DATA SCIENCE IN CLINICAL RESEARCH

Jeff Goldsmith
Biostatistics, Epidemiology and Research Design (BERD) Lecture

April 9, 2018
Data science is pretty new
Defining data science
More pictures

https://blog.zhaw.ch/datascience/the-data-science-skill-set/
Getting a little out of hand

I'm going to blame @drewconway for this

Data Science Is Multidisciplinary
MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand what a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xasS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decision and action
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Tableau, D3.js, Tableau

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand what a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xasS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decision and action
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Tableau, D3.js, Tableau
Why these definitions are bad

- “Data science is the combination of these 40 skills …” are unrealistic

Source: Drew Conway

https://www.youtube.com/watch?v=b9ZLXwAuUyw&app=desktop
Why these definitions are good

- Kinda like the blind men and the elephant – no one perspective is completely right or completely wrong, but piling them all up isn’t right either
- They give a sense of what is valued by the data science community – using data in a principled way and coding well
Why these definitions are good

- **Data science is interdisciplinary**
  - Does involve a breadth of statistical and computational skills
  - Curiosity and engagement is critical
  - You need some domain knowledge to be successful

https://www.xkcd.com/1831/
Is “data science” a buzzword?

• It is used to describe a frustratingly wide collection of ideas
• It is also used in different (sometimes contradictory) ways by different people

• Nonetheless, the popularity of “data science” reflects an increasingly data-centric reality
• Dismissing “data science” ignores this reality, and blinds people to the usefulness of a new perspective across disciplines
Reproducibility

• A concrete emphasis of data science is reproducibility
• Given the same data and the same code, anyone should be able to produce the same results
  – Openness is valuable – identify errors early and fix them quickly
  – Code is an important means of communication
  – New tools encourage reproducibility, but the concept is not platform-dependent
Replication

- Reproducibility is one part of replication …
Replication

• Reproducibility is one part of replication …
Replication

• Reproducibility is one part of replication …
Replication

- Reproducibility is one part of replication.

**Essay**

**Why Most Published Findings Are False**

*John P. A. Ioannidis*

**Is There a Replication Crisis?**

*By*

*Patil, Peng, and Leek 2016*
Statistical learning

• AKA machine learning, sometimes “AI”

• Becoming very popular…
Statistical learning

• AKA machine learning, sometimes “AI”

• Becoming very popular…

February 9, 2016

Machine Learning and the Profession of Medicine

Alison M. Darcy, PhD¹; Alan K. Louie, MD¹; Laura Weiss Roberts, MD, MA¹

» Author Affiliations  |  Article Information

Statistical learning

Previously unimaginable opportunities to apply machine learning to the care of individual patients in medical practice are currently within grasp.

Alison M. Darcy, PhD¹; Alan K. Louie, MD¹; Laura Weiss Roberts, MD, MA¹

» Author Affiliations  |  Article Information

Statistical learning

Previously unimaginable opportunities to apply machine learning to the care of individual patients in medical practice are currently within grasp.

Machine learning algorithms can accommodate different configurations of raw data, assign context weighting, and calculate the predictive power of every combination of variables available to assess diagnostic and prognostic elements. The applications of machine learning
Google flu trends

Google

Detecting influenza epidemics using search engine query data
Google flu trends

Detecting influenza epidemics using search engine query data
Google flu trends

Fever Peaks
A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.

Google’s algorithms overestimated peak flu levels this year.

When Google got flu wrong
US outbreak foxes a leading web-based method for tracking seasonal flu.
Statistical learning vs statistics

• Helpful to view statistical learning as part of a spectrum of tools
Statistical learning vs statistics

- Helpful to view statistical learning as part of a spectrum of tools

April 3, 2018

Big Data and Machine Learning in Health Care

Andrew L. Beam, PhD; Isaac S. Kohane, MD, PhD

» Author Affiliations  |  Article Information

Statistical learning vs statistics

• Helpful to view statistical learning as part of a spectrum of tools that perform on par with human physicians. Though machine learning and big data may seem mysterious at first, they are in fact deeply related to traditional statistical models that are recognizable to most clinicians. It is our hope...

