v20200618

Science Landscape

数式の解説

KOSHIBA Hitoshi

Science Landscape

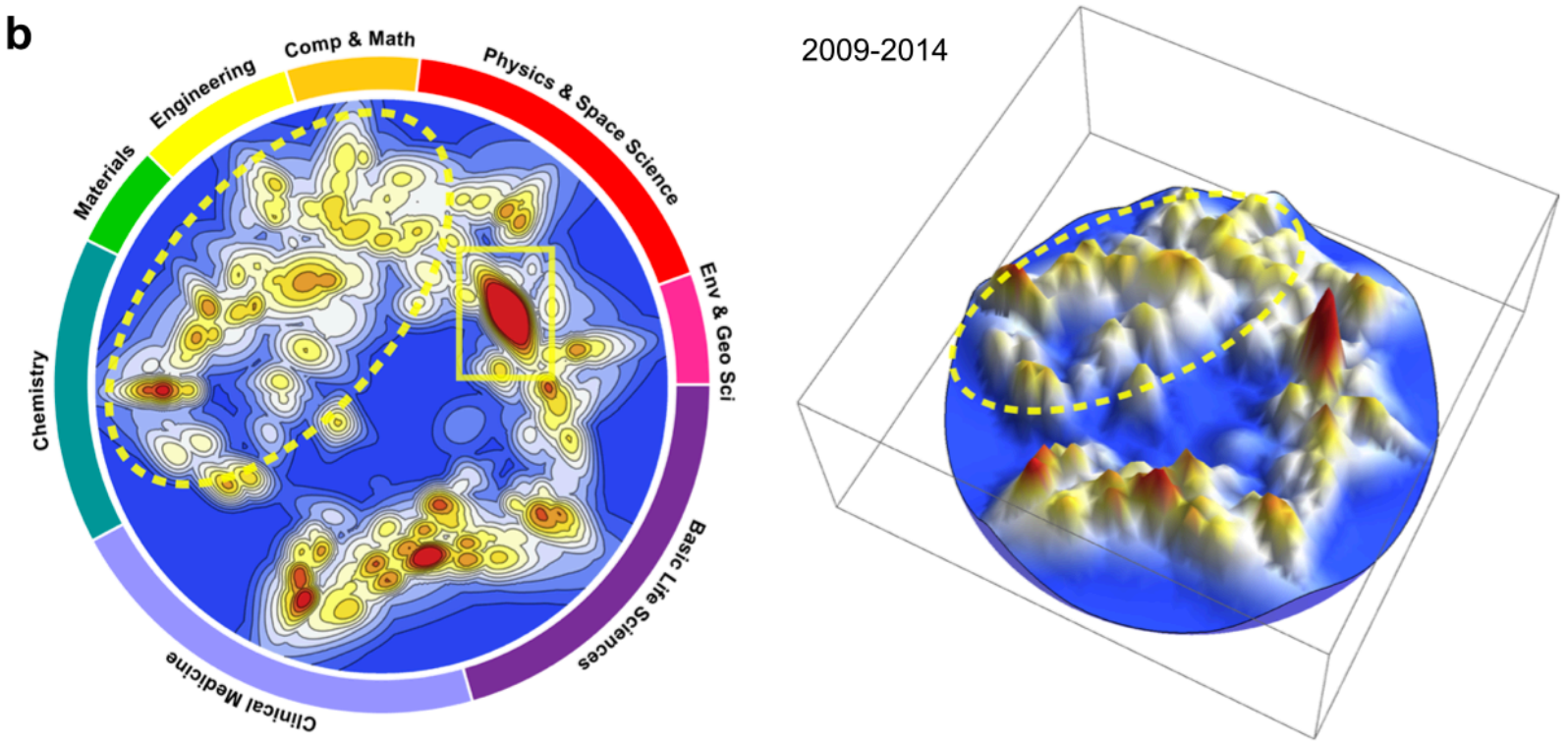
以下の論文で示されている，分野融合の度合いと被引用ベースのリサーチインパクトの可視化手法

- Top 1% 論文の Research Front ベースで，量ではなく“質”で可視化
- 元のデータは NISTEP の サイエンスマップ
- 分野の区切りは ESI22分類 をベースに NISTEP が設定した10分野
  - 手法自体はこれらの分類には特に関係しない

https://doi.org/10.1057/s41599-019-0352-4

Science Landscape

■ 可視化のイメージ



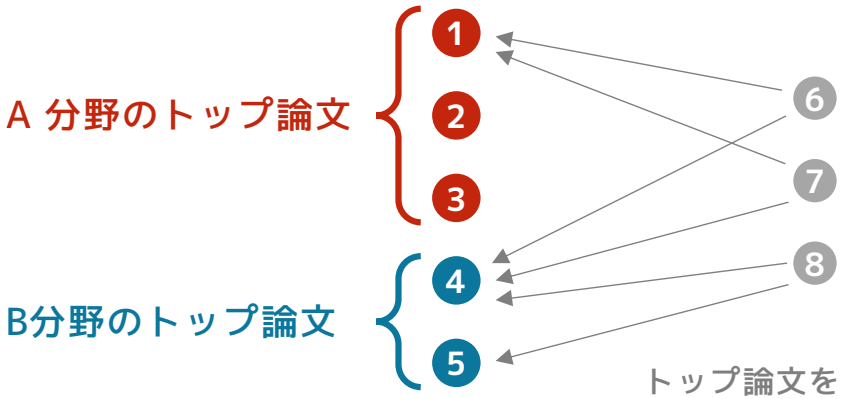
https://doi.org/10.1057/s41599-019-0352-4　fig.3 から抜粋

