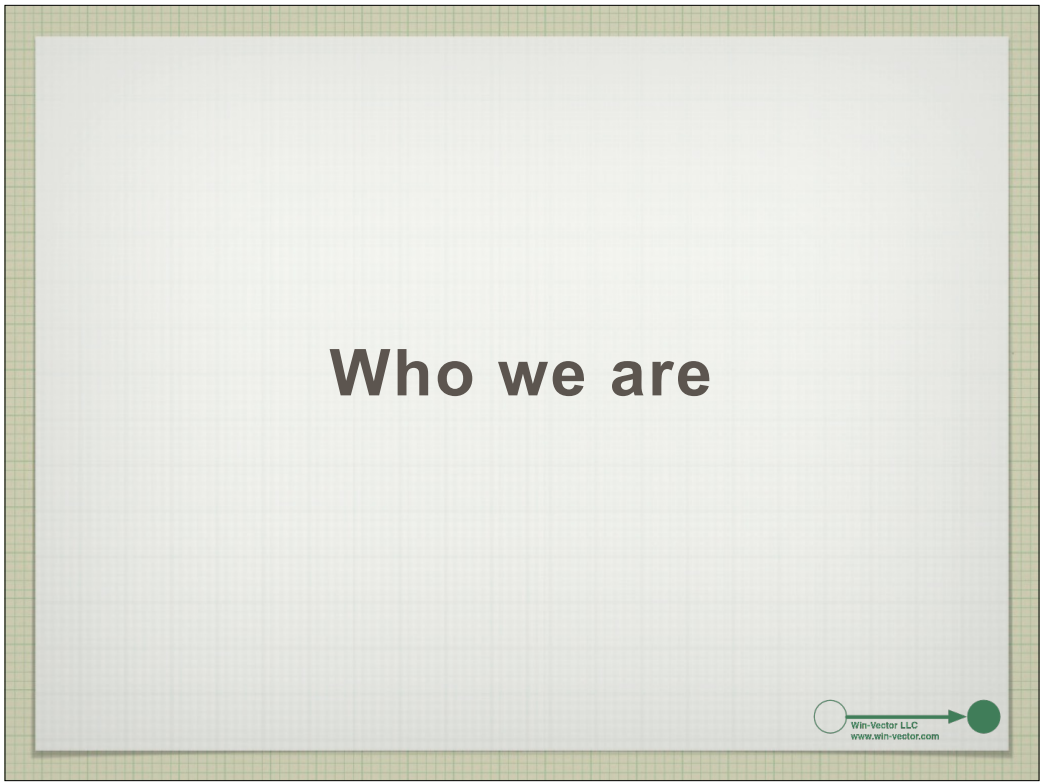
A quick introduction to practical data science

Nina Zumel
John Mount
Win-Vector LLC

Outline

- Who we are.
- A quick introduction to data science.
- A bit on our data science book: Practical Data Science with R



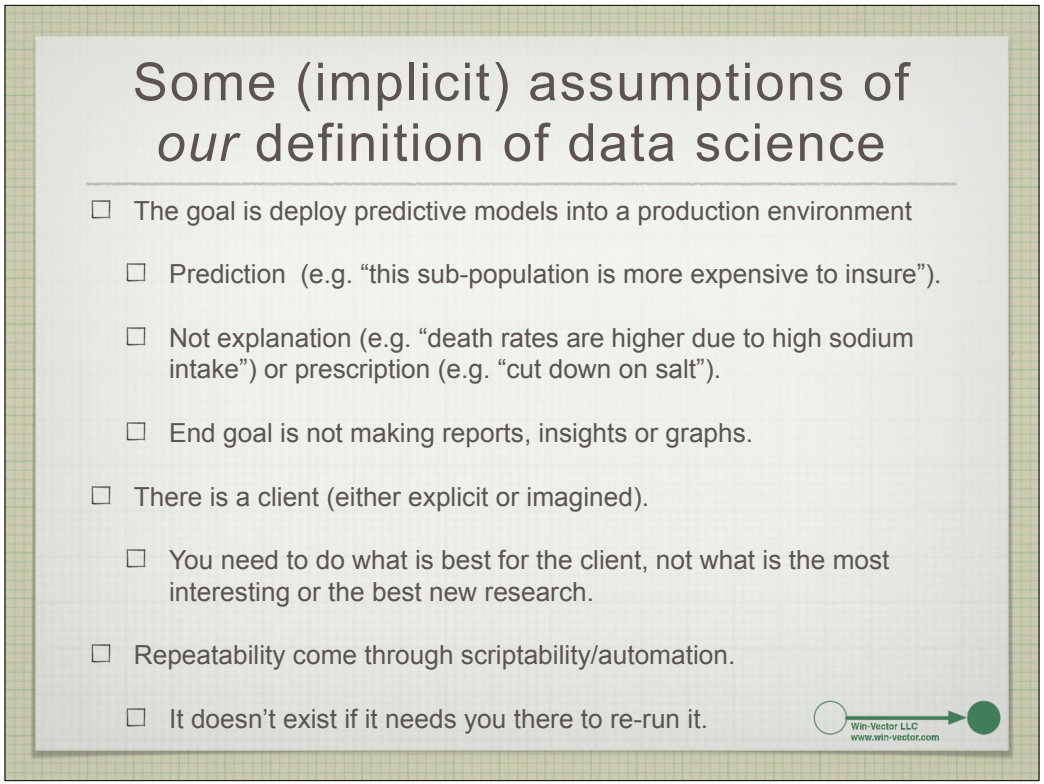
We are two consultants from Win-Vector LLC

What is Win-Vector LLC?

- Win-Vector, LLC is a two person data science consulting company: John Mount and Nina Zumel
  - Company started in 2007
  - Both PhDs; combined 30 years of experience in computer science and quantitative analytic work
  - We tend to work for other analysts and data scientists
- We write the Win-Vector blog
  - “Cool stuff” in CS, mathematics, statistics, and data science
  - Fairly technical

Unofficial theme of blog: “things it took me a long time to figure out, and now I want to write it down somewhere”

A quick introduction to data science

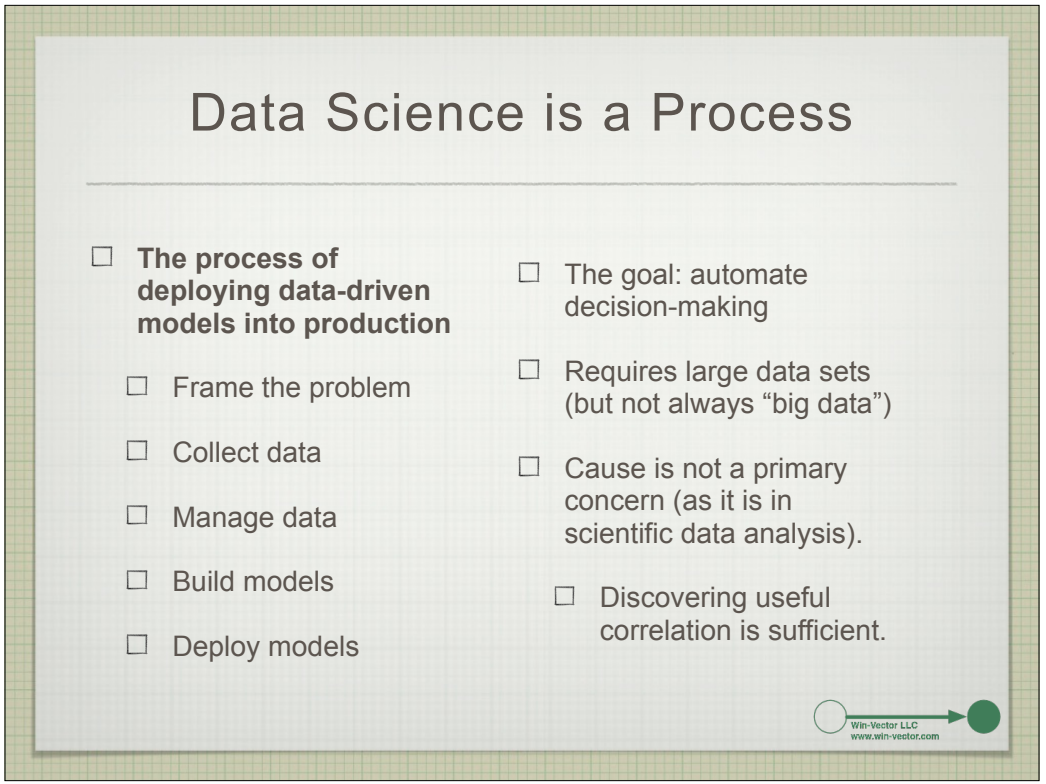


Notice how these may differ from a definition of statistics, so these are more argument the two fields are different. In our opinion data science takes a lot of ideas from agile development. Remember: data gets updated. You are always on your penultimate analysis.

