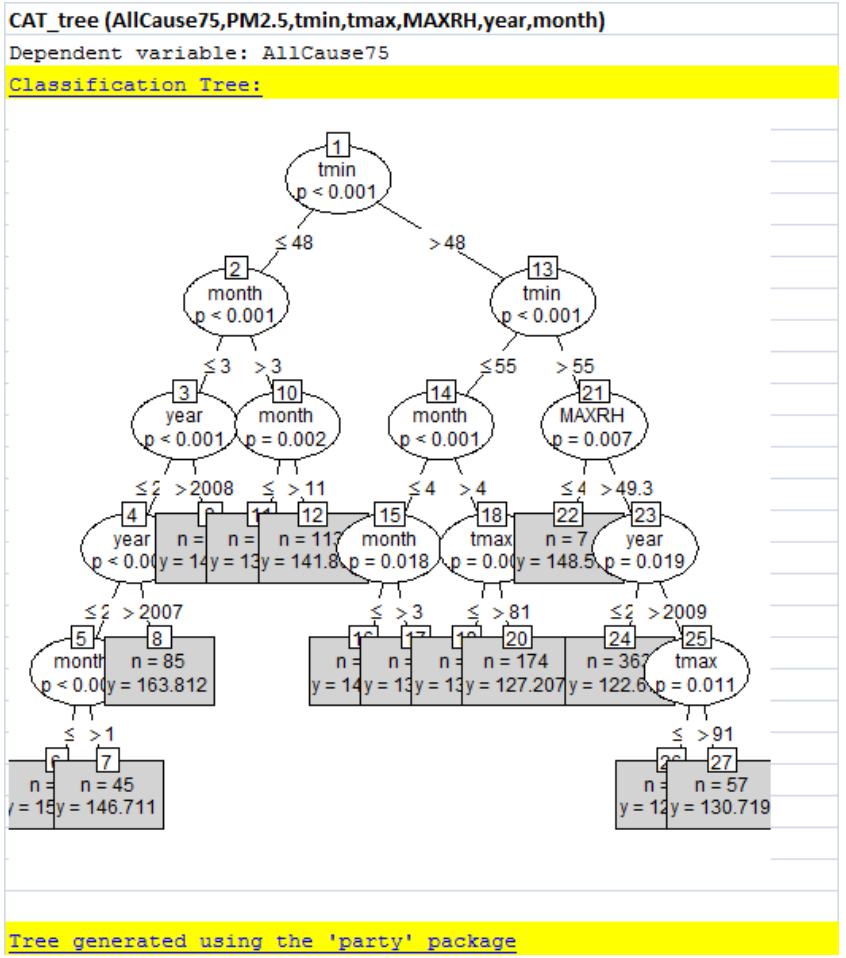
- Load data in Excel, click *Excel to R* to send it to R
  - Los Angeles air basin
  - 1461 days, 2007-2010 (Lopiano et al., 2015, thanks to Stan Young for data)
  - PM2.5 data from CARB
  - Elderly mortality ("AllCause75") from CA Department of Health
  - Daily min and max temps & max relative humidity from ORNL and EPA

year	month	day	AllCause75	PM2.5	tmin	tmax	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	160	19.1	41	76	40.9
2007	1	8	148	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4

Risk question: Does PM2.5 exposure increase elderly mortality risk? If so, by how much?

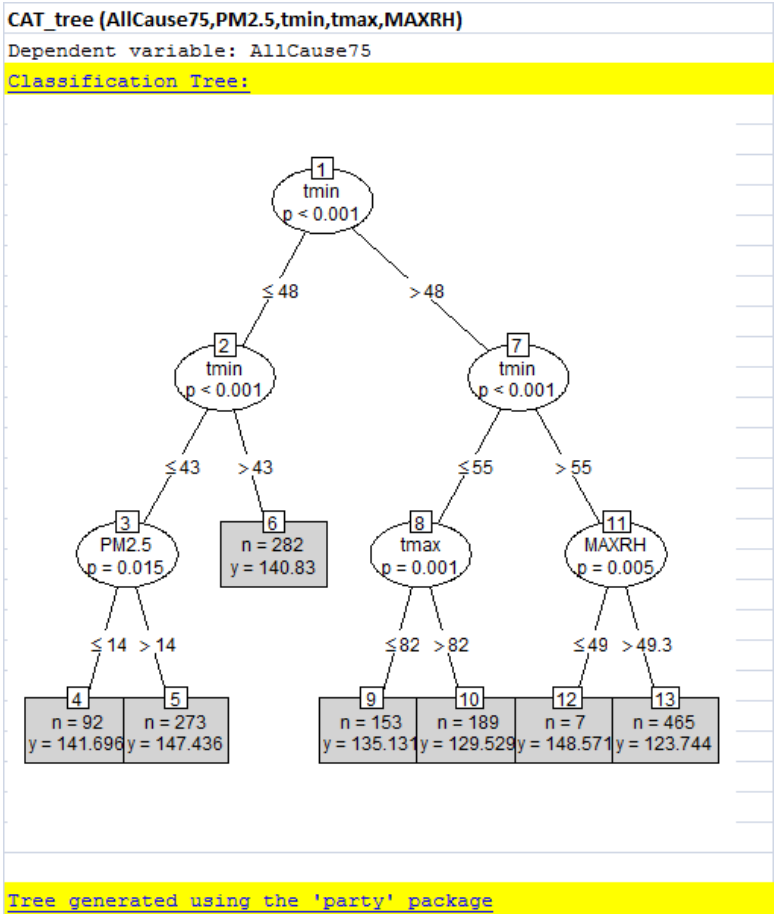
Air pollution example: Classification tree descriptive analytics

- tmin, tmax, month, year, MAXRH are potential predictors of AllCause75 (elderly mortality)
- PM2.5 does not appear in this tree
  - AllCause75 is *conditionally independent* of PM2.5 in this analysis, given the other variables in the tree
- Making *year* and *month* into categorical variables changes the tree but not this conclusion.



How a CART tree works

- Basic idea: Always ask the most informative question next, given answers so far.
  - *Questions* are represented by *splits* in tree
  - *Leaf nodes* show conditional means (or conditional distributions) of dependent variable
  - Internal nodes show significance level for split: how significant are differences between conditional distributions
- Reduces prediction error for dependent variable
- Stop this "recursive partitioning" when further questions (splits in tree) do not significantly improve prediction.
  - Classification & Regression Tree (CART) algorithm
- Some refinements:
  - Grow a large tree and prune back to minimize cross-validation error
  - fit multiple trees to random subsets of data and let them vote for best splits ("bagging")
  - over-train on mis-predicted cases ("boosting")
  - average predictions from many trees ("RandomForest" ensemble prediction)
  - Join prediction "patches" together smoothly (MARS)



Generalizations: Ensembles of trees (Random Forest)

- Averaging over hundreds of trees gives more robust results, reduces prediction errors
  - Random Forest ensemble is “go to” black-box method for predictive analytics

CAT_importance (tmax,MAXRH,year,month,AllCause75,PM2.5)

Dependent variable: tmax

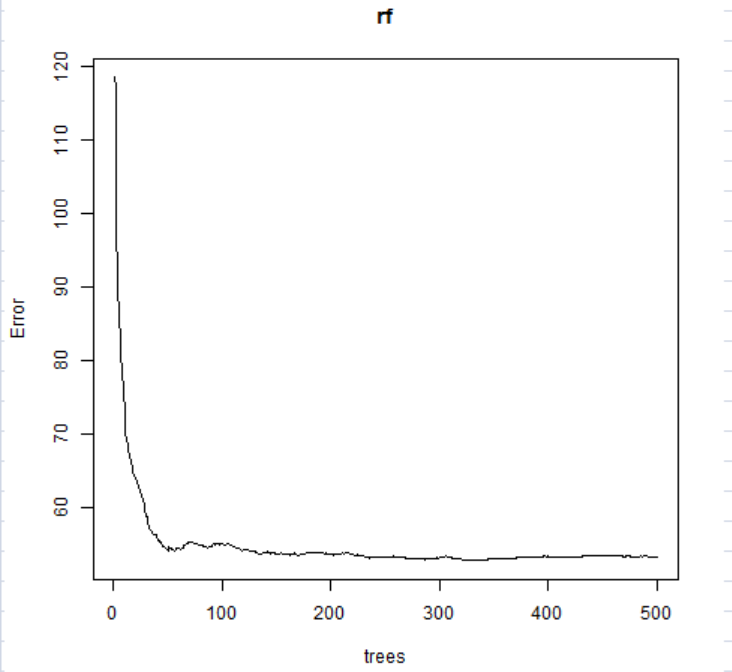
randomForest summary

No. of variables tried at each split: 1

Mean of squared residuals: 53.31128

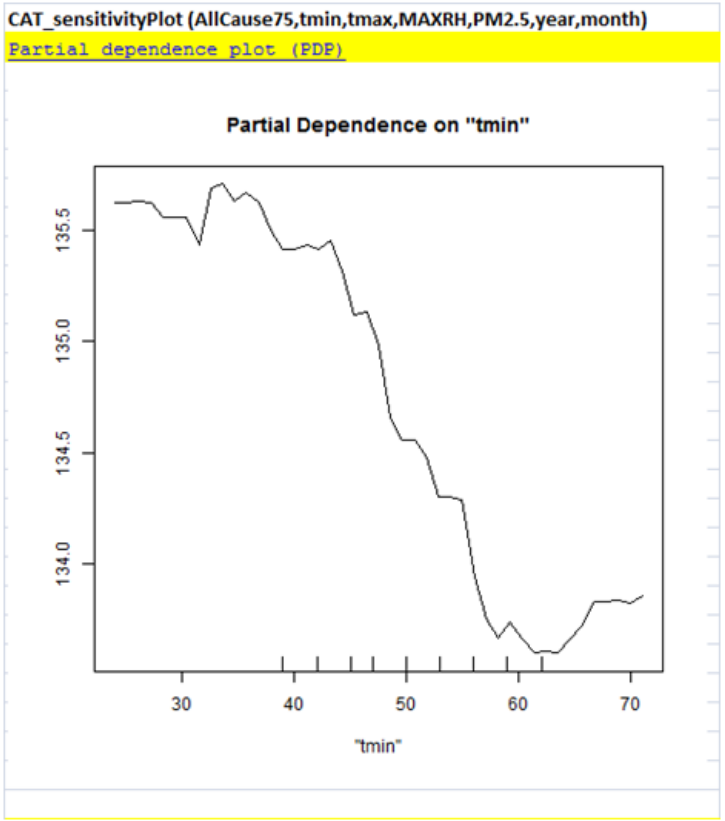
% Var explained: 69.36

randomForest accuracy vs. number of trees



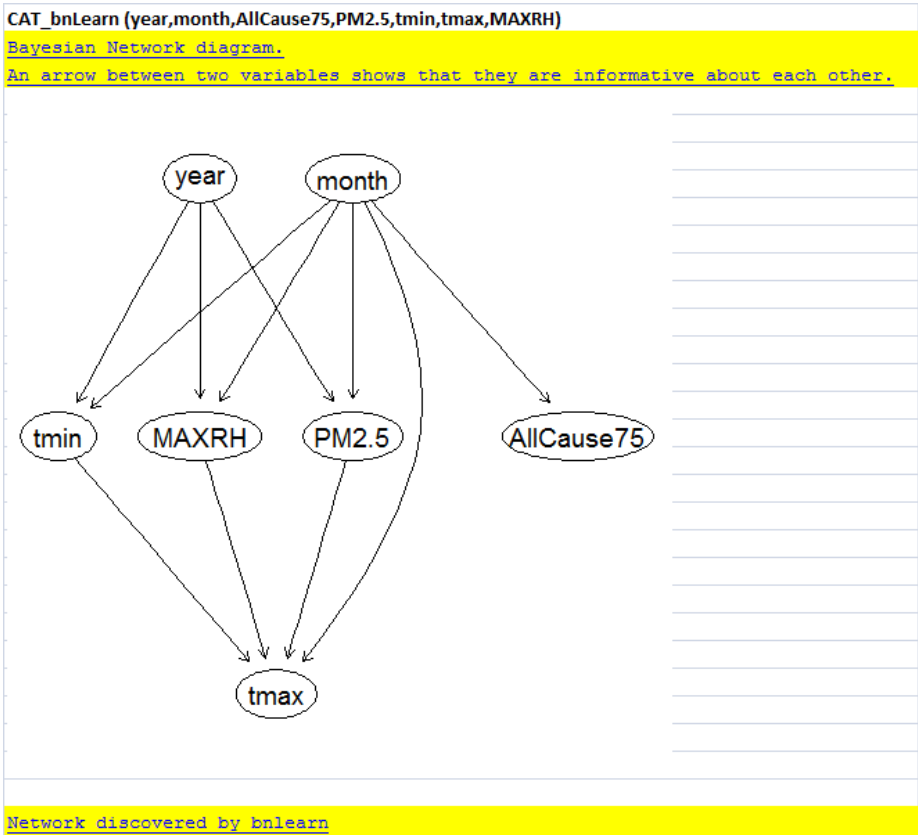
Generalizations: Ensembles of trees (Random Forest)

- Partial dependence plots summarize how dependent variable is predicted to change as one variable is changed, letting all other variables have their real values



Bayesian Networks (BNs) show information relations among variables

- BNs provides high-level roadmap for descriptive analytics
- Each node has a conditional probability table (CPT) (or regression model, CART tree, etc.) describing how the conditional probabilities of its values depend on other variables.
- If no arrow connects two variables, then they are *conditionally independent* of each other, given the other variables in the BN.
  - Omitted variables can create statistical dependencies
  - Conditioning on variables can also sometimes create dependencies
- Information principle for causality: Direct causes are not conditionally independent of their effects.

