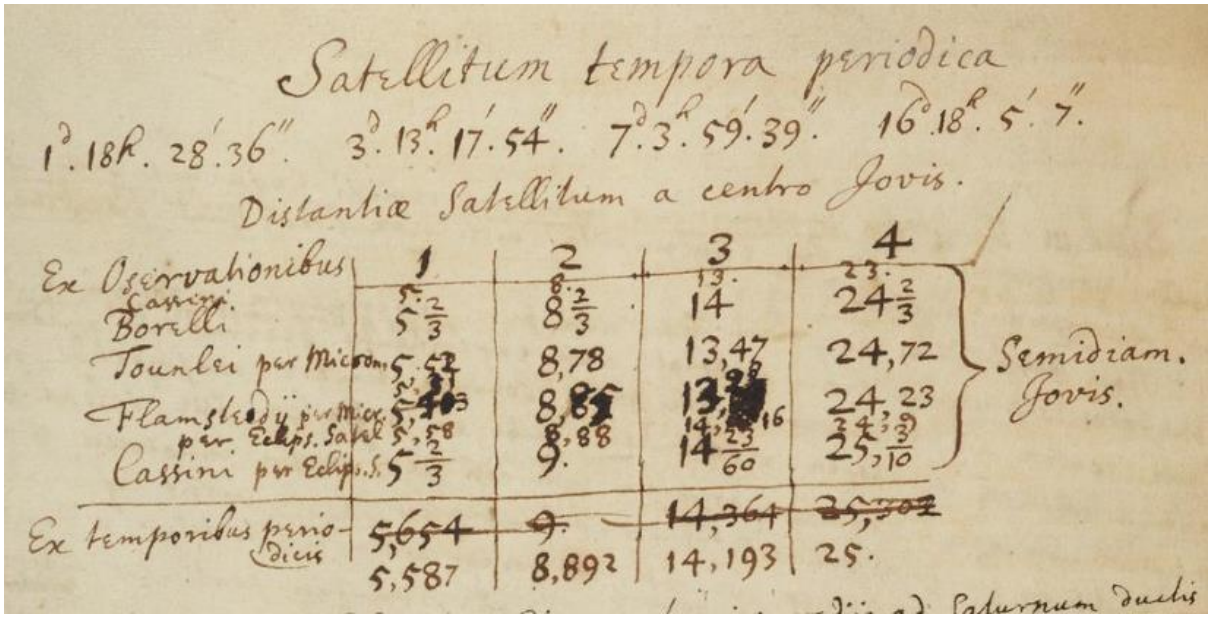
Quomodo dicitur «data science» Latine?[1]

Matthew A Jay and Mario Cortina Borja
University College London



Astronomical data from Isaac Newton's *Philosophæ Naturalis Principia Mathematica,* London: The Royal Society, 1687, hand-written by Newton on his own copy of the book.
https://cudl.lib.cam.ac.uk/view/PR-ADV-B-00039-00001/783

INTRODUCTION

Why should we care about translating a term into Latin which has only been very recently adopted in English? Many readers of *Significance* will perhaps be surprised to learn that there exists a growing community of Latin *speakers*[2] many of whom pursue non-classics professions and who speak the language for the sake of it. We would indeed not be surprised if there were at least a handful of readers who speak the language (*salvete vos, amicæ amicique!*). Nevertheless, for everyone, *Latine loquentes* or otherwise, the etymology of a word often provides interesting insights on its meaning and on how society interprets it. Usually this process travels forward in time, in the sense that a word firstly used in one language, e.g. Greek or Latin, finds its way into another. Sometimes,

[1] How do you say "data science" in Latin?
[2] By speakers, we mean speakers. There are many, MAJ among them, who use Latin in everyday social life as well as at conferences, in teaching and in publishing. See www.circuluslatinuslondiniensis.co.uk/vincula for a wealth of Latin resources for speaking Latin. For a Latin-language podcast about epidemiology, «Salvi Sitis!», see www.salvisitis.libsyn.com.

travelling in the reverse linguistic way may enrich our understanding of language. An interesting example of this is the work by Pigoli et al (2018) which applies complex statistical methods in phonetics with the aim to reconstruct "how the speakers of extinct[3] languages might have sounded." In this short article, we also move in the opposite direction to the usual Latin to English etymology, and look at philological and historical sources in an attempt to answer the question posed in the title, and in doing so, to discuss what do we mean by data science.

