

Data Pre-processing: Intro and Integration

Anran Zhao

Research Associate at UK Data Service

Email: anran.zhao@manchester.ac.uk

Table of Content

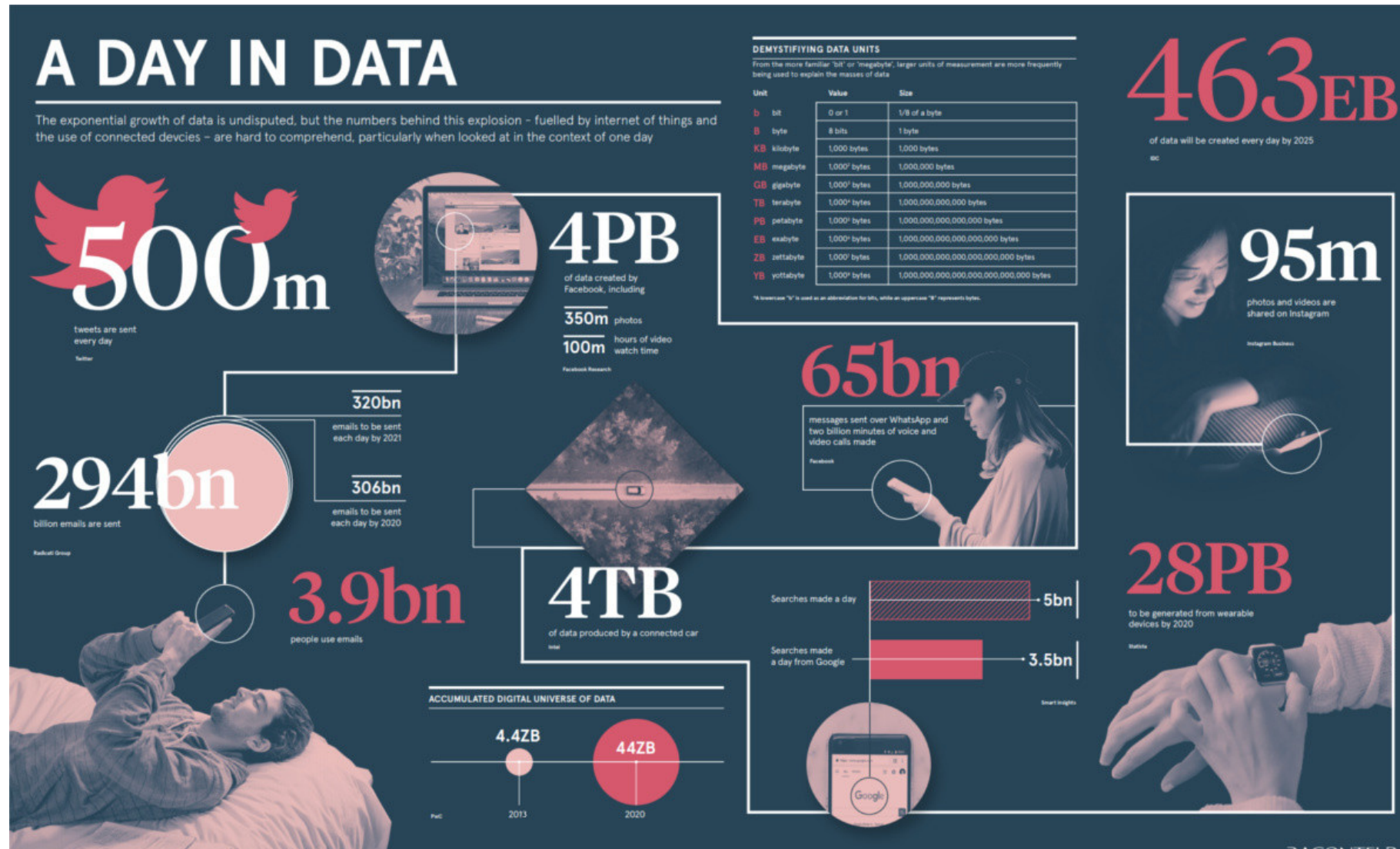
- **Definition, context, and collecting data**
- **Data integration (join tables)**
- Data cleaning (missing values, outliers, data types)
- Data reduction (correlation check and PCA)
- Data transformation (normalisation and one-hot encoding)

Data Pre-processing

- **Data pre-processing (a.k.a. data preparation)** is the process of manipulating or pre-processing raw data from one or more sources into a structured and clean data set for analysis. It is an important part of **Data Analytics**.
- Data pre-processing includes various tasks and considerations, e.g.
 - Selecting and acquiring data to use
 - Integrating different data sources together
 - Conducting exploratory analysis
 - Cleaning and repairing data, e.g., missing data, inconsistent data
 - Data reduction, transformation, etc.

