Careers in Data Science (You Know, Statistics)

Norm Matloff
University of California, Davis

Menlo-Atherton High School

May 24, 2016

http://heather.cs.ucdavis.edu/MenloAtherton.pdf

Confusing Terms

Confusing Terms

Confusing Terms

- We live in the Age of Buzzwords.

Confusing Terms

- We live in the Age of Buzzwords.
- Trust me:

Confusing Terms

- We live in the Age of Buzzwords.
- Trust me: Data Science = Statistics.

Confusing Terms

- We live in the Age of Buzzwords.
- Trust me: Data Science = Statistics.
- But isn't Data Science = Computer Science + Statistics?

Confusing Terms

- We live in the Age of Buzzwords.
- Trust me: Data Science = Statistics.
- But isn't Data Science = Computer Science + Statistics?
- I'm a computer scientist and a statistician — and I say No.

Confusing Terms

- We live in the Age of Buzzwords.
- Trust me: Data Science = Statistics.
- But isn't Data Science = Computer Science + Statistics?
- I'm a computer scientist and a statistician — and I say No.
- Statisticians have always had to be highly skilled with computers.

Norm Matloff
University of California, Davis

Menlo-Atherton High School

More Confusing Terms

More Confusing Terms

- *Big Data:*

More Confusing Terms

- *Big Data:*
  - Yes, have many huge data sets these days.

More Confusing Terms

- *Big Data:*
  - Yes, have many huge data sets these days.
  - Yes, typically requires parallel computation (one of my areas).

More Confusing Terms

- *Big Data:*
  - Yes, have many huge data sets these days.
  - Yes, typically requires parallel computation (one of my areas).
  - But still, not really a new paradigm.

More Confusing Terms

- *Big Data:*
  - Yes, have many huge data sets these days.
  - Yes, typically requires parallel computation (one of my areas).
  - But still, not really a new paradigm.
- *Machine Learning:*

More Confusing Terms

- *Big Data:*
  - Yes, have many huge data sets these days.
  - Yes, typically requires parallel computation (one of my areas).
  - But still, not really a new paradigm.
- *Machine Learning:*
  - Fancy new term for use of data for prediction.

More Confusing Terms

- *Big Data:*
  - Yes, have many huge data sets these days.
  - Yes, typically requires parallel computation (one of my areas).
  - But still, not really a new paradigm.
- *Machine Learning:*
  - Fancy new term for use of data for prediction. Statisticians have been doing that since 1804, thank you.

More Confusing Terms

- *Big Data:*
  - Yes, have many huge data sets these days.
  - Yes, typically requires parallel computation (one of my areas).
  - But still, not really a new paradigm.
- *Machine Learning:*
  - Fancy new term for use of data for prediction. Statisticians have been doing that since 1804, thank you.
  - Methods either invented by statisticians (e.g. Random Forests) or statistically motivated.

How to Become a GOOD Data Scientist

How to Become a GOOD Data Scientist

- I'll get to examples of Data Science shortly.

How to Become a GOOD Data Scientist

- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?

Norm Matloff
University of California, Davis

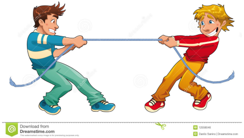
Menlo-Atherton High School

How to Become a GOOD Data Scientist

- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?
- Study Stat vs. CS?

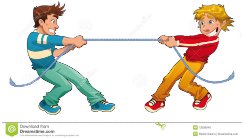
How to Become a GOOD Data Scientist

- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?
- Study Stat vs. CS?



How to Become a GOOD Data Scientist

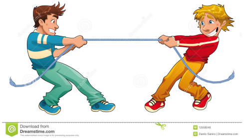
- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?
- Study Stat vs. CS?



- A bit of tug-of-war between these two fields these days.

How to Become a GOOD Data Scientist

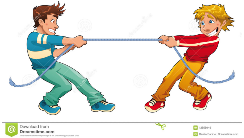
- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?
- Study Stat vs. CS?



- A bit of tug-of-war between these two fields these days. See Statistics Losing Ground to Computer Science, N. Matloff, *AMSTAT News*, Nov. 2014,

How to Become a GOOD Data Scientist

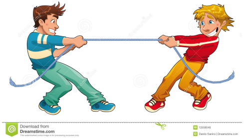
- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?
- Study Stat vs. CS?



- A bit of tug-of-war between these two fields these days. See Statistics Losing Ground to Computer Science, N. Matloff, *AMSTAT News*, Nov. 2014, and heated arguments on R vs. Python on Quora.

