

データサイエンティスト育成に向けて

和歌山大学 学長 瀧 寛和

2016.07.07
統計データ利活用シンポジウム



データサイエンティスト育成の必要性

1. 利用可能なデータの増大(オープンデータやWEB上のデータの活用)
2. データを活用して社会の状況を知る能力の必要性
 - 市場、自然現象、社会状況、最新の研究・科学を見通す能力
3. データに基づく客観的な判断能力の必要性
 - 得られるデータが少ないときは、勘や個人の経験で問題解決がなされてきた

⇒ 第4次産業革命を牽引する人材を育成

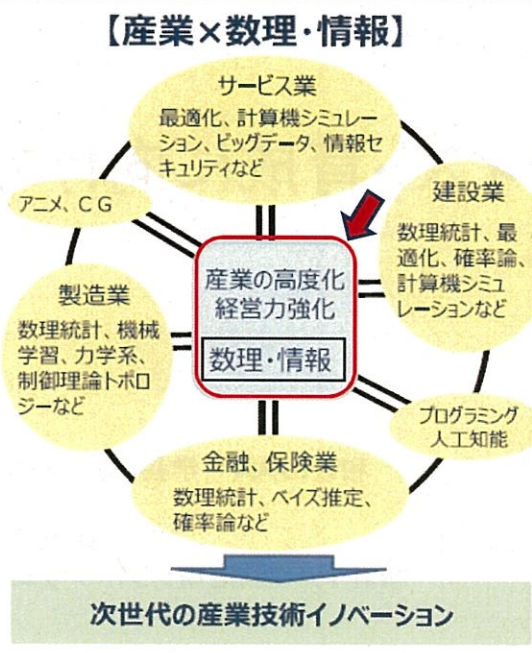


「第4次産業革命に向けた人材育成総合イニシアチブ」
～現状と課題①～

参考

数理・情報が第4次産業革命の鍵

米国では、好待遇の上位職種を数理・情報人材が独占。



2015年の米国内高待遇職種ランキング

※週刊「イブント」2016/1/23抜粋 (米Careercast.comより作成) 年収 (中位所得)

| | |
|---------------------|-----------|
| 1位 保険数理士 | 9万4209ドル |
| 2位 聴覚訓練士 | 7万1133ドル |
| 3位 数学者 | 10万2182ドル |
| 4位 統計家 | 7万9191ドル |
| 5位 生物医学エンジニア | 7万1133ドル |
| 6位 データサイエンティスト | 12万4149ドル |
| 7位 歯科衛生士 | 7万1002ドル |
| 8位 ソフトウェアエンジニア | 9万3113ドル |
| 9位 作業療法士 | 7万7114ドル |
| 10位 コンピュータシステムアナリスト | 8万1150ドル |

Best Jobs of 2014

- 1位 Mathematician (数学者)
 - 2位 Tenured University Professor
 - 3位 Statistician (統計家)
 - 4位 Actuary (保険数理士)
 - 5位 Audiologist
 - 6位 Dental Hygienist
 - 7位 Software Engineer (ソフトウェアエンジニア)
 - 8位 Computer Systems Analyst (コンピュータシステムアナリスト)
 - 9位 Occupational Therapist
 - 10位 Speech Pathologist
- ※米・求人情報サイトのキャリアキャスト・トップコム発表

(※文部科学省「第4次産業革命に向けた人材育成総合イニシアチブ」関連資料より)



「第4次産業革命に向けた人材育成総合イニシアチブ」
～③未来社会を創造するAI/IoT/ビッグデータ等を牽引するハイレベル人材育成の取組～

理化学研究所AIPセンター※における 研究開発・トップレベル人材育成

- ・ 組織においてデータ利活用を先導できる高度なレベルのデータサイエンティストを育成
- ・ 高度なセキュリティ知識と管理能力を持つサイバーセキュリティ人材育成
- ・ JSTの戦略的創造研究推進事業等と連携して若手研究者による研究を推進し、人材育成を加速

卓越研究員制度の活用

- ・ 優秀な若手研究者が安定かつ自立して活躍できる環境を推進
- ・ 産学官の研究機関をフィールドとして活躍できる若手研究者の新たなキャリアパスを開拓

ビッグデータ利活用プラットフォームに関する 国際研究拠点における若手人材育成

- ・ 我が国が強みを生かせる分野においてビッグデータを戦略的に利活用するプラットフォームを構築するための国際研究拠点を整備
- ※ 対象分野例：ナノテック・材料、健康・医療、防災・減災など
- ・ 研究拠点において、優秀な博士課程学生やポストク等に活躍の場を与え、データ利活用分野での専門人材の育成を加速

