Des données qui changent la donne

Nicolas Tremblay, agr.

Ph.D., chercheur scientifique

St-Jean-sur-Richelieu

AAC - St-Jean-sur-Richelieu

- Plateforme d'expertise pour AAC en AP
- Embauches de scientifiques
- Réalisation d'un "atelier transformateur"
- Recherche à la ferme, intelligence artificielle
- NT: president *Int. Soc. Precision Agriculture*


14e Conférence internationale sur l’agriculture de précision
Montréal, 24 au 27 juin 2018

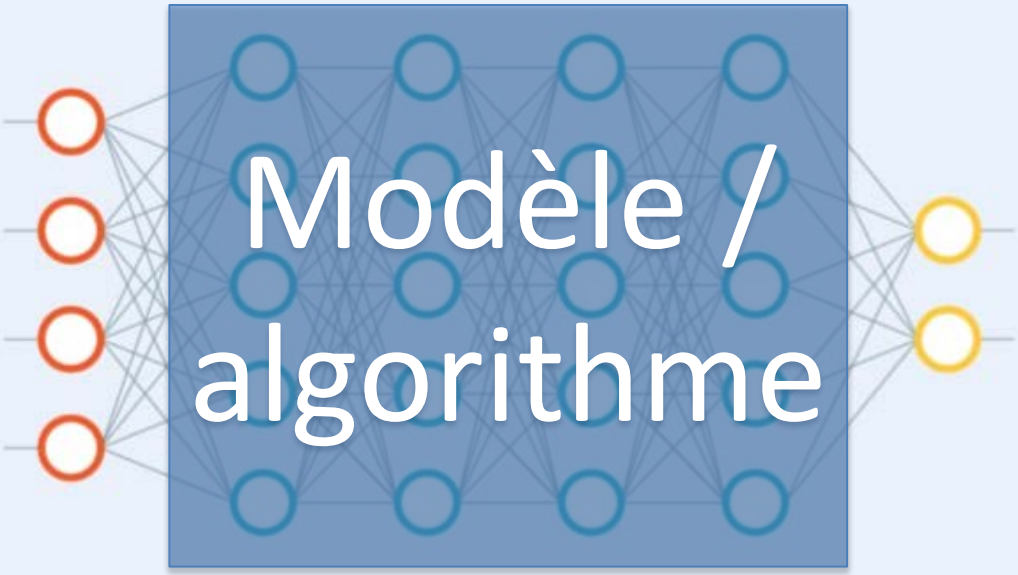

ISPA
International Society of Precision Agriculture
PRESENTER

INTELLIGENCE ARTIFICIELLE

Le «Nobel d'informatique» est décerné à Yoshua Bengio

Le professeur montréalais met la main sur le prix Turing avec deux de ses collègues.

Exemple de réseau de neurones profond



- Le modèle reçoit des informations de la part du programmeur : on lui présente la photo d'une personne.
- Le modèle crée ses propres connexions, qu'on appelle « couches cachées », sans intervention humaine : il apprend à reconnaître les composantes du visage.
- Le modèle arrive à une conclusion à partir des connexions qu'il a effectuées : il parvient à déterminer le nom de la personne apparaissant sur la photo.

Infographie *Le Devoir*

Humains, modèles et humilité

- L’esprit humain ne parvient pas à optimiser des décisions affectées par plusieurs dimensions
- Les décisions humaines sont menées par les émotions; pas tant par les faits
- Complexité↑; ordinateur meilleur que l’expert
- Mais sa logique ne peut pas être comprise
- Tout repose sur les données

WIRED MAGAZINE: 16.07

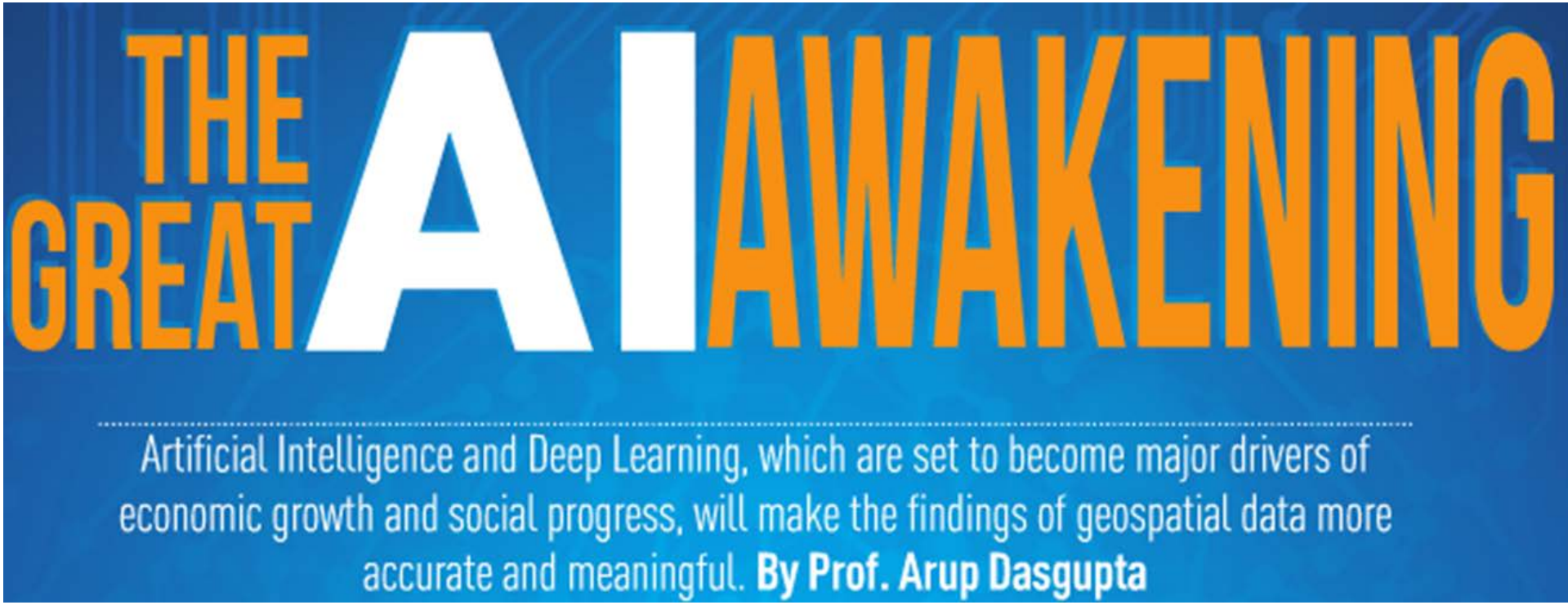
SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.



ET L'AGRICULTURE?

FROM ONE PIXEL TO ONE EARTH

The data platform is said to have analyzed over 2.8 quadrillion multispectral pixels for this. It enables processing at petabytes per day rates using multi-source data to produce calibrated, georeferenced imagery stacks at desired points in time and space that can be used for pixel level or global scale analysis or for visualizing or measure changes such as floods, or changes in the condition of crops.

