

Vers la sobriété numérique

Denis Trystram
LIG-Inria/DataMove
Denis.Trystram@univ-grenoble-alpes.fr

Journée de l'axe WAX – mai, 2022



L'hypothèse de base

- Les émissions de GES augmentent à un rythme sans précédent entraînant un important réchauffement climatique
 - C'est le symptôme d'une activité humaine incontrôlée
 - A un niveau ou un autre, tout le monde en a conscience¹ cela se voit dans les éco-gestes, dans les urnes, dans l'angoisse des gens, etc.
-
- La situation *impose* le HPC pour résoudre des problèmes complexes et l'IA pour prendre les bonnes décisions...

¹ou presque...

What's wrong?

Le numérique est en forte croissance, en particulier le HPC et l'IA.

- Comment prendre en compte cette dissonance dans nos vies professionnelles ? Spécialement dans un labo d'informatique...
- Témoignage des étapes de mon parcours
 - Prise de conscience
 - En adaptant l'objet de ses recherches
 - On passe à l'action
 - ...

La solution facile a été de partir d'une reconnaissance en optimisation garantie à faible coût et multi-objectifs...

- Un premier projet ANR en 2013
- Un second en 2018, puis un projet européen

Au mieux, le *numérique* ne peut qu'*optimiser*, mais rien remettre en cause fondamentalement.

L'envie d'y croire et le tremplin de la chaire MIAI *edge intelligence* :

- Définir l'apprentissage (distribué) frugal et le promouvoir
- Accompagner des entreprises motivées
- Faire de la médiation, lancer des alertes

Quel(s) effet(s) ?

Un problème d'apprentissage fédéré distribué

On se donne un ensemble convexe $\mathcal{K} \subseteq \mathbb{R}^d$ et un ensemble d'agents connectés via un réseau représenté par un graphe $G = (V, E)$ où $n = |V|$ est le nombre d'agents.

A chaque pas de temps $1 \leq t \leq T$, chaque agent $i \in V$ peut communiquer avec ses voisins immédiats dans G et prendre une décision $x_i^t \in \mathcal{K}$.

En conséquence, un batch de nouvelles données est révélé exclusivement à l'agent i et dans ce batch, on introduit localement une fonction de coût non convexe $f_i^t : \mathcal{K} \rightarrow \mathbb{R}$

Bien que l'agent i observe uniquement la fonction f_i^t , il est intéressé par le coût cumulé $F^t(\cdot)$ où $F^t(\cdot) := \frac{1}{n} \sum_{j=1}^n f_j^t(\cdot)$.

En particulier, au temps t , le coût est de i est x_i^t is $F^t(x_i^t)$.
L'objectif de chaque agent i est de minimiser le temps cumulé $\sum_{t=1}^T F^t(x_i^t)$.

Lorsque les fonctions de coût f_i^t sont convexes, une mesure standard est le *regret*.

Un algorithme on-line est $R(T)$ -*regret* si pour tout agent $1 \leq i \leq n$,

$$\frac{1}{T} \left(\sum_{t=1}^T F^t(x_i^t) - \min_{o \in \mathcal{X}} \sum_{t=1}^T F^t(o) \right) \leq R(T)$$

STOP !

Je peux vous en parler encore longtemps...

- C'est ce que je sais faire :
continuer comme avant en s'arrangeant avec sa conscience
- Par exemple : c'est mieux de travailler sur la logistique des vélos en centre ville que pour Total Energies



Je viens d'illustrer une approche *en rupture*... On arrête de faire comme avant et on change les façons de faire

La symphonie finale

- Je pensais être chef d'orchestre...
- Je n'étais que la baguette.

Conclusion : pas de rupture radicale mais des chemins de traverse

- Utiliser mes compétences et ma position pour convaincre
- Me rétracter dans la théorie
- Elargir mon champ multi-disciplinaire : la convivialité
- Sortir de la tour d'ivoire universitaire

Edge computing (et apprentissage embarqué)

L'idée de base

- Traiter localement les données, proche de là où elles sont produites².
- Faire l'inférence en local par federated/distributed learning, parfois aussi l'entraînement, voire *on-device*
- La meilleure pratique est toujours de se restreindre aux données les plus utiles et d'évaluer les bénéfices potentiels d'une telle approche avant de la développer !

Notons qu'il n'est pas si clair que ces approches soient vraiment plus vertueuses, il conviendrait de le prouver...

²<http://edge-intelligence@imag.fr>

Edge computing (et apprentissage embarqué)

L'idée de base

- Traiter localement les données, proche de là où elles sont produites².
- Faire l'inférence en local par federated/distributed learning, parfois aussi l'entraînement, voire *on-device*
- La meilleure pratique est toujours de se restreindre aux données les plus utiles et d'évaluer les bénéfices potentiels d'une telle approche avant de la développer !

Notons qu'il n'est pas si clair que ces approches soient vraiment plus vertueuses, **il conviendrait de le prouver...**

²<http://edge-intelligence@imag.fr>

Comment établir une recherche "vertueuse" ?

- Le thème de l'impact des technologies digitales sur le climat se diffuse rapidement dans la société.
Mais la prise de conscience est plus lente dans le secteur professionnel.
Elle est même un peu contradictoire dans les entreprises dont l'objectif premier est la production (et la rentabilité)...

Quel que soit le scénario...

Quel que soit le scénario...

Prévisions IoT par secteurs (1)

Prévisions par régions (2)

Prévisions IoT par secteurs (1)

Prévisions par régions (2)

Quel que soit le scénario...

Quel que soit le scénario...

Prévisions IoT par secteurs (1)

Prévisions par régions (2)

Historiquement, le edge était de déporter des traitements du cloud

Aujourd'hui, les systèmes edge embarquent stockage, traitements, énergie.

