Enjeux et défis de la recherche reproductible

Konrad HINSEN

Centre de Biophysique Moléculaire, Orléans, France
et
Synchrotron SOLEIL, Saint Aubin, France

23 mai 2019

Pourquoi tout le monde parle de reproductibilité ?

Royal Society of London for Improving Natural Knowledge (1663)

Pourquoi tout le monde parle de reproductibilité ?

Reproductibilité : preuve de **rigueur** qui inspire **confiance**

Un résultat non-reproductible suggère...

- ... une description incomplète
- ... une maîtrise insuffisante des techniques
- ... une erreur
- ... une fraude

Ou encore...

- ... une prise insuffisante sur les sujets d'étude (souris, étoiles, ...)

Le débat en cours

- Quelle est le rôle de la reproductibilité pour la méthode scientifique ?

Le débat en cours

- Quelle est le rôle de la reproductibilité pour la méthode scientifique ?
  Ma réponse : c'est un accélérateur important

Le débat en cours

- Quelle est le rôle de la reproductibilité pour la méthode scientifique ?
  Ma réponse : c'est un accélérateur important
- Quel effort est raisonnable pour améliorer la reproductibilité ?
- Quel effort est raisonnable pour vérifier la reproductibilité ?
- Comment motiver les chercheurs à faire plus d'effort ?
- Comment réduire l'effort nécessaire par la technologie ?

Du technique au scientifique

Reproductibilité

- Niveau technique
- Est-ce bien fait ?
- Évaluation simple
- Réponse simple : oui/non
- **Vérification**

Réplicabilité

- Niveau scientifique
- Est-ce la bonne chose à faire ?
- Évaluation laborieuse
- Réponse complexe : si...
- **Validation**

Du technique au scientifique

Reproductibilité

- Niveau technique
- Est-ce bien fait ?
- Évaluation simple
- Réponse simple : oui/non
- **Vérification**

Réplicabilité

- Niveau scientifique
- Est-ce la bonne chose à faire ?
- Évaluation laborieuse
- Réponse complexe : si...
- **Validation**

En calcul scientifique :

- Même logiciel
- Mêmes paramètres
- Mêmes données
- **Résultat identique ?**

- Nouveau logiciel
- Mêmes paramètres
- Mêmes données (ou pas)
- **Résultat équivalent ?**

Les enjeux de la reproductibilité en calcul scientifique

- Être sûr de ce qui a été calculé.
- Pouvoir vérifier le calcul.
- Pouvoir adapter le calcul.

Particularité du calcul : déterminisme

Exercices d'échauffement

Pour pouvoir vérifier et adapter un calcul, il faut évidement :

1. Avoir le code source et les données d'entrée.
2. Être sûr que c'est bien ce code source qui a été utilisé.
   $\longrightarrow$ pouvoir recompiler / réinstaller.

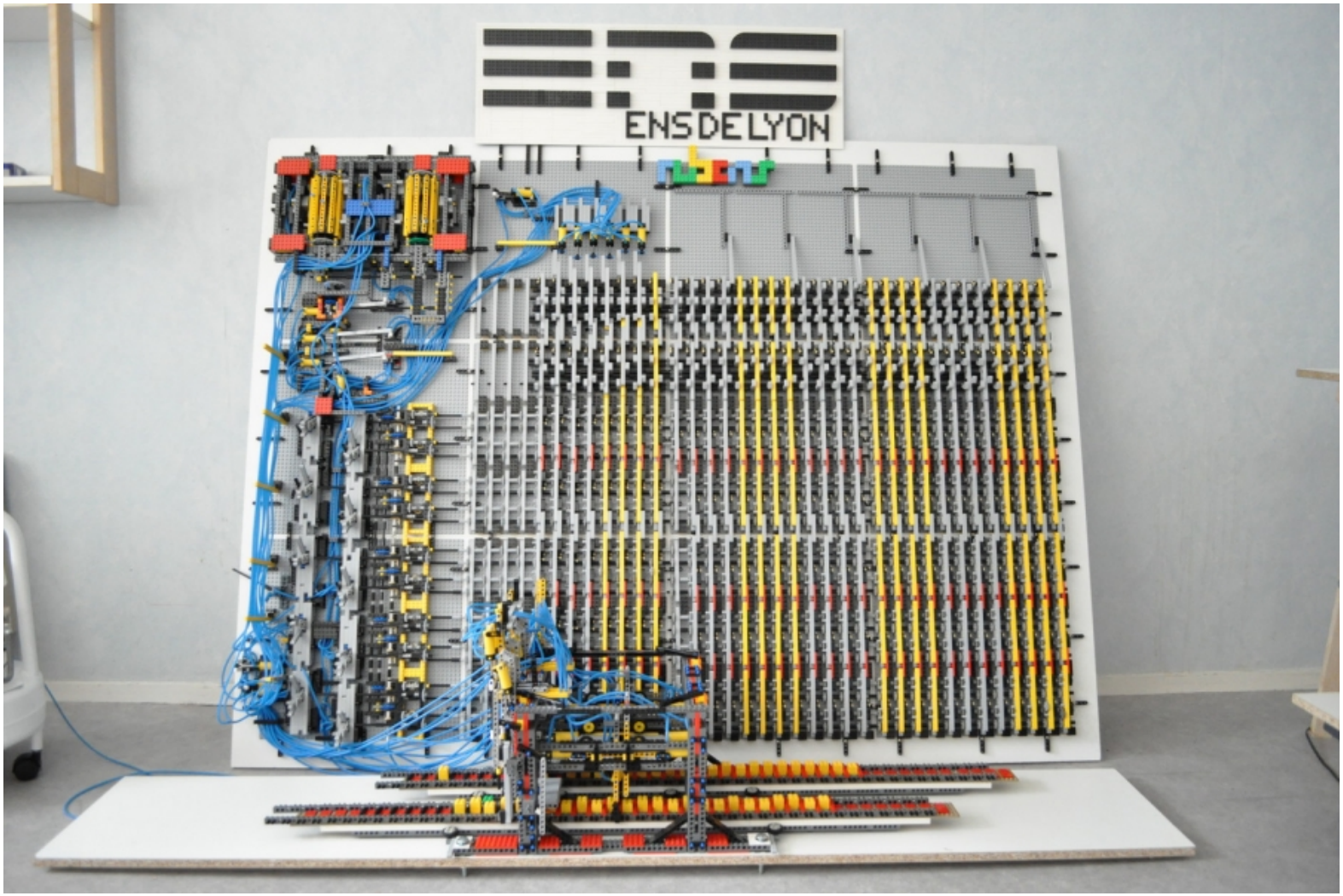
Il est avantageux d'avoir en plus

3. L'historique du code source
4. Des tests et exemples d'utilisation

Mots clés :