データサイエンティスト・キャリア開発支援

- ・ 博士課程学生やポストク等の若手人材に対して、各々の専門性を有しながら、企業が求めるデータサイエンス等の高度なスキルセット※を獲得する機会を提供
- ※ ITスキル、ビジネススキル、統計解析スキルの3領域を含む
- ・ 産業界とも連携し、問題解決型演習 (PBL)を活用した短期研修等により、データサイエンティストとしての能力をも身につけて活躍するためのキャリア開発を支援

(米国のInsight Data Science Fellows Programを参考)

※AIP:人工知能/ビッグデータ/IoT/サイバーセキュリティ統合プロジェクト

高等教育段階における取組
 > 全学的な数理・情報教育の強化
 > 数理・情報の専門人材の育成強化

(※文部科学省「第4次産業革命に向けた人材育成総合イニシアチブ」関連資料より)

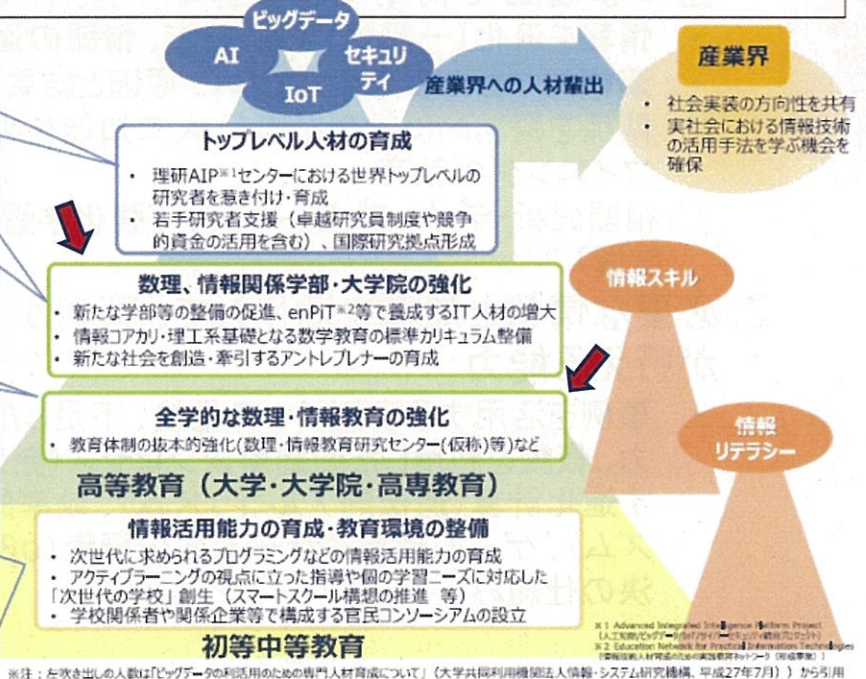


「第4次産業革命に向けた人材育成総合イニシアチブ」 ～未来社会を創造するAI/IoT/ビッグデータ等を牽引する人材育成総合プログラム～

- ・「第5期科学技術基本計画（平成28年1月閣議決定）」において謳われている「超スマート社会」の実現、及び「理工系人材育成に関する産学官円卓会議における行動計画」等を踏まえ、関連施策の一体的な推進が求められている
- ・生産性革命や第4次産業革命による成長の実現に向けて、情報活用能力を備えた創造性に富んだ人材の育成が急務
- ・日本が第4次産業革命を勝ち抜き、未来社会を創造するために、特に喫緊の課題であるAI、IoT、ビッグデータ、セキュリティ及びその基盤となるデータサイエンス等の人材育成・確保に資する施策を、初中教育、高等教育から研究者レベルでの包括的な人材育成総合プログラムとして体系的に実施

参考：必要とされるデータサイエンス人材数(※)

| |
|--|
| <ul style="list-style-type: none"> ・世界トップレベルの育成 (5人/年) ・業界代表レベルの育成 (50人/年) ・棟梁レベルの育成 (500人/年) |
| <ul style="list-style-type: none"> ・独り立ちレベルの育成 (5千人/年) ・見習いレベルの育成 (5万人/年) <small>目標 (MOJ, 2016)</small> <small>日本: 3,497人</small> <small>US: 257人, 中国: 177人</small> |
| <ul style="list-style-type: none"> ・リテラシーの醸成 (50万人/年) <small>大学入学者数: 約40万人</small> |
| <ul style="list-style-type: none"> ・小学校における体験的に学習する機会の確保、中学校におけるコンテンツに関するプログラミング学習、高等学校における情報科の共通必修科目化といった、発達段階に応じたプログラミング教育の必修化 ・全ての教科の課題発見・解決等のプロセスにおいて、各教科の特性に応じてICTを効果的に活用 ・文科省、経産省、総務省の連携により設立する官民コンソーシアムにおいて、優れた教育コンテンツの開発・共有等の取組を開始 <small>高等学校: 約137万人 (13年度)</small> <small>中学校: 約150万人 (13年度)</small> <small>小学校: 約600万人 (14年度)</small> |