Data Analytics – The Context

- ‘the sexiest job in the world in the 21st century’



Data Deluge

hospital patient registries
electronic point-of-sale data
stock trades OLTP telephone calls
catalog orders bank transactions tax returns
remote sensing images credit card charges
airline reservations social media commentary

“Data Deluge” –
the amount of
data being
generated is
overwhelming
the capacity of
organisations to
use them.

Big Data

Data Deluge

Data + preparation + analytics = actionable knowledge

the capacity of
organisations to
use them.

Big Data

Data Deluge

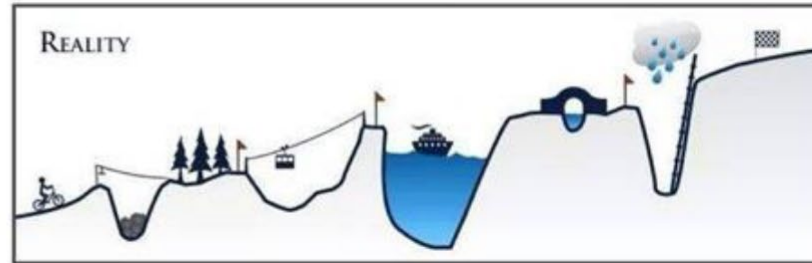
stories data calls turns rges

Data + preparation + analytics = actionable knowledge

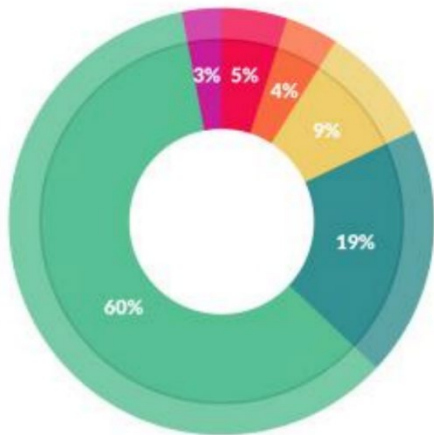
Google Trends, digital banking, Youtube analysis,
Amazon recommendations...

Data Analytics – The Reality

- Data Analytics: Expectation vs. Reality



- Forbes found that “Data preparation accounts for about 80% of the work of data scientists”



What data scientists spend the most time doing:

- building training sets 3%
- cleaning and organising data 60%
- collecting data sets 19%
- mining data for patterns 9%
- refining algorithms 4%
- other 5%.

Data Preparation

- Data preparation is crucial for getting meaningful results from data analytics.
- Insufficiently prepared data leads to 'Garbage in, Garbage out' (GIGO)

Do you trust your data?

If your business is a house, then data is its foundation.



How to check data quality?

- Meta data: “data about data”
- Understanding data availability, types, quantity, complexity, etc.
- Exploratory Data Analysis (EDA)
- Analyzing data sets to summarize their main characteristics, often with visual methods.



Collecting Data Methods

- Build private databases/data warehouses/etc.
- Scrape from websites
- From third-party platforms, e.g. Statista, World Bank, UKDS.
-

UKDS W



[Login](#) | [Contact Us](#) | [Guides](#) | [Visualisation](#) | [Usage Stats](#) | [Accessibility Statement](#) | [UK Data Service](#)

Data by Provider

Example Queries

Search Dataset Titles



Reset

All Providers

International Energy Agency

World Bank

OECD

United Nations

Human Rights Atlas

International Monetary Fund

Welcome to international aggregate data from the UK Data Service

We expect to run as normal a service as possible during this COVID-19 (Coronavirus) emergency. Please visit our [COVID-19](#) page for the latest information.

We host hundreds of economic and social datasets provided by the World Bank, OECD, International Monetary Fund, United Nations, and International Energy Agency. Datasets include World Energy Balances, World Development Indicators, World Development Indicators Statistics, Direction of Trade Statistics, World Economic Outlook, Main Economic Indicators, World Economic Outlook, Main Economic Indicators, and the Human Rights Atlas. Datasets also cover environment, education, health, and in depth regional

data provider facet on the left hand side of the screen, or [exploring and visualising data in UKDS.Stat](#) to get

Adobe Flash Player announced in July 2017, updates and support will end after December 31 2020 as detailed in [https://www.adobe.com/uk/products/flashplayer/end-of-life.html](#). The mapping and visualisation tool UKDS.Stat is based on Adobe Flash and will not be replaced due to the lack of a JavaScript based charting engine in the new version of the browser. Therefore charting and visualisations in the current version of UKDS.Stat have been disabled. If this is an issue for you please contact [support@ukdataservice.ac.uk](#) and we will advise you on alternatives.

