データに関わる人たちのすれちがい

データマイニングと統計数理研究会 (SIG-DMSM) 2008.7.23

データ分析からうまれる、広がる研究と交友の輪

産業技術総合研究所 神嶌 敏弘

http://www.kamishima.net/

データ分析手法の使われる範囲は広がっていると思います.
そのため，データ分析手法を作る側の各分野の間や，分析手法を作る人と使う人との間で接点が広がってきていると思います.
しかし，お互いに期待することが違うため，すれちがいが起きてしまうことがあるように思えます.

手法を作る人たち

データ分析分野のもともとの動機

機械学習

- 人が知識を学ぶように，機械も知識を学ぶ

統計

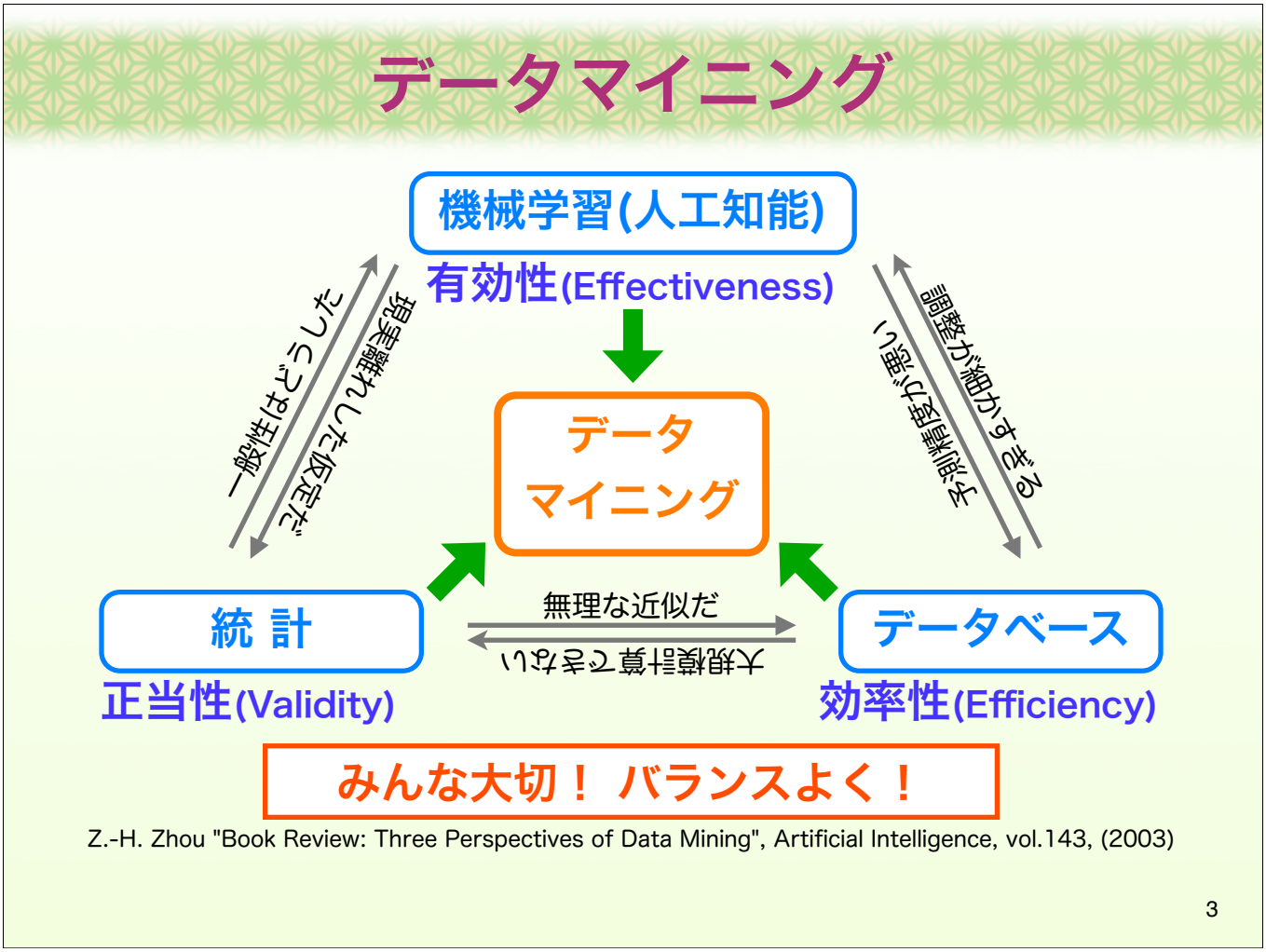
- データが示す事柄を明らかにする

データベース

- もっと大規模に，高速にデータを利用