How to Become a GOOD Data Scientist

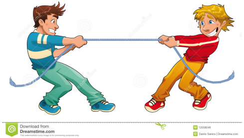
- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?
- Study Stat vs. CS?



- A bit of tug-of-war between these two fields these days. See Statistics Losing Ground to Computer Science, N. Matloff, *AMSTAT News*, Nov. 2014, and heated arguments on R vs. Python on Quora.
- My bias is Stat, with lots of CS, at least a CS minor.

How to Become a GOOD Data Scientist

- I'll get to examples of Data Science shortly. But first, what does one need to do Data Science well?
- Study Stat vs. CS?



- A bit of tug-of-war between these two fields these days. See Statistics Losing Ground to Computer Science, N. Matloff, *AMSTAT News*, Nov. 2014, and heated arguments on R vs. Python on Quora.
- My bias is Stat, with lots of CS, at least a CS minor. Also, advanced linear algebra (matrix theory).

What Really Counts

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization.

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization. **QUESTION THINGS!**

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization. **QUESTION THINGS!**
- E.g., if you've studied calculus: To minimize f, we set f' = 0.

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization. **QUESTION THINGS!**
- E.g., if you've studied calculus: To minimize f, we set f' = 0. Then check f" > 0.

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization. **QUESTION THINGS!**
- E.g., if you've studied calculus: To minimize f, we set f' = 0. Then check f" > 0. **BUT DO YOU KNOW WHY?**

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization. **QUESTION THINGS!**
- E.g., if you've studied calculus: To minimize f, we set f' $=$ 0. Then check f" $>$ 0. **BUT DO YOU KNOW WHY?**
- In statistics, in computing sample variance, we divide by $n-1$, not n:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization. **QUESTION THINGS!**
- E.g., if you've studied calculus: To minimize f, we set f' $=$ 0. Then check f" $>$ 0. **BUT DO YOU KNOW WHY?**
- In statistics, in computing sample variance, we divide by $n-1$, not n:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Why not divide by n? **IS THERE A GOOD REASON FOR THIS?**

What Really Counts

- FAR more important than Stat vs. CS: **Depth of insight**, not rote memorization. **QUESTION THINGS!**
- E.g., if you've studied calculus: To minimize f, we set f' = 0. Then check f" > 0. **BUT DO YOU KNOW WHY?**
- In statistics, in computing sample variance, we divide by $n-1$, not n:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

  Why not divide by n? **IS THERE A GOOD REASON FOR THIS?**
- Can you recognize Simpson's Paradox when you see it?

One More Slide on Prep for DS Career

One More Slide on Prep for DS Career

- GENERAL knowledge, awareness and insight are key!

One More Slide on Prep for DS Career

- GENERAL knowledge, awareness and insight are key!
- Do you know terms like *ameliorate*, *morbidity* and *elasticity of demand*?

One More Slide on Prep for DS Career

- GENERAL knowledge, awareness and insight are key!
- Do you know terms like *ameliorate*, *morbidity* and *elasticity of demand*?
- How about *FDA*, *HOV* and *CPI*?

One More Slide on Prep for DS Career

- GENERAL knowledge, awareness and insight are key!
- Do you know terms like *ameliorate*, *morbidity* and *elasticity of demand*?
- How about *FDA*, *HOV* and *CPI*?
- Can't be a good data miner without *understanding the data*!

One More Slide on Prep for DS Career

- GENERAL knowledge, awareness and insight are key!
- Do you know terms like *ameliorate*, *morbidity* and *elasticity of demand*?
- How about *FDA*, *HOV* and *CPI*?
- Can't be a good data miner without *understanding the data*! Ptolemy's epicycles fiasco.

All Right, Then, What Do DS People Do?

All Right, Then, What Do DS People Do?

- Example: Software running in a satellite notices a bright light in a forest.

All Right, Then, What Do DS People Do?

- Example: Software running in a satellite notices a bright light in a forest. Is it a fire? Maybe just a reflection?

All Right, Then, What Do DS People Do?

- Example: Software running in a satellite notices a bright light in a forest. Is it a fire? Maybe just a reflection? How can previous data be used here?

All Right, Then, What Do DS People Do?

- Example: Software running in a satellite notices a bright light in a forest. Is it a fire? Maybe just a reflection? How can previous data be used here?
- Example: You, humans, can spot the speaker:

All Right, Then, What Do DS People Do?