Research Fronts：RF

■ これまでになかった研究分野の検出に使える指標のひとつ

■ 引用関係の分析に基づく

- ◆ 雑誌，もしくは記事の分野分類を使用
- ◆ ある論文（基本的にはTop 1%,10%など）が引用されている記事（雑誌）の分野分類を調べて，その共起関係で規定
  - ▫ ある成果が，どういう分野に還元されていったか…みたいな指標
  - ▫ 「引用している」だと，何と何を混ぜ合わせたか，なので元の成分…みたいな指標
  - ▫ 実際には，共引用関係から設定



トップ論文のうち，1と4は一緒に引用される数が多い

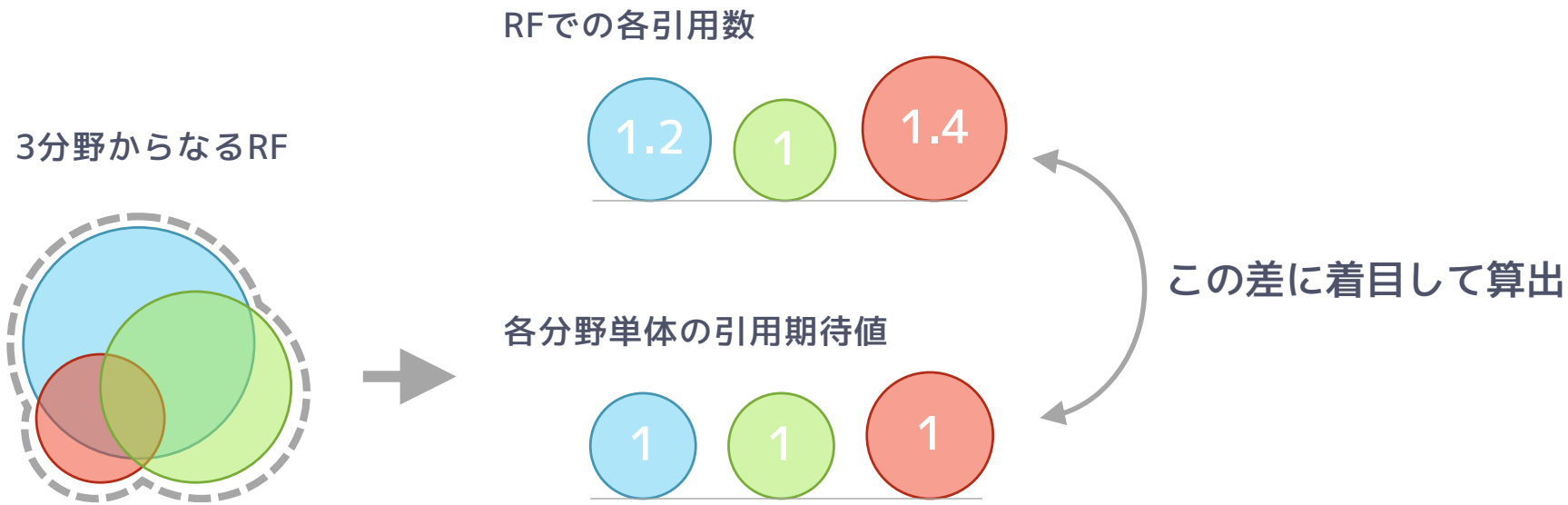
1と4はそれぞれ，A分野，B分野の論文

A分野 と B分野 の関係性に 1ポイント プラス

こんな感じの指標から算出

Research Impact：RI

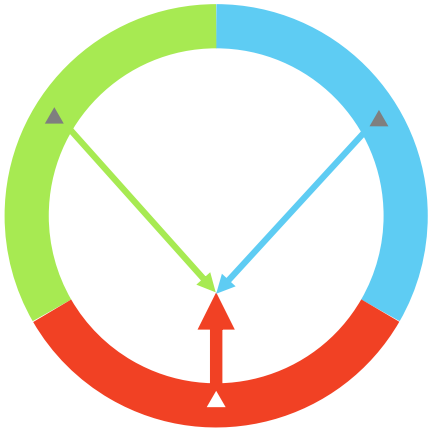
- これまでになかった研究分野の検出に使える指標のひとつ
- その組合せの“異質さ”を表すような指標
  - 単体での期待値と，RF中での実測値の違いで算出
  - ある種の情報量
    - 特定分野単体で平均的に200回引用，ある分野とセットの場合でも200回引用
    - 特定分野単体だと平均的に2回引用，ある分野とセットの場合200回引用 ← 異質！