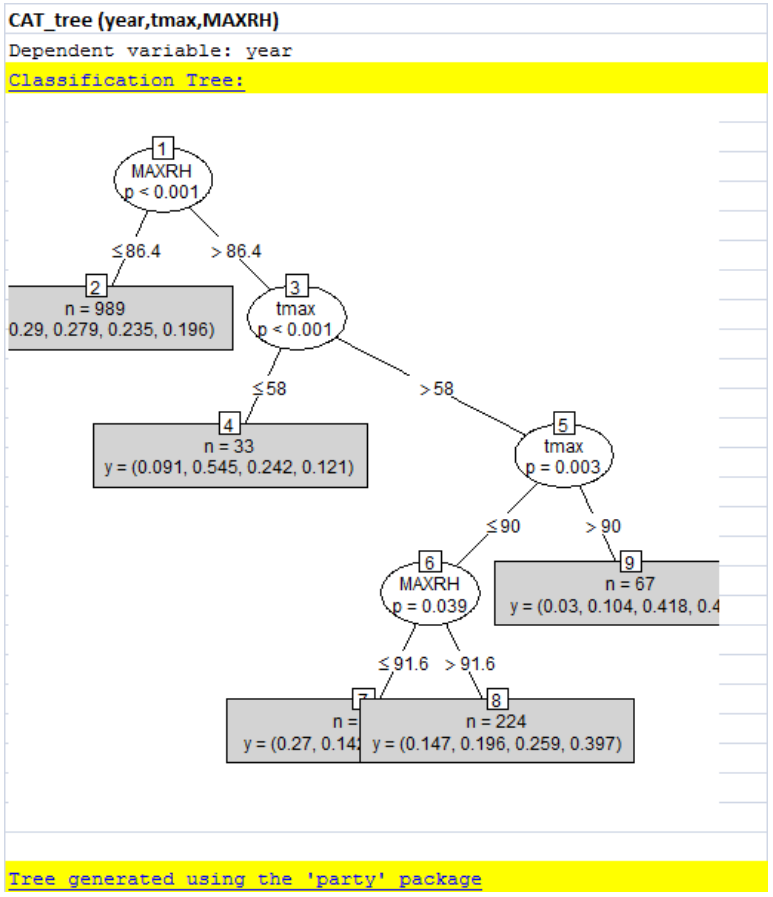


Automatically noticing and describing what matters

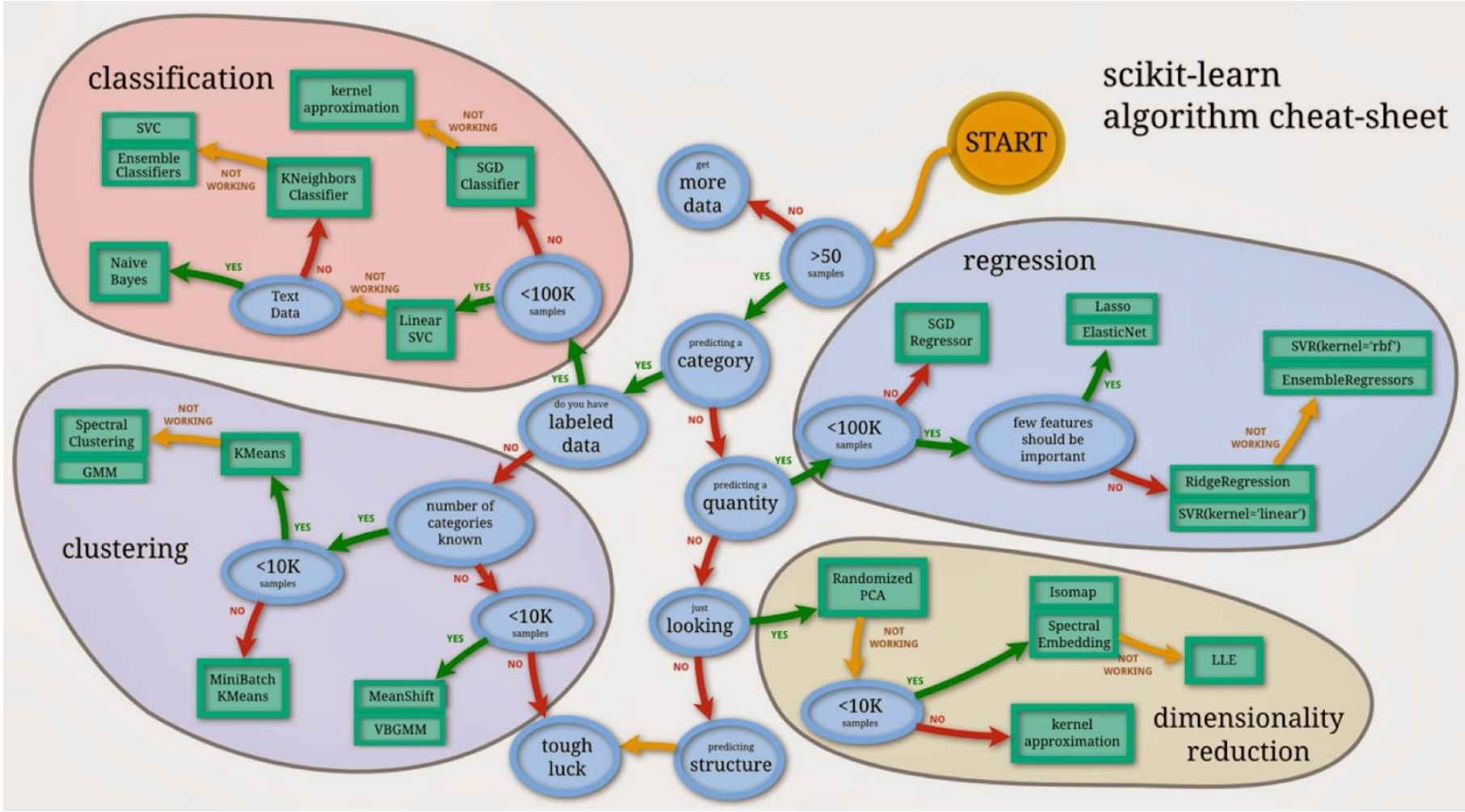
- Simple approach: Create binary indicator for "this period" vs. "recent periods"
- Treat indicator as dependent variable, find most parsimonious/best predictors in multivariate data
  - Show's what's different now
  - Highlights informative changes
- Embed key predictors in causal network model to explain and predict changes

Example: Change analysis of years 2007-2010

- Hot, high humidity days are more likely to occur in more recent years



Standard machine learning (ML) tools for descriptive analytics

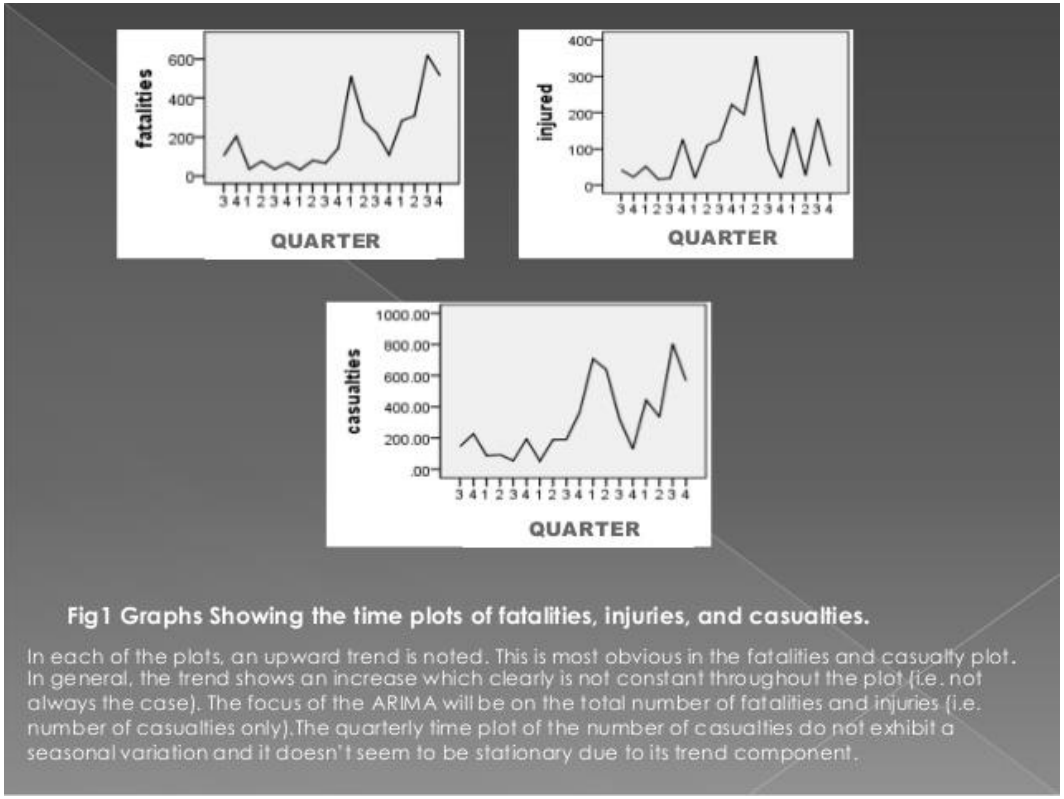


http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Predictive analytics

- What will happen if we do nothing?
- How sure can we be?

Example: Black-box ARIMA forecasting of losses due to terrorist attacks



http://www.slideshare.net/VictorOdutokun/arima-analysis-project-slide

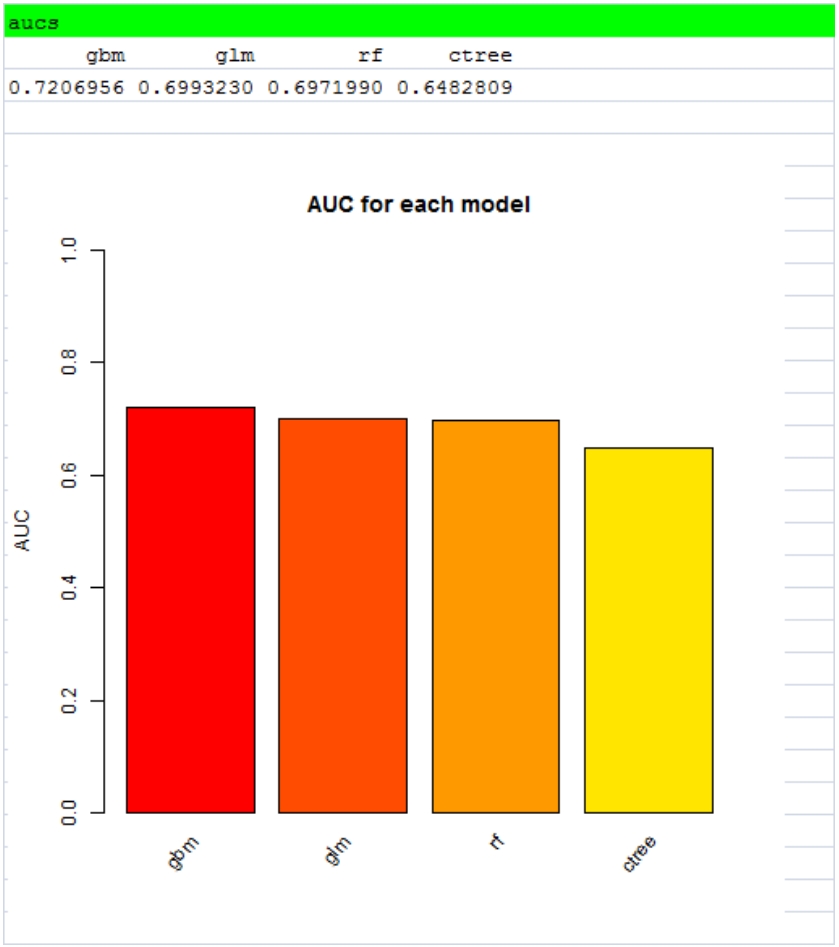
Predictive analytics techniques

- Forecasting: Pr(future outputs | past)
  - Dynamic Bayesian Networks (DBNs) forecast by conditioning on observed data
- Regression/classification: Pr(output | covariates)
  - R Caret package paradigm for training and testing
- Dynamic simulation: Pr(Δoutputs | Δinputs)
- Inference: Bayesian network (BN/ID) probabilities
  - Inference: Pr(outputs | observed inputs)
    - Monte-Carlo and exact inference algorithms
    - Structure learning and ensemble learning algs
  - Dynamic Bayesian Networks (DBNs)
    - Kalman filtering and extensions
    - Particle swarm optimization