We already have statistics, why do we need data science?

From “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics”, ISI Review, , 69, 21-26. W. S. Cleveland, 2001.

This document describes a plan to enlarge the major areas of technical work of the field of statistics. Because the plan is ambitious and implies substantial change, the altered field will be called “data science.”



This is our working definition for data science. data science is really the entire process. we want to give attention in the book to aspects of the process that often get ignored by people (on paper) who focus on specific sub-topics
Framing the problem: generate metrics for success in the context of the business/application area

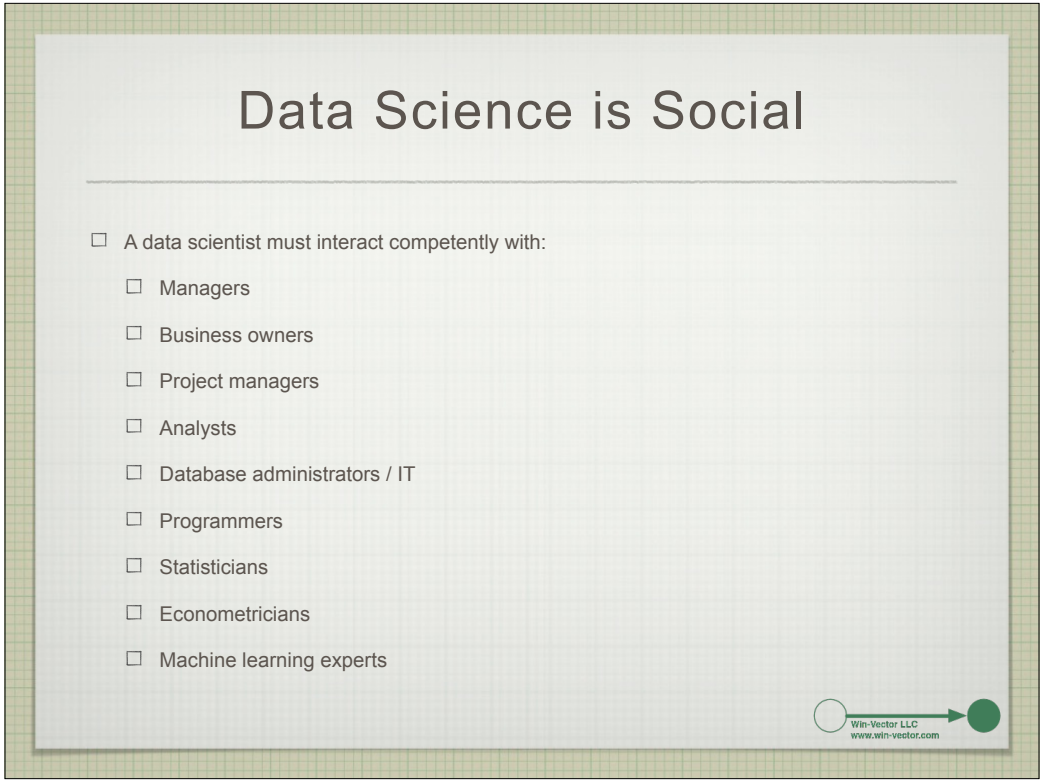
The activities encompass tasks businesses have always worried about:
loan default
anomaly detection
revenue management
consumer demand projection
social network discovery
-- to name a few

Now we can do it on a larger scale, and with faster turnarounds

Large data: asymptotic conditions have been met. worry less about statistical efficiency issues (different than algorithmic efficiency)

You need tools and habits to support process

- ☐ Data stores
- ☐ Source control
- ☐ Metadata store/inventory
- ☐ Documentation
- ☐ Permissioning
- ☐ Repeatable procedures



You have to be able to communicate with people in all these roles, which means that you have to be competent at some aspects of all of them.
You need to be good (or at least not bad) in all of these roles. At the roles you are “at least not bad” at you need to interact and recruit expert partners in your organization.

Data Science is Multidisciplinary

- Theoretical basis: statistics, machine learning, operations research
- Practical basis: computer and software engineering
- Data Science as we currently know it has evolved largely in Computer Science and Information Technology groups

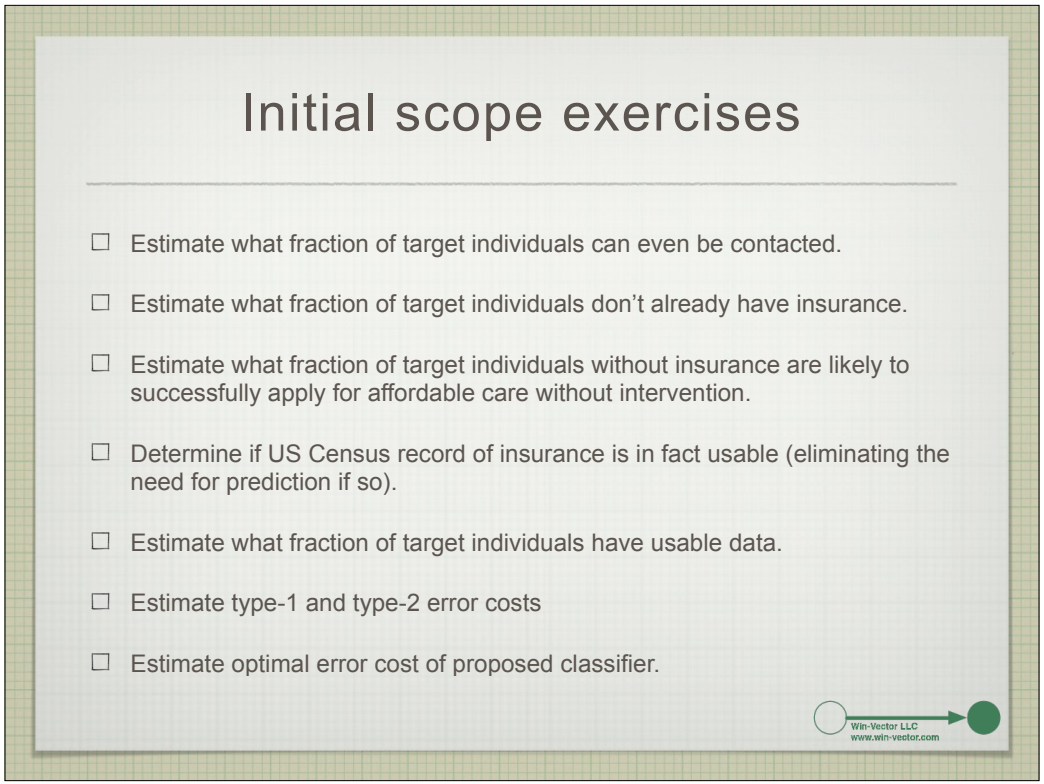
Machine learning to implement desired predictive models/ statistics to design data collection strategies and validate models
CSE to efficiently collect and manage high volume data for building models, and run desired computations in a timely manner.
Re pt 3: DS evolved in CS/IT groups because they had the data and the need (e.g. facilitate other groups who needed to answer the business questions)
And remember that machine learning, in its early years, was considered a sub-branch of Artificial Intelligence (a computer science topic), not of statistics.
However, bad software engineering is easier to fix than incorrect statistical inference (so we do pay attention to statistics).

