What data scientists do

March 19, 2012
My background

• Harvard Ph.D. in pure math, 1999
My background

- Harvard Ph.D. in pure math, 1999
- M.I.T. post doc, 1999-2005
My background

- Harvard Ph.D. in pure math, 1999
- M.I.T. post doc, 1999-2005
- Barnard Assistant Professor, 2005-2007
My background

- Harvard Ph.D. in pure math, 1999
- M.I.T. post doc, 1999-2005
- Barnard Assistant Professor, 2005-2007
- D.E. Shaw quant, 2007-2009
My background

• Harvard Ph.D. in pure math, 1999
• M.I.T. post doc, 1999-2005
• Barnard Assistant Professor, 2005-2007
• D.E. Shaw quant, 2007-2009
• RiskMetrics researcher, 2009-2011
Data Scientist

- I left finance in January 2011
Data Scientist

- I left finance in January 2011
- Got a data scientist job at Intent Media
Data Scientist

• I left finance in January 2011
• Got a data scientist job at Intent Media
• Started a blog
Data Scientist

• I left finance in January 2011
• Got a data scientist job at Intent Media
• Started a blog
• Started working with Occupy Wall Street
What do we do?

• I'll compare to being an academic

•

•
What do we do?

• I'll compare to being an academic
• Also to being a quant
What do we do?

• I'll compare to being an academic
• Also to being a quant
• What one prefers is a personality thing
What do we do?

• I'll compare to being an academic
• Also to being a quant
• What one prefers is a personality thing
• Job opportunities are not all created equal!
The academic's day

• Teach
• Talk to students
• Do research
• Go to seminars
• Organize and attend conferences
The quant's day

- Model (mostly by yourself)
- Talk to colleagues
- Keep an eye on the market and chatter
- Go to meetings
- Vacation in the Hamptons
The data scientist's day

• Model (mostly in collaboration)
• Talk to business people
• Keep an eye on domain news
• Go to meetings
• Organize and attend conferences
Talking to business people

• Teach: they are only somewhat quantitative
Talking to business people

• Teach: they are only somewhat quantitative
• Help them keep an eye on their business
Talking to business people

• Teach: they are only somewhat quantitative
• Help them keep an eye on their business
• Show them what models can do
Talking to business people

• Teach: they are only somewhat quantitative
• Help them keep an eye on their business
• Show them what models can do
• Estimate new business ideas (McKinsey)
Let's compare the modeling

<table>
<thead>
<tr>
<th></th>
<th>Finance</th>
<th>Internet</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td>big</td>
<td>huge</td>
</tr>
<tr>
<td><strong>Signal</strong></td>
<td>weak</td>
<td>strong</td>
</tr>
<tr>
<td><strong>Sophistication</strong></td>
<td>high</td>
<td>variable</td>
</tr>
<tr>
<td><strong>Forecasted</strong></td>
<td>return</td>
<td>behavior</td>
</tr>
</tbody>
</table>
Data

- In finance: flat files fed to python code
Data

• In finance: flat files fed to python code
• (except for high frequency trading)
Data

• In finance: flat files fed to python code
• (except for high frequency trading)
• In internet modeling: too much data
Data

• In finance: flat files fed to python code
• (except for high frequency trading)
• In internet modeling: too much data
• Need MapReduce methods in the cloud
Signals in Finance

- Returns have weak signals
Signals in Finance

- Returns have weak signals
- Otherwise we'd be stinking rich

-
Signals in Finance

- Returns have weak signals
- Otherwise we'd be stinking rich
- We use linear regression
Signals in Finance

• Returns have weak signals
• Otherwise we'd be stinking rich
• We use linear regression
• Complicated bayesian priors
Signals in Finance

- Returns have weak signals
- Otherwise we'd be stinking rich
- We use linear regression
- Complicated bayesian priors
- Complicated ways to avoid overfitting
Signals in machine learning

- Data scientists come from machine learning
Signals in machine learning

- Data scientists come from machine learning
- The signals are very strong
Signals in machine learning

- Data scientists come from machine learning
- The signals are very strong
- You can "see if the model is working"
Signals in machine learning

• Data scientists come from machine learning
• The signals are very strong
• You can "see if the model is working"
• That becomes kind of a mindset
Sophistication

• Neural networks, decision trees etc.
Sophistication

- Neural networks, decision trees etc.
- Not particularly subtle
Sophistication

- Neural networks, decision trees etc.
- Not particularly subtle
- They (mostly) don't need to be
Sophistication

• Neural networks, decision trees etc.
• Not particularly subtle
• They (mostly) don't need to be
• Seasonality, updating common to both
How clicks work

• Advertisers only pay when someone clicks
How clicks work

• Advertisers only pay when someone clicks
• Auctions to decide positions
How clicks work

• Advertisers only pay when someone clicks
• Auctions to decide positions
• Adrank is combo of bid and "quality score"
How clicks work

• Advertisers only pay when someone clicks
• Auctions to decide positions
• Adrank is combo of bid and "quality score"
• CTR, interconnectedness, spamminess
Internet ad models...

• Improve CTR through "A/B testing"
Internet ad models...

• Improve CTR through "A/B testing"
• Tailor ads through "segmentation models"
Internet ad models...

• Improve CTR through "A/B testing"
• Tailor ads through "segmentation models"
• Forecast behavior based on cookies, etc.
Internet ad models...

• Improve CTR through "A/B testing"
• Tailor ads through "segmentation models"
• Forecast behavior based on cookies, etc.
• Privacy issues obviously relevant here
Job opportunities

- Profs don't suggest students leave

-
Job opportunities

• Profs don't suggest students leave
• Snobbery and ignorance (personality)
Job opportunities

• Profs don't suggest students leave
• Snobbery and ignorance (personality)
• Finance is still shrinking somewhat
Job opportunities

• Profs don't suggest students leave
• Snobbery and ignorance (personality)
• Finance is still shrinking somewhat
• Internet stuff is exploding
Ethics

• In finance we go after "dumb money"
• "innovations" are weird instruments, contracts, and models
• These are highly scalable
• The market anonymizes everything
• The power structure and information asymmetry are key to this working
Ethics

• On the internet we go after users
• Our "innovations" are models which categorize and profile people
• These are highly scalable
• The internet anonymizes everything
• Trolled cookies make this work ("do not track" is a joke)
What we should understand

• The modeling death spirals
•
•
What we should understand

• The modeling death spirals
• Ex: insurance, credit scores
What we should understand

• The modeling death spirals
• Ex: insurance, credit scores
• These have an effect on our culture
What we should understand

• The modeling death spirals
• Ex: insurance, credit scores
• These have an effect on our culture
• They are becoming ever more prevalent
What we should do

• Modeler's Hippocratic Oath
What we should do

• Modeler's Hippocratic Oath
• Democratic society
What we should do

• Modeler's Hippocratic Oath
• Democratic society
• Quants need to work with ethicists
What we should do

• Modeler's Hippocratic Oath
• Democratic society
• Quants need to work with ethicists
• Formally and seriously address these issues