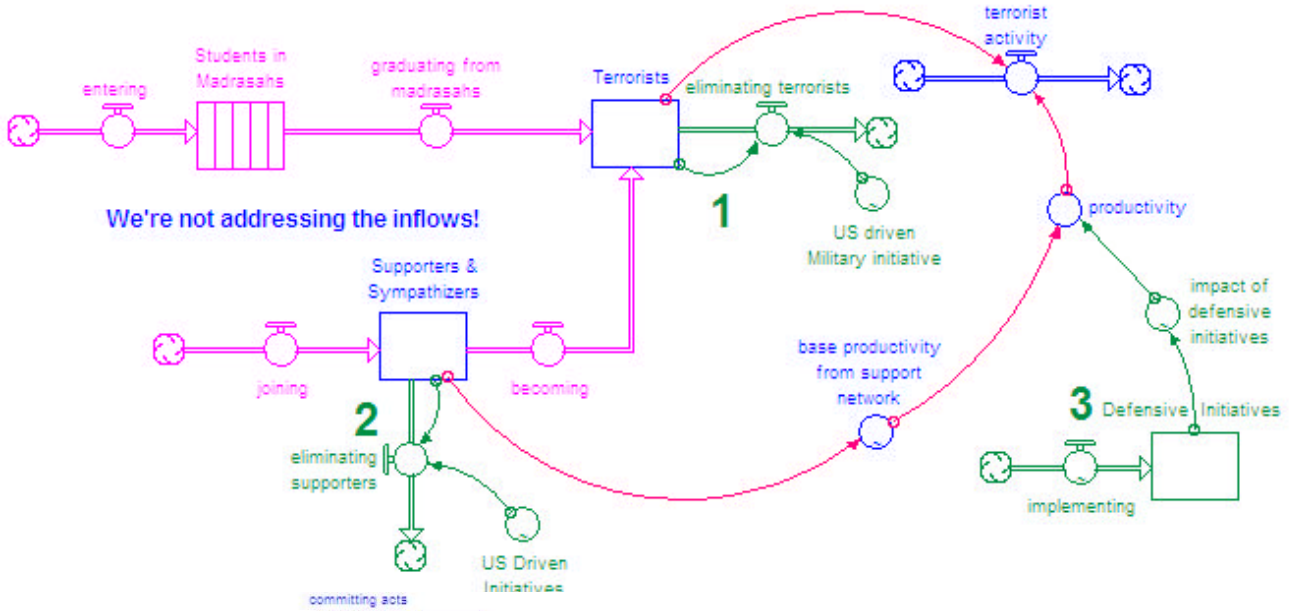
Breakthroughs in predictive analytics

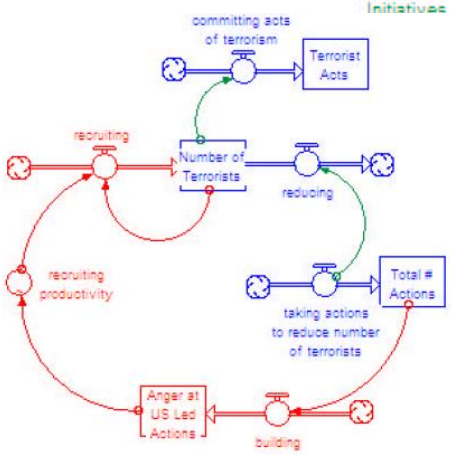
- Averaging predictions from multiple models improves predictions!
  - More accurate, less bias, more precise (lower error variance), less over-confidence (fewer type 1, type 2 errors)
- Ensemble methods improve forecasts
  - Random forest (rf)
  - Gradient boosting (gbm)
  - Cross-validation, BMA
  - Super-learning

```
aucs
       gbm       glm        rf     ctree
0.7206956 0.6993230 0.6971990 0.6482809
```



Systems dynamics simulations yield forecasts from inputs to causal models





http://www.systemdynamics.org/conferences/2003/proceed/PAPERS/417.pdf

Introduction to prescriptive analytics (decision analytics)

Introduction to evaluation analytics

Review of learning goals

- Learn how modern computational-statistical and machine-learning techniques can be used to implement information-based principles
- See how statistical and machine-learning methods support descriptive, predictive, causal, prescriptive, evaluation, and learning analytics
  - BN learning algorithms
  - CART trees
  - Influence diagram solution algorithms
- Be able to describe how causal analytics supports the rest of the risk management analytics cycle

Causal analytics informs the rest of the policy analytics cycle

Analytics Goal: Discover how to act more effectively

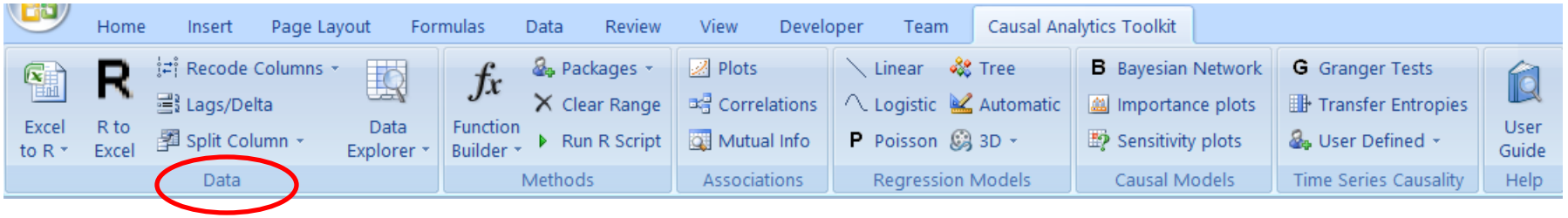
1. **Descriptive analytics:** What's happening? What's new? How have causes or effects changed? What to worry about?
2. **Predictive analytics:** What will (probably) happen next if we don't change what we're doing?
3. **Causal analytics:** What can we do about it? What will (probably) happen next if we do things differently?
4. **Prescriptive analytics:** What should we do?
5. **Evaluation analytics:** How well is it working?
6. **Learning analytics:** How to do better?
7. **Collaborative analytics:** How to do better together?

Making the algorithms useful: Netica®, R packages, and the Causal Analytics Toolkit (CAT)

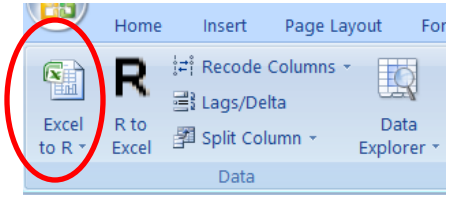
Learning goals for this section

- See how to apply R packages to carry out causal analytics based on information-theoretic principles and algorithms
  - Causal Analytics Toolkit (CAT) for R packages
  - BN learning algorithms
  - CART trees
  - randomForest ensembles
  - partial dependence plots
- Study practical application to an example data set

Causal Analytics Toolkit (CAT) software for advanced analytics



CAT uses data in Excel



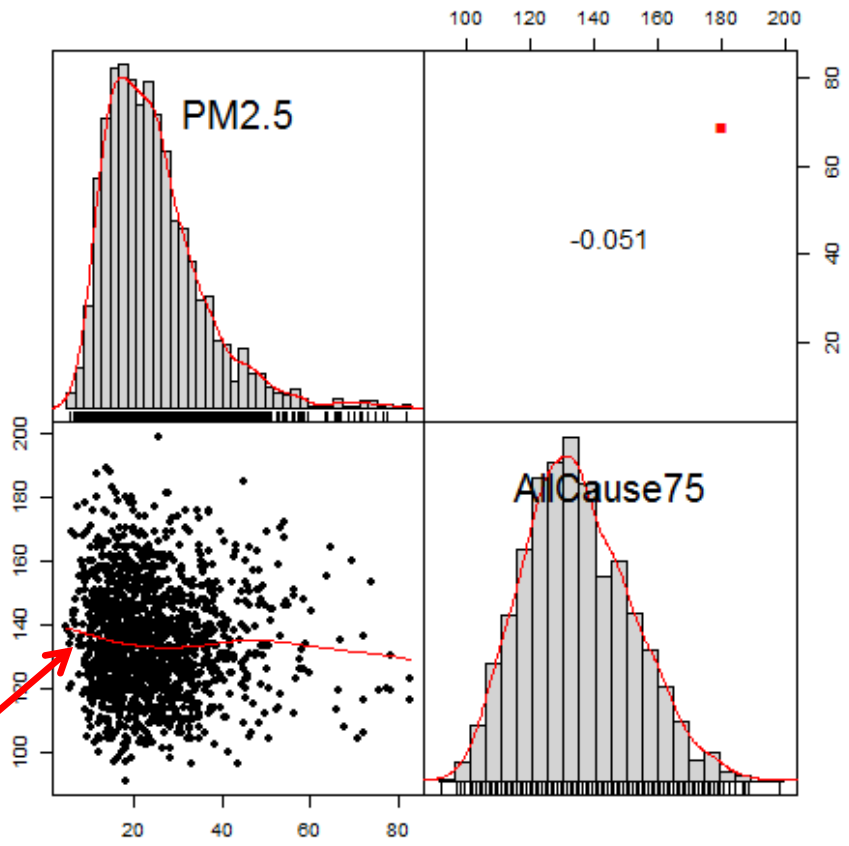
- Load data in Excel, click *Excel to R* to send it to R
  - Los Angeles air basin
  - 1461 days, 2007-2010 (Lopiano et al., 2015, thanks to Stan Young for data)
  - PM2.5 data from CARB
  - Elderly mortality ("AllCause75") from CA Department of Health
  - Daily min and max temps & max relative humidity from ORNL and EPA

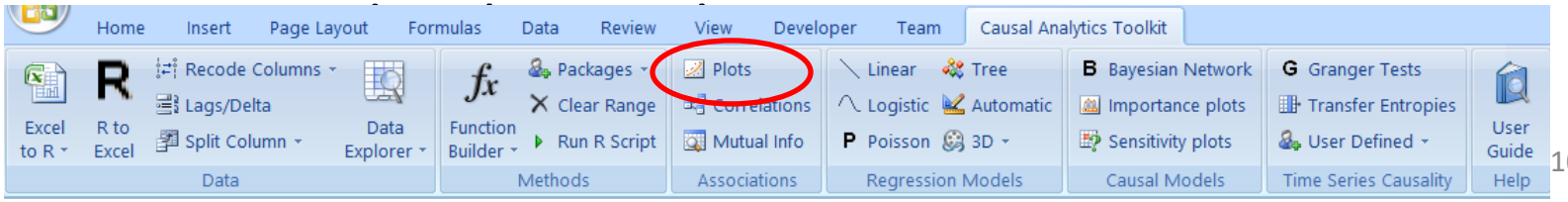
year	month	day	AllCause75	PM2.5	tmin	tmax	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	160	19.1	41	76	40.9
2007	1	8	148	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4

- *Risk question:* Does PM2.5 exposure increase elderly mortality risk? If so, how much?

