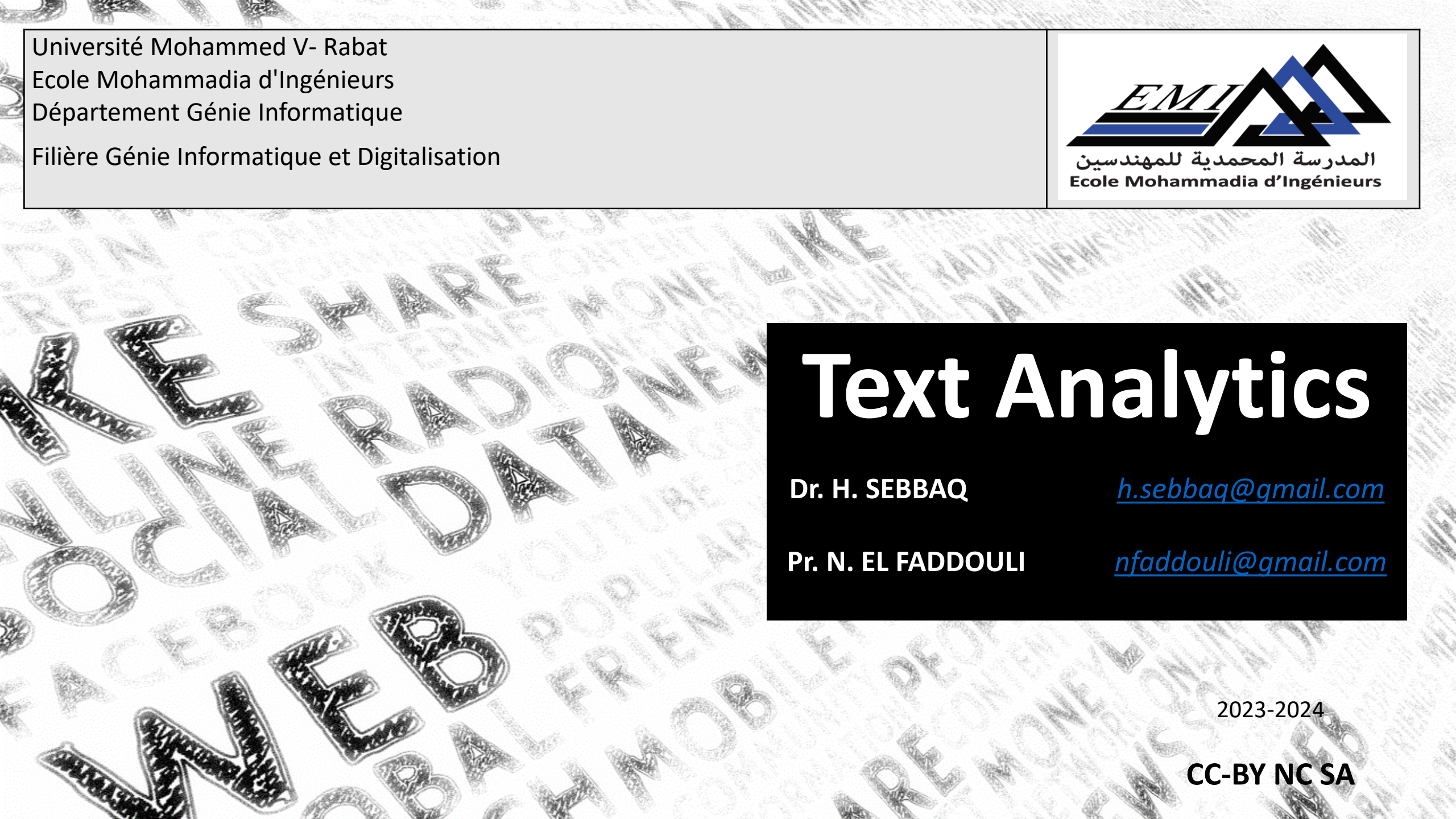
Université Mohammed V- Rabat
Ecole Mohammadia d'Ingénieurs
Département Génie Informatique

