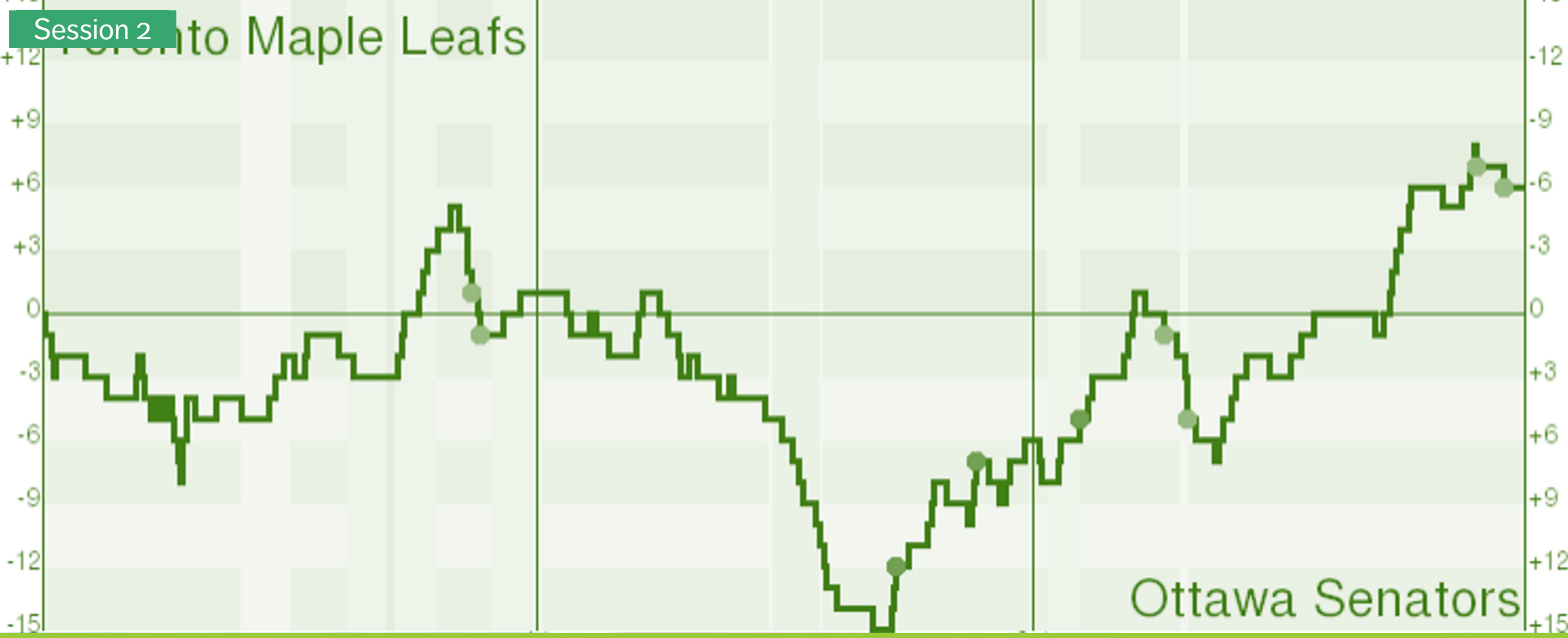
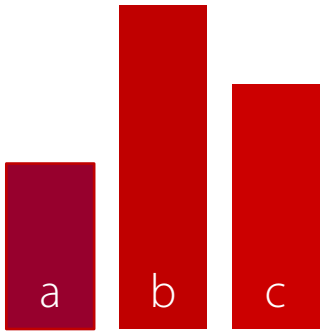


6. Les données et les renseignements

Modes d'analyse

Descriptive



Montrer **ce qui** s'est passé

Diagnostique

Expliquer **pourquoi** quelque chose s'est produit

Prédictive

Deviner **ce qui va** se passer

Prescriptive

Suggérer **ce qui devrait** se passer

Valeur faible
Faible difficulté

Valeur élevée
Difficulté élevée

Poser les bonnes questions

La science des données consiste à poser des questions et à y répondre :

- **Analytique :** "Combien de clics ce lien a-t-il obtenu ?"
- **La science des données :** "Sur la base de l'historique des achats précédents de cet utilisateur, puis-je prédire sur quels liens il va cliquer lors de son prochain accès au site ?"