There are more than 7 billion people on Earth now, and roughly one in eight people do not have enough to eat. According to the World Bank, the human population will hit an astounding 9 billion by 2050. With rapidly increasing population, the growing need for food is becoming a grave concern.

The burden is now on technology to make up for the looming food crises in the coming decades. But fortunately there is no shortage of ideas and innovative minds are seeking solutions to combat this problem.

Machine learning to the rescue

Descartes Labs, a Los Alamos, New Mexico-based start-up is using machine learning to analyze satellite imagery to predict food supplies months in advance of current methods employed by the US government, a technique that could help predict food crises before they happen.

Descartes Labs pulls images from public databases like NASA's Landsat and MODIS, ESA's Sentinel missions and other private satellite imagery providers, including Planet. It also keeps a check on Google Earth and Amazon Web Services public datasets. This continuous up-to-date imagery is referred to as the "Living Atlas of the Plant".

The commercial atlas, designed to provide real-time forecasts of commodity agriculture, uses decades of remotely sensed images stored on the Cloud to offer land use and land change analysis.

Descartes Labs cross-references the satellite information with other relevant data such as weather forecasts and prices of agricultural products. This data is then entered into the machine learning software, tracking and calculating future food supplies with amazing accuracy. By processing these images and data via their advanced machine learning algorithm, Descartes Labs collect remarkably in-depth information such as being able to distinguish individual crop fields and determining the specific field's crop by analyzing how the sun's light is reflecting off its surface. After the type of crop has been established, the machine learning program then monitors the field's production levels.

"With machine learning techniques, we look at tons of pixels from satellites, and that tells us what's growing," says Mark Johnson, CEO and Co-founder, Descartes Labs.

How to tackle a data deluge

The total database includes approximately a petabyte — or 10^{15} bytes — of data. Descartes has actually reprocessed the whole 40-year archive starting with the first Landsat satellite imagery to offer

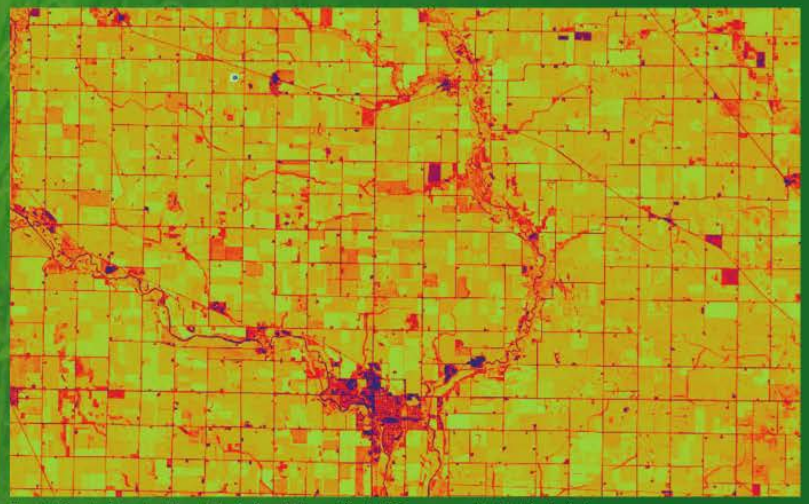
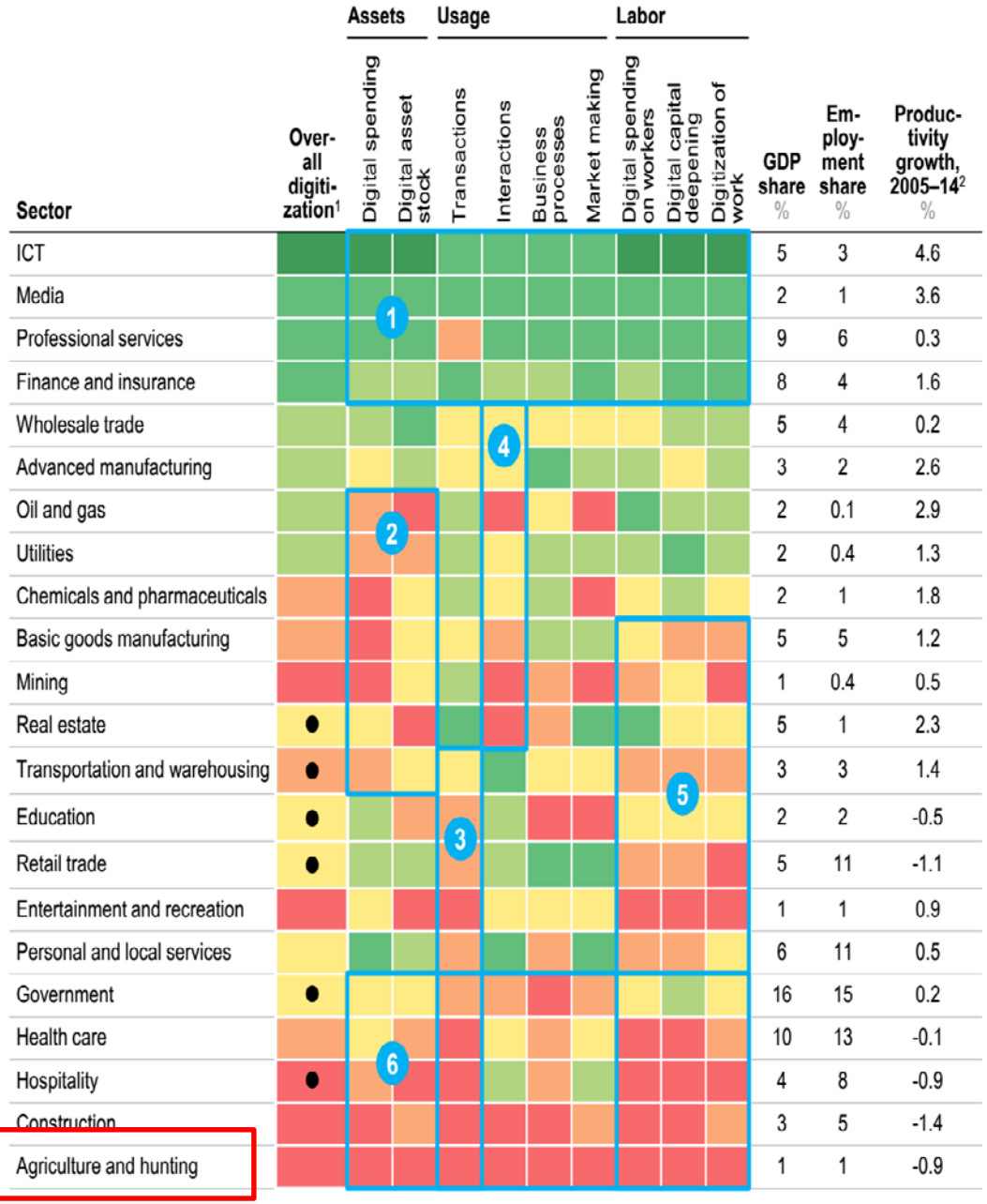