SciLand のイメージ

- RFの構成分野の割合で位置（重心）を設定
- RIで高さを設定

1 : 1 : 1



1 : 2 : 1

2 : 2 : 1

各分野の成分割合をバネの強さ，バネの始点を各分野の弧※の中心，にするイメージ

※ 弧の長さはコアペーパー数で規定

※ 中心値からのズレの程度は，分野融合度に一致

（実際にはいろいろ相違点があります，あくまでイメージです）

研究分野の分類

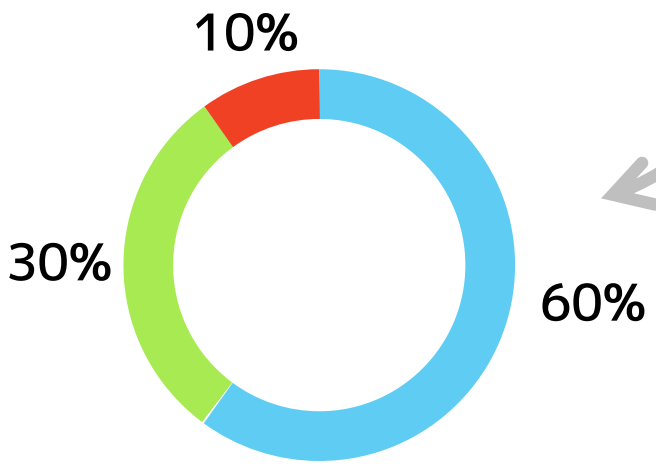
ESI 22 研究分野分類		10分野分類 (NISTEP)	
1.	環境/生態学	A.	環境・地球科学
2.	地球科学		
3.	物理学	B.	物理学
4.	宇宙科学		
5.	計算機科学	C.	計算機・数学
6.	数学		
7.	工学	D.	工学
8.	材料科学	E.	材料科学
9.	化学	F.	化学
10.	臨床医学	G.	臨床医学
11.	精神医学/心理学		
12.	農業科学	H.	基礎生命科学
13.	生物学・生化学		
14.	免疫学		
15.	微生物学		
16.	分子生物学・遺伝学		
17.	神経科学・行動学		
18.	薬理学・毒性学		
19.	植物・動物学		
20.	経済学・経営学	I.	経済・社会科学
21.	社会科学・一般		
22.	複合領域	J.	複合領域

