Scientific Data Management
Why is it so hard?

Nat Goodman

Institute for Systems Biology

September 2003

What is scientific data management?

- ~~What is science?~~

- What is scientific data?
  - Content areas
  - Bench/cage/bedside to publication and back
  - Small vs. big science

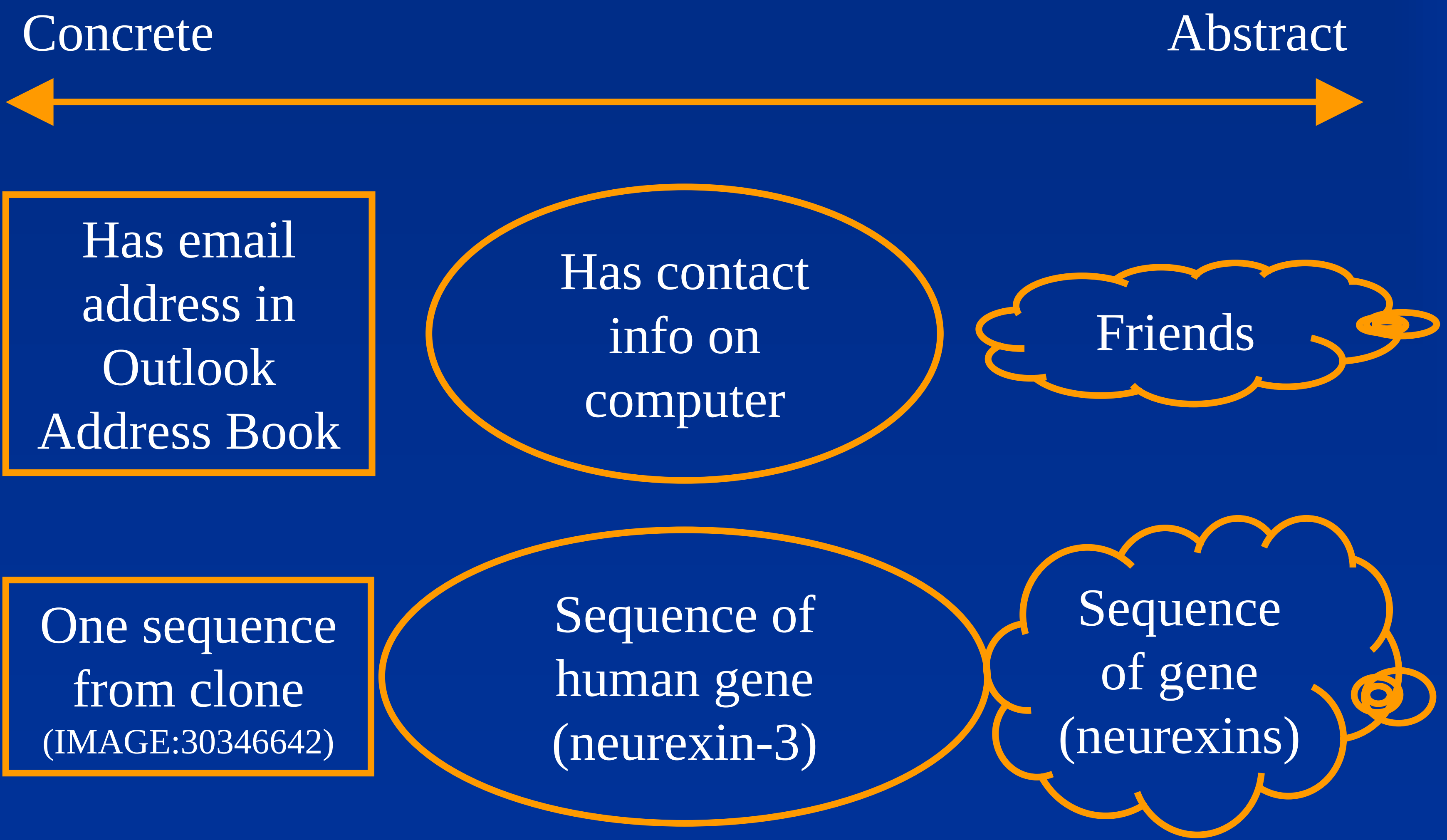
- What do we want to manage?
  - Store/fetch
  - Analysis
  - Sharing

