***NB: pour gagner du temps et faciliter la correction, traitez les parties A, B et C directement sur l'énoncé et placez-le à l'intérieur de la copie qui vous a été remise***

## A. Structures de la langue (4 points – niveau B1)

**1°) Usage des déterminants articles**: *corrigez 5 erreurs dans ce petit texte*
*(une correction juste rapporte 0,2 pt, une correction erronée coûte 0,1 pt)*

Big data starts with the fact that there is a lot more information floating around these days than ever before, and it is being put to an extraordinary new uses.
For most of the history, people have worked with relatively small amounts of data because the tools for collecting, organizing, storing and analyzing an information were poor. People reduced the information they relied on to the barest minimum so that they could examine it more easily.

**2°) Données quantitatives** *écrivez les nombres symboles et acronymes en toutes lettres, comme lorsque vous les prononcez à voix haute:*

***The technical environment has shifted 179 (one hundred [and] seventy-nine) degrees due to the*** ***2***
***development of the Internet in the 1990s (nineteen nineties).*** ***1***

***Today, every person alive could get 320 (three hundred and twenty) times as much*** ***1***

***information as was stored in the Library of Alexandria in the third century BC (Before Christ)*** ***2***

***an estimated 1,200 (twelve hundred / one thousand two hundred) exabytes' worth,*** ***2***
***with one exabyte being equal to 1,000,000,000 (one/a billion) GB (gigabytes).*** ***2***

**0,1 pt per item** **TOTAL = 10**

**3°) *Construire des phrases au passif avec les mots suivants aux formes appropriées :***

The World Wide Web – invent – Tim Berners-Lee

The World Wide Web was/has been invented by Tim Berners-Lee
Facebook users – give away – every day - enormous amounts of data

Every day, enormous amounts of data are given away by Facebook users
for the last ten years – sufficiently - personal data – not – protect

For the last ten years, personal data have not been sufficiently protected

Cukier – write – not – this article – alone

This article was not/has not been written by Cukier alone

## B. Décomposition et prononciation des mots (3 points – niveau B1)

Indiquer la catégorie (N=nom, V=verbe, A=adjectif, AV=adverbe) et la prononciation en notation API des mots suivants du texte (après les avoir éventuellement décomposés en écrivant les préfixes et suffixes EN MAJUSCULES et en parenthésant au besoin les étapes de la décomposition)
catégorie finale : mot = décomposition des préfixes et suffixes => /prononciation déduite/
notez les catégories avec N pour nom, V pour verbe, A pour adjectif et AV pour adverbe
par exemple: N : synthesizer = V: ( N: (SYN- + N: thesis) + -IZE) + -ER => /'sɪnθəsaizər/
N : inaccuracies = IN- + N : (A : accurate + -CY) + -S => /'ækjurəsiz/ 6 items
AV : carefully = A : (N : care + -FUL) + -LY => /'keəʳfu li/ 8 items
A : vibrational = N : (V : vibrate + -TION) + -AL => /vai'brei ʃᵊn ᵊl/ 10 items
N : failures = => V : fail + -URE + -S => /'feiləz/ 6 items

2 points pour 30 items

Dans l'extrait qui suit, cerclez précisément 8 occurrences de la voyelle /ɔ:/ (comme dans « door ») et de la diphtongue /ei/ (comme dans « rate ») :
These two shifts in how we think about data—from some to all and from clean to messy—give rise to a third change: from causation to correlation. This represents a move away from always trying to understand the deeper reasons behind how the world works to simply learning about an association among phenomena and using that to get things done. Of course, knowing the causes behind things is desirable.
1 point pour 8 items corrects sur les 12 possibles

## C. Compréhension écrite élémentaire (4 points – niveau B1)

***Les justifications doivent être de très courts extraits du texte judicieusement choisis***

a) Faux Vrai
□ ☑ on travaille par échantillonnage lorsque l'on est pressé d'avoir les résultats
… and time consuming, the sample was a savior
/ polls on election night ...

b) Faux Vrai
☑ □ il faut toujours définir à l'avance ses objectifs avant de collecter des données statistiques
we do not need to know beforehand what we plan to use it for

c) donnez deux synonymes en anglais de : « shortcoming » : deficiency limitation defect failure flaw lack

d) Faux Vrai
☑ □ un grand volume de données ne sert à rien dès qu'il contient des valeurs erronées
we can now allow some inaccuracies to slip in

e) Faux Vrai
□ ☑ rechercher des relations de causalité est un défaut typique de l'esprit humain
humans are conditioned to see causes even where none exist

f) Faux Vrai
☑ □ UPS ne change que les pièces réellement défectueuses
replace the part when it is convenient, instead of on the side of the road

g) Faux Vrai
☑ □ le suivi des paramètres vitaux permet d'identifier immédiatement les causes d'infection
knowing that something is likely to occur can be far more important than understanding exactly why

h) donnez un contraire en anglais de « overt » : concealed hidden masked

i) Faux Vrai
□ ☑ le texte est en anglais américain
savior / behavior

## D. Reformulation et synthèse (3 points – niveau B2)

*Answer each question in English in 5 to 10 lines ; carefully summarise and rephrase, and don't just copy-paste whole sentences from the original text: any select citation should be short and explicit (enclosed within quotation marks)*

1°) Why is a sampling approach frequently preferred to a complete systematic analysis?

4 reasons: data availability, data collection cost, processing complexity, time pressure (0,25 pt / argument)

2°) How do the authors consider the notion of causation? (their consider it as both a worthwhile pursuit and a deceptive one. Also the search for causes may lead to over-simplification and operational inefficiency in emergency situations.) (0,25 pt per item.)

