An introduction to Data Science and Big Data

Fabrice Rossi

January 12, 2018

Contents

1 General Introduction

1.1 Big Data? Data Science?

1.1.1 Who am I?

See my web site, especially my research pages. The page of this course is here.

1.1.2 What are Big Data and Data Science for you?

Introductory interactive discussion about Big Data. The goal is to introduce some real life day to day examples, such as:

- spam filtering
- recommendation
- credit scoring
- vocal recognition (a.k.a. Siri and co)
- etc.

1.1.3 Objectives of the course

- to clarify the concepts (what is big in big data?)
- to give a general introduction to the data science process
- to study in more details the mathematical aspects of some of the data science tools

1.2 Informal definitions

Data collections of measured and recorded values

Data Science tools, techniques, systems, etc. that extract information from data (data mining, business intelligence, knowledge discovery in databases, etc.)

Big Data data sets that are too big to be processed by a single computer or by a limited number of computers (say less than 10)

1.2.1 From data to information and knowledge

This is maybe the most important aspect of the data science paradigm. Measuring and recording data does not immediately produce usable information or knowledge. For instance, recording a person talking produces data but without specific processing it's just a series of numerical values. Turning the sound into a text as represented by a text file on a computer produces information. This is arguably the same data but the representation has changed in a drastic way. (In fact it's obviously not the same data because of a probable massive loss in information content, e.g. the tone of the person speaking.) Acting based on the text is yet another transformation. Think about this sequence:

1. you say to your phone "Call my mother"
2. and then you talk to your mother.

What happened:

1. the phone recorded your voice
2. the sound was turned into a series of words known as words (i.e., "understood" as such by the computer)
3. the order "Call" was recognized
4. the rest of the order was interpreted as a search request in your contacts
5. the contact "my mother" was chosen among the contacts, probably trough a fuzzy form of search
6. the full sentence was translated into simple action for the phone (dial a specific series of digits)

1.3 The Big Data confusion

The terms "Big Data" should be used only to refer to very large data sets and related applications. It should not be used to describe data science *in general*. This confusion is damaging because solutions designed for super large data sets are generally not adapted to normal size data sets.

1.3.1 Doug Laney's definition

Analyst at META group (now Gartner). Proposed in 2001 the 3 V's:

1. Volume: data size
2. Velocity: streaming context
3. Variety: text, image, video, etc.

Sometimes complemented by Veracity (data quality, confidence in the results). The really important points are Volume and Velocity. Variety has been around for ages and Veracity is addressed in other contexts. Keeping the 3/4 V's is a manifestation (among others) of the confusion.

1.3.2 Consequences of the confusion

Very high volume data sets cannot be stored, queried and processed by standard solutions (e.g. on a single yet powerful computer). Thus specialized solutions have been developed. Main example: the Map Reduce paradigm by Google, Hadoop as open source solution.

However, they have a quite high entry cost, both in human terms and in processing power terms. In other words, those tools are *not* adapted to small to medium scale data sets.

Standard confusion story:

1. we want to do predictive modelling but we confuse that with big data
2. we read tutorials and books about big data finding discussions about predictive modelling, reinforcing our confusion
3. we use the dominant open source solution, hadoop
4. performances are very bad so we add more computers, reinforcing our belief that we do big data magic
5. wash, rinse, repeat

1.3.3 What has changed?

Are there really a before and an after? an emergence of Big Data?

This is indeed the case with the massive use of computers to handle every day life. Computer use produces *traces* (a.k.a. activity log). The most basic example is maybe the logs of web servers which contain:

- the ip of the client (your computer) -> this gives *many* facts about you such as your location and your internet provider;
- the request date and time;
- the requested page;
- a user agent (a description of your web browser);
- the referrer (if available, this is where you come from).