"The promise of the Industrial Internet includes significant investment in the coming years. Management consulting firm, McKinsey, estimates that the Industrial IoT will create 7.5 Tdollars in value by 2025

The Industrial IoT brings together minds and machines connecting people to machine data that accelerate digital industrial transformation."

Evolutions directes du marché Capteurs et IoT

source IEEE Spectrum:

<https://www.marketsandmarkets.com/practices/sensors-control?>

Evolutions périphériques du marché Capteurs et IoT

Prévisions du marché

Prévisions du marché (par secteurs)

Prévisions du marché (par pays)

Conséquence(s) de cette évolution

Faut-il vraiment s'en réjouir ?

Le développement toujours plus massif de l'IoT pose la question de l'alimentation en énergie de ces objets connectés et nomades, capables de faire des calculs, de recevoir et/ou transmettre des informations, et que l'on veut également pouvoir rendre autonomes en énergie.

IoT et AI sont étroitement liés

- Pour résumer : Un marché en plein essor !
MAIS, la croissance, toujours la croissance...

Comment justifier le rapprochement des deux notions ?

- La vision de IIA sur gros calculateur nourrit le fantasme de BigBrother
Déporter les calculs, c'est s'en affranchir...
- Mettre de l'intelligence dans les objets
- Mettre en place des nouveaux paradigmes

IA nest pas un marché en soi, cest tentaculaire et cela ruisselle dans tous les secteurs !

S'acheter une bonne conscience

On distingue deux positions extrêmes

- ceux qui pensent que la High-tech et l'informatique pourra sauver la planète.
 - Par exemple, via des dispositifs IoT ou des algorithmes qui permettent de changer les pratiques et de se tourner vers l'écologie...
 - Est-il possible de construire un IoT sobre et comment faire la part entre les "bons" IoT et les "mauvais" ?
- Les alarmistes et techno sceptiques

Edge computing (et apprentissage embarqué)

L'idée de base

- Traiter localement les données, proche de là où elles sont produites³.
- Faire l'inférence en local par federated/distributed learning, parfois aussi l'entraînement, voire *on-device*
- La meilleure pratique est toujours de se restreindre aux données les plus utiles et d'évaluer les bénéfices potentiels d'une telle approche avant de la développer !

Notons qu'il n'est pas si clair que ces approches soient vraiment plus vertueuses, il conviendrait de le prouver...

³<http://edge-intelligence@imag.fr>

Edge computing (et apprentissage embarqué)

L'idée de base

- Traiter localement les données, proche de là où elles sont produites³.
- Faire l'inférence en local par federated/distributed learning, parfois aussi l'entraînement, voire *on-device*
- La meilleure pratique est toujours de se restreindre aux données les plus utiles et d'évaluer les bénéfices potentiels d'une telle approche avant de la développer !

Notons qu'il n'est pas si clair que ces approches soient vraiment plus vertueuses, **il conviendrait de le prouver...**

³<http://edge-intelligence@imag.fr>

Comment établir une recherche "vertueuse" ?

- Le thème de l'impact des technologies digitales sur le climat se diffuse rapidement dans la société.
Mais la prise de conscience est plus lente dans le secteur professionnel.
Elle est même un peu contradictoire dans les entreprises dont l'objectif premier est la production (et la rentabilité)...

Approches possibles

- 1 On optimise, et on continue en ciblant d'autres critères
- 2 Ajuster l'objet de nos recherche pour être moins en dissonance
- 3 Passer à l'action en rupture⁴

Le changement de comportement pourrait faire changer l'offre

C'est possible ! C'est là que se situe le levier.

Exempleä: l'application *Yuka*, adoptée par près d'un quart de la population française, qui a eu un effet sur la grande distribution...

⁴on peut aussi ne rien faire du tout !

Une autre façon d'envisager les choses

- L'approche scientifique classique se situe vers les avancées technologiques et les innovations.
- Elle suppose un accès illimité et pas cher aux ressources et à l'énergie

La convivialité de [Illich 1973]

"il est grand temps de **conduire une recherche à contre courant** sur la possibilité d'utiliser la technologie afin de créer des institutions au service des interactions personnelles, créatrices et autonomes et de permettre l'apparition de valeurs qui ne puissent pas être soumises aux règles des technocrates."

Une nouvelle perspective

- **Soutenabilité** sans effets rebonds ou induits, fiable, avec des ressources disponibles.
- **Equité** bien réparti dans les couches de la société qui génère de la stabilité
- **Résilience** capable de s'adapter aux situations de crises et catastrophes.
- **Utilité** démontrée des fruits de la recherche.
- **Ecologique** faible empreinte environnementale.
- **Maitrise** des outils, réparabilité, accroissement de la durée de vie, etc.

Purement fictif, imaginez le lancement d'un gros programme ANR.

Voiture autonome

- Doit-on candidater à ce programme ?
- Est-ce utile ou indispensable ?
- Quels besoins en calculs faudra-t-il ?
- Que peut-on faire localement ?
- Comment évaluer les coûts complets ?
- Est-ce que les citoyens perdront le contrôle d'une techno trop complexe ?

- Le véhicule autonome est essentiellement un petit data centre sur roues
- Le edge computing est au centre du développement
- GE Digital avec son partenaire Intel estiment qu'il faut embarquer plusieurs centaines de capteurs intelligents correspondant à 40TBytes de données pour 8 heures de route.
- Cela fait un gros paquet de données !
C'est peu sûr, impossible et inutile de les transmettre au cloud, même avec la "5G and beyond".

On reprend les critères précédents sur une échelle de 0 à 1

- Vélo-cargo en ville
- Voitures électriques
- Drônes spécialisés

Des réactions ?