(※文部科学省「第4次産業革命に向けた人材育成総合イニシアチブ」関連資料より)

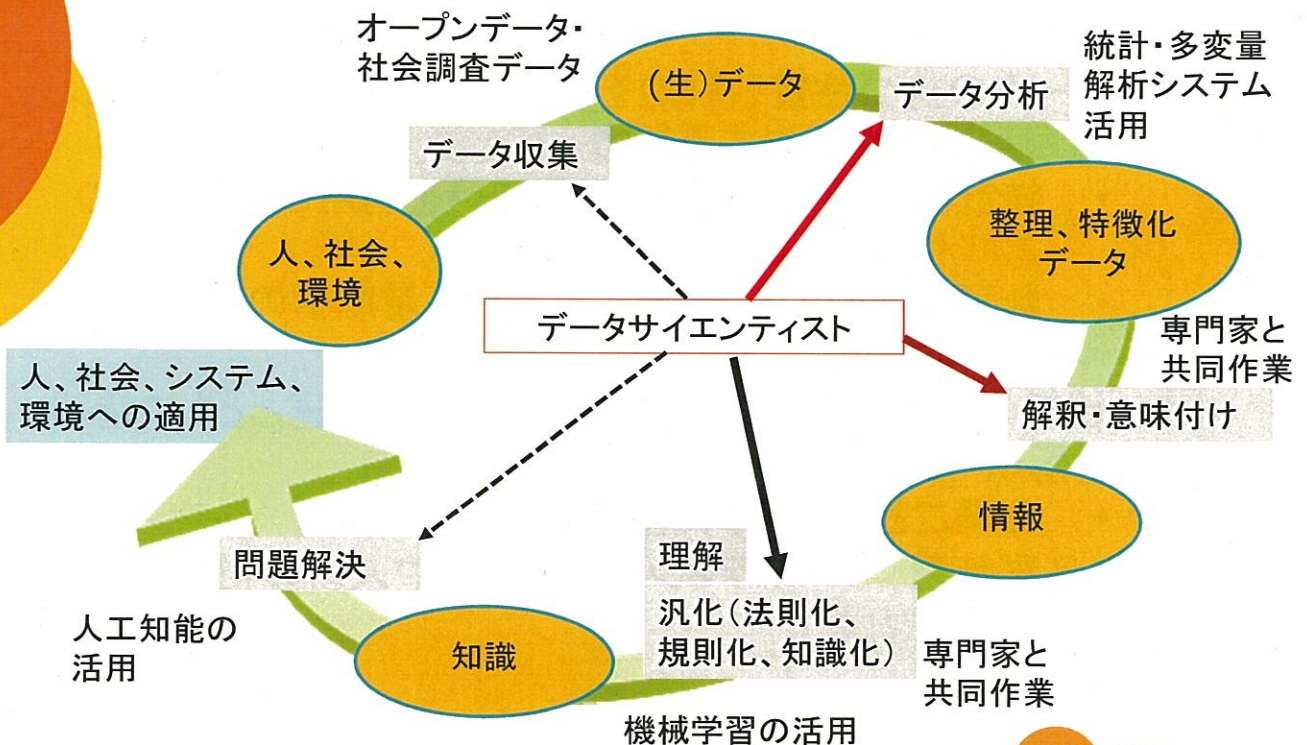
データを利用した問題解決のプロセス

1. データの持つ性質(特徴)を**分析**し、そのデータを**解釈**し、活用可能な情報にする能力
 - ・ 統計的なデータ分析と背景知識を利用したデータの**意味付け(情報化)**
 - ・ 数理・情報教育: 統計解析と**多変量解析**の教育 (分布・分散、クラスタリングなど)
2. 情報の**適用分野の基礎知識**を持ち、専門家とともに情報を**解釈・意味付け**する能力
 - ・ 分野ごとのデータに基づく**客観的な判断** (品質管理、マーケティング、高齢化や人口の把握、観光データ解釈など)