Scientific data sub-problems

- Basic vs. clinical
- For basic:

	Pre-publication	***Public***	***Post-publication***
Small science	???	Literature	Access to databases
Big science	LIMS Analysis pipelines	Databases	Data integration

Concrete is cheap - abstract is dear



Concrete

Abstract

Has email address in Outlook Address Book

Has contact info on computer

Friends

One sequence from clone (IMAGE:30346642)

Sequence of human gene (neurexin-3)

Sequence of gene (neurexins)

Human genes over time

1) name (HD), cytogenetic position (4p16.3), alleles

2) name, cytogenetic position,
mRNA sequence, coding region (➔ protein sequence)

3) name, cytogenetic position, splice forms

    splice form = mRNA sequence, coding region

☞ name, cytogenetic position, genomic coordinates, exons,
splice forms

    exon = genomic stretch

    splice form = (list of included exons (➔ mRNA sequence)
    OR mRNA sequence), coding region

    included exon = all or part of exon

5) name, genomic coordinates, exons, splice forms,
alleles, gene families (paralogs), othologs

Complications

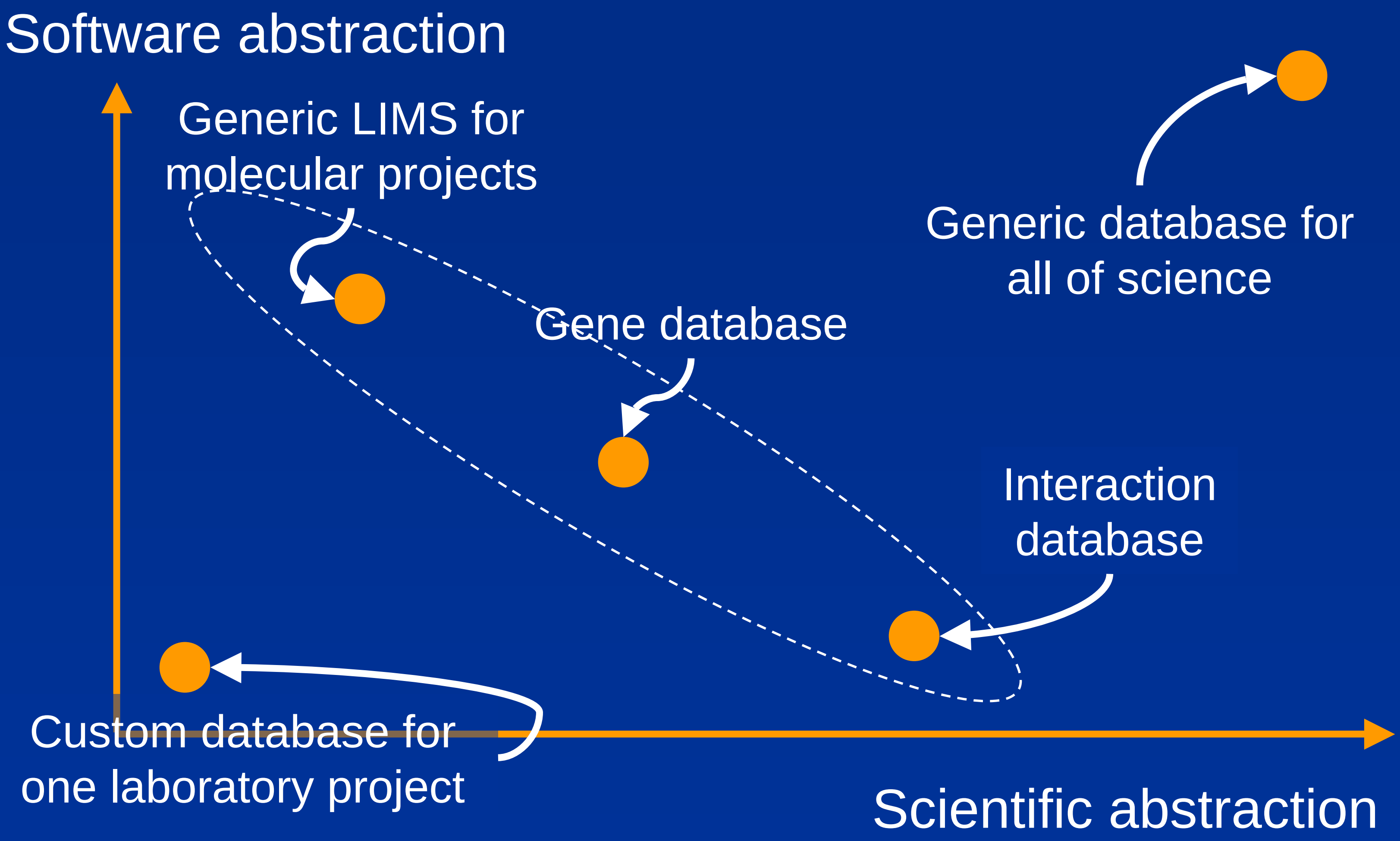
- Granularity issues
  - Some data abstract [Dean C et al. Nat Neurosci. 2003 Jul;6(7):708-16]
    “neurexins nucleate assembly of cytoplasmic scaffold ”
  - Some data specific (hypothetical example)
    mutation in specific base of human neurexin-3 affects specific splice form
- Attribution and evidence
- *Unstable* data
  - Conflicting data common
  - Errors common
    - Big databases notoriously error-prone
    - No one has ever calibrated literature – may be just as bad!
  - Abstract data evolves
- Versioning, esp., for data derived from public databases

