Battle of the Data Science Venn Diagrams

Previous post Next post

Tags: Data Science, Drew Conway, Venn Diagram

First came Drew Conway's data science Venn diagram. Then came all the rest. Read this comparative overview of data science Venn diagrams for both the insight into the profession and the humor that comes along for free.

comments

By David Taylor, Biotechnologist.

Data science is a rather fuzzily defined field; some of the definitions I've heard are:

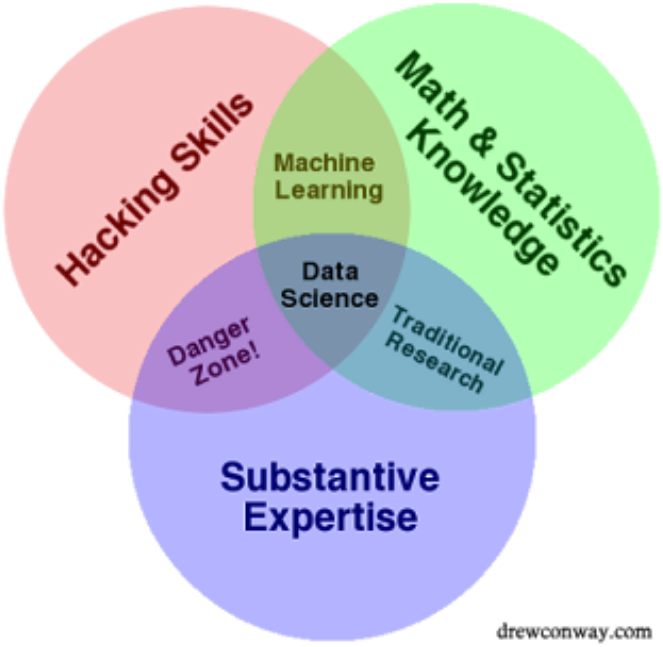
> "Work that takes more programming skills than most statisticians have, and more statistics skills than a programmer has."
>
> "Applied statistics, but in San Francisco."
>
> "The field of people who decide to print 'Data Scientist' on their business cards and get a salary bump."

Personally, I've recently decided to avoid the controversy by calling myself a data spelunker. (Data *miners* are out of vogue anyway.)

As a field in search of a definition, it's unsurprising that you can find a lot of different attempts to define it.

As a field full of data nerds with a penchant for visualization, it's also unsurprising that a lot of them use Venn diagrams. (Fun fact: John Venn, who invented the eponymous diagrams, and his son filed a patent in 1909 for an lawn bowling machine.)

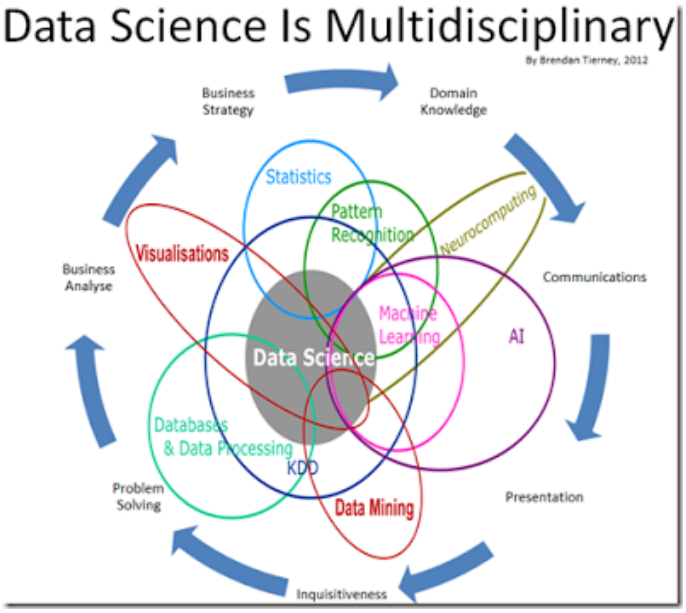
1. It all started with Drew Conway in 2010 (catching fire when he blogged it in 2013):



For Conway, the center of the diagram is *Data Science.* There's some controversy over what the bottom circle means (I'll address it farther down); all I can say, is if Conway meant something other than what I would call domain knowledge (e.g. physics), he chose the name *Substantive Expertise* very poorly. So assuming domain knowledge is at least part of what he meant, the idea is that a physicist, say, would have expertise in physics and math/stats knowledge, but lack hacking knowledge (I've met many physicists and I think that's less true than it used to be). *Machine Learning* experts tend to apply algorithms without an understanding of the domain they're analyzing (that sure as heck was my case when I first started building models in an industry that was totally new to me; I had to play a lot of catchup). And then people who can program and know their field but have no way to tell a statistically significant result from one arising from sheer coincidence are dangerous; they can arrive at some drastically wrong solutions and, for example, lose their companies lots of money.