データを利用した問題解決のプロセス

1. 様々な分野において、得られた利用可能な情報を理解して、色々な場面で利用できる**基準や規則化(知識化)を行う能力**
 - ・ 情報を汎化(一般化)することで、情報の適用範囲を広げる学習力
 - ・ 演繹的なものの考え方でなく、原因と結果の関係や、事象や事項の関係を**帰納的に規則を導く**、人工知能を活用した機械学習・データマイニングの教育
(相関分析・ディープ・ラーニング、強化学習、概念学習、説明に基づく学習)
2. 必要な情報と知識を活用して、**適切かつ効率的に問題解決**が行える能力
 - ・ 事例を活用する事例ベース推論、不足した情報のもとの仮説生成力、仮説を利用した問題解決プロセス(発想推論)、(準)最適解を探す進化計算(遺伝的アルゴリズム)、効率的な解の探索(A*アルゴリズム)、ゲーム理論やゲーム木の探索($\alpha\beta$ サーチ)、協調的問題解決の仕組み

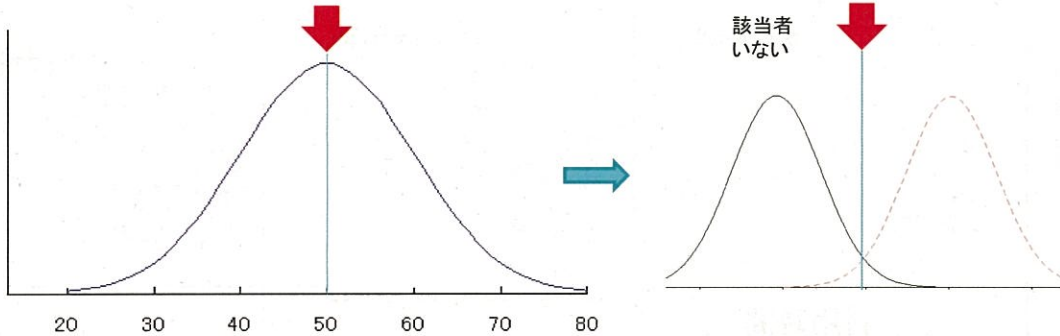
データを利用した問題解決のプロセス



統計的な考え方の大切さ

分布の理解の大切さ(統計量には、色々な見方や特徴をみる仕組みがある)

物事は平均値で考えがちである。試験の平均点が50点なので、講義内容を平均点の学生に合わせて行くと不満が続出(分からないと易し過ぎるの意見)。



正規分布をイメージしていたが。

成績はこんな分布だった

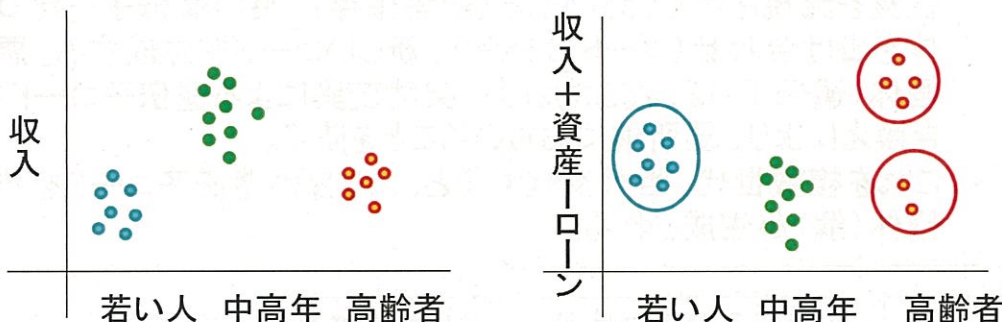
成績が2極化していることが分かり、易しい内容と高度な内容を講義して、学生の満足度向上

分布や分散、標準偏差など統計的な概念教育は大切

多変量解析の活用を理解する

適切なクラスタリングが大切、データの集まりのとらえ方で解釈に影響する(分析項目で変わる分け方)

収入でクラスタリングすると、中高年の収入が良いので、そこをターゲットに商品開発をしたが、全く売れなかった。



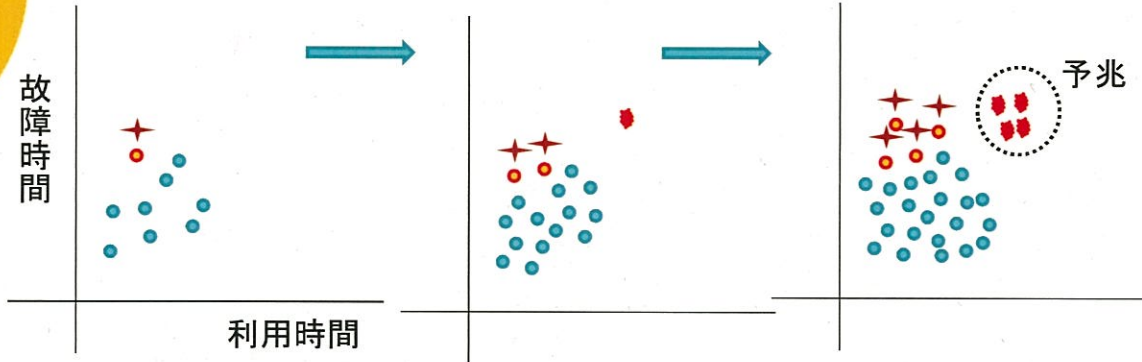
収入+資産-ローンの軸で見ると、中高年は低く、高齢者は2つのクラスになった。若い人用商品と裕福な高齢者商品で売り上げ倍増

