

V0A 2014-Schield-DSI-Stats-Curriculum 1

Business Analytics vs. Data Science

by
Milo Schield
*Member: International Statistical Institute
US Rep: International Statistical Literacy Project
Director, W. M. Keck Statistical Literacy Project*
**Presented at the
Annual Decision Sciences Institute Meeting
Tampa FL. Nov 22, 2014.**

Slides at: www.StatLit.org/pdf/2014-Schield-DSI2-Slides.pdf

V0A 2014-Schield-DSI-Stats-Curriculum 2

Data Science (DS), Data Analytics (DA), Business Analytics (BA)



In any new field, new terms are a bit vague. Distinctions are shades of grey; not black-white.

DS, DA, BA all involve some combination of mathematics, computer science and statistics.

Ideally, a DS major would take a substantial number of courses in all three areas. Ideally, they work from start-to-finish on a DS project.

Most students don't have time for this.

V0A 2014-Schield-DSI-Stats-Curriculum






Mathematics Aspect

2nd Stat courses can be classified by:

- Math pre-req: algebra, pre-calc or calculus.
- Topics: Just regression (Mendenhall-Sincich, Draper-Smith). Multilevel / hierarchical models (Gelman-Hill). Multivariate methods: cluster analysis, discriminant analysis, factor analysis, principle components, logistic regression, etc. (Sharma, Johnson-Wichern, Berenson-Levine-Goldstein)

V0A 2014-Schield-DSI-Stats-Curriculum

Computer Science Perspective

Data acquisition, manipulation & summarization are big topics in Computer Science.

Computer software is a big issue: SQL databases, SAS, R, Hadoop, etc.

V0A 2014-Schield-DSI-Stats-Curriculum 5

Data Science

Data science is dominated by computer scientists and mathematicians. The primary focus is on associations: correlations, models, prediction ...

Neither computer science nor mathematics has any language for causation. Both focus on what is necessary or sufficient.

Both mathematics and computer science focus on the form – and generally eschew the matter.

V0A 2014-Schield-DSI-Stats-Curriculum 6

Business Analytics

For science, the goal is truth – deep truths. For the physical sciences, the truth typically includes causal connections. For math and computer science, causation is conspicuously absent.

For business, the goal is create products and services that will be bought by customers at a price that generates a profit. Sometimes this involves prediction; other times is involves an intervention. Both of these involve causal connections.

V0A 2014-Schild-DSI-Stats-Curriculum 7

Four Big Ideas in Teaching Big-Data

- 1. Association** is not causation, but is often a sign of causation somewhere.
- 2. Confounding.** Why getting more data may not reduce confounding.
- 3. Coincidence:** Why coincidence increases as the amount of data (# of rows) increases.
- 4. Error:** Why errors (false positives) increase as the object of interest gets smaller (rarer).

V0A 2014-Schild-DSI-Stats-Curriculum 8

Statistical Literacy: Big Idea #1: Association

Just saying “Association is not Causation” exemplifies the “abstinence approach” to statistics.

Abstinence may be fine in a math class. It is not acceptable in a Business program where associations are typically a sign of causation somewhere.

Students should learn which statistical associations give stronger support for a causal connection.

V0A 2014-Schild-DSI-Stats-Curriculum 9

Statistical Literacy: Big Idea #2: Confounding

Confounders are related factors not taken into account in a study.

The influence of confounders [confounding] is omni-present in observational studies.

Simpson’s paradox (sign reversal or confounding)

- is incidental when modelling or forecasting,
- dominates when searching for causes.


V0A 2014-Schild-DSI-Stats-Curriculum 10

Statistical Literacy: Big Idea #3: Coincidence

Margin of error decreases as sample size increases.

The Law of Very Large Numbers: the unlikely becomes almost certain given enough tries.

Coincidence may be totally spurious or a sign of causation.



V0A 2014-Schild-DSI-Stats-Curriculum 11

Statistical Literacy: Big Idea #4: Tests

False positive are a constant problem in tests.

Qualitatively, the lower the prevalence of the group, the higher the chance of a false positive.

Quantitatively, if the prevalence of the group of interest is the same as the error rate in the test, then the prediction accuracy is always 50%.

The quantitative relationship is simple, memorable and helps in evaluating tests using Big Data.

V0A 2014-Schild-DSI-Stats-Curriculum 12

Conclusion

Business Analytics should focus on teaching the big ideas underlying the statistics produced by any analysis of observational data: big or small.

Business Analytics should help students see which associations give stronger support for a causal connection. They should be able to see the influence of confounders, of coincidence and Type-1 errors in big data.