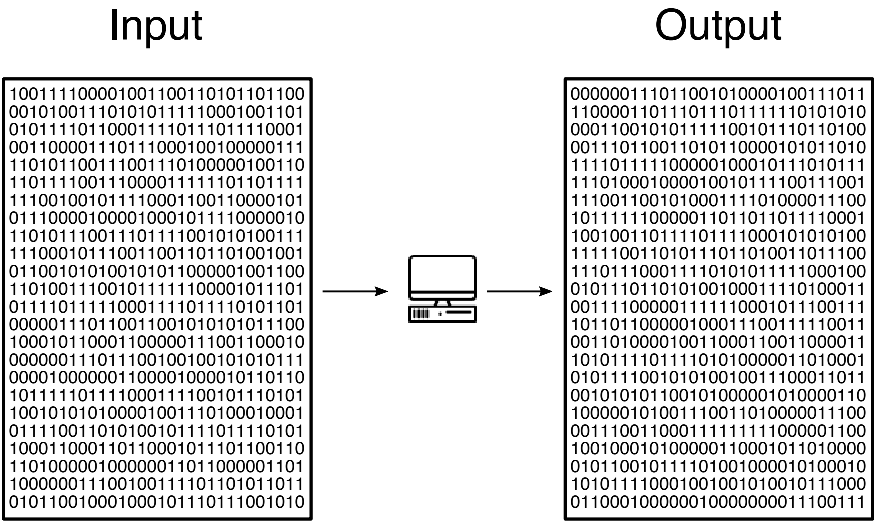
sauvegardes, archives (Zenodo, Software Heritage, ...),
gestion de versions, documentation, tests

C’est quoi un calcul ?



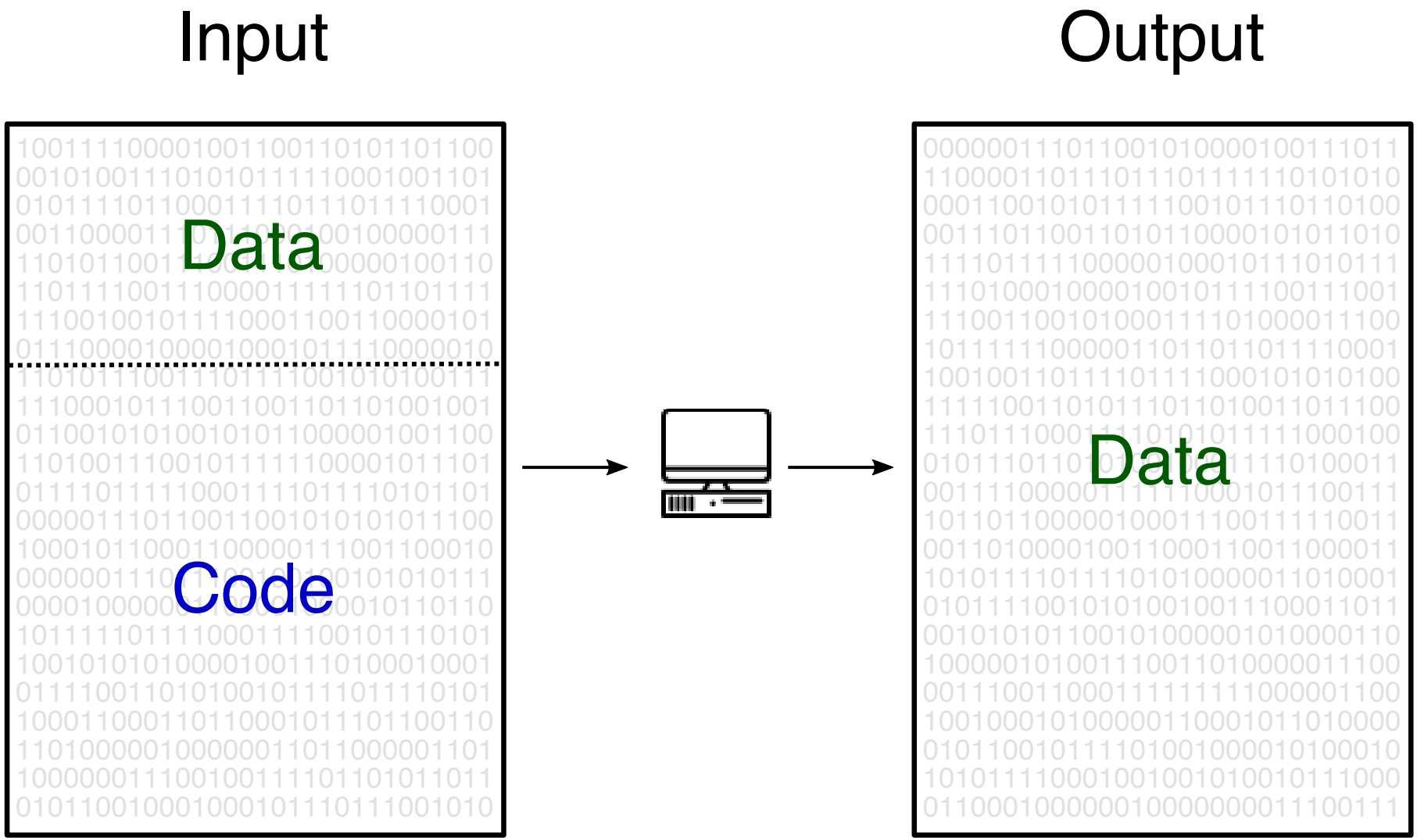
Vue globable de la machine de Turing en Légo du projet Rubens.

C'est quoi un calcul ?