Les modèles d'exploration de données/sciences sont généralement **prédictifs** (et non **explicatifs**) : ils montrent des connexions, mais ne révèlent pas **pourquoi** elles existent.

Attention : toutes les situations ne font pas appel à la science des données, à l'intelligence artificielle, à l'apprentissage automatique, aux statistiques, etc.

Les mauvaises questions

Trop souvent, les analystes posent les **mauvaises questions :**

- des questions **trop larges** ou **trop étroites**
- des questions **auxquelles aucune quantité de données ne pourra jamais répondre**
- les questions pour lesquelles **des données ne peuvent être obtenues**

Dans le meilleur des cas, les parties prenantes reconnaîtront que les réponses ne sont pas pertinentes.

Le **pire scénario** est qu'ils mettent en œuvre par erreur des politiques ou prennent des décisions sur la base de réponses qui n'ont pas été identifiées comme trompeuses ou inutiles.

Feuille de route

Comprendre le problème (opportunité vs problème)

Quelles hypothèses initiales ai-je sur la situation ?

Comment les résultats seront-ils utilisés ?

Quels sont les risques et/ou les avantages de répondre à cette question ?

Quelles questions des parties prenantes pourraient être soulevées en fonction des réponses ?

Ai-je accès aux données nécessaires pour répondre à cette question ?

Comment vais-je mesurer mes critères de "réussite" ?

Le piège du Oui/Non

Exemples de **mauvaises** questions :

- Nos revenus **augmentent-ils** d'une année sur l'autre ?
- La plupart de nos clients appartiennent-ils à **cette catégorie démographique** ?
- **Ce projet a-t-il des** ambitions valables pour l'ensemble du département ?
- Est-ce que notre équipe de succès de la clientèle, qui travaille dur, est **formidable**.
- À quelle fréquence **vérifiez-vous par trois fois** votre travail ?

Exemples de **bonnes** questions :

- Quelle est la **répartition** de nos revenus au cours des trois derniers mois ?
- D'où viennent nos **5** cohortes **les plus** dépensières ?
- Que sont les **différents avantages** de la poursuite de ce projet ?
- Que **sont trois bons** *et* **trois mauvais traits de** notre équipe de réussite client ?
- Avez-vous **tendance à** effectuer des tests d'assurance qualité sur vos livrables ?

Liste de contrôle

1. Ai-je évité de créer des questions de type oui/non ?
2. Est-ce que tous les membres de mon équipe/département comprendraient la question, indépendamment de leurs antécédents ?
3. La question nécessite-t-elle plus d'une phrase pour être exprimée ?
4. La question est-elle "équilibrée“ ? (champ d'application ni trop large pour une réponse, ni trop restreint au point de n'avoir qu'un impact minime)
5. La question est-elle orientée vers ce à quoi il est plus facile de répondre pour les compétences particulières de mon équipe ?

Contingence/Tableaux croisés

Tableau de contingence : examine la relation entre deux variables catégorielles

Tableau croisé dynamique : un tableau généré en appliquant des opérations (compte, moyenne, etc.) à des variables sur la base d'une autre variable.

Les tableaux de contingence sont des cas particuliers de tableaux croisés dynamiques (“pivot tables”).