- Example: Software running in a satellite notices a bright light in a forest. Is it a fire? Maybe just a reflection? How can previous data be used here?
- Example: You, humans, can spot the speaker:

But can **software** spot me?

All Right, Then, What Do DS People Do?

- Example: Software running in a satellite notices a bright light in a forest. Is it a fire? Maybe just a reflection? How can previous data be used here?
- Example: You, humans, can spot the speaker:

But can **software** spot me?

Norm Matloff
University of California, Davis

Menlo-Atherton High School

Detailed Example

Detailed Example

Goals:

Detailed Example

Goals:

- Show you something different from AP Stat.

Detailed Example

Goals:

- Show you something different from AP Stat.
- Show that serious math is involved (calculus, matrix theory).

Detailed Example

Goals:

- Show you something different from AP Stat.
- Show that serious math is involved (calculus, matrix theory).
- This will get a little technical; don't feel that you need to follow 100%.

Detailed Example

Goals:

- Show you something different from AP Stat.
- Show that serious math is involved (calculus, matrix theory).
- This will get a little technical; don't feel that you need to follow 100%.
- The Question: Will Mary like the movie *Captain America*?

Norm Matloff
University of California, Davis

Menlo-Atherton High School

Will Mary Like the Movie *Captain America*?

Will Mary Like the Movie *Captain America*?

- Mary hasn't seen the movie.

Will Mary Like the Movie *Captain America*?

- Mary hasn't seen the movie.
- But we have Mary's ratings on some other movies, and we have ratings of *Captain America* by some other people.

Will Mary Like the Movie *Captain America*?

- Mary hasn't seen the movie.
- But we have Mary's ratings on some other movies, and we have ratings of *Captain America* by some other people.
- How do we use this data to guess Mary's rating of this movie?

Norm Matloff University of California, Davis

Menlo-Atherton High School

Neighborhood-Based Approach

Neighborhood-Based Approach

- Find people ("neighbors") in our dataset whose movie tastes are similar to Mary's.

Neighborhood-Based Approach

- Find people ("neighbors") in our dataset whose movie tastes are similar to Mary's.
- Of those, focus on the ones that have seen *Captain America*.

Neighborhood-Based Approach

- Find people ("neighbors") in our dataset whose movie tastes are similar to Mary's.
- Of those, focus on the ones that have seen *Captain America*.
- Guess Mary's rating of the movie to be the mean of the ratings in that group.

A More Nuanced Model

Norm Matloff University of California, Davis

Menlo-Atherton High School

A More Nuanced Model

- Y_{ij} is user i's rating of movie j.

A More Nuanced Model

- Y_{ij} is user i's rating of movie j.
- We want to predict $Y_{Mary,Cpt.Am.}$.

A More Nuanced Model

- Y_{ij} is user i's rating of movie j.
- We want to predict $Y_{Mary,Cpt.Am.}$.
- Latent Factor Model:

A More Nuanced Model

- Y_{ij} is user i's rating of movie j.
- We want to predict $Y_{Mary,Cpt.Am.}$.
- Latent Factor Model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where

  - $\mu =$ mean ratings over all users and all movies.

A More Nuanced Model

- Y_{ij} is user i's rating of movie j.
- We want to predict $Y_{Mary,Cpt.Am.}$.
- Latent Factor Model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where

  - μ = mean ratings over all users and all movies.
  - α_i = tendency for user i to give higher/lower ratings than the typical user

Norm Matloff
University of California, Davis

Menlo-Atherton High School

A More Nuanced Model

- Y_{ij} is user i's rating of movie j.
- We want to predict $Y_{Mary,Cpt.Am.}$.
- Latent Factor Model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where

  - μ = mean ratings over all users and all movies.
  - α_i = tendency for user i to give higher/lower ratings than the typical user
  - β_j = tendency for movie j to be rated higher/lower than the typical movie

A More Nuanced Model

- Y_{ij} is user i's rating of movie j.
- We want to predict $Y_{Mary,Cpt.Am.}$.
- Latent Factor Model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where

- μ = mean ratings over all users and all movies.
- α_i = tendency for user i to give higher/lower ratings than the typical user
- β_j = tendency for movie j to be rated higher/lower than the typical movie
- ϵ_{ij} = sum of all unknown effects, e.g. user i's mood when viewing movie j

Where Does the Data Come in?

Where Does the Data Come in?

So, we really have a statistical estimation problem:

Where Does the Data Come in?