8分野のみ使用

弧の長さ，位置を決める

（RFに関係なく）コアペーパー数に応じた割合を算出

上の帯をぐるっと円にして配置



ここは地図の座標決めで，
RFとは関係ない点に注意
（あくまで，全体の割合）

※ ひとつ前の分野の長さ ＋ 自分の長さの半分 ＝ 自分の重心

弧の長さ，位置を決める

直線ベースでの分野Aの中心位置

$$\Theta_A = \sum_{B<A} \theta_B + \frac{1}{2}\theta_A$$

分野Aの長さの半分

Aより前の分野の長さ合計

$$\theta = 2\pi \frac{N_A}{N} \qquad N = \sum_{A \in \mathcal{R}} N_A$$

N_A ：分野Aのコアペーパー数

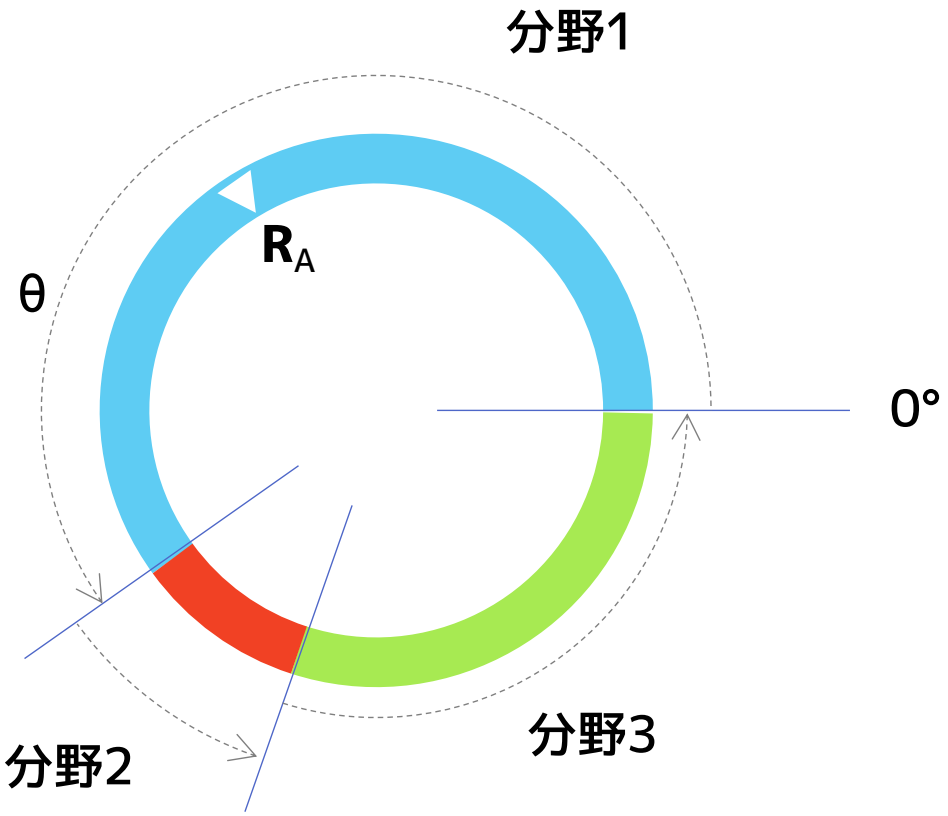
分野Aの円周上に占める中心位置

$$\boldsymbol{R}_A = \rho(\cos\Theta_A, \sin\Theta_A)^T$$

ρ：半径（定数）

弧の長さ，位置を決める

ちなみに，式からも明らかに，
実際の配置は以下



3時の方向から，反時計回りに配置

基準となる分野間の関係性を求める

ここからESIのRFの話

※ 実際に自分で実装することを試みる場合，
論文にある計算結果をそのまま使うので，
このセクションは読み飛ばしても良い

ある年月次でのRF全体 （n次元ベクトル）

$$S_{y/m} = \left\{ \mathrm{RF}_{n;y/m} \right\}$$

RFのID　　年月

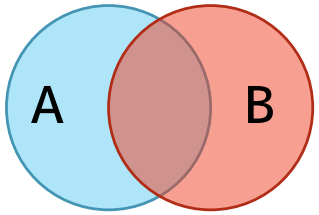
0か1を返す単純な2値関数

$$\sigma_{n;\mathrm{A}} \begin{cases} \sigma_{n,\mathrm{A}} = 0 & \cdots\ \mathrm{RF}_{n;y/m} \text{ に分野A が含まれない} \\ \sigma_{n,\mathrm{A}} = 1 & \cdots\ \mathrm{RF}_{n;y/m} \text{ に分野A が含まれる} \end{cases}$$

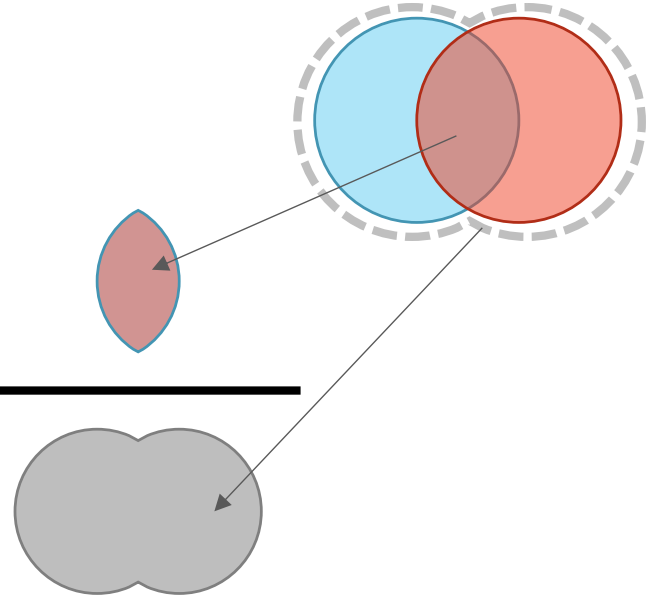
基準となる分野間の関係性を求める

Jaccard係数：
任意の集合間の重複度を測る指標

$$J = \frac{|A \cap B|}{|A \cup B|}$$



$$J = \frac{|A \cap B|}{|A \cup B|} =$$



全体に占める共通部分の割合

基準となる分野間の関係性を求める

$$T_{\mathrm{A};y/m} = \left\{ \mathrm{RF}_{n;y/m} \in S_{y/m} \mid \sigma_{n,\mathrm{A}} = 1 \right\}$$

ある 分野A を含む RF を 仮に T_A とする

S における ある分野A と B の重複度は…

$$J_{y/m}(\mathrm{A},\mathrm{B}) = \frac{\left| T_{\mathrm{A};y/m} \cap T_{\mathrm{B};y/m} \right|}{\left| T_{\mathrm{A};y/m} \cup T_{\mathrm{B};y/m} \right|}$$

$$= \frac{\left| \left\{ \mathrm{RF}_{n;y/m} \in S_{y/m} \mid \sigma_{n,\mathrm{A}} + \sigma_{n,\mathrm{B}} = 2 \right\} \right|}{\left| \left\{ \mathrm{RF}_{n;y/m} \in S_{y/m} \mid \sigma_{n,\mathrm{A}} + \sigma_{n,\mathrm{B}} \geq 1 \right\} \right|}$$

全部のRF間で，AとBの両方があるか，ないか調べて，ありなしの比を出す

基準となる分野間の関係性を求める

記号いっぱいなので，難しそうに見えますが…

$$J_{y/m}(\mathrm{A},\mathrm{B}) = \frac{\left|\left\{\mathrm{RF}_{n;y/m} \in S_{y/m} \mid \sigma_{n,\mathrm{A}} + \sigma_{n,\mathrm{B}} = 2\right\}\right|}{\left|\left\{\mathrm{RF}_{n;y/m} \in S_{y/m} \mid \sigma_{n,\mathrm{A}} + \sigma_{n,\mathrm{B}} \geq 1\right\}\right|}$$

$$= \frac{\text{分野A, B の両方を含む RF の数}}{\text{分野A, B の少なくとも一つを含む RF の数}}$$

… を，意味しています

基準となる分野間の関係性を求める

8分野あるので全部の組合せを計算して…

$$\boldsymbol{M}_{y/m} = \left[J_{y/m}(\mathrm{A}, \mathrm{B}) \right]$$

$$\boldsymbol{M} = \begin{bmatrix} J_{(1,1)} & J_{(1,2)} & \cdots & J_{(1,8)} \\ J_{(2,1)} & J_{(2,2)} & \cdots & J_{(2,8)} \\ \vdots & \vdots & \ddots & \vdots \\ J_{(8,1)} & J_{(8,2)} & \cdots & J_{(8,8)} \end{bmatrix}$$