Huge field, so let's limit down to a quick example of how to:

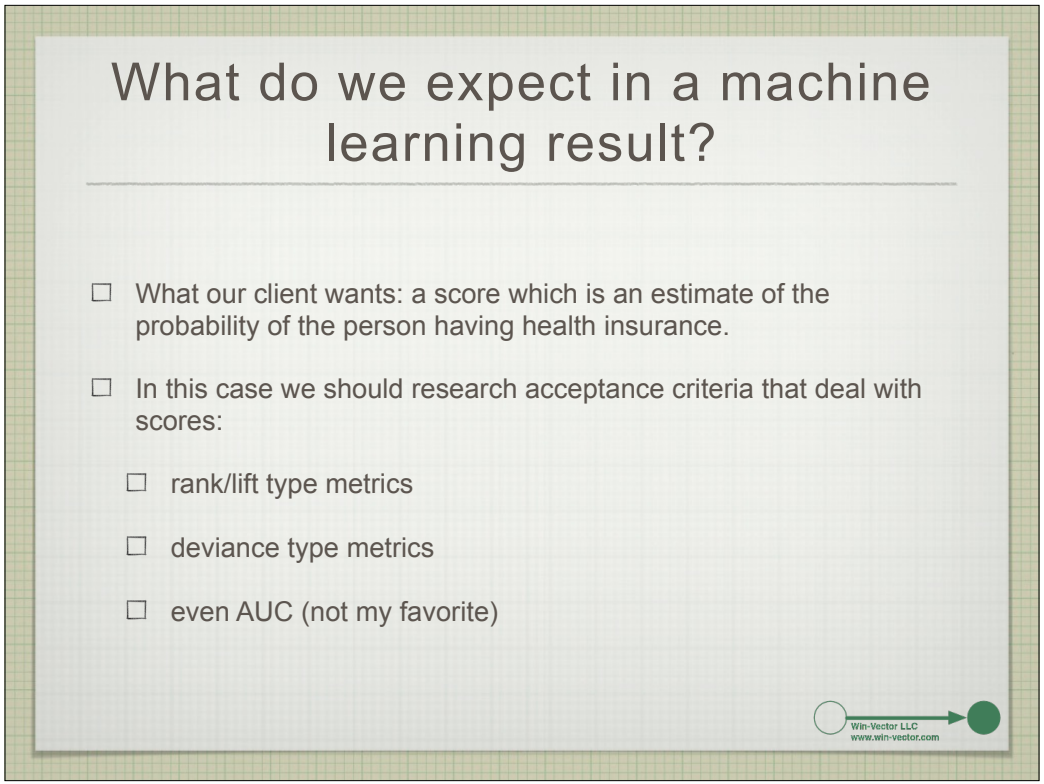
- ☐ Effectively scope and organize a data science project
- ☐ Be an effective consumer of machine learning results
- ☐ Choose among important model performance criteria
- ☐ Produce effective data visualizations

Our example problem

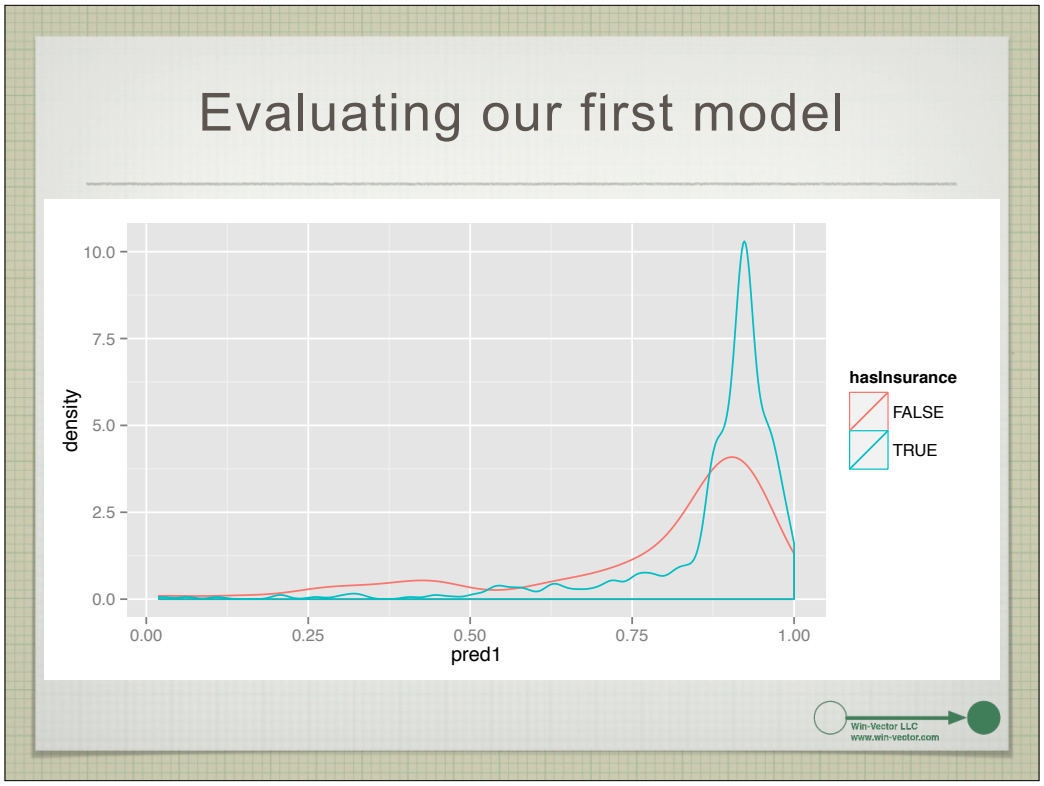
- Project Goal: contact people unlikely to have health insurance and try to help sign them up for affordable care.
- Client ask: a score that estimates the probability that a person has insurance based on a few census facts (age, income, occupation, and location).
  - Allows the client to set a threshold on who gets called and who doesn't get called after they see model results.



