INTRODUCTION:

WHAT IS DATA SCIENCE?

THE HYPE

BIG DATA
DATA SCIENCE

THE MOST BASIC TERMINOLOGIES

- **WHAT IS BIG DATA ?**
- **WHAT DOES DATA SCIENCE?**
- **WHAT IS THE RELATIONSHIP BETWEEN BIG DATA & DATA SCIENCE?**
    - Is data science "The science of big data" ?
    - Is data science only the stuff going on in companies like Google, Facebook, and Tech companies ?

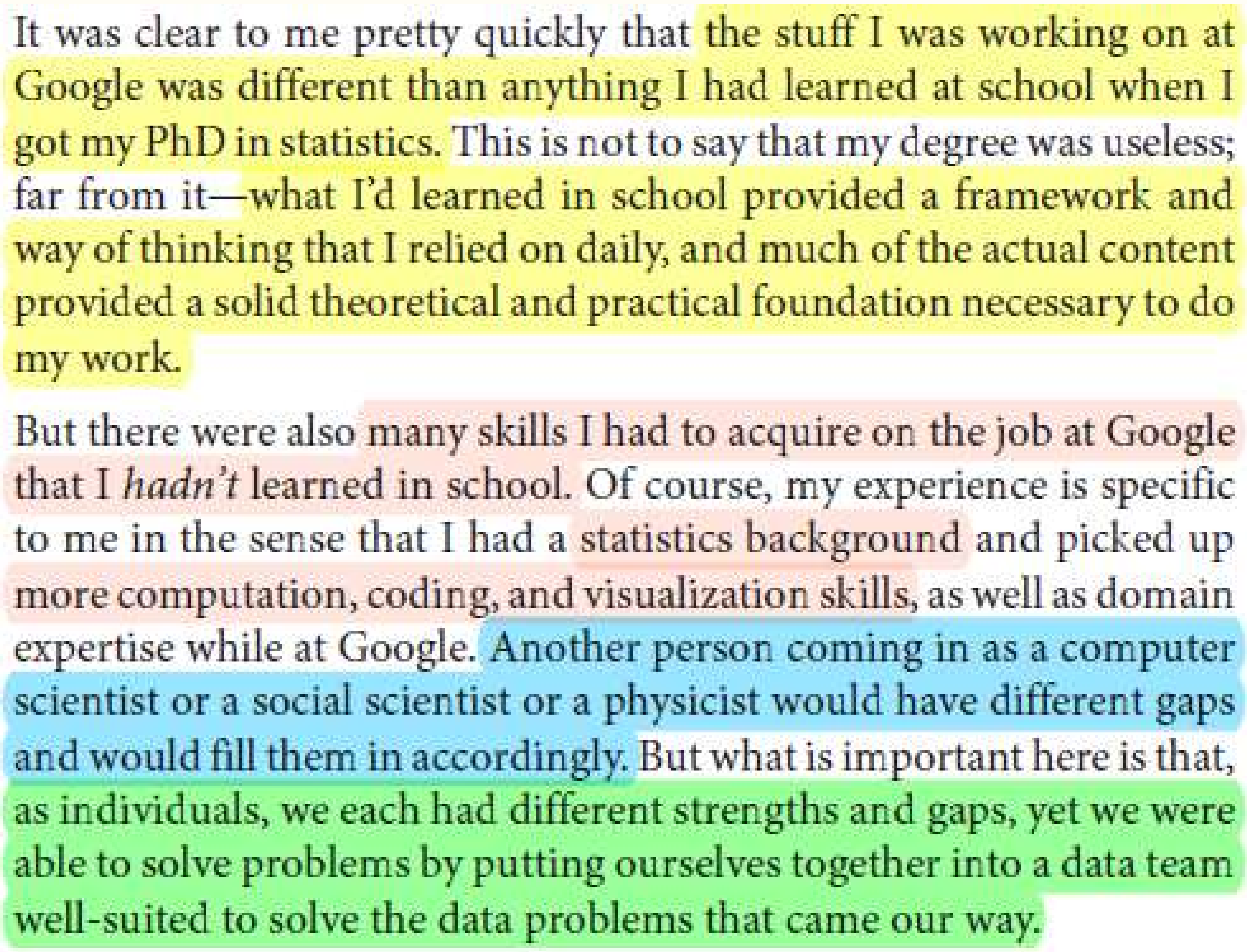
Statisticians already feel that they are studying and working on the "Science of Data."