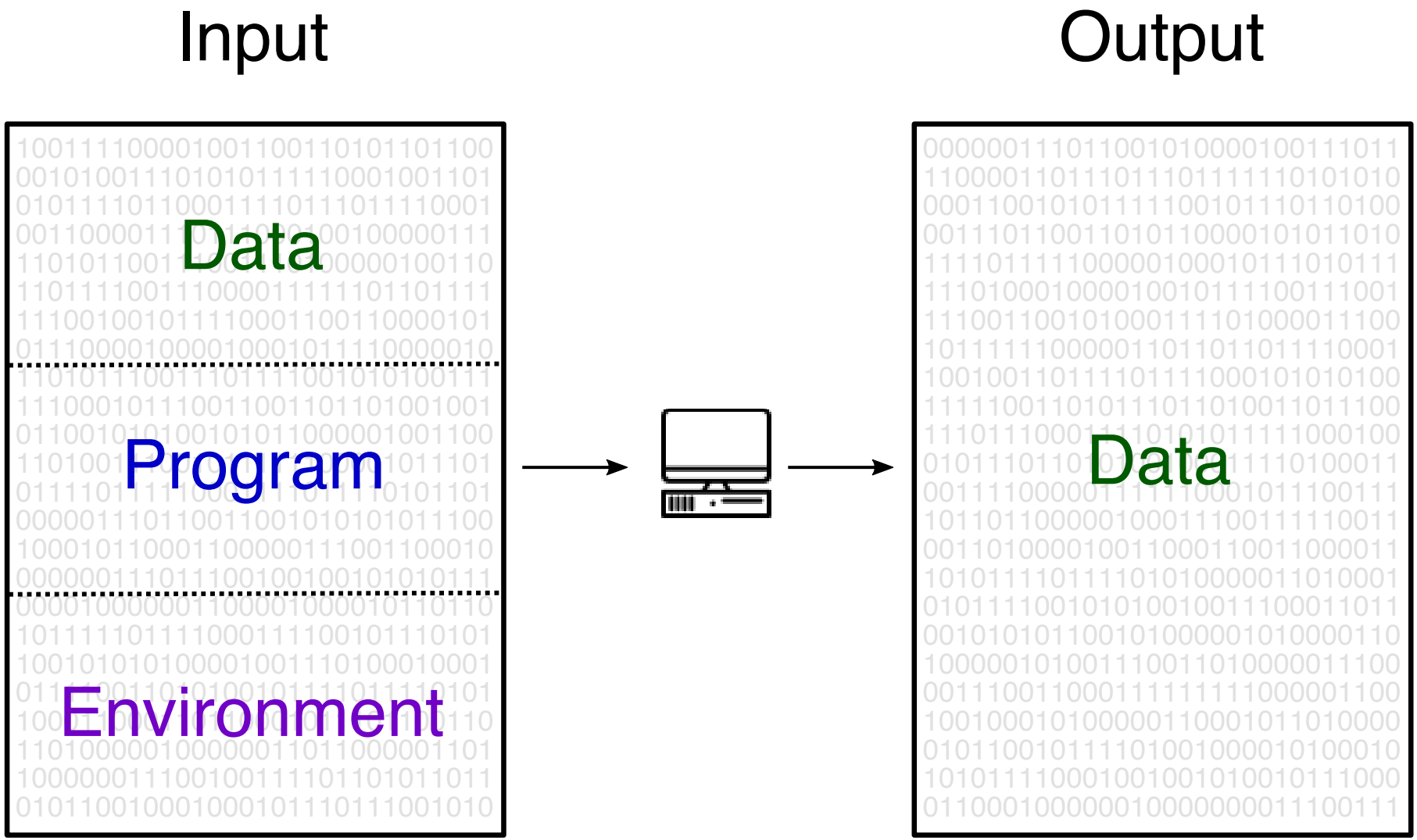
Computer by Creative Stall from the Noun Project

C’est quoi un calcul ?



Computer by Creative Stall from the Noun Project

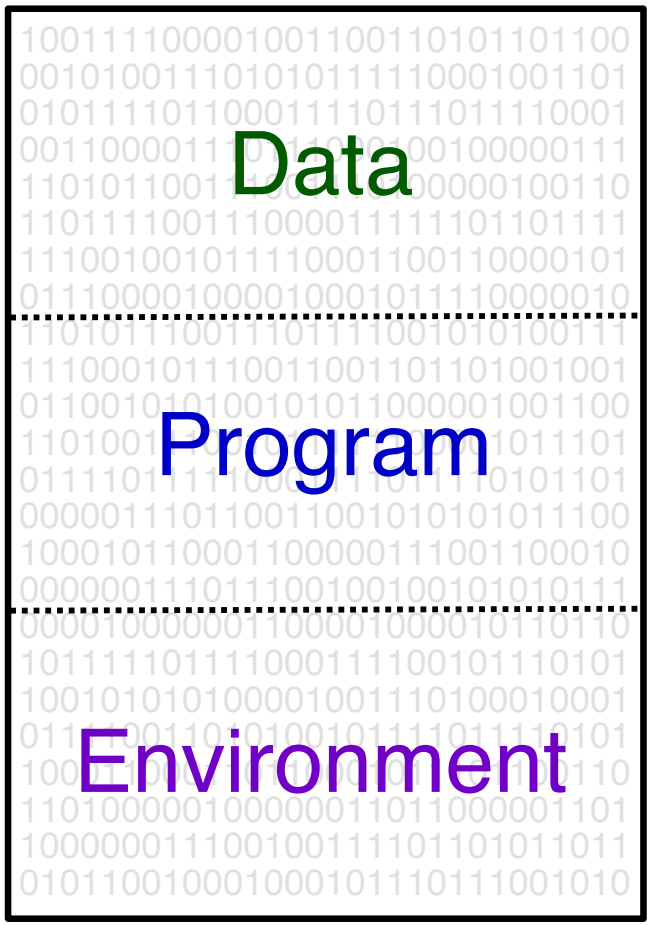
C’est quoi un calcul ?



Computer by Creative Stall from the Noun Project

Le point de vue du chercheur

Input



my research

my colleagues' code

stuff I don't care about

Que fait ce programme ?

data_analysis.py

```
from datalib import Dataset

points = [(1, 1), (-1, 1), (2, 4)]

data = Dataset()
for x, y in points:
    if x > 0:
        data.add_value(y)
print(data.average())
```

Que fait ce programme ?

data_analysis.py

```
from datalib import Dataset

points = [(1, 1), (-1, 1), (2, 4)]

data = Dataset()
for x, y in points:
    if x > 0:
        data.add_value(y)
print(data.average())
```

Réponse rapide :

Calculer la moyenne de y pour les points où x est positif.

Le résultat est 2.5.

Que fait ce programme ?

data_analysis.py

```
from datalib import Dataset

points = [(1, 1), (-1, 1), (2, 4)]

data = Dataset()
for x, y in points:
    if x > 0:
        data.add_value(y)
print(data.average())
```

Réponse correcte :

Ça dépend de `datalib`

Il faut bien connaître ses bibliothèques

datalib.py

```
class Dataset(object):

    def __init__(self):
        self.values = []

    def add_value(self, value):
        self.values = [value]

    def average(self):
        return sum(self.values, 0)/len(self.values)
```

Quelle surprise ! `add_value` ne garde que la dernière valeur !

Le résultat de `data_analysis.py` est donc 4.

Il faut bien connaître ses bibliothèques *et langages*

datalib.py

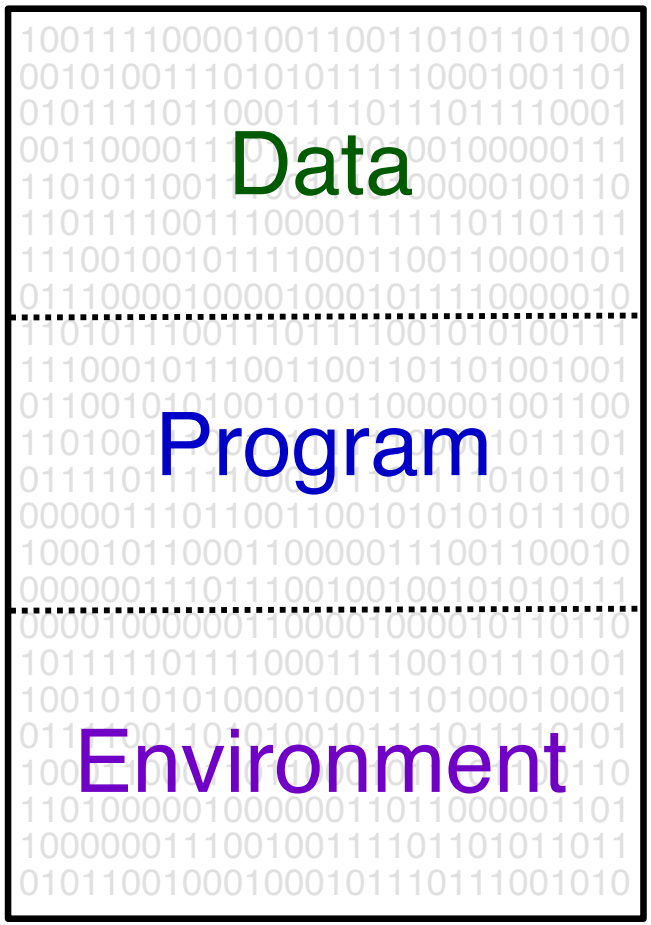
```
class Dataset(object):

    def __init__(self):
        self.values = []

    def add_value(self, value):
        self.values = [value]

    def average(self):
        return sum(self.values, 0)/len(self.values)
```