Figure1: A Descartes Labs estimate shows the health of corn fields around Humboldt, Iowa, as measured by their infrared reflectiveness

The MGI Industry Digitization Index

USA

2015 or latest available data



Relatively low digitization — Relatively high digitization

● Digital leaders within relatively undigitized sectors

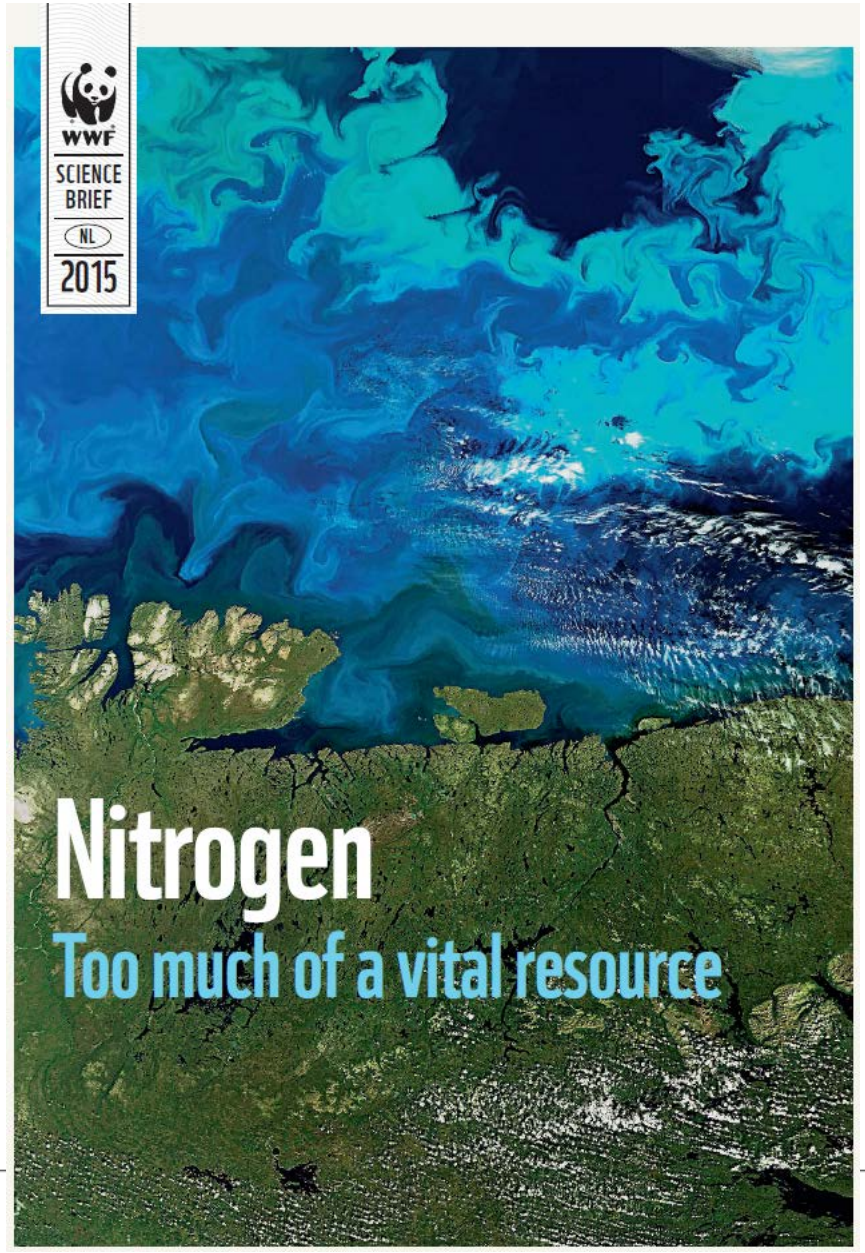
1. Knowledge-intensive sectors that are highly digitized across most dimensions
2. Capital-intensive sectors with the potential to further digitize their physical assets
4. B2B sectors with the potential to digitally engage and interact with their customers
5. Labor-intensive sectors with the potential to provide digital tools to their workforce
6. Quasi-public and/or highly localized sectors that lag across most dimensions

Options pour modèles

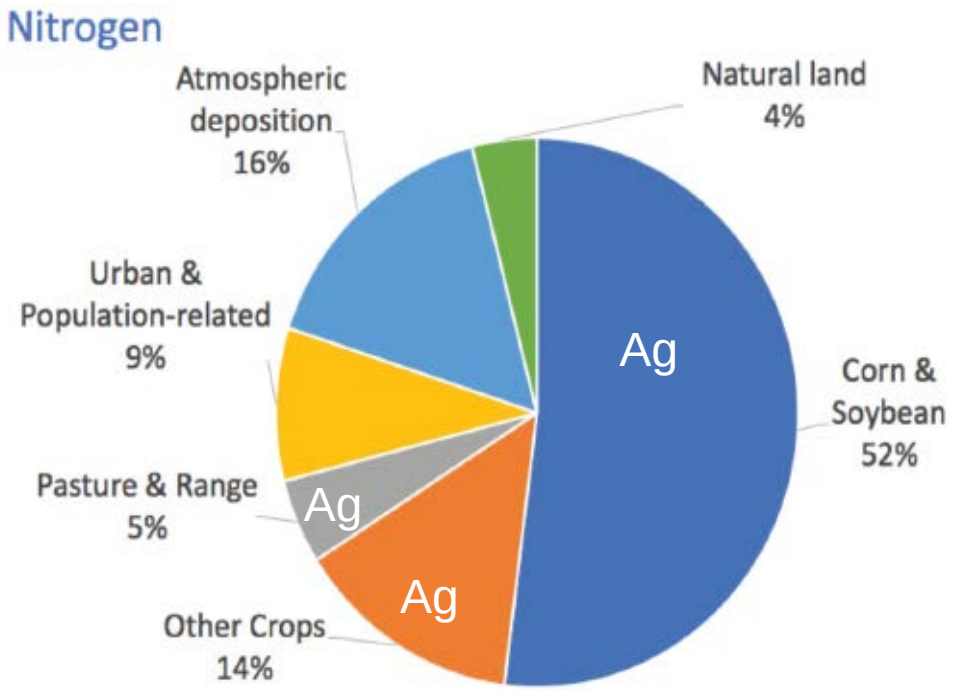
Simples ⟷ Panoplie d'algorithmes ⟷ **Compliqués**

- Régressions, factoriels
- Statistiques de Fisher
- Données limitées, contrôlées
- Domaine d'inférence limité
- Expert impliqué
- Communs en agriculture
- Peu performants