DATA SCIENCE IS NOT JUST A REBRANDING OF STATISTICS OR MACHINE LEARNING BUT RATHER A FIELD UNTO ITSELF

...Statistics + Machine Learning + TECH INDUSTRY...

What it represents may not be science but more of:

A CRAFT



It was clear to me pretty quickly that the stuff I was working on at Google was different than anything I had learned at school when I got my PhD in statistics. This is not to say that my degree was useless; far from it—what I'd learned in school provided a framework and way of thinking that I relied on daily, and much of the actual content provided a solid theoretical and practical foundation necessary to do my work.

But there were also many skills I had to acquire on the job at Google that I *hadn't* learned in school. Of course, my experience is specific to me in the sense that I had a statistics background and picked up more computation, coding, and visualization skills, as well as domain expertise while at Google. Another person coming in as a computer scientist or a social scientist or a physicist would have different gaps and would fill them in accordingly. But what is important here is that, as individuals, we each had different strengths and gaps, yet we were able to solve problems by putting ourselves together into a data team well-suited to solve the data problems that came our way.

Whenever you go from school to a real job, you realize there’s a gap between what you learned in school and what you do on the job.

… The difference between academic statistics and industry statistics …

WHAT FLOWERS ARE WORTH

WHY NOW ?

Doesn’t mean that there’s not new and exciting stuff going on, but we think it’s important to show some basic respect for everything that came than before.

Massive amounts of data about many aspects of our lives, and an abundance of inexpensive computing power

Shopping, communicating, reading news, listening to music, searching for information, expressing our opinions —all this is being tracked online,

The data itself, often in real time, becomes the building blocks of data products

Amazon recommendation systems, friend recommendations on Facebook, film and music recommendations

DATAFICATION of our offline behavior has started as well the online data

There’s a lot to learn about our behavior: finance, medical industry, pharmaceuticals, bioinformatics, social welfare, government, education, and retail

We’re witnessing the beginning of a massive, culturally saturated feedback loop where our behavior changes the product and the product changes our behavior.

Technology makes this possible:

Infrastructure for large-scale data processing, increased memory and bandwidth, cultural acceptance of technology

A process of “taking all aspects of life and turning them into data.”

“Google’s augmented-reality glasses datafy the gaze. Twitter datafies stray thoughts. LinkedIn datafies professional networks.”