	Large	Moyen	Petits
Fenêtre	1	32	31
Porte	14	11	0

Type	N	Signal moy	Signal ET
Bleu	4	4.04	0.98
Vert	1	4.93	N.A.
Orange	4	5.37	1.60

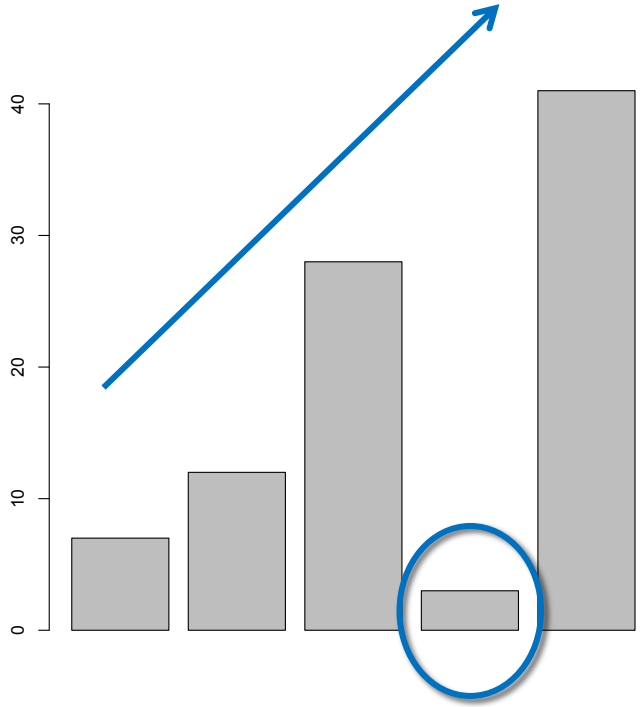
L'analyse par la visualisation

Analyse (au sens large) **:**

- identifier des modèles ou des structures
- ajouter du sens à ces modèles ou à cette structure en les interprétant dans le contexte du système.

Option 1 : utiliser des méthodes analytiques

Option 2 : visualiser les données et utiliser le pouvoir d'analyse du cerveau (perceptuel) pour tirer des conclusions significatives



Résumés numériques

Dans un premier temps, une variable peut être décrite selon 2 dimensions : la **centralité** et la **dispersion** (l'asymétrie et l'aplatissement sont aussi utilisés).

Les **mesures de centralité** comprennent :

- médiane, moyenne, mode (moins fréquemment)

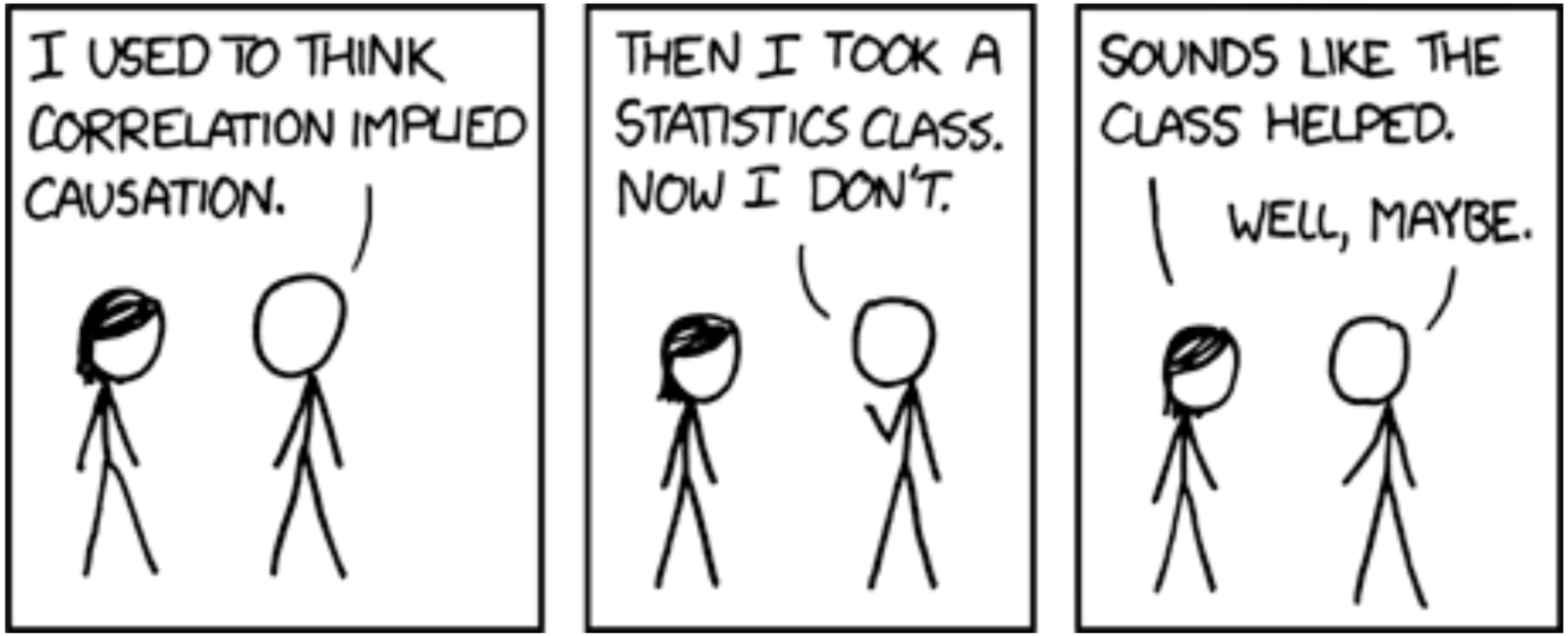
Les **mesures dispersion** (ou d'étalement) comprennent :