DATA SCIENCE

Whatever we mean by it, data science is enjoying a successful drift. Asking "what is data science?" in Google yields nearly half a million results, and the number of articles in the WoS database whose abstract includes the term has grown exponentially since 2010, reaching nearly 1000 articles published in 2018. *The Journal of Data Science*, whose first issue appeared in 2003, states in its scope that by *data science* it means "almost everything that has something to do with data: Collecting, analyzing, modeling... yet the most important part is its applications—all sorts of applications. This journal is devoted to applications of statistical methods at large". Another definition is given by UK Health Data Research: "Health data science combines maths, statistics and technology to study different types of health problems using data"[4]. Clearly, these descriptions from leading sources make it difficult to separate data science from applied statistics.

In order to translate *data science* into Latin, we need a clearer definition; we therefore review two possible origins of this term. Firstly, in 2000 Noburo Ohsumi, from the Institute of Statistical Mathematics, Tokyo, claimed that in 1992 he "argued the urgency of the need to grasp the concept "data science" and that Japanese researchers had by then collaborated with French colleagues "in the field of data science"; however, "the history of these exchanges is not widely known among statistical science researchers" (Ohsumi, 2000). For Ohsumi, data science includes "the most essential studies and concepts on *how to gather data*, including *how to design experiments in data gathering*, and *how to analyse the collected data*" (his italics). Ohsumi (2000) points out the important questions in data science are "what dataset is necessary to explicate a certain phenomenon, why is it necessary, how to design its acquisition, and how difficult the whole process is". Secondly, in 2001 WS Cleveland, from Bell Labs, published "an action plan to enlarge the technical work of the field of statistics" and since this is "ambitious and implies substantial change,

[3] An extinct language is one that no longer has any speakers. Latin is sometimes considered to be a dead language though it is arguably more accurate to consider it an endangered one as there does exist a community of speakers, including children.

[4] https://www.hdruk.ac.uk/ (accessed November 2019).

the altered field will be called 'data science'" (Cleveland, 2001). Whilst both papers are centred in what Cleveland explicitly names "computing with data", his paper defines data science by insisting that it is part of "the field of statistics" whilst Ohsumi sets a broader scope including the end goal of designing experiments in data gathering. These aspects play an important role in exploring answers to our title question.

TRANSLATING "DATA SCIENCE" INTO LATIN

Coining Latin neologisms requires us to consider the history of the concept itself as well as the words used to describe it. Springhetti (1954) warns that although this is a necessary task, we do not have free reign.[5] With something as complex as science, the task becomes that much harder. It is therefore beyond the scope of this article to conduct a survey of all Latin literature. Nor can this article be a study in the history of science. Both of these tasks would require the examination of a profoundly large literature in at least two languages. We hope therefore that the reader will forgive our translation for remaining to some degree *sub judice*.

A sensible place to start when translating *data science* is the two Latin words from which these words derive: *scientia* and *data*. Firstly, s*cientia* means *knowledge*, *expertness* or *skill*. Thus, to pick one convenient example, Quintilianus:

> There is nothing worse than they who, having themselves learned little more than the elementary rudiments, wrongly comport themselves as though they possess actual knowledge (*scientiæ*).[6]

This understanding of *scientia* is easily understood when observing that it derives from the verb *scio*–I know–whose present participle is *sciens*—knowing. The dictionaries do give this translation of *science* as an alternative but it is not straightforward. Firstly, the Oxford English Dictionary (OED)

[5] "Transitus ex unâ (primigeniâ) in aliam significationem analogicam in multis vocabulis jam factus est apud antiquos optimos scriptores, præsertim *translatione*; et cum pertineant ad thesaurum Latinitatis, nihil est contra dicendum. Non vero datur nunc etiam licentia, ut unusquisque ex linguæ imperitiâ, vocabula et constructiones ad novas significationes arbitrio suo detorqueat, vel ad aliquid exprimendum novas inducat formas non necessarias vel aliud significantes [italics in original]"—"Moving from one original meaning to an analogical one was something even the best ancient authors practised in many instances, especially through metaphor; and insofar as it pertains to the extant corpus of Latinity, nothing can be said against it. But we do not have free licence that each and everyone of us, through ignorance of the language and left each to our own devices, may torture words and phrases to give new meanings, or to force out new unnecessary forms or meanings." (translated by MAJ).

