Small Data

Examples for big and small data in your field:

BIG

Seequencing experiments

Mass-spectrometry experiments

Transcriptomis experiments
with few samples

Population-wide registry data

Images

MRI-scan and PET-scan

Pathology whole-slide sections

requires large data storage space
BIg data is not hypotheis driven compared to
traditonal prior hypotheis driven methods

Proteomic atlases

Electronic health records data bases
large cohort more than 10000 individuals

dynamic system with all protein-coding features

global structures like whole genome structure modeling
1 million single cell data

Metagenomics

small

Clinical data of rare diseases

Transcriptomis experiments
with few samples and pre-
selection of few interesting
genes

ex vivo drug screening (20 cell lines,
hundreds of drugs)

Rare disease data (10 patients),
3 time points

Proteomic cohort studies

less than 1TB of clinical data (after preprocessing
from raw)

few measurement time points per patient

single cell RNA data within 10000 cells

small number of individuals

How would you define small data? What characteristics would you consider?

Small n
https://en.wikipedia.org/wiki/Small_data

Number of observations vs number of variables

high-dimensional data in the sense that the number of parameters is large
compared to the number of samples/ examples (p>>n)

Depth of the research field and number of parameters/information associated to it

It has not so obvious patterns

too few data to directly fit a mechanistic model

low eterogeneity of the samples

Characteristic to consider the data as small data and not just 'not enough data':
no missing values, overall good quality of the data.

<10000 samples with simple explicit model

Differentiable Programming

What did you find most interesting/cool/surprising in the blogpost?

That big data techniques can be readapted to small data

Potential reuse or reproducibility and replication through modularity

The potential for increased interpretability of models

Differential equations fit to data via sensitivities -- as in physics,
epidemiology or pharmacodynamics -- are equivalent in all but terminology to neural networks.

That the deep learning field is rediscovering that it is useful to use what we know about
a problem (i.e. to include structure in models). ;-)

A statement you didn't understand or didn't agree with in the blogpost?

Redux

..."it's making a clear statement about what it thinks is happening outside"
The word THINKS is too strong for a model, I would say it's trying to guess or at best just
modeling and estimating parameters..

don't understand: Differential equations fit to data via sensitivities

differential equation fit to
data via sensitivities

How does the technology behind it work? What are the important ingredients?

Modeling

automatic differentiation

ODE system

Well-argued statistics

Algorithmic differentiation

Mechanistic understanding

Examples of applications where you would prefer "domain knowledge" models vs. black box models?

domain model

Anything with patient's data

metabolic network models

cellular signalling pathway models

with existing measurement calculation formulas

derivative to spatial/time variables

polymer model with physical models

whenever explanation/ causality/ interpretability is the goal

essential to advance science

signaling circuit

thermodynamic system as expression or translation process

black box / data-driven model

Quality control for MS/MS spectra matches or any other prediction for noisy data

Industry related, where you look for profit and no interpretability

CNN-based feature extraction and image classification

high-dimensional heterogeneous data

no empirical data

no labels

AlphaGo

sometimes, for prediction, a black-box model might be okay (but in the end we still want to understand why we predict what -> "explainable AI")