距離に変換したい & 幸い Jaccard は 0–1 なので…

$$\boldsymbol{D}_{y/m} = \mathbf{1} - \boldsymbol{M}_{y/m}$$

位置，高さを決める

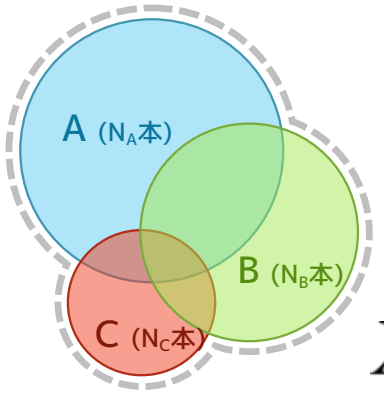
ここからサイエンスマップのRFの話

$N_{i,\mathrm{A}}$ ある RF_i における分野A の本数

$N_i = \sum_{\mathrm{A}\in\mathcal{R}} N_{i,\mathrm{A}}$ ある RF_i における論文数全体

X_i ある RF_i における被引用数

3分野からなるRF_i



X_i

RFの引用件数

$C_{\mathrm{A};y/m}$ ある分野A の，ある期間y/m での 平均的な被引用数

$\langle C_{\mathrm{A}} \rangle$ ある分野A の，全期間での 平均的な被引用数

$\langle \ldots \rangle$ <- 平均を取る…の意味で使用

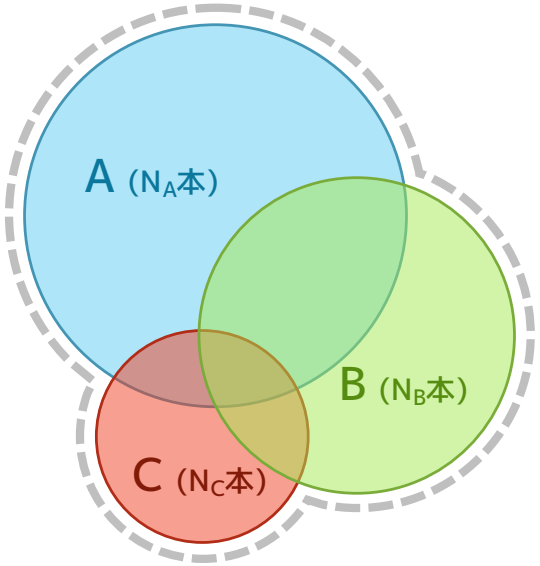
位置を決める

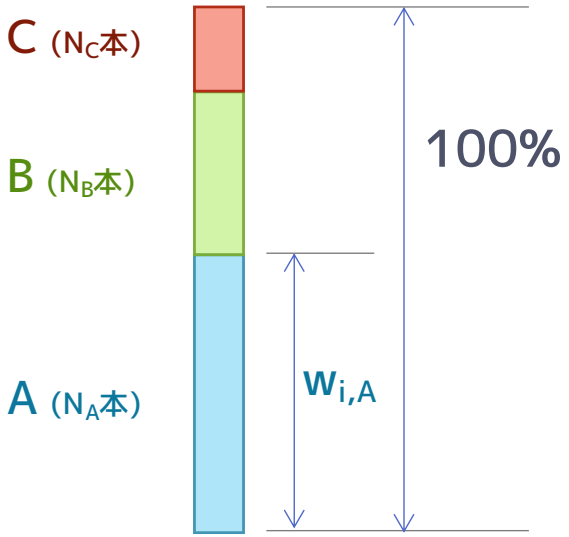
ある RF_i において分野A が本数に占める割合

$$w_{i,\mathrm{A}} = \frac{N_{i,\mathrm{A}}}{N_i}$$

$$\sum_{\mathrm{A}\in\mathcal{R}} w_{i,\mathrm{A}} = 1$$

割合なので全部足したら当然 1





位置を決める

ある RF_i において分野A が本数に占める割合　…を，すこし補正

$$\tilde{w}_{i,\mathrm{A}} = \frac{(w_{i,\mathrm{A}})^{\eta}}{\sum_{\mathrm{A}\in\mathcal{R}}(w_{i,\mathrm{A}})^{\eta}} \qquad \eta \in \mathbb{R}^{+}$$