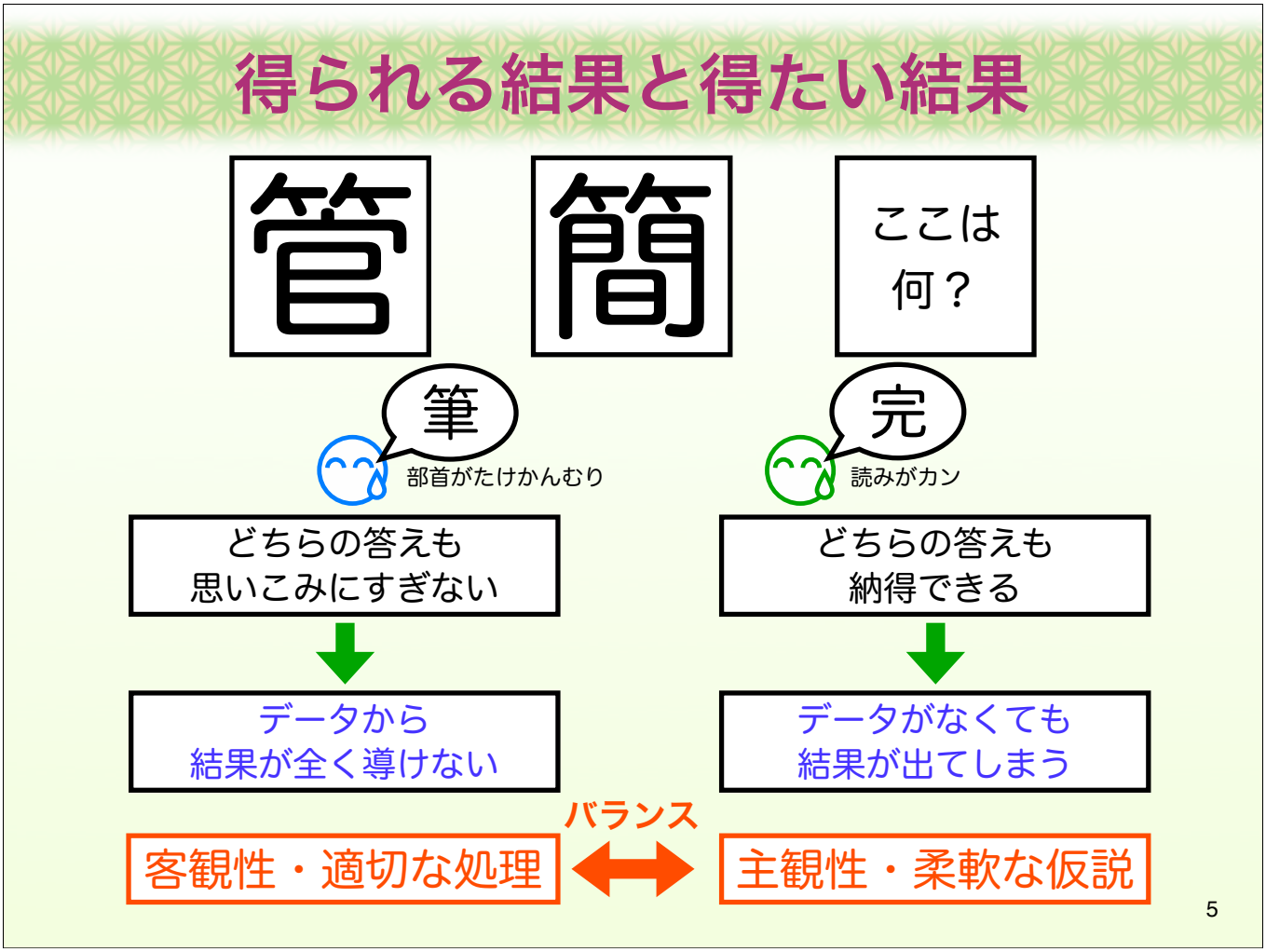
みんな，もっとデータから結果・知識を得たい

データ分析手法を作る側の，いろいろな分野の人たち関係から．
今はだいぶ変容しているが，各分野のもともとやりたかったことは…
・機械学習：人が学ぶように，機械にも，自分が観測したデータから自動的に知識を学ばせる
・統計：データを集め，そのデータが示すことを明らかにしたい
・データベース：データをもっと大規模に，もっと高速に利用できるようにしたい
本来は，データからもっと結果や知識を得たいという観点では一致しているはず．



データマイニング分野の三つの著名な教科書を題材に，データ分析手法の三つの分野の関係から，データマイニングを論じた書評が面白いので紹介．どの分野も，よりよく分析したいという点では一致していた．しかし，機械学習では予測精度，統計では結論の普遍性，データベースでは効率をそれぞれ重視し，互いになんとなく距離がでてきてしまった．確かに，いずれかの観点を重視すると，他の観点からは良くない分析になりがちである．しかし，実際にデータを分析するときには，どの観点もそこそこ重要で，各観点のバランスが大事だというかけ声がデータマイニングといえるだろう．各分野はバランスを意識しつつデータ分析にかかる必要があるのでは？

データ分析手法を作る人たちと使う人たちの間のすれちがいも大きいと思います.
データ分析手法の開発者は，手法が使われる目的と前提が想定されたとおりに使われなければ，意味のある結果ではないと考える.
データ分析手法の利用者は，手法の目的や前提がはずれていても，データから結果が得られることが重要.
ここですれちがいが生じる.



「管」「簡」の次は何か？「部首」に注目すると「筆」が，読み方に注目すると「完」などが適切.