VOA 2014-Schield-DSI-Stats-Curriculum 13

References

Berenson, (2013). Statistics Course for Big Data & Analytics. Slides at www.statlit.org/pdf/2013-Berenson-DSI-MSMESB-Slides.pdf

Berenson, (2013). Big Data Implications for Stat Analysis & Instruction. www.statlit.org/pdf/2013-Berenson2-DSI-MSMESB-Slides.pdf

Levine, Szabat & Stephan (2013). Data Discovery www.statlit.org/pdf/2013-Levine-Szabat-Stephan-DSI-MSMESB-Slides.pdf

Schield, M. (2014). Two Big Ideas for Teaching Big Data: ECOTS Paper at www.statlit.org/pdf/2014-Schield-ECOTS.pdf

Schield, M. (2014). Big Data: Coincidence. National Numeracy Network. www.statlit.org/pdf/2014-Schield-NNN1-Slides.pdf

Stine, D. (2013): Big Data Implications for intro stats. www.statlit.org/pdf/2013-Stine-DSI-MSMESB-Slides.pdf

Business Analytics vs. Data Science

by

Milo Schield

Member: International Statistical Institute

US Rep: International Statistical Literacy Project

Director, W. M. Keck Statistical Literacy Project

Presented at the

Annual Decision Sciences Institute Meeting

Tampa FL. Nov 22, 2014.

Slides at: www.StatLit.org/pdf/2014-Schield-DSI2-Slides.pdf

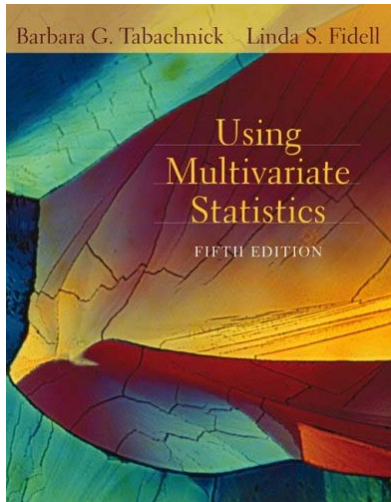
Data Science (DS), Data Analytics (DA), Business Analytics (BA)

In any new field, new terms are a bit vague. Distinctions are shades of grey; not black-white.

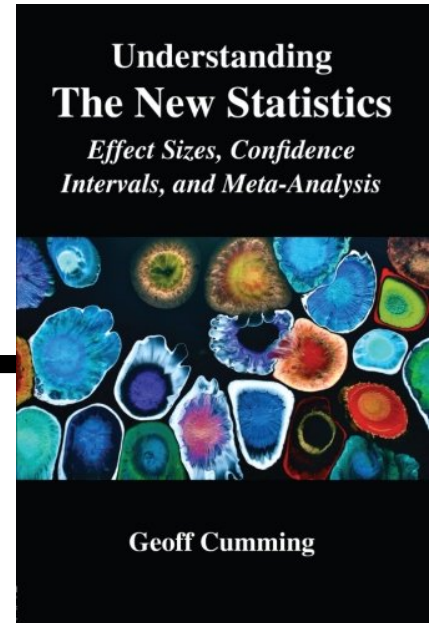
DS, DA, BA all involve some combination of mathematics, computer science and statistics.

Ideally, a DS major would take a substantial number of courses in all three areas. Ideally, they work from start-to-finish on a DS project.

Most students don't have time for this.

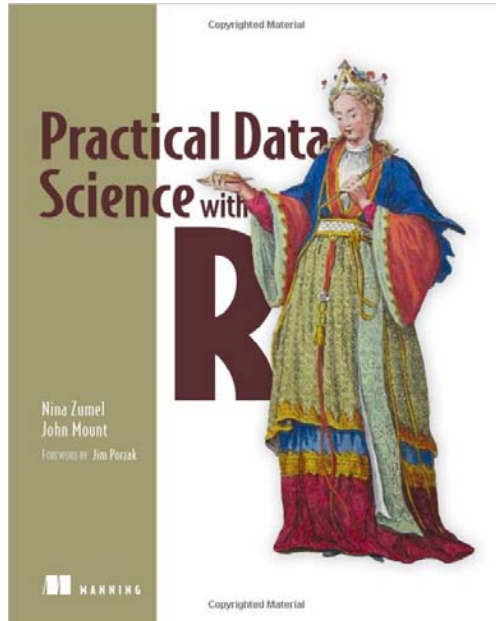


Mathematics Aspect

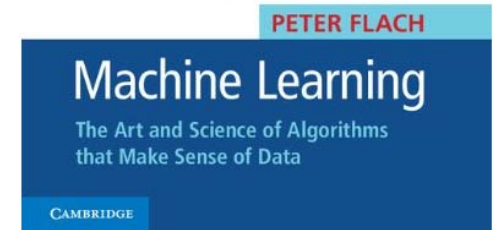
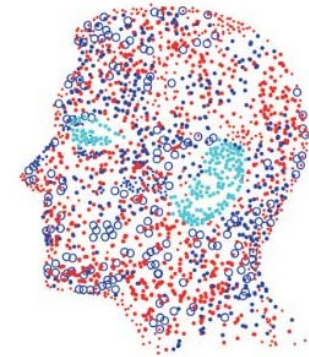