Open data

It is our explicit long-term goal to work with data owners to identify and remove all unnecessary barriers to access.

An increasing number of our datasets are available without registration or authentication using open data licences described in our [Data Access Policy](#). These data are for use with an open licence and are not classified as personal. We are also working to gain open data certification via the Open Data Institute.

We also provide links to other [open data resources](#) that may be of interest.

Census data

International
macrodata

Qualitative data

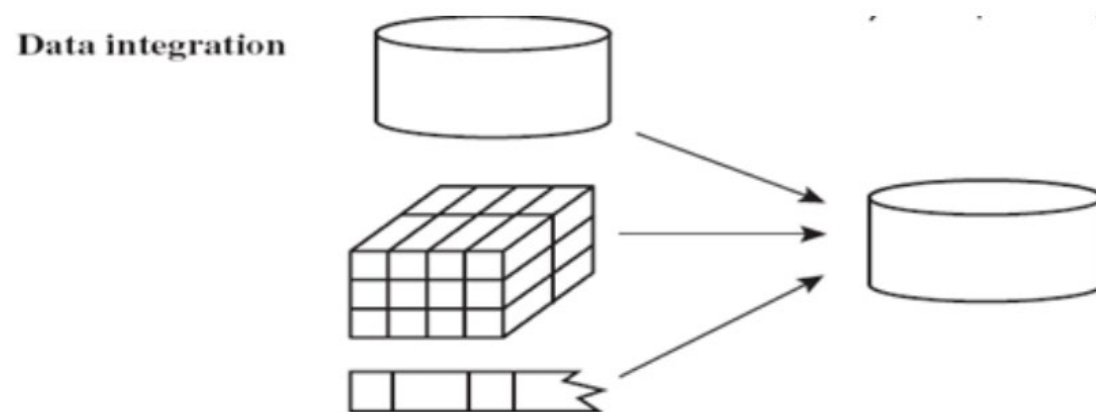
Survey data

Major Tasks in Data Pre-processing

- **Data integration** - integration of multiple databases, data cubes, or files
- **Data description**, **summarisation** and **visualisation**
- **Data cleaning** - fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- **Data reduction** - obtain reduced representation in volume but produces the same or similar analytical results
- **Data transformation** - normalisation and aggregation
- **Data discretisation** (for numerical data) and **generalisation**
- **There's no certain order of doing these steps!!**

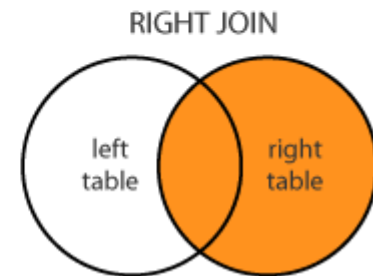
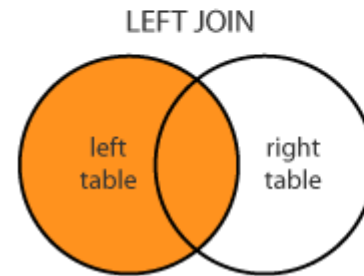
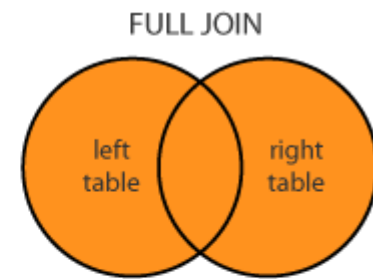
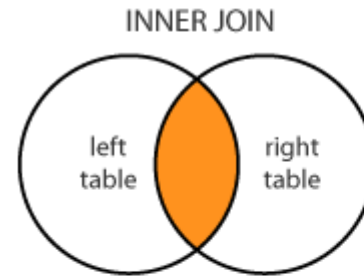
Data Integration

- Data integration: combining data from multiple sources into a unified view
 - To improve data quality
 - To enrich data with additional information
 - To allow reliable data analytics and beyond
- Integrating in-house data within data warehouse together is relatively straightforward (with common attributes and structures across schemas).



Manipulating Data - Joining

- Joining tables:
- Extract and simultaneously process data from more than one table.



Manipulating Data – Inner Join

By default, the joining query performs an *inner join*, which includes matching rows only in the results.