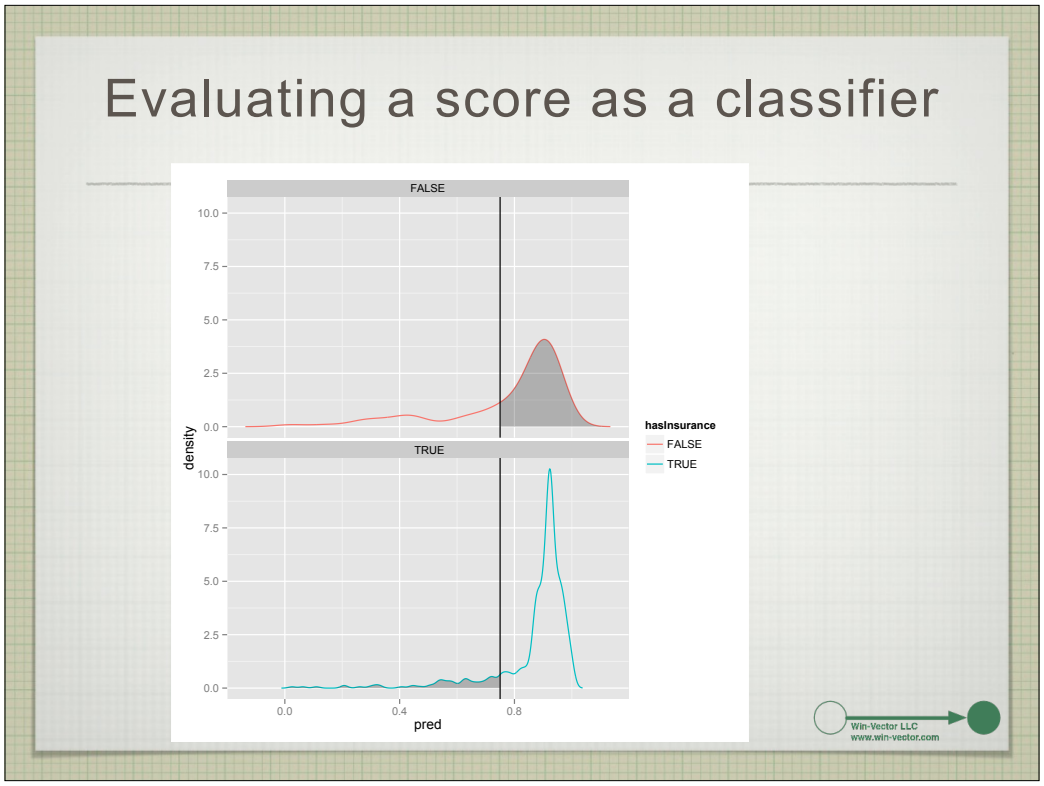
Notice what a machine learning researcher sees as the the scope: data size, and estimating the accuracy of the classifier (using their favorite method) isn't the first question to ask. Extreme answers to any of the earlier questions can invalidate the need for the project or the classifier at all. After project viability many of the scoping questions (type of data, volume of data, cost) are classic software engineering questions- so there are known good methods for front-loading and dealing with such uncertainty. A standard good idea: is can you fake/demonstrate the results using people and 3*5 cards; and if you did would the result be helpful?



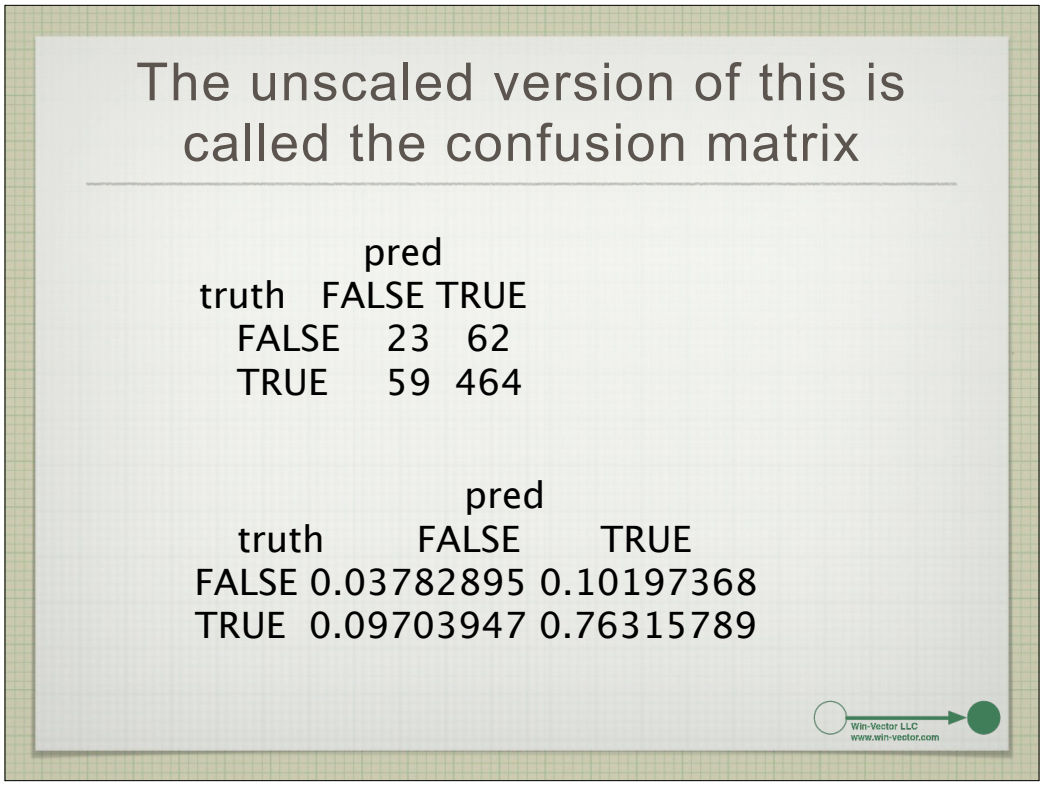
The client ask is a ranking or score. This means we have to return a scoring function and not a mere classifier (unless we have good reasons to convince the client classification is more appropriate for the overall project). Until we switch to a classification goal simple “accuracy” scores are not good enough to evaluate the project.



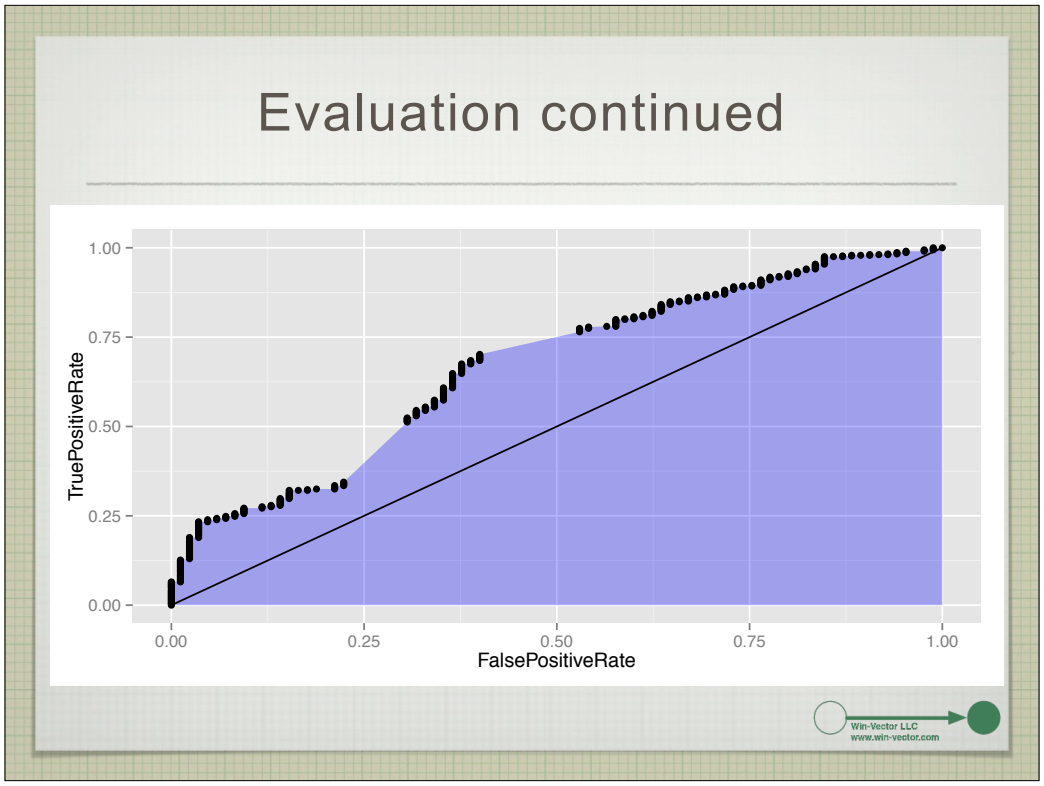
This what we call a double density plot. The x-axis is the model prediction score. For blue curve is the distribution of scores on example individuals known to have insurance and the blue curve is the distribution of scores on example individual known to not have insurance. Both curves are scaled to have total area of 1.0. An idea plot would have the red-curve as a sharp spike near 0.0 and the blue curve as a sharp spike near 1.0. The idea is to run Bayes law in your head: if the model score can separate classes then the model scores should look different conditioned on class. The result shown here is not idea: something you very much have to get used to in real world data science. The idea is to build best possible results in parallel with business policies and practices that work with the types of results you can produce. In this case first-level intervention should be cheap as the model isn't very selective at safe threshold levels. Each visualization has to answer questions (or there is no point). This visualization summarizes how score can be converted to a classifier.