論文では $\eta = 0.4$ を使用

円内での位置は，各分野の重心位置をベースに 分野の構成比と分野融合度で決定

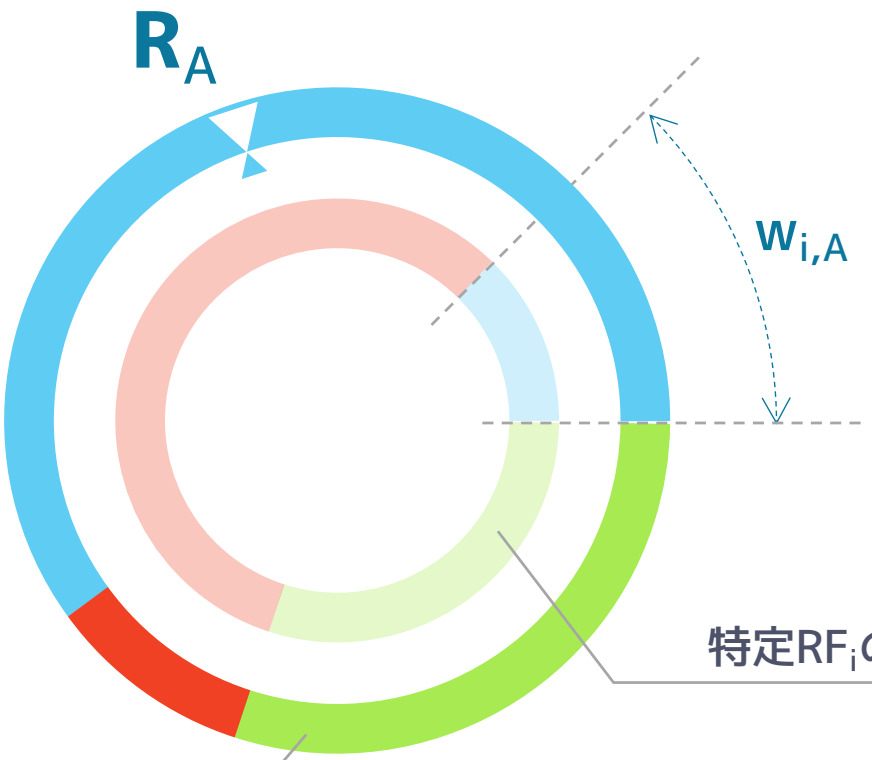
※ ここでは，円の中心から見てどの方向に位置するか

$$\boldsymbol{r}_{com,i} = \sum_{\mathrm{A}\in\mathcal{R}} \tilde{w}_{i,\mathrm{A}} \boldsymbol{R}_{\mathrm{A}}$$

$\mathbf{R}_{\mathrm{A}}$ ：全体での分野Aの重心位置

$w_{i,\mathrm{A}}$ ：RF_iにおける分野Aの論文割合

位置を決める



全体の分野比率の中で，
特定RFの比率がどうなっているか…
で，RF位置（方角）を決める

位置を決める

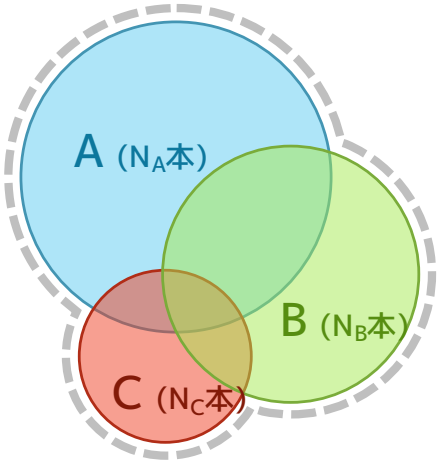
分野融合度 $\Delta_i = \left(\boldsymbol{w}_i{}^T \langle \boldsymbol{M} \rangle \boldsymbol{w}_i \right)^{-1}$

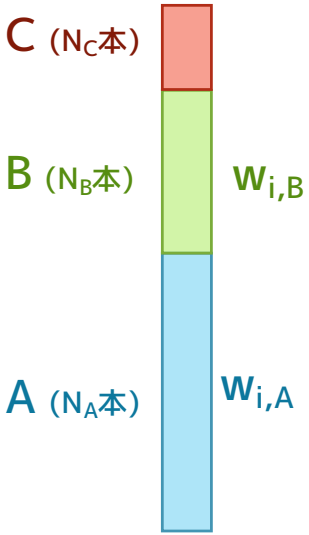
RF$_i$に占める各分野の割合と，分野間の近さから算出

$$w_{i,\mathrm{A}} = \frac{N_{i,\mathrm{A}}}{N_i}$$

← RF$_i$ に占める分野Aの論文数

$$\boldsymbol{w}_i{}^T = (w_{i,\mathrm{A}}, w_{i,\mathrm{B}}, \ldots)$$





$$\boldsymbol{M}_{y/m} = \left[J_{y/m}(\mathrm{A}, \mathrm{B}) \right]$$

$\langle \boldsymbol{M}_{\mathrm{AB}} \rangle$ は，分野A,Bに付いての全期間平均

位置を決める

ようやく，最終的な位置を確定

方角を決める

長さを決める

半径ρ から どれだけ中心に向けるか

$$\boldsymbol{r}_i = (1-\delta)(\rho - \Delta_i)\frac{\boldsymbol{r}_{com,i}}{\left|\boldsymbol{r}_{com,i}\right|}$$