主成分分析、因子分析、クラスター分析、重回帰分析などがある。
主成分分析では、分散(データが密集しない)するように、成分を決めることができる。

ビッグデータを扱うことの意味

ビッグデータで何が見えるか：

1. データ間の共起関係(同時に良く出現する組み合わせ)の精度が上がる
2. 希少データが検出できるようになる



✦ 赤○と星形の関係がより多く見つかり、関係の精度が増す。

データ数が増えるごとに、予兆が発見しやすくなる。

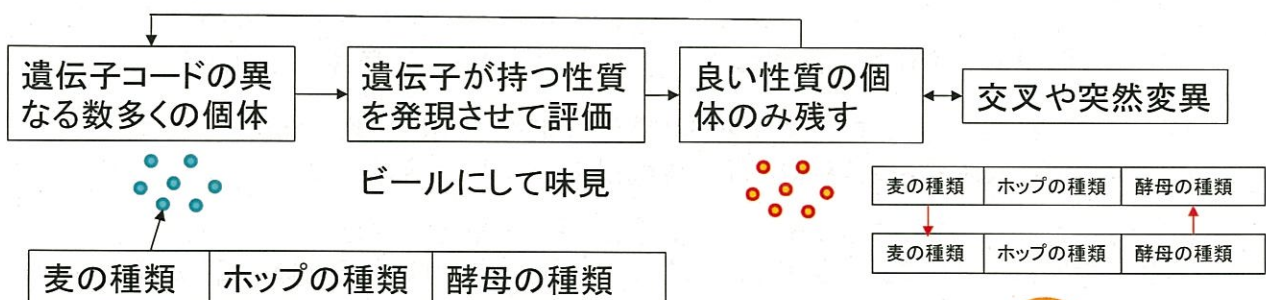
予兆： 故障の原因、新しい商品ニーズ、流行の兆し



人工知能の様々なアルゴリズムの活用

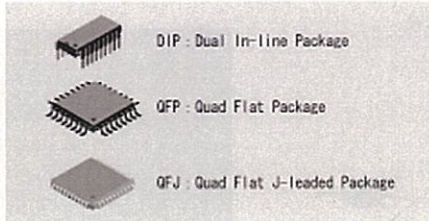
遺伝的アルゴリズムの活用

- 生物の進化を模倣したアルゴリズム
- 物事の性質を遺伝子のようなコードに変換、そのコードが持つ性質を発現させて、評価する(適者生存)。良い遺伝子を持つ個体を掛け合わせ(コードの交叉)、新しいコードを生成する。悪い個体(遺伝子)は、淘汰される。突然変異による遺伝子コード書き換えにより、局所解におちいることを防ぐ。
- これを複数世代、世代交代すると、より良い遺伝子コードを持つ個体(解)が生成される。

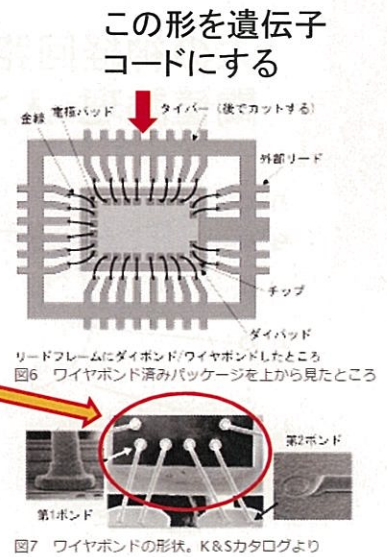


実例1 (遺伝的アルゴリズム<解の探索>の利用)

ICチップとリードフレーム(ICから出ている足の部分)を金線で配線する際に、最適なリードフレームの形(位置、サイズ)を考慮した設計ができるようになった。遺伝的アルゴリズムを利用して、金線の配線量を削減、費用削減につながった



金線が短くなる配置



Semiconductor Japan Net HP
<http://www.semiconductorjapan.net/serial/lesson/10.html>

機械学習(論理的手法)