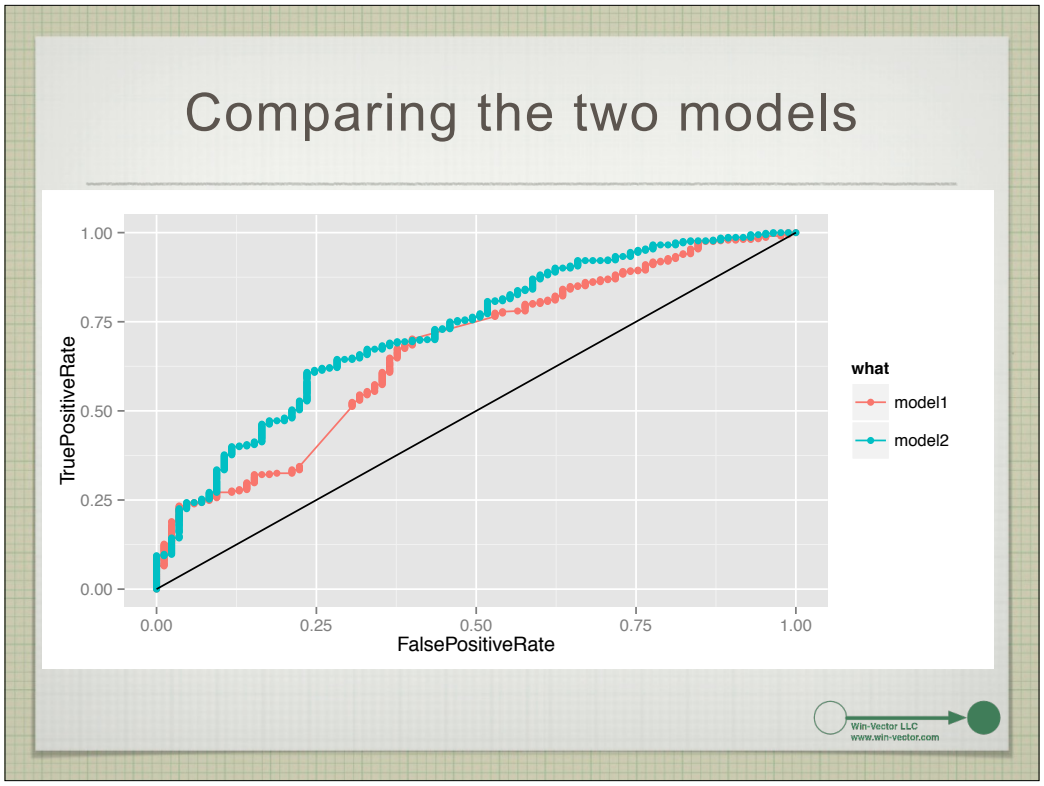
Any threshold converts a scoring model into a classifier. For example here we show (in the shaded) regions the proportions of data that are classified in error at a score threshold of 0.75.



Various scalings of the confusion matrix yield important statistics such as: true positive rate, false positive rate, true negative rate, false negative rate, sensitivity, specificity, precision, and recall. Pairs of these are to preferred to single classifier quality measures such as accuracy or F1.



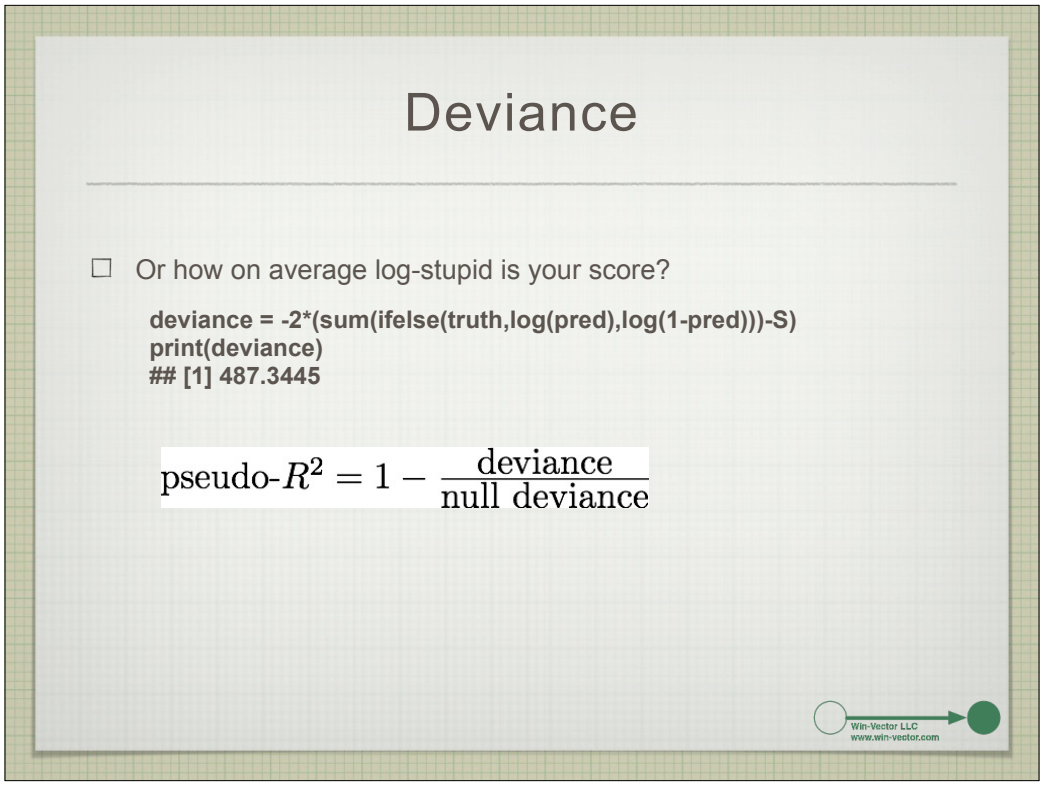
In the previous graph at any scoring threshold our model becomes a classifier that selects some fraction of the people with insurance (true positives: the fraction of people with insurance we correctly picked divided by the total number of people with insurance) and (false positives: the fraction of people without insurance that leak into the result over the number of people without insurance). When you realize that the actual score value returned by the classifier is an irrelevant feature you end up inventing the Receiver Operating Characteristic plot where for each FPR you plot what TPR rate it is achievable with (using your score). This graph gives a summary of all classifiers that can be derived for a given scoring system. We can see the subtended area (called AUC) and it is often used as an informal measure of model quality. However your client is likely going to pick a single threshold and use the model as a classifier at only one point on the curve: so are coming from regions where the client is not willing to use the model are irrelevant to model utility.



Notice model2 dominates model1 almost everywhere. We perceive area, so we see AUC.