Quelle surprise ! `add_value` ne garde que la dernière valeur !

Plus précisement : 4 en Python 2 mais 4.0 en Python 3.

Donner un sens aux bits

Input



zeros and ones

interpretation of the data

interpretation of the program

Parenthèse : la sécurité informatique

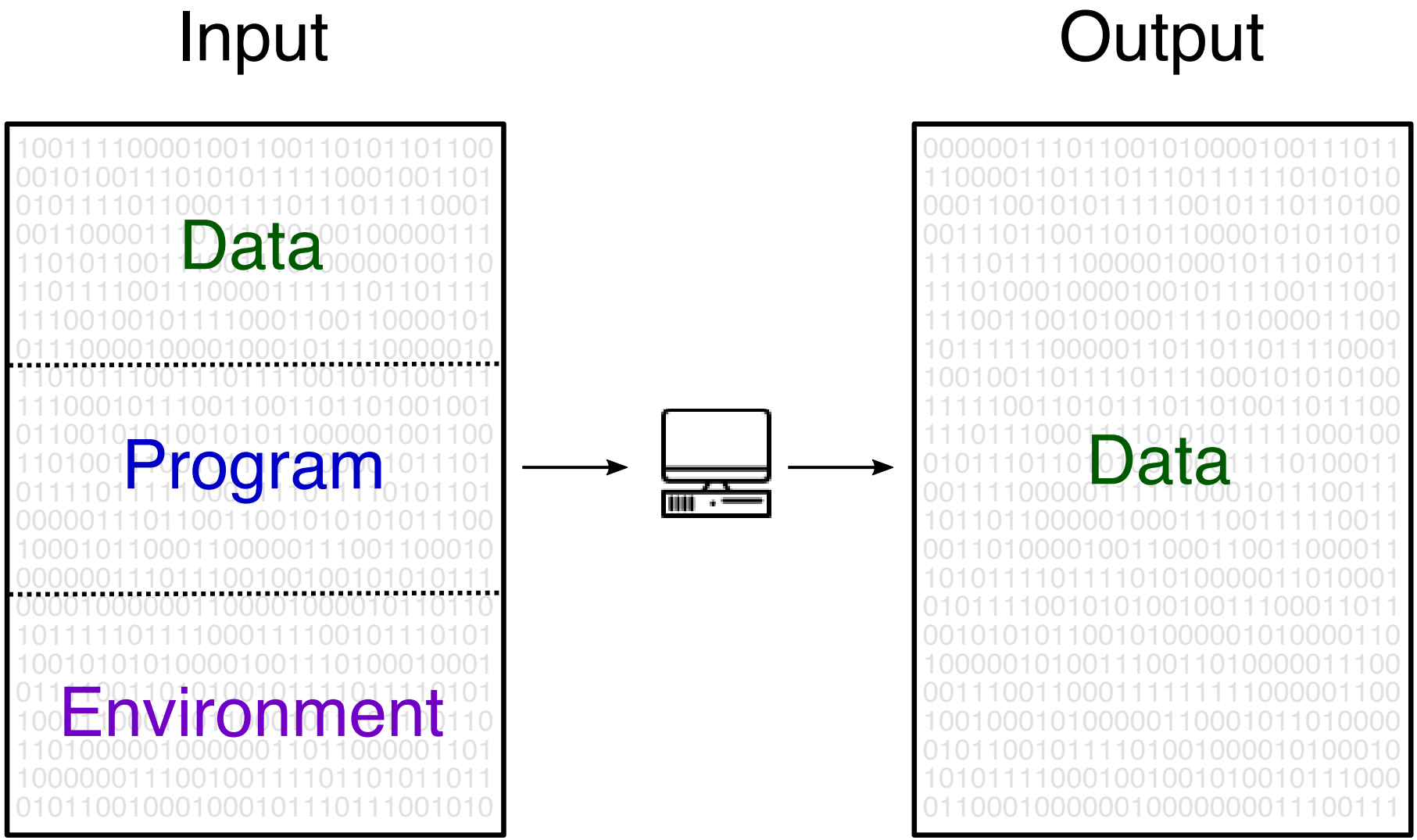
TURING AWARD LECTURE

Reflections on Trusting Trust

To what extent should one trust a statement that a program is free of Trojan horses? Perhaps it is more important to trust the people who wrote the software.