Filière Génie Informatique et Digitalisation



Text Analytics

Dr. H. SEBBAQ *h.sebbaq@gmail.com*

Pr. N. EL FADDOULI *nfaddouli@gmail.com*

2023-2024

Plan du chapitre

- **Introduction**
- **Aperçu sur l'analyse de texte**
- **Approches, tâches et applications de l'analyse de texte**
- **Processus d'analyse de texte**

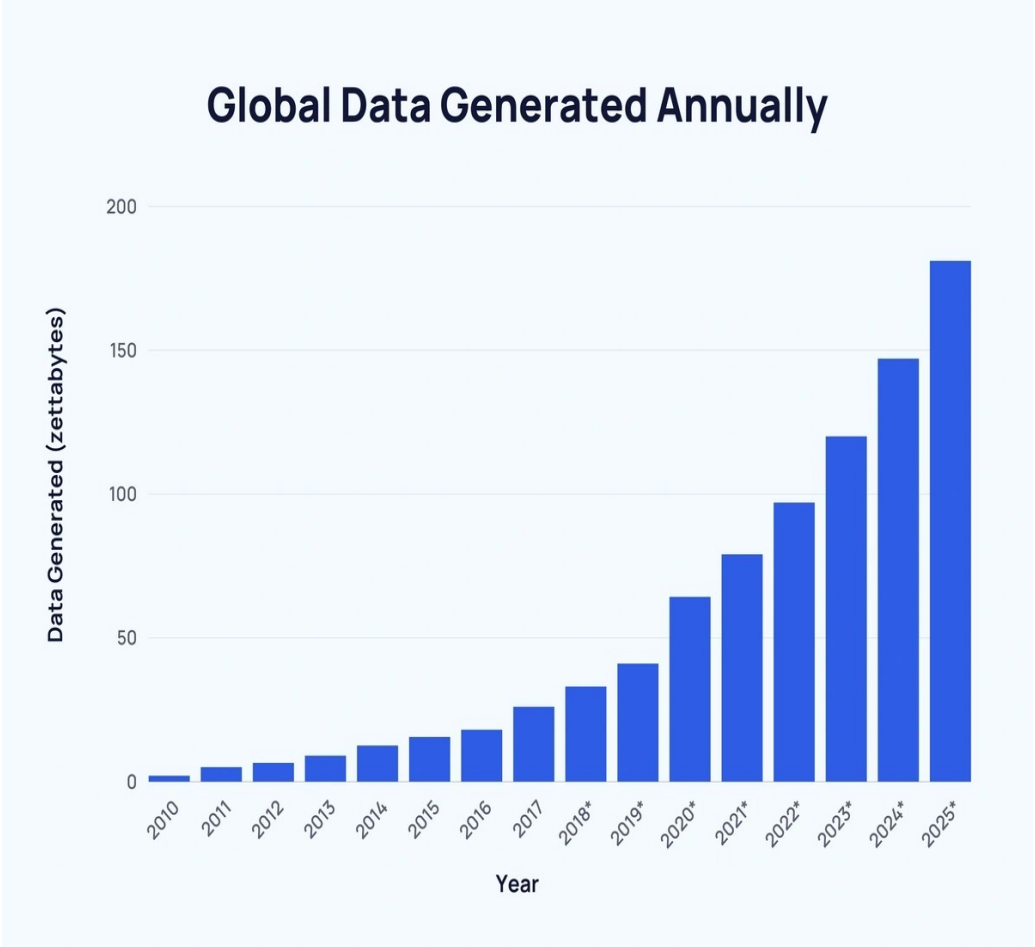
Plan Introduction

- **Explosion des données**
- **Données textuelles**
- **Données textuelles non structurées : Défi**