- écart-type (sd), variance, quartiles, écart interquartile (IQR), étendue (moins fréquemment)

La médiane, l'etendue, et les quartiles sont facilement calculés à partir de **listes ordonnées**.

[https://xkcd.com/552]

Corrélation



La corrélation n'implique pas la causalité, mais elle agite les sourcils de manière suggestive et fait des gestes furtifs en disant "regardez par là".

Régression linéaire

L'hypothèse de base de la **régression linéaire** est que la variable dépendante peut être approximée par combinaison linéaire des variables indépendantes :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

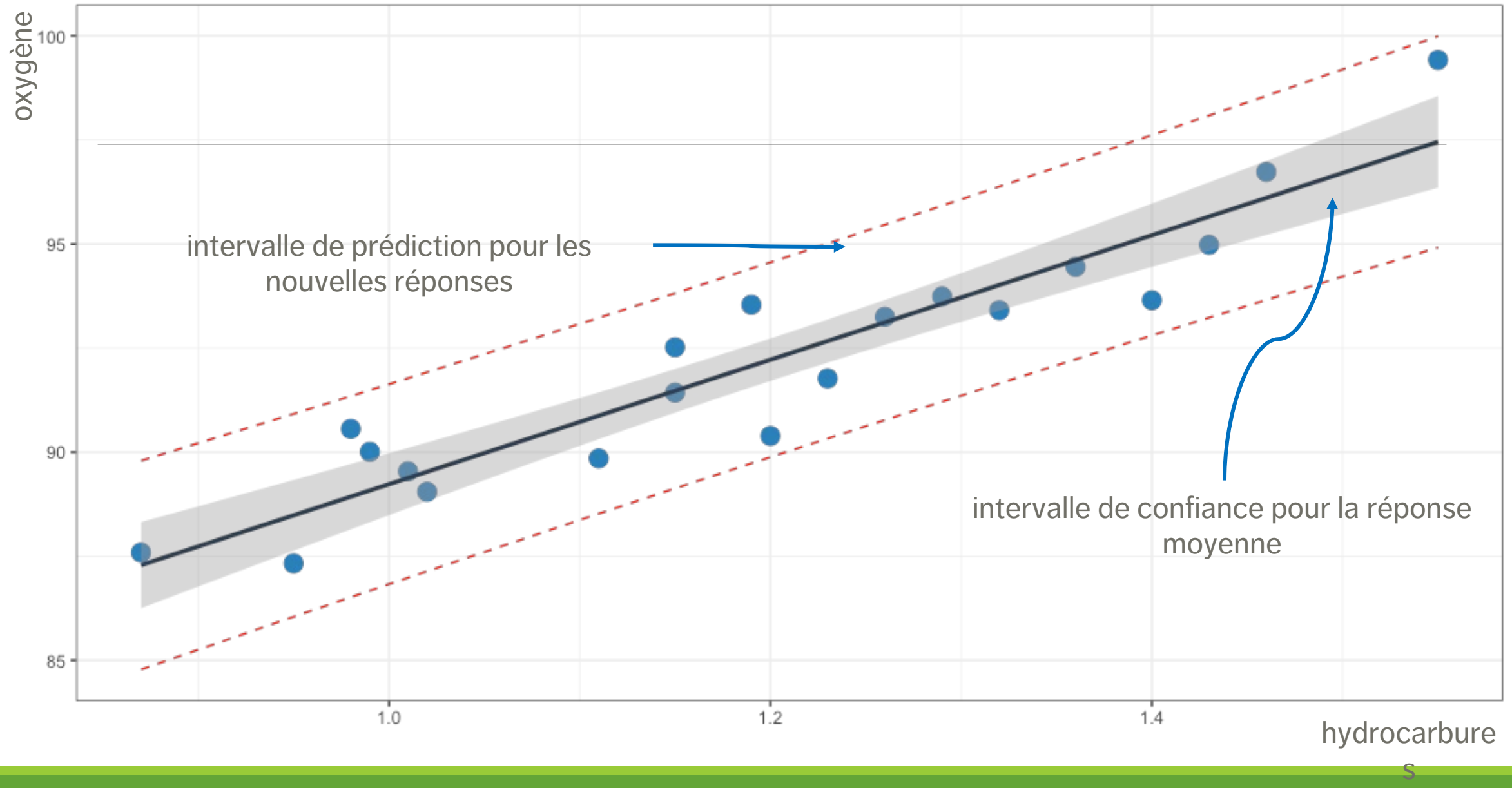
où $\boldsymbol{\beta} \in \mathbb{R}^p$ est déterminé sur la base de l'**ensemble d'apprentissage $\mathbf{X}$**, et