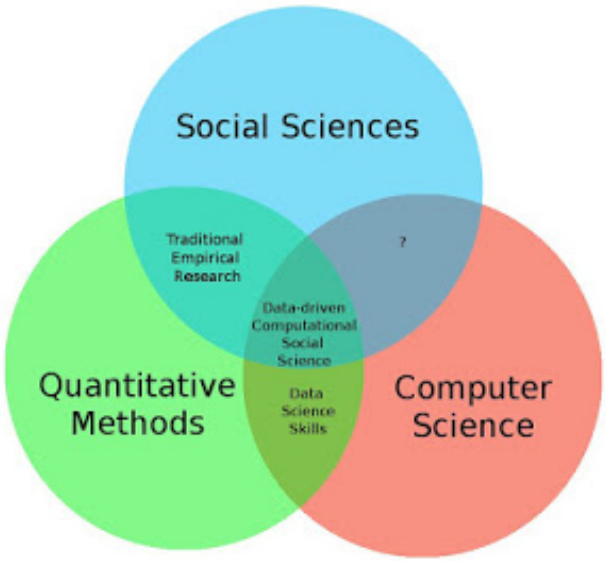
Note that **this isn't how a Venn diagram works.** *Hacking Skills,* for example, should apply to that **entire** circle, and the part that doesn't intersect with anything should be labeled, e.g. "hackers". But that's a fairly minor point, it's obvious what he's getting across.

2. After Conway's was made but before it was blogged, Brendan Tierney made a diagram in 2012 that's kinda Venn-ish.



It... sure is busy. *KDD* stands for *Knowledge Discovery and Data Mining,* by the way. Despite that, *Data Mining* also has its own circle. I do appreciate what he did here, though, implying what makes data science worthy of its own field is the breadth of its required skills. Apparently one of those skills is *Neurocomputing,* which seems a little... specific.

3. Quick on Conway's heels, Ulrich Matter blogged his riff on it later the same month in 2013:



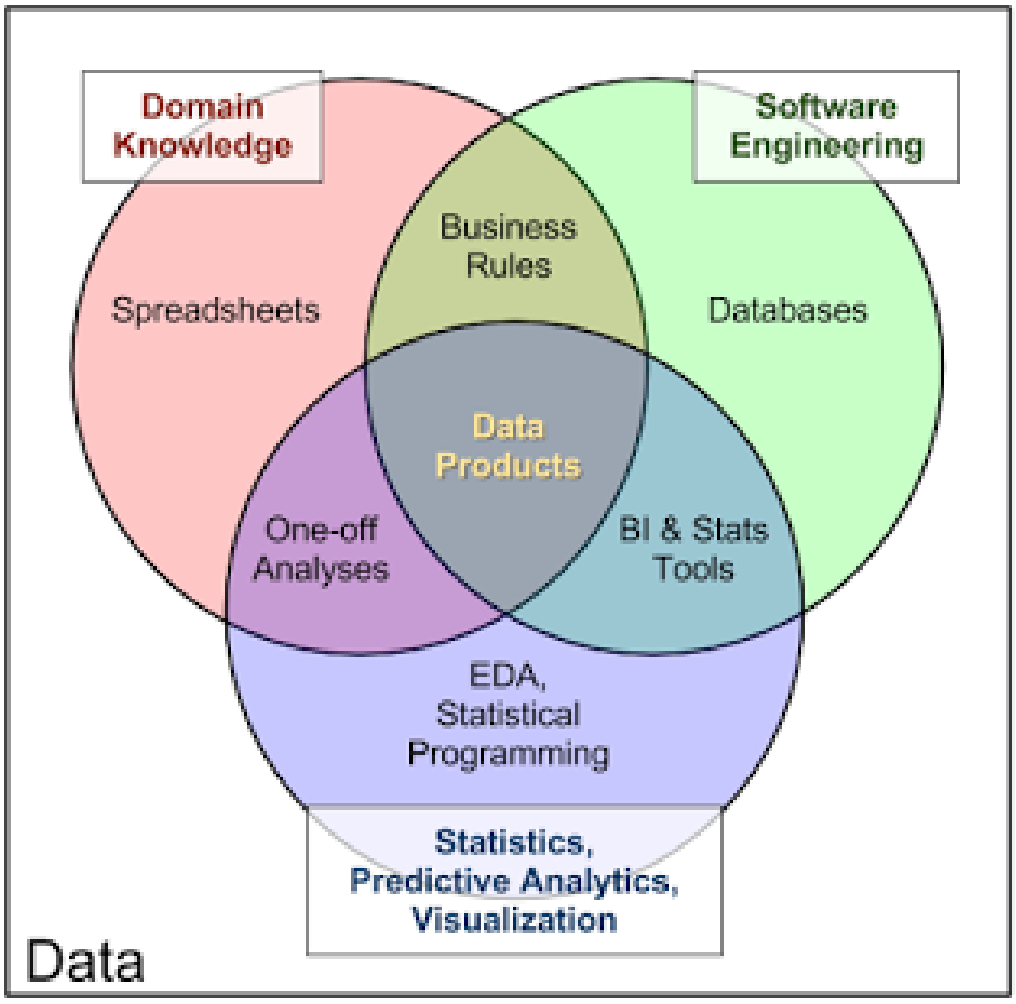
He's flipped it on the diagonal, specified the substantive expertise as *Social Sciences* (his field), changed hacking to computer science (you can see why someone would object to being characterized as a hacker, although I for one embrace it), and for some reason changed *Math & Stats* to *Quantitative Methods*. More importantly, he's moved *Data Science* where *Machine Learning* was in Conway's -- that's an interesting distinction, and one I've seen in the field. There are data scientists who specialize in one domain, and then there are generalists (who usually started out in one field but branched out, like me: I started in chemistry and now I'm in insurance). Also, he's apparently not comfortable