- ◆ 概念学習: その概念の正例と負例を与えて、その違いから正例を説明する性質(区別する基準)を学習する。
- ◆ 猫の概念を学習する

猫とは?
正例

区別する基準

猫とそれ以外を分ける

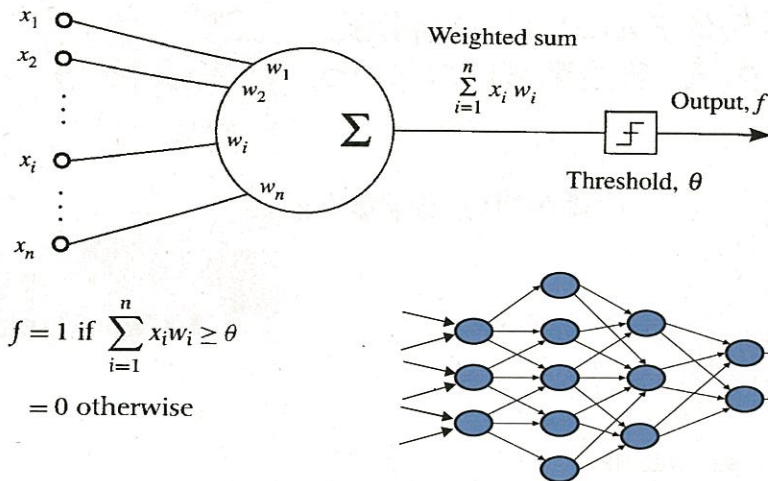
猫じゃない
負例

顔、鼻、目、鳴き声、耳。。。。

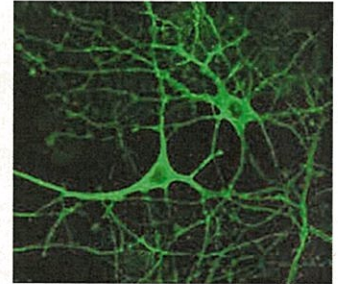
囲碁や将棋であれば、良い手と悪い手の違い(手を選ぶ基準)を学習する。

ニューラルネットワーク

人の神経回路網を模倣した演算素子である。
 閾値論理: 入力値の合計が閾値を超えると出力が1となる。

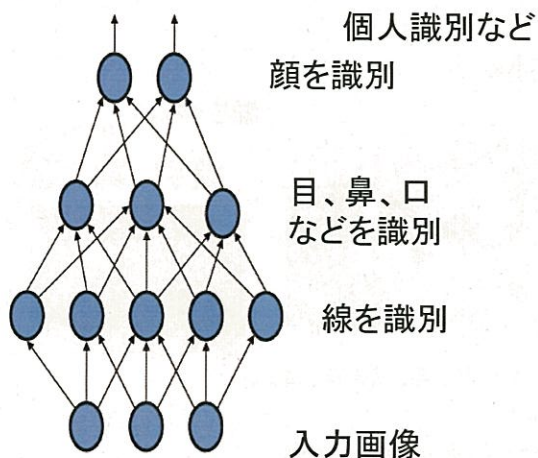


多層のニューラルネットワークで識別演算

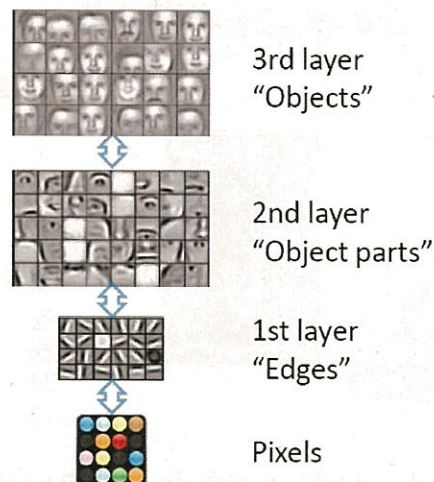


人工知能アルゴリズム (Deep Learning) <ニューラルネットワークによる機械学習>

神経回路網の学習



Feature representation



<https://deeplearningworkshopnips2010.files.wordpress.com/2010/09/nips10-workshop-tutorial-final.pdf>

データマイニング

- ◆ 単語間の関係を統計的に分析(共起確率)
- ◆ バケット分析:この買い物かごの組合せを分析
 - ❖ 例:コンビニやスーパーのレシートには、買い物かごの中の商品の組み合わせがデータとして残る。
 - (ワインとチーズの組み、スルメと日本酒の組みなど)
 - ❖ 結果例:米国:おむつを買う人は、ひげそりも。
- ◆ できるだけ多くのデータを分析する方がより興味深い関係が得られる=>ビッグデータ解析へ

囲碁でも人工知能がプロ棋士を凌駕 (打つ手の可能性が極めて多い問題)