[6] *Institutiones*, 1.1.8: *Nihil est pejus iis qui paulum aliquid ultra primas litteras progressi falsam sibi scientiæ persuasionem induerunt*. All translations in this article are MAJ's.

reveals that the original meaning of *science* in English is more akin to that of *scientia* in Latin: "The state or fact of knowing; knowledge or cognizance of something; knowledge as a personal attribute. Now archaic and rare." It is only at its 7th definition that the OED gives a definition close to what we think of as science: "A branch of study that deals with a connected body of demonstrated truths or with observed facts systematically classified and more or less comprehended by general laws, and incorporating trustworthy methods (now esp. those involving the scientific method and which incorporate falsifiable hypotheses) for the discovery of new truth in its own domain." This clearly doesn't equate to *knowledge*, *expertness* or *skill* and although there is an obvious nexus between knowledge and science, *scientia* is not used in Latin with this specialised meaning.

The traditional way of expressing in Latin what we consider science to be is *philosophia naturalis*—natural philosophy—and it is this term that a more purist rendering of *science* in Latin would require. It may sound odd to talk of philosophy as science but it must be remembered that the distinction between science and philosophy was only completed at the 19th Century, a long process studied by Frank (1952). We still see a very common vestige of this older understanding in the degree of *Philosophiæ Doctor* – Doctor of Philosophy.

Isidore of Seville, writing in the 6th Century CE says:

> The face of philosophy is tripartite: the first part is the *natural*, what in Greek is called *Physics*, in which is considered the investigation of nature; the second is moral, *Ethics* in Greek, which is about morals; the third is *rational*, which is called by its Greek word *Logic* and in which is disputed how truth itself be sought in the causes of things or morals of life. In *Physics*, therefore, resides the cause of inquiry, in *Ethics* the order of living and in *Logic* the rational method of understanding.[7]

It is therefore clear that, in pre-modern authors, *philosophia naturalis* is much closer to what we now think of as science. Of course, Isidore and the Romans before him never had the benefit of the modern scientific method but that is irrelevant: nobody would argue, for example, that Genome-

[7] *Origines*, 2.24.3: *Philosophiæ species tripertita est: una naturalis, quæ Græce Physica appellatur, in quá de naturæ inquisitione disseritur: altera moralis, quæ Græce Ethica dicitur, in quá de moribus agitur: tertia rationalis, quæ Græco vocabulo Logica appellatur, in quá disputatur quemadmodum in rerum causis vel vitæ moribus veritas ipsa quæratur. In Physicá igitur causa quærendi, in Ethicá ordo vivendi, in Logicá ratio intellegendi versatur.*

Wide Association Studies are not a part of science just because the method is new. What matters is that we are using the same word(s) for the same phenomenon: the investigation of nature.

We find the same three-part definition of *philosophia* in a letter by Seneca the Younger (4 – 65 CE) in which he discusses the point and proper place of the liberal arts. This letter is instructive for us for his discussion on the relationship between mathematics and natural philosophy. He first sets up a straw man who says that when we come to investigations of nature, we rely on geometry (e.g. physical measurements of size); therefore geometry is part of natural philosophy. Seneca responds:

> There are many things we rely on that are not parts of us and, in fact, if they were, they would not help us. Food is essential for the body but is by no means a part of it. Geometry is of some help to us: in the same way that the carpenter is necessary to geometry, geometry is necessary to philosophy. The carpenter, however, is not a part of geometry nor is geometry a part of philosophy. Each discipline has its own limits: the wise man seeks and discovers the causes of natural phenomena, the numbers and measurements of which the geometrician seeks and computes.[8]

The point here is to draw out the distinction between the investigation of natural phenomena (*philosophia naturalis*) and the technology we use to do so (here, *geometria*). This is a question that has direct relevance to us as modern natural, social and physical scientists, statisticians and data scientists.