with *Danger Zone*, changing it to... a question mark. But apparently what matters to Matter (so to speak) is in the center of the diagram: *Data-driven Computational [Social] Science.*

A... bit wordy, shall we say? He also made sure to insert *Empirical* into *Traditional Research.*

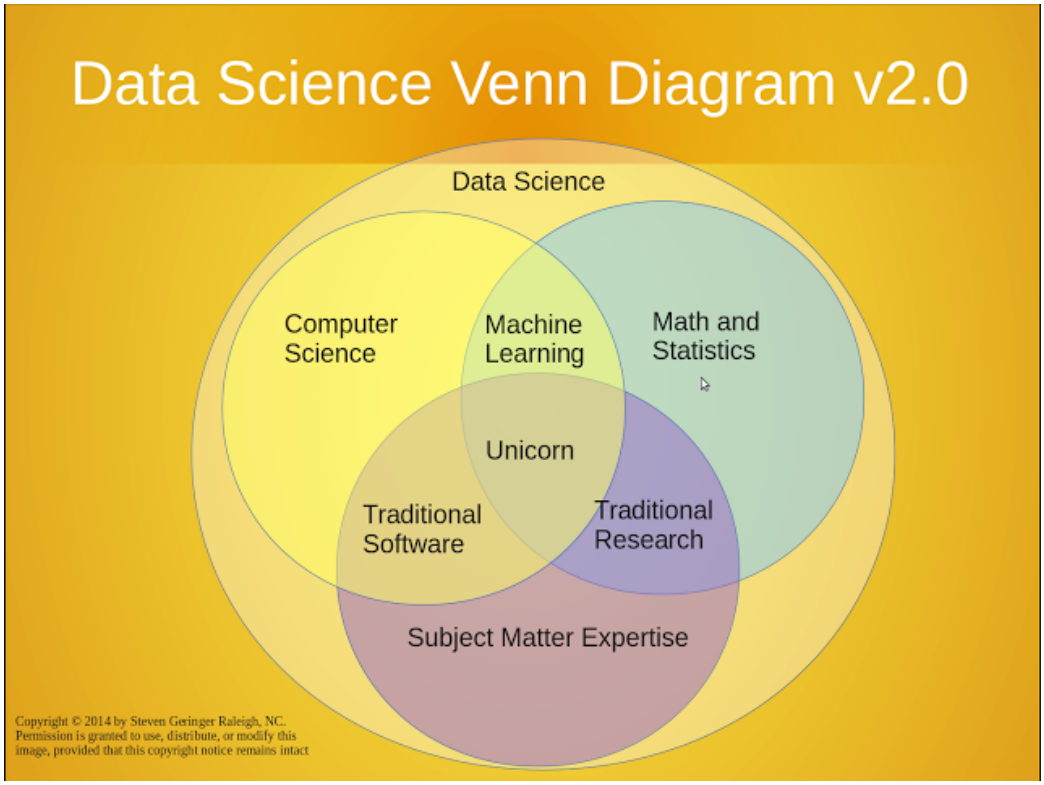
4. After the Edward Snowden news broke, Joel Grus supplied this tongue-in-cheek (*or is it?)* version. Now we're getting into more rarefied Venn territory, with four circles, the fourth being "evil".