Employee_Payroll

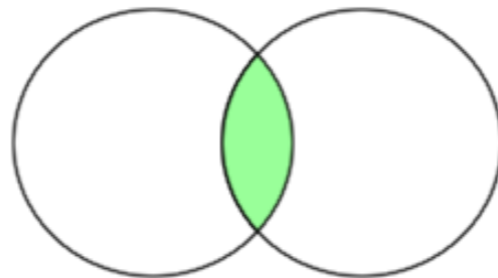
Employee_ID	Salary
120101	163040
120102	108255
120103	87975
120104	92500



Employee_Organization

Employee_ID	Department
120101	Sales Management
120102	Sales Management
120103	Engineering
120105	Administration

Employee_ID	Salary	Department
120101	163040	Sales Management
120102	108255	Sales Management
120103	87975	Engineering



Manipulating Data – Full Outer Join

- A **full outer join** includes all rows from both tables.

Employee_Payroll

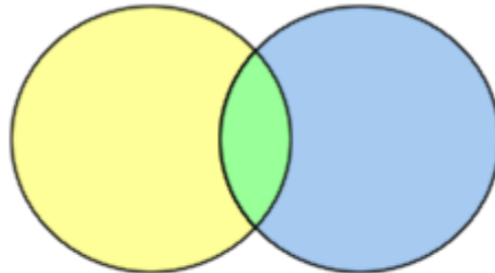
Employee_ID	Salary
120101	163040
120102	108255
120103	87975
120104	92500

Employee_Organization

Employee_ID	Department
120101	Sales Management
120102	Sales Management
120103	Engineering
120105	Administration



Employee_ID	Salary	Department
120101	163040	Sales Management
120102	108255	Sales Management
120103	87975	Engineering
120104	92500	
120105		Administration



Manipulating Data – Left Join

- A ***left join*** includes all rows from ***the left table***.

Employee_Payroll

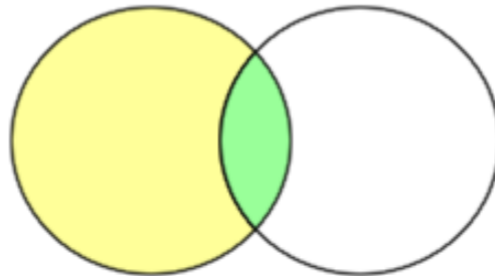
Employee_ID	Salary
120101	163040
120102	108255
120103	87975
120104	92500

Employee_Organization

Employee_ID	Department
120101	Sales Management
120102	Sales Management
120103	Engineering
120105	Administration



Employee_ID	Salary	Department
120101	163040	Sales Management
120102	108255	Sales Management
120103	87975	Engineering
120104	9250	



Difficulties of Integrating Data

- **Database heterogeneity**
 - **System Heterogeneity**: different operating system, hardware platforms
 - **Schematic or Structural Heterogeneity**: the native model or structure to store data
-
- **Data value conflicts**
 - E.g., metric vs. British units
-
- **Entity identification**
 - E.g., Bill Clinton = William Clinton



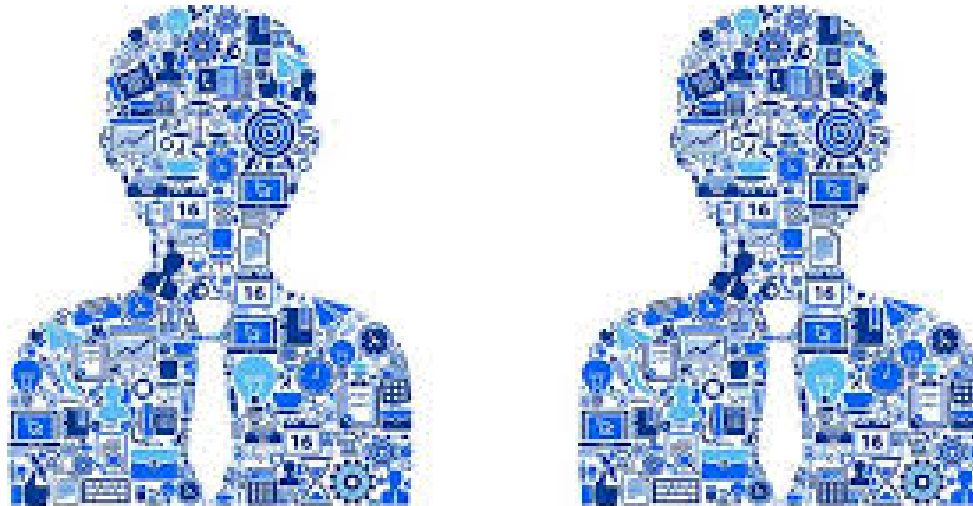
Example – Health Surveillance

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns of symptoms (in real time!)
- Various databases
- Many millions of records
- Privacy and confidentiality concerns