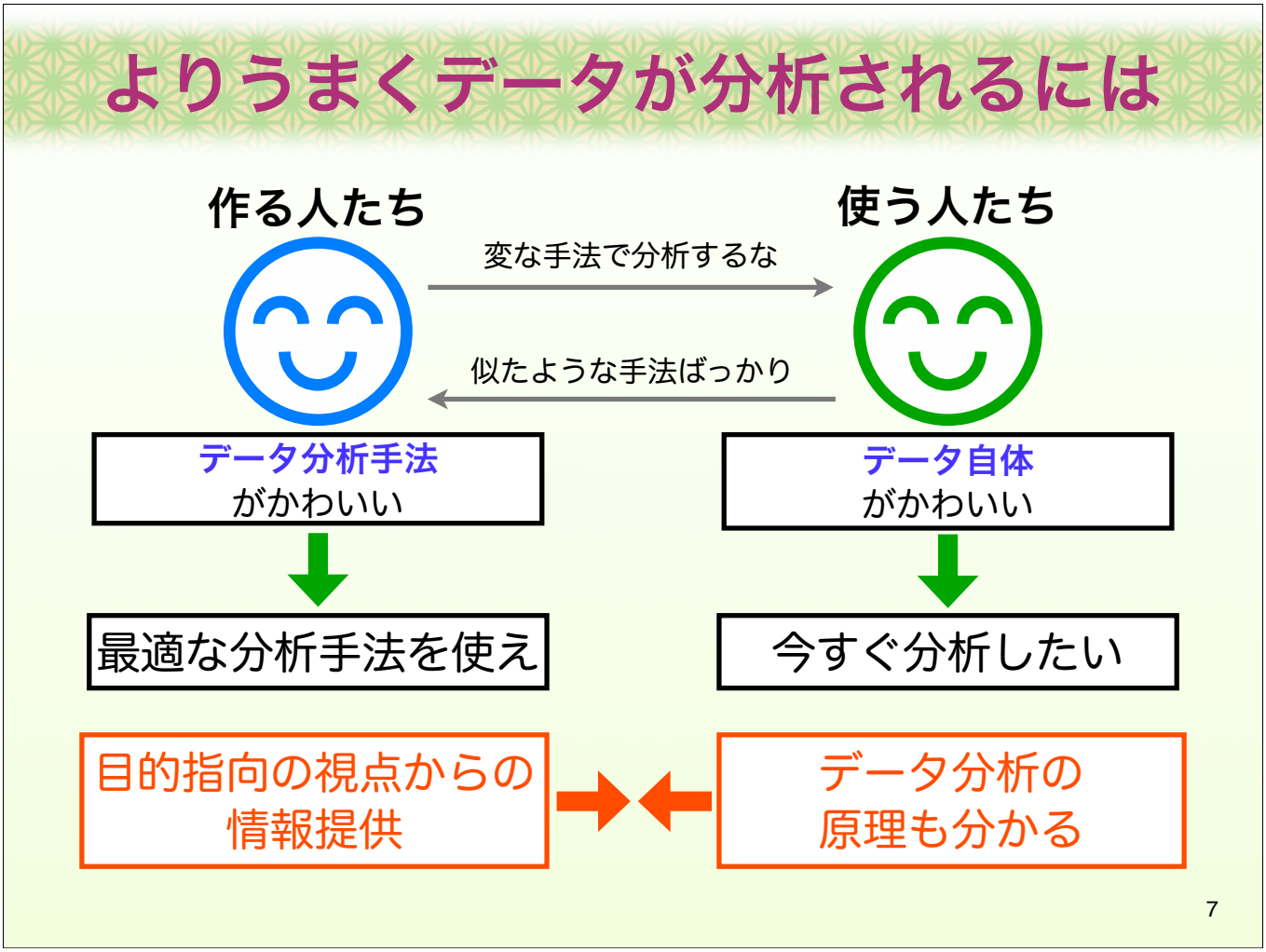
この注目点は，データ分析の言葉でいえば「モデル」や「背景知識」である．すなわち，データから導かれる結果は，ある観点の下でのみ正しい．これは，帰納的推論の限界.

データ解析手法の制作者の立場：こんなに少ないデータや，サンプルが偏ったデータからは，なんらかの知識を反映させた，いわば，思い込みの強いモデルなしには結果はでない．よって，弱い前提でのでの結果について保証はできないので，そういうことを求められても困る.

データ分析手法の利用者の立場：分析手法の細かいことはまかせるから，貴重なデータなのだから，とにかく結果がでないのでは困る.

現実には，客観性のある手法の前提を守った適用と，主観に基づいた仮説設定に基づく強引さを，状況に応じてバランスをとって結果を導く.

現実には，著名なKDDプロセスのように，適用の正しさと結果の良さが折り合うように，いろいろな手法やモデルを組み合わせつつ試行錯誤して結果を導く．データ分析をブラックボックスと考えて，データを入れるだけでは結果はでない．



すれちがいを埋めるにはどう歩み寄ればよいのか？
作る人たちは最適な選択ではなくても，そこそこに妥当な手法の選択でも受け入れる．教科書は後方参照が生じたりしないように，手法の中で使われる原理の関連によって分類されている．これを，各手法の特徴や用途の観点から整理した情報提供はできるか？
データ分析手法を使う側も，分析の段階をおまけと思わないことが必要ではないか？

さいごに

データ分析手法を作る人たちの間で

分析手法を作る人たちと使う人たちの間で

うまくバランスをとって

データを生かさなくてはいけない

そのために何をすればよい？