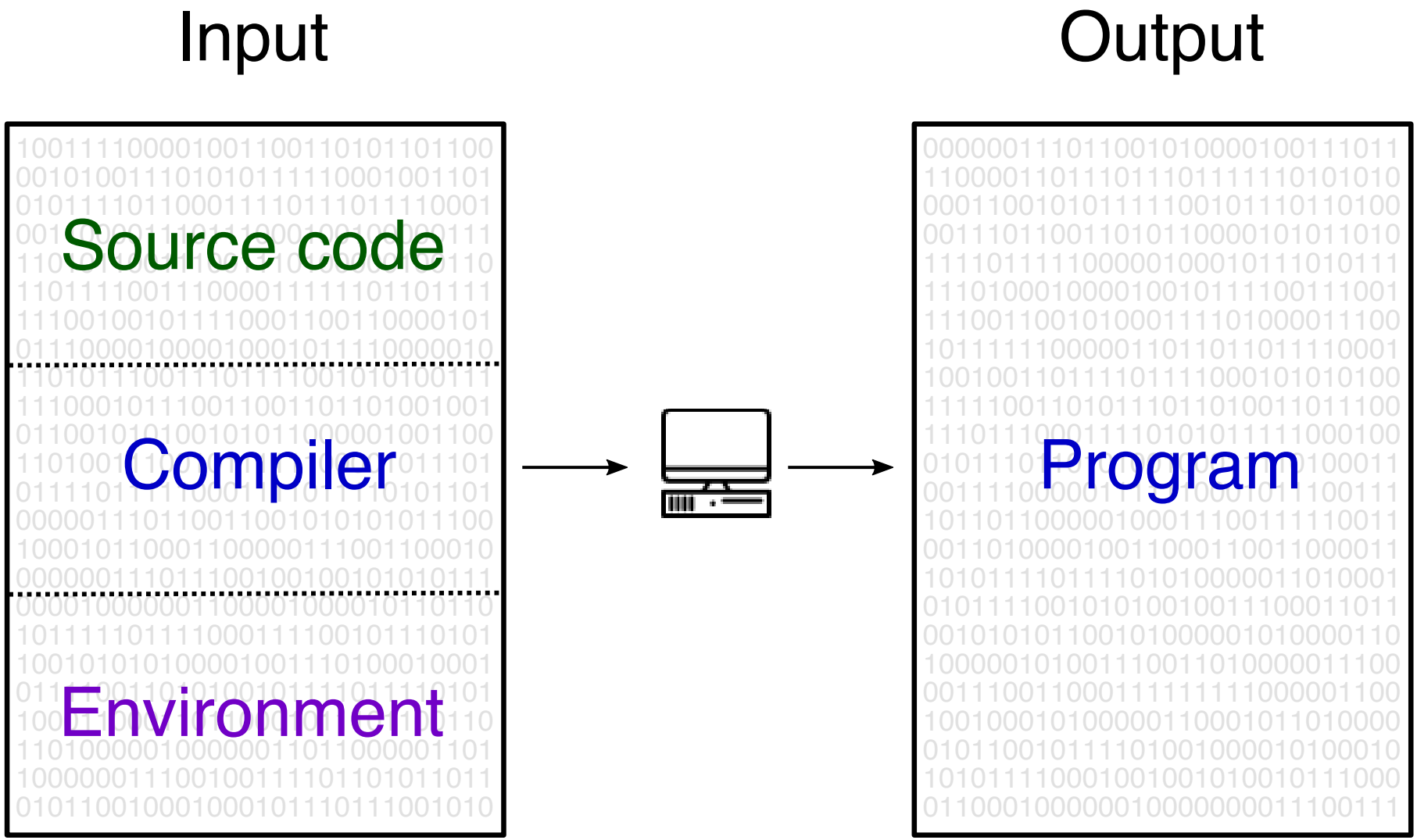
KEN THOMPSON

C’est quoi un calcul ?



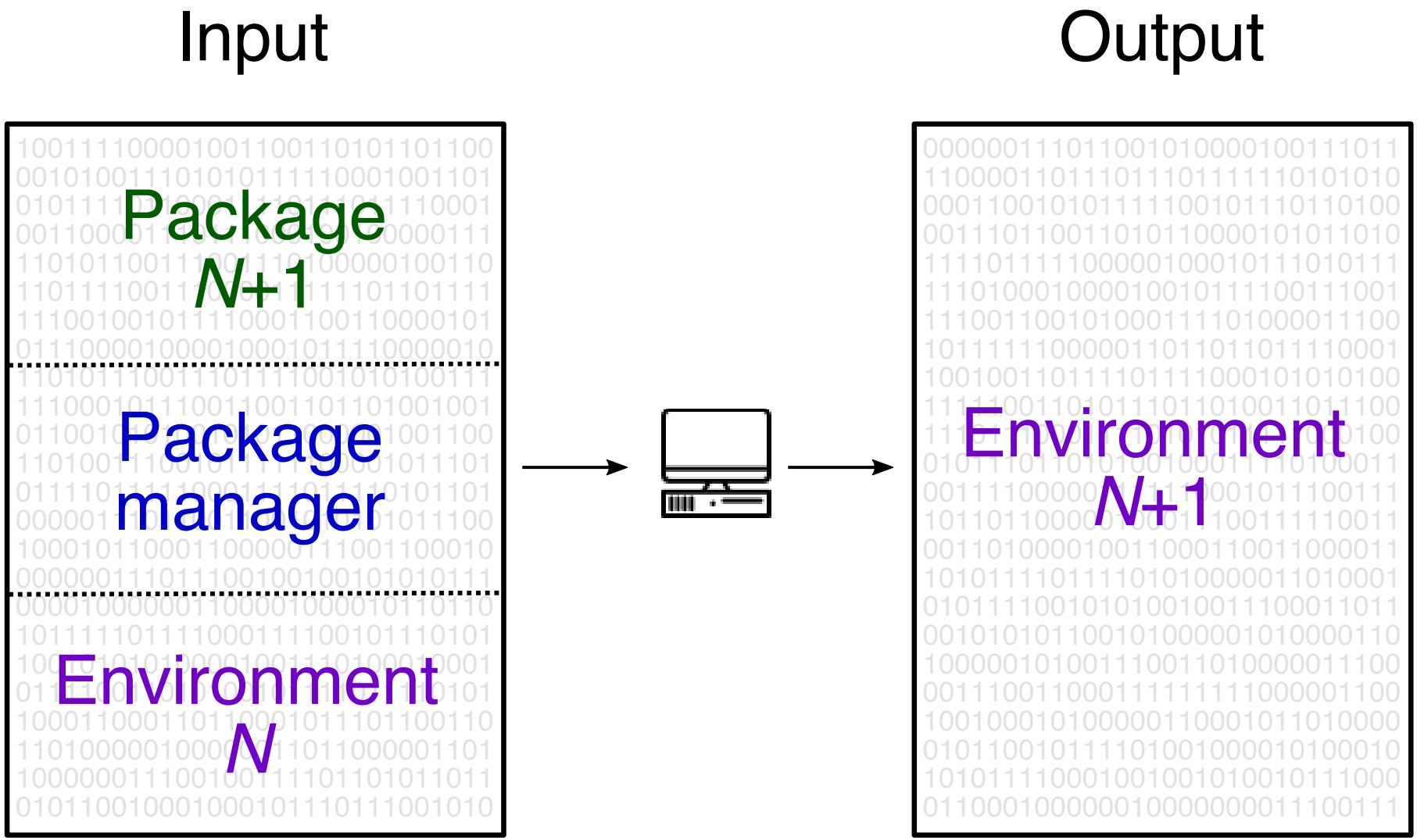
Computer by Creative Stall from the Noun Project

D’où vient le programme ?



