テーマ：
機械学習を用いた技術動向分析の試み
～R言語による視覚的な手法を用いた分析～

アジア特許情報研究会 知財情報解析チーム 四方

2018年11月29日

動機とテーマ

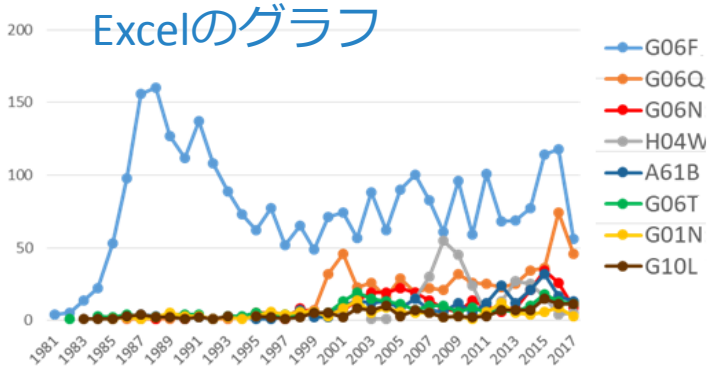
■動機

- ➢機械学習を自分でやってみたい
- ➢PythonやＲ言語等のOSS環境が充実
- ➢高価なツールがなくても分析可能に？（期待）

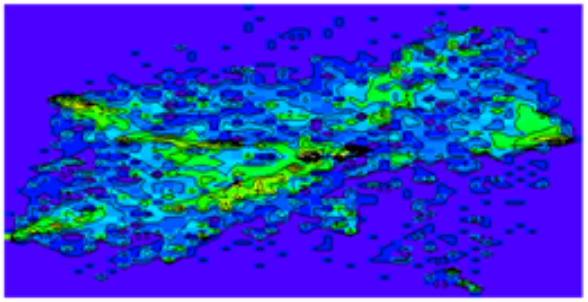
⇒ 今年４月頃からテキストマイニングの勉強を開始

■テーマ

- ➢視覚的な手法を用いた特許情報の分析



例：等高線／ヒートマップ



テーマ：

機械学習を用いた技術動向分析の試み
～R言語による視覚的な手法を用いた分析～

Ⅰ. 大規模データの等高線マップ作成方法の検討
Ⅱ. AI, ML, DL等の用語含む特許出願の動向分析
Ⅲ. GAFAの米国特許出願の動向分析

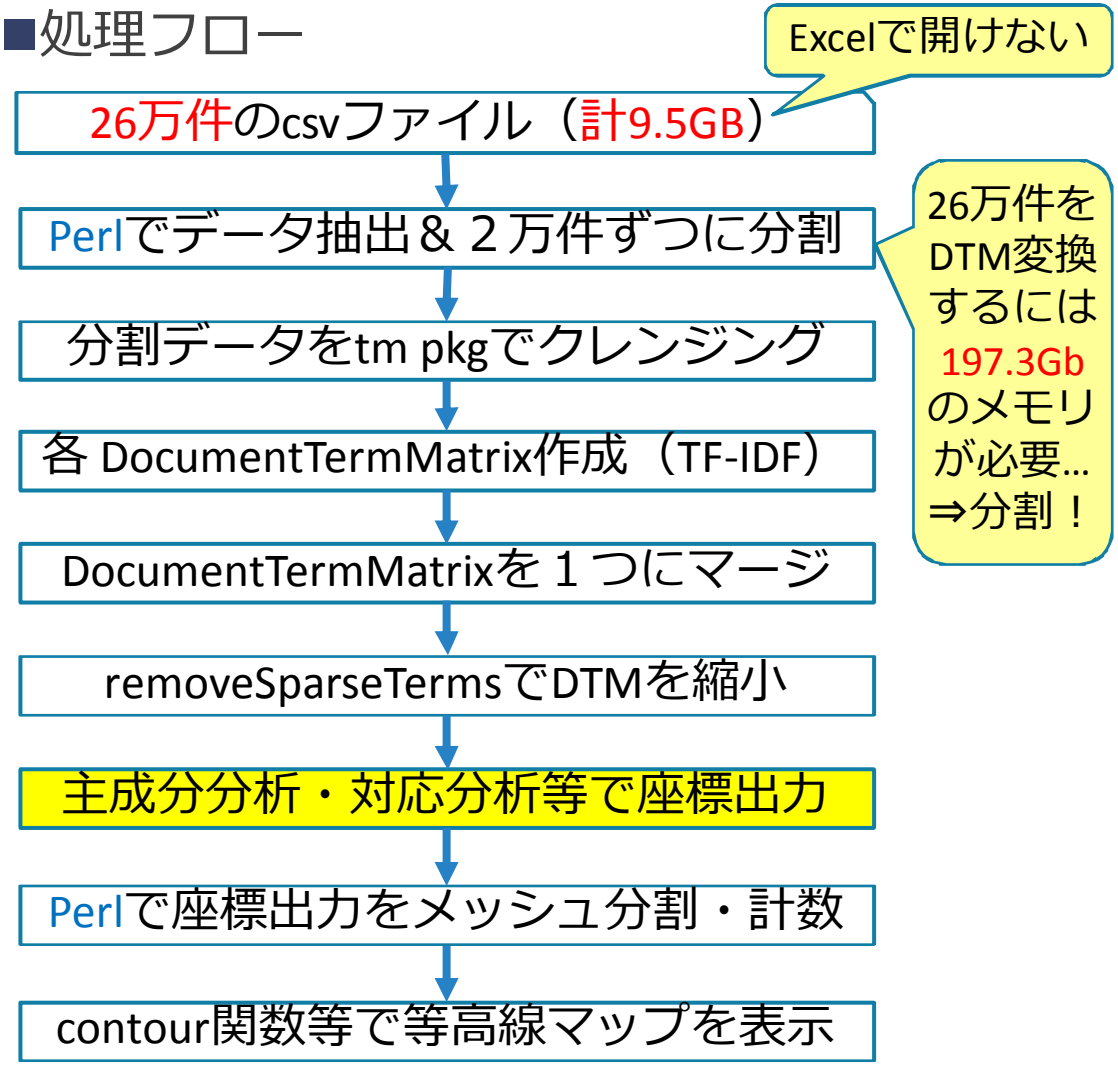
- 対象：“AI”, “ML”, “DL”等の用語含む出願
- 内容：等高線マップによる可視化手法の検討

Ⅰ. 対象データ、実行環境、処理フロー

■対象データ
- ➢“AI”の用語を含む特許出願
  - “artificial intelligence”
  - “machine learning”
  - “deep learning”
    （※ 全文検索のためノイズ含む）
- ➢期間：出願日が1960年代以降
- ➢対象：全世界の特許出願の英文タイトル＋英文要約
- ➢対象特許出願件数：約26万件
- ➢csvファイルのサイズ：約9.5GB
  ※抽出データの合計：約790MB

■実行環境（個人ノートPC）
- ➢PCスペック：
  Intel® Core™ i5-4200M
  メモリ 8.00GB
  Windows 10 Home (64bit)
- ➢R言語のバージョン：3.5.0

■処理フロー

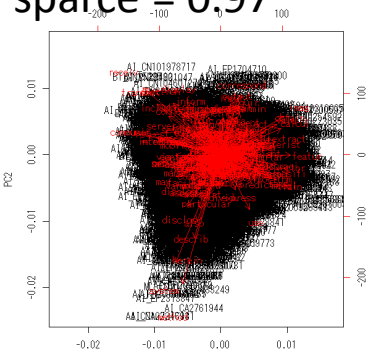


Ⅰ. 処理方法検討：主成分分析

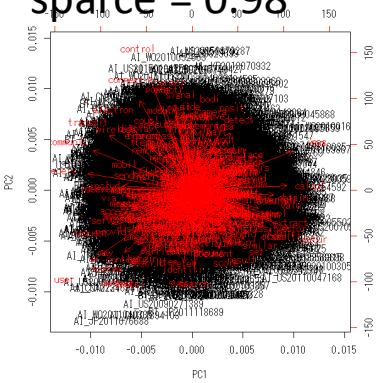
■”artificial intelligence”, “machine learning”等の用語を含む出願（英語文献）

主成分分析
stats: prcomp

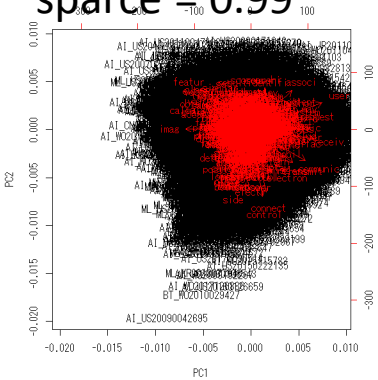
sparce = 0.97

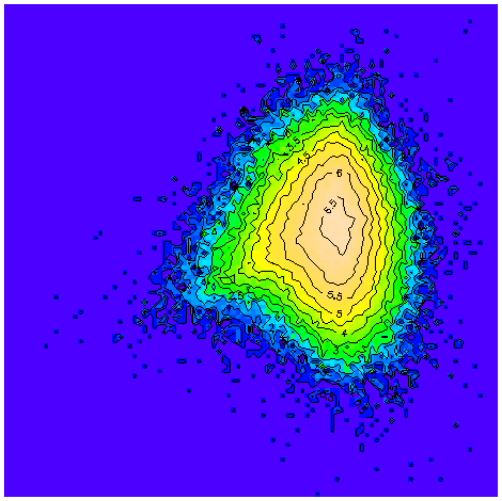


sparce = 0.98



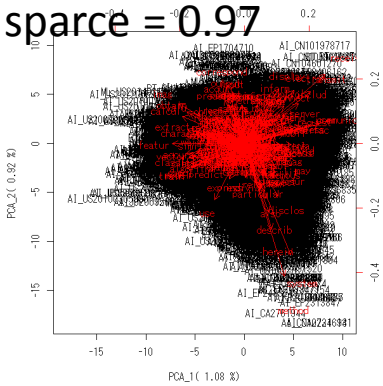
sparce = 0.99



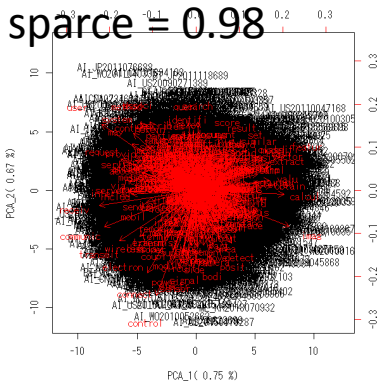


主成分分析
FactoMineR:PCA

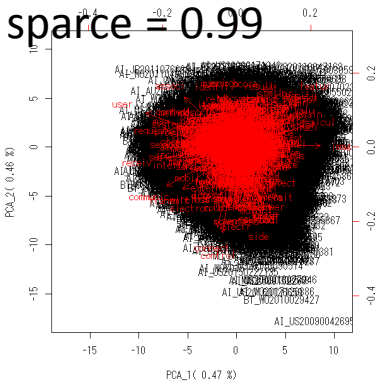
sparce = 0.97

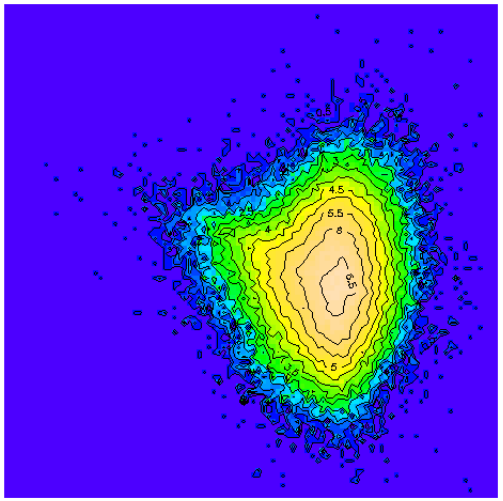


sparce = 0.98



sparce = 0.99



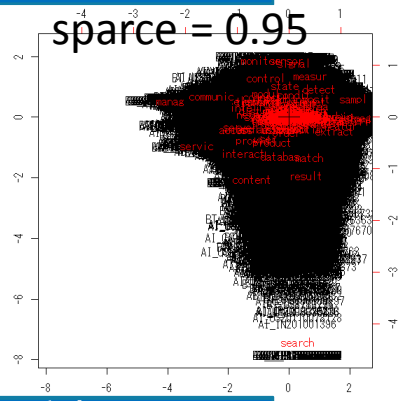


Ⅰ. 処理方法検討：対応分析（１）

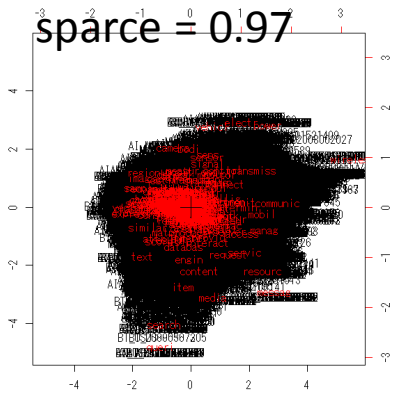
■”artificial intelligence”, “machine learning”等の用語を含む出願（英語文献）

対応分析
MASS: corresp

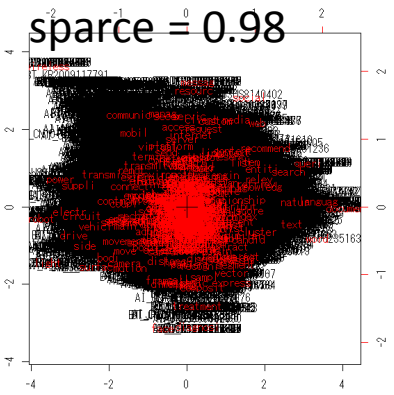
sparce = 0.95

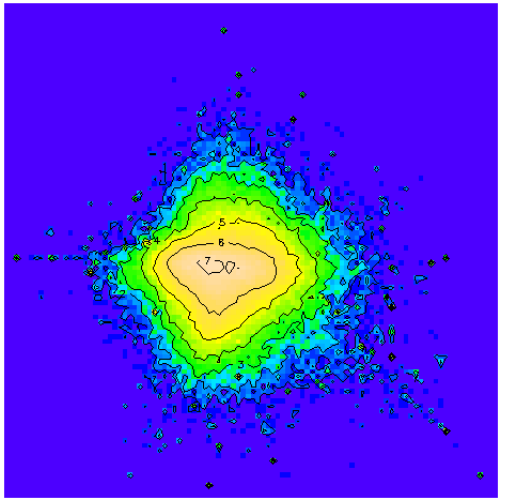


sparce = 0.97



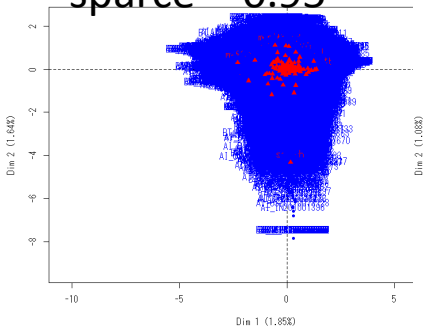
sparce = 0.98



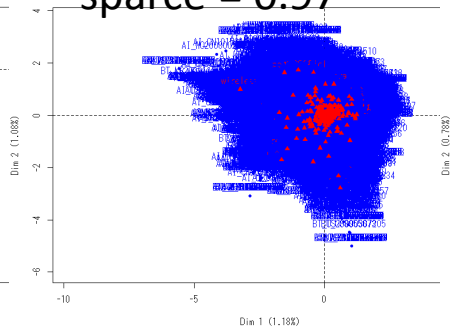


対応分析
FactoMineR: CA

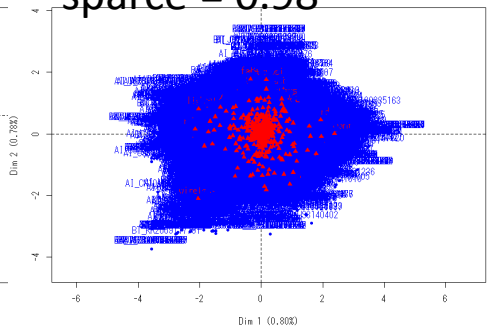
sparce = 0.95

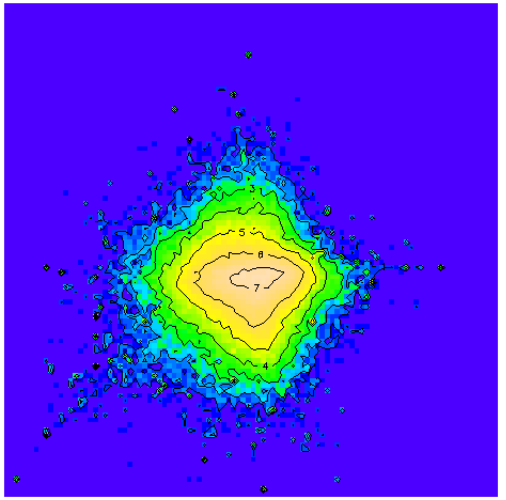


sparce = 0.97



sparce = 0.98



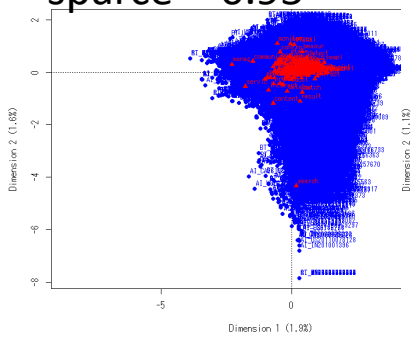


Ⅰ.処理方法検討：対応分析（２）

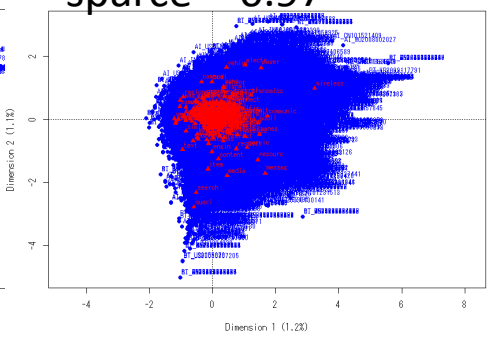
■”artificial intelligence”, “machine learning”等の用語を含む出願（英語文献）

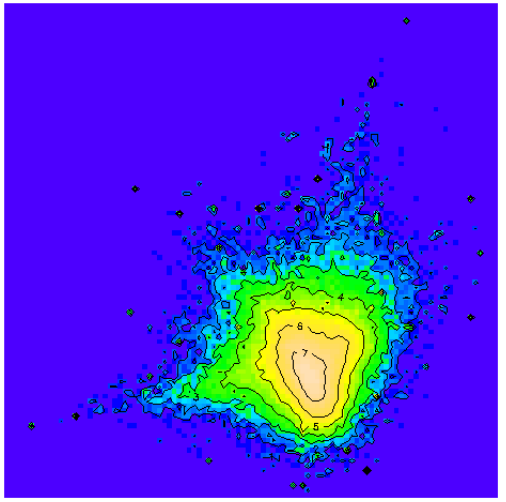
対応分析
ca: ca

sparce = 0.95



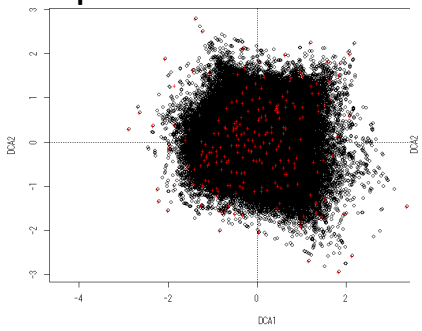
sparce = 0.97



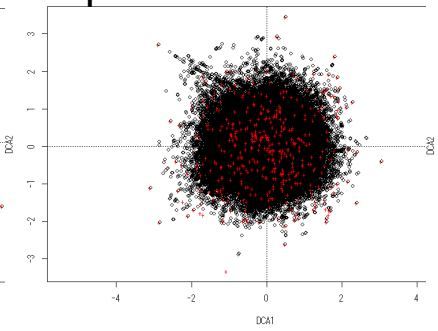


対応分析
vegan: decorana

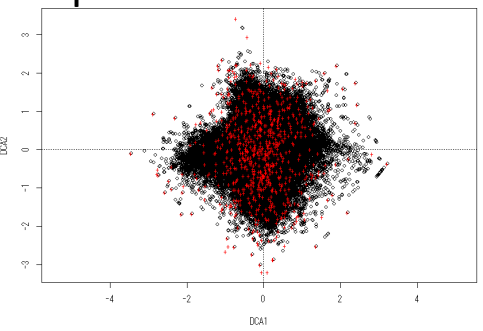
sparce = 0.97

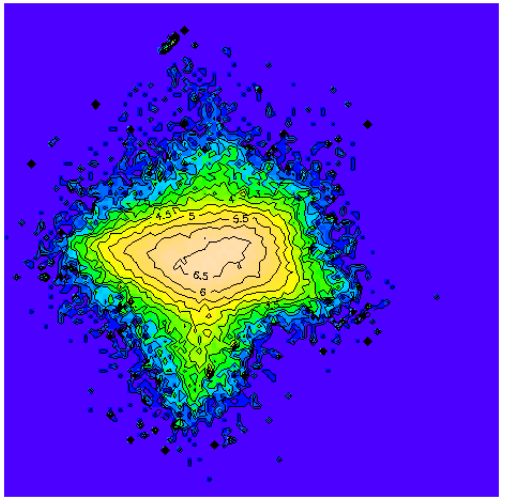


sparce = 0.98



sparce = 0.99





Ⅰ．トライアルした処理方法の一覧

■大規模な等高線表示用データ作成のためのコマンド比較

Library	command	DTM	sparce=0.95	sparce=0.97	sparce=0.98	sparce=0.99
stats	prcomp	TF-IDF	OK (22.7Mb)	OK (22.9Mb)	OK (23.0Mb)	OK (23.1Mb)
FactoMineR	PCA	TF-IDF	OK (375.5Mb)	OK (585.3Mb)	OK (874.2Mb)	OK (1.4Gb)
ca	ca	TF-IDF	OK (307.6Mb)	OK (516.8Mb)	Mem Error	Mem Error
FactoMineR	CA	TF-IDF	OK (654.6Mb)	OK (1.0Gb)	OK (1.6Gb)	Mem Error
MASS	corresp	TF-IDF	OK (301.7Mb)	OK (510.9Mb)	OK (799.5Mb)	Mem Error
vegan	decorana	TF-IDF	OK (28.5Mb)	OK (28.7Mb)	OK (28.9Mb)	OK (29.0Mb)
(計量MDS)	cmdscale	Tf/Tfidf	MemError(dist)	-		
(非計量MDS)	isoMDS	Tf/Tfidf	MemError(dist)	-		