5. In September 2013, Harlan Harris adapted this diagram to deal with data *products* instead of science.



The slices are no longer comparable to Conway because we've changed from science to products, but the categorizations are noteworthy (and they follow true Venn methodology, not being slices in themselves). *Domain Knowledge* remains, *Computer Science/Hacking* remains as *Software Engineering,* and crucially, Harris has added *Predictive Analytics* and *Visualization* to the*Statistics* circle. But not the actual tools they use, that's in the intersection with *Software Engineering.* Okay.

6. In January 2014, Steven Geringer provided a tweak that, instead of putting *Data Science* in the middle three-way intersection like Conway, calls *all* of it data science and calls the intersection *Unicorn* (i.e. a mythical beast with magical powers who's rumored to exist but is never actually seen in the wild.)



This is... a little weird, Venn-diagrammatically speaking. I think I know what he's getting at. When I first heard people referred to as data scientists, I often heard the riposte, "Aren't all scientists, by definition, data scientists?" True, there are no sciences that do not deal in data (insert psychiatry joke here), but still, data science, while quite nebulous, isn't just an umbrella term.

Plus, I'm sorry, but **you can see the screengrab of his mouse arrow in his diagram.**

Edit: An earlier version of this post omitted to give Geringer credit where credit is *definitely* due: **he was the first to remove the Danger Zone!** (Great, now that song is going to be in my head all day). Now people with subject matter expertise and computer skills can make *Traditional Software* without blowing the world up, or whatever. (My apologies to Mr. Geriner, and my thanks for his correction.)

Pages: **1** 2

Battle of the Data Science Venn Diagrams

Previous post Next post

Tags: Data Science, Drew Conway, Venn Diagram

First came Drew Conway's data science Venn diagram. Then came all the rest. Read this comparative overview of data science Venn diagrams for both the insight into the profession and the humor that comes along for free.