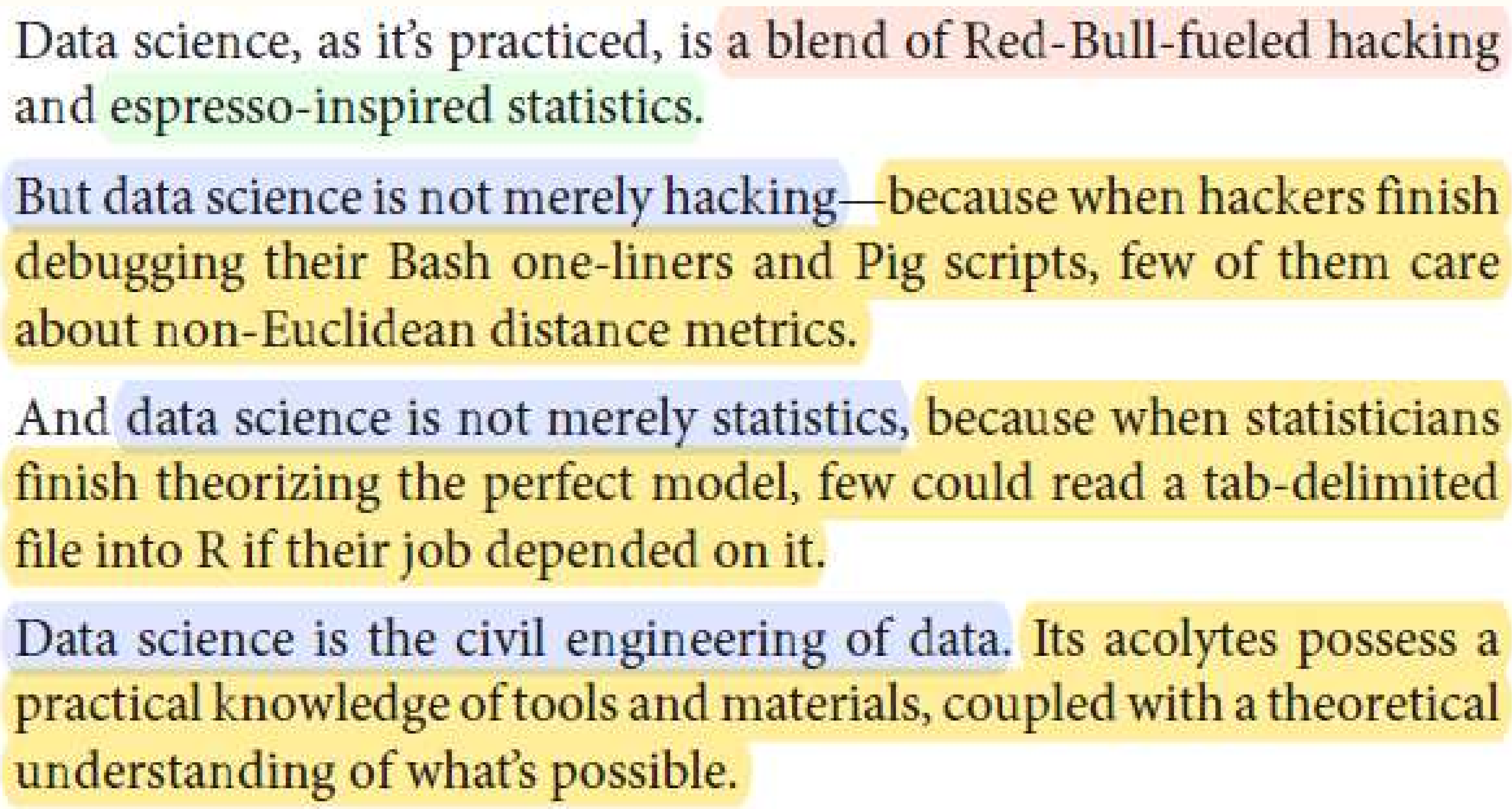
... "The Rise of Big Data" - Kenneth Neil Cukier and Viktor Mayer Schoenberger...

DATAFICATION

Once we datafy things, we can transform their purpose and turn the information into new forms of value.

... "The Rise of Big Data" - Kenneth Neil Cukier and Viktor Mayer Schoenberger...

DATAFICATION



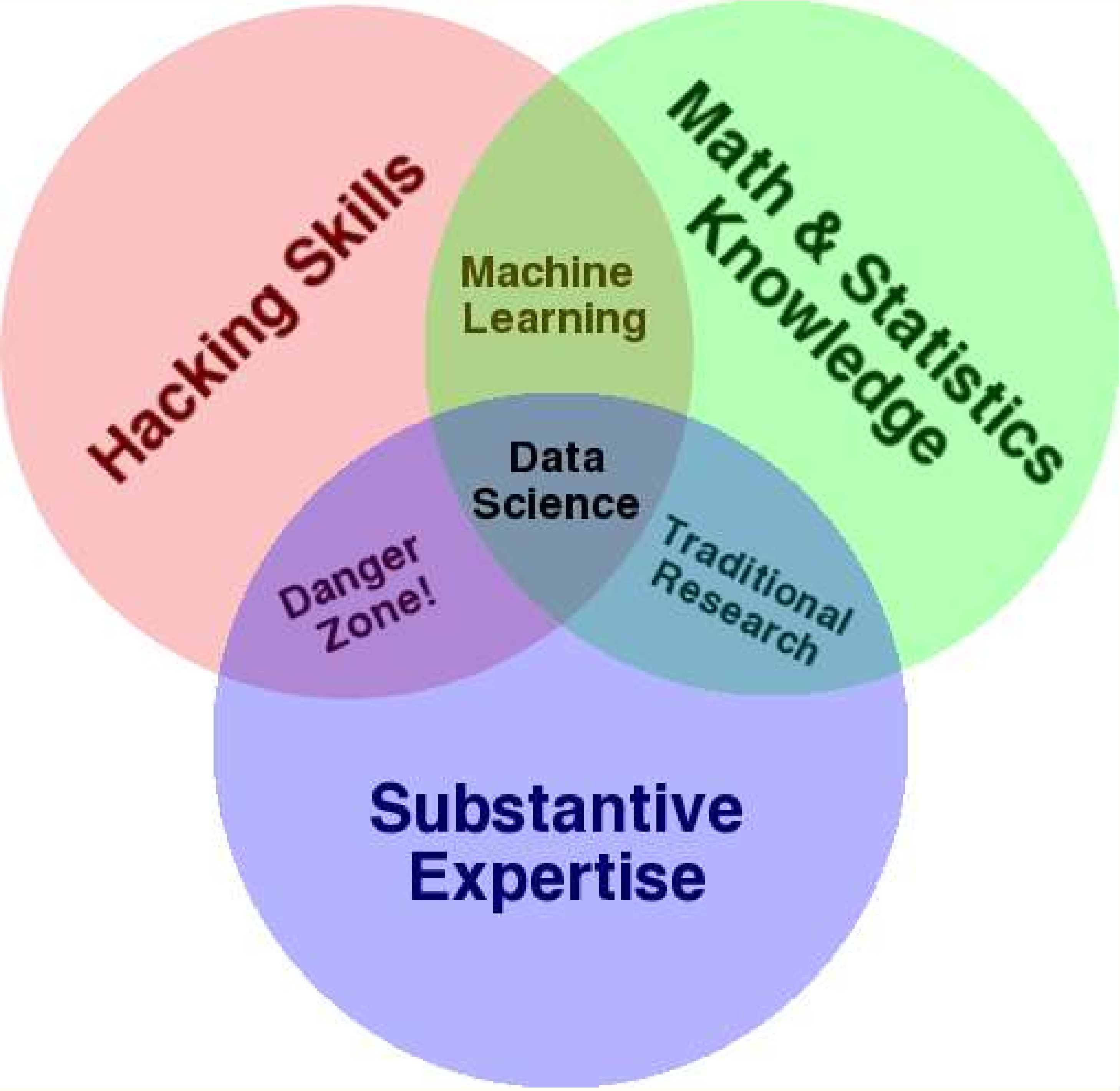
Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.