このデータで各手法の特徴を比較

※sparce：removeSparseTermsの”sparce”値，（）内は変数のMem使用量：例）res <- corresp(dtm)
※TF-IDF：dtm0x <- DocumentTermMatrix(corpus0x, control = list(weighting = weightTfIdf))
※decoranaはDCA(Detrended Correspondence Analysis)で、アーチ効果を除去する対応分析の方法

“dtm”のサイズ削減に関する項目	s=0.95	s=0.97	s=0.98	s=0.99
removeSparseTerms 実行後のDTMのmatrixサイズ	286.2 Mb	494.8 Mb	783.0 Mb	1.4 Gb
removeSparseTerms 実行後のDTMのTerm数	133	238	383	687

Ⅰ. 処理方法の評価：主成分分析

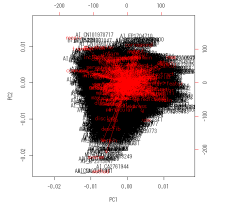
■主成分分析・対応分析の２次元出力結果を極座標系で評価

➢２次元ユークリッド空間での極座標（r, Θ）の算出

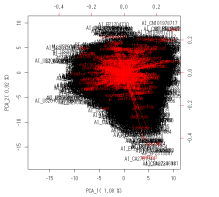
・$r = \sqrt{x^2 + y^2}$　　・$\Theta = \mathrm{sgn}\,(y) \arccos\left(x / \sqrt{x^2 + y^2} \right)$　（※ sgn は符号関数）

➢極座標系での各出力結果の比較

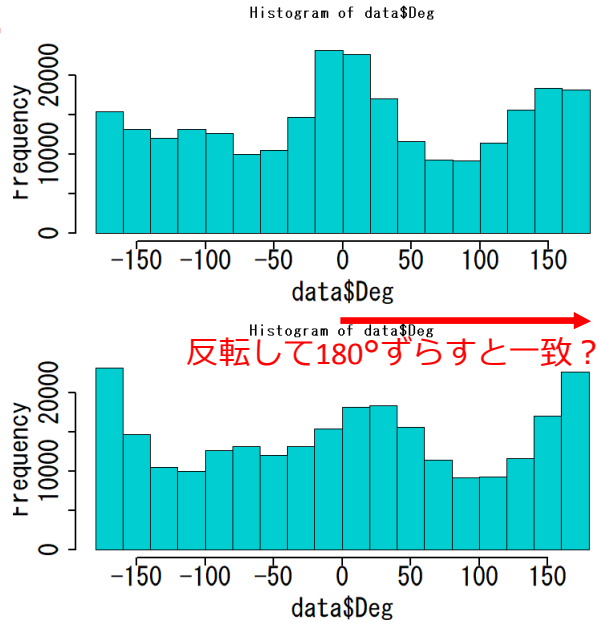
prcomp



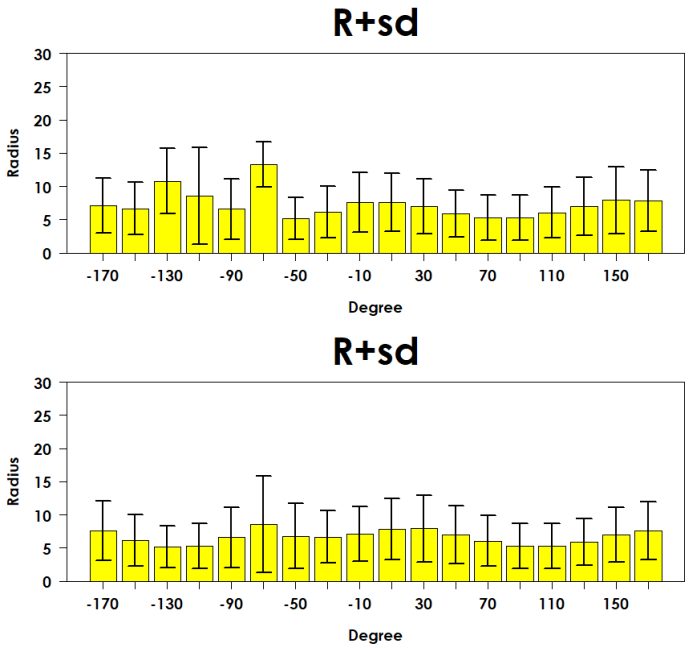
Facto:PCA



biplot図面



各角度での件数のヒストグラム



各角度での r の平均値と標準偏差

全角度での r の平均値と標準偏差

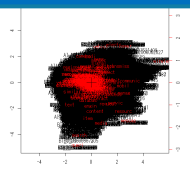
Rm=6.85
Rsd=4.51

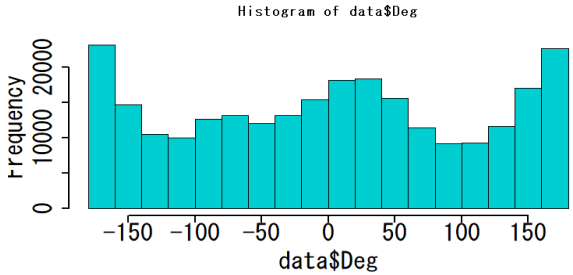
Rm=6.85
Rsd=4.51

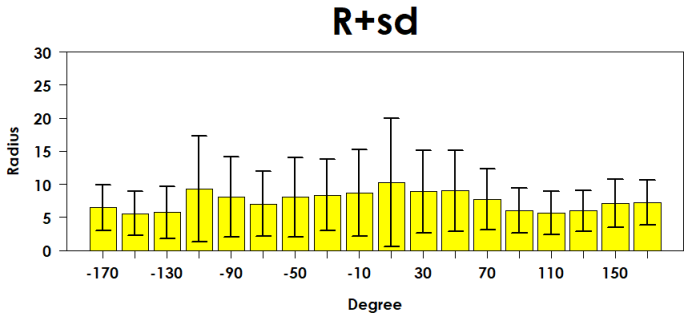
Ⅰ. 処理方法の評価：対応分析

- 極座標系での各出力結果の比較

MASS: corresp

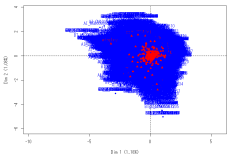


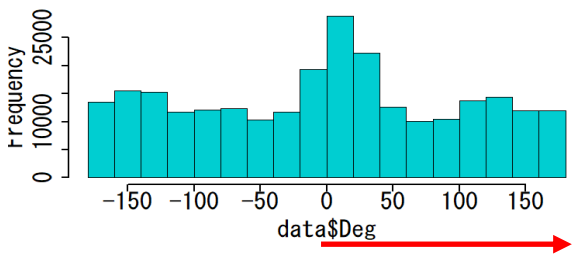


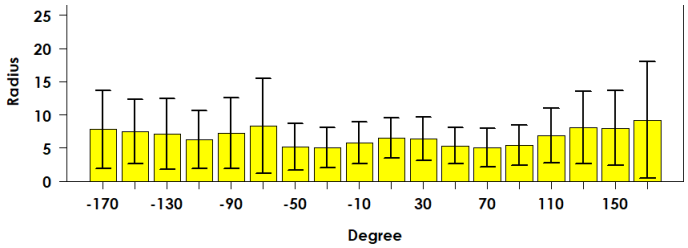


Rm=6.85
Rsd=4.51

Facto: CA

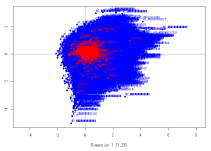




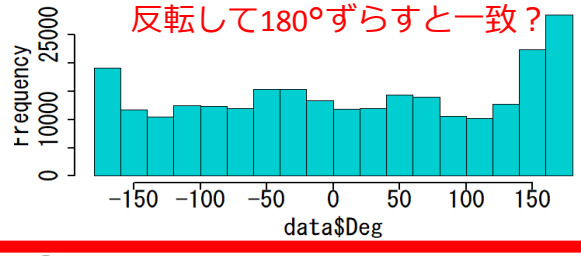


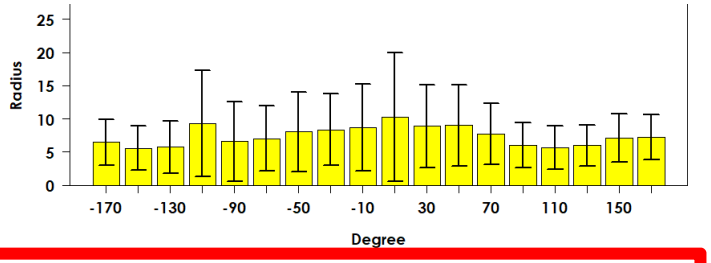
Rm=6.73
Rsd=4.81

ca: ca



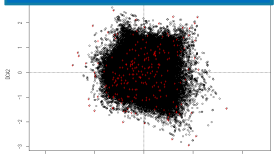
反転して180°ずらすと一致？

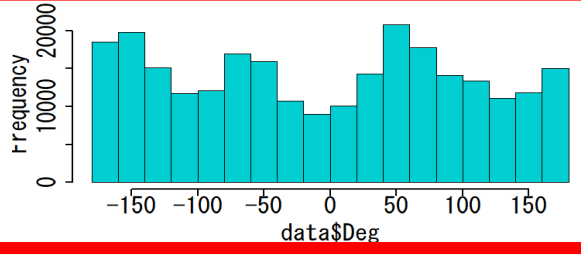


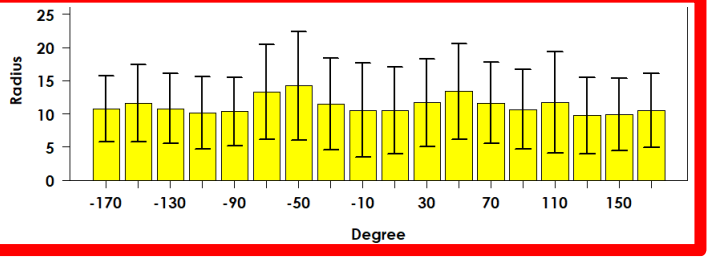


Rm=7.51
Rsd=5.37

decorana







選択

Rm=11.4
Rsd=6.48

テーマ：
機械学習を用いた技術動向分析の試み
～R言語による視覚的な手法を用いた分析～

Ⅰ. 大規模データの等高線マップ作成方法の検討
Ⅱ. AI, ML, DL等の用語含む特許出願の動向分析
Ⅲ. GAFAの米国特許出願の動向分析

- 対象：“AI”, “ML”, “DL”等の用語含む特許出願 26万件
- 内容：時系列動向を等高線マップ等で可視化
  - ① AI等の日米中の特許出願推移
  - ② 世界でのAI, ML, DLの内容の範囲比較
  - ③ 日米中での特許文献の被引用

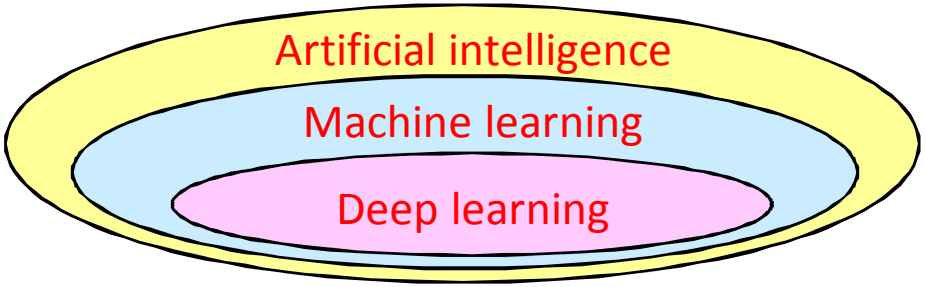
Ⅱ．AI, ML, DL等の用語の違いは？

■「人工知能」と「機械学習 」のどちらの用語が使用される？

- ■人工知能（artificial intelligence, AI）とは、「計算機（コンピュータ）による知的な情報処理システムの設計や実現に関する研究分野」を指す（wikipedia）
- ■機械学習（machine learning）とは、人工知能における研究課題の一つで、人間が自然に行っている学習能力と同様の機能をコンピュータで実現しようとする技術・手法（wikipedia）
- ■深層学習（deep learning）とは、多層のニューラルネットワーク（ディープニューラルネットワーク）による機械学習手法である（wikipedia）。

■日本のニュースでは明確な区別なく使われるが、違いは？日米中では？

機械学習の入門書やWeb記事の説明を参照すると・・・

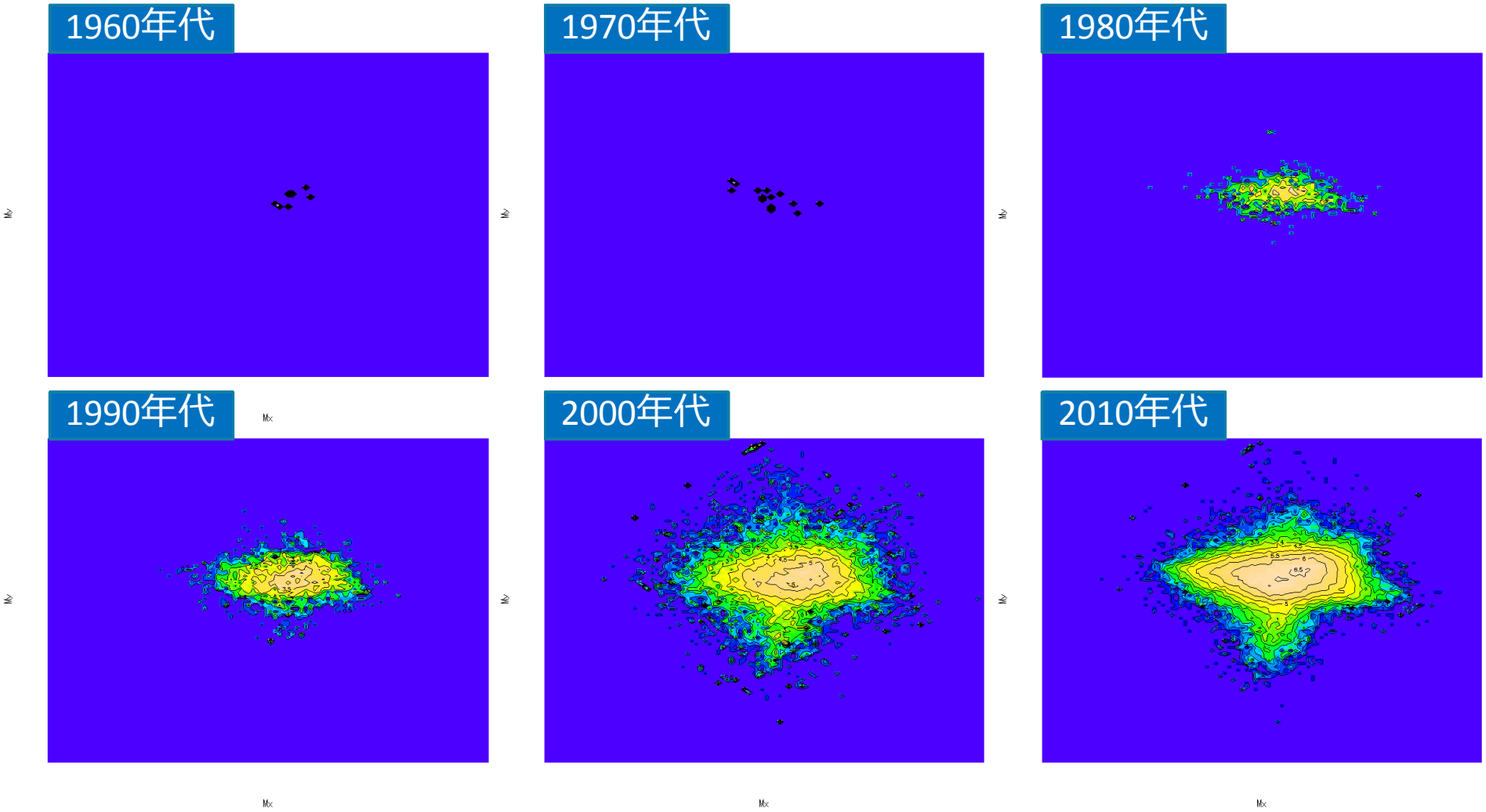


■特許出願の「タイトル」、「要約」のテキストマイニング

- ■等高線マップ上での国ごと（特に、日米中）のプロット範囲の変遷
- ■等高線マップ上でのAI, ML, DLのプロット範囲の変遷

Ⅱ．AI等の用語含む出願の等高線マップ推移

■全世界でAI, ML, DLの用語含む出願（英語文献）

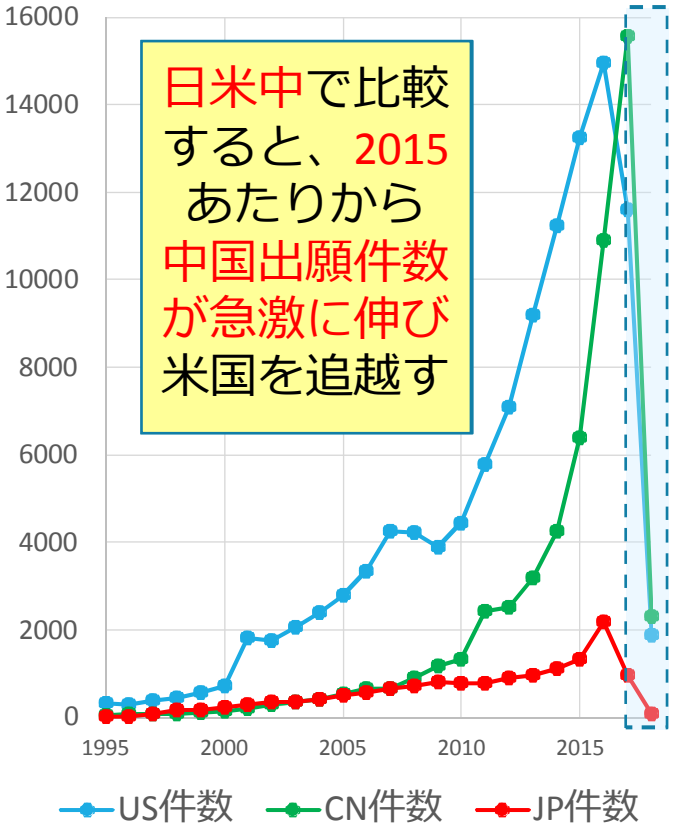


Ⅱ. ① AI等の用語含む特許出願の件数推移

■日米中でAI, ML, DLの用語を明細書中に含む出願（英語文献の全文検索※）
※ノイズ含む（以下同様）

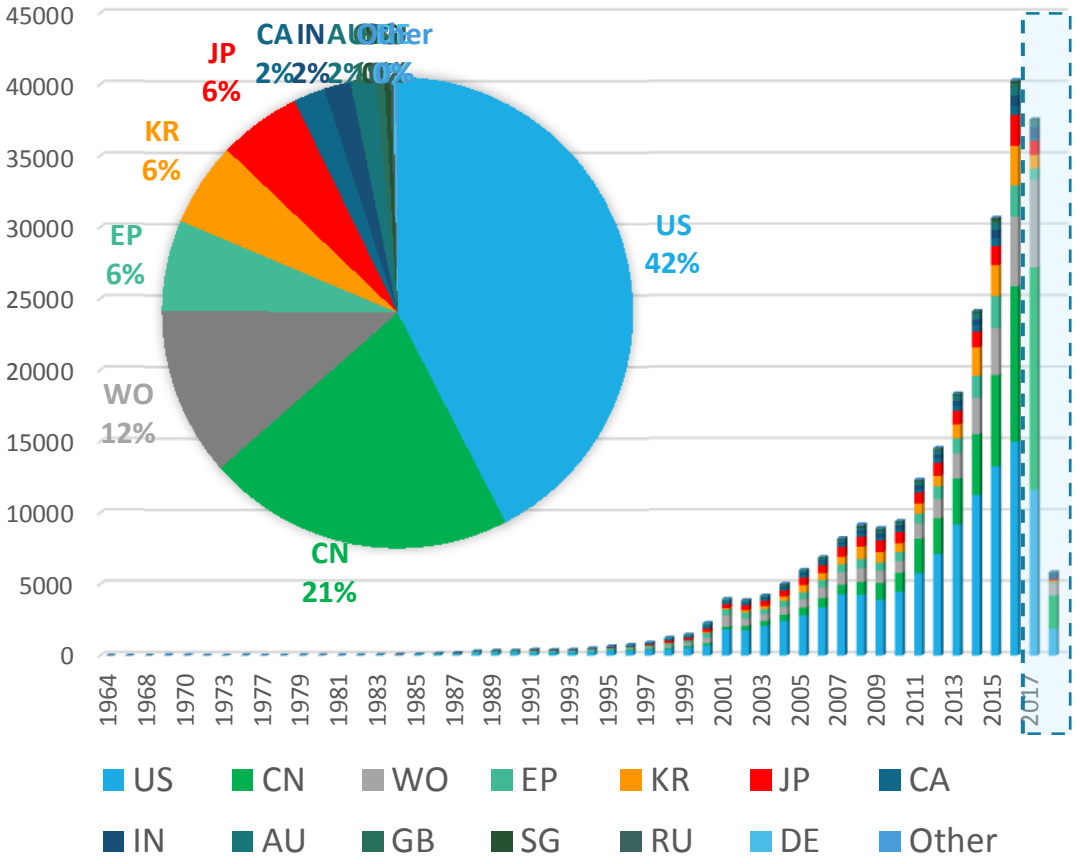
青：US出願　緑：CN出願　赤：JP出願

米国、中国、日本の出願推移



日米中で比較すると、2015あたりから中国出願件数が急激に伸び米国を追越す

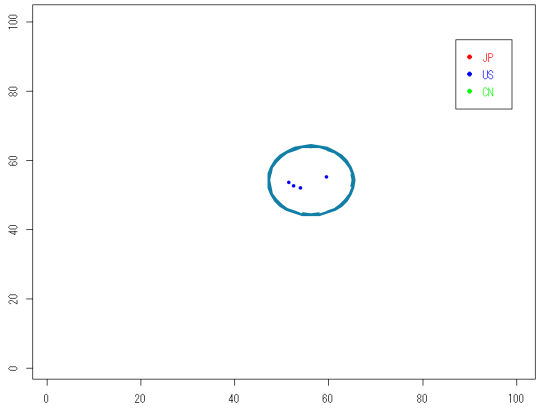
"artificial intelligence", "machine learning"含む特許出願