It is important to automate scoring

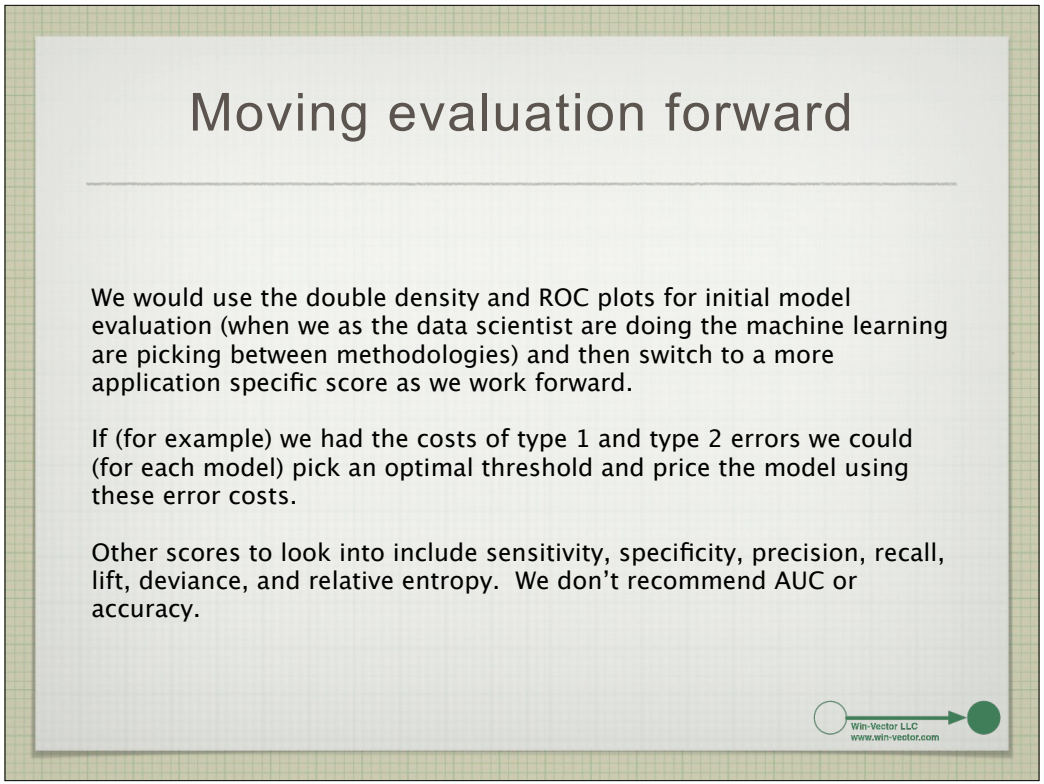
- So we prefer quantitative scores, like:
  - precision / recall
  - sensitivity / specificity
  - deviance / relative entropy



S is the log-likelihood of an ideal "saturated model" (usually taken to be 0). Perfect models have a deviance of 0

Deviance example

- 4 examples with truth-values= T , T , F , F
  - Model predicts p= 0.9, 0.1, 0.9, 0.1
  - Deviance= -2*(log(0.9) + log(0.1) + log(1-0.9) + log(1-0.1))
  - = -2*(-0.152 + -3.322 + -3.322 + -0.152)
  - okay ouch ouch okay



Not a closed set of procedures or answers: you have to research what is useful for your application.

Further writings on "statistics for engineers":

- Setting expectations in data science projects
  - http://wp.me/pgDAI-vA
- Statistics to English Translation, Part 1: Accuracy Measures
  - http://wp.me/pgDAI-gW
- A clear picture of power and significance in A/B tests
  - http://wp.me/pgDAI-GR
- Statistics to English Translation, Part 2a: 'Significant'
  - http://wp.me/pgDAI-j8
- Worry about correctness and repeatability, not p-values
  - http://wp.me/pgDAI-Cn
- Bayesian and Frequentist Approaches: Ask the Right Question
  - http://wp.me/pgDAI-CM

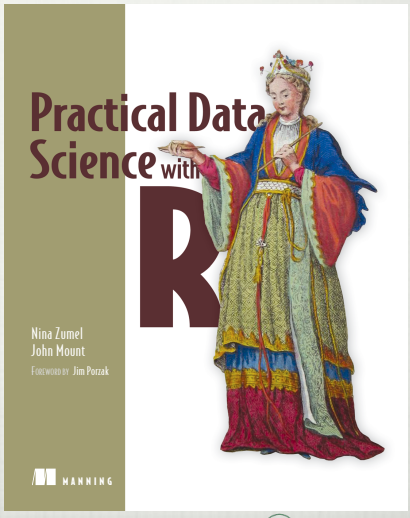
And some "engineering for statisticians"

- Minimal Version Control Lesson: Use It
  - http://wp.me/pgDAI-xF
- You don't need to understand pointers to program using R
  - http://wp.me/pgDAI-FZ
- R style tip: prefer functions that return data frames
  - http://wp.me/pgDAI-HG
- Trimming the Fat from glm() Models in R
  - http://wp.me/pgDAI-Ho

The book

The book: Practical Data Science with R

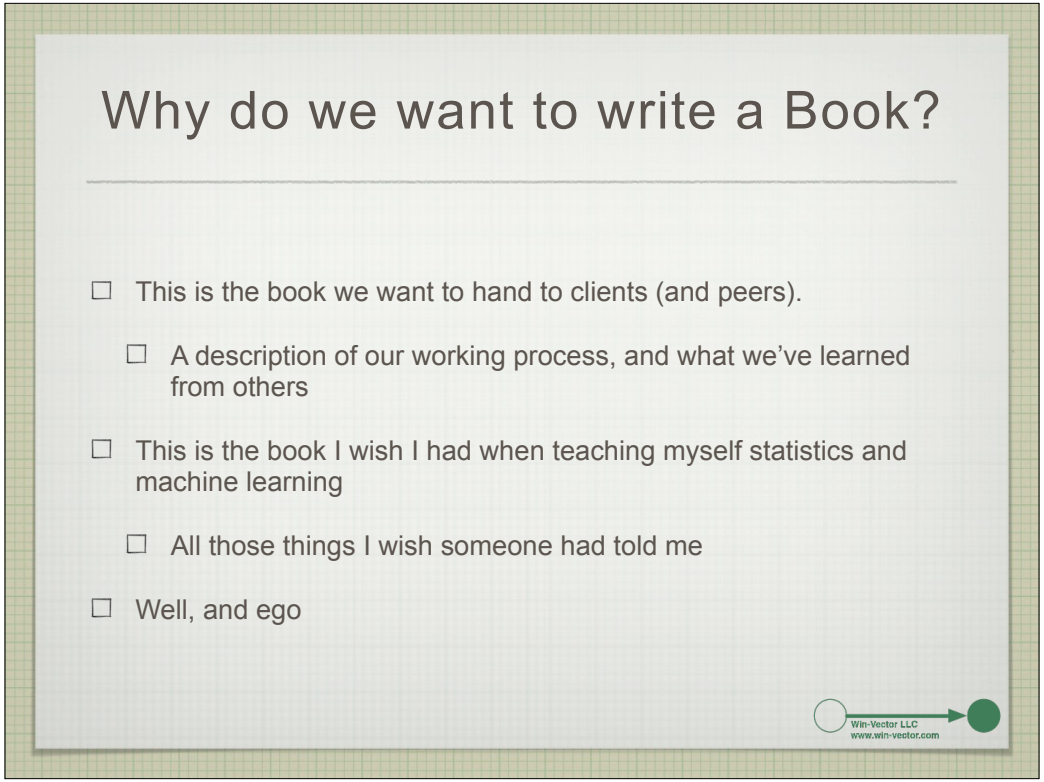
- Some color on:
  - The trade-offs we had to consider
  - The choices we've made, and why