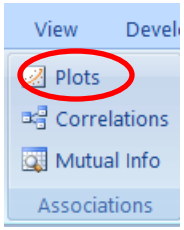
Using CAT to examine associations: Plotting the data

1. Send data from Excel to R
   - Highlight columns
   - Click on "*Excel to R*"
2. Select columns to analyze
   - Click on column headers
   - Cntrl-click toggles selectic
3. **Click on *Plots*** to view frequency distributions, scatter plots, correlation, smooth regression curves
   - PM2.5 is slightly negative

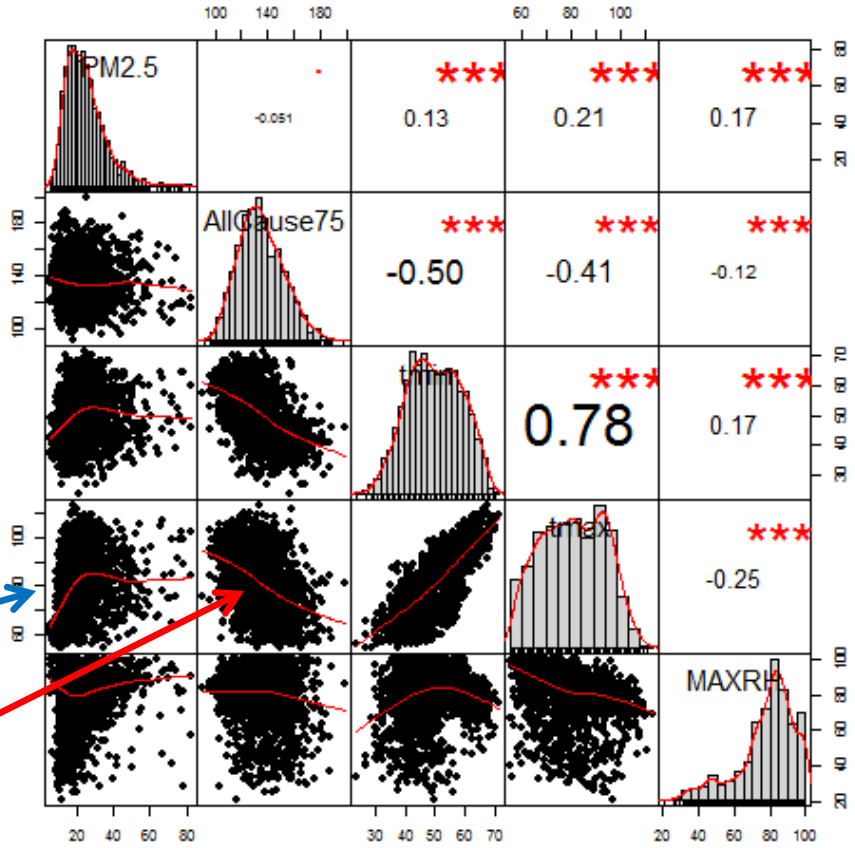




Using CAT to examine associations: Plotting more data



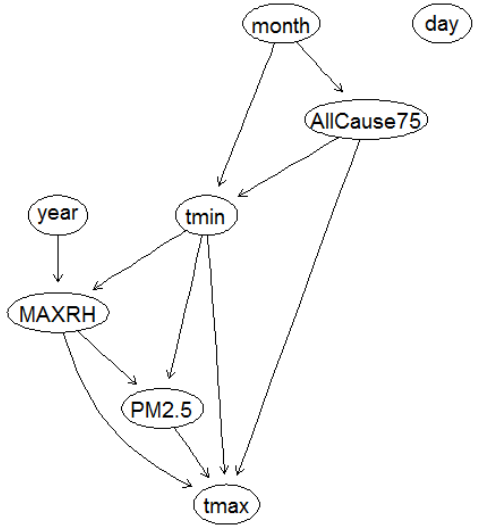
1. Send data from Excel to R
   - Highlight columns
   - Click on "*Excel to R*"
2. **Select columns**
   - Click on column heads
   - Cntrl-click toggles selection
3. **Click on *Plots*** to view frequency distributions, scatter plots, correlations, smooth regression curves
   - Temperature is positively associated with PM2.5
   - Temperature is negatively associated with mortality,

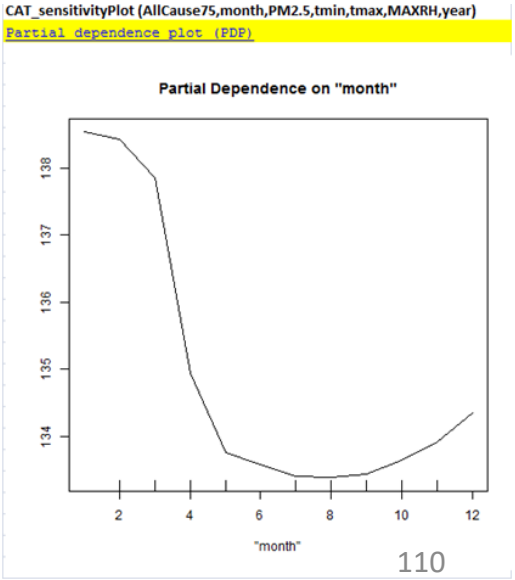


Basic ideas of Causal Analytics

- Use a network to show which variables provide direct information about each other
  - Arrows between variables show they are *informative about* each other, even given all other variables
  - Learn network structure directly from data
  - Carefully check conclusions
    - In non-parametric analyses we trust!
    - Do power analyses using simulation
  - Interpret neighbors in network as *potential direct causes* (satisfying necessary condition)

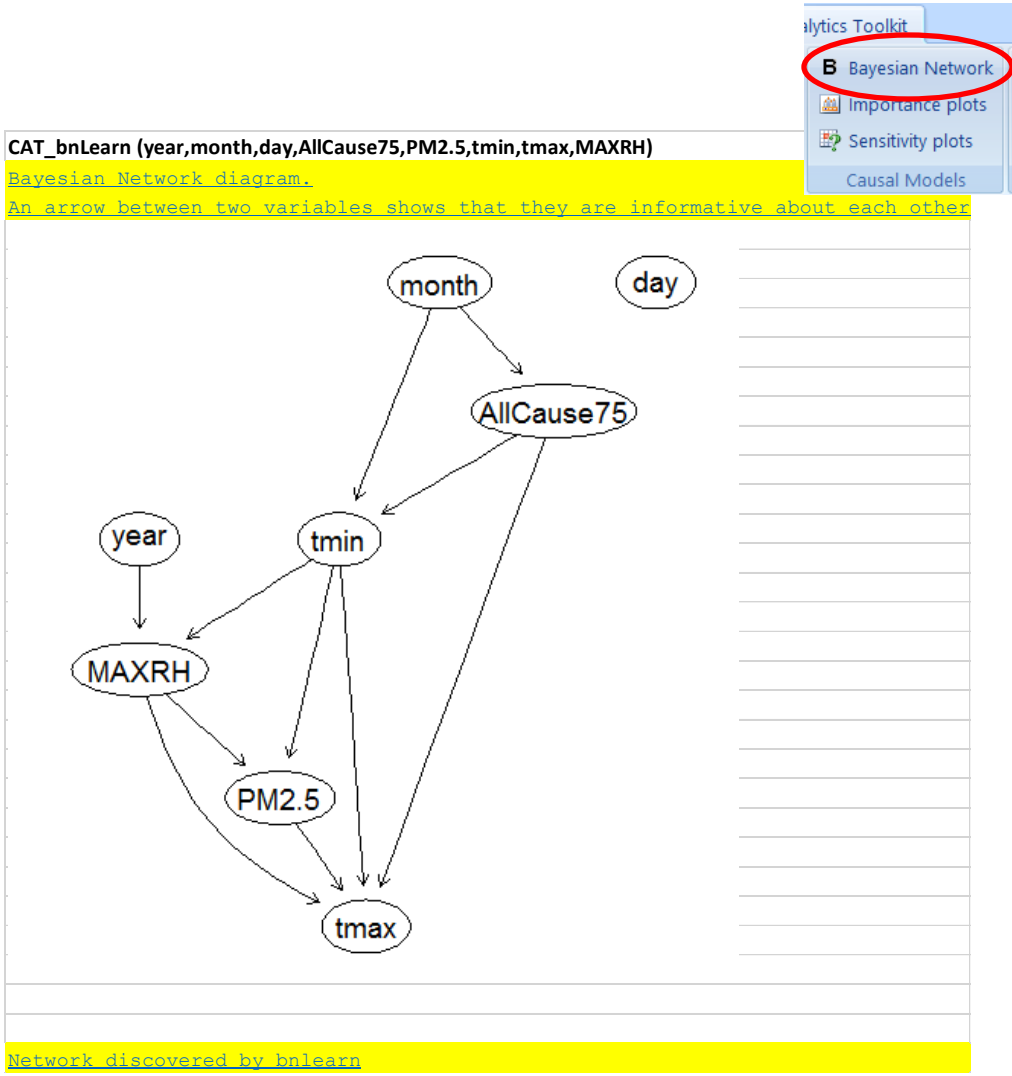
- Use sensitivity (partial dependence) graphs (based on averaging over many trees in randomForest ensemble to quantify relation between independent and dependent variables.





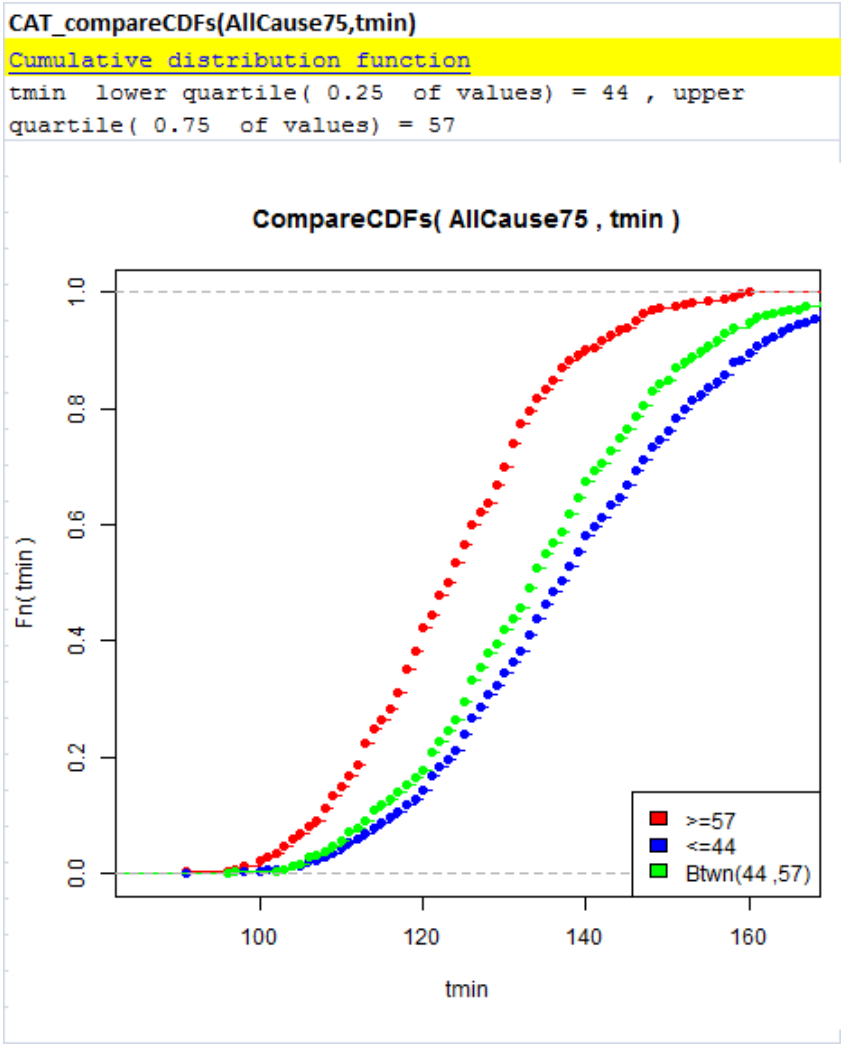
Run BN structure discovery algorithms

- Click **B** *Bayesian Network* to generate DAG structure.
  - Only variables connected to response variable by an arrow are identified as potential direct causes
  - Multiple pathways between two variables reveal potential direct and indirect effects
  - *Example:* Direct and indirect paths between tmax and AllCause75.