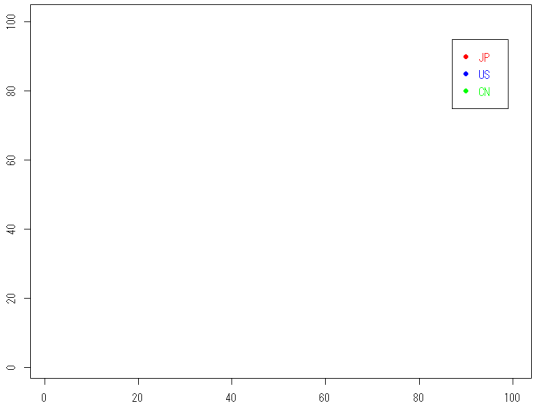
Ⅱ. ① AI等の用語含む日米中のプロット推移

■日米中でAI, ML, DLの用語を明細書中に含む出願（英語文献）

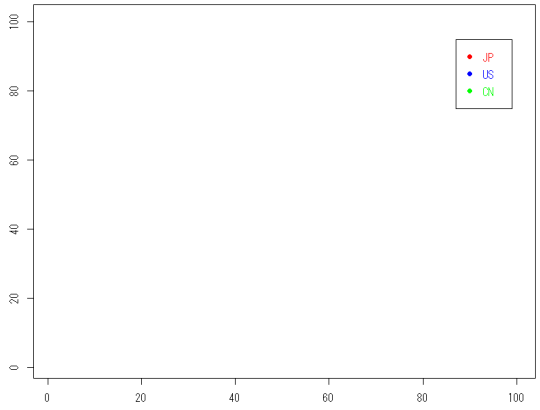
1960年代　青：US出願



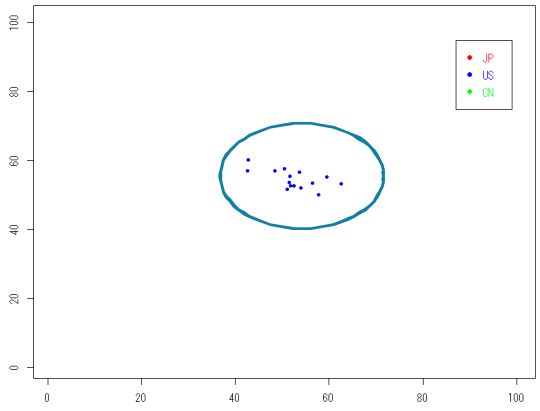
1960年代　緑：CN出願



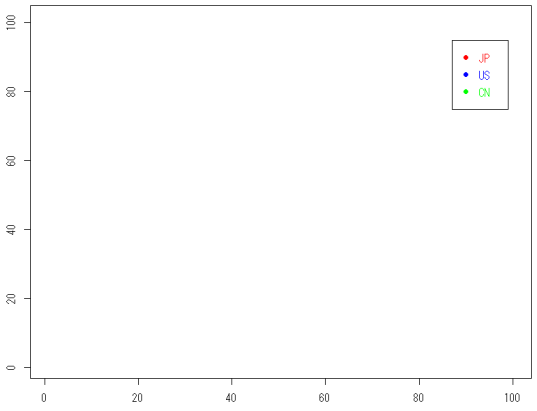
1960年代　赤：JP出願



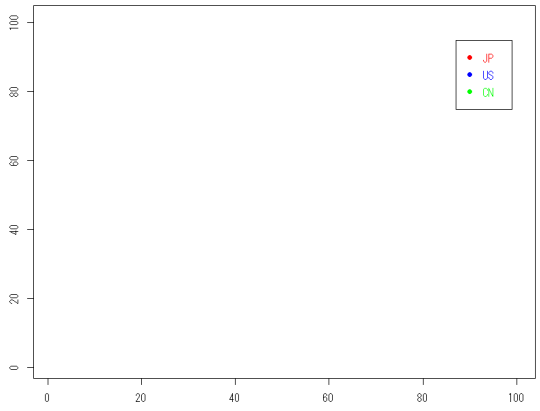
1970年代　青：US出願



1970年代　緑：CN出願



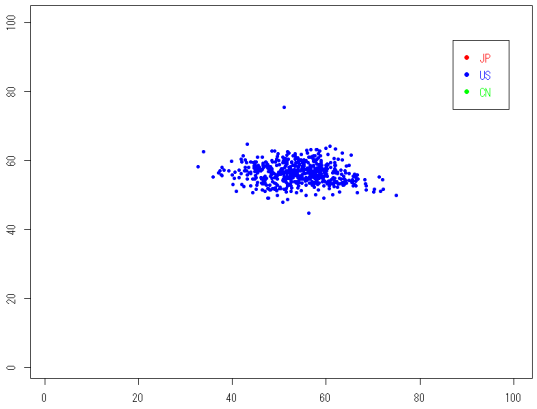
1970年代　赤：JP出願



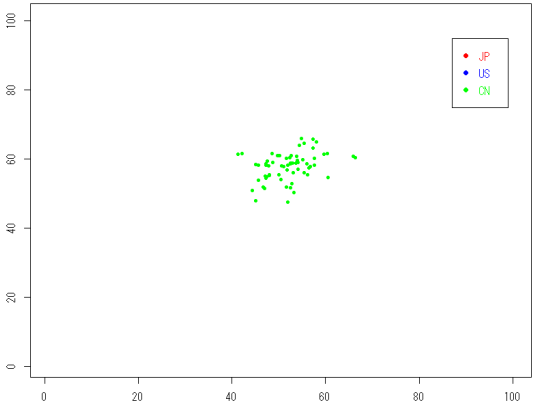
Ⅱ. ① AI等の用語含む日米中のプロット推移

■日米中でAI, ML, DLの用語を明細書中に含む出願（英語文献）

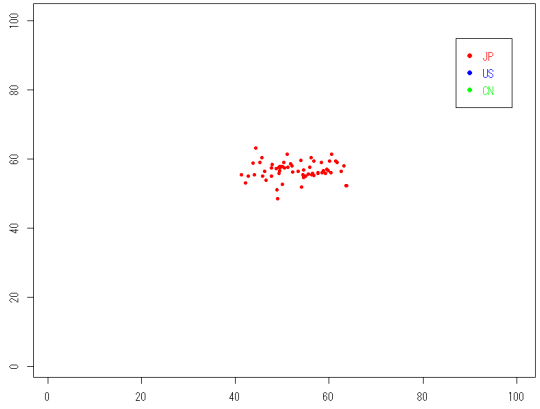
1980年代　青：US出願



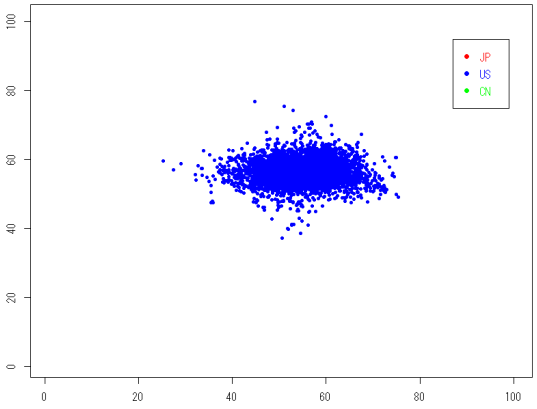
1980年代　緑：CN出願



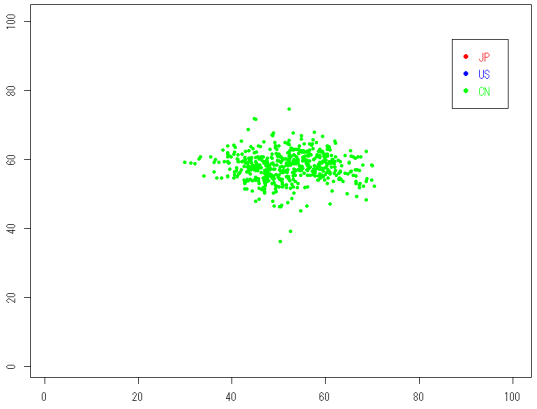
1980年代　赤：JP出願



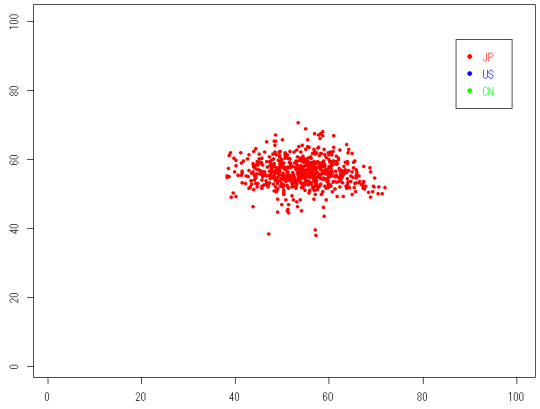
1990年代　青：US出願



1990年代　緑：CN出願



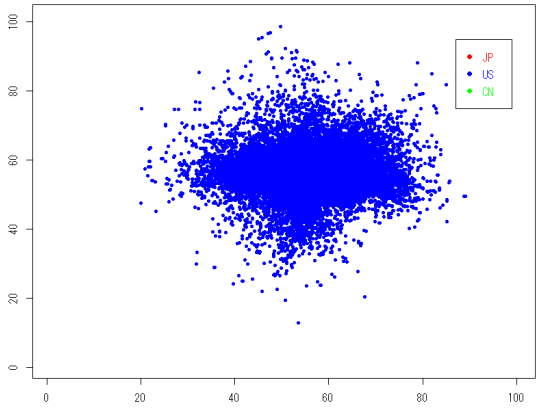
1990年代　赤：JP出願



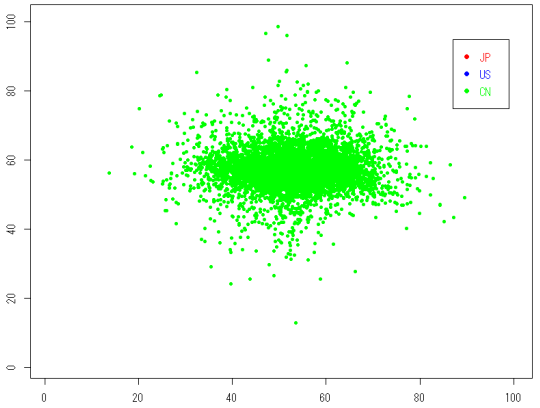
Ⅱ. ① AI等の用語含む日米中のプロット推移

■日米中でAI, ML, DLの用語を明細書中に含む出願（英語文献）

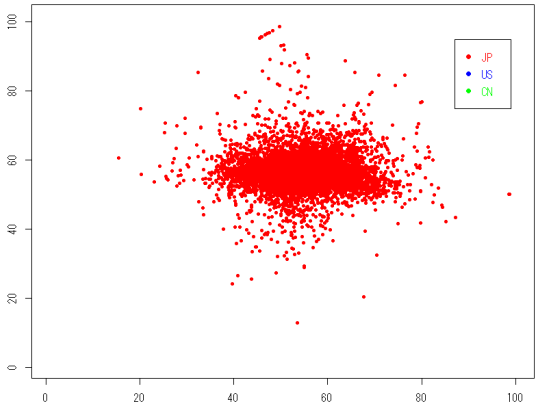
2000年代　青：US出願



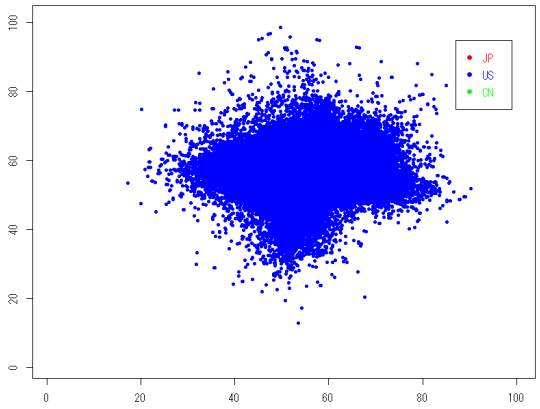
2000年代　緑：CN出願



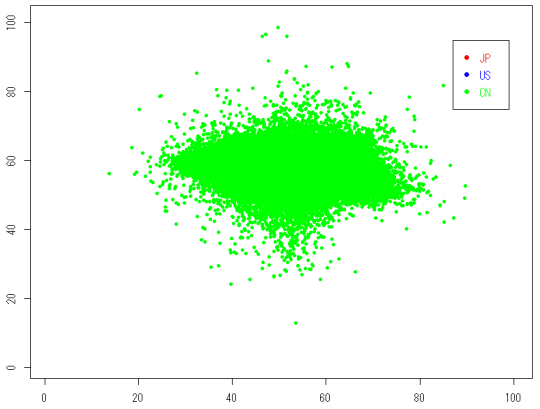
2000年代　赤：JP出願