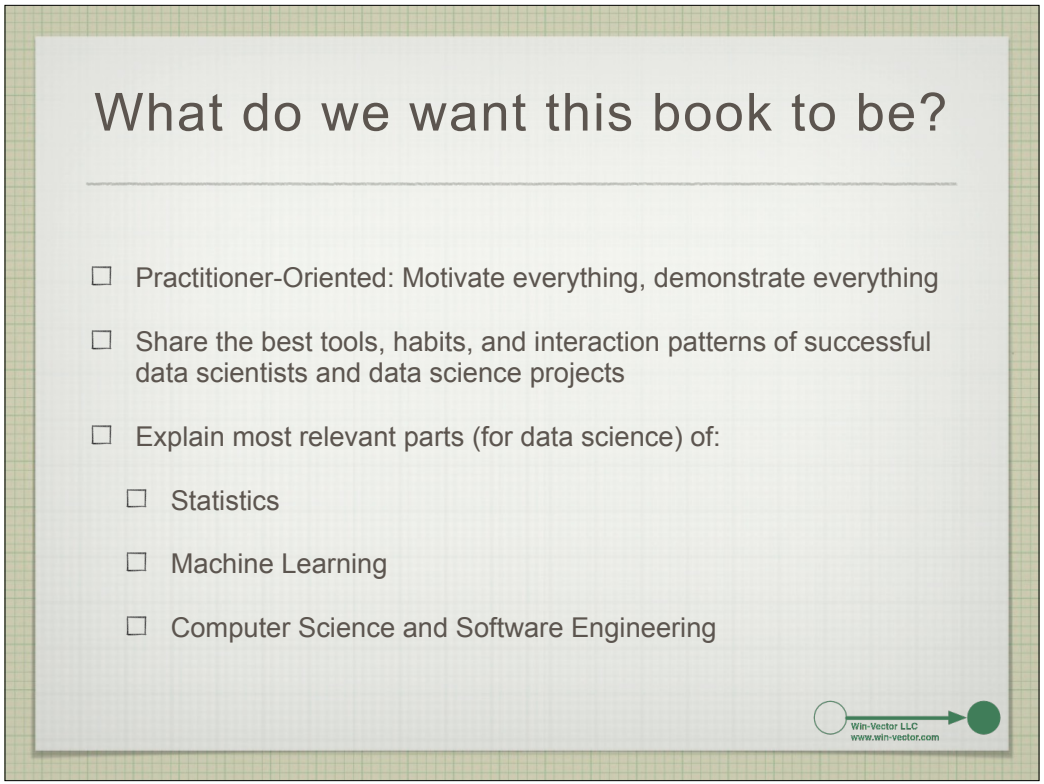
Win-Vector LLC
www.win-vector.com

Manning - primarily a technical manual publisher
Topic - data science from a practitioner's point of view

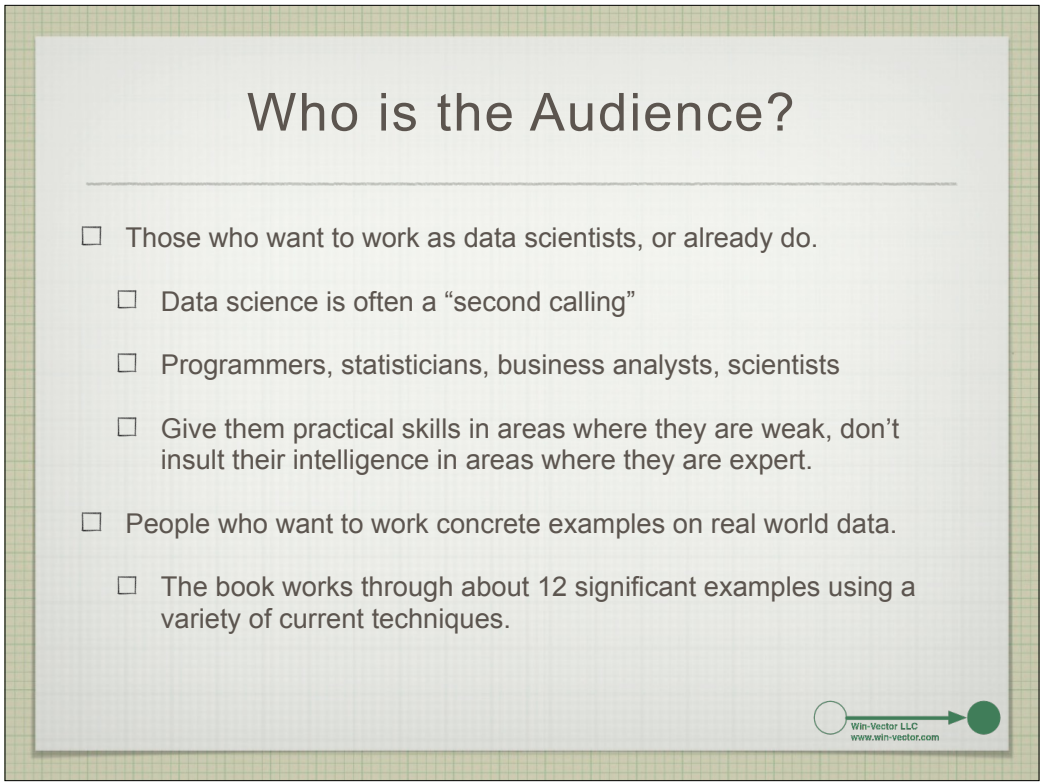
We weren't intimidated by the idea of a book going into it, because we felt that the blog gave us a lot of practice in writing to our target audience, and we'd given this topic a lot of thought.
If we can write the blog, how hard could the book be?



Let's start with why we wanted to write a book in the first place, since our motives color the choices we made.

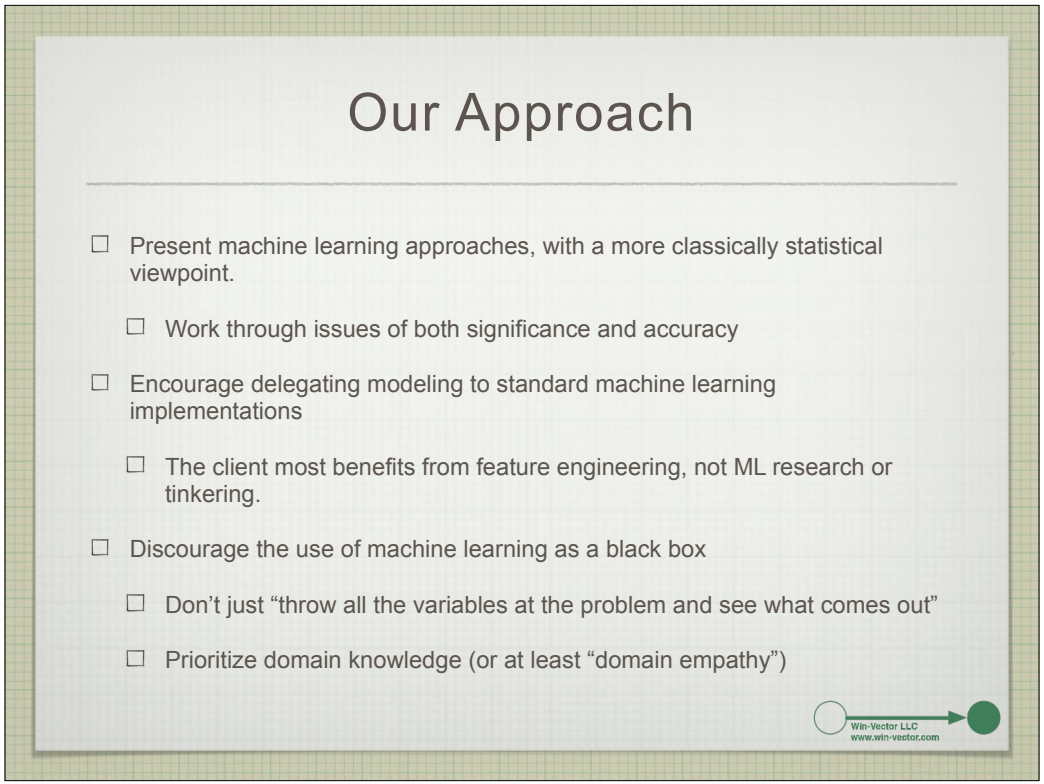


Practitioner-oriented: start with the problem, the move to the solution. less abstract, theory as appropriate
Best practices -- as we've learned the hard way, and as we've learned from others around us.
The FIRST reference book, but not the only one. Read us first; the deeper books will make more sense



There will be areas where the reader knows more that what we cover (or even what we know), and areas where their knowledge is thin.