Numerous other examples can be given. They all revolve around the following facts:

1. data collection is easy via internet services (web sites and specialized protocols and apps)
2. there is an incentive to provide accurate personal information (e.g. on facebook when discussing with your family and friends, on whatsapp/telegram where you give your phone number, on amazon where your provide your coordinates, etc.)
3. most of the providers are US based and do not have to satisfy very stringent data privacy rules (contrarily to e.g. in Germany or in France)
4. computer resources are *super cheap*, for instance storage:
   - in 1956 the first HDD by IBM costed 50 k $ (equivalent to roughly 430 $ K of 2013) and stored 5 MB
   - in 1995 a typical 1 GB HDD costed 850 $ (equivalent to 1300 $ in 2013)
   - in 2005 a typical 200 GB HDD costed 200 $ (equivalent to 120 $ in 2013)
   - in 2010 a typical 1 TB HDD costed 80 $ (equivalent to 85 $ in 2013)
   - in 2013 a typical 2 TB HDD costed 110 $
   - in 2018 a typical 4 TB HDD costs 110 $
   - the cost has dropped from 10000 $ per MB to 2.75 cents per GB!
   - even SSD are affordable (around 260 $ for 1 TB), ditto for super compact storage like micro sd card (around 150 $ for 256 GB)
5. computer resources are *available* especially via cloud computing systems (and their ancestors like grid computing systems)

Notice that while internet is major provider of personal data, big data appear in other contexts, such as:

- pre-internet services (e.g. credit/debit card, air company frequent flyer programs, etc.)
- physics: the Large Hadron Collider (LHC) and its ancestors
- biology:
  - in 2001, the estimated cost of sequencing the full human genome was estimated to have been around 2 billions of $ for 15 years of work
  - in 2009, it costed 100 k $ for 1 year of work
  - in 2014, it cost 1000 $ for 24 hours of work
- smart grids: ERDF linky
- the open data movement

1.4 Back to the objectives and take home message

Big Data should be used to refer to very large data sets and applications dealing with them. Technological and sociological changes have allowed the collection of enormous data sets, but the general idea of producing knowledge from data is anterior to the big data trend.

Unfortunately, "Big Data" is very frequently use to describe data science. This is probably the first pitfall of data science: one must learn to detect an improper use of "big data" and to adapt in consequence. Notice that "data science" is a fancy new name, but other names existed such as "data mining", "business intelligence", KDD, etc.

This course will not discuss much more the way to go "big" and will focus on the data science part. I will mention the limits of the presented methods in terms of data size but this will only provide rough guidelines. This course is mostly about some mathematical aspects of data science, not about its computational aspects. However to be a good data scientist, you need to have a good knowledge of both aspects.

2 The Data Science value chain

The value chain concept is useful to see the activation of an organization as a process which receives inputs and produces outputs in a series of steps that each increases the "value" of the objects they transform. While the concept is very business oriented is also gives some insights on the way data science operates. See [HWCL14].

In the context of data science, value is related to the information content of a data set and to the how "operable" the data are.

We will follow the value chain of some applications during the presentation:

1. spam filtering
2. product recommendation
3. fitness tracker

2.1 Data generation and acquisition

This can be seen as the input phase of the process, when the raw materials are obtained by the data scientist. Data collection is quite easy in the "big data" world:

1. data are produced essentially in digital format, nothing is analog anymore (music, image, video, text)
2. data collection is either automated or crowdsourced
3. mobile connected devices enable generation and collection from everywhere
4. part of the mobile evolution simplifies even further point 1, e.g. voice recognition, music indexing, etc.

The ongoing evolution will lead to even more data via the "internet of things", i.e. connected devices ranging from activity trackers to "smart" thermostats.

Data acquisition is generally part of the generation process, for instance a connection to a web site generates a log entry in the web server (ditto for e.g. e-commerce). For some applications, data acquisition can prove complex because of the needed transmission rates from distributed sensors (internet of things) and servers.