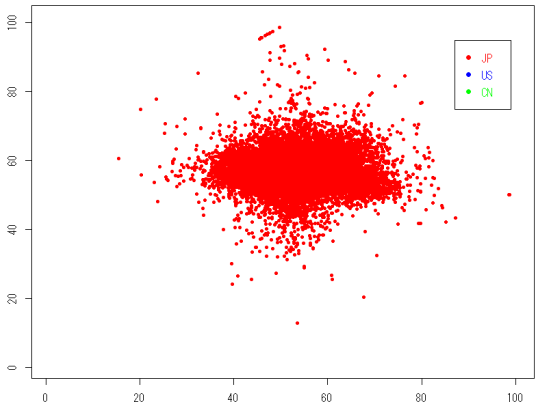
2010年代　青：US出願



2010年代　緑：CN出願



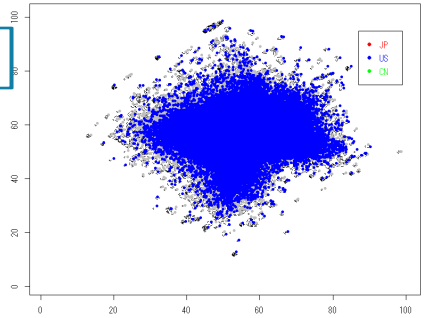
2010年代　赤：JP出願



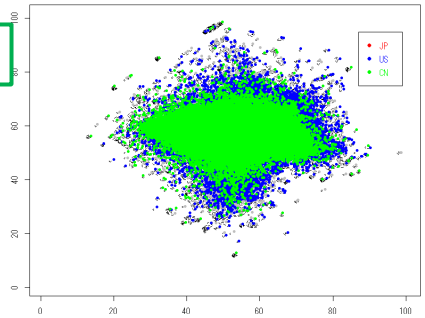
Ⅱ.① AI等の用語含む日米中のプロット範囲

■日米中でAI, ML, DLの用語を明細書中に含む出願（英語文献）

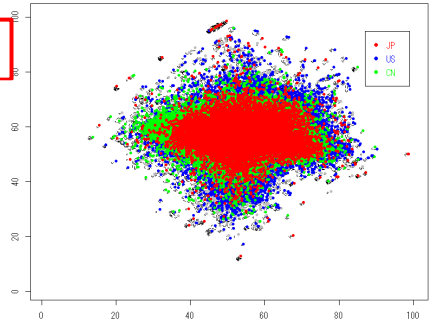
青：US出願

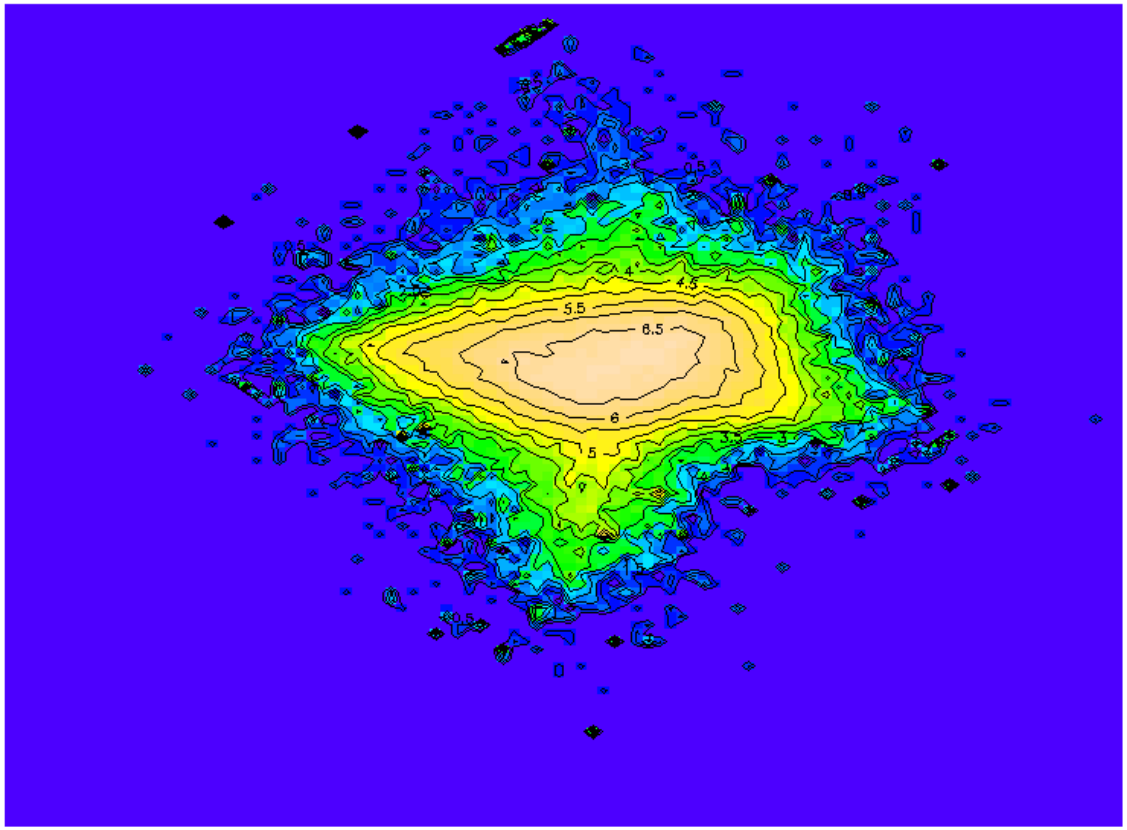


緑：CN出願



赤：JP出願



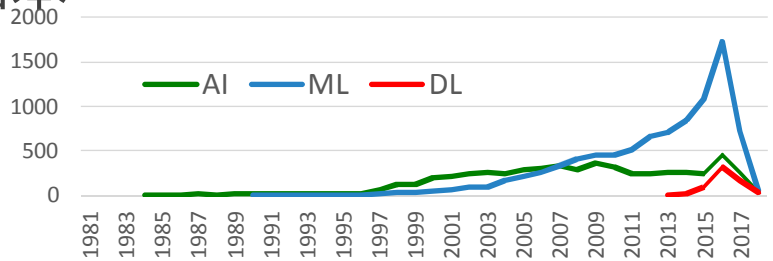


US出願が一番広い範囲、JP出願は一番狭い範囲

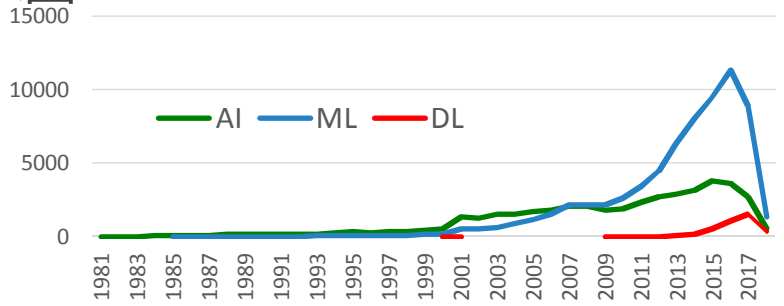
Ⅱ.② 世界でのAI, ML, DL毎の出願件数推移

■全文検索で”AI”, “ML”, “DL”を明細書中に含む日米中の特許出願推移

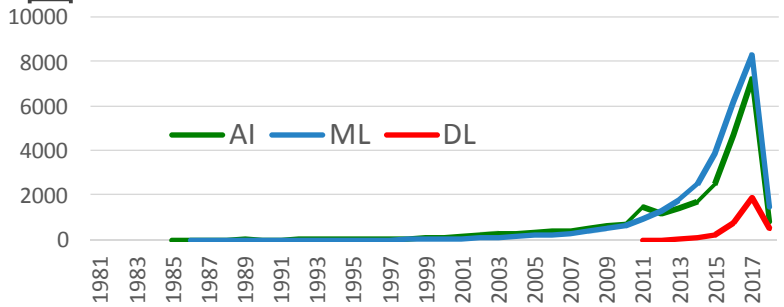
➢日本

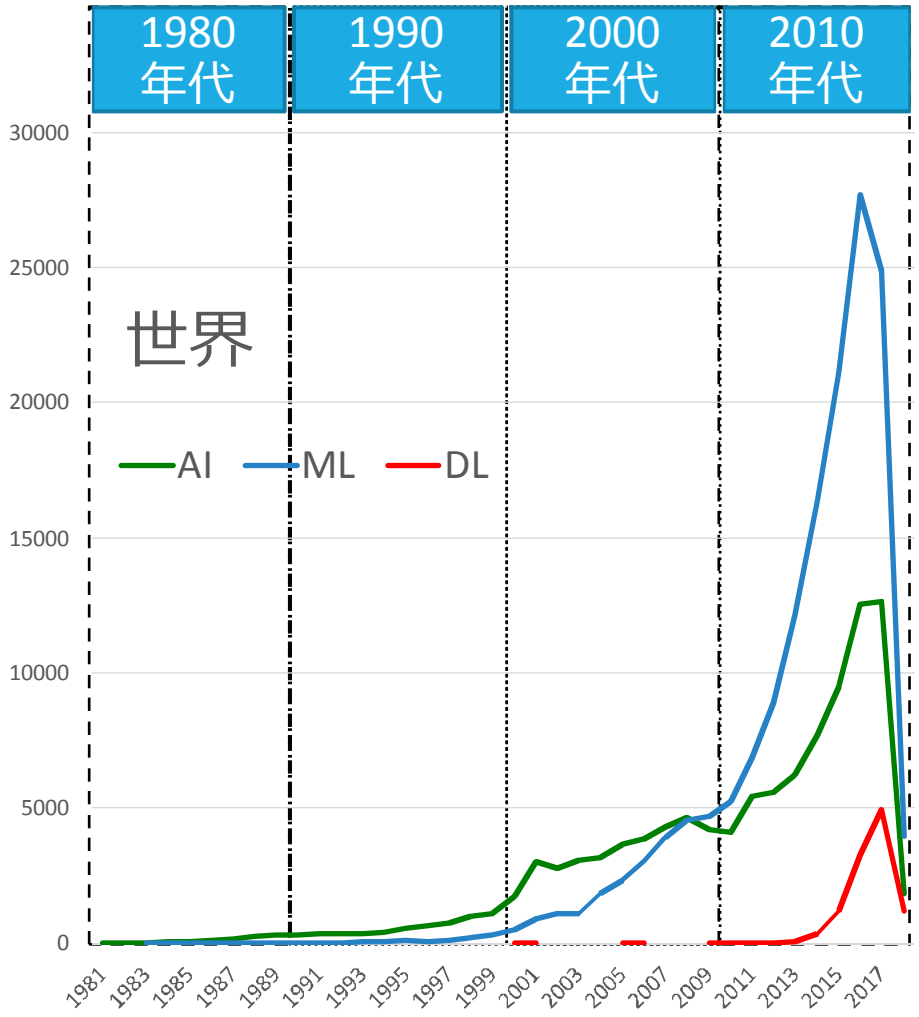


➢米国



➢中国

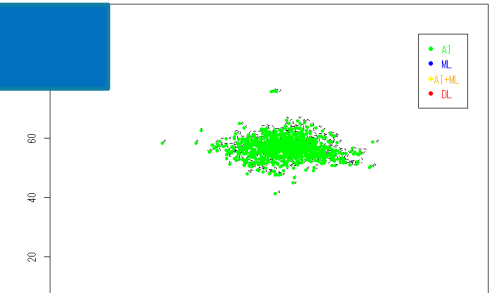




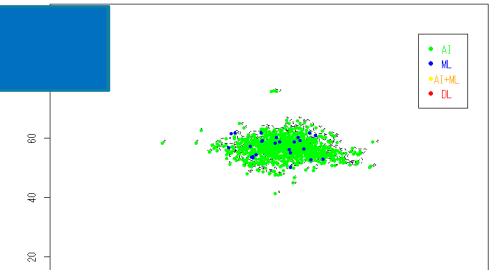
Ⅱ.② 世界でのAI, ML, DLの内容の範囲比較

■1980年代（ AI, ML, DLの用語を明細書中に含む出願）の色分けプロット

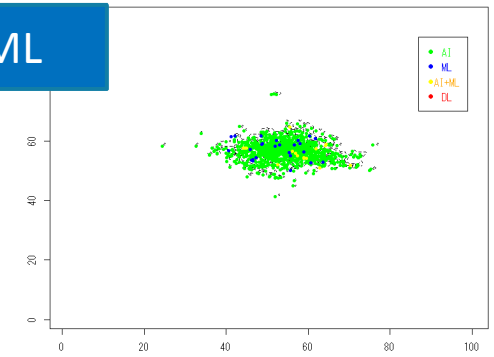
AI



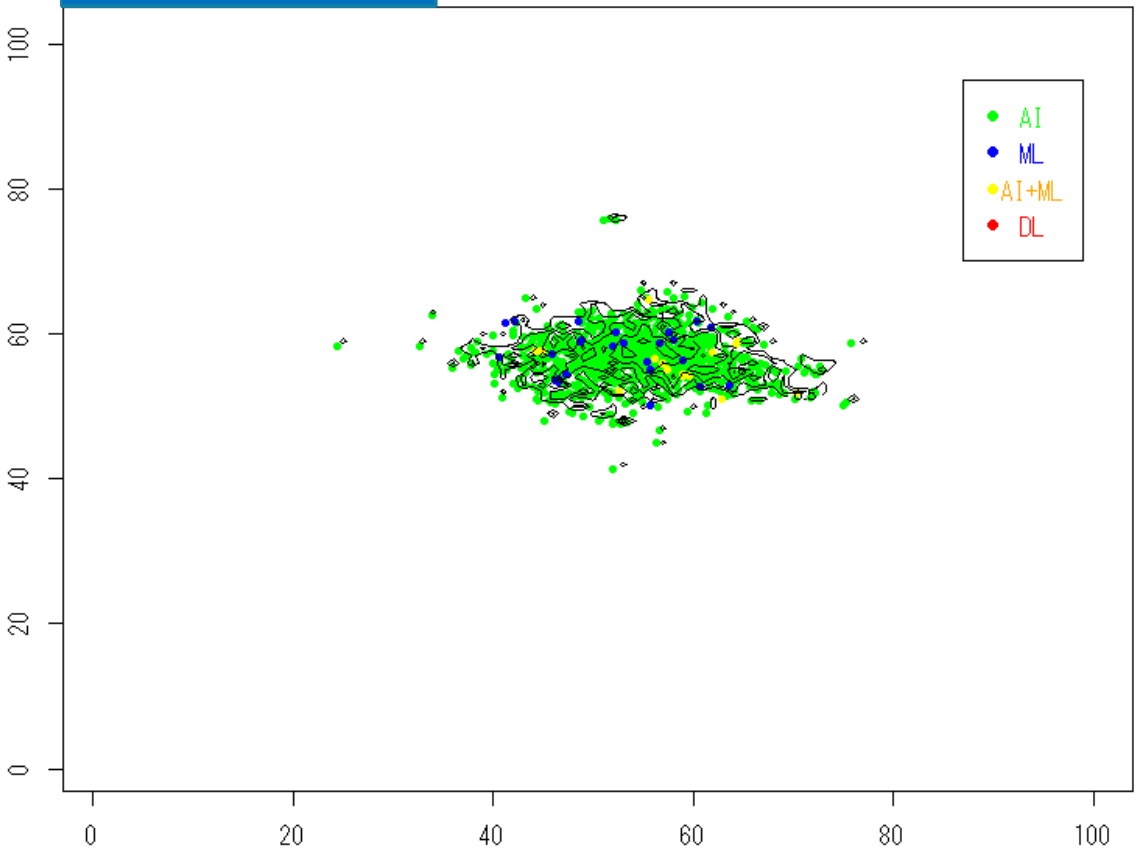
ML



AI & ML



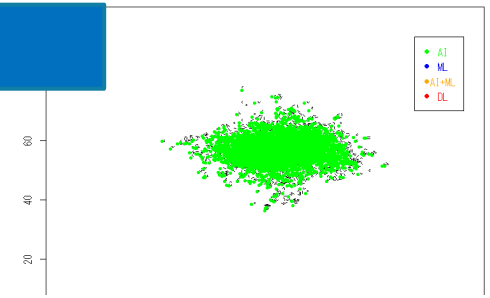
(AI or ML) & DL



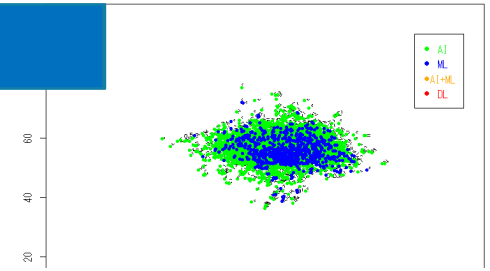
Ⅱ.② 世界でのAI, ML, DLの内容の範囲比較

■1990年代（ AI, ML, DLの用語を明細書中に含む出願）の色分けプロット

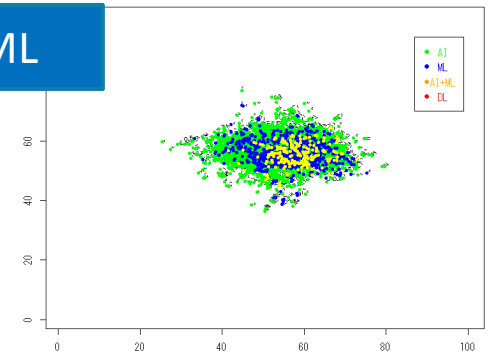
AI