- Apprentissage profond
- Réseaux de neurones
- Données abondantes, diversifiées
- Domaine d'inférence large
- Expert nuisible
- Absents en agriculture
- Succès maximal

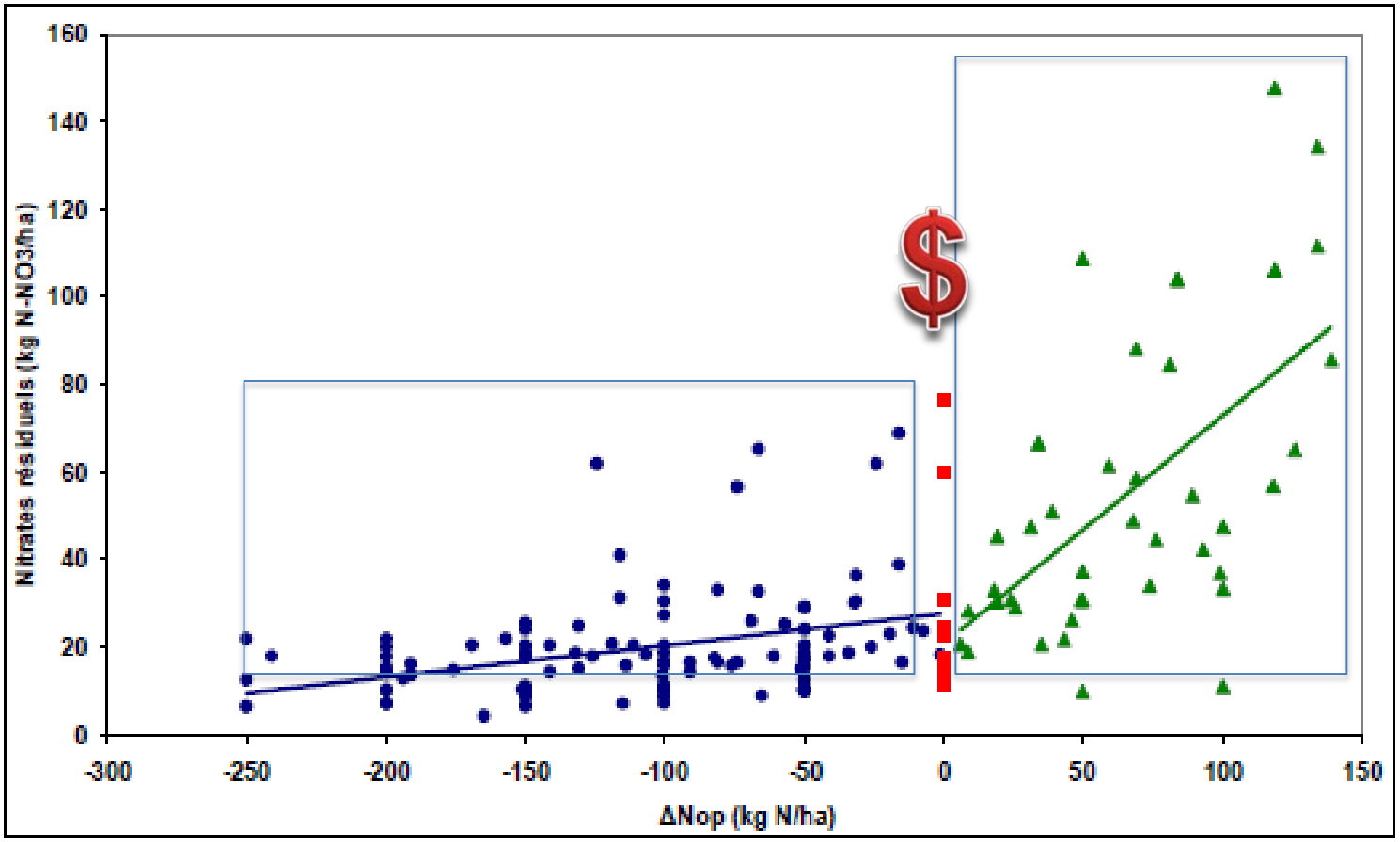


Fertilisation azotée



Estimated sources of N delivered to the northern Gulf of Mexico from the Mississippi River Basin (Alexander et al., 2008)

Cible: la dose optimale



Résultats du Québec (Giroux et al. 2009)

Résultats semblables pour N_2O (Van Groenigen et al. 2010)

Modèle industrie N_{opt} maïs

- *Modèle = kg N t^{-1} grains*
- *Simple…*
- *…mais inefficace, même ajustée par d'autres facteurs*
- *De plus en plus discréditée et retirée des systèmes d'aide à la décision aux É.-U.*

Lien RDT vs N_{opt}

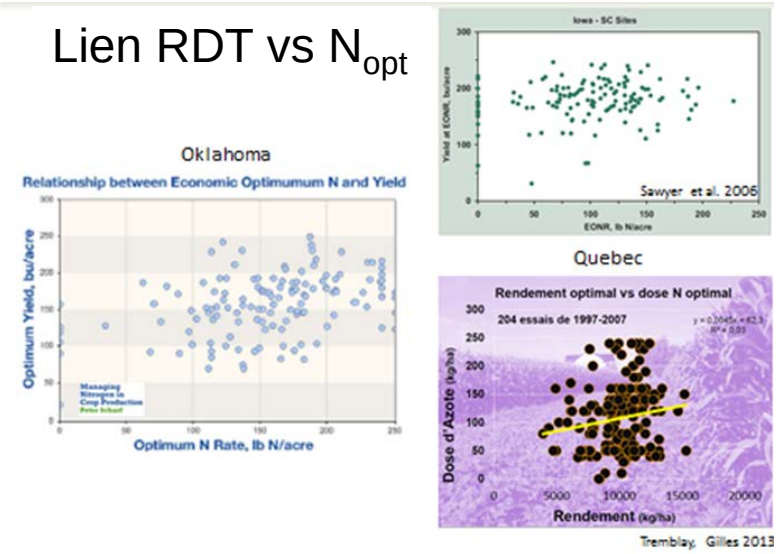
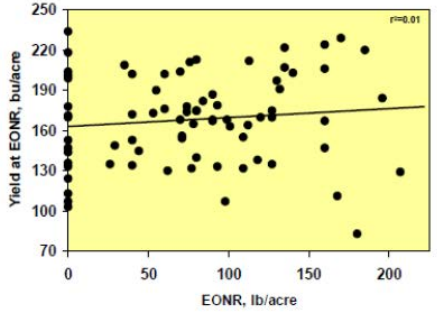
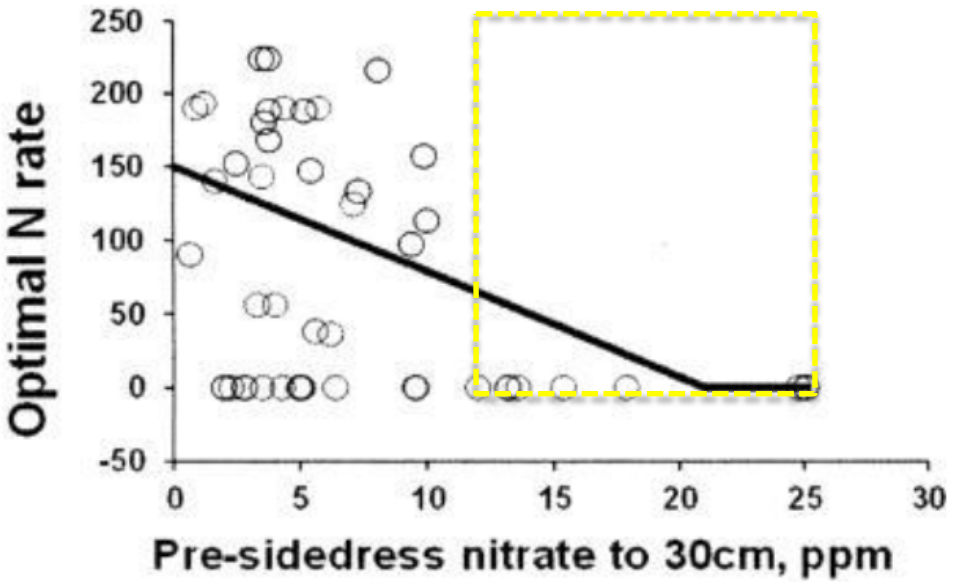


