テキストデータの統計科学入門

第1章 統計的テキストマイニング

新納浩幸

統計的テキストマイニング

データマイニング ← 定形データが対象

非定形データの代表例がテキスト

テキストマイニング

テキストを単語やフレーズに分解し、
それらの統計情報からデータマイニング

テキストマイニングの例

所信表明演説の語彙比較

安部	語彙	福田
16	改革	19
5	安定	11
2	安心	12

疑問

なぜこの3単語なの？

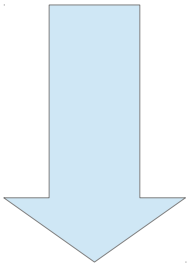
文書の長さを無視して頻度だけで大丈夫？

統計的テキストマイニングの小史

計量文体学（100年程前）

メンデルホーン

単語のスペクトル分析による著者推定 (1887年)



テキストマイニング（1990年代後半）

統計的テキストマイニングの諸分野

（1）計量文体学

（2）計量言語学とコーパス言語学

（3）情報･知識の発見と抽出

（4）ゲノムとテキスト解析

テキストマイニングの手順

(1)テキストの電子化
(2)クリーニング
(3)テキストの加工
　　記号･文字単位
　　形態素解析
　　構文解析
(4)データの抽出
(5)データの分析