Googleの囲碁人工知能・AlphaGo、李世ドル九段
第5局4勝1敗でAlphaGoの勝ち越し

圧倒的な学習力 AlphaGoのスペック、
CPU(中央計算処理装置)1202個、
GPU(グラフィックプロセッサー)176個
プロ棋士の棋譜を1日約3万局学習



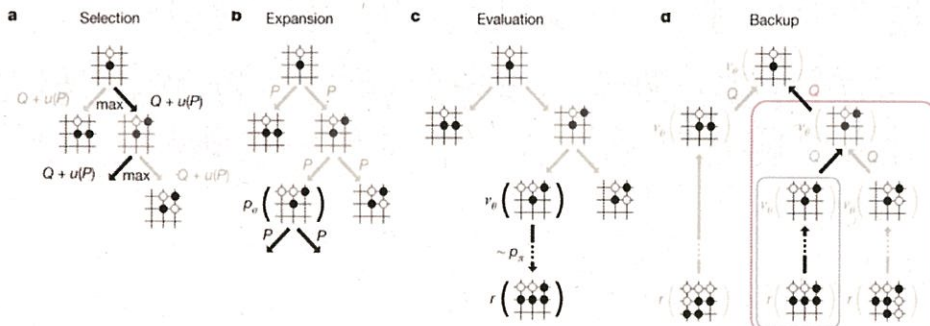
囲碁のプロ棋士に勝つ人工知能

1. 強化学習の利用(手を打ちながら学習)

- ・ 場面と良かった手の組みの点数を高くする(局面と打つ手の組)
<動的な学習>

2. ディープラーニングの利用(過去の棋譜を学習)

- ・ ある場面から次に良いとされる局面に導くパターンを学習し、実際に打つ手を選ぶ(局面から局面への展開への打つ手を選ぶ確率を計算)<局面と手の選択>



強化学習:うまいった流れを後ろから辿って、選びやすさの点を加算



国立大学法人
和歌山大学

データサイエンティストの育成

データサイエンティストは統計学と人工知能と専門分野(教育、経済、観光、工学)を学ぶことで育成できる(和歌山大学の教育を例として)

1. 和歌山大学学部・大学院の数理・情報教育の強化
2. 全学的なデータサイエンス、人工知能、IoT等の
教育研究強化に向けた整備→(例)サイバー知能教育研究部門
3. 経済学研究科(改組構想中)における統計データ活用教育強化
4. 社会人教育コースの活用