Figure 6. Relationship between economic optimum N rate (EONR) and corn yield from 77 sites in Wisconsin, 1992-2003.



Scharf N book 2015

N_{opt} maïs avec PSNT (simple)



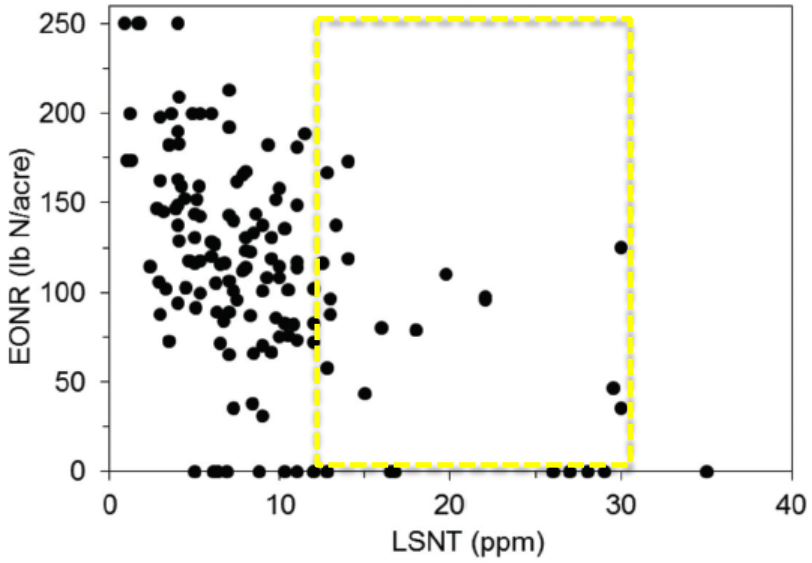


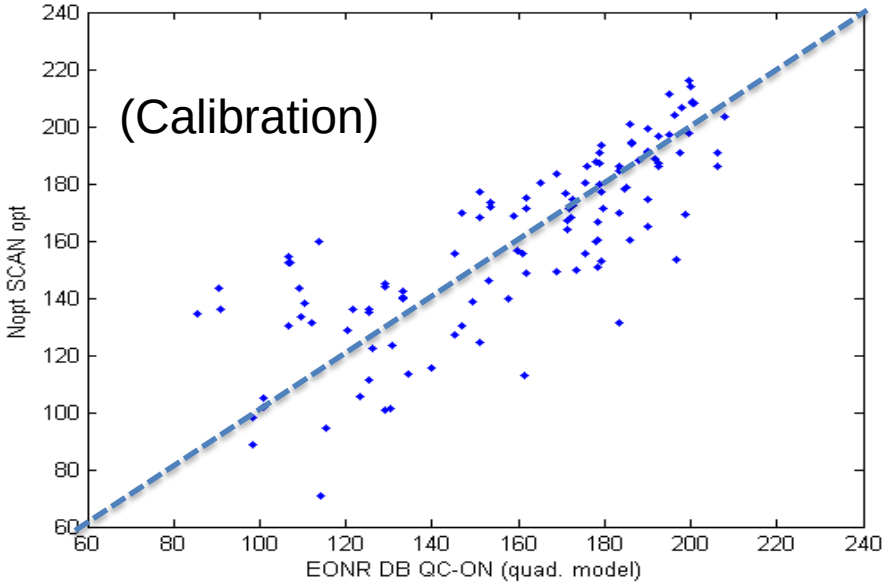
Figure 7. Relationship between the late spring soil nitrate test (LSNT) and the economic optimum N rate (EONR) across trials in multiple years with corn following soybean and continuous corn. Each point represents a trial presented in Figure 3 (J. Sawyer, 2015).

Morris et al. 2017. doi:10.2134/agronj2017.02.0112
https://dl.sciencesocieties.org/publications/aj/articles/110/1/1

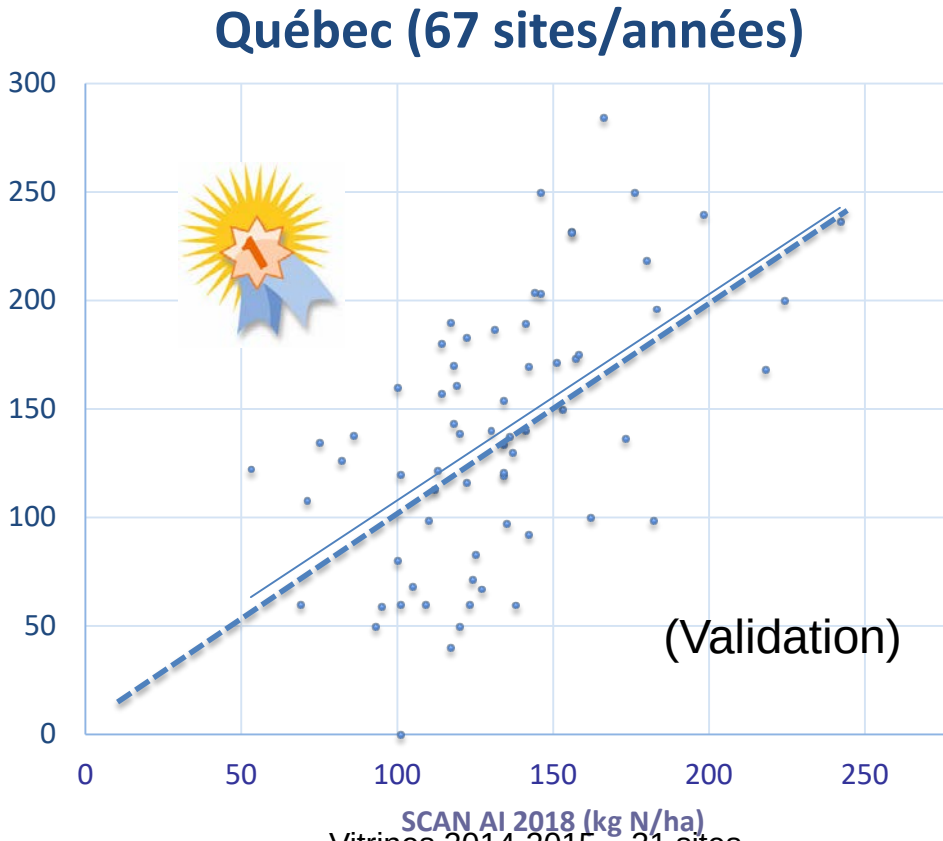
N_{opt} maïs avec SCAN AI (complexe)