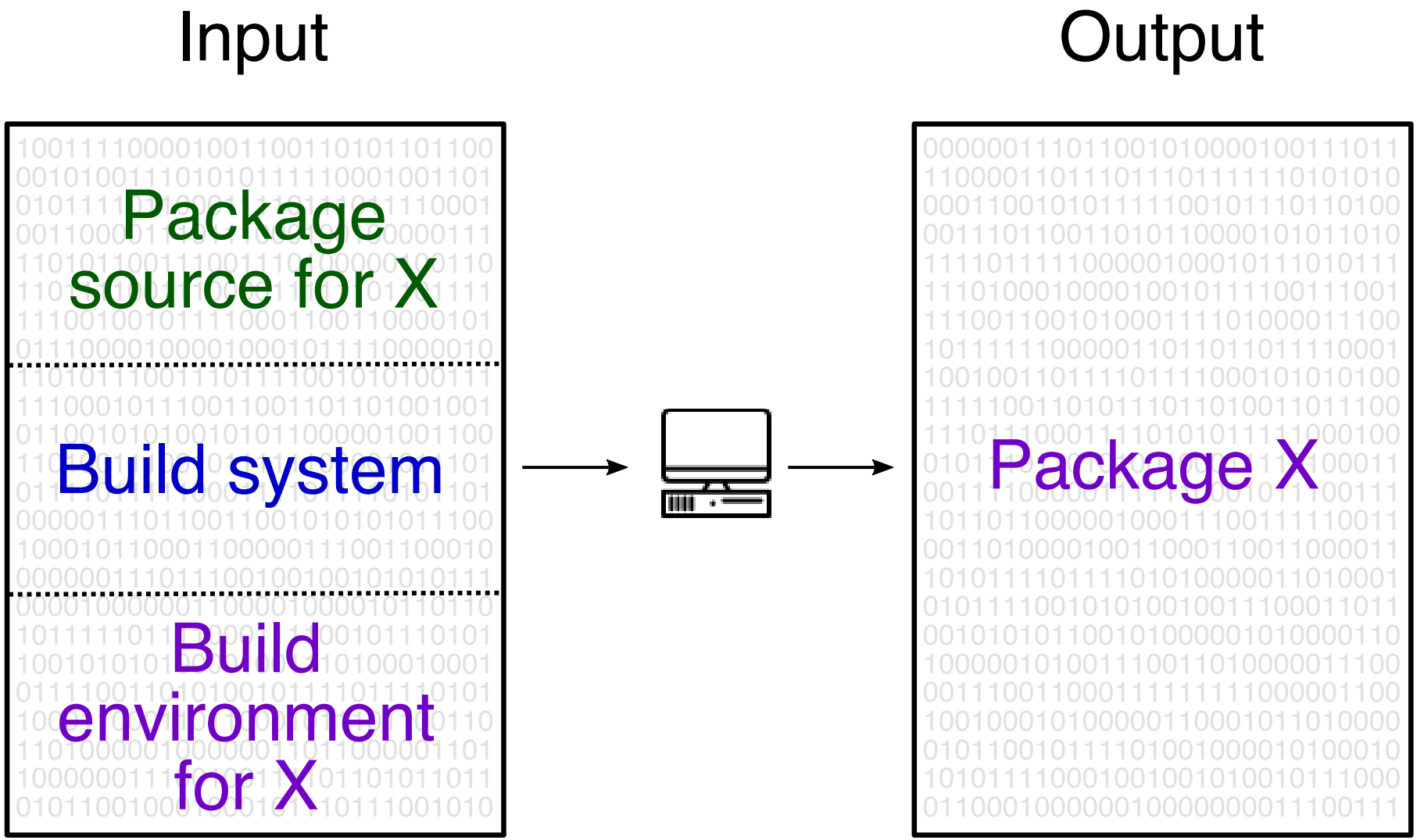
Computer by Creative Stall from the Noun Project

Et d'où vient l'environnement ?



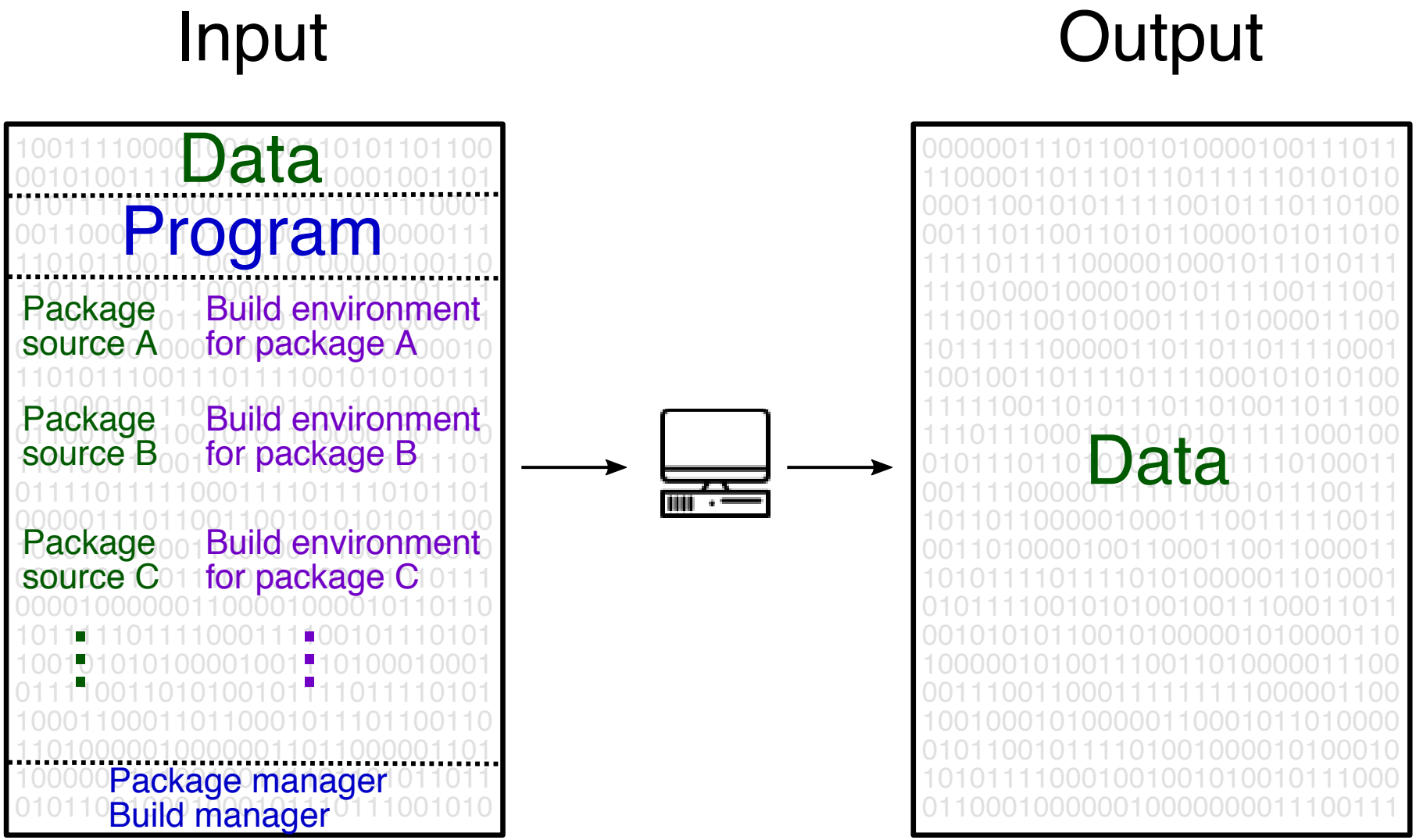
Computer by Creative Stall from the Noun Project

Bon, alors, d'où viennent les paquets ?