$$E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \qquad E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\mathbf{X}) = \sigma^2\mathbf{I}.$$

Généralement, les erreurs sont **distribuées selon une normale :**

$$\boldsymbol{\varepsilon}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

$$\text{oxygène} = 14.95 \times \text{hydrocarbure} + 74.28$$



Tâches d'apprentissage automatique

Classification et **estimation de la probabilité de classe :** quels clients sont susceptibles d'être des clients réguliers ?

Regroupement (“clustering”) **:** les clients forment-ils des groupes naturels ?

Règles d'association : quels livres sont couramment achetés ensemble ?

Autres :

profilage et description du comportement ; **prédiction des liens** ; **estimation de la valeur** (combien un client est-il susceptible de dépenser dans un restaurant) ; **mise en correspondance des similarités** (quels clients potentiels sont similaires aux meilleurs clients d'une entreprise ?); **réduction des données** ; **modélisation d'influence**, etc.

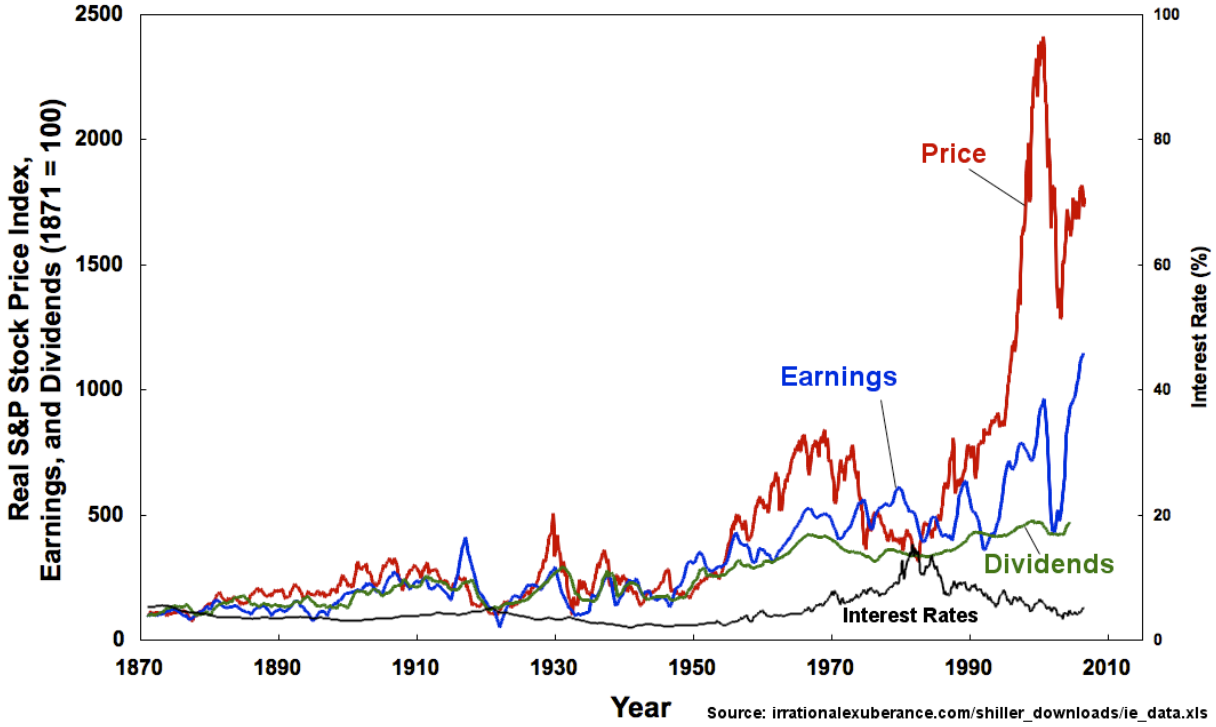
Analyse des séries temporelles

Une **série chronologique** simple **:**

- a deux variables : temps + 2nd variable
- la deuxième variable est *séquentielle*

Quel est le **comportement** de cette deuxième variable dans le temps ?

Pouvons-nous **prévoir le comportement futur** de la variable ?



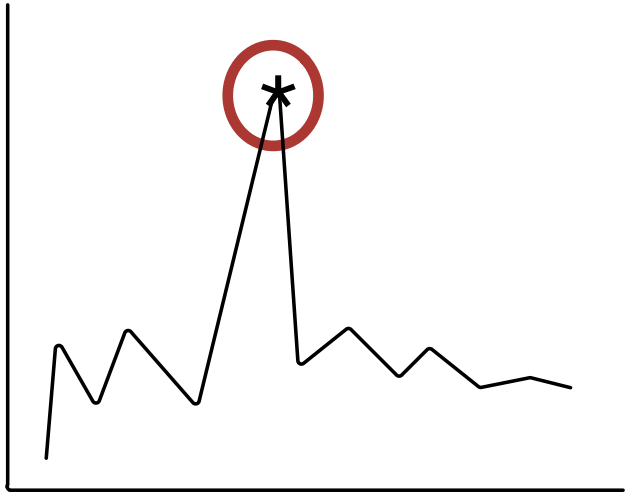
Détection d'anomalies