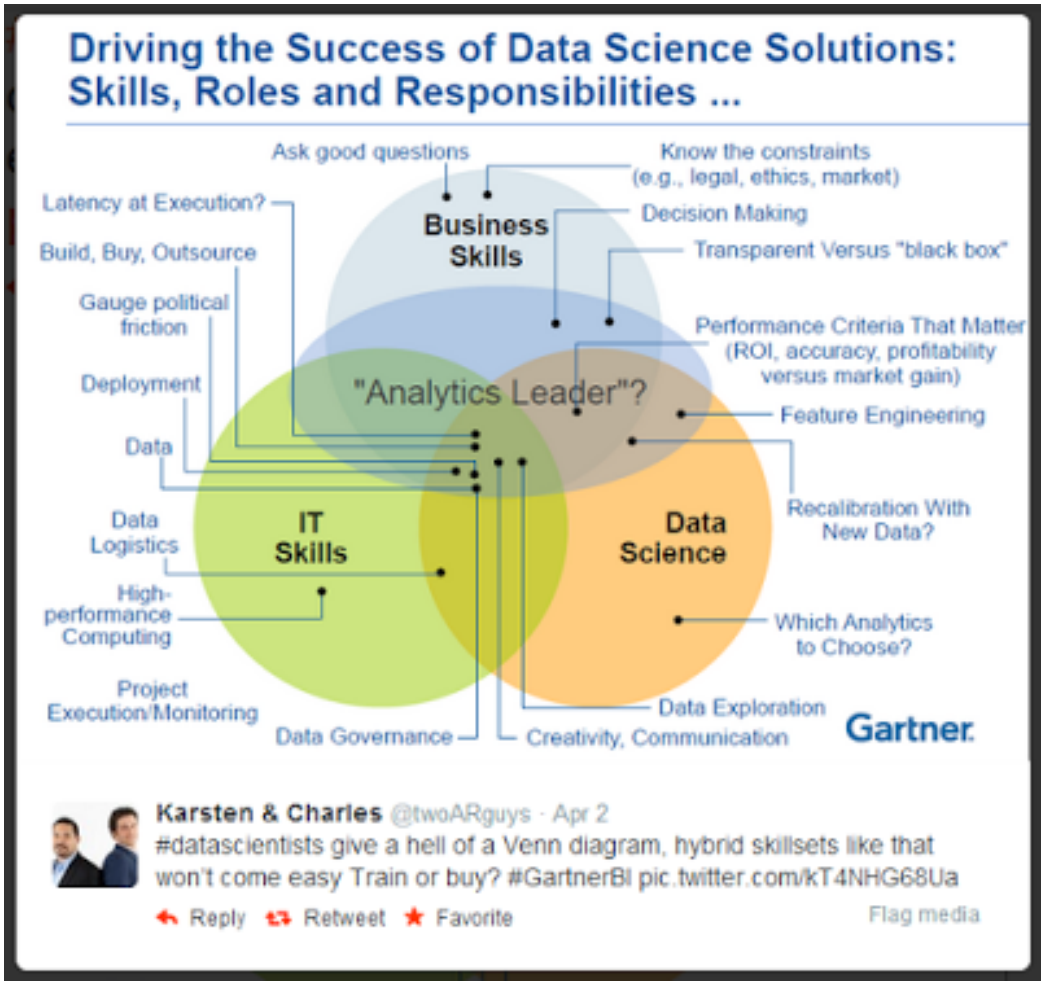
Pages: 1 **2**

---

7. In February 2014, Michael Malak added a fourth bubble, claiming Conway didn't mean domain knowledge when he said*Substantive Expertise*.

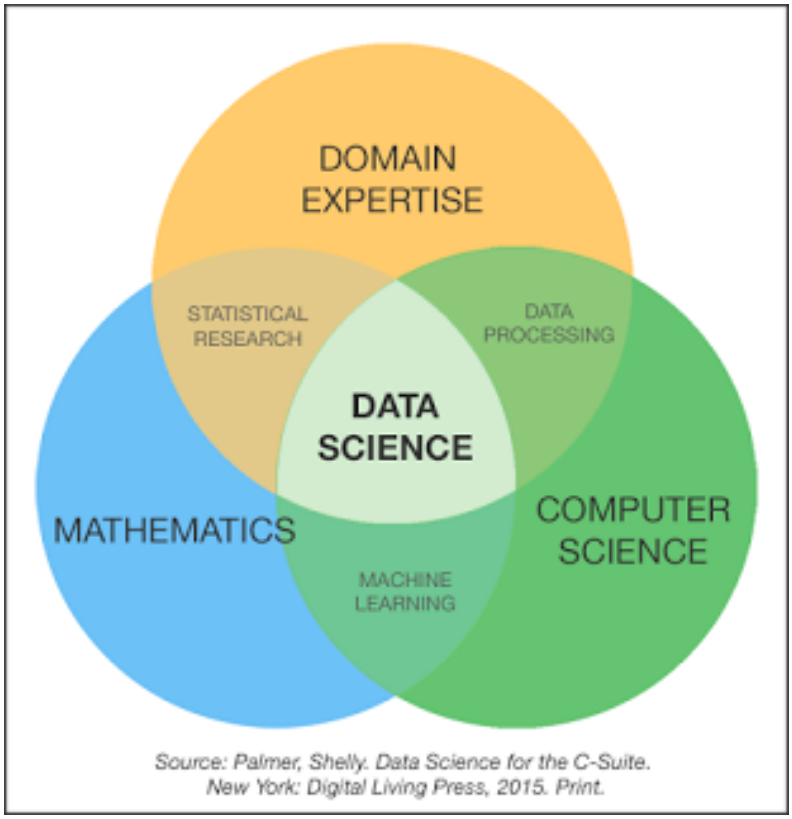
According to Malak, he's Inigo Montoya and we're all Vizzini when it comes to *Substantive Expertise*: "**You keep using that word. I do not think it means what you think it means**." Malak split it into *Domain Expertise*, and...er, knowledge of a domain, like *Social Sciences.* Maybe I'm dense, but I don't get the distinction. I'm also not sure what he's getting at with *Holistic Traditional Research* that, unlike *Traditional Research*, according to its placement doesn't include knowledge of the science you're researching? Am I reading that wrong? Holistic science is a thing, but it's not *that* thing. Anyways, *Data Science* is once again back in the unicorn position, and there are **three** danger zones (one of them **double danger**). Everyone be hatin' on the hackers.

8. My next example comes via Vincent Granville in April 2014, but he's reposting something by Gartner; I don't know the date of the original.