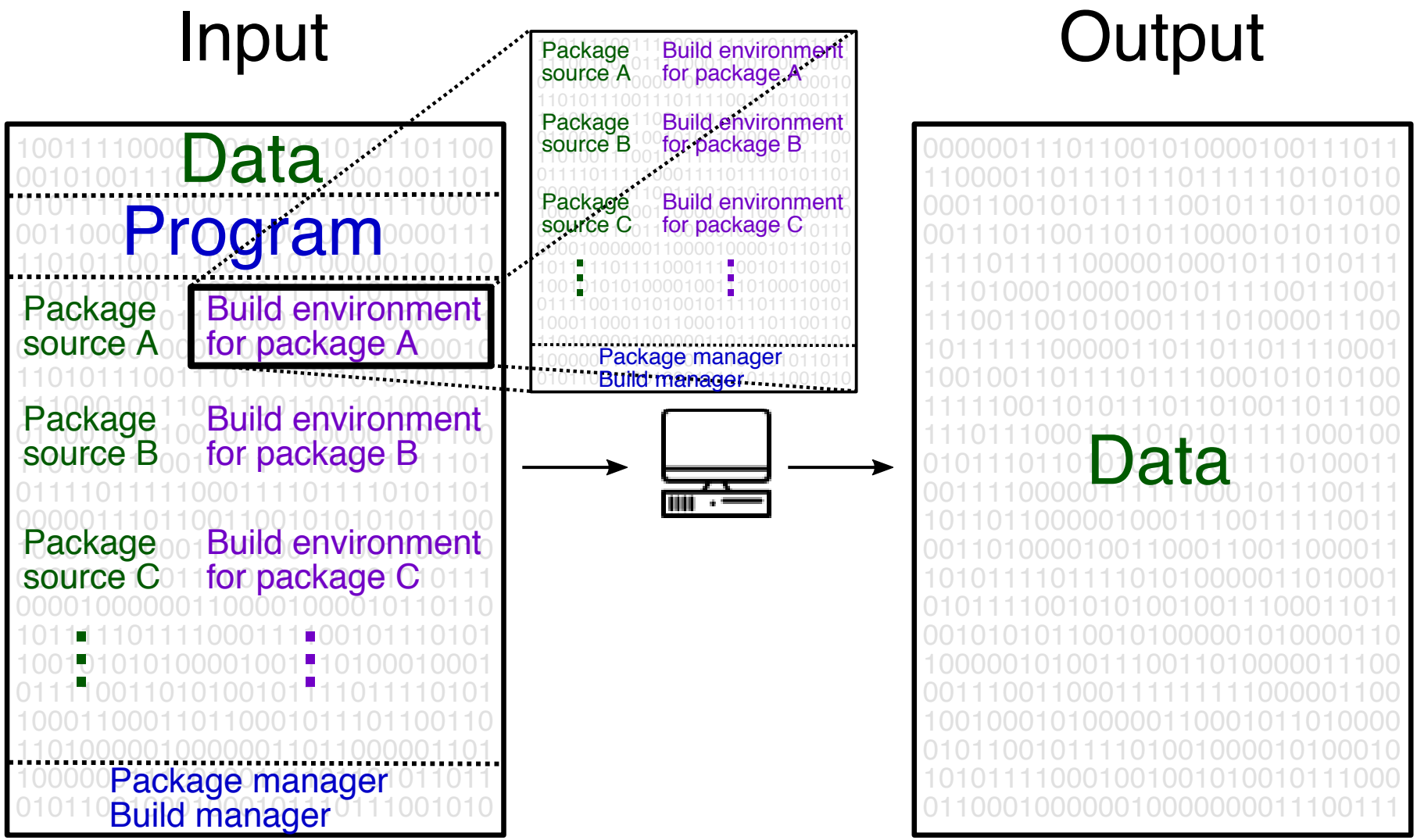
Computer by Creative Stall from the Noun Project

En résumé...



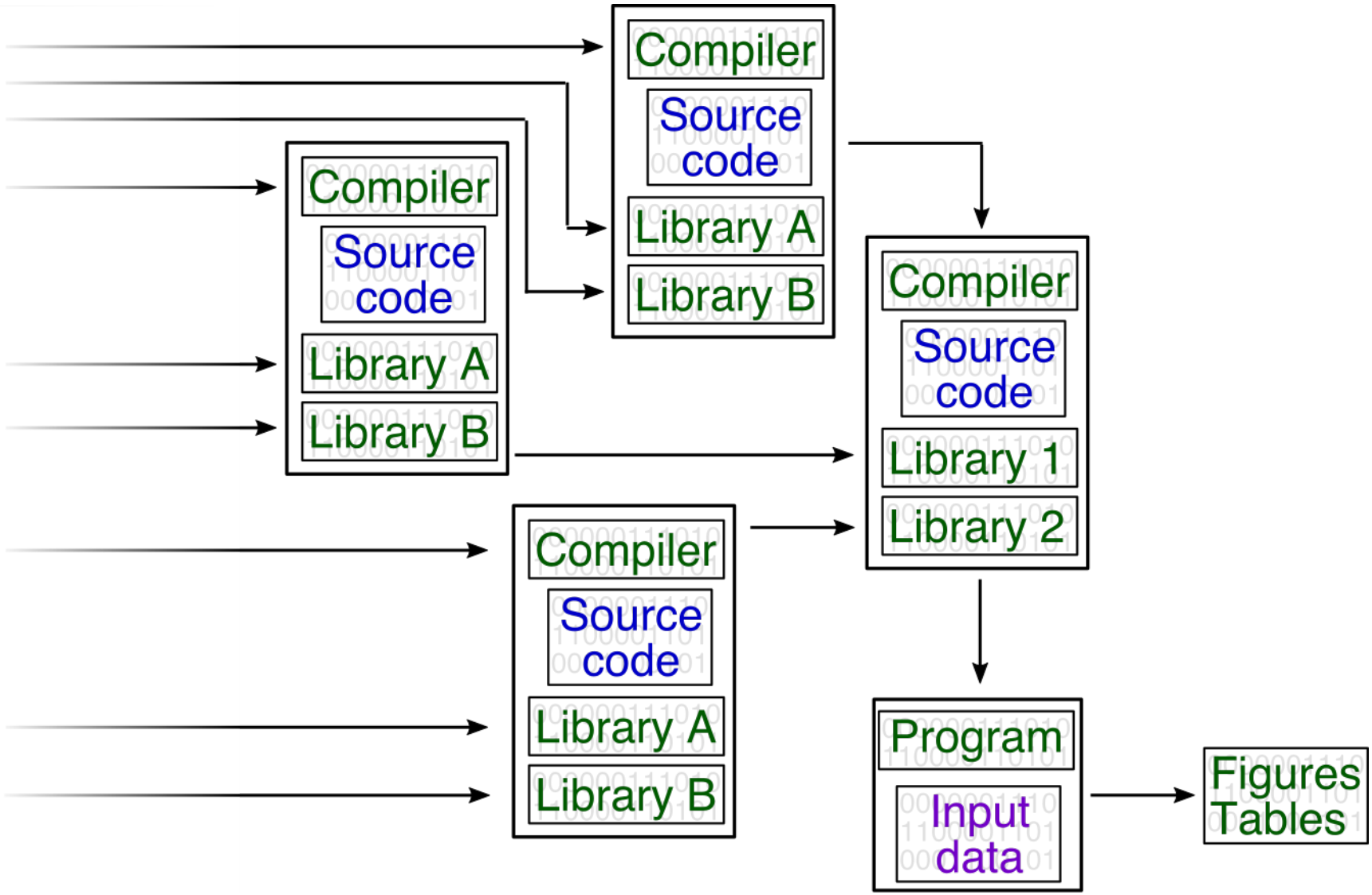
Computer by Creative Stall from the Noun Project

Sans oublier...