Both generation and acquisition are simplified by local processing (which could be included in the second step of the chain, the preprocessing one). For instance, activity trackers detect automatically the type of activity (sitting, standing, etc.) and can therefore transmit only time intervals. Music recognition (a.k.a. Shazam) is based on the local calculation of a signature. **Notice that this is already data science in the sense of information extraction.**

A very classical advice is to collect as much data as possible ("cast a wide net for data"). This is somewhat stupid: the more data you collect the more likely you are to find spurious correlations (see Figure 1). Assessing the significance of correlations is also difficult, especially in a big data context (see Figures 2 and 3), thus some care must be exercised.

Our examples:

spam filtering emails are collected as part of the email service. In addition, storing the emails in folder or tagging them provides metadata on those emails (categories). Many other things can be collected by a webmail (connection time, email reading time, etc.)

product recommendation carts are stored as part of the buying process (leading to co buying recommendations, for instance). Browsing behavior can be stored also, as well as comments, reviews, etc.

fitness tracker personal data (height, weight, sex) are obtained as part of the configuration process. Activity data include: hearth beat frequency, movements (via accelerometers), position (GPS), etc. Some local processing is done. Data are sent to the service provider via synchronization apps.

Notice that collecting data is already increasing their value as uncollected data simply vanish, especially when they have a high volume.

2.2 Data preprocessing

Realistic real world applications try to leverage multiple data sources, such as web based data, mobile application based data, or location data. Those data sources can be inconsistent and/or use incompatible native format. In addition, errors can occur in the collection process, sometimes in a deliberate way.

More generally, there is a whole range of data transformation methods that are applied prior to storing the data. Those preprocessing methods aim at simplifying the rest of the data science process without losing too much information. Examples include:

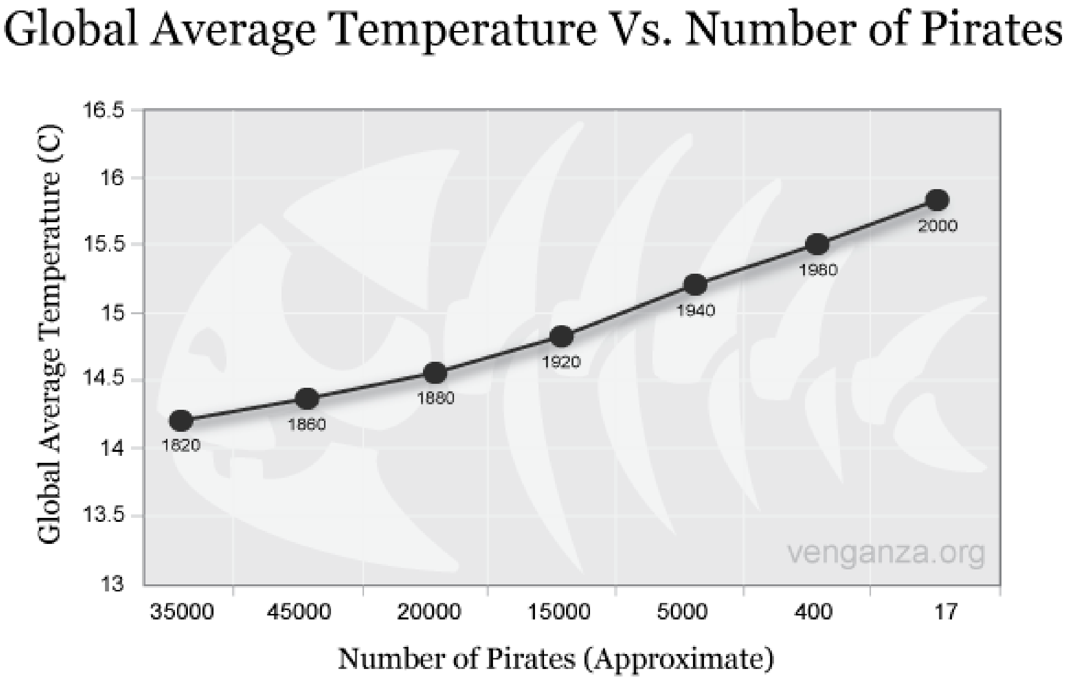

Figure 1: Global Warming is caused by the decrease in the number of pirates (or vice versa).

