We know now that there is a multitude of sources where data can be found. Each source has its own method of importing its data into Python or other programming languages, though the steps for data cleaning remain the same no matter the data source (i.e., a data frame imported from one data source can have the same steps applied to it as a frame imported from another source). Data can be found in three main locations: devices, the Internet, and servers. With all these data sources, how do we know which data we can trust? There are a few ways we can verify the authenticity of data from each source.

Devices: Any electronic device by default will generate data, such as computers, sensors, cameras, phones, televisions. Their data is usually stored in a local file format, but sometimes devices can upload data directly to the internet, such as with remote sensors. To verify local file authenticity, most computers will read file metadata such as the publisher, modification date, data type contained within, and sometimes hash functions and checksums (blocks of data derived from other data in the file). For example, the Luhn algorithm generates checksums from credit card numbers to verify if saved numbers are in fact valid. Just as well, the common “publisher not verified” error when downloading a third-party program arises when the computer cannot ascertain the program files’ publisher. Once a file’s credentials are verified, it is important to secure the device to keep the data authentic.

The Internet: The internet is like one giant editable book. Its data is stored in websites, or webpages, users can access the pages via a URL or ‘page number,’ and users can contribute to its data collective by “writing more” on the pages. This can be via manually uploading data to each page or having electronic devices automatically upload data to a page (such is the case with cloud-enabled sensors). However, with all the good webpages in the “book”, there are some websites that are not so good and try to get users to write personal data on their page. In addition, authentic data can get lost amidst all the inauthentic data. Just today, I was trying to find whether orange peel is acidic or basic, and several websites said different things!

This is where data authenticity verification comes into play. Some ways a web browser can verify the authenticity include:

1. Checking the connection type: HTTP or Hypertext Transfer Protocol is the standard connection type, but HTTPS (S for secure) ensures data is sent over a secure channel.
2. Verifying the website’s SSL certificate: This will ensure that the website’s data is secure from tampering. These certificates contain information such as the official website name, publisher, and e-signature to differentiate the data of that website from malicious sites, in essence a “website passport”
3. Using a third-party verification tool: Google hosts a URL checking tool that can check if a website’s data can be trusted. In addition and many APIs exist to “ping” websites for their security credentials.

Servers: While servers and the internet do go hand in hand, I’ve separated this category as sometimes information can be fetched from a server via an ad hoc connection (e.g. Bluetooth, IR). Servers, just like the internet, can store information, and devices can send requests to them to fetch information using APIs, just like websites. Servers contain their own credentials that can be fetched to determine integrity.

Authenticity verification, I would say, is much like how white blood cells filter germs. The body’s cells present molecules on their surface (their security credentials) to tell the white blood cells (the web browsers) that they are valid cells. Those cells that do not will be destroyed (website access blocked in the case of many browsers).

References:

User (2017). *Checking a file for integrity and authentication.* Stack Exchange. Retrieved on February 8, 2021 from https://crypto.stackexchange.com/questions/51265/checking-a-file-for-integrity-and-authentication.

Vocell, J. (2020). *A Beginner’s Guide to SSL.* HubSpot. Retrieved on February 8, 2021 from https://blog.hubspot.com/marketing/what-is-ssl.

Zheng, J. (2019). *10 Fool-Proof Ways to Verify Website Authenticity.* WhoisXMLAPI. Retrieved on February 8, 2021 from https://website-categorization.whoisxmlapi.com/blog/how-to-verify-website-authenticity.