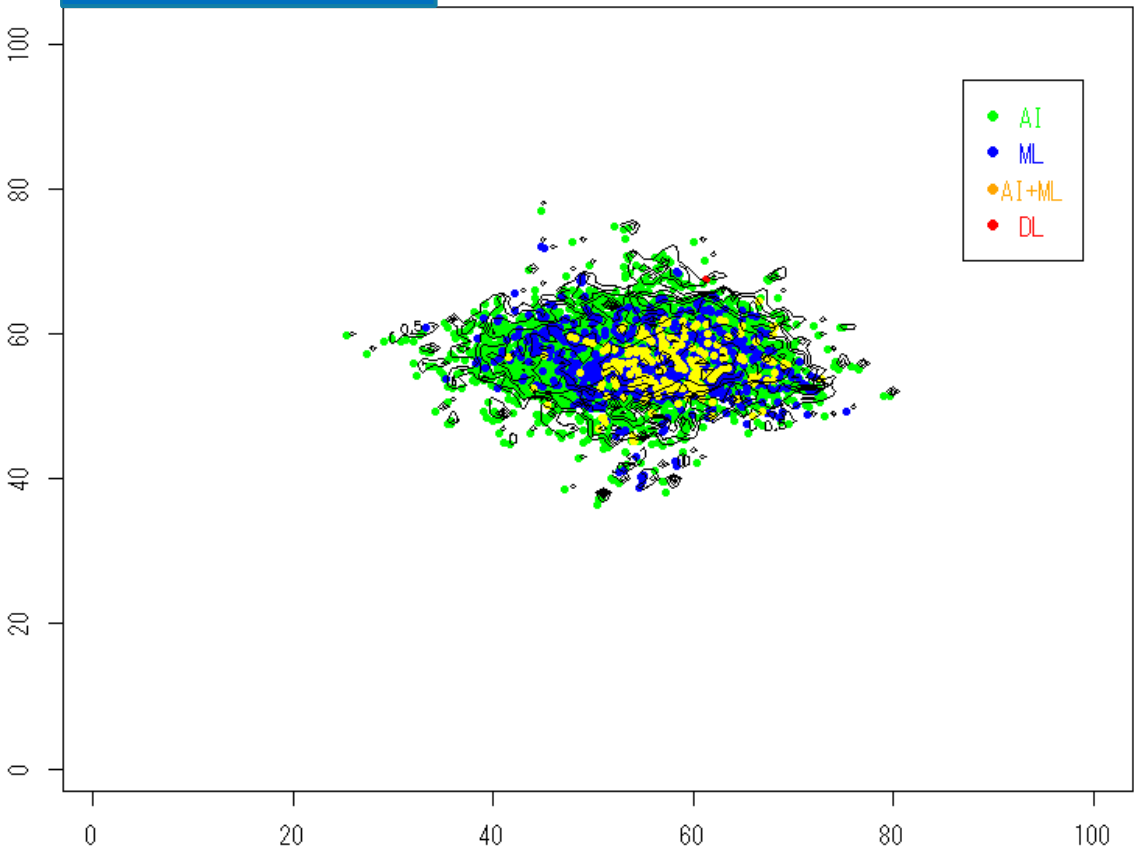
ML



AI & ML



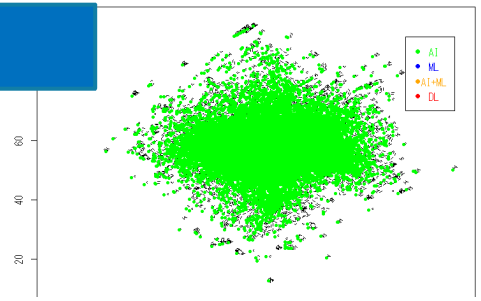
(AI or ML) & DL



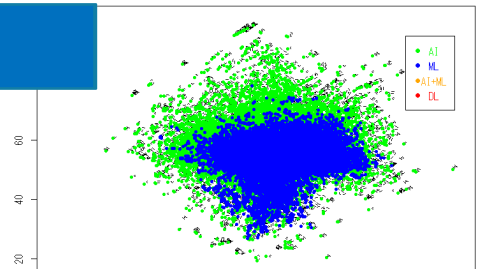
Ⅱ.② 世界でのAI, ML, DLの内容の範囲比較

■2000年代（ AI, ML, DLの用語を明細書中に含む出願）の色分けプロット

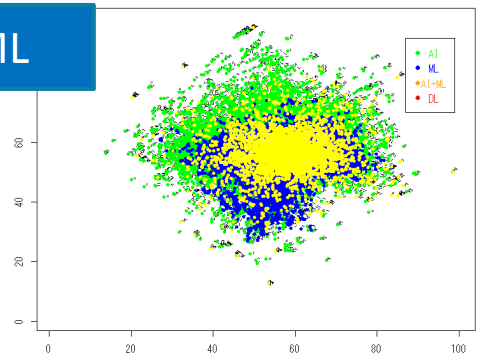
AI



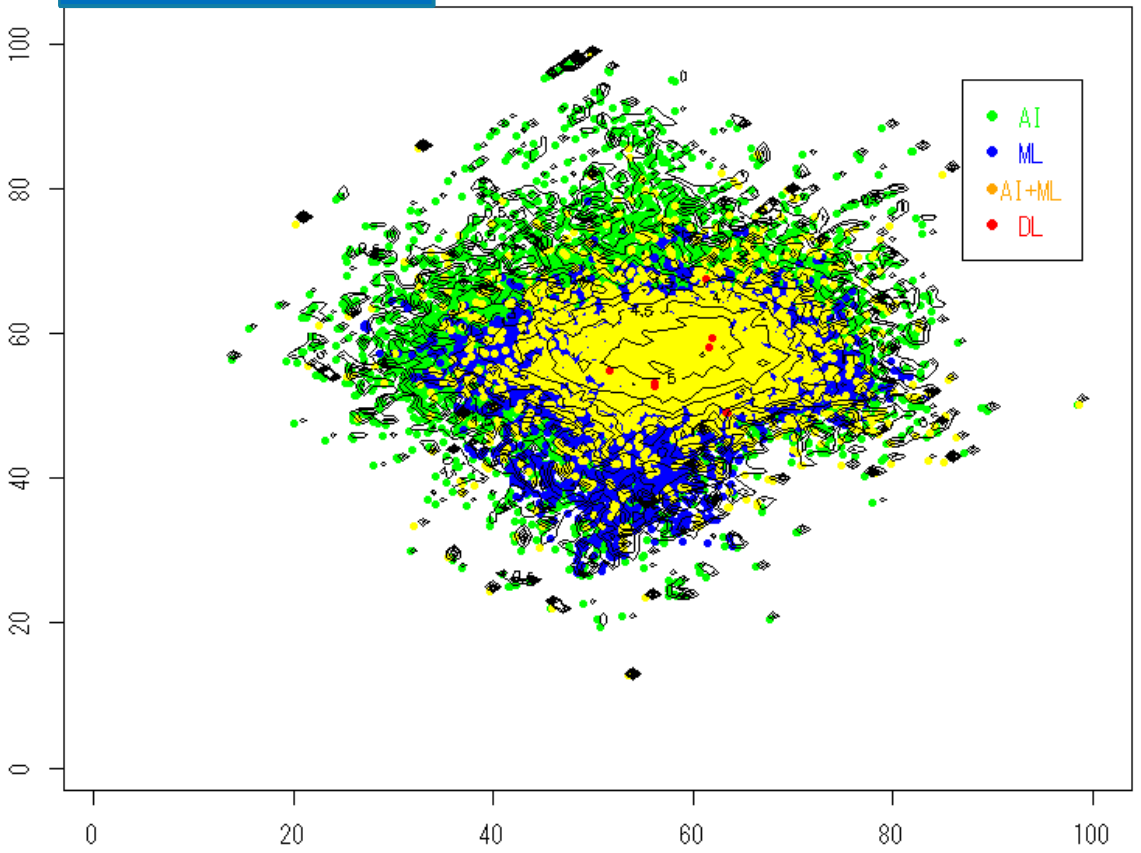
ML



AI & ML



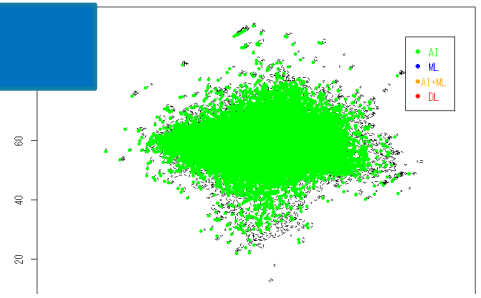
(AI or ML) & DL



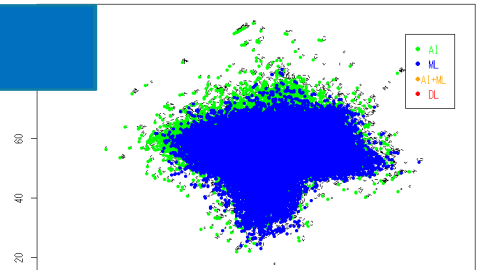
Ⅱ.② 世界でのAI, ML, DLの内容の範囲比較

■2010年代（ AI, ML, DLの用語を明細書中に含む出願）の色分けプロット

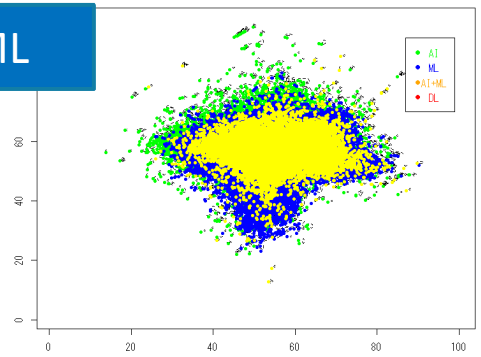
AI



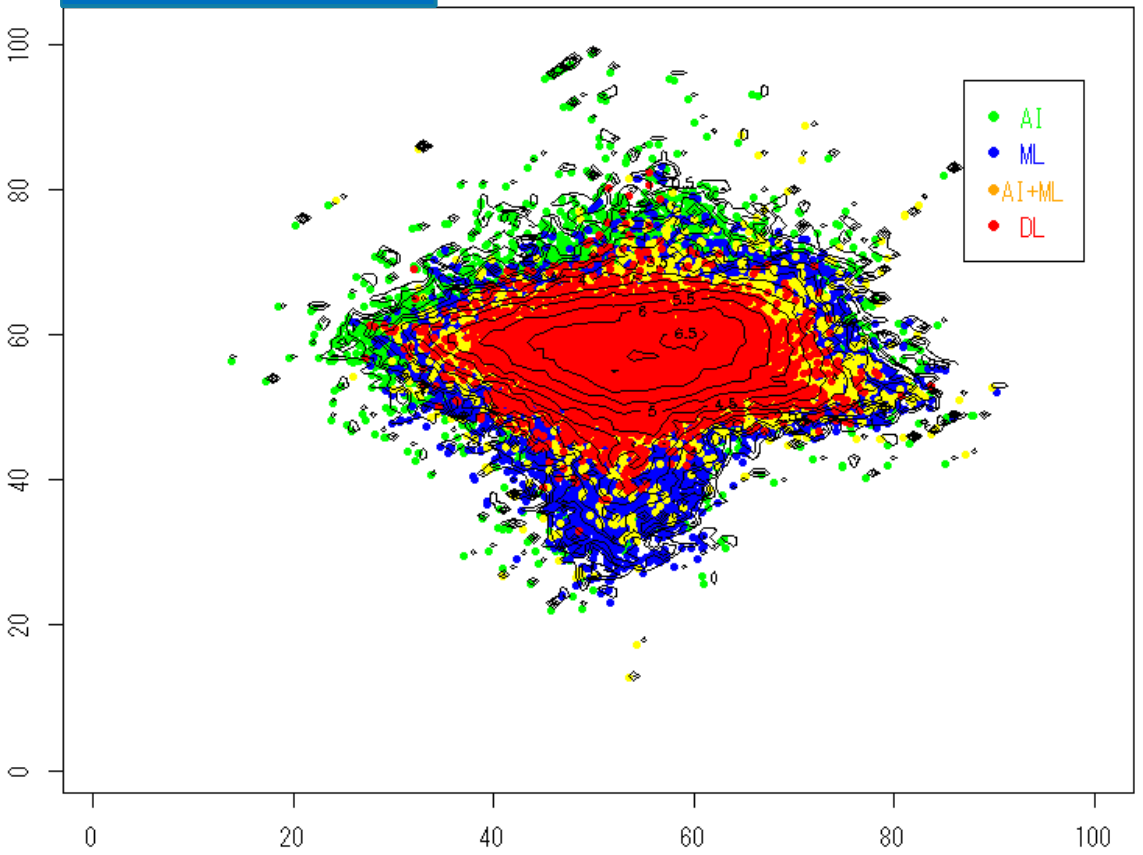
ML



AI & ML



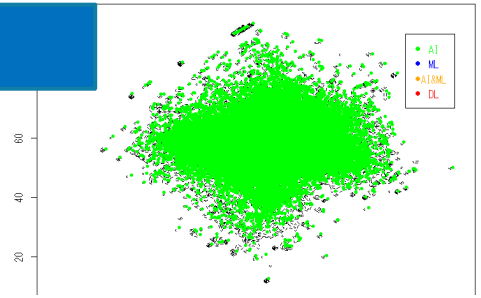
(AI or ML) & DL



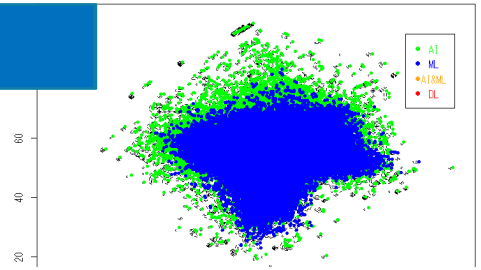
Ⅱ.② 世界でのAI, ML, DLの内容の範囲比較

■全年代（ AI, ML, DLの用語を明細書中に含む出願）の色分けプロット

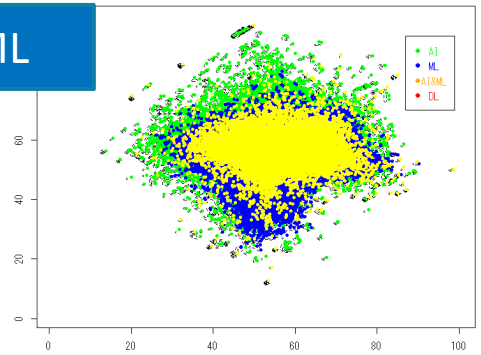
AI



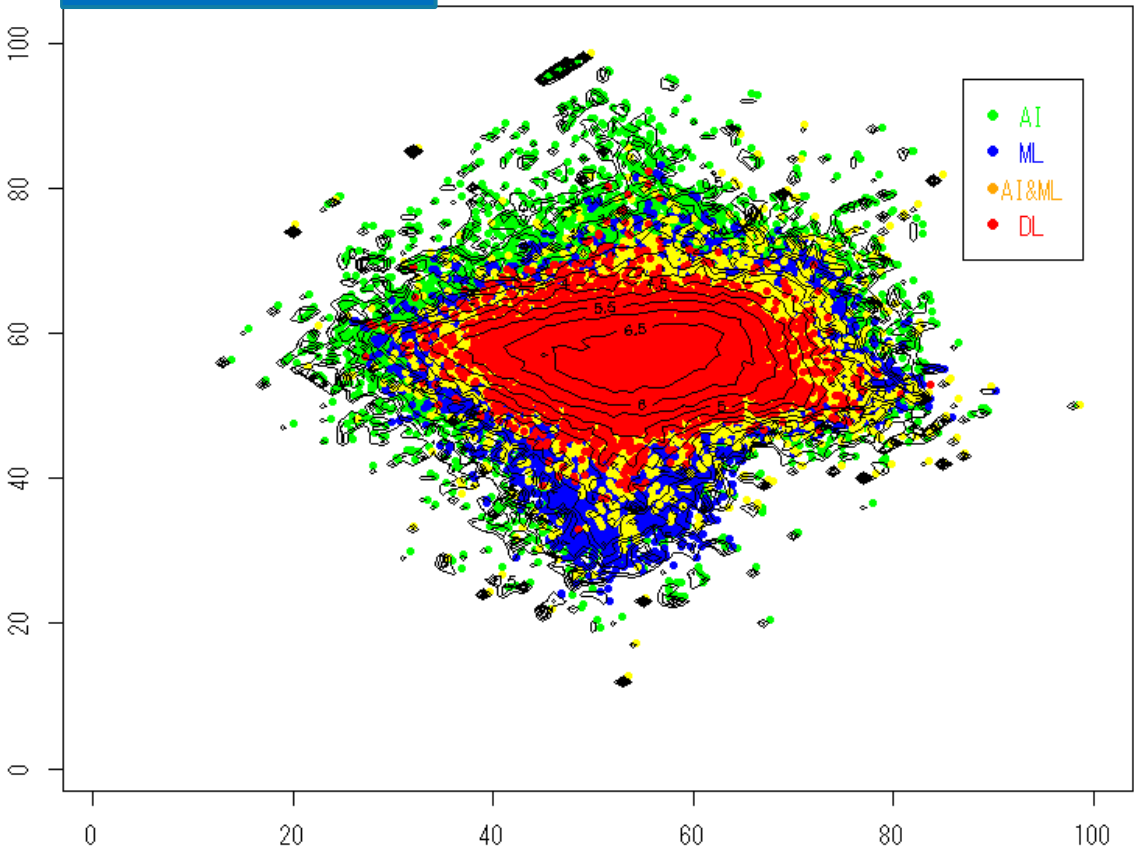
ML



AI & ML



(AI or ML) & DL

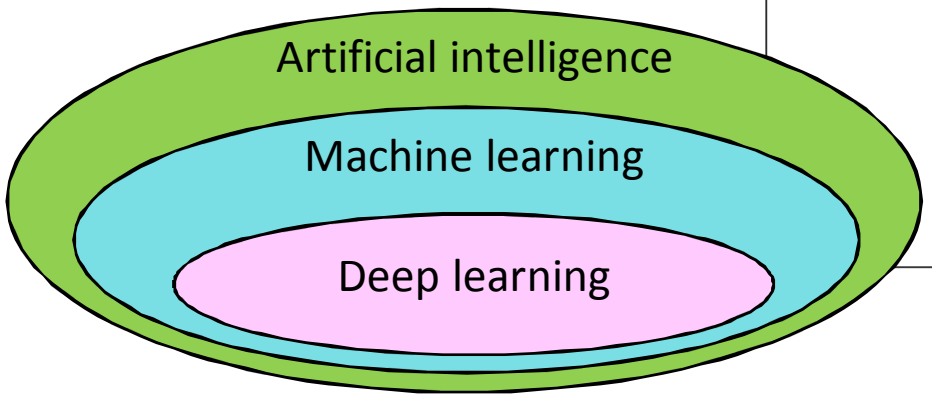


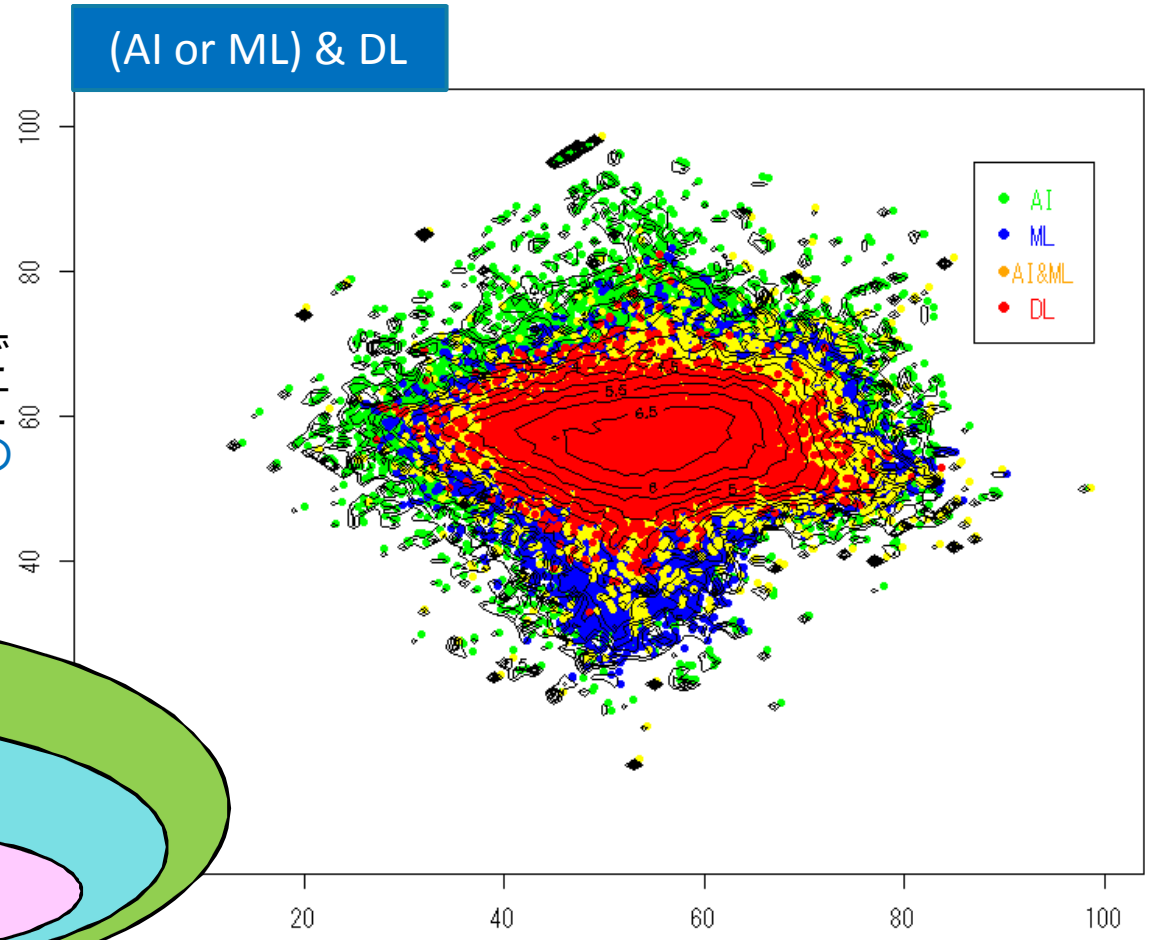
Ⅱ.② 世界でのAI, ML, DLの内容の範囲比較

■全年代（ AI, ML, DLの用語を明細書中に含む出願）の色分けプロット

■今回調査した用語：
“Artificial Intelligence”
“Machine Learning”
“Deep Learning”