Linking Data

- **Data linkage** is the process of bringing together information from two different records that are believed to belong to the same entity based on matching variables.
- Challenging if errors exists in the key variables



Linking Data

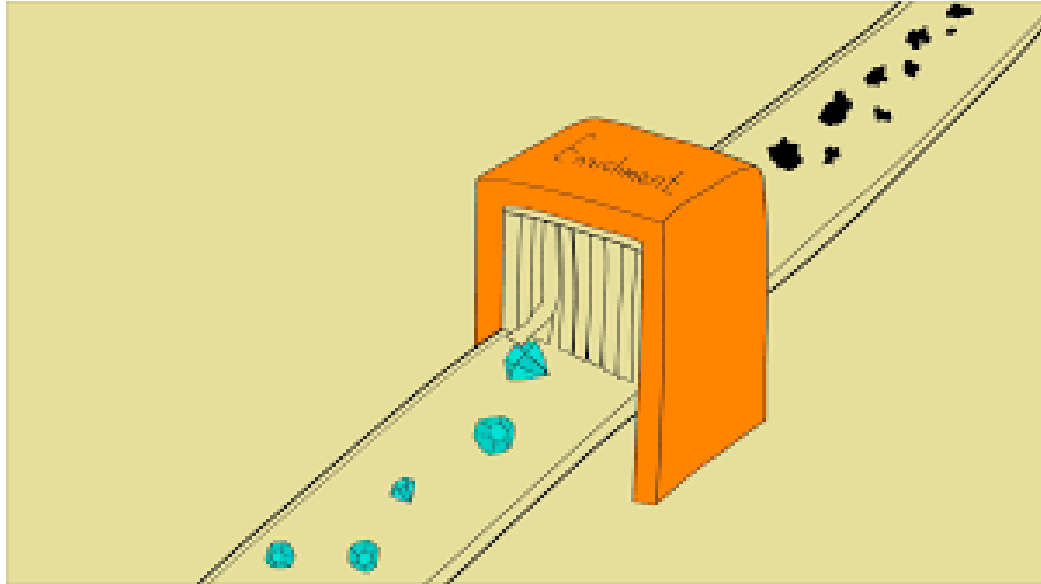
- **Deterministic linkage**
 - Agree exactly on the key, e.g. ID number
- **Probabilistic linkage**
 - Frequency analysis of data values.

Linking Data: Example

Which of these records represent the same person?

Data Set	#	SSN	Name	DOB	Sex	ZIP
Set A	1	000956723	Smith, William	1973/01/02	Male	94701
	2	000956723	Smith, William	1973/01/02	Male	94703
	3	000005555	Jones, Robert	1942/08/14	Male	94701
	4	123001234	Sue, Mary	1972/11/19	Female	94109
Set B	1	000005555	Jones, Bob	1942/08/14		
	2		Smith, Bill	1973/01/02	Male	94701

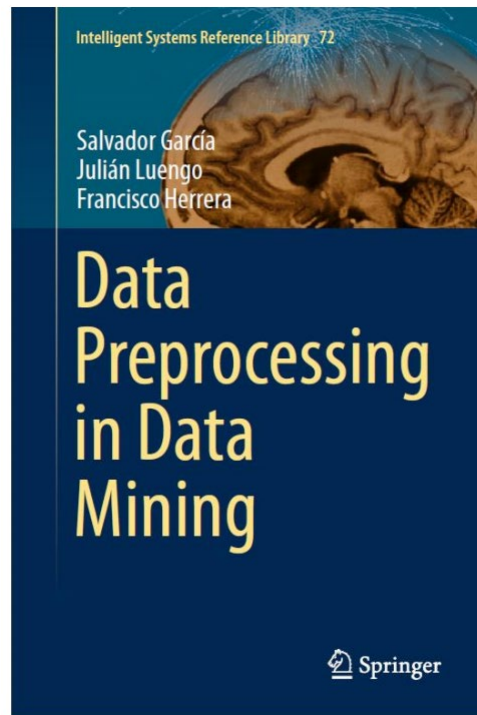
Enriching Data



- Data enrichment: the process of introducing more data
- Disparate data from other internal sources or third-party data from external sources
- Provide contextualization with additional data
- Help enrich or validate the data

Acknowledgement

- Some of the content is based on ...



García, S., Luengo, J. and Herrera, F., 2015. Data preprocessing in data mining (pp. 59-139). New York: Springer.)