Anomalie : un événement inattendu, inhabituel, atypique, ou statistiquement improbable.

Ne serait-il pas utile d'avoir un pipeline d'analyse de données qui vous alerte lorsque les choses sortent de l'ordinaire ?

Il y a plusieurs approches analytiques à adopter !

- regroupement
- classification
- techniques d'ensemble, etc.



Lectures suggérées

Les données et les renseignements

Data Understanding, Data Analysis, Data Science
Data Science Basics

Getting Insight From Data

- Asking the Right Questions
- Basic Data Analysis Techniques
- Common Statistical Procedures in R
- Quantitative Methods

***Probability and Applications** (advanced)

***Introductory Statistical Analysis** (advanced)

***Survey Sampling** (advanced)

***Regression Analysis** (coming soon)

Exercices

Les données et les renseignements

1. Faites l'exercice de la section Asking the Right Questions.

2. Recréez les exemples de Common Statistical Procedures in R.

3. Le fichier cities.txt contient des informations sur la population des villes d'un pays. Une ville est classée comme "petite" si sa population est inférieure à 75K, comme "moyenne" si elle se situe entre 75K et 1M, et comme "grande" autrement. Localisez et chargez le fichier dans l'espace de travail de votre choix. Combien de villes y a-t-il ? Combien y en a-t-il dans chaque groupe ? Affichez des statistiques démographiques sommaires pour les villes, à la fois globalement et par groupe.