■一般に言われる各用語の概念の包含関係と右の図で示される各用語が使われている特許出願の技術内容のplot範囲の包含関係が対応

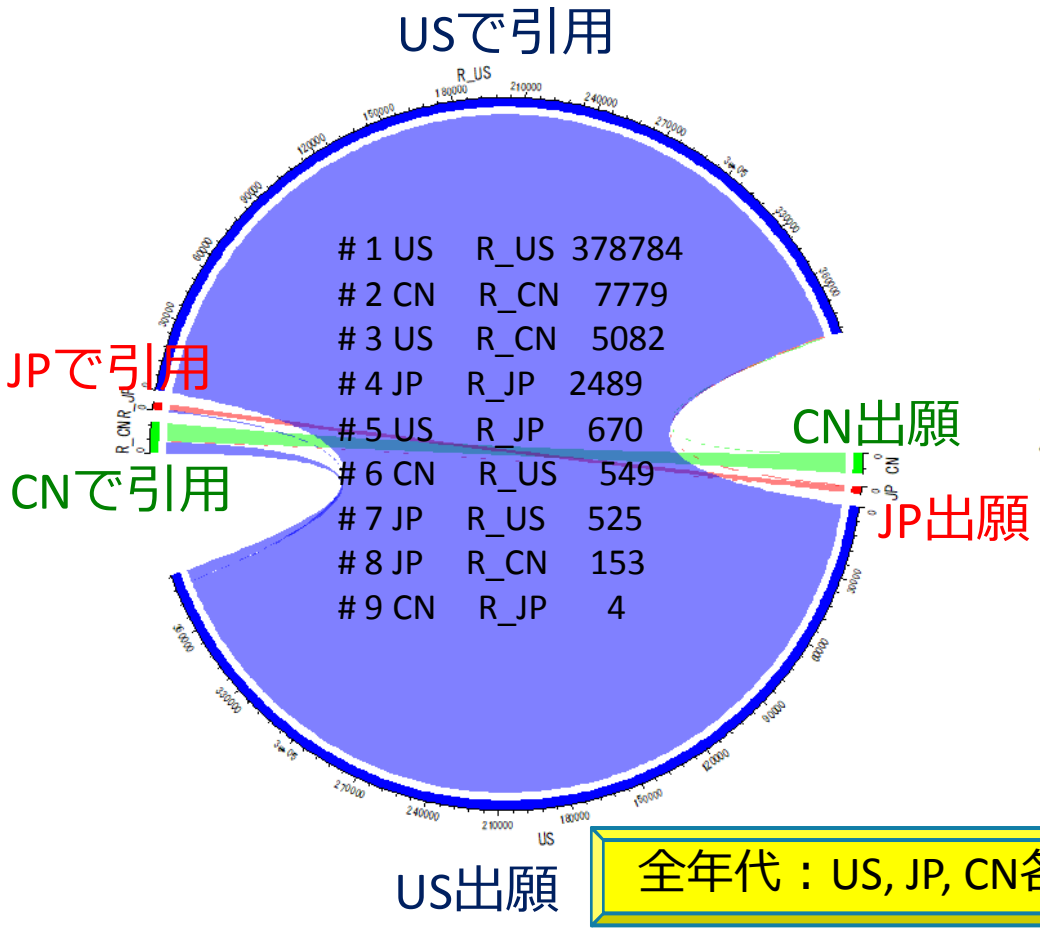




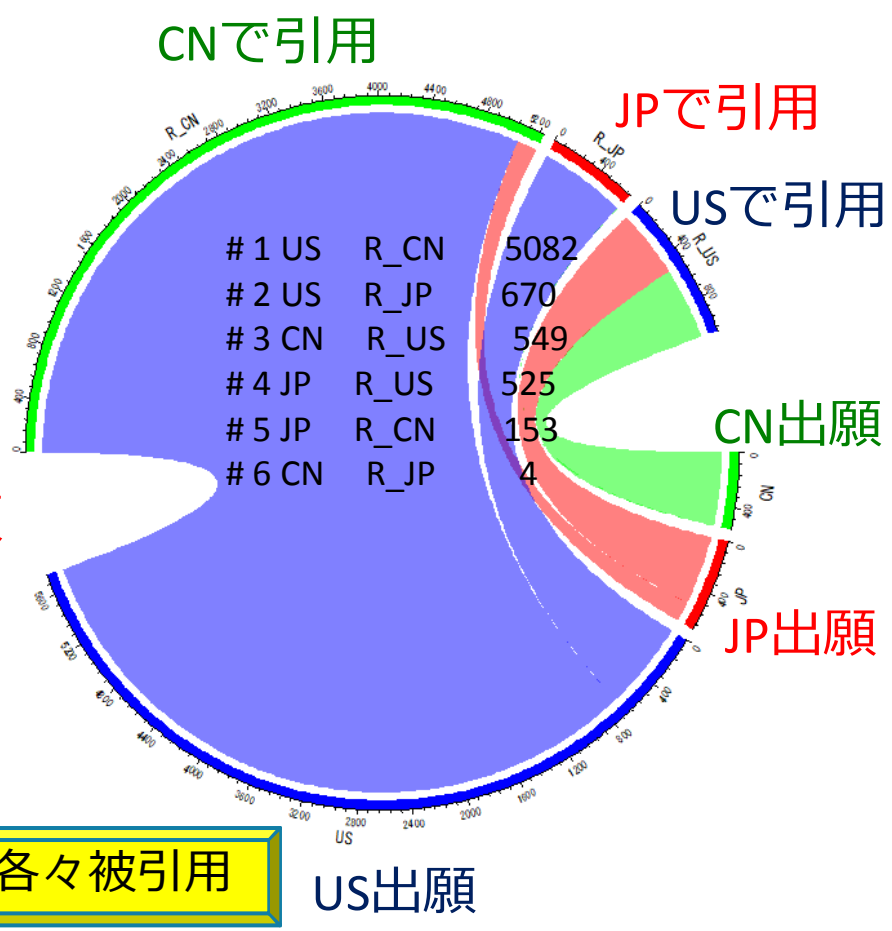
Ⅱ.③ 日米中での特許文献の被引用

■特許文献の被引用（「被引用特許（フォワード）」）

全年代の被引用（自国引用含む）



全年代の被引用（他国引用のみ）

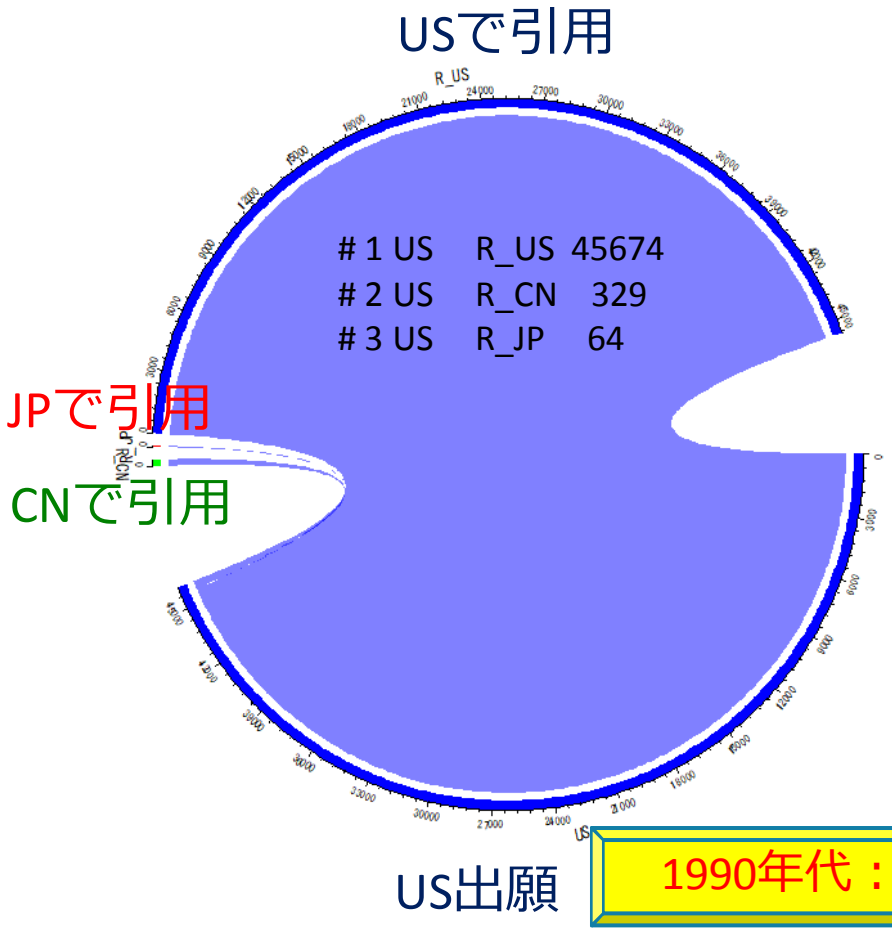


全年代：US, JP, CN各々被引用

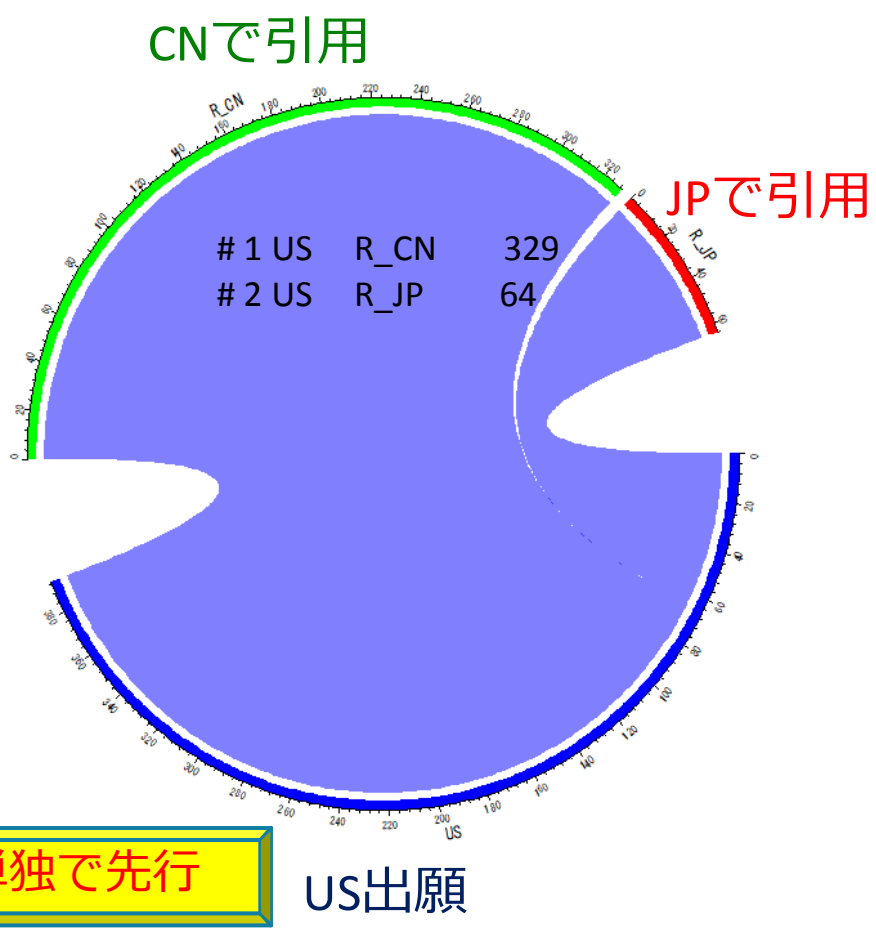
Ⅱ.③ 日米中での特許文献の被引用

■特許文献の被引用（「被引用特許（フォワード）」）

1990年代の被引用（自国引用含む）



1990年代の被引用（他国引用のみ）

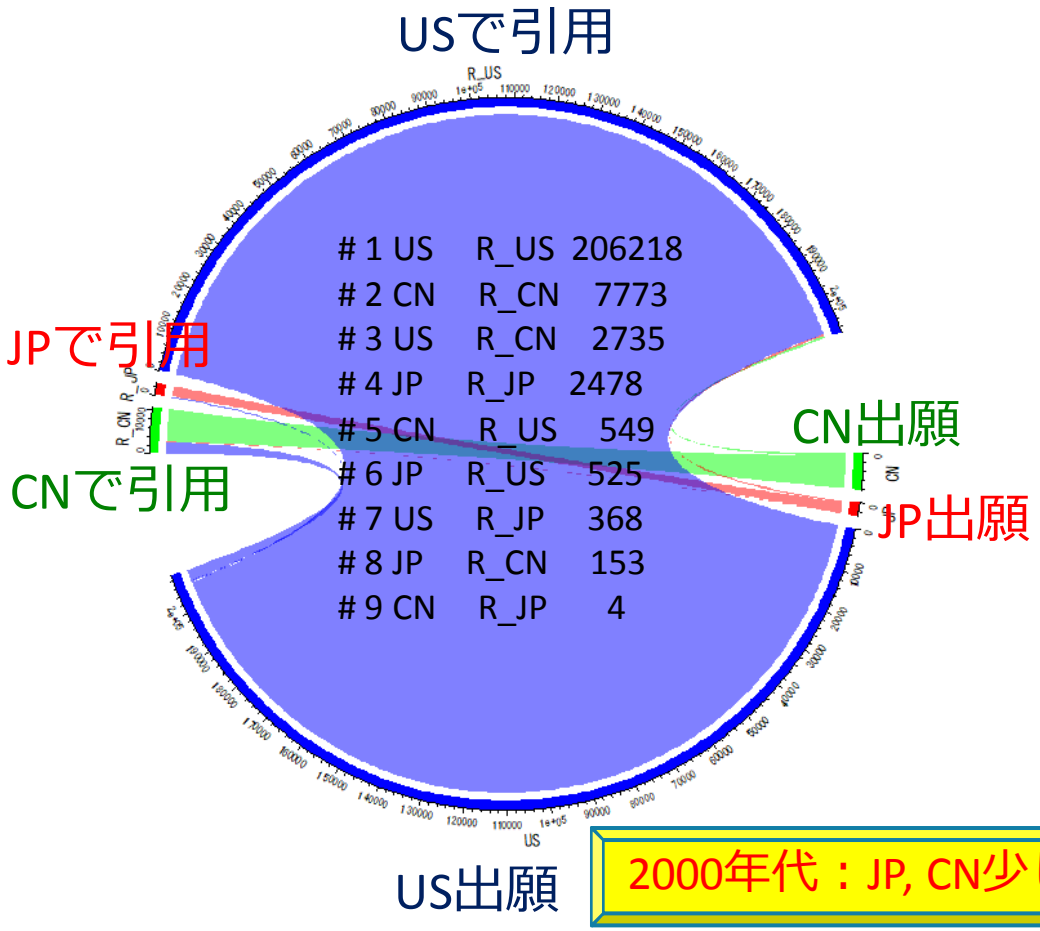


1990年代：米国単独で先行

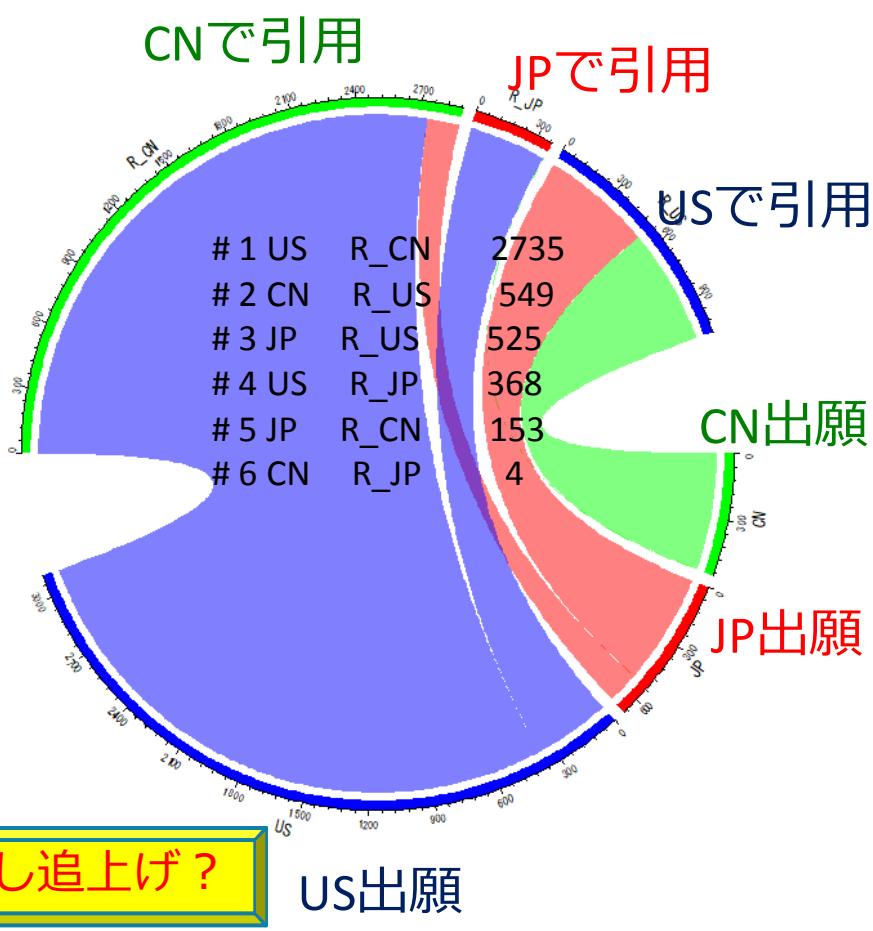
Ⅱ.③ 日米中での特許文献の被引用

■特許文献の被引用（「被引用特許（フォワード）」）

2000年代の被引用（自国引用含む）



2000年代の被引用（他国引用のみ）

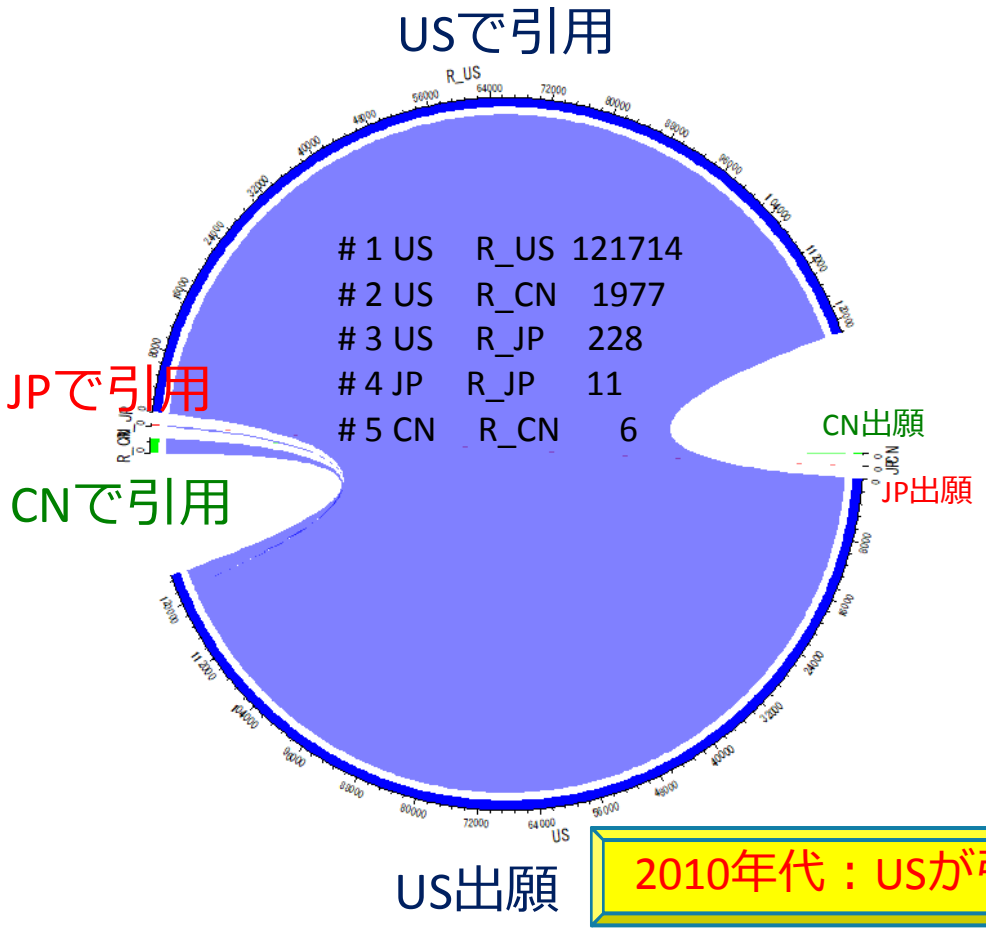


2000年代：JP, CN少し追上げ？

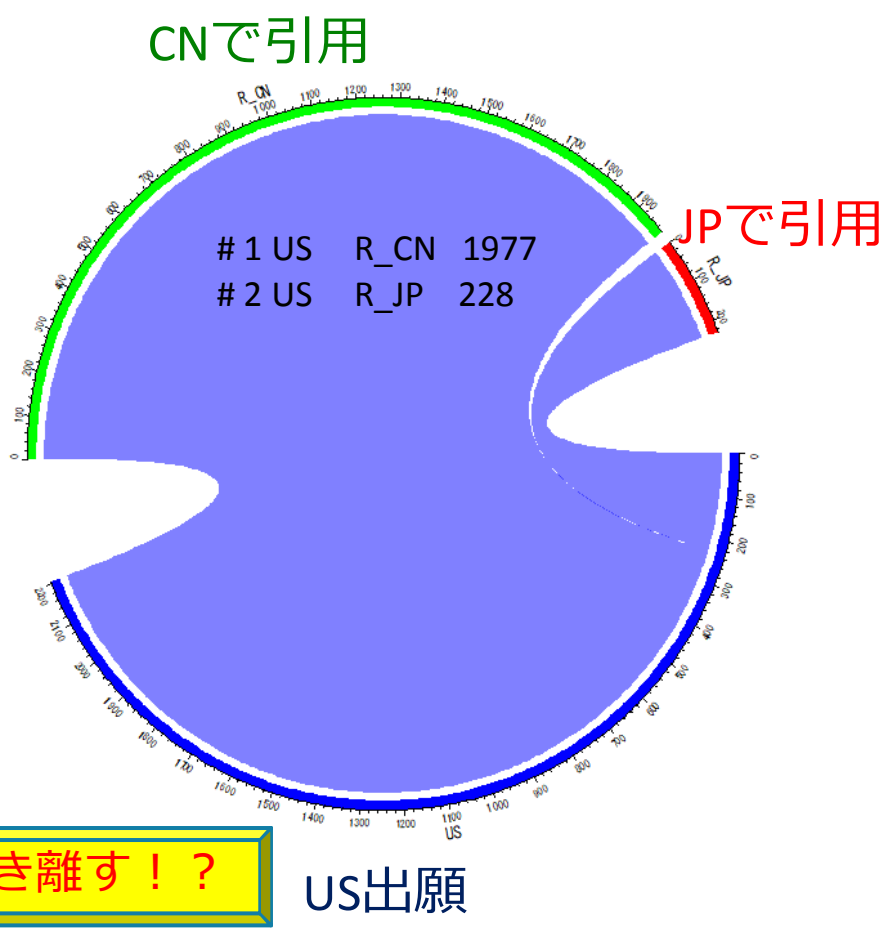
Ⅱ.③ 日米中での特許文献の被引用

■特許文献の被引用（「被引用特許（フォワード）」）

2010年代の被引用（自国引用含む）



2010年代の被引用（他国引用のみ）



2010年代：USが引き離す！？

テーマ：

機械学習を用いた技術動向分析の試み
～R言語による視覚的な手法を用いた分析～

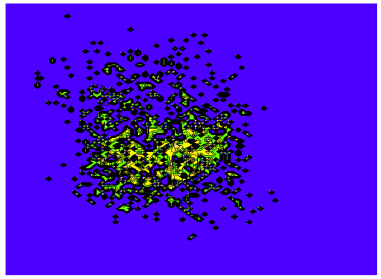
Ⅰ. 大規模データの等高線マップ作成方法の検討
Ⅱ. AI, ML, DL等の用語含む特許出願の動向分析
Ⅲ. GAFAの米国特許出願の動向分析