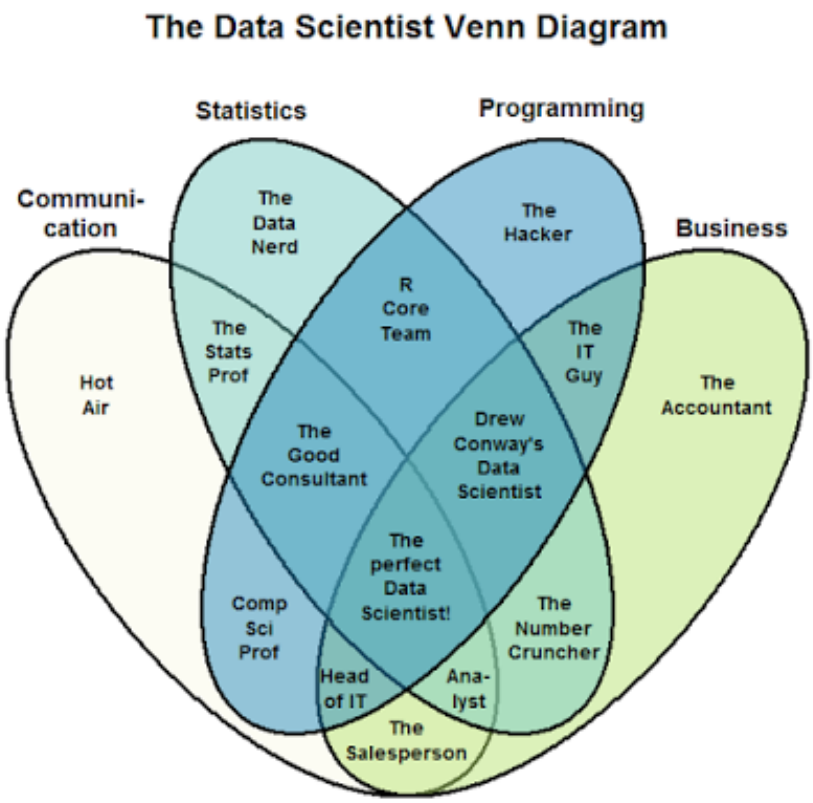
This is a Venn Diagram of *Data Science Solutions*, not data science itself; as such, *Data Science* is one of the circles, with other expertises (often not residing in the same person, but hopefully on the same team) being *IT Skills* and *Business Skills*. It kinda bothers me that the text labels are pointing to very specific positions in each slice, but the actual positions are arbitrary. That's business infographics for you.

9. Shelly Palmer guest-blogged for the Huffington Post in 2015, including this figure from a book he wrote:



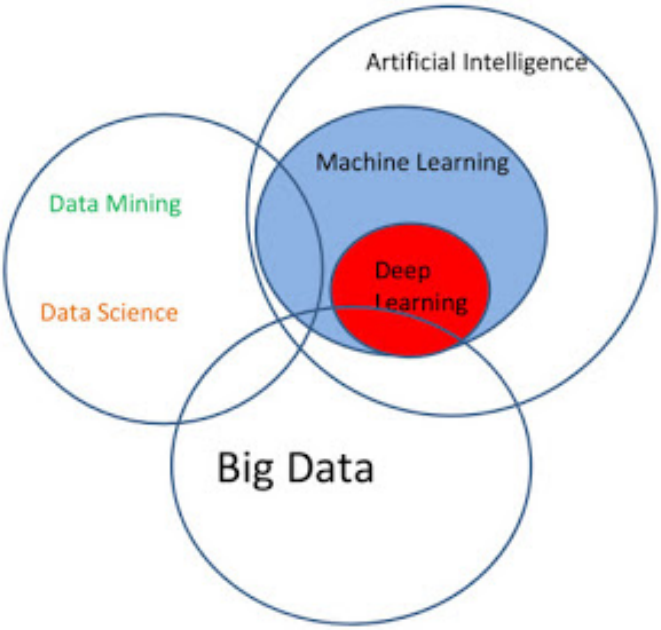
Pretty standard computer-math-domain triad straight from Conway, ~~but there's one revolutionary element:~~ **~~no danger zone~~**~~. Now computer-and-domain geeks without stats can do~~ *~~Data Processing~~* ~~without everything going all to hell. Seems reasonable.~~ *EDIT: Sorry Shelly, Geringer beat you to it, you're just not very noteworthy anymore.*

10. In November 2015, StackExchange Data Science user Stephan Kolassa came up with my personal favorite, adding*Communications* to Conway and changing his *Substantive Expertise* to *Business*:



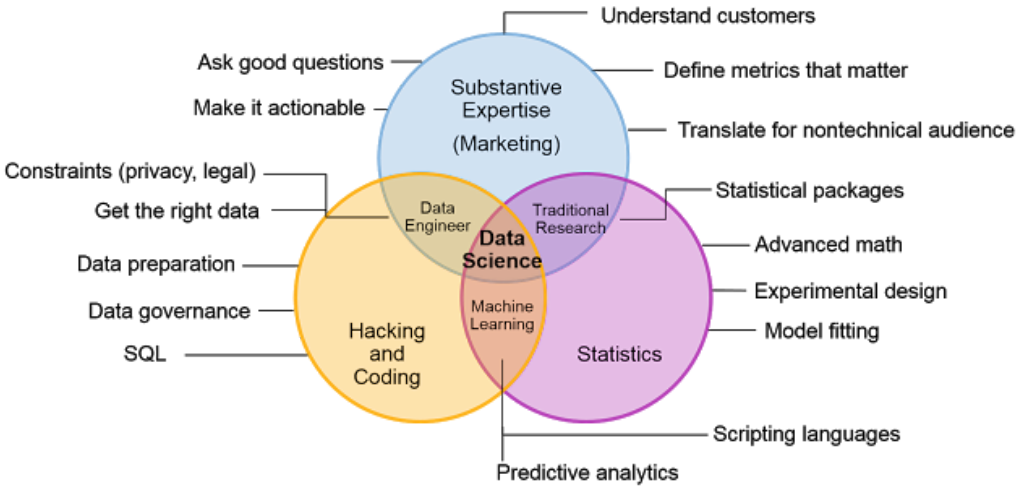
For all his effort, he was rewarded with only 21 (I'm one of them) upvotes in this beta-release forum. His categories are pretty good, too. I think I fall under *The Good Consultant*. Or possible *The Mediocre Consultant. The Consultant Who Tries Really Hard?* And yes, that's what a four-set Venn diagram looks like, not four circles like Malak's above, which does not contain all the combinations of intersections.

11. In 2016, Matthew Mayo blogged a diagram by Gregory Piatetsky-Shapiro:



Okay, this owes a debt to Tierney from four years prior, and although it purports to be a Venn diagram of data science, (a) it's not a Venn diagram, and (b) Data Science is *inside* one of the circles. It's good to see Big Data acknowledged, though. But...**Calibri**? Really? You went with the default font?

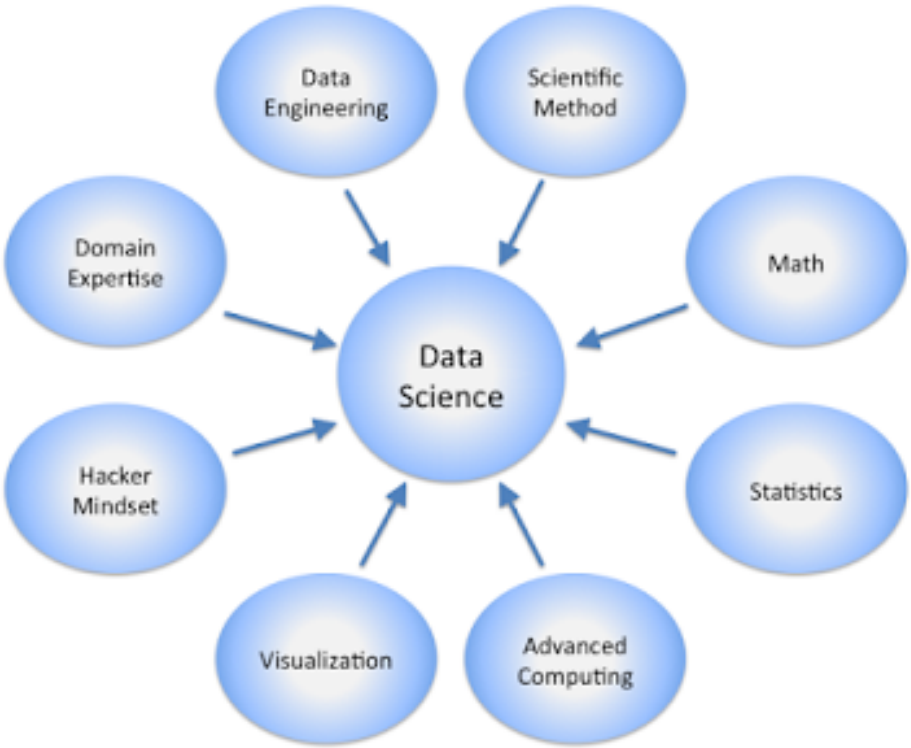
12. Finally (and I'm sure I don't have them all; If you know of any Venn diagrams I missed, please let me know!), later in 2016 Gartner redid their busy Data Solutions diagram, and made it prettier and confined to data science, as blogged by Christi Eubanks:



We've come full circle, back to Conway, except again *Danger Zone* is replaced, this time by *Data En gineer.* I like the callouts pointing to the edges better than their previous mess, as well.