Confirm or refute/refine BN structure with additional non-parametric tests

- Conditioning on very different values of a direct cause should cause the distribution of the response variable to change
- If the response variable does not change, then any association between them may be due to indirect pathways (e.g., confounding)



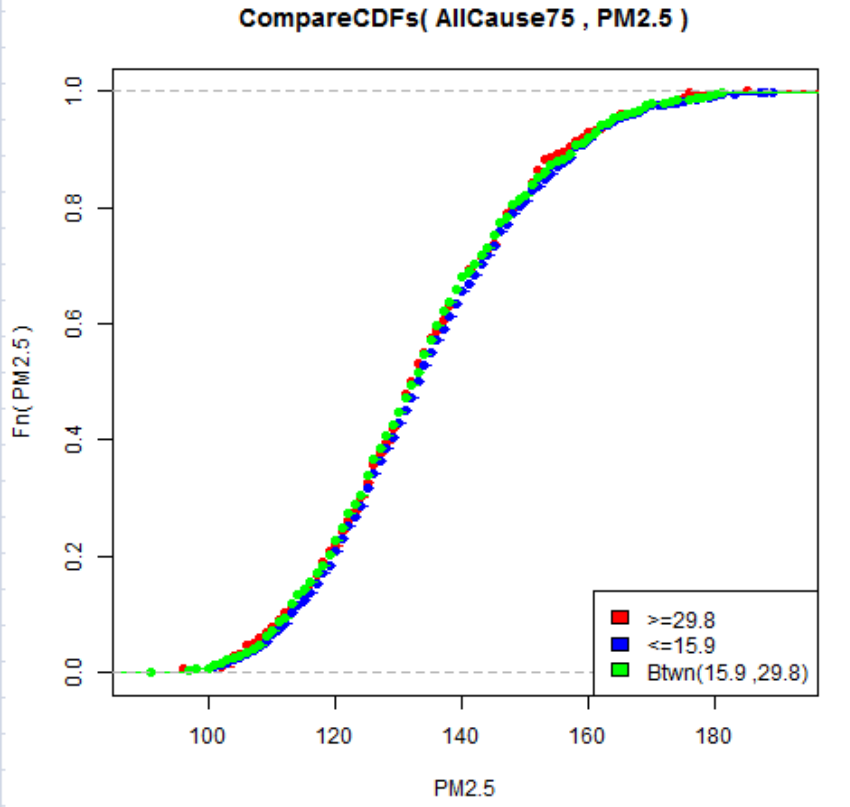
Confirm or refute/refine BN structure with additional non-parametric tests

- Conditioning on very different values of a direct cause should cause the distribution of the response variable to change
- If the response variable does not change, then any association between them may be due to indirect pathways (e.g., confounding)

CAT_compareCDFs(AllCause75,PM2.5)

Cumulative distribution function

PM2.5 lower quartile(0.25 of values) = 15.9 , upper quartile(0.75 of values) = 29.8



Discovering DAG structure resolves ambiguous associations

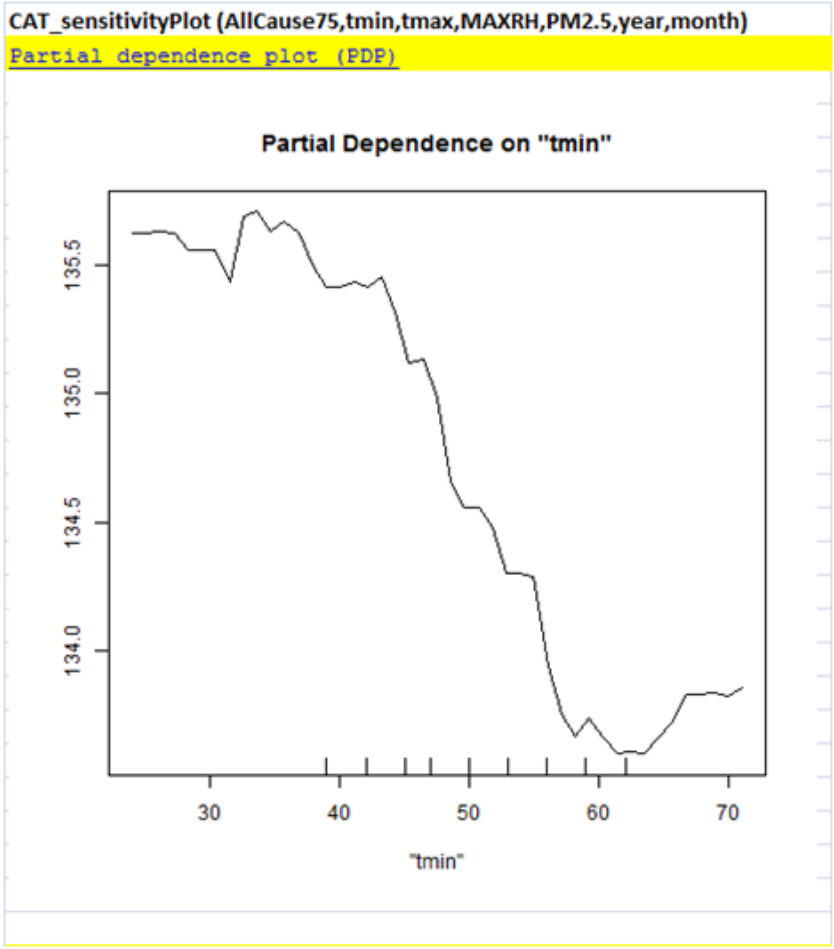
- How would cutting PM2.5 pollution in half affect future elderly mortalities per year?
  - No way to determine from association data

Community	PM2.5 in 1980 (μg/m^3)	Income	Elderly mortality rate in 1980
A	4	100	8
B	8	60	16
C	12	20	24

Model 1: *Income* $\rightarrow$ *PM2.5* $\rightarrow$ *Mortality:* mortality would be halved
Model 2: *PM2.5* $\rightarrow$ *Mortality* $\leftarrow$ *Income:* mortality would increase
Model 3: *PM2.5* $\leftarrow$ *Income* $\rightarrow$ *Mortality:* mortality would not change

Quantify direct causal relations

- *Procedure:* To quantify direct (potentially causal) relations after controlling for other variables and indirect pathways, estimate partial dependence graph for response R vs. (potential) cause C.
- *Rationale:* Screening and BN structure discovery have shown that the relation *might* be causal. Partial dependence estimates size of potential effect.



Validate quantified C-R relations in hold-out sample

- Current CAT uses bootstrap and cross-validation approaches for Random Forest ensembles
- Cross-validation and hold-out sample validation reports for regression and other analyses

CAT_importance (tmax,MAXRH,year,month,AllCause75,PM2.5)

Dependent variable: tmax

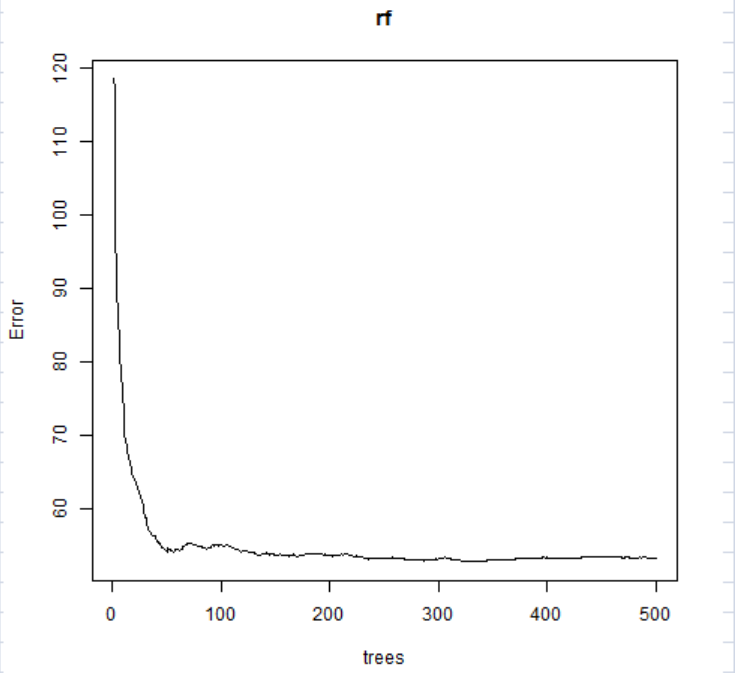
randomForest summary

No. of variables tried at each split: 1

Mean of squared residuals: 53.31128

% Var explained: 69.36

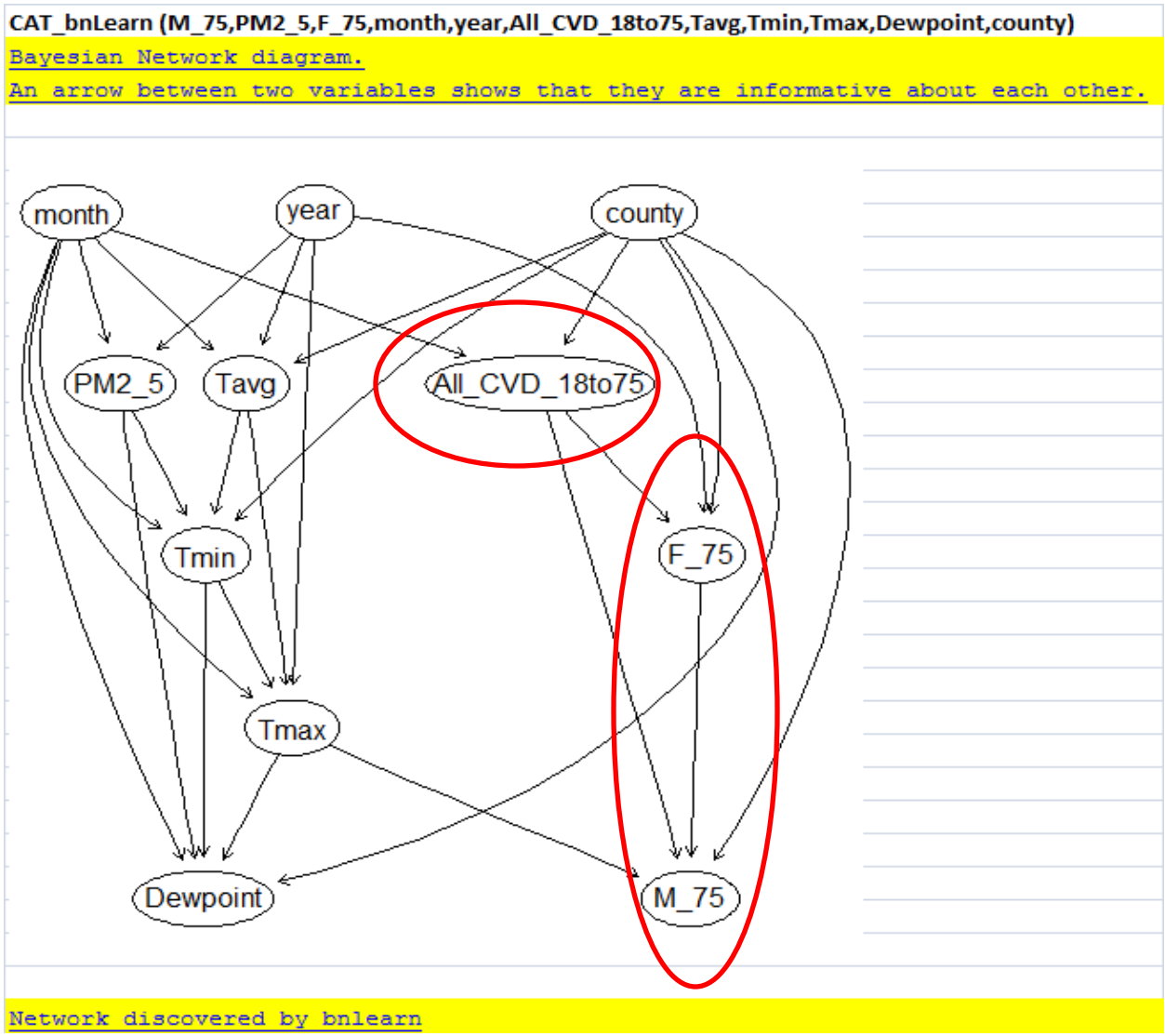
randomForest accuracy vs. number of trees



DAGs with hidden ("latent") variables: Testing for omitted confounders

- To test for effects of unobserved ("hidden" or "latent") confounders, partition study population into disjoint subsets
  - Men vs. women
  - Younger vs. older
- If mortality rate in one appears as direct cause of mortality in the other, then there is probably an omitted confounder that affects both.