- 対象：GAFA（GOOGLE, APPLE, FACEBOOK, AMAZON）
- 内容：重心の推移による動向分析

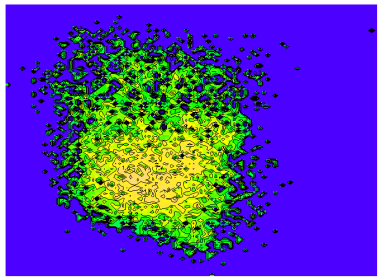
Ⅲ. GAFAの米国特許出願の等高線マップ

■GAFAの米国特許出願（英語文献、90,163件）の等高線マップ

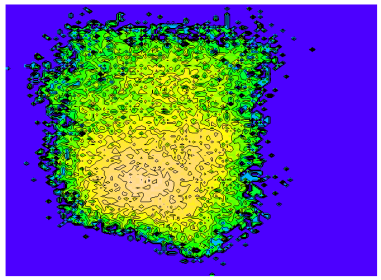
1990-1999出願



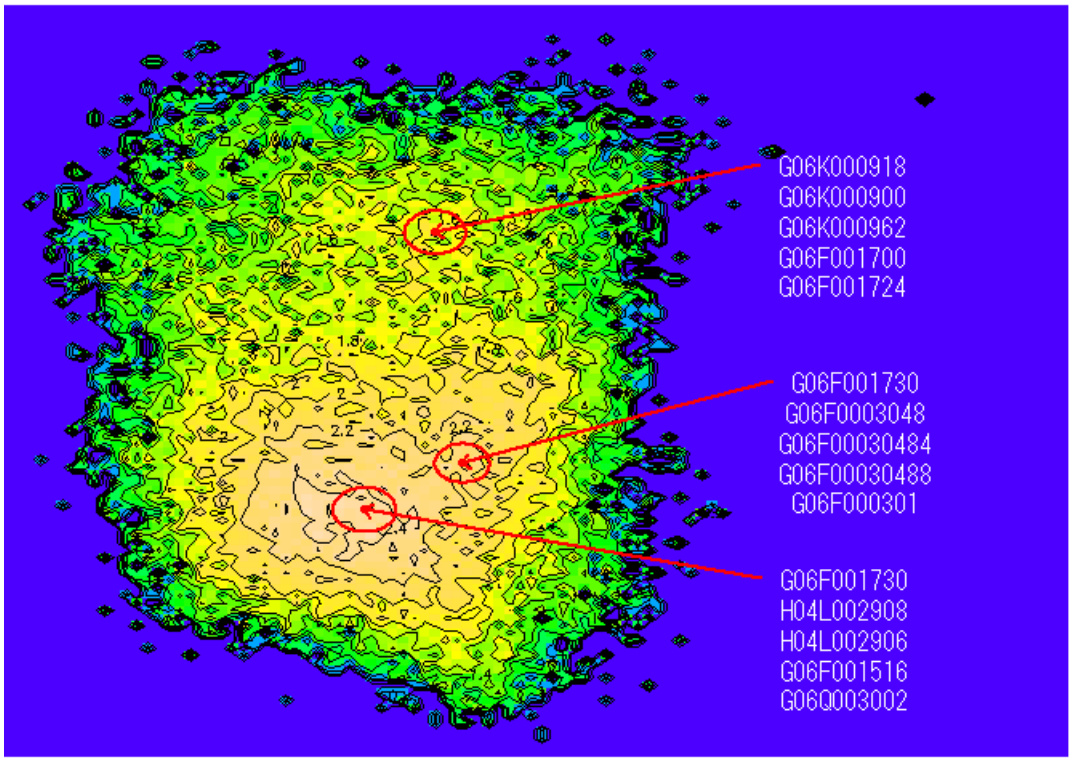
2000-2009出願



2010-2018出願

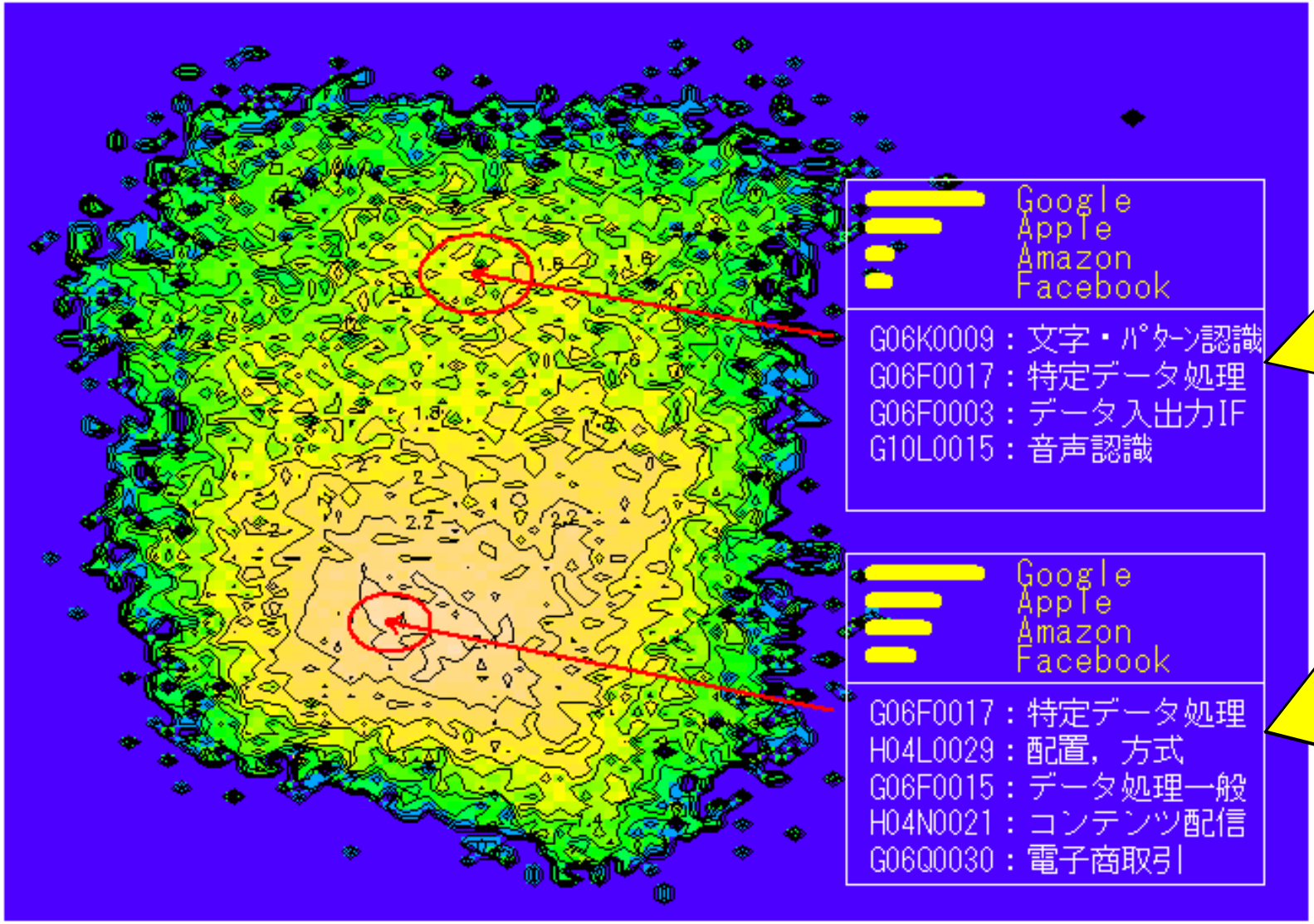


（対応分析：vegan::decorana, sparce=0.99）



（マウス選択領域に含まれる出願のIPCを抽出）

Ⅲ. GAFAの米国特許出願の等高線マップ



- 文字・パターン認識
- データ処理
- 入出力IF
- 音声認識

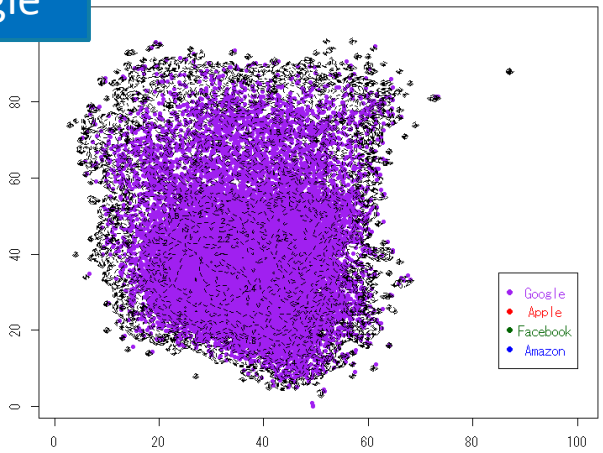
- データ処理
- 配置方式
- コンテンツ配信
- 電子商取引

※IPC上位8bit（メイングループ）

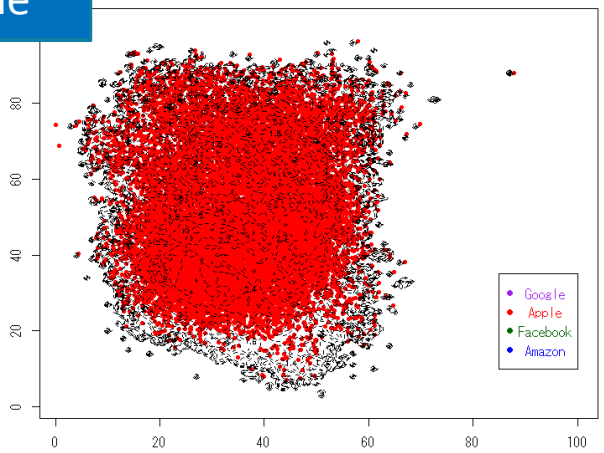
Ⅲ. GAFAの企業毎のプロット範囲

■GAFAの個別企業の等高線マップ上でのプロット

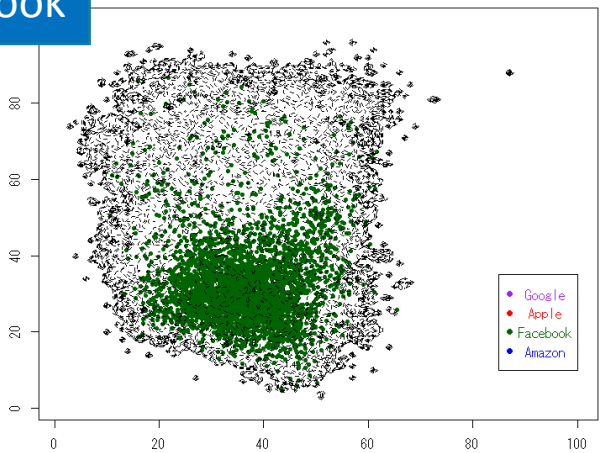
Google



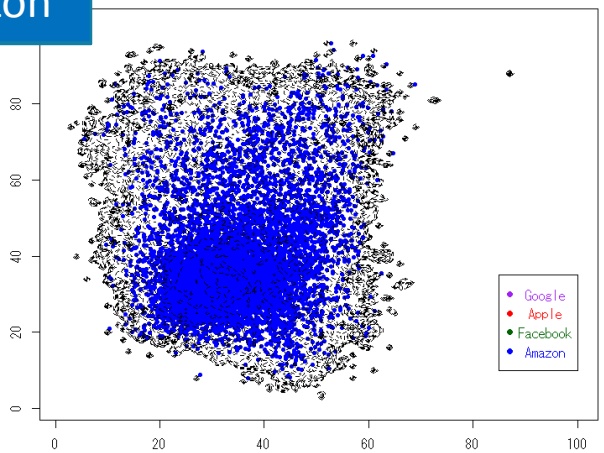
Apple



Facebook



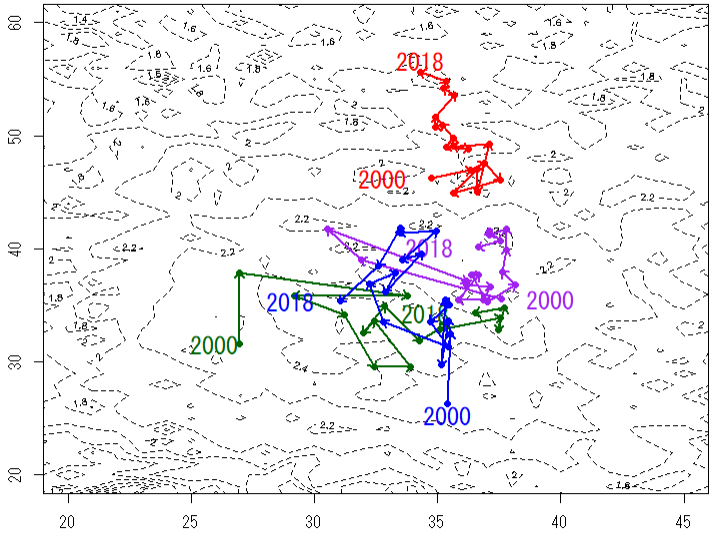
Amazon



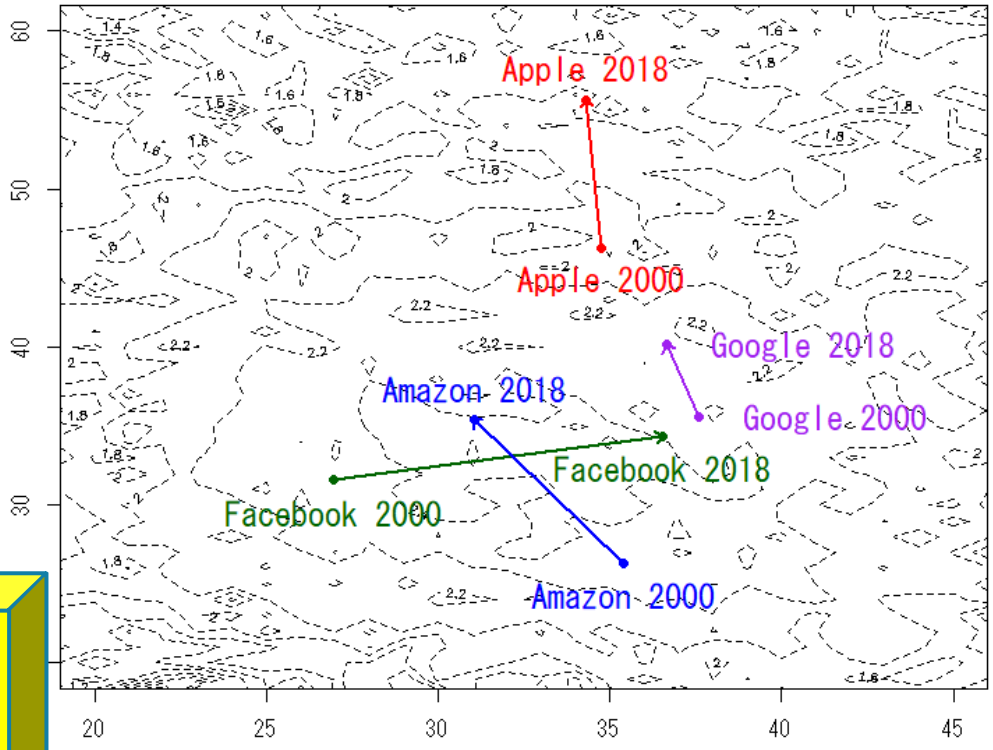
Ⅲ. GAFAの企業毎の重心点の推移

■GAFAの等高線マップ上での重心点の遷移

各年の遷移



最初と最後



疑問：
- 重心が近づくと内容も近づく？
- 重心の移動方向の意味は？

一つの切り口として特許分類で調査

Ⅲ. GAFA米国特許出願に含まれる特許分類

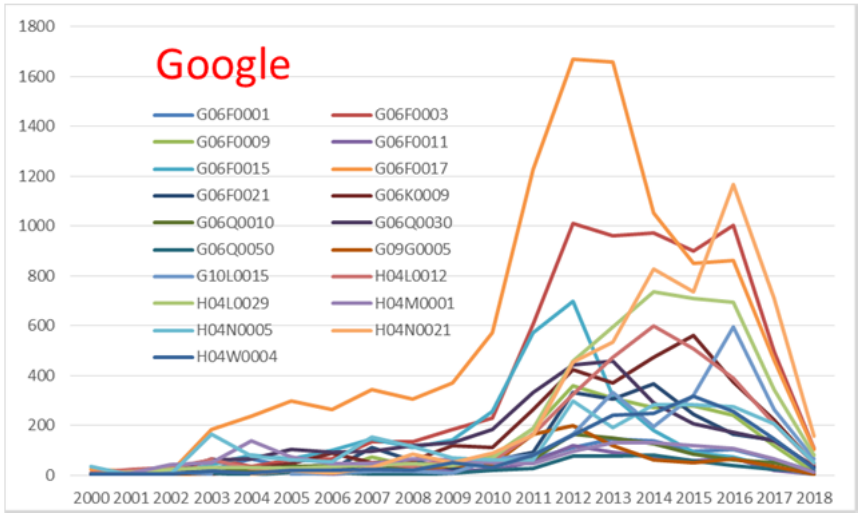
■GAFAの米国特許出願に多く含まれる特許分類 （IPC上位8bit：メイングループ）

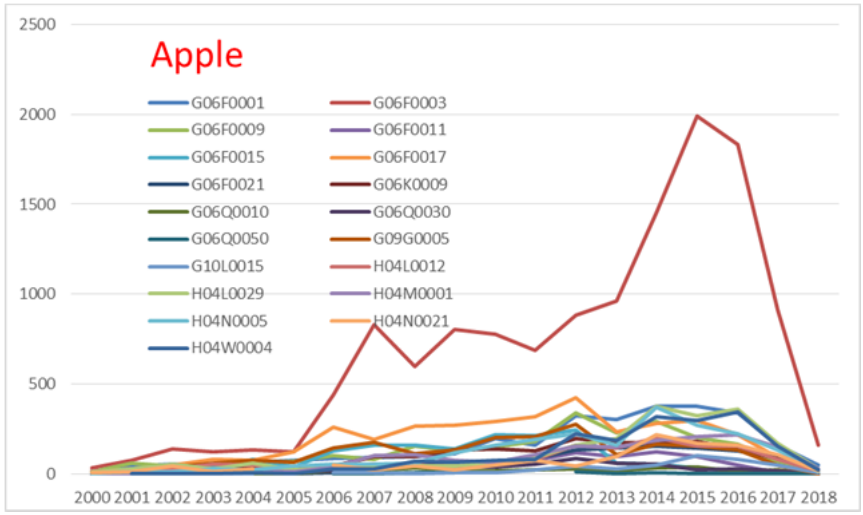
	Google IPC8	[%]	Apple IPC8	[%]	Facebook IPC8	[%]	Amazon IPC8	[%]
1	G06F0017：特定機能のデータ処理装置	10.3	G06F0003	12.7	H04L0029	12.3	H04L0029	9.0
2	G06F0003：データ入出力(I/F)装置	6.8	G06F0017	3.5	G06F0017	11.9	G06F0017	7.9
3	H04N0021：選択的なコンテンツ配信	5.0	G06F0001	2.9	G06F0003	8.5	H04L0012	6.0
4	H04L0029：デジタル情報の伝送方式	4.1	G06F0009	2.4	H04L0012	7.7	G06F0009	5.4
5	G06F0009：プログラム制御	3.2	H04L0029	2.1	G06F0015	5.6	G06F0003	5.3
6	H04L0012：データ交換ネットワーク	3.1	H04N0005	2.1	G06Q0030	5.1	G06Q0030	5.1
7	G06Q0030：商取引，例．電子商取引	2.9	G09G0005	2.1	H04N0021	4.9	G06F0015	4.9
8	G06F0015：データ処理装置一般	2.9	H04W0004	1.9	G06Q0050	4.3	G06F0021	3.1
9	H04N0005：テレビジョン方式の細部	2.4	G06K0009	1.7	H04W0004	4.0	G06F0011	2.8
10	G10L0015：音声認識	2.1	H04M0001	1.6	G06Q0010	2.6	H04N0021	2.7
11	G06F0009：プログラム制御	2.0	G06F0015	1.6	G06K0009	2.1	G06Q0010	2.3
12	G06F0021：セキュリティ装置	2.0	H04L0012	1.4	G06F0021	2.0	G10L0015	1.9
13	H04M0001：サブステーション装置	1.8	H04N0007	1.4	G06F0009	1.8	H04L0009	1.8
14	G06F0007：データの順序・内容の操作	1.7	G06F0012	1.3	H04M0003	1.2	G06K0009	1.8
15	H04W0004：無線通信ネットワーク	1.6	H04N0019	1.2	G06F0012	1.0	G06F0007	1.8

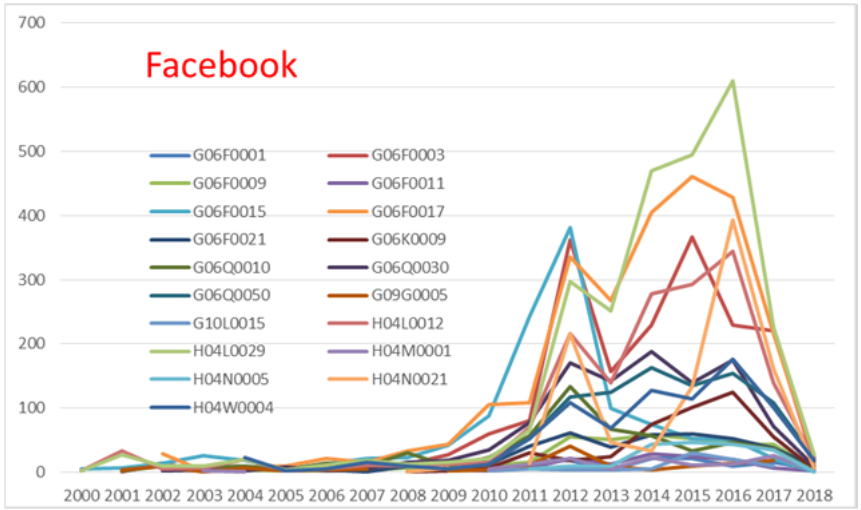
※2000年以降の特許出願が対象（１出願で複数IPCの記載がある場合、それらすべてをカウント）

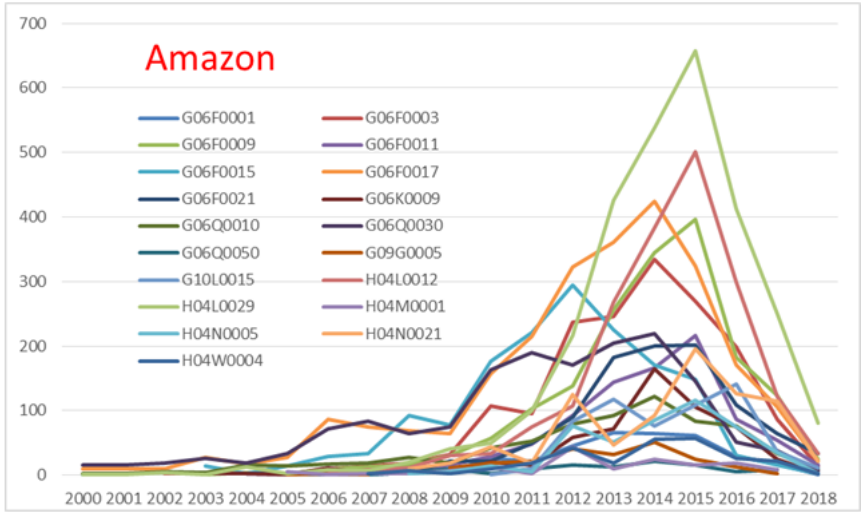
Ⅲ. GAFA米国特許出願の特許分類の推移

■GAFAの米国特許出願に多く含まれる特許分類 （IPC上位8bit：メイングループ）



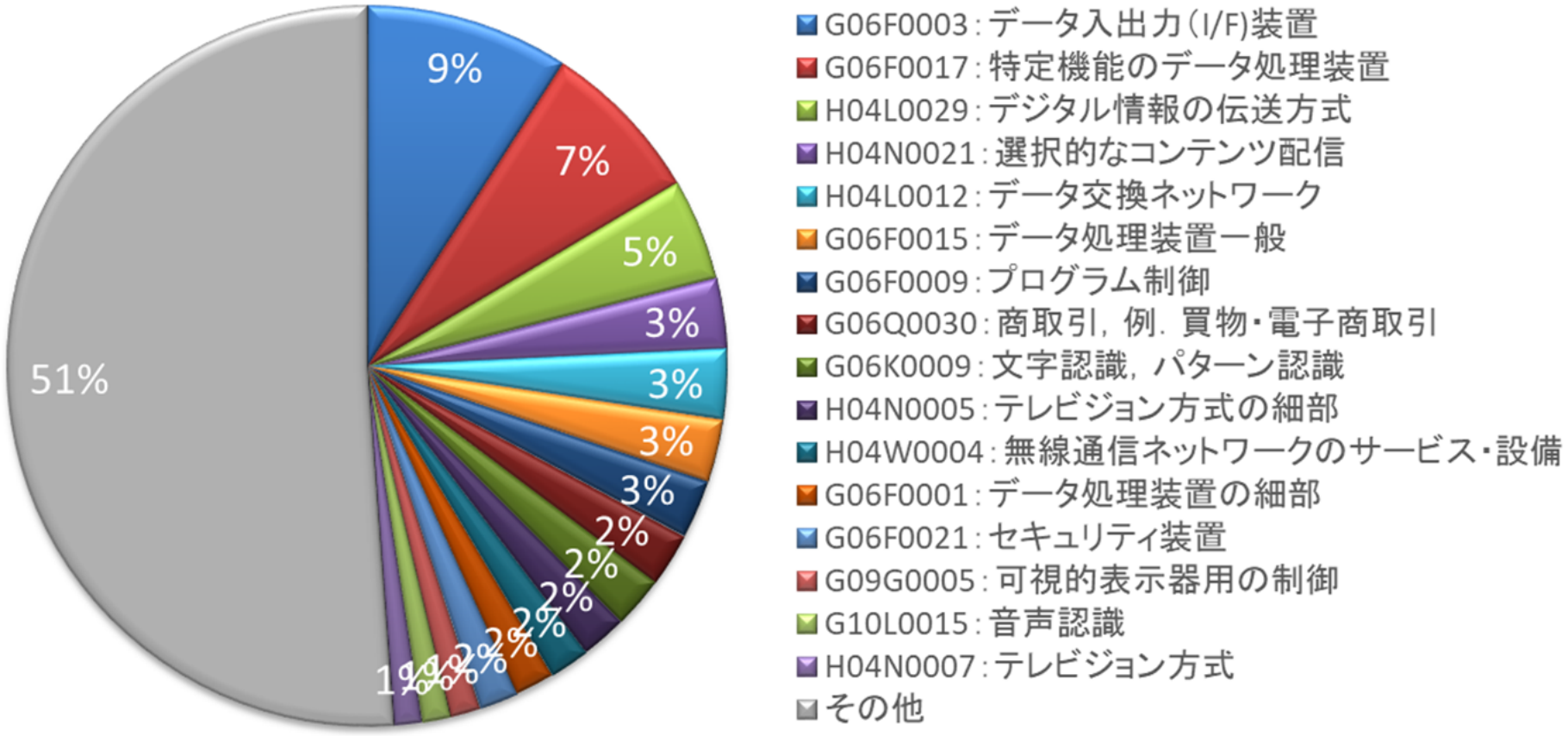






Ⅲ. GAFA米国特許出願に含まれる特許分類

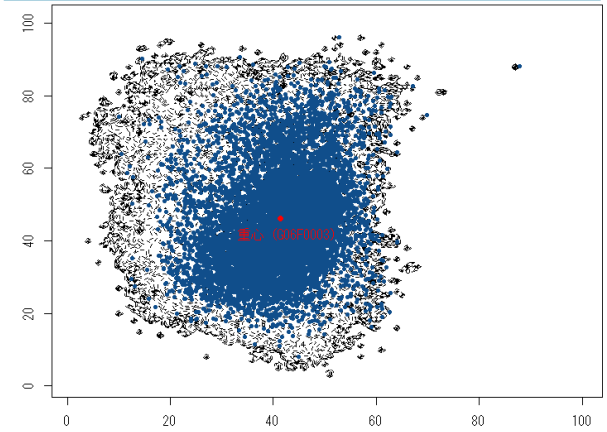
■GAFAの米国特許出願に多く含まれるIPC特許分類（上位8bit：メイングループ）



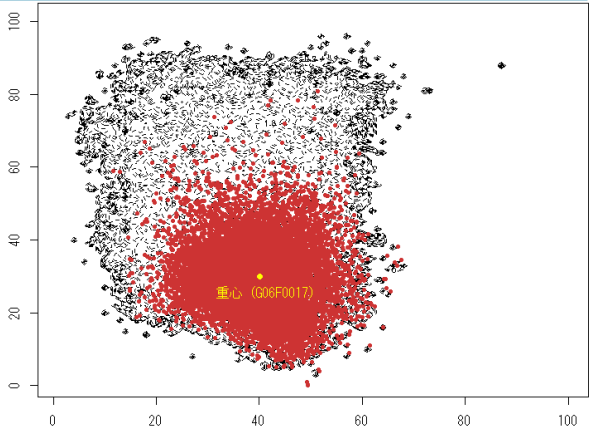
※2000年以降の特許出願が対象（１出願で複数IPCの記載がある場合、それらすべてをカウント）