But data science is not merely hacking—because when hackers finish debugging their Bash one-liners and Pig scripts, few of them care about non-Euclidean distance metrics.

And data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it.

Data science is the civil engineering of data. Its acolytes possess a practical knowledge of tools and materials, coupled with a theoretical understanding of what's possible.

... Metamarket CEO Mike Driscoll ...



... Drew Conwell (2010) ...

DATA SCIENCE VENN DIAGRAM

... Nathan Yau (2009) : Rise of The Data Scientist ...

SKILLS OF DATA GEEKS :

- **STATISTICS**
  Traditional analysis you're used to thinking about
- **DATA MUNGING**
  Parsing, scraping, and formatting data
- **VISUALIZATION**
  Graph, Tools, etc.

Much of the development of the field is happening in industry, not academia. That is, there are people with the job title data scientist in companies, but no professors of data science in academia.

(Though this may be changing.)

DATA SCIENCE PROFILE

SCORE YOURSELF: (1-4)

- Computer Science
- Math
- Statistics
- Machine Learning
- Domain Expertise
- Communication and Presentation skills
- Data Visualization

Who plans to become a data scientist?

an ACADEMIC DATA SCIENTIST is a scientist, trained in ANYTHING from social science to biology, who works with large amounts of data, and must grapple with computational problems posed by the structure, size, messiness, and the complexity and nature of the data, while SOLVING A REAL WORLD PROBLEM.

What about in Industry?

A chief data scientist should be setting the data strategy of the company, which involves a variety of things:

- Setting everything up from the engineering and infrastructure for collecting data and logging,
- Privacy concerns,
- Deciding what data will be user-facing, how data is going to be used to make decisions, and how it's going to be built back into the product.
- Manage a team of engineers,scientists, analysts and should communicate with leadership across the company, including the CEO, CTO, and product leadership.
- Concerned with patenting innovative solutions and setting research goals.

Generally… Statistics, and Software Engineering skills.

Someone who knows how to extract meaning from and interpret data, which requires both tools and methods from statistics and machine learning, as well as being human

Spends a lot of time in the process of collecting, cleaning, and munging data, because data is never clean

Also... Computer & Social science comprehension.

Once she gets the data into shape, a crucial part is exploratory data analysis, which combines visualization and data sense. Find patterns, build models, and algorithms

Design experiments, and decide a critical part of datadriven decision making. Then communicate with team members in clear language and with data visualizations