円周上にかぶらないように余白を取る

δ : 余白率

とりあえず正規化する

論文では $\delta = 0.2,\ \rho = 4.5$ を使用

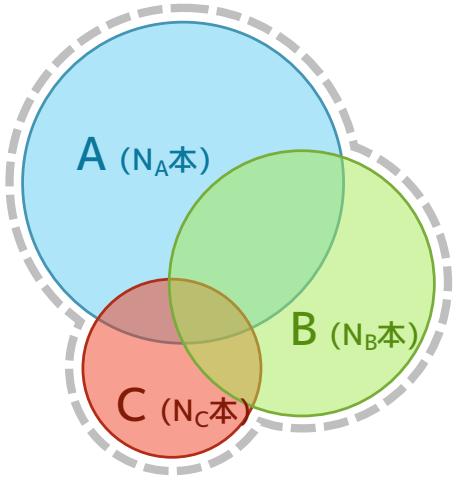
高さを決める

ある RF_i におけるインパクト

$$RI_i = \log\left\{\frac{X_i}{\sum_{A\in\mathcal{R}} N_{i,A}\langle C_A\rangle}\right\}$$

組合せなしの状態に比べて，どのくらい引用件数が多いか？

3分野からなるRF_i

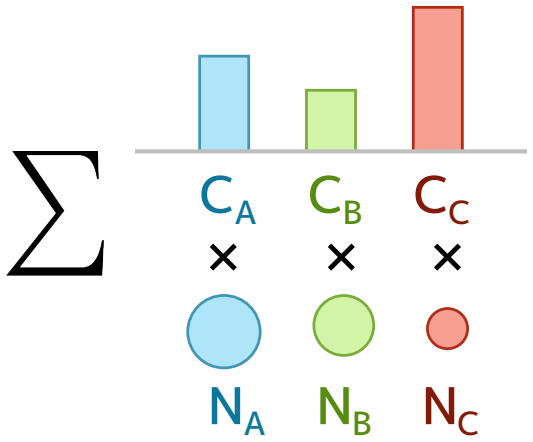


ここの比

X_i

RFの全引用件数

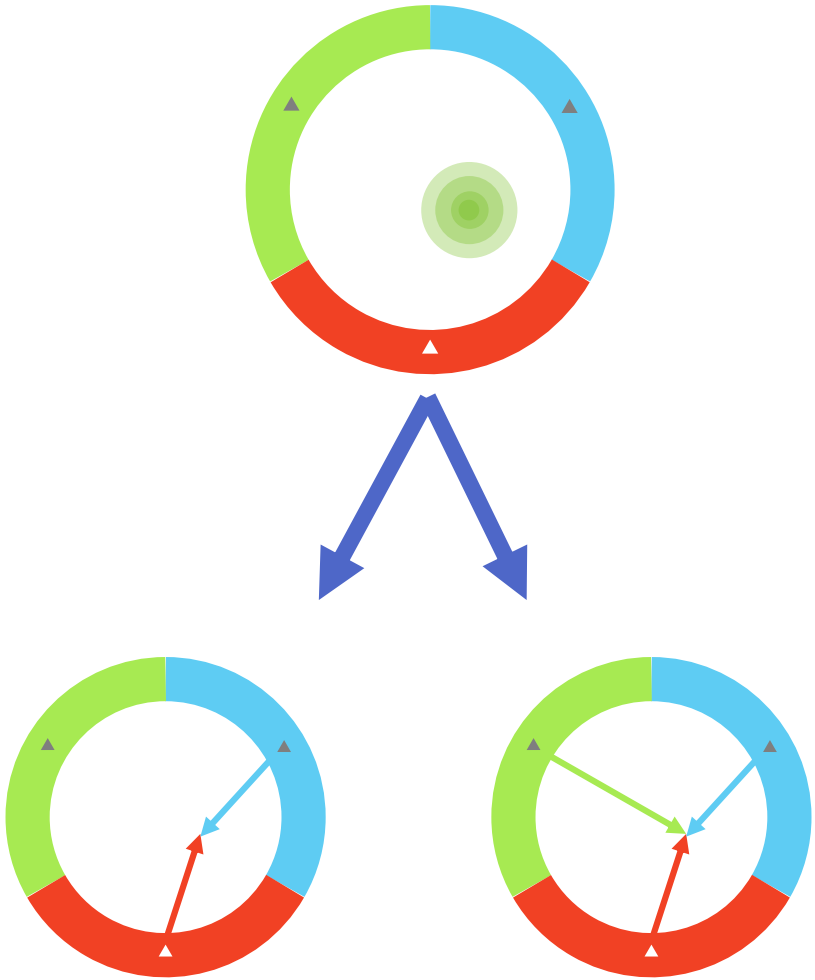
各分野単体の平均引用数



RFにおける各分野の本数

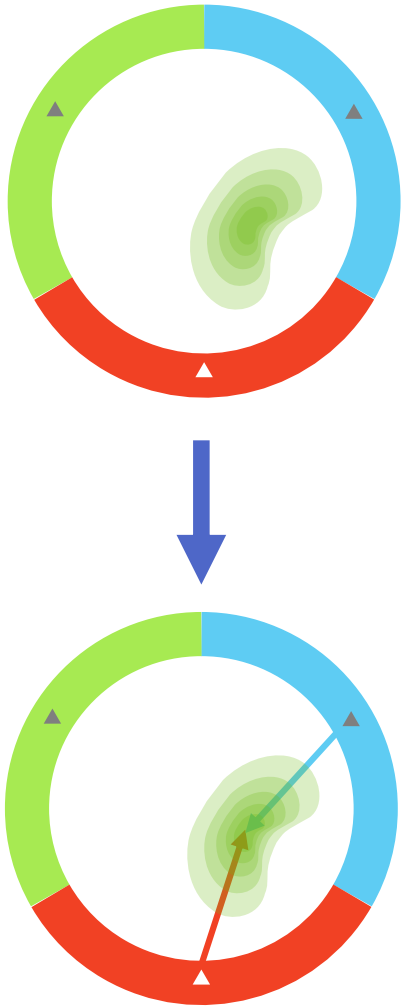
分布の形状を決める

単純な等高線だと…



元の構成要素が不明

成分方向に歪めると…



元の構成要素が推定できる

分布の形状を決める

稜線を作成する部分

$$v_{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{r}_i) = \frac{(\boldsymbol{R}_{\mathrm{A}} - \boldsymbol{r}_i)^T \boldsymbol{A} (\boldsymbol{r} - \boldsymbol{r}_i)}{|\boldsymbol{R}_{\mathrm{A}} - \boldsymbol{r}_i|}$$

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

ほぼ同じだが,
v は **A** がついている

$$u_{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{r}_i) = \frac{(\boldsymbol{R}_{\mathrm{A}} - \boldsymbol{r}_i)^T (\boldsymbol{r} - \boldsymbol{r}_i)}{|\boldsymbol{R}_{\mathrm{A}} - \boldsymbol{r}_i|}$$

$\mathbf{R}_{\mathrm{A}}$：全体での分野Aの重心位置（円周上）　$\mathbf{r}_i$：RF_iの位置（円内）　**r**：円内の任意の位置

$\boldsymbol{R}_{\mathrm{A}} - \boldsymbol{r}_i$　$\mathbf{r}_i$ から $\mathbf{R}_{\mathrm{A}}$ に向かうベクトル

$\boldsymbol{r} - \boldsymbol{r}_i$　$\mathbf{r}_i$ から任意の点 **r** に向かうベクトル

分布の形状を決める

ある分野A に付いて，任意の位置(**r**)の高さ(H_A)を，$\mathbf{r}_i$とそのRFの論文数（N_i）で決める

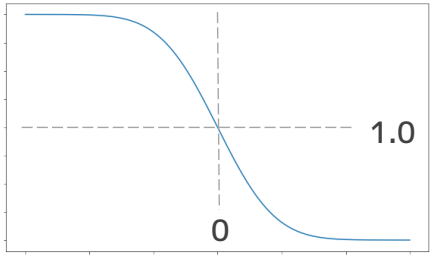