Ⅲ. GAFA出願の特許分類毎のプロット範囲

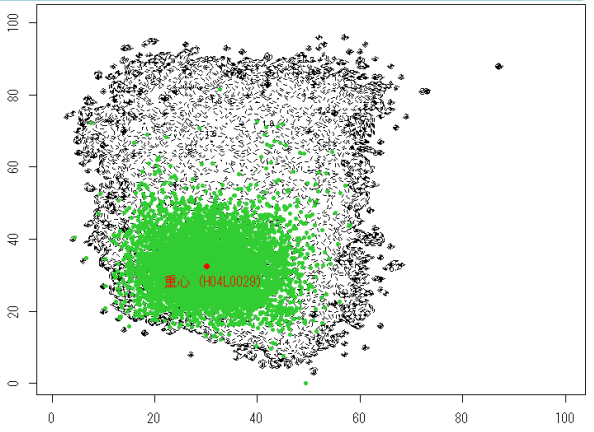
1.データ入出力装置：G06F0003



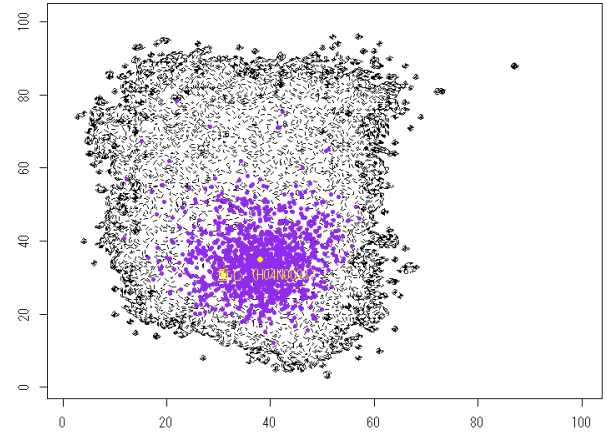
2.特定データ処理： G06F0017



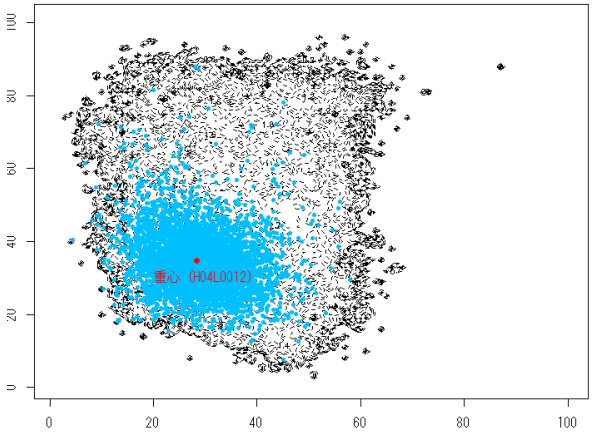
3.デジタル情報伝送： H04L0029



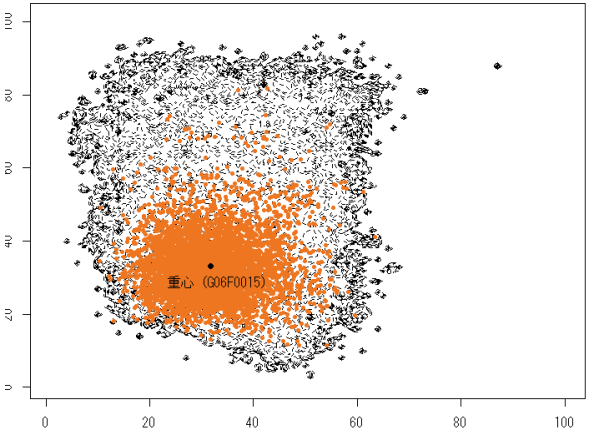
4.コンテンツ配信：H04N0021



5.データ交換NW： H04L0012



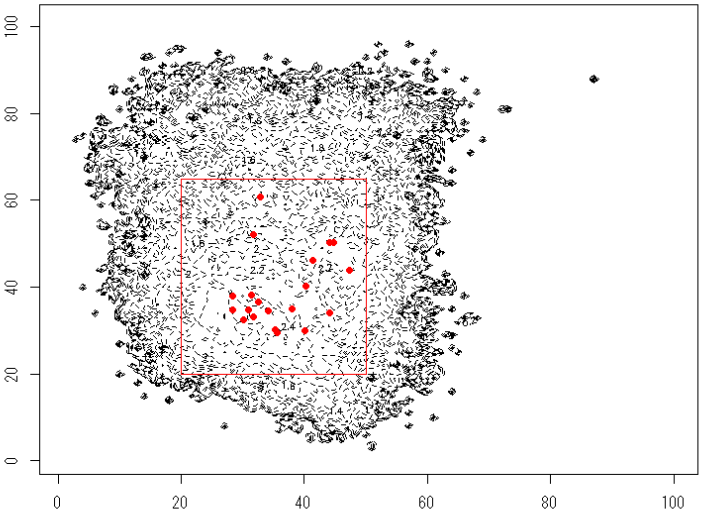
6.データ処理一般：G06F0015

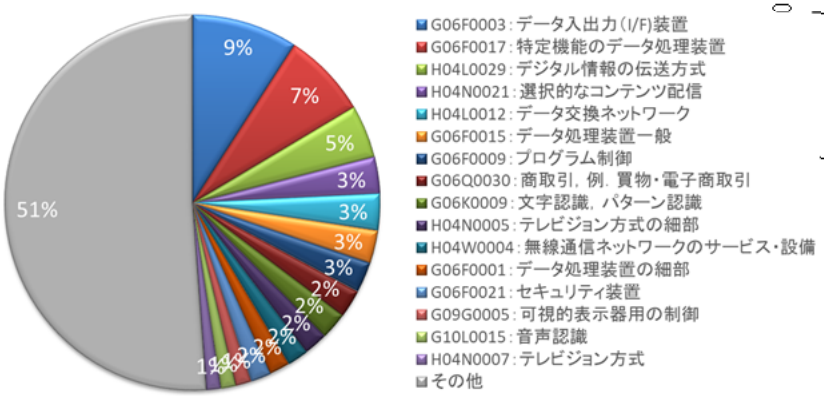


Ⅲ. GAFA出願の特許分類毎の重心点

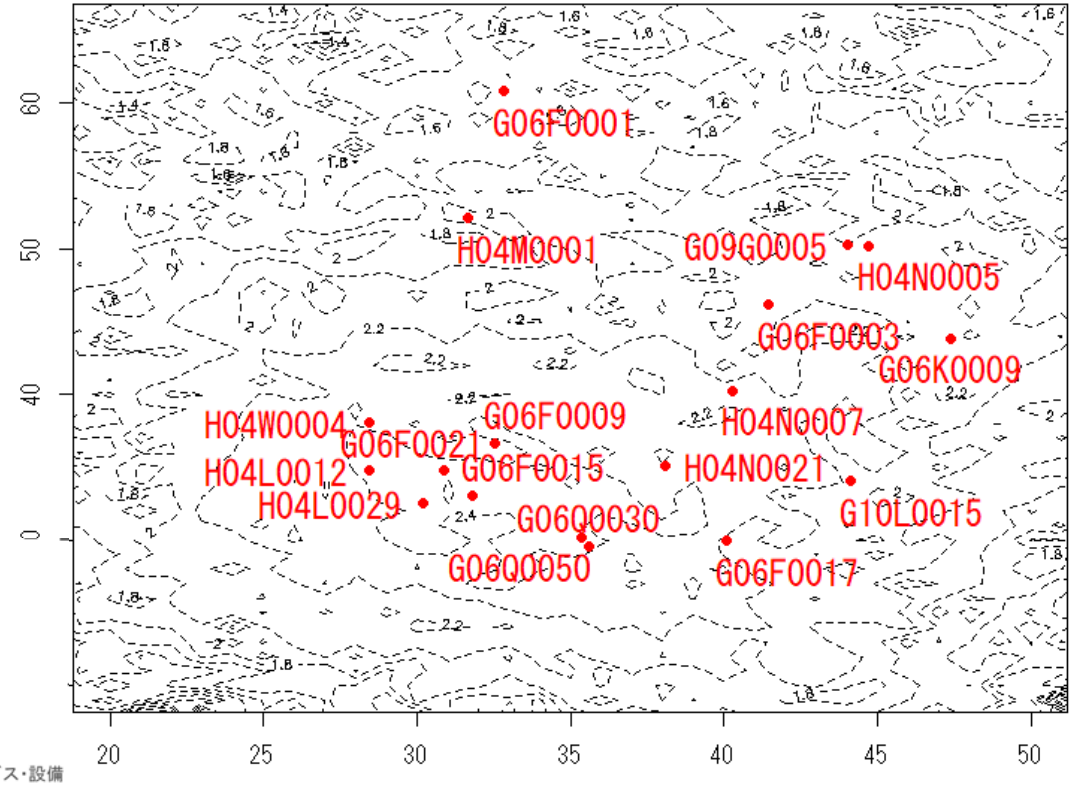
■GAFAの等高線マップ上での重心点の遷移

IPC 8bit 重心（全体）





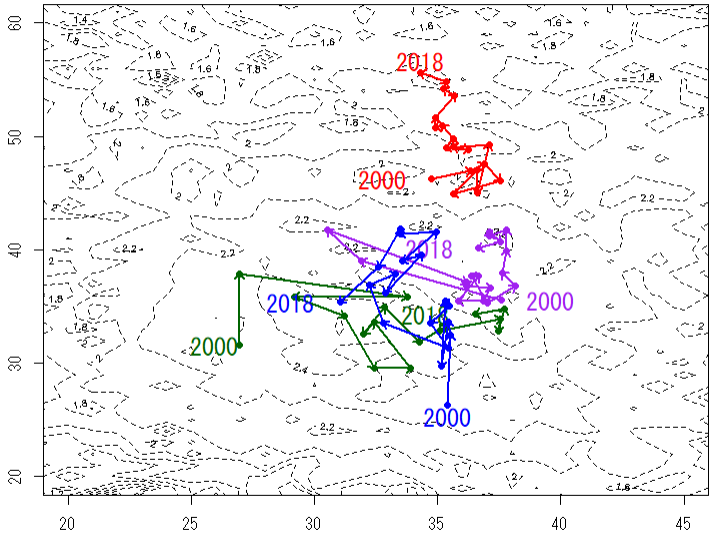
IPC 8bit 重心（拡大）

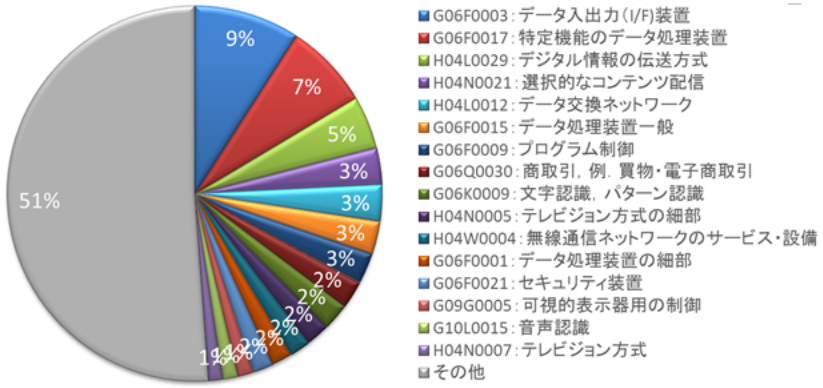


Ⅲ. GAFAの各企業と特許分類の重心点

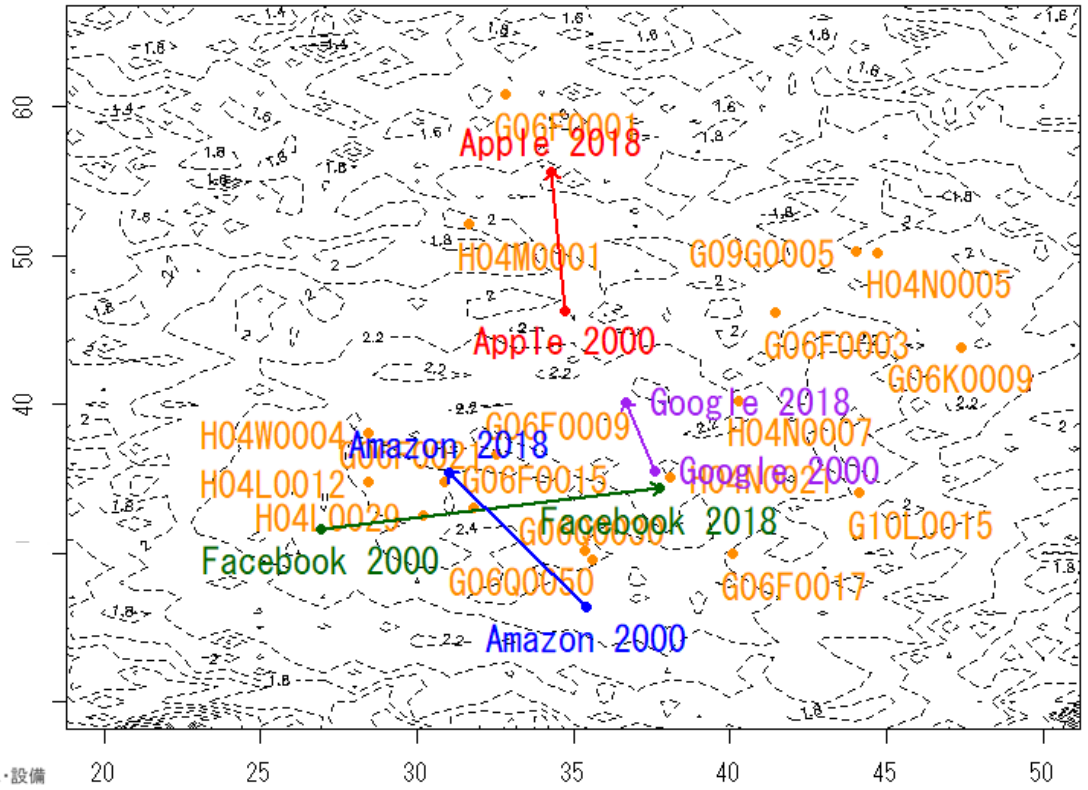
■GAFAの等高線マップ上での重心点の遷移

各年の遷移



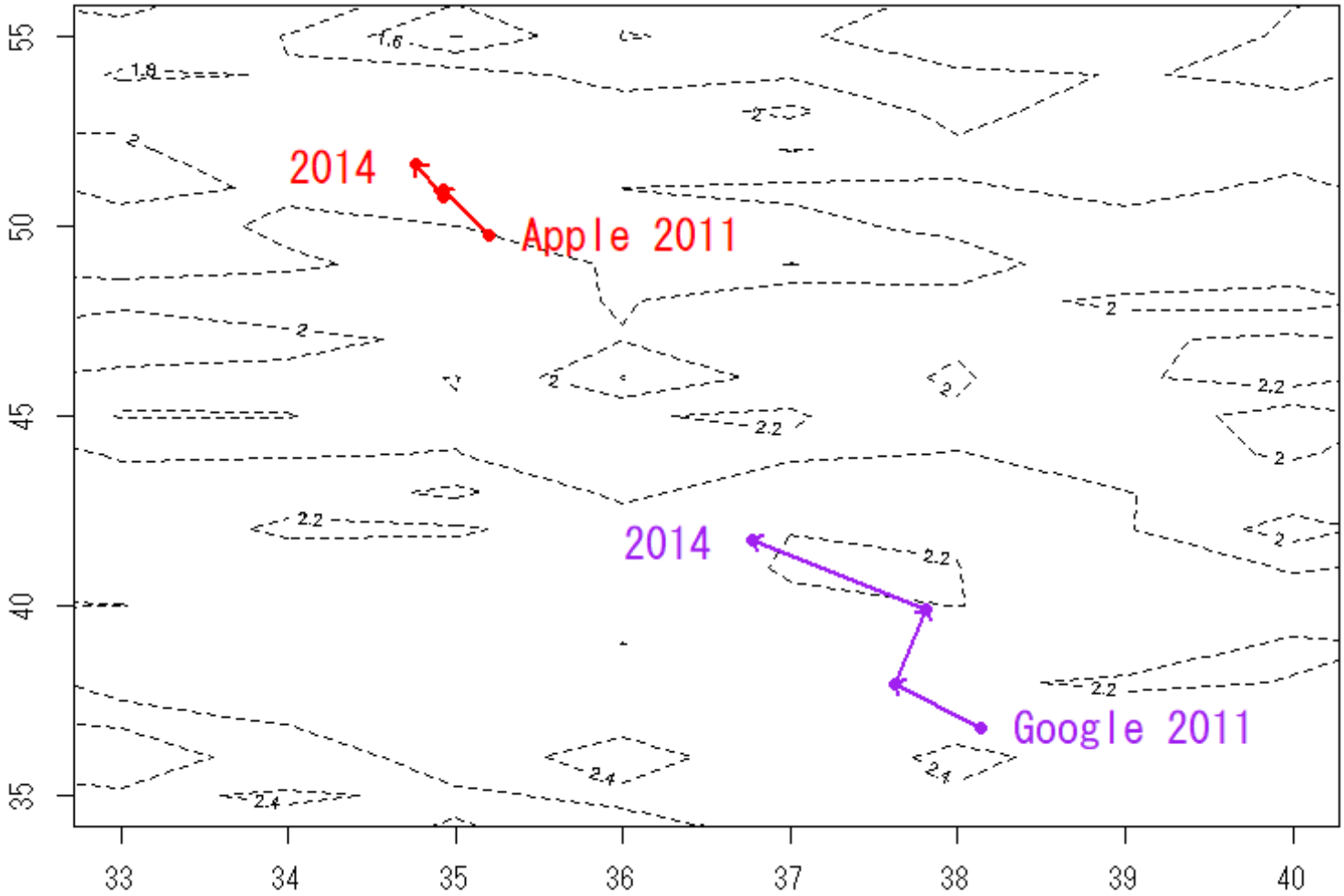


最初と最後



Ⅲ. Google, Appleの重心点推移（例）

■Google, Appleの重心点遷移（2011～2014年）



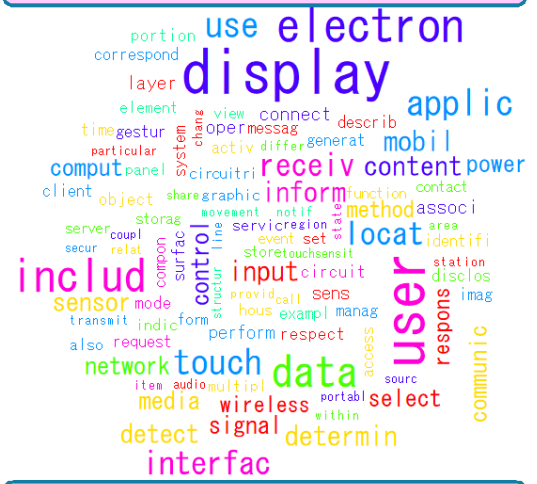
Ⅲ. Google, Appleの重心変化に寄与した分類

■IPC比率[%] 変化（2011⇒2014年）

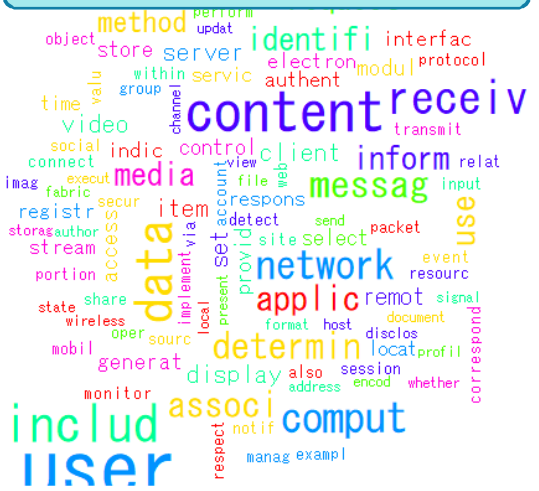
	Google	Apple
G06F0001：データ処理装置の細部	0.5	1.9
G06F0003：データ入出力	0.1	5.1
G06F0009：プログラム制御	−0.4	−0.4
G06F0011：エラー検出・訂正	−0.2	−0.2
G06F0015：データ処理一般	−10.3	−6.6
G06F0017：特定のデータ処理	−12.8	−5.5
G06F0021：セキュリティ	3.1	0.0
G06K0009：文字・パターン認識	0.7	−1.0
G06Q0010：管理；経営	0.3	−0.2
G06Q0030：電子商取引	−3.4	−0.8
G06Q0050：特定の業種のシステム	0.5	0.2
G09G0005：ディスプレイ制御	−2.9	−4.0
G10L0015：音声認識	1.0	0.2
H04L0012：データ交換ネットワーク	4.7	1.3
H04L0029：配置，装置，方式	6.2	4.4
H04M0001：サブステーション装置	0.8	−0.1
H04N0005：テレビジョン方式の細部	2.6	0.8
H04N0021：選択的なコンテンツ配信	8.0	1.6
H04W0004：無線通信ネットワーク	1.7	3.4

重心が同じ方向に移動してもその要因は必ずしも一致しない ⇒ 詳細に見ないと判断できない

Apple増加KeyWords



Google増加KeyWords

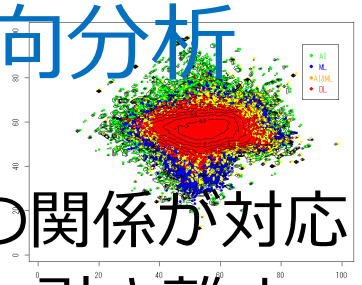


まとめ

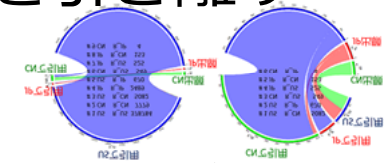
Ⅰ. 大規模データの等高線マップ作成方法の検討

- 26万件の大規模データを家庭用ノートPCで処理可能
- 今回の検討ではveganパッケージのdecoranaの対応分析を採用

Ⅱ. AI, ML, DL等の用語含む特許出願の動向分析

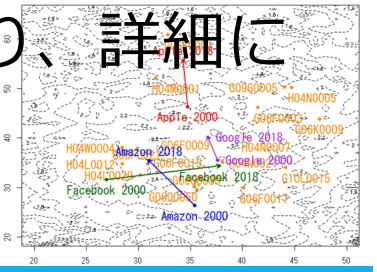


- 広い範囲の特許出願で米国が先行し、中国が追随
- AI, ML, DLの一般概念と等高線マップ上のplot範囲の関係が対応
- 日米中の被引用分析では近年、米国が再度日・中を引き離す



Ⅲ. GAFAの米国特許出願の動向分析

- 企業ごとの重心点の動向比較で、違う企業が同じ方向に移動しても、その要因が必ずしも一致しない場合があり、詳細に見ないと正確には動向の要因を判断できない



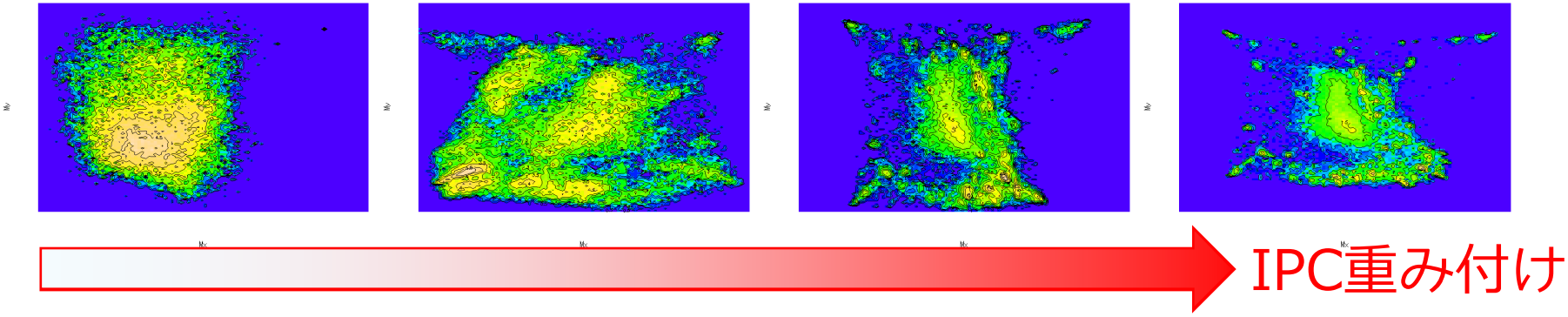
課題と今後のトライアルについて

■課題

- ➢今回、数百次元から２次元に落として重心分析：限界有り？
  - ・評価指標（特許分類等）による重みづけで分かりやすく？
  - ・２次元よりも高次元のまま分析？
- ➢大規模データではメモリが課題 ⇒ 次元圧縮手法の工夫必要

■今後行ってみたいトライアル

- ➢DeepLearningの分散表現ベクトルによる次元圧縮トライアル
- ➢様々な評価指標（特許分類等）による重みづけのトライアル



謝辞

本報告は2018年の「アジア特許情報研究会」の知財情報解析チームのワーキングの一環として報告するものです。

今年４月に機械学習の勉強を始めたばかりの初心者である筆者に様々なアドバイスをいただきました研究会の皆様に深く感謝申し上げます。