So, we really have a statistical estimation problem:

- We'll use our data to estimate μ, α_{Mary} and $\beta_{Cpt.Am.}$.

Where Does the Data Come in?

So, we really have a statistical estimation problem:

- We'll use our data to estimate μ, α_{Mary} and $\beta_{Cpt.Am.}$. This gives us (stat. notation) $\widehat{\mu}$ etc.

Where Does the Data Come in?

So, we really have a statistical estimation problem:

- We'll use our data to estimate μ, α_{Mary} and $\beta_{Cpt.Am.}$. This gives us (stat. notation) $\widehat{\mu}$ etc.
- We then guess Mary's rating of *Cpt. Am.* to be

$$\widehat{\mu} + \widehat{\alpha}_{Mary} + \widehat{\beta}_{Cpt.Am.}$$

Where Does the Data Come in?

So, we really have a statistical estimation problem:

- We'll use our data to estimate μ, α_{Mary} and $\beta_{Cpt.Am.}$. This gives us (stat. notation) $\widehat{\mu}$ etc.
- We then guess Mary's rating of *Cpt. Am.* to be

$$\widehat{\mu} + \widehat{\alpha}_{Mary} + \widehat{\beta}_{Cpt.Am.}$$

- But HOW will we get those estimates?

One Common Method

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.
- Mary has a row in the matrix.

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.
- Mary has a row in the matrix.
- *Captain America* has a column in the matrix.

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.
- Mary has a row in the matrix.
- *Captain America* has a column in the matrix.
- Unfortunately, the entry in Mary's row and *Captain America*'s column is unknown.

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.
- Mary has a row in the matrix.
- *Captain America* has a column in the matrix.
- Unfortunately, the entry in Mary's row and *Captain America*'s column is unknown.
- But some other entries in Mary's row *are* known. Same for *Captain America*'s column.

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.
- Mary has a row in the matrix.
- *Captain America* has a column in the matrix.
- Unfortunately, the entry in Mary's row and *Captain America*'s column is unknown.
- But some other entries in Mary's row *are* known. Same for *Captain America*'s column.
- Can use calculus, matrix theory to estimate the missing entries in the matrix.

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.
- Mary has a row in the matrix.
- *Captain America* has a column in the matrix.
- Unfortunately, the entry in Mary's row and *Captain America*'s column is unknown.
- But some other entries in Mary's row *are* known. Same for *Captain America*'s column.
- Can use calculus, matrix theory to estimate the missing entries in the matrix. Gives us matrices P and Q such that

$$A \approx PQ$$

One Common Method

A popular way to obtain those estimates is *matrix factorization*.

- Define A to be the matrix of the ratings — even the unknown ones.
- Mary has a row in the matrix.
- *Captain America* has a column in the matrix.
- Unfortunately, the entry in Mary's row and *Captain America*'s column is unknown.
- But some other entries in Mary's row *are* known. Same for *Captain America*'s column.
- Can use calculus, matrix theory to estimate the missing entries in the matrix. Gives us matrices P and Q such that

$$A \approx PQ$$

Details not shown. :-)

Summary

Summary

- Beware of the buzzwords like *data science*. Modern methodology is not really new.

Summary

- Beware of the buzzwords like *data science*. Modern methodology is not really new.
- But...things ARE different today.

Summary

- Beware of the buzzwords like *data science*. Modern methodology is not really new.
- But...things ARE different today. The applications are highly engaging, and we have powerful computers to handle the large volume of data.

Summary

- Beware of the buzzwords like *data science*. Modern methodology is not really new.
- But...things ARE different today. The applications are highly engaging, and we have powerful computers to handle the large volume of data.
- Don't fall into the trap of thinking that taking Course X or studying Major Y is sufficient!

Summary

- Beware of the buzzwords like *data science*. Modern methodology is not really new.
- But...things ARE different today. The applications are highly engaging, and we have powerful computers to handle the large volume of data.
- Don't fall into the trap of thinking that taking Course X or studying Major Y is sufficient!
- You need to be strong in math — this means INSIGHT, not just facility with equations — and be very AWARE of the world around you.

Summary

- Beware of the buzzwords like *data science*. Modern methodology is not really new.
- But...things ARE different today. The applications are highly engaging, and we have powerful computers to handle the large volume of data.
- Don't fall into the trap of thinking that taking Course X or studying Major Y is sufficient!
- You need to be strong in math — this means INSIGHT, not just facility with equations — and be very AWARE of the world around you.
- Good luck to you!