Detecting Hidden confounders



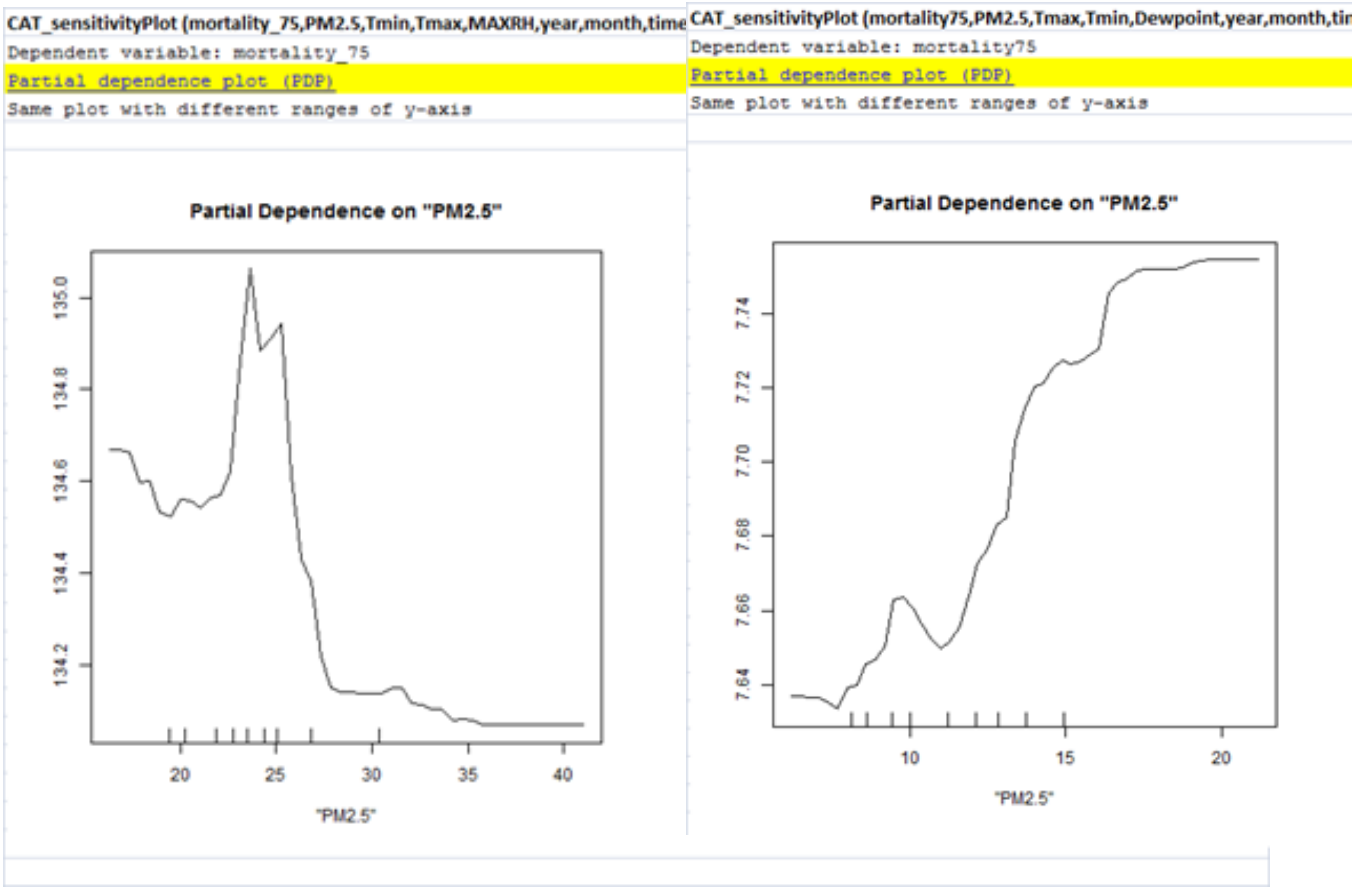
Transportability: Causal laws and mechanisms hold across settings

- Example model (or theory) structure for causes of response:

A directed acyclic graph (DAG) structure

- Quantify *Pr*(*mortality* | *age*, *sex*, *exposure*) ("CPT")
  - Conditional C-R relation, conditional probability table (CPT)
  - Response is *conditionally independent* of other variables, given the values of its direct parents in this network ("DAG model")
- A valid causal model or law (CPT) describing underlying mechanisms should be the same in all studies
  - Can be "transported" (generalized) across applications
  - Does not change based on arrows into *age*, *sex*, *exposure*
  - Otherwise, the causal theory needs to be expanded

Example: Testing transportability



Partial dependence relations between exposure (PM2.5) and mortality counts in two different cities look very different.

Summary of CAT's causal analytics

- Screen for total, partial, and temporal associations and information relations
- Learn BN network structure from data
- Estimate quantitative dependence relations among neighboring variables
  - Use partial dependence plots (Random Forest ensemble of non-parametric trees)
  - Use trees to quantify multivariate dependencies on multiple neighbors simultaneously
- Validate on hold-out samples
- Check internal consistency (dagitty, www.dagitty.net/dags.html), transportability, possible omitted variables

Review of learning goals

- See how to apply R packages to carry out causal analytics based on information-theoretic principles and algorithms
  - Causal Analytics Toolkit (CAT) for R packages
  - BN learning algorithms
  - CART trees
  - randomForest ensembles
  - partial dependence plots
- Study practical application to an example data set

Example applications: Law, regulation, science (toxicology, epidemiology), policy analysis

Learning goals for this section

- Examine some real-world applications and implications of causal challenges and techniques for science-policy practices
- Apply the concepts and methods we have learned to critical thinking about design and interpretation of real-world studies

IARC, 10-17-13

It has long been postulated that lung cancer may result from long-term exposure to ambient air pollution; the actual excess risk has nevertheless been estimated to be considerably less than that associated with tobacco smoking (Higgins, 1976; Pershagen, 1990). In confirmation of the early studies**, recent epidemiological investigations have observed an association between outdoor air pollution and lung cancer mortality**. It appears that particulate matter (PM), a complex mixture of airborne solid particles and aerosols, is the component causing serious health effects, for example mortality due to cardiovascular diseases and lung cancer (Dockery et al., 1993; Hemminki and Pershagen, 1994; Beeson et al., 1998; Abbey et al., 1999; Cohen, 2000; Pope et al., 2002; Vineis et al., 2004). In particular, long-term exposure to ambient fine particles (aerodynamic diameter < 2.5 µm [**PM2.5]) has been associated with lung cancer mortality** (or incidence) in studies carried out in different parts of the world and among nonsmokers (Dockery et al., 1993; Beeson et al., 1998; McDonnell et al., 2000; Pope et al., 2002, 2004; Laden et al., 2006; Beelen et al., 2008; Katanoda et al., 2011; Turner et al., 2011; Raaschou-Nielsen et al., 2011). **One extended follow-up study, the Harvard Six Cities Study from 1974–2009, demonstrated that the association between PM2.5 exposure and lung cancer mortality was statistically significant, with a linear concentration–response relationship without a threshold observed down to the PM2.5 level of 8 µm/m3 (Lepeule et al., 2012)**. In terms of lung cancer deaths, the annual contribution from ambient air pollution to lung cancer mortality has been estimated to be **responsible for more than 60 000 deaths worldwide, while more than 700 000 deaths are attributable to cardiac and non-malignant respiratory diseases (Cohen, 2003)**

Chronic Exposure to Fine Particles and Mortality: An Extended Follow-up of the Harvard Six Cities Study from 1974 to 2009

Johanna Lepeule[1], Francine Laden[1,2,3], Douglas Dockery[1,2,3], Joel Schwartz[1,2,3]

[1] Department of Environmental Health, and, [2] Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA, [3]Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA

Abstract

Background: Epidemiologic studies have reported associations between fine particles (aerodynamic diameter ≤ 2.5 µm; $PM_{2.5}$) and mortality. However, concerns have been raised regarding the sensitivity of the results to model specifications, lower exposures, and averaging time.

Objective: We addressed these issues using 11 additional years of follow-up of the Harvard Six Cities study, incorporating recent lower exposures.

Methods: We replicated the previously applied Cox regression, and examined different time lags, the shape of the concentration–response relationship using penalized splines, and changes in the slope of the relation over time. We then conducted Poisson survival analysis with time-varying effects for smoking, sex, and education.

Results: Since 2001, average $PM_{2.5}$ levels, for all six cities, were < 18 µg/m³. Each increase in $PM_{2.5}$ (10 µg/m³) was associated with an adjusted increased risk of all-cause mortality ($PM_{2.5}$ average on previous year) of 14% [95% confidence interval (CI): 7, 22], and with 26% (95% CI: 14, 40) and 37% (95% CI: 7, 75) increases in cardiovascular and lung-cancer mortality ($PM_{2.5}$ average of three previous years), respectively. The concentration–response relationship was linear down to $PM_{2.5}$ concentrations of 8 µg/m³. Mortality rate ratios for $PM_{2.5}$ fluctuated over time, but without clear trends despite a substantial drop in the sulfate fraction. Poisson models produced similar results.

Causal conclusion from non-causal data and analysis

Conclusions: These results suggest that further public policy efforts that reduce fine particulate matter air pollution are likely to have continuing public health benefits.

http://ehp.niehs.nih.gov/1104660/

Problem: Association is *not* causation. Evidence of association is *not* evidence of causation. (Confirmation bias makes this counterintuitive.)

- No matter how many adjectives (strong, consistent, etc.) apply, *association* does not necessarily reveal anything about *causation.*
  - Hill considerations misguide us
- Not only can confounders with time delays produce Hill-type *associations without causation*…
- But so can…
  - Data-, model-, and study-selection biases
  - Ignored model and exposure uncertainties
  - Multiple testing and multiple comparisons biases
  - Coincident historical trends

Association-based causal claims are inconclusive/unjustified

Pro (Claim)	Con (Caveat)
"Epidemiological evidence is used to quantitatively relate $PM_{2.5}$ exposure to risk of early death. We find **that UK combustion emissions cause 13,000 premature deaths in the UK per year**, while an additional 6000 deaths in the UK are caused by non-UK European Union (EU) combustion emissions" (Yim and Barrett, 2012).	"[A]lthough particulate matter has been associated with premature mortality in other studies, **a definitive cause-and-effect link has not yet been demonstrated**" (NHS, 2012)

Associations are inconclusive

Pro	Con
"**[A]bout 80,000 premature mortalities [per year] would be avoided by lowering PM2.5 levels** to 5 $\mu g/m^3$ nationwide" in the U.S. 2005 levels of PM2.5 caused about 130,000 premature mortalities per year among people over age 29, with a simulation-based 95% confidence interval of 51,000 to 200,000 (Fann et al., 2012).	"**Analysis assumes a causal relationship** between PM exposure and premature mortality based on strong epidemiological evidence… **However, epidemiological evidence alone cannot establish this causal link**" (EPA, 2011, Table 5-11).

Pro	Con
"[D]ata on the impact of improved air quality on children's health are provided, including… the **reduction in the rates of childhood asthma events during the 1996 Summer Olympics in Atlanta, Georgia, due to a reduction in local motor vehicle traffic**" (Buka et al., 2006). "During the Olympic Games, the **number of asthma acute care events decreased 41.6%** (4.23 vs 2.47 daily events) in the Georgia Medicaid claims file," coincident with significant reductions in ozone and other pollutants (Friedman et al., 2001).	"In their primary analyses, which were **adjusted for seasonal trends** in air pollutant concentrations and health outcomes during the years before and after the Olympic Games, the investigators **did not find significant reductions in the number of emergency department visits for respiratory or cardiovascular health outcomes in adults or children**." In fact, "relative risk estimates for the longer time series were actually suggestive of **increased ED [emergency department] visits** during the Olympic Games" (Health Effects Institute, 2010)

Associations are inconclusive

Pro	Con
“Our findings suggest that control of particulate air pollution in Dublin led to an immediate reduction in cardiovascular and respiratory deaths.” (Clancy et al., 2002) "The **results could not be more clear, reducing particulate air pollution reduces the number of respiratory and cardiovascular related deaths immediately**" (Harvard School of Public Health, 2002).	“Serious epidemics and pronounced trends feign excess mortality previously attributed to heavy black-smoke exposure” (Wittmaack, 2007).” “Thus**, a causal link between the decline in mortality and the ban of coal sales cannot be established**” (Pelucchi et al., 2009).

Benefit claims that probably are not true

- Banning passive smoking reduced heart attack risks among bar workers
- Reducing air pollution in Atlanta during Olympics reduced childhood asthma
- Banning coal-burning in Dublin reduced elderly mortality rates
- Red light cameras, flu shots, ….

Uncertain causation, regulation, and judicial review

- Causation is frequently poorly addressed in current regulatory practice and underlying science
  - Frequently conflated with association
  - Clear distinctions not made among associative, counterfactual, predictive, manipulative, and other types of causes
  - Tort-law's "but-for" causation not much help
- As a result, regulators may (and do) claim large benefits from regulations that do not necessarily cause them
  - Culture of true believers and judgment-centric determinations of causality favors exaggerated benefits estimates
  - Risk aversion for uncertain causality is ignored (Clean Air Act)
- Judicial review can increase net benefits from regulations by insisting on objective evidence of manipulative causation
  - Predictive causation is a useful, relatively objective data-driven screen
  - Otherwise, regulation is arbitrary and capricious

Example: Intervention study

Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study

Luke Clancy, Pat Goodman, Hamish Sinclair, Douglas W Dockery

Summary

Background Particulate air pollution episodes have been associated with increased daily death. However, there is little direct evidence that diminished particulate air pollution concentrations would lead to reductions in death rates. We assessed the effect of air pollution controls—ie, the ban on coal sales—on particulate air pollution and death rates in Dublin.

Methods Concentrations of air pollution and directly-standardised non-trauma, respiratory, and cardiovascular death rates were compared for 72 months before and after the ban of coal sales in Dublin. The effect of the ban on age-standardised death rates was estimated with an interrupted time-series analysis, adjusting for weather, respiratory epidemics, and death rates in the rest of Ireland.

Findings Average black smoke concentrations in Dublin declined by 35·6 μg/m³ (70%) after the ban on coal sales. Adjusted non-trauma death rates decreased by 5·7% (95% CI 4–7, p<0·0001), respiratory deaths by 15·5% (12–19, p<0·0001), and cardiovascular deaths by 10·3% (8–13, p<0·0001). Respiratory and cardiovascular standardised death rates fell coincident with the ban on coal sales. About 116 fewer respiratory deaths and 243 fewer cardiovascular deaths were seen per year in Dublin after the ban.