Par méta-analyse

- Texture du sol
- Pluviométrie
- Culture précédente
- Matière organique
- + rapport économique



BD Québec-Ontario

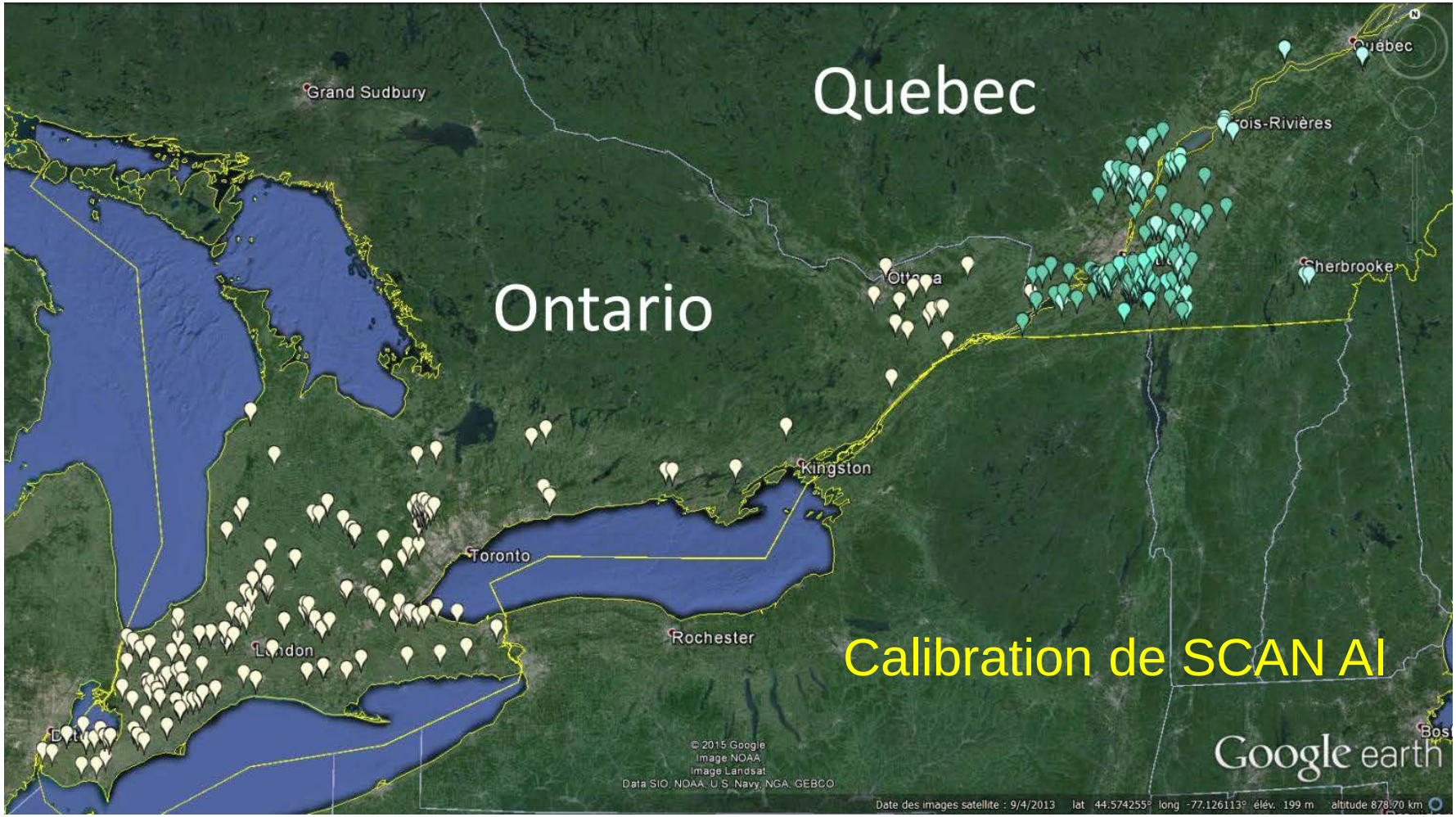


Vitrines 2014-2015 – 31 sites
Proconseil 2015-2017 – 18 sites
Pleine terre 2017 – 18 sites

SCAN AI

- Algorithme de logique floue (AI)
- Développé par AAC St-Jean-sur-Richelieu
- Contributions d'autres groupes de recherche et collaborateurs du milieu
- Meilleur modèle démontré et disponible
- Commercialisé
  - www.Fieldapex.com
  - Registre+ de Logiag
- Élaboré uniquement à partir de données