➔ What good is all this unless application programs can exploit?

Database of all genes

- Gene concept evolving at different rates in different organisms
- Database of all genes must transcend these differences
  1) Least common denominator
  2) Union of complexity
  3) Something fancy – variable complexity – abstraction!

Finding the sweet spot


Software abstraction
Generic LIMS for molecular projects
Generic database for all of science
Gene database
Interaction database
Custom database for one laboratory project
Scientific abstraction

Data models

- *Data model* can mean

  1) Design of specific database or data type
     e.g., GenBank sequence data model
  2) Computer language or system for encoding designs
     e.g., SQL data model
     ORACLE data model
  3) Formalism for expressing data designs
     e.g., relational data model
     object data model

- Different data models (in sense #2 or #3) have different expressive power

Data model history

- Relational model (Ted Codd, circa 1968)
  - Demolished all predecessors – start of modern era
- Entity-relationship, functional, many others
  - Emerged and competed briefly in 80s
- Object models
  - Emerged and competed briefly in 90s
  - Object-relational persists (“behind closed doors”)
- Meanwhile, in AI and knowledge representation
  - Semantic data models – much more powerful
  - Never accepted by database folks

Relational data model is wimpy

- Flat tables very limiting
- Current gene data model

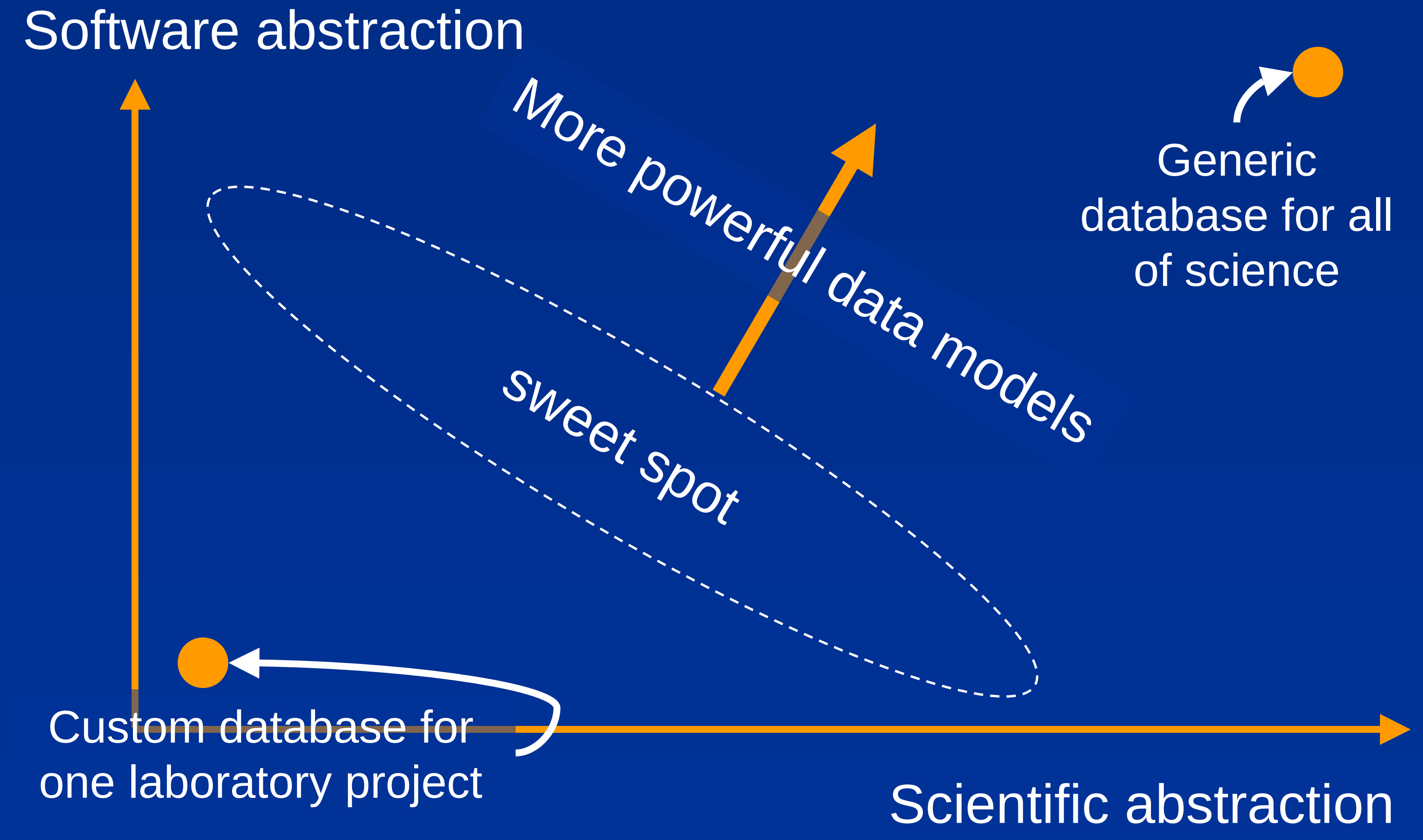
  name, cytogenetic position, genomic coordinates, exons, splice forms

  needs at least 5 tables

  Gene (id, name, cyto_position, chrom, start, length, strand)
  Exon (id, start, length)
  SpliceForm (id, name, gene, coding_start, coding_length)
  SpliceForm_type1 (spliceform, sequence)
  SpliceForm_type2 (spliceform, ordinal, exon, start, length)

  plus stored procedures or application code to "splice" exons and "translate" mRNA into protein
- Database of all genes needs several such models plus something that glues them together