3°) What makes UPS and Canadian doctors comparable?

Both have to solve concrete problems under severe safety and efficiency constraints and they privilege swift action and delay the in-depth analysis of the causes, even if they may sometimes over-react.
(0,2 pt per item.)

## *E. Production écrite élémentaire (3 points – niveau B1 - 150 mots)*

***NB : les étudiants des niveaux 2 & 3 doivent traiter ce sujet en priorité car l'évaluation de cette production portera davantage sur le contenu que sur la correction de la langue***

*Discuss some correlations (or absence of correlation) you observed on yourself between training and performance (either at university or in your favorite sport(s) or any other leisure activity).*

*Any example is acceptable providing the field is properly defined, the training activities and resulting performance clearly described, and a conclusion is drawn about the existence (or not) of some correlation between the two. Each complete example is worth 1 point.*

## *F. Production écrite avancée (3 points – niveau B2 – 250 mots)*

***NB : l'évaluation de cette production portera à la fois sur le contenu, le lexique et la syntaxe, ne la traitez que si vous avez fini tout le reste***

*Using various examples, explain the relationship between correlation and causation.*

*Causation implies correlation, but the reverse is not true : a correlation may be observed between two phenomena due to an external common cause (a classical example is the strong seasonal correlation between the sales of icecreams and sunglasses). However, a strong correlation between two series of data should lead to an investigation of possible causality links. When a causality link is known to exist, demonstrating its existence through the analysis of empirical data is not always possible because the expected correlation may be weakened by other interfering factors. Also, the relationship between the two sets of data can be distorted :*

*1°) Since the effect of a cause can include a time lag (eg by hysteresis or some other type of systemic inertia), two data series may not show a strong correlation, unless one of them undergoes a time-shift (eg the correlation between level of investment and growth rate).*

*In certain cases, the relation between effect and cause is not linear in magnitude, and some re-scaling of the raw data may be necessary (eg. The dose/response curves showing the effect of a drug)*

## Big Data

The way people handled the problem of capturing information in the past was through sampling. When collecting data was costly and processing it was difficult and time consuming, the sample was a savior. Modern sampling is based on the idea that, within a certain margin of error, one can infer something about the total population from a small subset, as long the sample is chosen at random. Hence, exit polls on election night query a randomly selected group of several hundred people to predict the voting behavior of an entire state. For straightforward questions, this process works well. But it falls apart when we want to drill down into subgroups within the sample. What if a pollster wants to know which candidate single women under 30 are most likely to vote for? How about university-educated, single Asian American women under 30? Suddenly, the random sample is largely useless, since there may be only a couple of people with those characteristics in the sample, too few to make a meaningful assessment of how the entire subpopulation will vote. But if we collect all the data — ”n = all,” to use the terminology of statistics — the problem disappears.

This example raises another shortcoming of using some data rather than all of it. In the past, when people collected only a little data, they often had to decide at the outset what to collect and how it would be used. Today, when we gather all the data, we do not need to know beforehand what we plan to use it for. Of course, it might not always be possible to collect all the data, but it is getting much more feasible to capture vastly more of a phenomenon than simply a sample and to aim for all of it. Big data is a matter not just of creating somewhat larger samples but of harnessing as much of the existing data as possible about what is being studied. We still need statistics; we just no longer need to rely on small samples.

There is a tradeoff to make, however. When we increase the scale by orders of magnitude, we might have to give up on clean, carefully curated data and tolerate some messiness. This idea runs counter to how people have tried to work with data for centuries. Yet the obsession with accuracy and precision is in some ways an artifact of an information-constrained environment. When there was not that much data around, researchers had to make sure that the figures they bothered to collect were as exact as possible. Tapping vastly more data means that we can now allow some inaccuracies to slip in (provided the data set is not completely incorrect), in return for benefiting from the insights that a massive body of data provides. [...]

These two shifts in how we think about data—from some to all and from clean to messy—give rise to a third change: from causation to correlation. This represents a move away from always trying to understand the deeper reasons behind how the world works to simply learning about an association among phenomena and using that to get things done. Of course, knowing the causes behind things is desirable. The problem is that causes are often extremely hard to figure out, and many times, when we think we have identified them, it is nothing more than a self-congratulatory illusion. Behavioral economics has shown that humans are conditioned to see causes even where none exist. So we need to be particularly on guard to prevent our cognitive biases from deluding us; sometimes, we just have to let the data speak.

Take UPS, the delivery company. It places sensors on vehicle parts to identify certain heat or vibrational patterns that in the past have been associated with failures in those parts. In this way, the company can predict a breakdown before it happens and replace the part when it is convenient, instead of on the side of the road. The data do not reveal the exact relationship between the heat or the vibrational patterns and the part’s failure. They do not tell us why the part is in trouble. But they reveal enough for the company to know what to do in the near term and guide its investigation into any underlying problem that might exist with the part in question or with the vehicle.

A similar approach is being used to treat breakdowns of the human machine. Researchers in Canada are developing a big-data approach to spot infections in premature babies before overt symptoms appear. By converting 16 vital signs, including heartbeat, blood pressure, respiration, and blood-oxygen levels, into an information flow of more than 1,000 data points per second, they have been able to find correlations between very minor changes and more serious problems. Eventually, this technique will enable doctors to act earlier to save lives. Over time, recording these observations might also allow doctors to understand what actually causes such problems. But when a newborn’s health is at risk, simply knowing that something is likely to occur can be far more important than understanding exactly why.

***Kenneth Cukier and Viktor Mayer-Schoenberger*** **The Rise of Big Data**
FOREIGN AFFAIRS May/June 2013 ● Volume 92 ● Number 3 (p 29-32)