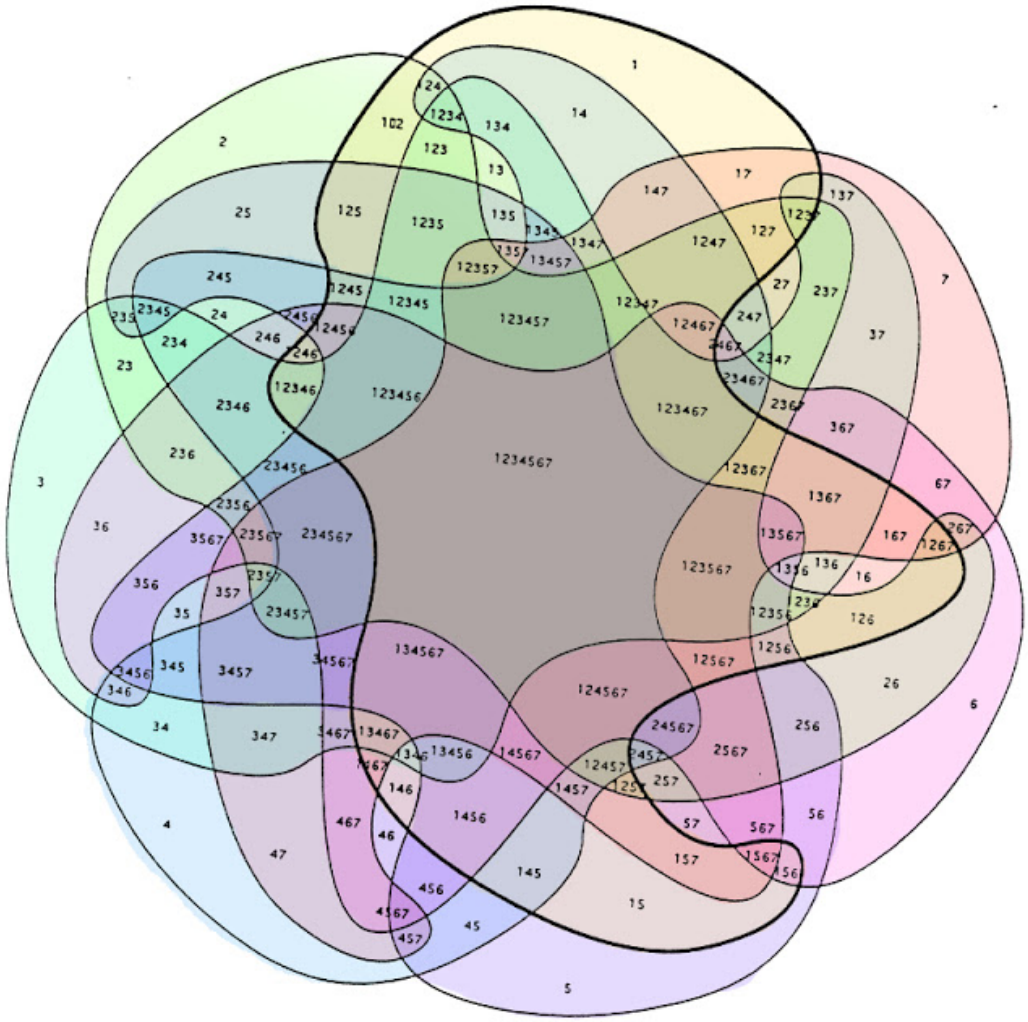
13. Data Science Venn diagrams of the future:

Wikipedia's page on data science has the following totally-not-a-Venn-diagram:



Really, in my opinion, this is the way to look at data science. Maybe not these exact skills, but it really is a synergy of different disciplines. Unfortunately, skill in one discipline can sometimes mask serious deficiencies in another and give data science a bad name. (I may or may not have contributed somewhat to this phenomena in my misspent youth, like, last year.)

Of course, then you'd need a really complicated Venn diagram. They do exist: here's one for seven sets:



Anyone want to give it a try?

Original. Reposted with permission.

Related:

- The (Not So) New Data Scientist Venn Diagram
- Does Data Scientist Mean What You Think It Means?
- The Data Science Puzzle, Explained