Remarks for Panel Discussion

Hugh Chipman, Acadia University

A looming shortage?

A looming shortage?

Yesterday we heard about the “4 V’s” of Big Data:

A looming shortage?

Yesterday we heard about the "4 V's" of Big Data:

- Volume
- Velocity
- Variety
- Veracity

A looming shortage?

Yesterday we heard about the "4 V's" of Big Data:

- Volume
- Velocity
- Variety
- Veracity

And a few new ones:

- Value
- Visualization
- Variance

A looming shortage?

Yesterday we heard about the "4 V's" of Big Data:

- Volume
- Velocity
- Variety
- Veracity

And a few new ones:

- Value
- Visualization
- Variance

However, without a bigger list, big data research and education are doomed.

Some proposals to close the “V gap”

Some proposals to close the "V gap"

- Versatility

Some proposals to close the “V gap”

- Versatility
- Vigilance

Some proposals to close the “V gap”

- Versatility
- Vigilance
- Vectorized

Some proposals to close the “V gap”

- Versatility
- Vigilance
- Vectorized
- Verisimilitude

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)
- Very big data

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)
- Very big data
- Very very big data

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)
- Very big data
- Very very big data
- Very very very big data

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)
- Very big data
- Very very big data
- Very very very big data
- Very very very very big data

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)
- Very big data
- Very very big data
- Very very very big data
- Very very very very big data

However, asymptotic analysis (of the kind Andrew Rau-Chaplin mentioned) reveals a serious problem:

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)
- Very big data
- Very very big data
- Very very very big data
- Very very very very big data

However, asymptotic analysis (of the kind Andrew Rau-Chaplin mentioned) reveals a serious problem:

We're going to eventually run out of words beginning with "V".

Some proposals to close the "V gap"

- Versatility
- Vigilance
- Vectorized
- Verisimilitude
- Vanilla?
- Voluptuousness (is this a word?)
- Very big data
- Very very big data
- Very very very big data
- Very very very very big data

However, asymptotic analysis (of the kind Andrew Rau-Chaplin mentioned) reveals a serious problem:

We're going to eventually run out of words beginning with "V".

Is it too late to switch to "T"?

But seriously, folks....

My background is statistics.

I'll make a few remarks on what we teach our undergrads and masters students about computing.

The short answer is: not enough

We're still trying to get statistics students to do basics:

- Can I write code to estimate this model?
- How do I write it?
- Will the code run? (hint: not the first time)
- Are the results I get correct? (hint: probably not)

These are all questions that I suspect are pretty typical for a course in programming or software engineering.

Are the results I get correct? (probably not)

This last question is one that I think is important and probably one of the harder ones to teach.

Output from code that manipulates data is "random":

- Input data are random, so output will be random
- This makes debugging and checking for correctness difficult.
- Debugging/checking is a good test of statistical understanding.
- That is, "Can you generate some data where the result of your code should be obvious (without running it)?"
- Once such a "hello world" simulation is made, then additional random (simulated) test cases can be run through the code.
- Similar consistency checking of code should be applied to the analysis of data with existing tools.

Are the results I get correct? (probably not)

The **interplay** between programming and ideas of randomness is key: We understand a model better once we implement it as code, and we have confidence in the code once we see it agrees with multiple different random scenarios.

Additional remark: sophisticated approaches to scaling up code for very big data are another matter entirely.

- Some statisticians will learn to take it to this level.
- I suspect many students will become users of High Performance Computing, rather than producers of HPC code.

Stepping back from statistics, it is safe to say that students in all areas of "Big Data" should understand the randomness of data *and* be able to write code to explore it.