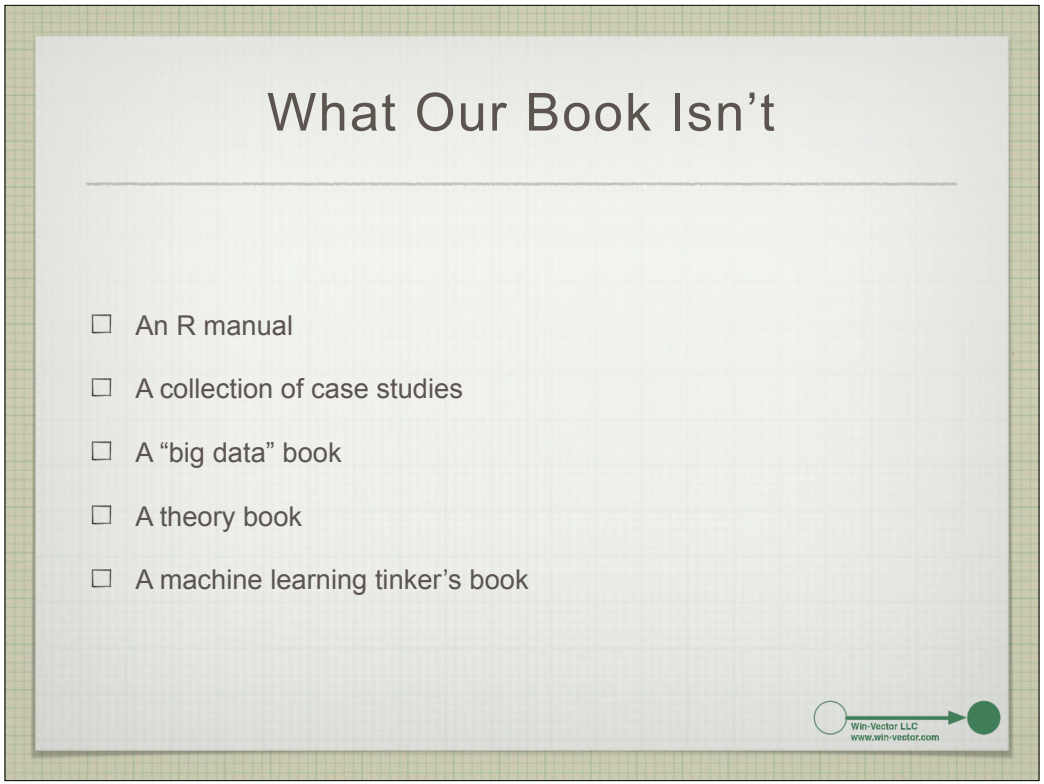
The need to have confidence that the book gives competent instruction in all areas -- and they will judge the book based on the topics that they know.ß



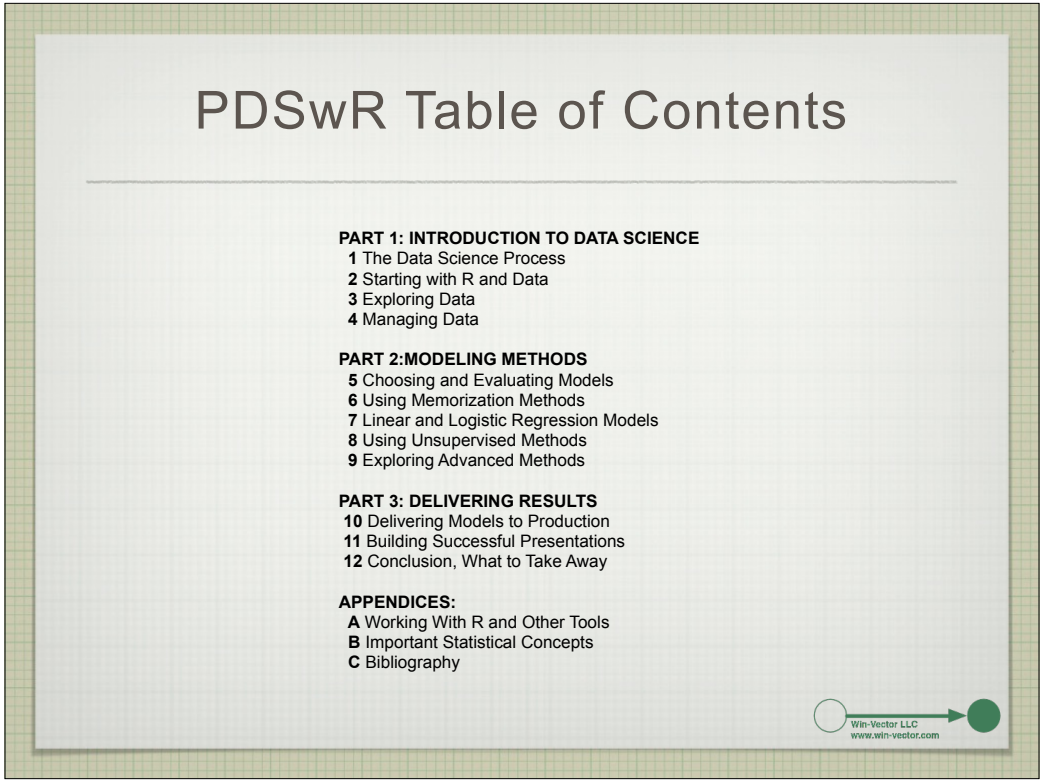
Implicit assumption in the book: there is a client.

For the book, we present several of the common machine learning algorithms, but we want to focus a bit more attention on issues of inference and significance, so we've tried to frame our discussion from a more classical statistical viewpoint. For every algorithm we try to explain how to evaluate the learned models for their reality, not just their accuracy.
We also discourage the use of machine learning as a black box.... and emphasize the importance of domain knowledge in the variable selection process.

other examples; masking effects, omitted variable bias, collinearity, dependent/correlated inputs



These are all fine things, and in-demand things; we just don't want to write any of them.



From all of the points we wanted to cover, we eventually settled on this structure for the book. [Pt1, 2...] Part 3 is about effectively disseminating project results and deploying the models into production
* William Cleveland: visualization as instrument for statistical analysis. Different visualizations expose different relationships in data.
[End]A big chunk of the book is about interactions between the data scientist and other interested parties, Chapter 5 in particular runs through some exercises on how to convert customer interviews into appropriate model scoring criteria. We try to emphasize how access to domain knowledge is a necessary component of selecting appropriate variables, interpreting observations and results.

Part 1 talks about the data science process, and about framing the problem, identifying the need, and evaluating the data in light of the need. This includes data visualization (Chapter 3): we follow William Cleveland's philosophy of visualization as an instrument for statistical analysis, and we focus on how different types of visualzations expose specific types of relationships in data.

Thank you

These slides:
https://github.com/WinVector/Examples/tree/master/DSTalk

For more information please try our blog:
http://www.win-vector.com/blog/
and the book
"Practical Data Science with R"
http://practicaldatascience.com .

43% off on Practical Data Science with R with code:
pdswr614 at www.manning.com/zumel/

Please contact us with comments, questions,
ideas, projects at:
jmount@win-vector.com
nzumel@win-vector.com

Follow us on Twitter: @WinVectorLLC