Introduction : Explosion des données

- La quantité de données générées annuellement a augmenté d'année en année depuis 2010.
- Alors qu'il n'était que de **2 zettaoctets en 2010**, en l'espace de 13 ans, ce chiffre a été multiplié par 60
- En fait, on estime que **90 % des données mondiales ont été générées au cours des deux dernières années seulement.**
- Selon les estimations, Les **120 zettaoctets générés en 2023** devraient augmenter de plus de 150 % en 2025, pour atteindre **181 zettaoctets**.

https://explodingtopics.com/blog/data-generated-per-day



Introduction : Explosion des données - Facteurs clés

- **Connectivité croissante**
  - Adoption répandue d'Internet, smartphones, générant un flux massif de données.
  - Interactions des utilisateurs (médias sociaux, forums), capteurs, transactions en ligne, streaming et contenu numérique.

- **Transformation numérique majeure**
  - Industries, entreprises et gouvernements subissent des changements numériques importants.
  - Génération de données à partir de processus, transactions, communications et opérations.

- **Domaines de recherche et IoT**
  - Recherche scientifique, santé, astronomie et surveillance de l'environnement contribuent à une explosion des données.
  - Volumes massifs via expériences, simulations, observations et capteurs IoT.

Introduction : Données textuelles

❑ Données structurées

- **Format organisé :** Les données structurées sont organisées selon un schéma fixe avec des formats prédéfinis. Elles sont souvent stockées dans des bases de données relationnelles.
- **Facilité d'analyse :** En raison de leur organisation précise, ces données sont plus faciles à interroger et à analyser à l'aide de requêtes SQL ou d'autres langages de base de données.
- **Stockage :** Elles sont généralement stockées dans des systèmes de gestion de bases de données relationnelles (SGBDR).
- **Exemples :** Informations transactionnelles, données sur les clients (nom, adresse, numéro de téléphone), données financières (comptes, transactions, soldes), données d'inventaire, etc.

Introduction : Données textuelles

❑ Données non structurées

- **Absence de format prédéfini :** Ces données ne suivent pas un schéma fixe et n'ont pas de format prédéfini → ne peuvent pas être facilement organisées dans les bases de données traditionnelles.
- **Variété de formats :** Comprend des informations non linéaires et hétérogènes →
- **Complexité de l'analyse :** Le traitement est complexe et nécessite souvent des techniques d'analyse avancées
- **Stockage :** Ces types de données sont souvent stockés dans des systèmes de fichiers, des plateformes de stockage en Cloud ou des entrepôts de données spécialisés pour les données non structurées.
- **Exemples de données non structurées :** E-mails, Documents, Flux de médias sociaux, Chat logs, Documents juridiques, Documents et articles de recherche, Transcriptions de discours, Notes et mémos, Critiques de produits, Articles d'actualité et blogs

Introduction : Données textuelles

❑ Données semi-structurées

- **Un mélange des deux mondes :** Une certaine structure mais pas aussi rigides que les données structurées. Elles peuvent contenir des éléments structurels ou des balises.
- **Adaptabilité :** Souvent utilisées dans les applications web, les services API et d'autres domaines où la flexibilité est nécessaire pour représenter des informations complexes.
- **Exemples :**
  - JSON (JavaScript Object Notation) → Echange de données entre applications
  - XML (eXtensible Markup Language), HTML (Hypertext Markup Language) → Définit structure des pages web

Introduction : Données textuelles non structurées - Défi

- Les données textuelles structurées ont tendance à avoir une organisation préexistante ou des schémas définis → Simplifie leur traitement.
- En revanche, les données textuelles non structurées en raison de leur nature non uniforme et variée → Nécessitent souvent des **étapes préliminaires importantes de prétraitement et d'analyse avancée pour extraire des informations exploitables**
- Les données non structurées représentent 80 % des informations disponibles, créant un défi majeur pour leur exploitation.