compression this a key step for media applications (audio, image and video) as the expected apparent quality in this context is very high. This is incompatible with the bandwidth and the storage generally available on the devices that produce those multimedia data. Thus those data are compressed prior their local storage.

data cleaning sensors are noisy and produce inconsistent data (e.g. GPS positions). Many data acquisition systems include some consistency based cleaning systems (bogus GPS positions are removed from trajectories, for instance). Users tend also to include bogus values (like fake postal codes). Some of those values can be detected at with low level checks (again, fake postal codes) and can thus be either deleted or marked as bogus.

data integration this is a database oriented concept which covers the general idea of using some common "format" for the data (for instance converting all time data to UTC).

It should be noted that the concept of preprocessing is a blurry one. Some preprocessing are quite obvious (conversion to UTC) others are very complex (activity extraction from an activity tracker). Data science contributes to the design of advanced preprocessing techniques and thus the latter stage of the data science value chain can be pushed back to the initial part at some point.

Our examples:

spam filtering data integration consists in linking actions in the client (web mail) with the emails in a way that will allow sophisticated queries (in stage 4 of the pipeline) combining criterion such as the content of the email and the time spent reading it.

product recommendation there is a strong need for detecting "equivalent" products in this application fields (softcover and hardcover editions of the same novel, for instance).

fitness tracker as pointed out above, preprocessing is crucial in this domain for activity reporting, position tracking, etc.

2.3 Data storage

In fact, collecting data raises their value only if they are stored, as least "for some time". Because of the trade off between the distributed collection and the bandwidth that would be needed to transmit all data to a central server, some part of the data is frequently lost.

Apart from that, basic storage is not very costly and is considered more or less as a solved problem in data science, even for very large data sets. However, as pointed out above, do not mistake "big data" as in data science and "big data" as in really big data. In the latter case, some complex distributed file system should be used, for instance HDFS.

For our examples, there is nothing particular to discuss here.

2.4 Data management and querying

The low level storage part is mostly a solved problem, but putting data on a hard drive is not the best way to make sense of them. A crucial part of data science is to store the data in an "operable format". The standard way of doing that is to use a relational database management system (RDBMS). The so-called NoSQL movement provides alternative solutions.

On the one hand, RDBMS seem to be the best solution, notably via ACID transactions: they should provide the properties needed in order to ensure that parallel data production, access, modification, etc. do not break the whole system. On the other hand, the CAP "theorem" says that they can't provide what's needed, thus alternative should be explored.

This area is plagued by confusions around the NoSQL term and around the CAP "theorem". It is mostly a computer science and computer engineering set of problems. While a data scientist *must* know some SQL and, more generally, the principles of the relational model, those aspects will be studied in the present course.

2.4.1 CAP

The CAP "theorem" is in fact a general observation/principle made Eric Brewer in 1998-2000 about distributed systems. It says, in essence, that because traditional RDBMS operate under the ACID principle, they cannot be partition tolerant. One of the basis of the NoSQL movement is then to move away from the RDBMS model in a tentative to be more tolerant to the partition issue. As pointed out in [Bre12], the actual issue is whether *during a partition event* one should favor consistency or availability, and then how to recover a fully consistent system when the partition disappears. All of this applies only to very large data sets or inherently distributed schemes (e.g. ATM).

2.4.2 NoSQL

The common point of most of the NoSQL DBMS is that they do not use the relational model but other forms of organizations (which are not adapted to the SQL language, hence the NoSQL acronym).

This does not mean that all NoSQL DBMS are distributed. In particular, the key-value database model, which is a typical non relational database structure, can be implemented as a local system or in a distributed fashion.