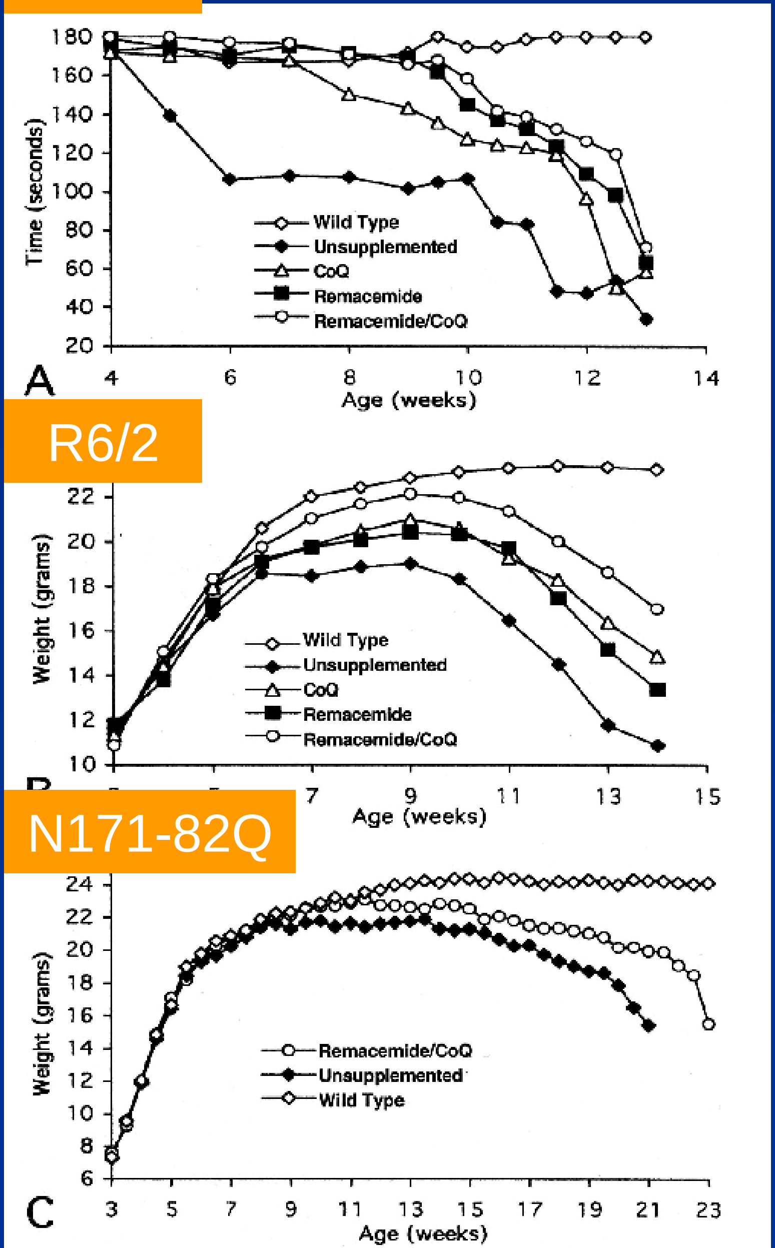
Better data models → better databases


Software abstraction
More powerful data models
Generic database for all of science
sweet spot
Custom database for one laboratory project
Scientific abstraction

Therapeutic agents in mouse models

- Agents: drugs, lifestyle, environment
- Mouse models
- Study design
  - Treatment arms
    - agent, formulation, route of administration, dose, schedule
  - Measures
    - protocol, schedule, reporting
  - Endpoints
- Study itself, i.e., one running of the design
  - How many animals per arm
  - Deviations from design
  - Results

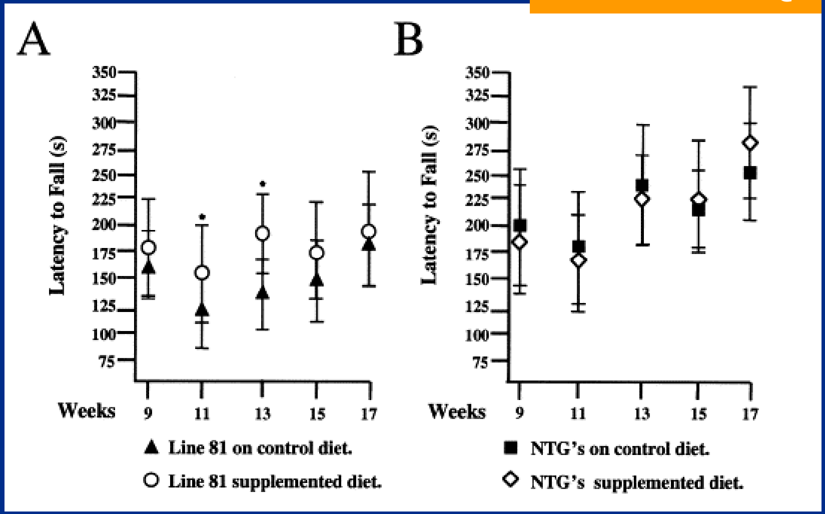
R6/2



R6/2

N171-82Q

N171-82Q



Fancy database can't compensate for incomparable data

Data from

Schilling G et al. Neurosci Lett 2001 Nov 27;315(3):149-53.

Ferrante RJ et al. J Neurosci 2002 Mar 1;22(5):1592-9

Why is it so hard?

- Vast field
- Complex, abstract, conflicting, evolving, incomparable data
- Relational data model barely copes
- Smarter data models may be futile unless application programs get smarter, too