Interpretation Reductions in respiratory and cardiovascular death rates in Dublin suggest that control of particulate air pollution could substantially diminish daily death. The net benefit of the reduced death rate was greater than predicted from results of previous time-series studies.

Lancet 2002; **360:** 1210–14
See Commentary page 1184

Introduction

Results of many epidemiological studies have suggested an association between particulate air pollution and daily deaths.[1–3] Despite these findings, it does not follow that a reduction in particulate air pollution would diminish daily deaths or increase life-expectancy.[4] Great improvements in air quality in Dublin after the introduction of domestic coal-burning regulations offered an opportunity to assess the effects of reduced particulate air pollution on death rates in the general population.

Dublin's air quality deteriorated in the 1980s after a switch from oil to cheaper and more readily available solid fuels, mainly bituminous coal for domestic space and water heating.[5] Periods of high air pollution were associated with increased in-hospital respiratory deaths.[6]

On Sept 1, 1990, the Irish Government banned the marketing, sale, and distribution of bituminous coals within the city of Dublin.[7] The effect of this intervention was an immediate and permanent reduction in average monthly particulate concentrations.[8] We assessed the effect of the ban of coal on death in Dublin.

Methods

Procedures

We compared air pollution, weather, and deaths for 72 months before (Sept 1, 1984, to Aug 31, 1990) and after (Sept 1, 1990, to Aug 31, 1996) the ban, by seasons. We defined spring as March–May, summer as June–August, autumn as September–November, and winter as December–February. We calculated mean daily air pollution (black smoke and sulphur dioxide) concentrations with measurements from six residential monitoring stations in the city of Dublin (Dublin County Borough).[8] We obtained mean daily temperatures (°C) and mean daily relative humidity (%) from Dublin airport. We calculated the change in mean air pollution and weather variables

Example: Intervention study

Effect of air-pollution control on death rates in Dublin, Ireland: an intervention study

In conclusion, the ban on coal sales within Dublin County Borough led to a substantial decrease in concentration of black smoke particulate air pollution. After adjustment for age-distribution of the population, known predictors of death (including temperature, humidity, and respiratory epidemics), and death rates in the rest of Ireland as an index of unmeasured secular changes in deaths, we estimated that there were about 243 fewer cardiovascular deaths and 116 fewer respiratory deaths per year in Dublin after the ban on coal sales. These changes were seen immediately in the winter after introduction of the ban. Our findings suggest that control of particulate air pollution in Dublin led to an immediate reduction in cardiovascular and respiratory deaths. These data lend support to a relation between cause and the reported increase in acute mortality associated with daily particulate air pollution. Moreover, our data suggest time-series studies could be underestimating the benefits of particulate air pollution controls.

By what test?

Example: Intervention study

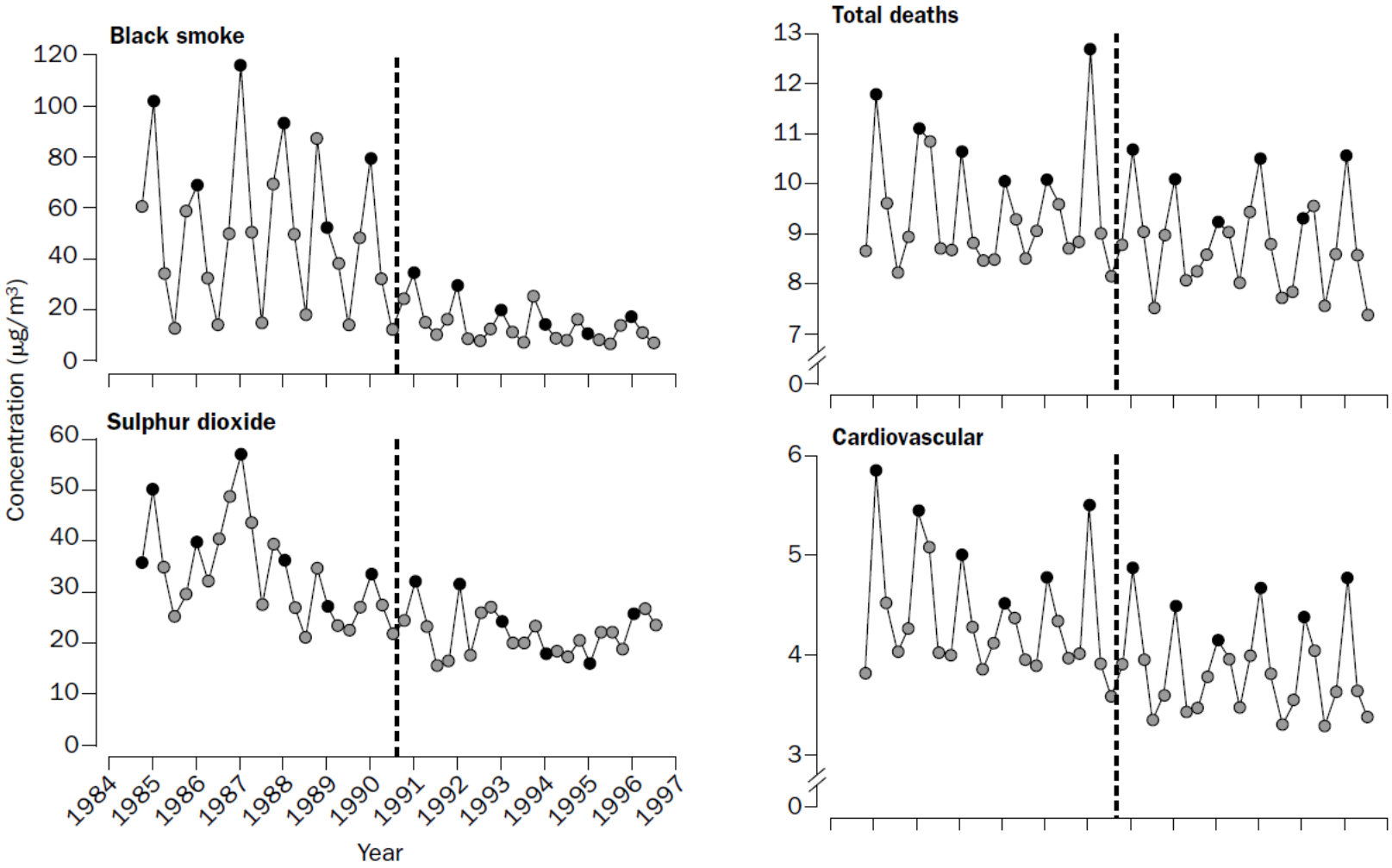


Figure 1: **Seasonal mean black smoke (upper) and sulphur dioxide (lower) concentrations, September 1984–96**
Vertical line shows date sale of coal was banned in Dublin County Borough. Black circles represent winter data.

Example: Intervention study

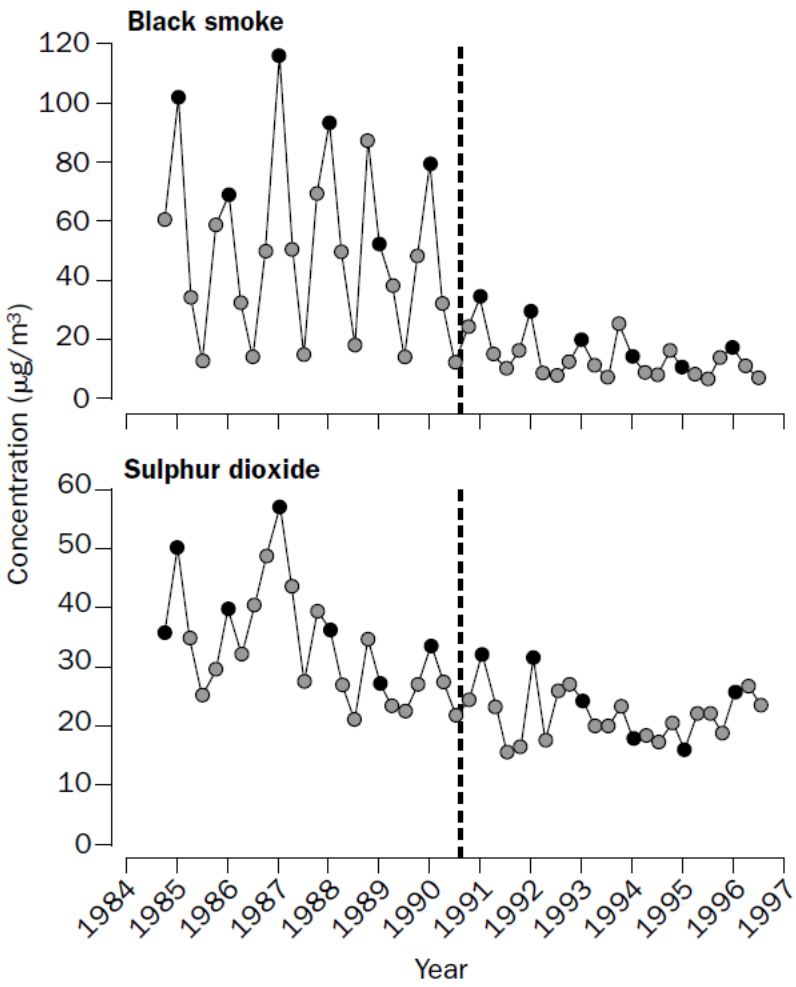
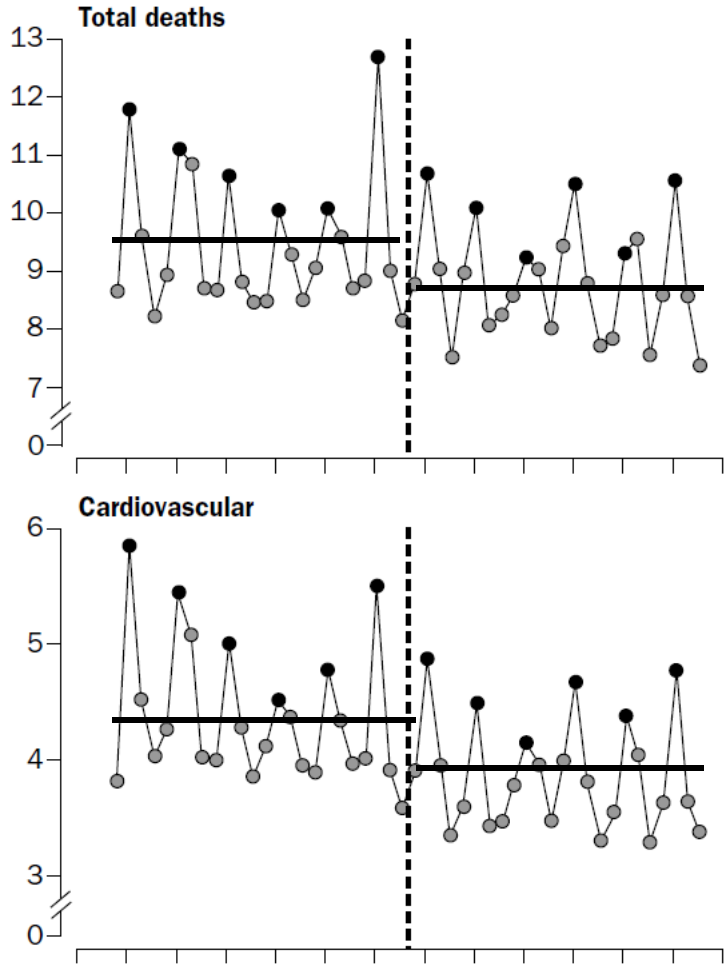


Figure 1: **Seasonal mean black smoke (upper) and sulphur dioxide (lower) concentrations, September 1984–96**
Vertical line shows date sale of coal was banned in Dublin County Borough. Black circles represent winter data.