※学部や大学院教育では、教育から人材輩出までにタイムラグがあるため、社会人に向けた教育コースの活用

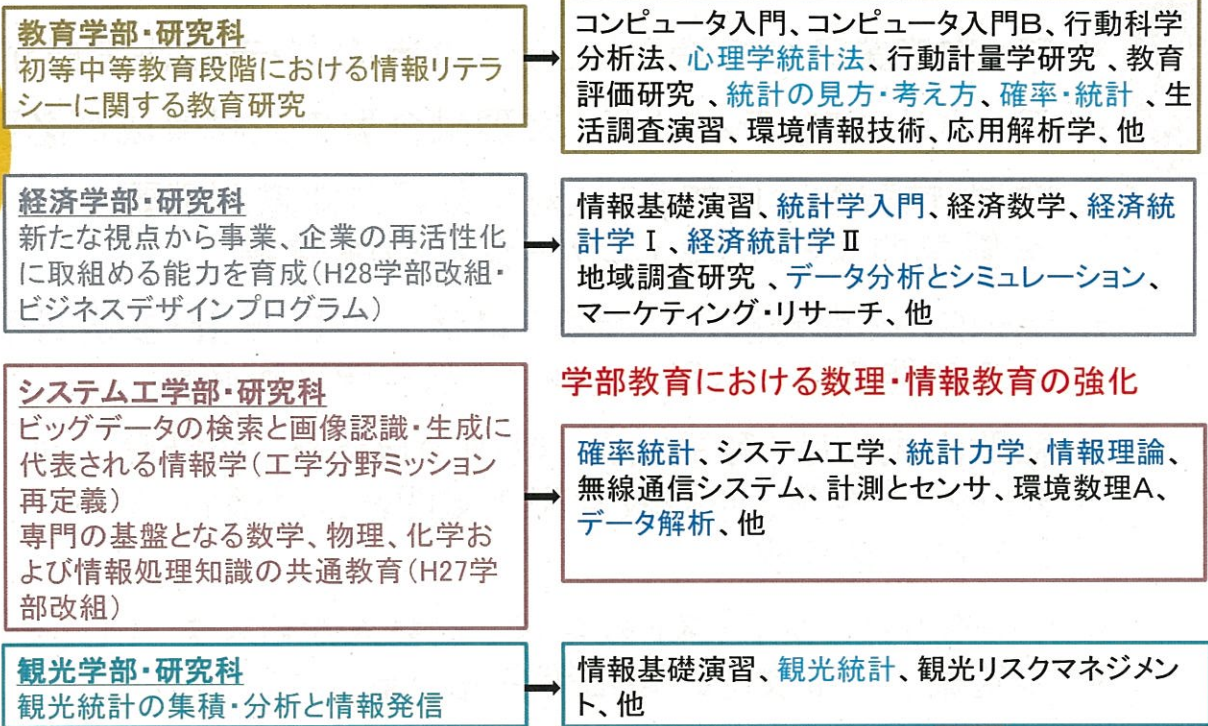
【例:経済学研究科】

科目等履修生として授業科目を履修し、その授業科目の試験に合格することで、単位を修得できます。そののち、社会人特別選抜などの入学試験を経て、経済学研究科の正規の課程に進学すると、このコースで修得した単位を正規の課程の修了に必要な単位として参入することができます。

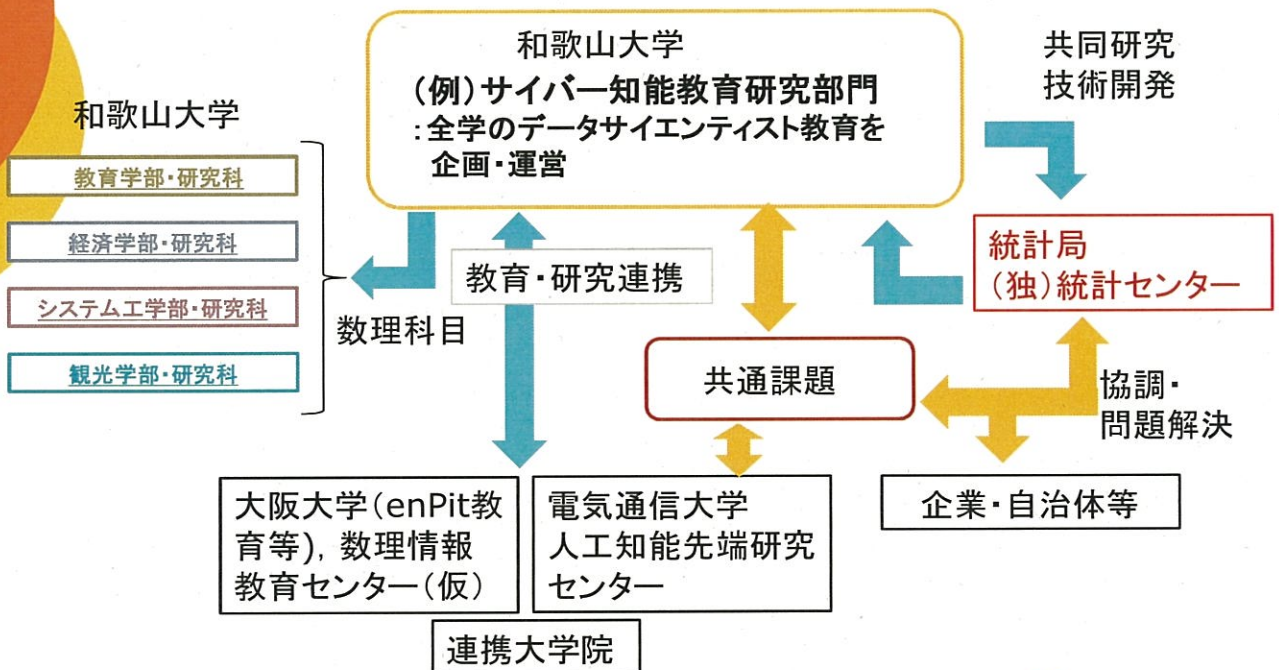


国立大学法人
和歌山大学

データサイエンティストの育成(数理・情報教育)



データサイエンス教育研究強化に向けた整備



まとめ

- 第4次産業革命の牽引に、データサイエンティスト育成が急務
- データサイエンティストは、文理融合型の人材育成が必要
- 数理・情報教育に加えて、対象分野の知識が必要
分野専門家とのコミュニケーション力も重要
- データから問題解決に至るプロセスでは、統計・多変量解析・人工知能の各システムを使いこなす能力が必要
- 和歌山大学は、和歌山県、県内の自治体、統計局、(独)統計センター、企業、他の教育研究機関との連携で、データサイエンティストを育成
- データサイエンティスト教育は、社会人からスタートし、研究科の整備、学部教育の強化等を進める