Based on Ohsumi's and Cleveland's definitions of data science, we submit that it is very clear that data science falls very much within the technology side. For instance, Healy (1978) argues that statistics is likewise a technology rather than a science. Data science is not the investigation of natural phenomena or their causes, it is part of the process we use to measure those phenomena. In Seneca's reasoning, data science is necessary to science (as in, natural philosophy) but is no more a constitutive part of it than is a scatter plot of data science. In other words, *«data science» non est philosophia naturalis*: data science is not science.

---

[8] *Ad Lucilium*, 88.25: *Multa adjuvant nos nec ideo partes nostri sunt; immo si partes essent, non adjuvarent. Cibus adjutorium corporis nec tamen pars est. Aliquod nobis præstat geometria ministerium: sic philosophiæ necessaria est quomodo ipsi faber, sed nec hic geometriæ pars est nec illa philosophiæ. Præterea utraque fines suos habet; sapiens enim causas naturalium et quærit et novit, quroum numeros mensurasque geometres persequitur et supputat.*

It can be argued that such a purist translation ignores the historical processes that occurred up to the 19th Century and beyond that shaped what we now call and think of science: after all, *natural philosophy* was used in English as well and it is arguable that modern science is indeed distinct from it. If Latin is to be used as a tool of modern communication, we cannot ignore these processes. It is true that the phrase *philosophia naturalis* is not confined to the ancients. Newton used it in titling one of the most important scientific treatises ever penned: *Philosophiæ Naturalis Principia Mathematica*—the Mathematical Principles of Natural Philosophy—which perhaps also represents the same division between science (natural philosophy) and technology (the mathematical principles) that Seneca and Healy noted. Also, Gauss (1830) himself uses the phrase when discussing the movement of molecules as described by Laplace. The phrase *philosophia naturalis* is therefore not confined solely to situations devoid of modern scientific principles. However, these works predate the separation of science from philosophy. It is for these reasons that many Latin speakers do use the word *scientia* for science. It is also for these reasons that one possible translation of the *science* part of *data science* is in fact *scientia*.

If for the moment, however, we wish to put *scientia* aside for purist reasons. What, then? Of all the possible candidates, we would argue that the word which bears closest resemblance to the most clearly defined understandings of data science is *principia* (principles / elements / foundations).[9] What are "the most essential studies and concepts on *how to gather data*, including *how to design experiments in data gathering*, and *how to analyse the collected data*" (Ohsumi, 2000) if not *principia*? Using this word also enables us to explicate the relationship between these *principia*—and the next question is: *principia* of what?—and science in a rather pleasing way: "*Philosophiæ Naturalis Principia*…" or "*Scientiæ Principia…*" We now turn to the "Of what?"—the data.

Some researchers delight in enlightening us that *data*, in English, must be a plural countable noun because it derives from the nominative plural neuter perfect passive participle of the Latin *dare*—to give. Essentially, it means *things given* so we're using it in English in the sense of *things taken to be true*. The problem with this tiresome pedantry is that, firstly, some of the earliest attested uses of the word *data* in English treat it as a countable singular noun. For instance, in the OED we find this sentence, used in 1645: "The verticall Angles, according to the diversity of the three Cases being by the foresaid Datas thus obtained" From 1702, we find an uncountable, mass, noun; for example: "And by this Data there are twelve Problems resolved." More importantly, the word *data* never

[9] Other possibilites include *historia naturalis* (natural history), *physica* (physics), *studium* (in its primary sense, zeal or interest, but also application to learning or study), and *disciplina* (discipline).

seems to have been used in the sense of *information* in Latin. The closest we get is the sense of conceding a point in philosophical discourse, as we find in Cicero:

> … if you should have granted (*dederis*) the first propositions, all others must be granted (*danda*).[10]

This is certainly close to the way in which we use the word *data* in English—as something given in the metaphorical sense—but is not what we have in mind when talking about data. The three definitions in the OED are "An item of information," "Related items of (chiefly numerical) information considered collectively, typically obtained by scientific work and used for reference, analysis, or calculation" and "Computing. Quantities, characters, or symbols on which operations are performed by a computer, considered collectively. Also (in non-technical contexts): information in digital form." We have moved firmly into the territory of recorded information.