Nevertheless, there are numerous large scale distributed data store that can be used in big data contexts, both to store the data and to query them, albeit in a limited way compared to SQL possibilities. There are also data store dedicated to some types of data, for instance for documents (XML, JSON, etc.) and for graphs.

2.4.3 Take home message

The database aspect of big data is a very active research area. The movement has been originally to move away from traditional relational DBMS to the NoSQL distributed DBMS. However, many of them offer little or no consistency guarantee, leading to e.g. data loss. Thus, many real world solutions are based on hybrid systems that combine relational DBMS for crucial data with NoSQL distributed DBMS for other data. The current trend in research seems to be the NewSQL class, exemplified by Google's Spanner. It's a move back to almost relational systems with SQL interface and strong consistency guarantees. At a smaller scale, distributed versions of standard RDBMS seem to work quite well (e.g. MySQL Cluster).

For our examples, we can distinguish two aspects:

1. classical data (e.g., personal data, carts, etc.) are generally stored in (distributed) relational DBMS;

2. less "structured" data are generally stored in dedicated format. For instance emails are frequently stored via the Maildir format. This allows software to offer specialized data access/query operations.

Integration of the two parts are extremely important to achieve a high "value": if the usage data collected on the webmail are not linked to the emails, potential correlations will be missed.

2.5 Data mining

The last step of the data science value chain is the "data mining" phase also called "data analysis" or "analytics". As explained before, the main goal is to extract information and knowledge from the data. The application range is enormous and contains, among many others:

- system monitoring (dashboard): this is very important for large web sites, for instance, and more generally in order to understand how the users are actually using the service (see the retweet interface added by Tweeter, for instance)
- user modeling for:
  - recommendations: amazon, netflix, etc.
  - targeted advertisement: google, facebook, etc.

In addition to those obvious applications, it should be noted that the data are also used to build models that are in turn used to offer services to users. For instance (recommendation is in between):

- content searching (recommendation outside of the commercial context)
- machine translation
- face detection/recognition (in picture)
- automatic summarizing
- etc.

The main goal of this course if to explore the mathematical aspects of this data mining/machine learning part.

Our examples:

spam filtering the obvious goal is here to automatically classify emails into categories. The classifier program is designed automatically by another program from the collected data. This is an example of **supervised learning**.

product recommendation another typical example of supervised learning were the goal is to **rank** objects according to the tastes of a user. Frequent item set discovery is also a typical application in this situation. It belongs more to **unsupervised learning**.

fitness tracker the goals are far less clear here. Beyond personalized advises (which could be a typical case of traditional artificial intelligence), the providers of the tracking service could be interested in "understanding" their user base in order to offer targeted advertisement for instance. This is a typical case of **unsupervised learning**.

2.6 Take home message

A data scientist should be able to handle the full data science value chain and should be a specialist of the final stages of this chain. This implies to master some aspects of computer science, of computer engineering and of mathematics (probability, statistics and optimization, mainly).

In general, data acquisition, integration, and storage is handled by a dedicated computer engineer team. Interaction with the team is needed to increase the coverage of the data collection, to improve data integration and cleaning, etc. The main job of the data scientist starts after that and consists in understanding the data via interactive queries and domain knowledge acquisition. This allows one to build dashboards and other visual tools that help monitoring the state of a system. From this, predictive modelling can be done.

References

[Bre12] Eric Brewer. Cap twelve years later: How the "rules" have changed. *Computer*, 45(2):23–29, 2012.

[HWCL14] Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li. Toward scalable systems for big data analytics: A technology tutorial. *Access, IEEE*, 2:652–687, 2014.