$$H_{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{r}_i; N_i) = \operatorname{erfc}\left(\frac{u_{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{r}_i)}{a(N_i)}\right) \operatorname{erfc}\left(-\frac{u_{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{r}_i)}{b(N_i)}\right) \exp\left(-\frac{v_{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{r}_i)^2}{c(N_i)^2}\right)$$

※ 単純な拡散に対して，任意方向に強めたり弱めたり

ここで…

$$\operatorname{erfc}(x)$$

普通の相補差関数



$$a(N_i) = 0.64\sqrt{2}\,N_i^{1/4}$$

$$b(N_i) = 0.12\sqrt{2}\,N_i^{1/4}$$

$$c(N_i) = 0.16\sqrt{2}\,N_i^{1/4}$$

分布の形状を決める

前頁の H_A は，ある分野A の成分だけを取りだして考えたもの

▼

実際には複数の分野があるので，ある地点の高さはそれらの合成

全体を揃えるために，正規化を行う

$$\tilde{H}_{\mathrm{A}} = \frac{H_{\mathrm{A}}(\boldsymbol{r})}{\max_{\boldsymbol{r}}\{H_{\mathrm{A}}(\boldsymbol{r})\}}$$

分布の形状を決める

ある地点 **r** の高さ（ K(**r**) ）は，各年のRF全体において
各分野の高さ成分を合成したもののうちの最大値とする

$$K_{\mathrm{YEAR}}(\boldsymbol{r}) = K_0 \max_{i \in S_{\mathrm{year}}} \left\{ RI_i \sum_{\mathrm{A} \in \mathcal{R}} w_{i,\mathrm{A}}\, \tilde{H}_{\mathrm{A}}(\boldsymbol{r}, \boldsymbol{r}_i; N_i) \right\}$$

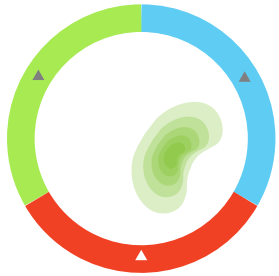
論文では $K_0 = 1$ を使用

$$w_{i,\mathrm{A}} = \frac{N_{i,\mathrm{A}}}{N_i}$$

← RF_i に占める
分野Aの論文数

$$RI_i = \log \left\{ \frac{X_i}{\sum_{\mathrm{A} \in \mathcal{R}} N_{i,\mathrm{A}} \langle C_{\mathrm{A}} \rangle} \right\}$$

← リサーチインパクト
引用数に関する偏差値的な



完成！

まとめ

SciLandscape は 分野の融合度合いを可視化したもの

- RI という指標と，ヒートマップ時に裾野の形状を少し変えるところがポイント

ひつようなもの

- 各分野のコアペーパー数
- RF ごとの，各分野の割合

…あとは，この2つを操作して配置を決定