Given that this sense of *data* in English is first attested in the 17th Century, it is best to seek usage in later writers. There are a number of possibilities other than the word *data*, but none are entirely satisfactory given the specialised meaning that the word now has. In William Harvey's 1628 treatise *Exercitatio Anatomica De Motu Cordis et Sanguinis*—An Anatomical Study on the Movement of the Heart and Blood—we frequently find *observatio*, observation. Thus:

> I therefore trust that, out of these and like observations (*observationibus*), the motion of the heart has been discovered to occur thusly.[11]

An English writer today could easily replace "these and like observations" with the "data" and the meaning would remain unaltered. We also find *observatio* in Galileo (1610) and Newton (1686). Perhaps closest is Newton (1686) who even lays out tables with numerical entries, which he refers to as *observationes* (see the table at the head of this paper). However, it is possible to make an observation without recording it and the word is clearly being used in the literal sense of observing. Thus in Galileo's 1610 *Sidereus Nuncius*—the Sidereal Messenger:

---

[10] *De Finibus*, 5.28.83: *…prima si dederis, danda sunt omnia.*

[11] Harvey (1628): *Ego vero ex his tandem, & hujusmodi observationibus repertum iri confido, motum Cordis ad hunc modum fieri.*

> A recent astronomical message declaring *observations* achieved by means of a new telescope of the face of the moon...[12]

The other *observationes* in Newton (1686) and Harvey (1628) are of the same kind. Among Gauss (1799) and Newton (1686) we find the words *quantitas* (quantity) and in Newton (1686) *qualitas* (quality) and *phænomena* (phenomenon) but these are considered properties of the things observed, not the recorded observations themselves. The word *mensura* (measurement) occurs frequently in Newton (1686). It is probably possible to argue that all data are measurements in some way or another: even a sequential index integer is a measurement of the order in which rows are added. But, like *observatio*, it is possible to make a measurement without recording it.

In truth, the Latin word *data* is probably the best candidate even if there is no attested usage in classical or later writers. It is legitimate even in classical Latin to coin neologisms by analogy, as illustrated by Springhetti (1954). The sense that Cicero had in mind in the quote above is an analogical leap from the literal sense of *dare*—to give—to the philosophical one and it is not a huge leap thence to what we think of as data (this is after all the analogical leap that authors in the 17th Century made when adopting *data* into English). Using *data* also has the advantage that it is understood by modern speakers. The "Of What?" we posed above can be answered by *data*. Possession in Latin is expressed in the genitive case which for *data* is *datorum.*

It seems to us, therefore, that the best candidate translations into Latin of *data science* are *Principia Datorum* or *Scientia Datorum*. These are, of course, neologisms. It would have been impossible to find the phrase *data science* in the established corpus of Latinity and even finding the word *data*, despite its long pedigree in English, is difficult. However, they are neologistic phrases that are not borne *ex nihilo* and we submit that they are, at least as starting points, a good rendering in the Eternal Language, of this very modern thing.

CONCLUSION

Our discussion has led us to answer our title's question as *Principia Datorum* or *Scientia Datorum*. It has forced us to carefully confront exactly what the technology of data science is, what science is, and the relationship between the two, a distinction which the ancients were attuned to long before either modern science or data science were ever conceived of. The Romans were excellent

[12] Galileo (1610): *Astronomicus nuncius observationes recens habitas novi perspicilii beneficio in lunæ facie […] declarans.*

technologists in the sense explained by Healy (1978) and they were well versed in both collecting and analysing empirical data, and producing official statistics. An example of the former comes from the politician, Frontinus (40 – 103 CE), who was put in charge of the aqueducts (Rodgers, 1986). He first observed the impossibility that, according to imperial records, more water was coming out of the aqueducts than was coming in. Compelled to investigate, he found that the records were wrong: the amount going in was actually higher, far higher than could reasonably accounted for by expected leakage. It turned out that public officials had been deliberately using smaller pipes at the exits in an attempt to conceal illegal tapping of the water supply.[13] The latter can be illustrated by an example well known to actuaries: the life table (Figure 1) compiled by the jurist Ulpian (d 223 CE) which allowed computing the tax due on an annuity depending on the subject's age.