Computer by Creative Stall from the Noun Project

Un autre point de vue



La provenance des données numériques

Observations, expériences

- Source : instruments scientifiques
- Enregistrement avec des méta-données (date, instrument, ...)

Apports humains

- Modèles, codes source, paramètres, ...
- Traçables par la gestion de versions

Résultats de calcul

- Données scientifiques mais aussi logiciels en binaire
- Reproductibles à partir des entrées

Les gestionnaires de paquets classiques

Linux : dpkg/apt, RPM, pacman, ...
macOS : Homebrew, MacPorts, Fink
Multi-plateforme : conda, Spack, ...

- L'environnement est découpé en "paquets" qui sont assemblés sur un serveur puis téléchargés pour l'installation.
- On perd la trace des environnements de construction des paquets individuels.

L'environnement n'est pas reproductible à partir du code source
...sans mesures complémentaires

Machines virtuelles, conteneurs

Machines virtuelles : QEMU, VirtualBox, VMWare, JVM, ...
Conteneurs : Docker, Singularity, ...

- Paquets gros-grain
- On perd la trace des environnements de construction.
- Il faut assurer la reproductibilité par d'autres moyens.

L'environnement n'est pas reproductible à partir du code source

...sans mesures complémentaires

Comment faire mieux ?

- Nix, Guix $\longrightarrow$ Ludovic Courtès
- Reproducible Builds

  “a set of software development practices that create an independently-verifiable path from source to binary code”

Les trois défis de la reproductibilité numérique

- L'arithmétique à virgule flottante
- Le calcul parallèle
- Comment réconcilier la manipulation interactive avec la reproductibilité ?

L’arithmétique à virgule flottante

- Standardisé en 1985 norme IEEE 754
- Universellement acceptée aujourd’hui.
- Ses opérations sont précisément spécifiées...
- ...et parfaitement déterministes.
- Donc : aucune particularité pour la reproductibilité !

L’arithmétique à virgule flottante

- Standardisé en 1985 norme IEEE 754
- Universellement acceptée aujourd’hui.
- Ses opérations sont précisément spécifiées...
- ...et parfaitement déterministes.
- Donc : aucune particularité pour la reproductibilité !
- Aucun langage de programmation donne un accès direct aux opérations IEEE 754.
- Les optimisations modifient les résultats.
- Le programmeur n’a pas le contrôle complet sur les résultats.
- Il devient obligatoire d’intégrer la compilation dans le protocole de reproductibilité : même compilateur, mêmes options.

Le calcul parallèle

Pacte avec le Diable :

- calcul plus rapide
- perte de contrôle sur l'ordre des opération
  $\longrightarrow$ perte de reproductibilité

Pas de vraie solution,
on ne peut que réduire l'impact des aléas.
Exemple : https ://bebop.cs.berkeley.edu/reproblas/

Chaque bit compte !

- Reproductibilité = résultat égal.
- Pas d'exception pour l'arithmétique à virgule flottante.
- Au niveau des bits, le "presque" n'existe pas.
- Le "suffisament proche" relève de la validation / réplicabilité : c'est un jugement scientifique

L’avenir du calcul reproductible

Source : PinClipart.com

L’avenir du calcul reproductible

La reproductibilité du calcul sera assurée automatiquement par l’ordinateur.
(sauf pour le calcul parallèle et le travail interactif)

Technologies qui rendent les données numériques traçables :

- content-addressed storage
- blockchain

Déjà utilisés par :

- Git, Software Heritage
- Nix, Guix
- IPFS / IPLD

Au-delà de la reproductibilité

Que fait ce programme parfaitement reproductible ?

```
#define _ F-->00 || F-OO--;
long F=00,OO=00;
main(){F_OO();printf("%1.3f\n", 4.*-F/OO/OO);}F_OO()
{
             _-_-_-_
        _-_-_-_-_-_-_-_-_
     _-_-_-_-_-_-_-_-_-_-_-_
   _-_-_-_-_-_-_-_-_-_-_-_-_-_
  _-_-_-_-_-_-_-_-_-_-_-_-_-_-_
  _-_-_-_-_-_-_-_-_-_-_-_-_-_-_
_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_
_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_
_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_
_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_-_
  _-_-_-_-_-_-_-_-_-_-_-_-_-_-_
  _-_-_-_-_-_-_-_-_-_-_-_-_-_-_
   _-_-_-_-_-_-_-_-_-_-_-_-_-_
     _-_-_-_-_-_-_-_-_-_-_-_
        _-_-_-_-_-_-_-_-_
             _-_-_-_
}
```

International Obfuscated C Code Contest, 1988

Résumé

- La reproductibilité est une source de confiance dans un résultat.

Résumé

- La reproductibilité est une source de confiance dans un résultat.
- Il y en a d'autres, complémentaires.

Résumé

- La reproductibilité est une source de confiance dans un résultat.
- Il y en a d'autres, complémentaires.
- Pour le calcul, elle sera assurée automatiquement...

Résumé

- La reproductibilité est une source de confiance dans un résultat.
- Il y en a d'autres, complémentaires.
- Pour le calcul, elle sera assurée automatiquement...
- ... si nous faisons l'effort pour y arriver.

Résumé

- La reproductibilité est une source de confiance dans un résultat.
- Il y en a d'autres, complémentaires.
- Pour le calcul, elle sera assurée automatiquement...
- ... si nous faisons l'effort pour y arriver.
- Deux défis se posent par la suite :
  - le travail interactif
  - la compréhensibilité

Résumé

- La reproductibilité est une source de confiance dans un résultat.
- Il y en a d'autres, complémentaires.
- Pour le calcul, elle sera assurée automatiquement...
- ... si nous faisons l'effort pour y arriver.
- Deux défis se posent par la suite :
  - le travail interactif
  - la compréhensibilité
- Chercher le salut dans les artefacts binaires est sans espoir,

Résumé

- La reproductibilité est une source de confiance dans un résultat.
- Il y en a d'autres, complémentaires.
- Pour le calcul, elle sera assurée automatiquement...
- ... si nous faisons l'effort pour y arriver.
- Deux défis se posent par la suite :
  - le travail interactif
  - la compréhensibilité
- Chercher le salut dans les artefacts binaires est sans espoir,
- ...mais ils peuvent être des outils efficaces dans le cadre d'une approche plus large.