Beaucoup de données diversifiées



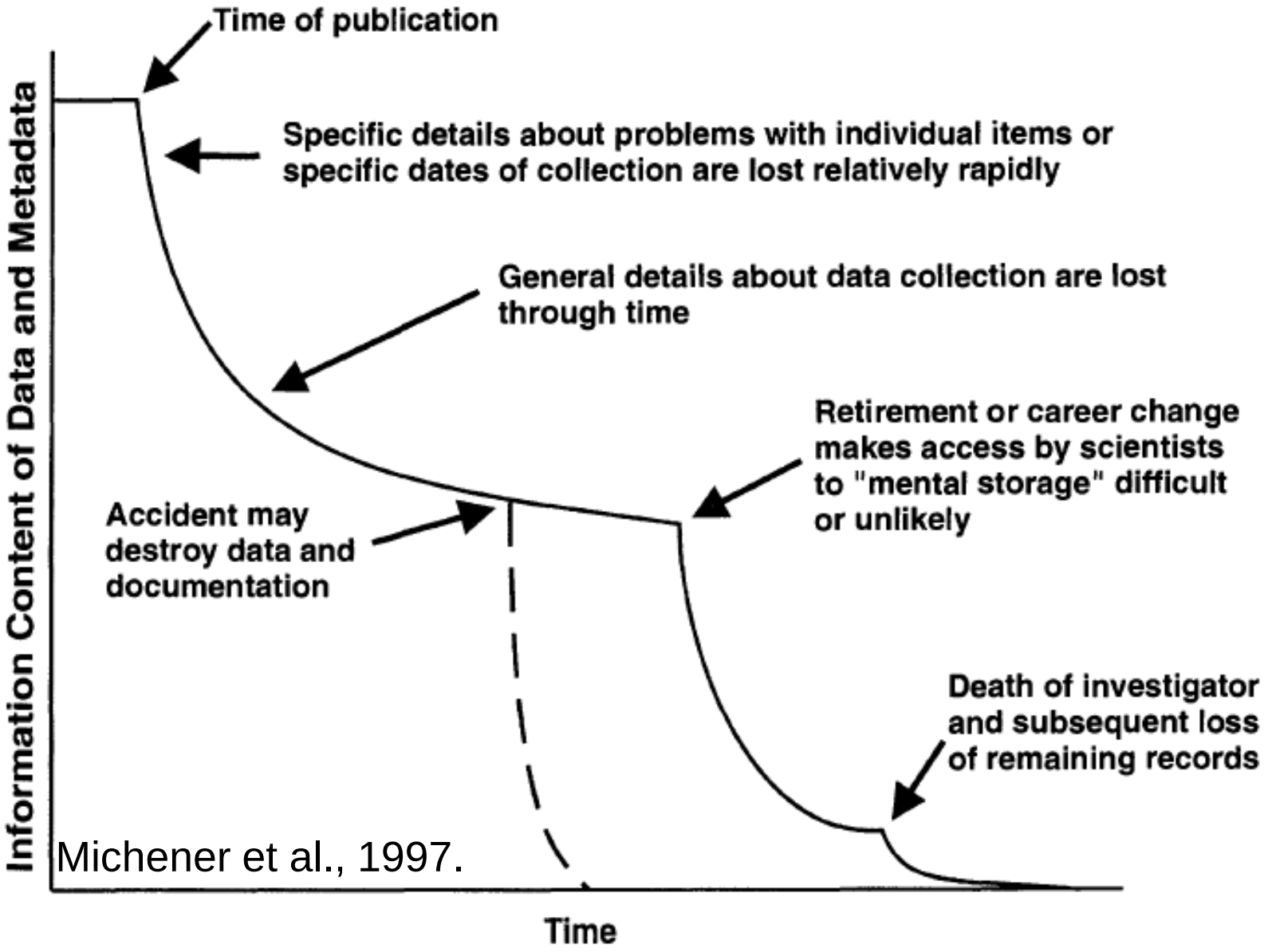
Leçons apprises

- *More isn't just more. More is different* (Wired 2008)
  - La vérité est dans les grands nombres
  - Contrastes et étendue dynamique
  - G x E x M
- L'importance des métadonnées (contexte)
  - Localisation (GPS), dates / stades
  - Travail du sol? Qualité du sol?

Partager mes données ou pas?

- Incitatifs à développer
- L'analogie de l'*Apple Watch*
- Les résultats de recherche ne s'appliquent bien qu'aux réalités dont ils sont issus

Dégradation normale du contenu en information



Michener et al., 1997.

OPEN DATA 5 ★

Tim Berners-Lee, principal inventeur du Web et instigateur des *Linked Data*, a suggéré un programme de déploiement en 5 étoiles pour l'Open Data. Voici quelques exemples qui détaillent chacune des étapes que représentent ces étoiles, avec les coûts et bénéfices qui les accompagnent.

https://5stardata.info/fr/

Ressources

- Les mécanismes du partage
  - Licences
  - Politiques données ouvertes
  - Données fournies avec articles scientifiques
- Spécialistes

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

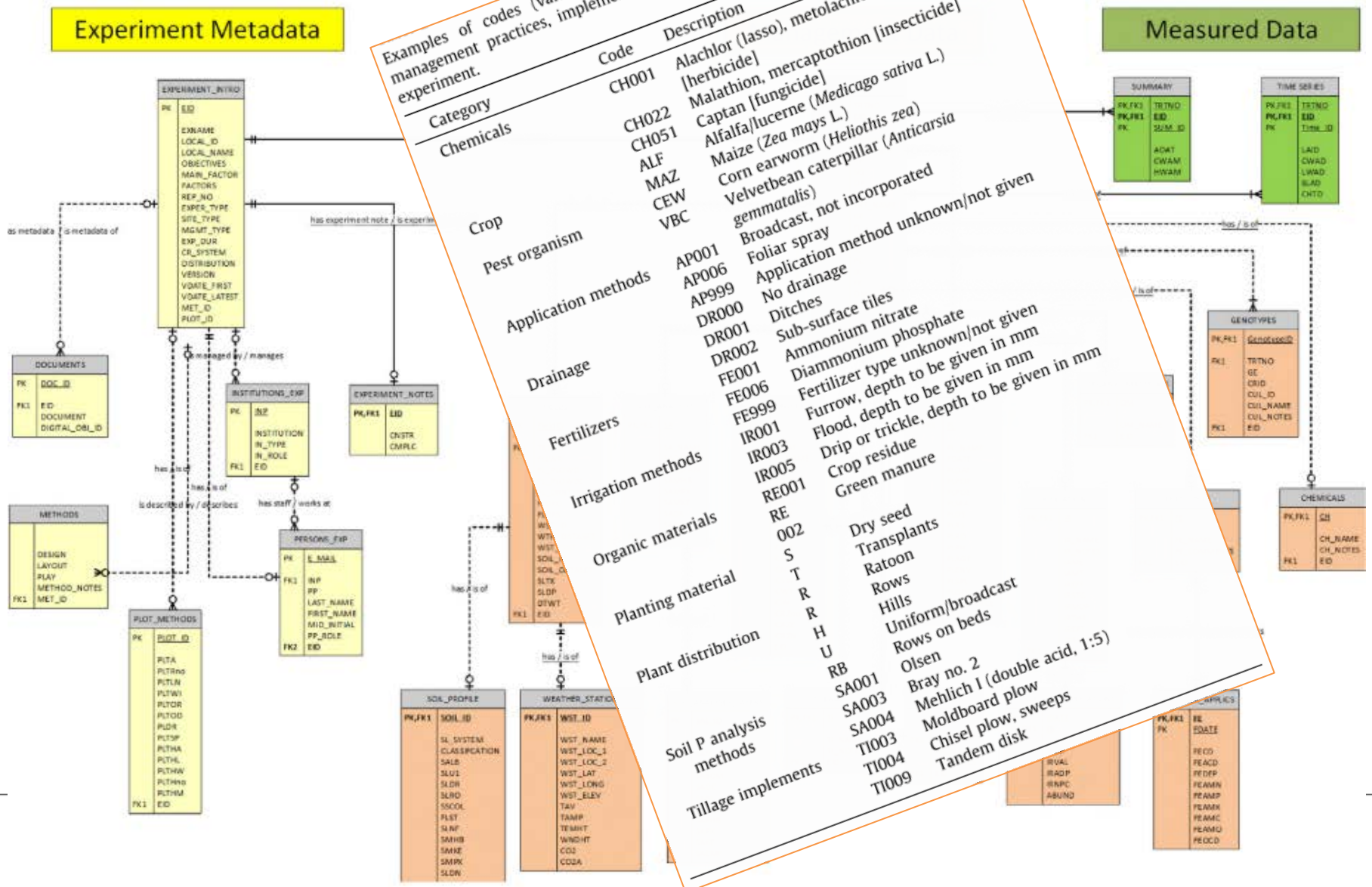
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Standardisation