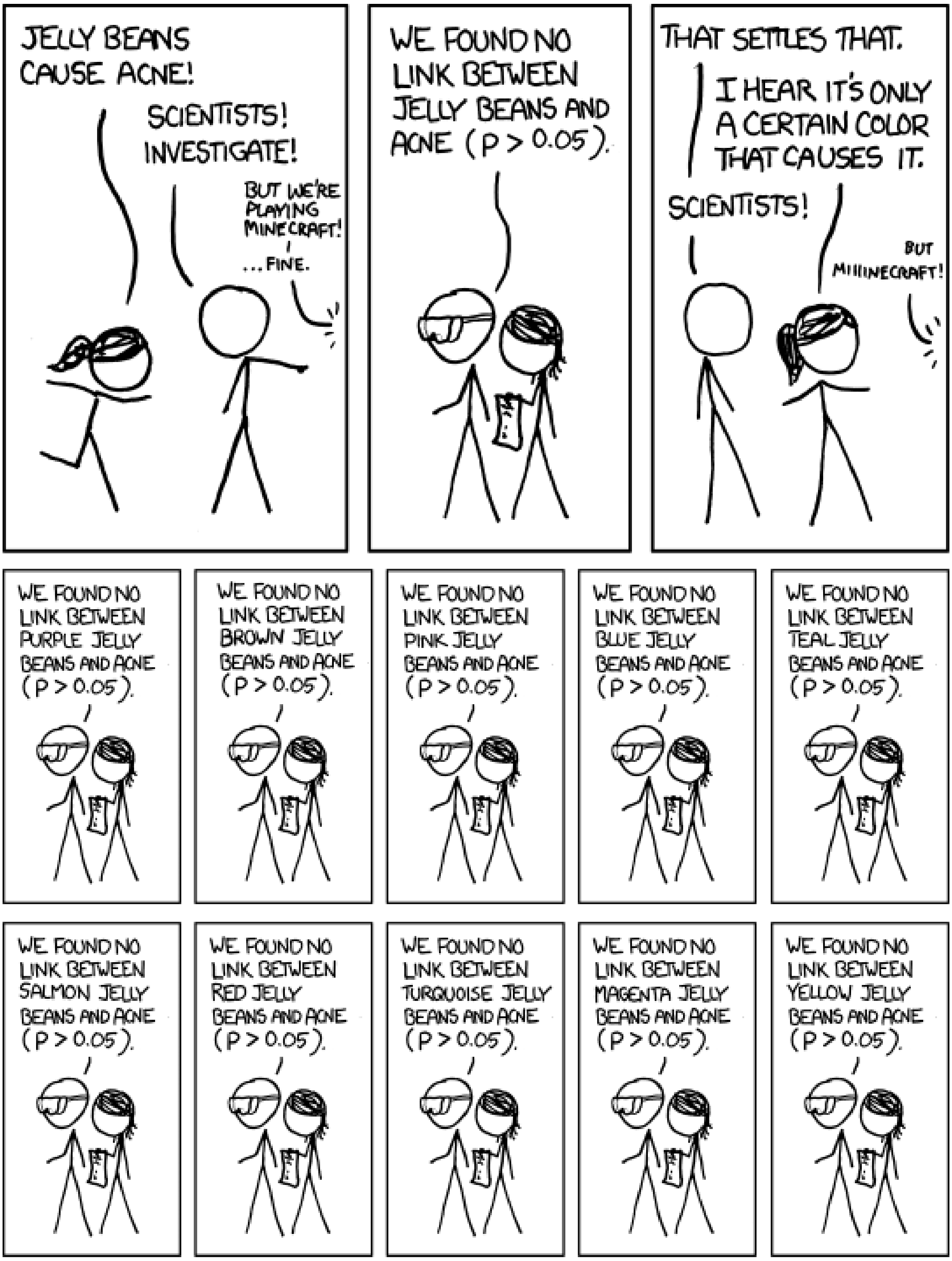
JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

Figure 2: Randall Munroe's take on tests (part 1)