Table 1. Ulpian's life table and the customary life table (years)

A. Annuitant's present age	B. Corresponding figure	
	1. Ulpian	2. Customary
Birth–19	30	30
20–24	28	
25–29	25	
30–34	22	(60–x)
35–39	20	
40–49	(60–x–1)	
50–54	9	
55–59	7	
60 onward	5	0 (?)

Figure 1: A modern representation of Ulpian's life table (after Frier (1982))

Neither science nor data are new, though are ways of doing them are. We in the sciences are often guilty of neglecting what has been published even the in recent past. Although we have here barely scratched the surface, thinking about data science in Latin takes us back to the foundational *principia* on which all our endeavours are built.

POSTSCRIPTUM: ADHORTATIO AD LECTORES AMABILES

Matthæus lectricibus lectoribusque optimis suís sal.

Magnum mihi est gratum vos ad hunc locum pervenisse post symbolam longam super rebus spinosis. Ut diximus, locutiones a nobis propositæ sunt nec firmæ nec fixæ. Timeo ne sint loca a me prætermissa ubi illud *datum* ipsum invenitur aut, quod est proximum, verba locutiove ad quæstionem nostram magis pertinentes inveniuntur. Si jam vobis sunt ulla nota, quæso mihi mittas apud matthew.jay.15@ucl.ac.uk.

[13] *De Aquaductu*, 2.64-77.

Curate ut valeatis quam optime!

ACKNOWLEDGEMENTS

We thank A. Gratius Avitus for his discussion on the question of the distinction between science and natural philosophy. Gratias A. Gratio Avito agere velimus quod collocutus cum Matthæo est super eo quod intersit inter scientiam et philosophiam naturalem.

REFERENCES

Cleveland WS (2001) An Action Plan for Expanding the Technical Areas of the Field of Statistics. *International Statistical Review / Revue Internationale de Statistique*, **69**, 21-26.

Frank P (1952) The origin and the separation between science and philosophy. *Proceedings of the American Academy of Arts and Sciences*, **80**, 115-139.

Frier B (1982) Roman life expectancy: Ulpian's evidence. *Harvard Studies in Classical Philology*, **86**, 213-251.

Galilei G (1610) *Siderus Nuncius*. Venice: Thomas Baglionum.

Gauss CF (1799) *Demonstratio Nova Theorematis Omnen Functionem Algebraicam Rationalem Integram Unius Variabilis in Factores Reales Primi vel Secundi Gradús Resolvi Posse*. Helmstadt: Fleckeisen.

Gauss CF (1830) *Principia Generalia Theoriæ Figuræ Fluidorum in Statu Æquilibrii*. Gottingen: Dieterich.

Harvey W (1628) *Exercitatio Anatomica De Motu Cordis et Sanguinis in Animalibus*. Frankfurt: Guilielmus Fitzerus.

Healy MJR (1978) Is Statistics a Science*? Journal of the Royal Statistical Society, series A*, **141**, 385-393.

Newton I (1687) *Philosophiæ Naturalis Principia Mathematica*. London: The Royal Society.

Ohsumi N (2000) From data analysis to data science, pp 329-334 in Kiers HAL; Rasson J-P; Groenen PJF; Schader M (eds) (2000) *Data Analysis, Classification and Related Methods*. Springer: Berlin. (Proceedings of the 7th Conference of the International Federation of Classification Societies).

Pigoli D; Hadjipantelis PZ; Coleman JS; Aston JAD (2018) The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages (with discussion). *Applied Statis*tics, **67**, 1103–1145.

Rodgers RH (1986) *Copia Aquarum*: Frontinus' Measurements and the Perspective of Capacity. *Transactions of the American Philological Association*, **116**, 353-360.

Springhetti A (1954) *Institutiones Stili Latini*. Rome: Pontifica Universitas Gregoriana.