- La tour de Babel
- Principe FAIR
  - *Findable, accessible, interoperable, reusable*
- *Research Data Alliance*
  - Élaboration des méthodes
- Pour l'aspect agronomique
  - ICASA (*International Consortium for Agricultural Systems Applications*)

ICASA



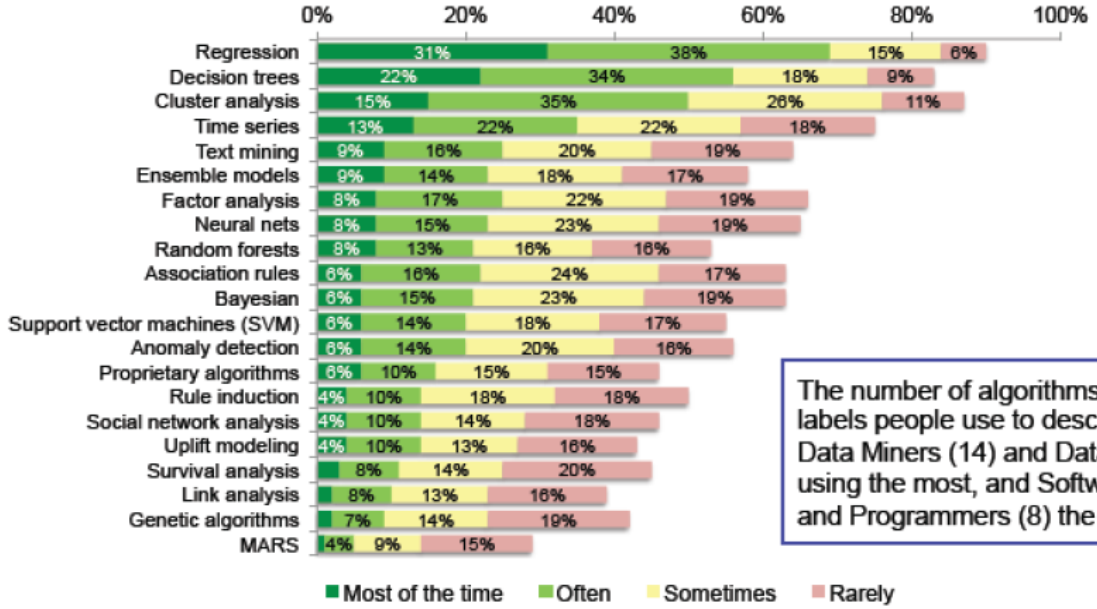
Examples of codes (variable codes) used to represent specific inputs, crops, management practices, implements or other aspect necessary to characterize an experiment.

Category	Code	Description
Chemicals	CH001	Alachlor (lasso), metolachlor (dual) [herbicide]
	CH022	Malathion, mercaptothion [insecticide]
	CH051	Captan [fungicide]
	ALF	Alfalfa/lucerne (*Medicago sativa* L.)
Crop	MAZ	Maize (*Zea mays* L.)
Pest organism	CEW	Corn earworm (*Heliothis zea*)
	VBC	Velvetbean caterpillar (*Anticarsia gemmatalis*)
Application methods	AP001	Broadcast, not incorporated
	AP006	Foliar spray
	AP999	Application method unknown/not given
Drainage	DR000	No drainage
	DR001	Ditches
	DR002	Sub-surface tiles
Fertilizers	FE001	Ammonium nitrate
	FE006	Diammonium phosphate
	FE999	Fertilizer type unknown/not given
Irrigation methods	IR001	Furrow, depth to be given in mm
	IR003	Flood, depth to be given in mm
	IR005	Drip or trickle, depth to be given in mm
Organic materials	RE001	Crop residue
	RE	Green manure
Planting material	002	Dry seed
	S	Transplants
	T	Ratoon
	R	Rows
Plant distribution	R	Hills
	H	Uniform/broadcast
	U	Rows on beds
	RB	Olsen
Soil P analysis methods	SA001	Bray no. 2
	SA003	Mehlich I (double acid, 1:5)
	SA004	Moldboard plow
Tillage implements	TI003	Chisel plow, sweeps
	TI004	Tandem disk
	TI009	

Analyse de données

Algorithms

- Regression, decision trees, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent since the first Data Miner Survey in 2007.
- The average respondent reports typically using 12 algorithms. People with more years of experience use more algorithms, and consultants use more algorithms (13) than people working in other settings (11).



The number of algorithms used varies by the labels people use to describe themselves, with Data Miners (14) and Data Scientists (14) using the most, and Software Developers (9) and Programmers (8) the fewest.

Question: What algorithms / analytic methods do you TYPICALLY use? (Select all that apply)

Source: Saporta

AAC

- Nouveaux chercheurs en AI = ressources
- Science et données ouvertes
- Laboratoires vivants
- Recherche observationnelle (*on-farm*)
- Agriculture de précision
  - Télédétection, capteurs, senseurs, IoT
  - Source de données abondantes
- Partenaires / collaborateurs

Réflexions

- Modèle
  - Au mieux, on dispose d'une calibration
  - Validation? Rarissime
- Conséquence: faible succès des décisions
- Le secteur agricole doit faire mieux
- Les données changent la donne
  - Les clubs ont des atouts majeurs
  - Penser localement, agir globalement

Data Sharing Benefits

Kitchen et al. 2016

- Provides more comprehensive understanding of agronomy practices over diverse environments for the purpose of developing better solutions
  - **G x E x M**
  - **G_{ST} x E_{ST} x M_{ST}**
- Creates "understanding" connections between the groups with competing interests, researcher and the farmer, relevance, (Cook et al)
- When you share it; you create an opportunity to grow it (conditional on completeness, quality, and standards)
- Data can be re-used to uncover new hypotheses when aggregated over space and time.
- As organizations develop capacity and encourage data sharing, recognition, credit, and awards will follow

Merci de votre attention