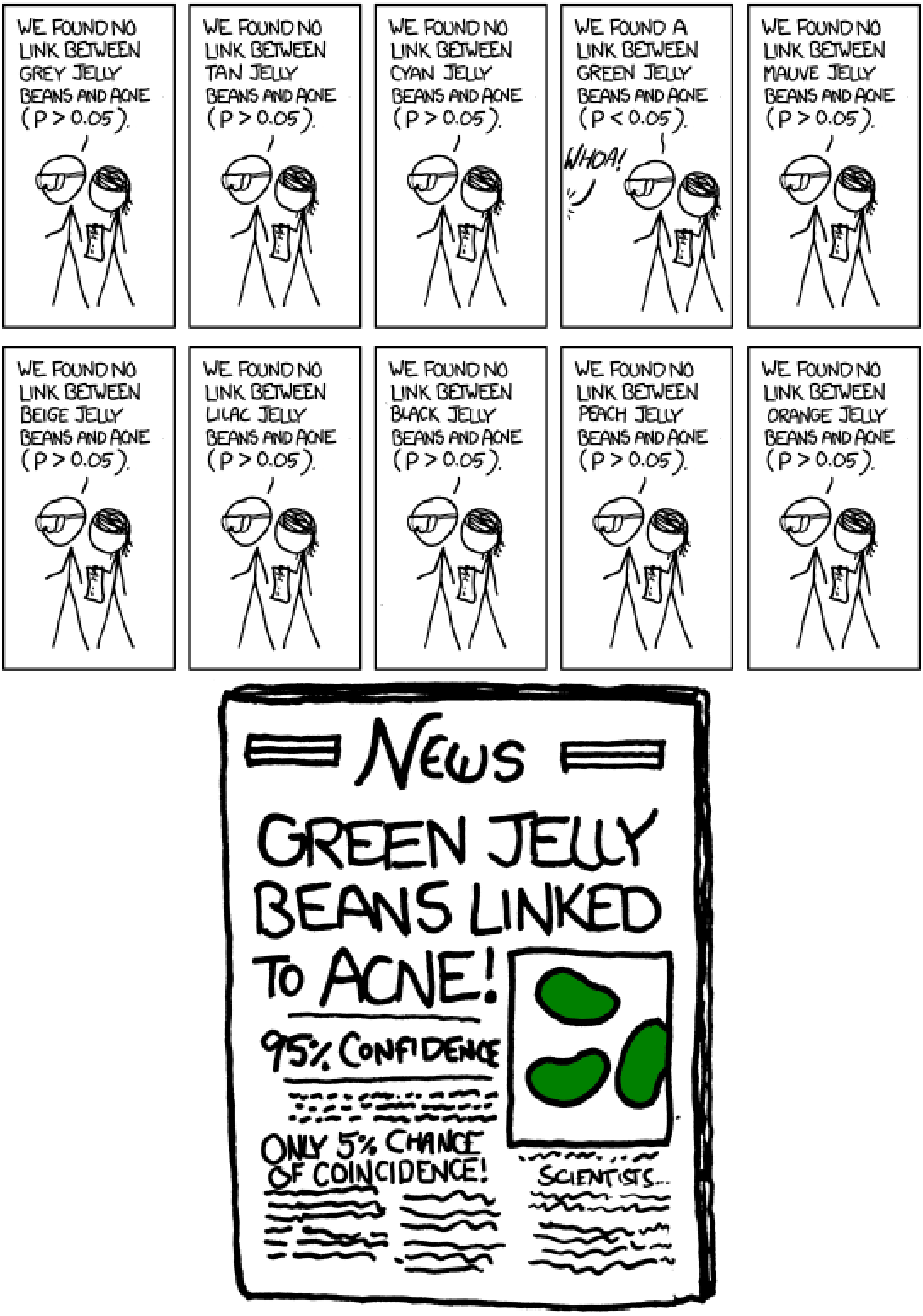
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Figure 3: Randall Munroe's take on tests (part 2)