講義日：10月4日（火） はじめに：計算生命科学の概要

講師：神戸大学大学院　科学技術イノベーション研究科　特命教授　森　一郎

質　問　事　項	回　答
質問者：製薬会社勤務 製薬会社に勤めていますが、現状ではデータサイエンスを使う場面が ありません。今後、データサイエンスが普及していくにあたり、どのような準備 をしておくべきか知りたいです。	理論も大切ですが，実践も大事です． 実際にデータに触ると教科書に載っていないケースに多くぶち当たります． 現在，身近にデータは溢れています． 企業内でも聞いて回ると，研究から営業までデータサイエンスのニーズは多く存在します． ただ，それをデータサイエンスの枠として仕事が確立されているかどうかは難しいところです． また，組織内にデータがなくとも，例えばオープンデータ使いながら，実践を積むことも一つだと思います． 自発報告データFAERSやJADER，Clinical Trials.govなどのデータ， さらには論文やSNS等のテキストデータにも目を向けるとデータ量は膨大になります． それらデータを扱うためにはビッグデータをハンドリングする技術が必要ですし， さらに何らかのルール，仮説を見つけるのに役立つ機械学習の習得が必要です．
質問者：研究所勤務 先生の講義での質問、間に合わなかったのですが1点ございます。 八尾先生のレポートの　データに関する考え方を変える必要　についてです。 第3　因果関係でなく、相関関係を重視する とありました。 因果関係が分からない場合にも、偶然でも相関があれば、全く新規のデータ （化学物質）の（活性等の）予測が可能なのでしょうか？ また、扱えるデータの適用範囲の概念の定義は今後必要なくなるのでしょうか？	相関関係を検出したデータを吟味しない限り， それが因果関係（方向も含む）なのか，というのは判断できません． それは，偶然の相関関係かもしれません． 例えば，部分集団での偶然の相関かもしれません． しかしながら，相関関係を検出したデータが母集団と等価であれば， それは因果方向や交絡の如何に関わらず，何らかのルールとして見ることが可能となります． これがビッグデータ時代になり，相関関係に重きを置き始めた理由となります． ただ，化合物の活性予測は勝手が少し異なります． ケミカルスペースの全体をカバーした母集団データは， 少なくとも現時点では存在せず，予測モデルのベースとなったケミカルスペースを把握し， 適用データがそのスペースのどの辺に位置するか，を確認する作業が必要になるかと思います．