"Adjusted non-trauma death rates decreased by 5.7% (95% CI 4-7, $p < 0.0001$)"

Example: Intervention study

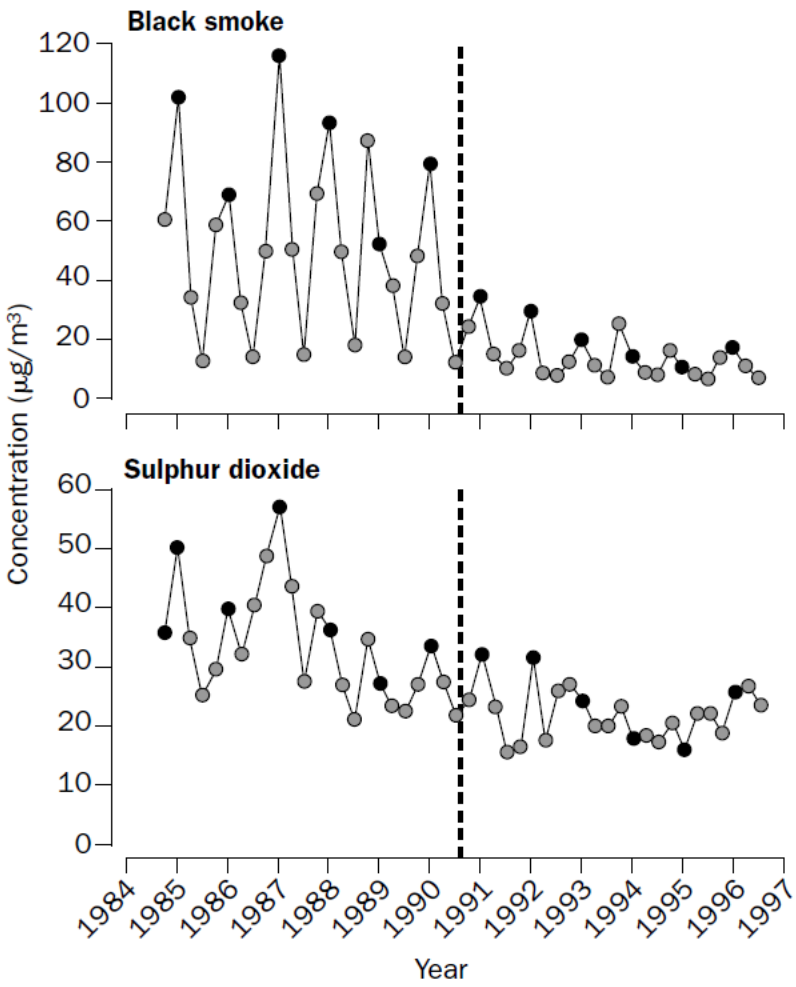
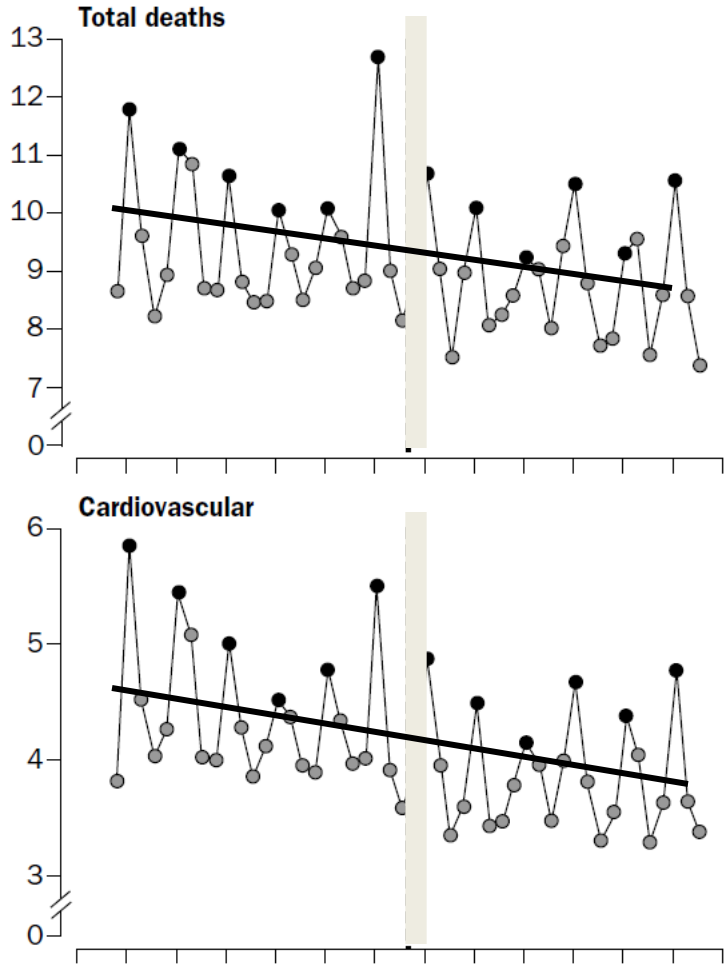


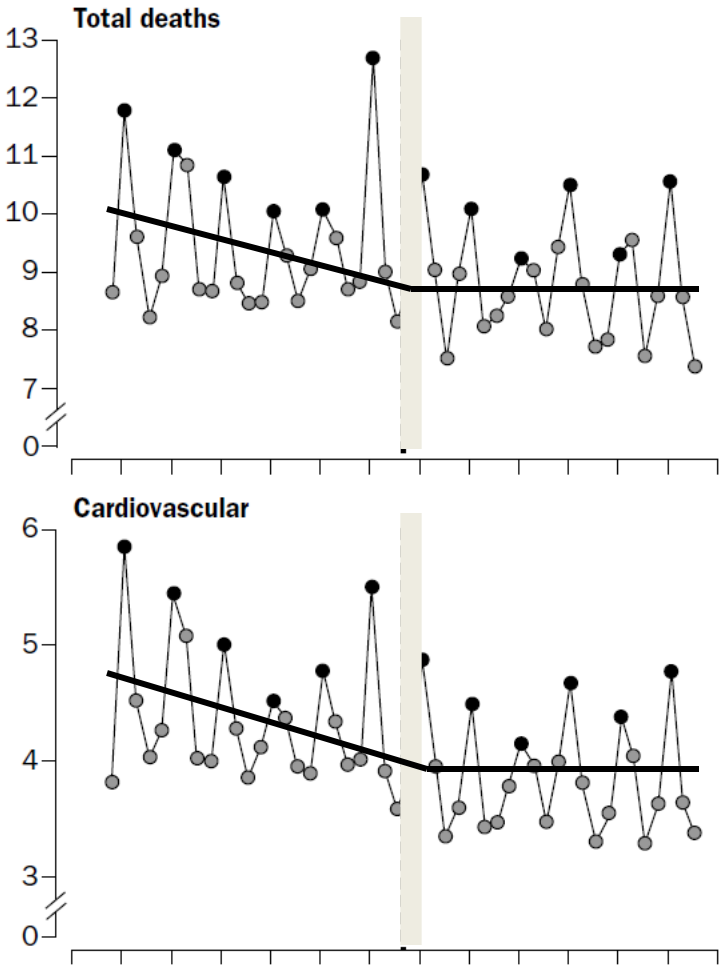
Figure 1: **Seasonal mean black smoke (upper) and sulphur dioxide (lower) concentrations, September 1984–96**
Vertical line shows date sale of coal was banned in Dublin County Borough. Black circles represent winter data.



"No significant reduction was found in total death rates" (Dockery et al., 2013)

Did the ban stop progress?

- Informal causal conclusions are just subjective opinions, with no known validity.
- Since 1960s, the quasi-experimental "O X O" one-group pretest post-test design has been cited as an example of a design that is not valid for causal inference (Campbell and Stanley, 1963, p. 7)
- What's missing: Learning by using information from *control groups* outside the ban area



Inhal Toxicol. 2007 Apr;19(4):343-50.

The big ban on bituminous coal sales revisited: serious epidemics and pronounced trends feign excess mortality previously attributed to heavy black-smoke exposure.

Wittmaack K. SF-National Research Centre for Environment and Health, Institute of Radiation Protection, Neuherberg, Germany.

Abstract

The effect of banning bituminous coal sales on the black-smoke concentration and the mortality rates in Dublin, Ireland, has been analyzed recently. Based on the application of standard epidemiological procedures, the authors concluded that, as a result of the ban, the total nontrauma death rate was reduced strongly (-8.0% unadjusted, -5.7% adjusted). The purpose of this study was to reanalyze the original data with the aim of clarifying the three most important aspects of the study, (a) the effect of epidemics, (b) the **trends in mortality rates** due to advances in public health care, and (c) the correlation between mortality rates and black-smoke concentrations. Particular attention has been devoted to a detailed evaluation of the time dependence of mortality rates, stratified by season. Death rates were found to be strongly enhanced during three severe **pre-ban winter-spring epidemics**. The cardiovascular mortality rates exhibited a continuous decrease over the whole study period, in general accordance with trends in the rest of Ireland**.** These two effects can **fully account for the previously identified apparent correlation between reduced mortality and the very pronounced ban-related lowering of the black-smoke concentration**. The third important finding was that in nonepidemic pre-ban seasons **even large changes in the concentration of black smoke had no detectable effect on mortality rates.**

Claimed health benefits vanish when control group information is used

Res Rep Health Eff Inst. 2013 Jul;(176):3-109.

Effect of air pollution control on mortality and hospital admissions in Ireland.

Dockery DW[1], Rich DQ, Goodman PG, Clancy L, Ohman-Strickland P, George P, Kotlov T; HEI Health Review Committee.

Abstract

During the 1980s the Republic of Ireland experienced repeated severe pollution episodes. Domestic coal burning was a major source of this pollution. In 1990 the Irish government introduced a ban on the marketing, sale, and distribution of coal in Dublin. The ban was extended to Cork in 1995 and to 10 other communities in 1998 and 2000. … In comparisons with the pre-ban periods, **no significant reduction was found in total death rates** associated with the 1990 (1% reduction), 1995 (4% reduction), or 1998 (0% reduction) bans, nor for cardiovascular mortality (0%, 4%, and 1% reductions for the 1990, 1995, and 1998 bans, respectively). The successive coal bans resulted in immediate and sustained decreases in particulate concentrations … but **no detectable improvement in cardiovascular mortality**.

Too late to change perceptions and policy

- "We intend to extend the health and environmental benefits of the ban on smoky coal, currently in place in our cities and large towns, **to the entire country**. …
- Benefits of a smoky coal ban include **very significant reductions in respiratory problems and indeed mortalities** from the effects of burning smoky coal. The original ban in Dublin has been **cited widely as a successful policy intervention** and has become something of an icon of **best practice** within the international clean air community. …
- **Research indicated that the ban in Dublin resulted in over 350 fewer annual deaths**. An estimate of these benefits in monetary terms put the value at over 20m euro."

www.housing.gov.ie/environment/air-quality/coal/smoky-coal-ban, 2015

Wishful thinking leads to more optimistic but unwarranted conclusions

www.slideshare.net/stmslide/patrick-goodman-dublin-technical

How the coal ban dealt with Dublin's burning issue

The prohibition of 'smoky' coal in 1990 resulted in 350 fewer annual deaths in city

Sat, Sep 26, 2015, 01:00

Olivia Kelly

In September 1990, following a series of winters during which Dublin city was engulfed in thick black smog, a ban on the sale, marketing, and distribution of bituminous or "smoky" coal was introduced in Dublin.

The results were dramatic with the city's caustic winter air pollution disappearing almost immediately.

It has since been reckoned the prohibition **resulted in 350 fewer annual deaths in the capital**.

Mary Harney, who helped push through the ban on 'smoky coal' in Dublin in 1990. Photograph: Aidan Crawley

In monetary terms it has had an **estimated benefit of more than €20 million**.

Despite the clear causative link between household coal burning and smog, there was strong resistance to the ban. Just one year previously Fianna Fáil environment minister Pádraig Flynn had ruled out a ban, claiming it would hurt widows and old-age pensioners.

However later in 1989 he got a new junior minister in Progressive Democrat Mary Harney, who was determined to see the ban through.

http://www.irishtimes.com/news/environment/how-the-coal-ban-dealt-with-dublin-s-burning-issue-1.2367021

Interpretation of “evidence” is not uniform

- Pope, 2009, “Evaluating the effectiveness of air quality regulations: A review of accountability studies and frameworks”: Intervention studies such as the Dublin air ban study “have provided additional evidence of adverse human health effects of air pollution… How many other opportunities such as the Dublin coal ban (Clancy et al., 2002) are being missed?
- Wittmaack, 2007: “The cardiovascular mortality rates exhibited a continuous decrease over the whole study period, in general accordance with trends in the rest of Ireland. ”

Review of learning goals

- Examine some real-world applications and implications of causal challenges and techniques for science-policy practices
- Apply the concepts and methods we have learned to critical thinking about design and interpretation of real-world studies

Goals for this workshop

- Introduce algorithms and principles for identifying approximately correct causal models from data
  - Using objective (assumption-free, modeler-independent) machine-learning methods where possible
- Distinguish between
  - (a) *statistical* associations, inferences, and models; and
  - (b) *causal* models to support/improve policy decisions
- Distinguish among different types of causality
  - Associational, counterfactual, predictive, manipulative, mechanistic/explanatory
- Fit causal analytics into larger analytics framework
- Introduce main concepts and software tools currently available to solve causal analytics problems

Causal analytics informs the rest of the analytics cycle

Analytics Goal: Discover how to act more effectively

1. **Descriptive analytics:** What's happening? What's new? How have causes or effects changed? What to worry about?
2. **Predictive analytics:** What will (probably) happen next if we don't change what we're doing?
3. **Causal analytics:** What can we do about it? What will (probably) happen next if we do things differently?
4. **Prescriptive analytics:** What should we do?
5. **Evaluation analytics:** How well is it working?
6. **Learning analytics:** How to do better?
7. **Collaborative analytics:** How to do better together?