2nd Stat courses can be classified by:

- Math pre-req: algebra, pre-calc or calculus.
- Topics: Just regression (Mendenhall-Sincich, Draper-Smith). Multilevel / hierarchical models (Gelman-Hill). Multivariate methods: cluster analysis, discriminant analysis, factor analysis, principle components, logistic regression, etc. (Sharma, Johnson-Wichern, Berenson-Levine-Goldstein)



Computer Science Perspective



Data acquisition, manipulation & summarization are big topics in Computer Science.

Computer software is a big issue: SQL databases, SAS, R, Hadoop, etc.

Data Science

Data science is dominated by computer scientists and mathematicians. The primary focus is on associations: correlations, models, prediction ...

Neither computer science nor mathematics has any language for causation. Both focus on what is necessary or sufficient.

Both mathematics and computer science focus on the form – and generally eschew the matter.

Business Analytics

For science, the goal is truth – deep truths. For the physical sciences, the truth typically includes causal connections. For math and computer science, causation is conspicuously absent.

For business, the goal is create products and services that will be bought by customers at a price that generates a profit. Sometimes this involves prediction; other times is involves an intervention. Both of these involve causal connections.

Four Big Ideas in Teaching Big-Data

1. **Association** is not causation, but is often a sign of causation somewhere.
2. **Confounding.** Why getting more data may not reduce confounding.
3. **Coincidence:** Why coincidence increases as the amount of data (# of rows) increases.
4. **Error:** Why errors (false positives) increase as the object of interest gets smaller (rarer).

Statistical Literacy:

Big Idea #1: Association

Just saying “Association is not Causation” exemplifies the “abstinence approach” to statistics.

Abstinence may be fine in a math class. It is not acceptable in a Business program where associations are typically a sign of causation somewhere.

Students should learn which statistical associations give stronger support for a causal connection.

Statistical Literacy:

Big Idea #2: Confounding

Confounders are related factors not taken into account in a study.

The influence of confounders [confounding] is omni-present in observational studies.

Simpson's paradox (sign reversal or confounding)

- is incidental when modelling or forecasting,
- dominates when searching for causes.

Statistical Literacy:

Big Idea #3: Coincidence

Margin of error decreases as sample size increases.

The Law of Very Large Numbers: the unlikely becomes almost certain given enough tries.

Coincidence may be totally spurious or a sign of causation.



Statistical Literacy:

Big Idea #4: Tests

False positive are a constant problem in tests.

Qualitatively, the lower the prevalence of the group, the higher the chance of a false positive.

Quantitatively, if the prevalence of the group of interest is the same as the error rate in the test, then the prediction accuracy is always 50%.

The quantitative relationship is simple, memorable and helps in evaluating tests using Big Data.

Conclusion

Business Analytics should focus on teaching the big ideas underlying the statistics produced by any analysis of observational data: big or small.

Business Analytics should help students see which associations give stronger support for a causal connection. They should be able to see the influence of confounders, of coincidence and Type-1 errors in big data.

References

Berenson, (2013). Statistics Course for Big Data & Analytics. Slides at www.statlit.org/pdf/2013-Berenson-DSI-MSMESB-Slides.pdf

Berenson, (2013). Big Data Implications for Stat Analysis & Instruction. www.statlit.org/pdf/2013-Berenson2-DSI-MSMESB-Slides.pdf

Levine, Szabat & Stephan (2013). Data Discovery www.statlit.org/pdf/2013-Levine-Szabat-Stephan-DSI-MSMESB-Slides.pdf

Schield, M. (2014). Two Big Ideas for Teaching Big Data: ECOTS Paper at www.statlit.org/pdf/2014-Schield-ECOTS.pdf

Schield, M. (2014). Big Data: Coincidence. National Numeracy Network. www.statlit.org/pdf/2014-Schield-NNN1-Slides.pdf

Stine, D. (2013): Big Data Implications for intro stats. www.statlit.org/pdf/2013-Stine-DSI-MSMESB-Slides.pdf