Author Affiliations | Article Information

Statistical learning vs statistics

- Helpful to view statistical learning as part of a spectrum of tools that perform on par with human physicians. Though machine learning and big data may seem mysterious at first, they are in fact deeply related to traditional statistical models that are recognizable to most clinicians. It is our hope

Consequently, these algorithms often require hundreds of thousands of examples to extract the salient image features that are correlated with the outcome of interest. Higher placement on the machine learning spec-
Statistical learning spectrum

**Risk calculators**

17. CHA$_2$DS$_2$-VASc Score for atrial fibrillation stroke risk (2017)
18. MELD end-stage liver disease risk score (2001)

**Other**

22. Clinical wisdom

**Randomized Clinical Trials**


Beam and Kohane, 2018
Regression

• Regression (linear, logistic, etc) is interested in the conditional distribution of an outcome $Y$ given some predictors $x$

• Common form (continuous outcome):
  $$E(Y|x) = b_0 + b_1 x$$

• Regression has a lot of benefits, including:
  – Common understanding
  – Interpretable coefficients
  – Inference / p-values
Regression → Lasso

• One drawback of regression is lack of scalability
  – When you have some covariates, you have model-building options
  – When you have a lot of covariates, you have fewer options

• Lasso is useful when you have a lot of coefficients and few strong hypotheses
  – Goal is a regression-like model that “automatically” selects variables
Regression → Lasso

• Regression is estimated using the data likelihood:

\[ \min_b [RSS(b)] \]

• Lasso adds a penalty on the sum of all coefficients

\[ \min_b [RSS(b) + \lambda p(b)] \]

• Estimation is now a balance between overall fit and coefficient size
  – Roughly the same is true in other regression models
Lasso

\[
\min_b \left[ \text{RSS}(b) + \lambda p(b) \right]
\]

- Penalized estimation forces some coefficients to be 0, which effectively removes some covariates from the model.
- Result has a similar form to regression:
  - Can get predicted values based on covariates
- There are also some drawbacks:
  - No inference / p-values
  - Very different interpretation (if any)
  - Have to choose the tuning parameter
  - Coefficients for included covariates is not the same as in a regression using only those covariates

These drawbacks are roughly similar across statistical learning methods
Tuning parameter selection

\[ \min_b [RSS(b) + \lambda p(b)] \]

- For any tuning parameter value, Lasso returns coefficient estimates
- These can be used to produce predicted values based on covariates

- Tuning parameters are frequently chosen using cross validation
  - Split the data into training and testing sets
  - Fit Lasso for a fixed tuning parameter using training data
  - Compare observations to predictions on testing data
  - Repeat for many possible tuning parameter values
  - Pick the tuning parameter that gives the best predictions
Lasso example

- Toy example using public data
- 72 Leukemia patients
  - 47 ALL
  - 25 AML
- 3571 genes

- Regression is not really helpful here
- Lasso can search among genes for the most predictive of ALL vs AML
Lasso example

• Implemented using the \texttt{glmnet} R package
  – R Markdown file available
  – CV implemented as part of the package
• Lasso selects 5 of the 3571 genes
Statistical learning won’t fix bad data

From “Total Survey Error: Past, Present, and Future” (Groves and Lyberg)
via “Data Alone Isn’t Ground Truth” by Angela Bassa
Statistical learning won’t fix bad data

Untrustworthy Data Will Lead to Poor Conclusions

Tacking all data as if it were fact is a dangerous proposition.

So, can any data ever be trusted?

The short answer is… it depends. Skepticism is not a free pass to disregard data you disagree with. It’s a tool to ensure that the conclusions derived from data are reliable and do, in fact, reflect reality.

You also shouldn’t trust data just because it “proves” a point that you’re already inclined to believe. It’s probably even more important to be skeptical of extraordinary claims with which your heuristics already naturally align.

From “Total Survey Error: Past, Present, and Future” (Groves and Lyberg) via “Data Alone Isn’t Ground Truth” by Angela Bassa
Statistical learning won’t fix bad data

Untrustworthy Data Will Lead to Poor Conclusions

Trusting all data as if it is valuable is a recipe for disaster.

So, can any data ever be trusted?

The short answer is… it depends. We have to interpret data you disagree with. It’s unlikely that all data are reliable and do, in fact, reflect reality.

You also shouldn’t trust data just because it “proves” a point that you’re already inclined to believe. It’s probably even more important to be skeptical of extraordinary claims with which your heuristics already naturally align.

From “Total Survey Error: Past, Present, and Future” (Groves and Lyberg)
via “Data Alone Isn’t Ground Truth” by Angela Bassa

Beam and Kohane, 2018
Data Science Resources

- BERD R Short Course
- Data Science I (www.jeffgoldsmith.com/DSI)
Thanks!

- jeff.goldsmith@columbia.edu
- jeffgoldsmith.com
- github.com/jeff-goldsmith/

**BERD EDU link:**
http://irvinginstitute.columbia.edu/resources/biostat_educational_initiatives.html